Published online 21 October 2020 Nucleic Acids Research, 2021, Vol. 49, Database issue **D687–D693**
doi: 10.1093/nar/gkaa890

DualSeqDB: the host–pathogen dual RNA sequencing database for infection processes

Javier Macho Rendón[1,†], Benjamin Lang[2,3,†], Marc Ramos Llorens[1], Gian Gaetano Tartaglia[2,4,5,*] and Marc Torrent Burgas[1,*]

[1]Systems Biology of Infection Lab, Department of Biochemistry and Molecular Biology, Biosciences Faculty, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Spain, [2]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain, [3]Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA, [4]Department of Neuroscience and Brain Technologies, Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genoa, Italy and [5]Department of Biology 'Charles Darwin', Sapienza University of Rome, Ple A. Moro 5, 00185 Rome, Italy

Received August 12, 2020; Revised September 27, 2020; Editorial Decision September 28, 2020; Accepted September 29, 2020

ABSTRACT

Despite antibiotic resistance being a matter of growing concern worldwide, the bacterial mechanisms of pathogenesis remain underexplored, restraining our ability to develop new antimicrobials. The rise of high-throughput sequencing technology has made available a massive amount of transcriptomic data that could help elucidate the mechanisms underlying bacterial infection. Here, we introduce the DualSeqDB database, a resource that helps the identification of gene transcriptional changes in both pathogenic bacteria and their natural hosts upon infection. DualSeqDB comprises nearly 300 000 entries from eight different studies, with information on bacterial and host differential gene expression under *in vivo* and *in vitro* conditions. Expression data values were calculated entirely from raw data and analyzed through a standardized pipeline to ensure consistency between different studies. It includes information on seven different strains of pathogenic bacteria and a variety of cell types and tissues in *Homo sapiens*, *Mus musculus* and *Macaca fascicularis* at different time points. We envisage that DualSeqDB can help the research community in the systematic characterization of genes involved in host infection and help the development and tailoring of new molecules against infectious diseases. DualSeqDB is freely available at http://www.tartaglialab.com/dualseq.

INTRODUCTION

During infection, pathogens trigger the expression of unique genes that ensure their survival and allow replicating within the host. In turn, the host activates complex mechanisms to recognize and kill pathogens. Hence, the simultaneous detection of host and pathogen transcripts during the infection process can provide deeper insights into the host–pathogen interaction than those detected from the host or pathogen in isolation. The term 'dual RNA-seq' refers to the process of simultaneously analyzing RNA-seq data of a pathogenic bacteria and the infected host (1). Dual RNA-seq has become a leading approach to uncover the intricate relationship between pathogen and host interactions allowing researchers to identify 'molecular phenotypes' that would otherwise remain undetected (2–4).

In a typical dual RNA-seq experiment, either animals are inoculated with a defined load of bacteria (*in vivo*) or relevant cell culture models are inoculated with bacteria at a defined multiplicity of infection (*in vitro*). After inoculation, samples are taken over time to determine the time response. At each time point, infected cells are lysed, RNA is isolated and the cDNA library is prepared and sequenced using high-throughput sequencing technologies, which generates large amounts of data. RNA-seq data of mock-infected host cells and initial bacterial cultures are used as control conditions for expression analysis. Dual RNA-seq experiments have several technical difficulties, including the different nature and content of RNA between bacteria and eukaryotic cells, the larger proportion of RNA from eukaryotic cells and the need to account for the prevalence of rRNA transcripts and variable infection rates (1,5). Usually, such limitations can be solved using high-depth sequencing, pathogen and host rRNA depletion, and enrichment

[*]To whom correspondence should be addressed. Tel: +34 93 581 4147; Email: marc.torrent@uab.cat
Correspondence may also be addressed to Gian Gaetano Tartaglia. Tel: +34 93 316 0116; Email: gian.tartaglia@crg.eu
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Table 1. List of dual RNA-seq studies included in DualSeqDB

