Dual RNA-Seq meta-analysis in *Plasmodium* infection

Parnika Mukherjee[1,2,3] and Emanuel Heitlinger[*1,2]

[1]Department of Molecular Parasitology, Humboldt University Berlin, Philippstraße 13, Haus 14, 10115 Berlin, Germany

[2]Research Group Ecology and Evolution of Molecular Parasite-Host Interactions, Leibniz-Institut für Zoo- und Wildtierforschung (IZW), Alfred-Kowalke-Straße 17, 10315 Berlin, Germany

[3]Department of Immunology and Infectious Disease, The John Curtin School of Medical Research, The Australian National University, Canberra, Australia

Abstract

Dual RNA-Seq is the simultaneous retrieval and analysis of host and pathogen transcriptomes. It follows the rationale that cross-species interactions determine both pathogen virulence and host tolerance, resistance or susceptibility. Correlated gene expression might help identify interlinked signaling, metabolic or gene regulatory pathways in addition to potentially physically interacting proteins. Numerous studies have used RNA-Seq to investigate *Plasmodium* infection with a focus on only one organism, either the host or the parasite.

Here we propose a meta-analysis approach for dual RNA-Seq. We screened malaria transcriptome experiments for gene expression data from both *Plasmodium* and its host. Out of 105 malaria studies in *Homo sapiens, Macaca mulatta* and *Mus musculus*, we identified 56 studies with the potential to provide host and parasite data. While 15 studies (1,935 total samples) of these 56 explicitly aimed to generate dual RNA-Seq data, 41 (1,129 samples) had an original focus on either the host or the parasite. We show that a total of up to 2,530 samples are suitable for dual RNA-Seq analysis providing an unexplored potential for meta-analysis.

We argue that the multitude of variations in experimental conditions should help narrow down a conserved core of cross-species interactions. Different hosts are infected by evolutionarily diverse species of the genus *Plasmodium*. We propose to overlay interaction networks of different host-parasite systems based on orthologous genes. This might allow us to gauge the applicability of model systems for different pathways in malaria infection and to address the evolution of parasite-host interactions.

Keywords: dual RNA-Seq, meta-analysis, malaria, correlated expression, host-parasite interactions

Word count: 2836, Target journal word limit: 3000

[*]Corresponding author: Emanuel Heitlinger, emanuel.heitlinger@hu-berlin.de

Introduction

Transcriptomes are often analyzed as a first attempt to understand cellular and organismic events, because a comprehensive profile of RNA expression can be obtained at a reasonable cost and with high technical accuracy [37]. Microarrays dominated transcriptomics since 1995 [11, 22, 44]. Microarrays quantify gene expression based on hybridization of a target sequence to an immobilized probe of known sequence. Technical difficulties associated with microarrays lie in probe-selection, cross-hybridization, and design cost of custom chips [51]. RNA sequencing (RNA-Seq) eliminates these difficulties and provides deep and accurate expression estimates for all RNAs in a sample. RNA-Seq has thus replaced microarrays as the predominant tool for transcriptomics [37, 50]. RNA-Seq assesses host and parasite transcriptomes simultaneously, if RNA of both organisms is contained in a sample. It has been proposed to analyze the transcriptomes of both organisms involved in an infection as pathogen virulence in an infectious disease is often a result of interlinked processes of host and pathogen (host-pathogen interactions) [32, 50, 51].

In case of malaria, unlike in bacterial infections, both the pathogen and the host are eukaryotic organisms with similar transcriptomes. Host and parasite mRNA is selected simultaneously when poly-dT priming is used to amplify polyadenylated transcripts [50, 51]. This means that most malaria RNA-Seq datasets potentially contain transcripts of both species, making them suitable for dual RNA-Seq analysis. In the mammalian intermediate host, *Plasmodium* first invades liver and then red blood cells (RBCs) for development and asexual expansion. While the nuclear machinery of cells from both host and parasite produces mRNA in the liver, RBCs are enucleated and transcriptionally inactive in the mammalian host [39]. In blood stage of the infection, leukocytes are, thus, the source of host mRNA. Malaria research, especially transcriptomics, is traditionally designed to target one organism, either the host or the parasite. Expression of mRNA, for example, can be compared between different time points in the life cycle of *Plasmodium* or between different drug treatment conditions. Researchers conducting a targeted experiment might consider transcripts from the non-target organism "contamination". Nevertheless, expression of those transcripts potentially corresponds to response to stimuli during the investigation. Additionally, some recent studies on malaria make intentional use of dual RNA-Seq. Malaria is the most thoroughly investigated disease caused by a eukaryotic organism and the accumulation of these two kinds of studies, RNA-Seq with "contaminants" and intentional dual RNA-Seq, provides a rich resource for meta-analysis.

Such a meta-analysis can use co-regulated gene expression to infer host-parasite interactions. Correlation of mRNA expression can be indicative of different kinds of biological "interactions": On one

hand, protein products could be directly involved in the formation of complexes and might therefore be produced in quantities varying similarly under altered conditions. On the other hand, involvement in the same biological pathways can result in co-regulated gene expression without physical interaction. This broad concept of interaction has long been exploited in single organisms (e.g. [10, 33, 42, 46]). We (and others before [41]) propose to extrapolate this to interactions between a host and its pathogen. It can be expected that a stimulus presented by the parasite to its host elicits host immune response and that the parasite, in turn, tries to evade this response, creating a cascade of genes co-regulated at different time points or under different conditions.

