dRNASb: a systems biology approach to decipher dynamics of host-pathogen interactions using temporal dual RNA-seq data

Mojdeh Dinarvand1,*, Forrest C. Koch1, Daniel Al Mouiee1,2,3, Kaylee Vuong1, Abhishek Vijayan1, Afia Fariha Tanzim1, A. K. M. Azad4, Anahit Penesyan5, Natalia Castaño-Rodríguez1 and Fatemeh Vafaee1,3,*

Abstract

Infection triggers a dynamic cascade of reciprocal events between host and pathogen wherein the host activates complex mechanisms to recognise and kill pathogens while the pathogen often adjusts its virulence and fitness to avoid eradication by the host. The interaction between the pathogen and the host results in large-scale changes in gene expression in both organisms. Dual RNA-seq, the simultaneous detection of host and pathogen transcripts, has become a leading approach to unravelling complex molecular interactions between the host and the pathogen and is particularly informative for intracellular organisms. The amount of in vitro and in vivo dual RNA-seq data is rapidly growing, which demands computational pipelines to effectively analyse such data. In particular, holistic, systems-level, and temporal analyses of dual RNA-seq data are essential to enable further insights into the host-pathogen transcriptional dynamics and potential interactions. Here, we developed an integrative network-driven bioinformatics pipeline, dRNASb, a systems biology-based computational pipeline to analyse temporal transcriptional clusters, incorporate molecular interaction networks (e.g. protein-protein interactions), identify topologically and functionally key transcripts in host and pathogen, and associate host and pathogen temporal transcriptome to decipher potential between-species interactions. The pipeline is applicable to various dual RNA-seq data from different species and experimental conditions. As a case study, we applied dRNASb to analyse temporal dual RNA-seq data of Salmonella-infected human cells, which enabled us to uncover genes contributing to the infection process and their potential functions and to identify putative associations between host and pathogen genes during infection. Overall, dRNASb has the potential to identify key genes involved in bacterial growth or host defence mechanisms for future uses as therapeutic targets.

DATA SUMMARY

The data set used to showcase the pipeline is available in NCBI's Gene Expression Omnibus under accession number GSE60144. The pipeline code is available at the GitHub repository https://github.com/VafaeeLab/dRNASb. The authors confirm all supporting data and code have been provided within the article and the GitHub repository or through supplementary data.

INTRODUCTION

Infectious diseases continue to be a major threat worldwide, with devastating impacts on lives and economies around the globe. A better understanding of microbial infection mechanisms is central to the development of new antimicrobial therapies [1]. During bacterial infections, a complex interplay is engaged by both bacterium and the host to negotiate the respective survival and defence strategies [2]. Unravelling pathogen and host regulatory interactions, virulence mechanisms and innate...
Impact Statement

Dual RNA sequencing is an important emerging technique to study large-scale transcriptional changes in host and pathogen simultaneously during infection. However, computational tools for the downstream analyses of such data, past the initial read mapping steps, particularly at the systems level are still lacking. This study is aimed to address this methodological need. Here, we describe the development and application of a bioinformatics framework for the analysis of temporal dual RNA-seq profiles of pathogens and their host cells during intracellular infection. We established a general, network-driven pipeline for the analyses of such data with wide applicability and impact on the microbial genomics community.

responses has led to the understanding of the dynamics of infectious processes and the development of therapeutics [3, 4]. In studying infections, analysing the transcriptomes of both organisms involved is key to better understand pathogenesis and disease progression [5, 6]. Dual RNA sequencing (dual RNA-seq) is an emerging powerful tool to capture whole transcriptomes, including coding and non-coding RNAs, of both host and pathogen simultaneously in order to dissect the host-pathogen interplay and reveal the impacts that both organisms exert over each other during infection [4, 7–9]. Dual RNA-seq experiments are often designed to capture transcriptomics profiles of both organisms at different time-points after infection to enable the study of the molecular dynamics underlying host response and bacterial fitness [10]. To understand the strategies employed by a pathogen for controlling or progressing infection, a growing number of dual RNA-seq experiments have been conducted under in vivo and in vitro conditions [2–6] for pathogens such as Yersinia pseudotuberculosis [11], Pseudomonas aeruginosa [12, 13], and Mycobacterium leprae [14], as well as intracellular bacteria such as Mycobacterium tuberculosis [7, 15] and Piscirickettsia salmonis [16].

Dual RNA-seq experiments inherently generate a mixture of transcripts from both host and pathogen, requiring computational approaches to sort this mixture into the components associated with each organism. Multiple analytical workflows have been developed to align this read mixture to the reference genomes of the host and pathogen in model organisms [4, 7, 17] as well as non-model systems by leveraging the genomic resources of closely-related species [18]. The majority of dual RNA-seq bioinformatics pipelines generated to date have been focused on upstream analyses (e.g. quality control, alignment, read filtration, and normalisation) [18], with downstream analyses often limited to differential expression and functional analyses [7, 17]. Hence, there is a need for computational resources that offer more holistic and temporal downstream analyses of host and pathogen transcriptome, enabling further insights into the host-pathogen transcriptional dynamics and potential interactions. In particular, systems biology approaches which integrate information on molecular interactions with gene expression information [19, 20], can provide a more holistic understanding of molecular intricacies underlying the infection process. However, the application of systems biology is still limited in this context.

To address this resource need, we developed an integrative network-driven bioinformatics pipeline, dRNASb (dual RNA-seq Systems biology-based analysis), to analyse temporal transcriptional clusters, incorporate molecular interaction networks (e.g. protein–protein interactions), identify topologically and functionally key transcripts in host and pathogen, and associate host and pathogen temporal transcriptome to decipher potential between-species interactions. The pipeline is applicable to various dual RNA-seq data from different species and experimental conditions. As a case study, we applied dRNASb to analyse dual RNA-seq data of Salmonella-infected human cells [21] at six time-points post-infection (pi), which enabled us to uncover genes contributing to Salmonella infection and to identify their potential functions.

