A virus associated with the zoonotic pathogen *Plasmodium knowlesi* causing human malaria is a member of a diverse and unclassified viral taxon

Mary E. Petrone[1,2*†], Justine Charon[3*], Matthew J Grigg[4,5], Timothy William[5,6], Giri S Rajahram[5,7], Jacob Westaway[4], Kim A Piera[4], Mang Shi[8,9,10,11], Nicholas M. Anstey[4,5], Edward C. Holmes[2,12]

[1]Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia.
[2]Laboratory of Data Discovery for Health Limited, Hong Kong SAR, China.
[3] Fruit Biology and Pathology Unit, University of Bordeaux, INRAE, Bordeaux, France
[4]Global and Tropical Health Division, Menzies School of Health Research, Charles Darwin University, Darwin, Australia.
[5]Infectious Diseases Society Kota Kinabalu Sabah - Menzies School of Health Research Clinical Research Unit, Kota Kinabalu, Malaysia.
[6]Subang Jaya Medical Centre, Subang Jaya, Malaysia.
[7]Queen Elizabeth Hospital II, Ministry of Health Malaysia, Kota Kinabalu, Sabah, Malaysia.
[8]State Key Laboratory for Biocontrol, School of Medicine, Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China.
[9]National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, Sun Yat-sen University, Shenzhen, China.
[10]Shenzhen Key Laboratory for Systems Medicine in Inflammatory Diseases, Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, China.
[11]Guangdong Provincial Center for Disease Control and Prevention, Guangzhou, China.
[12]School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia.

*Contributed equally
[†]Correspondence: mary.petrone@sydney.edu.au

ABSTRACT

Apicomplexa are single-celled eukaryotes that can infect humans and include the mosquito-borne parasite *Plasmodium*, the cause of malaria. Increasing rates of drug resistance in human-only *Plasmodium* species are reducing the efficacy of control efforts and antimalarial treatments. There are also rising cases of *P. knowlesi*, the only zoonotic *Plasmodium* species that causes severe disease and death in humans. Thus, there is a need to develop additional innovative strategies to combat malaria. Viruses that infect non-*Plasmodium* spp. disease-causing protozoa have been shown to affect pathogen life cycle and disease outcomes. However, only one virus (Matryoshka RNA virus 1) has been identified in *Plasmodium*, and none have been identified in zoonotic *Plasmodium* species. The rapid expansion of the known RNA virosphere using structure- and artificial intelligence-based methods suggests that this dearth is due to the divergent nature of RNA viruses that infect protozoa. We leveraged these newly uncovered data sets to explore the virome of human-infecting *Plasmodium* species collected in Sabah, east (Borneo) Malaysia. We identified a highly divergent RNA virus in two

human-infecting *P. knowlesi* isolates that is related to the unclassified group 'ormycoviruses'. By characterising fifteen additional ormycoviruses identified in the transcriptomes of arthropods we show that this group of viruses exhibits a complex ecology at the arthropod-mammal interface. Through the application of artificial intelligence methods, we then demonstrate that the ormycoviruses are part of a diverse and unclassified viral taxon. This is the first observation of an RNA virus in a zoonotic *Plasmodium* species. By linking small-scale experimental data to large-scale virus discovery advances, we characterise the diversity and genomic architecture of an unclassified viral taxon. This approach should be used to further explore the virome of disease-causing Apicomplexa and better understand how protozoa-infecting viruses may affect parasite fitness, pathobiology, and treatment outcomes.

INTRODUCTION

Parasitic protozoa are a highly diverse collection of single-celled eukaryotes that can cause disease in many vertebrates. Organisms belonging to the phylum Apicomplexa are associated with a range of human diseases including malaria (*Plasmodium*), inflammation of the brain (*Toxoplasma*)[1], diarrhea (*Cryptosporidium*)[2], and severe anaemia (*Babesia*)[3]. *Plasmodium* is the leading cause of death from Apicomplexa in humans worldwide[4]. This mosquito-borne infection is estimated to have caused over 240 million cases of malaria and to have killed over 600,000 people in 2022 alone[5].

Efforts to control and treat malaria are challenged by the complex ecology of this parasite[6] and mounting antimalarial resistance[7,8]. Of the five human-only infecting species of *Plasmodium* (*P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale wallikeri*, *and P. ovale curtisii*), *P. falciparum* and *P. vivax* cause the greatest morbidity and mortality, with *P. falciparum* accounting for more than 95% of malaria fatalities[4]. Partial resistance of *P. falciparum* to artemisinin is entrenched in the Greater Mekong subregion of Southeast Asia[9,10] and has now emerged independently in Africa[5,11,8]. Eight additional *Plasmodium* species can cause human malaria through zoonotic transmission via mosquito vectors[12]. Among these, *P. knowlesi* is the only species to cause severe disease and death in humans[4,13-15]. Although predominating in Malaysian Borneo[16,17], *P. knowlesi* is now recognised as a significant cause of malaria across Southeast Asia[18], in association with changing land use and deforestation[18-21], and in areas with declining incidence of the cross-protective species, *P. vivax*[22]. Thus, innovative strategies are needed to combat and control *Plasmodium* as the efficacy of accessible treatments declines in the human-only species, and changes in land-use cause greater numbers of zoonotic malaria cases.

One potential approach for malaria control involves the use of viruses that infect disease-causing protozoa. In a similar manner to how bacteriophage have been leveraged to combat drug-resistant bacterial infections[23-25], protozoa-infecting viruses have been proposed as a potential new avenue for therapeutics[26,27]. These parasitic protozoan viruses (PPVs)[27] have been identified in *Giardia*, *Leishmania*, *Cryptosporidum*[28,29], *Eimeria*[30-38], *Toxoplasma*[39], *P. vivax*[40,41], and *Babesia*[42,43]. They are of particular interest because some impact the parasite life cycle and modulate disease outcomes in the parasite host. Notably, *Leishmania* species that harbour Leishmania RNA virus 1 have been associated with an increased risk of treatment failure in humans[44] and more severe disease outcomes in mice[45]. Similarly, it has been proposed that infection of *Toxoplasma* with the recently characterised apocryptoviruses (*Narnaviridae*) may be associated with increased disease severity in humans[39], although this has yet to be formally tested. Cryptosporidium parvum virus 1 modulates the interferon response in *Cryptosporidium*-infected mammals[46]. To date, however, only one virus, Matryoshka RNA virus 1, has been identified in a *Plasmodium* species (*P. vivax*)[40,41], and it is not known whether this virus impacts *Plasmodium* fitness or disease pathogenesis in humans.

Extending the known diversity of PPVs requires innovative approaches to virus discovery because both protozoa and the viruses that infect them are likely ancient and often highly divergent. As a case in point, the ormycoviruses were first identified in parasitic protozoa and fungi using structure-based methods[47] and have since been identified in kelp (Stramenopila)[48], ticks[49], palm[50], and additional fungal species[51,52]. This group of bi-segmented RNA viruses shares no measurable phylogenetic relationship to known viral taxa, rendering it invisible to sequence-based discovery methods[47]. Little else is known about ormycoviruses including their complete host range or whether they encode positive- or negative-sense genomes. The application of artificial intelligence-based methods[53], in addition to large-scale sampling of aquatic environments[54], has further uncovered previously inaccessible virus diversity, including entirely novel “supergroups” of unclassified viral taxa[53]. These tools and the data they have generated can be leveraged to explore the viromes of disease-causing protozoa including *Plasmodium*.

In this study, we combine these facets of virus discovery to characterise a divergent virus associated with human-infecting *P. knowlesi* isolates. We contextualise this virus within the vast viral diversity revealed through large-scale virus discovery studies. We also explore the complex ecology of viruses that infect parasites and can be transmitted as passengers to mammalian hosts. Our findings extend the diversity of known *Plasmodium*-associated viruses and highlight the importance of integrating large- and small-scale virus discovery research to better understand viruses that infect these ancient, microscopic hosts.

RESULTS

Identification of a divergent RNA virus associated with human-infecting *Plasmodium knowlesi*

To extend the known diversity of RNA viruses in disease-causing Apicomplexa, we analysed the metatranscriptomes of 18 human blood samples with PCR-confirmed *Plasmodium* infections and six uninfected human controls, collected in Sabah, east (Borneo) Malaysia between 2013 and 2014. These samples are the same as those previously described[40]. Of the patients with malaria, seven were infected with *P. vivax*, six with *P. knowlesi*, and five with *P. falciparum*[40]. Sequencing libraries were pooled according to *Plasmodium* species as were the negative controls, resulting in four libraries (SRR10448859-62, BioProject PRJNA589654). Matryoshka RNA virus 1 was previously found exclusively in all seven *P. vivax* isolates (SRR10448862)[40].

