Publicly available transcriptomes provide the opportunity for dual RNA-Seq meta analysis in *Plasmodium* infection

Parnika Mukherjee[1,2,3] , Gaétan Burgio[3] and Emanuel Heitlinger[1,2]

[1]Department of Molecular Parasitology, Humboldt University Berlin, Germany

[2]Research Group Ecology and Evolution of Molecular Parasite-Host Interactions, Leibniz-Institute for Zoo and Wildlife Research (IZW), Berlin, Germany

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

Abstract

Dual RNA-Seq is the simultaneous analysis of host and parasite transcriptomes. It can potentially identify host-parasite interactions by correlated gene expression. Co-expression might highlight interlinked signalling, metabolic or gene regulatory pathways in addition to potentially physically interacting proteins. Numerous studies on malaria focus on one organism – either the host or the parasite – and the other is considered contaminant. Here we assess the applicability of 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 171 malaria studies in *Homo sapiens, Macaca mulatta* and *Mus musculus*, we identified 63 studies with the potential to provide host and parasite data. While 16 studies (1950 total samples) explicitly aimed to generate dual RNA-Seq data, 47 (1398 samples) had an original focus on either the host or the parasite. We show that a total of up to 727 samples from blood and liver studies are suitable for dual RNA-Seq analysis. As a proof-of-principle, we conceive and apply a method for meta-analysis linking host-parasite systems via orthologs. Our approach recovered broad processes known to be interlinked between host and parasites in malaria in addition to individual associations between host and parasite proteins. We suggest these for further experimental investigation.

We argue that the multitude of variations in experimental conditions found in the selected studies should help narrow down a conserved core of cross-species interactions. In the future, detailed analyses building on the datasets and concepts conceived here, conserved sets of core interacting pathways and co-regulated genes across study systems might be identified. This might also provide the opportunity to gauge the applicability of model systems for different pathways in malaria studies.

Introduction

Transcriptomes are often analysed in a first attempt to understand cellular and organismic events. A comprehensive profile of RNA expression can be obtained using high-throughput sequencing of cDNA from reverse transcription of expressed RNA. Such RNA-sequencing (RNA-Seq) provides high technical accuracy at reasonable cost, making it the current method of choice for transcriptomics [26,27].

In an infection experiment, RNA-Seq can assess host and pathogen transcriptomes simultaneously, if RNA from both organisms is contained in a sample. On the one hand, it has been proposed to analyse transcriptomes of both organisms involved in an infection to obtain a more complete understanding of an infectious disease[1]–[3], such as virulence of a pathogen resulting from interlinked processes of both host and pathogen (host-pathogen interactions). This approach is called dual RNA-Seq. Some recent studies on malaria pathogenesis make an intentional use of dual RNA-Seq to study the host and the parasite simultaneously. On the other hand, researchers intending to study one of the two organisms, the target, might consider transcripts from the non-target organism “contamination”. Malaria research is indeed traditionally designed to target one organism, either the host or the parasite. Nevertheless, expression of “contaminant” transcripts potentially corresponds to response to stimuli during the investigation.

In case of malaria, unlike in bacterial or viral infections, both the parasite, *Plasmodium* sp*.*, and the host are eukaryotic organisms with similar transcriptomes [2]. Their messenger RNA have poly-A tails next to the 3’-end. Host and parasite mRNA is selected simultaneously when poly-dT priming is used to amplify polyadenylated mRNA transcripts [2], [3]. This makes most malaria transcriptome datasets potentially suitable for dual RNA-Seq analysis.

In the mammalian intermediate host, *Plasmodium* sp*.* first multiply in the liver and then invade Red Blood Cells (RBCs) for development and asexual expansion. While the nuclear machinery from both host and parasite cells produces mRNA in the liver, RBCs are enucleated and transcriptionally inactive in the mammalian host. In the blood stage of the infection, leukocytes or White Blood Cells (WBCs) provide most of the transcriptomic response and are, thus, the source of host mRNA. Malaria is the most thoroughly investigated disease caused by a eukaryotic organism and the accumulation of these two types 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 types of biological interactions: protein products could be directly involved in the formation of complexes and might therefore be produced at quantities varying similarly under altered conditions. Alternatively, 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. [4], [5], [6]). We (and others before [7][28]) propose to extrapolate this to interactions between the host and parasite. It can be expected that a stimulus presented by the parasite to a host causes host immune response and 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 research article we show that existing raw sequencing read datasets collectively present an unexplored potential to answer questions that have not been investigated by individual studies on malaria. We explore first steps in a comparative meta-analysis of dual RNA-Seq transcriptomes of *Plasmodium* and its hosts. Since rodent and simian malaria are often used as laboratory models for human malaria, we demonstrate the availability and suitability of mRNA sequencing data from three evolutionarily close hosts - *Homo sapiens, Macaca mulatta* and *Mus musculus* - and their respective *Plasmodium* parasite species. We summarise available data and conceive an approach to elucidate host-parasite interactions using orthologs across different host-parasite pairs. We demonstrate that this approach provides meaningful results by cross-validation of information content from networks from different host-parasite systems. We find that these networks highlight prominent known functional determinants of host-parasite interactions in broad categories such as invasion and calcium ion homeostasis for the parasite and host immune response and cadherin mediated cell-cell adhesion on the host side. We also find novel interactions at a finer scale potentially worth further investigation.

Methods

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) [13], EBI European Nucleotide Archive (ENA) [48] and DDBJ Sequence Read Archive (DRA) [49]. We used SRAdb v1.36.0 [11], a Bioconductor/R package[12], to query SRA [13] 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 “RNA-Seq” and with “Plasmodium” in the study title, abstract or sample attributes fields using the function dbGetQuery(). 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. During this search, we disregarded 121 studies, providing data from vectors and non-target hosts (e.g. avian malaria). Following this step, the main criterion for including studies is that the study design must allow it the potential to include both host and *Plasmodium* mRNA. 108 more studies were excluded because their gene expression data was derived from infection experiments of the blood stage in culture media, eg., erythrocytes or RPMI 1640 (Roswell Park Memorial Institute 1640), and thus can be expected to be devoid of nucleated host cells and corresponding mRNA. We then used SRAdb, and the prefetch and fastq-dump functions from SRAtoolkit (v2.8.2), to download all replicate samples (called runs in the databases) of the selected studies. The curation and the download of the studies was performed on 21 January, 2019 and updated on 24 July, 2020.

Mapping and quantification of gene expression

To assess the presence of host and parasite RNA sequencing reads in each study and sample we mapped sequencing reads onto concatenated host and parasite reference genomes using STAR v2.6.0c[19], [20] with default parameters. Host and *Plasmodium* sp. genome assemblies and gene annotation files were downloaded from Ensembl version 43. 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 v3.7[21] and calculated the proportion of reads mapping to host and parasite genes.

Identification of co-expressed genes via correlation techniques

Reid and Berriman [7] recommended using empirical p-values for the analysis of gene co-expression. This allows not only to scrutinise housekeeping genes likely to show almost uniform expression under different experimental conditions but also requires few assumptions about the quality of the input data. One can compute empirical correlation indices, similar to p-values, for any gene-pair. Empirical p-values 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)[23], [24]. We constructed null distributions using permutations of the given data instead of assuming a theoretical null distribution. To obtain p-values corrected for multiple testing, since host and parasite genomes total at nearly 30,000 genes, the number of permutations would have to be around 10^{12} for a resolution of 0.1% FDR. As computational costs for these permutations are too high we used uncorrected p-values to rank genes as initially proposed by Reid and Berriman [7]. In the proof-of-principle presented here we only consider uncorrected p-values of 10^{-5} (in 100,000 permutations) as evidence of co-regulation.

