Matryoshka RNA virus 1: a novel RNA virus associated with *Plasmodium* parasites in human malaria

Justine Charon[1], Matthew J. Grigg[2,3], John-Sebastian Eden[1,4], Kim A. Piera[2], Timothy William[3,5,6], Karrie Rose[7], Miles P. Davenport[8], Nicholas M. Anstey[2,3] and Edward C. Holmes[1]*

[1]Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, School of Life and Environmental Sciences and Sydney Medical School, The University of Sydney, Sydney, NSW 2006, Australia.

[2]Global and Tropical Health Division, Menzies School of Health Research and Charles Darwin University, Darwin, NT 0810, Australia.

[3]Infectious Disease Society Kota Kinabalu Sabah – Menzies School of Health Research Clinical Research Unit, Kota Kinabalu, Sabah, Malaysia.

[4]Westmead Institute for Medical Research, Centre for Virus Research, Westmead NSW, 2145, Australia.

[5]Clinical Research Centre – Queen Elizabeth Hospital, Kota Kinabalu, Sabah, Malaysia

[6]Gleneagles Hospital, Kota Kinabalu, Sabah, Malaysia.

[7]Australian Registry of Wildlife Health, Taronga Conservation Society Australia, Mosman, NSW 2088, Australia.

[8]Kirby Institute for Infection and Immunity, University of New South Wales, Sydney, NSW 2052, Australia.

*Corresponding author:
Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre,
School of Life and Environmental Sciences and Sydney Medical School,
The University of Sydney,
Sydney, NSW 2006, Australia.
Tel: +61 2 9351 5591
Email: edward.holmes@sydney.edu.au

Abstract

Parasites of the genus *Plasmodium* cause human malaria. Yet nothing is known about the viruses that infect these divergent eukaryotes. We investigated the *Plasmodium* virome by performing a meta-transcriptomic analysis of blood samples from malaria patients infected with *P. vivax*, *P. falciparum* or *P. knowlesi*. This revealed a novel bi-segmented narna-like RNA virus restricted to *P. vivax* and named Matryoshka RNA virus 1 (MaRNAV-1) to reflect its "Russian doll" nature: a virus, infecting a parasite, infecting an animal. MaRNAV-1 was abundant in geographically diverse *P. vivax* from humans and mosquitoes. Notably, a related virus (MaRNAV-2) was identified in Australian birds infected with a *Leucocytozoon* - eukaryotic parasites that group with *Plasmodium* in the Apicomplexa subclass hematozoa. This is the first report of a *Plasmodium* virus. As well as broadening our understanding of the eukaryotic virosphere, the restriction to *P. vivax* may help understand *P. vivax*-specific biology in humans and mosquitoes.

Introduction

Viruses are the most abundant biological entities on Earth, replicating in diverse host organisms (*Forterre 2010*). Although there has been an expansion of metagenomic studies dedicated to exploring this immense virosphere (*Angly et al., 2006; Culley et al., 2006; Desnues et al., 2008; Paez-Espino et al., 2016; Suttle 2005*), our knowledge of the viral universe remains limited, with only a minute fraction of eukaryotic species sampled to date (*Zhang et al., 2018*). This knowledge gap is especially wide in the case of unicellular eukaryotes (i.e. protists), including those responsible for parasitic disease in humans, on which only a small number of studies have been performed.

Viral-like particles in parasites were first observed by electron microscopy as early as the 1960's in various protozoa from the apicomplexan and kinetoplastid phyla (reviewed in *Miles 1988*). The first molecular evidence for the presence of protozoan viruses was obtained in the late 1980s, resulting in the characterization of double-strand (ds) RNA viruses in the human parasites *Giardia*, *Leishmania*, *Trichomonas* and *Cryptosporidium* (*Khramtsov et al., 1997; Miller et al., 1988; Tarr et al., 1988; Wang and Wang 1985; Wang and Wang 1986; Widmer et al., 1989)*. More recently, single-stranded narnavirus-like and bunyavirus-like RNA viruses were identified in trypanosomatid parasites, including *Leptomonas seymouri, Leptomonas moramango, Leptomonas pyrrhocoris* and *Crithidia* sp. (*Akopyants et al., 2016; Grybchuk et al. 2018; Lye et al., 2016; Sukla et al., 2017*). However, our knowledge of protozoan viruses is clearly limited, with many of those identified stemming from fortuitous discovery.

The identification and study of protozoan viruses is also important for our understanding of so-called "Russian doll" ("Matryoshka" in Russian) infections (*Padma 2015*), in which parasites are themselves infected by other microbes. A key question here is whether viruses of parasites can in turn have an impact on aspects of parasite pathogenesis? An increasing number of studies have demonstrated that dsRNA viruses of protozoa can affect key aspects of parasite biology, including their virulence, in a variety of ways (*Gómez-Arreaza et al., 2017*). For instance, data from *Leishmania guyanensis* and *Trichomonas vaginalis* strongly suggest a link between parasite pathogenesis and the presence of Leishmania RNA virus 1 (LRV1) and Trichomonas vaginalis virus, respectively (*Fichorova et al., 2012; Ito et al., 2015; Ives et al., 2011*). By increasing the inflammatory response in the host these viruses could in theory enhance human pathogenesis (*Brettmann et al., 2016; Zangger et al., 2014*). Interestingly, associations have also been observed between LRV1-

infected *L. guyanensis* or *L. braziliensis* and treatment failure in patients with leishmaniasis (*Adaui et al., 2016; Bourreau et al., 2016*).

Viral co-infection also has the potential to alter protozoan biology and/or attenuate the mammalian host response, leading to greater replication or persistent protozoan infection, in turn promoting ongoing parasite transmission. Persistence (i.e. avirulent infection) has been proposed in the case of Cryptosporidium parvum virus 1 (CSpV1) that infects the apicomplexan *Cryptosporidium* (*Nibert et al., 2009*), and increased *C. parvum* fecundity has been demonstrated in isolates experiencing viral co-infection (*Jenkins et al., 2008*). Viral infections may also have a deleterious effect on parasite biology, adversely impacting such traits as growth and adhesion in the case of axenic cultures of *Giardia lamblia* (*Miller et al., 1988*). Clearly, the effects and underlying molecular basis of any consequences that protozoal viruses have on their hosts, including in the context of pathogenesis, requires rigorous investigation. Documenting novel protozoal viruses is an obvious first step in this process.

Nothing is known about those viruses that infect species of *Plasmodium* (order Haemosporida) - obligate apicomplexan parasites of vertebrates and insects. In vertebrate hosts, these protozoa first infect the liver cells as sporozoites where they mature into schizonts. Resulting merozoites are then released into the bloodstream to undergo asexual multiplication in red blood cells. A portion of these replicating asexual forms can differentiate into gametocytes which, following ingestion by blood-feeding female *Anopheles* mosquitoes, develop into sporozoites and are transmitted to another host via mosquito saliva. The genus *Plasmodium* currently comprises approximately 100 species that infect various mammals, birds and reptile hosts. Among these, six species commonly infect humans and are important causative agents of human malaria: *P. falciparum, P. vivax, P. malariae, P. ovale curtisi, P. ovale wallikeri*, and *P. knowlesi*. Despite an early observation of viral-like particles in cytoplasmic vacuoles of simian *P. cynomolgi* sporozoites (*Garnham et al., 1962*), no viruses have been discovered in the parasites responsible for malaria.

With 219 million cases reported in 2017 in 90 countries around the world, malaria continues to be the most important protozoan disease affecting humans (*WHO 2018*). Despite ongoing and considerable global public health efforts, recent progress in reducing the disease burden due to malaria has stalled. Reasons include the emergence of resistance to insecticides in the mosquito vectors, and parasite resistance to antimalarial drugs in

humans. In addition, the large number of asymptomatic and/or submicroscopic *Plasmodium* infections in peripheral blood are an important source of transmission, and pose a major challenge to control and eradication strategies (*Bousema and Drakeley 2011*). This is compounded by the ability of some *Plasmodium sp.*, including *P. vivax, P. ovale curtisi,* and *P. ovale wallikeri,* to form latent liver stages and later relapse. They also illustrate the need for approaches targeting the human parasite reservoir rather than treating only those with clinical disease.

There is an obvious interest in identifying viruses associated with human *Plasmodium* species from both an evolutionary and clinical perspective. The presence of RNA viruses infecting hematozoa parasites have been largely overlooked, although their divergent position in the eukaryotic phylogeny means that they may constitute a valuable source of information to help understand early events in the evolution of eukaryotic RNA viruses. Knowledge of *Plasmodium*-specific viral infection may also provide insights into parasite biology in humans and mosquitoes, with the potential for identifying preventative or therapeutic strategies.

Results

Plasmodium-infected human samples

To investigate the virome of *Plasmodium* parasites that infect and cause disease in humans, we performed a meta-transcriptomic study of three species - *P. vivax* (hereby denoted Pv), *P. knowlesi* (Pk) and *P. falciparum* (Pf). These samples were obtained from 7, 6 and 5 malaria patients, respectively, at different locations in the state of Sabah, east Malaysia (Table S1) (*Grigg et al., 2018*). All patients with malaria had uncomplicated disease. An additional library of 6 non-infected patients were also included as a negative control. All infected blood samples were validated for their corresponding *Plasmodium* species (Table S1). Microscopic parasite counts from peripheral blood films revealed similar densities (i.e. no significant differences, p-value=0.7) between the three *Plasmodium* species, with parasitemia centered around 6000-8000/µL (Figure S1).

Sample processing

Homogenous and equimolar ratios of each of the total RNA samples were used to prepare RNA-Seq libraries. Sequencing depth was similar for all samples, with 17±0.5 million reads obtained (Figure 1.A). The human rRNA read depletion drastically reduced the number of

reads in both the non-infected and Pf data sets (93 and 81% of reads filtered, respectively) (Figure 1.B) and to a lesser extent in Pk and Pv (only 42-57% of reads removed). Pf transcripts were less abundant in libraries than those in Pk and Pv. Finally, the contig assemblies performed on each library depleted for rRNA, human and *Plasmodium* reads were almost equally successful for all libraries, with a similar contig length distribution between the data sets (Figure S2).