In this paper, we explore first steps in a comparative meta-analysis of dual RNA-Seq transcriptomes. Existing raw read datasets collectively present an unexplored potential to answer questions that have not been investigated by individual studies. Meta-analyses increase the number of observations and statistical power and help eliminate false positives and true negatives, which may otherwise conceal important biological inferences [16, 17, 21]. Since mice and macaques are used as laboratory models for human *Plasmodium* infection, we analyzed the availability and suitability of mRNA sequencing data from three evolutionarily close hosts - *Homo sapiens, Macaca mulatta* and *Mus musculus* - and their associated *Plasmodium* parasites. Here we summarize available data, outline challenges and show exemplary approaches to uncover host-parasite interactions. We discuss orthology across different host-parasite systems as a means to enrich information.

Data review and curation of potentially suitable studies

Sequence data generated in biological experiments is submitted to one of the three mirroring databases of the International Nucleotide Sequence Database Collaboration (INSDC): NCBI Sequence Read Archive (SRA), EBI Sequence Read Archive (ERA) and DDBJ Sequence Read Archive (DRA). Comprehensive query tools to access these databases via web interfaces and programmatically via scriptable languages exist (for example, SRAdb [53], ENAbrowseR [47]). In these databases, all experiments submitted under a single accession are given a single "study accession number" and are collectively referred to as a "study" here onwards.

We used SRAdb, a Bioconductor/R package [24, 25], to query SRA [34] for malaria RNA-Seq studies with the potential to provide host and *Plasmodium* reads for our meta-analysis. We first selected studies with "library_strategy" given as "RNA-Seq" and "*Plasmodium*" in study title, abstract or sample attributes using the "dbGetQuery" function. Then we used the "getSRA" function with the

query "(malaria OR *Plasmodium*) AND RNA-Seq". This function searches all fields. We manually curated the combined results and added studies based on a literature review using the terms described for the "getSRA" function in SRA, PubMed and Google Scholar. From this search resulting in 196 studies, we disregarded 91 studies, all of which provide data from vectors and non-target hosts (e.g. avian malaria). 49 more studies were excluded because their gene expression data was derived from *Plasmodium. spp* in blood-stage *Plasmodium* cultures and thus can be expected to be devoid of host mRNA. We then used SRAdb and the prefetch and fastq-dump functions from SRAtoolkit [6], to download all replicate samples (experimental replicate or "run" in the jargon of sequencing databases) of the selected studies. The final curation of studies and download of runs was performed on 21 January, 2019.

In total, we found 56 potentially suitable studies in this database and literature review (Figure 1a). *Homo sapiens* was the host organism for 22 studies, *Mus musculus* for 24 and *Macaca mulatta* for 10. Human studies featured *P. falciparum* and *P. vivax*. *P. yoelii*, *P. chabaudi* and *P. berghei* were used in mouse infection studies. *P. berghei* was additionally used to infect human liver cell culture. Macaque studies included *P. cynomolgi* and *P. coatneyi* infections(table 1).

We note that 20 of the 56 studies depleted (or enriched, respectively) specific classes of cells from their samples. The low number of parasites cells at the physiologically asymptomatic liver stage [30], on one hand, gives the opportunity to test the effect of sporozoite-derived vaccines [9, 23, 45], but on the other hand, makes it difficult to study *Plasmodium* transcriptomes in this stage. To reduce overwhelming host RNA levels, 3 out of 10 liver studies sorted infected hepatoma cells from uninfected cells. Similarly, 17 other studies have depleted or enriched host white blood cells (WBCs or leukocytes) to focus expression analysis on *Plasmodium* or the host immune system, respectively. In all these scenarios, we suspect depletion to be imperfect and thus the samples to potentially include mRNA of both organisms.

For 15 out of the 56 studies, the authors state that they intended to simultaneously study host and parasite transcriptomes ("dual RNA-Seq"). This includes 8 studies from MaHPIC (Malaria Host-Pathogen Interaction Center), based at Emory University, that made extensive omics measurements in macaque malaria. The original focus of the remaining 41 studies was on the parasite in 20 and on the host in 21 cases.

Besides invading liver and RBCs, *Plasmodium* parasites sequester in bone marrow, adipose tissue, lung, spleen and brain (during cerebral malaria) [12, 18]. To study a comprehensive spectrum of host-parasite interactions, it would be optimal to have data from these different tissues. Our collection of

studies comprise data derived from blood and liver for all three host organisms. In addition, we have found seven transcritpomic studies on spleen [1–3,7,19,29,36] and two studies on cerebral malaria [5,48] from mice. MaHPIC offers a collection of blood and bone marrow studies in macaques.

Experiments performed on mouse blood focus on the parasite instead of the host (11 vs. 0). Studies on human blood infection focus more often on the host immune response than on the parasite (9 vs. 5). Liver and spleen studies focus on host and parasite almost equally as often, with sources for host tissue in this case being either mice (*in vivo*) or hepatoma cell cultures (*in vitro*).

Below, we argue that small clusters of genes co-expressed across several such diverse conditions might help point towards potentially novel core host-parasite interactions. In addition, the extent of agreement in gene expression correlation between *in vivo* and *in vitro* systems, different tissues and organisms, might help us assess how far different biological pathways of these systems have model characteristics for human malaria.

Dual RNA-Seq suitability analysis

A sample suitable for dual RNA-Seq analysis must provide "sufficient" gene expression from both host and parasite. To assess the proportion for host and parasite RNA sequencing reads in each study and sample, we mapped sequencing reads onto concatenated host and parasite reference genomes using STAR [14, 15]. Simultaneous mapping against both genomes should avoid non-specific mapping of reads in regions conserved between host and parasites. We quantified the sequencing reads mapped to exons using the "countOverlaps" function of the GenomicRanges package [31] and calculated the proportion of reads mapping to host and parasite genes.

