Development and application of bioinformatics tools for analysis of dual RNA-seq experiments

Entwicklung und Anwendung von Bioinformatikwerkzeugen für die Analyse von dualen RNA-seq Experimenten

Doctoral thesis for a doctoral degree
at the Graduate School of Life Sciences,
Julius-Maximilians-Universität Würzburg,
Section: Infection and Immunity
submitted by

Bożena Mika-Gospodorz

from

Mikołów, Poland

Würzburg 2022

Submitted on:

Members of the Thesis Committee

Chairperson: Prof. Dr. Thomas Dandekar

Primary Supervisor: Jun.-Prof. Dr. Lars Barquist

Supervisor (Second): Jun.-Prof. Dr. Alexander Westermann

Supervisor (Third): Prof. Dr. Petra Dersch

Supervisor (Fourth): Prof. Dr. Vera Kozjak-Pavlovic

Date of Public Defence:

Date of Receipt of Certificates:

"Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more, so that we may fear less."

Maria Skłodowska-Curie

Summary

RNA-seq is an efficient technique for measuring the global abundance of transcripts present at a given time and condition for any species. As infection is a dynamic process involving at least two agents, capturing their direct relationship requires less disruptive experimental protocols than relying on the physical separation of the interacting organisms; dual RNA-seq is such a method. Extraction of total RNA of the host-pathogen system and *in silico* separation of their transcriptomes allows the simultaneous profiling of the gene expression of these organisms and the analysis of their direct interactions occurring during infection.

This work presents an application of dual RNA-seq to explore processes occurring during the infection of the human endothelial cells with *Orientia tsutsugamushi* (*Ot*) — the causative agent of scrub typhus. The study aimed to investigate the biology of this obligate intracellular pathogen and the host response to this bacterium. Therefore, Human Umbilical Vein Endothelial Cells (HUVECs) were infected with two clinical isolates of *Ot* that differ in virulence, UT176 and Karp, followed by isolation and sequencing of the RNA of the system five days post-infection. As a part of my doctoral work, I analyzed the obtained RNA-seq data by applying various bioinformatics approaches to gain a deeper insight into the bacterial RNA biology and transcriptional profile of the *Ot*-HUVEC system. Combining comparative genomics, transcriptomics, and proteomics, we investigated the transcriptional architecture of *Ot* and identified: (i) several dozen potential small RNAs and housekeeping transcripts including transfer-messenger RNA (tmRNA), ribonuclease P (RNase P), signal recognition particle (SRP), and 5S rRNA; (ii) conserved operons between two *Ot* strains; (iii) and widespread antisense transcription, that may have a role in regulation of repetitive genes that are abundant in the *Ot* genome. In addition, the comparative analysis of bacterial transcriptomes allowed us to investigate factors that drive the difference in virulence between Karp and UT176. It indicated that genes encoding virulence-associated surface and effector proteins are upregulated in Karp during infection. Meanwhile, besides a proinflammatory antiviral response activated upon infection of HUVECs with either strain, the comparison of the host response to the infection with Karp and UT176 uncovered unique immune regulatory networks altered by each strain. While Karp induced an IL33-NOS3-FAS response, UT176 upregulated the IL6-mediated proinflammatory gene network. These results also correlate with differences in disease severity in a murine infection model of scrub typhus.

The *Ot* dual RNA-seq data analysis illustrated the lack of a robust pipeline for processing dual RNA-seq data. Therefore, I developed Dualrnaseq, a workflow for quantifying dual RNA-seq data. The pipeline is developed in Nextflow and uses container technology, such that it can easily be executed on different computing platforms and facilitates reproducibility of the analysis. Within the pipeline, host and pathogen sequencing reads are processed together. After quality control and

trimming, they are mapped onto a chimeric reference comprising the bacterial and eukaryotic genomes or transcriptomes. I implemented three read quantification strategies. The first is an alignment-based mapping of reads onto the chimeric genome with STAR followed by quantification with HTSeq — a widely used tool for estimating gene expression from uniquely mapped reads. Considering the importance of multi-mapped reads that may originate from different gene isoforms in the host and repetitive elements which are highly abundant in some bacterial genomes, a fast transcriptome quantification method handling multi-mapped reads (Salmon with Selective Alignment) was applied as a second strategy. The third approach (Salmon alignment-based mode) uses a STAR-derived alignment combined with Salmon quantification. In addition, the Dualrnaseq pipeline has become a part of the nf-core repository: https://nf-co.re/dualrnaseq. Initial simulation-based benchmark analysis of host-pathogen systems was performed to investigate optimal read quantification strategies for dual RNA-seq data.

Overall, the work described in this thesis provides new insight into the biology of a genetically intractable bacterium, *Orientia tsutsugamushi*, serving as an example for other obligate pathogens. It also provides evidence for a widespread post-transcriptional regulatory role of antisense transcription in *Ot*, which has not been observed on such a scale before. Finally, the Dualrnaseq pipeline presented in the second part of this work is the first publicly available, highly reproducible, scalable, and user-friendly workflow developed for processing dual RNA-seq data.

Zusammenfassung

RNA-seq ist ein effizientes Verfahren zur Messung der globalen Häufigkeit von Transkripten, die zu einem bestimmten Zeitpunkt und unter bestimmten Bedingungen bei einer beliebigen Spezies vorhanden sind. Da die Infektion ein dynamischer Prozess ist, an dem mindestens zwei Spezies beteiligt sind, erfordert die Erfassung ihrer direkten Beziehung weniger störende Versuchsprotokolle als die physische Trennung der interagierenden Organismen; duale RNA-seq ist eine solche Methode. Die Extraktion der Gesamt-RNA des Wirt-Pathogen-Systems und die in-silico-Auftrennung ihrer Transkriptome ermöglicht die gleichzeitige Erstellung von Genexpressionsprofilen dieser Organismen und die Analyse ihrer direkten Interaktionen während der Infektion.

In dieser Arbeit wird eine Anwendung der dualen RNA-seq vorgestellt, um Prozesse zu untersuchen, die während der Infektion menschlicher Endothelzellen mit *Orientia tsutsugamushi* (*Ot*) — dem Erreger von Scrub-Typhus — auftreten. Ziel der Studie war es, die Biologie dieses obligat intrazellulären Erregers und die Reaktion des Wirts auf dieses Bakterium zu untersuchen. Zu diesem Zweck wurden humane Nabelvenenendothelzellen (HUVECs) mit zwei klinischen Isolaten von *Ot* infiziert, die sich in ihrer Virulenz unterscheiden — UT176 und Karp — gefolgt von der Isolierung und Sequenzierung der RNA des Systems fünf Tage nach der Infektion. Im Rahmen meiner Doktorarbeit analysierte ich die gewonnenen RNA-seq-Daten mit Hilfe verschiedener bioinformatischer Ansätze, um einen tieferen Einblick in die bakterielle RNA-Biologie und das Transkriptionsprofil des *Ot*-HUVEC-Systems zu gewinnen. Durch die Kombination von vergleichender Genomik, Transkriptomik und Proteomik haben wir die transkriptionelle Architektur von *Ot* untersucht und identifiziert: (i) mehrere Dutzend potenzielle kleine RNAs und Transkripte von Haushaltsgenen, darunter Transfer-Messenger-RNA (tmRNA), Ribonuklease P (RNase P), Signalerkennungspartikel (SRP) und 5S rRNA; (ii) konservierte Operons zwischen zwei *Ot*-Stämmen; (iii) und eine weit verbreitete Antisense-Transkription, die möglicherweise eine Rolle bei der Regulierung von repetitiven Genen spielt, die im *Ot*-Genom häufig vorkommen. Darüber hinaus ermöglichte uns die vergleichende Analyse der bakteriellen Transkriptome die Untersuchung der Faktoren, die für die unterschiedliche Virulenz von Karp und UT176 verantwortlich sind. Es zeigte sich, dass Gene, die virulenzassoziierte Oberflächen- und Effektorproteine kodieren, bei Karp während der Infektion hochreguliert sind. Neben einer proinflammatorischen antiviralen Reaktion, die bei der Infektion von HUVECs mit einem der beiden Stämme ausgelöst wird, wurden beim Vergleich der Wirtsreaktion auf die Infektion mit Karp und UT176 unterschiedliche immunregulatorische Netzwerke aufgedeckt, die durch jeden Stamm verändert werden. Während Karp eine IL33-NOS3-FAS-Reaktion auslöste, regte UT176 das IL6-vermittelte

proinflammatorische Gennetzwerk an. Diese Ergebnisse korrelieren auch mit Unterschieden in der Schwere der Erkrankung in einem Mausinfektionsmodell für Scrub-Typhus.

Die Analyse der dualen RNA-seq-Daten von *Ot* machte deutlich, dass es keine robuste Pipeline für die Verarbeitung dualer RNA-seq-Daten gibt. Daher habe ich Dualrnaseq entwickelt, einen Arbeitsablauf zur Quantifizierung dualer RNA-seq-Daten. Die Pipeline wurde in Nextflow implementiert und nutzt die Container-Technologie, sodass sie leicht auf verschiedenen Computerplattformen ausgeführt werden kann und die Reproduzierbarkeit der Analyse erleichtert wird. Innerhalb der Pipeline werden die Wirts- und Pathogen-Sequenzierungsdaten gemeinsam verarbeitet. Nach der Qualitätskontrolle und dem Trimmen werden sie auf eine chimäre Referenz abgebildet, die das bakterielle und eukaryotische Genom oder Transkriptom umfasst. Ich habe drei Read-Quantifizierungsstrategien eingebaut. Die erste ist eine auf dem Alignment basierende Zuordnung von Reads auf das chimäre Genom mit STAR, gefolgt von einer Quantifizierung mit HTSeq — einem weit verbreiteten Tool zur Schätzung der Genexpression aus eindeutig zugeordneten Reads. In Anbetracht der Bedeutung von mehrfach gemappten Reads, die von verschiedenen Isoformen von Genen im Wirt und repetitiven Elementen stammen können, die in einigen bakteriellen Genomen sehr häufig vorkommen, wurde als zweite Strategie eine schnelle Transkriptom-Quantifizierungsmethode angewandt, die mit mehrfach gemappten Reads arbeitet (Salmon mit Selektivem Alignment). Der dritte Ansatz (Salmon alignment-based mode) verwendet ein von STAR abgeleitetes Alignment in Kombination mit der Salmon-Quantifizierung. Darüber hinaus wurde die Dualrnaseq-Pipeline in das nf-core-Repository aufgenommen: https://nf-co.re/dualrnaseq. Erste simulationsbasierte Benchmark-Analysen von Wirt-Pathogen-Systemen wurden durchgeführt, um die optimalen Lesequantifizierungsstrategien für duale RNA-seq-Daten zu untersuchen.

Insgesamt bietet diese Arbeit neue Einblicke in die Biologie des genetisch schwer zu erschließenden Bakteriums, *Orientia tsutsugamushi*, das als Beispiel für andere obligate Krankheitserreger dient. Darüber hinaus liefert sie Beweise für eine weit verbreitete posttranskriptionelle regulatorische Rolle der Antisense-Transkription in *Ot*, die in diesem Ausmaß bisher noch nicht beobachtet worden ist. Schließlich ist die im zweiten Teil dieser Arbeit vorgestellte Dualrnaseq-Pipeline der erste öffentlich verfügbare, hochgradig reproduzierbare, skalierbare und benutzerfreundliche Workflow, der für die Verarbeitung dualer RNA-seq-Daten entwickelt wurde.

List of Figures

List of Tables

Abbreviation index

A	Adenine
ANK	Ankyrin repeat
asRNA	Antisense RNA
ATP	Adenosine triphosphate
BAM	Binary Alignment Map
BLAST	Basic local alignment search tool
bp	Base pair
C	Cytosine
CAGE	Cap analysis gene expression
cDNA	Complementary DNA
CDS	Coding DNA sequence
c.f.u.	Colony-forming units
COG	Clusters of Orthologous Groups
CPM	Counts per million
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dsRNA	Double-stranded RNA
E. coli	*Escherichia coli*
EM	Expectation-maximization
ER	Endoplasmic Reticulum
FACS	Fluorescence-activated cell sorting
FDR	False discovery rate
G	Guanine
GEO	Gene Expression Omnibus
GLM	Generalized linear model
GO	Gene ontology
h	Hour
HPC	High-Performance Computing
HUVEC	Human Umbilical Vein Endothelial Cells
ICAM	Intercellular adhesion molecule
ICE	Integrative and conjugative element
IGR	Intergenic region
IGV	Integrative Genomics Viewer
IL	Interleukin

IFN	Interferon
IPA	Ingenuity Pathway Analysis
IRF	Interferon regulatory factor
KEGG	Kyoto encyclopedia of genes and genomes
k-mer	k-long nucleotide subsequence
LFQ	Label-free quantification value
lncRNA	Long non-coding RNA
logFC	log_2 fold-change
LPS	Lipopolysaccharide
Mbp	Megabase pair
MDS	Multidimensional scaling
min	Minute
miRNA	MicroRNA
misc_RNA	Miscellaneous RNA
mitoRNA	Mitochondrial RNA
MMP	Maximal Mappable Prefix
MOI	Multiplicity of infection
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
ncRNA	Non-coding RNA
NFκB	Nuclear factor-kappa B
NGS	Next-Generation Sequencing
NOD	Nucleotide oligomerization domain
NR	Number of reads
nt	Nucleotide
NTS	Non-typhoidal *Salmonella*
ORF	Open reading frame
Ot	*Orientia tsutsugamushi*
PAMP	Pathogen-associated molecular pattern
PCR	Polymerase chain reaction
p.i.	Post-infection
PRR	Pattern recognition receptors
QC	Quality control
RAGE	Rickettsial-amplified genetic element
qPCR	Quantitative PCR
qRT-PCR	Quantitative real-time PCR
RNA	Ribonucleic acid

RNAP	RNA polymerase
RNase	Ribonuclease
RNA-seq	RNA sequencing
rRNA	Ribosomal ribonucleic acid
SA	Selective-Alignment algorithm
SAGE	Serial analysis of gene expression
SAM	Sequence Alignment Map
snoRNA	Small nucleolar RNA
SNP	single nucleotide polymorphism
snRNA	Small nuclear RNA
SPI	*Salmonella* pathogenicity island
spp.	Species
sRNA	Bacterial small RNA
SRP	Signal recognition particle
STAT	Signal transducer and activator of transcription
S. Typhimurium	*Salmonella enterica* serovar Typhimurium
T	Thymine
T1SS	Type I secretion system
T3SS	Type III secretion system
T4SS	Type IV secretion system
TLR	Toll-like receptor
tmRNA	Transfer-messenger RNA
TPM	Transcripts per million
TPR	Tetratricopeptide repeat
TraDIS	Transposon-directed insertion-site sequencing
tRNA	Transfer RNA
V	Volt
vol	Volume
WGCNA	Weighted-gene correlation analysis

Table of contents

1 Introduction

1.1 The complexity of host-pathogen interactions

Infection is a process that involves two interacting agents: a pathogen that invades and replicates either intracellularly or extracellularly in the host, which in turn responds to this action. Various pathogens ranging from eukaryotic fungi and protozoa to viruses and bacteria have evolved different routes to invade diverse hosts and disseminate across various cell types. For instance, *Salmonella*, a bacterial model of pathogenesis, has diverged into several species, subspecies, and serovars that exhibit diversity in the host range and manifestation of the infection (Hurley et al., 2014). While some serovars are host restricted, others can infect a broad range of hosts leading to acute and chronic infections (Gal-Mor et al., 2014). For example, non-typhoidal *Salmonella* (NTS) serovars cause acute self-limiting gastroenteritis in humans, cattle, swine, and poultry. In contrast, typhoid fever occurs only in humans infected with host-restricted *Salmonella* Typhi. These differences in host-adaptation and infection outcomes are shaped mainly by a divergent repertoire of genes present in the pathogens' genomes (Tanner & Kingsley, 2018). However, the clinical manifestation also depends on the host's susceptibility. NTS can cause bacteremia and systemic infection in both immunocompromised patients and very young or older individuals. Thus, investigating key features that drive the differences in the host adaptation of pathogens and the host response requires in-depth studies.

Researchers have investigated the complex host-pathogen interactions using various infection models. However, ethical concerns about experiments on humans necessitate the use of non-human models. *Salmonella* Typhimurium (*S*. Typhimurium) serves here as an example to show the challenges in finding universal systems to study infection processes. This gram-negative bacterium became a widely used model of pathogenesis as it belongs to the same family as *Escherichia coli,* Enterobacteriaceae, which allowed the use of well-established molecular tools. In addition, *S.* Typhimurium is an example of a serovar that can infect a wide range of species, allowing the use of animal models to study the infection processes (Santos et al., 2001). Murine infection models have become the gold standard in infection research as they are relatively cheap and approximate the human immune response. Also, in the case of *Salmonella*, mouse models have provided great insight into infection processes involving this pathogen (Tsolis et al., 2011). Although murine systems are easy to manipulate genetically, they are different from humans, so they cannot fully mimic human infections. For example, in well-established mice models, BALB/c and C57BL/6, *S.* Typhimurium causes symptoms similar to human typhoid fever instead of the gastroenteritis

manifested in people infected with this serovar. Further development of infection models has been motivated by finding cheaper and easier to manipulate animal systems and a desire to explore more specific questions (Bender et al., 2013; Raffatellu et al., 2008; Roux et al., 2010; Tenor et al., 2004; Van Der Sar et al., 2003). Certainly, animal models significantly impact our understanding of how pathogens disseminate throughout the host and what virulence determinants they utilize to enhance their spread. However, they also have shortcomings: they are expensive, difficult to maintain, and cannot fully mimic infections in humans.

Thus, *in-vitro* experiments are exploited to study the molecular details of pathogenesis in a particular host cell type using less complex and cheaper models than animals. The development of various techniques applicable to cell culture has remarkably impacted our understanding of the course of infection processes at the cellular level, including the presence of heterogeneous bacterial populations and varied host responses activated in different cell types (Castanheira & García-Del Portillo, 2017). Overall, both *in-vitro* and *in-vivo* systems have allowed exploration of the pathogens' lifestyle within the host, showing high complexity of interactions between these organisms that can occur at many developing stages of the infection. For instance, several routes of *Salmonella* dissemination within the host that involve various cell types have been identified (Watson & Holden, 2010). During this journey, the pathogen exploits a number of virulence factors that facilitate invasion of the host, replication, and survival within both non-phagocytic epithelial cells and phagocytic dendritic cells and macrophages (Haraga et al., 2008; Jennings et al., 2017; LaRock et al., 2015). On the other hand, host determinants also play a role in this interaction since specialized defense mechanisms against each bacterial subpopulation are activated. In short, infection is a dynamic and complex process. Pathogens adapt to the constantly changing environment to overcome the immune system and disseminate further in the host, whereas the host employs an expanded strategy to stop the intruder and ensure homeostasis. Each transition to a new infection stage involves the adaptation of interacting organisms to changing conditions by modulating the gene expression. Therefore, exploring the global transcriptional profile of both host and pathogen can help uncover new virulence factors and host pathways that play a role in the infection, ultimately leading to a deeper understanding of host-pathogen interactions during this process.

The expression of genes in a given condition can be measured using different techniques. Reverse transcriptase quantitative PCR (RT-qPCR) is an example of a method measuring the expression of small subsets of transcripts by detecting each of them using a pair of specific oligonucleotides (primers). However, an investigation of all factors that play a role in the infection process requires higher-throughput techniques. Microarray is an example of such a technology, but it is limited to a defined set of transcripts. In contrast, RNA-seq allows measuring the expression of the transcriptome as a whole, and dual RNA-seq extends it to simultaneous analysis of both host and pathogen transcriptomes (Westermann et al., 2012). The rapid expansion of high-throughput methods has allowed the development of approaches that provide a massive amount of data while enabling

transcriptome-wide analysis of different systems. In the upcoming section, I describe the techniques that have provided insight into the host-pathogen interactions at the transcriptomic level, improving our understanding of the infection process.

1.2 High-throughput methods to study host-pathogen interactions

Measuring a transcriptome, which represents the set of transcripts present in a cell and their quantity, is crucial for understanding the molecular mechanisms of an organism's adaptation to a given condition. After isolating RNA from the biological systems, their transcriptome can be explored using various high-throughput methods (Lowe et al., 2017). Microarrays were the first technology that allowed high-throughput analysis starting the era of transcriptomics in the middle of the 90s. Further advances led to RNA sequencing (RNA-seq). Many protocols have been developed to study more specific aspects, for example, capturing direct host-pathogen interactions with dual RNA-seq.

1.2.1 Microarrays

Microarrays revolutionized molecular biology by allowing transcriptome analyses on a large scale. Instead of measuring the expression of a few genes at a time, tens of thousands of genes can be profiled in a single experiment. Briefly, RNA isolated from cells of interest is either labeled directly or converted into labeled complementary DNA (cDNA) and then hybridized to an array that consists of spots of short nucleotide oligomers (probes). After washing, the fluorescent signal is measured at each spot. The intensity at each probe location captures the expression level of the gene with a complementary sequence to the probe (Jaksik et al., 2015). From different types of arrays (Bumgarner, 2013; Miller & Tang, 2009), two gained high popularity: low-density spotted arrays and high-density in situ-synthesized arrays, which differ in the fabrication method. The fundamental goal of most microarray experiments is to identify factors that differ between two biological groups. The comparison of the signal intensities between samples of different groups enables the identification of transcripts that are either upregulated, downregulated or unchanged between tested conditions. Thus, advances in the development of array-based approaches and bioinformatics data analysis methods allowed transcriptome-wide analysis of responses to biological perturbations.

Gene expression analysis using microarrays was successfully applied in various fields, including infection research involving diverse organisms. For example, the gene expression profile of a pathogen within the context of an *in-vitro* infection model was investigated for various bacteria such as *Mycobacterium tuberculosis* (Mangan et al., 1999), *S.* Typhimurium (Eriksson et al., 2003; Hautefort et al., 2008), *Chlamydia trachomatis* (Belland et al., 2003), *Chlamydophila pneumoniae* (Mäurer et al., 2007), and others (La et al., 2008; Waddell et al., 2007). Furthermore, in some studies,

microarrays were applied to explore *in-vivo*-regulated genes of pathogens during infection. For instance, genome-wide analysis of *Plasmodium falciparum* extracted from blood samples from infected patients helped identify factors necessary for the pathogen to survive in the human host (Daily et al., 2005). In addition, the whole-genome expression profile of *Vibrio cholerae* recovered from the human stools indicated genes that characterize a hypervirulent state and contribute to epidemic spread (Merrell et al., 2002). Microarrays have also enabled large-scale comparative gene expression analysis of *Borrelia burgdorferi* under different conditions, indicating critical factors that foster bacterial survival in various environments (Revel et al., 2002). Identification of host responses across multiple infection systems was another important aspect investigated using this high-throughput method (Hossain et al., 2006; Jenner & Young, 2005). Thus, examples presented here show that the application of microarrays has facilitated the initial discovery of a repertoire of genes that play a role in the infection processes.

However, many studies have focused on exploring gene expression of one agent, either host or pathogen, such that the complex interactions between these organisms could not be investigated. Dual transcriptomics, the simultaneous analysis of host and pathogen gene expression, has allowed identification of genome-wide responses of interacting organisms. This approach was established for the human fungal pathogen *Aspergillus fumigatus* co-incubated with human airway epithelial cells using human and fungal microarrays (Oosthuizen et al., 2011). Interestingly, a single array that captures both host and pathogen responses was constructed for the murine malaria model with *Plasmodium berghei*, which helped profile their gene expression in a time- and cost-effective manner (Lovegrove et al., 2006). In another example, Motley et al. (2004) analyzed gene expression simultaneously for *E. coli* and the murine granulomatous pouch infection model and identified host-pathogen cross-talk. The authors showed that the host limits iron availability in response to the infection while the pathogen adapts to this new condition by reprogramming the transcriptional machinery. Overall, the parallel examination of two interacting organisms' gene expression seems crucial to explore host-pathogen interactions fully. However, the application of microarrays is limited in this context due to some technical aspects.

One of the biggest drawbacks of microarrays is their restriction to organisms with known sequences because the probe design step relies on annotations. Moreover, the design and production of custom arrays for novel applications, e.g. non-model organisms, increase experiment costs (Westermann et al., 2012). Also, inaccurate annotations for some probes may be sources of errors in the data, leading to wrong conclusions (Zhao et al., 2014). Another challenging aspect involves cross-hybridization between target molecules and array probes with similar but not identical sequences (C. Wu, 2005), which may become more problematic in a mixed host-pathogen RNA sample (Westermann et al., 2012). In addition, both background and saturation of the signal limit the dynamic detection range, and in consequence, microarrays lack sensitivity to detect differences

in very low and high abundance transcripts. Hence new techniques that tackle some of these limitations were needed.

Examples include tiling arrays, another hybridization-based technique built of oligonucleotide probes covering the entire genomic sequence instead of selected gene fragments (Mockler et al., 2005). Their application uncovered the functional importance of untranslated regions, including small RNAs, in bacterial pathogens: *Listeria monocytogenes* (Toledo-Arana et al., 2009), *Streptococcus* (R. Kumar et al., 2010; Perez et al., 2009), and *Mycobacterium leprae* (Akama et al., 2009). In addition, tag-based methods such as Serial Analysis of Gene Expression (SAGE) and Cap Analysis of Gene Expression (CAGE) provide quantitative gene expression levels (Harbers & Carninci, 2005). In contrast to microarrays, they offer a wider dynamic range of detection without the requirement for reference genome availability. Briefly, short cDNA "tag" fragments from the 5′ end (CAGE) or 3′ end (SAGE) are concatemerized and sequenced, followed by evaluation of the frequency of each tag, which captures the expression level of the gene which the tag comes from. Although such techniques were successfully applied to various pathogens (Kronstad, 2006), their usage is limited as they are labor-intensive and expensive. Other limitations of array-based and tag-based methods relevant to studying various species are reviewed in (Westermann et al., 2012). Due to all these obstacles, the biggest revolution in transcriptome studies began with the appearance of RNA sequencing technology.

1.2.2 RNA-seq

Sequencing technologies and their constant development have revolutionized many fields, including microbiology (Loman et al., 2012; Loman & Pallen, 2015; Saliba et al., 2017). Their application helped improve annotations of model organisms like human, mouse, and *E. coli* and allowed the exploration of non-model system genomes in a cost-effective manner (Croucher & Thomson, 2010; Morozova & Marra, 2008). The RNA-seq technique combines high-throughput sequencing methods with computational approaches to determine transcripts present in a sample.

The RNA-seq protocol consists of several steps (Van den Berge et al., 2019). First, following RNA isolation from a sample, enrichment of specific RNAs can be performed to increase the sensitivity of an experiment. It can be done by either capturing poly(A) to extract polyadenylated RNAs or depleting ribosomal RNA to discard highly abundant transcripts that are not of primary interest in many studies. Second, the RNA molecules are fragmented and then reverse-transcribed to create stable double-stranded cDNAs. Next, adapter sequences are ligated to one or both ends of the double-stranded cDNA, which act as sites for primer binding in the following amplification step performed by polymerase chain reaction (PCR). The library is then sequenced on a high-throughput platform to obtain short sequences from one end (single-end reads) or both ends (paired-end reads). The final stage involves computational analysis of the read count data to quantify the gene

expression. As diverse variations of these steps can be introduced in RNA-seq protocols, researchers have optimized these strategies. They have evaluated each stage of the sequencing assay to maximize performance and examine their impact on the outcome. Optimal read length (Chhangawala et al., 2015), strand-specificity (Sarantopoulou et al., 2019), sequencing depth (Tarazona et al., 2011), and the number of biological replicates (Rapaport et al., 2013; Robles et al., 2012) are examples of aspects that need to be considered while designing the experiment. However, the main advantage of this technology is a lack of reliance on existing knowledge about the genome sequence in the design of the experiment, which was a limiting factor in microarrays.

RNA-seq offers several other advantages over microarrays (Z. Wang et al., 2009; Zhao et al., 2014). First, in the absence of annotations, it can detect all transcripts and reveal sequence variation, such as single nucleotide polymorphisms (SNPs) and other alterations in transcribed regions. Second, RNA-seq ensures single-nucleotide resolution allowing precise location of transcription boundaries, including 5' and 3' ends of transcripts, that facilitates detection of novel features of gene organization, such as alternative splicing sites in eukaryotes and operon structures in prokaryotes. In addition, RNA-seq enables qualitative and quantitative investigation of novel coding and non-coding transcripts encoded in both sense and antisense direction. It also has a wide dynamic range for detection including features with low and high expression levels and the expression estimates are highly reproducible (Marioni et al., 2008; SEQC, 2014). Moreover, RNA-seq demands less input RNA, in the range of nanograms instead of micrograms. Finally, in contrast to microarrays, which require the design of new chips for different organisms, RNA-seq is a species-independent platform facilitating application to both model and non-model organisms (Westermann et al., 2012).

In particular, various adaptations of RNA-seq protocols have allowed exploring the transcriptome and its regulatory complexity at different levels in diverse bacterial species. For instance, such a global transcriptional map has been generated for *S.* Typhimurium (Kröger et al., 2012, 2013; Srikumar et al., 2015), uncovering factors involved in the regulation of gene expression in response to infection-relevant conditions. Overall, RNA-seq has gained popularity to define the transcriptome of many pathogens, including intracellular and extracellular bacteria, viruses, fungi, and parasites (reviewed in Colgan et al., 2017; Westermann et al., 2012). Although many such studies were performed in *in-vitro* or *in-vivo* conditions using various infection models, they usually focused on the pathogen and neglected the host response. Obtaining the complete picture of infection mechanisms requires a simultaneous investigation of processes occurring in both organisms.

1.2.3 Dual RNA-seq

Dual RNA-seq allows the parallel analysis of host and pathogen transcripts from the same sample without the physical separation of the interacting organisms. In this method, their RNA is extracted as a whole and further subjected to characterization at the computational level. For the first time, such a global transcriptome profiling of a host-pathogen system was established for the opportunistic fungal pathogen *Candida albicans* co-cultured with mouse bone marrow-derived dendritic cells (BMDC) (Tierney et al., 2012). Other examples include examining interactions between parasites and their hosts at different developmental stages (Choi et al., 2014; Pittman et al., 2014). Although some RNA-Seq protocols can be applied to identify low-abundant pathogen transcripts such as viral particles in patient-derived samples (Strong et al., 2013; Wesolowska-Andersen et al., 2017; Maulding et al., 2022) enabling global assessment of both host and pathogen transcriptomes, the application of polyA-based RNA-protocols is limited as they capture only polyadenylated transcripts. As bacterial mRNAs lack polyadenylation, simultaneous analysis of the bacterial and eukaryotic transcriptomes requires techniques that overcome such differences (Westermann et al., 2012, 2017).

The problem with the high underrepresentation of intracellular bacterial reads in the total RNA of infected cells can be tackled in several ways. For instance, increased sensitivity for *S.* Typhimurium reads was obtained by separating infected host cells (HeLa epithelial cells or macrophage-like cell lines) from uninfected using FACS and depleting highly abundant uninformative transcripts — rRNAs (Westermann et al., 2016). Other methods that provide a higher bacterial to host RNA ratio were applied in the studies of *Mycobacterium* spp. These include enrichment for bacterial transcripts (Rienksma et al., 2015; Zimmermann et al., 2017), FACS sorting of infected cells combined with microbial enrichment (Pisu et al., 2020), and employment of an infection model that provides a high bacterial load (Montoya et al., 2019). A recently developed method, that performs both enrichment for bacterial mRNAs and depletion of structural RNAs (rRNAs and tRNAs) simultaneously using bacterial transcriptome-specific probes (Betin et al., 2019), was successfully applied to a dual RNA-seq library composed of transcriptomes of *Pseudomonas aeruginosa* and human bladder epithelial cells (Penaranda et al., 2021). Although different experimental set-ups may not be representative of natural infection and can affect the infection progress or RNA composition detected in the study (Hayward et al., 2020), the sequencing of total RNA of the host-pathogen system as a whole opens new opportunities.

Dual RNA-seq was applied to explore host-pathogen systems involving several obligate intracellular pathogens that cannot replicate extracellularly of the host cells. Examples include protozoa such as *Toxoplasma gondii* (Pittman et al., 2014), *Leishmania* spp. (Aoki et al., 2019; Dillon et al., 2015; Fernandes et al., 2016; Orikaza et al., 2020), and *Plasmodium* spp. (Bradwell et al., 2020; LaMonte et al., 2019; H. J. Lee et al., 2018; Yamagishi et al., 2014) as well as bacteria: *Chlamydia*

trachomatis (Hayward et al., 2020; Humphrys et al., 2013), a pathogen that causes common sexually transmitted bacterial infections in humans; *Mycobacterium leprae* (Montoya et al., 2019) causing leprosy; and *Lawsonia intracellularis* which is responsible for porcine proliferative enteropathy (Vannucci et al., 2013). In addition, dual RNA-seq experiment for the intracellular model pathogen *S.* Typhimurium during macrophage infection performed by Stapels et al. (2018) indicated that persisters are metabolically active, which is contrary to the state of *in-vitro* generated non-growing bacteria. This study shows that persisters translocate SPI-2 T3SS effectors that suppress proinflammatory responses and induce anti-inflammatory macrophage polarization, promoting pathogen survival under antibiotic treatment. Therefore, extraction of host and pathogen total RNA without physical separation of the organisms before sequencing preserves the direct interactions captured in conditions closer to the natural environment for the intracellular pathogens.

To better mimic and explore infection processes, dual RNA-seq has been applied to diverse *in-vivo* models. For instance, infection of *Yersinia pseudotuberculosis* growing extracellularly in Peyer's patches of mice facilitated the identification of transcripts playing a role in stress response and metabolic adaptation to the conditions present in the host (Nuss et al., 2017). Also, dual RNA-seq of two macrophage subpopulations isolated directly from lungs of *Mycobacterium tuberculosis*-infected mice helped to investigate the gene expression unique to the *in-vivo* environment, nutrients necessary for pathogen survival within host cells, and the molecular basis of phenotypic differences in *Mycobacterium tuberculosis* growth between macrophage subpopulations (Pisu et al., 2020). The simultaneous analysis of transcriptomes isolated from various *in-vivo* models was also performed for *Pseudomonas aeruginosa* causing lung infections in humans. The first study reveals a battle for iron between *Pseudomonas aeruginosa* and the murine host (Damron et al., 2016). Another analysis shows *in-vivo*-induced changes in bacterial stress responses and metabolism, representing an adaptive mechanism to strong conditions in cystic fibrosis lungs (Rossi et al., 2018). In the third study, the authors investigated interactions of the pathogen with the innate immunity of the zebrafish infection model (S. S. Kumar et al., 2018). The applications of dual RNA-seq to another extracellular pathogen — *Streptococcus pneumoniae* — enabled the identification of genes relevant for establishing an infection in different sites (pleura and lungs) in a murine model (Ritchie & Evans, 2019) and shared and strain/organ-specific responses to infection in the bacteria and mice (D'Mello et al., 2020). Furthermore, an SNP in the raffinose pathway transcriptional regulator gene rafR, present between blood and ear isolates, was shown to shape different disease outcomes — localized and systemic infection (Minhas et al., 2020). Another dual RNA-seq study involved a human-specific *Streptococcus pyogenes* and a nonhuman primates infection model of necrotizing myositis (Kachroo et al., 2020), where an integration of the RNA-seq and transposon-directed insertion site sequencing (TraDIS) data allowed determining new virulence factor candidates. Furthermore, analysis of host-pathogen transcriptomic data from a *Staphylococcus aureus* infection of mice has shown that the level of host resistance impacts the pathogen's expression profile, and ultimately affects

the effectiveness of the anti-virulence strategy to this antibiotic-resistant bacteria (Thänert et al., 2017). Thänert et al. (2019) also performed dual RNA-seq on biopsies of patients suffering from necrotizing soft tissue infections (NSTIs) and investigated differences between monomicrobial NSTI, caused predominantly by *Streptococcus pyogenes*, and polymicrobial NSTI. Their analysis indicated factors that mediate each of these infections and distinct host responses, that may facilitate faster diagnosis. Although only a few studies have employed dual RNA-seq to analyze patient-derived samples that can fully capture the conditions of human infections (Bradwell et al., 2020; Griesenauer et al., 2019; H. J. Lee et al., 2018; Thänert et al., 2019; Montoya et al., 2019; Pérez-Losada et al., 2015; Rossi et al., 2018; Wesolowska-Andersen et al., 2017; Yamagishi et al., 2014), the application of dual RNA-seq in various systems has shown the complexity of factors that shape host-pathogen cross talk.

Complete understanding of the interactions between host and pathogen requires the identification of both coding and non-coding RNAs (ncRNAs). In particular, high-resolution data generated with dual RNA-seq enabled identification of PinT small RNA (sRNA) that is an important virulence factor in *S.* Typhimurium, regulating genes involved in both invasion and intracellular replication states (Westermann et al., 2016). In the same study, the interspecies correlation revealed the PinT-dependent induction of the host immune response necessary to establish the intracellular replicative niche of the pathogen. In another research, Westermann et al. (2019) applied dual RNA-seq to investigate the role of ProQ, one of the major bacterial RNA-binding proteins involved in post-transcriptional regulation of many infection-relevant mRNAs, often in conjunction with sRNAs. The analysis indicated the importance of this particular protein in regulating bacterial genes involved in motility, chemotaxis, and virulence. In addition, it revealed the impact of ProQ-mediated changes of bacterial gene expression on the host response and discovered a novel sRNA repressing the mRNA of one of the magnesium ion transporters in a ProQ-dependent manner. The role of other global ncRNA regulators, Hfq and Crp, and novel virulence-relevant ncRNAs were identified for *Yersinia pseudotuberculosis* (Nuss et al., 2017). Dual RNA-seq facilitated the discovery of ncRNAs, potentially crucial in pathogenesis or housekeeping functions, also for other bacteria, including *Streptococcus pneumoniae* (Ritchie & Evans, 2019), *Mycobacterium leprae* (Montoya et al., 2019), *Pseudomonas aeruginosa* (Damron et al., 2016), *Haemophilus influenzae* (Baddal et al., 2015). In addition, several dual RNA-seq studies examined the role of the host non-coding transcripts (Kachroo et al., 2020, Hayward et al., 2020, Baddal et al., 2015, Fabozzi et al., 2018, Lisnic et al., 2013). Thus, by capturing the expression of various host and pathogen RNA species, dual RNA-seq data may provide comprehensive information about the infection process at the transcriptomic level.

A complete picture of infection may be obtained by generating time-resolved dual RNA-seq data sets that capture both host and pathogen responses over time. Investigating dynamic alterations in gene expression may help to understand the transcriptional adaptation of a pathogen to changing

conditions in the host and the host reaction to this action (Aprianto et al., 2016; Baddal et al., 2015; Dillon et al., 2015; Fabozzi et al., 2018; Farrer et al., 2018; Fernandes et al., 2016; Juranic Lisnic et al., 2013; C. H. Mavromatis et al., 2015; Westermann et al., 2016). In addition, dual RNA-seq data enables exploration of host-pathogen interactions using either correlation analysis (Bradwell et al., 2020; Stapels et al., 2018; Westermann et al., 2016) or more sophisticated network analysis methods. Weighted-gene correlation analysis (WGCNA) (Langfelder & Horvath, 2008) is one of the widely used approaches (Kachroo et al., 2020; H. J. Lee et al., 2018; Montoya et al., 2019; Wesolowska-Andersen et al., 2017) that groups correlated genes into modules and identifies the association between bacterial and host responses. A bipartite network, which comprises correlated changes in the host and pathogen gene levels upon infection, was applied to create the first host-pathogen interaction network from the dual RNA-seq data of human-derived biopsies (Griesenauer et al., 2019). Another approach that uses ordinary differential equations to model gene expression kinetics and infers inter-species gene regulatory networks from time-series dual RNA-seq data (Schulze et al., 2015), predicted interactions for *Candida albicans* and murine dendritic cells (Tierney et al., 2012). In addition, dual RNA-seq data supported with other data sets may provide a better explanation of the system behavior. For instance, Zimmermann et al. (2017) integrated metabolic and transcriptomic data and generated a genome-wide reaction pair network to identify host-pathogen interaction subnetworks of both enzymes with significantly affected transcription and metabolites with altered levels. This analysis helped them to identify the complex metabolic adaptation of *Mycobacterium tuberculosis* during the infection. Thus, integrating dual RNA-seq with other multi-omics data sets may provide a more in-depth exploration of host-pathogen interactions.

The continuous development of the sequencing techniques accompanied with reduction in costs has expanded the applicability of dual RNA-seq. This method also has gained popularity in studying other host-pathogen interactions unrelated to human health. For example, simultaneous transcriptome profiling of the honey bee, an ecologically important pollinator, and the *Lotmaria passim* parasite, one of the agents responsible for the drastic decline of bee populations, indicated how the parasite adapts to the new environment to establish and maintain infection, and how the host modifies its gene expression in response to this (Q. Liu et al., 2020). Dual RNA-seq was also employed to study economically significant pathogens that cause diseases in farmed animals (Botwright et al., 2021; L. Huang et al., 2019; Park et al., 2015; Valenzuela-Miranda & Gallardo-Escárate, 2018) or agriculturally important plants (Balsells-Llauradó et al., 2020; Kawahara et al., 2012; Z.-X. Liao et al., 2019; W. Li et al., 2019; Lundén et al., 2015; Musungu et al., 2020; Q. Wang et al., 2021; Yazawa et al., 2013). Beside parastatic interactions, several studies investigated other types of relationships. Examples include the interplay between symbionts (Mohamed et al., 2020), human-infecting pathogen and its vector (S. K. Buddenborg et al., 2017), or opportunistic pathogens that are frequently isolated from co-infections (Doing et al., 2020). Analysis of more complex inter-species interactions involves assessment of the microbiota composition and its

functional diversity in the nasal epithelial cell samples of either asthmatic or non-asthmatic patients, providing evidence for microbe-host interactions and their role in developing asthma (Pérez-Losada et al., 2015). Also, recently developed triple RNA-seq (Seelbinder et al., 2020) allowed simultaneous detection of transcriptomes of multiple organisms in co-infection settings, namely monocyte-derived dendritic cells infected with two pathogens known to affect the lungs of immunosuppressed patients, *Aspergillus fumigatus* and human cytomegalovirus. Furthermore, single-cell sequencing shifts the analysis of host-pathogen interactions to the cellular level allowing investigating heterogeneity in gene expression among individual host cells and pathogens (Avital et al., 2017; Patir et al., 2020; Steuerman et al., 2018). In sum, the adoption of the main idea behind the dual RNA-seq technique, which is the replacement of the physical isolation with *in-silico* separation of the interacting agents, sheds new light on the interactions of various biological systems.

1.3 Aims of the study

Dual RNA-seq opens up new opportunities for studying obligate intracellular pathogens and their interactions with the hosts. *Orientia tsutsugamushi* is an example of a pathogen responsible for severe morbidity accounting for a large number of deaths in humans. The work presented in chapter 2 aims at exploring the biology of this poorly characterized intracellular pathogen and the host response stimulated during the infection with this bacterium. Investigation of these aspects involves analyzing a dual RNA-seq data set generated for two clinical isolates of *Orientia tsutsugamushi* infecting human endothelial cells. Additionally, deeper characterization of the transcriptional architecture of this pathogen includes a prediction of operon structures, non-coding RNAs, and examination of antisense regulatory mechanisms. Overall, the motivation of this study was to take advantage of the dual RNA-seq protocol and, for the first time, explore the whole infection system involving this genetically intractable pathogen.

The host and pathogen transcriptional profiles in each dual RNA-seq study are obtained *in-silico* by applying tools developed for RNA-seq data analysis. Nevertheless, processing total reads from the host-pathogen system requires additional steps that simultaneously establish transcriptomes from two diverse organisms. However, the lack of a robust pipeline for dual RNA-seq data processing was a motivation to create a workflow which I called Dualrnaseq. This Nextflow-based workflow provides all essential steps of sequencing read processing for dual RNA-seq data. In chapter 3, I present this user-friendly pipeline that supports reproducibility, portability and provides three mapping and quantification strategies. The benchmark analysis of the employed methods will give recommendations ensuring accurate estimation of host and pathogen transcript expression. The Dualrnaseq workflow is publicly available, serving as a tool for processing raw dual RNA-seq data of any eukaryotic and bacterial organisms with a reference genome and annotation.

2 Dual RNA-seq of *Orientia tsutsugamushi* informs on host-pathogen interactions for this neglected intracellular human pathogen

This chapter is a modified version of the previously published article (Mika-Gospodorz et al., 2020). This work is a result of collaboration with Suparat Giengkam (Mahidol University, Bangkok, Thailand), who established a dual RNA-seq protocol for *Orientia tsutsugamushi* using human endothelial cells as an infection model. Names of other co-authors and their contribution are highlighted in section 2.2. The work was supervised by Jeanne Salje (Mahidol-Oxford Tropical Medicine Research Unit, Rutgers University, University of Oxford) and Lars Barquist (Helmholtz Institute for RNA-based Infection Research and University of Würzburg).

2.1 Introduction

Orientia tsutsugamushi is the causative agent of scrub typhus, a disease endemic to South-East Asia. In this chapter, I describe a study applying dual RNA-seq to the infection model involving this pathogen. Dual RNA-seq has provided an opportunity to investigate both the differences in gene expression between two *Orientia* strains, Karp and UT176, during the infection of human endothelial cells and the joint and strain-specific host response. Furthermore, the transcriptional architecture of the pathogen is explored through integrating RNA-seq, comparative genomics, proteomics, and machine learning. This includes identification of operon structure, non-coding RNAs, and providing evidence for wide-spread post-transcriptional antisense regulation.

2.1.1 *Orientia tsutsugamushi*: a neglected obligate intracellular bacterial pathogen

Orientia tsutsugamushi (Ot) is a Gram-negative bacterium belonging to the family Rickettsiaceae of the class Alphaproteobacteria. It is an obligate intracellular pathogen that causes a severe mite-borne disease in humans, scrub typhus. The bacterium lives in trombiculid mites called '*tsutsugamushi*' in Japanese, and it is transmitted to the next tick generation by transovarial transmission (Wongsantichon et al., 2020). The feeding larvae (chiggers) contain *Ot* in salivary glands and transfer the bacterium to humans and other hosts. Symptoms in patients usually begin 7–14 days after the inoculation and include fever, headache, rash, stupor, myalgia, and lymphadenopathy. *Ot* infections can be effectively treated using antibiotics such as doxycycline, azithromycin, chloramphenicol, and rifampicin. However, the unspecific nature of scrub typhus symptoms hinders diagnosis, and untreated, the disease can progress to cause complications, including multiorgan

failure and death. Originally the endemic area was associated with Asia-Pacific regions, but recent reports suggest a wider distribution of the disease beyond Asia, including in Africa and South America (Bonell et al., 2017). In general, morbidity and mortality from scrub typhus are higher in developing countries with limited access to healthcare, diagnostics, and treatment. The etiological agent of scrub typhus, *Orientia* is estimated to infect at least one million people per year (Taylor et al., 2015).

Despite increasing awareness in endemic regions and an expanding global presence, scrub typhus remains a neglected tropical disease. Evidence for resistance to common classes of antibiotics and lack of a preventative vaccine against *Ot* indicate an urgent need to investigate the biology of this pathogen to develop new therapeutic and preventive strategies. However, the lack of genetic tools for such pathogens has restricted research on *Ot*. Nevertheless, based on the current knowledge, we can see how interesting and unusual this bacterium is. *Ot* is characterized by a complex infection cycle that involves several hosts and cell types. Different studies, ranging from *in-vivo* and *in-vitro* infection models, have identified the main routes of *Orientia* dissemination that cover a wide range of human cells the pathogen may infect (reviewed in Díaz et al., 2018). Briefly, after inoculation into the skin, other host cells including dendritic cells, monocytes, macrophages, and endothelial cells compose the main targets. Furthermore, *Ot* can disseminate via the blood and lymphatic system to multiple organs, including the liver, spleen, skin, heart, lung, kidney, pancreas, and brain.

At the cellular level, the life cycle of all Rickettsiaceae starts from entering the host cells using a 'zipper-like' mechanism of induced endocytosis (Salje, 2021). Shortly after the entry, the bacteria escape from the endolysosomal pathway and replicate directly in the host cell cytoplasm. *Ot* replicates as a microcolony adjacent to the host nucleus, whereas *Rickettsia* spp., bacteria of another genus of the Rickettsiaceae family, undergo replication while distributed throughout the cytoplasm (Salje, 2017, 2021). Another aspect that distinguishes *Ot* from closely related bacteria of the *Rickettsia* genus is their movement mechanism. Due to the lack of flagella, both species use the host cell cytoskeleton. However, *Ot* employs microtubule-driven processes instead of the actin-based motility common in *Rickettsia* spp. To exit infected host cells, *Ot* uses a virus-like budding mechanism, whereas *Rickettsia* spp. move directly into adjacent cells or uses a host cell lysis strategy. *Ot* also has an unusual cell membrane and cell wall structure. In contrast to most Gram-negative bacteria, it possesses a minimal peptidoglycan-like cell wall, and its outer membrane is not equipped with lipopolysaccharide (LPS). All these extraordinary characteristics of the *Ot* intracellular lifestyle and its cell structure are encoded in the genome which also shows some unusual characteristics in this pathogen.

The *Orientia* genome is a single circular chromosome with a length of 1.93–2.47 Mbp (Batty et al., 2018), which is relatively large compared to other obligate intracellular bacterial genomes, including the most closely related rickettsial species (McLeod et al., 2004). Such bacteria usually have undergone genome reduction as a consequence of adaptation to the intracellular lifestyle. Also,

physical isolation from other bacteria limits the acquisition of new genes. However, *Ot* likely has acquired some sequences by horizontal gene transfer from other bacterial species (K. Nakayama et al., 2010). Importantly, the *Ot* genome expanded due to widespread amplification of repetitive elements and gene duplications. As a result, this pathogen possesses one of the most highly repetitive bacterial genomes known to date. Almost 50% of its genome consists of repeated DNA elements, including short repetitive sequences, transposable elements, and integrative and conjugative elements (ICEs) called the rickettsial-amplified genetic elements (RAGEs) (Salje, 2017). RAGEs are present in multiple partially degraded copies and encode integrases, transposases, *tra* genes typical of Type IV secretion systems, and some potential effector proteins such as ankyrin repeat-containing proteins, histidine kinases, and tetratricopeptide repeat domain-containing proteins. Due to the massive amplification of these repetitive elements and chromosomal rearrangements that have taken place in the evolutionary course of *Orientia*, there is limited synteny between the eight *Ot* genomes. The core genome of *Orientia* is relatively small and contains 657 genes grouped into 51 conserved genomic islands separated by repeat regions (Batty et al., 2018).

These unusual factors that characterize *Ot* may shape not only its unique obligate intracellular lifestyle but also the host immune response. Unfortunately, natural immunity to scrub typhus is poor and high levels of antigenic diversity among *Orientia* strains driven by frequent genetic recombination have caused difficulties in finding a universal vaccine. Nevertheless, *Ot* possesses three immunogenic surface proteins: TSA56 (also known as OmpA), TSA22, and TSA47 (also called HtrA). The first one is the most well-characterized antigen and a strongly immunogenic surface protein involved in the binding and entry of the pathogen into the host cells (B.A. Cho et al., 2010; J.-H. Lee et al., 2008). The sequence of this major outer membrane protein has served as the principal target for serological classification of *Ot* strains into several subgroups (Kelly et al., 2009). It is also known that another surface protein — TSA47 — is involved in the budding-like host cell exit of *Ot* (M.J. Kim et al., 2013), whereas the function of the TSA22 antigen is unknown. Therefore, TSA56 and TSA47 have been investigated as candidates for vaccine development. Another type of proteins present in the *Ot* outer membrane are autotransporters. There are five autotransporters (ScaA, ScaB, ScaC, ScaD, and ScaE) associated with the Type V secretion system. It has been shown that ScaC and ScaA are used by *Ot* to adhere to non-phagocytic mammalian cells (Ha et al., 2011, 2015), and ScaA has been evaluated as a vaccine candidate (Ha et al., 2016). Overall, *Ot* possesses several known outer membrane proteins that mediate the internalization into the host cell and stimulate the immune system.