METHODS

Data pre-processing and differential gene expression analysis

The pipeline accepts a dual RNA-seq tab-delimited input file where read counts from both the host and the bacterium are collapsed into a single matrix of genes (rows) and time-points (columns). Rows with more than 50% missing values were filtered out. Raw counts were normalised using the trimmed mean of M-values (TMM), which corrects for differences in RNA composition and sample outliers while providing better comparability across samples. Normalised data were then log-transformed. Before differential expression analysis, reads were separated to the pathogen and host genome using the respective reference genomes (provided as input). Then, the limma package in R was used to identify differentially expressed (DE) genes, which uses linear modelling to describe expression data [22]. The output is a list of DE genes compared with the baseline (i.e. 0 h). Genes with a false discovery rate (FDR) cut-off ≤0.05 (i.e. 5% false positives) and fold-change of at least two-fold in either direction (upregulation/downregulation), i.e. log2(fold-change)>1, were considered as statistically significant DE genes. These lists can then be used as an input for the analysis of temporal gene expression profiles.
Temporal analyses of gene expression profiles
To unravel transcriptional dynamics, fuzzy clustering was conducted on DE genes of at least one time-point. We used the Mfuzz package in R [23], which is developed for noise-robust soft clustering of gene expression time-series data. The number of clusters should be first identified for which we performed repeated soft clustering for a range of cluster numbers using 'Dmixn' function and reported the minimum centroid distance (MCD), i.e. the minimum distance between two cluster centres. The optimal cluster number was then determined by plotting MCD against cluster numbers and selecting a number where there is a drop of MCD and a slower decrease afterwards (a.k.a. the elbow rule). After soft clustering, each gene is assigned a membership score representing the degree of its association to each cluster that reflects a particular temporal pattern. The objective of Mfuzz is to compute membership values to minimise the within-cluster distances and maximise the between-cluster distances. The output of Mfuzz is a membership list for each gene class. The membership value ranges from 0 to 1, with higher values indicating that a gene is more likely to belong to a particular class.

Pathway enrichment analyses for host and pathogen
To identify potential functions associated with DE genes in each temporal cluster, the Cluster Evaluation R (clueR) package in R [24] was used to assess the enrichment of different functional terms by DE genes in each cluster, using Fisher's exact test with hypergeometric null distribution to calculate the probability of pathways associated with a set of DE genes within a cluster; adjusted P<0.05 was considered to indicate statistically significant associations. Enrichment analysis relies on a reference dataset that annotates genes based on associated pathways or other functional terms. A database of pathways, regulons, and genomic islands was constructed using information obtained from the BioCyc.org Database for the bacterium (i.e. Salmonella enterica SL1344). BioCyc.org is unique in integrating a diverse range of data and providing a high level of curation for important microbes as well as model eukaryotic organisms and Homo sapiens [25]. The Kyoto Encyclopaedia of Genes and Genomes (KEGG) and Gene Ontology Consortium (Amigo: http://amigo.geneontology.org) were used to retrieve functional annotations for the host (i.e. human).

Gene ontology enrichment analysis and visualisation for host and pathogen
Gene Ontology (GO) is an annotation database structured as a hierarchical directed acyclic graph where each node represents a class of gene function (GO term), and the connection between two GO terms indicates different relationships, such as ‘is a’ or ‘part of’. GO terms are categorised into biological processes (BPs), molecular functions (MFs) and cellular components (CCs) [26]. Statistical enrichment of GO terms was conducted using the Enrichr tool which comprise an up-to-date libraries of diverse gene sets including gene ontology BP, MF and CC gene sets [27]. Similarly, statistical overrepresentation analyses were performed using Fisher's exact test with the hypergeometric null distribution. Visualisation of enrichment was accomplished via bar plots and bubble plots using R scripts.

Protein-protein interaction network construction, analyses, and visualisation
A protein-protein interaction (PPI) databases for the pathogen (e.g. SalmoNet, salmonet.org [28] for the pathogen of interest in this study) and host (e.g. human STRING, string-db.org [29]) were retrieved, and the corresponding network was constructed using the igraph package [30] for network topology analyses. The STRING database can be used to retrieve PPIs for several other host and/or pathogen species of interest. Its current version (v11) covers functional and physical interactions among 24.6 million proteins from 5090 organisms [31, 32]. Multiple network measures, including node degrees, betweenness centrality [33], closeness centrality and modularity (via community detection) were identified. The functions of the hub genes (i.e. those genes with a high degree of connectivity) were identified as described before. Community structure detection was constructed using the cluster edge betweenness function in the igraph package to cluster similar groups of hub genes (i.e. those which have the highest similarity degree) [34].

Inter and intraspecies correlation analysis of pathogen and host transcriptome
Pearson correlation coefficients were calculated between and across host and pathogen DE genes, and P values were calculated using the function cor.test in R. To account for a possible temporal delay between Salmonella expression changes and effect manifestation in the host cell, a time-shift was allowed. This means the expressions of pathogen genes at each time-point were compared to the host expression at the subsequent time point. Once the correlation network was built (correlation coefficient <-0.7 or >0.7, and adjusted P<0.05), densely connected genes were identified via Louvain clustering, which can quickly find clusters with high modularity in large networks [35]. Host and pathogen genes within each cluster were further investigated to identify genes potentially important during the early stage of infection.