We searched each library for divergent viruses using the RdRp-scan bioinformatic pipeline[55]. This revealed a putative, highly divergent RNA-dependent RNA polymerase (RdRp) that was 3,177nt in length with a complete open reading frame (ORF) and robust sequencing coverage in the *P. knowlesi* library (SRR10448860) (**Fig. 1a**). No identical or related sequences were found in the other three libraries. The transcript was relatively abundant (1.4% of non-rRNA reads), and we confirmed the presence of this putative RdRp in two of the six isolates in the pool using RT-PCR (**Table S1, Fig. S1**). Both patients with putative virus-infected *P. knowlesi* isolates were from Kota Marudu district residing in villages approximately 30km apart. There was three months difference in the date of hospital presentation. Both had uncomplicated malaria with parasitemia of 7,177 and 41,882 parasites/μL, respectively, which were higher than the median parasitemia found in the *P. knowlesi* infections that lacked the putative virus (4,518/μL). Parasitemia was correlated with the RdRp signals we observed with PCR (**Fig. S1**).

Further inspection indicated that this putative virus was a bi-segmented ormycovirus likely infecting the *Plasmodium*. The divergent RdRp shared low but detectable similarity with that of seven previously identified viruses, of which six were ormycoviruses (**Table S2**). We identified a putative second segment of unknown function, 1,721nt in length sharing 22.8% identity (e-value = 3.14×10^{-15}) with the hypothetical protein of Erysiphe lesion-associated ormycovirus 1 (USW07196). The structures of the putative and known hypothetical proteins were significantly similar (p-value = 1.62×10^{-2}) when predicted with AlphaFold2[56,57] and compared by pairwise alignment with FATCAT[58] (**Fig. 1b**). Similar transcripts were not identified in the ormycovirus-negative libraries from the same BioProject. Analysis of the library

156 composition with CCMetagen[59] and the KMA database[60] did not reveal plausible host
157 candidates aside from the *Plasmodium*, which comprised 24% of non-rRNA reads. The
158 remainder aligned to the *Hominidae*, reflecting that the *Plasmodium* were themselves infecting
159 humans. We assumed that the host range of the ormycoviruses likely did not extend to
160 vertebrates, consistent with their absence in the humans without *Plasmodium* infection. Unlike
161 its closest relatives, the *P. knowlesi*-associated RdRp encoded GDD in motif C of its palm
162 domain rather than NDD (**Fig. 1c**).
163
164 To assess the prevalence of this and other ormycoviruses in *P. knowlesi*, we screened 1,470
165 *P. knowlesi* RNA SRA libraries (**Supp. Data 1**) with a custom ormycovirus database. This
166 returned no additional ormycovirus candidates. However, all 1,470 libraries were generated
167 from only seven BioProjects, and only the library we generated was derived from human-host
168 *P. knowlesi* infections. The majority (n = 1,356) were generated from macaque-host *P.*
169 *knowlesi* infections, and all of these were generated by a single contributor from a small set of
170 laboratory-maintained Rhesus macaques (PRJNA508940, PRJNA526495, and
171 PRJNA524357). Sixty-one libraries were derived from cell culture, and the source of 52
172 (BioProject PRJEB24220) could not be determined. Thus, an accurate prevalence estimate of
173 the *P. knowlesi*-associated ormycovirus could not be obtained from this data set.
174
175 We next investigated the prevalence of ormycoviruses more broadly in disease-causing
176 Apicomplexa by screening 2,898 RNA SRA libraries (*Cryptosporidum*, *Coccidia*, *Toxoplasma*,
177 *Babesia*, and *Theileria*) (**Supp. Data 2**). This yielded identical ormyco-like RdRp segments in
178 the transcriptomes of 22 Coccidia (*Cystoisospora suis*) libraries, 21 of which belonged to the
179 same BioProject (PRJEB52768)[61]. The remaining library (SRR4213142) was published by the
180 same authors, suggesting that all 22 libraries were generated from the same source[62]. The
181 transcripts of the *Cystoisospora*-associated virus encoded complete ORFs with an NDD motif
182 C and were ~3.1kb in length (range: 3009-3203) (**Table S3**). This virus was highly divergent,
183 sharing only 32.5% identity (e-value = 6×10^{-37}) with its closest blast hit (Wildcat Canyon virus,
184 WZL61396.1). It was also at low abundance across the 22 libraries (range: 0.01-0.08% of non-
185 rRNA reads). We could not conclude that *C. suis* was the host because fungi represented
186 4.6% of the non-rRNA reads in a representative library (ERR9846867). Regardless, the
187 prevalence of ormycoviruses was 100% among *Cystoisospora suis* libraries but otherwise
188 very low in this data set (0.76%).
189
190 Phylogenetic analysis placed both Apicomplexa-associated viruses in the “Alpha” clade of the
191 ormycoviruses (**Fig. 1d**). The topology of the inferred phylogenies was stable across six
192 combinations of alignment and trimming methods and recapitulated the three main

ormycovirus clades “Alpha”, “Beta”, and “Gamma”[47] with strong support (**Fig. S2**). Viruses did not cluster by host. For example, viruses associated with the fungal species *Erysiphe* fell across all three clades and encoded three different catalytic triads (**Fig. 1d, *icons***), and the Apicomplexa-associated viruses were not closely related within the Alpha clade.

We concluded that the *P. knowlesi*-associated virus represents the first evidence of an RNA virus associated with *P. knowlesi* and constitutes only the second instance of an RNA virus associated with any *Plasmodium* species. We have provisionally named it “Selindung RNA virus 1” because it appeared to be concealed (“terselindung”, Bahasa Malaysia) within the *Plasmodium* parasite, and we will use this name herein.