Effector proteins are another type of bacterial molecule that play an essential role in pathogenesis by manipulating host cell activity. Examples include ankyrin repeat-containing proteins, one of the RAGE elements, that are present in *Ot* in great numbers compared to other microbes (Jernigan & Bordenstein, 2014). Those proteins contain 33-residue ankyrin repeats (ANKs) — the most common protein-protein interaction motif in nature — and have diverse

functions in different organisms (Al-Khodor et al., 2010). During infection of mammalian cells, *Ot* produces various ANK-containing proteins that traffic to distinct subcellular localizations, including the Endoplasmic Reticulum (ER), Golgi apparatus, nucleus, or remain in the cytosol (Beyer et al., 2017; Min et al., 2014; VieBrock et al., 2015). They facilitate survival in the host cells. Ank1 and Ank6 reduce the nuclear accumulation of both NF-κB subunit and p65, and inhibit NF-κB-dependent gene expression, an important element of the antimicrobial host defense (Evans et al., 2018). Another example is Ank4 which is used for acquisition of amino acids necessary for the intracellular growth, by activating mechanisms that normally are induced by ER stress caused by accumulation of misfolded proteins (Rodino et al., 2018). Finally, Ank9 employs separate eukaryotic-like domains to modulate multiple host cell processes. It targets the Golgi apparatus followed by binding to COPB2 to facilitate retrograde trafficking to the ER, which in turn, disturbs the Golgi apparatus and ER structures, induces ER stress, and inhibits protein secretion (Beyer et al., 2017). Moreover, Ank9 can also interact with a subunit of the SCF1 ubiquitin ligase complex, which usually catalyzes the ubiquitination of proteins destined for proteasomal degradation in eukaryotic cells (Beyer et al., 2015). Another effector interacting with ubiquitin ligase complex is ANK13 — a nucleomodulin that downregulates expression of many host genes including those involved in transcriptional control and the inflammatory response (Adcox et al., 2021). In addition to ANK-containing proteins, the RAGE also encodes proteins composed of eukaryotic-like protein-protein interaction motifs, that serve as virulence factors in bacteria (Cerveny et al., 2013). These are tetratricopeptide repeats (TPRs) that function as a scaffold for assembling multiprotein complexes. In *Ot*, the TPR-containing proteins inhibit translation in the host cells through interaction with the DDX3 RNA helicase that is involved in multiple RNA metabolic processes (Bang et al., 2016). Other known *Ot* effector proteins include otDUB deubiquitylase, an enzyme that cleaves ubiquitin chains from target proteins (Berk et al., 2020). The effector proteins are translocated into host cells through secretion systems. Examples include the conjugative type IV secretion system (T4SS) that translocates DNA and protein molecules through a channel assembly that connects two cells during the horizontal gene transfer or an infection process. The high conservation of organization of T4SS genes among *Rickettsia* and *Orientia* strains suggests an essential role of this secretion system in establishing the intracellular niche for these pathogens (K. Nakayama et al., 2008). In addition, RAGE elements contain another T4SS composed of *tra* and *trb* genes (Gillespie et al., 2015; N.H. Cho et al., 2007). However, it is unknown if the RAGE T4SS is functional. Type 1 secretion system (T1SS) is another system utilized by *Ot* to translocate effector proteins, particularly some ANK-containing proteins, directly from the cytoplasm to the extracellular environment in a one-step process (VieBrock et al., 2015). In general, secretion systems and effector proteins are essential to modulate eukaryotic cell processes to the pathogen's advantage.

In response to the pathogen action, eukaryotes have evolved various defense mechanisms. Their effectiveness, however, is dependent on multiple factors. Infection with *Ot* impacts many

aspects of the human immune reaction, but the fate of the host seems to be determined by the early inflammatory responses (Jerrells & Osterman, 1981). A repertoire of the proinflammatory molecules expressed during *Ot* infection has been identified in various studies using different cell lines (K.A. Cho et al., 2010; N. H. Cho et al., 2000, 2001; Tantibhedhyangkul et al., 2011, 2013), mouse infection models (Koh et al., 2004; Yun et al., 2005), and patient-derived samples (Chierakul et al., 2004; Chung et al., 2008; Tantibhedhyangkul et al., 2011). Production of these inflammatory mediators is induced by pattern recognition receptors (PRRs) for example, that detect pathogen-associated molecular patterns (PAMPs). Examples of PRRs recognizing *Ot* cell elements include Toll-like receptor 2 (TLR2). Although activation of this membrane receptor mediates immediate antimicrobial responses *in-vitro*, an *in-vivo* study has shown that it is one of the host factors that increase the severity of the *Ot* infection (Gharaibeh et al., 2016). The *Ot* peptidoglycan-like structure also stimulates an intracellular PRR — the Nucleotide-binding oligomerization domain-containing protein 1 (NOD1). K. A. Cho et al. (2010) showed that the activation of this receptor in endothelial cells leads to increased production of IL-32 followed by both secretion of proinflammatory cytokines IL-1β, IL-6, IL-8, and affected expression of intercellular adhesion molecule 1 ICAM-1.

However, induction of IL-10 in the early stages of macrophage cell infection suppresses the expression of proinflammatory cytokines by inhibiting the NF-κB signaling path, which results in the proliferation of the intracellular bacteria (M.J. Kim et al., 2006). Tsai et al. (2016) showed that while the infection is progressing, reaching a high number of infecting bacteria, a low level of IL-10 leads to increased production of proinflammatory cytokines through NF-κB activation. Their analysis also implies that high expression of microRNA-155, on the other hand, prevents a cytokine storm — the main cause of severe complications in scrub typhus patients. Also, upregulation of some cytokines may cause pathological changes, e.g., extensive tissue damage was attributed to high levels of IL-33 in the mouse model of scrub typhus (Shelite et al., 2016). In addition to genes encoding inflammatory cytokines and chemokines, live *Ot* cells alter the expression of genes involved in an antiviral type I interferon (IFN) pathway (Tantibhedhyangkul et al., 2011, 2013). Type I INFs are a group of cytokines mainly represented by IFNα and IFNβ, which are activated through interferon α/β receptor (IFNAR) and induce a diverse set of interferon-stimulated genes that play a role in stimulating host cell death, activating innate immune cells, promoting the development of the adaptive immune response, and activating the antiviral inflammatory gene program to interrupt the viral life cycle. However, in bacterial infection, activation of Type I INFs modulates different outcomes, beneficial either to the host or the pathogen (Boxx & Cheng, 2016). Its role in *Ot* infection is unclear, but it has been reported that sensitivity to IFN-mediated inhibition is dependent on the *Ot* strain and the genetic background of the host cells (Hanson, 1991).

Although there is still a lot to discover about the *Ot* infection, the current knowledge already shows the complexity of processes involved in host response and the pathogen adaptation. During

infection, bacteria express a complex repertoire of genes to establish their own survival and replication within the host. Those genes encode for both proteins and various non-coding transcripts that together create coordinated transcriptional and regulatory systems.

2.1.2 Transcriptional architecture and regulatory mechanisms in bacteria

Bacteria have an ability to adjust to changing conditions, e.g., those within the host, by remodeling their gene expression in response to external stimuli. Although bacteria are characterized by complex transcriptional architecture, their coordinated regulatory programs precisely tune gene expression (Mejía-Almonte et al., 2020). Bacterial genes are organized in clusters known as operons enabling their transcription as a single mRNA. These structures can be composed of either a single gene (monocistronic operons) or multiple genes (polycistronic operons) arranged under a common promoter (a DNA sequence recognized by RNA polymerase initiating transcription) and regulated by a common operator (site of binding of a repressor, which prevents transcription by blocking the attachment of RNA polymerase to the promoter). Other operon regulators include activators that increase the transcription by facilitating RNA polymerase binding to the promoter, corepressors that activate repressors, and inducers that repress or activate transcription by interacting with an activator or a repressor, respectively. Specific transcription initiation processes can also be controlled by sigma factors that enable specific binding of RNA polymerase (RNAP) to gene promoters. The process of dissociating RNA polymerase at the end of the transcription unit, called transcript termination, is mediated by either physical modification of RNA structure (intrinsic termination) or the Rho factor (Rho-dependent termination) (Ray-Soni et al., 2016). Some operons may contain multiple promoters and terminators, and each alternative promoter-terminator pair facilitates transcription of a transcription unit comprising adjacent genes (Conway et al., 2014). Overall, there has been shown higher complexity in the operon structure and transcription regulation, also involving non-coding RNAs (Bossi et al., 2012; Sedlyarova et al., 2016; Silva et al., 2019).

Non-coding RNAs (ncRNAs) are untranslated transcripts and some of them have various functions essential for cellular processes. In addition to rRNAs and tRNAs, there are non-coding transcripts that modulate the activity of proteins by mimicking secondary structures of other RNA or DNA molecules in bacteria (Gottesman & Storz, 2011). Examples include the 6S RNA that forms a double-stranded (ds)RNA hairpin with a single-stranded central bubble that mimics an open DNA promoter and interacts with the RNA polymerase regulating transcription. Other important well-conserved non-coding transcripts in bacteria include the ribonuclease P (RNase P) RNA that is the catalytic component of the ribonucleoprotein catalyzing maturation of the 5′ end of tRNA (Kazantsev & Pace, 2006); the 4.5S RNA of the signal recognition particle (SRP), a protein-RNA complex that recognizes and delivers specific proteins to their cellular destination — the plasma membrane in bacteria (Akopian et al., 2013); transfer-messenger RNA (tmRNA), which has

properties of tRNA (binds a stalled ribosome) and mRNA (encodes a short ORF ended with a termination codon), and rescues stalled ribosomes releasing the defective mRNA and incomplete mRNA polypeptide tagged for degradation (Withey & Friedman, 2003). tmRNA also contributes to virulence in some pathogens (Julio et al., 2000; Svetlanov et al., 2012).

In addition, bacterial genomes are abundant in small RNAs (sRNAs) that play an important role in the post-transcriptional regulation of gene expression (Storz et al., 2011; Wagner & Romby, 2015). The regulatory mechanism of many sRNAs is mediated by base-pairing with the target gene. For instance, trans-acting sRNAs are encoded within intergenic regions and act on genes located at distant genomic positions. Due to their partial nucleotide complementarity with their targets, the interaction between the sRNA and target mRNA is mediated by RNA chaperons, e.g. Hfq and ProQ. On the other hand, cis-acting RNAs originate from the antisense strand of protein-coding genes and have the potential to base pair with the corresponding sequence with nearly perfect nucleotide complementarity. The possible mechanisms of antisense RNA (asRNA) action include RNase-dependent degradation of the asRNA-mRNA duplexes and alteration of both transcription termination and mRNA translation (Georg & Hess, 2011; Wade & Grainger, 2014). Importantly, antisense RNAs are widespread in bacteria, but the regulatory role of only some individual asRNAs has been discovered (Georg & Hess, 2018; Thomason & Storz, 2010; Millar & Raghavan, 2021). The function of most of them is unknown. Moreover, pervasive antisense transcription has been thought to be noise (Lloréns-Rico et al., 2016), as many asRNAs might arise from spurious promoters and are likely nonfunctional. This hypothesis has been supported by the weak evolutionary conservation of antisense promoters (Raghavan et al., 2012; Shao et al., 2014). Thus, the function of many non-coding transcripts remains an open question and, together with other regulatory and transcriptional systems, is a broad area of research for different bacteria.

Technological advancements and the development of experimental methods and computational tools have led to the identification and characterization of elements of regulatory systems in various bacteria (Hör et al., 2018; Barquist & Vogel, 2015). However, while different aspects of the regulation of gene expression were studied in *Rickettsia* species, including transcription termination (Woodard & Wood, 2011) and sRNAs (Schroeder et al., 2015, 2016), such investigations have not been carried out for *Ot*.

2.2 Methods

Suparat Giengkam established the cell culture and the dual RNA-seq protocol for two *Ot* strains infecting HUVEC cells described here. After RNA extraction, generation of the cDNA libraries for Illumina sequencing was performed by Vertis Biotechnologie AG, Freising-Weihenstephan, Germany. Alexander J. Westermann validated the RNA-seq data using the Northern blot approach

and confirmed differentially expressed genes by qRT-PCR. Lars Barquist performed an initial RNA-seq read processing using the READemption pipeline and preliminary data analysis. Jantana Wongsantichon and Loo Chien Wang prepared proteomic samples. Radoslaw Sobota carried out the mass spectrometry and, together with Jantana Wongsantichon and Loo Chien Wang, processed raw spectra. I performed re-quantification of reads mapped to the *Ot* genomes using Salmon to improve quantification of the repetitive sequences and the downstream analysis of the RNA-seq host-pathogen data, including (i) prediction of ncRNAs and operon structures in bacteria; (ii) application of a machine learning approach to investigate antisense regulatory mechanisms in the *Ot* Karp strain; (iii) differential gene expression and pathway enrichment analysis for both host and pathogen. Jeanne Salje retrieved specific annotations for *Ot* ankyrin and tetrapeptide repeat proteins, and surface proteins using BLAST search. Selvakumar Subbian applied Ingenuity Pathway Analysis (IPA) to identify host gene networks activated during the *Ot*-infection, and Sandy Pernitzsch re-drew them. Piyanate Sunyakumthorn performed experiments on mice. Suthida Chuenklin did the tissue extraction and qPCR for the mouse experiments. The laboratory experiments are described briefly here; a more detailed description is available in (Mika-Gospodorz et al., 2020).

2.2.1 Growth of *Ot* and isolation of RNA

Two clinical isolate strains of *Ot*, Karp and UT176, were propagated in a confluent monolayer of Human Umbilical Vein Endothelial Cells (HUVEC; Gibco C0035C) for 5 days at MOI 100:1. Cells were cultured using Media200 (ThermoFisher; M200-500) supplemented with LVES media (ThermoFisher) at 35 °C and 5% CO_2. For RNA isolation, bacteria from frozen stocks were first pregrown in HUVEC cells in a T25 culture flask. After 5 days, they were harvested and inoculated onto a fresh lawn of host cells, with each condition filling 2 × 6-well plates for the second round of growth. The MOI of infections for RNA isolation was established by measuring the number of bacteria in the pregrowth supernatant one day before harvesting the bacteria. Using qPCR of the inoculum sample, it was determined that the MOI for infection was 35:1 and 32:1 bacteria:host for Karp and UT176, respectively, although it is expected that a smaller number of bacteria entered into host cells. The washing step was performed with fresh media 3 h p.i. The host cells (infected and uninfected) were harvested by incubation on ice. Subsequently, they were resuspended in RNAprotect Bacteria Reagent (Qiagen), followed by storage at −80 °C. RNA was extracted using the Qiagen RNeasy Plus kit.

The bacterial growth curve was prepared for the pregrowth of the bacteria in host cells following the above protocol and then grown in 24-well plates. The obtained MIO was 8:1 and 25:1 for UT176 and Karp, respectively. At each time point, bacterial DNA was isolated using alkaline lysis extraction, and qPCR was performed (Giengkam et al., 2015).

2.2.2 RNA processing and sequencing

Assessment of the integrity of the DNase-treated RNA samples was performed in a Bioanalyzer (Agilent), with RNA integrity number values ≥ 8.0 for all samples. Ribo-Zero Gold (epidemiology) kit (Illumina) was used to remove rRNAs following the manufacturer's instructions. rRNA-depleted RNA was precipitated in ethanol for 3 h at −20 °C.

For cDNA library preparation, ultrasound sonication shearing of rRNA-free RNA samples (four 30-s pulses at 4 °C) was performed to generate on average 200- to 400-nt fragments. After removing fragments of 20 nt (Agencourt RNAClean XP kit, Beckman Coulter Genomics), the Illumina TruSeq adapter was ligated to the 3′ ends of the remaining fragments, and further served as a primer in the first-strand cDNA synthesis step performed using M-MLV reverse transcriptase (NEB). Following the purification of the first-strand cDNA, the 5′ Illumina TruSeq sequencing adapter was ligated to the 3′ end of the antisense cDNA, and PCR amplification of resulting cDNAs was performed to obtain 10-20 ng/µl using a high-fidelity DNA polymerase. 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 size-fractionation of the cDNA pool in the 200-600 bp size range was performed 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). 75-bp long single-end reads were generated on the NextSeq 500 platform (Illumina) at Vertis Biotechnologie AG, Freising-Weihenstephan, Germany. The raw sequencing data are available in GEO with accession number GSE139498.

2.2.3 Northern blots

15 µg of total RNA were loaded per lane and separated on 6% (vol/vol) polyacrylamide–7 M urea gels, electro-blotted (1 h, 50 V, 4 °C) onto Hybond XL membranes (Amersham) in a tank blotter (Peqlab), cross-linked with UV light, and hybridized with gene-specific 32P-end-labeled DNA oligonucleotides (Table S6) in Hybri-Quick buffer (Carl Roth AG) at 42 °C. Typhoon FLA 7000 phosphorimager (GE Healthcare) was used for the readout.

2.2.4 qRT-PCR

qRT-PCR of selected host and pathogen expressed genes was performed with the Power SYBR Green RNA-to-CT1-Step kit (Applied Biosystems) following 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 $2(-\Delta\Delta Ct)$ method (Livak

& Schmittgen, 2001). In addition, the specificity of primer sequences (Table S6) has been confirmed using Primer-BLAST (NCBI).

2.2.5 RNA-seq read processing and quantification

The raw reads were processed as described in a previous dual RNA-seq study (Westermann et al., 2016). Removal of adapter sequences and low-quality read ends was performed with Cutadapt using a minimum read quality of 20 (M. Martin, 2011). The READemption pipeline (v0.4.363) (Förstner et al., 2014) and segemehl (Otto et al., 2014) with the lack remapper (v0.2.064) (Hoffmann et al., 2014) were used to map reads against the human (GRCh38) and *Ot* (UT176 accession: LS398547.1; Karp accession: LS398548.1) reference genomes, followed by removal of cross-mapped reads. For the host, only uniquely mapped reads were quantified.

For more accurate quantification of repetitive sequences, reads mapped to the *Ot* genomes were re-mapped and quantified with Salmon (v0.9.1) (Patro et al., 2017) in quasi-mapping mode (*-type quasi*). The *Ot* transcriptome references were created using an in-house script that extracts gene sequences from the genome fasta files based on the gene coordinates from the GenBank annotation files. The quantification was performed by setting stranded forward library type (*-lSF*) and removing incompatible mappings (*–incompatPrior 0.0*). Identical gene repeats detected by Salmon were collected in 218 and 127 groups for Karp and UT176, respectively (Appendix 1), and a single gene from each group was retained for the quantification. Antisense reads were quantified following the above steps using reverse complemented transcript sequences.

2.2.6 Gene annotation

The gene names, gene products, and amino acid sequences were obtained from the GenBank annotation. In addition, gene names, COG functional categories (Galperin et al., 2015), KEGG pathways (Kanehisa et al., 2017) and GO terms were predicted using eggNOG-mapper (Huerta-Cepas et al., 2016). Surface antigen encoding proteins were manually identified using BLAST. Also, the KEGGREST (v1.18.1) (Tenenbaum, D., 2019) and GO.db (v3.5.0) (Carlson M, 2019) R packages were used to retrieve KEGG and GO terms, respectively. Additionally, specific annotations for ANK- and TPR-containing proteins were obtained through manual comparison using BLAST search to annotations in the *Ot* Ikeda strain [GenBank assembly number GCA_000010205.1]. The key *Ot* surface proteins TSA56, TSA47, TSA22, ScaA, ScaC, ScaD, and ScaE were also manually annotated using BLAST.

2.2.7 Gene expression analysis

For analysis of *Ot* gene expression, genes were classified as expressed or highly expressed if the mean TPM value from three replicates of each strain was greater than 10 or 50, respectively.

A hypergeometric test using the *phyper* function of the stats R package with a Benjamini-Hochberg correction for multiple testing (*p.adjust* function) was applied to test for enrichment of KEGG pathways in either expressed or highly expressed bacterial genes. Only gene sets that consist of more than 10 genes were considered in the analysis.

2.2.8 Non-coding RNA prediction

Non-coding RNAs were annotated using various tools: Rockhopper (v2.03) (McClure et al., 2013), ANNOgesic (v0.7.17) (Yu et al., 2018), and Infernal (v1.1.2) (Nawrocki & Eddy, 2013) which searches sequences against the Rfam database (Kalvari et al., 2018). Due to inconsistent predictions of intergenic sRNAs, they were further manually curated by visual comparison of the predicted coordinates with the read coverage in the Integrative Genomics Viewer (IGV) (v2.5.2) (J. T. Robinson et al., 2011). Core housekeeping ncRNAs, including tmRNA, RNase P, SRP, and 5S rRNA were predicted with Infernal. The quantification of the bacterial transcriptomes complemented with predicted ncRNAs was performed with Salmon.

2.2.9 Genomic alignment

The genomic comparisons in Figure 2.5A and C were generated in Easyfig (Sullivan et al., 2011). *Escherichia coli* K-12 MG1655 (accession number U00096) and *Salmonella enterica* serovar Typhimurium SL1344 (accession number FQ312003) were used as comparators for synteny analysis.

2.2.10 Orthology and conserved operon prediction

The orthologous genes between the two *Orientia* strains were predicted using Poff (included in ProteinOrtho, v5.16) (Lechner et al., 2014) with default parameters in synteny mode. An in-house script was used to identify conserved operons by combining the information on orthologous genes with the operon structures predicted in each strain by Rockhopper (Tjaden, 2015). Visual evaluation of read coverage in the IGV genome browser (v2.5.2) (J. T. Robinson et al., 2011) allowed manually extending some operons by adding genes or merging two operons into one. Operons with missing genes in one strain were classified as partially conserved operons.

2.2.11 Differential gene expression and pathway analysis

The bacterial differential gene expression analysis was performed between orthologous genes identified for Karp and UT176 by Poff (Lechner et al., 2014). Genes predicted as an orthologous group (more than two genes) and duplicates (transcripts with perfectly identical sequences identified by Salmon in either strain) were removed from the analysis. The differential gene expression analysis, for both host and pathogen was performed with the edgeR package (v3.20.9)

(M. D. Robinson et al., 2010) using robust quasi-likelihood estimation (Y. Chen et al., 2016), including genes with CPM (counts per million) > 10 (for *Orientia*) or CPM > 1 (for host) in at least three replicates. To investigate biological processes that differ between Karp and UT176, the *fry* test from the edgeR package was used to perform gene set enrichment analysis of KEGG and GO terms that contain at least four expressed genes. Three additional gene sets were created manually: RAGE pathway consisting of genes that encode conjugal transfer proteins, transposases, integrases, and hypothetical proteins; surface proteins and adhesins including TSA22, TSA47, TSA56, ScaA, and ScaC genes; and secreted effector proteins consisting of both ANK- and TPR-containing proteins.

2.2.12 Proteomic sample preparation

Bacteria were propagated in the HUVEC cell line at an MOI 70:1 and 159:1 for Karp and UT176, respectively. After 5 days p.i., cells were harvested, and *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 and vacuum dried.

2.2.13 Mass spectrometry

The dried samples were resuspended in 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 (ThermoScientific) in single-column setup using 0.1% formic acid in water and 0.1% formic acid in acetonitrile as mobile phases using reversed-phase C18 column (EASY-Spray LC Column, 75 µm inner diameter × 50 cm, 2 µm particle size) (ThermoScientific). The samples were injected and separated on an 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.

To obtain the data, an Orbitrap Fusion mass spectrometer in data-dependent mode was used. Samples were ionized with 2.5 kV and 300 °C at the nanospray source and positively-charged precursor MS1 signals were detected by setting an Orbitrap analyzer to the resolution of 60,000, automatic gain control (AGC) target of 400,000 ions, and maximum injection time (IT) of 50 ms. Precursors with charges 2–7 and 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. To avoid re-sampling of

high abundance peptides, precursors used for MS2 scans were excluded for 90 s. The MS1–MS2 cycles were repeated every 3 s until completion of the run.

Proteins were identified with MaxQuant (v1.5.5.1). Raw mass spectra were searched against *Ot* primary protein sequences derived from complete genome data for the Karp and UT176 strains. Human whole proteome sequences obtained from Uniprot were 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 used as the digestion enzyme and up to three missed cleavage sites were allowed. 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. The proteomics data are available in jPOSTrepo with accession number PXD017956.

2.2.14 Proteomic data analysis

Each protein was classified as either detected or undetected; if at least two peptides were identified in at least two biological replicates, a protein was defined as detected and represented by the mean label-free quantification values (LFQs) across the three replicates; otherwise, the protein was undetected, and the LFQ value was set to zero. 97 proteins assigned to one of the 23 protein groups that could not be resolved were removed from the analysis.

2.2.15 Transcript classification

Within the analysis, 318 genes were classified as detected in proteomics, and 1608 genes were classified as undetected. Further, each gene was represented by the sense expression defined by the mean TPM value across replicates and the antisense/sense ratio, which was calculated as the ratio of mean read counts assigned to the antisense and sense strand of coding annotations. Additionally, ncRNAs and duplicates identified by Salmon (Appendix 1) were removed from the analysis. The Spearman's correlation between TPMs and LFQs for genes with detected proteins indicates a weak positive correlation (the coefficient equal to 0.33); however, Pearson's correlation coefficient equal to 0.04 indicates a lack of a linear association. Transcripts with sense expression >10 TPMs (previously defined as the expression threshold) were selected for further analysis.

2.2.16 Logistic regression models

A machine learning approach was employed to test whether antisense-sense ratios are predictive of protein expression. Logistic regression was applied to model the probability of a binary response, i.e. whether a protein is expressed or not. Three models were generated. The first model makes predictions of the protein expression based solely on TPM values of the sense strand:

$$P = (1 + \exp(-(\beta_0 + \beta_1 TPM_{sense})))^{-1};$$

model 2 makes predictions solely on the antisense-sense read count ratio:

$$P = (1 + \exp(-(\beta_0 + \beta_1 \frac{NR_{antisense}}{NR_{sense}})))^{-1},$$

where NR represents the number of reads; model 3 uses both sense transcription and the antisense-sense ratio to make predictions:

$$P = (1 + \exp(-(\beta_0 + \beta_1 TPM_{sense} + \beta_2 \frac{NR_{antisense}}{NR_{sense}})))^{-1}.$$

As the data was highly imbalanced, 316 transcripts with detected proteins and 915 with no detected peptide products, a balanced data set for the model training step was created by applying a downsampling procedure using the *downSample()* function from the caret R package (Kuhn & Others, 2008). Next, the *glm()* function with a logit link function from the caret package was used to fit the regression models. To assess the predictive power of the models, initially they were trained on the reduced data set containing 632 genes, and then tested on the complete data set. More rigorous evaluation of this result was performed by applying 500-fold cross-validation. For each fold, the data was split randomly into training and testing data sets, which included 1171 and 60 genes, respectively. Each time the new training data set was reduced to 602 genes with the downsampling procedure and then used to estimate the model parameters. The evaluation of model performance was performed on the small testing data set and used a variety of measures. This included precision which is defined as the ratio of correctly predicted elements from the class of detected proteins to all those classified as protein detected. Furthermore, we used the recall, also known as sensitivity, which measures the ability of the model to predict the transcripts accurately from the class of detected proteins. Since the testing data set contained different numbers of elements from each class in each cross-validation fold, metrics robust to the imbalanced data sets were also employed for model evaluation, e.g. balanced accuracy that presents an average of the proportion of correct predictions of each class individually, ROC curves illustrating the ability of the classifier to distinguish between two classes at various discrimination threshold (pROC, v1.14.0) (Robin et al., 2011), and the area under the ROC curve (AUC).

2.2.17 Host networks/pathway analysis

Host pathways affected by both Karp and/or UT176 were investigated by analyzing genes differentially expressed at an adjusted *p*-value of <0.05 using Ingenuity Pathway Analysis (IPA) software (Krämer et al., 2014; Subbian et al., 2013). Interesting pathways were selected based on the enrichment *p*-values and activation z-scores, and served as the basis for Figure 2.9, Figure 2.10, Figure S3, Figure S4, Figure S5, Figure S6.

2.2.18 Mice and ethics statement

The research performed on mice was performed under a protocol approved by the Armed Forces Research Institute of Medical Sciences (AFRIMS) Animal Care and Use Committee and carried out in accordance with Thai law, the Animal Welfare Act, and all applicable U.S. Department of Agriculture, Office of Laboratory Animal Welfare and U.S. Department of Defense guidelines. The number of the protocol is PN16-05. Female C57BL/6NJcl mice at the age of 6–8 weeks were used in the experiments. Two groups of mice (8 mice per group) were intravenously injected in the tail vein with 1.25 × 106 genome copies of either Karp or UT176. The *Ot* inoculum was derived from *Ot*-infected L929 cells. Mice were monitored for disease symptoms for 12 days and euthanized with CO_2 inhalation 12 days post-inoculation. Blood and tissue samples (lungs, liver, spleen, and kidneys) were collected for bacteria quantification (qPCR) and histopathology. Tissues were stained by hematoxylin and eosin, followed by histopathological scoring of the extent of tissue damage.

2.3 Results

2.3.1 Dual RNA-seq of *Orientia tsutsugamushi* infecting endothelial cells

In this study, an *in-vitro* dual RNA-seq protocol was established for two clinical isolates of *Orientia tsutsugamushi*, Karp (Enatsu et al., 1999) and UT176 (Paris et al., 2009), both infecting Human umbilical vein endothelial cells (HUVEC). The HUVEC cells were selected as host cells due to their similarity to cell types involved in early and advanced infections. They were infected with bacteria at different MOI, 32:1 for UT176 and 35:1 for Karp, and were grown for five days (Figure 2.1A). Uninfected HUVEC cells were grown in parallel. After five days, when host cells were heavily loaded with bacteria (Figure 2.1B), total RNA was isolated, depleted for rRNA, converted to cDNA,

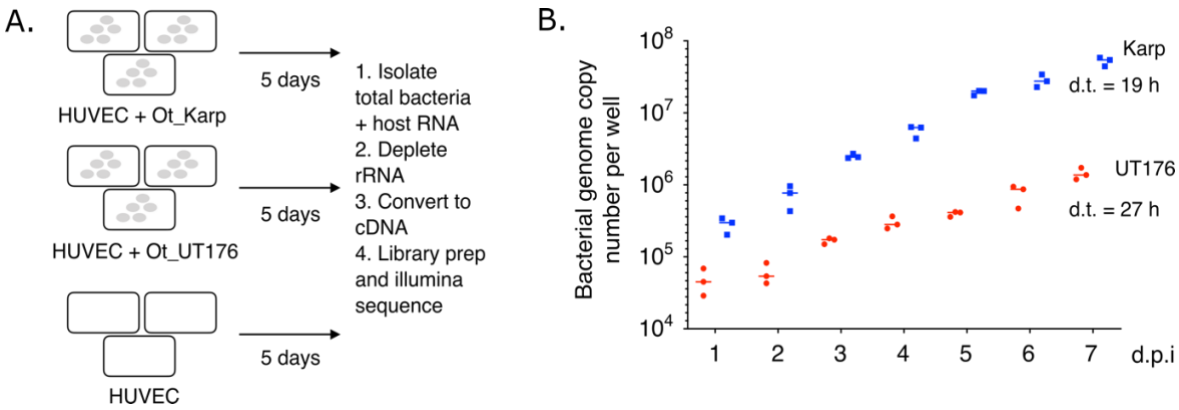

Figure 2.1 Establishment of the dual RNA-seq protocol for Ot. A) Overview of the dual RNA-seq protocol applied in this study; HUVEC - human umbilical vein endothelial cell. B) Growth curve showing replication of Ot in cultured HUVEC cells. Mean and SD from three independent replicates are shown. This figure from A to B was created by Jeanne Salje.

and sequenced to ~35 million reads per library using Illumina technology.

The sequenced reads were initially mapped to the complete genomes of Karp, UT176 (Batty et al., 2018), and, in parallel, to the human genome, following the bioinformatic protocol established in a previous dual RNA-seq study (Westermann et al., 2016). The READemption pipeline (Förstner et al., 2014) with segemehl mapper (Otto et al., 2014) and Lack (Hoffmann et al., 2014) supporting splice junction site recognition, were used. Reads that mapped equally well to the bacterial and host genomes, were defined as cross-mapped and removed.

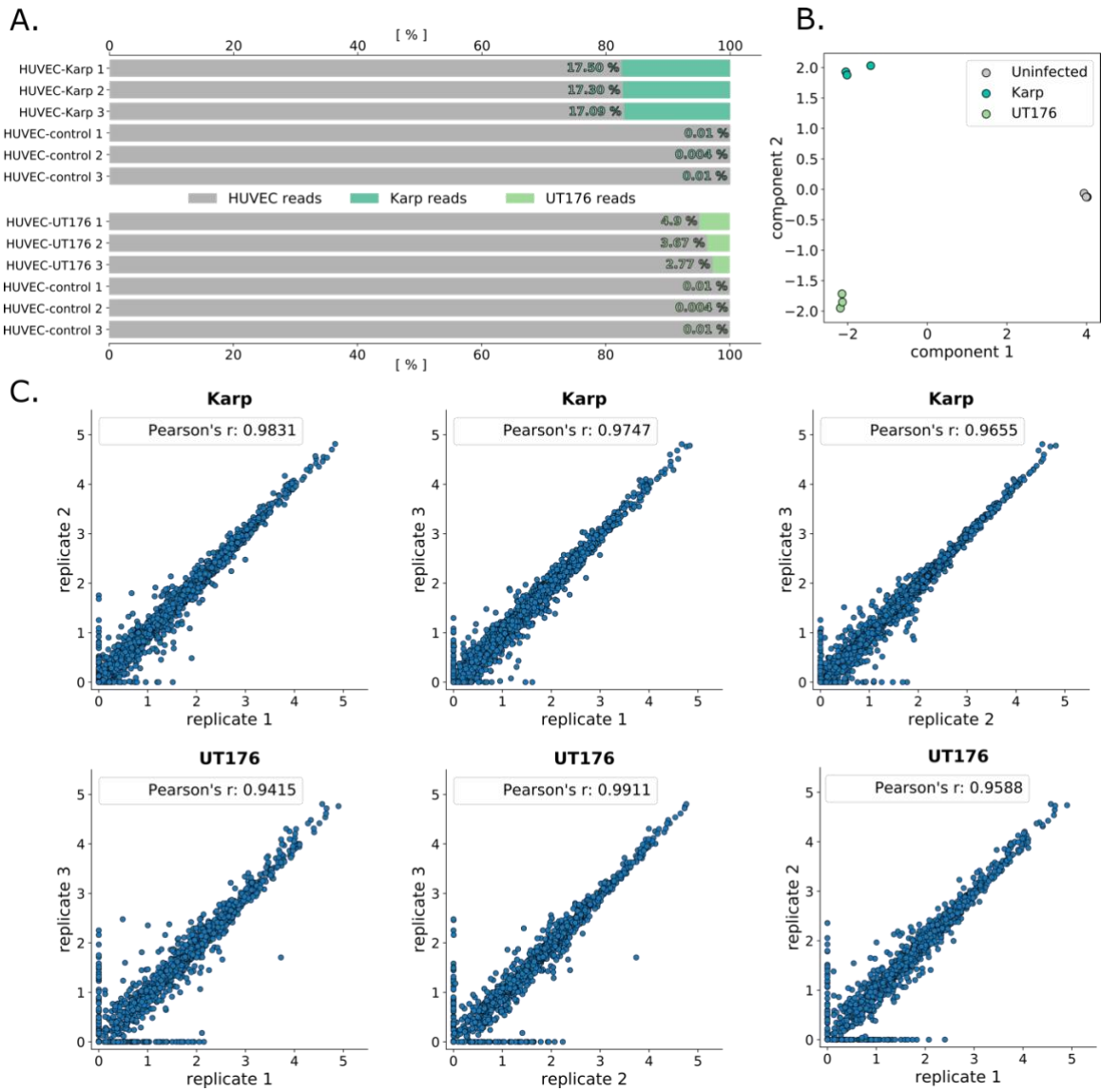

Figure 2.2 Quality control and overview of mapping and quantification results. A) RNA mapping statistics showing the fraction of host and Ot RNA for each individual sample. B) MDS plot of the host data shows a grouping of the samples by the condition (uninfected host cells or HUVECs infected with either Karp or UT176), indicating good quality of the data. C) Summary of data reproducibility indicating comparability of the transcript abundances between the samples. Comparison of the replicate log_{10}(TPM) values. Pearson correlation coefficient was calculated for untransformed TPM values.

In total, 17.1–17.5% Karp reads and 2.8–4.9% UT176 reads were identified in the infected HUVEC cells, respectively (Figure 2.2A). The difference in the number of detected bacterial reads

may reflect differences in both host cell entry efficiency and growth rate between these two *Ot* strains, which have doubling times of 19 and 27 h in HUVEC, respectively (Figure 2.1B). For the host, only uniquely mapped reads were further processed for quantification. The good quality of the data and a lack of batch effects were confirmed by a multidimensional scaling (MDS) plot (Figure 2.2B), indicating the condition as the greatest source of variation in the normalized host data.

As the *Orientia* genome is repeat-rich, reads mapped to the *Ot* genomes were re-mapped to bacterial transcriptomes and quantified using Salmon in quasi-mapping mode (Patro et al., 2017). A model-based approach employed in Salmon allows estimation of transcript abundances using both uniquely and multi-mapped reads. Alignment-free methods such as the one implemented in Salmon are discussed in more detail in Section 3.1.1 and 3.1.6 in the next chapter. Additionally, comparing the bacterial transcript abundances between the replicates shows a strong linear correlation (Figure 2.2C) indicating good data quality.

The RNA mapping statistics (Figure 2.3; Appendix 1) show that all major bacterial and eukaryotic transcript classes were detected in this study. Coding sequences (CDSs) are one of most the abundant classes identified in both the HUVEC (54% of all host-mapped reads across all samples) and the bacterial data (35% of the Karp- and 38% of the UT176-mapped reads). Using the relative abundance of transcripts in TPM (transcripts per million) units and thresholds for expressed (TPM > 10) and highly expressed genes (TPM > 50), these two sets of genes were explored for each *Ot* strain. While 1422 expressed and 856 highly expressed genes were identified in Karp, UT176 harbors 1244 and 766 expressed and highly expressed genes during the infection, respectively (Appendix 1). The pathway analysis identified KEGG gene sets involved in gene expression regulation, virulence, and metabolism as statistically significant pathways enriched in expressed genes in both Karp (Table S1, Table S2) and UT176 (Table S3, Table S4). Further exploration of COG functional categories (Galperin et al., 2015) confirmed the above-identified functions of strongly activated genes during the infection (Figure S1). The higher number of highly expressed genes come from COG categories such as *Energy production and conversion* (C), *Translation* (J), *Replication and repair* (L), and *Intracellular trafficking and secretion* (U). These results indicate an active lifestyle of the bacteria within the HUVEC cells. In particular, secretion of effectors and host-dependent nutrient acquisition and metabolism are the main characteristics of the obligate intracellular pathogens. Genes encoding proteins with such functionalities usually are conserved within species. Out of the previously identified core *Ot* genes (Batty et al., 2018), 599 of them are expressed, and 369 are highly expressed in this analysis (Appendix 1).

Besides coding transcripts, dual RNA-seq also detected various ncRNA classes from both the host and the bacteria (Figure 2.3; Appendix 1). The low number of detected human rRNA reads indicates efficient depletion of ribosomal transcripts in the host transcriptome. On the other hand, reads mapping to rRNA were 32% and 44% of the total in Karp and UT176, respectively. Since most of these bacterial ribosomal reads were derived from the 5S rRNA, this indicates the divergence of

5S rRNA sequences between *Ot* and bacterial model organisms used for the optimization of the Ribo-Zero approach (https://emea.illumina.com/products/selection-tools/ribo-zero-kit-species-compatibility.html?langsel=/de/). Analysis of dual RNA-seq data has also enabled the identification of other novel bacterial ncRNAs.

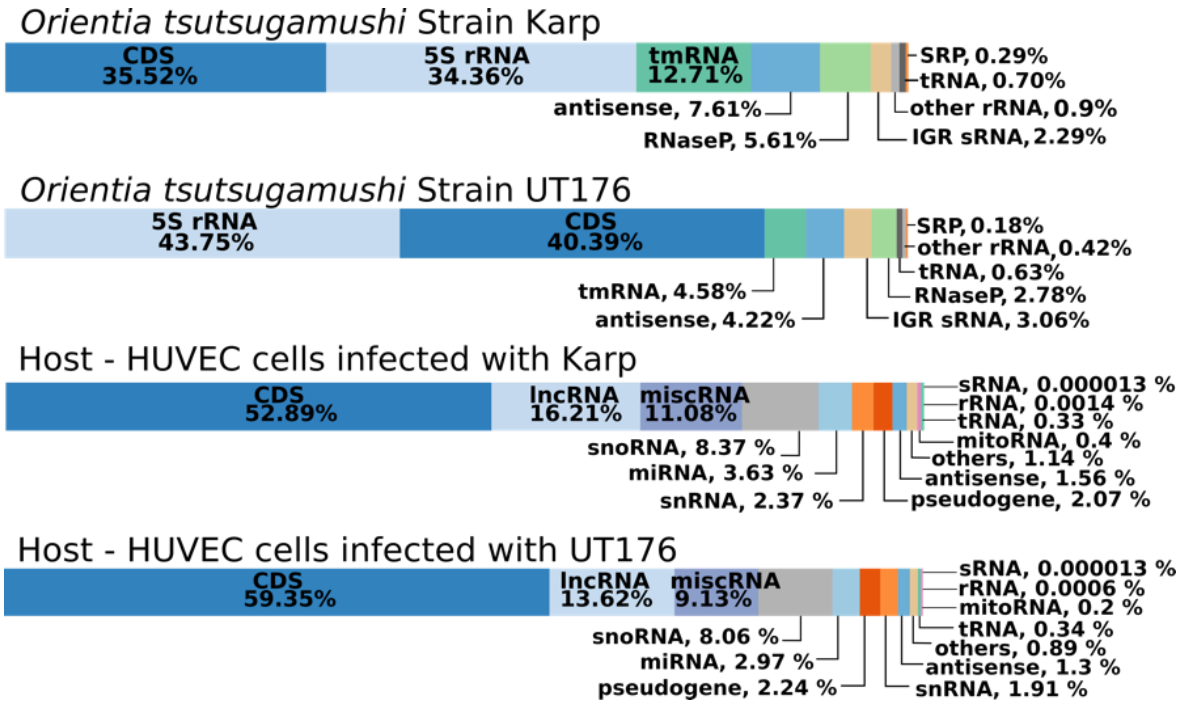

Figure 2.3 Percentage of RNA-seq reads assigned to different RNA classes in Karp, UT176, and HUVEC; First replicates are shown; CDS – coding sequence, IGR – intergenic region, antisense – reads that mapped antisense to CDS.

2.3.2 *Ot* ncRNAs and evidence for tmRNA processing

Bacterial genomes encode various non-coding transcripts, and many of them play essential roles for bacterial survival in different environments. From the most conserved housekeeping ncRNAs, RNA components of RNase P, SRP, and tmRNA were identified for each *Ot* strain (Figure 2.3; see Appendix 1 for genome coordinates). Furthermore, the detected housekeeping ncRNAs in Karp were confirmed by Northern blot (Figure 2.4A). The M1 RNA of the RNase P and 4.5S RNA of the SRP ran at their expected lengths of ~385 and ~100 nt, respectively. Moreover, there is evidence for a precursor-M1, since a second stronger band indicates a length of ~450 nt. Interestingly, the identified high level of tmRNA expression, contributing between 4.6-13% of total bacterial reads (Figure 2.3; Appendix 1) may suggest an important role in *Ot* survival in mammalian cells. tmRNA has undergone a circular permutation in some clades of bacteria (Mao et al., 2009), including the Alphaproteobacteria (Keiler et al., 2000) producing a two-piece form. A tRNA-like (acceptor) domain is encoded upstream of a mRNA-like coding domain, and the precursor transcript is processed into separate, base-pairing acceptor and coding RNA chains (Gaudin et al., 2002; Sharkady, 2004) (Figure 2.4B). Three *Ot* tmRNA forms were detected here using Northern blot:

(i) a long precursor tmRNA (372 nt); (ii) a 5′ fragment of ~80 nt, the acceptor domain; (iii) and the 3′ coding domain of ~240 nt (Figure 2.4A). 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 (Figure 2.4C).

In addition to these highly conserved housekeeping ncRNAs, examination of results from three different tools for ncRNA prediction and manual curation led to the identification of 55 and 81 intergenic (IGR) sRNAs for Karp and UT176, respectively (Figure 2.3; see Appendix 1 for genome coordinates). Although intergenic sRNAs are trans-acting, *Ot* is not known to encode for any chaperone molecule, like hfq, that would facilitate the interaction with target genes. However, when normalized to the genome size of *Ot*, the number of identified IGRs is consistent with the number of sRNAs identified in model bacterial pathogens (Albrecht et al., 2011; Kröger et al., 2012; Sharma et al., 2010; Toledo-Arana et al., 2009; Vogel, 2003).

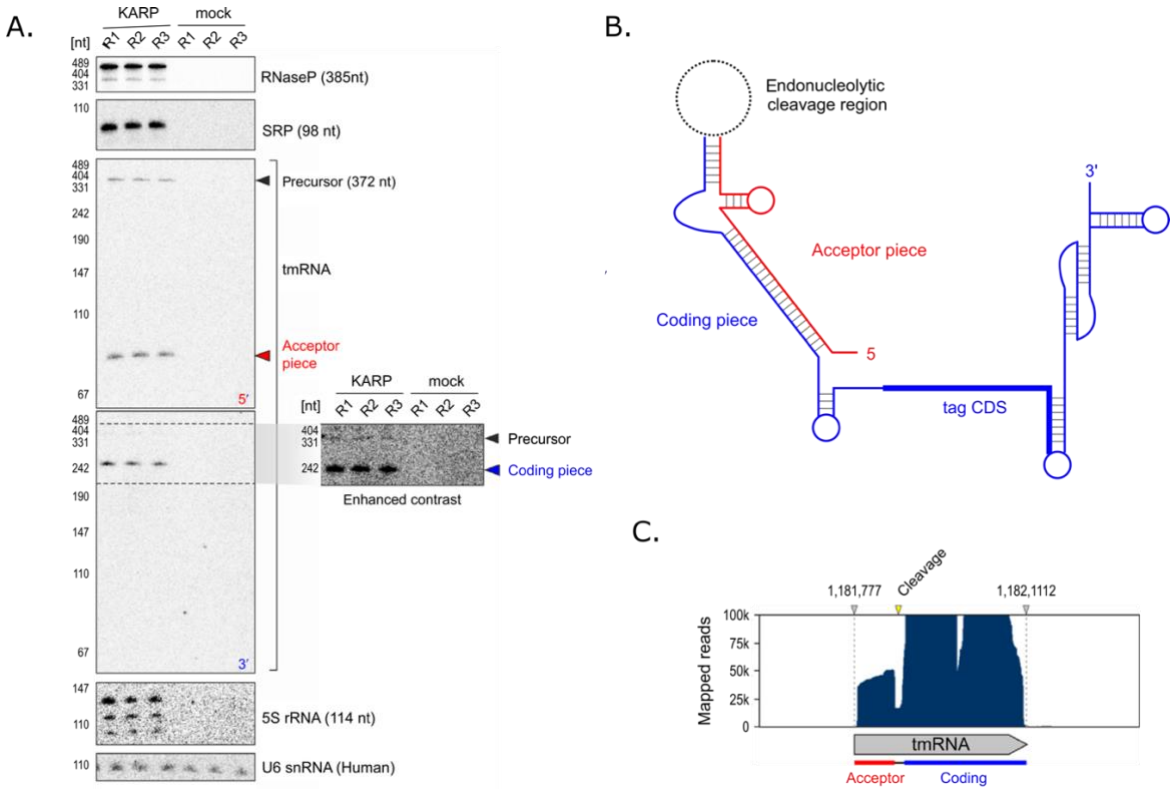

Figure 2.4 Experimental validation of identified housekeeping non-coding RNAs in Ot. A) Northern blot analysis of identified RNAs in Karp. B) Structure of the two-piece tmRNA identified in Ot. C) RNA-seq read coverage over the tmRNA gene mirrors cleavage observed by Northern blot. This figure from A to C was generated by Alexander J. Westermann.

2.3.3 Conserved operons in a dynamic genome

The two *Ot* strains studied here contain chromosomes diverse in size: 2 469 803 bp in Karp and 1 932 116 bp long in UT176 (Batty et al., 2018). The repetitive elements occupy 33% and 49% of the UT176 and Karp genomes, respectively, representing 38-47% of the genes in these genomes. Because of the massive amplification of the repetitive elements, extensive genome

shuffling has occurred between the *Ot* strains. There is a small correspondence between the positions of analogous genes in Karp and UT176. Figure 2.5A shows only minimal collinearity between these two genomes when compared to the high degree of synteny preserved between two bacteria from different genera, *E. coli* and *S.* Typhimurium. As bacterial genomes are organized into operons that facilitate co-transcription of functionally related genes, the aim of this analysis was to identify them and evaluate their conservation in such a dynamic genome. Sets of adjacent genes expressed as a continuous transcript were explored using Rockhopper (Tjaden, 2015) and manual curation. Combining the information on operon structures in each strain and predicted orthologous genes between Karp and UT176, 131 operons were identified as fully conserved with all genes expressed in both strains (Figure 2.5B; Appendix 1). In the case of seven operons only a partial set of their genes was expressed in both strains (Table S5). Batty et al. (Batty et al., 2018) identified 51 universally conserved genomic islands for eight *Ot* strains, that include 35 potential collinear gene clusters consisting of two to thirteen genes; operonic transcripts may originate from 24 of these (Appendix 1). In addition, 212 and 192 identified operons are present solely in Karp or UT176, respectively (Appendix 1). 73% of Karp- and 93% of UT176-specific operons consist of RAGE genes, whereas only 14% of conserved operons are expressed from these repetitive elements (Figure 2.5B). The length of conserved operons ranges from 2 to 30 genes, though 84% of them consist of only two or three genes (Figure 2.5D; Appendix 1). Longer co-transcribed gene clusters tend to encode for core cellular processes. For instance, two of them located in distinct loci contain 6 and 5 genes and encode for portions of the NADH–ubiquinone oxidoreductase complex organized

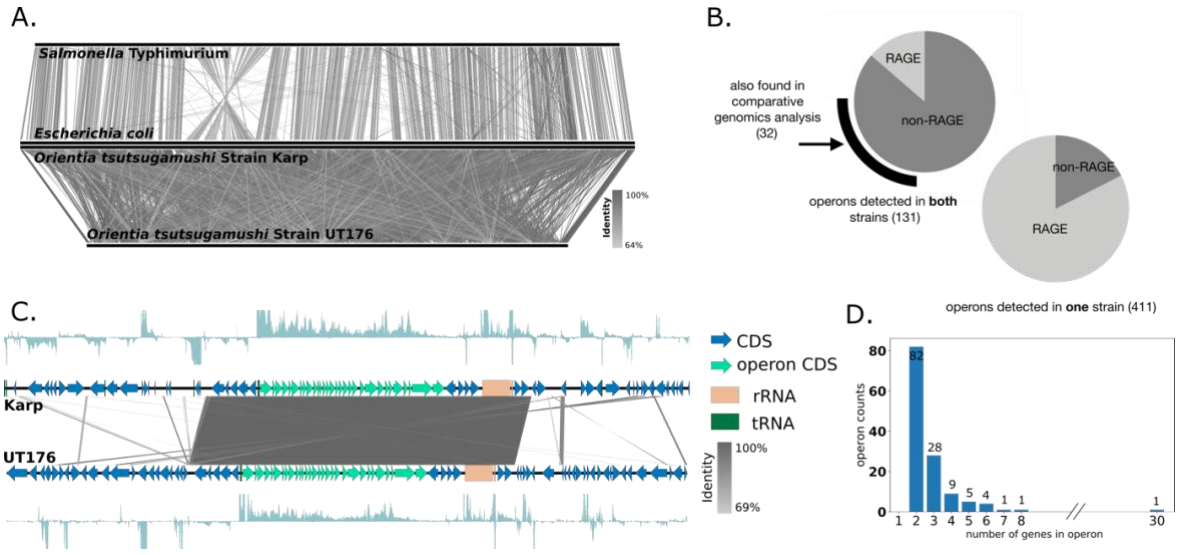

Figure 2.5 Comparative genomics and identification of operon structures in two Ot strains. A comparison of genomic synteny of two species within the same family (E. coli and S. Typhimurium, top), with synteny between the two Ot strains from this study (bottom). B) Representation of relative abundance of RAGE genes in both conserved (top pie chart) and strain-specific (bottom pie chart) operons, and the number of conserved operons identified within the conserved Ot genomic islands (Batty et al., 2018). This figure was created by Jeanne Salje. C) Visualization of the longest conserved operon identified in Karp and UT176 that encodes ribosomal genes; shown with RNA-seq coverage in both strains. D) Lengths of conserved operons identified in Karp and UT176.

similarly to that observed in *Rickettsia prowazekii* and eukaryotic mitochondria (Andersson et al., 1998). In addition, the 8 gene-long operon is involved in iron-sulfur cluster assembly. Finally, the longest operon that contains 30 genes encodes almost half of ribosomal proteins present in the *Ot* genome, which are proximal to co-transcribed ribosomal RNA and 5S rRNA genes (Figure 2.5C). In summary, identifying the identification of operons in a genome as highly dynamic as that of Orientia indicates a strong selection for co- transcription of those genes due to their involvement in the same pathways and likely shared regulation.

2.3.4 Evidence for *Ot* RAGE regulation by antisense RNA

Analysis of the RNA-seq data uncovered that almost half of the most highly expressed genes of Karp and UT176 belong to RAGE repetitive elements (defined in the present study as a set of integrases, transposases, conjugal transfer genes, and hypothetical proteins). Moreover, these genes were also highly expressed in the antisense direction in *Ot* (Figure 2.6A). This was confirmed by an enrichment analysis of antisense transcription in the RAGE elements in both genomes (Figure 2.6B), leading to the hypothesis that the repetitive elements may be regulated by antisense gene expression.