Pathogen	Host organism	Tissue/cell type	Condition	GEO code	Reference
Streptococcus pyogenes	*Macaca fascicularis*	Skeletal muscle tissue	*In vivo*	GSE144100	(9)
Salmonella enterica serovar Typhimurium SL1344	*Homo sapiens*	HeLa-S3 cells	*In vitro*	GSE60144	(4)
Salmonella enterica serovar Typhimurium SL1344	*Homo sapiens*	Endothelial cells Epithelial cells Monocytic cells NK cells	*In vitro*	GSE136717	(10)
Yersinia pseudotuberculosis IP 32953	*Mus musculus*	Lymphoid tissue	*In vivo*	PRJEB14242 (ENA)	(3)
Pseudomonas aeruginosa PA01	*Mus musculus*	Lung tissue	*In vivo*	SRP090213 (SRA)	(11)
Haemophilus ducreyi 35000HP	*Homo sapiens*	Skin tissue	*In vivo*	GSE130901	(12)
Mycobacterium tuberculosis ATCC 35733	*Homo sapiens*	THP-1 cells	*In vitro*	PRJEB6552 (ENA)	(13)
Streptococcus pneumoniae D39	*Homo sapiens*	A549 cells	*In vitro*	GSE79595	(2)

of samples for infected host cells by fluorescence-activated cell sorting (6).

Dual RNA-seq is a mixture of host and pathogen transcripts where different RNA samples may contain variable proportions of pathogen to host reads (7,8). These transcripts can be sorted into the corresponding organisms by different computational strategies (1,2) and the accuracy of differential expression values depends on the analytical method used. To circumvent these biases, we need a standard pipeline to compare data from different sources.

Despite the increasing availability of raw sequencing data from dual RNA-seq experiments, the existence of multiple analysis pipelines may hinder the comparison between datasets. Dual RNA-seq pipelines are very sensitive to software selection and parameter definition. This lack of standardization motivated us to create DualSeqDB, a user-friendly platform to search for changes in gene expression levels during infection at both pathogen and host levels. To build this database, we analyzed raw sequencing data from heterogeneous dual RNA-seq studies using a well-defined pipeline, to generate comparable gene expression data. This setup allows DualSeqDB to compare across multiple species and experimental conditions.

MATERIALS AND METHODS

Processing sequencing data

To build DualSeqDB, we reprocessed raw data from available studies (Table 1) and used a well-defined pipeline to provide robust and homogeneous information in our database (Figure 1). To this end, we selected only dual RNA-seq studies containing at least two biological replicates and only when data were available for infected and control conditions for both pathogen and host (2–4,9–13. For each study, genome and annotation files were downloaded from the NCBI Reference Sequence Database (RefSeq) (14). Bacterial and eukaryotic genome indices were created with Bowtie2 (15) and HISAT2 (16), respectively. HISAT2 can take into account alternative splicing of genes and was used for eukaryotic genome indexing. For each biological replicate, raw sequencing reads in FastQ format were trimmed with Trimmomatic (17) to remove adapter content. During this process, reads that are <36 bases long are dropped from the analysis. Afterward, surviving reads were mapped to host genome index with HISAT2. Mapped reads were stored as BAM files, and unmapped reads were kept in a separate FastQ file. FeatureCounts (18), together with the host annotation file, was used for gene counting, and a matrix of read counts was generated where each row represents an annotated gene and each column represents a different condition or biological replicate. Unmapped reads from the previous mapping step were then mapped back to the bacterial genome index with Bowtie2, and a matrix of read counts was produced similarly by using the bacterial annotation file and FeatureCounts. HISAT2 and Bowtie2 are run with default parameters as a way to simplify and standardize criteria when analyzing data coming from heterogeneous sources. Finally, to calculate gene expression changes in treated against control conditions, differential expression analysis was performed separately for the bacterial and the host matrices by using the DESeq2 R Package (19). For this, the Wald test was used under the null hypothesis that there is no differential expression between the control and the treated samples. The estimated gene expression change value [measured in $\log_2$ fold change (FC)] and its associated *P*-value were generated for each annotated gene with detected reads in at least one condition (Figures 1 and 2). *P*-values were corrected for multiple testing using the Benjamini–Hochberg method. All additional information such as bacterial ID, host ID, time point, experimental condition (*in vivo*/*in vitro*) and cell type/tissue was added to each gene to create the final format as displayed in DualSeqDB.

Technical aspects

DualSeqDB was built using PHP on an Apache web server with a MySQL database backend. Sequence identifiers and cross-references were obtained from UniProt and the NCBI RefSeq, Gene and Genome resources (20). DualSeqDB stores no user data, except for the anonymous caching of BLAST search results for a given sequence in order to greatly speed up repeated searches. The open-source Bootstrap library was used to allow display on devices of any screen size, including mobile devices. Several icons were included from Font Awesome and the Noun Project, and a number of JavaScript libraries are used for table export and sorting. Please see the About section of the website for detailed attributions.