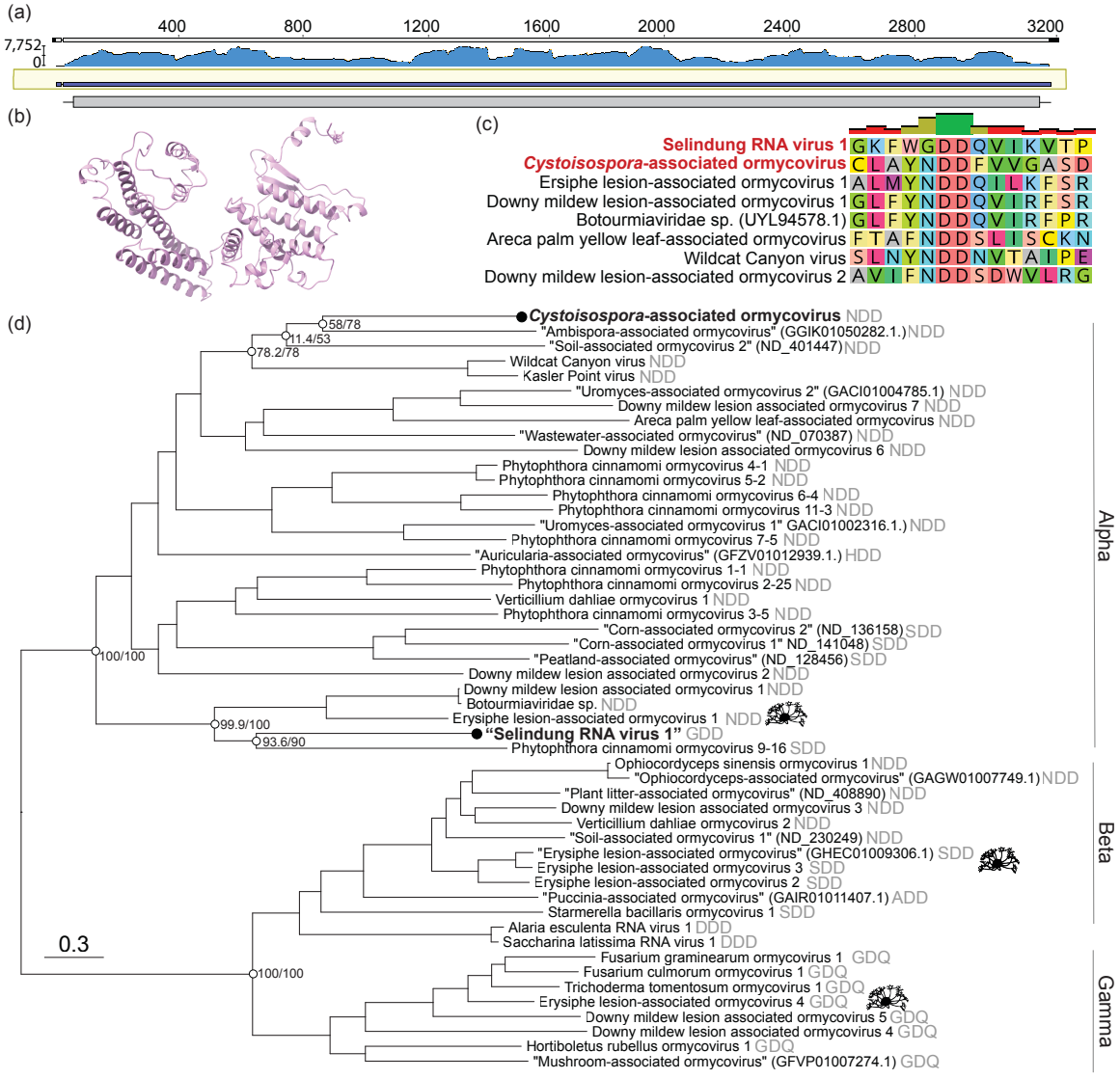

Figure 1. A divergent RNA virus associated with human-infecting *P. knowlesi* is a member of the unclassified group ‘ormycovirus’. (a) Sequencing coverage of the RNA-dependent RNA polymerase (RdRp) of a *P. knowlesi*-associated viral contig. Trimmed reads were mapped to the assembled contig using BBMap[63] and visualised with Geneious Prime

v2024.0.7. (b) The predicted structure of the putative hypothetical protein of Selindung RNA virus 1. (c) MAFFT alignment of motif C in the palm domain of the *P. knowlesi*- and *Cystoisospora*-associated viruses and representative ormycoviruses. (d) Phylogenetic inference of the ormycoviruses aligned with MAFFT. The positions of *Erysiphe*-associated viruses are denoted with black icons (source: phylopic.org). Black tip dots indicate viruses identified in this study. Tips with names in quotes were previously identified but not named[47]. Their corresponding NCBI or RVMT accession is shown in parentheses. The catalytic triad encoded in each palm domain is denoted in grey. Support values are shown at select nodes as sh-aLRT/UFBoot. Tree branches are scaled to amino acid substitutions.

Ormycoviruses are associated with arthropod metatranscriptomes

In addition to expanding the diversity of *Plasmodium*-associated RNA viruses, Selindung RNA virus 1 was of particular interest because it had evidently been transmitted along with its *Plasmodium* host to a human via a mosquito vector. Taking this together with the detectable phylogenetic relationship of this virus and two viruses recovered from tick metagenomes (Wildcat Canyon virus and Kasler Point virus), we posited that ormycoviruses might exhibit a complex ecology at the arthropod-mammal interface. We therefore sought to further extend the known host range of ormycoviruses to the transcriptomes of the arthropods that indirectly transmit them.

We screened the 4,864 arthropod libraries available on NCBI Transcriptome Shotgun Assemblies (TSA) as of August 2024, initially using Kasler Point virus (WZL61394) as input and then following an iterative process (see *Methods*). In this way we identified 15 putative viruses associated with three of the four extant subphyla of the Arthropoda: Chelicerata (n = 1), Crustacea (n = 1), and Hexapoda (n = 13) (**Table S4**). All shared detectable but minimal sequence similarity with published ormycoviruses (range: 27.1-41.0%, **Table S4**). Two encoded GDD at motif C like Selindung RNA virus 1, while the remainder had NDD at this position.

Phylogenetic analysis again supported the conclusion that these viruses are part of the ormycovirus group (**Fig. 2a**). All viruses identified in this study fell in the Alpha clade. Selindung RNA virus 1 formed a group with the other two GDD-encoding viruses (Beetle-associated ormycovirus 1 and Bristletail-associated ormycovirus 1). This placement was consistent across all six iterations of phylogenetic inference (**Fig. S3**). However, aside from this instance and the Gamma clade (GDQ), minimal clustering of motifs was observed. In addition, although the host organisms had been collected from all six inhabited continents, there was no clustering of viruses by geographic region of sampling (**Fig. 2a**).

We concluded that at least some of these viruses were likely infecting single-celled organisms rather than the arthropods themselves for two reasons. First, assessment of each library composition revealed instances of parasitic hosts. Contigs mapping to alveolates accounted for more than one tenth of one Hexapoda (GDXN01) and the only crustacean (GFJG01) library (13% *Gregarinidae* and 12% Ciliophora, respectively) (**Fig 2b**). Similarly, the mite assembly (GEYJ01) included 25% of contigs mapping to fungi (**Fig. 2b**). Second, the virus identified in GBHO01 (*Lygus hesperus*) likely utilised the ciliate genetic code (i.e., only a truncated ORF could be recovered with the standard genetic code) yet fell within the diversity of the taxon (**Fig. 2a, Fig. S3, *arrow***). Identical amino acid translations of the crustacean-associated virus were produced when either the standard or the ciliate genetic code were used.

As with the *P. knowlesi* library, we searched these assemblies for hypothetical proteins. From this, we identified a putative second segment in the *Machilis pallida* (Hexapoda) assembly HBDP01 containing Bristletail-associated ormycovirus 1 that was 1,619bp in length and encoded a partial ORF (HBDP01002991.1). We could not recover candidates corresponding to the remaining libraries or assemblies.