Identification of orthologs

Orthologs are genes in different species that have been passed on from a common ancestral allele by vertical descent. This means that orthologous genes from different species, in general, perform the same function [31]. One of the reasons it is important to study malaria model systems next to human malaria datasets is that therapeutic measures are first tested on the model systems before commencing human clinical trials. To recover core host-parasite interactions, it is therefore important to understand how orthologous genes in the hosts - models and humans - and the parasites interact.

Whole genome based predictions of the proteome for host species were downloaded from Ensembl release 43 and for *Plasmodium* from PlasmoDB release 37. Orthologs were identified using OrthoFinder v2.2.7 [25] wherein blastp [29] results for all-versus-all protein comparisons are clustered using the MCL clustering algorithm [30] (tools bear default parameters for Orthofinder). One-to-one ortholog clusters, or orthogroups, were recovered for hosts and for parasites. Host

clusters were given names starting with “h_OG” and parasite groups with “p_OG”, for example “h_OG000034” and “p_OG002367”.

Network and functional analyses

Pairwise gene correlations performed on matrices with host and parasite gene expression produce three different categories - host-host, parasite-parasite and host-parasite correlations. Host-parasite gene expression correlations are of immediate interest to us as they directly convey interactions between host and parasite.

Bipartite networks/graphs are graphs where the nodes of the graph can be divided into two independent sets, and the nodes from one set connect only with nodes in the other set [52,53]. To visualise host-parasite interactions, bipartite networks were constructed with the R package igraph 1.2.5 [32], where the nodes were the genes and their interaction was represented with an edge. The UpSetR package (v1.4.0) [33] in R was used to visualise intersection sizes of overlapping edges between networks.

For functional analysis of the genes found in host-parasite interactions from co-expression analysis, their Gene Ontology [34] terms were found using the package topGO 2.36.0 and enriched using Kolmogorov-Smirnov test [35]. The GO terms help us get an idea of the characteristics of the genes that have been found and also possible functions and locations of uncharacterised genes (~40% genes in *Plasmodium* spp.). PlasmoDB [36] and Malaria.tools [37] databases were used for looking up gene functions and stage-associated transcript expression.

In general, R (v3.4.3 - 4.0.0)[50] was used for analysis. All code is available at DOI 10.5281/zenodo.3987157 (as used in this manuscript) and Github repository https://github.com/parnika91/CompBio-Dual-RNAseq-Malaria (under further development).

Results

Suitable studies for human, mouse and monkey malaria

Comprehensive query tools to access sequence databases like SRA via web interfaces and programmatically via scriptable languages exist. 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.

In total we found 63 potentially suitable studies (listed in Table S1) querying this database and performing a literature review. The host organism for 27 studies was *Homo sapiens*, for 26, *Mus musculus* and for 10, *Macaca mulatta*. The corresponding infecting parasites were *P. falciparum, P. vivax* and *P. berghei* in human studies (including four artificial infections of human liver cell culture with *P. berghei*), *P. yoelii, P. chabaudi* and *P. berghei* in mouse studies and *P. cynomolgi* and *P. coatneyi* in macaque studies (Table 1). For 16 out of the 63 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 simian malaria. The original focus of the remaining 47 studies was on the parasite in 23 and on the host in 24 cases, but they sequenced their samples without the purposeful removal of one or the other organism. Our collection of studies comprises data derived from blood and liver for all three host organisms human, mouse and macaque. Experiments performed on mouse blood focus on the parasite instead of the host (12 vs. 0). Studies on human blood infection focus more often on the host immune response than on the parasite (10 vs. 6). 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).

We note that 21 of the 63 studies depleted (or enriched, respectively) specific classes of cells from their samples. Seventeen blood-stage studies depleted or enriched host WBCs to focus expression analysis on *Plasmodium* or the host immune system, respectively. Assuming this depletion was imperfect, we tested whether such samples present mRNA from both organisms. The physiologically asymptomatic liver stage [14] is an appealing target for sporozoite-derived vaccines and hence attracts research interest [15-17], but on the other hand, low parasite numbers make it difficult to study *Plasmodium* transcriptomes for this stage. To reduce overwhelming host RNA levels, 10 out of the 14 liver studies sorted infected hepatoma cells from uninfected cells. Three out of the remaining 4 studies were among the ones interested in the host expression during the infection and 1 enriched for host cells.

We postulate 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.

Blood and liver samples from different studies and host parasite systems are suitable for Dual RNA-Seq analysis

Our ability to infer co-regulation of transcripts from host and parasite depends on the detection of the expressed transcriptome of both organisms. A sample must provide sufficient gene expression from both host and parasite to be suitable for dual RNA-Seq co-expression analysis.

Analysing the proportions of sequencing reads mapping to the host and parasite transcriptomes, respectively, we found that the original focus of the study was not always reflected (Fig 1a). In many native samples (not enriched or depleted for host material) the number of host reads was found to be overwhelming, in a few cases, however (probably when parasitemia was very high), parasite transcriptomes were still recovered. Some examples are replicate runs in the studies SRP032775, SRP029990 and ERP106769. Similarly, many studies using depletion or enrichment of a certain cell type prior to RNA sequencing (enriched/depleted in Fig 1a) show considerable expression of the non-target organism ('E' in Fig 1a). Examples are the studies ERP023892 ERP002273 ERP004598 ERP005730 ERP110375 and SRP112213. 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. Although we note that this might come with the caveat that host expression might be biased by unequal depletion of particular cell types.

The number of sequencing reads mapping onto host or parasite transcriptome is a major determinant of the proportion of the organisms transcriptome detected as expressed (transcripts with at least one read mapped; Fig 1b and c). As expected, only runs that were sequenced deeply were able to capture the expression of a high proportion of the transcriptome. For both host and parasite, this plateaus at the total number of genes expressed in the transcriptome of the respective tissue in these organisms. For further analysis, runs with an acceptable proportion of the transcriptome detected as expressed were chosen in our proof-of-principle analysis below. We propose that thresholds based on the proportion of the transcriptome detected as expressed are most suitable for this selection of samples/runs (Table 2a and b).

Suitable blood-stage runs are identified from native (not enriched) human-*P. falciparum* (SRP032775) and human-*P.vivax* (SRP056443) systems and studies with enrichment from mouse-*P.berghei* (ERP004598) and mouse-*P. chabaudi* (ERP110375, ERP002273) systems. Liver runs from human-*P.berghei*, mouse-*P. chabaudi* and monkey-*P.cynomolgi* systems are also identified as suitable. One human liver study (ERP105548) and the monkey liver study (SRP096160) were originally targeted towards parasite biology but contained host gene expression because of the disparity between the amount of host cells and parasite cells in liver stage for bulk RNA-sequencing as discussed before.

Deciding on the stringency of the thresholds applied to a dual RNA-Seq meta-analysis requires additional analysis. We here propose to test different thresholds and to select one that maximises the signal common between studies - the number of edges shared between correlation networks for different studies (Fig 2). We can also conclude from this analysis that the edges shared between different studies within and among host-parasite systems (see below for an exception regarding human *P. falciparum* studies) outnumber random expectations highly significantly ($p<0.001$, Fisher's exact test; between all selected studies), indicating a biological signal shared between datasets. We suggest using the required computational resources for optimizing the overlapping edge sets sizes instead of setting a fixed (e.g. intermediate) threshold and continued analysis with the optimal version (selection of runs) of each study.

Across different studies, across different host-parasite systems

Knowledge of one-to-one orthologs between different hosts and different parasite species can be used in the next steps to integrate across different host-parasite systems. Humans share 18179 1:1 orthologous genes with macaques and 17089 with mice. 13986 genes are 1:1:1 orthologs among the three host species. Similarly, 4005 one-to-one groups of orthologous genes were found among the *Plasmodium* species.