As the general role of antisense transcription is controversial, mass spectrometry was performed to explore both the relationship between the transcriptomics and proteomics of *Ot* and the potential regulatory role of the antisense transcripts. Because the higher bacterial load makes detection of bacterial proteins more likely, Karp was chosen to represent *Ot* strains in this investigation. Comparison of RNA-seq with the proteomics data set showed that significantly fewer RAGE gene products were detected by mass spectrometry (Figure 2.6A). Overall, only 318 proteins were detected within the analyzed data set, and their genes had higher transcript expression on average than those not detected by proteomics (Figure 2.6C). However, there were also highly expressed transcripts that were not observed by proteomics. Thus, considering previous results, the question arises whether antisense transcription would affect the protein expression. Interestingly, all genes with detected protein products had an antisense-sense read count ratio of <1, whereas a higher ratio was observed only for genes without the detected proteins (Figure 2.6D), suggesting antisense transcripts may play a role in inhibiting translation.

This hypothesis was verified using a machine learning approach. Three logistic regression models that predict protein detection based on the transcriptomics data were created; the first model uses TPM expression values of the sense strand to make the predictions; the second uses only the antisense-sense read count ratio as a predictor; the third model combines both features. For a first indication of whether any of these models are predictive, the models were trained on the reduced balanced data set (see 2.2.16 method section) and then tested on the complete data set. Evaluation of the models indicated that antisense transcription is predictive of protein expression (Figure 2.7).

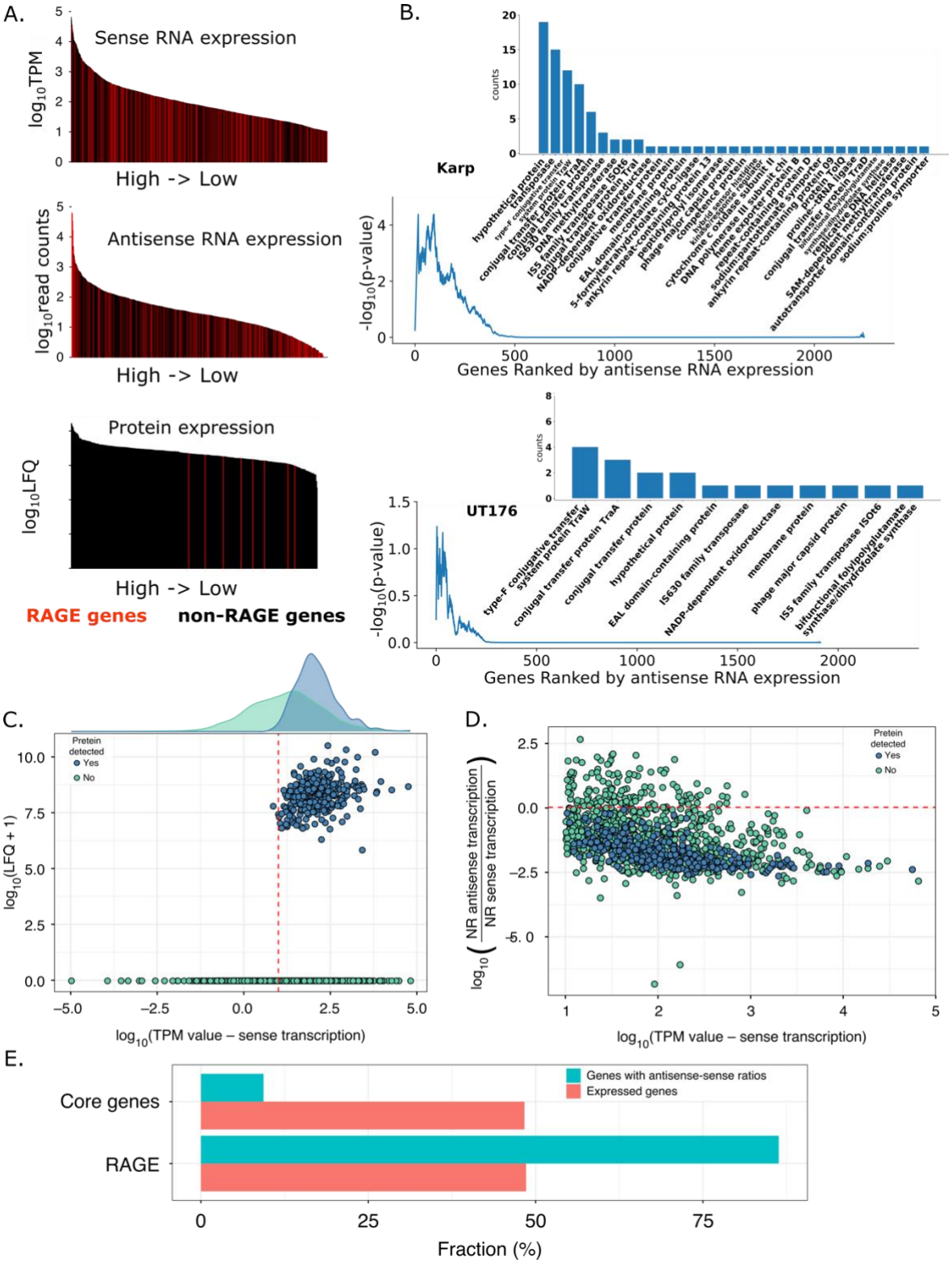

Figure 2.6 Antisense transcription in Ot. A) Visualization of the expression levels from the sense and antisense strands and protein expression over genes ranked from high to low levels in Karp; RAGE genes are marked in red. This figure was generated by Lars Barquist B) Enrichment of the antisense expression in RAGE genes in both Ot strains. Genes are ranked by the antisense expression from highest to lowest, and the distribution of one-sided hypergeometric p-values calculated for the enrichment of RAGE genes is shown. Inset plots show representative annotations for the top genes of the ranked list. C) Visualization of the relationship between protein expression, defined by LFQs, and transcript expression, defined by TPMs, indicating two gene clusters separated by the protein expression. The red line indicates the threshold for expressed genes (TPM value equal to 10). D) Sense transcription and antisense-sense ratio showing the classification based on proteomics detection. The red line indicates the sense-antisense ratio (1.06) above which no protein was detected by mass spectrometry; NR - number of reads. E) Comparison of the fraction of core genes and RAGE genes in the set of genes with high antisense–sense ratios and all expressed genes.

Model 1, which depends only on the sense expression, did little better than chance at predicting protein detection. Incorporating the antisense-sense ratio (models 2 and 3) increased the predictive power. Also, an application of 500-fold cross-validation (see 2.2.16 method section) confirmed the results and indicated that antisense transcription has a widespread regulatory role in *Ot*. In particular, antisense transcripts may control the expression of selfish genetic elements at the protein level, as the RAGE genes were significantly enriched among those with high antisense-sense ratios (Figure 2.6E). The antisense-sense ratio greater than 1 was also observed for thirty-one core genes (Appendix 1). These include genes encoding the chromosomal replication initiator protein dnaA, DNA polymerase subunit III, glutamine synthetase, the outer membrane autotransporter protein scaD, two transporters, the protein export protein secB, and 13 hypothetical proteins. None of the tested models achieved more than 65% balanced accuracy, which may indicate other post-transcriptional regulation mechanisms and/or a lack of sensitivity in the proteomics.

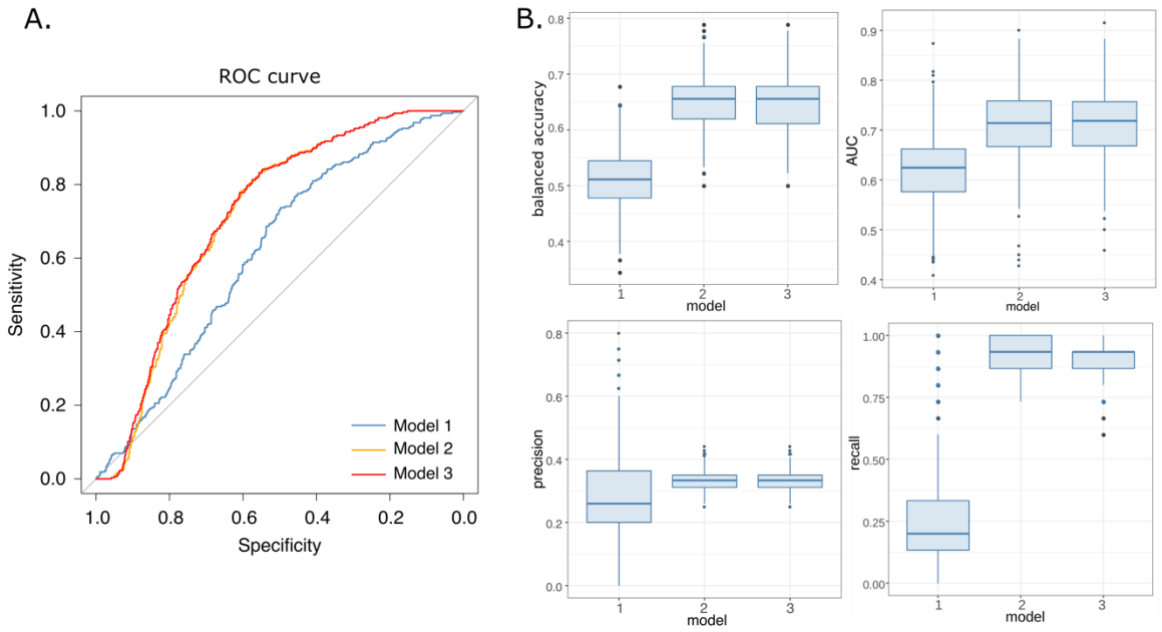

Figure 2.7 Evaluation of logistic regression models' performance. A) ROC (receiver operating characteristic) curves evaluating the performance of the models to predict protein expression from RNA-seq read counts. Model 1 strictly uses sense expression, model 2 the antisense–sense ratio, and model 3 uses both. Incorporating antisense expression clearly improves model performance. B) Performance measures of logistic regression models in 500 fold cross-validation. The box plot hinges represent the first and third quartiles of the data distribution, while the central line is the median. Upper and bottom whiskers are composed of the points between the maximum and minimum of the data sets within the 1.5x interquartile range (IQR) and the upper and bottom hinges, respectively. The individual points show outliers.

2.3.5 Differential expression of genes in Karp and UT176

As *Ot* is genetically intractable, the identification of virulence factors has been limited so far. Here, the aim was to apply dual RNA-seq for two different *Ot* strains and identify variations in gene expression that may discriminate between Karp and UT176 transcriptomes during the infection. Taking the sequencing data of these strains at 5 days after infection of HUVEC cells, differential expression analysis was performed between orthologous protein-coding genes of these two *Ot*

strains. Next, gene set analysis of KEGG pathways, GO terms, and manually created gene sets (see 2.2.11 method section) allowed a broader indication of biological processes that differ between Karp and UT176. Most of the enriched pathways were upregulated in Karp compared to UT176 (Figure 2.8; Appendix 1). These include gene sets involved in metabolism and DNA replication, which is consistent with Karp's higher growth rate observed in this study (Figure 2.1B). Others consist of the RAGE elements, secreted effector proteins, surface proteins and adhesins.

At the gene level (Figure 2.8B; Appendix 1), most genes with high differential expression belong to RAGE elements. In addition, several surface and effector proteins (ANK-containing proteins) were differentially regulated between the two strains. Of the known outer membrane proteins, the most differentially expressed genes were *scaE*, *tsa56*, and *tsa22* showing 1.40, 3.08, and 3.96 logFC difference in Karp over UT176, respectively (Appendix 1, confirmed by qRT-PCR in Figure S2). In contrast, *scaD* expression was higher in UT176 but to a lesser degree (0.99 logFC in UT176 over Karp). It is likely that different expression levels of these surface proteins will affect interactions with host cells (through stronger binding of host cell receptors or activation of innate immune receptors) or affect the induced adaptive immune response in animals.

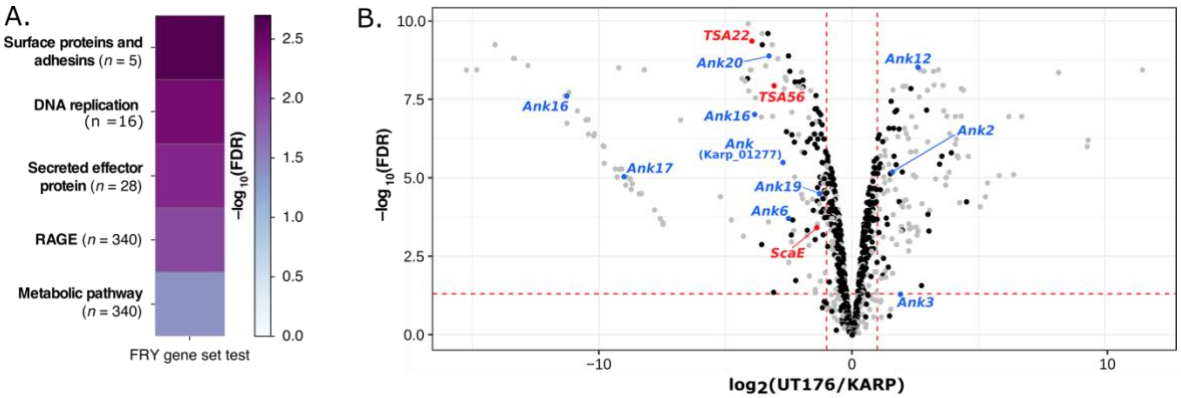

Figure 2.8 Differential gene expression analysis between Karp and UT176 infecting HUVECs. A) Bacterial gene sets enriched in differentially expressed genes. All illustrated pathways are induced higher in Karp than UT176. FDR-corrected p-values were calculated using the fry gene set enrichment test implemented in the edgeR R package (M. D. Robinson et al., 2010). B) Volcano plot representing differentially expressed bacterial genes in Karp and UT176. Bacterial surface genes (red) and ankyrin-repeat-containing effector proteins (blue) with $\log_2$ fold change ≥1 are highlighted. Gray dots indicate RAGE genes. FDR-corrected two-sided p-values were calculated using the quasi-likelihood F-test in the edgeR R package.

Other types of *Orientia* gene products involved in the infection process are effector proteins such as ANK- and TPR-containing proteins. Each of the *Ot* genomes studied here encodes multiple copies of genes encoding those effectors. While the Karp genome contains 33 and 29 copies of Ank and TPR genes, respectively, the UT176 genome contains 21 Ank and 22 TPR genes (Batty et al., 2018). Several of them were identified as differentially expressed (Figure 2.8B; Appendix 1) here. The higher expressions in UT176 than Karp were observed for *ank2*, *ank3*, *ank12*, and two copies of *tpr8* (logFC >1.5 in UT176 vs. Karp), whereas six Anks (*ank6*, *ank20*, *ank17*, and two copies of *ank16*) and three TPR genes (*tpr1*, *tpr3*, and *tpr5*) were upregulated with a logFC greater than 1.5 in

Karp. The function of most of these effectors is unknown, except for Ank6 that interferes with NF-κB translocation to the nucleus and inhibits NF-κB-dependent transcription (Evans et al., 2018). In addition, most of the protein products of these genes were not detected in the Karp proteomics data set, suggesting that either their transcripts were not translated or that the proteins were secreted and lost during purification. As secreted effectors traffic to various subcellular compartments of the host cells and interact directly with its proteins (Bang et al., 2016; Beyer et al., 2017; Min et al., 2014; VieBrock et al., 2015), their differential expression may lead to downstream differences in the host response.

2.3.6 Karp and UT176 induce a proinflammatory response

To understand how *Ot* affects the transcriptional program of HUVEC cells, gene expression of host cells infected with either Karp or UT176 was compared to that of unstimulated cells. The results showed a clear core response to both strains dominated by a type-I interferon proinflammatory response (Figure 2.9A, red; Appendix 1). Induction of this signaling has been previously identified in both cultured monocytes and human-derived macrophages infected with *Ot* (Tantibhedhyangkul et al., 2011, 2013). Here, genes of the type-I interferon pathway were upregulated in HUVEC cells stimulated with both Karp (Figure 2.9B) and UT176 (Figure 2.9C). These include *IFNB1* (interferon beta), genes involved in regulating the type-I interferon response — *IRF9* (interferon-regulatory factor 9) and *STAT1/2* — as well as various interferon-stimulated genes such as interferon-induced proteins with tetratricopeptide repeats (*IFIT*) genes and 2′-5′-oligoadenylate synthase 1 (*OAS1*). In addition, infection with either strain induces high expression of proinflammatory chemokine genes, including *CXCL10*, *CXCL11*, and for cytokine receptors *IL13RA2*, *IL7R*, *IL15RA*, and IL3RA (Appendix 1).

The upstream signals leading to activation of these genes are unknown but in addition to already identified PRRs stimulated by *Ot* — NOD1 and TLR2 (K.A. Cho et al., 2010; Gharaibeh et al., 2016) — the upregulation of *TLR3* in HUVEC cells infected with both *Ot* strains was identified in this study (Appendix 1). As this receptor recognizes viral double-stranded (ds)RNA (Kawai & Akira, 2007), it is possible that it detects cytosolic dsRNA of *Ot*. Induction of the *IRF7* transcription factor, known to respond to stimulation from membrane-bound TLRs, further supports the role of TLR2 and TLR3 in the detection of this pathogen.

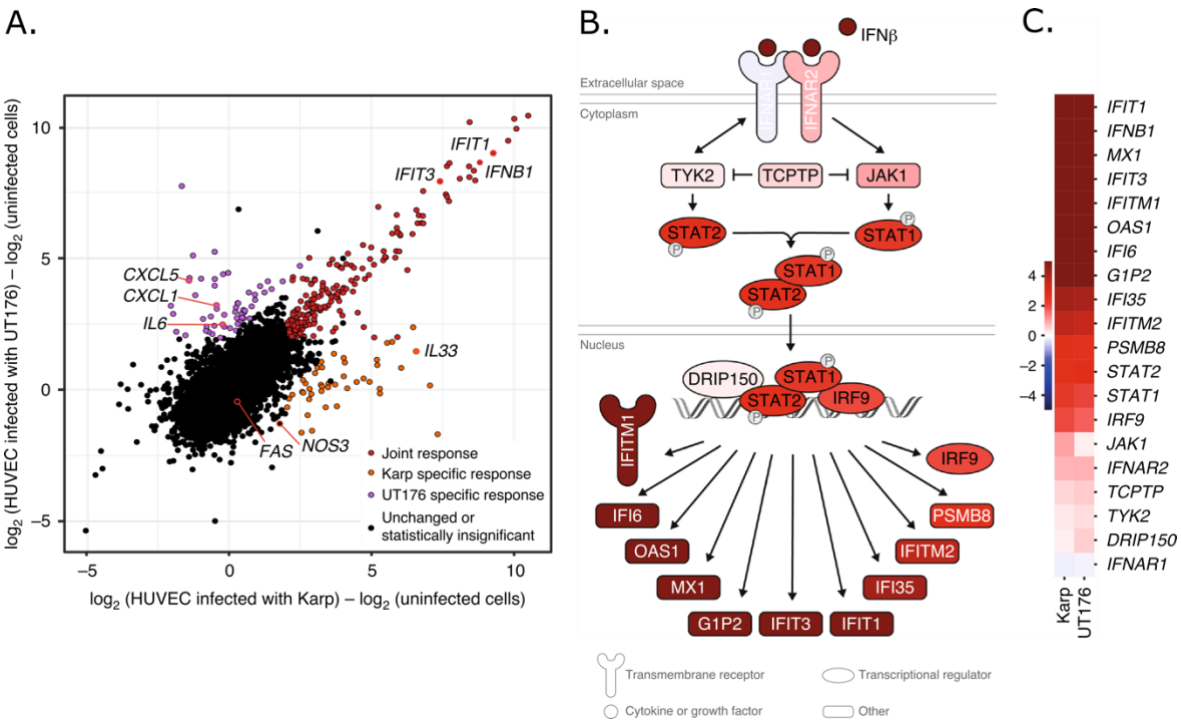

Figure 2.9 Common host response to the infection with Ot strains. Joint and strain-specific host responses. The joint response is defined as genes with a logFC > 2 and FDR-corrected p-value < 0.01 for HUVECs infected with both Karp and UT176. Strain-specific host responses are composed of genes with a logFC > 2 and FDR-corrected p-value < 0.01 for HUVECs infected with either Karp or UT176, excluding genes already specified in the joint response. FDR-corrected two-sided p-values were calculated using the quasi-likelihood F-test in the edgeR R package (M. D. Robinson et al., 2010). B) Gene network of the canonical interferon signaling pathway activated in Karp-infected HUVEC cells compared with uninfected host cells. The network was generated by Selvakumar Subbian and re-drawn by Sandy Pernitzsch. C) Heatmap with genes of the interferon signaling pathway induced in the HUVECs infected with two Ot strains compared with uninfected host cells. The color scale represents the logFC in gene expression.

2.3.7 Differential host responses to Karp and UT176

Although comparing the transcriptional profiles of *Ot* infected endothelial cells with uninfected cells allowed the identification of common responses induced after the infection with both strains (Figure 2.9A, C), a differential gene expression analysis between Karp and UT176-infected cells has also uncovered responses unique to each strain. The results show that UT176 induces higher expression of multiple proinflammatory cytokines, chemokines, and cytokine receptors compared to Karp (Figure 2.10A; Appendix 1); these include *CXCL8*, *CXCL1*, *CXCL2*, *CXCL10*, *IL6*, *IL1RL1*, and *IL18R1*. In addition, expression of genes encoding cytokine-inducible surface adhesion molecules — *VCAM1* and *ICAM1*, upregulated upon endothelial cell activation — was also higher in HUVECs infected with UT176 than Karp (Appendix 1). Although HUVEC cells expressed *TLR3* in response to infection with either strain, higher induction of *TLR3* was observed in UT176-infected cells than those infected with Karp. The IPA analysis (see 2.2.17 method section) also indicated that UT176 stimulates higher expression of genes involved in the NF-κB pathway and NOS2 production compared to Karp (Figure S4). Also, stronger induction of expression of genes associated with leukocyte proliferation and mononuclear leukocyte differentiation was observed in host cells infected

with UT176 than with Karp (Figure S5). Thus, UT176 seems to induce a stronger proinflammatory response that may lead to more efficient pathogen clearance (Figure 2.1B).

In contrast to UT176, Karp induced higher expression of the proinflammatory cytokine involved in the scrub typhus pathogenicity — *IL33* (Shelite et al., 2016) (5 logFC difference between Karp- and UT176-infected cells; Appendix 1). Moreover, most genes involved in the IL33-FAS network also reached higher mRNA levels in the HUVEC cells stimulated with KARP (Figure 2.10B, Figure S3). In contrast to UT176, Karp induced networks of genes involved in (i) organismal growth failure, (ii) organismal morbidity and mortality, (iii) and organismal death (Figure S6).

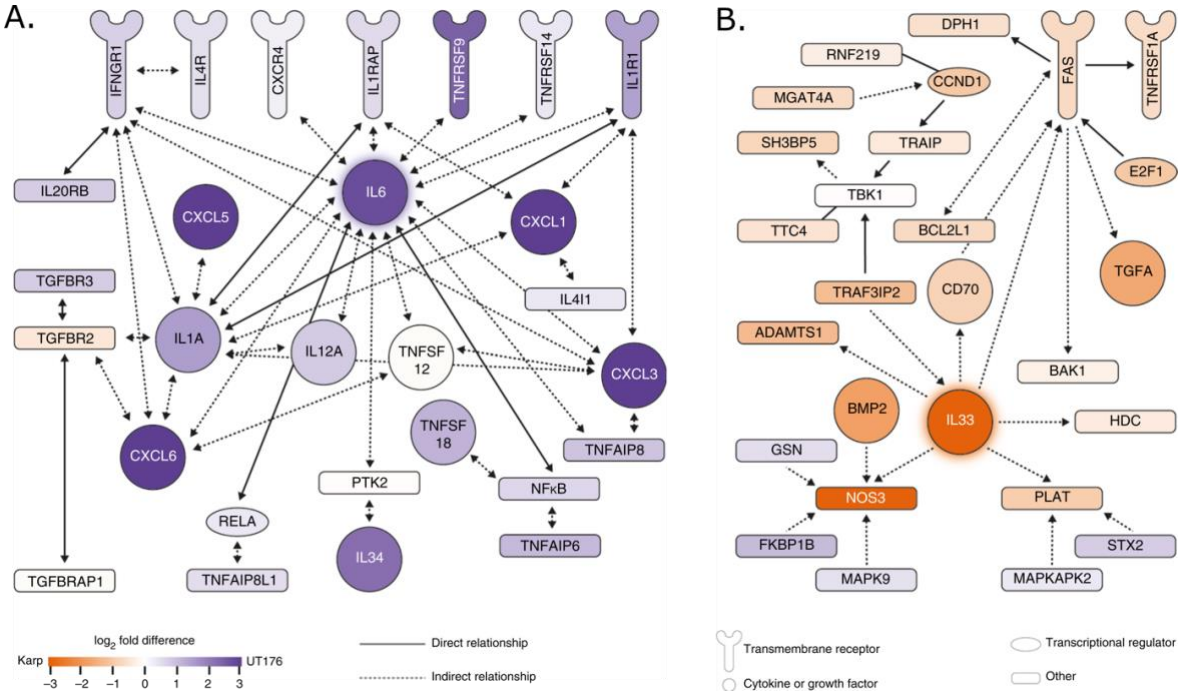

Figure 2.10 Differently induced host gene networks in response to Karp and UT176. A) A network map of proinflammatory chemokines and cytokines upregulated in HUVECs infected with UT176 compared to Karp-infected host cells. B) IL33-FAS-mediated anoikis network induced in HUVEC cells stimulated with Karp. The networks were generated by Selvakumar Subbian and re-drawn by Sandy Pernitzsch.

2.3.8 Two *Ot* strains differ in virulence in a mouse model

The relative virulence of the two strains was also tested in a murine infection model (see 2.2.18 method section). The results of the experiment showed that Karp-infected mice exhibit lower weight gain (Figure 2.11A) and more severe clinical symptoms (Figure 2.11B) than UT176-infected animals. The bacterial load in the blood (Figure 2.11C) and tissues (Figure 2.11D) was also higher in mice infected with Karp than with UT176. In addition, the histopathological analysis indicated more severe disease manifestation in lung, liver, kidney, and spleen of Karp-infected mice (Figure 2.11E-F). The diffuse thickening of alveolar septa and infiltration of macrophages and lymphocytes were more evident in mice infected with Karp.

Overall, the experiment on the murine infection model confirmed that Karp is more virulent than UT176. Although these results are consistent with the observations in HUVECs, the outcome

observed here may not translate to humans. Using only a single time point (12 days) is another limitation of this study. As the disease dynamics might differ between the two strains, different results could be observed at different times after infection. Thus, the differential host response between mice infected with Karp and UT176, including the roles of the adaptive immune response and dissemination kinetics within the host, requires further investigation.

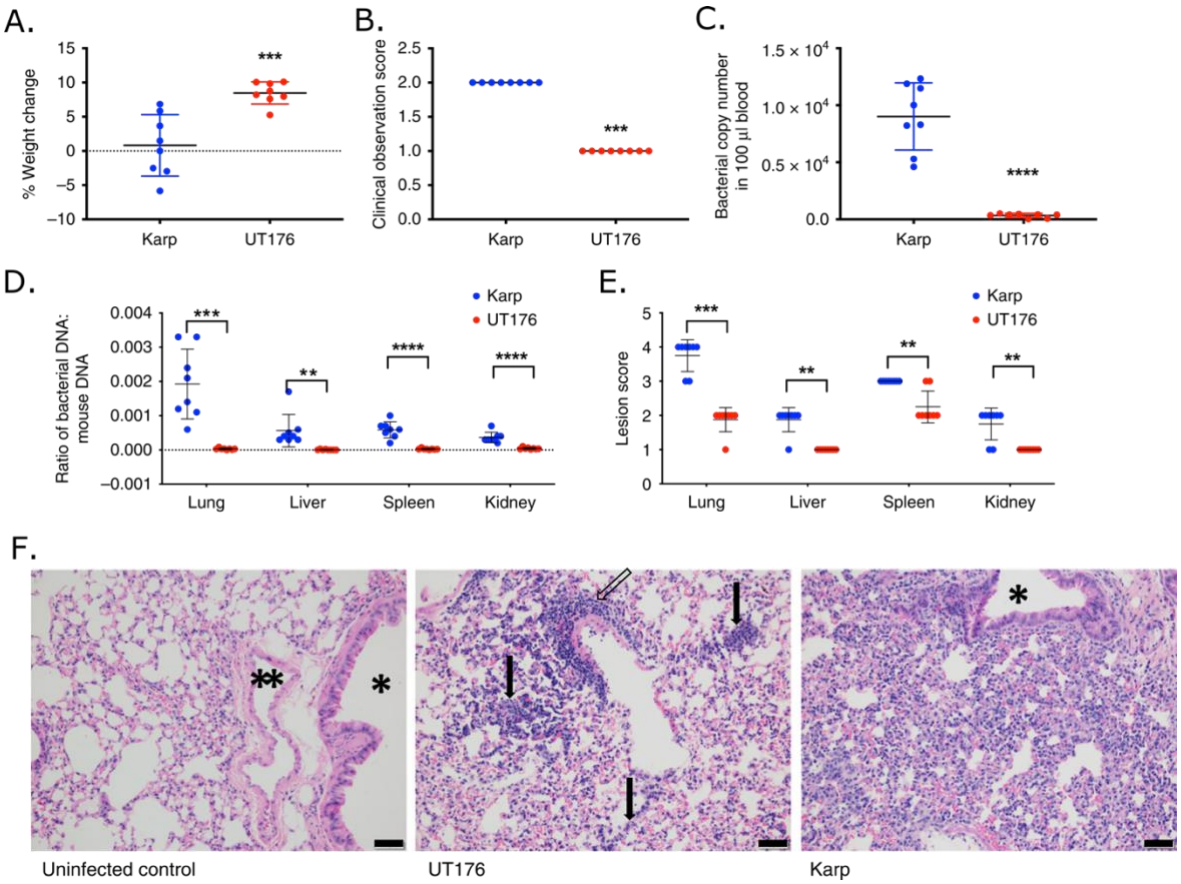

*Figure 2.11 Assessment of severity of infection with two Ot strains in mice. Weight change of Ot-infected mice over 12 days of infection. B) Clinical observation score of mice determined based on their appetite, activity, and hair coat 12 days p.i.; higher numbers represent low appetite, low activity, and ruffled fur. Details on the scoring system are provided in Figure S7. C) Bacterial genome copy number detected in 100 µl blood of euthanized mice 12 days p.i. D) The ratio of bacterial genome copy number to mouse genome copy number in organs of euthanized mice 12 days p.i. E) Lesion scores of stained organs of euthanized mice 12 days p.i. Lesion score ranges from 0 (normal tissue) to 5 (severe lesion damage). More details are provided in Figure S7. Mean and standard deviation is shown in all graphs. Statistical significance was calculated using unpaired t-test (**$p \leq 0.01$ ***$p \leq 0.001$ ****$p \leq 0.0001$). F) Images of stained lung tissue of mice infected with either buffer, UT176, or Karp. Scale bars = 50 µm. * indicates airway and ** indicates blood vessels. Arrows indicate diffuse thickening and infiltration of alveolar septa with a mixed population of macrophages and lymphocytes in UT176-infected lungs; open arrow - mild perivascular lymphohistiocytic inflammation. Diffuse moderate thickening and infiltration of alveolar septa with a mixed population of macrophages and lymphocytes occurs in Karp-infected lungs. The airway (*) is unaffected and normal. This figure from A to F was generated by Jeanne Salje and Piyanate Sunyakumthorn.*

2.4 Discussion

Both the obligate intracellular lifestyle and the complexity of the genome architecture make *Ot* difficult to study. In particular, the genome instability makes investigation of its transcriptional

architecture challenging. In this study, dual RNA-seq has allowed us to profile gene expression of two *Ot* strains and identify housekeeping ncRNAs, putative intergenic sRNAs, and operonic transcripts. Interestingly, among the conserved operons between Karp and UT176, few contain more than two or three genes, and these encode for proteins involved in core cellular processes, e.g. respiration and translation. Another aspect investigated in this study was antisense transcription of *Ot* and its role in the control of protein expression.

The Karp genome contains only 12 known transcription factors and 3 sigma factors, which is a relatively small number compared to *E. coli* which possesses 300 transcription and 7 sigma factors. This may indicate that transcription is not strictly controlled, and an alternative mechanism of protein expression regulation may play a dominant role in *Ot*. This study has shown that antisense expression is partially predictive of protein expression in the Karp strain (Figure 2.7). In addition, enrichment of the antisense expression of RAGE genes in Karp and UT176 (Figure 2.6B) suggests that such regulation occurs also in other *Ot* strains, especially for this particular type of repetitive elements (conjugal transfer proteins, transposases, integrases, and hypothetical proteins). Although antisense regulation of transposons was already identified in several bacterial species (Ellis & Haniford, 2016), it has never been observed at the scale implied by this study. Such antisense regulation could arise spontaneously through capture of transcriptional noise, providing a parsimonious alternative to transcriptional control (Jose et al., 2019). The functionality of these untranslated transcripts in *Ot* is unclear. They might also be selfish elements that the pathogen is unable to eliminate because of its small population size. However, this regulatory mechanism would provide dsRNAs upon intracellular bacterial lysis, which may explain the observed induction of TLR3 and an antiviral immune response.

As genetic manipulation of this obligate intracellular pathogen is currently not possible, the identification of factors that drive virulence differences between two strains is challenging. In this study, dual RNA-seq captured gene expression of *Ot* Karp and UT176 at a similar infection condition, which allowed us to perform differential gene expression analysis, leading to the identification of upregulated genes encoding virulence-associated surface proteins in Karp. Comparison of the host response to the infection with either strain indicated differential activation of the immune reaction, which correlated with differential outcomes in a murine infection model of scrub typhus. Beside an antiviral proinflammatory response activated upon the infection of endothelial cells with either strain, our analysis indicated induction of an IL6-mediated proinflammatory response after infection with UT176, and an IL33-NOS3-FAS response in Karp-infected cells. These differences may partly explain the differences in virulence of these strains. As IL-33 has already been linked with severe outcome of Karp infection in mice, decreased levels of IL-33 upon infection with the less virulent UT176 strain supports its role in pathogenesis of scrub typhus.

However, the present study has several limitations. First, it is impossible to distinguish between differential host responses due to actively replicating bacteria versus non-replicating bacteria, nor between *Ot* specific versus non-specific uptake outcomes. Secondly, a higher growth rate (Figure 2.1B) leading to higher read counts of Karp compared to UT176 (Figure 2.2A) limits the interpretability of our data. Although the bacterial differential gene expression analysis is unlikely to be affected by the differences in bacterial growth rate as normalization between samples was applied, it is difficult to assess how these differences in bacterial growth affect the differences in bacterial virulence on the host response. Finally, in order to obtain enough reads, a relatively high MOI (~30:1) was applied in the RNA-seq experiments, which does not reflect the physiological infection conditions where the bacterial cell number is likely lower; this may affect the immune response of host cells. In addition, it would also be interesting to further investigate these findings at other time points and in different cell types.

In summary, dual RNA-seq combined with proteomics and bioinformatics has allowed us to investigate the transcriptome structure and gene regulation mechanism in the genetically intractable obligate intracellular human pathogen. We identified conserved operons between Karp and UT176, ncRNAs that might serve a regulatory role in either strain, and widespread post-transcriptional antisense regulation, particularly for the RAGE genes. In addition, differential expression analysis of bacterial genes has indicated factors that may drive virulence in host cells. Finally, comparative analysis of host transcriptomic data provided evidence for a connection between disease severity and the relative induction of IL33- and IL6-based gene networks. This study forms the foundation for further investigations of gene expression regulation in *Ot* and provides new perspectives on the mechanisms of pathogenesis. Importantly, it may also serve as an example for further characterization of other obligate intracellular bacterial pathogens.

3 Dualrnaseq: a Nextflow-based workflow for host-pathogen dual RNA-seq analysis

The work presented here was performed in collaboration with Regan Hayward (Helmholtz Institute for RNA-based Infection Research, Würzburg, Germany) and Lars Barquist (Helmholtz Institute for RNA-based Infection Research, University of Würzburg, Germany)

3.1 Introduction

Dual RNA-seq captures both host and pathogen transcriptomes at the site of infection, facilitating an exploration of processes that play an essential role in pathogenesis and the host defense. RNA-seq generates a pool of reads whose origin must be established *in-silico*, and their assessment allows for quantitative description of transcripts expressed in a given condition. This chapter describes a workflow, Dualrnaseq, for dual RNA-seq data analysis. As this experimental technique extends the standard RNA-seq protocol, the computational methods in this workflow applied for RNA abundance estimation of two interacting organisms are inherited from RNA-seq-based approaches. Thus, this chapter briefly describes the common steps of RNA-seq data processing and widely used methods. Next, I review computational techniques applied to various dual RNA-seq studies to illustrate the challenges that simultaneous analysis of two different organisms brings to the field. This review also uncovers another problem: poor descriptions of the technical aspects in many publications, hindering reproducibility in science. I give an overview of solutions supporting the reproducibility of projects widely used in the computational field. Next, the introduction to different pipeline frameworks that facilitate the creation of user-friendly workflows highlights the advantages of Nextflow, a platform for Dualrnaseq. Description of approaches applied in the pipeline and the presentation of adopted data processing steps gives a comprehensive picture of the Dualrnaseq workflow. Finally, a simulation-based benchmark analysis compares the employed methods in the context of various host-pathogen systems and provides advice regarding their application.

3.1.1 RNA-seq data processing methods

RNA-seq data analysis consists of several major steps performed sequentially, leading to the quantification of the transcriptome of a studied biological system (Conesa et al., 2016; Van den Berge et al., 2019). A sequencing platform generates pieces of nucleotide sequences, known as reads, and the information about their sequences and quality is stored in FASTQ files (Cock et al., 2010). The quality of each base in a read is represented by a PHRED score describing a probability of being

called incorrectly. In the first phase of the bioinformatics analysis, the raw sequencing data are subjected to quality control (QC) which aims to generate: (i) overview statistics of a total number of sequenced reads, their length and %GC content, enabling to check if these values are as expected; (ii) a read quality assessment necessary to identify problems occurring during sequencing; (iii) and a report of overrepresented sequences helping to identify contamination in the data like adapters. All these aspects can be evaluated using FastQC (Andrews S., 2010), one of the most popular Illumina read quality control packages. In addition, several tools such as NGS QC Toolkit (Patel & Jain, 2012), FastqPuri (Pérez-Rubio et al., 2019), fastp (S. Chen et al., 2018), HTQC (Yang et al., 2013), and AfterQC (S. Chen et al., 2017), also offer QC of raw reads. Applying such approaches is crucial to identify potential problems in sequencing data that may affect the downstream analysis. For instance, low-quality read bases, i.e., bases with high error probability, may hinder the identification of the read position on the genome, decreasing fragment mappability (Del Fabbro et al., 2013). Also, underrepresentation of fragments with low or high GC-content, mainly attributed to inefficient PCR amplification of such sequences during library preparation, results in underestimated abundances of transcripts from which these fragments originate (Love et al., 2016). Therefore, assessing sequencing data and identifying technical biases is essential as it allows for further adaptation of strategies improving the RNA-seq data quality.

The first step to improve raw sequencing data quality is trimming of both adapter sequences and low-quality read ends. The low-quality bases can be removed using, e.g., a sliding window-based algorithm implemented in one of the popular trimming tools — Trimmomatic (Bolger et al., 2014). This tool scans read beginning at the 5' end and removes bases from the 3' end if the average quality within the window is lower than a user-defined threshold. Another method is a running sum used by Cutadapt (M. Martin, 2011), where a user-defined cut-off is subtracted from all base quality scores to create a sequence of values, for which the partial sum is calculated in the next step. The minimum partial sum value indicates the position at which the read sequence is clipped. After the trimming of low-quality read ends, Cutadapt finds user-specified adapter sequences, using a semi-global alignment, and removes them from the reads. However, this step in Cutadapt requires knowledge about adapters used in the library preparation step. If the adapter sequences are unknown, tools that use an adapter database can be employed. For instance, Trimmomatic removes Illumina-specific sequences, and BBDuk (B. Bushnell, 2020) cuts adapters from various library preparation protocols. Moreover, BBDuk is faster than Cutadapt and Trimmomatic, as it matches adapter k-mers to the reads instead of performing alignment (Guzman & D'Orso, 2017). Some tools integrate QC and trimming steps. Examples include FastqPuri (Pérez-Rubio et al., 2019), fastp (S. Chen et al., 2018), AfterQC (S. Chen et al., 2017), and QC-Chain (Zhou et al., 2013). Although the goal of applying such tools is to improve the quality of the data, this process needs to be approached with caution. Aggressive quality-based trimming may influence the identification of short or lowly expressed

transcripts, affecting the estimation of transcript expression levels and the downstream analysis, e.g., identification of differentially expressed genes (Macmanes, 2014; Williams et al., 2016).

When the adapter sequences are removed, and the subsequent QC confirms that the trimming step improved the overall data quality, the origin of the reads needs to be established. If an organism of interest is well characterized, reads can be mapped to its available references, i.e., genome, which is a representative DNA sequence of an organism, or a transcriptome containing a set of transcript sequences — the products of the transcription and, in eukaryotes, splicing process. However, mapping reads only to the transcriptome prevents the detection of novel transcripts and allows quantification of only already identified gene isoforms. In contrast, mapping to the genome sequence involves aligning reads onto the whole DNA sequence independently from the gene annotations providing coordinates for known genes and transcripts. However, sequencing data obtained from eukaryotic organisms can contain reads that span exon boundaries. Therefore, mapping to a reference genome, which includes both introns and exons, requires tools that allocate pieces of reads across splice junctions. Spliced alignment, which is an alignment of reads against an unspliced genomic sequence can be performed with the most popular mappers: TopHat (D. Kim et al., 2013) and STAR (Dobin et al., 2013). However, such tools are usually computationally expensive as they perform additional steps to handle the alignment of reads to exons separated by introns. In contrast, transcriptome mapping does not require any sophisticated read mappers because gene isoforms are present in the reference and reads can be aligned directly to them. Although the number of computationally demanding steps is reduced in this approach, a problem with allocating reads that originate from exons present in different isoforms arises. In this case, a read may be mapped equally well to multiple gene isoforms that share an exon with a sequence equivalent to the read. Such a read would have been mapped to a unique locus on the genome. Each strategy has advantages and disadvantages, but the choice between transcriptome or genome is mainly driven by the availability of references and the study goal. Genome mapping facilitates the identification of novel transcripts from RNA-seq data, which is impossible for transcriptome mapping.

The reference-based mapping methods can be divided further into alignment-based and alignment-free strategies. Although these approaches first pre-process (index) a reference sequence to reduce its size and make queries fast, the major difference between these strategies occurs in further steps. The alignment-based method, implemented in, e.g., TopHat (D. Kim et al. 2013) and STAR (Dobin et al. 2013), performs alignment of reads to the best matching piece of the reference sequence. Next, a quantification tool assigns a read to either genes or transcripts and estimates their abundance. In contrast, alignment-free tools perform both mapping and quantification computationally efficiently. Instead of mapping entire reads, alignment-free tools split reads into k-mers (k-long nucleotide subsequences) and match to the pre-indexed transcriptome, allowing the assignment of reads directly to transcripts. The key concept behind this approach, pioneered in Sailfish (Patro et al., 2014), is the observation that counts of k-mers occurring in reads are sufficient

to accurately estimate transcript coverage, that is the average number of reads mapped to a given position in the reference sequence. Other alignment-free methods include pseudo-alignment present in Kallisto (Bray et al. 2016) and quasi-mapping implemented in both Sailfish from version 0.7 (Srivastava et al., 2016) and Salmon (Patro et al., 2017). Salmon with the quasi-mapping algorithm assigns a read to a transcript by looking for a chain of exact unique matches between the read and the transcript that cannot be extended in either direction. As these matches cover a unique sequence fragment that distinguishes them from other matches, reads that would have been aligned to multiple loci are assigned to a single transcript. However, if sequenced reads vary from the reference, their actual origin may be missed leading to misquantification. The solution is a Selective-Alignment (SA) algorithm that performs alignment-free mapping and selects the best transcript candidate for a read by computing an edit distance quality score for each pair of ambiguous matches (Sarkar et al., 2018). Nevertheless, alignment-free techniques utilize the transcriptome solely, which increases the risk of mapping reads to annotated regions when they originate from unannotated genomic loci. To avoid assigning the reads to the wrong transcripts when the relevant genomic region is unavailable, Salmon with SA provides a possibility to define decoy sequences (Srivastava et al., 2020) that represent either a set of sequences extracted from the genome that are similar to sequences present in the transcriptome reference; or a whole-genome sequence. Thus, Salmon with Selective-Alignment and decoy sequences has become an ultra-fast method to accurately determine transcript abundance. However, the biggest drawback of both the alignment-free methods and the SA algorithm is using the transcriptome as a reference, which prevents the identification of novel transcripts. For this purpose, alignment-based tools using the genome as a reference may be a better choice.

Novel transcripts can also be identified with transcriptome assembly methods. Depending on the availability of the reference genome sequence, there are two main strategies: a reference-based approach and *de novo* transcript assembly (J. A. Martin & Wang, 2011). In the first procedure, a splice-aware aligner is employed to generate read-to-genome alignments, then used to create a graph of all base connections in the transcriptome and identify possible alternative splicing events. In the final phase, the graph is parsed into individual transcripts. Such a strategy can be performed using Cufflinks for example (Trapnell et al., 2010). However, reference-based transcriptome identification methods are limited when a high-quality reference genome is unavailable. In that scenario, the *de novo* transcriptome assembly technique implemented for example, in Trinity (Grabherr et al., 2011) may be a solution. In this method, transcripts are reconstructed directly from sequencing reads using a De Bruijn graph-based approach without prior-defined genome annotations. In this way, assembly methods have significantly impacted the investigation of transcriptomes of non-model organisms and improved their annotations (Mahmood et al., 2020; Rana et al., 2016).

The reference-based assembly methods have several advantages over *de novo* techniques. They are more computationally efficient and sensitive enough to assemble low abundant transcripts. They are also less sensitive to contamination and sequencing artifacts because of the slight chance

of aligning the contaminating reads to the reference genome. However, the accuracy of their assemblies depends on the genetic divergence between the reads and the reference, which is not the case for *de novo* methods. Therefore, to exploit the benefits of each of the transcriptome assembly methods, one can use a reference-based approach, bringing high sensitivity, followed by *de novo* assembly of unaligned reads to detect novel transcripts. This method is called align-then-assemble and can be also applied to filter out reads from an irrelevant organism before assembling the reads of the target organism. Another technique, assemble-then-align, can be applied if the quality of the reference genome is low. In this case, *de novo* assembled reads are scaffolded and then extended based on the reference genome allowing full-length transcript reconstruction. Regardless of which approach is employed to identify transcripts, either reference-based, *de novo* transcriptome assembly, or sequential application of these two, the next step involves quantification.

Once reads are mapped onto the genome or transcriptome, or the transcriptome is assembled, a quantification approach is applied to determine transcript or gene-level abundances. Unfortunately, assigning reads or read pairs to annotated locations brings about some challenges. In particular, reads generated from repetitive sequences, shared domains of paralogous genes, or different gene isoforms create ambiguity regarding the true locus of origin. Such reads are referred to as multi-mapped. Although each quantifier allocates multi-mapped reads among transcripts or genes using different strategies (Deschamps-Francoeur et al., 2020), the standard approach relies on counting reads assigned to only one position in the reference, called uniquely mapped reads, and discarding those mapped to multiple regions. HTSeq (Anders et al., 2015) and featureCounts (Y. Liao et al., 2014) are the most popular tools that employ this strategy by default. Nevertheless, quantification of multi-mapped reads requires more sophisticated methods than simple counting, such as model-based approaches. For instance, RSEM (B. Li & Dewey, 2011; B. Li et al., 2010) utilizes a generative model and the Expectation-Maximization (EM) algorithm to handle read mapping uncertainty and estimate relative transcript abundances. Model-based quantification is still an open area of research, and alignment-free methods also offer several variations of the statistical models (Bray et al., 2016; Patro et al., 2017; Srivastava et al., 2016).

Regardless of the chosen scenario of the read quantification, the standard output is a matrix containing either read counts that represent the number of reads that have been assigned to each gene or the estimated relative abundance levels of transcripts. Those values serve as a basis for downstream analyses, including differential gene expression analysis, pathway analysis, or network analysis to explore differences between studied conditions or identify gene-level interactions. With the remarkable evolution of high-throughput technologies, bioinformatics tools have been developed continuously to create accurate computational approaches that exploit the potential of each innovation in the sequencing field. Each year, the number of computational methods is constantly increasing (Deshpande et al., 2020), expanding their application to explore novel biological

questions. Since dual RNA-seq extends the RNA-seq protocol, existing RNA-seq methods need to be adapted to simultaneously identify transcriptomes of two interacting organisms.

3.1.2 Review of dual RNA-seq analysis workflows

The number of dual RNA-seq applications has increased over the last few years. In this section, I review methods applied to analyze various dual transcriptomes, mainly focusing on research performed on bacteria (see Table 3.1). In the text, I also mention studies of other biological systems if they use an approach of interest. Overall, some studies aimed to perform complex downstream analysis to capture direct host-pathogen interactions, but others focused on identifying factors that differ between conditions in either organism. Nonetheless, some common steps are applied in these analyses, including processing raw sequencing reads to estimate the gene expression of interacting organisms.

In each RNA-seq sequencing study, the processing of raw reads should start from their quality evaluation and removal of both adapters and low-quality nucleotide bases. However, many reviewed publications did not provide information about applied QC and trimming steps at all (see Table 3.1). If applied, FastQC (Andrews S., 2010) and Trimmomatic (Bolger et al., 2014) were the most widely used approaches for QC and trimming, respectively. As discussed above, QC is an essential step that guarantees the identification of potential artifacts in the data and helps to take appropriate steps to eliminate them, e.g., in the trimming step. Next, either trimmed or raw reads are used to identify expressed transcripts.

The choice of the method for transcriptome investigation in the selected publications was motivated mainly by the type of the explored host-pathogen system. Analyses involving organisms with comprehensive references and annotations, e.g., human, mouse, *E. coli*, and *S.* Typhimurium used popular reference-based tools — STAR (Dobin et al., 2013), TopHat (D. Kim et al., 2013; Trapnell et al., 2009), and Bowtie (Langmead et al., 2009; Langmead & Salzberg, 2012) — followed by the quantification of uniquely-mapped reads. HTSeq (Anders et al., 2015) was the preferred tool for counting such reads. Only a few studies included multi-mapped reads using a model-based approach implemented in RSEM (B. Li & Dewey, 2011; S. K. Buddenborg et al., 2017; Farrer et al., 2018; Mohamed et al., 2020). On top of that, quantification with alignment-free methods is unpopular among the dual RNA-seq analyses. Only a single study used Kallisto to map and quantify host transcripts solely (Bray et al. 2016; Thänert et al., 2019). In general, many dual RNA-seq analyses investigated the expression of known transcripts by applying methods that rely on annotations.

Several studies identified transcripts by adopting an align-then-assemble method (Aoki et al., 2019; Griesenauer et al., 2019; S. S. Kumar et al., 2018; Q. Liu et al., 2020; W. Li et al., 2019; C. Mavromatis et al., 2015; Pérez-Losada et al., 2015; Yazawa et al., 2013) following the Cufflinks protocol (Trapnell et al., 2012). Others used rnaSPAdes (Bushmanova et al., 2019;

Ritchie & Evans, 2019), Velvet (Zerbino & Birney, 2008; Wesolowska-Andersen et al., 2017), or SOAPdenovo2 (Luo et al., 2012; Fabozzi et al., 2018) to reconstruct pathogen transcripts from unmapped reads combined with read selection supported by pathogen-related databases. On the other hand, the absence of host references in studies on snail-parasite interactions (S. K. Buddenborg et al., 2017) and bacterial infection of fish (L. Huang et al., 2019) resulted in employing the Trinity *de novo* assembly tool (Haas et al., 2013; Grabherr et al., 2011) to identify host transcripts from the pool of unmapped reads to the pathogen reference. A complete set of sequencing reads was used to *de novo* assemble both host and pathogen transcripts of the Norway spruce tree and its pathogenic fungi that lack comprehensive references (Lundén et al., 2015). Overall, assembly methods allow studying non-model organisms, thus expanding the application of the dual RNA-seq technique — at least from the computational point of view.

Unfortunately, none of the reviewed studies used a customized computational workflow that integrates third-party tools into a single piece of software to analyze dual RNA-seq data. Although several studies employed programs that support the processing of sequencing reads, neither of these tools was developed explicitly for dual RNA-seq data, using a workflow management system that facilitates designing and executing multiple processes (in section 3.1.5 I elaborate more on this topic). For instance, READemption (Förstner et al., 2014), employed in two analyses (Westermann et al., 2016, 2019) is a tool with a command-line interface integrating several steps of RNA-seq data processing. Some studies used commercial software with a graphical interface — CLC Genomics Workbench — which also supports sequencing data analysis but works under an expensive license (Damron et al., 2016; Zimmermann et al., 2017; Ritchie & Evans, 2019; Doing et al., 2020; Schulte et al., 2020; Camilios-Neto et al., 2014; Musungu et al., 2020; Valenzuela-Miranda & Gallardo-Escárate, 2018). Likely, other analyses have been performed directly from the command line or using not published in-house scripts. Thus, it may be challenging to reproduce their results. Finally, this review has also indicated new challenges researchers face while analyzing dual RNA-seq data.