Pipeline access and usage notes
The pipeline code is available at GitHub repository https://github.com/VafaeeLab/dRNASb. The pipeline is reusable and applicable to different dual RNA-seq datasets. Users need to provide related input files, including a phenotype file (specifying samples and
groups/conditions), expression data (read counts of transcripts in host and pathogen), functional annotations, and PPIs of the species of interest. The repository also provides example input files as well as self-contained documentation describing the pipeline, scripts, and folders. The package is available as a library that can be installed on the end-user machines as per the instructions available.

RESULTS AND DISCUSSION

The dRNASb bioinformatics pipeline comprises the following steps: filtering and normalisation of count data, identification of DE genes in the host and pathogen, temporal gene expression clustering, functional enrichment analyses, network analysis, visualisation, and inter/intraspecies correlation analyses (Fig. 1). The pipeline, as mentioned above, is coded in R (version-4.0). The model system used in this work to study the bioinformatic pipeline comprised human cervix carcinoma cells (HeLa-S3; ATCC CCL-2.2) (host) infected with Salmonella enterica serovar Typhimurium strain SL1344 (bacterium), which is a well-defined host–bacteria system [36]. The pipeline, however, can be repurposed to analyse temporal dual RNA-seq data from other host-bacteria systems of interest.

The dual RNA-seq read count matrices were downloaded from Gene Expression Omnibus (accession number GSE60144) [21], quantifying the expression level of host and pathogen transcripts across six time-points (0, 2, 4, 8, 16, and 24 h). Details of upstream analyses including read processing and mapping are available in the original paper [25]. Gene expression levels in the host and pathogen at different times compared to the start of the experiment (0h) were evaluated, and DE genes (adjusted-P-value <0.05 and |log2(fold-change)|>1) were identified. In total, 6248 and 831 DE genes were found in the host and pathogen, respectively, that were differentially expressed at least in one time-point Table S1. The overlap of upregulated and downregulated DE genes in the host and pathogen is displayed in Figs 2 and 3, respectively. As can be seen, the number of downregulated genes in the host (Fig. 2a) and the pathogen (Fig. 3a) was significantly increased at 2h and 24h pi (compared to pre-infection), respectively. A higher number of upregulated genes in both organisms was observed at 24h pi. Venn diagrams in Figs 2 and 3 also present the number of genes dysregulated specific to each time point.

In general, genes with similar expression patterns which form temporally co-expressed clusters have the potential to exhibit similar cellular functions as guided by the ‘guilt-by-association’ assumption [37]. Accordingly, fuzzy clustering algorithm, which produces overlapping clusters was applied to the 6248 host and 831 pathogen genes with expression time series of length 6. The optimal number of clusters was ten and determined as previously explained (see Methods). The membership degree was set to ≥ 0.5 to select genes that tightly follow the average expression pattern of each cluster and remove genes which do not strongly belong to any cluster.

Accordingly, 4893 out of 6248 genes grouped into ten clusters (Fig. 4a) and the most populated expression pattern observed as upregulation and downregulation activities peak at 24h (pi) in clusters six and ten, respectively (Table 1). On the other hand, the result of pathogen clustering showed 690 out of 831 pathogen genes were grouped into ten clusters (Fig. 4b); the most populated expression pattern observed as an up-regulation and down-regulation activities peak at 24h (pi) in clusters three and ten, respectively (Table 2).

The functional annotation and enrichment analysis on clusters of DE genes using GO terms and KEGG allowed us to obtain functions enriched by temporal clusters of DE genes in host (Table S3) and pathogen (Table S2) during the infection process.

Overall, 20 genes from four pathogen clusters were enriched with two main classes of cellular functions important for pathogenesis (Table S2). The first class was related to the transmembrane transporter activity, electron transport chain, and the biosynthesis of a micronutrient such as biotin that acts as a co-factor for bacterial metabolic activities. The role of biotin biosynthesis and transport functions in bacterial pathogens have been well studied [38–40]. Therefore, understanding the transcriptional mechanisms regulating essential elements of biosynthesis and transport will extend our knowledge on bacterial survival and metabolic adaptation during pathogenesis.

The peptidoglycan biosynthetic pathway is a critical process in the bacterial cell [41] and the interplay between peptidoglycan biology and pathogenesis has been previously reported [42]. The enrichment analysis has shown that the second class of functions overrepresented by pathogen DE clusters (i.e. cluster 8, genes tig (E1W8V7) and ftsK (E1W6Z2)) included the regulation of cell shape and peptidoglycan synthesis that enable bacteria to maintain, modify, and reshape the peptidoglycan layer without risking its essential functions in cell shape and cellular integrity. DE analysis shows that these genes are downregulated during the first 8h and upregulated afterwards. Peptidoglycan is generally considered one of the main determinants of the cell shape maintenance in bacteria [43]. The outcome of this study supports this function: during earlier hours of infection, expressions of these genes were downregulated, which may indicate that the bacteria’s shape is changing. After the bacterial invasion of host cells, the expression of these genes was sharply upregulated, which can be the result of duplication of bacteria and the initiation of infection.

During the initial steps of infection, pathogens manipulate and modify host cell biology to ensure conditions best suited for colonisation [44]. This process can stimulate the response of the innate immune system [45]. GO enrichment analysis has shown...
that 380 genes across host clusters were enriched with functions that play essential roles in mediating cellular entry of pathogens (e.g. epidermal growth factor receptor (EGFR) signalling pathway [46], growth hormone receptor signalling pathway [47], mitochondrial ATP synthesis coupled electron transport [48], protein targeting to the endoplasmic reticulum [49]), communication of bacteria with the host environment (e.g. positive regulation of establishment of protein localisation [50]), control of infection (e.g. regulation of protein metabolic process [51]), invading pathogen (e.g. response to laminar fluid shear stress [52]), and co-opting host factors (e.g. viral gene expression [53]) that can be considered as potential targets in designing antibacterial therapies. Enriched functions were mostly observed at 24h and 4h pi in the host and pathogen, respectively, with mainly downregulated expression.
Integration of biological networks, such as protein-protein interactions, has shown to be a useful approach to correlating gene expression changes with changing conditions [54, 55] and can provide information about the most significant connections among classified DE genes in the host and pathogen during early stages of infection. We, therefore, overlaid DE genes across different clusters on the protein-protein interaction (PPI) network of the host and pathogen to show how genes and their products of different clusters can possibly cross talk. We also constructed PPI networks among all DE genes (not exclusive to those genes that appeared in temporal clusters) in the host and pathogen that, in this study, are referred to as the host gene network (HGN) and the pathogen gene network (PGN), respectively.