We have identified two different but interlinked workflows to reconstruct a consensus network of expression correlation (Fig 3). A first approach integrates data from different studies of one host-parasite system by simply appending expression profiles of their runs. Expression data across host-parasite systems is combined in orthogroups and correlations of gene expression are computed for all samples simultaneously. We refer to the results of this approach as the "overall"

network below. It has the benefits that it uses and weights all the information sources in one concise process.

Alternatively, we search for consensus networks by comparison of individual networks from different studies. Similar to the approach taken to optimise expression thresholds for the inclusion of runs (Fig 3), we compare overlapping edges (co-expression between two genes) in a multilayer network analysis. Sets of overlapping edges from multiple studies and host-parasite systems offer more control when querying for similar correlation in different layers representing different host-parasite systems. Similarly, correlations from different types of tissues (blood and liver) could be combined as multilayer networks in future work.

As proof-of-principle, we combined networks from two human, two monkey and two mouse studies of blood stage *Plasmodium* infection - the “overall” dataset. The results show a larger network form the combined approach compared to overlaps from the same individual studies (Fig 5). In detail, 590,000 bipartite edges are found only in the "overall" dataset. 13 - 43% of edges from the overall dataset are recovered in each individual dataset (median 18.5%), indicating a substantial contribution from each individual dataset to the overall network. The most dominant study (presenting 43% of the edges in the overall network in it’s own sub-network) was a mouse study [44] in which parasite gene expression was the primary focus. 91% of this study’s bipartite edges are shared with at least one other individual study in addition to the overall network. The second most influential study was a dual RNA-Seq study (on monkey [53]) that shares 89% of its bipartite edges with at least one other study (Fig 4). This not only illustrates the contributions of studies that initially focused on one organism (“with contaminants”) in dual RNA-Seq analysis, but also that information can be pooled across different studies with our approach. Surprisingly, there were no common edges found between the two human studies ([40] on Thai patients and [1] on Gambian patients, both infected with *P. falciparum*), possibly indicating differences in sampling from patients. Alternatively, this might hint towards idiosyncrasies in human blood studies using native samples with high parasitemia. The potential difficulty to apply RNA-Seq meta-analysis to these human samples directly highlights the need to transfer such data across model systems while controlling for concordance with the human system.

We show that correlation between host and parasite transcript expression can highlight known processes important in host-parasite interactions (Fig 5). Biological processes for hosts consistent among almost all networks include “Cell adhesion by cadherin” and “Calcium-dependent cell

adhesion". Cadherins are cell-adhesion proteins that depend on calcium. *Plasmodium* infection causes systemic endothelial activation of host blood vessels when the infected RBCs sequester on the lining of blood vessels. This is accompanied with an increased interaction of endothelial cells with WBCs, as cell adhesion molecules on the vessel lining direct WBC trafficking to infected RBC sequestration sites [45]. Another host biological process commonly found in our interaction networks is the response to oxidative stress combined with response to hydrogen peroxide and nitric oxide mediated signalling. It has been found that infected RBCs produce twice as many free OH- radicals and H2O2 than uninfected RBCs [44][46]. This mechanism is believed to be a defence mechanism to abate the infection, although the exact mechanism of parasite killing is still unclear. Nitric oxide (NO) production by the host is also believed to promote parasite death in malaria [43]. An interesting GO term observed in the results was "Killing by host of symbiont cells". This term included genes that were all involved in defence response towards microbes by invoking immune cells (eg, Neutrophil cytosol factor 1 and Cathepsin G), respiratory burst (Ncf1) and blood coagulation (eg, prothrombin).

Host-parasite interaction networks from all infection systems were consistently enriched for the parasite processes, "Calcium ion homeostasis" and "Late endosome to vacuole transport". Calcium ion homeostasis regulated by calmodulins and Ca2+-dependent kinases (CDPKs) play roles in complex signalling pathways and are important for apicomplexan parasite virulence [41][42]. Vacuolar protein sorting-associated proteins-1, -2 and -46 (VPS1, 2 and 46) along with serine/threonine protein kinase (VPS15) underly the enrichment of "Late endosome to vacuole transport" process and are interlinked with host immune reaction genes in our interaction networks. The endomembrane system is important for the parasite in order to invade the host cell, to establish the parasitophorous vacuole and to obtain nutrients. The Endosomal Sorting Complex Required for Transport (ESCRT) is involved in late endosome formation as endosomes exist at different stages of formation - early endosome, multivesicular bodies, late endosome, before finally fusing with the lysosome or vacuole (reviewed in [47]). The interrelation of the immune system and endosomal formation and sorting events should be a future focus of research including our deeper studies of correlated gene expression networks. (See Table S2 for a list of enriched GO terms)

An interaction of a single gene pair was conserved across 5 individual networks and the "overall" network (Fig 4): expression of perforin-like protein 5 (PLP5) in *Plasmodium* and palmitoyltransferase ZDHHC4 (zinc finger DHHC-type palmitoyltransferase 4) in the hosts is

correlated in all these datasets. PLP5 is a pore-forming protein and is associated with “entry into host cell” (the GO term GO:0030260). It has previously been studied in relation to its function in mosquito mid-gut epithelium and it is believed to not be expressed in the blood stage [reviewed in 38], [39]. However, more recent transcriptomics studies in fact have found PLP5 transcripts expressed in several blood stages, especially at the trophozoite stage (also see PlasmoDB [36] and Malaria.tools [37]). Host palmitoyltransferase function in malaria has - to our knowledge - not yet been studied, but ZDHHC4 is a membrane targeted protein (GO term GO:0006612). We propose that interaction between these two proteins (direct, physical protein-protein interaction) or interlinkage of associated pathways may be worth further scrutiny in mechanistic investigations.

Our proof-of-principle analysis recovers both very broad and well known processes involved in host-parasite interaction as well as a strong interaction signal for a single gene pair. This indicates that finer analysis of interaction networks might uncover novel links between host and parasite processes helping to focus further research into detailed mechanisms of host-parasite interaction. Further analyses building on the methodology and concepts conceived here will involve networks derived from more studies curated here, including those from liver as a tissue in which host and parasite cells interact more directly than in the blood. Finer analyses of these multilayer networks will involve linking of host and parasite network modules and combined functional annotation and enrichment analyses for both organisms. Meta-analysis of co-expression networks can assess whether host-parasite interactions are shared among different studies and treatments, different tissues and different host-parasite systems. This will allow the identification of processes at different scales of molecular pathway organisation: from single gene pairs interacting between host and parasite to broad processes as depicted above.

Moreover, the sharing of pathways among co-expression networks from different host-parasite systems might indicate how easily results on those pathways can be translated between different host-parasite systems. Meta-analyses of correlated gene expression across host-parasite systems can thus provide insights into how easily and for which pathways observations made in malaria models can be translated to human malaria.

Conflict of Interest: The authors declare that there is no conflict of interest.

Author Contribution: EH and PM designed the study. PM obtained and analysed the data. EH and PM wrote the manuscript. All authors contributed to the final version of the manuscript.

Acknowledgement: This research project was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

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. G.B.’s work is supported by the National Health and Medical Research Council of Australia (Grant APP1143008), the Australian Research Council (Grant DP180101494) and the National Natural Science Foundation of China (Grant 81772214)

References:

[1] H. J. Lee *et al.*, “Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria,” *Sci. Transl. Med.*, vol. 10, no. 447, p.eaar3619, Jun. 2018.

[2] A. J. Westermann, S. A. Gorski, and J. Vogel, “Dual RNA-seq of pathogen and host,” *Nat. Rev. Microbiol.*, vol. 10, no. 9, pp. 618–630, Sep. 2012.