Table 3.1 Overview of selected dual RNA-seq studies. The table contains information on the pathogen and host used in a study, experiment design, software used to process sequencing reads, availability of code and data, and information on the software versions. The mapping procedure describes steps of the read mapping procedure applied in the corresponding study; "pat."- pathogen; p.i - post—infection; h – hours; min. – minuts; d. – days

Reference	pathogen	host	*in-vitro/ in-vivo*	time points	library type	quality control software	trimming software	mapping procedure	mapping software	quantification software	code availability	tools' version availability	raw data availability
(Humphrys et al. 2013)	*Chlamydia trachomatis*; obligate intracellular; G-	human HEp-2 cell line	*in-vitro*	1, 24 h p.i.	paired-end 100 bp	FastQC (Andrews S., 2010)	in-house quality control pipeline	1. mapping to pat. genome 2. mapping of unmapped reads to host genome 3. analysis of cross-mapped reads	pat.: Bowtie (Langmead et al. 2009) host: TopHat (Trapnell et al. 2009)	HTSeq (Anders et al. 2015)	no	yes	par-tial
(Baddal et al. 2015)	*Haemophilus influenzae;* intracellular; G-	primary normal human bronchial epithelial cells	*in-vitro*	1,6, 24,72 h p.i.	paired-end 75 bp reads			1. alignment to host genome 2. alignment of unmapped reads to pat. genome	host: STAR (Dobin et al. 2013) pat.: Rsubread (Liao et al. 2013)	host: HTSeq (Anders et al. 2015), pat.: featureCounts (Liao et al. 2014)	no	par-tial	yes
(Pérez-Losada et al. 2015)	microbiome community	human; nasal epithelial cell brushings from asthmatics and healthy people	*in-vivo*		single-end 100 bp reads		PRINSEQ-lite (Chen et al. 2018)	1. alignment to host genome 2. host transcriptome assembly 3. alignment of unmapped reads against the NCBI-NR protein reference database	alignment to host genome: Bowtie2 (Langmead and Salzberg 2012), TopHat2 (Kim et al. 2013) transcriptome assembly: Cufflinks2 (Trapnell et al. 2012) microbiome read alignment: DIAMOND (Buchfink et al. 2015)	host: Cufflinks2 (Roberts et al. 2011)	no	par-tial	yes
(Mavromatis et al. 2015)	*Escherichia coli*; intracellular; G-	Mouse bone marrow-derived macrophages (BMMs)	*in-vitro*	1, 2, 24 h p.i.	paired-end 75 bp reads			1. alignment to host and pat. genome 2. transcriptome assembly	TopHat	Cufflinks	no	no	no
(Westermann et al. 2016)	*Salmonella* Typhimurium*;* intracellular; G-	Human cervix carcinoma cells (HeLa-S3)	*in-vitro*	0,2,4, 8,16, 24 h p.i.	75 bp single-end	FastQC (Andrews S., 2010)	fastq_quality_trimmer; http://hannonlab.cshl.edu/fastx_toolkit/	1. parallel mapping to host. and pat. genomes 2. removal of cross-mapped reads	READemption pipeline (Förstner et al. 2014) with segemehl (Otto et al. 2014) and the remapper lack (Hoffmann et al. 2014)	READemption pipeline, uniquely mapped reads	yes	yes	yes

Table 3.1 continued

Reference	pathogen	host	*in-vitro/ in-vivo*	time points	library type	quality control software	trimming software	mapping procedure	mapping software	quantification software	code availability	tools' version availability	raw data availability
(Aprianto et al. 2016)	*Streptococcus pneumoniae*; extracellular; G+	Human type II lung epithelial cell line A549	*in-vitro*	30,60,120, 240 min. p.i.	single-end 75 bp reads	FastQC (Andrews S., 2010)	Trimmomatic (Bolger et al. 2014)	1. mapping to chimeric genome	STAR (Dobin et al. 2013)	featureCounts (Liao et al. 2014)	yes	no	yes
(Thänert et al., 2017)	*Staphylococcus aureus;* intracellular; G+	kidneys of infected mice	*in-vivo*	48 h p.i.	single-end 50 bp reads	fastq-mcf (Aronesty, 2013)		1. alignment of reads to both host and pathogen	STAR (Dobin et al. 2013)	HTSeq (Anders et al. 2015)	no	no	yes
(Nuss et al. 2017)	*Yersinia pseudotubercu-losis*; extracellular; G-	murine Peyer's patches	*in-vivo*	3 days p.i.	single-end 50 bp reads	FastQC (Andrews S., 2010)	fastx_trimmer http://hannonlab.cshl.edu/fastx_toolkit/	1. mapping to pat. genome 2. unmapped reads classified as host reads 3. aligning the pat.- classified reads back to the host genome 4. removal of cross-mapped reads 5. alignment of pat.-reads to pat. genome, and host-classified reads to host genome	bacteria: Bowtie2 (Langmead and Salzberg 2012) host: TopHat2 (Kim et al. 2013)	HTSeq (Anders et al. 2015)	no	no	yes
(Rossi et al. 2018)	*Pseudomonas aeruginosa*; extracellular; G-	CF sputum samples from chronically infected patients	*in-vivo*		single-end and paired-ends 75 bp reads		Trimmomatic (Bolger et al. 2014), SortMeRNA (Kopylova et al. 2012)	1.mapping to host genome 2. pan-genome analysis using unmapped reads	host: BWA bacteria: Roary (Page et al. 2015)	HTSeq (Anders et al. 2015)	yes	yes	yes
(Stapels et al. 2018)	*Salmonella Typhimurium*; intracellular; G-	mouse bone marrow–derived macrophages	*in-vitro*		single-end 75 bp		Trimmomatic (Bolger et al. 2014)	1. alignment of trimmed reads to pat. assembly 2. alignment of unmapped reads to host genome 3. alignment of trimmed reads to host genome 4. alignment of unmapped reads to pat. assembly	TopHat	HTSeq (Anders et al. 2015)	no	yes	yes

Table 3.1 continued

Reference	pathogen	host	*in-vitro/ in-vivo*	time points	library type	quality control software	trimming software	mapping procedure	mapping software	quantification software	code availability	tools' version availability	raw data availability
(Westermann et al. 2019)	*Salmonella* Typhimurium; intracellular; G-	Human cervix carcinoma HeLa cells	*in-vitro*	8, 16 h p.i.	single-end 75 bp reads	-	Cutadapt (Martin 2011)	1. parralel mapping to host and pathogen genomes	READemption pipeline (Förstner et al. 2014) with segemehl (Otto et al. 2014) and the remapper lack (Hoffmann et al. 2014)	READemption pipeline, uniquely mapped reads	yes	yes	yes
(Montoya et al. 2019)	*Mycobacterium leprae*; obligate intracellular; G+	patient leprosy skin biopsy specimens	*in-vivo*		single-end 50 bp		Trim Galore!	1. mapping to host genome 2. mapping of unmapped reads to pat. genome	STAR (Dobin et al. 2013)	HTSeq (Anders et al. 2015)	no	no	yes
(Griesenauer et al. 2019)	*Haemophilus ducreyi*; extracellular; G-	RNA isolated from wounds of human volunteers infected with	*in-vivo*		paired-end 75 bp reads	-	-	1. mapping to host and pat. genome	TopHat	Cufflinks	no	no	yes
(Thänert et al., 2019)	microbial community	human tissue biopsies	*in-vivo*		single-end 50 bp reads	-	-	1. mapping to host genome 2. metatranscriptome analysis of unmapped reads	Kallisto (Bray et al. 2016)		par-tial	no	yes
(Pisu et al. 2020)	*Mycobacterium tuberculosis*; intracellular	mice; two Mtb-infected macrophage subpopulations isolated from lungs	*in-vivo*	14 days		FastQC (Andrews S., 2010)	Flexbar (Roehr et al. 2017)	1. removal of rRNAs in Bowtie2 (Langmead and Salzberg 2012) 2. separation into species-specific fastq files in Bowtie2 3. alignment of reads to respective transcriptomes	Hisat2 (Kim et al. 2015)	HTSeq (Anders et al. 2015)	yes	yes	yes
(D'Mello et al. 2020)	*Streptococcus pneumonia;* extracellular; G+	Blood, kidneys, lungs, and nasopharynx from infected mice	*in-vivo*	7 d.p.i/ 2 d.p.i./ 30-36 h p.i.	paired-end 150bp reads			1. mapping to host and pathogen	host: HISAT pat.: Bowtie	HTSeq (Anders et al. 2015)	yes	no	yes

Table 3.1 continued

Reference	pathogen	host	*in-vitro/ in-vivo*	time points	library type	quality control software	trimming software	mapping procedure	mapping software	quantification software	code availability	tools' version availability	raw data availability
(Minhas et al. 2020)	*Streptococcus pneumoniae*; extracellular; G+	Lungs of infected mice	*in-vivo*	6 h p.i.	single-end 85 bp reads	FastQC (Andrews S., 2010)	Trimmomatic (Bolger et al. 2014)	1. alignment onto chimeric genome	STAR (Dobin et al. 2013)	featureCount (Liao et al. 2014)	no	yes	yes
(Kachroo et al. 2020)	*Streptococcus pyogenes;* extracellular; G+	nonhuman primate (NHP) skeletal muscle samples	*in-vivo*	24 h p.i.	single-end 75 bp reads	FastQC (Andrews S., 2010)	Trimmomatic (Bolger et al. 2014)	1. mapping to host genome 2. mapping of host-derived reads to pat. genome 3. removal of cross-mapped reads 4. mapping to pat. genome	host: STAR (Dobin et al. 2013) pathogen: EDGE-Pro http://ccb.jhu.edu/software/EDGE-pro/		no	no	yes
(Penaranda et al. 2021)	*Pseudomonas aeruginosa*; intracellular; G-	human bladder epithelial cells	*in-vitro*	2 h p.i.	paired-end reads			1. read alignment to pat. genome and host transcriptome	pat.: BWA host: BBMap	in-house script	no	no	yes
(Hayward et al. 2020)	*Chlamydia trachomatis*; obligate intracellular G-	HEp-2 epithelial cells	*in-vitro*	1, 24 h p.i.	paired-end 100 bp reads	FastQC, Trim Galore!, BEDtools, Bamtools	Trim Galore!	1. mapping to pat. genome 2. mapping to host genome 3. removal of cross-mapped reads	pat.: Bowtie2 (Langmead and Salzberg 2012) host: STAR (Dobin et al. 2013)	featureCounts (Liao et al. 2014)	yes	yes	yes

3.1.3 Challenges of dual RNA-seq data analysis

The adaptation of existing RNA-seq methods to simultaneous identification of transcriptomes from two interacting organisms has brought new challenges to the field (Westermann et al., 2017). Several aspects need to be considered while processing sequencing reads obtained from a dual RNA-seq experiment. First of all, the choice of strategy mostly depends on the organisms of interest. As presented in a previous section, studies on biological systems with known annotations give more flexibility; there is a possibility to take advantage of alignment-free methods. Their computational efficiency (Sahraeian et al., 2017) might be significant while working on more than one organism. On the other hand, visualization of alignment data provided by genome alignment-based methods enables the evaluation of the mapping results and investigation of unannotated regions. In particular, this strategy can be helpful in studies on non-model organisms with poor annotations. Exploring unannotated areas with high read coverage, one can identify novel transcripts or genomic regions expressed under the studied condition. Likewise, an application of one of the transcriptome assembly methods can improve the annotations. Since dual RNA-seq is often performed on organisms with different genomes, e.g. smaller bacterial and larger eukaryotic genomes, their key features need to be considered. For instance, for data that contains eukaryotic reads spanning exon-exon junctions, an accurate genomic alignment can only be performed with a mapper that allows for splice junction recognition. Thus, it is crucial to select a mapping method appropriate to the studied organisms.

When both host and pathogen references are available, the reads can be mapped against them. Marsh et al. (2018) suggested aligning reads separately — first to the host and then re-mapping unmapped fragments against the bacterial genome. This strategy was implemented in several host-pathogen analyses (Baddal et al., 2015; Farrer et al., 2018; LaMonte et al., 2019; Z.-X. Liao et al., 2019; Montoya et al., 2019; Orikaza et al., 2020). Only a few mapping procedures were performed onto genomes in the opposite order (Humphrys et al., 2013; Rienksma et al., 2015). Another approach involves parallel mapping of reads to both references, which allows the identification of cross-mapped reads aligned equally well to both organisms (Westermann et al., 2016, 2019; Aprianto et al., 2016; Choi et al., 2014; H. J. Lee et al., 2018; Minhas et al., 2020; Mohamed et al., 2020; Maulding et al., 2022). The last strategy seems to be the most accurate as mapping to the genomes separately may bias the read assignment in favor of the first reference (Espindula et al., 2019; Z. Liu et al., 2019). It means that reads originating from the second organism can map to the first one if the sequence similarities satisfy the alignment conditions. Therefore, more reliable gene expression estimates may be obtained using concatenated reference files into a chimeric genome or transcriptome as each read has access to the host and pathogen references at the same time during the mapping procedure.

The choice of quantification tool is another critical aspect. Eukaryotic and bacterial genomes and transcriptomes harbor repeat sequences that complicate the quantification step. For example,

eukaryotic mRNA isoforms produced from the same locus through alternative splicing may share some exons being a source of ambiguous reads. In addition, many genomic repeats are a consequence of recombination or transposition. Bacterial genomes also contain repetitive sequences, some to a very high extent as the genome of *Orientia tsutsugamushi* presented in the previous chapter. The expression of genes/transcripts that share high sequence similarity can be incorrectly estimated due to the removal of multi-mapped reads (Robert & Watson, 2015). Thus, applying a quantification tool that handles multi-mapped reads, we can investigate the importance of different gene isoforms in the eukaryotic host and repetitive elements in pathogen genomes. Therefore, selecting a tool that quantifies multi-mapped reads may be essential for accurate analysis in many dual RNA-seq applications.

Besides the challenges mentioned above, another aspect that should be considered is a detailed report of steps taken during the dual RNA-seq data analysis. Most of the published studies do not provide comprehensive information about applied strategies (see Table 3.1). Therefore, reproduction of their results would be difficult or even impossible, and poor reporting has become one of the sources of the reproducibility crisis in science.

3.1.4 Reproducibility in computational research

Reproducibility of a project is achieved if one can obtain the same outcome as in the original research using the same data and protocol. Unfortunately, recreating someone's results is challenging in the case of many studies, especially for novices with minimal expertise (Baker, 2016; Garijo et al., 2013). Today many computer-based projects are getting more complex, so it is essential to provide detailed descriptions of applied methods to be able to repeat them later. Sometimes, every tiny detail is crucial. For instance, variations solely in software versions or operating systems may lead to different outcomes in independently performed analyses (Y.M. Kim et al., 2018). Thus, in NGS-based projects, each piece of information on the technical aspect, e.g., parameter settings or the version of used references, may be crucial to evaluate publications' outcomes and draw conclusions from those that show conflicting results (Nekrutenko & Taylor, 2012). If employed techniques and procedures are incompletely described, the verification of findings from such studies may be complicated, cost a lot of effort, or even be impossible. Therefore, transparent reporting of applied methods and making both protocols and raw data publicly available is very important. However, it may still be insufficient to fully reproduce a study. Thus, the application of some tools that facilitate the reproducibility of bioinformatics analysis may be helpful.

Several approaches support both software development and transparency of the study. For instance, Version Control Systems (VCS) have gained high popularity to manage projects (repositories) and their accessibility. Such systems store files of the project and record changes made in them, allowing comparison of different versions and reverting to a specific state. Moreover,

Distributed Version Control Systems (DVCSs), e.g., Git, facilitate collaboration with other developers. Everyone involved in a project works independently on a copy of the repository and can merge provided changes with the central archive at any time (Chacon & Straub, 2014). Furthermore, web-based applications, such as GitHub (https://github.com/), provide open access to the whole repository enhancing the project transparency. Overall, Git has a great potential to support the reproducibility of research projects (Ram, 2013), but other technical aspects need to be also considered for full reproducibility.

Although it is easier to reproduce a study if details on used tools are provided, installing and maintaining both software and their dependencies on different platforms sometimes becomes problematic and consumes valuable time. Some technologies can facilitate software installation and usage. These are: virtual machines which create a virtual version of a computer system; and containers that provide an abstract operating system leveraging features of the user's operating system to isolate processes from the rest of the system. They encapsulate tools and their dependencies in an image package, which can be shared with others and executed on different computational environments, such as local computers, clouds, and clusters. While virtual machines also include the whole operating system, containers are dependent on the user's system, ultimately reducing the weight of their packages (Piccolo & Frampton, 2016). Widely used containers include Docker (Boettiger, 2015) and Singularity (Kurtzer et al., 2017), and effortless installation and high portability are the most significant advantages of these technologies.

All aspects mentioned above should become a standard in the age of complex research projects. A reliable code with a clean structure and accessible detailed documentation, e.g., through GitHub, helps to easily repeat someone's analysis. Together with the availability of data and an image package encapsulating tools, they may foster reproducibility. However, it may still be insufficient for performing bioinformatics analyses in an efficient and user-friendly manner.

3.1.5 Bioinformatics pipeline frameworks

The complexity of many studies involving data analysis emerges from a series of computationally intensive steps executed one after another. Thus, scalable pipelines have been gaining increased attention. They allow combining third-party tools and in-house scripts to process hundreds of millions of short RNA-seq reads and provide human-readable outputs. Since high-throughput technologies have become a standard method in many areas, user-friendly tools and data processing workflows are needed to handle large datasets.

A common approach to combine tools in a workflow involves a command-line environment and bash scripting. However, a primary limitation is the lack of structure that facilitates parallel computing or resume mechanisms. Fortunately, various frameworks have been developed to create pipelines for high-performance automated analysis, including Galaxy (Goecks et al., 2010),

Snakemake (Köster & Rahmann, 2012), Nextflow (Di Tommaso et al., 2017), Toil (Vivian et al., 2017) and many others (reviewed in Leipzig, 2017). The main difference between them is their design philosophy (Wratten et al., 2021; Ahmed et al., 2021), which mostly affects the workflow development step (Jackson et al., 2021). For instance, some workflow frameworks limit the further evolution of the pipelines, including an integration of new tools in intermediate steps; others support the execution of serial and parallel steps also in distributed architectures like clusters, offer the ability to integrate containers, and provide a resume mechanism for failed tasks.

Nextflow is one such workflow management system. Nextflow language simplifies the development of data-intensive computational pipelines. Scripts created in any language can be integrated and efficiently executed in a parallel manner on different platforms, including High-Performance Computing (HPC) clusters and clouds (Di Tommaso et al., 2017). Moreover, by supporting container technologies such as Singularity (Kurtzer et al., 2017) and Docker (Boettiger, 2015), Nextflow ensures reproducibility of results and platform independence. Additional integration with SingularityHub (https://singularity-hub.org/) or Docker Hub (https://hub.docker.com/) that automatically build image packages based on a recipe file containing software adopted in the pipeline, makes the containers easily accessible from different platforms. Also, hosting the pipeline on GitHub makes it widely available. In addition, the nf-core community project (P. A. Ewels et al., 2020) collects reviewed high-quality nextflow-based pipelines that fulfill development good-practice guidelines. The community provides appropriate tools and requirements that exploit Nextflow capabilities to build user-friendly standardized workflows. Thus, I have decided to develop the Dualrnaseq pipeline in Nextflow and integrate it into the nf-core project.

3.1.6 Choice of tools employed in the Dualrnaseq pipeline

The selection of an appropriate framework for workflow development is an important step as it should support developers and potential pipeline users. However, the major concern is the choice of data processing tools, because they influence the analysis results and define a range of possible workflow applications. In the Dualrnaseq pipeline, the quality of the sequencing reads is evaluated using one of the widely used QC tools - FastQC (Andrews S., 2010) followed by adapter and quality trimming performed by either Cutadapt (M. Martin, 2011) or BBDuk (B. Bushnell, 2020). While Cutadapt enables trimming of a single provided sequence, BBDuk uses a database of adapter sequences. For the next step of RNA-seq data analysis — mapping — both alignment-based and Selective-Alignment methods were incorporated to exploit their unique advantages in dual RNA-seq studies.

The alignment-based method is employed with STAR (Dobin et al., 2013), one of the splice-aware alignment tools characterized by high mapping accuracy coupled with computational efficiency (Baruzzo et al., 2017; Teng et al., 2016). These features are a result of

the STAR algorithm, which consists of two steps. First, the seed searching phase looks for the longest sequences of the reads, called Maximal Mappable Prefixes (MMPs), that exactly match one or more genomic regions of the reference. If a read cannot be contiguously mapped, after finding MMPs for the first part of the read, STAR searches for MMPs for the unmapped fragment. If exact matching sequences cannot be found because of mismatches or indels, the MMPs are extended. The different parts of a non-contiguously mapped read are called seeds, which are stitched together in the next phase. A read's final alignment is chosen based on an alignment score that considers penalties for mismatches, indels, and splice junction gaps. The sequential application of MMP search for different seeds allows for detecting splice junction regions and performing read mapping in a computationally efficient manner. Such efficient alignment is one of the features that may be relevant for analyzing complex data sets.

The alignment SAM/BAM file generated by STAR and containing information on the likely origin of reads is used to assign the reads and estimate gene abundance by another software. The most popular count-based tool for gene quantification, also applied in the workflow, is HTSeq (Anders et al., 2015; Deshpande et al., 2020). It counts the number of aligned reads overlapping exons or other user-specified features for each gene. It also gives several choices on how to handle reads that cover more than one feature. However, to avoid false positives for genes that share similar sequences, it is recommended to count only uniquely mapped reads, which may underestimate expression of many genes.

The second option for mapping implemented in the pipeline involves a very fast and accurate quantification method — Salmon with Selective-Alignment and a genome sequence defined as a decoy (Srivastava et al., 2020). The speed of mapping in this tool is ensured by indexing that involves data structures (suffix arrays and hash tables) allowing efficient item retrieval (Patro et al., 2017). The accuracy is an effect of the SA algorithm, consisting of several steps. First, the maximal exact matches (uni-MEMs) between the sequenced reads and index are collected. Next, for each read, a set of potential transcripts from which they might have originated is extracted. Only mappings compatible with a defined library type are selected. Finally, to resolve the position of a read along each transcript, the chaining score of Minimap2 (H. Li, 2018) is calculated for all possible mappings. The next phase involves the evaluation of the potential transcript candidates based on the optimal alignment score, which is also calculated for decoy sequences. If the best alignment score is higher for a decoy sequence than the annotated transcripts, the fragment’s mappings are discarded. All mappings that pass the filtration steps are subjected to quantification in the next phase.

As a transcriptome can be defined as a set of expressed transcripts and their frequencies at a given time, RNA-seq data enables one to estimate the relative expression level of transcripts in a sample. In Salmon, the relative transcript abundances are estimated using a probabilistic model. Given a nucleotide fraction (η) that represents all nucleotides in a sample originating from a transcript, the relative expression value of this transcript can be computed by normalizing η by

the transcript length (B. Li et al., 2010). Salmon seeks to infer η quantities from the sequencing data for a given set of transcripts using a two-phase inference procedure. Initial expression levels and parameters of sample-specific bias models are estimated in the online phase, then the expression values are refined in an offline step. In order to optimize memory and speed up the inference procedure, Salmon creates equivalence classes for fragments that map to the same set of transcripts.

Salmon SA is a relatively new approach. However, its fast mapping step combined with an improved algorithm for identifying read locations and model-based quantification is a promising alternative for alignment-free techniques. However, the main drawback of this approach is a lack of alignment file generation, limiting the evaluation of the mapping process. Fortunately, Salmon offers an alternative — alignment-based mode — which is also available through the Dualrnaseq workflow. In this method, Salmon performs quantification using alignments generated by other tools. It calculates conditional probabilities for alignments, which help estimate the probability of fragments originating from transcripts. In the Dualrnaseq pipeline, a transcriptome BAM file produced by STAR is used. In this way, Salmon alignment-based mode completes the set of mapping strategies that can be desirable in many applications of dual RNA-seq.

3.2 Materials and Methods

The Dualrnaseq pipeline was developed by me at the beginning. Regan Hayward (Helmholtz Institute for RNA-based Infection Research, Würzburg, Germany) helped me establish the final version of the workflow that fulfills the requirements of the nf-core community, and now the pipeline is accessible from their repository https://nf-co.re/dualrnaseq. Regan Hayward is involved in further development of the pipeline and benchmark analysis, though his work is not presented here. The work was performed under the supervision of Lars Barquirst (Helmholtz Institute for RNA-based Infection Research and University of Würzburg, Germany).

This section describes: (i) data sets used for developing and testing the pipeline (3.2.1); (ii) information about the Dualrnaseq structure, applied third-party tools and their parameter settings (3.2.2); and the tools used to benchmark the methods implemented in the workflow (3.2.3).

3.2.1 Data sets

3.2.1.1 Dual RNA-seq reads

The pipeline was developed and tested using the following dual RNA-seq data sets: *S.* Typhimurium infecting HeLa cells (Westermann et al., 2016), *Streptococcus pneumoniae* infecting Human lung epithelial cells (Aprianto et al., 2016), *Mycobacterium leprae* isolated with patient leprosy skin

biopsies (Montoya et al., 2019), and two *Ot* strains infecting human endothelial cells (Mika-Gospodorz et al., 2020). Further evaluation of the implemented methods in the workflow was performed using *S.* Typhimurium - HeLa and *Ot* - HUVEC data sets.

3.2.1.2 References and annotation files

The testing data comes from studies which were either performed on patient-acquired samples or human-derived cell lines. Therefore, the host reference (human genome fasta file) and annotations (comprehensive gene annotation covering all regions and tRNA gff file) were obtained from the GENCODE project (Frankish et al., 2019) from GRCh38.p13 reference assembly.

The *S.* Typhimurium fasta and annotation gff files were acquired from NCBI using the following accession numbers: FQ312003 for the genome, and HE654724, HE654725, HE654726 for the plasmids. The genome and plasmid gff files were combined into a single file. Further modifications of the bacterial gff file included replacing the original ncRNA annotations with a list of in-house ncRNA annotations, removal of annotations containing "gbkey=tRNA" and "gbkey=rRNA," and the addition of tRNA and rRNA annotations generated from the GenBank file downloaded under the FQ312003.1 accession number.

References of the two *Ot* strains were obtained from the NCBI under the LS398547.1 and LS398548.1 accession numbers for UT176 and Karp, respectively. Additionally, the gff files were supplemented with ncRNAs annotations identified by us (Mika-Gospodorz et al., 2020).

Other bacterial genome fasta and annotation gff files were obtained following the authors' description (Aprianto et al., 2016; Montoya et al., 2019).

3.2.2 Implementation and content of the Dualrnaseq pipeline

The Dualrnaseq pipeline was implemented in Nextflow (v20.10.0.5430) and integrated into the nf-core community. Therefore, the workflow was built using the nf-core standardized pipeline template, which provides the initial nextflow script, configuration files, documentation, and environment files. The project is also integrated with Git and hosted on the nf-core GitHub repository (https://github.com/nf-core/dualrnaseq).

The pipeline consists of the main.nf nextflow script describing processes and data workflow, scripts supporting the main workflow stored in the bin folder, configuration files specifying parameters and an environment file promoting container integration. All files and settings described here come from the first release of the nf-core/dualrnaseq pipeline (1.0.0) called Dualrnaseq here.

3.2.2.1 Test data set

Nf-core pipelines contain a pre-defined test profile for automated testing of the workflows using parameters specified in a *test.config* file (Figure S9B) and a test data set stored on the nf-core/test-datasets Github project. The Dualrnaseq test data set consists of Human (GRCh38.p13) and *S.* Typhimurium (FQ312003) genome fasta and gff annotation subsets and simulated paired-end reads.

The test sequencing reads for the Dualrnaseq pipeline were generated in R using the *simulate_experiment()* function provided by the Polyester package (v1.24.0) (Frazee et al, 2021). As an input, the function accepted a chimeric transcriptome generated from Human and *Salmonella* subsets of the genomes and annotation files using an in-house script. Following the read simulation, fasta files with read sequences were converted into FASTQ files using reformat.sh script from the BBTools package (v38.79) (Bushnell, B., 2014).

The test paired-end reads, the reference files, and the code used to simulate the reads are stored on the nf-core/test-datasets repository under the *dualrnaseq* branch https://github.com/nf-core/test-datasets/tree/dualrnaseq . This data set can be utilized under the test profile (*-profile test*) of the Dualrnaseq pipeline, which may be helpful for both developers to evaluate the performance of the pipeline and users to establish the workflow and environment before running the analysis on an actual data set. There is also a possibility to run the pipeline with the full test data set (*-profile test_full*) that contains RNA-seq data of HUVEC cells infected with the *Ot* strain Karp (Mika-Gospodorz et al., 2020).

3.2.2.2 Configuration Files

The default parameters of the pipeline were established using different host-pathogen systems (see 3.2.1), and those that can be customized are defined in configuration files. Options available for all environments are specified in the *nextflow.config* file (Figure S8). *Base.config* (Figure S9A) provides additional parameters for high-performance computing environments, whereas settings for testing are stored in the *test.config* (Figure S9B) and *test_full.config*. The last configuration file, *genomes.config* (Figure S9C), provides directories for references and annotations of different organisms. All these parameters can be overwritten through the command line or direct modification of the configuration files.

3.2.2.3 Environment and containerization

The pipeline uses fixed versions of third-party tools and packages obtained from software distributions like conda-forge (https://conda-forge.org/) and Bioconda (https://bioconda.github.io/). All tools employed within the pipeline are listed in the environment.yml file (Figure 3.1A). These

packages and their dependencies are installed through the Conda package management system (https://docs.conda.io/en/latest/), which builds an environment within a Docker image following the instructions described in the recipe file (Figure 3.1B). Moreover, the Dualrnaseq pipeline is connected to Docker Hub, a cloud-based repository that builds a container using the recipe file (Figure 3.1B) stored on the nf-core/dualrnaseq Github project (https://hub.docker.com/r/nfcore/dualrnaseq). Nextflow also supports Singularity container technology; the pipeline converts the Docker image into the Singularity image. Conda is the third option that can be used to create an environment for tools from the environment.yml file, however, it is not recommended due to poor reproducibility compared to container technologies.

Only one of the aforementioned technologies can be specified using the *-profile* option (*-profile <docker,singularity,conda>*) while running the pipeline. Although all of them automatically install tools used by the pipeline, the workflow was developed using Singularity. In contrast to Docker, Singularity can be executed as a non-root user, which is useful in HPC environments.

A. environment.yml

```
# This file contains all of the dependencies to setup the dualrnaseq pipeline
# It can also be used to setup an environment to replicate the pipeline (for testing, local analysis etc)
# To do this, use the following command: conda env create -f environment.yml
name: nf-core-dualrnaseq-1.0.0
channels:
  - conda-forge
  - bioconda
  - defaults
dependencies:
  - bioconda::fastqc=0.11.9
  - bioconda::samtools=1.9
  - bioconda::salmon=1.3.0
  - bioconda::STAR=2.7.3a
  - bioconda::gffread=0.12.1
  - bioconda::multiqc=1.8
  - bioconda::htseq=0.12.4
  - bioconda::bbmap=38.87
  - bioconda::cutadapt=3.2
  - bioconda::pysam=0.15.3
  - bioconda::bioconductor-rtracklayer=1.48.0
  - bioconda::bioconductor-tximport=1.16.0
  - conda-forge::python=3.7.6
  - conda-forge::r-plyr=1.8.6
  - conda-forge::markdown=3.1.1
  - conda-forge::pymdown-extensions=6.0
  - conda-forge::numpy=1.19.5
  - conda-forge::pandas=1.1.5
  - conda-forge::biopython=1.78
  - conda-forge::seaborn=0.10.0
  - conda-forge::matplotlib=3.1.1
```

B. Dockerfile

```
FROM nfcore/base:1.12.1

LABEL authors="Bozena Mika-Gospodorz, Regan Hayward" \
      description="Docker image containing all software requirements for the nf-core/dualrnaseq pipeline"

# Install the conda environment
COPY environment.yml /
RUN conda env create --quiet -f /environment.yml && conda clean -a

# Add conda installation dir to PATH (instead of doing 'conda activate')
ENV PATH /opt/conda/envs/nf-core-dualrnaseq-1.0.0/bin:$PATH

# Dump the details of the installed packages to a file for posterity
RUN conda env export --name nf-core-dualrnaseq-1.0.0 > nf-core-dualrnaseq-1.0.0.yml

# Instruct R processes to use these empty files instead of clashing with a local version
RUN touch .Rprofile
RUN touch .Renviron
```

Figure 3.1 Tools and environment setup of the Dualrnaseq pipeline. A) Environment file B) Docker recipe file.

3.2.2.4 Quality control and trimming

Within the pipeline, the quality control of reads is performed by FastQC (v0.11.9) (Andrews S., 2010) allowing for parallel computing through *--threads* parameter, suppressing progress messages on stdout with the *--quiet* option and preventing uncompression of output files by the *--noextract* flag.

Adapter sequences from the reads can be removed using Cutadapt (v3.2) (M. Martin, 2011). By default, the Dualrnaseq pipeline trims TruSeq Illumina adapter sequences with the *-a* option for the regular 3' adapter set to "*AGATCGGAAGAGCACACGTCTGAACTCCAGTCA*", and the *-A* flag, for paired-reads, defined as "*AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT*". It also performs quality trimming through the *-q* option, set to 10 in the pipeline (the *--quality_cutoff*

parameter of the Dualrnaseq workflow). The *-j* option ensures parallel computing and the *-m* flag set to 1 prevents writing empty reads to the output.

In addition, the Dualrnaseq pipeline provides the BBDuk trimming tool, which is a part of the BBtools package (v38.87) (B. Bushnell, 2020). The list of adapter sequences provided with the package is enclosed with the pipeline in the assets/adapters.fa file (*--adapters* flag is by default set to "*projectDir/assets/adapters.fa*"). For BBDuk, the memory usage is restricted to 1 GB with the *-Xmx1g* option defined within the pipeline. Additional parameters specified by the Dualrnaseq workflow include *minlen* set to 18, which removes reads shorter than this value after trimming and *qtrim* defined as *r* to trim the right ends of reads with quality below 10 according to the *trimq* option. In addition, the following parameters concern adapter trimming: *ktrim* is set to *r*, *k* to 17, *mink* to 11, and *hdist* is defined as 1.

3.2.2.5 Mapping and quantification

The Dualrnaseq pipeline can be used with the STAR alignment tool (v2.7.3a) (Dobin et al., 2013) in two modes, quantifying uniquely mapped reads with HTSeq (Anders et al., 2015) or also estimating multi-mapped reads with Salmon (Patro et al., 2017).

Both approaches use a chimeric genome created by concatenating host and pathogen fasta files for indexing. Since only the human genome gff annotation file is passed to the STAR index step in the pipeline, the *--sjdbGTFfeatureExon* option is set to *exon*, and *--sjdbGTFtagExonParentTranscript* to *Parent* by default. Another common parameter for indexing is *--sjdbOverhang*, which defines the sequence length on each side of the annotated junctions used in constructing the splice junctions database, and it is set to 100.

The alignment procedure is performed with the recommended ENCODE options (See STAR documentation for release 2.7.3a stored with the source code on the GitHub https://github.com/alexdobin/STAR/releases/tag/2.7.3a). In contrast to the ENCODE options, the Dualrnaseq pipeline does not restrict multi-mappings by switching off filters for a maximum number of both multiple alignments and loci anchors (*--outFilterMultimapNmax* and *--winAnchorMultimapNmax* are set to 999). Tests on the *Mycobacterium leprae* data set showed a higher mapping rate for this non-laboratory bacteria strain when filtering the number of mismatches relative to read length (*--outFilterMismatchNoverReadLmax*) was specified to 1. While the ENCODE project (Djebali et al., 2012) was performed on the cell lines whose RNA sequences may not vary significantly from the references, discarding this particular filter may be relevant for clinical samples rich in sequence variants. Moreover, in Dualrnaseq, STAR generates a SAM file that provides standard attributes (*--outSAMattributes Standard*) and contains unmapped reads (*--outSAMunmapped Within*). Also, the workflow does not limit RAM for sorting through the *--limitBAMsortRAM* option defined as 0, and if applicable, it runs STAR in parallel.

The following default parameters specified in the Dualrnaseq pipeline differ between the STAR-HTSeq and STAR-Salmon modes. In the strategy with quantification of uniquely mapped reads, the Dualrnaseq pipeline provides additional options for alignment with STAR, including *--outWigType* set to *None* and *--outWigStrand* defined as *Stranded*, by default. In addition, in the mode with HTSeq, an alignment file is sorted by coordinates (*--outSAMtype BAM SortedByCoordinate*). In the Salmon mode, STAR generates an unsorted BAM file (*--outSAMtype BAM Unsorted*) and converts it into transcript coordinates (*--quantMode TranscriptomeSAM*), allowing for insertions, deletions and soft-clips (*--quantTranscriptomeBan Singleend*). In the alignment phase, STAR also utilizes gff files. Thus, in the STAR-HTSeq strategy, Dualrnaseq transfers the same gff used in the indexing step to the alignment process and defines the following options: *--sjdbGTFfeatureExon* as *exon* and *--sjdbGTFtagExonParentTranscript* as *Parent*. For Salmon, STAR translates alignments into transcriptome coordinates, based on a chimeric reference file which contains both *quant* in place of gene features to be processed by the pipeline (*--sjdbGTFfeatureExon quant*) and *parent* as a gene attribute (*--sjdbGTFtagExonParentTranscript parent*).

Quantification of uniquely mapped reads is performed with HTSeq (v0.12.4) (Anders et al., 2015), which processes BAM files (*-f bam*) sorted by alignment position (*-r pos*). The quantification results are estimated for transcripts defined in the chimeric gff file with the *quant* gene feature (*-t quant*). By default, these are exons and tRNAs for host (*--gene_feature_gff_to_quantify_host "[exon, tRNA]"*) as well as genes, sRNAs, tRNAs and rRNAs for the pathogen (*--gene_feature_gff_to_quantify_pathogen "[gene, sRNA, tRNA, rRNA]"*), in the workflow. The counts are further aggregated into gene-level estimates using an attribute defined through *--host_gff_attribute* set to *gene_id*. In the Dualrnaseq pipeline, the pathogen attribute (*--pathogen_gff_attribute locus_tag*) is replaced by the host attribute for the chimeric gff file for consistency. In addition, HTSeq expects stranded library type (*-s, --stranded yes*), removes reads with alignment quality lower than 10 (*-a, --minaqual* 10), and limits the number of maximum reads staying in memory until the mates are found to be 30000000 (*--max-reads-in-buffer* option in HTSeq, *--max_reads_in_buffer* 30000000). Also, the quantification of uniquely mapped reads within Dualrnaseq can be performed in parallel (*-n*).

Another tool employed for mapping and quantification in the Dualrnaseq pipeline is Salmon (v1.3.0) (Patro et al., 2017), which uses the decoy-aware transcriptome index in Selective-Alignment (SA) mode (*--validateMappings*) or the alignment file obtained from STAR. In each of the modes, Salmon considers mappings or alignments that are compatible with the library type (*--incompatPrior* 0.0) specified by a user through the *--libtype* parameter of the Dualrnaseq pipeline (*-l* option in Salmon). In addition, Salmon creates an index of the decoy-aware transcriptome in SA mode, consisting of the chimeric transcriptome and host genome provided as a decoy. The pipeline passes the k-mer size set to 21 (*--kmer_length*) to the salmon index parameters, as the application of

k-mers of length 21 has shown a better accuracy in quantifying total RNA, including sRNA transcripts (D. C. Wu et al., 2018). Finally, the host transcript-level estimates obtained from Salmon in both modes are aggregated into gene-level estimates using Tximport (v1.16.0) (Soneson et al., 2016).

3.2.2.6 Data processing

The pipeline also consists of scripts written in R (v4.0.2), Python (v3.7.6), and Bash for processing different files and data sets. While the Pandas (1.1.5) (McKinney, 2010) python package is used to collect and manipulate different data sets, the Biopython (v1.78) (Cock et al., 2009) package is applied to handle fasta files. Furthermore, cross-mapped reads from BAM files are extracted using the Pysam (v0.15.3) python package and Samtools (v1.9) (Danecek et al., 2021). Applied R packages include Rtracklayer (v1.48.0) (Lawrence et al., 2009) that handles gff annotation files and extracts gene lengths for calculating TPM values from HTSeq counts. In addition, Gffread (v0.12.1) is employed to generate a transcriptome from a host genome gff file.

3.2.2.7 Summary statistics and reports

The workflow generates statistics and visualizes the results using third-party software or scripts written in Python. For example, summary reports of all processes are produced by MultiQC (v.1.8) (P. Ewels et al., 2016), and html reports are generated using the Markdown (v3.1.1) package. The other Python packages, including Pandas (1.1.5) (McKinney, 2010), NumPy (v1.19.5) (Harris et al., 2020), and SciPy (v1.5.3) (Virtanen et al, 2020), are employed to collect data, transform them and calculate statistics of mapping, RNA classes, and Pearson correlation coefficients, whereas Matplotlib (3.1.1) (Hunter, 2007) and Seaborn (0.10.0) (Waskom, 2021) are used for plotting the statistics.

3.2.3 Benchmark analysis

3.2.3.1 Read simulation

The evaluation of the performance of each mapping and quantification strategy implemented in the pipeline was performed using a simulation-based approach. For the dual RNA-seq data sets, including *Salmonella* - HeLa dual RNA-seq data (Westermann et al., 2016) and HUVEC cells infected with either *Ot* strain (Mika-Gospodorz et al., 2020), the quantification tables were generated in Salmon alignment-based mode. Next, utilizing count tables for each sample, and a chimeric host-pathogen transcriptome created by the Dualrnaseq pipeline, the reads that serve as a ground truth were simulated using the *simulate_experiment_countmat()* function of the Polyester package

(v1.24.0) (Frazee et al, 2021). The conversion of the obtained fasta files into FASTQ files was performed using the reformat.sh script from the BBTools package (v38.79) (Bushnell, B., 2014).

3.2.3.2 Benchmark analysis

The quantification statistics plot shown in Figure 3.4 was generated in python (v3.8.8) using the following packages: Pandas (v1.2.4) (McKinney, 2010), NumPy (v1.20.1) (Harris et al., 2020), and altair (v4.1.0) (VanderPlas et al., 2018). The correlation analysis between the number of reads' estimates and the ground truth was computed with *stats.spearmanr()* function of the SciPy (v1.6.2) (Virtanen et al, 2020) python package and visualized (Figure 3.5) using Pandas (v1.2.4), NumPy (v1.20.1), Matplotlib (v3.3.4) (Hunter, 2007), and Seaborn (v0.11.1) (Waskom, 2021) packages.

3.3 Results

3.3.1 Dualrnaseq: a Nextflow-based pipeline

Dualrnaseq is a robust, user-friendly pipeline created in Nextflow and comprises an assembly of different tools and scripts to process dual RNA-seq data. It performs basic steps of RNA-seq data analysis for both host and pathogen simultaneously (Figure 3.2A). After the QC and trimming, the sequencing reads are mapped onto two different references, the bacterial and eukaryotic genomes or transcriptomes, followed by quantifying either uniquely mapped reads alone or together with multi-mapped reads. Regardless of the chosen strategy, simultaneous processing of data from two different organisms requires adaptation of the annotation and reference files.

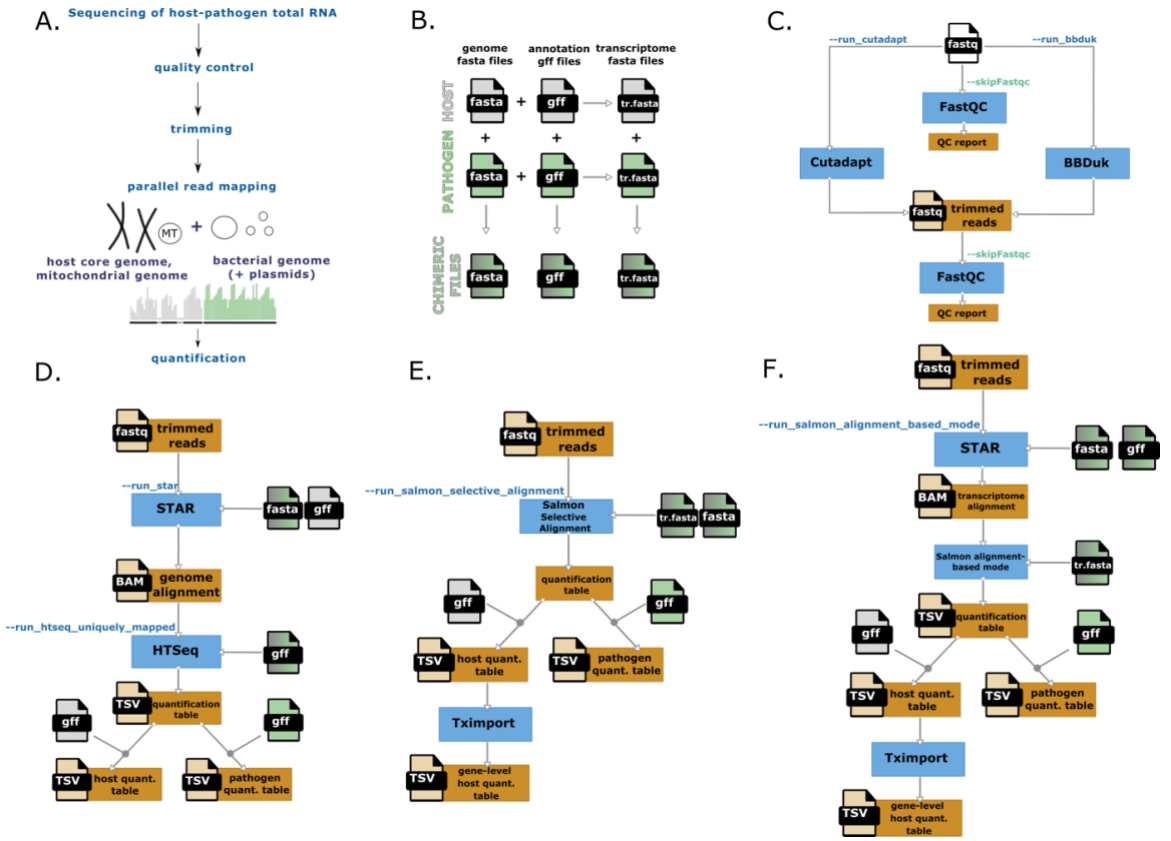

Figure 3.2 Steps of the Dualrnaseq pipeline. A) Scheme of the dual RNA-seq data analysis (Westermann et al., 2017). B) Host and pathogen reference and annotation files used in the pipeline to create chimeric files utilized further by various mapping and quantification strategies implemented in the Dualrnaseq pipeline. C) QC and trimming step of the workflow. D) Illustration of the STAR-HTSeq mode implemented in Dualrnaseq. E) Mapping and quantification performed with Salmon in SA mode. F) Salmon alignment-based strategy utilizing transcriptome BAM file created by STAR. In figures C-F, blue boxes represent third-party tools employed in the pipeline; dark yellow boxes illustrate outputs of the tools; blue command-line flags represent options that must be specified to run a tool; green command switches off a tool that is executed within the pipeline by default.

3.3.1.1 Reference and annotation files

In the Dualrnaseq pipeline, host and pathogen RNA-seq reads are processed together. Therefore, the workflow creates chimeric references and annotations by concatenating fasta and gff files of two organisms (Figure 3.2B). Thus, both genome fasta and annotation gff files must be provided as input using the following parameters: *--fasta_host*, *--fasta_pathogen*, *--gff_host*, *--gff_pathogen*, and optionally *--gff_host_tRNA*. However, instead of defining the paths for each file on the command line separately, they can also be provided in the *genomes.config* (Figure S9C). In addition, the pipeline standardizes the annotation files coming from two different organisms. Gene features specified through *--gene_feature_gff_to_quantify_host* (default: *exon, tRNA*) and *--gene_feature_gff_to_quantify_pathogen* (default: *gene, sRNA, tRNA, rRNA*) are replaced with a *quant* feature in the gff files and only those elements are further quantified. In addition, the pathogen gene attribute (*--pathogen_gff_attribute*, default: *locus_tag*) is replaced by the host attribute defined by *--host_gff_attribute* (default: *gene_id*) to keep the chimeric gff file consistent.

The annotation file is also necessary for the Salmon mode to generate transcriptome fasta files. The sequences are extracted from the genome fasta file based on gff coordinates of entries defined by *--gene_feature_gff_to_create_transcriptome_pathogen* (default: *gene, sRNA, tRNA, rRNA*) and *--gene_feature_gff_to_create_transcriptome_host* (default: *exon, tRNA*). The name of each transcript, on the other hand, is extracted from the gff attributes specified with *--gene_attribute_gff_to_create_transcriptome_pathogen* (default: *locus_tag*), and *--gene_attribute_gff_to_create_transcriptome_host* (default: *transcript_id*). However, Dualrnaseq also provides the possibility to read custom transcriptome files by setting *--read_transcriptome_fasta_host_from_file*, and *--read_transcriptome_fasta_pathogen_from_file* options, and giving paths to the files through *--transcriptome_host* and *--transcriptome_pathogen* flags or *genomes.config*. Such prepared chimeric references and annotations are employed in further steps of the dual RNA-seq data analysis.

3.3.1.2 Quality control and trimming

The Dualrnaseq pipeline can process both compressed and uncompressed FASTQ files containing raw sequencing reads. In the first step, the QC of those files with FastQC is performed by default (Figure 3.2C) (Andrews S., 2010). For each sample, FastQC produces a Quality Control report containing statistics summarizing the analyzed files, evaluation of read quality, GC content, and presence of overrepresented sequences. Additional FastQC-related parameters (3.2.2.4) can be specified by the *--fastqc_params* option. If desired, the whole process can be omitted through the *--skipFastqc* flag.

Simultaneously with the QC, adapter trimming and quality read trimming can be performed using either Cutadapt (M. Martin, 2011) or BBDuk (B. Bushnell, 2020) (Figure 3.2C). In addition to default parameters defined in the Dualrnaseq pipeline (3.2.2.4), tool-specific options can be specified by the *--cutadapt_params* or *--BBDuk_params* flags, respectively. After read trimming, the QC reports are generated for new FASTQ files unless the *--skipFastqc* flag is defined, and reads are submitted for mapping and quantification.

3.3.1.3 Mapping and quantification

Dualrnaseq provides three strategies to map and quantify dual RNA-seq data. The first option is alignment-based mapping of reads onto the chimeric genome with STAR (Dobin et al., 2013), followed by counting of uniquely mapped reads with HTSeq (Anders et al., 2015) (Figure 3.2D). For the second fast transcriptome quantification method that handles multi-mapped reads — Salmon with Selective-Alignment is employed (Figure 3.2E). In the third strategy, a transcriptome alignment is created with STAR and used by Salmon to estimate transcript abundance (Figure 3.2F).

To perform mapping and quantification in STAR-HTSeq mode, the *--run_star* and *--run_htseq_uniquely_mapped* options need to be specified (Figure 3.2D). In this strategy, except from the chimeric genome references and annotation files that serve as inputs for each of the tool, human genome gff annotations are passed to STAR to be used to identify and correctly map spliced alignments across annotated splice junctions (Dobin & Gingeras, 2015). Also, beside default options set in the workflow (3.2.2.5), other options can be defined for the STAR indexing process and its alignment step using *--star_index_params* and *--star_alignment_params* flags, respectively. Additional HTSeq parameters can be specified with *--htseq_params*. Following the quantification performed by HTSeq, each resulting count table is split into two matrices based on organism-specific genes. Next, the pipeline provides TPM (transcripts per million) values obtained by normalizing the gene counts for gene length (the length of the gene exon union by default) and sequencing depth, to allow for a between-sample comparison of the gene abundances.

Another mapping and quantification strategy implemented in the pipeline involves the Salmon with Selective-Alignment algorithm (*--run_salmon_selective_alignment*), which maps reads to an index consisting of the chimeric transcriptome and the host genome (Figure 3.2E). The latter sequence is marked as a decoy, so it is handled differently during alignment scoring (for more details see 3.1.6) (Srivastava et al., 2020). The pathogen genome is not defined as a decoy, because the lack of precise coordinates of bacterial transcripts in the transcriptome leads to mapping of reads to the decoy and filtering them out from the analysis (investigation performed by Regan Hayward; data not shown here). Next, the Salmon quantification provides transcript-level abundance estimates in both TPM and Number of reads units. After splitting the resulting table into the host and pathogen quantification files, gene-level estimates for the host are computed with tximport (Soneson et al., 2016). This tool takes advantage of the transcript-level estimates to compute gene-level abundances and provides average transcript length correction terms (offsets). The offsets can be used in the subsequent analysis to account for differences in isoform usage between the samples. Salmon-specific parameters for indexing and mapping can be defined using *--salmon_sa_params_index*, and *--salmon_sa_params_mapping*, respectively.

Salmon can also perform quantification utilizing alignments in either BAM or SAM format generated with any preferable transcript alignment tool. In the Dualrnaseq pipeline, STAR alignment files are translated into transcript coordinates and used as an input for Salmon (Figure 3.2F). This option can be activated by setting both *--run_salmon_alignment_based_mode* and *--run_star*. Additional tool-specific parameters can be specified through *--STAR_salmon_index_params* (for the STAR index process), *--STAR_salmon_alignment_params* (for STAR alignment), and *--salmon_alignment_based_params* (for Salmon). As STAR returns both genome and transcriptome BAM files, evaluation of unannotated genome regions is possible in this mode. In addition, the assessment of performed processes can also be performed by analyzing summary statistics generated by the pipeline for each of the three mapping and quantification strategies.