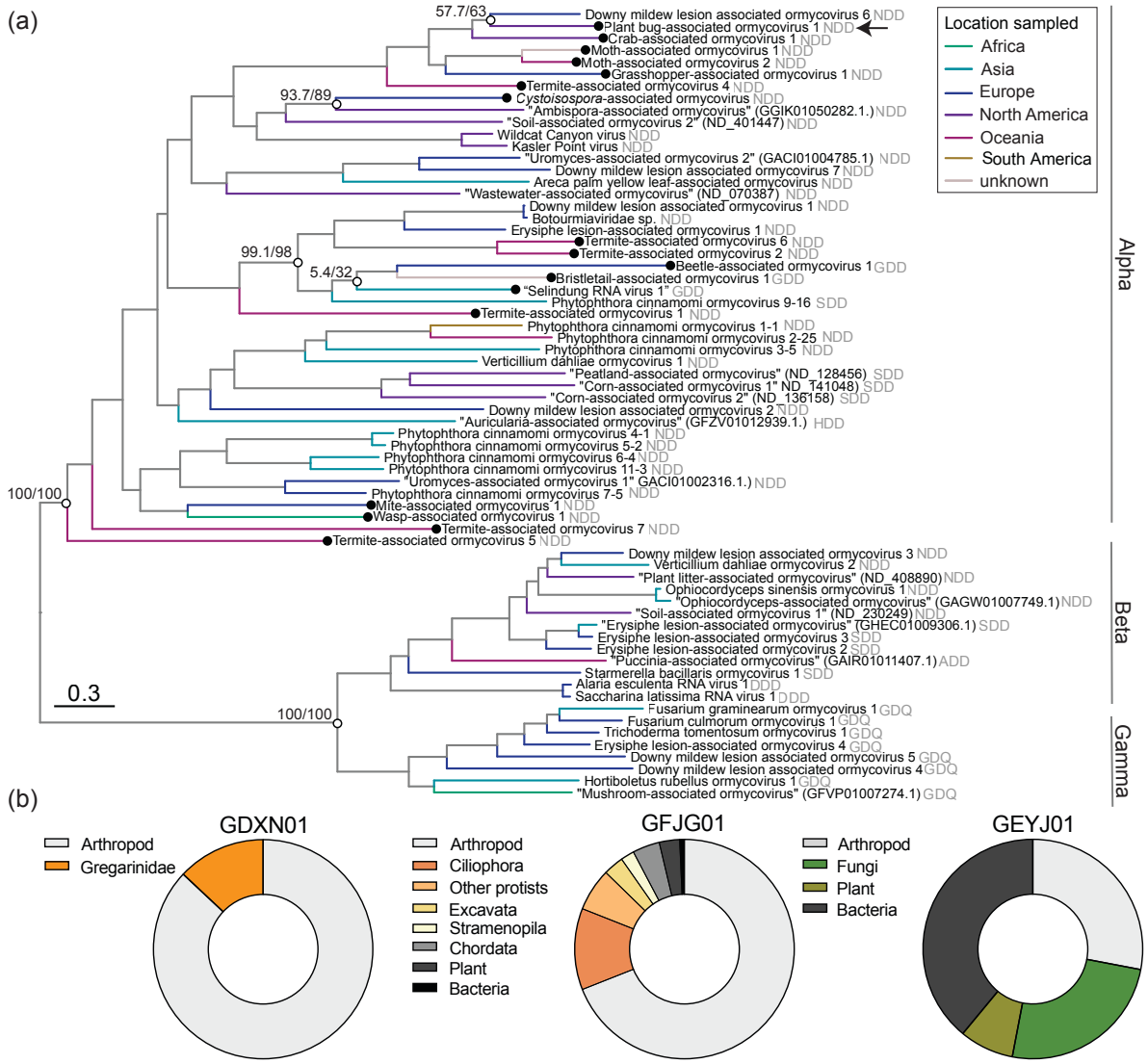

Figure 2. Ormycoviruses are associated with arthropod transcriptomes. (a) Phylogenetic inference of the extended diversity of ormycoviruses. Viruses identified in this study are indicated by black circles. The arrow indicates the position of the virus that appears to use the ciliate genetic code. Clades are annotated according to designations established by Forgia *et al*.[47]. The catalytic triad encoded in each palm domain is denoted in grey. The tip labelling scheme for unnamed viruses (denoted by quotation marks) is the same as in Fig. 1. Support values are shown at select nodes as sh-aLRT/UFBoot. Tree branches are coloured by the location where each tip was sampled, and they are scaled by amino acid substitutions. (b) Library composition of select arthropod assemblies. The graph labels correspond to the TSA project ID.

The ormycoviruses are members of a diverse and unclassified viral taxon

The wide host range of the ormycoviruses, spanning Alveolata, Stramenopila, and Opisthokonta (Fungi), suggested that this unclassified group harboured unrealised viral diversity. We therefore aimed to contextualise the diversity of the ormycoviruses within unclassified taxa identified in virus discovery studies. To do this, we assembled a custom database of the viruses identified using an artificial intelligence-based method[53] and screened

the ormycoviruses against it using DIAMOND Blastx[64]. This approach placed ormycoviruses within an unclassified taxon referred to in the original study as the proposed "SuperGroup 024"[53], a name which we will use herein.

Phylogenetic analysis illustrated that the current set of ormycoviruses represent only a fraction of the total diversity of this group as they fell throughout the phylogeny. Interestingly, the addition of the SuperGroup 024 viruses expanded the diversity of the Alpha group, scattering the original members across three sections of the tree (**Fig. 3a, *blue branches***). The Beta and Gamma clades were unchanged and characterised by a long branch at their shared base (**Fig. 3a, *green and yellow branches***).

Members of SuperGroup 024 encoded a more diverse set of catalytic triads at the motif C palm domain compared to the original ormycovirus data set[47] (**Fig 3b**). However, their addition did not lead to observable clustering of discrete motif sequences, as flexibility was observed throughout the phylogeny. Selindung RNA virus 1 again fell in a section predominated by GDD at that position (**Fig. 3a, *arrow***).

We searched the libraries containing SuperGroup 024 RdRp segments for ormycovirus hypothetical proteins. Of the 259 SRA libraries in which SuperGroup 024 RdRps were detected and assembled, we recovered hypothetical protein candidates at least 1000bp in length in 190 (73.4%). It was not possible to assign hypothetical proteins to corresponding RdRps as many libraries contained multiple RdRp segments. Despite this, our finding supports the conclusion that bisegmentation is a characteristic of viruses in this taxon and that ormycoviruses and SuperGroup 024 are one and the same.

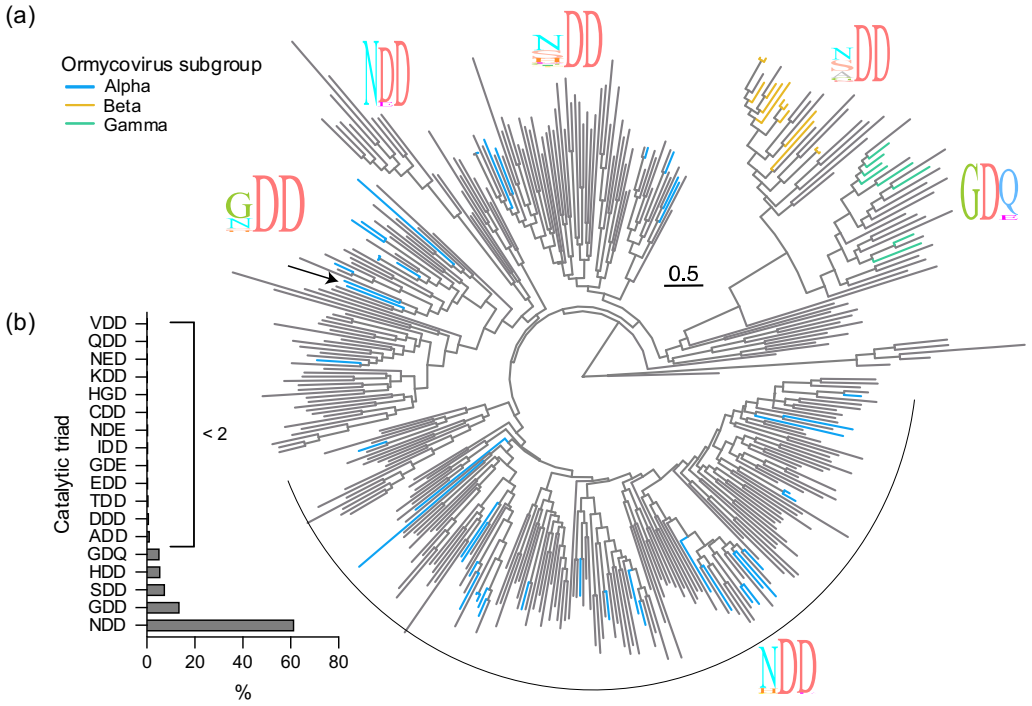

Figure 3. Ormycoviruses are members of a diverse and unclassified viral taxon with a flexible motif C in its palm domain. (a) Phylogenetic inference of viruses in SuperGroup 024[53]. Branches are coloured by their placement in the ormycovirus-only phylogenetic tree (Fig. 1 and 2). Grey tree branches indicate that those tips were not previously recognised as ormycoviruses. The icons show the proportion of individual amino acids at each position of the catalytic triad in motif C of the RdRp palm domain for the corresponding clades. The arrow indicates the topological position of Selindung RNA virus 1. Tree branches are scaled according to amino acid substitutions. (b) Distribution of catalytic triads encoded by members of SuperGroup 024. The x-axis shows the percentage that each triad comprises among all known SuperGroup 024 species.

DISCUSSION

This study expands the diversity of *Plasmodium*-associated RNA viruses and presents the first evidence of an RNA virus associated with zoonotic transmission of *P. knowlesi*. Previously, only Matryoshka RNA virus 1 (*Narnaviridae*) had been identified in a *Plasmodium* species (the human-only *Plasmodium* species, *P. vivax*)[40,41]. Although it is not possible to conclusively establish that Selindung RNA virus 1 was infecting the parasite from metatranscriptomic data alone, lines of indirect evidence suggest that it was. Most notably, no other probable hosts, including fungi, were identified in the library, and the RdRp contig was relatively abundant (1.4% of non-rRNA reads). Contamination was an unlikely source because neither the putative RdRp nor the second segment were detected in the other three libraries extracted and sequenced at the same time. In addition, we were able to confirm the presence of the RdRp segment in two of the six *P. knowlesi* isolates using PCR. We therefore concluded that Selindung RNA virus 1 most likely represents an RNA virus in a second *Plasmodium* species.

Robust sampling of natural *P. knowlesi* infections is needed to evaluate the prevalence and pathobiology of Selindung RNA virus 1. We observed one instance of Selindung RNA virus 1 among 1,470 SRA libraries, which suggests that associations occur infrequently and contrasts with the identification of Matryoshka RNA virus 1 in 13 of 30 *P. vivax* SRA libraries[40]. However, ours was the only library to have been generated from isolates collected from naturally infected humans, while most of the publicly available data were derived from laboratory experiments. The detection of the RdRp segment in two of six isolates in our library could indicate that associations are more frequent in natural infections in Sabah, but our study was not powered to assess this. Similarly, whether the observation that presence of the virus was correlated with higher parasitemia is meaningful requires further epidemiological investigation.