[3] A. J. Westermann, L. Barquist, and J. Vogel, “Resolving host–pathogen interactions by dual RNA-seq,” *PLOS Pathog.*, vol. 13, no. 2, p. e1006033, Feb. 2017.

[4] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, “A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules,” *Science (80-.).*, vol. 302, no. 5643, pp. 249 LP – 255, Oct. 2003.

[5] H. Bono and Y. Okazaki, “Functional transcriptomes: comparative analysis of biological pathways and processes in eukaryotes to infer genetic networks among transcripts,” *Curr. Opin. Struct. Biol.*, vol. 12, no. 3, pp. 355–361, Jun. 2002.

[6] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, “Coexpression analysis of human genes across many microarray data sets,” *Genome Res.*, vol. 14, no. 6, pp. 1085–1094, Jun. 2004.

[7] A. J. Reid and M. Berriman, “Genes involved in host–parasite interactions can be revealed by their correlated expression,” *Nucleic Acids Res.*, vol. 41, no. 3, pp. 1508–1518, Dec. 2012.

[8] P. S. Usama Fayyad, Gregory Piatetsky-Shapiro and B. Goethals, “From Data Mining to knowledge Discovery in Databases,” *AI Mag.*, vol. 17, no. 3, 1996.

[9] R. H. Fagard, “Advantages and disadvantages of the meta-analysis approach,” *J. Hypertens. Suppl.*, vol. 14, no. 2, 1996.

[10] A. B. Haidich, “Meta-analysis in medical research,” *Hippokratia*, vol. 14, no. Suppl 1, pp. 29–37, Dec. 2010.

[11] Y. Zhu, R. M. Stephens, P. S. Meltzer, and S. R. Davis, “SRAdb: query and use public next-generation sequencing data from within R,” *BMC Bioinformatics*, vol. 14, no. 1, p. 19, 2013.

[12] W. Huber *et al.*, “Orchestrating high-throughput genomic analysis with Bioconductor,” *Nat. Methods*, vol. 12, no. 2, pp. 115–121, 2015.

[13] R. Leinonen, H. Sugawara, M. Shumway, and on behalf of the I. N. S. D. Collaboration, “The Sequence Read Archive,” *Nucleic Acids Res.*, vol. 39, no. suppl_1, pp. D19–D21, Nov. 2010.

[14] G. M. LaMonte *et al.*, “Dual RNA-seq identifies human mucosal immunity protein Mucin-13 as a hallmark of *Plasmodium* exoerythrocytic infection,” *Nat. Commun.*, vol. 10, no. 1, pp. 1–13, 2019.

[15] M. R. Hollingdale and M. Sedegah, “Development of whole sporozoite malaria vaccines,” *Expert Rev. Vaccines*, vol. 16, no. 1, pp. 45–54, Jan. 2017.

[16] R. A. Seder *et al.*, “Protection Against Malaria by Intravenous Immunization with a Nonreplicating Sporozoite Vaccine,” *Science (80-.).*, vol. 341, no. 6152, pp. 1359 LP – 1365, Sep. 2013.

[17] E. M. Bijker, S. Borrmann, S. H. Kappe, B. Mordmüller, B. K. Sack, and S. M. Khan, “Novel approaches to whole sporozoite vaccination against malaria,” *Vaccine*, vol. 33, no. 52, pp. 7462–7468, Dec. 2015.

[18] B. Franke-Fayard, J. Fonager, A. Braks, S. M. Khan, and C. J. Janse, “Sequestration and Tissue Accumulation of Human Malaria Parasites: Can We Learn Anything from Rodent Models of Malaria?,” *PLOS Pathog.*, vol. 6, no. 9, p. e1001032, Sep. 2010

[19] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Oct. 2012.

[20] A. Dobin and T. R. Gingeras, “Mapping RNA-seq Reads with STAR,” *Curr. Protoc. Bioinforma.*, vol. 51, pp. 11.14.1-11.14.19, Sep. 2015.

[21] M. Lawrence *et al.*, “Software for Computing and Annotating Genomic Ranges,” *PLOS Comput. Biol.*, vol. 9, no. 8, p. e1003118, Aug. 2013.

[22] G. Deschamps-Francoeur, J. Simoneau, and M. S. Scott, “Handling multi-mapped reads in RNA-seq,” *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1569–1576, 2020.

[23] T. A. Knijnenburg, L. F. A. Wessels, M. J. T. Reinders, and I. Shmulevich, “Fewer permutations, more accurate P-values,” *Bioinformatics*, vol. 25, no. 12, pp. 161–168, 2009.

[24] B. Phipson and G. K. Smyth, “Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn,” *Stat. Appl. Genet. Mol. Biol.*, vol. 9, no. 1, pp. 1–12, 2010.

[25] D. M. Emms and S. Kelly, “OrthoFinder: phylogenetic orthology inference for comparative genomics,” *Genome Biol.*, vol. 20, no. 1, p. 238, 2019.

[26] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nat. Rev. | Genet.*, vol. VOLUME 10, no. jANUARy 2009, pp. 57–63, 2009.

[27] R. Stark, M. Grzelak, and J. Hadfield, “RNA sequencing: the teenage years,” *Nat. Rev. Genet.*, vol. 20, no. 11, pp. 631–656, 2019.

[28] B. M. Musungu *et al.*, “A network approach of gene co-expression in the Zea mays/Aspergillus flavus pathosystem to map host/pathogen interaction pathways,” *Front. Genet.*, vol. 7, no. NOV, pp. 1–14, 2016.

[29] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, 1990.

[30] S. van Dongen and C. Abreu-Goodger, “Using MCL to Extract Clusters from Networks BT - Bacterial Molecular Networks: Methods and Protocols,” J. van Helden, A. Toussaint, and D. Thieffry, Eds. New York, NY: Springer New York, 2012, pp. 281–295.

[31] E. V Koonin, “An apology for orthologs - or brave new memes,” *Genome Biol.*, vol. 2, no. 4, p. comment1005.1, 2001.

[32] G. Csardi & T. Nepusz (2006). “The Igraph Software Package for Complex Network Research.” InterJournal 2006, Complex Systems, 1695.

[33] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, “UpSet: Visualization of Intersecting Sets,” *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1983–1992, Dec. 2014.

[34] P. Gaudet, N. Škunca, J. C. Hu, and C. Dessimoz, “Primer on the Gene Ontology,” *Methods Mol. Biol.*, vol. 1446, pp. 25–37, 2017.

[35] A. Alexa and J. Rahnenfuhrer, “Gene set enrichment analysis with topGO,” Jan. 2007.

[36] A. Bahl *et al.*, “PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data,” *Nucleic Acids Res.*, vol. 31, no. 1, pp. 212–215, Jan. 2003.

[37] Q. W. Tan and M. Mutwil, “Malaria.tools—comparative genomic and transcriptomic database for *Plasmodium* species,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D768–D775, Aug. 2019.

[38] A. J. Guerra and V. B. Carruthers, “Structural Features of Apicomplexan Pore-Forming Proteins and Their Roles in Parasite Cell Traversal and Egress,” *Toxins (Basel).*, vol. 9, no. 9, p. 265, Aug. 2017.

[39] J. Tavares, R. Amino, and R. Ménard, “The Role of MACPF Proteins in the Biology of Malaria and Other Apicomplexan Parasites BT - MACPF/CDC Proteins - Agents of Defence, Attack and Invasion,” G. Anderluh and R. Gilbert, Eds. Dordrecht: Springer Netherlands, 2014, pp. 241–253.

[40] J. Yamagishi *et al.*, “Interactive transcriptome analysis of malaria patients and infecting *Plasmodium* falciparum,” *Genome Res.*, vol. 24, no. 9, pp. 1433–1444, Sep. 2014.