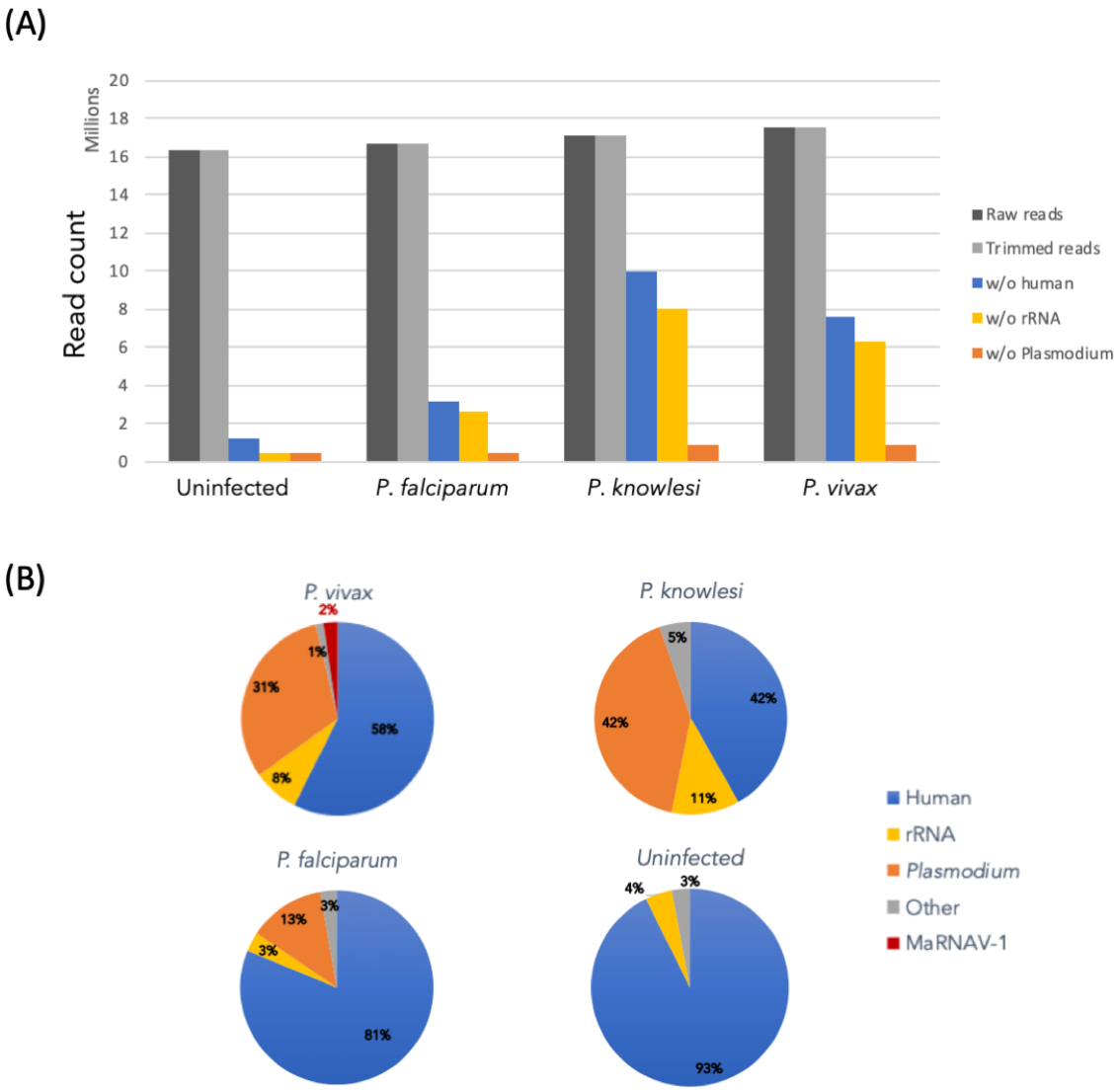

Figure 1. Host read depletion in RNA-Seq libraries. Reads were mapped against rRNA SILVAdb (SortMeRNA tool), the human genome, and the genomes of three *Plasmodium* species. (A) Efficiency of read filtering (rRNA and host sorting). (B) Proportion of major host transcripts in each data set. The number of reads mapping to the human genome, *Plasmodium* sp. genomes and MaRNAV-1 are expressed as the percentage of trimmed reads for each library.

Virus discovery in *Plasmodium vivax*-infected blood samples

Discovery of a bi-segmented RNA Narna-like virus

Ribosomal RNA-depleted data sets were submitted to Blastx against a database containing all the RNA-dependent RNA polymerase (RdRp) protein sequences available at the NCBI. We focused on this protein as it is the mostly highly conserved among RNA viruses and hence constitutes the best marker for detecting their presence and performing expansive phylogenetic analyses. False-positive hits (i.e. non-viral contigs) were discarded by using a second round Blast against the nr database and removing contigs with non-viral top hits. Notably, true-positive RdRp signals were only found in the Pv library (Table 1).

Table 1. Results of the RdRp Blastx analysis.

Contig query	Length	estimated count	Blastn	Blastx best hit	%ID	e-value	taxID	Virus
Pv_1_DN5867_c0_g1_i1	**2924**	**234605.7**	No hit	**YP_009388589.1 RdRp**	43	1.30E-170	2010280	**Wilkie narna-like virus 1**
Pv_1_DN5867_c0_g1_i2	**3023**	**77610.78**	No hit	**YP_009388589.1 RdRp**	43	1.10E-170	2010280	**Wilkie narna-like virus 1**
Pv_1_DN5867_c0_g1_i3	**3023**	**286828.4**	No hit	**YP_009388589.1 RdRp**	43	3.70E-184	2010280	**Wilkie narna-like virus 1**
Pv_1_DN5867_c0_g1_i4	**2141**	**105799.2**	No hit	**YP_009388589.1 RdRp**	43	1.50E-185	2010280	**Wilkie narna-like virus 1**
Pv_1_DN5867_c0_g1_i5	**3023**	**1834.39**	No hit	**YP_009388589.1 RdRp**	43	1.10E-170	2010280	**Wilkie narna-like virus 1**
Pv_1_DN5867_c0_g1_i6	**2045**	**79319.65**	No hit	**YP_009388589.1 RdRp**	43	6.60E-171	2010280	**Wilkie narna-like virus 1**
Pv_1_DN5867_c0_g1_i7	**1496**	**13751.44**	No hit	**YP_009388589.1 RdRp**	42.8	8.30E-171	2010280	**Wilkie narna-like virus 1**

These contigs all correspond to variants of the same gene from the Trinity assembles, and all share their highest sequence similarity score (between 42.6 and 42.9% identity, Table 1) with the RdRp of Wilkie narna-like virus 1, an unclassified virus related to the narnaviruses recently identified in mosquitoes samples (*Shi et al., 2017*). The mapping of Pv-reads against this newly-described viral-like genome revealed that the virus-like contig was highly

abundant in the Pv library, comprising approximately 1.6% of all the non-rRNA (Figure 1). A more detailed characterization of this virus is presented below. To detect homologs to this newly identified RNA-virus-like contig in the other *Plasmodium* species, DN5867 contigs were used as the reference for another round of Blastx: this analysis revealed no matches in either the Pf or Pk data sets.

The apparent bias in virus composition between libraries likely reflects differences in their virome composition rather than experimental bias, since the quality of samples, the depth of sequencing, and the contig assembly were similar among the four libraries. However, it is also possible that it in part reflects the limits of Blastn/Blastx sequence-based homology detection methods to identify highly divergent RNA viruses. To help overcome this limitation, and try to identify any highly divergent RNA viruses, we performed a Blastn/Blastx search on the contigs using the nt and nr databases, respectively. All the "orphan" contigs (i.e. those without any match in any of the nt or nr databases) were sorted depending on their length (Figure S3). Assuming that RNA virus-like contigs would be of a certain minimum length, only those larger than 1000 nt were used for further analysis (Table 2). To identify remote virus signal from these sequences, a second round of Blastx search was conducted with lower levels of stringency: this revealed no clear hits to RNA viruses.

Finally, we employed a structural-homology based approach to virus discovery, utilizing information on protein 3D structure rather than primary amino acid sequence, as the former can be safely assumed to be more conserved than the latter (*Illergård et al., 2009*) and is therefore predicted to be better able to reveal distant evolutionary relationships. Accordingly, hypothetical ORFs were predicted from orphan contigs and Hidden Markov model (HMM) searches combined with 3D-structure modelling were performed on the corresponding amino acid sequences using the Phyre2 web portal (*Kelley et al., 2015*). Again, this revealed no reliable signal for the presence of highly divergent viruses in the RNA sequences obtained here.

Notably, with 122,452 reads mapped to it, the Pv unknown contig retained is highly expressed, possessing a similar level of abundance as the newly-identified Pv RdRp-like contig. Specifically, the abundance of these two contigs were in the same range, with Reads per Kilobase per Million (RPKM) of the Pv RdRp-like and Pv unknown contigs of 2.9 and 2.6, respectively (Table 2). Such a similarity in abundance levels supports the existence of a bi-segmented RNA virus. Finally, the 3kb RdRp-segment described in our *P. vivax*

samples is also within the range of the genome lengths seen in other members of the *Narnaviridae* (2.3 to 3.6 kb).

Narna-like virus genome and protein annotation

The two segments of our putative narnavirus were both validated by RT-PCR in each of the seven *P. vivax* samples used for this study (Figure 2), but were not found in the *P. knowlesi*, *P.falciparum* and non-infected samples. Corresponding amplicons were then Sanger-sequenced to define both the intra and inter-sample sequence diversity. We named this new virus Matryoshka RNA virus 1 (MaRNAV-1) because of its "Russian doll" composition, reflecting a virus that infects a parasite (protist) that infects an animal.

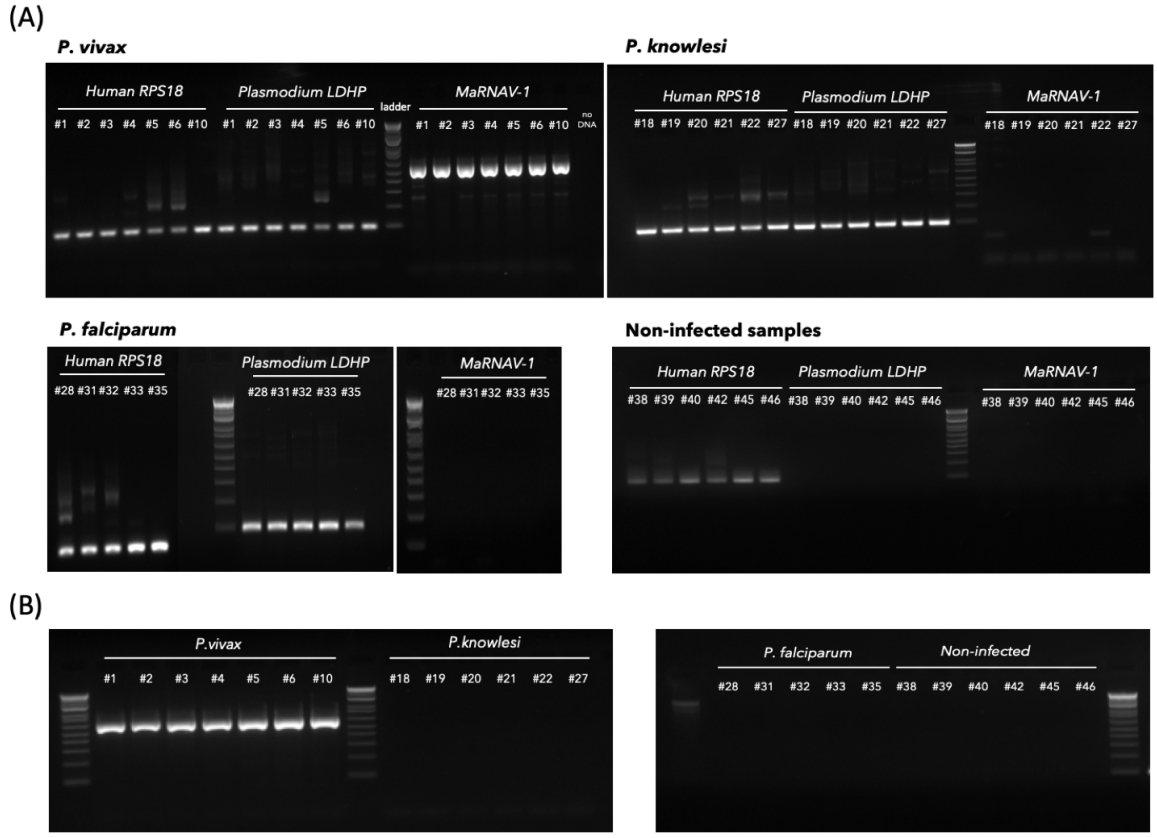

Figure 2. RT-PCR confirmation of host and virus-like sequences in all *Plasmodium*-infected and non-infected samples used in this study. (A) RT-PCR of each samples using human RPS18 primers, *Plasmodium* LDHP primers and MaRNAV-1 primers (segment I). (B) Detection of MaRNAV-1 segment II via RT-PCR.