3.3.1.4 Mapping and quantification statistics

After running all processes, a comprehensive summary report is generated with MultiQC (P. Ewels et al., 2016). In addition, the Dualrnaseq pipeline also generates statistics and plots for host-pathogen data in both tabular and graphical form. The set of graphics may help evaluate the data quality and the results. For instance, the reproducibility among replicates can be assessed based on scatter plots generated for TPM estimates between replicates (Figure 3.3A). Importantly, the number of reads processed in the trimming, mapping and quantification steps can be also evaluated visually (Figure 3.3B). In addition, the pipeline provides a summary of RNA class composition in both host and pathogen (Figure 3.3C). Evaluation of the biotype composition may be investigated as a quality control of the RNA-seq experiment, especially the RNA purification step. Although such outputs are necessary for basic evaluation of the data quality, these results might also be helpful for investigating differences between outcomes from various mapping and quantification methods.

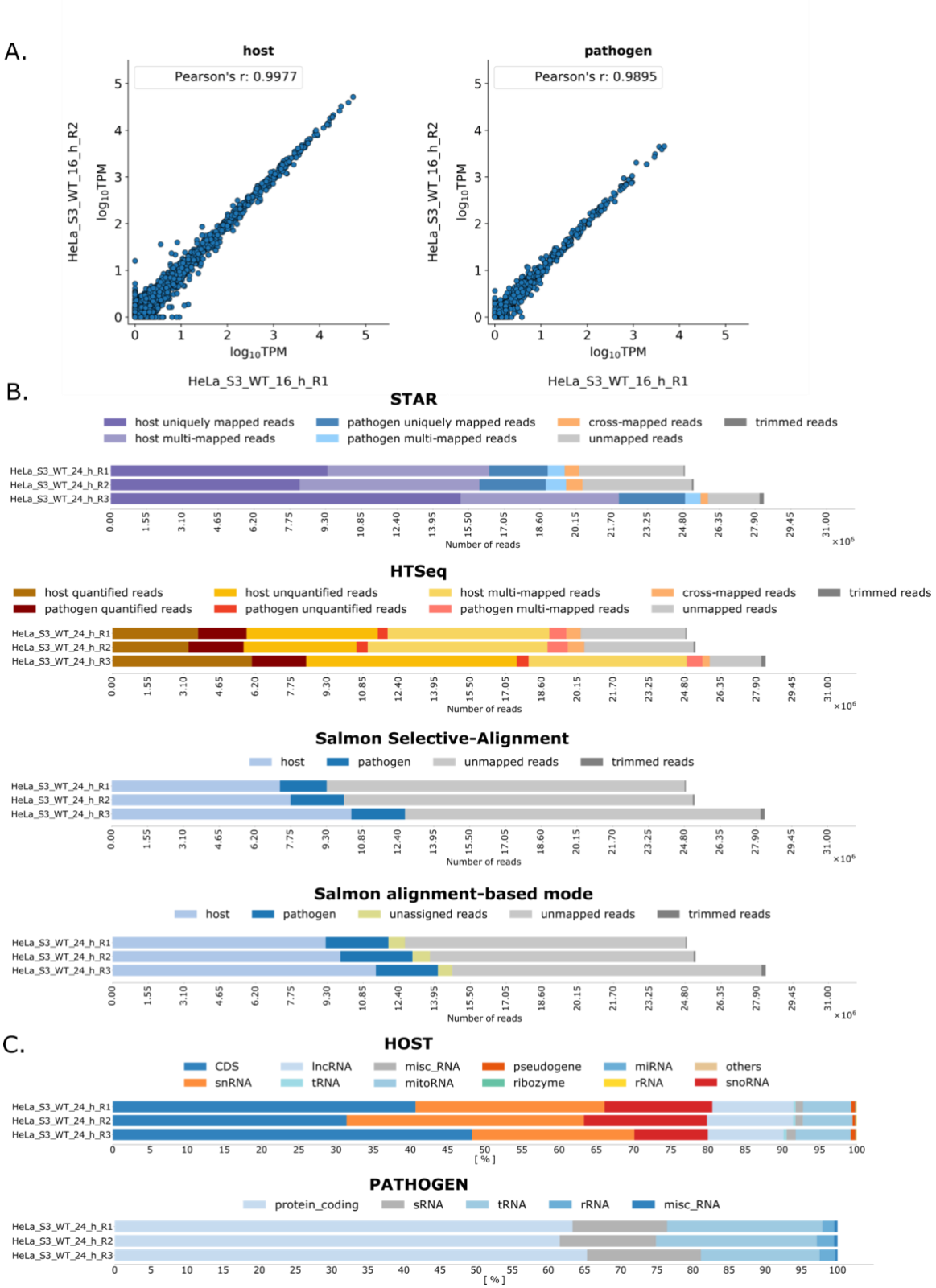

Figure 3.3 Mapping and quantification statistics generated by the Dualrnaseq workflow. A) Example of scatter plots of gene expression values (TPM) derived from RNA-seq of the host and pathogen. The Pearson correlation coefficient is calculated for untransformed TPM values B) Summary statistics of the following: mapping to the genome with STAR; both mapping performed with STAR and counting with HTSeq; quantification results of Salmon in SA mode; results of both transcriptome mapping performed with STAR and quantification with Salmon. C) RNA class statistics plots generated for the host and pathogen.

3.3.2 Benchmark analysis of the mapping and quantification methods

The evaluation of the methods employed in the pipeline was performed on the data sets generated from human cells infected with either *S.* Typhimurium or two *Ot* strains (for details see 3.2.1 section). The initial analysis of the number of reads assigned to either host or pathogen indicated differences in the performance of the three mapping-quantification strategies (Figure 3.4). Overall, a significantly lower number of reads was assigned onto the host references in the STAR-HTSeq strategy. The results also indicate a slightly higher number of reads quantified with the alignment-based mode of Salmon than Selective-Alignment, except for Karp and UT176 samples. Although the number of reads assigned to *S.* Typhimurium is increasing over time reflecting the intracellular replication of the pathogen, their abundance varies between the three strategies with the highest number of reads assigned to *Salmonella* transcripts in STAR-Salmon mode; opposite to *Ot* samples. However, investigating the accuracy of obtained outcomes requires the ground truth that represents a real quantity of known sequences among the sequenced transcripts.

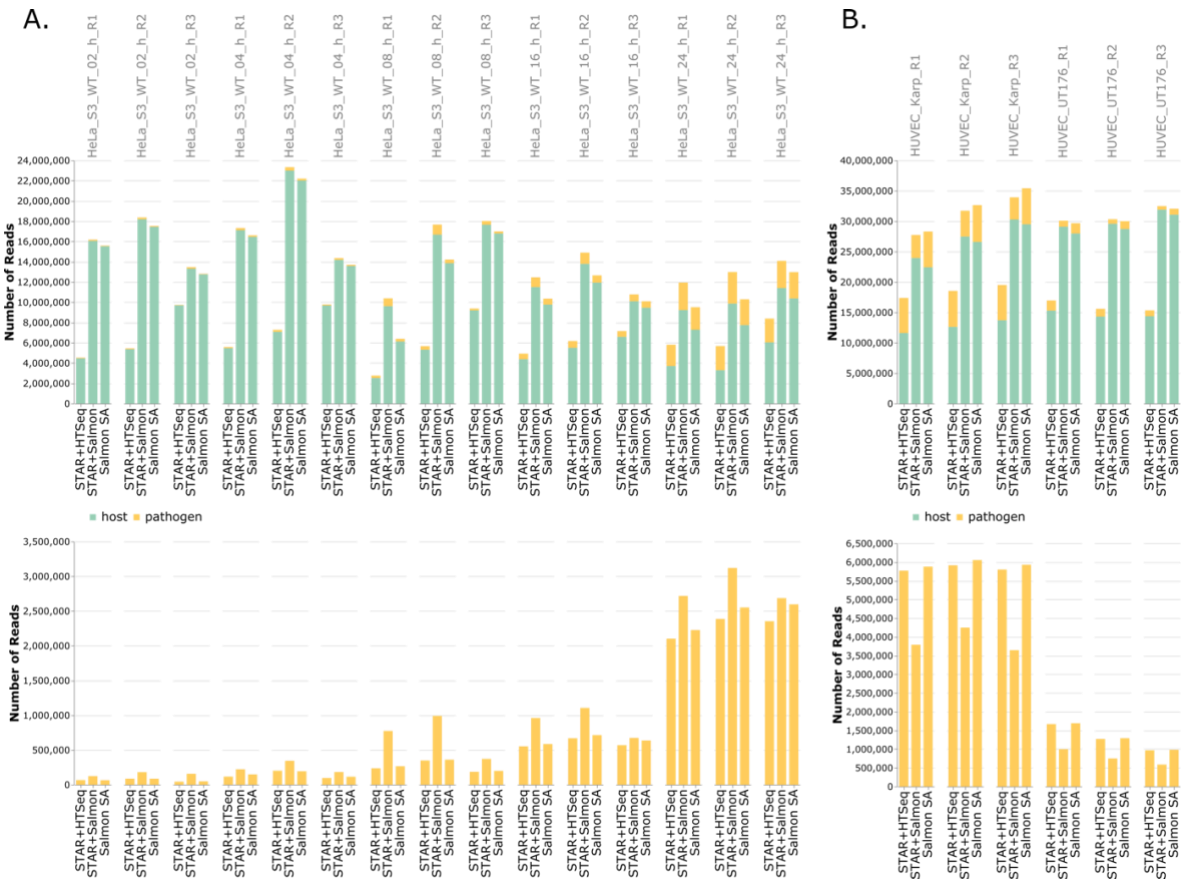

Figure 3.4 Quantification statistics for dual RNA-seq data involving various host-pathogen systems. Plots represent a summary of reads assigned to either the host or pathogen for A) the S. Typhimurium-Hela dual RNA-seq data set. B) two Ot strains infecting HUVEC cells, using three mapping-quantification strategies implemented in the Dualrnaseq workflow. The upper plots show both host and pathogen results; the bottom represent reads assigned to pathogens solely.

The benchmark analysis of the mapping and quantification methods was performed using the original dual RNA-seq data sets (see 3.2.3.1 for details). The new sets of sequencing reads were simulated from the HeLa-*Salmonella* and HUVEC-*Orientia* chimeric transcriptomes, and the salmon quantification tables functioned as ground truth. For the simulated time course dual RNA-seq data of *Salmonella*, high dispersion of the correlation coefficients between the pathogen gene expression estimates and the ground truth was observed in the STAR-HTSeq strategy (Figure 3.5A; Table S7). This behavior depends on the time point: the later the time point, the more bacterial reads in the pool, and the worse the HTSeq estimates of the pathogen gene expression. Salmon quantifies both host and pathogen genes with high accuracy for all Hela-*Salmonella* samples. Considering the RNA species, tRNAs and rRNAs are the most challenging to quantify with any of the three strategies (Figure 3.5A, Appendix 2), which might be driven by high sequence fragment similarity among tRNAs and rRNA transcripts. However, Salmon estimates showed a higher correlation with ground truth than HTSeq. For another host-pathogen system — *Ot* infecting human endothelial cells — Salmon also outperformed HTSeq (Figure 3.5A; Table S7). Interestingly, results for *Ot* RNA species have indicated a lower correlation between ground truth and estimates of Karp and UT176 ncRNAs and protein-coding sequences compared to tRNAs and rRNAs (Figure 3.5A, Appendix 2). This might be caused by a relatively small number of annotated tRNAs and rRNAs used in this analysis, and a high proportion of repetitive elements in the *Ot* genome annotated as CDS. Therefore, the evaluation of the mapping-quantification methods was also performed for RAGE elements in *Ot* str. Karp, defined in the previous chapter as a set of integrases, transposases, conjugal transfer genes, and hypothetical proteins. The comparison of the ground truth and the number of read estimates for RAGE genes indicates a higher accuracy of Salmon which handles multi-mapped reads unlike HTSeq which only counts uniquely-mapped reads (Figure 3.5B; Table S8). Overall, Salmon showed a good performance with both strategies, Selective-Alignment and alignment-based quantification. Thus, the mapping strategy does not play as important a role as the method of quantification.

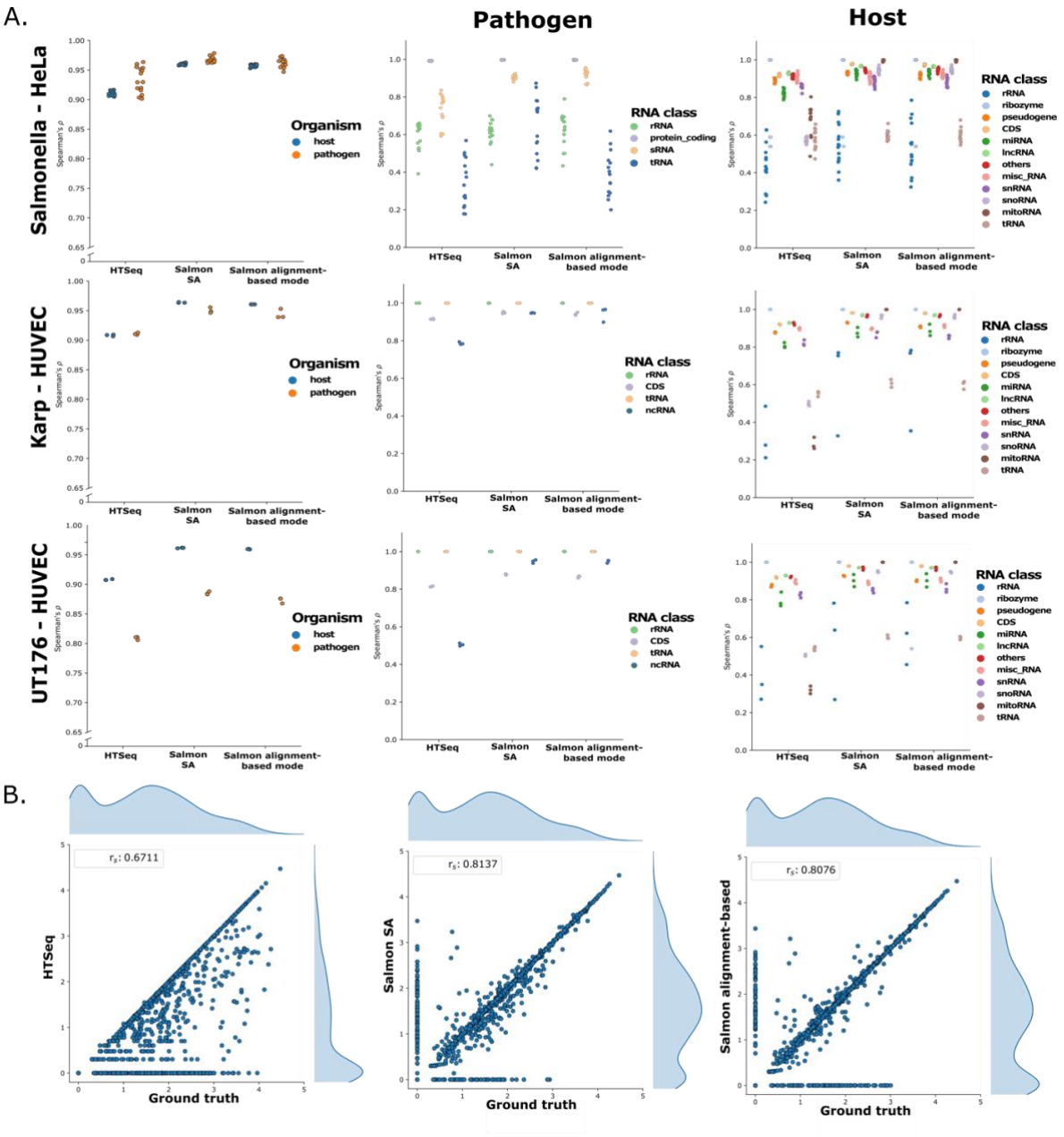

Figure 3.5 Benchmark analysis of various host-pathogen systems. Each dot represents a Spearman's rank correlation coefficient between the Number of reads (NR) and the ground truth for a sample obtained from S. Typhimurium-Hela dual RNA-seq and Orientia tsutsugamushi str. Karp/UT176–HUVEC RNA-seq. From left, the plots represent correlation coefficients obtained for each sample considering host and pathogen genes separately; the middle plots show correlation coefficients for pathogen in respect to various RNA species, and the right plots represent correlations obtained for the host RNA classes. B) Comparison of log_{10} NR estimates obtained for Karp RAGE elements (hypothetical proteins, integrases, transposases and conjugal transfer proteins) with the ground truth. Spearman's rank correlation coefficient was calculated for untransformed NR values; first replicates are shown.

3.4 Discussion

Dual RNA-seq data analysis involves various tools to handle diverse genomes and transcriptomes, making the sequencing data processing more complex than in standard RNA-seq. Furthermore, the lack of any user-friendly workflow that would facilitate efficient analysis of such data and reproducibility makes it more challenging. To my knowledge, the pipeline presented in this study is the first publicly available workflow created explicitly for processing dual RNA-seq data. Dualrnaseq includes various strategies of mapping and quantification, allowing extensive exploration of different host-pathogen systems. Notably, the workflow can quantify the whole range of RNA species, enabling the identification of both coding and non-coding transcripts expressed during the infection. Although the tools employed in the pipeline are widely used across the bioinformatics community and show good performance in various benchmark analyses, their default parameters were optimized for a particular type of genome, e.g., STAR for mammalian genome (Dobin et al., 2013). Thus, their application to various references should be performed with caution, especially when using bacterial or viral genomes which drastically differ from the eukaryotic. Although the testing of the Dualrnaseq pipeline on various transcriptomic data has enabled exploration of some parameter values, further improvement of the pipeline performance and the evaluation of the accuracy of the three mapping-quantification strategies require in-depth benchmark analysis for this particular type of data.

The initial simulation-based analysis implied higher importance of the quantification method compared to the mapping technique for obtaining accurate expression estimates, which is consistent with previous reports (Fonseca et al., 2014; Teng et al., 2016). Overall, analyses benefit from quantifying multi-mapped reads using a model-based approach (Soneson et al., 2016), providing high-quality expression estimates that allow accurate identification of differentially expressed genes (Sahraeian et al., 2017). Thus, the STAR-Salmon strategy is recommended instead of STAR-HTSeq if an alignment-based method is desired within the Dualrnaseq pipeline. The relatively high number of assigned reads to the host with Salmon (Figure 3.4) raises the question of what genes are identified as expressed in this approach but would have been missed quantifying uniquely-mapped reads only. Likely, most of these genes are rRNAs, pseudogenes, snRNAs, miscRNAs, and snoRNAs, as these biotypes were shown to display high sequence similarity to other genes in the human genome (Deschamps-Francoeur et al., 2020). Although Salmon quantifies most RNA species very well in both Selective-Alignment and alignment-based modes (Figure 3.5A), assessment of the mapping-quantification methods on various RNA classes requires additional inspection. It has already been shown that alignment-free methods outperform HTSeq in quantifying some non-coding transcripts (Zheng et al., 2019), but they are limited in quantifying very short or lowly-expressed RNAs (D. C. Wu et al., 2018; Everaert et al., 2017). Many benchmark analyses of RNA-seq tools

have neglected the non-coding transcripts, but their significance in many biological processes indicates the importance of such investigations.

A further simulation-based analysis would evaluate various sequencing library types to test the hypothesis if the long paired-end reads show more accurate results than short single-end reads providing information on the minimum read length that gives satisfactory results. For instance, 50 bp single-end reads have been shown to be sufficient for accurate differential expression analysis unless splice-junction detection is necessary (Chhangawala et al., 2015). Such an analysis would identify the impact of strandedness and read length for dual transcriptomic data, particularly on the number of multi-mapped reads, expression of ncRNAs, and overall performance of the approaches implemented in the Dualrnaseq pipeline. Evaluation of execution time and memory usage of the workflow would help investigate the resource limitations for each of the Dualrnaseq modes. Such a comprehensive benchmark study will provide concrete recommendations for dual RNA-seq data analysis.

As Dualrnaseq is built on the Nextflow framework, it is easily extensible to integrate additional tools and scripts, expanding its applicability. The current version of the pipeline is developed and tested explicitly for eukaryotic-bacterial systems. However, as the application of dual RNA-seq ranges from viruses to parasites, the pipeline should be further adapted to analyze other types of interspecies interactions, including non-pathogenic relationships such as commensalism, and mutualism (Wolf et al., 2018). Depending on the system and their reference accessibility, the user should also have a possibility to decide if, in Salmon Selective-Alignment mode, the decoy sequence will consist of two, one, or neither genomes. The current version of this strategy in the pipeline limits the analysis of an organism with an incomplete genome sequence as the workflow requires this reference. On the other hand, the lack of the decoy for a pathogen would reduce the accuracy of the analysis if the precise transcript annotations for this organism are available. Moreover, integration of assembly methods, e.g. the align-then-assemble technique by combining STAR and Cufflinks (Dobin & Gingeras, 2015), would allow generating a transcriptome assembly for two interacting organisms in each condition, facilitating transcript discovery. Such a transcriptome reconstruction may help to complete transcript annotations, improving the specificity of RNA quantification (Zheng et al., 2019). In addition, the *de novo* assembly of unmapped reads would help to identify sequence variations in the studied organisms different from the reference sequences, contamination, or co-infections relevant to tissue-derived samples. Investigation of the presence of other species in the samples may also be possible by integrating tools such as FastQ Screen (Wingett & Andrews, 2018), which screens the sequencing data against a set of databases of viral, bacterial, or other sequences.

Dualrnaseq has gained many benefits by being written in Nextflow. It is straightforward to run: after cloning it from the nf-core Github repository https://github.com/nf-core/dualrnaseq to any location and installing Nextflow and one of the container technologies (Docker or Singularity),

Dualrnaseq can be run with a single command. Different user-customized profiles and configuration files allow the reliable execution of the workflow for transcriptomic data of various host-pathogen systems. As analysis of the dual RNA-seq data involves several computationally demanding steps, handy features of Nextflow include (i) ability to execute the workflow on multicore systems; (ii) reentrancy feature using caching, which enables the pipeline to recover from the nearest checkpoint if the execution was interrupted, avoiding overwriting already generated files. Also, supporting containerization and providing reports of file paths, pipeline parameters, and software versions with each run, Dualrnaseq is fully reproducible. As the workflow is a part of nf-core, a continuously growing community that constantly improves the pipeline template bringing new functionalities that allow building highly standardized, reproducible, scalable, and robust pipelines, I believe the Dualrnaseq workflow can improve the reproducibility of dual RNA-seq studies.

4 Future perspectives

Dual RNA-seq opens new possibilities for investigating complex host-pathogen interactions that would have been difficult to explore with traditional experimental methods. Examples include the study on the obligate intracellular pathogen *Orientia tsutsugamushi* described in the second chapter of this work. Although high-quality sequencing data and continued development of the computational methods provide opportunities to investigate both model and non-model organisms, the study of many infectious agents is still limited at the experimental stage. Therefore, the development of new infection models, protocols that capture the total RNA of the interacting partners, and methods that improve the detection of all important transcripts present in the sample is a further direction (reviewed in Westermann & Vogel, 2021). For instance, novel *in-vitro* systems, e.g., organoids (Yin & Zhou 2018) or more complex 3D tissue infection systems (Schulte et al., 2020) will enable studying the host-pathogen interactions in conditions that better mimic the physiological infection process in humans compared to standard cell cultures or animal models. Application of dual RNA-seq to such systems will allow examining the interaction of pathogens with different host cell types providing a better understanding of the molecular strategies employed by the pathogen and the response activated by various host cell types (Schulte et al., 2020) or genetically diverse host cells (Saxena et al., 2017). Simultaneous profiling of three interacting partners is a step toward studying more complex inter-species interplays, i.e. triple RNA-seq which identified synergies between viral and fungal co-infection of immune cells (Seelbinder et al., 2020). Further examination of other inter-species relationships may involve commensal bacteria as they play a key role in pathogenesis through competition with pathogens in the host.

Recent studies have implied the importance of the microbiome for human health and disease (Fan & Pedersen, 2021; Nejman et al., 2020; Wirbel et al, 2019), or correlated pathogens with non-infectious diseases (Hatta et al., 2021). Therefore, such a simultaneous identification of host sequencing reads and unknown a priori microbes or viruses (Simon et al., 2018; Xu et al., 2014; Zapatka et al., 2020) may generate new hypotheses regarding their role in human diseases. *In-silico* investigation of the host transcriptome and microbial metatranscriptome may help to identify microbial composition and metabolic functions as well as the host immune signaling pathways associated with a disease (Ren et al., 2018; Pérez-Losada et al., 2015). As RNA-seq does not rely on annotations, it has a great potential to be successfully applied in the clinics for detecting pathogens in human-derived samples and identifying their interaction with the host, without requirement to know a priori a composition of samples (Wesolowska-Andersen et al., 2017). Although the triggered transcriptional program of the infected host can be predictive of the viral and bacterial infection (He et al., 2021; Mayhew et al., 2020), considering the impact of the host genetic background on the host susceptibility and microbiome and/or pathogen transcriptomes together with their interactions will

help to identify the risk factors of developing more severe symptoms and guide potential therapeutic interventions or predict responsiveness to therapy.

Dual RNA-seq can help to uncover host factors with biomarker potential by profiling expression of RNA extracted from, e.g. human tissues (Thänert et al., 2019). However, the bulk RNA-seq of tissues or organs represents the sum of the measurements collected from different host cell types. Thus, the detection of pathogens at the beginning of the infection and examination of cell-type specific host-pathogen interactions and bystander effects requires more sensitive methods. Dual RNA-seq with cell type-specific antibody staining (Frönicke et al. 2018; Pisu et al. 2020) or single-cell transcriptomics (Avraham et al. 2015; Saliba et al. 2016; Blattman et al. 2020; Imdahl et al. 2020; Kuchina et al. 2021) enables profiling of gene expression considering cellular heterogeneity. In addition, a comprehensive view on the dynamics of the system can be obtained by performing time course dual RNA-seq experiments. Such simultaneous profiling of temporal changes in gene expression of multiple interacting organisms will uncover complex interactions, and novel virulence factors and host pathways activated during infection. Further development and broader application of methods to analyze temporal RNA-seq data (Spies et al. 2019; Liang & Kelemen 2018; Kaur et al., 2020; Pierrelée et al., 2021) will provide new discoveries in gene regulatory interactions in the host-pathogen systems. Additionally, as non-coding transcripts play a number of key regulatory roles, capturing their expression profile will complete the picture of the infection processes at the transcriptomic level uncovering some ncRNAs that may serve as novel biomarkers of the ongoing infection or therapeutic targets used to modify the pathogen's activity within the host.

With the advent of the development of high throughput methods (Stark et al., 2019), dual RNA-seq will help to design more efficient diagnostic and therapeutic strategies. Further integration of dual RNA-seq data with other types of omics, e.g., TraDIS data (Cainet al., 2020) or CRISPR-pooled screen data (Yeung et al., 2019; Lai et al., 2021) will support an efficient finding of novel virulence determinants and potential therapeutic targets. Also, as high transcript expression does not necessarily imply protein production, the integration of proteomics or further validation should be performed to determine the functionality of expressed transcripts identified by RNA-seq. Therefore, investigation of complex interactions, such as the interplay between two or more diverse organisms requires employment of various methods. For instance, complementing traditional laboratory experiments with mathematical modeling allows investigating such spatially heterogeneous and dynamic processes in a time and cost-effective manner (Schulze et al., 2016; Ewald et al. 2020; Dühring et al. 2015; Seal et al. 2011; Peer & An 2014), and dual RNA-seq will be a great source of hypotheses and valuable information necessary for modeling. Thus, the synergy of experimental work with various computational analyses creates a powerful tool for understanding the biology of many infectious agents and the processes associated with the infection, which may lead to novel or improved treatment strategies.

5 References

Adcox, H. E., Hatke, A. L., Andersen, S. E., Gupta, S., Otto, N. B., Weber, M. M., Marconi, R. T., & Carlyon, J. A. (2021). Orientia tsutsugamushi Nucleomodulin Ank13 Exploits the RaDAR Nuclear Import Pathway To Modulate Host Cell Transcription. *mBio*, 12(4), e0181621.

Ahmed, A. E., Allen, J. M., Bhat, T., Burra, P., Fliege, C. E., Hart, S. N., Heldenbrand, J. R., Hudson, M. E., Istanto, D. D., Kalmbach, M. T., Kapraun, G. D., Kendig, K. I., Kendzior, M. C., Klee, E. W., Mattson, N., Ross, C. A., Sharif, S. M., Venkatakrishnan, R., Fadlelmola, F. M., & Mainzer, L. S. (2021). Design considerations for workflow management systems use in production genomics research and the clinic. *Scientific reports*, 11(1), 21680.

Akama, T., Suzuki, K., Tanigawa, K., Kawashima, A., Wu, H., Nakata, N., Osana, Y., Sakakibara, Y., & Ishii, N. (2009). Whole-genome tiling array analysis of Mycobacterium leprae RNA reveals high expression of pseudogenes and noncoding regions. *Journal of Bacteriology*, 191(10), 3321–3327.

Akopian, D., Shen, K., Zhang, X., & Shan, S.-O. (2013). Signal Recognition Particle: An Essential Protein-Targeting Machine. *Annual Review of Biochemistry*, 82(1), 693–721.

Albrecht, M., Sharma, C. M., Dittrich, M. T., Müller, T., Reinhardt, R., Vogel, J., & Rudel, T. (2011). The transcriptional landscape of Chlamydia pneumoniae. *Genome Biology*, 12(10), R98.

Al-Khodor, S., Price, C. T., Kalia, A., & Abu Kwaik, Y. (2010). Functional diversity of ankyrin repeats in microbial proteins. *Trends in Microbiology*, 18(3), 132–139.

Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Pontén, T., Alsmark, U. C., Podowski, R. M., Näslund, A. K., Eriksson, A. S., Winkler, H. H., & Kurland, C. G. (1998). The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature*, *396*(6707), 133–140.

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* , *31*(2), 166–169.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Aoki, J. I., Muxel, S. M., Zampieri, R. A., Müller, K. E., Nerland, A. H., & Floeter-Winter, L. M. (2019). Differential immune response modulation in early Leishmania amazonensis infection of BALB/c and C57BL/6 macrophages based on transcriptome profiles. *Scientific Reports*, *9*(1), 19841.

Aprianto, R., Slager, J., Holsappel, S., & Veening, J.-W. (2016). Time-resolved dual RNA-seq reveals extensive rewiring of lung epithelial and pneumococcal transcriptomes during early infection. *Genome Biology*, *17*(1), 198.

Aronesty, E. (2013). Comparison of sequencing utility programs. *The open bioinformatics journal*, 7(1).

Avital, G., Avraham, R., Fan, A., Hashimshony, T., Hung, D. T., & Yanai, I. (2017). scDual-Seq: mapping the gene regulatory program of Salmonella infection by host and pathogen single-cell RNA-sequencing. *Genome Biology*, *18*(1), 200.

Avraham, R., Haseley, N., Brown, D., Penaranda, C., Jijon, H. B., Trombetta, J. J., ... & Hung, D. T. (2015). Pathogen cell-to-cell variability drives heterogeneity in host immune responses. *Cell*, 162(6), 1309-1321.

Baddal, B., Muzzi, A., Censini, S., Calogero, R. A., Torricelli, G., Guidotti, S., Taddei, A. R., Covacci, A., Pizza, M., Rappuoli, R., Soriani, M., & Pezzicoli, A. (2015). Dual RNA-seq of Nontypeable Haemophilus influenzae and Host Cell Transcriptomes Reveals Novel Insights into Host-Pathogen Cross Talk. *mBio*, *6*(6), e01765–15.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454.

Balsells-Llauradó, M., Silva, C. J., Usall, J., Vall-Llaura, N., Serrano-Prieto, S., Teixidó, N., Mesquida-Pesci, S. D., de Cal, A., Blanco-Ulate, B., & Torres, R. (2020). Depicting the battle between nectarine and Monilinia laxa: the fruit developmental stage dictates the effectiveness of the host defenses and the pathogen's infection strategies. *Horticulture Research*, *7*, 167.

Bang, S., Min, C.-K., Ha, N.-Y., Choi, M.-S., Kim, I.-S., Kim, Y.-S., & Cho, N.-H. (2016). Inhibition of eukaryotic translation by tetratricopeptide-repeat proteins of Orientia tsutsugamushi. *Journal of Microbiology* , *54*(2), 136–144.

Barquist, L., & Vogel, J. (2015). Accelerating Discovery and Functional Analysis of Small RNAs with New Technologies. *Annual Review of Genetics*, *49*, 367–394.

Baruzzo, G., Hayer, K. E., Kim, E. J., Di Camillo, B., FitzGerald, G. A., & Grant, G. R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods*, *14*(2), 135–139.

Batty, E. M., Chaemchuen, S., Blacksell, S., Richards, A. L., Paris, D., Bowden, R., Chan, C., Lachumanan, R., Day, N., Donnelly, P., Chen, S., & Salje, J. (2018). Long-read whole genome sequencing and comparative analysis of six strains of the human pathogen Orientia tsutsugamushi. *PLoS Neglected Tropical Diseases*, *12*(6), e0006566.

Belland, R. J., Zhong, G., Crane, D. D., Hogan, D., Sturdevant, D., Sharma, J., Beatty, W. L., & Caldwell, H. D. (2003). Genomic transcriptional profiling of the developmental cycle of Chlamydia trachomatis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(14), 8478–8483.

Bender, J. K., Wille, T., Blank, K., Lange, A., Gerlach, R. G., & Junior Research Group. (2013). LPS structure and PhoQ activity are important for Salmonella Typhimurium virulence in the Gallleria mellonella infection model. *PloS one*, 8(8), e73287.

Berk, J. M., Lim, C., Ronau, J. A., Chaudhuri, A., Chen, H., Beckmann, J. F., Loria, J. P., Xiong, Y., & Hochstrasser, M. (2020). A deubiquitylase with an unusually high-affinity ubiquitin-binding domain from the scrub typhus pathogen Orientia tsutsugamushi. *Nature Communications*, *11*(1), 2343.

Betin, V., Penaranda, C., Bandyopadhyay, N., Yang, R., Abitua, A., Bhattacharyya, R. P., Fan, A., Avraham, R., Livny, J., Shoresh, N., & Hung, D. T. (2019). Hybridization-based capture of pathogen mRNA enables paired host-pathogen transcriptional analysis. *Scientific Reports*, *9*(1), 19244.

Beyer, A. R., Rodino, K. G., VieBrock, L., Green, R. S., Tegels, B. K., Oliver, L. D., Jr, Marconi, R. T., & Carlyon, J. A. (2017). Orientia tsutsugamushi Ank9 is a multifunctional effector that utilizes a novel GRIP-like Golgi localization domain for Golgi-to-endoplasmic reticulum trafficking and interacts with host COPB2. *Cellular Microbiology*, *19*(7), e12727.

Beyer, A. R., VieBrock, L., Rodino, K. G., Miller, D. P., Tegels, B. K., Marconi, R. T., & Carlyon, J. A. (2015). Orientia tsutsugamushi Strain Ikeda Ankyrin Repeat-Containing Proteins Recruit SCF1 Ubiquitin Ligase Machinery via Poxvirus-Like F-Box Motifs. *Journal of Bacteriology*, *197*(19), 3097–3109.

Blattman, S. B., Jiang, W., Oikonomou, P., & Tavazoie, S. (2020). Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. *Nature Microbiology*, 5(10), 1192–1201.

Boettiger, C. (2015). An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review*, *49*(1), 71–79.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* , *30*(15), 2114–2120.

Bonell, A., Lubell, Y., Newton, P. N., Crump, J. A., & Paris, D. H. (2017). Estimating the burden of scrub typhus: A systematic review. *PLoS Neglected Tropical Diseases*, *11*(9), e0005838.

Bossi, L., Schwartz, A., Guillemardet, B., Boudvillain, M., & Figueroa-Bossi, N. (2012). A role for Rho-dependent polarity in gene regulation by a noncoding small RNA. *Genes & development*, *26*(16), 1864–1873.

Botwright, N. A., Mohamed, A. R., Slinger, J., Lima, P. C., & Wynne, J. W. (2021). Host-Parasite Interaction of Atlantic salmon (Salmo salar) and the Ectoparasite Neoparamoeba perurans in Amoebic Gill Disease. *Frontiers in Immunology*, *12*, 672700.

Boxx, G. M., & Cheng, G. (2016). The Roles of Type I Interferon in Bacterial Infection. *Cell Host & Microbe*, *19*(6), 760–769.

Bradwell, K. R., Coulibaly, D., Koné, A. K., Laurens, M. B., Dembélé, A., Tolo, Y., Traoré, K., Niangaly, A., Berry, A. A., Kouriba, B., Plowe, C. V., Doumbo, O. K., Lyke, K. E., Takala-Harrison, S., Thera, M. A., Travassos, M. A., & Serre, D. (2020). Host and Parasite Transcriptomic Changes upon Successive Plasmodium falciparum Infections in Early Childhood. *mSystems*, *5*(4).

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*, 525.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60.

Buddenborg, S. K., Bu, L., Zhang, S.-M., Schilkey, F. D., Mkoji, G. M., & Loker, E. S. (2017). Transcriptomic responses of Biomphalaria pfeifferi to Schistosoma mansoni: Investigation of a neglected African snail that supports more S. mansoni transmission than any other snail species. *PLoS Neglected Tropical Diseases*, *11*(10), e0005984.

Bumgarner, R. (2013). Overview of DNA microarrays: types, applications, and their future. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, *Chapter 22*, Unit 22.1.

Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D. (2019). rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, *8*(9).

Bushnell, B. BBduk version 38.87.[software] 2020. https://jgi.doe.gov/data-and-tools/bbtools/

Bushnell, B. (2014). BBTools software package. URL http://sourceforge.net/projects/bbmap, 578, 579.

Cain, A. K., Barquist, L., Goodman, A. L., Paulsen, I. T., Parkhill, J., & van Opijnen, T. (2020). A decade of advances in transposon-insertion sequencing. Nature Reviews Genetics, 21(9), 526-540.

Camilios-Neto, D., Bonato, P., Wassem, R., Tadra-Sfeir, M. Z., Brusamarello-Santos, L. C. C., Valdameri, G., Donatti, L., Faoro, H., Weiss, V. A., Chubatsu, L. S., Pedrosa, F. O., & Souza, E. M. (2014). Dual RNA-seq transcriptional analysis of wheat roots colonized by Azospirillum brasilense reveals up-regulation of nutrient acquisition and cell cycle genes. *BMC Genomics*, *15*, 378.

Carlson M. (2019). GO.db: A set of annotation maps describing the entire Gene Ontology. R package version 3.5.0.

Castanheira, S., & García-Del Portillo, F. (2017). Salmonella Populations inside Host Cells. *Frontiers in Cellular and Infection Microbiology*, *7*, 432.

Cerveny, L., Straskova, A., Dankova, V., Hartlova, A., Ceckova, M., Staud, F., & Stulik, J. (2013). Tetratricopeptide repeat motifs in the world of bacterial pathogens: role in virulence mechanisms. *Infection and Immunity*, *81*(3), 629–635.

Chacon, S., & Straub, B. (2014). *Pro git*. Springer Nature.

Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M., & Gu, J. (2017). AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC bioinformatics, 18(3), 91-100.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884-i890.

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., Zhang, X., Wang, J., Yang, H., Fang, L., & Chen, Q. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience*, *7*(1), 1–6.

Chen, Y., Lun, A. T. L., & Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, *5*.

Chhangawala, S., Rudy, G., Mason, C. E., & Rosenfeld, J. A. (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biology*, *16*, 131.

Chierakul, W., de Fost, M., Suputtamongkol, Y., Limpaiboon, R., Dondorp, A., White, N. J., & van der Poll, T. (2004). Differential expression of interferon-γ and interferon-γ-inducing cytokines in Thai patients with scrub typhus or leptospirosis. *Clinical Immunology , 113*(2), 140–144.

Choi, Y. J., Aliota, M. T., Mayhew, G. F., Erickson, S. M., & Christensen, B. M. (2014). Dual RNA-seq of parasite and host reveals gene expression dynamics during filarial worm–mosquito interactions. *PLoS neglected tropical diseases*, 8(5), e2905.

Cho, B.A., Cho, N.-H., Seong, S.-Y., Choi, M.-S., & Kim, I.-S. (2010). Intracellular invasion by Orientia tsutsugamushi is mediated by integrin signaling and actin cytoskeleton rearrangements. *Infection and Immunity*, *78*(5), 1915–1923.

Cho, K.A., Jun, Y. H., Suh, J. W., Kang, J.-S., Choi, H. J., & Woo, S.-Y. (2010). Orientia tsutsugamushi induced endothelial cell activation via the NOD1-IL-32 pathway. *Microbial Pathogenesis*, *49*(3), 95–104.

Cho, N. H., Kim, H. R., Lee, J. H., Kim, S. Y., Kim, J., Cha, S., ... & Kim, I. S. (2007). The Orientia tsutsugamushi genome reveals massive proliferation of conjugative type IV secretion system and host–cell interaction genes. *Proceedings of the National Academy of Sciences*, 104(19), 7981-7986.

Cho, N. H., Seong, S. Y., Choi, M. S., & Kim, I. S. (2001). Expression of chemokine genes in human dermal microvascular endothelial cell lines infected with Orientia tsutsugamushi. *Infection and Immunity*, *69*(3), 1265–1272.

Cho, N. H., Seong, S. Y., Huh, M. S., Han, T. H., Koh, Y. S., Choi, M. S., & Kim, I. S. (2000). Expression of chemokine genes in murine macrophages infected with Orientia tsutsugamushi. *Infection and Immunity*, *68*(2), 594–602.

Chung, D. R., Lee, Y. S., & Lee, S. S. (2008). Kinetics of inflammatory cytokines in patients with scrub typhus receiving doxycycline treatment. *The Journal of Infection*, *56*(1), 44–50.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422-1423.

Colgan, A. M., Cameron, A. D., & Kröger, C. (2017). If it transcribes, we can sequence it: mining the complexities of host-pathogen-environment interactions using RNA-seq. *Current Opinion in Microbiology*, *36*, 37–46.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13.

Conway, T., Creecy, J. P., Maddox, S. M., Grissom, J. E., Conkle, T. L., Shadid, T. M., Teramoto, J., San Miguel, P., Shimada, T., Ishihama, A., Mori, H., & Wanner, B. L. (2014). Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. *mBio*, *5*(4), e01442–14.

Croucher, N. J., & Thomson, N. R. (2010). Studying bacterial transcriptomes using RNA-seq. *Current Opinion in Microbiology*, *13*(5), 619–624.

Daily, J. P., Le Roch, K. G., Sarr, O., Ndiaye, D., Lukens, A., Zhou, Y., Ndir, O., Mboup, S., Sultan, A., Winzeler, E. A., & Wirth, D. F. (2005). In vivo transcriptome of Plasmodium falciparum reveals overexpression of transcripts that encode surface proteins. *The Journal of Infectious Diseases*, *191*(7), 1196–1203.

Damron, F. H., Oglesby-Sherrouse, A. G., Wilks, A., & Barbier, M. (2016). Dual-seq transcriptomics reveals the battle for iron during Pseudomonas aeruginosa acute murine pneumonia. *Scientific Reports*, *6*, 39172.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, 10(2), giab008.

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PloS One*, *8*(12), e85024.

Deschamps-Francoeur, G., Simoneau, J., & Scott, M. S. (2020). Handling multi-mapped reads in RNA-seq. *Computational and Structural Biotechnology Journal*, 18, 1569-1576.

Deshpande, D., Chhugani, K., Chang, Y., Karlsberg, A., Loeffler, C., Zhang, J., Muszynska, A., Rotman, J., Tao, L., Martin, L. S., Balliu, B., Tseng, E., Eskin, E., Zhao, F., Mohammadi, P., Labaj, P. P., & Mangul, S. (2020). A comprehensive overview of computational tools for RNA-seq analysis. arXiv e-prints, arXiv-2010.

Díaz, F. E., Abarca, K., & Kalergis, A. M. (2018). An Update on Host-Pathogen Interplay and Modulation of Immune Responses during Orientia tsutsugamushi Infection. *Clinical Microbiology Reviews*, *31*(2).

Dillon, L. A. L., Suresh, R., Okrah, K., Corrada Bravo, H., Mosser, D. M., & El-Sayed, N. M. (2015). Simultaneous transcriptional profiling of Leishmania major and its murine macrophage host cell reveals insights into host-pathogen interactions. *BMC Genomics*, *16*, 1108.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., … Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, *489*(7414), 101–108.

D'Mello, A., Riegler, A. N., Martínez, E., Beno, S. M., Ricketts, T. D., Foxman, E. F., ... & Tettelin, H. (2020). An in vivo atlas of host–pathogen transcriptomes during Streptococcus pneumoniae colonization and disease. *Proceedings of the National Academy of Sciences*, 117(52), 33507-33518.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , *29*(1), 15–21.

Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq reads with STAR. *Current protocols in bioinformatics*, 51(1), 11-14.

Doing, G., Koeppen, K., Occipinti, P., Harty, C. E., & Hogan, D. A. (2020). Conditional antagonism in co-cultures of Pseudomonas aeruginosa and Candida albicans: An intersection of ethanol and phosphate signaling distilled from dual-seq transcriptomics. *PLoS Genetics*, *16*(8), e1008783.

Dühring, S., Germerodt, S., Skerka, C., Zipfel, P., Dandekar, T., & Schuster, S. (2015). Host-pathogen interactions between the human innate immune system and Candida albicans—understanding and modeling defense and evasion strategies. *Frontiers in Microbiology*, 6, 625.

Ellis, M. J., & Haniford, D. B. (2016). Riboregulation of bacterial and archaeal transposition. *Wiley Interdisciplinary Reviews. RNA*, *7*(3), 382–398.

Enatsu, T., Urakami, H., & Tamura, A. (1999). Phylogenetic analysis of Orientia tsutsugamushi strains based on the sequence homologies of 56-kDa type-specific antigen genes. *FEMS Microbiology Letters*, *180*(2), 163–169.

Eriksson, S., Lucchini, S., Thompson, A., Rhen, M., & Hinton, J. C. D. (2003). Unravelling the biology of macrophage infection by gene expression profiling of intracellular Salmonella enterica. *Molecular Microbiology*, *47*(1), 103–118.

Espindula, E., Sperb, E. R., Bach, E., & Passaglia, L. M. P. (2019). The combined analysis as the best strategy for Dual RNA-Seq mapping. *Genetics and Molecular Biology*, *42*(4).

Evans, S. M., Rodino, K. G., Adcox, H. E., & Carlyon, J. A. (2018). Orientia tsutsugamushi uses two Ank effectors to modulate NF-κB p65 nuclear transport and inhibit NF-κB transcriptional activation. *PLoS Pathogens*, *14*(5), e1007023.

Everaert, C., Luypaert, M., Maag, J. L., Cheng, Q. X., Dinger, M. E., Hellemans, J., & Mestdagh, P. (2017). Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Scientific reports*, 7(1), 1-11.

Ewald, J., Sieber, P., Garde, R., Lang, S. N., Schuster, S., & Ibrahim, B. (2020). Trends in mathematical modeling of host–pathogen interactions. *Cellular and Molecular Life Sciences*: CMLS, 77(3), 467–480.

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., ... & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology*, 38(3), 276-278.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* , *32*(19), 3047–3048.

Fabozzi, G., Oler, A. J., Liu, P., Chen, Y., Mindaye, S., Dolan, M. A., ... & Subbarao, K. (2018). Strand-specific dual RNA sequencing of bronchial epithelial cells infected with influenza A/H3N2 viruses reveals splicing of gene segment 6 and novel host-virus interactions. *Journal of virology*, 92(17), e00518-18.

Fan, Y., & Pedersen, O. (2021). Gut microbiota in human metabolic health and disease. *Nature Reviews. Microbiology*, 19(1), 55–71.

Farrer, R. A., Ford, C. B., Rhodes, J., Delorey, T., May, R. C., Fisher, M. C., Cloutman-Green, E., Balloux, F., & Cuomo, C. A. (2018). Transcriptional Heterogeneity of Cryptococcus gattii VGII Compared with Non-VGII Lineages Underpins Key Pathogenicity Pathways. *mSphere*, *3*(5).

Fernandes, M. C., Dillon, L. A. L., Belew, A. T., Bravo, H. C., Mosser, D. M., & El-Sayed, N. M. (2016). Dual Transcriptome Profiling of Leishmania-Infected Human Macrophages Reveals Distinct Reprogramming Signatures. *mBio*, *7*(3).

Fonseca, N. A., Marioni, J., & Brazma, A. (2014). RNA-Seq gene profiling--a systematic empirical comparison. *PloS One*, *9*(9), e107026.

Förstner, K. U., Vogel, J., & Sharma, C. M. (2014). READemption—a tool for the computational analysis of deep-sequencing–based transcriptome data. *Bioinformatics* , *30*(23), 3421–3423.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773.

Frazee AC, Jaffe AE, Kirchner R, Leek JT (2021). polyester: Simulate RNA-seq reads.

Frönicke, L., Bronner, D. N., Byndloss, M. X., McLaughlin, B., Bäumler, A. J., & Westermann, A. J. (2018). Toward Cell Type-Specific In Vivo Dual RNA-Seq. *Methods in Enzymology*, 612, 505–522.

Gal-Mor, O., Boyle, E. C., & Grassl, G. A. (2014). Same species, different diseases: how and why typhoidal and non-typhoidal Salmonella enterica serovars differ. *Frontiers in Microbiology*, *5*, 391.

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, 43(D1), D261-D269.

Garijo, D., Kinnings, S., Xie, L., Xie, L., Zhang, Y., Bourne, P. E., & Gil, Y. (2013). Quantifying reproducibility in computational biology: the case of the tuberculosis drugome. *PloS One*, *8*(11), e80278.

Gaudin, C., Zhou, X., Williams, K. P., & Felden, B. (2002). Two-piece tmRNA in cyanobacteria and its structural analysis. *Nucleic acids research*, 30(9), 2018–2024.

Georg, J., & Hess, W. R. (2011). cis-antisense RNA, another level of gene regulation in bacteria. *Microbiology and Molecular Biology Reviews: MMBR*, *75*(2), 286–300.

Georg, J., & Hess, W. R. (2018). Widespread Antisense Transcription in Prokaryotes. *Microbiology Spectrum*, *6*(4).

Gharaibeh, M., Hagedorn, M., Lilla, S., Hauptmann, M., Heine, H., Fleischer, B., & Keller, C. (2016). Toll-Like Receptor 2 Recognizes Orientia tsutsugamushi and Increases Susceptibility to Murine Experimental Scrub Typhus. *Infection and Immunity*, *84*(12), 3379–3387.

Giengkam, S., Blakes, A., Utsahajit, P., Chaemchuen, S., Atwal, S., Blacksell, S. D., Paris, D. H., Day, N. P. J., & Salje, J. (2015). Improved Quantification, Propagation, Purification and Storage of the Obligate Intracellular Human Pathogen Orientia tsutsugamushi. *PLoS Neglected Tropical Diseases*, *9*(8), e0004009.

Gillespie, J. J., Kaur, S. J., Rahman, M. S., Rennoll-Bankert, K., Sears, K. T., Beier-Sexton, M., & Azad, A. F. (2015). Secretome of obligate intracellular Rickettsia. *FEMS Microbiology Reviews*, *39*(1), 47–80.

Goecks, J., Nekrutenko, A., Taylor, J., & Galaxy Team. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*(8), R86.

Gottesman, S., & Storz, G. (2011). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harbor perspectives in biology*, 3(12), a003798.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652.

Griesenauer, B., Tran, T. M., Fortney, K. R., Janowicz, D. M., Johnson, P., Gao, H., Barnes, S., Wilson, L. S., Liu, Y., & Spinola, S. M. (2019). Determination of an Interaction Network between an Extracellular Bacterial Pathogen and the Human Host. *mBio*, *10*(3).

Guzman, C., & D'Orso, I. (2017). CIPHER: a flexible and extensive workflow platform for integrative next-generation sequencing data analysis and genomic regulatory element prediction. *BMC Bioinformatics*, *18*(1), 363.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512.

Hanson, B. (1991). Comparative susceptibility to mouse interferons of Rickettsia tsutsugamushi strains with different virulence in mice and of Rickettsia rickettsii. *Infection and Immunity*, *59*(11), 4134–4141.

Ha, N.-Y., Cho, N.-H., Kim, Y.-S., Choi, M.-S., & Kim, I.-S. (2011). An autotransporter protein from Orientia tsutsugamushi mediates adherence to nonphagocytic host cells. *Infection and Immunity*, *79*(4), 1718–1727.

Ha, N.-Y., Sharma, P., Kim, G., Kim, Y., Min, C.-K., Choi, M.-S., Kim, I.-S., & Cho, N.-H. (2015). Immunization with an autotransporter protein of Orientia tsutsugamushi provides protective immunity against scrub typhus. *PLoS Neglected Tropical Diseases*, *9*(3), e0003585.