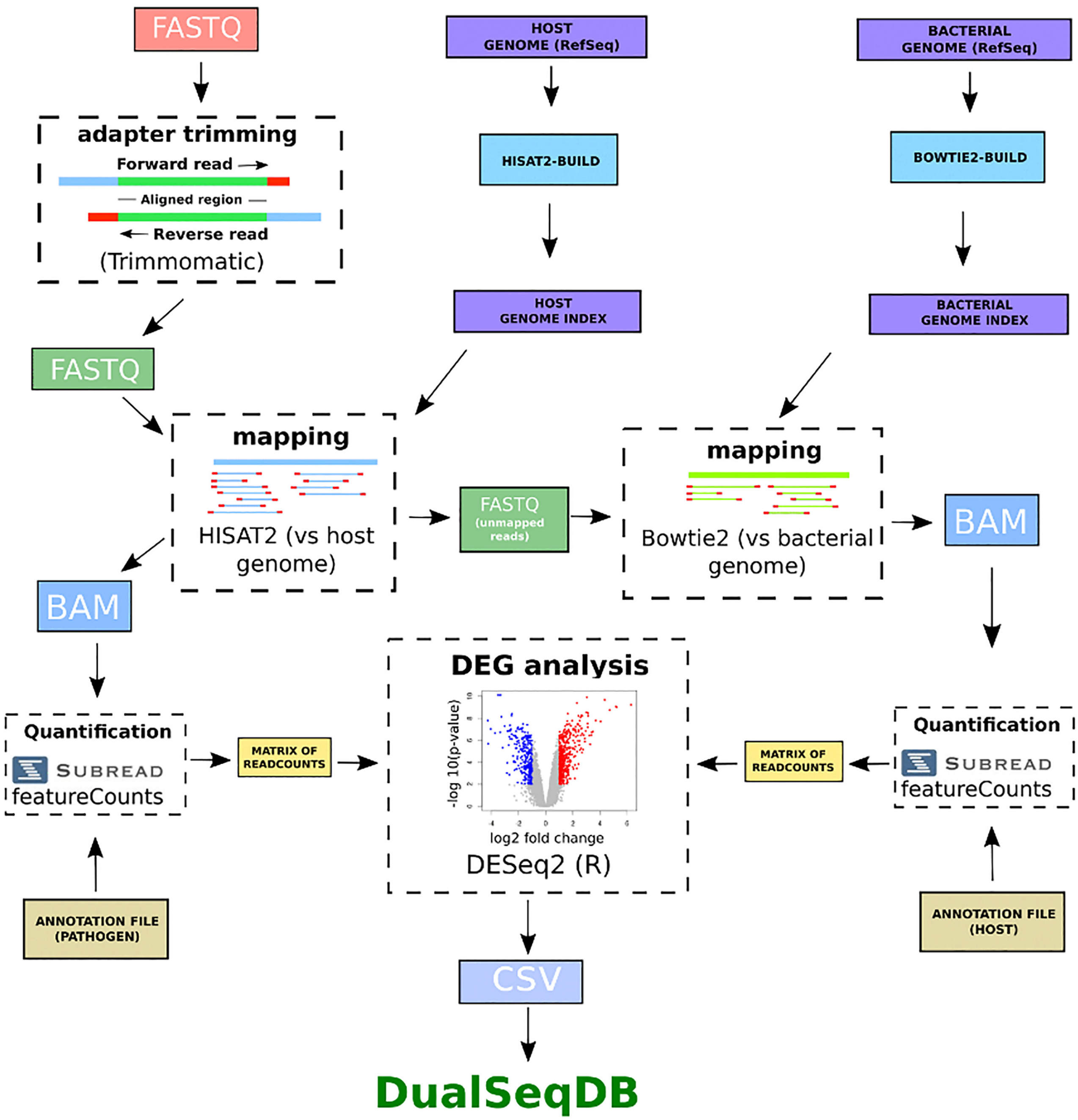


Figure 1. Pipeline used to process raw sequencing data from dual RNA-seq studies. Raw sequencing data were downloaded from the corresponding repository in FastQ format and adapter sequences were removed with Trimmomatic. Pathogen and host genomes and annotations were downloaded to build genome indices. Trimmed FastQ files were then mapped to the host index genome with HISAT2 and the unmapped reads were subsequently mapped to the pathogen index genome using Bowtie2. From this point onward, pathogen and host reads were analyzed in parallel. Mapped reads were quantified with FeatureCounts and their respective annotation files, creating a matrix of read counts. This matrix containing control and treated samples is then used as input for the DESeq2 R package to perform a differential expression analysis. The differential gene expression changes (measured as $\log_2$ FC) and corresponding *P*-values (Benjamini–Hochberg correction for multiple testing) were calculated using DESeq2.