A very high level of sequence conservation was observed at both the intra and inter-sample scales in the RdRp-encoding segment ("RdRp-segment"; Figure 3A, left). Indeed, very few

nucleotide polymorphisms were observed between those viruses collected from samples 1, 3, 4, 5, 6 and 10 (effectively 100%). Across the data set as a whole, only two polymorphic sites were observed at the intra-sample level and this was restricted to sample #10. In contrast, the MaRNAV-1 from sample #2 was more distant, with 93% identity to the other sequences.

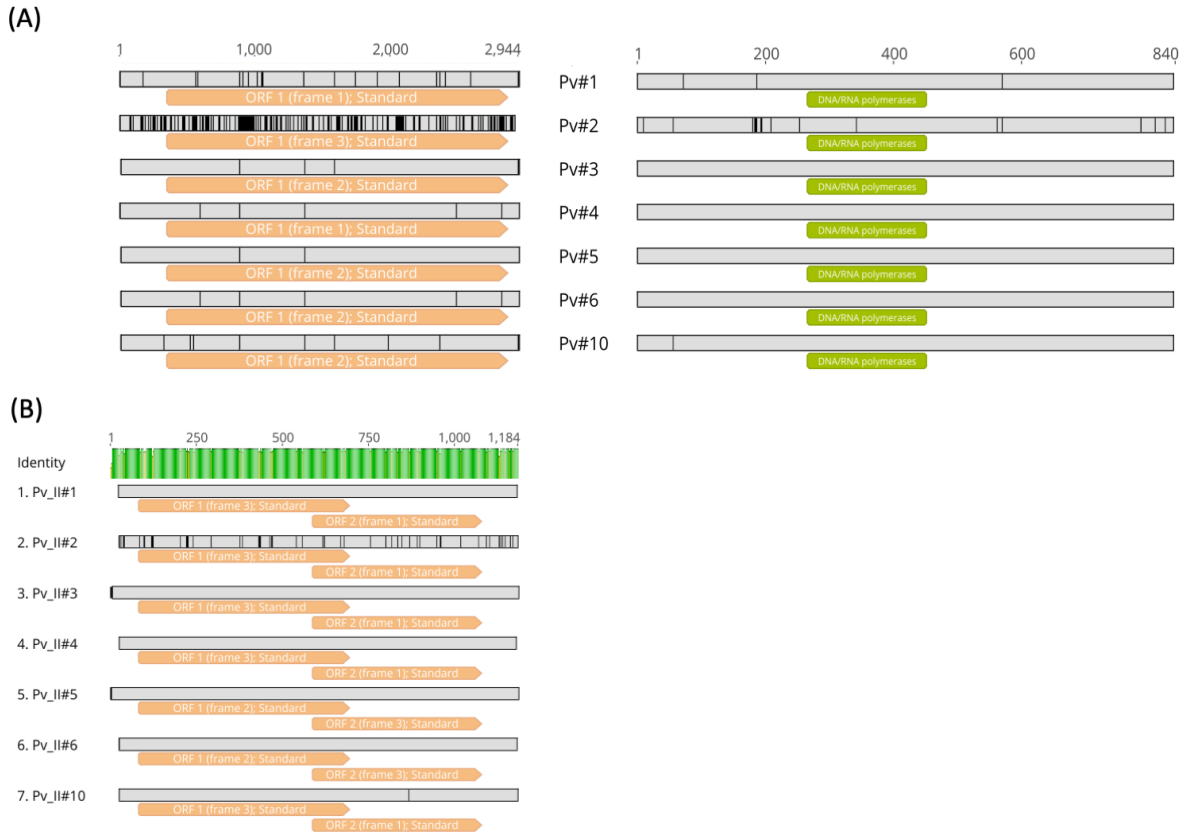

Figure 3. Multiple sequence alignment of MaRNAV-1 from each of the 7 *P. vivax*-infected blood samples. (A) RdRp-segment analysis. Left: Nucleotide sequence alignment and ORF prediction (orange boxes). Right: Protein sequence alignment and InterPro domains predicted (green boxes). Sequence polymorphisms are highlighted in black. (B) Analysis of segment II. Nucleotide alignment and ORFs predicted from the unknown segment in *P. vivax samples* (top). Distance matrix with percentage of identity at the nucleotide level (bottom).

Virus ORFs were predicted using the ORFinder NCBI tool and corresponding amino acid sequences were obtained and aligned (Figure 3A, right). This revealed that the nucleotide polymorphisms described above were also present at the amino acid level, even though these sequences were still highly conserved (98-100%), especially in the RdRp. An

additional attempt at functional annotation was performed, but did not reveal any additional functional motifs or domains aside from the RdRp.

The second segment, the presence of which distinguishes MaRNAV-1 virus from all other narnaviruses, was also highly conserved between *P. vivax* samples (between 95 and 100% identity at the nucleotide level) and is likely to encode two protein products of 205 and 163 amino acids in length through two overlapping ORFs (Figure 3B). Unfortunately, the level of sequence divergence between this second segment and all other sequences available at NCBI meant that no functional annotations were possible.

Phylogenetic analysis of MaRNAV-1

To link the newly-identified *Plasmodium vivax* virus to the diversity of RNA viruses already available, we performed a phylogenetic analysis with the sequence newly acquired here and the closest relatives identified with Blastx (Figure 4). As noted above, this is the first description of a virus infecting *Plasmodium* spp. and few apicomplexan-related dsDNA viruses have been isolated to date. Hence, it is not surprising that only low levels of amino acid sequence similarity (between 15 and 54%) were found in comparisons between MaRNAV-1 and the closest related RNA viruses available at NCBI. Importantly, however, the most closely related viruses were consistently classified as members of the family *Narnaviridae* (genus *Narnavirus*) - a group of single-stranded positive-sense RNA viruses known to infect such host species as fungi, plants and protists (Figure 4). The most closely related virus - Wilkie narna-like virus 1 - was recently identified in a large-scale survey of mosquitoes (*Shi et al., 2017*) and is yet to be formally taxonomically assigned. Although the low abundance of this virus meant that no host could be conclusively assigned, a preliminary study suggested that it was unlikely to be a virus of mosquitoes, such that it could, in theory, infect a protozoan within the mosquitoes. In addition, two of the other narnaviruses most closely related to MaRNAV-1 virus - *Leptomonas seymouri arna-like virus 1* (LepseyNLV1) and *Phytophtora infestans RNA virus 4* - have been described as infecting unicellular eukaryotes, *Leptomonas seymouri* and *Phytophtora infestans*, respectively.

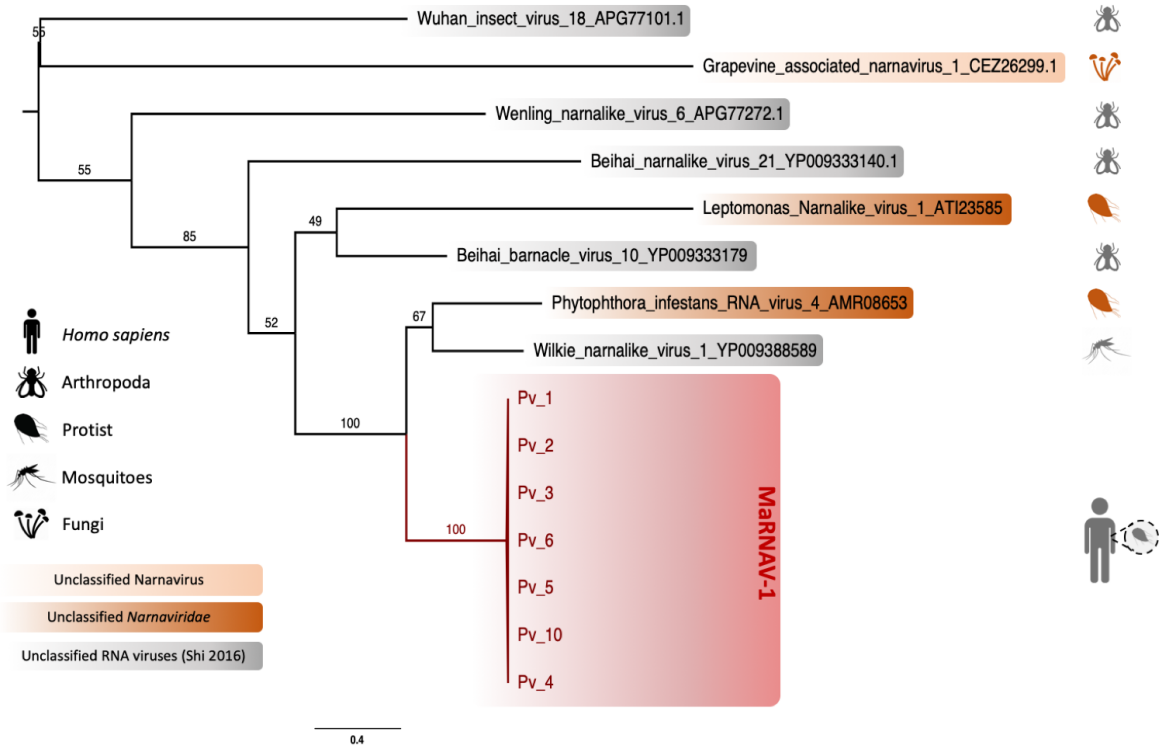

Figure 4. Phylogenetic analysis of a novel narna-like virus - MaRNAV-1 - associated with *Plasmodium vivax*. Boxes refer to the newly-described MaRNAV-1 viral sequences obtained in this study (red) or to RNA viruses classified as members of the *Narnaviridae* (dark orange), genus *Narnavirus* (light orange) or currently unclassified (grey). Taxa corresponding to the validated (coloured icons, right) and non-validated (grey icons, right) hosts are reported on the left part of the tree. Bootstrap values are indicated on each branch. The tree is mid-point rooted for clarity only.

The second putative segment found in all the *P. vivax* samples described here also aligned with the second segment present in *LepseyNLV1* (*Lye et al., 2016*) which similarly encodes two overlapping ORFs (KU935605.1), even though they share little sequence identity (only 14-18% identity at the amino acid level for ORF1 and ORF2, respectively). This high divergence explains why this sequence was not identified in previous Blastx analyses and precluded more detailed phylogenetic analysis.

Virus-host assignment

A major challenge for all metagenomic studies is accurately assigning viruses to hosts as they could in reality be associated with host diet, the environment surrounding the sampling site, or a co-infecting micro-organism. In assigning hosts we assumed: (i) that a virus with a

high abundance is likely to be infecting a host found also in high abundance, (ii) a virus consistently found in association with one particular host is likely to infect that host, (iii) a virus that is phylogenetically related to those previously identified as infecting a particular host taxa is likely to infect a similar range of host taxa, and (iv) a virus and a host that share identical genetic code and/or similar codon usage or dinucleotide compositions are likely to have adapted and co-evolved, indicative of a host-parasite interaction.

Eukaryotic host read profiling

To initially assess whether MaRNAV-1 is likely to infect *Plasmodium* rather than other intra-host microbes and co-infecting parasites, the host taxonomy of the Blastn/x top hits for each contig of the human and *Plasmodium*-depleted *P. vivax* library were compared to their respective size and abundance. However, this analysis revealed only a small number of short contigs associated with fungi and bacteria (Figure S4). This result is of note as the usual hosts associated with narnaviruses are fungi, and the closest relatives have been found in mosquito samples, although the true host of this virus could in theory be protozoal. Among the Metazoa identified, all the contigs were associated with vertebrates, rather than members of the Arthropoda or Nematoda. Hence, *Plasmodium vivax* appears to be the most likely host of the newly-identified MaRNAV-1 virus.