In order to detect key genes and clusters of genes with functionally enriched pathways during early stages of infection, a gene co-expression network analysis using igraph package was performed. Co-expression modules identified by clustering are often large, and therefore, it is important to identify key gene(s) in each module. A widely used approach is to identify highly connected genes in a co-expression network (i.e. hub genes). A hub gene is a gene associated with many other genes in gene networks.

**Fig. 2.** Visualisations of overlaps among host genes upregulated (3686 genes) or downregulated (2628 genes) across five different time points post-infection compared to baseline (0 h) (adjusted $P<0.05$, $|\log_2$(fold-change)|>1) using Venn diagrams (a and c) and UpSet plots (b and d). While Venn diagrams are visually more familiar, UpSet plots [86] are better suited for the quantitative analysis of data with more than three sets by visualising set intersections in a matrix layout.
Additionally, centrality measures, mainly ‘betweenness centrality’, are often used to identify topologically important genes in a biological network [56]. High betweenness centrality of a gene indicates that the node/gene serves as a connector in the network and lies in many shortest paths that connect together different parts of the network [57]. Genes with high degree and betweenness centrality may play a key role in biological processes and gene regulation [58] and tend to be highly relevant to the functionality of biological networks. Many biological networks are organised into connected substructures or modules which carry particular functions within the system [59]. Hub genes can be identified within a module (rather than the entire network), that is referred to as intra-modular hubs and can be central to a specific function carried by the module. Correlation analysis has shown that hub genes, i.e. nodes with degree >10 for pathogen and >100 for host, separately identified in the host and pathogen co-expression networks (129 and 293 hubs in host networks, respectively), are significantly correlated. We also applied network community detection (Louvain clustering algorithm) on hub genes and identified four inter-connected modules in negatively correlated (corr <−0.7) (Fig. 5a and Tables S4–S7) network and three modules in positively correlated (corr >0.7) (Fig. 5b and Table S8-S10) network. These modules can potentially be involved in the same or similar biological processes.

Fig. 3. Visualisations of overlaps among pathogen genes upregulated (426 genes) or downregulated (458 genes) across five different time points post-infection compared to baseline (0 h) (adjusted \(P<0.05, \log_{2}(\text{fold-change})>1\) ) using Venn diagrams (a and c) and UpSet plots (b and d).
To identify potential cellular functions and diseases associated with genes contained in each module, we performed KEGG pathway, KEGG disease, and GO enrichment analyses on the set of genes contained in each of the negatively and positively correlated modules. Analyses have shown that the host and pathogen gene sets in the negatively and positively correlated modules were enriched with similar as well as host-/pathogen-specific functions (Table 3). For instance, the most frequently enriched functions specific to the host gene set in the first negatively-correlated module were involved in processes underlying the protection against infectious diseases, and the innate immune system which employs germline-encoded pattern-recognition receptors.
(PRRs) and Toll-like receptors [60–62]. These distinct processes are activated upon detection of the initial natural infection mediated by a pathogen or immunisation.

The mitochondria metabolic process was the main function enriched by host genes in the second negatively correlated module. This function has a central role in regulating cellular activities [63] and directly controls host stimulated responses against infections [63]. Yet, multiple pathogens have the ability to manipulate mitochondrial dynamics and functions as well as cell metabolism and immune responses [64, 65].

The Wnt signalling pathway was a dominant pathway enriched by the host gene set in the third negatively correlated module. This pathway contributes to the cell cycle control, cytoskeleton reorganisation during phagocytosis and cell migration, autophagy, apoptosis, and a number of other inflammation-related events [66]. It has also emerged as an integral component of host responses to infection, but its function in the context of immune responses is not fully understood [66–68]. Recent studies have shown that this pathway can control host processes related to bacterial infection; however, pathogens have evolved strategies to manipulate Wnt-associated processes in order to enhance infection and survival within the human host [66–68]. The mitogen-activated protein kinase (MAPK) signalling pathway was the second most enriched pathway by this set of genes. This pathway plays a central role in host-pathogen complex interactions and is pivotal for triggering the host immune response against pathogens [69, 70].

In the fourth negatively correlated module, lipid droplets (LDs) were enriched by the host gene set, which are the choke point in the conflict between host and pathogen for nutrients [71, 72]. Previous studies on the nutritional status of the infected host have discussed the malfunctioning or lack of LDs as a key player in the regulation of systemic innate immunity [72]. On the other hand, bacteria can uncouple LDs from mitochondria and increase host-pathogen contacts by reducing fatty acid metabolism [73]. Yet, the nature of the distinct junction between the front line of the body defence (LDs) and bacteria is not well known [72]. Furthermore, recent studies have revealed pathogen ability in the manipulation of the host membrane to facilitate pathogenic entry and recruitment of specific host lipids for the maintenance of favourable intracellular niche to augment their survival and proliferation [71, 72, 74].

In the first positively correlated module, the host genes enriched the Notch signalling pathway that is critically involved in developmental processes and regulating immune responses during infections [75]. Two main signalling pathways enriched by the host gene set in the second positive module included G protein-coupled receptors and epidermal growth factor receptors (EGFR). These are involved in the pathogen pathophysiological processes [76, 77] and in mediating cellular entry of numerous pathogen cells [46], respectively. Generally, the functions related to this module have a role in immune pathogenesis by activities of innate and adaptive immune systems.