Ha, N.-Y., Shin, H. M., Sharma, P., Cho, H. A., Min, C.-K., Kim, H.-I., Yen, N. T. H., Kang, J.-S., Kim, I.-S., Choi, M.-S., Kim, Y. K., & Cho, N.-H. (2016). Generation of protective immunity against Orientia tsutsugamushi infection by immunization with a zinc oxide nanoparticle combined with ScaA antigen. *Journal of Nanobiotechnology*, *14*(1), 76.

Haraga, A., Ohlson, M. B., & Miller, S. I. (2008). Salmonellae interplay with host cells. *Nature Reviews. Microbiology*, *6*(1), 53–66.

Harbers, M., & Carninci, P. (2005). Tag-based approaches for transcriptome research and genome annotation. *Nature Methods*, *2*(7), 495–502.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357-362.

Hatta, M. N. A., Mohamad Hanif, E. A., Chin, S.-F., & Neoh, H.-M. (2021). Pathogens and Carcinogenesis: A Review. *Biology*, 10(6), 533.

Hautefort, I., Thompson, A., Eriksson-Ygberg, S., Parker, M. L., Lucchini, S., Danino, V., Bongaerts, R. J. M., Ahmad, N., Rhen, M., & Hinton, J. C. D. (2008). canscriptional adaptation that results in simultaneous expression of three type 3 secretion systems. *Cellular Microbiology*, *10*(4), 958–984.

Hayward, R. J., Humphrys, M. S., Huston, W. M., & Myers, G. S. (2021). Dual RNA-seq analysis of in vitro infection multiplicity and RNA depletion methods in Chlamydia-infected epithelial cells. *Scientific Reports*, 11(1), 1-14.

Heller, M. J. (2002). DNA microarray technology: devices, systems, and applications. *Annual Review of Biomedical Engineering*, *4*, 129–153.

He, Y. D., Wohlford, E. M., Uhle, F., Buturovic, L., Liesenfeld, O., & Sweeney, T. E. (2021). The Optimization and Biological Significance of a 29-Host-Immune-mRNA Panel for the Diagnosis of Acute Infections and Sepsis. *Journal of Personalized Medicine*, 11(8).

Hoffmann, S., Otto, C., Doose, G., Tanzer, A., Langenberger, D., Christ, S., Kunz, M., Holdt, L. M., Teupser, D., Hackermüller, J., & Stadler, P. F. (2014). A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biology*, *15*(2), R34.

Hossain, H., Tchatalbachev, S., & Chakraborty, T. (2006). Host gene expression profiling in pathogen–host interactions. *Current Opinion in Immunology*, *18*(4), 422–429.

Hör, J., Gorski, S. A., & Vogel, J. (2018). Bacterial RNA biology on a genome scale. *Molecular cell*, 70(5), 785-799.

Huang, L., Zhao, L., Liu, W., Xu, X., Su, Y., Qin, Y., & Yan, Q. (2019). Dual RNA-Seq Unveils Pseudomonas plecoglossicida htpG Gene Functions During Host-Pathogen Interactions With Epinephelus coioides. *Frontiers in Immunology*, *10*, 984.

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. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, *44*(D1), D286–D293.

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. A. (2013). Simultaneous transcriptional profiling of bacteria and their host cells. *PloS One*, *8*(12), e80597.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(3), 90–95.

Hurley, D., McCusker, M. P., Fanning, S., & Martins, M. (2014). Salmonella-host interactions - modulation of the host innate immune system. *Frontiers in Immunology*, *5*, 481.

Imdahl, F., Vafadarnejad, E., Homberger, C., Saliba, A.-E., & Vogel, J. (2020). Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. *Nature Microbiology*, 5(10), 1202–1206.

Jackson, M., Kavoussanakis, K., & Wallace, E. (2021). Using prototyping to choose a bioinformatics workflow management system. *PLoS computational biology*, 17(2), e1008622.

Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J., & Kimmel, M. (2015). Microarray experiments and factors which affect their reliability. *Biology Direct*, *10*, 46.

Jenner, R. G., & Young, R. A. (2005). Insights into host responses against pathogens from transcriptional profiling. *Nature Reviews. Microbiology*, *3*(4), 281–294.

Jennings, E., Thurston, T. L. M., & Holden, D. W. (2017). Salmonella SPI-2 Type III Secretion System Effectors: Molecular Mechanisms And Physiological Consequences. *Cell Host & Microbe*, *22*(2), 217–231.

Jernigan, K. K., & Bordenstein, S. R. (2014). Ankyrin domains across the Tree of Life. *PeerJ*, *2*, e264.

Jerrells, T. R., & Osterman, J. V. (1981). Host defenses in experimental scrub typhus: inflammatory response of congenic C3H mice differing at the Ric gene. *Infection and Immunity*, *31*(3), 1014–1022.

Jose, B. R., Gardner, P. P., & Barquist, L. (2019). Transcriptional noise and exaptation as sources for bacterial sRNAs. *Biochemical Society Transactions*, 47(2), 527-539.

Julio, S. M., Heithoff, D. M., & Mahan, M. J. (2000). ssrA (tmRNA) Plays a Role in Salmonella enterica Serovar Typhimurium Pathogenesis. *Journal of Bacteriology*, *182*(6), 1558–1563.

Juranic Lisnic, V., Babic Cac, M., Lisnic, B., Trsan, T., Mefferd, A., Das Mukhopadhyay, C., Cook, C. H., Jonjic, S., & Trgovcich, J. (2013). Dual analysis of the murine cytomegalovirus and host cell transcriptomes reveal new aspects of the virus-host cell interface. *PLoS Pathogens*, *9*(9), e1003611.

Ju, X., Li, D., & Liu, S. (2019). Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. *Nature microbiology*, 4(11), 1907-1918.

Kachroo, P., Eraso, J. M., Olsen, R. J., Zhu, L., Kubiak, S. L., Pruitt, L., Yerramilli, P., Cantu, C. C., Ojeda Saavedra, M., Pensar, J., Corander, J., Jenkins, L., Kao, L., Granillo, A., Porter, A. R., DeLeo, F. R., & Musser, J. M. (2020). New Pathogenesis Mechanisms and Translational Leads Identified by Multidimensional Analysis of Necrotizing Myositis in Primates. *mBio*, *11*(1).

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., Bateman, A., Finn, R. D., & Petrov, A. I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*, *46*(D1), D335–D342.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361.

Kaur, S., Peters, T. J., Yang, P., Luu, L. D. W., Vuong, J., Krycer, J. R., & O'Donoghue, S. I. (2020). Temporal ordering of omics and multiomic events inferred from time-series data. *NPJ Systems Biology and Applications*, 6(1), 22.

Kawahara, Y., Oono, Y., Kanamori, H., Matsumoto, T., Itoh, T., & Minami, E. (2012). Simultaneous RNA-seq analysis of a mixed transcriptome of rice and blast fungus interaction. *PloS One*, *7*(11), e49423.

Kawai, T., & Akira, S. (2007). Signaling to NF-kappaB by Toll-like receptors. *Trends in Molecular Medicine*, *13*(11), 460–469.

Kazantsev, A. V., & Pace, N. R. (2006). Bacterial RNase P: a new view of an ancient enzyme. *Nature Reviews. Microbiology*, *4*(10), 729–740.

Keiler, K. C., Shapiro, L., & Williams, K. P. (2000). tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: A two-piece tmRNA functions in Caulobacter. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(14), 7778–7783.

Kelly, D. J., Fuerst, P. A., Ching, W. M., & Richards, A. L. (2009). Scrub typhus: the geographic distribution of phenotypic and genotypic variants of Orientia tsutsugamushi. *Clinical Infectious Diseases*, *48*(Supplement_3), S203-S230.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36.

Kim, M. J., Kim, M. K., & Kang, J. S. (2006). Orientia tsutsugamushi inhibits tumor necrosis factor α production by inducing interleukin 10 secretion in murine macrophages. *Microbial pathogenesis*, 40(1), 1-7.

Kim, M. J., Kim, M. K., & Kang, J. S. (2013). Involvement of lipid rafts in the budding-like exit of Orientia tsutsugamushi. *Microbial Pathogenesis*, *63*, 37–43.

Kim, Y. M., Poline, J. B., & Dumas, G. (2018). Experimenting with reproducibility: a case study of robustness in bioinformatics. *GigaScience*, *7*(7).

Koh, Y. S., Yun, J. H., Seong, S.-Y., Choi, M. S., & Kim, I. S. (2004). Chemokine and cytokine production during Orientia tsutsugamushi infection in mice. *Microbial Pathogenesis*, *36*(1), 51–57.

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* , *28*(24), 3211–3217.

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* , *28*(19), 2520–2522.

Krämer, A., Green, J., Pollard, J., Jr, & Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* , *30*(4), 523–530.

Kröger, C., Colgan, A., Srikumar, S., Händler, K., Sivasankaran, S. K., Hammarlöf, D. L., Canals, R., Grissom, J. E., Conway, T., Hokamp, K., & Hinton, J. C. D. (2013). An infection-relevant transcriptomic compendium for Salmonella enterica Serovar Typhimurium. *Cell Host & Microbe*, *14*(6), 683–695.

Kröger, C., Dillon, S. C., Cameron, A. D., Papenfort, K., Sivasankaran, S. K., Hokamp, K., ... & Hinton, J. C. (2012). The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *Proceedings of the National Academy of Sciences*, 109(20), E1277-E1286.

Kronstad, J. W. (2006). Serial analysis of gene expression in eukaryotic pathogens. *Infectious Disorders-Drug Targets (Formerly Current Drug Targets-Infectious Disorders)*, 6(3), 281-297.

Kuchina, A., Brettner, L. M., Paleologu, L., Roco, C. M., Rosenberg, A. B., Carignano, A., Kibler, R., Hirano, M., DePaolo, R. W., & Seelig, G. (2021). Microbial single-cell RNA sequencing by split-pool barcoding. *Science*, 371(6531).

Kuhn, M., & Others. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26.

Kumar, R., Shah, P., Swiatlo, E., Burgess, S. C., Lawrence, M. L., & Nanduri, B. (2010). Identification of novel non-coding small RNAs from Streptococcus pneumoniae TIGR4 using high-resolution genome tiling arrays. *BMC Genomics*, *11*, 350.

Kumar, S. S., Tandberg, J. I., Penesyan, A., Elbourne, L. D. H., Suarez-Bosche, N., Don, E., Skadberg, E., Fenaroli, F., Cole, N., Winther-Larsen, H. C., & Paulsen, I. T. (2018). Dual Transcriptomics of Host-Pathogen Interaction of Cystic Fibrosis Isolate Pseudomonas aeruginosa PASS1 With Zebrafish. *Frontiers in Cellular and Infection Microbiology*, *8*, 406.

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PloS One*, *12*(5), e0177459.

Lai, Y., Cui, L., Babunovic, G. H., Fortune, S. M., Doench, J. G., & Lu, T. K. (2021). High-Throughput CRISPR Screens To Dissect Macrophage-Shigella Interactions. *mBio*, 12(6), e0215821.

LaMonte, G. M., Orjuela-Sanchez, P., Calla, J., Wang, L. T., Li, S., Swann, J., Cowell, A. N., Zou, B. Y., Abdel-Haleem Mohamed, A. M., Villa Galarce, Z. H., Moreno, M., Tong Rios, C., Vinetz, J. M., Lewis, N., & Winzeler, E. A. (2019). Dual RNA-seq identifies human mucosal immunity protein Mucin-13 as a hallmark of Plasmodium exoerythrocytic infection. *Nature Communications*, *10*(1), 488.

La, M.-V., Raoult, D., & Renesto, P. (2008). Regulation of whole bacterial pathogen transcription within infected hosts. *FEMS Microbiology Reviews*, *32*(3), 440–460.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25.

LaRock, D. L., Chaudhary, A., & Miller, S. I. (2015). Salmonellae interactions with host processes. *Nature Reviews. Microbiology*, *13*(4), 191–205.

Lawrence M, Gentleman R, Carey V (2009). "rtracklayer: an R package for interfacing with genome browsers." *Bioinformatics*, 25, 1841-1842.

Lechner, M., Hernandez-Rosales, M., Doerr, D., Wieseke, N., Thévenin, A., Stoye, J., Hartmann, R. K., Prohaska, S. J., & Stadler, P. F. (2014). Orthology detection combining clustering and synteny for very large datasets. *PloS One*, *9*(8), e105015.

Lee, H. J., Georgiadou, A., Walther, M., Nwakanma, D., Stewart, L. B., Levin, M., Otto, T. D., Conway, D. J., Coin, L. J., & Cunnington, A. J. (2018). Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria. *Science Translational Medicine*, *10*(447).

Lee, J.-H., Cho, N.-H., Kim, S.-Y., Bang, S.-Y., Chu, H., Choi, M.-S., & Kim, I.-S. (2008). Fibronectin facilitates the invasion of Orientia tsutsugamushi into host cells through interaction with a 56-kDa type-specific antigen. *The Journal of Infectious Diseases*, *198*(2), 250–257.

Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics*, *18*(3), 530–536.

Liang, Y., & Kelemen, A. (2018). Dynamic modeling and network approaches for omics time course data: overview of computational approaches and applications. *Briefings in Bioinformatics*, 19(5), 1051–1068.

Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, *41*(10), e108.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* , *30*(7), 923–930.

Liao, Z.-X., Ni, Z., Wei, X.-L., Chen, L., Li, J.-Y., Yu, Y.-H., Jiang, W., Jiang, B.-L., He, Y.-Q., & Huang, S. (2019). Dual RNA-seq of Xanthomonas oryzae pv. oryzicola infecting rice reveals novel insights into bacterial-plant interaction. *PloS One*, *14*(4), e0215039.

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323.

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* , *26*(4), 493–500.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* , *34*(18), 3094–3100.

Liu, Q., Lei, J., Darby, A. C., & Kadowaki, T. (2020). Trypanosomatid parasite dynamically changes the transcriptome during infection and modifies honey bee physiology. *Communications Biology*, 3(1), 51.

Liu, Z., Li, Y., Pan, Y., Wang, L., Ouellet, T., & Fobert, P. (2019). Strategy in wheat-Fusarium dual-genome RNA-seq data processing. *bioRxiv*.

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. *methods*, 25(4), 402-408.

Li, W., Wang, X., Li, C., Sun, J., Li, S., & Peng, M. (2019). Dual species transcript profiling during the interaction between banana (Musa acuminata) and the fungal pathogen Fusarium oxysporum f. sp. cubense. *BMC Genomics*, *20*(1), 519.

Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H., Bork, P., Glass, J. I., Serrano, L., & Lluch-Senar, M. (2016). Bacterial antisense RNAs are mainly the product of transcriptional noise. *Science Advances*, *2*(3), e1501363.

Loman, N. J., Constantinidou, C., Chan, J. Z. M., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R., & Pallen, M. J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews. Microbiology*, *10*(9), 599–606.

Loman, N. J., & Pallen, M. J. (2015). Twenty years of bacterial genome sequencing. *Nature Reviews. Microbiology*, *13*(12), 787–794.

Lovegrove, F. E., Peña-Castillo, L., Mohammad, N., Liles, W. C., Hughes, T. R., & Kain, K. C. (2006). Simultaneous host and parasite expression profiling identifies tissue-specific transcriptional programs associated with susceptibility or resistance to experimental cerebral malaria. *BMC genomics*, 7(1), 1-17.

Love, M. I., Hogenesch, J. B., & Irizarry, R. A. (2016). Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature Biotechnology*, *34*(12), 1287–1291.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, *13*(5), e1005457.

Lunden, K., Danielsson, M., Durling, M. B., Ihrmark, K., Gorriz, M. N., Stenlid, J., ... & Elfstrand, M. (2015). Transcriptional Responses Associated with Virulence and Defence in the Interaction between Heterobasidion annosum s. s. and Norway Spruce. *PLoS One*, 10(7), e0131182.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., ... & Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1), 2047-217X.

McKinney, W. (2010, June). Data structures for statistical computing in python. *In Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51-56).

Macmanes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*, *5*, 13.

Mahmood, K., Orabi, J., Kristensen, P. S., Sarup, P., Jørgensen, L. N., & Jahoor, A. (2020). De novo transcriptome assembly, functional annotation, and expression profiling of rye (Secale cereale L.) hybrids inoculated with ergot (Claviceps purpurea). *Scientific Reports*, *10*(1), 13475.

Mangan, J. A., Monahan, I. M., Wilson, M. A., Schnappinger, D., Schoolnik, G. K., & Butcher, P. D. (1999). The expression profile of Mycobacterium tuberculosis infecting the human monocytic cell line THP-1 using whole genome microarray analysis. *Nature Genetics*, *23*(3), 61–61.

Mao, C., Bhardwaj, K., Sharkady, S. M., Fish, R. I., Driscoll, T., Wower, J., Zwieb, C., Sobral, B. W. S., & Williams, K. P. (2009). Variations on the tmRNA gene. *RNA Biol.*, 6, 355–361.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9), 1509-1517.

Marsh, J. W., Hayward, R. J., Shetty, A. C., Mahurkar, A., Humphrys, M. S., & Myers, G. S. A. (2018). Bioinformatic analysis of bacteria and host cell dual RNA-sequencing experiments. *Briefings in Bioinformatics*, *19*(6), 1115–1129.

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews. Genetics*, *12*(10), 671–682.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12.

Mäurer, A. P., Mehlitz, A., Mollenkopf, H. J., & Meyer, T. F. (2007). Gene Expression Profiles of Chlamydophila pneumoniae during the Developmental Cycle and Iron Depletion–Mediated Persistence. *PLoS Pathogens*, *3*(6), e83.

Maulding, N. D., Seiler, S., Pearson, A., Kreusser, N., & Stuart, J. M. (2022). Dual RNA-Seq analysis of SARS-CoV-2 correlates specific human transcriptional response pathways directly to viral expression. *Scientific reports*, 12(1), 1329.

Mavromatis, C. H., Bokil, N. J., Totsika, M., Kakkanat, A., Schaale, K., Cannistraci, C. V., Ryu, T., Beatson, S. A., Ulett, G. C., Schembri, M. A., Sweet, M. J., & Ravasi, T. (2015). The co-transcriptome of uropathogenic Escherichia coli-infected mouse macrophages reveals new insights into host-pathogen interactions. *Cellular Microbiology*, *17*(5), 730–746.

Mayhew, M. B., Buturovic, L., Luethy, R., Midic, U., Moore, A. R., Roque, J. A., Shaller, B. D., Asuni, T., Rawling, D., Remmel, M., Choi, K., Wacker, J., Khatri, P., Rogers, A. J., & Sweeney, T. E. (2020). A generalizable 29-mRNA neural-network classifier for acute bacterial and viral infections. *Nature Communications*, 11(1), 1177.

McClure, R., Balasubramanian, D., Sun, Y., Bobrovskyy, M., Sumby, P., Genco, C. A., Vanderpool, C. K., & Tjaden, B. (2013). Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Research*, *41*(14), e140.

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., … Weinstock, G. M. (2004). Complete genome sequence of Rickettsia typhi and comparison with sequences of other rickettsiae. *Journal of Bacteriology*, *186*(17), 5842–5855.

Mejía-Almonte, C., Busby, S. J. W., Wade, J. T., van Helden, J., Arkin, A. P., Stormo, G. D., Eilbeck, K., Palsson, B. O., Galagan, J. E., & Collado-Vides, J. (2020). Redefining fundamental concepts of transcription initiation in bacteria. *Nature Reviews. Genetics*, *21*(11), 699–714.

Merrell, D. S., Butler, S. M., Qadri, F., Dolganov, N. A., Alam, A., Cohen, M. B., Calderwood, S. B., Schoolnik, G. K., & Camilli, A. (2002). Host-induced epidemic spread of the cholera bacterium. *Nature*, *417*(6889), 642–645.

Mika-Gospodorz, B., Giengkam, S., Westermann, A. J., Wongsantichon, J., Kion-Crosby, W., Chuenklin, S., Wang, L. C., Sunyakumthorn, P., Sobota, R. M., Subbian, S., Vogel, J., Barquist, L., & Salje, J. (2020). Dual RNA-seq of Orientia tsutsugamushi informs on host-pathogen interactions for this neglected intracellular human pathogen. *Nature Communications*, *11*(1), 3363.

Millar, J. A., & Raghavan, R. (2021). Modulation of Bacterial Fitness and Virulence Through Antisense RNAs. *Frontiers in cellular and infection microbiology*, 10, 596277.

Miller, M. B., & Tang, Y.-W. (2009). Basic concepts of microarrays and potential applications in clinical microbiology. *Clinical Microbiology Reviews*, *22*(4), 611–633.

Min, C.-K., Kwon, Y.-J., Ha, N.-Y., Cho, B.-A., Kim, J.-M., Kwon, E.-K., Kim, Y.-S., Choi, M.-S., Kim, I.-S., & Cho, N.-H. (2014). Multiple Orientia tsutsugamushi ankyrin repeat proteins interact with SCF1 ubiquitin ligase complex and eukaryotic elongation factor 1 α. *PloS One*, *9*(8), e105652.

Minhas, V., Aprianto, R., McAllister, L. J., Wang, H., David, S. C., McLean, K. T., Comerford, I., McColl, S. R., Paton, J. C., Veening, J.-W., & Trappetti, C. (2020). In vivo dual RNA-seq reveals that neutrophil recruitment underlies differential tissue tropism of Streptococcus pneumoniae. *Communications Biology*, *3*(1), 293.

Mockler, T. C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S. E., & Ecker, J. R. (2005). Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, *85*(1), 1–15.

Mohamed, A. R., Andrade, N., Moya, A., Chan, C. X., Negri, A. P., Bourne, D. G., Ying, H., Ball, E. E., & Miller, D. J. (2020). Dual RNA-sequencing analyses of a coral and its native symbiont during the establishment of symbiosis. *Molecular Ecology*, *29*(20), 3921–3937.

Montoya, D. J., Andrade, P., Silva, B. J. A., Teles, R. M. B., Ma, F., Bryson, B., Sadanand, S., Noel, T., Lu, J., Sarno, E., Arnvig, K. B., Young, D., Lahiri, R., Williams, D. L., Fortune, S., Bloom, B. R., Pellegrini, M., & Modlin, R. L. (2019). Dual RNA-Seq of Human Leprosy Lesions Identifies Bacterial Determinants Linked to Host Immune Response. *Cell Reports*, *26*(13), 3574–3585.e3.

Morozova, O., & Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5), 255-264.

Motley, S. T., Morrow, B. J., Liu, X., Dodge, I. L., Vitiello, A., Ward, C. K., & Shaw, K. J. (2004). Simultaneous analysis of host and pathogen interactions during an in vivo infection reveals local induction of host acute phase response proteins, a novel bacterial stress response, and evidence of a host-imposed metal ion limited environment. *Cellular microbiology*, 6(9), 849-865.

Musungu, B., Bhatnagar, D., Quiniou, S., Brown, R. L., Payne, G. A., O'Brian, G., Fakhoury, A. M., & Geisler, M. (2020). Use of Dual RNA-seq for Systems Biology Analysis of Zea mays and Aspergillus flavus Interaction. *Frontiers in Microbiology*, *11*, 853.

Nakayama, K., Kurokawa, K., Fukuhara, M., Urakami, H., Yamamoto, S., Yamazaki, K., Ogura, Y., Ooka, T., & Hayashi, T. (2010). Genome comparison and phylogenetic analysis of Orientia tsutsugamushi strains. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *17*(5), 281–291.

Nakayama, K., Yamashita, A., Kurokawa, K., Morimoto, T., Ogawa, M., Fukuhara, M., ... & Hayashi, T. (2008). The whole-genome sequencing of the obligate intracellular bacterium Orientia tsutsugamushi revealed massive gene amplification during reductive genome evolution. *DNA research*, 15(4), 185-199.

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933–2935.

Nejman, D., Livyatan, I., Fuks, G., Gavert, N., Zwang, Y., Geller, L. T., Rotter-Maskowitz, A., Weiser, R., Mallel, G., Gigi, E., Meltser, A., Douglas, G. M., Kamer, I., Gopalakrishnan, V., Dadosh, T., Levin-Zaidman, S., Avnet, S., Atlan, T., Cooper, Z. A., Arora, R., … Straussman, R. (2020). The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science*, 368(6494), 973–980.

Nekrutenko, A., & Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature Reviews. Genetics*, *13*(9), 667–672.

Nuss, A. M., Beckstette, M., Pimenova, M., Schmühl, C., Opitz, W., Pisano, F., Heroven, A. K., & Dersch, P. (2017). Tissue dual RNA-seq allows fast discovery of infection-specific functions and riboregulators shaping host-pathogen transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(5), E791–E800.

Oosthuizen, J. L., Gomez, P., Ruan, J., Hackett, T. L., Moore, M. M., Knight, D. A., & Tebbutt, S. J. (2011). Dual organism transcriptomics of airway epithelial cells interacting with conidia of Aspergillus fumigatus. *PloS One*, *6*(5), e20527.

Orikaza, C. M., Pessoa, C. C., Paladino, F. V., Florentino, P. T., Barbiéri, C. L., Goto, H., ... & Real, F. (2020). Dual host-intracellular parasite transcriptome of enucleated cells hosting Leishmania amazonensis: control of half-life of host cell transcripts by the parasite. *Infection and immunity*, 88(11), e00261-20.

Otto, C., Stadler, P. F., & Hoffmann, S. (2014). Lacking alignments? The next-generation sequencing mapper segemehl revisited. *Bioinformatics* , *30*(13), 1837–1843.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* , *31*(22), 3691–3693.

Paris, D. H., Aukkanit, N., Jenjaroen, K., Blacksell, S. D., & Day, N. P. J. (2009). A highly sensitive quantitative real-time PCR assay based on the groEL gene of contemporary Thai strains of Orientia tsutsugamushi. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *15*(5), 488–495.

Park, S.-J., Kumar, M., Kwon, H.-I., Seong, R.-K., Han, K., Song, J.-M., Kim, C.-J., Choi, Y.-K., & Shin, O. S. (2015). Dynamic changes in host gene expression associated with H5N8 avian influenza virus infection in mice. *Scientific Reports*, *5*, 16512.

Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS One*, *7*(2), e30619.

Patir, A., Gossner, A., Ramachandran, P., Alves, J., Freeman, T. C., Henderson, N. C., Watson, M., & Hassan, M. A. (2020). Single-cell RNA-seq reveals CD16- monocytes as key regulators of human monocyte transcriptional response to Toxoplasma. *Scientific Reports*, *10*(1), 21047.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419.

Patro, R., Mount, S. M., & Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, *32*(5), 462–464.

Peer, X., & An, G. (2014). Agent-based model of fecal microbial transplant effect on bile acid metabolism on suppressing Clostridium difficile infection: an example of agent-based modeling of intestinal bacterial infection. *Journal of Pharmacokinetics and Pharmacodynamics*, 41(5), 493–507.

Penaranda, C., Chumbler, N. M., & Hung, D. T. (2021). Dual transcriptional analysis reveals adaptation of host and pathogen to intracellular survival of Pseudomonas aeruginosa associated with urinary tract infection. *PLoS Pathogens*, *17*(4), e1009534.

Pérez-Losada, M., Castro-Nallar, E., Bendall, M. L., Freishtat, R. J., & Crandall, K. A. (2015). Dual Transcriptomic Profiling of Host and Microbiota during Health and Disease in Pediatric Asthma. *PloS One*, *10*(6), e0131819.

Perez, N., Treviño, J., Liu, Z., Ho, S. C. M., Babitzke, P., & Sumby, P. (2009). A genome-wide analysis of small regulatory RNAs in the human pathogen group A Streptococcus. *PloS One*, *4*(11), e7668.

Pérez-Rubio, P., Lottaz, C., & Engelmann, J. C. (2019). FastqPuri: high-performance preprocessing of RNA-seq data. *BMC Bioinformatics*, *20*(1), 226.

Piccolo, S. R., & Frampton, M. B. (2016). Tools and techniques for computational reproducibility. *GigaScience*, *5*(1), 30.

Pierrelée, M., Reynders, A., Lopez, F., Moqrich, A., Tichit, L., & Habermann, B. H. (2021). Introducing the novel Cytoscape app TimeNexus to analyze time-series data using temporal MultiLayer Networks (tMLNs). *Scientific Reports*, 11(1), 13691

Pisu, D., Huang, L., Grenier, J. K., & Russell, D. G. (2020). Dual RNA-Seq of Mtb-Infected Macrophages In Vivo Reveals Ontologically Distinct Host-Pathogen Interactions. *Cell Reports*, *30*(2), 335–350.e4.

Pittman, K. J., Aliota, M. T., & Knoll, L. J. (2014). Dual transcriptional profiling of mice and Toxoplasma gondii during acute and chronic infection. *BMC Genomics*, *15*, 806.

Raffatellu, M., Santos, R. L., Verhoeven, D. E., George, M. D., Wilson, R. P., Winter, S. E., Godinez, I., Sankaran, S., Paixao, T. A., Gordon, M. A., Kolls, J. K., Dandekar, S., & Bäumler, A. J. (2008). Simian immunodeficiency virus–induced mucosal interleukin-17 deficiency promotes Salmonella dissemination from the gut. *Nature Medicine*, *14*(4), 421–428.

Raghavan, R., Sloan, D. B., & Ochman, H. (2012). Antisense transcription is pervasive but rarely conserved in enteric bacteria. *MBio*, 3(4), e00156-12.

Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine*, *8*(1), 7.

Rana, S. B., Zadlock, F. J., 4th, Zhang, Z., Murphy, W. R., & Bentivegna, C. S. (2016). Comparison of De Novo Transcriptome Assemblers and k-mer Strategies Using the Killifish, Fundulus heteroclitus. *PloS One*, *11*(4), e0153104.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., & Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, *14*(9), R95.

Ray-Soni, A., Bellecourt, M. J., & Landick, R. (2016). Mechanisms of Bacterial Transcription Termination: All Good Things Must End. *Annual Review of Biochemistry*, *85*, 319–347.

Ren, L., Zhang, R., Rao, J., Xiao, Y., Zhang, Z., Yang, B., Cao, D., Zhong, H., Ning, P., Shang, Y., Li, M., Gao, Z., & Wang, J. (2018). Transcriptionally Active Lung Microbiome and Its Association with Bacterial Biomass and Host Inflammatory Status. *mSystems*, 3(5).

Revel, A. T., Talaat, A. M., & Norgard, M. V. (2002). DNA microarray analysis of differential gene expression in Borrelia burgdorferi, the Lyme disease spirochete. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(3), 1562–1567.

Rienksma, R. A., Suarez-Diez, M., Mollenkopf, H.-J., Dolganov, G. M., Dorhoi, A., Schoolnik, G. K., Martins Dos Santos, V. A., Kaufmann, S. H., Schaap, P. J., & Gengenbacher, M. (2015). Comprehensive insights into transcriptional adaptation of intracellular mycobacteria by microbe-enriched dual RNA sequencing. *BMC Genomics*, *16*, 34.

Ritchie, N. D., & Evans, T. J. (2019). Dual RNA-seq in Streptococcus pneumoniae Infection Reveals Compartmentalized Neutrophil Responses in Lung and Pleural Space. *mSystems*, *4*(4).

Van den Berge, K., Hembach, K. M., Soneson, C., Tiberi, S., Clement, L., Love, M. I., ... & Robinson, M. D. (2019). RNA sequencing data: Hitchhiker's guide to expression analysis. *Annual Review of Biomedical Data Science*, 2, 139-173.

Robert, C., & Watson, M. (2015). Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biology*, *16*, 177.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, *12*(3), R22.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* , *26*(1), 139–140.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., & Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics*, *13*, 484.

Rodino, K. G., VieBrock, L., Evans, S. M., Ge, H., Richards, A. L., & Carlyon, J. A. (2017). Orientia tsutsugamushi modulates endoplasmic reticulum-associated degradation to benefit its growth. Infection and immunity, 86(1), e00596-17.

Roehr, J. T., Dieterich, C., & Reinert, K. (2017). Flexbar 3.0--SIMD and multicore parallelization. *Bioinformatics* , *33*(18), 2941–2942.

Rossi, E., Falcone, M., Molin, S., & Johansen, H. K. (2018). High-resolution in situ transcriptomics of Pseudomonas aeruginosa unveils genotype independent patho-phenotypes in cystic fibrosis lungs. *Nature Communications*, *9*(1), 3459.

Roux, C. M., Butler, B. P., Chau, J. Y., Paixao, T. A., Cheung, K. W., Santos, R. L., Luckhart, S., & Tsolis, R. M. (2010). Both hemolytic anemia and malaria parasite-specific factors increase susceptibility to Nontyphoidal Salmonella enterica serovar typhimurium infection in mice. *Infection and Immunity*, *78*(4), 1520–1527.

Sáenz-Lahoya, S., Bitarte, N., García, B., Burgui, S., Vergara-Irigaray, M., Valle, J., Solano, C., Toledo-Arana, A., & Lasa, I. (2019). Noncontiguous operon is a genetic organization for coordinating bacterial gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 116(5), 1733–1738.

Sahraeian, S. M. E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P. T., Au, K. F., Bani Asadi, N., Gerstein, M. B., Wong, W. H., Snyder, M. P., Schadt, E., & Lam, H. Y. K. (2017). Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nature Communications*, *8*(1), 59.

Saliba, A.-E., C Santos, S., & Vogel, J. (2017). New RNA-seq approaches for the study of bacterial pathogens. *Current Opinion in Microbiology*, *35*, 78–87.

Saliba, A.-E., Li, L., Westermann, A. J., Appenzeller, S., Stapels, D. A. C., Schulte, L. N., Helaine, S., & Vogel, J. (2016). Single-cell RNA-seq ties macrophage polarization to growth rate of intracellular Salmonella. *Nature Microbiology*, 2, 16206.

Salje, J. (2017). Orientia tsutsugamushi: A neglected but fascinating obligate intracellular bacterial pathogen. *PLoS Pathogens*, *13*(12), e1006657.

Salje, J. (2021). Cells within cells: Rickettsiales and the obligate intracellular bacterial lifestyle. *Nature Reviews Microbiology*, 1-16.

Santos, R. L., Zhang, S., Tsolis, R. M., Kingsley, R. A., Adams, L. G., & Bäumler, A. J. (2001). Animal models of Salmonella infections: enteritis versus typhoid fever. *Microbes and Infection / Institut Pasteur*, *3*(14-15), 1335–1344.

Sarantopoulou, D., Tang, S. Y., Ricciotti, E., Lahens, N. F., Lekkas, D., Schug, J., Guo, X. S., Paschos, G. K., FitzGerald, G. A., Pack, A. I., & Grant, G. R. (2019). Comparative evaluation of RNA-Seq library preparation methods for strand-specificity and low input. *Scientific Reports*, *9*(1), 13477.

Sarkar, H., Zakeri, M., Malik, L., & Patro, R. (2018). Towards Selective-Alignment: Bridging the Accuracy Gap between Alignment-Based and Alignment-Free Transcript Quantification. *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '18*, 27–36.

Saxena, K., Simon, L. M., Zeng, X.-L., Blutt, S. E., Crawford, S. E., Sastri, N. P., Karandikar, U. C., Ajami, N. J., Zachos, N. C., Kovbasnjuk, O., Donowitz, M., Conner, M. E., Shaw, C. A., & Estes, M. K. (2017). A paradox of transcriptional and functional innate interferon responses of human intestinal enteroids to enteric virus infection. *Proceedings of the National Academy of Sciences of the United States of America*, 114(4), E570–E579.

Schroeder, C. L. C., Narra, H. P., Rojas, M., Sahni, A., Patel, J., Khanipov, K., Wood, T. G., Fofanov, Y., & Sahni, S. K. (2015). Bacterial small RNAs in the Genus Rickettsia. *BMC Genomics*, *16*, 1075.

Schroeder, C. L. C., Narra, H. P., Sahni, A., Rojas, M., Khanipov, K., Patel, J., Shah, R., Fofanov, Y., & Sahni, S. K. (2016). Identification and Characterization of Novel Small RNAs in Rickettsia prowazekii. *Frontiers in Microbiology*, *7*, 859.

Schulte, L. N., Schweinlin, M., Westermann, A. J., Janga, H., Santos, S. C., Appenzeller, S., Walles, H., Vogel, J., & Metzger, M. (2020). An Advanced Human Intestinal Coculture Model Reveals Compartmentalized Host and Pathogen Strategies during Salmonella Infection. *mBio*, *11*(1).

Schulze, S., Henkel, S. G., Driesch, D., Guthke, R., & Linde, J. (2015). Computational prediction of molecular pathogen-host interactions based on dual transcriptome data. *Frontiers in Microbiology*, *6*, 65.

Schulze, S., Schleicher, J., Guthke, R., & Linde, J. (2016). How to Predict Molecular Interactions between Species? *Frontiers in Microbiology*, *7*, 442.

Seal, J. B., Alverdy, J. C., Zaborina, O., & An, G. (2011). Agent-based dynamic knowledge representation of Pseudomonas aeruginosa virulence activation in the stressed gut: Towards characterizing host-pathogen interactions in gut-derived sepsis. *Theoretical Biology & Medical Modelling*, 8, 33.

Sedlyarova, N., Shamovsky, I., Bharati, B. K., Epshtein, V., Chen, J., Gottesman, S., Schroeder, R., & Nudler, E. (2016). sRNA-Mediated Control of Transcription Termination in E. coli. *Cell*, *167*(1), 111–121.e13.

Seelbinder, B., Wallstabe, J., Marischen, L., Weiss, E., Wurster, S., Page, L., Löffler, C., Bussemer, L., Schmitt, A.-L., Wolf, T., Linde, J., Cicin-Sain, L., Becker, J., Kalinke, U., Vogel, J., Panagiotou, G., Einsele, H., Westermann, A. J., Schäuble, S., & Loeffler, J. (2020). Triple RNA-Seq Reveals Synergy in a Human Virus-Fungus Co-infection Model. *Cell Reports*, *33*(7), 108389.

SEQC Consortium. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature biotechnology*, 32(9), 903.

Shao, W., Price, M. N., Deutschbauer, A. M., Romine, M. F., & Arkin, A. P. (2014). Conservation of transcription start sites within genes across a bacterial genus. *mBio*, *5*(4), e01398–14.

Sharkady, S. M., & Williams, K. P. (2004). A third lineage with two-piece tmRNA. *Nucleic acids research*, 32(15), 4531-4538.

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. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, *464*(7286), 250–255.

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. (2016). IL-33-Dependent Endothelial Activation Contributes to Apoptosis and Renal Injury in Orientia tsutsugamushi-Infected Mice. *PLoS Neglected Tropical Diseases*, *10*(3), e0004467.

Silva, I. J., Barahona, S., Eyraud, A., Lalaouna, D., Figueroa-Bossi, N., Massé, E., & Arraiano, C. M. (2019). SraL sRNA interaction regulates the terminator by preventing premature transcription termination of rho mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(8), 3042–3051.

Simon, L. M., Karg, S., Westermann, A. J., Engel, M., Elbehery, A. H., Hense, B., ... & Theis, F. J. (2018). MetaMap: an atlas of metatranscriptomic reads in human disease-related RNA-seq data. Gigascience, 7(6), giy070.

Soneson, C., Love, M. I., & Robinson, M. D. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4.

Spies, D., Renz, P. F., Beyer, T. A., & Ciaudo, C. (2019). Comparative analysis of differential gene expression tools for RNA sequencing time course data. *Briefings in Bioinformatics*, 20(1), 288–298.

Srikumar, S., Kröger, C., Hébrard, M., Colgan, A., Owen, S. V., Sivasankaran, S. K., Cameron, A. D. S., Hokamp, K., & Hinton, J. C. D. (2015). RNA-seq Brings New Insights to the Intra-Macrophage Transcriptome of Salmonella Typhimurium. *PLoS Pathogens*, *11*(11), e1005262.

Srivastava, A., Malik, L., Sarkar, H., Zakeri, M., Almodaresi, F., Soneson, C., Love, M. I., Kingsford, C., & Patro, R. (2020). Alignment and mapping methodology influence transcript abundance estimation. *Genome Biology*, *21*(1), 239.

Srivastava, A., Sarkar, H., Gupta, N., & Patro, R. (2016). RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics* , *32*(12), i192–i200.

Stapels, D. A. C., Hill, P. W. S., Westermann, A. J., Fisher, R. A., Thurston, T. L., Saliba, A.-E., Blommestein, I., Vogel, J., & Helaine, S. (2018). Salmonella persisters undermine host immune defenses during antibiotic treatment. *Science*, *362*(6419), 1156–1160.

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews. Genetics*, *20*(11), 631–656.

Steuerman, Y., Cohen, M., Peshes-Yaloz, N., Valadarsky, L., Cohn, O., David, E., Frishberg, A., Mayo, L., Bacharach, E., Amit, I., & Gat-Viks, I. (2018). Dissection of Influenza Infection In Vivo by Single-Cell RNA Sequencing. *Cell Systems*, *6*(6), 679–691.e4.

Storz, G., Vogel, J., & Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Molecular Cell*, *43*(6), 880–891.

Strong, M. J., Xu, G., Coco, J., Baribault, C., Vinay, D. S., Lacey, M. R., Strong, A. L., Lehman, T. A., Seddon, M. B., Lin, Z., Concha, M., Baddoo, M., Ferris, M., Swan, K. F., Sullivan, D. E., Burow, M. E., Taylor, C. M., & Flemington, E. K. (2013). Differences in gastric carcinoma microenvironment stratify according to EBV infection intensity: implications for possible immune adjuvant therapy. *PLoS Pathogens*, *9*(5), e1003341.

Subbian, S., Bandyopadhyay, N., Tsenova, L., O'Brien, P., Khetani, V., Kushner, N. L., Peixoto, B., Soteropoulos, P., Bader, J. S., Karakousis, P. C., Fallows, D., & Kaplan, G. (2013). Early innate immunity determines outcome of Mycobacterium tuberculosis pulmonary infection in rabbits. *Cell Communication and Signaling: CCS*, *11*, 60.

Sullivan, M. J., Petty, N. K., & Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* , *27*(7), 1009–1010.

Svetlanov, A., Puri, N., Mena, P., Koller, A., & Karzai, A. W. (2012). Francisella tularensis tmRNA system mutants are vulnerable to stress, avirulent in mice, and provide effective immune protection. *Molecular Microbiology*, *85*(1), 122–141.

Tanner, J. R., & Kingsley, R. A. (2018). Evolution of Salmonella within hosts. Trends in microbiology, 26(12), 986-998.

Tantibhedhyangkul, W., Amara, A. B., Textoris, J., Gorvel, L., Ghigo, E., Capo, C., & Mege, J. L. (2013). Orientia tsutsugamushi, the causative agent of scrub typhus, induces an inflammatory program in human macrophages. *Microbial pathogenesis*, 55, 55-63.

Tantibhedhyangkul, W., Prachason, T., Waywa, D., El Filali, A., Ghigo, E., Thongnoppakhun, W., Raoult, D., Suputtamongkol, Y., Capo, C., Limwongse, C., & Mege, J.-L. (2011). Orientia tsutsugamushi stimulates an original gene expression program in monocytes: relationship with gene expression in patients with scrub typhus. *PLoS Neglected Tropical Diseases*, *5*(5), e1028.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Research*, *21*(12), 2213–2223.

Taylor, A. J., Paris, D. H., & Newton, P. N. (2015). A Systematic Review of Mortality from Untreated Scrub Typhus (Orientia tsutsugamushi). *PLoS Neglected Tropical Diseases*, *9*(8), e0003971.

Tenenbaum, D. (2019). KEGGREST: Client-side REST access to KEGG. R package version 1.18.1.

Teng, M., Love, M. I., Davis, C. A., Djebali, S., Dobin, A., Graveley, B. R., Li, S., Mason, C. E., Olson, S., Pervouchine, D., Sloan, C. A., Wei, X., Zhan, L., & Irizarry, R. A. (2016). A benchmark for RNA-seq quantification pipelines. *Genome Biology*, *17*, 74.

Tenor, J. L., McCormick, B. A., Ausubel, F. M., & Aballay, A. (2004). Caenorhabditis elegans-based screen identifies Salmonella virulence factors required for conserved host-pathogen interactions. *Current Biology: CB*, *14*(11), 1018–1024.

Thänert, R., Goldmann, O., Beineke, A., & Medina, E. (2017). Host-inherent variability influences the transcriptional response of Staphylococcus aureus during in vivo infection. *Nature Communications*, *8*, 14268.

Thänert, R., Itzek, A., Hoßmann, J., Hamisch, D., Madsen, M. B., Hyldegaard, O., ... & Pieper, D. H. (2019). Molecular profiling of tissue biopsies reveals unique signatures associated with streptococcal necrotizing soft tissue infections. *Nature communications*, *10*(1), 1-15.

Thomason, M. K., & Storz, G. (2010). Bacterial antisense RNAs: how many are there, and what are they doing?. *Annual review of genetics*, 44, 167-188.

Tierney, L., Linde, J., Müller, S., Brunke, S., Molina, J. C., Hube, B., Schöck, U., Guthke, R., & Kuchler, K. (2012). An Interspecies Regulatory Network Inferred from Simultaneous RNA-seq of Candida albicans Invading Innate Immune Cells. *Frontiers in Microbiology*, *3*, 85.

Tjaden, B. (2015). De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biology*, *16*, 1.

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. (2009). The Listeria transcriptional landscape from saprophytism to virulence. *Nature*, *459*(7249), 950–956.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* , *25*(9), 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–578.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515.

Tsai, M.-H., Chang, C.-H., Tsai, R.-K., Hong, Y.-R., Chuang, T.-H., Fan, K.-T., Peng, C.-W., Wu, C.-Y., Hsu, W.-L., Wang, L.-S., Chen, L.-K., & Yu, H.-S. (2016). Cross-Regulation of Proinflammatory Cytokines by Interleukin-10 and miR-155 in Orientia tsutsugamushi-Infected Human Macrophages Prevents Cytokine Storm. In *Journal of Investigative Dermatology* (Vol. 136, Issue 7, pp. 1398–1407).

Tsolis, R. M., Xavier, M. N., Santos, R. L., & Bäumler, A. J. (2011). How to become a top model: impact of animal experimentation on human Salmonella disease research. *Infection and Immunity*, *79*(5), 1806–1814.

Valenzuela-Miranda, D., & Gallardo-Escárate, C. (2018). Dual RNA-Seq Uncovers Metabolic Amino Acids Dependency of the Intracellular Bacterium Piscirickettsia salmonis Infecting Atlantic Salmon. *Frontiers in Microbiology*, *9*, 2877.

Van Der Sar, A. M., Musters, R. J. P., Van Eeden, F. J. M., Appelmelk, B. J., Vandenbroucke-Grauls, C. M. J. E., & Bitter, W. (2003). Zebrafish embryos as a model host for the real time analysis of Salmonella typhimurium infections. *Cellular Microbiology*, *5*(9), 601–611.

VanderPlas, J., Granger, B., Heer, J., Moritz, D., Wongsuphasawat, K., Satyanarayan, A., … Sievert, S. (2018). Altair: Interactive statistical visualizations for python. *Journal of Open Source Software*, 3(32), 1057.

Vannucci, F. A., Foster, D. N., & Gebhart, C. J. (2013). Laser microdissection coupled with RNA-seq analysis of porcine enterocytes infected with an obligate intracellular pathogen (Lawsonia intracellularis). *BMC Genomics*, *14*, 421.

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. (2015). Orientia tsutsugamushi ankyrin repeat-containing protein family members are Type 1 secretion system substrates that traffic to the host cell endoplasmic reticulum. *Frontiers in Cellular and Infection Microbiology*, *4*, 186.

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., Pfeil, J., Narkizian, J., Deran, A. D., Musselman-Brown, A., Schmidt, H., Amstutz, P., Craft, B., Goldman, M., Rosenbloom, K., Cline, M., O'Connor, B., Hanna, M., Birger, C., … Paten, B. (2017). Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology*, *35*(4), 314–316.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272

Vogel, J. (2003). RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. In *Nucleic Acids Research* (Vol. 31, Issue 22, pp. 6435–6443).

Waddell, S. J., Butcher, P. D., & Stoker, N. G. (2007). RNA profiling in host-pathogen interactions. *Current Opinion in Microbiology*, *10*(3), 297–302.

Wade, J. T., & Grainger, D. C. (2014). Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nature Reviews. Microbiology*, *12*(9), 647–653.

Wagner, E. G. H., & Romby, P. (2015). Small RNAs in bacteria and archaea: who they are, what they do, and how they do it. *Advances in genetics*, 90, 133-208.

Wang, Q., Shakoor, N., Boyher, A., Veley, K. M., Berry, J. C., Mockler, T. C., & Bart, R. S. (2021). Escalation in the host-pathogen arms race: A host resistance response corresponds to a heightened bacterial virulence response. *PLoS Pathogens*, *17*(1), e1009175.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63.

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.

Watson, K. G., & Holden, D. W. (2010). Dynamics of growth and dissemination of Salmonella in vivo. *Cellular Microbiology*, *12*(10), 1389–1397.

Wesolowska-Andersen, A., Everman, J. L., Davidson, R., Rios, C., Herrin, R., Eng, C., Janssen, W. J., Liu, A. H., Oh, S. S., Kumar, R., Fingerlin, T. E., Rodriguez-Santana, J., Burchard, E. G., & Seibold, M. A. (2017). Dual RNA-seq reveals viral Liu respiratory illness which are associated with changes in the airway transcriptome. *Genome Biology*, *18*(1), 12.

Westermann, A. J., Barquist, L., & Vogel, J. (2017). Resolving host–pathogen interactions by dual RNA-seq. *PLoS Pathogens*, *13*(2), e1006033.

Westermann, A. J., Förstner, K. U., Amman, F., Barquist, L., Chao, Y., Schulte, L. N., Müller, L., Reinhardt, R., Stadler, P. F., & Vogel, J. (2016). Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. *Nature*, *529*(7587), 496–501.

Westermann, A. J., Gorski, S. A., & Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nature Reviews. Microbiology*, *10*(9), 618–630.

Westermann, A. J., Venturini, E., Sellin, M. E., Förstner, K. U., Hardt, W.-D., & Vogel, J. (2019). The Major RNA-Binding Protein ProQ Impacts Virulence Gene Expression in Salmonella enterica Serovar Typhimurium. *mBio*, *10*(1), e02504–e02518.

Westermann, A. J., & Vogel, J. (2021). Cross-species RNA-seq for deciphering host–microbe interactions. *Nature Reviews. Genetics*, 22(6), 361–378.

Williams, C. R., Baccarella, A., Parrish, J. Z., & Kim, C. C. (2016). Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics*, *17*, 103.

Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, 7.

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J. S., Voigt, A. Y., Palleja, A., Ponnudurai, R., Sunagawa, S., Coelho, L. P., Schrotz-King, P., Vogtmann, E., Habermann, N., Niméus, E., Thomas, A. M., Manghi, P., Gandini, S., Serrano, D., … Zeller, G. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nature medicine*, 25(4), 679–689.

Withey, J. H., & Friedman, D. I. (2003). A salvage pathway for protein synthesis: tmRNA and trans-translation. *Annual Reviews in Microbiology*, 57(1), 101-123.

Wolf, T., Kämmer, P., Brunke, S., & Linde, J. (2018). Two's company: studying interspecies relationships with dual RNA-seq. *Current Opinion in Microbiology*, *42*, 7–12.

Wongsantichon, J., Jaiyen, Y., Dittrich, S., & Salje, J. (2020). Orientia tsutsugamushi. *Trends in Microbiology*, *0*(0).

Woodard, A., & Wood, D. O. (2011). Analysis of convergent gene transcripts in the obligate intracellular bacterium Rickettsia prowazekii. *PloS One*, *6*(1), e16537.

Wratten, L., Wilm, A., & Göke, J. (2021). Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature methods*, 18(10), 1161–1168.

Wu, C., Carta, R., & Zhang, L. (2005). Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic acids research*, 33(9), e84-e84.

Wu, D. C., Yao, J., Ho, K. S., Lambowitz, A. M., & Wilke, C. O. (2018). Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics*, *19*(1), 510.

Xu, G., Strong, M. J., Lacey, M. R., Baribault, C., Flemington, E. K., & Taylor, C. M. (2014). RNA CoMPASS: a dual approach for pathogen and host transcriptome analysis of RNA-seq datasets. *PloS One*, 9(2), e89445.

Yamagishi, J., Natori, A., Tolba, M. E. M., Mongan, A. E., Sugimoto, C., Katayama, T., Kawashima, S., Makalowski, W., Maeda, R., Eshita, Y., Tuda, J., & Suzuki, Y. (2014). Interactive transcriptome analysis of malaria patients and infecting Plasmodium falciparum. *Genome Research*, *24*(9), 1433–1444.

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., Zhao, F., & Zhu, B. (2013). HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics*, *14*, 33.

Yazawa, T., Kawahigashi, H., Matsumoto, T., & Mizuno, H. (2013). Simultaneous transcriptome analysis of Sorghum and Bipolaris sorghicola by using RNA-seq in combination with de novo transcriptome assembly. *PloS One*, *8*(4), e62460.

Yeung, A. T., Choi, Y. H., Lee, A. H., Hale, C., Ponstingl, H., Pickard, D., ... & Dougan, G. (2019). A genome-wide knockout screen in human macrophages identified host factors modulating Salmonella infection. *MBio*, 10(5), e02169-19

Yin, Y., & Zhou, D. (2018). Organoid and enteroid modeling of Salmonella infection. *Frontiers in cellular and infection microbiology*, 8, 102.