Comparison of codon usage and dinucleotide composition

Host adaptation can result in specific patterns of codon and dinucleotide usage. We compared the codon usage observed in MaRNAV-1 to those of the potential host organisms. The Codon Adaptation Index (CAI) measures the similarity in synonymous codon usage between a gene and a reference set, and was assessed for MaRNAV-1 using *H. sapiens*, *P. vivax* and *Anopheles* genera mosquitoes (pool of 7 species) as reference data sets. However, as none of the CAI/eCAI values obtained were significant (<1) (Figure S5), no conclusions could be drawn regarding the potential host of MaRNAV-1. In a complementary approach, we compared the dinucleotide composition bias between the newly identified virus and the potential hosts (*Di Giallonardo et al., 2017*). Again, the dinucleotide frequencies in the two potential hosts *An. gambiae* and *P. vivax* revealed strong similarities that prevented us from identifying any signature of viral adaptation (Figure S6).

Investigation of the MaRNAV-1 and Plasmodium sp. association using the Sequence Read Archive (SRA)

To further test for an association between MaRNAV-1 and *Plasmodium* parasites, we performed a wider investigation of the occurrence of this virus in *Plasmodium*-infected samples and other *Plasmodium* species for which RNA-Seq data were available on the SRA. These data sets comprised *P. chabaudi, P. cynomolgi, P. falciparum, P. yoelli, P. knowlesi* and *P. berghei* (the relevant Bioprojects are listed in Table S2).

In total, 1682 RNA-Seq data sets from *Plasmodium*-related projects on the SRA were screened for the presence of MaRNAV-1 using Blastx. This analysis identified reads mapping to MaRNAV-1 in 45 libraries, all belonging to *P. vivax* (Figure 5). Among the *P. vivax*-related runs (Table S3), none of the 31 non-infected or *P. falciparum*-infected samples contained MaRNAV-1 (Chi-squared test, p-value = 0, Figure S7.A). This pattern is strongly suggestive of a specific association between MaRNAV-1 and *P. vivax*, rather than the result of experimental bias or contamination introduced during RNA extraction or sequencing. Moreover, MaRNAV-1 was found in 43% (13 out of 30) of the *P. vivax*-infected SRA samples investigated here (biological replicates omitted).

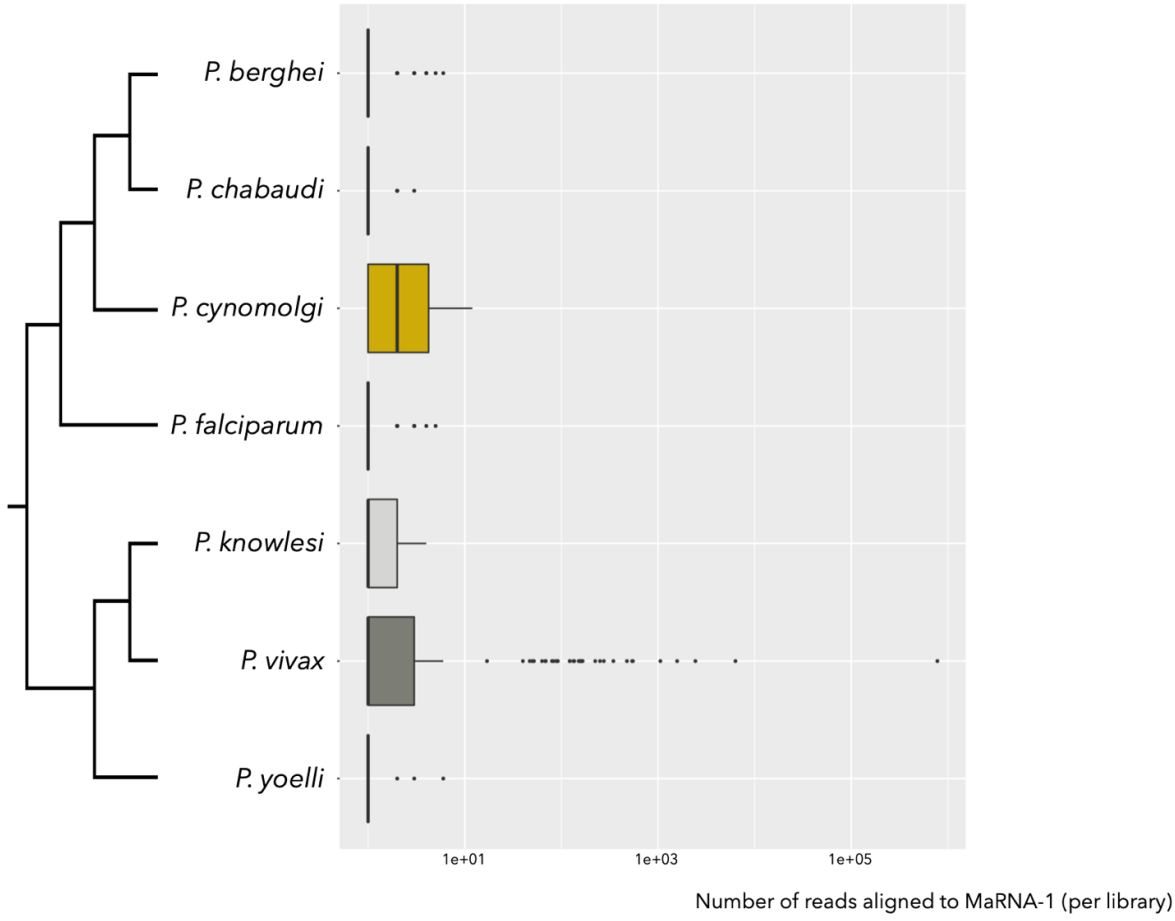

Figure 5. Number of *Plasmodium* SRA reads aligning with the MaRNAV-1 sequence (RdRp-segment) using Blastx (cut-off 1e-5).

Finally, the location of the *P. vivax* isolates from the SRA-based studies (Colombia, Cambodia and Thailand) and the multiple sample types (human blood, mosquito dissected salivary glands or *ex vivo* cultures) were found to be independent of MaRNAV-1 detection (Chi-squared tests, p-values > 0.05, Figure S7.C and D). Hence, together, these results strongly support a specific association between MaRNAV-1 and *P. vivax*.

Analysis of SRA-derived MaRNAV-1 virus-like genomes

Narnavirus positive *P. vivax* data sets were further analyzed following the same workflow as described above. Hence, contigs were *de novo* assembled and re-submitted to Blastx to extract full-length contigs corresponding to MaRNAV-1. The genomes obtained were validated and quantified using read mapping and overlapping contigs were merged to obtain full-length viral genomes.

A phylogenetic analysis of these sequences containing MaRNAV-1 was performed at the nucleotide level (Figure 6). Importantly, phylogenetic position was strongly associated with sampling location rather than the nature of the samples (i.e. human blood or *Anopheles sp.* mosquito salivary glands). This again reinforces the idea that these sequences come from a RNA virus infecting *Plasmodium sp.* rather than human or *Anopheles sp.* hosts. Despite this geographical association, all these newly identified RdRp-encoding sequences shared a high level of sequence nucleotide identity (85-100%). Promisingly, the sequence of the second segment identified in this study is also found in *P. vivax* SRA data sets and is strongly associated (R> 0.95) with the presence of the RdRP-encoding segment (Figure S8).

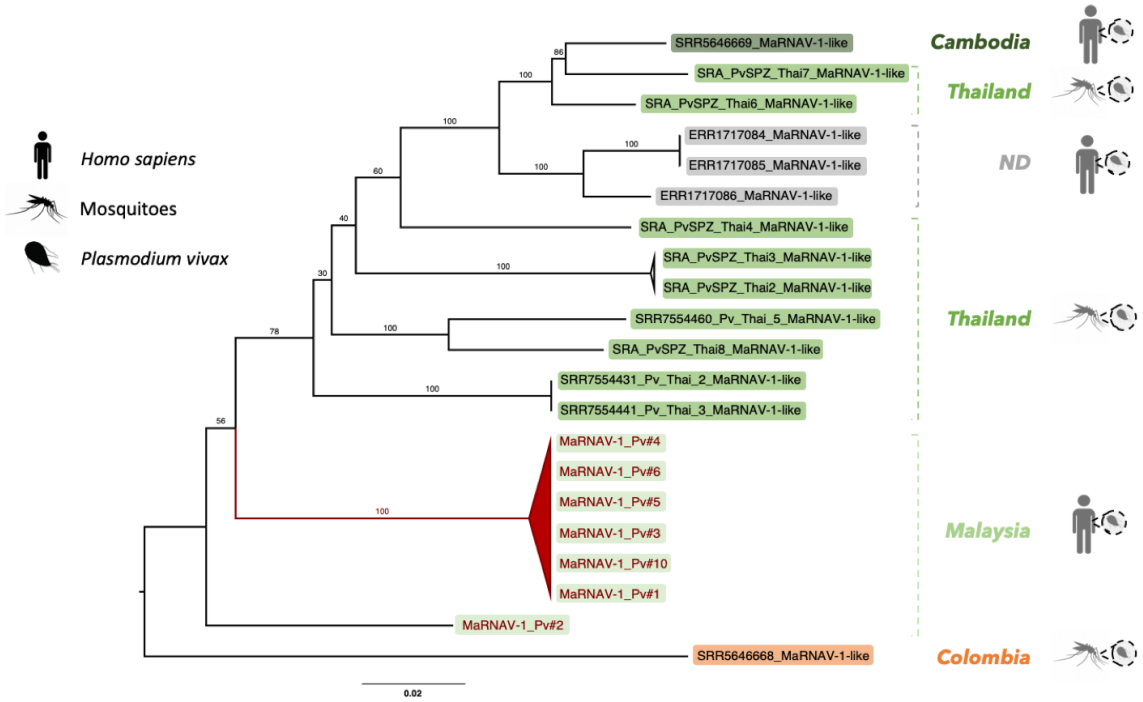

Figure 6. Phylogenetic analysis, based on the RdRp, of the MaRNAV-1 documented here and from the *P. vivax* sequences available on the SRA. Those viruses obtained in this study are shown in red while those from the SRA are shown in black. Sampling location and host characteristics (i.e. human-infected or mosquito-infected samples) are indicated on the right. Colored boxes indicate the samples collected in Asia (green), in South America (orange) or from unknown location (grey; ND : non-determined). The tree is mid-point rooted for clarity only.

Detection of MaRNAV-2 in *Leucocytozoon* parasites infecting birds

To investigate whether these narna-like sequences might infect a wider taxonomic distribution of hosts, we performed a complementary analysis of bird samples infected by members of the genus *Leucocytozoon*: apicomplexan parasites that belong to the same hematozoa subclass (of the Apicomplexa) as *Plasmodium*. These complementary studies were conducted on available RNA-Seq data previously obtained from liver, brain, heart and kidney tissues from Australian Magpie, Pied currawong and Raven birds collected at various time points in New South Wales, Australia (Table S4). Using the newly-discovered MaRNAV-1 viral segments as references, a Blastx analysis was performed on RNA-Seq data previously obtained for these samples. A first segment encoding a single predicted ORF of 859 amino acid long and containing the RdRp domain motif was retrieved and compared to the *P. vivax* MaRNAV-1 sequences (Figure S9A). A relatively high level of sequence similarity (73% identity at the amino acid level) was observed between the

Leucocytozoon-infected birds and the viral sequences found in the *P. vivax* infected-humans. A second segment was also extracted from these avian libraries that exhibited strong similarities in terms of length, genome organization and sequence identity with the prediction of two overlapping ORFs, denoted ORF1 and ORF2, that encode proteins of 246 and 198 amino acids, respectively, and that share 48-52% amino acid identity with the MaRNAV-1 segment II ORFs (Figure S9.B). The relative abundance of the *Leucocytozoon* and MaRNAV-1 like transcripts were assessed and showed a strong correlation in all the five RNA-Seq libraries (Figure 7).