Arthropods are a powerful tool for measuring the prevalence of viruses in nature, particularly when sampling from humans or other vertebrates is not feasible. The identification of ormycoviruses in arthropod metatranscriptomes and in a human blood sample suggests that these viruses represent a unique type of arbovirus that can be transmitted as a passenger between arthropods and mammals. Mosquito-based surveillance methods have been proposed for tracking the incidence and spread of human pathogens[65,66]. Unlike cell culture or primary samples, which rely on symptomatic individuals with access to diagnostic testing, arthropod-based surveillance would be relatively unbiased, enabling more accurate estimates of protozoan virus prevalence and diversity within communities. When combined with cell culture data, this approach could also be used to parse arthropod- and protozoan-infecting viruses. Because they can be indirectly transmitted by arthropods, it may be that other protozoan viruses have already been identified, but their relationship to their protozoan host was obscured because they were part of an arthropod metatranscriptome.

An incidental and surprising finding was the identification of an ormycovirus that appears to use a non-standard genetic code (Plant bug-associated ormycovirus 1). Despite this difference, the virus fell within the diversity of the ormycoviruses and SuperGroup 024. As RNA viruses are reliant on host machinery for translation, it was previously proposed that the evolution of alternative genetic codes was an antiviral defence[67]. Under this assumption, the use of host-specific genetic codes by RNA viruses would imply a long-term virus-host coevolutionary relationship, and we would not expect to find viral taxa in which members use different genetic codes. Genetic code switching has been observed infrequently in the *Picornavirales* and *Lenarviricota*[68]. Whether these select instances are an aberration in an otherwise broadly held rule of virology requires further investigation. However, we posit that there may be many more instances of code switching within known viral taxa that have been overlooked as a consequence of inadequate bioinformatic workflows. For example, if we had

used an automated pipeline that filtered out contigs that did not produce an ORF with the standard genetic code, Plant bug-associated ormycovirus 1 would have been removed from our data set. We therefore advocate for the inclusion of multiple genetic codes when searching for divergent RNA viruses.

That Selindung RNA virus 1 does not belong to a known viral taxon is notable because it demonstrates that parasitic protozoa likely harbour currently unrealised diversity, and additional discoveries may be imminent as new bioinformatic tools are developed to explore the RNA virosphere. However, the discovery of the ormycoviruses highlights the importance of linking large-scale metatranscriptomic data to smaller-scale experimental work when searching for protozoan viruses. Large-scale virus discovery studies often prioritise environmental samples such as water[54], sediment[53], and soil[68] because these biodiverse sources are rich with RNA viruses. Yet, this approach cannot distinguish between bacterial-, archaeal-, and eukaryotic-infecting RNA viruses. Without the discovery of the ormycoviruses and the experimental validation by Forgia *et al*.[47], SuperGroup 024 would have been overlooked as a potential source of protozoan virus candidates. Similarly, large-scale studies are not equipped to distinguish segmented from non-segmented viruses because they necessarily focus on detecting RdRps, rendering them "blind" to segmentation. The molecular characterisation of the ormycoviruses again demonstrates this limitation because their hypothetical protein does not share detectable sequence or structural similarity with known viral proteins. Without the incidental finding by Forgia *et al*.[47], we would not have been able to infer that the ormycoviruses and the members of SuperGroup 024 are likely segmented viruses.

This, as with other metagenomic studies, primarily serves to generate hypotheses and raise questions about RNA virus evolution and biology that require additional experimental data to answer. It is not known whether the ormycoviruses are positive- or negative-sense viruses. Forgia *et al*. proposed that they are negative-sense because they observed a higher proportion of negative-sense RNA in their samples[47]; however, this is not definitive. The presence of both SDD and GDD catalytic triads in motif C in the palm domain counters the hypothesis that SDD is specific to segmented negative-sense RNA viruses[69], although it is possible that ormycoviruses do indeed fall into this category. The flexibility of the catalytic triad also raises the question of whether individual triads have a detectable impact on the biology of the virus and why flexibility is permitted in otherwise highly conserved region of the virus genome. From a global health perspective, the most important questions to address include how viral infection of *Plasmodium* affects onward *Plasmodium* transmission and the pathobiology of *Plasmodium* in humans. Additionally, which part of the parasite the virus infects and whether this could be

used as a potential drug target remain unanswered. It has already been shown that viruses can serve as a weapon against drug-resistant bacterial infections[23-25]. Whether a similar approach could be deployed to combat malaria and other disease-causing Apicomplexa should be a research priority.

METHODS

Human malaria isolates

Plasmodium RNA was isolated from cryopreserved red cells collected from 18 patients with acute malaria, enrolled in Kudat Division, Sabah, Malaysia in 2013 and 2014[15]. PCR was used to confirm *Plasmodium* species as *P. knowlesi* (n=6), *P. vivax* (n=7) and *P. falciparum* (n=5), as previously reported[40].

SRA library data sets

BioProject PRJNA589654 libraries

Plasmodium SRA libraries in BioProject PRJNA589654 (n = 4) (i.e., the BioProject that contained Matryoshka RNA virus 1) were downloaded from NCBI. Nextera adapters were trimmed using Cutadapt v.1.8.3[70] with the parameters removing 5 bases from the beginning and end of each read, a quality cutoff of 24, and a minimum length threshold of 25. The quality of trimming was assessed using FastQC v0.11.8[71]. rRNA reads were removed using SortMeRNA v4.3.3[72], and non-rRNA reads were assembled using MEGAHIT v1.2.9[73].

Disease-causing Apicomplexa libraries

We downloaded all *P. knowlesi* RNA SRA libraries of at least 0.5Gb in size available on NCBI as of August 2024 (n = 1,470). We also downloaded all RNA SRA libraries for *Cryptosporidium, Coccidia, Toxoplasmosis, Babesia,* and *Theileria* available on NCBI as of March 2024 that are at least 0.5Gb in size and generated on the Illumina platform (n = 3,162).

SuperGroup 024 libraries

To analyse the libraries containing RdRp segments of so-called SuperGroup 024[53], we first downloaded all of the contigs designated in this group by Hou *et al*.[53] (http://47.93.21.181/). We then extracted the corresponding SRA libraries from each sequence header and removed duplicates (n = 273). All but one were downloaded from NCBI. The library SRR1027962 failed repeated attempts to download, likely due to its size (99.8Gb).

Arthropod Transcriptome Shotgun Assemblies (TSA) screen

We began by screening all arthropod TSA (n = 4,864) available in August 2024, using Kasler Point virus (a tick-associated ormycovirus) as input. This screen was performed with tBLASTn

implemented in the NCBI Blast web interface (https://blast.ncbi.nlm.nih.gov/Blast.cgi). All hits were reviewed and filtered according to three criteria: (1) the contig was at least 800bp in length, (2) the contig encoded an uninterrupted ORF, (3) the contig did not return any hits to cellular genes when screened against the NCBI non-redundant (nr) database. We then aligned our filtered data set using MAFFT[74] with default parameters, and selected the most divergent virus according to the distance matrix. This virus was then used as input for an additional screen of the arthropod TSA. This process was repeated until no new contigs were identified.

Library processing

Contig assembly

For all data sets obtained from the SRA, Nextera adapters were trimmed using Cutadapt v.1.8.3[70] with the parameters described above. The efficacy of trimming was assessed using FastQC v0.11.8[71]. In total, 1,470 *P. knowlesi* libraries, 2,898 additional Apicomplexa libraries, and 259 libraries included by Hou *et al.*,[53] were successfully assembled using MEGAHIT v1.2.9[73].

Abundance estimates

The expected count of putative viral transcripts was inferred using RSEM v1.3.0[75]. For the *P. knowlesi* library containing the ormycovirus (SRR10448860), reverse-strandedness was specified to match the sequencing protocol. Default parameters were used for the remaining libraries. To infer the proportion of reads of each putative viral transcript, we calculated the total expected count for the isoforms in each library and used this value as the denominator to measure the percentage that putative viral reads comprised in the library. This analysis was performed in R v4.4.0.