Leucine-rich repeats (LRRs) were enriched by the host gene set in the third positively correlated module. The leucine-rich repeat domain presents in several innate immune receptors, where it is frequently responsible for sensing danger signals and specific pathogen-associated molecules. It will be active the innate immune system as a host defence system, where they sense specific pathogen-associated molecules [78]. Fc-gamma and cell surface receptors were the second signalling pathways which

| Table 1. Summary of DEGs distribution in the host clusters |
|----------------------------------------------------------|
| Timepoint (pi) | Total genes | Upregulated | Downregulated |
|----------------|-------------|-------------|---------------|
| 2h             | 1646        | 492         | 1154          |
| 4h             | 1413        | 449         | 964           |
| 8h             | 2164        | 702         | 1462          |
| 16h            | 3620        | 1151        | 2469          |
| 24h            | 4328        | 1412        | 2916          |
were enriched by the host gene set. These receptors are an essential component in many immune system functions such as the phagocytosis and have the ability to stimulate innate immune responses, releasing inflammatory mediators [44, 79].

Vascular endothelial growth factor (VEGF), also enriched in this module, is known to play crucial roles in endothelial cell proliferation, migration, angiogenesis, vascular permeability, inhibition of apoptosis, and pathogen infection. This pathway is also involved in the interaction between pathogens and host cells as pathogens are able to encode VEGF homologs which bind to the VGFR receptor (VEGFR-2) of the host cells and then contribute to the infection process [80].

Another interesting pathway was Pattern Recognition Receptors (PRRs) which are proteins found in the host as part of the innate immune system and can also impact adaptive immunity [81]. PRRs were enriched in both positively and negatively correlated modules, however, via different genes, as detailed in Table 3. Other functions related to this module have roles in the protein modification process, cell junction assembly and substrate adhesion-dependent cell spreading. Most functions in this module are involved in infection development and stress responses.

Interestingly, an important function frequently enriched by both host and pathogen gene sets across different modules was related to the regulation of metal ions transmembrane transporters. Metal ions are essential for living organisms and play a vital role in the regulation of both microbial virulence and host immune responses. Recent studies have shown a role for metal ions in protection of the host against infections [82–85]. Therefore, it could be beneficial for the invading pathogen to express a variety of such genes in order to steal metal ions from the host.
Overall, we identified co-expression modules which are potentially involved in specific biological processes during early stages of infection. We then further investigated the inter-module relationships to identify genes that are key in interconnected different modules. We accordingly first selected host-pathogen correlated genes to further stringent the subsequent analyses to modules whose constituent host/pathogen genes have a very similar expression pattern during infection. This resulted in 54 genes with positive or negative correlations (Fig. 6) with corr $<-0.9$ or corr $>0.9$ forming six negative (Fig. 7a) and
Fig. 6. Inter- and Intra-species correlations among pathogen and host genes (Pearson correlation coefficient $<-0.9$ or $>0.9$) obtained by the ‘corrplot’ function in R resulted in 54 genes (comprising 15 host and 39 pathogen genes) which were used in the subsequent analyses to obtain co-expressed clusters.
positive modules (Fig. 7b). We computed network-based metrics on genes within these modules and selected genes connecting different modules with high betweenness (≥40) and degree (≥10) as key inert-module connectors. The analyses uncovered the role of E1W9K0 (kdpA), CAV1, E1WA57 (mrdE) and SKP1 in interactions between four negative correlated modules (Tables S11 and S12). Additionally, E1WAC0 (hilD), ATP6V1A, ILK, E1W6Z0 (trxB), E1W8W1 (hupB), and E1W9A9 (ahpC) were identified as key genes to interconnect positive correlated modules (Tables S13 and S14). Further, DE analyses have shown downregulation activities of EEF1A1, E1W8W1 (hupB), and CALM2 switching off the expression of E1W9H8 (nagB) in module one. Furthermore, the downregulation of ILK switches off the expression of E1W9H8 (nagB) in module three. Enrichment analyses have shown pathogen genes involved with oxidation-reduction process and potassium ion transport. Host genes mainly related to regulation of cellular response to inflammation, cellular iron ion homeostasis, and regulation of epithelial cell differentiation. Together, these observations support interactions of these genes in the early stage of infection.
CONCLUSION

Here, we present the dRNASb pipeline for the network-level analyses of dual RNA-seq data. The pipeline was showcased on a Salmonella enterica serovar Typhimurium-infected HeLa cell model, with a detailed description of analyses performed, and results achieved. The proposed network-based analyses provide complementary information on infection processes not obtainable otherwise. In a nutshell, combining statistical analyses with network-level and functional analyses, as presented in this study, can elucidate the interplay between host and pathogen during bacterial infection and provide valuable information on which genes in the host or pathogen can be potential molecular targets for therapeutic interventions. Development of such pipelines can enhance system-level understanding of host-pathogen transcriptomic dynamics and can aid in the identification of key genes or functional interpretation of poorly annotated genes. In the future, dRNASb can be further enhanced by inclusion of machine learning models to further integrate complementary information for accurate prediction of host-pathogen interactions during the infection process.

Funding Information
NCR is supported by a Cancer Institute NSW Early Career Fellowship (2019/ECF1082) and a UNSW Scientia Fellowship. Otherwise, not externally funded.

Acknowledgements
We thank Dr Susan Corley, Dr Sara Loo, and Dr Xabier Vázquez-Campos for useful discussions and support.