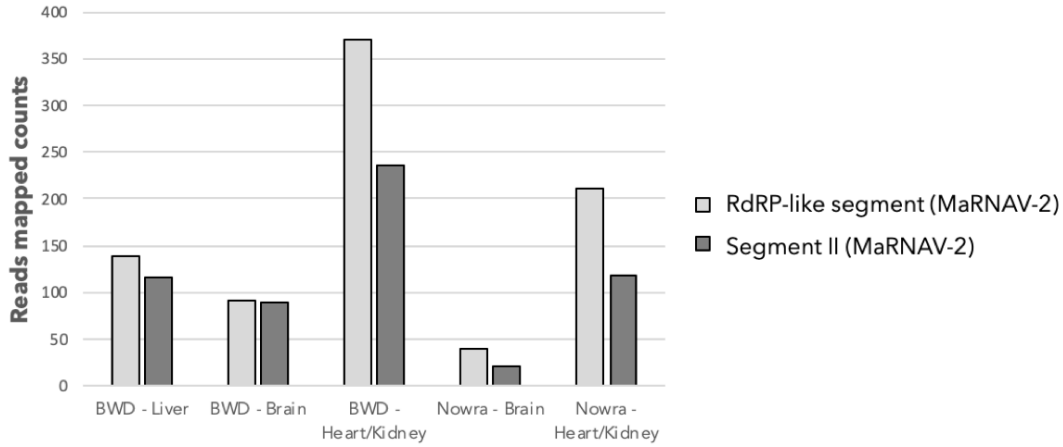

Figure 7. Comparative abundance of MaRNAV-2 in *Leucocytozoon*-infected avian RNA-Seq libraries.

We also explored the association between the presence of the *Leucocytozoon* parasites and the MaRNAV-1 virus homologs by performing RT-PCR on each individual sample previously used to perform RNA-Seq. First, the two viral segments were always found as both-present or both-absent for all of the 12 avian samples analyzed (Table S4, Figure S10). In addition, in the majority of samples (25 of 27), the presence of the viral segments was directly linked to the presence of the parasite: that is, the virus was present only when the parasite was detected and absent in parasite-free samples (Table S4). This supports the idea that the viral sequences screened are infecting the *Leucocytozoon* parasite rather than the bird carrying it. Because of its similarity to *P. vivax* MaRNAV-1, we term this *Leucocytozoon* parasite MaRNAV-2.

MaRNAV-2 sequences were highly prevalent and detected in 6 of the 7 individual birds carrying the *Leucocytozoon*, independently of the tissue, date of sampling or bird species

collected (Table S4). Interestingly, the only *Leucocytozoon* parasites free of MaRNAV-2 (sample 9585.3 collected from an Australian Magpie) may belong to a different *Leucocytozoon* species as it is phylogenetically distinct from the cluster formed by all the other *Leucocytozoon* populations in an analysis of the cytB gene (Figure S11.A).

A phylogenetic analysis of the *Leucocytozoon* MaRNAV-2 amino acid sequences revealed a strong clustering of the RdRp-segment with the *P. vivax*-infecting MaRNAV-1 viruses (Figure 8). Together, these *Plasmodium* and *Leucocytozoon*-associated viruses appear to belong to a newly-described viral clade infecting haematozoa parasites. In addition, for both segment I (Figure S11.B) and segment II (Figure S11.C), the viral sequence variability between samples reflects the bird species rather than the location or the date of sampling. Interestingly, the overall level of divergence is similar between the two putative segments (between 86 and 100% identical nucleotide sites).

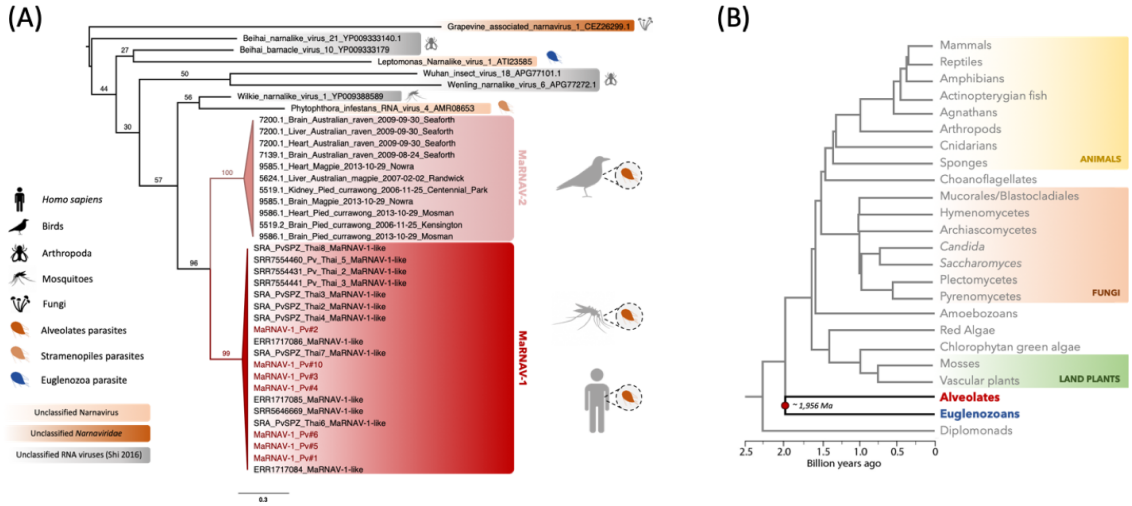

Figure 8. Evolutionary relationships of the newly-identified hematozoa viral sequences (MaRNAV-1 and MaRNAV-2). (A) Phylogeny of all the newly-identified viral sequences. Red box: *P. vivax* viruses MaRNAV-1 (human or mosquitoes infection stage). Pink box: *Leucocytozoon sp.* MaRNAV-2 (bird infection stage). MaRNAV-1 viruses identified from *P. vivax* samples from this study are highlighted in red. Putative protozoan hosts are coloured depending on their belonging to the Alveolates (orange dark), Stramenopiles (light orange) and Euglenozoa (blue) major eukaryotic groups. Numbers indicate the branch support from 1000 bootstrap replicates. The virus tree is mid-point rooted for clarity only. (B) Eukaryotic host evolution and timescale, adapted from (*Hedges et al., 2004*). The two major groups

Alveolates (red) and Euglenozoa (blue) are basal and their separation potentially occurred approximately two billion years ago (*Hedges et al., 2004*).

Discussion

Our meta-transcriptomic study of human blood samples infected with three major *Plasmodium* species revealed the presence of a highly abundant and geographically dispersed bi-segmented RNA virus in *P. vivax* that we named Matryoshka RNA virus 1 (MaRNAV-1). To the best of our knowledge this is the first documented RNA virus in the genus *Plasmodium,* and more broadly in parasites of the Apicomplexa subclass hematozoa. An additional investigation of complementary data sets from the SRA similarly provided strong evidence for the presence of MaRNAV-1 in *P. vivax*, but not in any of the other *Plasmodium* species analyzed. Notably, MaRNAV-1 is both highly conserved and at high prevalence in *P. vivax* populations from both South East Asia and South America. Finally, we documented closely related-viral sequences - MaRNAV-2 - in avian samples infected with *Plasmodium*-related *Leucocytozoon* parasites. The divergent nature of the *Plasmodium* and *Leucocytozoon* viruses identified here, including the unique presence of a second genome segment, raises the possibility that they should be classified as a new genus or family, although this may require additional sampling.

The first segment of MaRNAV-1 encodes a single ORF containing the conserved RdRp-motif that is related to those found in the *Narnaviridae*, while the second segment, which is not characteristic of narnaviruses, encodes two overlapping ORFs of unknown function. The family *Narnaviridae* comprises a capsid-less viral family that infects plants, fungi and protists. Interestingly, no sequences associated with fungi were observed in our samples, suggesting that this virus is indeed likely to infect *Plasmodium*. In addition, the closest RNA virus homologs were also observed in protozoans, or in arthropods that could conceivably be infected by protozoan parasites (*Shi et al., 2016; Shi et al., 2017*). Such a strong virus-protist association evidence was reinforced by the consistent link between this virus and *P. vivax*, rather than to the metazoan hosts (mosquitoes and human) from which the samples were extracted, or the other *Plasmodium* species.

The evolutionary origin of these novel *Plasmodium* and *Leucocytozoon* viruses is less clear, but can be framed as two hypotheses: (i) an ancient virus-host co-divergence between the Euglenozoa (e.g. *Leptomonas*) and Alveolate (including the hematozoa) groups of eukaryotes at almost two billion years ago (*Hedges et al., 2004*) (Figure 9A), or (ii) horizontal

virus transfer events between either a secondary (likely invertebrate) host and protozoan parasites, or among two protozoan parasites co-infecting the same secondary host, over an unknown time-scale (Figure 9B). Given the very high rates of evolutionary change in RNA viruses (*Duffy et al., 2008*), the recognizable sequence similarity between the narna-like RNA viruses from *Plasmodium* (Alveolates, Apicomplexa) and *Leptomonas* (Euglenozoa, kinetoplastids) suggests that they are unlikely to have arisen from a common ancestor that lived approximately two billion years ago (Figure 8.B, Figure 9). Some Euglenozoa and Alveolates independently evolved a parasitic lifestyle by infecting invertebrates and, more recently, vertebrate hosts. Hence, it is more likely that the protozoan narnavirus-like similarities reflects viral cross-species transmission between two protozoan parasites during mixed-infection in either vertebrate or invertebrate hosts (Figure 9B). The wide distribution and prevalence of MaRNAV-1 in *P. vivax* populations, as well as in the different species of *Leishmania* parasites investigated previously, supports the idea that this host jumping event is relatively ancient, although the exact time-scale is difficult to determine. As previously demonstrated, invertebrates play a key role in RNA virus evolution by feeding on many different hosts and transmitting viruses, fungi and protozoa among both plants and vertebrates (*Li et al., 2015; Shi et al., 2016*). This may also explain why narnaviruses or closely related RNA virus have been able to spread to such a diverse range of eukaryotes, including Fungi, Stramenopiles, Alveolates and Euglenozoa. Moreover, the recent characterization of a narnavirus in the plant-infecting trypanosomatid *Phytomonas serpens* (*Akopyants et al., 2016*) suggests that vertebrates are not likely to be the hosts where the horizontal virus transfer occurred.