Yun, J.-H., Koh, Y.-S., Lee, K.-H., Hyun, J.-W., Choi, Y.-J., Jang, W.-J., Park, K.-H., Cho, N.-H., Seong, S.-Y., Choi, M.-S., & Kim, I.-S. (2005). Chemokine and cytokine production in susceptible C3H/HeN mice and resistant BALB/c mice during Orientia tsutsugamushi infection. *Microbiology and Immunology*, *49*(6), 551–557.

Yu, S. H., Vogel, J., & Förstner, K. U. (2018). ANNOgesic: a Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. *GigaScience*, 7(9), giy096.

Zapatka, M., Borozan, I., Brewer, D. S., Iskar, M., Grundhoff, A., Alawi, M., ... & Lichter, P. (2020). The landscape of viral associations in human cancers. *Nature genetics*, 52(3), 320-330.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. In *PLoS ONE* (Vol. 9, Issue 1, p. e78644).

Zheng, H., Brennan, K., Hernaez, M., & Gevaert, O. (2019). Benchmark of long non-coding RNA quantification for RNA sequencing of cancer samples. *GigaScience*, *8*(12).

Zhou, Q., Su, X., Wang, A., Xu, J., & Ning, K. (2013). QC-Chain: fast and holistic quality control method for next-generation sequencing data. *PloS One*, *8*(4), e60234.

Zimmermann, M., Kogadeeva, M., Gengenbacher, M., McEwen, G., Mollenkopf, H.-J., Zamboni, N., Kaufmann, S. H. E., & Sauer, U. (2017). Integration of Metabolomics and Transcriptomics Reveals a Complex Diet of Mycobacterium tuberculosis during Early Macrophage Infection. *mSystems*, *2*(4).

Appendix 1 Supplementary data for chapter 2

Supplementary data associated with this study can be found on the CD attached to this thesis or in the online version of the article (Mika-Gospodorz et al., 2020) with doi:10.1038/s41467-020-17094-8. The excel-formatted "Supplementary Data 1-22" file (41467_2020_17094_MOESM4_ESM.xlsx) can be found at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7335160/bin/41467_2020_17094_MOESM4_ESM.xlsx.

The following tables are a part of this thesis and can be found as sheets in the 41467_2020_17094_MOESM4_ESM.xlsx file:

- *Mapping statistics*: percentage of RNA-seq reads assigned to different RNA classes in Karp, UT176 and HUVEC in three replicates;
- *Karp duplicates*: groups of the Karp genes identified as duplicates by Salmon (Patro et al., 2017);
- *UT176 duplicates*: groups of the UT176 genes identified as duplicates by Salmon (Patro et al., 2017);
- *Karp expressed*: list of all genes expressed in Karp strain grown in HUVEC cells;
- *Karp highly exp*: list of all genes highly expressed in Karp strain grown in HUVEC cell;
- *Karp all summary*: list of all quantified genes in Karp strain grown in HUVEC cells;
- *UT176 all summary*: list of all quantified genes in UT176 strain grown in HUVEC cells;
- *Proteomics*: proteomics data of both human and Karp;
- *Core genes*: list of core genes identified in (Batty et al., 2018), their transcript expression and presence or absence in the proteomics dataset;
- *Pred ncRNA Karp*: genome coordinates of predicted ncRNAs in Karp;
- *Pred ncRNA UT176*: genome coordinates of predicted ncRNAs in UT176;
- *Conserved operons*: list of conserved operons identified in both Karp and UT176;
- *Karp operons*: list of operons identified only in Karp;
- *UT176 operons*: list of operons identified only in UT176;
- *Conserved islands*: list of conserved islands identified in (Batty et al., 2018);
- *Bac diff exp long*: results of the differential expression of bacterial genes in HUVEC cells;
- *Bac diff exp short*: results of the differential expression analysis for non-repetitive bacterial genes with abs(logFC) $\geq$1.0;
- *Diff exp ank tpr*: results of the differential expression of ankyrin-repeat and TPR-repeat bacterial genes of Karp and UT176 in HUVEC cells;
- *Enrichment anal*: results of the bacterial gene set enrichment analysis;

- *Joint response*: genes upregulated in Karp- and UT176-infected HUVEC cells compared with uninfected HUVEC cells;
- *Host diff exp long*: full list of host genes differentially expressed in response to Karp or UT176;
- *Host diff exp short*: selected host genes differentially expressed by HUVEC in response to UT176 or Karp and uninfected HUVEC cells.

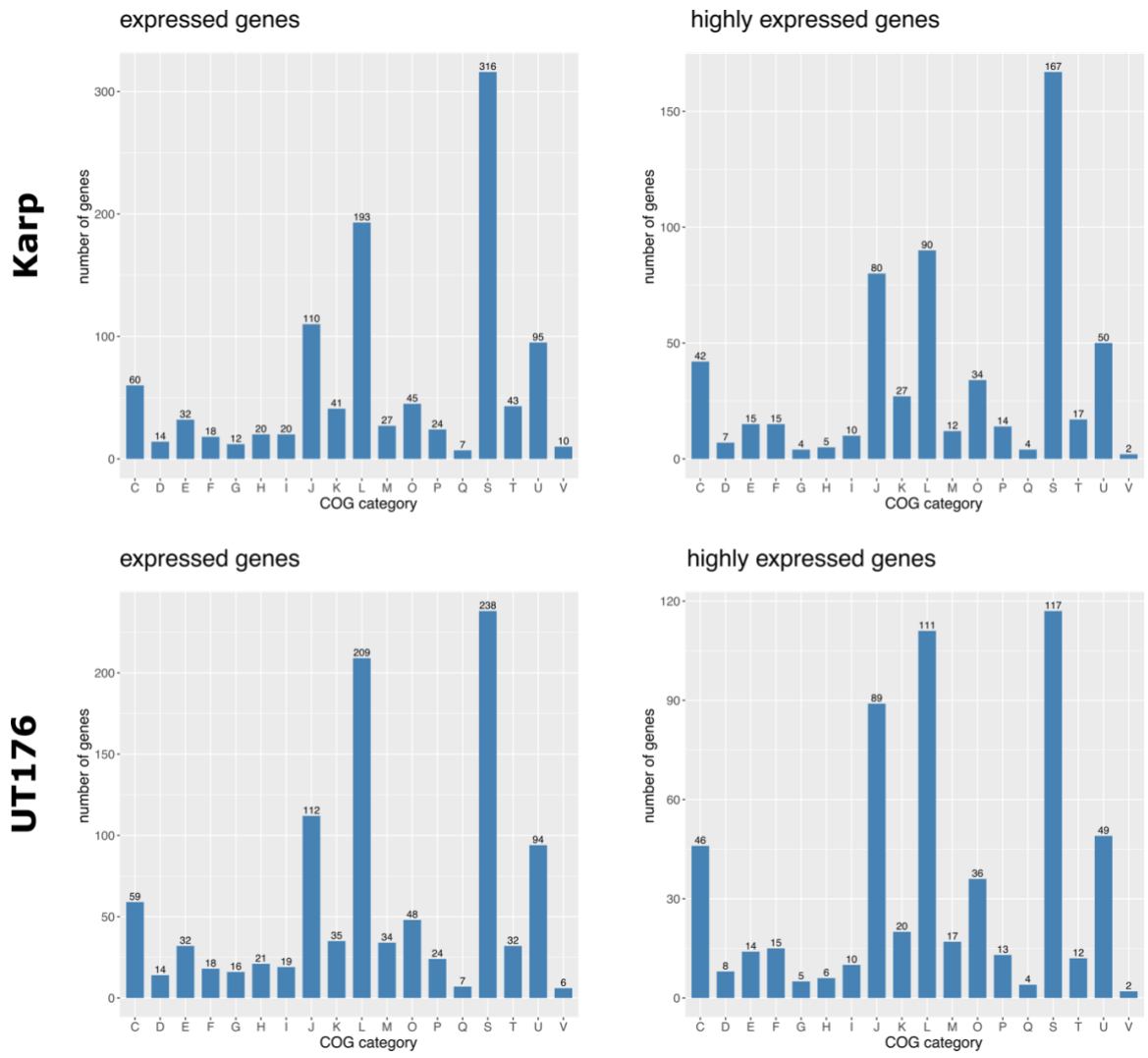

Figure S1 Expressed and highly expressed genes of Ot strains classified into COG categories: Energy production and conversion (C), Cell cycle control and mitosis (D), Amino Acid metabolism and transport (E), Nucleotide metabolism and transport (F), Carbohydrate metabolism and transport (G), Coenzyme metabolism (H), Lipid metabolism (I), Translation (J), Transcription (K), Replication and repair (L), Cell wall/membrane/envelop biogenesis (M), Cell motility (N), Post-translational modification, protein turnover, chaperone functions (O), Inorganic ion transport and metabolism (P), Secondary Structure (Q), Function Unknown (S), Signal Transduction (T), Intracellular trafficking and secretion (U), Intracellular trafficking and secretion (V).

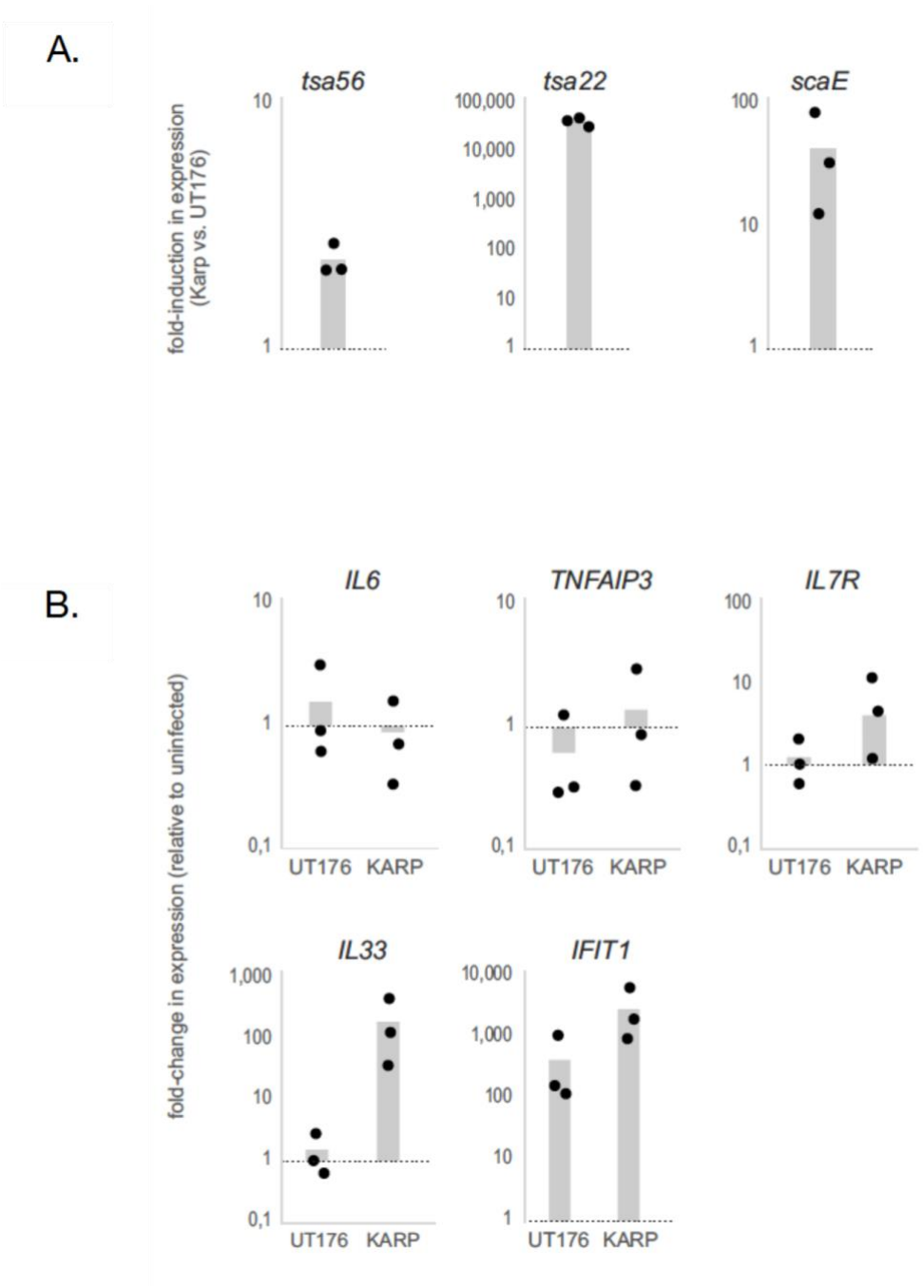

Figure S2 Differential analysis of bacterial and host genes. qRT-PCR of selected bacterial genes in HUVECs infected with Karp or UT176. The expression level was normalized using 7S mRNA. B) qRT-PCR of host genes in HUVEC cells infected with Karp or UT176. The individual values and mean of three biologically independent replicates are shown. This figure from A to B was generated by Alexander J. Westermann.

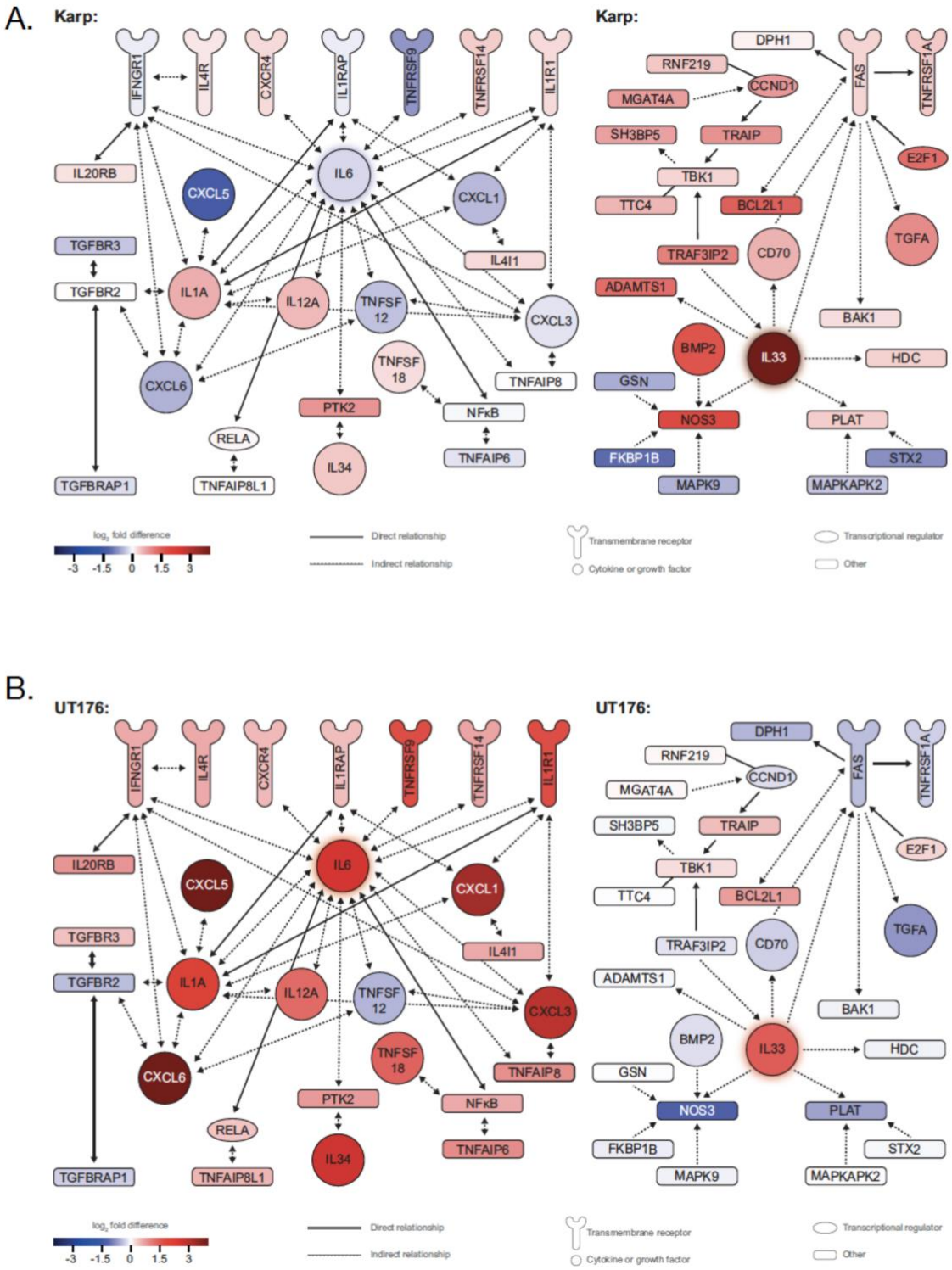

Figure S3 Karp and UT176 infection lead to up-regulation of distinct networks in HUVEC cells. Network map of proinflammatory chemokines and cytokines and IL33-FAS-mediated anoikis network induced in HUVECs infected by A) Karp or B) UT176, in comparison to uninfected cells. The networks were generated by Selvakumar Subbian and re-drawn by Sandy Pernitzsch.

A.

Symbol	Entrez Gene Name	Expr Log Ratio (UT176 vs	Expr Log Ratio (KARP vs	Location	Gene ID-Huma	Gene ID-Mous
Cot	mitogen-activated protein kinase kinase kinase 8	1.70782	0.144809	Cytoplasm	1326	26410
RelB	RELB proto-oncogene, NF-kB subunit	1.28783	-0.03004	Nucleus	5971	19698
MEKK1	mitogen-activated protein kinase kinase kinase 1	0.92341	-0.0359	Cytoplasm	4214	26401
JNK1	mitogen-activated protein kinase 8	0.817456	0.423134	Cytoplasm	5599	26419
NF-κB2 p100	nuclear factor kappa B subunit 2	0.766144	-0.39373	Nucleus	4791	18034
NF-κB1	nuclear factor kappa B subunit 1	0.599586	-0.03203	Nucleus	4790	18033
β-TrCP	beta-transducin repeat containing E3 ubiquitin protein ligase	0.465374	0.146903	Cytoplasm	8945	12234
MALT1	MALT1 paracaspase	0.404342	0.115903	Cytoplasm	10892	240354
p65/RelA	RELA proto-oncogene, NF-kB subunit	0.400752	0.054767	Nucleus	5970	19697
LTBR	lymphotoxin beta receptor	-0.3528	-0.22677	Plasma Membrane	4055	17000
IKKα	conserved helix-loop-helix ubiquitous kinase	-0.39103	0.0463	Cytoplasm	1147	12675
Bcl10	B cell CLL/lymphoma 10	-0.67364	-0.04115	Cytoplasm	8915	12042

Expression of NFkB pathway genes

B.

Symbol	Entrez Gene Name	Expr Log Ratio (UT176 vs UnInf)	Expr Log Ratio (KARP vs UnInf)	Location	Gene ID-Human	Gene ID-Mouse
PPARα	peroxisome proliferator activated receptor alpha	0.955636	0.256633	Nucleus	5465	19013
SIRPα	signal regulatory protein alpha	0.855516	0.231361	Plasma Membrane	140885	19261
MEK1	mitogen-activated protein kinase kinase 1	0.726498	0.240664	Cytoplasm	5604	26395
CBP	CREB binding protein	0.365126	0.187724	Nucleus	1387	12914

Expression of genes involved in NOS2 production

Figure S4 Differential regulation of inflammatory pathways by UT176 and Karp. A) expression of NFκB pathway genes in UT176- and Karp-infected host cells. B) expression of host genes associated with NOS2 production. Red indicates increased expression relative to uninfected cells, blue indicates decreased expression. This figure was created by Selvakumar Subbian.

A. Expression of host genes associated with differentiation of mononuclear leukocytes

ID	Genes in dataset	Expr Log Ratio (UT176 vs Uninf)	Expr Log Ratio (KARP vs Uninf)	Prediction in UT176 (based on measurement direction)
ENSG00000171855.6	*IFNB1*	8.664	8.790328	Increased
ENSG00000164400.5	*CSF2*	7.764	-1.667	Increased
ENSG00000184979.9	*USP18*	5.078	5.464608	Increased
ENSG00000108342.12	*CSF3*	5.037	2.804455	Increased
ENSG00000114251.13	*WNT5A*	4.976	3.61382	Increased
ENSG00000271503.5	*CCL5*	4.801	4.102429	Increased
ENSG00000137462.6	*TLR2*	4.416	-0.21786	Increased
ENSG00000163735.6	*CXCL5*	4.174	-1.41573	Increased
ENSG00000057657.14	*PRDM1*	3.961	1.080978	Increased
ENSG00000128917.6	*DLL4*	3.512	1.850144	Increased
ENSG00000168685.14	*IL7R*	3.5	3.802424	Increased
ENSG00000121858.10	*TNFSF10*	3.376	3.567363	Increased
ENSG00000163739.4	*CXCL1*	3.236	-0.44897	Increased
ENSG00000137752.22	*CASP1*	3.15	2.684897	Increased
ENSG00000130775.15	*THEMIS2*	3.135	3.231943	Increased
ENSG00000185507.19	*IRF7*	3.022	3.875153	Increased
ENSG00000169429.10	*CXCL8*	2.907	-1.96917	Increased
ENSG00000028277.20	*POU2F2*	2.764	0.453238	Increased
ENSG00000115604.10	*IL18R1*	2.736	0.499293	Increased
ENSG00000138378.17	*STAT4*	2.695	2.145555	Increased
ENSG00000157368.10	*IL34*	2.595	0.369739	Increased
ENSG00000170298.15	*LGALS9B*	2.528	2.965212	Increased
ENSG00000136244.11	*IL6*	2.503	-0.23512	Increased
ENSG00000134363.11	*FST*	2.481	1.895928	Increased
ENSG00000109906.13	*ZBTB16*	2.408	0.917825	Increased
ENSG00000089041.16	*P2RX7*	2.275	0.379212	Increased
ENSG00000204632.11	*HLA-G*	2.231	2.459558	Increased
ENSG00000184371.13	*CSF1*	2.22	1.125217	Increased
ENSG00000134470.19	*IL15RA*	2.14	2.119752	Increased
ENSG00000115415.18	*STAT1*	2.106	2.370024	Increased
ENSG00000115008.5	*IL1A*	2.073	0.577578	Increased
ENSG00000102524.11	*TNFSF13B*	2.069	1.73795	Increased
ENSG00000164330.16	*EBF1*	2.065	2.951442	Increased
ENSG00000123496.7	*IL13RA2*	2.024	5.897182	Increased
ENSG00000125347.13	*IRF1*	2.022	1.618451	Increased
ENSG00000134321.11	*RSAD2*	10.321	9.995906	Affected
ENSG00000169245.5	*CXCL10*	10.195	8.430441	Affected
ENSG00000107201.9	*DDX58*	4.771	3.772748	Affected
ENSG00000164342.12	*TLR3*	4.216	2.7201	Affected
ENSG00000168961.16	*LGALS9*	3.756	4.338328	Affected
ENSG00000152689.17	*RASGRP3*	3.317	2.743657	Affected
ENSG00000048052.21	*HDAC9*	2.949	1.379593	Affected
ENSG00000170581.13	*STAT2*	2.748	2.681712	Affected
ENSG00000134215.15	*VAV3*	2.603	2.987545	Affected
ENSG00000169248.12	*CXCL11*	4.66	4.371531	Decreased
ENSG00000240065.7	*PSMB9*	3.567	3.780511	Decreased
ENSG00000166923.10	*GREM1*	3.214	-2.05055	Decreased
ENSG00000115602.16	*IL1RL1*	2.959	0.520912	Decreased
ENSG00000118503.14	*TNFAIP3*	2.674	1.151952	Decreased
ENSG00000140464.19	*PML*	2.167	1.947424	Decreased

B. Expression of host genes associated with leukocyte proliferation

ID	Genes in dataset	Expr Log Ratio (UT176 vs UnInf)	Expr Log Ratio (KARP vs UnInf)	Prediction in UT176 (based on measurement direction)
ENSG00000171855.6	*IFNB1*	8.664	8.790328	Increased
ENSG00000164400.5	*CSF2*	7.764	-1.667	Increased
ENSG00000187608.8	*ISG15*	5.293	5.416531	Increased
ENSG00000108342.12	*CSF3*	5.037	2.804455	Increased
ENSG00000271503.5	*CCL5*	4.801	4.102429	Increased
ENSG00000137462.6	*TLR2*	4.416	-0.21786	Increased
ENSG00000172183.14	*ISG20*	4.384	4.897459	Increased
ENSG00000164342.12	*TLR3*	4.216	2.7201	Increased
ENSG00000168685.14	*IL7R*	3.5	3.802424	Increased
ENSG00000172348.14	*RCAN2*	2.909	1.89052	Increased
ENSG00000100234.11	*TIMP3*	2.788	1.065608	Increased
ENSG00000028277.20	*POU2F2*	2.764	0.453238	Increased
ENSG00000138378.17	*STAT4*	2.695	2.145555	Increased
ENSG00000134215.15	*VAV3*	2.603	2.987545	Increased
ENSG00000157368.10	*IL34*	2.595	0.369739	Increased
ENSG00000170091.10	*NSG2*	2.557	3.988829	Increased
ENSG00000136244.11	*IL6*	2.503	-0.23512	Increased
ENSG00000026950.16	*BTN3A1*	2.475	2.699843	Increased
ENSG00000170166.5	*HOXD4*	2.286	2.00549	Increased
ENSG00000089041.16	*P2RX7*	2.275	0.379212	Increased
ENSG00000184371.13	*CSF1*	2.22	1.125217	Increased
ENSG00000185950.8	*IRS2*	2.142	-0.59384	Increased
ENSG00000134470.19	*IL15RA*	2.14	2.119752	Increased
ENSG00000113319.11	*RASGRF2*	2.099	1.634606	Increased
ENSG00000115008.5	*IL1A*	2.073	0.577578	Increased
ENSG00000102524.11	*TNFSF13B*	2.069	1.73795	Increased
ENSG00000114251.13	*WNT5A*	4.976	3.61382	Affected
ENSG00000162692.10	*VCAM1*	4.467	-0.03961	Affected
ENSG00000168961.16	*LGALS9*	3.756	4.338328	Affected
ENSG00000152689.17	*RASGRP3*	3.317	2.743657	Affected
ENSG00000177409.11	*SAMD9L*	3.277	2.927656	Affected
ENSG00000115602.16	*IL1RL1*	2.959	0.520912	Affected
ENSG00000023445.13	*BIRC3*	2.889	1.028695	Affected
ENSG00000163734.4	*CXCL3*	2.872	-0.17656	Affected
ENSG00000152217.16	*SETBP1*	2.574	1.333599	Affected
ENSG00000140464.19	*PML*	2.167	1.947424	Affected
ENSG00000169245.5	*CXCL10*	10.195	8.430441	Decreased
ENSG00000175899.14	*A2M*	5.199	4.163204	Decreased
ENSG00000131203.12	*IDO1*	4.034	2.870018	Decreased
ENSG00000057657.14	*PRDM1*	3.961	1.080978	Decreased
ENSG00000173193.13	*PARP14*	3.713	3.461054	Decreased
ENSG00000115009.11	*CCL20*	3.616	-1.19333	Decreased
ENSG00000128917.6	*DLL4*	3.512	1.850144	Decreased
ENSG00000121858.10	*TNFSF10*	3.376	8.790328	Decreased
ENSG00000169429.10	*CXCL8*	2.907	-1.667	Decreased
ENSG00000118503.14	*TNFAIP3*	2.674	5.416531	Decreased
ENSG00000170298.15	*LGALS9B*	2.528	2.804455	Decreased
ENSG00000109906.13	*ZBTB16*	2.408	4.102429	Decreased
ENSG00000204632.11	*HLA-G*	2.231	-0.21786	Decreased
ENSG00000115415.18	*STAT1*	2.106	4.897459	Decreased

Figure S5 Differential activation of inflammatory pathways by UT176 and Karp A) Expression of host genes associated with differentiation of mononuclear leukocytes B) Expression of host genes associated with leukocyte proliferation. The predicted effect on leukocyte differentiation/proliferation is shown based on the gene expression in response to UT176 infection. Red indicates increased expression relative to uninfected cells; blue indicates decreased expression. This figure from A to B was created by Selvakumar Subbian.

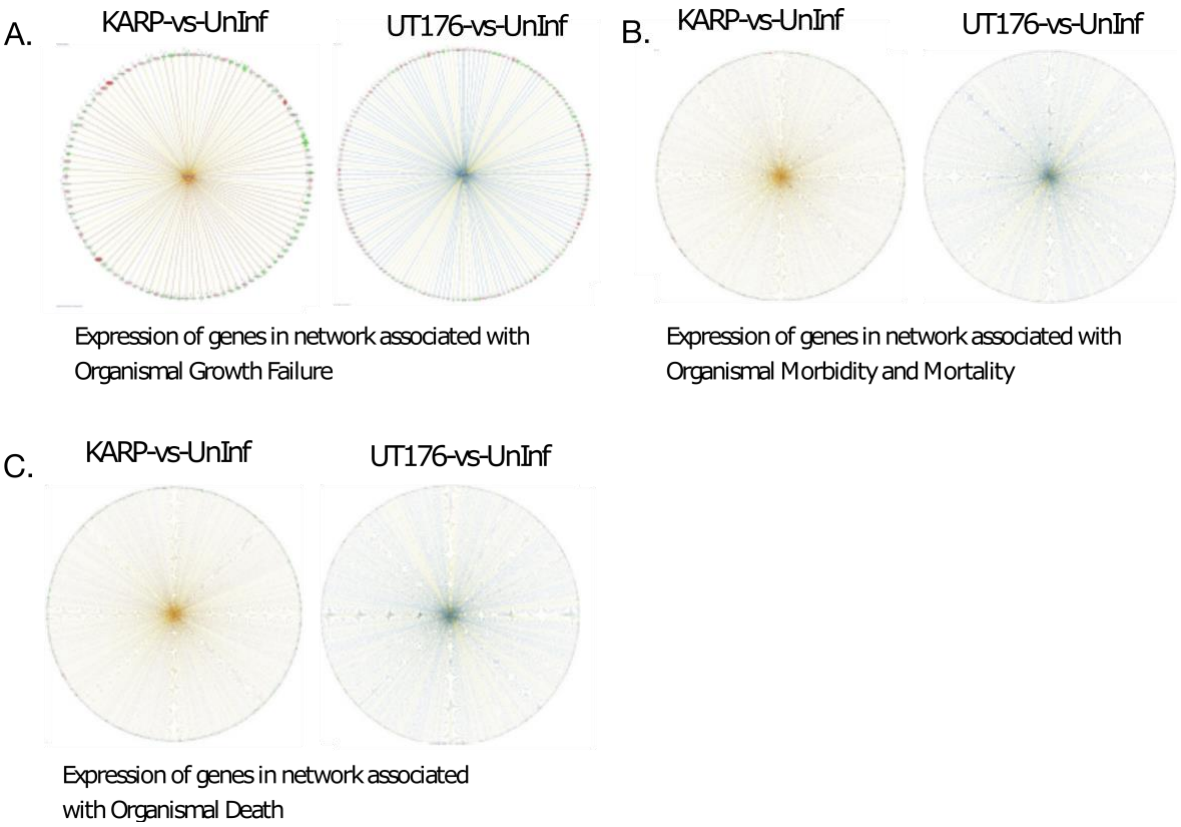

Figure S6 Host networks upregulated after the infection with Karp. They are associated with (A) organismal growth failure, (B) morbidity and mortality and (C) death. Connections colored in red are up-regulated whilst connections in blue are down-regulated. This figure from A to C was generated by Selvakumar Subbian.

A. Mouse clinical observation scoring system

Parameter		CODE	SCORE
Group Appetite	ate all 6-12 chows	0	0
	ate 1-5 chows	1	1
	ate 0 to 1/2 chow	2	2
Activity	move around	0	0
	locomotion after slight stimulation	A1	1
	move slowly after moderate stimulation	A2	2
	unable to move	A3	3
Hair coat	well-groomed hair coat	0	0
	rough hair coat	R1	1
	ungroomed, very rough hair coat, and dirty	R2	2
Total score	7		

B.

Lesion scoring system

0 = Normal tissue
1 = Minimal lesion severity and extent
2 = Mild
3 = Moderate
4 = Marked
5 = Severe

Figure S7 Scoring system used in mice experiments for evaluation of (A) clinical observations and (B) Lesions in hematoxylin and eosin-stained tissue sections. The experiments were performed by Piyanate Sunyakumthorn.

Table S1 KEGG gene sets enriched in expressed genes in Karp. Results of the hypergeometric test. No.- number; expr – expressed.

KEGG ID	Name of a KEGG pathway	No. of expr. genes in pathway	No. of genes in a pathway	No. of genes not in a pathway	No. of expr. genes	p_value	FDR
map01100	Metabolic pathways	144	156	2095	1422	4,73E-18	6,53E-16
map03010	Ribosome	50	50	2201	1422	7,69E-11	5,30E-09
map00190	Oxidative phosphorylation	38	38	2213	1422	2,19E-08	1,01E-06
map01110	Biosynthesis of secondary metabolites	56	61	2190	1422	2,37E-07	8,18E-06
map00970	Aminoacyl-tRNA biosynthesis	27	27	2224	1422	3,75E-06	1,04E-04
map00240	Pyrimidine metabolism	24	25	2226	1422	0,0001508	0,0034682
map00020	Citrate cycle (TCA cycle)	17	17	2234	1422	0,0003922	0,0067656
map00910	Nitrogen metabolism	22	23	2228	1422	0,0003534	0,0067656
map03070	Bacterial secretion system	29	32	2219	1422	0,0004505	0,0069070
map00230	Purine metabolism	23	25	2226	1422	0,0011541	0,0159261
map01120	Microbial metabolism in diverse environments	38	46	2205	1422	0,0032002	0,0401483
map01230	Biosynthesis of amino acids	18	20	2231	1422	0,0077041	0,0885969
map03018	RNA degradation	9	9	2242	1422	0,0158728	0,1684963
map00260	Glycine, serine and threonine metabolism	8	8	2243	1422	0,0251788	0,1930372
map02010	ABC transporters	12	13	2238	1422	0,0216110	0,1930372
map01210	2-Oxocarboxylic acid metabolism	8	8	2243	1422	0,0251788	0,1930372
map00564	Glycerophospholipid metabolism	8	8	2243	1422	0,0251788	0,1930372
map00195	Photosynthesis	8	8	2243	1422	0,0251788	0,1930372
map00780	Biotin metabolism	7	7	2244	1422	0,0399301	0,2623980
map05010	Alzheimer disease	7	7	2244	1422	0,0399301	0,2623980
map05016	Huntington disease	7	7	2244	1422	0,0399301	0,2623980
map03430	Mismatch repair	13	15	2236	1422	0,0457365	0,2820591
map00860	Porphyrin and chlorophyll metabolism	10	11	2240	1422	0,0470098	0,2820591
map05012	Parkinson disease	6	6	2245	1422	0,0633073	0,3640171
map03410	Base excision repair	9	10	2241	1422	0,0686950	0,3791962
map00720	Carbon fixation pathways in prokaryotes	14	17	2234	1422	0,0771857	0,4096778
map02020	Two-component system	13	16	2235	1422	0,1034267	0,4757629
map00300	Lysine biosynthesis	8	9	2242	1422	0,0996262	0,4757629
map00670	One carbon pool by folate	5	5	2246	1422	0,1003446	0,4757629
map00650	Butanoate metabolism	5	5	2246	1422	0,1003446	0,4757629
map05152	Tuberculosis	4	4	2247	1422	0,1590086	0,5774523
map00900	Terpenoid backbone biosynthesis	4	4	2247	1422	0,1590086	0,5774523
map05134	Legionellosis	4	4	2247	1422	0,1590086	0,5774523
map05200	Pathways in cancer	4	4	2247	1422	0,1590086	0,5774523
map00130	Ubiquinone and other terpenoid-quinone biosynthesis	7	8	2243	1422	0,1431898	0,5774523
map03420	Nucleotide excision repair	4	4	2247	1422	0,1590086	0,5774523
map04626	Plant-pathogen interaction	4	4	2247	1422	0,1590086	0,5774523

Table S1 continued

map00623	Toluene degradation	4	4	2247	1422	0,1590086	0,5774523
map03060	Protein export	13	17	2234	1422	0,1887101	0,6677433
map05120	Epithelial cell signaling in Helicobacter pylori infection	3	3	2248	1422	0,2519037	0,6705943
map00450	Selenocompound metabolism	3	3	2248	1422	0,2519037	0,6705943
map00010	Glycolysis / Gluconeogenesis	6	7	2244	1422	0,2035704	0,6705943
map00280	Valine, leucine and isoleucine degradation	4	6	2245	1422	0,6099330	0,6705943
map00620	Pyruvate metabolism	6	8	2243	1422	0,3847123	0,6705943
map00550	Peptidoglycan biosynthesis	8	12	2239	1422	0,5298805	0,6705943
map00061	Fatty acid biosynthesis	6	7	2244	1422	0,2035704	0,6705943
map04141	Protein processing in endoplasmic reticulum	2	2	2249	1422	0,3989658	0,6705943
map00270	Cysteine and methionine metabolism	3	3	2248	1422	0,2519037	0,6705943
map00330	Arginine and proline metabolism	5	6	2245	1422	0,2855307	0,6705943
map00360	Phenylalanine metabolism	2	2	2249	1422	0,3989658	0,6705943
map00400	Phenylalanine, tyrosine and tryptophan biosynthesis	1	1	2250	1422	0,6317192	0,6705943
map00401	Novobiocin biosynthesis	1	1	2250	1422	0,6317192	0,6705943
map00950	Isoquinoline alkaloid biosynthesis	1	1	2250	1422	0,6317192	0,6705943
map00960	Tropane, piperidine and pyridine alkaloid biosynthesis	1	1	2250	1422	0,6317192	0,6705943
map04115	p53 signaling pathway	1	1	2250	1422	0,6317192	0,6705943
map04210	Apoptosis	1	1	2250	1422	0,6317192	0,6705943
map05014	Amyotrophic lateral sclerosis (ALS)	1	1	2250	1422	0,6317192	0,6705943
map05145	Toxoplasmosis	1	1	2250	1422	0,6317192	0,6705943
map05161	Hepatitis B	1	1	2250	1422	0,6317192	0,6705943
map05164	Influenza A	1	1	2250	1422	0,6317192	0,6705943
map05168	Herpes simplex virus 1 infection	1	1	2250	1422	0,6317192	0,6705943
map05210	Colorectal cancer	1	1	2250	1422	0,6317192	0,6705943
map05222	Small cell lung cancer	1	1	2250	1422	0,6317192	0,6705943
map05416	Viral myocarditis	1	1	2250	1422	0,6317192	0,6705943
map00630	Glyoxylate and dicarboxylate metabolism	6	9	2242	1422	0,5628122	0,6705943
map00680	Methane metabolism	3	4	2247	1422	0,5305890	0,6705943
map00710	Carbon fixation in photosynthetic organisms	4	6	2245	1422	0,6099330	0,6705943
map00051	Fructose and mannose metabolism	3	4	2247	1422	0,5305890	0,6705943
map00640	Propanoate metabolism	4	6	2245	1422	0,6099330	0,6705943
map00660	C5-Branched dibasic acid metabolism	2	2	2249	1422	0,3989658	0,6705943
map00460	Cyanoamino acid metabolism	1	1	2250	1422	0,6317192	0,6705943
map04260	Cardiac muscle contraction	3	3	2248	1422	0,2519037	0,6705943
map04070	Phosphatidylinositol signaling system	2	2	2249	1422	0,3989658	0,6705943
map00440	Phosphonate and phosphinate metabolism	1	1	2250	1422	0,6317192	0,6705943
map00311	Penicillin and cephalosporin biosynthesis	1	1	2250	1422	0,6317192	0,6705943
map00312	beta-Lactam resistance	1	1	2250	1422	0,6317192	0,6705943
map03020	RNA polymerase	3	3	2248	1422	0,2519037	0,6705943
map00561	Glycerolipid metabolism	1	1	2250	1422	0,6317192	0,6705943

Table S1 continued

map00480	Glutathione metabolism	3	3	2248	1422	0,2519037	0,6705943
map00980	Metabolism of xenobiotics by cytochrome P450	1	1	2250	1422	0,6317192	0,6705943
map00982	Drug metabolism - cytochrome P450	1	1	2250	1422	0,6317192	0,6705943
map05204	Chemical carcinogenesis	1	1	2250	1422	0,6317192	0,6705943
map00030	Pentose phosphate pathway	1	1	2250	1422	0,6317192	0,6705943
map00290	Valine, leucine and isoleucine biosynthesis	3	3	2248	1422	0,2519037	0,6705943
map04146	Peroxisome	2	2	2249	1422	0,3989658	0,6705943
map03008	Ribosome biogenesis in eukaryotes	1	1	2250	1422	0,6317192	0,6705943
map05205	Proteoglycans in cancer	1	1	2250	1422	0,6317192	0,6705943
map00071	Fatty acid degradation	1	1	2250	1422	0,6317192	0,6705943
map00791	Atrazine degradation	1	1	2250	1422	0,6317192	0,6705943
map00983	Drug metabolism - other enzymes	3	3	2248	1422	0,2519037	0,6705943
map05133	Pertussis	2	2	2249	1422	0,3989658	0,6705943
map00365	Furfural degradation	1	1	2250	1422	0,6317192	0,6705943
map05142	Chagas disease (American trypanosomiasis)	1	1	2250	1422	0,6317192	0,6705943
map05143	African trypanosomiasis	1	1	2250	1422	0,6317192	0,6705943
map01040	Biosynthesis of unsaturated fatty acids	1	1	2250	1422	0,6317192	0,6705943
map04151	PI3K-Akt signaling pathway	1	1	2250	1422	0,6317192	0,6705943
map04612	Antigen processing and presentation	1	1	2250	1422	0,6317192	0,6705943
map04621	NOD-like receptor signaling pathway	1	1	2250	1422	0,6317192	0,6705943
map04914	Progesterone-mediated oocyte maturation	1	1	2250	1422	0,6317192	0,6705943
map04915	Estrogen signaling pathway	1	1	2250	1422	0,6317192	0,6705943
map05215	Prostate cancer	1	1	2250	1422	0,6317192	0,6705943
map04122	Sulfur relay system	3	3	2248	1422	0,2519037	0,6705943
map00500	Starch and sucrose metabolism	1	1	2250	1422	0,6317192	0,6705943
map00730	Thiamine metabolism	2	2	2249	1422	0,3989658	0,6705943
map05211	Renal cell carcinoma	2	2	2249	1422	0,3989658	0,6705943
map00770	Pantothenate and CoA biosynthesis	2	2	2249	1422	0,3989658	0,6705943
map04940	Type I diabetes mellitus	1	1	2250	1422	0,6317192	0,6705943
map00760	Nicotinate and nicotinamide metabolism	1	1	2250	1422	0,6317192	0,6705943
map00380	Tryptophan metabolism	2	2	2249	1422	0,3989658	0,6705943
map00627	Aminobenzoate degradation	1	1	2250	1422	0,6317192	0,6705943
map00643	Styrene degradation	1	1	2250	1422	0,6317192	0,6705943
map04723	Retrograde endocannabinoid signaling	1	1	2250	1422	0,6317192	0,6705943
map00785	Lipoic acid metabolism	2	2	2249	1422	0,3989658	0,6705943
map00363	Bisphenol degradation	1	1	2250	1422	0,6317192	0,6705943
map00591	Linoleic acid metabolism	1	1	2250	1422	0,6317192	0,6705943
map00625	Chloroalkane and chloroalkene degradation	1	1	2250	1422	0,6317192	0,6705943
map00473	D-Alanine metabolism	2	2	2249	1422	0,3989658	0,6705943
map00520	Amino sugar and nucleotide sugar metabolism	3	3	2248	1422	0,2519037	0,6705943
map00790	Folate biosynthesis	1	1	2250	1422	0,6317192	0,6705943

Table S1 continued

map05111	Biofilm formation - Vibrio cholerae	1	1	2250	1422	0,6317192	0,6705943
map00310	Lysine degradation	2	2	2249	1422	0,3989658	0,6705943
map00362	Benzoate degradation	1	1	2250	1422	0,6317192	0,6705943
map00281	Geraniol degradation	1	1	2250	1422	0,6317192	0,6705943
map00626	Naphthalene degradation	1	1	2250	1422	0,6317192	0,6705943
map00750	Vitamin B6 metabolism	1	1	2250	1422	0,6317192	0,6705943
map00903	Limonene and pinene degradation	1	1	2250	1422	0,6317192	0,6705943
map04066	HIF-1 signaling pathway	1	1	2250	1422	0,6317192	0,6705943
map00521	Streptomycin biosynthesis	1	1	2250	1422	0,6317192	0,6705943
map03013	RNA transport	1	1	2250	1422	0,6317192	0,6705943
map00430	Taurine and hypotaurine metabolism	1	1	2250	1422	0,6317192	0,6705943
map00250	Alanine, aspartate and glutamate metabolism	2	3	2248	1422	0,6930900	0,7245941
map00562	Inositol phosphate metabolism	2	3	2248	1422	0,6930900	0,7245941
map00471	D-Glutamine and D-glutamate metabolism	1	2	2249	1422	0,8644727	0,8969717
map04112	Cell cycle - Caulobacter	19	55	2196	1422	0,9999966	0,9999997
map03030	DNA replication	19	55	2196	1422	0,9999966	0,9999997
map03440	Homologous recombination	30	81	2170	1422	0,9999997	0,9999997
map00350	Tyrosine metabolism	1	4	2247	1422	0,9816884	0,9999997
map00624	Polycyclic aromatic hydrocarbon degradation	1	4	2247	1422	0,9816884	0,9999997

Table S2 KEGG gene sets enriched in highly expressed genes in Karp. Results of the hypergeometric test. No.- number; h. expr – highly expressed.

KEGG ID	Name of a KEGG pathway	No. of h.expr. genes in pathway	No. of genes in a pathway	No. of genes not in a pathway	No. of h.expr. genes	p_value	FDR
map03010	Ribosome	50	50	2201	856	4,07E-22	4,35E-20
map00190	Oxidative phosphorylation	28	38	2213	856	7,32E-06	0,000392
map01100	Metabolic pathways	83	156	2095	856	4,71E-05	0,001680
map00020	Citrate cycle (TCA cycle)	13	17	2234	856	0,001409	0,037689
map00230	Purine metabolism	16	25	2226	856	0,007287	0,111385
map00910	Nitrogen metabolism	15	23	2228	856	0,007264	0,111385
map00195	Photosynthesis	7	8	2243	856	0,006061	0,111385
map00240	Pyrimidine metabolism	15	25	2226	856	0,020724	0,246873
map00720	Carbon fixation pathways in prokaryotes	11	17	2234	856	0,023072	0,246873
map00623	Toluene degradation	4	4	2247	856	0,020821	0,246873
map03070	Bacterial secretion system	18	32	2219	856	0,026787	0,260567
map03020	RNA polymerase	3	3	2248	856	0,054872	0,451638
map00520	Amino sugar and nucleotide sugar metabolism	3	3	2248	856	0,054872	0,451638
map01120	Microbial metabolism in diverse environments	23	46	2205	856	0,063650	0,486466
map05016	Huntington disease	5	7	2244	856	0,078083	0,494398
map00630	Glyoxylate and dicarboxylate metabolism	6	9	2242	856	0,078549	0,494398
map00650	Butanoate metabolism	4	5	2246	856	0,072526	0,494398
map05120	Epithelial cell signaling in Helicobacter pylori infection	2	3	2248	856	0,323770	0,565132
map03018	RNA degradation	5	9	2242	856	0,226344	0,565132
map05152	Tuberculosis	3	4	2247	856	0,157025	0,565132
map03060	Protein export	9	17	2234	856	0,153738	0,565132
map00250	Alanine, aspartate and glutamate metabolism	2	3	2248	856	0,323770	0,565132
map00360	Phenylalanine metabolism	2	2	2249	856	0,144505	0,565132
map00400	Phenylalanine, tyrosine and tryptophan biosynthesis	1	1	2250	856	0,380275	0,565132
map00401	Novobiocin biosynthesis	1	1	2250	856	0,380275	0,565132
map00950	Isoquinoline alkaloid biosynthesis	1	1	2250	856	0,380275	0,565132
map00960	Tropane, piperidine and pyridine alkaloid biosynthesis	1	1	2250	856	0,380275	0,565132
map01110	Biosynthesis of secondary metabolites	28	61	2190	856	0,125466	0,565132
map03410	Base excision repair	5	10	2241	856	0,318040	0,565132
map04115	p53 signaling pathway	1	1	2250	856	0,380275	0,565132
map04210	Apoptosis	1	1	2250	856	0,380275	0,565132
map05010	Alzheimer disease	4	7	2244	856	0,252316	0,565132
map05012	Parkinson disease	4	6	2245	856	0,152754	0,565132
map05014	Amyotrophic lateral sclerosis (ALS)	1	1	2250	856	0,380275	0,565132
map05134	Legionellosis	3	4	2247	856	0,157025	0,565132
map05145	Toxoplasmosis	1	1	2250	856	0,380275	0,565132
map05161	Hepatitis B	1	1	2250	856	0,380275	0,565132

Table S2 continued

map05164	Influenza A	1	1	2250	856	0,380275	0,565132
map05168	Herpes simplex virus 1 infection	1	1	2250	856	0,380275	0,565132
map05210	Colorectal cancer	1	1	2250	856	0,380275	0,565132
map05222	Small cell lung cancer	1	1	2250	856	0,380275	0,565132
map05416	Viral myocarditis	1	1	2250	856	0,380275	0,565132
map00620	Pyruvate metabolism	4	8	2243	856	0,360396	0,565132
map00670	One carbon pool by folate	3	5	2246	856	0,283772	0,565132
map00564	Glycerophospholipid metabolism	5	8	2243	856	0,144236	0,565132
map00640	Propanoate metabolism	4	6	2245	856	0,152754	0,565132
map00660	C5-Branched dibasic acid metabolism	2	2	2249	856	0,144505	0,565132
map00460	Cyanoamino acid metabolism	1	1	2250	856	0,380275	0,565132
map00311	Penicillin and cephalosporin biosynthesis	1	1	2250	856	0,380275	0,565132
map00312	beta-Lactam resistance	1	1	2250	856	0,380275	0,565132
map04626	Plant-pathogen interaction	3	4	2247	856	0,157025	0,565132
map00561	Glycerolipid metabolism	1	1	2250	856	0,380275	0,565132
map00480	Glutathione metabolism	2	3	2248	856	0,323770	0,565132
map00980	Metabolism of xenobiotics by cytochrome P450	1	1	2250	856	0,380275	0,565132
map00982	Drug metabolism - cytochrome P450	1	1	2250	856	0,380275	0,565132
map05204	Chemical carcinogenesis	1	1	2250	856	0,380275	0,565132
map00030	Pentose phosphate pathway	1	1	2250	856	0,380275	0,565132
map03008	Ribosome biogenesis in eukaryotes	1	1	2250	856	0,380275	0,565132
map05205	Proteoglycans in cancer	1	1	2250	856	0,380275	0,565132
map00983	Drug metabolism - other enzymes	2	3	2248	856	0,323770	0,565132
map05142	Chagas disease (American trypanosomiasis)	1	1	2250	856	0,380275	0,565132
map05143	African trypanosomiasis	1	1	2250	856	0,380275	0,565132
map00730	Thiamine metabolism	2	2	2249	856	0,144505	0,565132
map04122	Sulfur relay system	2	3	2248	856	0,323770	0,565132
map04940	Type I diabetes mellitus	1	1	2250	856	0,380275	0,565132
map00627	Aminobenzoate degradation	1	1	2250	856	0,380275	0,565132
map00643	Styrene degradation	1	1	2250	856	0,380275	0,565132
map04723	Retrograde endocannabinoid signaling	1	1	2250	856	0,380275	0,565132
map00473	D-Alanine metabolism	2	2	2249	856	0,144505	0,565132
map05111	Biofilm formation - Vibrio cholerae	1	1	2250	856	0,380275	0,565132
map00362	Benzoate degradation	1	1	2250	856	0,380275	0,565132
map00430	Taurine and hypotaurine metabolism	1	1	2250	856	0,380275	0,565132
map00330	Arginine and proline metabolism	3	6	2245	856	0,414789	0,607979
map00900	Terpenoid backbone biosynthesis	2	4	2247	856	0,490516	0,690594
map05200	Pathways in cancer	2	4	2247	856	0,490516	0,690594
map00680	Methane metabolism	2	4	2247	856	0,490516	0,690594
map00010	Glycolysis / Gluconeogenesis	3	7	2244	856	0,536644	0,745727
map02020	Two-component system	6	16	2235	856	0,610741	0,757666

Table S2 continued

map04141	Protein processing in endoplasmic reticulum	1	2	2249	856	0,616046	0,757666
map00970	Aminoacyl-tRNA biosynthesis	10	27	2224	856	0,614381	0,757666
map05211	Renal cell carcinoma	1	2	2249	856	0,616046	0,757666
map00380	Tryptophan metabolism	1	2	2249	856	0,616046	0,757666
map00471	D-Glutamine and D-glutamate metabolism	1	2	2249	856	0,616046	0,757666
map00785	Lipoic acid metabolism	1	2	2249	856	0,616046	0,757666
map05133	Pertussis	1	2	2249	856	0,616046	0,757666
map04146	Peroxisome	1	2	2249	856	0,616046	0,757666
map00310	Lysine degradation	1	2	2249	856	0,616046	0,757666
map01210	2-Oxocarboxylic acid metabolism	3	8	2243	856	0,642393	0,781092
map01230	Biosynthesis of amino acids	7	20	2231	856	0,689990	0,829538
map00710	Carbon fixation in photosynthetic organisms	2	6	2245	856	0,735123	0,845786
map00300	Lysine biosynthesis	3	9	2242	856	0,729771	0,845786
map00280	Valine, leucine and isoleucine degradation	2	6	2245	856	0,735123	0,845786
map00550	Peptidoglycan biosynthesis	4	12	2239	856	0,730916	0,845786
map00270	Cysteine and methionine metabolism	1	3	2248	856	0,762184	0,858460
map04260	Cardiac muscle contraction	1	3	2248	856	0,762184	0,858460
map00061	Fatty acid biosynthesis	2	7	2244	856	0,814514	0,898484
map00780	Biotin metabolism	2	7	2244	856	0,814514	0,898484
map00350	Tyrosine metabolism	1	4	2247	856	0,852740	0,921649
map00624	Polycyclic aromatic hydrocarbon degradation	1	4	2247	856	0,852740	0,921649
map00130	Ubiquinone and other terpenoid-quinone biosynthesis	2	8	2243	856	0,871887	0,923683
map00260	Glycine, serine and threonine metabolism	2	8	2243	856	0,871887	0,923683
map03430	Mismatch repair	4	15	2236	856	0,882517	0,925778
map04112	Cell cycle - Caulobacter	7	55	2196	856	0,999994	0,999997
map03030	DNA replication	7	55	2196	856	0,999994	0,999997
map03440	Homologous recombination	13	81	2170	856	0,999997	0,999997
map00860	Porphyrin and chlorophyll metabolism	1	11	2240	856	0,994899	0,999997
map02010	ABC transporters	1	13	2238	856	0,998053	0,999997

Table S3 KEGG gene sets enriched in expressed genes in UT176. Results of the hypergeometric test. No.- number; expr – expressed.