The proportions of host and parasite reads for each run does not always reflect the original focus of a study (fig. 1a). Studies without depletion or enrichment give us an idea how skewed overall RNA expression towards one organism is to be expected under native conditions. In studies originally designed to use a dual RNA-Seq approach on blood stages, samples in which parasitemia is very high are used. Samples with lower parasitemia, such as many of those for which, the original focus is mostly on immune gene expression from leukocytes, the number of host reads is often overwhelming and these samples are mostly not suitable for dual RNA-Seq analysis (table 1).

Many studies using depletion or enrichment prior to RNA sequencing ("enriched/depleted" in fig. 1a) still show considerable expression of the non-target organism. Studies on liver infection, for example, [4] and [20], comprise several runs with balanced proportions of host and parasite reads. This

is a result of infected liver cells being sorted from uninfected cells in culture. Studies depleting whole blood from leukocytes to focus on parasite transcriptomes still show considerable host gene expression and provide principally suitable runs for the analysis of blood infection at lower intensities. The latter comes with the caveat that host expression might be biased by unequal depletion of particular cell types. This could be the case if incomplete depletion affected different types of WBCs differentially and hence biased the detectable host mRNA expression in the direction of less depleted cell types. For similar reasons, controlling for parasite depletion might be challenging.

To establish suitability thresholds for inclusion of individual samples in further analysis, we plotted the number of host and parasite reads against the number of host and parasite genes expressed(fig. 1b and fig. 1c). For runs with high sequencing depth, the total number of expressed genes of the host and parasite approaches the number of all annotated genes: around 30,000 for the mammalian host and around 4,500 for *Plasmodium*. When sequencing depth is lower, the number of genes detected as "expressed" is lower and a decrease in sensitivity can be expected to prevent analysis of poorly expressed genes. We propose four parameters for suitability thresholds in dual RNA-Seq analysis: the number reads mapping to (1)host and (2)parasite genes and the number of genes these reads map to (expressed genes) in (3)host and (4)parasite. In table 1, we give the number of runs considered suitable for three different combinations of thresholds. Without claiming a particular threshold to be ideal, we propose to use thresholds to avoid uninformative runs in further processing to reduce the unnecessary computational burden of co-expression analysis.

Suitable runs bearing "contaminants" at the thresholds chosen here are identified from human-*P. falciparum*, macaque-*P. cynomolgi*, human-*P. berghei* and mouse-*P. berghei* systems. Unfortunately, with current thresholds and currently available data, we highly under-represent human-*P. vivax* and human-*P. berghei* systems, the two liver *in vitro* models. This outcome is understandable owing to the low parasitemia in liver cultures [13]. We note that the thresholds could further be made lenient enough to include more runs for these systems at the cost of analyzing only the most highly expressed parasite genes. An alternative approach relies on using depleted/enriched samples for these systems. For further analysis, however, including depleted/enriched samples could prove challenging as discussed before. Analysis approaches such as multilayer networks (see below) might help to gauge problems with such runs for the inference of co-expression in further steps of analysis.

Identification of co-expressed genes via correlation techniques

Some genes are likely to show almost uniform expression under different experimental conditions (e.g., "housekeeping genes"). Naïve assessments of correlation could, however, identify pairs of such genes as highly correlated. An analysis of co-expression can deal with this challenge in two different ways:

Firstly, the most variable genes within and across studies can be selected and other genes discarded. While requiring little computational time and resources, exclusion of genes with too little variance in expression from downstream analysis should be performed with caution, as seemingly small variations might result in a suitable signal over a large set of runs. To select only variable genes, one option is to compute their variance across all samples (in one or multiple studies). Genes with variance below a threshold may then be excluded from further analysis. As variance increases with the mean for gene expression data, the Biological Coefficient of Variation (BCV) [38,43] may provide a more robust threshold.

Secondly, one can compute empirical correlation indices, similar to p-values, for any gene-pair. Empirical correlation indices are a robust way to estimate whether gene-pairs are correlated because of specific events (treatment condition, time point) and not by chance (e.g., housekeeping genes) [27,40]. These methods construct a null distribution using permutations of the given data instead of assuming a null distribution in advance. Since host and parasite genomes total nearly 35,000 genes, the number of permutations has to be around 10^{12} (for a resolution of 0.1% FDR) to be suitable for corrections for multiple-testing [8]. Alternatively, as computational costs for these permutations can be expected to be too high for datasets with thousands of samples, uncorrected "p-values" may be considered ranking for host-parasite gene correlation, following the suggestion of Reid and Berriman [41]. Nevertheless, reliance on empirical computation of p-values without prior variance/BCV filtering might become impracticable for very large datasets in the proposed meta-analysis.

We consider partial correlation as an additional approach that could be combined with the above methods. Partial correlation can control pairwise correlations for the influence of other genes [26]. In transcriptomic applications, full-conditioned partial correlation is computationally very expensive. Some studies therefore resort to second-order partial correlation (relationship between two genes independent of two other genes) [54, 55]. A suitable pipeline might first use zero-order partial, i.e., "regular" correlation with empirical p-values to remove constitutively expressed genes. For all correlations with an empirical "p-value" below a certain threshold, one could compute e.g. first-order partial correlations, thus, reducing the number of computations. Iterations of such an approach with higher-order partial correlations are then possible.