[41] C. R. S. Garcia, “Calcium Homeostasis and Signaling in the Blood-stage Malaria Parasite,” *Parasitol. Today*, vol. 15, no. 12, pp. 488–491, Dec. 1999.

[42] S. Lourido and S. N. J. Moreno, “The calcium signaling toolkit of the Apicomplexan parasites Toxoplasma gondii and *Plasmodium* spp,” *Cell Calcium*, vol. 57, no. 3, pp. 186–193, 2015.

[43] S. Percário *et al.*, “Oxidative stress in malaria,” *Int. J. Mol. Sci.*, vol. 13, no. 12, pp. 16346–16372, Dec. 2012.

[44] D. S. Guttery *et al.*, “Genome-wide Functional Analysis of *Plasmodium* Protein Phosphatases Reveals Key Regulators of Parasite Development and Differentiation,” *Cell Host Microbe*, vol. 16, no. 1, pp. 128–140, Jul. 2014.

[45] G. D. Turner *et al.*, “Systemic endothelial activation occurs in both mild and severe malaria. Correlating dermal microvascular endothelial cell phenotype and soluble cell adhesion molecules with disease severity,” *Am. J. Pathol.*, vol. 152, no. 6, pp. 1477–1487, Jun. 1998.

[46] H. Atamna and H. Ginsburg, “Origin of reactive oxygen species in erythrocytes infected with Plasmodium falciparum,” *Mol. Biochem. Parasitol.*, vol. 61, no. 2, pp. 231–241, 1993.

[47] E. Jimenez-Ruiz, J. Morlon-Guyot, W. Daher, and M. Meissner, “Vacuolar protein sorting mechanisms in apicomplexan parasites,” *Mol. Biochem. Parasitol.*, vol. 209, no. 1–2, pp. 18–25, 2016.

[48] R. Leinonen *et al.*, “The European Nucleotide Archive,” *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D28–D31, Jan. 2011.

[49] O. Ogasawara, Y. Kodama, J. Mashima, T. Kosuge, and T. Fujisawa, “DDBJ Database updates and computational infrastructure enhancement,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D45–D50, Jan. 2020.

[50] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[51] G. A. Pavlopoulos, P. I. Kontou, A. Pavlopoulou, C. Bouyioukos, E. Markou, and P. G. Bagos, “Bipartite graphs in systems biology and medicine: a survey of methods and applications,” *Gigascience*, vol. 7, no. 4, Feb. 2018.

[52] J. Li, S. Wang, Z. Chen, and Y. Wang, “A Bipartite Network Module-Based Project to Predict Pathogen–Host Association,” *Front. Genet.*, vol. 10, no. January, pp. 1–9, 2020.

[53] Y. Tang *et al.*, “Metabolic modeling helps interpret transcriptomic changes during malaria,” *Biochim. Biophys. Acta - Mol. Basis Dis.*, vol. 1864, no. 6, Part B, pp. 2329–2340, 2018.

Table 1. Potentially available host-parasite systems, number of studies and runs in different tissues after database querying

	Blood	Spleen	Liver	Brain	Bone marrow	Lungs
Human - *P. falciparum*	14 / 855	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
Human - *P. vivax*	6 / 145	0 / 0	1 / 4	0 / 0	0 / 0	0 / 0
Human - *P. berghei*	0 / 0	0 / 0	6 / 145	0 / 0	0 / 0	0 / 0
Mouse - *P. yoelii*	3 / 19	2 / 10	3 / 20	0 / 0	1 / 4	1 / 6
Mouse - *P. chabaudi*	5 / 204	5 / 795	2 / 46	0 / 0	0 / 0	0 / 0
Mouse - *P. berghei*	3 / 44	0 / 0	0 / 0	3 / 83	0 / 0	0 / 0
Monkey - *P. cynomolgi*	5 / 811	0 / 0	2 / 59	0 / 0	0 / 0	0 / 0
Monkey - *P. coatneyi*	1 / 35	0 / 0	0 / 0	0 / 0	2 / 66	0 / 0

Table 2. Number of studies for each host-parasite pair and suitability analysis of their experimental replicates (runs). a) blood b) liver [* = after applying a quality threshold of a unique read mapping proportion of 70%, yellow: int, green: str]

a) Blood samples

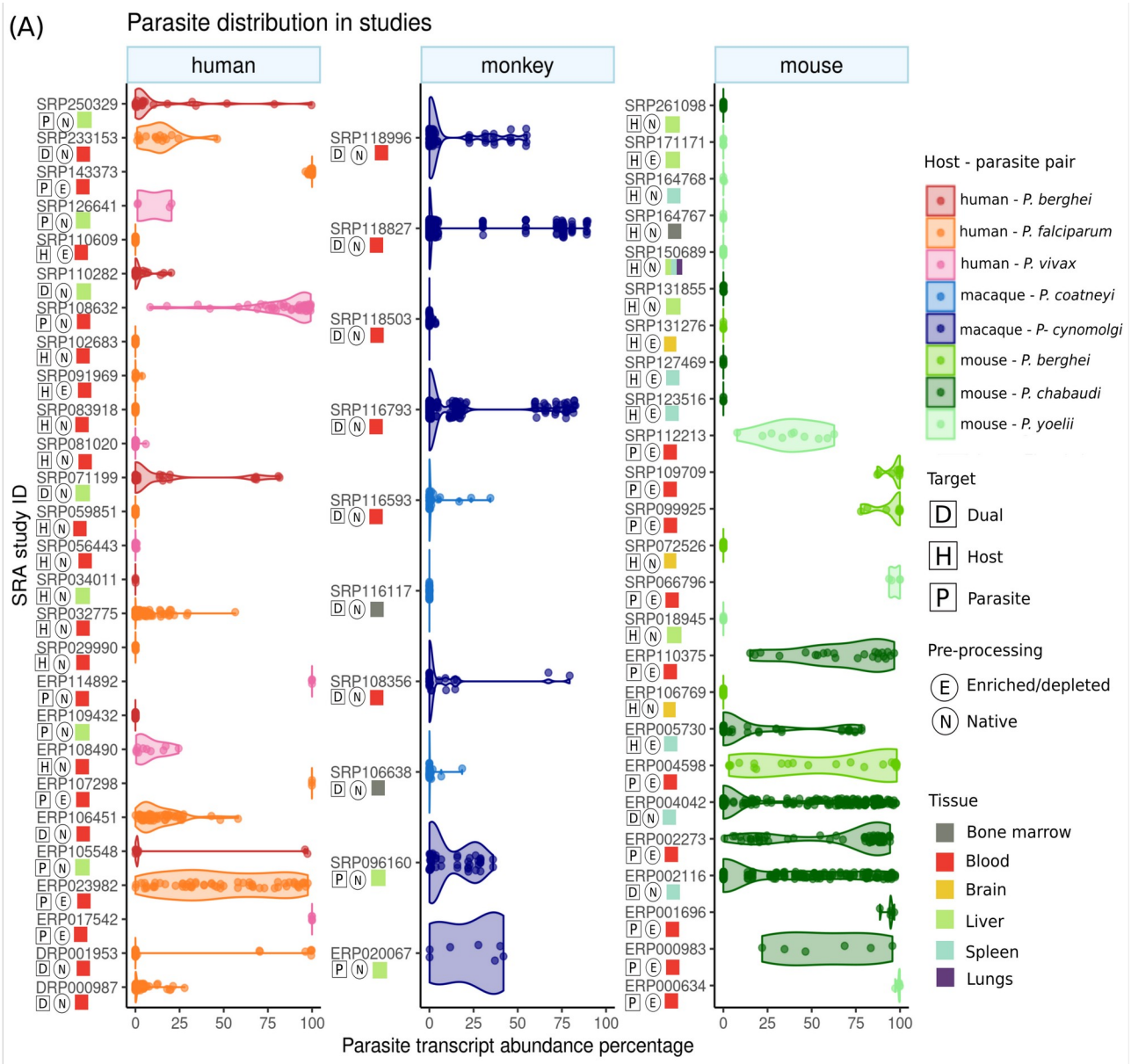
Host-parasite	# studies*	# runs*	Threshold: % transcriptome expressed	# studies with runs meeting threshold	# runs meeting threshold
Human – *P. falciparum*	13	621	30	8	180
			40	8	152
			50	6	131
			60	5	112
			70	4	101
Human – *P. vivax*	6	105	30	3	11
			40	3	10
			50	2	6
			60	2	5
			70	1	2
Monkey – *P. cynomolgi*	5	811	30	5	535
			40	5	500