Author contributions
Conceptualisation: F.V., M.D. Methodology: M.D., F.V. Software: A.V., M.D. Validation: M.D., N.C.R., A.P. Formal Analysis: M.D., F.V., F.C.K., D.A.M., A.F.T., K.V., A.A. Investigation: F.V., M.D. Resources: F.V. Data Curation: M.D. Writing – Original Draft Preparation: M.D., F.V. Writing – Review and Editing: all authors; Conceptualisation: F.V., M.D. Methodology: M.D., F.V. Software: A.V., M.D. Validation: M.D., N.C.R., A.P. Formal Analysis: M.D., F.V., F.C.K., D.A.M., A.F.T., K.V., A.A. Investigation: F.V., M.D. Resources: F.V. Data Curation: M.D. Writing – Original Draft Preparation: M.D., F.V. Writing – Review and Editing: all authors; Visualisation: M.D.; Supervision: F.V. Funding: F.V., N.C.R.

Conflicts of interest
Authors declare no conflicts of interest.

References
1. Crua Asensio N, Muñoz Giner E, de Groot NS, Torrent Burgas M. Centrality in the host-pathogen interactome is associated with pathogen fitness during infection. Nat Commun 2017;8:14092.
2. Macho Rendón J, Lang B, Ramos Llorens M, Gaetano Tartaglia G, Torrent Burgas M. DualSeqDB: the host–pathogen dual RNA sequencing database for infection processes. Nucleic Acids Res 2020;49:D687–D693.
3. Marsh JW, Hayward RJ, Shetty AC, Mahurkar A, Humphrys MS, et al. Bioinformatic analysis of bacteria and host cell dual RNA-sequencing experiments. Brief Bioinform 2018;19:1115–1129.
4. Mika-Gospodorz B, Giengkam S, Westermann AJ, Wongantichon J, Kion-Crosby W et al. Dual RNA-seq of Orientia tsutsugamushi informs on host-pathogen interactions for this neglected intracellular human pathogen. Nat Commun 2020;11:3363.
5. Mukherjee P, Burgio G, Heitlinger E. Dual RNA sequencing meta-analysis in Plasmodium infection identifies host-parasite interactions. mSystems 2021;6:e00182-21.
6. Marsh JW, Humphrys MS, Myers GSA. A laboratory methodology for dual RNA-sequencing of bacteria and their host cells In Vitro. Front Microbiol 2017:8:1830.
7. Pisu D, Huang L, Grenier JK, Russell DG. Dual RNA-Seq of Mtb-infected macrophages in vivo reveals Ontologically distinct host-pathogen interactions. Cell Rep 2020;30:335–350.
8. Westermann AJ, Barquist L, Vogel J. Resolving host-pathogen interactions by dual RNA-seq. PLoS Pathog 2017;13:e1006033.
9. Keiler KC. Bacterial Regulatory RNA. In: Methods and Protocols. Totowa, NJ: Humana Press, 2012.
10. Russell DG, Huang L, VanderVen BC. Immunometabolism at the interface between macrophages and pathogens. Nat Rev Immunol 2019;19:291–304.
11. Nuss AM, Beckstette M, Pimenova M, Schmüll C, Opitz W, et al. Tissue dual RNA-seq allows fast discovery of infection-specific functions and riboregulators shaping host-pathogen transcriptomes. Proc Natl Acad Sci USA 2017;114:E791–E800.
12. Damron FH, Oglesby-Sherrouse AG, Wilks A, Barbier M. Dual-seq transcriptomics reveals the battle for iron during Pseudomonas aeruginosa acute murine Pneumonia. Sci Rep 2016;6:39172.
13. Kumar SS, Tandberg JI, Penesyan A, Elbourne LDH, Suarez-Bosche N, et al. Dual transcriptomics of host-pathogen interaction of Cystic Fibrosis isolate Pseudomonas aeruginosa PASS1 With Zebrafish. Front Cell Infect Microbiol 2018;8:406.
14. Montoya DJ, Andrade P, Silva BJA, Teles RMB, Ma F, et al. Dual RNA-Seq of human leprosy lesions identifies bacterial determinants linked to host immune response. Cell Rep 2019;26:3574–3585.
15. Huang L, Nazarova EV, Tan S, Liu Y, Russell DG. Growth of Mycobacterium tuberculosis in vivo segregates with host macrophage metabolism and ontogeny. J Exp Med 2018;215:1135–1152.
16. Valenzuela-Miranda D, Gallardo-Escárate C. Dual RNA-Seq uncovers metabolic amino acids dependency of the intracellular bacterium Piscirickettsia salmonis infecting Atlantic salmon. Front Microbiol 2018;9:2877.
17. Chung M, Bruno VM, Rasko DA, Cuomo CA, Muñoz JF, et al. Best practices on the differential expression analysis of multi-species RNA-seq. Genome Biol 2021;22:121.
18. O’Keeffe KR, Jones CD, Kembel S. Challenges and solutions for analysing dual RNA -seq data for non-model host–pathogen systems. Methods Ecol Evol 2019:10:401–414.
19. Greene CS, Troyanskaya OG. Integrative systems biology for data-driven knowledge discovery. Semin Nephrol 2010;30:443–454.
20. Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. J R Soc Interface 2015;12:112.
21. Westermann AJ, Förstner KU, Amman F, Barquist L, Chao Y, et al. Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. Nature 2016;529:496–501.
22. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
23. Kumar L, E Futschik M. Muuzz: a software package for soft clustering of microarray data. Bioinformation 2007;2:5–7.
24. Wirbel J, Cutillas P, Saez-Rodriguez J. Phosphoproteomics–based profiling of kinase activities in cancer cells. Methods Mol Biol 2018;1711:103–132.
25. Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, et al. BioCyc: a genomic and metabolic web portal with multiple omics analytical tools. *FASEB J* 2019;33:51.