Across different studies; across different host-parasite systems

Gene × *gene* matrices obtained from correlation analysis can be visualized and analyzed as interaction networks. We have identified different but interlinked workflows to reconstruct a consensus network of expression correlation (fig. 2). A first approach (fig. 2(a)) integrates data from different studies of one host-parasite system by simply appending expression profiles of their runs.

Knowledge of 1:1 orthologs [28] between different host and different parasite species can be used in the next steps to integrate across different host-parasite systems. Humans and macaques share 18,179 1:1 orthologous genes, humans and mice share 17,089 orthologous genes and 14,776 genes are 1:1:1 orthologous among all three species. Similarly, 7,600 groups of orthologous genes exist among the *Plasmodium* species.

A simple approach to combine data across host-parasite systems could again append these orthologs in the original datasets before correlations of gene expression. Alternatively, to construct a consensus network involving all hosts and parasites, a multi-layer network analysis could align networks by orthologous genes. This approach (fig. 2(b)) can offer more control when looking for correlations consistent in different layers representing different host-parasite systems. Similarly, more insight could be gained when correlations from different types of tissues are combined as multilayer networks. This would require the construction of networks for each study in a single host-parasite system followed by a multi-layered network analysis on these networks.

We anticipate that correlation between host and parasite transcript expression will highlight host-parasite interactions worth scrutiny of further focused research. As a second goal, meta-analysis involving different host-parasite systems could give insights into how easily observations made in malaria models can be translated to human malaria. If, e.g., certain groups of pathways show lower evolutionary conservation in host-parasite co-expression networks, one could expect results on those pathways to be more difficult to translate between systems. Finally, one can ask whether expression correlation between host and parasite species is more or less evolutionarily conserved than within host species in malaria [35, 49, 52].

Acknowledgements

We thank Gaétan Burgio, Alyssa Ingmundson, Kai Matuschewski and Alice Balard for comments on earlier versions of this manuscript.

Funding

This work was supported by the Alliance Berlin Canberra "Crossing Boundaries: Molecular Interactions in Malaria", which is co-funded by a grant from the Deutsche Forschungsgemeinschaft (DFG) for the International Research Training Group (IRTG) 2290 and the Australian National University.

Conflict of Interest Statement

The authors declare that there is no conflict of interest.

Author Contribution

EH and PM designed the study. PM obtained and analyzed the data. EH and PM contributed to the final manuscript.

References

[1] ERP002116. https://www.ncbi.nlm.nih.gov/sra/?term=ERP002116. Accessed: 21-01-2019.

[2] ERP004042. https://www.ncbi.nlm.nih.gov/sra/?term=ERP004042. Accessed: 21-01-2019.

[3] ERP005730. https://www.ncbi.nlm.nih.gov/sra/?term=ERP005730. Accessed: 21-01-2019.

[4] ERP023982. https://www.ncbi.nlm.nih.gov/sra/?term=ERP023982. Accessed: 21-01-2019.

[5] ERP106769. https://www.ncbi.nlm.nih.gov/sra/?term=ERP106769. Accessed: 21-01-2019.

[6] SRAtoolkit. http://ncbi.github.io/sra-tools/. Accessed: 21-01-2019.

[7] SRP164768. https://www.ncbi.nlm.nih.gov/sra/?term=SRP164768. Accessed: 21-01-2019.

[8] Benjamini, Y., and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*, 1 (1995), 289–300. URL: http://www.jstor.org/stable/2346101.

[9] Bijker, E. M., Borrmann, S., Kappe, S. H., Mordmüller, B., Sack, B. K., and Khan, S. M. Novel approaches to whole sporozoite vaccination against malaria. *Vaccine 33*, 52, 7462 – 7468. doi: https://doi.org/10.1016/j.vaccine.2015.09.095.

[10] Bono, H., and Okazaki, Y. Functional transcriptomes: comparative analysis of biological pathways and processes in eukaryotes to infer genetic networks among transcripts. *Current Opinion in Structural Biology 12*, 3 (2002), 355 – 361. doi: https://doi.org/10.1016/S0959-440X(02)00335-4.

[11] Craig, A., and Scherf, A. Transcriptional analysis in *Plasmodium falciparum*. *Trends in Microbiology 8*, 8 (2000), 350 – 351. doi: https://doi.org/10.1016/S0966-842X(00)01808-4.

[12] David, P. H. e. a. Parasite sequestration in *Plasmodium falciparum* malaria: spleen and antibody modulation of cytoadherence of infected erythrocytes. *Proceedings of the National Academy of Sciences of the United States of America 80* (1983). doi: https://www.ncbi.nlm.nih.gov/pubmed/6348780.

[13] Derbyshire, E. R., Prudêncio, M., Mota, M. M., and Clardy, J. Liver-stage malaria parasites vulnerable to diverse chemical scaffolds. *Proceedings of the National Academy of Sciences 109*, 22 (2012), 8511–8516. doi: https://doi.org/10.1073/pnas.1118370109.

[14] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics 29*, 1 (2013), 15–21. doi: https://doi.org/10.1093/bioinformatics/bts635.

[15] Dobin, A., and Gingeras, T. R. Mapping RNA-seq reads with STAR. *Current Protocols in Bioinformatics 51*, 1, 11.14.1–11.14.19. doi: https://doi.org/10.1002/0471250953.bi1114s51.

[16] Fagard, R. H., Staessen, J. A., and Thijs, L. Advantages and disadvantages of the meta-analysis approach. *Journal of hypertension 14*, 2 (Oct 1996). doi: https://doi.org/10.1097/00004872-199609002-00004.