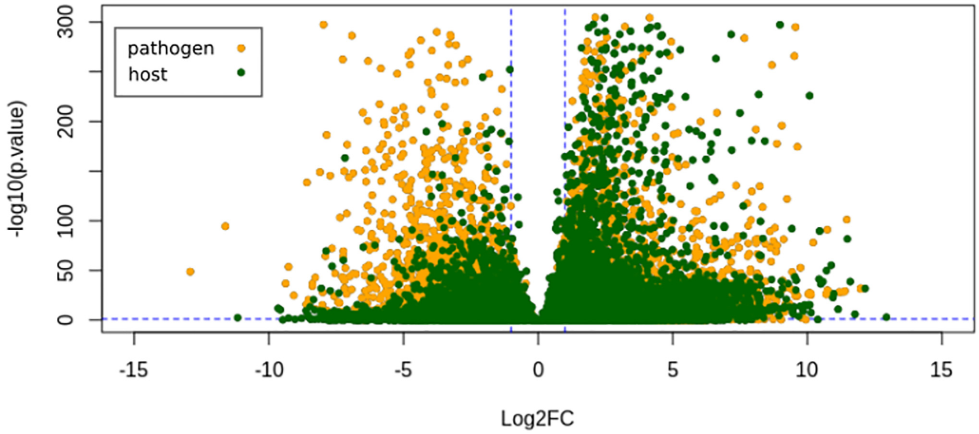


Figure 2. Visualization of overall statistical significance (*P*-value) and magnitude of change ($\log_2$ FC) of all entries in DualSeqDB. Gene expression changes were considered significant when $\log_2$ FC > 1 (upregulated) or $\log_2$ FC < −1 (downregulated) when the corresponding *P*-value was <0.05 (dashed blue lines). Pathogen genes are labelled in yellow and host genes in green.

BLAST search

The NCBI BLAST suite version 2.9.0+ (March 2019) (21) is used to search by sequence similarity. The BLASTP program is used for amino acid sequences, and BLASTX for nucleic acid (coding) sequences. BLAST search results are cached for each unique sequence, which means that re-running a search using the same sequence will yield results nearly instantaneously. As on all other pages, results from the BLAST search page can be linked to and shared with other researchers using the 'Link to these results' link at the bottom of the page. For sequences above a URL length of 2000 characters, this link uses a sequence hash identifying the cached sequence, rather than the sequence itself.

Protein visualization

For UniProt proteins, a protein visualization is automatically generated and displayed by ProViz (22). ProViz is an interactive exploration tool for investigating the structural, functional and evolutionary features of proteins and is likely to be particularly helpful for analyzing uncharacterized proteins.

USING DUALSEQDB

DualSeqDB consists of several elements: (i) a text search function to find specific eukaryotic and bacterial genes; (ii) a BLAST search function to find genes similar to a protein or nucleic acid sequence of interest; (iii) a Browse function to quickly identify genes up- and downregulated during infection; and (iv) a Tutorial section to get started quickly by following a step-by-step guide. DualSeqDB relies on JavaScript; therefore, users will need to enable this in their web browser for full functionality.

Search function

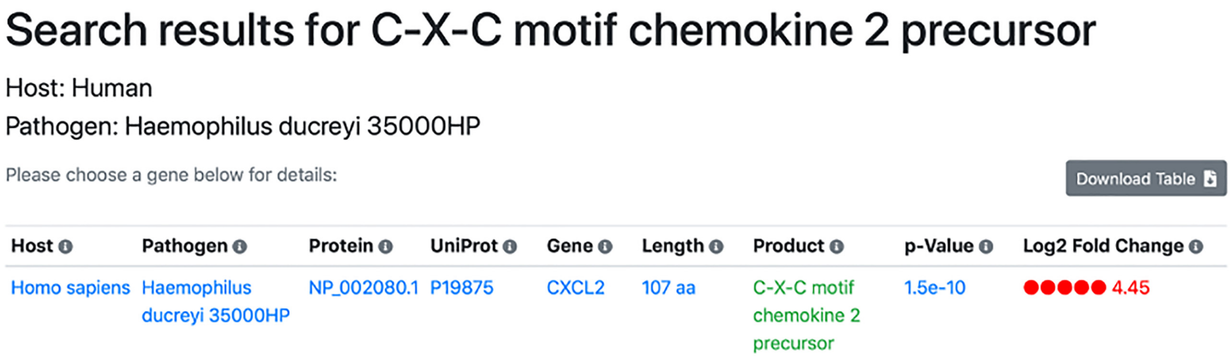
To search for a gene or protein, users simply need to type its name or identifier. Any of the following options are available: gene symbols, gene locus identifiers, NCBI protein identifiers, UniProt protein accessions or a free-text search in the gene product's description (Figure 3). To search within a particular host and/or pathogen, users can select the pathogen and/or host name in the drop-down menu. If no gene or protein name is given, the output will display a complete list of genes, similar to the Browse view (described below).