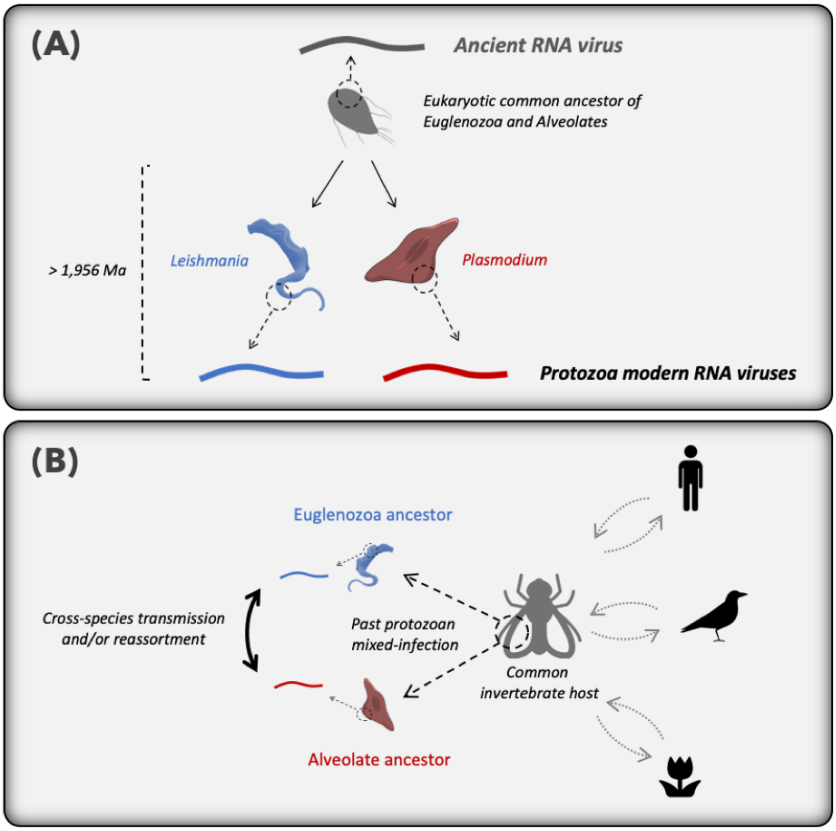

Figure 9. Hypothetical scenarios for the origin and evolution of MaRNAV-1 and MaRNAV-2 and relatives among parasites belonging to the Alveolates and Euglenozoa.

The RdRp segment of MaRNAV-1 documented here is clearly related to the narnavirus RdRp, although we are unable to identify a clear homolog for the second, divergent segment. Hence, as previously hypothesized for the tri-segmented plant RNA virus ourmiaviruses that combine a *Narnaviridae*-like RdRp and *Tombusviridae*-like movement and capsid proteins (*Rastgou et al., 2009*), our newly-described viruses may have evolved by reassortment of different RNA viruses during co-infection, resulting in the combination of RdRp from Narnavirus and another two ORFs from an undescribed yet RNA virus family or families (Figure 9B). Further investigation of the origins and functions of these hypothetical proteins clearly need to be conducted, for both the understanding of the virus biological cycle and its evolutionary history. Indeed, capsid-less elements cannot exist in an extracellular state and are necessarily transmitted via intracellular mode (cell division or cell mating) (*Dolja and Koonin 2012; Hillman and Cai 2013*).

Narnaviridae comprise some of the simplest viruses described to date, containing a single segment encoding a single replicase. Despite this, they are still able to impact their hosts in profound ways. For example, a reduction in host virulence (hypovirulence) has been documented in the case of mitoviruses (a genus of *Narnaviridae* infecting Fungi) (*Hillman and Cai 2013*). Combined with the previously reported impacts of dsRNA viral infections on the biology, pathogenesis and treatment response of the human parasite *Leishmania* (*Gómez-Arreaza et al., 2017*), investigating the effects of newly discovered viruses on *P. vivax* biology and pathogenesis is clearly an area of interest. Intuitively, the biological consequences of the high prevalence of this virus in *P. vivax*-infected individuals must represent an important avenue for future research. More broadly, the characterization of the viral cycle of MaRNAV-1, their biology, and interactions with its host may also help to better understand the biology and life-cycle of *P. vivax* parasites, as well as the modulation of host and parasite responses leading to immunoevasion, pathogenesis and transmission. In particular, future work should focus on the impact of virus infection on the hypnozoite liver stage of *P. vivax*, which is not present among the other, virus-negative, *Plasmodium* species assessed in this study, and is responsible for *P. vivax* infection relapses in human hosts and ongoing transmission in the absence of specific liver treatment. Similarly, it will be important to determine whether MaRNAV-1 or a related virus infects the other human *Plasmodium spp.* with the hypnozoite life-cycle stage - *P. ovale curtisi* and *P. ovale wallikeri* (*White 2016*) - as well as the possible role of viral infection in promoting immunoevasion, such as asymptomatic infection (*Pava et al., 2016*) or pathogenesis (*Barber et al., 2015*).

Materials and Methods

Ethics statement

The study was approved by the ethics committees of the Malaysian Ministry of Health (NMRR-10-754-6684) and Menzies School of Health Research (HREC 2010-1431).

Biological samples

Whole blood samples (1 ml) were collected at district hospitals from healthy and *Plasmodium*-infected patients in the state of Sabah, east Malaysia in 2013-14. Patients had a clinical illness consistent with malaria, with blood collected prior to antimalarial treatment. Parasite density was quantified by research microscopy using pre-treatment slides and reported as the number of parasites per 200 leukocytes from thick blood film. This was converted into the number of parasites per microliter using the patient's leukocyte count

565 from their hospital automated hematology result. Remaining blood samples were stored in
566 RNAlater and conserved at -80°C until RNA extraction. Sampling locations, sampling dates,
567 *Plasmodium* species validation and parasite counts are reported in Table S1.
568
569 PCR validation for *P. vivax* and *P. falciparum* were conducted following *Padley et al. (2003*).
570 A single-round PCR was performed using one single reverse primer in combination with
571 species-specific forward primers (Table S5). The *P. knowlesi*-infected sample validation
572 were conducted following *Imwong et al. (2009*) using a nested-PCR strategy with two
573 primer couples: rPLU3 and rPLU4 for the first PCR, and PkF1140 and PkR1150 for the
574 nested PCR (see Table S5 for the corresponding sequences).
575
576 **Total RNA extraction and RNA sequencing**
577 Total RNAs were extracted from blood samples using the Qiagen® RNeasy Plus Universal
578 MIDI kit and following manufacturer's instructions. Importantly, randomized and serial
579 extractions were conducted to prevent potential experimental biases and to facilitate the
580 detection of kit, columns, reagents or extraction-specific contamination from the
581 corresponding meta-transcriptomic data.
582
583 Total RNA samples were grouped by *Plasmodium* species and pooled in equimolar ratios
584 into a single sample. Quality assessments were then conducted and four TruSeq stranded
585 libraries were synthetized by the Australian Genome Research facility (AGRF), including a
586 rRNA and globin mRNA depletion using RiboZero and globin depletion kit from Illumina®.
587 Resulting libraries were run on HiSeq2500 (paired-end, 100bp) from the AGRF platform
588 (RNA samples quality and the features of each library are described in Table S6).
589
590 **rRNA and host read depletion**
591 Raw reads were first trimmed using the Trimmomatic software (*Bolger et al., 2014*) to
592 remove Illumina adapters and low-quality bases. Human, ribosomal RNA (rRNA) and
593 *Plasmodium* associated reads were removed from the data sets by successively mapping
594 the trimmed reads to the latest versions of each corresponding reference sequence
595 databases (see Table S7 for more details) with either SortmeRNA (*Kopylova et al., 2012*) or
596 Bowtie2 software and by applying local and very-sensitive options for the alignment
597 (*Langmead and Salzberg 2012*). All corresponding databases and the software used for the
598 host analyses and rRNA depletions are summarized in Table S7.
599

Contig assembly and counting

Depleted read data sets were assembled into longer contigs using the Trinity software (*Grabherr et al., 2011*). The resulting contig abundances were estimated using the RSEM software (*Li and Dewey 2011*).

Virus discovery

A global sequence-based homology detection was performed using Blastn and Diamond Blastx (*Buchfink et al., 2015*) against the entire non-redundant nucleotide (nt) or protein (nr) databases with using e-values of 1e-10 and 1e-5, respectively. Profiling plots were obtained by clustering contigs based on the taxonomy of their best Blastn and/or Blastx hits (highest Blast score) and plotting their respective length and abundance using ggplot2 (*Wickham 2009*).

In parallel, a RNA virus-specific sequence-based homology detection was conducted by first aligning our data sets to a viral RdRp database using Diamond Blastx. To ensure the removal of false-positives, a second Blastx round using exhaustive hits output parameters was performed on each RdRP-matched contigs to discard contigs which are more likely from a non-viral source. True-positive viral contigs were merged when possible and further analysed using Geneious 11.1.4 software (*Kearse et al., 2012*).

“Unknown contigs” (i.e. contigs with no Blastx hit) longer than 1kb were retained and submitted to a second round of Blastx using low-stringent cut-off of 1e-4. HMM-profile and structural-based homology searches were also performed on these unknown contigs using the normal mode search of the Protein Homology/analogY Recognition Engine v 2.0 (Phyre2) web portal (*Kearse et al., 2012; Kelley et al., 2015*). Briefly, the amino acid sequences of predicted ORFs were first compared to a curated non-redundant nr20 data set (comprising only sequences with <20% mutual identity) using HHblits (*Remmert et al., 2012*). Secondary structures were predicted from the multiple sequence alignment and this information was converted into a Hidden Markov model (HMM). This HMM was then used as a query against a HMM database built from proteins of known 3D-structures and using HHsearch (*Söding 2005*). Finally, a 3D-structure modelling step was performed using HHsearch hits as templates, following the method described in *Remmert et al. (2012*).

Virus-like sequences were further experimentally confirmed in total RNA samples by performing cDNA synthesis using the SuperScript™ IV reverse transcriptase (Invitrogen™,

Catalog number: 11756500) and PCR amplification with virus candidate specific primers using the Platinum™ SuperFi™ DNA polymerase (Invitrogen™, Catalog number: 12359010). Amplified products were Sanger sequenced using intermediary primers enabling a full-length coverage (all primers are listed in Table S5).

Host-virus assignment

To help assign a virus to a specific host (i.e. determining which host organism these viruses likely infect), we analyzed both codon usage bias and genomic dinucleotide composition (*Di Giallonardo et al., 2017; Su et al., 2009*). Accordingly, the average codon usage of *H. sapiens* and *P. vivax* were retrieved from the Codon Usage Database (available at http://www.kazusa.or.jp/codon/) and the codon adaptation index (CAI) and associated expected-CAI (eCAI) were determined using the CAIcal web-server, available at http://genomes.urv.es/CAIcal/E-CAI/ (*Puigbò et al., 2008*). The most prevalent *Anopheles* mosquito vector in the Sabah region of Malaysia (*Anopheles balabacensis*) did not have a codon usage table available. Hence, in this case we retrieved the codon usage from seven other *Anopheles* species (*A. dirus*, *A. minimus*, *A. cracens*, *A. gambiae*, *A. culicifacies*, *A. merus* and *A. stephensi*) which included major vectors of malaria in South East Asia.

As well as codon bias, we determined the dinucleotide composition of MaRNAV-1 and compared to that of *Anopheles gambiae* (RefSeq | GCF_000005575.2) and *P. vivax* (RefSeq | GCF_000002415.2). The match between host and virus was then calculated using the method described in *Di Giallonardo et al. (2017*) by calculating the f/ffratio from the MaRNAV-1 sequences obtained by Sanger sequencing (see above).

Virus genome characterization

Validated virus-like genomes were further characterized using both genome/protein annotation programs, including InterProScan for protein domain, Sigcleave and Fuzzpro from EMBOSS package for signal cleavage sites and motifs, and TMHMM for transmembrane regions (*Krogh et al., 2001; Mulder and Apweiler 2007*).

Mining of the Sequence Read Archive (SRA)

To identify homologs of MaRNAV-1, the newly identified Narna-like virus sequence was used as a reference in both Magic-blast blastn (default parameters) (*Boratyn et al., 2018*) and Diamond blastx (cut-off 1e-5) (*Buchfink et al., 2015*) analyses of several RNA-seq data sets obtained from *Plasmodium sp*. available on the NCBI SRA using the NCBI SRA toolkit

v2.9.2. The list of the corresponding BioProjects, runs and references are provided in Table S2.