[17] Fayyad, Usama, P.-S., Gregory, S., and Padhraic. From data mining to knowledge discovery in databases. *AI Magazine 17* (1996). doi: https://doi.org/10.1609/aimag.v17i3.1230.

[18] Franke-Fayard, B., Fonager, J., Braks, A., Khan, S. M., and Janse, C. J. Sequestration and tissue accumulation of human malaria parasites: Can we learn anything from rodent models of malaria? *PLOS Pathogens 6*, 9 (09 2010), 1–10. doi: https://doi.org/10.1371/journal.ppat.1001032.

[19] Gabryšová, L., Alvarez-Martinez, M., Luisier, R., Cox, L. S., Sodenkamp, J., Hosking, C., Pérez-Mazliah, D., Whicher, C., Kannan, Y., Potempa, K., Wu, X., Bhaw, L., Wende, H., Sieweke, M. H., Elgar, G., Wilson, M., Briscoe, J., Metzis, V., Langhorne, J., Luscombe, N. M., and O'Garra, A. c-Maf controls immune responses by regulating disease-specific gene networks and repressing IL-2 in CD4+ T cells. doi: https://doi.org/10.1038/s41590-018-0083-5.

[20] Guttery, D. S., Poulin, B., Ramaprasad, A., Wall, R. J., Ferguson, D. J., Brady, D., Patzewitz, E.-M., Whipple, S., Straschil, U., Wright, M. H., Mohamed, A. M., Radhakrishnan, A., Arold, S. T., Tate, E. W., Holder, A. A., Wickstead, B., Pain, A., and Tewari, R. Genome-wide functional analysis of *Plasmodium* protein phosphatases reveals key regulators of parasite development and differentiation. *Cell Host Microbe 16*, 1 (2014), 128 – 140. doi: https://doi.org/10.1016/j.chom.2014.05.020.

[21] Haidich, A. Meta-analysis in medical research. *Hippokratia 14* (Dec 2011). URL: https://www.ncbi.nlm.nih.gov/pubmed/21487488.

[22] Hayward, R. E., DeRisi, J. L., Alfadhli, S., Kaslow, D. C., Brown, P. O., and Rathod, P. K. Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. *Molecular Microbiology 35*, 1 (2000), 6–14. doi: https://doi.org/10.1046/j.1365-2958.2000.01730.x.

[23] Hollingdale, M. R., and Sedegah, M. Development of whole sporozoite malaria vaccines. *Expert Review of Vaccines 16*, 1 (2017), 45–54. doi: https://doi.org/10.1080/14760584.2016.1203784.

[24] Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence,

M., Love, M. I., MacDonald, J., Obenchain, V., Ole's, A. K., Pag'es, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., and Morgan, M. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods 12*, 2 (2015), 115–121. doi: https://doi.org/10.1038/nmeth.3252.

[25] Ihaka, R., and Gentleman, R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics 5*, 3 (1996), 299–314. doi: https://doi.org/10.1080/10618600.1996.10474713.

[26] Khanin, R., and Wit, E. *Construction of Malaria Gene Expression Network Using Partial Correlations*, vol. 5. 02 2007, pp. 75–88. doi: https://doi.org/10.1007/978-0-387-34569-7_6.

[27] Knijnenburg, T. A., Wessels, L. F. A., Reinders, M. J. T., and Shmulevich, I. Fewer permutations, more accurate P-values. *Bioinformatics 25*, 12 (2009), i161–i168. doi: https://doi.org/10.1093/bioinformatics/btp211.

[28] L, L., Christian J. Stoeckert, J., and Roos, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research 13* (2003). doi: https://doi.org/10.1101/gr.1224503.

[29] Lai, S. M., Sheng, J., Gupta, P., Renia, L., Duan, K., Zolezzi, F., Karjalainen, K., Newell, E. W., and Ruedl, C. Organ-specific fate, recruitment, and refilling dynamics of tissue-resident macrophages during blood-stage malaria. *Cell Reports 25*, 11 (2018), 3099 – 3109.e3. doi: https://doi.org/10.1016/j.celrep.2018.11.059.

[30] LaMonte, G., Orjuela-Sanchez, P., Wang, L., Li, S., Swann, J., Cowell, A., Zou, B. Y., Abdel-Haleem Mohamed, A., Villa-Galarce, Z., Moreno, M., Tong-Rios, C., Vinetz, J., Lewis, N., and Winzeler, E. A. Dual RNAseq identifies human mucosal immunity protein muc13 as a hallmark of *Plasmodium* exoerythrocytic infection. *Nature Communications 10* (2019). doi: https://doi.org/10.1038/s41467-019-08349-0.

[31] Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., and Carey, V. Software for computing and annotating genomic ranges. *PLoS Computational Biology 9* (2013). doi: https://doi.org/10.1371/journal.pcbi.1003118.

[32] Lee, H. J., Georgiadou, A., Walther, M., Nwakanma, D., Stewart, L. B., Levin, M., Otto, T. D., Conway, D. J., Coin, L. J., and Cunnington, A. J. Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria. *Science Translational Medicine 10*, 447 (2018). doi: https://doi.org/10.1126/scitranslmed.aar3619.

[33] Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. Coexpression analysis of human genes across many microarray data sets. *Genome Research 14* (2004). doi: https://doi.org/10.1101/gr.1910904.

[34] Leinonen, R., Sugawara, H., Shumway, M., and on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Research 39*, supplement1 (2011), D19 – D21. doi: https://doi.org/10.1093/nar/gkq1019.

[35] Liao, B.-Y., and Zhang, J. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Molecular Biology and Evolution 23*, 3 (11 2005), 530–540. doi: https://doi.org/10.1093/molbev/msj054.