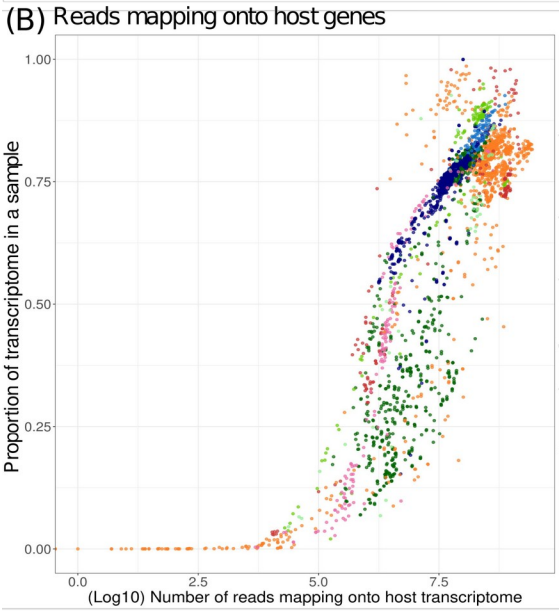
			50	5	478
			60	5	370
			70	5	237
Monkey – *P. coatneyi*	1	34	30	1	21
			40	1	20
			50	1	18
			60	1	16
			70	1	13
Mouse – *P. berghei*	3	44	30	3	28
			40	3	23
			50	3	20
			60	1	13
			70	1	1
Mouse – *P. chabaudi*	4	91	30	4	31
			40	4	20
			50	3	8
			60	1	2
			70	0	0
Mouse – *P. yoelii*	3	14	30	1	5
			40	1	4
			50	1	1

b) Liver samples

Host-parasite	# studies*	# runs*	Threshold: % transcriptome expressed	# studies with runs meeting threshold	# runs meeting threshold
Human – *P. falciparum*	0	0	30	0	0
Human – *P. vivax*	1	2	30	0	0
Human – *P. berghei*	6	125	30	4	51
			40	3	36
			50	3	29

			60	3	28
			70	3	28
Monkey – *P. cynomolgi*	2	48	30	3	45
			40	3	40
			50	2	36
			60	1	26
			70	1	20
Monkey – *P. coatneyi*	0	0	30	0	0
Mouse – *P. chabaudi*	2	46	30	1	5
			40	1	3
			50	0	0
mouse – *P. yoelii*	3	19	30	0	0
			20	1	1
Mouse – *P. berghei*	0	0	30	0	0
Blood and liver:					
All host-parasite pairs	49	1960	50	26	727


(A)
Parasite distribution in studies
human
monkey
mouse
SRA study ID
Parasite transcript abundance percentage
Host - parasite pair
human - P. berghei
human - P. falciparum
human - P. vivax
macaque - P. coatneyi
macaque - P- cynomolgi
mouse - P. berghei
mouse - P. chabaudi
mouse - P. yoelii
Target
D Dual
H Host
P Parasite
Pre-processing
E Enriched/depleted
N Native
Tissue
Bone marrow
Blood
Brain
Liver
Spleen
Lungs


(B) Reads mapping onto host genes
Proportion of transcriptome in a sample
(Log10) Number of reads mapping onto host transcriptome

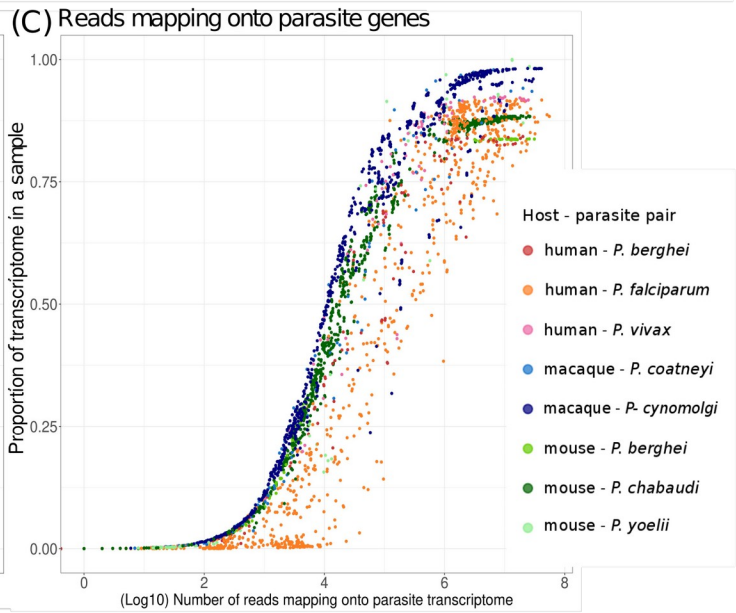

(C) Reads mapping onto parasite genes
Proportion of transcriptome in a sample
(Log10) Number of reads mapping onto parasite transcriptome
Host - parasite pair
human - P. berghei
human - P. falciparum
human - P. vivax
macaque - P. coatneyi
macaque - P- cynomolgi
mouse - P. berghei
mouse - P. chabaudi
mouse - P. yoelii

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=63, total runs=3348) mapped against appropriate host and parasite genomes. (a) The percentage of parasite reads (x-axis) is plotted for runs 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. We also plot the number of reads mapped against the number of expressed genes for host (a) and parasite (b). The proportion of transcriptome detected as expressed increases with sequencing depth towards the maximum of all expressed in the transcriptome.