P. vivax SRA library information (i.e. host, location and biological replicates) was manually retrieved from the corresponding papers (Table S3). When possible, this information was used to assess whether such variables were associated with the detection of narna-like viruses by performing Chi-squared tests using the SPSS Statistics software (IBM®). SRA runs positive for homologs to MaRNAV-1 (number of read blast hits >100) were imported locally and assembled following the same workflow as previously used to extract homologous full-length contigs and to calculate their relative abundance in the samples.

To further assess host assignments, the same SRA data sets were subjected to Magic-Blast using *Plasmodium* and mosquito and human specific housekeeping gene transcripts (LDH-P gb|M93720.1 and RSP-7 gb|L20837.1, respectively). This large-scale analysis may necessarily result in false-negative results because of the idiosyncrasies of the experimental procedures used, such as the depletion of human reads or *Plasmodium* RNA enrichment, both of which can bias such host read counting. Moreover, some samples come from the same biological replicate and hence cannot be counted as independent. Such potential biases forced us to manually retrieve all information for each *P. vivax* SRA library using related publications (Table S3).

Phylogenetic analysis

Predicted ORFs containing the viral RNA-dependent (RdRp) domain from both the SRA and human blood and bird associated sequences (see below) were translated and aligned using the E-INS-I algorithm in MAFFT v7.309 (*Katoh and Standley 2013*). To place the newly identified viruses into a more expansive phylogeny of RNA viruses, reference protein sequences of the closest homologous viral families or genera were retrieved from NCBI and incorporated to the amino acid sequence alignment. The alignments were then trimmed with Gblocks under the lowest stringency parameters (*Castresana 2000*). The best-fit amino acid substitution models were then inferred from each curated protein alignment using either the Smart model selection (SMS) (*Lefort et al., 2017*) or ModelFinder (*Kalyaanamoorthy et al., 2017*), and maximum likelihood phylogenetic trees were then estimated with either PhyML (*Guindon et al., 2009*) or IQ-tree (*Nguyen et al., 2015*) with bootstrapping (1000 replicates) used to assess node support. For clarity, all phylogenetic trees were midpoint rooted.

705

706 **Analysis of avian meta-transcriptomic libraries**

707 To supplement our analysis of human *Plasmodium* samples, we analyzed four meta-
708 transcriptomic libraries sampled from four bird species (*Gymnorhina tibicen, Strepera*
709 *graculina, Corvus coronoides* and *Grallina cyanoleuca*) in New South Wales, Australia.
710 Sampling details are reported in Table S4. The RNA-Seq data analysis and the Blastx
711 detection of MaRNAV-1 homologs from bird sample data sets were conducted as
712 described above.

713

714 The PCR-based detection of both Narna-like viruses and *Leucocytozoon* parasites were
715 conducted using newly-designed primers corresponding to the *Leucocytozoon* homologs
716 of the MaRNAV-1 RdRp and segment II (primers are described in Table S5), and following
717 the same PCR protocol as described above. PCR-based *Leucocytozoon* detections were
718 performed using primers targeting the *Leucocytozoon* mitochondrial cytochrome B oxidase
719 gene as described in (*Pacheco et al., 2018*). All additional analyses of the bird data sets
720 were performed utilizing the software and tools described above.

721

722 **Acknowledgments**

723 This work was supported by an Australian Research Council Australian Laureate Fellowship
724 awarded to ECH (FL170100022), the Australian National Health and Medical Research
725 Council (grants #1037304 and #1045156; Fellowships to NMA [#1042072] and MJG
726 [#1138860]), and the Australian Centre of Research Excellence in Malaria Elimination. We
727 acknowledge the Sydney Informatics Hub and the University of Sydney's high performance
728 cluster Artemis for providing the computational resources required for the RNA-seq data
729 processing, and Wei-shan Chang for providing technical assistance with PCR validation.
730 We also thank the Director-General, Ministry of Health, Malaysia, for permission to publish
731 this manuscript.

References

Adaui V, Lye L-F, Akopyants NS, Zimic M, Llanos-Cuentas A, Garcia L, Maes I, De Doncker S, Dobson DE, Arevalo J, Dujardin J-C, Beverley SM. 2016. Association of the endobiont double-stranded RNA virus LRV1 with treatment failure for human Leishmaniasis caused by *Leishmania braziliensis* in Peru and Bolivia. *Journal of Infectious Diseases* **213**:112-121. doi: 10.1093/infdis/jiv354.

Akopyants NS, Lye L-F, Dobson DE, Lukeš J, Beverley SM. 2016a. A novel bunyavirus-like virus of trypanosomatid protist parasites. *Genome Announcements* **4**:e00715-16. doi: 10.1128/genomeA.00715-16.

Akopyants NS, Lye L-F, Dobson DE, Lukeš J, Beverley SM. 2016b. A narnavirus in the trypanosomatid protist plant pathogen *Phytomonas serpens*. *Genome Announcements* **4**:e00711-16. doi: 10.1128/genomeA.00711-16.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F. 2006. The marine viromes of four oceanic regions. *PLoS Biology* **4**:e368. doi: 10.1371/journal.pbio.0040368.

Barber BE, William T, Grigg MJ, Parameswaran U, Piera KA, Price RN, Yeo TW, Anstey NM. 2015. Parasite biomass-related inflammation, endothelial activation, microvascular dysfunction and disease severity in vivax malaria. *PLoS Pathogens* **11**:e1004558. doi: 10.1371/journal.ppat.1004558.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120. doi: 10.1093/bioinformatics/btu170.

Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL. 2018. Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads. *bioRxiv* doi: 10.1101/390013.

Bourreau E, Ginouves M, Prévot G, Hartley M-A, Gangneux J-P, Robert-Gangneux F, Dufour J, Sainte-Marie D, Bertolotti A, Pratlong F, Martin R, Schütz F, Couppié P, Fasel N, Ronet C. 2016. Presence of Leishmania RNA Virus 1 in *Leishmania guyanensis* increases the risk of first-line treatment failure and symptomatic relapse. *Journal of Infectious Diseases* **213**:105–111. doi: 10.1093/infdis/jiv355.

Bousema T, Drakeley C. 2011. Epidemiology and infectivity of *Plasmodium falciparum* and *Plasmodium vivax* gametocytes in relation to malaria control and elimination. *Clinical Microbiology Reviews* **24**:377-410. doi:10.1128/cmr.00051-10.

Brettmann EA, Shaik JS, Zangger H, Lye L-F, Kuhlmann FM, Akopyants NS, Oschwald DM, Owens KL, Hickerson SM, Ronet C, Fasel N, Beverley SM. 2016. Tilting the balance

767 between RNA interference and replication eradicates Leishmania RNA virus 1 and
768 mitigates the inflammatory response. *Proceedings of the National Academy of*
769 *Sciences USA* **113**:11998–12005. doi: 10.1073/pnas.1615085113.
770 Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND.
771 *Nature Methods* **12**:59–60. doi: 10.1038/nmeth.3176.
772 Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, Crabtree J, Angiuoli SV,
773 Merino EF, Amedeo P, Cheng Q, Coulson RMR, Crabb BS, Del Portillo HA, Essien K,
774 Feldblyum TV, Fernandez-Becerra C, Gilson PR, Gueye AH, Guo X, Kang'a S, Kooij
775 TWA, Korsinczky M, Meyer EV-S, Nene V, Paulsen I, White O, Ralph SA, Ren Q,
776 Sargeant TJ, Salzberg SL, Stoeckert CJ, Sullivan SA, Yamamoto MM, Hoffman SL,
777 Wortman JR, Gardner MJ, Galinski MR, Barnwell JW, Fraser-Liggett CM. 2008.
778 Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*.
779 *Nature* **455**:757–763. doi: 10.1038/nature07327.
780 Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in
781 phylogenetic analysis. *Molecular Biology & Evolution* **17**:540–552. doi:
782 10.1093/oxfordjournals.molbev.a026334.
783 Culley AI, Lang AS, Suttle CA. 2006. Metagenomic analysis of coastal RNA virus
784 communities. *Science* **312**:1795-1798. doi: 10.1126/science.1127404.
785 Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M,
786 Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Thurber RV, Reid RP,
787 Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F. 2008. Biodiversity
788 and biogeography of phages in modern stromatolites and thrombolites. *Nature*
789 **452**:340–343. doi: 10.1038/nature06735.
790 Di Giallonardo F, Schlub TE, Shi M, Holmes EC. 2017. Dinucleotide composition in animal
791 RNA Viruses is shaped more by virus family than by host species. *Journal of Virology*
792 **91**:e02381-16. doi: 10.1128/JVI.02381-16.
793 Dolja VV, Koonin EV. 2012. Capsid-less RNA viruses. *eLS*. doi:
794 10.1002/9780470015902.a0023269.
795 Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns
796 and determinants. *Nature Reviews Genetics* **9**:267–276. doi: 10.1038/nrg2323.
797 Fichorova RN, Lee Y, Yamamoto HS, Takagi Y, Hayes GR, Goodman RP, Chepa-Lotrea X,
798 Buck OR, Murray R, Kula T, Beach DH, Singh BN, Nibert ML. 2012. Endobiont viruses
799 sensed by the human host - beyond conventional antiparasitic therapy. *PLoS One*
800 **7**:e48418. doi: 10.1371/journal.pone.0048418.

801 Forterre P. 2010. Defining life: the virus viewpoint. *Origins of Life & Evolution of the*
802 *Biosphere* **40**:151–160. doi: 10.1007/s11084-010-9194-1.
803 Garnham PC, Bird RG, Baker JR. 1962. Electron microscope studies of motile stages of
804 malaria parasites. III. The ookinetes of *Haemamoeba* and *Plasmodium*. *Transactions of*
805 *the Royal Society of Tropical Medicine and Hygiene* **56**:116–120. doi: 10.1016/0035-
806 9203(62)90137-2.
807 Gómez-Arreaza A, Haenni A-L, Dunia I, Avilán L. 2017. Viruses of parasites as actors in the
808 parasite-host relationship: A "ménage à trois." *Acta Tropica* **166**:126-
809 132.doi:10.1016/j.actatropica.2016.11.028.
810 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L,
811 Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma
812 F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length
813 transcriptome assembly from RNA-Seq data without a reference genome. *Nature*
814 *Biotechnology* **29**:644–652. doi: 10.1038/nbt.1883.
815 Grigg MJ, William T, Barber BE, Rajahram GS, Menon J, Schimann E, Piera K, Wilkes CS,
816 Patel K, Chandna A, Drakeley CJ, Yeo TW, Anstey NM. 2018. Age-related clinical
817 spectrum of *Plasmodium knowlesi* malaria and predictors of severity. *Clinical Infectious*
818 *Diseases* **67**:350–359. doi: 10.1093/cid/ciy065.
819 Grybchuk D, Akopyants NS, Kostygov AY, Konovalovas A, Lye L-F, Dobson DE, Zangger H,
820 Fasel N, Butenko A, Frolov AO, Votýpka J, d'Avila-Levy CM, Kulich P, Moravcová J,
821 Plevka P, Rogozin IB, Serva S, Lukeš J, Beverley SM, Yurchenko V. 2018. Viral
822 discovery and diversity in trypanosomatid protozoa with a focus on relatives of the
823 human parasite. *Proceedings of the National Academy of Sciences USA* **115**:E506–
824 E515. doi: 10.1073/pnas.1717806115.
825 Guindon S, Delsuc F, Dufayard J-F, Gascuel O. 2009. Estimating Maximum Likelihood
826 Phylogenies with PhyML. *Methods in Molecular Biology* **537**:113-137. doi:10.1007/978-
827 1-59745-251-9_6.
828 Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote
829 evolution and the rise of complex multicellular life. *BMC Evolutionary Biology* **4**:2. doi:
830 10.1186/1471-2148-4-2.
831 Hillman BI, Cai G. 2013. The family *Narnaviridae*: simplest of RNA viruses. *Advances in*
832 *Virus Research* **86**:149–176. doi: 10.1016/B978-0-12-394315-6.00006-4.
833 Illergård K, Ardell DH, Elofsson A. 2009. Structure is three to ten times more conserved
834 than sequence-a study of structural response in protein cores. *Proteins* **77**:499–508.
835 doi: 10.1002/prot.22458.