[36] Mamedov, M. R., Scholzen, A., Nair, R. V., Cumnock, K., Kenkel, J. A., Oliveira, J. H. M., Trujillo, D. L., Saligrama, N., Zhang, Y., Rubelt, F., Schneider, D. S., hsiu Chien, Y., Sauerwein, R. W., and Davis, M. M. A macrophage colony-stimulating-factor-producing $\gamma\delta$ T cell subset prevents malarial parasitemic recurrence. *Immunity 48*, 2 (2018), 350 – 363.e7. doi: https://doi.org/10.1016/j.immuni.2018.01.009.

[37] Martin, S., Dehler, C., and Krol, E. Transcriptomic responses in the fish intestine. *Developmental Comparative Immunology 64* (03 2016). doi: https://doi.org/10.1016/j.dci.2016.03.014.

[38] McCarthy, J., D., Chen, Yunshun, Smyth, and K., G. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research 40*, 10 (2012), 4288–4297. doi: https://doi.org/10.1093/nar/gks042.

[39] Minot, C.-S. Morphology of the blood corpuscles. *The American Naturalist 24*, 287 (1890), 1020–1023. URL: http://www.jstor.org/stable/2452319.

[40] Phipson, B, S., and G. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology 9*, 1 (2010). doi: https://doi.org/10.2202/1544-6115.1585.

[41] Reid, A. J., and Berriman, M. Genes involved in host–parasite interactions can be revealed by their correlated expression. *Nucleic Acids Research 41*, 3 (2013), 1508–1518. doi: https://doi.org/10.1093/nar/gks1340.

[42] Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology 23* (2005). doi: https://doi.org/10.1038/nbt1103.

[43] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics 26*, 1 (2010), 139–140. doi: https://doi.org/10.1093/bioinformatics/btp616.

[44] Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science 270*, 5235 (1995), 467–470. doi: https://doi.org/10.1126/science.270.5235.467.

[45] Seder, R. A., Chang, L.-J., Enama, M. E., Zephir, K. L., Sarwar, U. N., Gordon, I. J., Holman, L. A., James, E. R., Billingsley, P. F., Gunasekera, A., Richman, A., Chakravarty, S., Manoj, A., Velmurugan, S., Li, M., Ruben, A. J., Li, T., Eappen, A. G., Stafford, R. E., Plummer, S. H., Hendel, C. S., Novik, L., Costner, P. J. M., Mendoza, F. H., Saunders, J. G., Nason, M. C., Richardson, J. H., Murphy, J., Davidson, S. A., Richie, T. L., Sedegah, M., Sutamihardja, A., Fahle, G. A., Lyke, K. E., Laurens, M. B., Roederer, M., Tewari, K., Epstein, J. E., Sim, B. K. L., Ledgerwood, J. E., Graham, B. S., Hoffman, S. L., and The VRC 312 strudy team. Protection against malaria by intravenous immunization with a nonreplicating sporozoite vaccine. *Science 341*, 6152 (2013), 1359–1365. doi: https://doi.org/10.1126/science.1241800.

[46] Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science 302*, 5643 (2003), 249–255. doi: https://doi.org/10.1126/science.1087447.

[47] Stubben, C. *ENAbrowseR: Search and retreive data from the ENA Browser REST URL*, 2016.

[48] Torre, S., Polyak, M. J., Langlais, D., Fodil, N., Kennedy, J. M., Radovanovic, I., Berghout, J., Leiva-Torres, G. A., Krawczyk, C. M., Ilangumaran, S., Mossman, K., Liang, C., Knobeloch, K.-P., Healy, L. M., Antel, J., Arbour, N., Prat, A., Majewski, J., Lathrop, M., Vidal, S. M., and Gros, P. USP15 regulates type I interferon response and is required for pathogenesis of neuroinflammation. *Nature Immunology 18*, 54 (2016). doi: https://doi.org/10.1038/ni.3581.

[49] Tsaparas, P., Mariño-Ramírez, L., Bodenreider, O., Koonin, E. V., and Jordan, I. K. Global similarity and local divergence in human and mouse gene co-expression networks. *BMC Evolutionary Biology 6*, 1 (Sep 2006). doi: https://doi.org/10.1186/1471-2148-6-70.

[50] Westermann, A., A Gorski, S., and Vogel Prof. Dr, J. Dual RNA-Seq of pathogen and host. *Nature reviews. Microbiology 10* (08 2012), 618–30. doi: https://doi.org/10.1038/nrmicro28520.

[51] Westermann, A. J., Barquist, L., and Vogel, J. Resolving host–pathogen interactions by dual RNA-Seq. *PLOS Pathogens 13*, 2 (02 2017), 1–19. doi: https://doi.org/10.1371/journal.ppat.1006033.

[52] Yanai, I., Graur, D., and Ophir, R. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS: A Journal of Integrative Biology 8*, 1 (2004), 15–24. doi: https://doi.org/10.1089/153623104773547462.

[53] Zhu, Y., Stephens, R. M., Meltzer, P. S., and Davis, S. R. SRAdb: query and use public next-generation sequencing data from within R. *BMC bioinformatics 14*, 1 (2013), 19. doi: https://doi.org/10.1186/1471-2105-14-19.

[54] Zuo, Y., Yu, G., Tadesse, M. G., and Ressom, H. W. Reconstructing biological networks using low order partial correlation. In *2013 IEEE International Conference on Bioinformatics and Biomedicine* (Dec 2013), pp. 171–175. doi: https://doi.org/10.1109/BIBM.2013.6732482.