Identification of divergent viruses

Polymerase segment identification

We identified Selindung RNA virus 1 using the RdRp-scan workflow[55]. Briefly, we screened the protein sequence and HMM-profile of assembled contigs from each library against a viral RdRp database. To search for additional divergent viruses, we screened all SRA libraries against the RdRp-scan database[55] and a custom database containing known ormycoviruses using DIAMOND Blastx v2.0.9[64] and the setting 'ultra-sensitive'. This database included the 39 published ormycoviruses and the Selindung RNA virus 1 RdRp segment. Only hits with e-values below 1e-07 were retained for further analysis. Contigs with hits to this database were then screened against the NCBI nr protein database to remove false positives, again using DIAMOND Blastx v2.0.9[64] and an e-value threshold of 1e-07. The parameter 'very-sensitive' was specified. Contigs that shared detectable sequence similarity to cellular genes were

485 excluded from further analysis. Nucleotide sequences were translated using Expasy
486 (https://web.expasy.org/translate/). The standard genetic code was used by default. Contigs
487 that did not return an ORF in any frame with this code were checked manually using all codes
488 available in Expasy.
489
490 *Second segment identification*
491 We first used blastn to screen libraries for contigs sharing conserved 5' and 3' termini of the
492 corresponding ormycovirus RdRp. When this did not reveal any candidates, we compiled a
493 database of all known ormycovirus second segments and used this to screen all SRA libraries
494 using DIAMOND Blastx v2.0.9[64]. Contigs that had statistically significant hits to this database
495 were checked against the NCBI nr protein database to remove false positives (i.e., cellular
496 genes). Nucleotide sequences were either translated individually with Expasy
497 (https://web.expasy.org/translate/) or with InterProScan v5.65-97.0. For sequences processed
498 with the latter, the longest translated ORFs were used for downstream analysis. To tally the
499 number of SuperGroup 024 libraries with detectable hypothetical proteins, we cross-checked
500 the presence of RdRp segments and hypothetical protein segments in each library using R
501 v4.4.0.
502
503 For the primary *P. knowlesi* library, we searched for similar sequences to those at the 5' and
504 3' termini of the RdRp segment in other contigs in the library. To do this, we extracted these
505 regions from the RdRp segment and used each as input for tblastn against the assembled
506 library (SRR10448860). To ensure that the putative Selindung RNA virus 1 hypothetical protein
507 was not present in other libraries in the same BioProject, we used this sequence as input for
508 tblastn against the three remaining libraries.
509
510 Both tblastn screens were implemented in Geneious Prime v2024.0.7 and default parameters
511 were used.
512
513 *PCR validation*
514 We first generated cDNA from the isolates using the SuperScript IV reverse transcriptase
515 (Invitrogen). These products were then used as templates for amplification with PCR.
516 Reactions were carried out in a total volume of 50ul, of which 25ul was SuperFi II (Invitrogen)
517 master mix and 1ul was the cDNA template. 2.5ul of forward and reverse primers were used
518 (**Table S1**). Reactions were performed on a thermocycler with the following conditions: 98°C
519 for 1 min followed by 35 cycles of 98°C for 10s, 60°C for 10s, 72°C for 1 min, and 72°C for 5
520 min. The PCR products were analysed on an agarose gel. We used *Plasmodium* LDHP
521 primers as the positive control.

522

523 **Library composition analysis**

524 *CCMetagen*

525 The composition of individual sequencing libraries was assessed using ccmetagen v1.2.4[59]
526 and kma v1.3.9a[60] using assembled contigs as input. The results presented in **Fig. 2b** were
527 visualised with Prism v.10.3.0.

528

529 **Protein structure inference**

530 The structure of the putative hypothetical proteins of Selindung RNA virus 1 and Erysiphe
531 lesion-associated ormycovirus 1 were predicted using AlphaFold2[56,57] implemented in the
532 Google Colab cloud computing platform. The confidence (as measured by pIDDT) of the
533 prediction was compared across five models, and the highest performing models (Selindung
534 RNA virus 1: #2, Erysiphe lesion-associated ormycovirus 1: #4) were selected for downstream
535 analysis (**Fig. S4**). To assess structural similarity, we performed a pairwise alignment of the
536 resulting pdb files of each predicted structure using FatCat[58]. All pdb files were visualised in
537 ChimeraX v1.7.1[76].

538

539 **Functional domain inference**

540 Several approaches were used to infer functional domains in the hypothetical protein, although
541 none were successful. We first performed a preliminary check with InterProScan[77], screening
542 against the CDD, NCBIfam, and TMHMM databases. This approach was implemented in
543 Geneious Prime v2024.0.7. We then employed Phyre2[78] and HHPred[79] using PDB. Finally, we
544 used the predicted structure of the hypothetical protein of Selindung RNA virus 1 as input for
545 FoldSeek[80], implemented on the Foldseek Server.

546

547 **Phylogenetic analysis**

548 To assess the phylogenetic relationships of the ormycoviruses identified in this study with
549 those documented previously, we compiled a data set of all known ormycoviruses. This
550 comprised 36 ormycoviruses[47-52] and unclassified or misclassified ormycoviruses that shared
551 detectable sequence similarity with known ormycoviruses: Wildcat Canyon virus (WZL61396),
552 Kasler Point virus (WZL61394), and a fungus-associated "Botourmiaviridae" (UYL94578). For
553 the SuperGroup 024 analysis, we utilised the data set featured in the phylogenetic analysis
554 presented by Hou *et al.*[53].

555

556 We first added the *P. knowlesi*- and *Cystoisospora*-associated viruses identified in this study
557 to the ormycovirus data set and aligned with MAFFT v7.490[81] and MUSCLE v5.1[82].
558 Ambiguities in each alignment were considered in three ways using trimAl v1.4.1[83]: (i) no

ambiguities were removed; (ii) ambiguities were removed using a gap threshold of 0.5 and a conservation percentage of 50; (iii) ambiguities were removed using the parameter "gappyout". Phylogenetic trees for these six alignments were inferred using ModelFinder and IQ-TREE v1.6.12[84]. To quantify support for the topology, we again used 1000 ultra-fast bootstraps and 1000 SH-aLRT bootstrap replicates.

To infer the pan-SuperGroup 024 phylogeny, all amino acid sequences were aligned using both MAFFT v7.490[81] and MUSCLE v5.1[82]. Ambiguities were removed using trimAl v1.4.1[83] and the parameter -gappyout. The phylogenetic tree was inferred using IQ-TREE v1.6.12[84] with ModelFinder limited to LG. Support values were measured with 1000 ultra-fast bootstraps (UFboot) and 1000 sh-aLRT bootstrap replicates.

All trees were visualised with ggtree[85,86] (implemented in R v4.4.0) and Adobe Illustrator v26.4.1.

Motif C tally and visualisation

The catalytic triad encoded by each virus in SuperGroup 024 was recorded and tabulated using R v4.4.0. The results were visualised with Prism v.10.3.0.

Sequences from individual clades were extracted from the SuperGroup 024 phylogeny by selecting individual nodes using the function "extract.clade()" implemented in the R package ape. Sequences from each clade were then realigned with MAFFT and the motif C logos were generated according to the consensus sequence in Geneious Prime v2024.0.7.

DATA AVAILABILITY

All sequencing data analysed in this study are publicly available on NCBI (ormycoviruses) and an independent repository (http://47.93.21.181/, SuperGroup 024). Assembled contigs for the viruses identified in this study, the custom database used to screen libraries, alignments, and tree files are available on GitHub (https://github.com/mary-petrone/Plasmodium_ormyco).

ACKNOWLEDGMENTS

This work was funded by a National Health and Medical Research Council (NHMRC) Investigator award (MJG), AIR@InnoHK administered by the Innovation and Technology Commission, Hong Kong Special Administrative Region, China (ECH), a Sydney ID Seed Funding Award (MEP), and the National Institutes of Health, USA R01 AI160457-01 and Malaysia Ministry of Health Grant BP00500/117/1002 (GSR). We thank the Director General of Health, Malaysia for the permission to publish this article.

We also thank Jon Mifsud for his BatchArtemisSRAMiner pipeline (https://github.com/JonathonMifsud/BatchArtemisSRAMiner), which we used for all SRA screens, and Dr. Alvin Kuo Jing Teo for suggesting the name Selindung RNA virus 1.

AUTHOR CONTRIBUTIONS

M.E.P., J.C., N.M.A., and E.C.H. designed the study. M.J.G, T.W., G.R., J.W., and K.A.P. designed the original malaria studies and collected the samples. M.E.P., J.C., and M.S. performed the experiments and analyses. M.E.P. wrote the initial manuscript draft. All authors reviewed and edited the manuscript.