Tables on DualSeqDB: sorting, downloading and linking to results

After selecting a gene of interest, a view will open with all the information available for the corresponding gene (Figure 4). The heading of this page provides information on the selected protein: protein and host/pathogen name, length, gene name and UniProt ID. In the table, all available experimental data are listed: tissue of the host organism, tissue condition (whether the experiment was carried out *in vivo* or *in vitro*), time after infection, differential expression gene data, including the $\log_2$ FC and the associated *P*-value, a note giving information on the growth conditions of control bacteria (including temperature and growth phase, whenever specified in its study, otherwise it is shown as 'none') and the reference to the original paper where the data were published.

A brief description on the meaning of $\log_2$ FC and *P*-value is also available as mouse-over explanation on the column headers. For any proteins in UniProt, a protein visualization is automatically provided by ProViz from the Davey lab. Proteins >5000 amino acids are not displayed due to display speed limitations. ProViz is an interactive exploration tool that allows inspection of the structural, functional and evolutionary features of proteins, including Pfam domains and transmembrane regions. This tool is particularly useful for unknown and uncharacterized proteins.

Alternatively, the protein's FASTA sequence can be displayed by pressing the 'Show protein sequence' button, along with a 'Copy' link in the top right corner to copy and paste the protein's sequence into other research tools, or into the DualSeqDB BLAST Search to search for similar proteins. You can also immediately search for similar



Search results for C-X-C motif chemokine 2 precursor

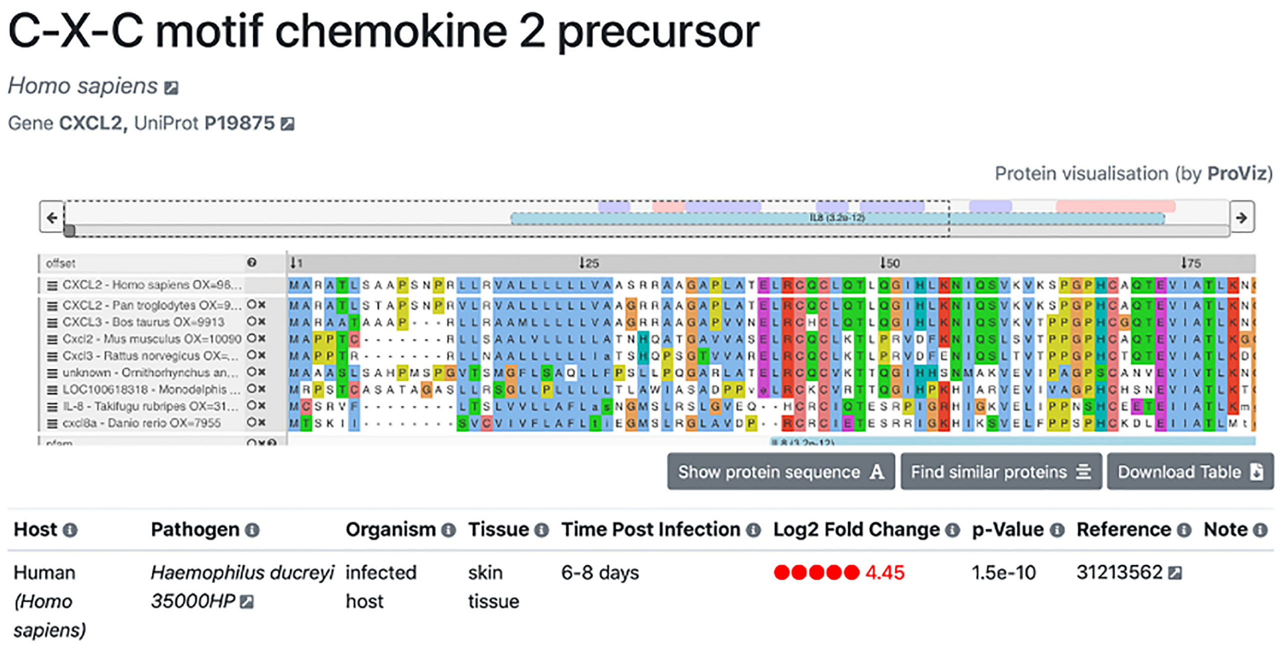
Host: Human

Pathogen: Haemophilus ducreyi 35000HP

Please choose a gene below for details:

Host	Pathogen	Protein	UniProt	Gene	Length	Product	p-Value	Log2 Fold Change
Homo sapiens	Haemophilus ducreyi 35000HP	NP_002080.1	P19875	CXCL2	107 aa	C-X-C motif chemokine 2 precursor	1.5e-10	●●●●● 4.45

Figure 3. Search results example summary. The search results page displays a list of any host or bacterial genes matching the search term. It also displays information on the infected host species and its associated pathogenic bacteria, the NCBI protein identifier, the UniProt protein accession code and the gene symbol of the gene for which the expression change was measured. This preview section also shows a description of the gene product and its length, together with the expression change value (measured as $\log_2$ FC) and the corresponding *P*-value. In this example, we show the case of CXCL2, a chemoattractant chemokine with pro-inflammatory function, involved in many immune responses, such as cancer metastasis, wound healing or angiogenesis. The results collected in DualSeqDB show that, upon infection of skin tissue human cells with the pathogen *Haemophilus ducreyi*, the human gene CXCL2 increases its expression levels, as indicated by $\log_2$ FC > 4.



C-X-C motif chemokine 2 precursor

Homo sapiens

Gene **CXCL2**, UniProt **P19875**

Protein visualisation (by **ProViz**)
IL8 (3.2e-12)
offset
1
25
50
75
CXCL2 - Homo sapiens OX=96...
CXCL2 - Pan troglodytes OX=9...
CXCL3 - Bos taurus OX=9913
Cxcl2 - Mus musculus OX=10090
Cxcl3 - Rattus norvegicus OX=...
unknown - Ornithorhynchus an...
LOC100618318 - Monodelphis ...
IL-8 - Takifugu rubripes OX=31...
cxcl8a - Danio rerio OX=7955
pfam
IL8 (3.2e-12)
Show protein sequence
Find similar proteins
Download Table

Host	Pathogen	Organism	Tissue	Time Post Infection	Log2 Fold Change	p-Value	Reference	Note
Human *(Homo sapiens)*	*Haemophilus ducreyi 35000HP*	infected host	skin tissue	6-8 days	●●●●● 4.45	1.5e-10	31213562	

Figure 4. Detailed view of gene expression changes. The detailed view page displays all the information available for a host or bacterial gene, along with a ProViz visualization of the protein's sequence and structural features, showing sequence conservation with similar proteins. It also shows the $\log_2$ FC and the associated *P*-value for each entry, together with all the details of the experiment: host name, pathogen name, organism (indicating whether the measured gene belongs to the host or the pathogen), cell type/tissues, post-infection time points, as well as the PMID reference with a PubMed link to the original study, and a note column, specifying the bacterial growth conditions if listed in the original study.

proteins via BLAST (see below for more details) by pressing the 'Find similar proteins' button.

To sort the table as desired, users can select any of the column headers. The current table can be downloaded as a comma-separated CSV file for export into spreadsheet software such as Microsoft Excel using the 'Download Table' button in the top right. An appropriate readable file name is automatically generated. The results can also be linked to and shared with other researchers by right-clicking and copying the 'Link to these results' link at the bottom of the page.

BLAST search

The BLAST Search tab provides a search by sequence similarity. When the protein of interest is not in our database, the user may search for similar proteins using BLAST sequence alignment. Finding a similar protein with a high variation in $\log_2$ FC and low *P*-value is a strong indication that the query sequence may be relevant during infection. To search for similar proteins in our database using BLAST, protein or coding sequences in FASTA format can be used and have to be properly identified in the drop-down menu. When the

BLAST alignment is ready, a search results page will open with the following information:

1. *Identity*: The percentage of sequence identity between query and target in the successfully aligned region.
2. *Aligned*: The total number of amino acids that were successfully aligned between query and target.
3. *Bit score*: The required size of a sequence database in which the current match could be found just by chance. The bit score is a $\log_2$-scaled and normalized raw score, meaning that each increase by one doubles the required database size.
4. *E-value*: The number of expected hits of similar quality (score) that could be found in the BLAST sequence database just by chance.