836 Imwong M, Tanomsing N, Pukrittayakamee S, Day NPJ, White NJ, Snounou G. 2009.
837 Spurious amplification of a *Plasmodium vivax* small-subunit RNA gene by use of
838 primers currently used to detect *P. knowlesi*. *Journal of Clinical Microbiology* **47**:4173-
839 4175. doi:10.1128/jcm.00811-09.
840 Ito MM, Catanhêde LM, Katsuragawa TH, da Silva Junior CF, Camargo LMA, de Godoi
841 Mattos R, Vilallobos-Salcedo JM. 2015. Correlation between presence of Leishmania
842 RNA virus 1 and clinical characteristics of nasal mucosal leishmaniosis. *Brazilian*
843 *Journal of Otorhinolaryngology* **81**:533-540. doi:10.1016/j.bjorl.2015.07.014
844 Ives A, Ronet C, Prevel F, Ruzzante G, Fuertes-Marraco S, Schutz F, Zangger H, Revaz-
845 Breton M, Lye L-F, Hickerson SM, Beverley SM, Acha-Orbea H, Launois P, Fasel N,
846 Masina S. 2011. Leishmania RNA virus controls the severity of mucocutaneous
847 leishmaniasis. *Science* **331**:775–778. doi: 10.1126/science.1199326.
848 Jenkins MC, Higgins J, Abrahante JE, Kniel KE, O'Brien C, Trout J, Lancto CA,
849 Abrahamsen MS, Fayer R. 2008. Fecundity of *Cryptosporidium parvum* is correlated
850 with intracellular levels of the viral symbiont CPV. *International Journal of Parasitology*
851 **38**:1051–1055. doi: 10.1016/j.ijpara.2007.11.005.
852 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder:
853 fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**:587–589.
854 doi: 10.1038/nmeth.4285.
855 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
856 improvements in performance and usability. *Molecular Biology & Evolution* **30**:772–780.
857 doi: 10.1093/molbev/mst010.
858 Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A,
859 Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious
860 Basic: An integrated and extendable desktop software platform for the organization
861 and analysis of sequence data. *Bioinformatics* **28**:1647-1649.
862 doi:10.1093/bioinformatics/bts199.
863 Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal
864 for protein modeling, prediction and analysis. *Nature Protocols* **10**:845–858. doi:
865 10.1038/nprot.2015.053.
866 Khramtsov NV, Woods KM, Nesterenko MV, Dykstra CC, Upton SJ. 1997. Virus-like,
867 double-stranded RNAs in the parasitic protozoan *Cryptosporidium parvum*. *Molecular*
868 *Microbiology* **26**:289–300. 10.1046/j.1365-2958.1997.5721933.x.

869 Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal
870 RNAs in metatranscriptomic data. *Bioinformatics* **28**:3211–3217. doi:
871 10.1093/bioinformatics/bts611.
872 Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane
873 protein topology with a hidden Markov model: application to complete genomes.
874 *Journal of Molecular Biology* **305**:567–580. doi: 10.1006/jmbi.2000.4315.
875 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature*
876 *Methods* **9**:357–359. doi: 10.1038/nmeth.1923.
877 Lefort V, Longueville J-E, Gascuel O. 2017. SMS: Smart Model Selection in PhyML.
878 *Molecular Biology & Evolution* **34**:2422–2424. doi: 10.1093/molbev/msx149.
879 Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or
880 without a reference genome. *BMC Bioinformatics* **12**:323. doi: 10.1186/1471-2105-12-
881 323.
882 Li C-X, Shi M, Tian J-H, Lin X-D, Kang Y-J, Chen L-J, Qin X-C, Xu J, Holmes EC, Zhang Y-
883 Z. 2015. Unprecedented genomic diversity of RNA viruses in arthropods reveals the
884 ancestry of negative-sense RNA viruses. *eLife* **4**:e05378. doi:10.7554/eLife.05378.
885 Lye L-F, Akopyants NS, Dobson DE, Beverley SM. 2016. A narnavirus-like element from the
886 trypanosomatid protozoan parasite *Leptomonas seymouri*. *Genome Announcements* **4**:
887 e00713-16. doi:10.1128/genomeA.00713-16.
888 Miles MA. 1988. Viruses of parasitic protozoa. *Parasitology Today* **4**:289–290. doi:
889 10.1016/0169-4758(88)90023-3.
890 Miller RL, Wang AL, Wang CC. 1988. Purification and characterization of the *Giardia lamblia*
891 double-stranded RNA virus. *Molecular & Biochemical Parasitology* **28**:189–195.
892 10.1016/0166-6851(88)90003-5.
893 Mulder N, Apweiler R. 2007. InterPro and InterProScan: Tools for protein sequence
894 classification and comparison. *Methods in Molecular Biology* **396**:59-70. doi:
895 10.1007/978-1-59745-515-2_5.
896 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective
897 stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology*
898 *& Evolution* **32**:268–274. doi: 10.1093/molbev/msu300.
899 Nibert ML, Woods KM, Upton SJ, Ghabrial SA. 2009. *Cryspovirus*: a new genus of
900 protozoan viruses in the family *Partitiviridae*. *Archives of Virology* **154**:1959-1965.
901 doi:10.1007/s00705-009-0513-7.
902 Pacheco MA, Cepeda AS, Bernotienė R, Lotta IA, Matta NE, Valkiūnas G, Escalante AA.
903 2018. Primers targeting mitochondrial genes of avian haemosporidians: PCR detection

and differential DNA amplification of parasites belonging to different genera. *International Journal of Parasitology* **48**:657–670. doi: 10.1016/j.ijpara.2018.02.003.

Padley D, Moody AH, Chiodini PL, Saldanha J. 2003. Use of a rapid, single-round, multiplex PCR to detect malarial parasites and identify the species present. *Annals of Tropical Medicine & Parasitology* **97**:131-137. doi:10.1179/000349803125002977.

Padma TV. 2015. Russian doll disease is a virus inside a parasite inside a fly. New Scientist. August 10th, 2015. https://www.newscientist.com/article/dn28020-russian-doll-disease-is-a-virus-inside-a-parasite-inside-a-fly/.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's virome. *Nature* **536**:425–430. doi: 10.1038/nature19094.

Pava Z, Burdam FH, Handayuni I, Trianty L, Utami RAS, Tirta YK, Kenangalem E, Lampah D, Kusuma A, Wirjanata G, Kho S, Simpson JA, Auburn S, Douglas NM, Noviyanti R, Anstey NM, Poespoprodjo JR, Marfurt J, Price RN. 2016. Submicroscopic and asymptomatic *Plasmodium* parasitaemia associated with significant risk of anaemia in Papua, Indonesia. *PLoS One* **11**:e0165340. doi: 10.1371/journal.pone.0165340.

Puigbò P, Bravo IG, Garcia-Vallve S. 2008. CAIcal: a combined set of tools to assess codon usage adaptation. *Biology Direct* **3**:38. doi: 10.1186/1745-6150-3-38.

Rastgou M, Habibi MK, Izadpanah K, Masenga V, Milne RG, Wolf YI, Koonin EV, Turina M. 2009. Molecular characterization of the plant virus genus *Ourmiavirus* and evidence of inter-kingdom reassortment of viral genome segments as its possible route of origin. *Journal of General Virology* **90**:2525–2535. doi: 10.1099/vir.0.013086-0.

Remmert M, Biegert A, Hauser A, Söding J. 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9**:173–175. doi:10.1038/nmeth.1818.

Shi M, Lin X-D, Tian J-H, Chen L-J, Chen X, Li C-X, Qin X-C, Li J, Cao J-P, Eden J-S, Buchmann J, Wang W, Xu J, Holmes EC, Zhang Y-Z. 2016. Redefining the invertebrate RNA virosphere. *Nature* **540**:539-543. doi: 10.1038/nature20167.

Shi M, Neville P, Nicholson J, Eden J-S, Imrie A, Holmes EC. 2017. High-resolution metatranscriptomics reveals the ecological dynamics of mosquito-associated RNA viruses in Western Australia. *Journal of Virology* **91**:e00680-17. doi:10.1128/JVI.00680-17.

Söding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**:951–960. doi: 10.1093/bioinformatics/bti125.

938 Sukla S, Roy S, Sundar S, Biswas S. 2017. Leptomonas seymouri narna-like virus 1 and not
939 leishmaniaviruses detected in kala-azar samples from India. *Archives of Virology*
940 **162**:3827–3835. doi: 10.1007/s00705-017-3559-y.
941 Su M-W, Lin H-M, Yuan HS, Chu W-C. 2009. Categorizing host-dependent RNA viruses by
942 principal component analysis of their codon usage preferences. *Journal of*
943 *Computational Biology* **16**:1539–1547. doi: 10.1089/cmb.2009.0046.
944 Suttle CA. 2005. Viruses in the sea. *Nature* **437**:356–361. doi:10.1038/nature04160.
945 Tarr PI, Aline RF Jr, Smiley BL, Scholler J, Keithly J, Stuart K. 1988. LR1: a candidate RNA
946 virus of *Leishmania*. *Proceedings of the National Academy of Sciences USA* **85**:9572–
947 9575. doi: 10.1073/pnas.85.24.9572.
948 Wang AL, Wang CC. 1986. Discovery of a specific double-stranded RNA virus in *Giardia*
949 *lamblia*. *Molecular & Biochemical Parasitology* **21**:269–276. doi: 10.1016/0166-
950 6851(86)90132-5.
951 Wang AL, Wang CC. 1985. A linear double-stranded RNA in *Trichomonas vaginalis*. *Journal*
952 *of Biological Chemistry* **260**:3697–3702. doi: 10.1073/pnas.83.20.7956.
953 White NJ. 2016. Why do some primate malarias relapse? *Trends in Parasitology* **32**:918–
954 920. doi: 10.1016/j.pt.2016.08.014.
955 WHO. 2018. World Health Organization. *World Malaria Report 2018*.
956 Widmer G, Comeau AM, Furlong DB, Wirth DF, Patterson JL. 1989. Characterization of a
957 RNA virus from the parasite Leishmania. *Proceedings of the National Academy of*
958 *Sciences USA* **86**:5979–5982. doi: 10.1073/pnas.86.15.5979.
959 Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer.
960 Zangger H, Hailu A, Desponds C, Lye L-F, Akopyants NS, Dobson DE, Ronet C, Ghalib H,
961 Beverley SM, Fasel N. 2014. *Leishmania aethiopica* field isolates bearing an
962 endosymbiontic dsRNA virus induce pro-inflammatory cytokine response. *PLoS*
963 *Neglected Tropical Diseases* **8**:e2836. doi: 10.1371/journal.pntd.0002836.
964 Zhang Y-Z, Shi M, Holmes EC. 2018. Using metagenomics to characterize an expanding
965 virosphere. *Cell* **172**:1168–1172. doi: 10.1016/j.cell.2018.02.043.