REFERENCES

1 Aguirre, A. A. *et al.* The One Health Approach to Toxoplasmosis: Epidemiology, Control, and Prevention Strategies. *Ecohealth* **16**, 378-390 (2019). https://doi.org:10.1007/s10393-019-01405-7

2 Ryan, U., Fayer, R. & Xiao, L. Cryptosporidium species in humans and animals: current understanding and research needs. *Parasitology* **141**, 1667-1685 (2014). https://doi.org:10.1017/s0031182014001085

3 Krause, P. J. Human babesiosis. *Int J Parasitol* **49**, 165-174 (2019). https://doi.org:10.1016/j.ijpara.2018.11.007

4 Poespoprodjo, J. R., Douglas, N. M., Ansong, D., Kho, S. & Anstey, N. M. Malaria. *Lancet* **402**, 2328-2345 (2023). https://doi.org:10.1016/S0140-6736(23)01249-7

5 *Malaria*, <https://www.who.int/news-room/fact-sheets/detail/malaria> (2023).

6 Lee, W. C. *et al.* Plasmodium knowlesi: the game changer for malaria eradication. *Malar J* **21**, 140 (2022). https://doi.org:10.1186/s12936-022-04131-8

7 Wellems, T. E. & Plowe, C. V. Chloroquine-Resistant Malaria. *The Journal of Infectious Diseases* **184**, 770-776 (2001). https://doi.org:10.1086/322858

8 Rosenthal, P. J. *et al.* The emergence of artemisinin partial resistance in Africa: how do we respond? *Lancet Infect Dis* (2024). https://doi.org:10.1016/s1473-3099(24)00141-5

9 Blasco, B., Leroy, D. & Fidock, D. A. Antimalarial drug resistance: linking Plasmodium falciparum parasite biology to the clinic. *Nat Med* **23**, 917-928 (2017). https://doi.org:10.1038/nm.4381

10 Ehrlich, H. Y., Jones, J. & Parikh, S. Molecular surveillance of antimalarial partner drug resistance in sub-Saharan Africa: a spatial-temporal evidence mapping study. *Lancet Microbe* **1**, e209-e217 (2020). https://doi.org:10.1016/s2666-5247(20)30094-x

11 Conrad, M. D. *et al.* Evolution of Partial Resistance to Artemisinins in Malaria Parasites in Uganda. *N Engl J Med* **389**, 722-732 (2023). https://doi.org:10.1056/NEJMoa2211803

12 Fornace, K. M. *et al.* No evidence of sustained nonzoonotic Plasmodium knowlesi transmission in Malaysia from modelling malaria case data. *Nature Communications* **14**, 2945 (2023). https://doi.org:10.1038/s41467-023-38476-8

13 Anstey, N. M. *et al.* Knowlesi malaria: Human risk factors, clinical spectrum, and pathophysiology. *Adv Parasitol* **113**, 1-43 (2021). https://doi.org:10.1016/bs.apar.2021.08.001

14 Rajahram, G. S. *et al.* Deaths From Plasmodium knowlesi Malaria: Case Series and Systematic Review. *Clin Infect Dis* **69**, 1703-1711 (2019). https://doi.org:10.1093/cid/ciz011

15 Grigg, M. J. *et al.* Age-Related Clinical Spectrum of Plasmodium knowlesi Malaria and Predictors of Severity. *Clin Infect Dis* **67**, 350-359 (2018). https://doi.org:10.1093/cid/ciy065

16 Cox-Singh, J. *et al.* Plasmodium knowlesi Malaria in Humans Is Widely Distributed and Potentially Life Threatening. *Clinical Infectious Diseases* **46**, 165-171 (2008). https://doi.org:10.1086/524888

17 Cooper, D. J. *et al.* Plasmodium knowlesi Malaria in Sabah, Malaysia, 2015-2017: Ongoing Increase in Incidence Despite Near-elimination of the Human-only Plasmodium Species. *Clin Infect Dis* **70**, 361-367 (2020). https://doi.org:10.1093/cid/ciz237

18 Tobin, R. J. *et al.* Updating estimates of Plasmodium knowlesi malaria risk in response to changing land use patterns across Southeast Asia. *PLoS Negl Trop Dis* **18**, e0011570 (2024). https://doi.org:10.1371/journal.pntd.0011570

19 Brock, P. M. *et al.* Predictive analysis across spatial scales links zoonotic malaria to deforestation. *Proc Biol Sci* **286**, 20182351 (2019). https://doi.org:10.1098/rspb.2018.2351

20 Fornace, K. M. *et al.* Environmental risk factors and exposure to the zoonotic malaria parasite Plasmodium knowlesi across northern Sabah, Malaysia: a population-based cross-sectional survey. *Lancet Planet Health* **3**, e179-e186 (2019). https://doi.org:10.1016/S2542-5196(19)30045-2

21 Grigg, M. J. *et al.* Individual-level factors associated with the risk of acquiring human Plasmodium knowlesi malaria in Malaysia: a case-control study. *Lancet Planet Health* **1**, e97-e104 (2017). https://doi.org:10.1016/S2542-5196(17)30031-1

22 Anstey, N. M. & Grigg, M. J. Zoonotic Malaria: The Better You Look, the More You Find. *J Infect Dis* **219**, 679-681 (2019). https://doi.org:10.1093/infdis/jiy520

23 Hatfull, G. F., Dedrick, R. M. & Schooley, R. T. Phage Therapy for Antibiotic-Resistant Bacterial Infections. *Annu Rev Med* **73**, 197-211 (2022). https://doi.org:10.1146/annurev-med-080219-122208

24 Kortright, K. E., Chan, B. K., Koff, J. L. & Turner, P. E. Phage Therapy: A Renewed Approach to Combat Antibiotic-Resistant Bacteria. *Cell Host Microbe* **25**, 219-232 (2019). https://doi.org:10.1016/j.chom.2019.01.014

25 Strathdee, S. A., Hatfull, G. F., Mutalik, V. K. & Schooley, R. T. Phage therapy: From biological mechanisms to future directions. *Cell* **186**, 17-31 (2023). https://doi.org:10.1016/j.cell.2022.11.017

26 Barrow, P. *et al.* Viruses of protozoan parasites and viral therapy: Is the time now right? *Virol J* **17**, 142 (2020). https://doi.org:10.1186/s12985-020-01410-1

27 Zhao, Z. *et al.* Multiple Regulations of Parasitic Protozoan Viruses: A Double-Edged Sword for Protozoa. *mBio* **14**, e0264222 (2023). https://doi.org:10.1128/mbio.02642-22

28 Adjou, K. T. *et al.* First identification of Cryptosporidium parvum virus 1 (CSpV1) in various subtypes of Cryptosporidium parvum from diarrheic calves, lambs and goat kids from France. *Vet Res* **54**, 66 (2023). https://doi.org:10.1186/s13567-023-01196-4

29 Berber, E., Şimşek, E., Çanakoğlu, N., Sürsal, N. & Gençay Göksu, A. Newly identified Cryptosporidium parvum virus-1 from newborn calf diarrhoea in Turkey. *Transbound Emerg Dis* **68**, 2571-2580 (2021). https://doi.org:10.1111/tbed.13929

30 Ellis, J. & Revets, H. Eimeria species which infect the chicken contain virus-like RNA molecules. *Parasitology* **101 Pt 2**, 163-169 (1990). https://doi.org:10.1017/s0031182000063198

31 Han, Q. *et al.* Virus-like particles in Eimeria tenella are associated with multiple RNA segments. *Exp Parasitol* **127**, 646-650 (2011). https://doi.org:10.1016/j.exppara.2010.12.005

32 Lee, S. & Fernando, M. A. Viral double-stranded RNAs of Eimeria spp. of the domestic fowl: analysis of genetic relatedness and divergence among various strains. *Parasitol Res* **86**, 733-737 (2000). https://doi.org:10.1007/pl00008560

33 Lee, S. & Fernando, M. A. Intracellular localization of viral RNA in Eimeria necatrix of the domestic fowl. *Parasitol Res* **84**, 601-606 (1998). https://doi.org:10.1007/s004360050458

34 Lee, S., Fernando, M. A. & Nagy, E. dsRNA associated with virus-like particles in Eimeria spp. of the domestic fowl. *Parasitol Res* **82**, 518-523 (1996). https://doi.org:10.1007/s004360050155

35 Revets, H. *et al.* Identification of virus-like particles in Eimeria stiedae. *Mol Biochem Parasitol* **36**, 209-215 (1989). https://doi.org:10.1016/0166-6851(89)90168-0

36 Roditi, I., Wyler, T., Smith, N. & Braun, R. Virus-like particles in Eimeria nieschulzi are associated with multiple RNA segments. *Mol Biochem Parasitol* **63**, 275-282 (1994). https://doi.org:10.1016/0166-6851(94)90063-9

37 Wu, B. *et al.* Eimeria tenella: a novel dsRNA virus in E. tenella and its complete genome sequence analysis. *Virus Genes* **52**, 244-252 (2016). https://doi.org:10.1007/s11262-016-1295-0