The meaning of the Host, Pathogen, Locus, Protein, Gene, Product, *P*-value and $\log_2$ FC columns can be found in the Browse tab section below, or via the mouse-over information symbols in the top row of any table. By default, BLAST matches with the highest bit scores are shown first and matches with 100% sequence identity will be highlighted in green. Tables can also be sorted as desired using the column headers. As for all tables, the results can be downloaded as a comma-separated CSV file for export into spreadsheet software such as Microsoft Excel using the 'Download Table' button in the top right corner. An appropriate readable file name is automatically generated. The results can also be linked to and shared with other researchers by right-clicking and copying the 'Link to these results' link at the bottom of the results table.

Browsing the entire database

The Browse tab provides an overview of all entries in the DualSeqDB database. A pathogenic species or a host of interest can be chosen in the selection element at the top. This table is sorted by significance and $\log_2$ FC. It displays pathogen/host genes with a high and significant change in expression during infection at the top, followed by insignificant genes by decreasing $\log_2$ FC in absolute value. Genes with very little expression changes and high *P*-value are listed at the very end of the table. Arrows next to each field provide links to useful external databases:

1. *Pathogen/Host*: Links out to the NCBI Taxonomy database, a comprehensive taxonomic database.
2. *Locus*: Links out to the Ensembl database, which provides genome annotation for all species included in the database.
3. *Protein*: Links out to the NCBI Protein database, which provides protein sequences and information.
4. *UniProt Accession and Gene Symbol*: Links out to the UniProt Knowledgebase, which provides comprehensive protein annotation.

Users can select the Locus, Protein, and UniProt Accession and Gene Symbol entries to view details for the given protein in the external databases. This information is also available as a mouse-over explanation in the Browse tab.

DISCUSSION

The development of new antimicrobial therapies heavily relies on our knowledge of the mechanisms of bacterial infection (23–25). Therefore, it is crucial to understand how bacterial infection develops and which bacterial genes are required to infect a host . The use of high-throughput sequencing technologies has unveiled new levels of complexity in the transcriptomic response of pathogens and hosts during infection. In the last few years, dual RNA-seq has become the leading approach to uncover the intricate relationship between pathogen and host interactions. Hence, dual RNA-seq could be used to define host–pathogen interactions or identify potential biomarkers of infection (26). At present, dual RNA-seq data are disseminated in multiple locations and incompatible formats and are therefore not accessible to the scientific community without specialized tools and knowledge.

DualSeqDB intends to be a valuable central resource for the systematic identification of proteins that are crucial for successful infection, aimed to understand how the bacterial and host transcriptomes change and interact during infection. In this context, we envisage that DualSeqDB will facilitate the finding of interspecies relationships between pathogen and host and will help us to uncover new mechanisms of infection. The analysis of the results included in DualSeqDB may inspire the design of new therapeutic interventions aimed to prevent the spread of infection. Given the current *momentum* of sequencing technologies in research and clinics, we expect that our database will grow continuously and become a comprehensive repository that will help us in the fight against infectious diseases.

DATA AVAILABILITY

To download the entire DualSeqDB database for local analysis, please click the link available under the Download tab. Currently, DualSeqDB v1 is available, and will be upgraded with new data as they become available.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Norman E. Davey for his kind support in providing ProViz for protein visualization.

FUNDING

Ministerio de Ciencia, Innovación y Universidades [SAF2015-72518-EXP, SAF2017-82158-R, BFU2017-86970-P and RYC-2012-09999]; European Research Council [RIBOMYLOME 309545]; Horizon 2020—Research and Innovation Framework Programme [727658, 793135 to B.L.]; Ministry of Economy and Competitiveness [BFU2017-86970-P]; Centro de Excelencia Severo Ochoa 2013–2017; Generalitat de Catalunya; European Society of Clinical Microbiology and Infectious Diseases Research Grant 2016. Funding for open access charge: Ministerio de Ciencia, Innovación y Universidades [BFU2017-86970-P].

Conflict of interest statement. None declared.