[55] Zuo, Y., Yu, G., Tadesse, M. G., and Ressom, H. W. Biological network inference using low order partial correlation. *Methods 69*, 3 (2014), 266 – 273. doi: https://doi.org/10.1016/j.ymeth.2014.06.010.

Figures and tables

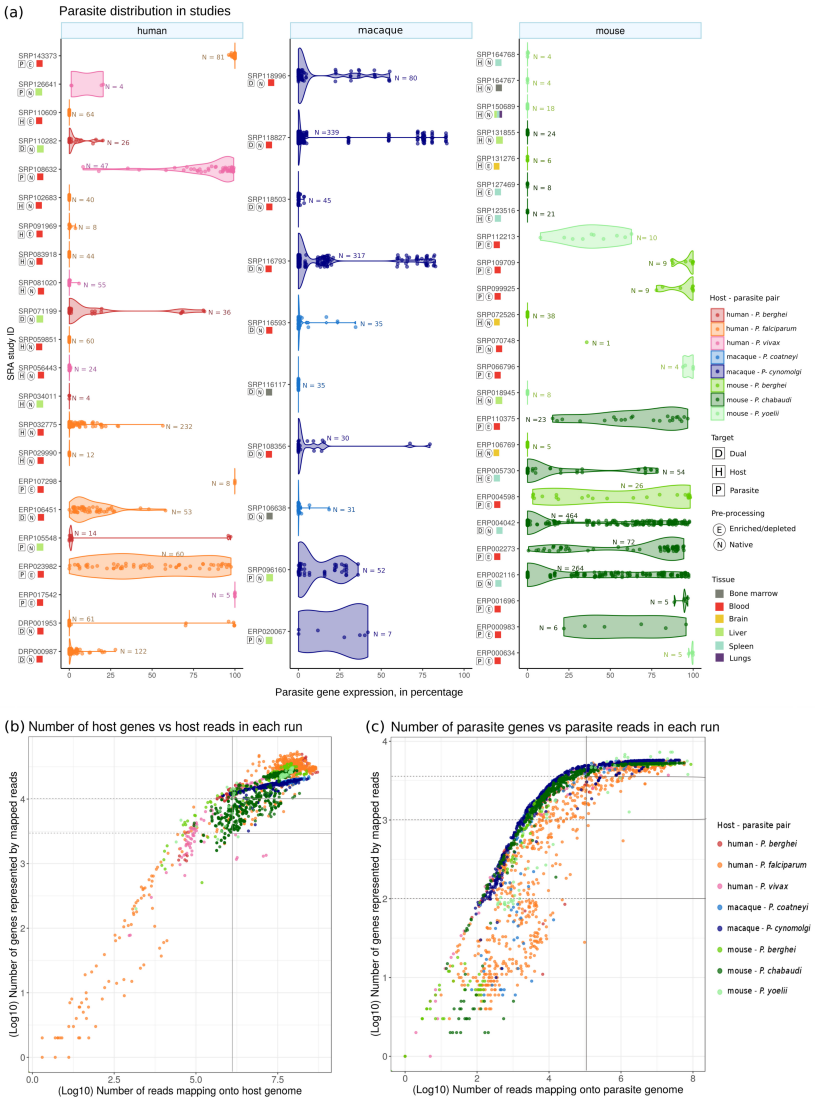

Figure 1: Proportion and number of sequencing reads and expressed genes from parasite and host in selected malaria RNA-Seq studies. We mapped sequencing reads from studies selected for their potential to provide both host and parasite gene expression data (total studies=56, total runs=3064; references in Supplementary File 1) against appropriate host and parasite genomes. (a) The percentage of parasite reads (x-axis) is plotted for each run in each study. The studies are categorised according to the host organisms and labeled "enriched/depleted" to indicate enrichment of infected hepatocytes or depletion of leukocytes from blood. Studies labeled "dual" were originally intended to simultaneously assess host and parasite transcriptomes. The colored area gives an estimation of the density of runs at a particular host-parasite percentage. The number of runs, "N", in each study are displayed. We also plot the number of reads mapped in each run against the number of expressed genes for (b)host and (c)parasite estimated based on this mapping. The number of expressed genes increases with sequencing depth towards the maximum of all annotated genes for the respective organism. The vertical lines indicate a threshold of 1,000,000 and 100,000 reads for host and parasite, respectively. The horizontal lines correspond to thresholds on the number of expressed genes at 10,000 and 3,000 for host and at 3,000, 1,000 and 100 for the parasite. At such exemplary thresholds, data could be considered sufficient for dual RNA-Seq analysis on both organisms.

Table 1: Number of studies for each host-parasite pair and suitability analysis of their runs.

						Run suitability analysis from all studies					
Host	Parasite	# total studies	# total runs	# enriched/ depletion studies	# enriched/ depletion runs	# host genes	# parasite genes	no threshold on read count		prc[a] = 10^5; hrc[b] = 10^6	
								# studies	# runs	# studies	# runs
Human	*P. falciparum*	13	840	5	216	3000	100	11	523	6	270
						3000	1000	11	362	6	268
						10000	3000	7	219	5	207
Human	*P. berghei*	4	77	4	77	3000	100	3	60	2	16
						3000	1000	3	42	2	16
						10000	3000	3	25	2	16
Human	*P. vivax*	5	135	1	5	3000	100	3	39	1	1
						3000	1000	3	36	1	1
						10000	3000	2	4	1	1
Mouse	*P. berghei*	7	128	4	50	3000	100	6	74	4	23
						3000	1000	6	54	4	23
						10000	3000	4	24	4	22
Mouse	*P. chabaudi*	10	926	7	174	3000	100	8	869	7	341
						3000	1000	7	801	7	341
						10000	3000	7	252	6	109
Mouse	*P. yoelii*	7	65	3	19	3000	100	7	34	2	8
						3000	1000	6	21	2	8
						10000	3000	2	7	2	6
Macaque	*P. coatneyi*	3	70	0	0	3000	100	3	74	3	26
						3000	1000	3	53	3	26
						10000	3000	3	33	3	26
Macaque	*P. cynomolgi*	7	849	0	0	3000	100	7	857	7	375
						3000	1000	7	706	7	357
						10000	3000	7	514	7	365