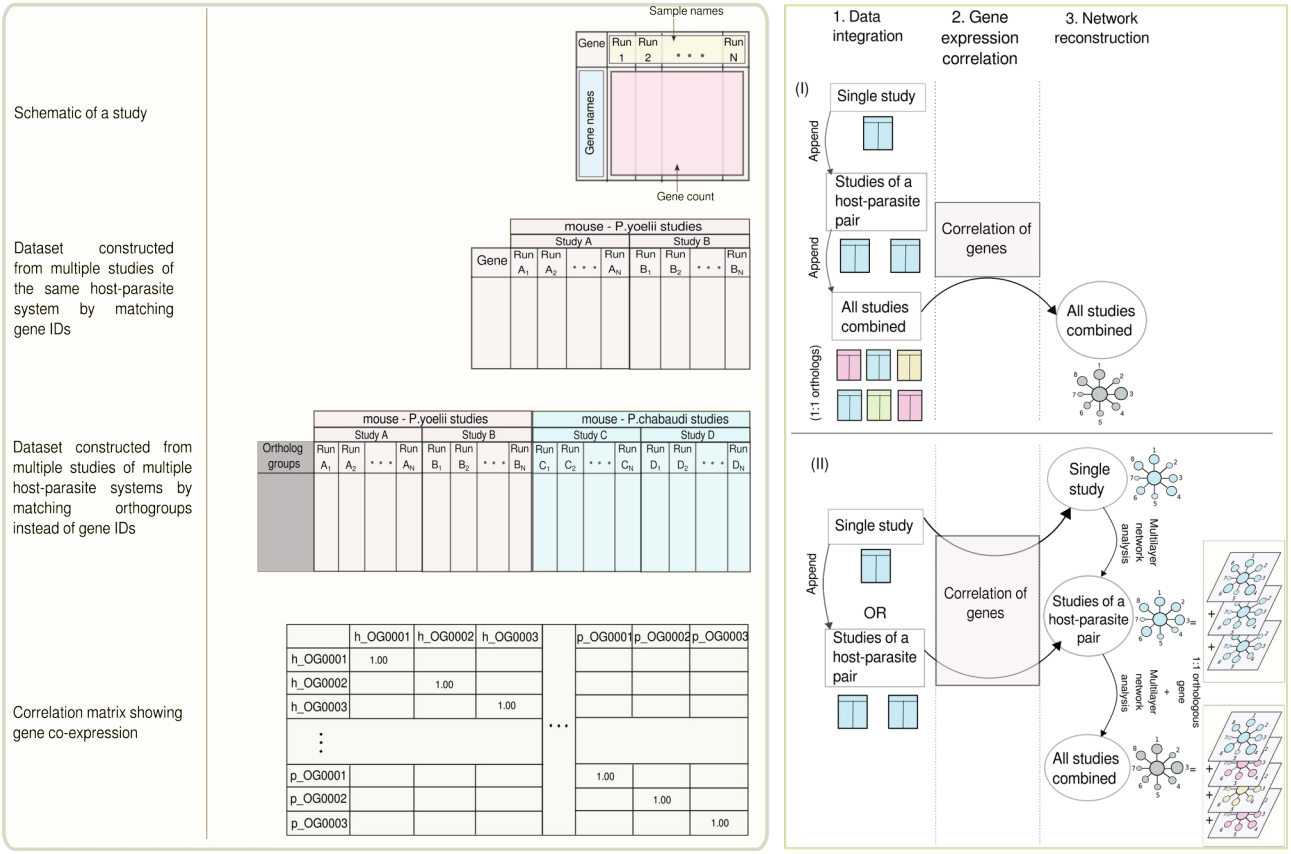


Figure 2.Two alternative strategies to reconstruct host-parasite interaction networks. We envisage two approaches to obtain a consensus network involving multiple hosts-parasite systems. The first approach (a) constructs an appended RNA-Seq dataset before correlation techniques are used to obtain a consensus network. 1:1 orthologous genes for all hosts and all parasites are used in the process. Alternatively (b) correlation techniques are applied to individual datasets or datasets appended for one host-parasite pair. Networks are merged in a multi-layer network analysis to obtain a consensus network. The combination of several layers of individual networks enhances control over how individual studies and parasite-systems change the final consensus network.

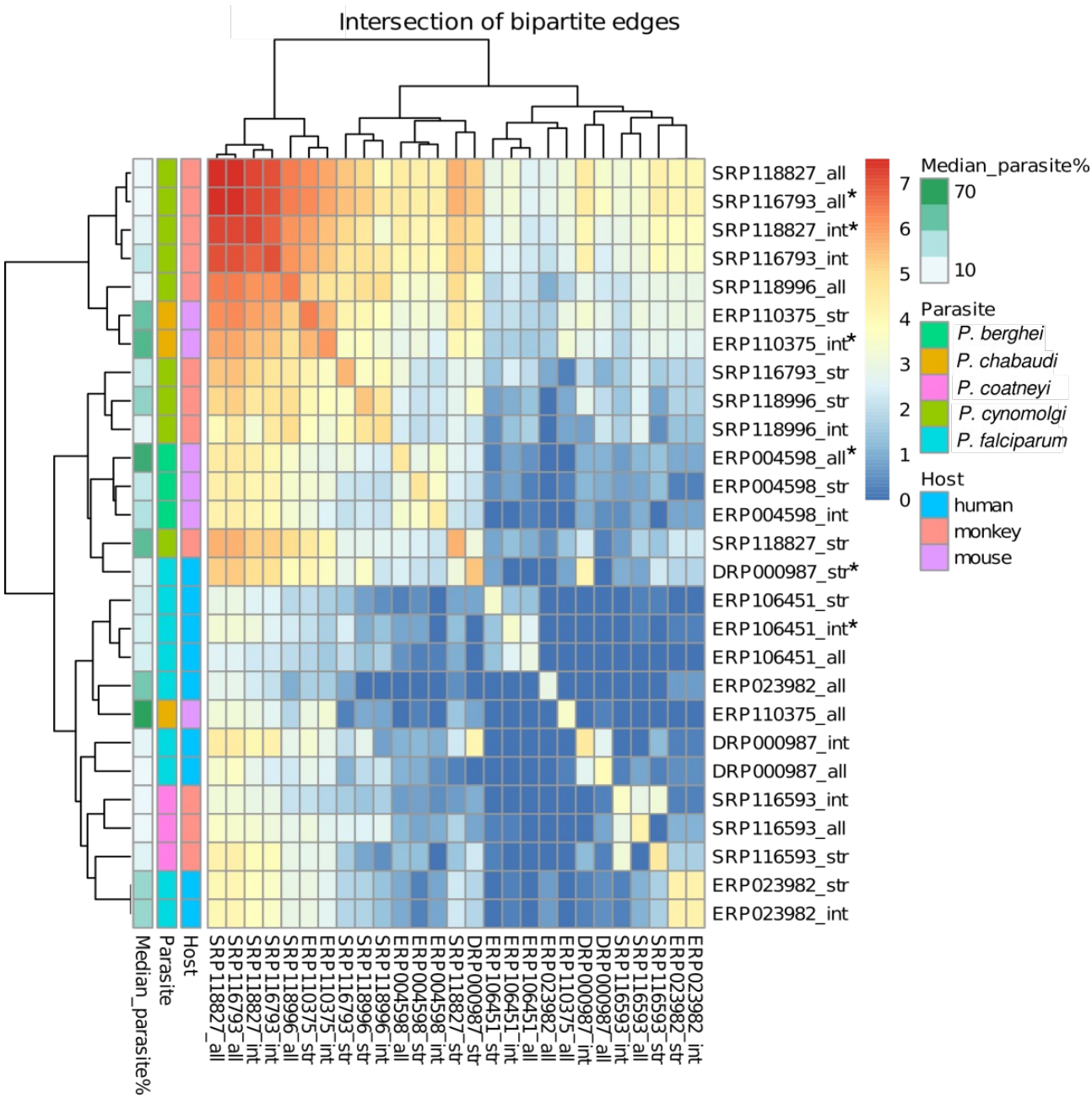


Fig 3. Overlapping bipartite edges between datasets across host-parasite systems

Based on the proportion of transcriptome expressed in each study, three thresholds were implemented on the selection of runs for co-expression analysis. The first, "all", includes all runs. The intermediate ("int") and stringent ("str") thresholds, include runs that express approximately 50% (“int”) and 70% (“str”) of the transcriptome of both host and the parasite, respectively. Nine studies covering most of our host-parasite systems were considered for this experiment and were subsetted according to the aforementioned thresholds, producing a total of 27 datasets. For each dataset, the median parasite percentage was calculated as the median of parasite read proportion (as defined for Fig 1a) in the runs of that dataset. The heatmap shows the log10 transformed number of bipartite edges for each study on the diagonal. The set size of the overlapping bipartite edges between corresponding datasets are displayed in the remaining fields. The datasets are clustered based on their euclidean distance between these set sizes. The size of the overlapping set of bipartite edges is not determined by the median parasite proportion of a dataset - even datasets with low number of reads compared to the host do have a high number of bipartite edges found in common with other datasets. The optimal overlap set size does not suggest a single threshold - all three thresholds are found to have a high number of overlapping edges with other datasets. The versions of the datasets used for further analysis are marked with an asterix.