REFERENCES

1. Westermann,A.J., Gorski,S.A. and Vogel,J. (2012) Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.*, **10**, 618–630.
2. Aprianto,R., Slager,J., Holsappel,S. and Veening,J.W. (2016) Time-resolved dual RNA-seq reveals extensive rewiring of lung epithelial and pneumococcal transcriptomes during early infection. *Genome Biol.*, **17**, 198.
3. Nuss,A.M., Beckstette,M., Pimenova,M., Schmuhl,C., Opitz,W., Pisano,F., Heroven,A.K. and Dersch,P. (2017) Tissue dual RNA-seq allows fast discovery of infection-specific functions and riboregulators shaping host–pathogen transcriptomes. *Proc. Natl Acad. Sci. U.S.A.*, **114**, E791–E800.
4. Westermann,A.J., Forstner,K.U., Amman,F., Barquist,L., Chao,Y., Schulte,L.N., Muller,L., Reinhardt,R., Stadler,P.F. and Vogel,J. (2016) Dual RNA-seq unveils noncoding RNA functions in host–pathogen interactions. *Nature*, **529**, 496–501.
5. Westermann,A.J., Barquist,L. and Vogel,J. (2017) Resolving host–pathogen interactions by dual RNA-seq. *PLoS Pathog.*, **13**, e1006033.
6. Saliba,A.E., Cantos,S.C. and Vogel,J. (2017) New RNA-seq approaches for the study of bacterial pathogens. *Curr. Opin. Microbiol.*, **35**, 78–87.
7. Baddal,B., Muzzi,A., Censini,S., Calogero,R., Torricelli,G., Guidotti,S. and Paxxicoli,A. (2015). Dual RNA-seq of nontypeable *Haemophilus influenzae* and host cell transcriptomes reveals novel insights into host–pathogen cross talk. *mBio*, **6**, e0176515.
8. Choi,Y.-J., Aliota,M. T., Mayhew,G. F., Erickson,S. M. and Christensen,B. M. (2014). Dual RNA-seq of parasite and host reveals gene expression dynamics during filarial worm–mosquito interactions. *PLOS Negl. Trop. Dis.*, **8**, e2905.
9. 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. *et al.* (2020) New pathogenesis mechanisms and translational leads identified by multidimensional analysis of necrotizing myositis in primates. *mBio*, **11**, e03363-19.
10. Schulte,L.N., Schweinlin,M., Westermann,A.J., Janga,H., Santos,S.C., Appenzeller,S., Walles,H., Vogel,J. and Metzger,M. (2020) An advanced human intestinal coculture model reveals compartmentalized host and pathogen strategies during *Salmonella* infection. *mBio*, **11**, e03348.
11. Damron,F.H., Oglesby-Sherrouse,A.G., Wilks,A. and Barbier,M. (2016) Dual-seq transcriptomics reveals the battle for iron during *Pseudomonas aeruginosa* acute murine pneumonia. *Sci. Rep.*, **6**, 39172.
12. Griesenauer,B., Tran,T.M., Fortney,K.R., Janowicz,D.M., Johnson,P., Gao,H., Barnes,S., Wilson,L.S., Liu,Y. and Spinola,S.M. (2019) Determination of an interaction network between an extracellular bacterial pathogen and the human host. *mBio*, **10**, e01193.
13. 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. and Gengenbacher,M. (2015) Comprehensive insights into transcriptional adaptation of intracellular mycobacteria by microbe-enriched dual RNA sequencing. *BMC Genomics*, **16**, 34.
14. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
15. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
16. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.
17. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
18. Liao,Y., Smyth,G.K. and Shi,W. (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.*, **47**, e47.
19. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
20. NCBI,Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **44**, D7–D19.
21. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
22. Jehl,P., Manguy,J., Shields,D.C., Higgins,D.G. and Davey,N.E. (2016) ProViz: a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. *Nucleic Acids Res.*, **44**, W11–W15.
23. Crua Asensio,N., Muñoz Giner,E., de Groot,N.S. and Torrent Burgas,M. (2017) Centrality in the host–pathogen interactome is associated with pathogen fitness during infection. *Nat. Commun.*, **8**, 14092.
24. de Groot,N.S. and Torrent Burgas,M. (2019) ACoordinated Response at The Transcriptome and Interactome Level is Required to Ensure Uropathogenic. *Microorganisms*, **7**, 292.
25. Rendón,J.M., Lang,B., Tartaglia,G.G. and Burgas,M.T. (2020) BacFITBase: a database to assess the relevance of bacterial genes during host infection. *Nucleic Acids Res.*, **48**, D511–D516.
26. Wolf,T., Kammer,P., Brunke,S. and Linde,J. (2018) Two's company: studying interspecies relationships with dual RNA-seq. *Curr. Opin. Microbiol.*, **42**, 7–12.