[a]prc: parasite read count
[b]hrc: host read count

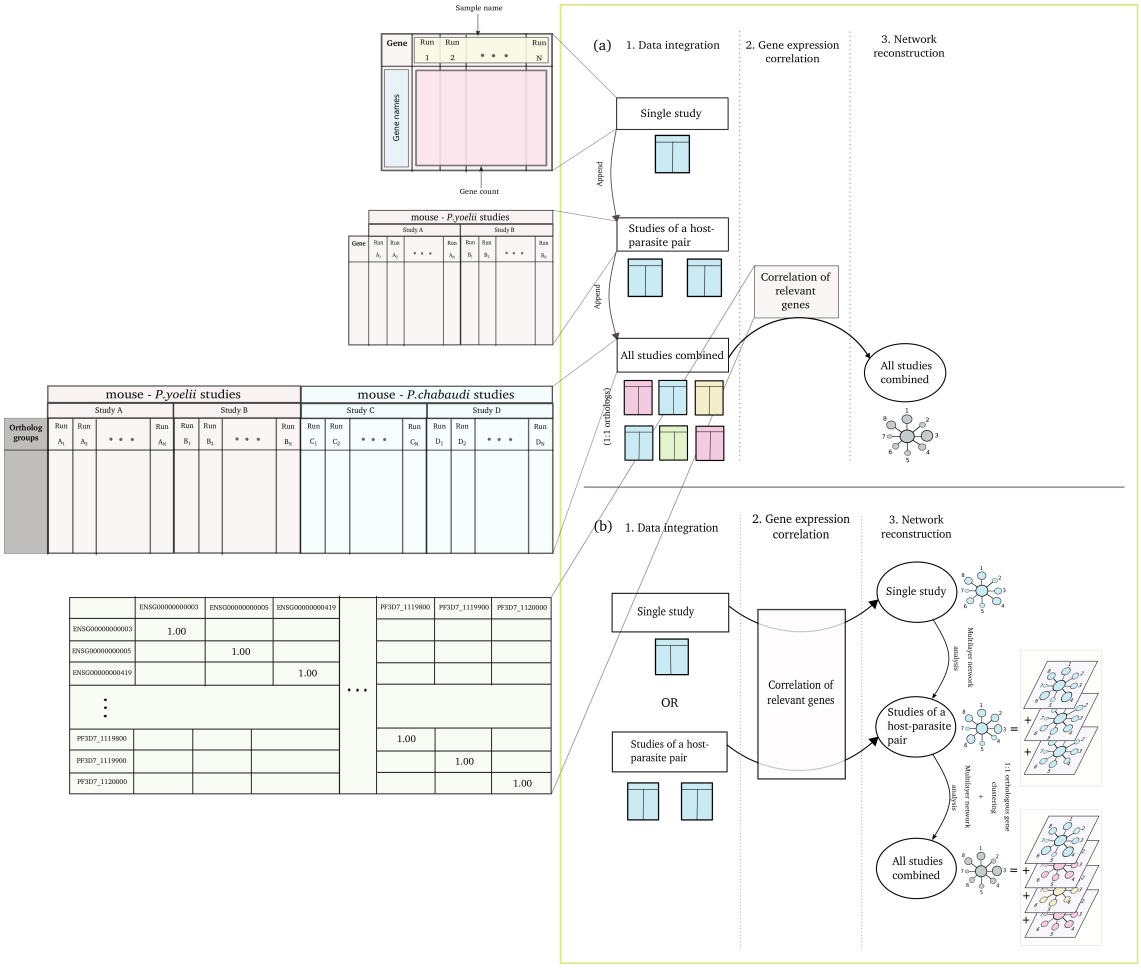

Figure 2: Two strategies identified to reconstruct host-parasite interaction networks from SRA. We identified two approaches to obtain a consensus network involving multiple hosts and multiple parasites. We selected appropriate studies from SRA for this analysis. The aim is to find a set of important interactions in malaria using co-regulated gene expression and visualizing this information as a biological network. Using the first approach (figure (a)), we form single networks from single RNA-Seq datasets or single networks from all studies of a host-parasite system appended one after the other, using cross-species gene correlation. To obtain a consensus network for all hosts and all parasites, we use 1:1 orthologous genes names for all hosts and all parasites, rename these genes to show their equivalency and append them to form one big dataset. Next, we perform pairwise correlation of genes and finally, a network that will represent the direct interactions among orthologous genes. In (b), the second approach, we implement multi-layered network analysis to obtain a consensus network from several layers of individual networks. In this approach, we make single networks for individual RNA-seq datasets. To obtain a network for a host-parasite system, we either append all datasets of the host-parasite system with each other and form a network, or we apply multi-layered network analysis on single networks to get the consensus. To reconstruct a network involving multiple host-parasite systems, we rename orthologous genes in each layer and then look for overlapping communities.