26. Chen Z, Lin Y, Gao J, Lin S, Zheng Y, et al. Identification of key candidate genes for colorectal cancer by bioinformatics analysis. *Oncof Lett* 2019;18:6583–6593.

27. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;44:W90–7.

28. Métris A, Sudhakar P, Fazekas D, Demeter A, Ari E, et al. SalmoNet, an integrated network of ten *Salmonella enterica* strains reveals common and distinct pathways to host adaptation. *NPJ Syst Biol Appl* 2017;3:31.

29. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015;43:D447–52.

30. Csardi G, Nepusz T. *IJ* complex systems. The igraph software package for complex network research. *Inter J Complex Syst* 2006;1695:1–9.

31. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;47:D607–D613.

32. Bajpai AK, Davuluri S, Tiwary K, Narayanam S, Oguru S, et al. Systematic comparison of the protein-protein interaction databases from a user’s perspective. *J Biomed Inform* 2020;103:103380.

33. Oldham S, Fulcher B, Parkes L, Arnatkevíc ļūtė A, Suo C, et al. Consistency and differences between centrality measures across different classes of networks. *PLoS One* 2019;14(4):e0220861.

34. Li Vigni M, Durante C, Cocchi M. Chapter 3 - exploratory data analysis. In: *Marini F* (eds). *Data Handling in Science and Technology*: Elsevier. 2013. pp 55–126.

35. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech* 2013:2008:P10008.

36. Walch P, Selkirk J, Knodler LA, Rettel M, Stein F, et al. Global mapping of *Salmonella enterica*-host protein-protein interactions during infection. *Cell Host Microbe* 2021;29:1316–1332.

37. Oliver S. Guilt-by-association goes global. *Nature* 2000;403:601–603.

38. Satiaputra J, Shearwin KE, Booker GW, Polyak SW. Mechanisms of biofilm-regulated gene expression in microbes. *Synth Syst Biototechnol* 2016;1:17–24.

39. Kurayama R, Ito N, Nishibori Y, Fukuhara D, Akimoto Y, et al. Role of amino acid transporter LAT2 in the activation of mTORC1 pathway and the pathogenesis of crescentic glomerulonephritis. *Lab Invest* 2011;91:992–1006.

40. Kaila VRI,Wikström M. Architecture of bacterial respiratory chains. *Nat Rev Microbiol* 2021;19:319–330.

41. Lovering AL, Safadi SS, Strynadka NCJ. Structural perspective of peptidoglycan biosynthesis and assembly. *Annu Rev Biochem* 2012;81:451–478.

42. Graham CLB, Newman H, Gillett FN, Smart K, Briggs N, et al. A dynamic network of proteins facilitate cell envelope biogenesis in gram-negative Bacteria. *Int J Mol Sci* 2021;22:23.

43. Rohs PDA, Bernhardt TG. Growth and division of the peptidoglycan matrix. *Annu Rev Microbiol* 2021;75:315–336.

44. Alberts B JA, Lewis J. *Molecular Biology of the Cell*. 4th edition. Cell Biology of Infection, 2002.

45. Janeway CA JT, Walport M. *Immunobiology: The Immune System in Health and Disease*. 5th edition. Part V, The Immune System in Health and Disease, 2001.

46. Gru W, Chang L, He Y, Zhu L. The role of epidermal growth factor receptor signaling pathway during bovine herpesvirus 1 productive infection in cell culture. *Viruses* 2020;12:E927.

47. Strous GJ, Almeida ADS, Putters J, Schantl J, Sedeck M, et al. Growth hormone receptor regulation in cancer and chronic diseases. *Front Endocrinol* 2011;11:397977.

48. Escoll P, Platon L, Buchrieser C. Roles of mitochondrial respiratory complexes during infection. *Immunometabolism* 2019;1:e190011.

49. Celi J, Tsolis RM. Bacteria, the endoplasmic reticulum and the unfolded protein response: friends or foes? *Nat Rev Microbiol* 2015;13:71–82.

50. Wettstadt S, Llamas MA. Role of regulated proteolysis in the communication of bacteria with the environment. *Front Mol Biosci* 2020;7:586497.

51. Tomé D. Amino acid metabolism and signalling pathways: potential targets in the control of infection and immunity. *Eur J Clin Nutr* 2021;75:1319–1327.

52. Lupu F, Kinasewitz G, Dormer K. The role of endothelial shear stress on haemodynamics, inflammation, coagulation and glyco-calyx during sepsis. *J Cell Mol Med* 2020;24:12258–12271.

53. Ho JSY, Zhu Z, Marazzi I. Unconventional viral gene expression mechanisms as therapeutic targets. *Nature* 2021;593:362–371.

54. Silverbush D, Sharan R. A systematic approach to orient the human protein-protein interaction network. *Nat Commun* 2019;10:3015.

55. Métris A, Sudhakar P, Fazekas D, Demeter A, Ari E, et al. SalmoNet, an integrated network of ten *Salmonella enterica* strains reveals common and distinct pathways to host adaptation. *NPJ Syst Biol Appl* 2017;3:31.

56. Ashtiani M, Salehzadeh-Yazdi A, Razaghi-Moghadam Z, Hennig H, Wolkenhauer O, et al. A systematic survey of centrality measures for protein-protein interaction networks. *BMC Syst Biol* 2018;12:80.

57. Wang CCN, Li CY, Cai J-H, Sheu PC-Y, Tsai JJP, et al. Identification of prognostic candidate genes in breast cancer by integrated bioinformatic analysis. *J Clin Med* 2019;8:E1160.

58. Yao Q, Song Z, Wang B, Qin G, Zhang J-A. Identifying key genes and functionally enriched pathways in *Sjögren’s syndrome* by weighted gene co-expression network analysis. *Front Genet* 2019;10:1142.