KEGG ID	Name of a KEGG pathway	No. of expr. genes in pathway	No. of genes in a pathway	No. of genes not in a pathway	No. of expr. genes	p_value	FDR
map01100	Metabolic pathways	148	154	1760	1244	4,03E-22	5,65E-20
map03010	Ribosome	50	50	1864	1244	3,10E-10	2,17E-08
map01110	Biosynthesis of secondary metabolites	58	60	1854	1244	2,13E-09	9,94E-08
map00190	Oxidative phosphorylation	38	38	1876	1244	6,34E-08	2,22E-06
map00970	Aminoacyl-tRNA biosynthesis	26	26	1888	1244	1,24E-05	0,000348
map03070	Bacterial secretion system	29	30	1884	1244	3,77E-05	0,000880
map01120	Microbial metabolism in diverse environments	40	45	1869	1244	0,000234	0,004678
map00240	Pyrimidine metabolism	24	25	1889	1244	0,000284	0,004967
map00910	Nitrogen metabolism	22	23	1891	1244	0,000629	0,009789
map00230	Purine metabolism	23	25	1889	1244	0,002023	0,028328
map00020	Citrate cycle (TCA cycle)	16	17	1897	1244	0,006516	0,082931
map01230	Biosynthesis of amino acids	18	20	1894	1244	0,011769	0,137310
map03430	Mismatch repair	14	15	1899	1244	0,013889	0,149572
map00720	Carbon fixation pathways in prokaryotes	13	14	1900	1244	0,020165	0,191217
map03018	RNA degradation	9	9	1905	1244	0,020488	0,191217
map00260	Glycine, serine and threonine metabolism	8	8	1906	1244	0,031593	0,221153
map00130	Ubiquinone and other terpenoid-quinone biosynthesis	8	8	1906	1244	0,031593	0,221153
map00564	Glycerophospholipid metabolism	8	8	1906	1244	0,031593	0,221153
map02010	ABC transporters	12	13	1901	1244	0,029154	0,221153
map00195	Photosynthesis	8	8	1906	1244	0,031593	0,221153
map02020	Two-component system	14	16	1898	1244	0,044404	0,252546
map05010	Alzheimer disease	7	7	1907	1244	0,048705	0,252546
map05016	Huntington disease	7	7	1907	1244	0,048705	0,252546
map00620	Pyruvate metabolism	7	7	1907	1244	0,048705	0,252546
map00010	Glycolysis / Gluconeogenesis	7	7	1907	1244	0,048705	0,252546
map00780	Biotin metabolism	7	7	1907	1244	0,048705	0,252546
map00550	Peptidoglycan biosynthesis	11	12	1902	1244	0,041946	0,252546
map04112	Cell cycle - Caulobacter	24	30	1884	1244	0,057089	0,284497
map00860	Porphyrin and chlorophyll metabolism	10	11	1903	1244	0,060020	0,284497
map03060	Protein export	13	15	1899	1244	0,060964	0,284497
map05012	Parkinson disease	6	6	1908	1244	0,075064	0,318454
map00710	Carbon fixation in photosynthetic organisms	6	6	1908	1244	0,075064	0,318454
map00061	Fatty acid biosynthesis	6	6	1908	1244	0,075064	0,318454
map03030	DNA replication	23	30	1884	1244	0,121739	0,460634
map00300	Lysine biosynthesis	8	9	1905	1244	0,120439	0,460634
map00670	One carbon pool by folate	5	5	1909	1244	0,115656	0,460634
map00650	Butanoate metabolism	5	5	1909	1244	0,115656	0,460634

Table S3 continued

map01210	2-Oxocarboxylic acid metabolism	7	8	1906	1244	0,168489	0,519596
map00900	Terpenoid backbone biosynthesis	4	4	1910	1244	0,178147	0,519596
map03410	Base excision repair	7	8	1906	1244	0,168489	0,519596
map05134	Legionellosis	4	4	1910	1244	0,178147	0,519596
map05152	Tuberculosis	4	4	1910	1244	0,178147	0,519596
map05200	Pathways in cancer	4	4	1910	1244	0,178147	0,519596
map00680	Methane metabolism	4	4	1910	1244	0,178147	0,519596
map03420	Nucleotide excision repair	4	4	1910	1244	0,178147	0,519596
map00051	Fructose and mannose metabolism	4	4	1910	1244	0,178147	0,519596
map04626	Plant-pathogen interaction	4	4	1910	1244	0,178147	0,519596
map00623	Toluene degradation	4	4	1910	1244	0,178147	0,519596
map05120	Epithelial cell signaling in Helicobacter pylori infection	3	3	1911	1244	0,274327	0,662168
map00270	Cysteine and methionine metabolism	3	3	1911	1244	0,274327	0,662168
map04260	Cardiac muscle contraction	3	3	1911	1244	0,274327	0,662168
map00640	Propanoate metabolism	3	3	1911	1244	0,274327	0,662168
map00562	Inositol phosphate metabolism	3	3	1911	1244	0,274327	0,662168
map00280	Valine, leucine and isoleucine degradation	3	3	1911	1244	0,274327	0,662168
map03020	RNA polymerase	3	3	1911	1244	0,274327	0,662168
map00983	Drug metabolism - other enzymes	3	3	1911	1244	0,274327	0,662168
map00290	Valine, leucine and isoleucine biosynthesis	3	3	1911	1244	0,274327	0,662168
map00520	Amino sugar and nucleotide sugar metabolism	3	3	1911	1244	0,274327	0,662168
map00330	Arginine and proline metabolism	5	6	1908	1244	0,318614	0,710880
map00360	Phenylalanine metabolism	2	2	1912	1244	0,422313	0,710880
map00400	Phenylalanine, tyrosine and tryptophan biosynthesis	1	1	1913	1244	0,649948	0,710880
map00401	Novobiocin biosynthesis	1	1	1913	1244	0,649948	0,710880
map00950	Isoquinoline alkaloid biosynthesis	1	1	1913	1244	0,649948	0,710880
map00960	Tropane, piperidine and pyridine alkaloid biosynthesis	1	1	1913	1244	0,649948	0,710880
map04115	p53 signaling pathway	1	1	1913	1244	0,649948	0,710880
map04210	Apoptosis	1	1	1913	1244	0,649948	0,710880
map05014	Amyotrophic lateral sclerosis (ALS)	1	1	1913	1244	0,649948	0,710880
map05145	Toxoplasmosis	1	1	1913	1244	0,649948	0,710880
map05161	Hepatitis B	1	1	1913	1244	0,649948	0,710880
map05164	Influenza A	1	1	1913	1244	0,649948	0,710880
map05168	Herpes simplex virus 1 infection	1	1	1913	1244	0,649948	0,710880
map05210	Colorectal cancer	1	1	1913	1244	0,649948	0,710880
map05222	Small cell lung cancer	1	1	1913	1244	0,649948	0,710880
map05416	Viral myocarditis	1	1	1913	1244	0,649948	0,710880
map04066	HIF-1 signaling pathway	1	1	1913	1244	0,649948	0,710880
map00624	Polycyclic aromatic hydrocarbon degradation	3	4	1910	1244	0,562865	0,710880
map05133	Pertussis	2	2	1912	1244	0,422313	0,710880
map00791	Atrazine degradation	1	1	1913	1244	0,649948	0,710880

Table S3 continued

map00071	Fatty acid degradation	1	1	1913	1244	0,649948	0,710880
map03008	Ribosome biogenesis in eukaryotes	1	1	1913	1244	0,649948	0,710880
map05205	Proteoglycans in cancer	1	1	1913	1244	0,649948	0,710880
map00770	Pantothenate and CoA biosynthesis	2	2	1912	1244	0,422313	0,710880
map00785	Lipoic acid metabolism	2	2	1912	1244	0,422313	0,710880
map05211	Renal cell carcinoma	2	2	1912	1244	0,422313	0,710880
map00790	Folate biosynthesis	1	1	1913	1244	0,649948	0,710880
map00473	D-Alanine metabolism	2	2	1912	1244	0,422313	0,710880
map04940	Type I diabetes mellitus	1	1	1913	1244	0,649948	0,710880
map05111	Biofilm formation - Vibrio cholerae	1	1	1913	1244	0,649948	0,710880
map00310	Lysine degradation	2	2	1912	1244	0,422313	0,710880
map00380	Tryptophan metabolism	2	2	1912	1244	0,422313	0,710880
map04146	Peroxisome	2	2	1912	1244	0,422313	0,710880
map04141	Protein processing in endoplasmic reticulum	1	1	1913	1244	0,649948	0,710880
map04151	PI3K-Akt signaling pathway	1	1	1913	1244	0,649948	0,710880
map04612	Antigen processing and presentation	1	1	1913	1244	0,649948	0,710880
map04621	NOD-like receptor signaling pathway	1	1	1913	1244	0,649948	0,710880
map04914	Progesterone-mediated oocyte maturation	1	1	1913	1244	0,649948	0,710880
map04915	Estrogen signaling pathway	1	1	1913	1244	0,649948	0,710880
map05215	Prostate cancer	1	1	1913	1244	0,649948	0,710880
map04122	Sulfur relay system	3	4	1910	1244	0,562865	0,710880
map00480	Glutathione metabolism	3	4	1910	1244	0,562865	0,710880
map00760	Nicotinate and nicotinamide metabolism	1	1	1913	1244	0,649948	0,710880
map00363	Bisphenol degradation	1	1	1913	1244	0,649948	0,710880
map00591	Linoleic acid metabolism	1	1	1913	1244	0,649948	0,710880
map00625	Chloroalkane and chloroalkene degradation	1	1	1913	1244	0,649948	0,710880
map00500	Starch and sucrose metabolism	1	1	1913	1244	0,649948	0,710880
map00643	Styrene degradation	1	1	1913	1244	0,649948	0,710880
map04723	Retrograde endocannabinoid signaling	1	1	1913	1244	0,649948	0,710880
map01040	Biosynthesis of unsaturated fatty acids	1	1	1913	1244	0,649948	0,710880
map05142	Chagas disease (American trypanosomiasis)	1	1	1913	1244	0,649948	0,710880
map05143	African trypanosomiasis	1	1	1913	1244	0,649948	0,710880
map00365	Furfural degradation	1	1	1913	1244	0,649948	0,710880
map00908	Zeatin biosynthesis	1	1	1913	1244	0,649948	0,710880
map00250	Alanine, aspartate and glutamate metabolism	2	3	1911	1244	0,718286	0,767634
map00340	Histidine metabolism	2	3	1911	1244	0,718286	0,767634
map00730	Thiamine metabolism	2	3	1911	1244	0,718286	0,767634
map03440	Homologous recombination	27	46	1868	1244	0,855835	0,902208
map00350	Tyrosine metabolism	3	6	1908	1244	0,882875	0,902208
map00471	D-Glutamine and D-glutamate metabolism	1	2	1912	1244	0,877582	0,902208
map00980	Metabolism of xenobiotics by cytochrome P450	1	2	1912	1244	0,877582	0,902208

Table S3 continued

map00982	Drug metabolism - cytochrome P450	1	2	1912	1244	0,877582	0,902208
map05204	Chemical carcinogenesis	1	2	1912	1244	0,877582	0,902208
map00903	Limonene and pinene degradation	1	3	1911	1244	0,957231	0,957231
map00627	Aminobenzoate degradation	1	3	1911	1244	0,957231	0,957231
map00362	Benzoate degradation	1	3	1911	1244	0,957231	0,957231

Table S4 KEGG gene sets enriched in highly expressed genes in UT176. Results of the hypergeometric test. No.- number; h. expr – highly expressed.

KEGG ID	Name of a KEGG pathway	No. of h.expr. genes in pathway	No. of genes in a pathway	No. of genes not in a pathway	No. of expr. genes	p_value	FDR
map03010	Ribosome	50	50	1864	766	4,84E-21	5,71E-19
map00190	Oxidative phosphorylation	32	38	1876	766	2,02E-08	1,19E-06
map01100	Metabolic pathways	87	154	1760	766	0,000012	0,000485
map00910	Nitrogen metabolism	18	23	1891	766	0,000201	0,005921
map03070	Bacterial secretion system	21	30	1884	766	0,000792	0,018686
map00195	Photosynthesis	7	8	1906	766	0,008433	0,165858
map00230	Purine metabolism	16	25	1889	766	0,012745	0,167105
map00240	Pyrimidine metabolism	16	25	1889	766	0,012745	0,167105
map00020	Citrate cycle (TCA cycle)	12	17	1897	766	0,010354	0,167105
map05016	Huntington disease	6	7	1907	766	0,018720	0,215207
map05152	Tuberculosis	4	4	1910	766	0,025533	0,215207
map03018	RNA degradation	7	9	1905	766	0,024840	0,215207
map04626	Plant-pathogen interaction	4	4	1910	766	0,025533	0,215207
map00623	Toluene degradation	4	4	1910	766	0,025533	0,215207
map05012	Parkinson disease	5	6	1908	766	0,040815	0,321080
map02020	Two-component system	10	16	1898	766	0,057754	0,377303
map05120	Epithelial cell signaling in Helicobacter pylori infection	3	3	1911	766	0,063950	0,377303
map00720	Carbon fixation pathways in prokaryotes	9	14	1900	766	0,057838	0,377303
map03020	RNA polymerase	3	3	1911	766	0,063950	0,377303
map00520	Amino sugar and nucleotide sugar metabolism	3	3	1911	766	0,063950	0,377303
map00650	Butanoate metabolism	4	5	1909	766	0,086919	0,488404
map05010	Alzheimer disease	5	7	1907	766	0,096053	0,515194
map01110	Biosynthesis of secondary metabolites	29	60	1854	766	0,115345	0,591771
map00250	Alanine, aspartate and glutamate metabolism	2	3	1911	766	0,352226	0,599260
map00330	Arginine and proline metabolism	4	6	1908	766	0,179127	0,599260
map00360	Phenylalanine metabolism	2	2	1912	766	0,160042	0,599260
map00400	Phenylalanine, tyrosine and tryptophan biosynthesis	1	1	1913	766	0,400209	0,599260
map00401	Novobiocin biosynthesis	1	1	1913	766	0,400209	0,599260
map00950	Isoquinoline alkaloid biosynthesis	1	1	1913	766	0,400209	0,599260
map00960	Tropane, piperidine and pyridine alkaloid biosynthesis	1	1	1913	766	0,400209	0,599260
map01120	Microbial metabolism in diverse environments	20	45	1869	766	0,320578	0,599260
map01210	2-Oxocarboxylic acid metabolism	4	8	1906	766	0,406278	0,599260
map04115	p53 signaling pathway	1	1	1913	766	0,400209	0,599260
map04210	Apoptosis	1	1	1913	766	0,400209	0,599260
map05014	Amyotrophic lateral sclerosis (ALS)	1	1	1913	766	0,400209	0,599260
map05134	Legionellosis	3	4	1910	766	0,179200	0,599260
map05145	Toxoplasmosis	1	1	1913	766	0,400209	0,599260

Table S4 continued

map05161	Hepatitis B	1	1	1913	766	0,400209	0,599260
map05164	Influenza A	1	1	1913	766	0,400209	0,599260
map05168	Herpes simplex virus 1 infection	1	1	1913	766	0,400209	0,599260
map05200	Pathways in cancer	3	4	1910	766	0,179200	0,599260
map05210	Colorectal cancer	1	1	1913	766	0,400209	0,599260
map05222	Small cell lung cancer	1	1	1913	766	0,400209	0,599260
map05416	Viral myocarditis	1	1	1913	766	0,400209	0,599260
map00670	One carbon pool by folate	3	5	1909	766	0,317621	0,599260
map00970	Aminoacyl-tRNA biosynthesis	12	26	1888	766	0,326116	0,599260
map04260	Cardiac muscle contraction	2	3	1911	766	0,352226	0,599260
map00460	Cyanoamino acid metabolism	1	1	1913	766	0,400209	0,599260
map00130	Ubiquinone and other terpenoid-quinone biosynthesis	4	8	1906	766	0,406278	0,599260
map00640	Propanoate metabolism	2	3	1911	766	0,352226	0,599260
map00660	C5-Branched dibasic acid metabolism	2	2	1912	766	0,160042	0,599260
map00564	Glycerophospholipid metabolism	4	8	1906	766	0,406278	0,599260
map03060	Protein export	8	15	1899	766	0,212730	0,599260
map00311	Penicillin and cephalosporin biosynthesis	1	1	1913	766	0,400209	0,599260
map00312	beta-Lactam resistance	1	1	1913	766	0,400209	0,599260
map00561	Glycerolipid metabolism	1	1	1913	766	0,400209	0,599260
map00030	Pentose phosphate pathway	1	1	1913	766	0,400209	0,599260
map00430	Taurine and hypotaurine metabolism	1	1	1913	766	0,400209	0,599260
map03008	Ribosome biogenesis in eukaryotes	1	1	1913	766	0,400209	0,599260
map05205	Proteoglycans in cancer	1	1	1913	766	0,400209	0,599260
map00473	D-Alanine metabolism	2	2	1912	766	0,160042	0,599260
map04940	Type I diabetes mellitus	1	1	1913	766	0,400209	0,599260
map05111	Biofilm formation - Vibrio cholerae	1	1	1913	766	0,400209	0,599260
map04146	Peroxisome	2	2	1912	766	0,160042	0,599260
map04141	Protein processing in endoplasmic reticulum	1	1	1913	766	0,400209	0,599260
map04151	PI3K-Akt signaling pathway	1	1	1913	766	0,400209	0,599260
map04612	Antigen processing and presentation	1	1	1913	766	0,400209	0,599260
map04621	NOD-like receptor signaling pathway	1	1	1913	766	0,400209	0,599260
map04914	Progesterone-mediated oocyte maturation	1	1	1913	766	0,400209	0,599260
map04915	Estrogen signaling pathway	1	1	1913	766	0,400209	0,599260
map05215	Prostate cancer	1	1	1913	766	0,400209	0,599260
map00480	Glutathione metabolism	3	4	1910	766	0,179200	0,599260
map00760	Nicotinate and nicotinamide metabolism	1	1	1913	766	0,400209	0,599260
map00730	Thiamine metabolism	2	3	1911	766	0,352226	0,599260
map00643	Styrene degradation	1	1	1913	766	0,400209	0,599260
map04723	Retrograde endocannabinoid signaling	1	1	1913	766	0,400209	0,599260
map05142	Chagas disease (American trypanosomiasis)	1	1	1913	766	0,400209	0,599260
map05143	African trypanosomiasis	1	1	1913	766	0,400209	0,599260

Table S4 continued

map00365	Furfural degradation	1	1	1913	766	0,400209	0,599260
map00983	Drug metabolism - other enzymes	2	3	1911	766	0,352226	0,599260
map00630	Glyoxylate and dicarboxylate metabolism	3	6	1908	766	0,456114	0,648451
map00710	Carbon fixation in photosynthetic organisms	3	6	1908	766	0,456114	0,648451
map00061	Fatty acid biosynthesis	3	6	1908	766	0,456114	0,648451
map00900	Terpenoid backbone biosynthesis	2	4	1910	766	0,525252	0,720694
map00680	Methane metabolism	2	4	1910	766	0,525252	0,720694
map04122	Sulfur relay system	2	4	1910	766	0,525252	0,720694
map04112	Cell cycle - Caulobacter	12	30	1884	766	0,570540	0,748162
map01230	Biosynthesis of amino acids	8	20	1894	766	0,585429	0,748162
map00620	Pyruvate metabolism	3	7	1907	766	0,580780	0,748162
map00010	Glycolysis / Gluconeogenesis	3	7	1907	766	0,580780	0,748162
map00780	Biotin metabolism	3	7	1907	766	0,580780	0,748162
map05133	Pertussis	1	2	1912	766	0,640376	0,748162
map00785	Lipoic acid metabolism	1	2	1912	766	0,640376	0,748162
map00471	D-Glutamine and D-glutamate metabolism	1	2	1912	766	0,640376	0,748162
map00770	Pantothenate and CoA biosynthesis	1	2	1912	766	0,640376	0,748162
map00310	Lysine degradation	1	2	1912	766	0,640376	0,748162
map05211	Renal cell carcinoma	1	2	1912	766	0,640376	0,748162
map00380	Tryptophan metabolism	1	2	1912	766	0,640376	0,748162
map00980	Metabolism of xenobiotics by cytochrome P450	1	2	1912	766	0,640376	0,748162
map00982	Drug metabolism - cytochrome P450	1	2	1912	766	0,640376	0,748162
map05204	Chemical carcinogenesis	1	2	1912	766	0,640376	0,748162
map03410	Base excision repair	3	8	1906	766	0,685481	0,793007
map00270	Cysteine and methionine metabolism	1	3	1911	766	0,784451	0,857086
map00562	Inositol phosphate metabolism	1	3	1911	766	0,784451	0,857086
map00550	Peptidoglycan biosynthesis	4	12	1902	766	0,775889	0,857086
map00627	Aminobenzoate degradation	1	3	1911	766	0,784451	0,857086
map00362	Benzoate degradation	1	3	1911	766	0,784451	0,857086
map00290	Valine, leucine and isoleucine biosynthesis	1	3	1911	766	0,784451	0,857086
map00051	Fructose and mannose metabolism	1	4	1910	766	0,870851	0,934186
map00624	Polycyclic aromatic hydrocarbon degradation	1	4	1910	766	0,870851	0,934186
map00260	Glycine, serine and threonine metabolism	2	8	1906	766	0,894334	0,950734
map03030	DNA replication	9	30	1884	766	0,908121	0,950918
map03430	Mismatch repair	4	15	1899	766	0,910624	0,950918
map00350	Tyrosine metabolism	1	6	1908	766	0,953685	0,987147
map03440	Homologous recombination	8	46	1868	766	0,999777	0,999777
map00860	Porphyrin and chlorophyll metabolism	1	11	1903	766	0,996455	0,999777
map02010	ABC transporters	1	13	1901	766	0,998735	0,999777
map00300	Lysine biosynthesis	1	9	1905	766	0,990080	0,999777

Table S5 Partly-conserved operons identified in both Karp and UT176. 'gene_name' represents a gene name derived from the GenBank annotation; 'predicted_gene' is a gene name predicted by EggNOG-mapper; 'gene_function' describes a gene product retrieved from GenBank annotation; If an identified operon was also identified in the comparative genomics analyses (Batty et al., 2018), the column 'Islands' defines a number of the island described in the conserved islands table present in the "Supplementary Data 1-22" excel file.

Orthologs	Locus tag UT176	Gene name UT176	Predicted gene UT176	Gene function UT176	Locus tag Karp	Gene name Karp	Predicted gene Karp	Gene function Karp	Islands
UT176_00524 - Karp_00794	UT176_00524			integrase	Karp_00794			integrase	
UT176_00525 - Karp_00795	UT176_00525			integrase	Karp_00795			integrase	
UT176_00527 - Karp_00796	UT176_00526			hypothetical protein					
	UT176_00527			integrase					
Orthologs	Locus tag UT176	Gene name UT176	Predicted gene UT176	Gene function UT176	Locus tag Karp	Gene name Karp	Predicted gene Karp	Gene function Karp	Islands
UT176_00894 - Karp_02135	UT176_00894		ILVE	branched chain amino acid aminotransferase	Karp_02135		ILVE	branched chain amino acid aminotransferase	
UT176_00895 - Karp_02134	UT176_00895		DAPA	4-hydroxy-tetrahydrodipicolinate synthase	Karp_02134		DAPA	4-hydroxy-tetrahydrodipicolinate synthase	
UT176_00896 - Karp_02133	UT176_00896		SMPB	SsrA-binding protein	Karp_02133		SMPB	SsrA-binding protein	
UT176_00897 - Karp_02132	UT176_00897		SPPA	signal peptide peptidase SppA	Karp_02132		SPPA	signal peptide peptidase SppA	
UT176_00898 - Karp_02131	UT176_00898		LIPA	lipoyl synthase	Karp_02131		LIPA	lipoyl synthase	
UT176_00899 - Karp_02130	UT176_00899		YFJG	ubiquinone-binding protein	Karp_02130		YFJG	ubiquinone-binding protein	
					Karp_02129		PURC	phosphoribosylaminoimidazolesuccinocarboxamide synthase	
Orthologs	Locus tag UT176	Gene name UT176	Predicted gene UT176	Gene function UT176	Locus tag Karp	Gene name Karp	Predicted gene Karp	Gene function Karp	Islands
UT176_01555 - Karp_01515	UT176_01555		ATPD	ATP synthase subunit beta	Karp_01515		ATPD	ATP synthase subunit beta	
UT176_01556 - Karp_01514	UT176_01556		ATPC	ATP synthase subunit epsilon	Karp_01514		ATPC	ATP synthase subunit epsilon	
UT176_01557 - Karp_01147	UT176_01557			IS110 family transposase					

Table S5. continued

Orthologs	Locus tag UT176	Gene name UT176	Predicted gene UT176	Gene function UT176	Locus tag Karp	Gene name Karp	Predicted gene Karp	Gene function Karp	Islands
UT176_01151 - Karp_01773	UT176_01151			integrase	Karp_00789			integrase	
UT176_01152 - Karp_00789	UT176_01152			integrase	Karp_00790			integrase	
UT176_01153 - Karp_00790	UT176_01153			integrase					
Orthologs	Locus tag UT176	Gene name UT176	Predicted gene UT176	Gene function UT176	Locus tag Karp	Gene name Karp	Predicted gene Karp	Gene function Karp	Islands
UT176_01015 - Karp_02348	UT176_01015		TRAD	conjugal transfer protein	Karp_02343			conjugal transfer protein	
UT176_01016 - Karp_02349	UT176_01016		TRAD	conjugal transfer protein	Karp_02344		TRAH	conjugal transfer protein TraH	
					Karp_02345		TRAH	conjugal transfer protein TraH	
					Karp_02346		TRAG	conjugal transfer protein TraG	
					Karp_02347			hypothetical protein	
					Karp_02348		TRAD	conjugal transfer protein	
					Karp_02349		TRAD	conjugal transfer protein	
Orthologs	Locus tag UT176	Gene name UT176	Predicted gene UT176	Gene function UT176	Locus tag Karp	Gene name Karp	Predicted gene Karp	Gene function Karp	Islands
UT176_01319 - Karp_02278	UT176_01319			hypothetical protein	Karp_02278			hypothetical protein	
UT176_01320 - Karp_02279	UT176_01320			conjugative transfer protein	Karp_02279			conjugal transfer protein TraE	
					Karp_02280			conjugative transfer protein	
					Karp_02281			hypothetical protein	

Table S5. continued

Orthologs	Locus tag UT176	Gene name UT176	Predicted gene UT176	Gene function UT176	Locus tag Karp	Gene name Karp	Predicted gene Karp	Gene function Karp	Islands
					Karp_01353			hypothetical protein	
UT176_01721 - Karp_01352	UT176_01721		ERUM6880	hypothetical protein	Karp_01352		ERUM6880	hypothetical protein	13
UT176_01722 - Karp_01351	UT176_01722		YGFA	5-formyltetrahydrofolate cyclo-ligase	Karp_01351		YGFA	5-formyltetrahydrofolate cyclo-ligase	13

Table S6 Summary of primers and probes used in this study. The experiments were performed by Alexander J. Westermann and Suparat Giengkam.

Northern blot probes	name	sequence (5'->3')	length [nt]	description	reference
	AWO-009	TACCTCTATTCTTAA TAAAACTTATTGCC	29	antisense Northern probe against Orientia tmRNA (5')	this study
	AWO-010	TGATTTCCTTAAGCT GCTAATG	22	antisense Northern probe against Orientia tmRNA (3')	this study
	AWO-011	GGACTTTCCTCACA AATCTAT	21	antisense Northern probe against Orientia RNaseP RNA	this study
	AWO-013	GTTGATGCCTACGC CAGTTA	20	antisense Northern probe against Orientia SRP RNA	this study
	AWO-022	CTCTCCCATGTTTA AACATA	20	antisense Northern probe against Orientia 5S rRNA	this study
	JVO-7672	ATATGGAACGCTTC ACGAATTTG	23	antisense Northern probe against human U6 snRNA	PMID:26789254
qRT-PCR primers	**name**	**sequence (5'->3')**	**length [nt]**	**description**	**reference**
	JVO-8896	TCGGTACATCCTCG ACGG	18	sense qPCR primer against human IL6 mRNA	this study
	JVO-8897	TGTTTTCTGCCAGT GCCTC	19	antisense qPCR primer against human IL6 mRNA	this study
	JVO-14331	GCTGTGAAGATACG GGAGAGAAC	23	sense qPCR primer against human TNFAIP3 mRNA	this study
	JVO-14332	CCTGGATGTTTCTG TCGATGAG	22	antisense qPCR primer against human TNFAIP3 mRNA	this study
	JVO-9476	AAGTGGCTATGCTC AAAATG	20	sense qPCR against human IL7R mRNA	PMID:21307942
	JVO-9477	TTCAGGCACTTTAC CTCCAC	20	antisense qPCR against human IL7R mRNA	PMID:21307942
	AWO-007	CAAACGATAGGCTC AAACACT	22	sense qRT-PCR oligo for human IL33 mRNA	this study
	AWO-008	TGAGTGTTGCCTAA GACATC	20	antisense qRT-PCR oligo for human IL33 mRNA	this study
	JVO-13531	ATGCAGGAAGAACA TGACAACC	22	sense qRT-PCR oligo for human IFIT1 mRNA	this study
	JVO-13532	TCTGGACACTCCAT TCTATAGCG	23	antisense qRT-PCR oligo for human IFIT1 mRNA	this study
	JVO-7673	GCTTCGGCAGCAC ATATACTAAAAT	25	sense qPCR primer against human U6 snRNA	PMID:26789254
	JVO-7672	ATATGGAACGCTTC ACGAATTTG	23	antisense qPCR primer against human U6 snRNA	PMID:26789254
RT-PCR primers	**name**	**sequence (5'->3')**	**length [nt]**	**description**	**reference**
	47kda FW	TCCAGAATTAAAT GAGAATTTAGGAC	26	Amplification of 47kda gene in Ot for bacterial quantification	PMID: 26317517
	47 kda RV	TTAGTAATTACATC TCCAGGAGCAA	25	Amplification of 47kda gene in Ot for bacterial quantification	PMID: 26317517
	47 kda PROBE	FAM-TTCCACATTGTGC TGCAGATCCTTC-	25	Amplification of 47kda gene in Ot for bacterial quantification	PMID: 26317517

Appendix 2 Supplementary data for chapter 3

Supplementary data associated with this study can be found on the CD attached to this thesis.

The *Benchmark_analysis_RNA_classes.csv* file contains results of the benchmark analysis of various strategies implemented in the Dualrnaseq pipeline in quantifying RNA classes of different host-pathogen systems.

```
/*
 * -------------------------------------------------
 *  nf-core/dualrnaseq Nextflow config file
 * -------------------------------------------------
 * Default config options for all environments.
 */

// Global default params, used in configs
params {

  //--------
  // Workflow flags:
  //--------
  genome_host = false
  genome_pathogen = false
  input = "data/*{1,2}.fastq.gz"
  single_end = false
  outdir = './results'
  publish_dir_mode = 'copy'

  //--------
  // Cutadapt:
  //--------
  run_cutadapt = false
  a = "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA"
  A = "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT"
  quality_cutoff = "10"
  cutadapt_params = ""


  //--------
  // BBDuk:
  //--------
  run_bbduk = false
  minlen = "18"
  qtrim = "r"
  trimq = "10"
  ktrim = "r"
  k = "17"
  mink = "11"
  hdist = "1"
  adapters = "$projectDir/assets/adapters.fa"
  bbduk_params = ""

  //--------
  // FastQC:
  //--------
  skip_fastqc = false
  fastqc_params = ""

  //--------
  // Salmon - general options available for both modes, Selective
Alignment and alignment-based mode
  //--------
  libtype = ''
  generate_salmon_uniq_ambig = false
  incompatPrior = 0.0
  gene_attribute_gff_to_create_transcriptome_host = "transcript_id"
  gene_feature_gff_to_create_transcriptome_host = ["exon", "tRNA"]
  gene_attribute_gff_to_create_transcriptome_pathogen = "locus_tag"
  gene_feature_gff_to_create_transcriptome_pathogen =
["gene","sRNA","tRNA","rRNA"]
  read_transcriptome_fasta_host_from_file = false
  read_transcriptome_fasta_pathogen_from_file = false

  //--------
  // Salmon Selective Alignment
  //--------
  run_salmon_selective_alignment = false
  kmer_length = 21
  writeUnmappedNames = false
  softclipOverhangs = false
  dumpEq = false
  writeMappings = false
  keepDuplicates = false
  salmon_sa_params_index = ""
  salmon_sa_params_mapping = ""


  //--------
  // STAR - general options available for both modes, genome mapping
with HTSeq quantification and salmon - alignment-based mode:
  //--------
  run_star = false
  outSAMunmapped = "Within"
  outSAMattributes = "Standard"
  outFilterMultimapNmax = 999
  outFilterType = "BySJout"
  alignSJoverhangMin = 8
  alignSJDBoverhangMin = 1
  outFilterMismatchNmax = 999
  outFilterMismatchNoverReadLmax = 1
  alignIntronMin = 20
  alignIntronMax = 1000000
  alignMatesGapMax = 1000000
  limitBAMsortRAM = 0
  winAnchorMultimapNmax = 999
  sjdbOverhang = 100


  //--------
  // STAR - additional options available only for genome mapping
with HTSeq quantification mode
  //--------
  outWigType = "None"
  outWigStrand = "Stranded"
  star_index_params = ""
  star_alignment_params = ""
  //--------
  // STAR - additional options available only for Salmon -
alignment-based mode:
  //--------
  quantTranscriptomeBan = "Singleend"
  star_salmon_index_params = ""
  star_salmon_alignment_params = ""

  //--------
  // Salmon - alignment-based mode:
  //--------
  run_salmon_alignment_based_mode = false
  salmon_alignment_based_params = ""

  //--------
  // HTSeq:
  //--------
  run_htseq_uniquely_mapped = false
  stranded = "yes"
  max_reads_in_buffer = 30000000
  minaqual = 10
  gene_feature_gff_to_quantify_host = ["exon","tRNA"]
  gene_feature_gff_to_quantify_pathogen = ["gene", "sRNA", "tRNA",
"rRNA"]
  host_gff_attribute = "gene_id"
  pathogen_gff_attribute = "locus_tag"
  htseq_params = ""

  //--------
  // mapping statistics:
  //--------
  mapping_statistics = false
  rna_classes_to_replace_host = "$projectDir/data/
RNA_classes_to_replace.tsv"



  //--------
  // Options: Custom config:
  //--------
  custom_config_version = 'master'
  custom_config_base = "https://raw.githubusercontent.com/nf-core/
configs/${params.custom_config_version}"
  hostnames = false
  config_profile_description = false
  config_profile_contact = false
  config_profile_url = false


  //--------
  // Options: Other:
  //--------
  name = false
  multiqc_config = false
  email = false
  email_on_fail = false
  max_multiqc_email_size = 25.MB
  plaintext_email = false
  monochrome_logs = false
  help = false
  tracedir = "${params.outdir}/pipeline_info"

  // Defaults only, expecting to be overwritten
  max_memory = 128.GB
  max_cpus = 16
  max_time = 240.h

}


//--------
// Work directory:
//--------
//This location stores the working temporary files and downloaded
images
//Note: If commented out - errors sometimes occur re failing to
publish files (although files are published correctly)
workDir = "$projectDir/work"


//--------
// Container:
//--------
process.container = 'nfcore/dualrnaseq:1.0.0'

// Load base.config by default for all pipelines
includeConfig 'conf/base.config'


//Option to use a custom configuration file (which is included in
conf/genomes.conf)
//false is default and thus the config file will be used.
//To not use, you can simply pass --genomes_ignore on the command
line
genomes_ignore = false

// Load genomes.config if required
if (!params.genomes_ignore) {
  includeConfig 'conf/genomes.config'
}
```

Figure S8 Fragment of the nextflow.config file containing all default parameters of the Dualrnaseq pipeline.

A.

```
/*
 * -------------------------------------------------
 *  nf-core/dualrnaseq Nextflow base config file
 * -------------------------------------------------
 * A 'blank slate' config file, appropriate for general
 * use on most high performace compute environments.
 * Assumes that all software is installed and available
 * on the PATH. Runs in `local` mode - all jobs will be
 * run on the logged in environment.
 */

process {

  cpus = { check_max( 2 * task.attempt, 'cpus' ) }
  memory = { check_max( 10.GB * task.attempt, 'memory' ) }
  time = { check_max( 4.h * task.attempt, 'time' ) }

//  errorStrategy = { task.exitStatus in [143,137,104,134,139] ?
'retry' : 'finish' }
  errorStrategy = 'retry'
  maxRetries = 3
  maxErrors = '-1'

  // Process-specific resource requirements
  withLabel:process_low {
    cpus = { check_max( 2 * task.attempt, 'cpus' ) }
    memory = { check_max( 14.GB * task.attempt, 'memory' ) }
    time = { check_max( 6.h * task.attempt, 'time' ) }
  }
  withLabel:process_medium {
    cpus = { check_max( 15 * task.attempt, 'cpus' ) }
    memory = { check_max( 30.GB * task.attempt, 'memory' ) }
    time = { check_max( 8.h * task.attempt, 'time' ) }
  }
  withLabel:process_high {
    cpus = { check_max( 20 * task.attempt, 'cpus' ) }
    memory = { check_max( 30.GB * task.attempt, 'memory' ) }
    time = { check_max( 10.h * task.attempt, 'time' ) }
  }
  withLabel:process_long {
    time = { check_max( 20.h * task.attempt, 'time' ) }
  }
  withName:get_software_versions {
    cache = false
  }
}
```

B.

```
/*
 * -------------------------------------------------
 *  Nextflow config file for running tests
 * -------------------------------------------------
 * Defines bundled input files and everything required
 * to run a fast and simple test. Use as follows:
 *   nextflow run nf-core/dualrnaseq -profile test,<docker/
singularity>
 */

params {
  config_profile_name = 'Test profile'
  config_profile_description = 'Minimal test dataset to check
pipeline function'

  // Limit resources so that this can run on GitHub Actions
  max_cpus = 2
  max_memory = 6.GB
  max_time = 48.h

  // Input data
  single_end = false


  // Run processes when the profile is set to 'test'
  genomes_ignore = true
  run_bbduk = true
  qtrim = "rl"
  run_salmon_selective_alignment = true
  libtype = "ISF" //Salmon - paired-end reads
  mapping_statistics = true


  input_paths = [
    ['sample_R1', ['https://raw.githubusercontent.com/nf-core/test-
datasets/dualrnaseq/PE_reads/sample_R1_1.fq', 'https://
raw.githubusercontent.com/nf-core/test-datasets/dualrnaseq/PE_reads/
sample_R1_2.fq']],
    ['sample_R2', ['https://raw.githubusercontent.com/nf-core/test-
datasets/dualrnaseq/PE_reads/sample_R2_1.fq', 'https://
raw.githubusercontent.com/nf-core/test-datasets/dualrnaseq/PE_reads/
sample_R2_2.fq']],
    ['sample_R3', ['https://raw.githubusercontent.com/nf-core/test-
datasets/dualrnaseq/PE_reads/sample_R3_1.fq', 'https://
raw.githubusercontent.com/nf-core/test-datasets/dualrnaseq/PE_reads/
sample_R3_2.fq']]
  ]


// Genome references
  genomes {
    'test_host' {
      fasta_host  = "https://github.com/nf-core/test-datasets/raw/
dualrnaseq/references/GRCh38.p13_sub.fasta"
      gff_host = "https://github.com/nf-core/test-datasets/raw/
dualrnaseq/references/Human_gencode.v33_sub.gff3"
      gff_host_tRNA = ""
    }
    'test_pathogen' {
      fasta_pathogen  = "https://github.com/nf-core/test-datasets/
raw/dualrnaseq/references/SL1344_sub.fasta"
      gff_pathogen = "https://github.com/nf-core/test-datasets/raw/
dualrnaseq/references/SL1344_sub.gff3"
    }
  }
genome_host = 'test_host'
genome_pathogen = 'test_pathogen'
}
```

C.

```
/*
 * -------------------------------------------------
 *  Nextflow config file for host/pathogen reference paths
 * -------------------------------------------------
 * This file allows you to define host and pathogen references and
create a permanent link to the files.
 * Below, we show the general style that should be used. This file
can be populated with a large number of different hosts and
pathogens.
 * For an example usage: to use the human and salmonella references
here, add --genome_host GRCh38 --genome_pathogen SL1344 to the
command line
 * Annotation files are required to be in the GFF3 format - not GTF
 */

params {
  genomes {
    'GRCh38' {
      fasta_host          = "path_to_references/human/
GRCh38.genome.fa"
      gff_host            = "path_to_references/human/
GRCh38.annotation.gff3"
      gff_host_tRNA       = "path_to_references/human/
GRCh38.gencode.tRNAs.gff3"
      transcriptome_host  = "path_to_references/human/
GRCh38.gencode.transcripts.fa"
    }

    'SL1344' {
      fasta_pathogen      = "path_to_references/Salmonella/
SL1344.fasta"
      gff_pathogen        = "path_to_references/Salmonella/
SL1344.gff"
    }

    'My_Bacteria' {
      fasta_pathogen      = "path_to_references/My_Bacteria/
My_Bacteria.fasta"
      gff_pathogen        = "path_to_references/My_Bacteria/
My_Bacteria.gff"
    }
}
}
```

Figure S9 Configuration files of the Dualrnaseq pipeline. A) base.config file, B) test.config file, C) genome.config file.

Table S7 Benchmark analysis of various mapping-quantification strategies using different host-pathogen systems. Spearman's rank correlation coefficients and associated p-value computed between the Number of reads estimated with one of the mapping-quantification strategies and the ground truth for either host or pathogen from various dual RNA-seq samples.

Sample name	r_s STAR - HTSeq	p-value STAR - HTSeq	r_s STAR - Salmon	p-value STAR - Salmon	r_s Salmon SA	p-value Salmon SA	Host- pathogen system	organism
HeLa_S3_WT_02_h_R1	0.917	0.0	0.959	0.0	0.961	0.0	Hela - Salmonella	host
HeLa_S3_WT_02_h_R2	0.915	0.0	0.959	0.0	0.962	0.0	Hela - Salmonella	host
HeLa_S3_WT_02_h_R3	0.915	0.0	0.960	0.0	0.963	0.0	Hela - Salmonella	host
HeLa_S3_WT_04_h_R1	0.916	0.0	0.959	0.0	0.9601	0.0	Hela - Salmonella	host
HeLa_S3_WT_04_h_R2	0.914	0.0	0.96	0.0	0.961	0.0	Hela - Salmonella	host
HeLa_S3_WT_04_h_R3	0.910	0.0	0.957	0.0	0.961	0.0	Hela - Salmonella	host
HeLa_S3_WT_08_h_R1	0.906	0.0	0.953	0.0	0.958	0.0	Hela - Salmonella	host
HeLa_S3_WT_08_h_R2	0.907	0.0	0.957	0.0	0.96	0.0	Hela - Salmonella	host
HeLa_S3_WT_08_h_R3	0.91	0.0	0.958	0.0	0.960	0.0	Hela - Salmonella	host
HeLa_S3_WT_16_h_R1	0.907	0.0	0.956	0.0	0.958	0.0	Hela - Salmonella	host
HeLa_S3_WT_16_h_R2	0.905	0.0	0.955	0.0	0.958	0.0	Hela - Salmonella	host
HeLa_S3_WT_16_h_R3	0.911	0.0	0.955	0.0	0.959	0.0	Hela - Salmonella	host
HeLa_S3_WT_24_h_R1	0.906	0.0	0.957	0.0	0.959	0.0	Hela - Salmonella	host
HeLa_S3_WT_24_h_R2	0.907	0.0	0.959	0.0	0.96	0.0	Hela - Salmonella	host
HeLa_S3_WT_24_h_R3	0.907	0.0	0.957	0.0	0.959	0.0	Hela - Salmonella	host
HeLa_S3_WT_02_h_R1	0.96	0.0	0.971	0.0	0.965	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_02_h_R2	0.955	0.0	0.972	0.0	0.967	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_02_h_R3	0.964	0.0	0.974	0.0	0.965	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_04_h_R1	0.954	0.0	0.965	0.0	0.963	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_04_h_R2	0.951	0.0	0.964	0.0	0.963	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_04_h_R3	0.949	0.0	0.967	0.0	0.962	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_08_h_R1	0.93	0.0	0.96	0.0	0.968	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_08_h_R2	0.929	0.0	0.954	0.0	0.967	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_08_h_R3	0.944	0.0	0.965	0.0	0.962	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_16_h_R1	0.92	0.0	0.959	0.0	0.966	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_16_h_R2	0.9157	0.0	0.955	0.0	0.969	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_16_h_R3	0.919	0.0	0.959	0.0	0.964	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_24_h_R1	0.903	0.0	0.965	0.0	0.979	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_24_h_R2	0.902	0.0	0.965	0.0	0.975	0.0	Hela - Salmonella	pathogen
HeLa_S3_WT_24_h_R3	0.908	0.0	0.947	0.0	0.972	0.0	Hela - Salmonella	pathogen
HUVEC_Karp_R1	0.907	0.0	0.960	0.0	0.962	0.0	Karp - HUVEC	host
HUVEC_Karp_R2	0.907	0.0	0.959	0.0	0.961	0.0	Karp - HUVEC	host
HUVEC_Karp_R3	0.909	0.0	0.96	0.0	0.962	0.0	Karp - HUVEC	host
HUVEC_Karp_R1	0.811	0.0	0.876	0.0	0.884	0.0	Karp - HUVEC	pathogen
HUVEC_Karp_R2	0.806	0.0	0.868	0.0	0.888	0.0	Karp - HUVEC	pathogen
HUVEC_Karp_R3	0.811	0.0	0.876	0.0	0.883	0.0	Karp - HUVEC	pathogen
HUVEC_UT176_R1	0.909	0.0	0.961	0.0	0.963	0.0	UT176 - HUVEC	host
HUVEC_UT176_R2	0.909	0.0	0.961	0.0	0.965	0.0	UT176 - HUVEC	host
HUVEC_UT176_R3	0.907	0.0	0.961	0.0	0.963	0.0	UT176 -HUVEC	host
HUVEC_UT176_R1	0.91	0.0	0.939	0.0	0.949	0.0	UT176 - HUVEC	pathogen
HUVEC_UT176_R2	0.910	0.0	0.953	0.0	0.956	0.0	UT176 -HUVEC	pathogen
HUVEC_UT176_R3	0.913	0.0	0.939	0.0	0.946	0.0	UT176- HUVEC	pathogen

Table S8 Benchmark analysis of RAGE genes of Ot str. Karp. Spearman's rank correlation coefficients and associated p-value computed between the Number of reads estimated with one of the mapping-quantification strategy and the ground truth for the RAGE genes of Karp.

Sample name	r_s STAR - HTSeq	p-value STAR - HTSeq	r_s STAR - Salmon	p-value STAR - Salmon	r_s Salmon SA	p-value Salmon SA
Karp_R1	0.671	1.299e-182	0.808	1.17e-320	0.814	0.0
Karp_R2	0.665	2.531e-178	0.796	1.952e-304	0.819	0.0
Karp_R3	0.676	1.436e-186	0.812	0.0	0.808	1.204e-320

Curriculum vitae

Publication list

Mika-Gospodorz, **B.**, Giengkam, S., Westermann, A. J., Wongsantichon, J., Kion-Crosby, W., Chuenklin, S., Wang, L. C., Sunyakumthorn, P., Sobota, R. M., Subbian, S., Vogel, J., Barquist, L., & Salje, J. (2020). Dual RNA-seq of Orientia tsutsugamushi informs on host-pathogen interactions for this neglected intracellular human pathogen. *Nature Communications*, *11*(1), 3363.

Manuscripts in preparation:

Mika-Gospodorz, B., Hayward, R., Barquist, L. (TBD)
Dualrnaseq: a Nextflow-based workflow for host-pathogen dual RNA-seq analysis. (TBD)

Unpublished research projects conducted during the PhD work:

- Analysis of dual RNA-seq data from *Salmonella* Typhimurium (WT and *ΔgtgE*) and *Salmonella* Typhi (WT and strain expressing *gtgE*) infecting Bone marrow-derived macrophages.
  Supervised by Jun.-Prof. Dr. Lars Barquist.
  Collaboration with Natalia Cattelan (University of Aberdeen).
- Development of a tool for design of guide RNA library for CRISPR base editing of bacterial genomes.
  Supervised by Jun.-Prof. Dr. Lars Barquist.
  Collaboration with Prof. Dr. Chase Beisel (Helmholtz Institute for RNA-based Infection Research), Scott Collins (North Carolina State University), Yanying Yu (Helmholtz Institute for RNA-based Infection Research).
- Initial analysis of time series dual RNA-seq data of *Salmonella* Typhimurium infecting epithelial cells (HeLa cell line) and macrophages (Bone marrow-derived macrophages).
  Supervised by Jun.-Prof. Dr. Lars Barquist.
  Collaboration with Jun.-Prof. Dr. Alexander Westermann.

Acknowledgements

Foremost, I would like to thank my supervisor Jun.-Prof. Lars Barquist for giving me an excellent possibility to open my horizons to a new area of research and work on such interesting projects that allow me to grow as a scientist. Thank you for always being available for discussions. Your continuous guidance and open-mindedness have allowed me to accomplish my PhD work successfully.

I would like to acknowledge my thesis committee members — Jun.-Prof. Alexander Westerman, Prof. Petra Dersch, and Prof. Vera Kozjak-Pavlovic — for all discussions on my work, advice, and positivity during our meetings. I thank to Prof. Thomas Dandekar for chairing the doctoral committee and the defense of this thesis.

I am thankful to my colleagues from the Barquist lab for creating a friendly atmosphere. Special thanks to Laura Jenniches for her companionship and support from the beginning of my PhD; Shuba Varshini Alampalli for all discussions on coding problems; and Regan Hayward for the amazing collaboration that resulted in the first release of the nf-core/Dualrnaseq pipeline before my leave.

I thank all people I shared the office with for making it a pleasant and friendly workplace.

I thank Michael Kütt for IT technical support. Also, I would like to thank Alice Hohn and Julia Mendorff for helping me with the administrative work.

I thank the Graduate School of Life Sciences (GSLS) of the University of Würzburg for giving me the possibility to become part of their outstanding, interdisciplinary program.

I am thankful for all proof-readers of this thesis, especially Laura Jenniches and Willow Kion-Crosby.

I express my gratitude to the people who guided and inspired me to choose the direction in my career that led me to this place.

Na koniec chciałabym bardzo podziękować mojej rodzinie za ogromne wsparcie i miłość, które dają mi siłę pokonywać wszelkie trudności. W szczególności, jestem wdzięczna moim rodzicom za wszelką pomoc i zapewnienie warunków, które umożliwiły mój rozwój i podążanie za marzeniami. Chciałabym również podziękować mojemu mężowi, za wsparcie niezależnie od ilości dzielących nas kilometrów. Dziękuję Ci za wyrozumienie i motywowanie mnie do dalszego działania.

Affidavit

I hereby confirm that my thesis entitled “Development and application of bioinformatics tools for analysis of dual RNA-seq experiments” is the result of my own work. I did not receive any help or support from commercial consultants. All sources and/or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

_____ _____

Place, Date Bożena Mika-Gospodorz

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation “Entwicklung und Anwendung von Bioinformatikwerkzeugen für die Analyse von dualen RNA-seq Experimenten”, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

_____ _____

Ort, Datum Bożena Mika-Gospodorz