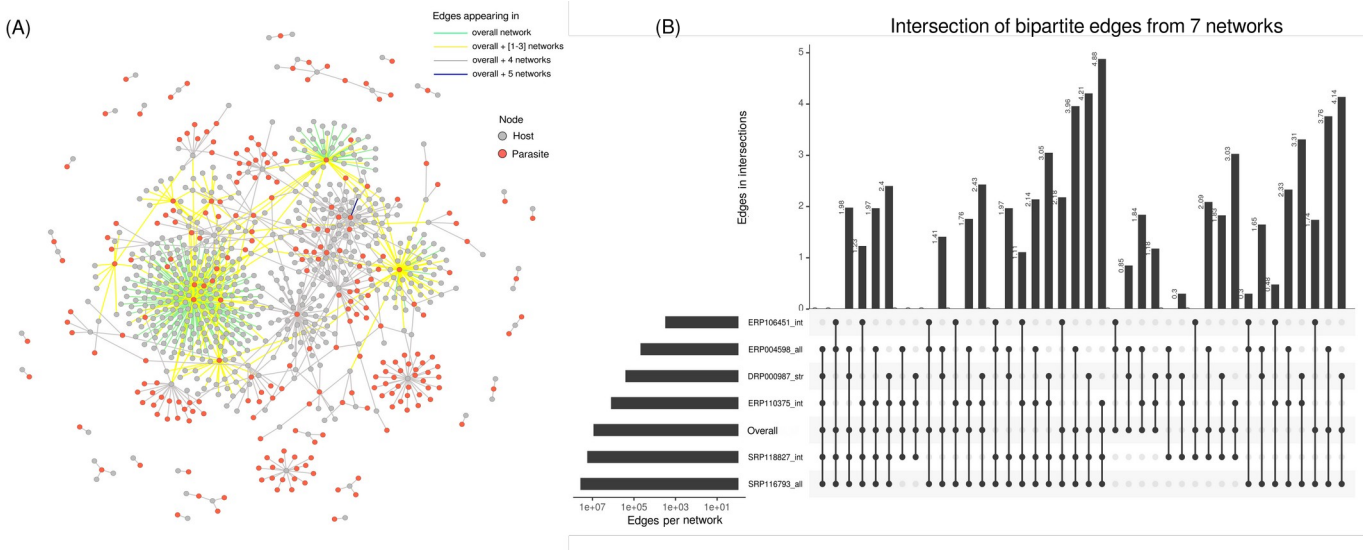


Fig 4. Multi-layer network of host-parasite interactions across host-parasite systems

Following the methodology described in fig 2, we performed pairwise gene correlation tests on individual datasets (see fig 3) and on an "overall" dataset constructed by concatenating the same six datasets (two from each host, maximizing information overlap between sample selection thresholds). Fig 4A displays a random subset of the “overall” network. Starting from a “core edge” found in most networks (overall + 5), the edges were randomly selected in equal numbers, where available, based on how many networks, out of 7, they were found in. The number of networks an edge is found in is represented by the edge colour. This figure shows several edges that were found in common to multiple datasets across host-parasite systems and also edges that were reported only by the consensus "overall" dataset. Fig 4B is an UpSetR plot showing the (log 10 transformed) number of edges in (horizontal bars) and shared (vertical bars) among the 7 networks.

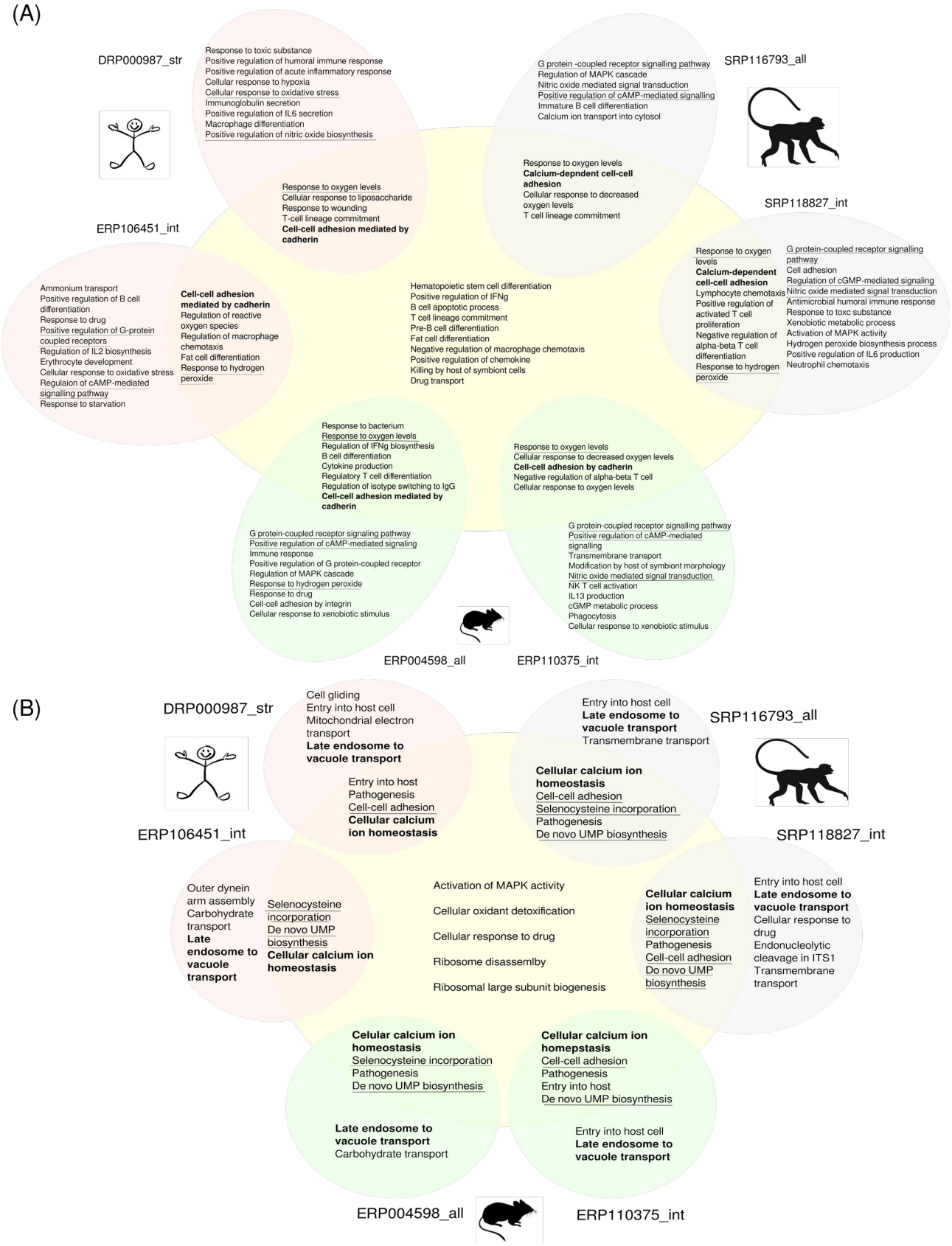


Fig 5. Biological Processes (BP) Gene Ontology (GO) terms shared between datasets across host-parasite systems. We performed GO term enrichment analysis for the "overall" dataset and its 6 constituent datasets (see fig 4) for (A) host and (B) parasite genes. Enriched GO terms (p-value >= 0.05) are presented

in sets coloured in red, green and blue for human, mouse and macaque studies. The GO terms of the "overall" dataset are displayed in the central yellow set. GO terms overlapping between each individual dataset and the "overall" dataset are shown in the respective areas. Set overlaps are not illustrated between the individual datasets but respective terms are underlined (shared between 4-5 studies) or bold (shared between all 6 studies).