59. Barabási A-L, Oltvai ZN. Network biology: understanding the cell’s functional organization. *Nat Rev Genet* 2004;5:101–113.

60. Tangye SG, Pelham SJ, Deenick EK, Ma CS. Cytokine-mediated regulation of human lymphocyte development and function: insights from primary immunodeficiencies. *J Immunol* 2017;199:1949–1958.

61. Kawasaki T, Kawai T. Toll-like receptor signaling pathways. *Front Immunol* 2014;5:461.

62. Li D, Wu M. Pattern recognition receptors in health and diseases. *Signal Transduct Target Ther* 2021;6:291.

63. Tiku V, Tan M-W, Dicik I. Mitochondrial functions in infection and immunity. *Trends Cell Biol* 2020;30:263–275.

64. Elesela S, Lukacs NW. Role of mitochondria in viral infections. *Life* 2021:11:232.

65. Andrieux P, Chevillard C, Cunha-Neto E, Nunes JPS Mitochondria as a cellular hub in Inflammation and Infection. *Int J Mol Sci* 2021;22:21.

66. Rogan MR, Patterson LL, Wang JY, McBride JW. Bacterial manipulation of Wnt signaling: a host-pathogen tug-of-Wnt. *Front Immunol* 2019;10:2390.

67. Ljungberg JK, Kling JC, Tran TT, Blumenthal A. Functions of the WNT signaling network in shaping host responses to infection. *Front Immunol* 2019;10:2521.

68. Karmam A, Bonam SR, Rambabu N, Wong SSW, Aimananda V, et al. Wnt-β-catenin signaling in human dendritic cells mediates regulatory T-cell responses to fungi via the PD-L1 pathway. *mBio* 2021;12:e0282421.

69. Guo Z, Kang S, Wu Q, Wang S, Crickmore N, et al. The regulation landscape of MAPK signaling cascade for thwarting *Bacillus thuringiensis* infection in an insect host. *PLoS Patog* 2021;17:e1009917.

70. He X, Wang C, Wang H, Li L, Wang C. The function of MAPK cascades in response to various stresses in horticultural plants. *Front Plant Sci* 2020;11:952.
1. Dadhich R, Kapoor S. Various facets of pathogenic lipids in infectious diseases: exploring virulent lipid-host interactome and their druggability. J Membr Biol 2020;253:399–423.

2. Bosch M, Sweet MJ, Parton RG, Pol A. Lipid droplets and the host-pathogen dynamic: FATAl attraction? J Cell Biol 2021;220:e202104005.

3. Bosch M, Sánchez-Álvarez M, Fajardo A, Kapetanovic R, Steiner B, et al. Mammalian lipid droplets are innate immune hubs integrating cell metabolism and host defense. Science 2020;370:eaay8085.

4. González-Fernández C, Basauri A, Fallanza M, Bringas E, Oostenbrink C, et al. Fighting against bacterial Lipopolysaccharide-caused infections through molecular dynamics simulations: a review. J Chem Inf Model 2021;61:4839–4851.

5. Castro RC, Gonçales RA, Zambuzi FA, Frantz FG. Notch signaling pathway in infectious diseases: role in the regulation of immune response. Inflamm Res 2021;70:261–274.

6. Venkatesh SR, Singh V. G protein-coupled receptors: he choreographers of innate immunity in Caenorhabditis elegans. PLOS Pathog 2021;17:e1009151.

7. Zeng Z, Mukherjee A, Varghese AP, Yang X-L, Chen S, et al. Roles of G protein-coupled receptors in inflammatory bowel disease. World J Gastroenterol 2020;26:1242–1261.

8. Kedzierski Ł, Montgomery J, Curtis J, Handman E. Leucine-rich repeats in host-pathogen interactions. Arch Immunol Ther Exp 2004;52:104–112.

9. Bourantas S, Gupta A, Ravetch JV. The role of IgG Fc receptors in antibody-dependent enhancement. Nat Rev Immunol 2020;20:633–643.

10. Li S, Wang Z, Li F, Yu K, Xiang J. A novel vascular endothelial growth factor receptor participates in white spot syndrome virus infection in Litopenaeus vannamei. Front Immunol 2017;8:1457.

11. Amarante-Mendes GP, Adjemian S, Branco LM, Zanetti LC, Weinlich R, et al. Pattern recognition receptors and the host cell death molecular machinery. Front Immunol 2018;9:2379.

12. Liu X, Zhang K, Liu Y, Zou D, Wang D, et al. Effects of calcium and signal sensing systems on Azorhizobium caulinodans biofilm formation and host colonization. Front Microbiol 2020;11:563367.

13. King MM, Kayastha BB, Franklin MJ, Patrauchan MA. Calcium regulation of bacterial virulence. Adv Exp Med Biol 2020;1131:827–855.

14. Kehl-Fie TE, Skaar EP. Nutritional immunity beyond iron: a role for manganese and zinc. Curr Opin Chem Biol 2010;14:218–224.

15. Xia P, Lian S, Wu Y, Yan L, Quan G, et al. Zinc is an important inter-kingdom signal between the host and microbe. Vet Res 2021;52:39.

16. Lex A, Gehlenborg N, Strobelt H, Vuilleumot R, Pfister H. UpSet: visualization of intersecting sets. IEEE Trans Vis Comput Graph 2014;20:1983–1992.

---

**Five reasons to publish your next article with a Microbiology Society journal**

1. When you submit to our journals, you are supporting Society activities for your community.
2. Experience a fair, transparent process and critical, constructive review.
3. If you are at a Publish and Read institution, you’ll enjoy the benefits of Open Access across our journal portfolio.
4. Author feedback says our Editors are ‘thorough and fair’ and ‘patient and caring’.
5. Increase your reach and impact and share your research more widely.

Find out more and submit your article at microbiologyresearch.org.