38 Xin, C. *et al.* Complete genome sequence and evolution analysis of Eimeria stiedai RNA virus 1, a novel member of the family Totiviridae. *Arch Virol* **161**, 3571-3576 (2016). https://doi.org:10.1007/s00705-016-3020-7

39 Gupta, P. *et al.* A parasite odyssey: An RNA virus concealed in Toxoplasma gondii. *Virus Evolution* **10** (2024). https://doi.org:10.1093/ve/veae040

40 Charon, J. *et al.* Novel RNA viruses associated with Plasmodium vivax in human malaria and Leucocytozoon parasites in avian disease. *PLoS Pathog* **15**, e1008216 (2019). https://doi.org:10.1371/journal.ppat.1008216

41 Kim, A., Popovici, J., Menard, D. & Serre, D. Plasmodium vivax transcriptomes reveal stage-specific chloroquine response and differential regulation of male and female gametocytes. *Nat Commun* **10**, 371 (2019). https://doi.org:10.1038/s41467-019-08312-z

42 Hotzel, I. *et al.* Extrachromosomal nucleic acids in bovine Babesia. *Mem Inst Oswaldo Cruz* **87 Suppl 3**, 101-102 (1992). https://doi.org:10.1590/s0074-02761992000700014

43 Johnston, R. C. *et al.* A putative RNA virus in Babesia bovis. *Molecular and Biochemical Parasitology* **45**, 155-158 (1991). https://doi.org:https://doi.org/10.1016/0166-6851(91)90037-7

44 Heeren, S. *et al.* Diversity and dissemination of viruses in pathogenic protozoa. *Nat Commun* **14**, 8343 (2023). https://doi.org:10.1038/s41467-023-44085-2

45 Atayde, V. D. *et al.* Exploitation of the Leishmania exosomal pathway by Leishmania RNA virus 1. *Nat Microbiol* **4**, 714-723 (2019). https://doi.org:10.1038/s41564-018-0352-y

46 Deng, S. *et al.* Cryptosporidium uses CSpV1 to activate host type I interferon and attenuate antiparasitic defenses. *Nat Commun* **14**, 1456 (2023). https://doi.org:10.1038/s41467-023-37129-0

47 Forgia, M. *et al.* Three new clades of putative viral RNA-dependent RNA polymerases with rare or unique catalytic triads discovered in libraries of ORFans from powdery mildews and the yeast of oenological interest Starmerella bacillaris. *Virus Evolution* **8** (2022). https://doi.org:10.1093/ve/veac038

48 Dekker, R. J. *et al.* Discovery of novel RNA viruses in commercially relevant seaweeds Alaria esculenta and Saccharina latissima. *bioRxiv*, 2024.2005.2022.594653 (2024). https://doi.org:10.1101/2024.05.22.594653

49 Martyn, C. *et al.* Metatranscriptomic investigation of single Ixodes pacificus ticks reveals diverse microbes, viruses, and novel mRNA-like endogenous viral elements. *mSystems* **9**, e00321-00324 (2024). https://doi.org:doi:10.1128/msystems.00321-24

50 Niu, X. *et al.* A Putative Ormycovirus That Possibly Contributes to the Yellow Leaf Disease of Areca Palm. *Forests* **15**, 1025 (2024).

51 Pagnoni, S. *et al.* A collection of Trichoderma isolates from natural environments in Sardinia reveals a complex virome that includes negative-sense fungal viruses with unprecedented genome organizations. *Virus Evolution* **9** (2023). https://doi.org:10.1093/ve/vead042

52 Sahin, E., Edis, G., Keskin, E. & Akata, I. Molecular characterization of the complete genome of a novel ormycovirus infecting the ectomycorrhizal fungus Hortiboletus rubellus. *Arch Virol* **169**, 110 (2024). https://doi.org:10.1007/s00705-024-06027-1

53 Hou, X. *et al.* Artificial intelligence redefines RNA virus discovery. *bioRxiv*, 2023.2004.2018.537342 (2023). https://doi.org:10.1101/2023.04.18.537342

54 Zayed, A. A. *et al.* Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156-162 (2022). https://doi.org:doi:10.1126/science.abm5847

55 Charon, J., Buchmann, J. P., Sadiq, S. & Holmes, E. C. RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. *Virus Evol* **8**, veac082 (2022). https://doi.org:10.1093/ve/veac082

56 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org:10.1038/s41586-021-03819-2

57 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679-682 (2022). https://doi.org:10.1038/s41592-022-01488-1

58 Li, Z., Jaroszewski, L., Iyer, M., Sedova, M. & Godzik, A. FATCAT 2.0: towards a better understanding of the structural diversity of proteins. *Nucleic Acids Research* **48**, W60-W64 (2020). https://doi.org:10.1093/nar/gkaa443

59 Marcelino, V. R. *et al.* CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biol* **21**, 103 (2020). https://doi.org:10.1186/s13059-020-02014-2

60 Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* **19**, 307 (2018). https://doi.org:10.1186/s12859-018-2336-6

61 Cruz-Bustos, T. *et al.* The transcriptome from asexual to sexual in vitro development of Cystoisospora suis (Apicomplexa: Coccidia). *Sci Rep* **12**, 5972 (2022). https://doi.org:10.1038/s41598-022-09714-8

62 Palmieri, N. *et al.* The genome of the protozoan parasite Cystoisospora suis and a reverse vaccinology approach to identify vaccine candidates. *Int J Parasitol* **47**, 189-202 (2017). https://doi.org:10.1016/j.ijpara.2016.11.007

63 Bushnell, B., Rood, J. & Singer, E. BBMerge - Accurate paired shotgun read merging via overlap. *PLoS One* **12**, e0185056 (2017). https://doi.org:10.1371/journal.pone.0185056

64 Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**, 366-368 (2021). https://doi.org:10.1038/s41592-021-01101-x

65 Fauver, J. R. *et al.* Xenosurveillance reflects traditional sampling techniques for the identification of human pathogens: A comparative study in West Africa. *PLoS Negl Trop Dis* **12**, e0006348 (2018). https://doi.org:10.1371/journal.pntd.0006348

66 Grubaugh, N. D. *et al.* Xenosurveillance: a novel mosquito-based approach for examining the human-pathogen landscape. *PLoS Negl Trop Dis* **9**, e0003628 (2015). https://doi.org:10.1371/journal.pntd.0003628

67 Shackelton, L. A. & Holmes, E. C. The role of alternative genetic codes in viral evolution and emergence. *J Theor Biol* **254**, 128-134 (2008). https://doi.org:10.1016/j.jtbi.2008.05.024

68 Chen, Y. M. *et al.* RNA viromes from terrestrial sites across China expand environmental viral diversity. *Nat Microbiol* **7**, 1312-1323 (2022). https://doi.org:10.1038/s41564-022-01180-2

69 Venkataraman, S., Prasad, B. & Selvarajan, R. RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution. *Viruses* **10** (2018). https://doi.org:10.3390/v10020076

70 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3 (2011). https://doi.org:10.14806/ej.17.1.200

71 Andrews, S. *FastQC*, <https://github.com/s-andrews/FastQC> (2023).

72 Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211-3217 (2012). https://doi.org:10.1093/bioinformatics/bts611

73 Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674-1676 (2015). https://doi.org:10.1093/bioinformatics/btv033

74 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013). https://doi.org:10.1093/molbev/mst010

75 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011). https://doi.org:10.1186/1471-2105-12-323

76 Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci* **32**, e4792 (2023). https://doi.org:10.1002/pro.4792

77 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014). https://doi.org:10.1093/bioinformatics/btu031

78 Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858 (2015). https://doi.org:10.1038/nprot.2015.053

79 Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology* **430**, 2237-2243 (2018). https://doi.org:https://doi.org/10.1016/j.jmb.2017.12.007

80 van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* (2023). https://doi.org:10.1038/s41587-023-01773-0

81 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011). https://doi.org:10.1038/msb.2011.75

82 Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat Commun* **13**, 6968 (2022). https://doi.org:10.1038/s41467-022-34630-w

83 Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009). https://doi.org:10.1093/bioinformatics/btp348

84 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020). https://doi.org:10.1093/molbev/msaa015

85 Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinformatics* **69**, e96 (2020). https://doi.org:10.1002/cpbi.96

86 Yu, G., Lam, T. T., Zhu, H. & Guan, Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Mol Biol Evol* **35**, 3041-3043 (2018). https://doi.org:10.1093/molbev/msy194