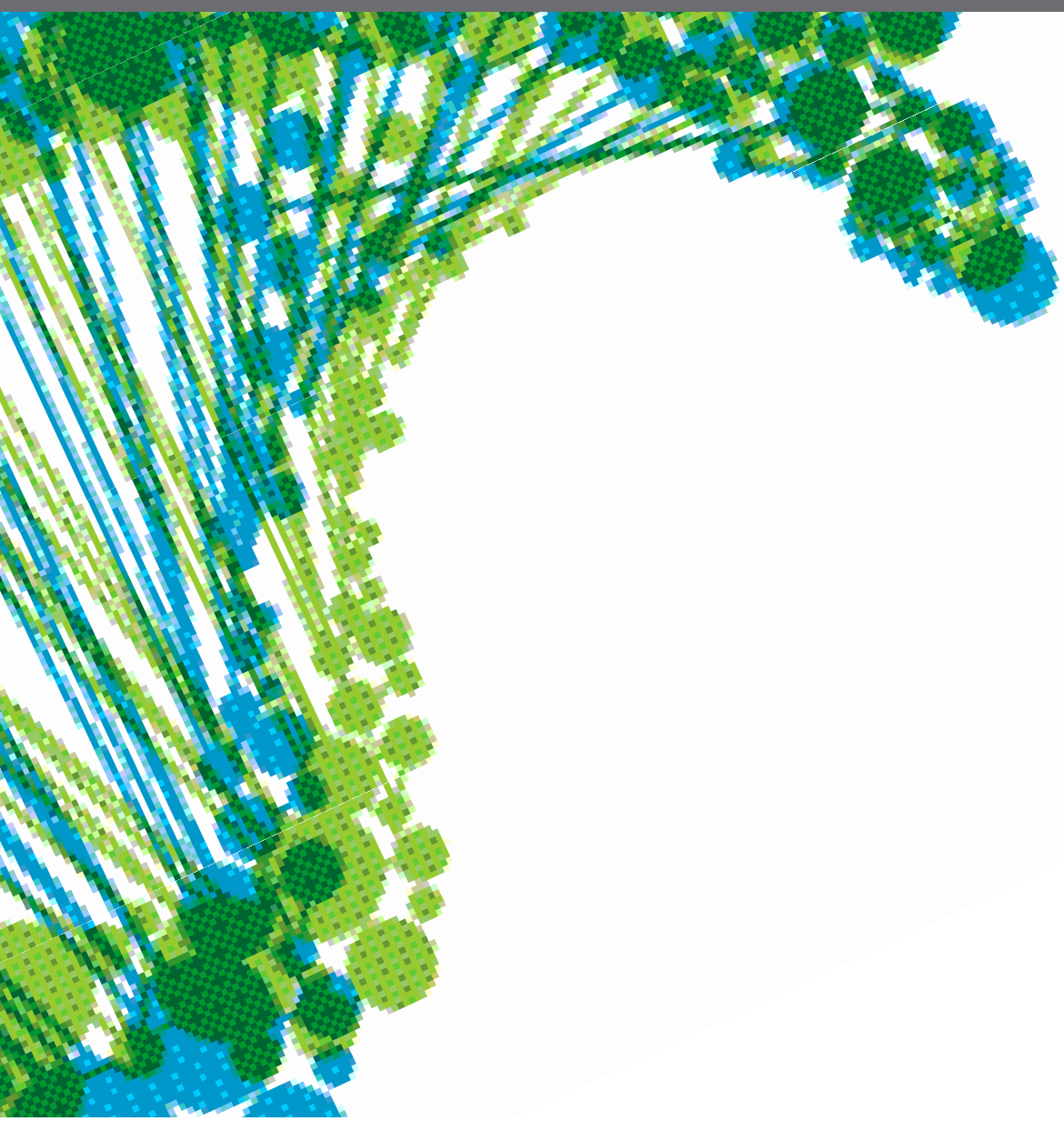
GENETICS OF APICOMPLEXANS AND APICOMPLEXAN-RELATED PARASITIC DISEASES

EDITED BY: Moses Okpeku, Matthew Adekunle Adeleke and Jun-Hu Chen
PUBLISHED IN: Frontiers in Genetics and Frontiers in Microbiology



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-814-3
DOI 10.3389/978-2-88974-814-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

GENETICS OF APICOMPLEXANS AND APICOMPLEXAN-RELATED PARASITIC DISEASES

Topic Editors:
Moses Okpeku, University of KwaZulu-Natal, South Africa
Matthew Adekunle Adeleke, University of KwaZulu-Natal, South Africa
Jun-Hu Chen, National Institute of Parasitic Diseases (China), China

Citation: Okpeku, M., Adeleke, M. A., Chen, J.-H., eds. (2022). Genetics of Apicomplexans and Apicomplexan-Related Parasitic Diseases. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-814-3

Table of Contents

frontiers in Genetics

EDITORIAL
published: 10 March 2022
doi: 10.3389/fgene.2022.875778

Editorial: Genetics of Apicomplexans and Apicomplexan-Related Parasitic Diseases

Moses Okpeku[1]*, *Matthew A. Adeleke*[1]* and *Jun-Hu Chen*[2,3,4,5]*

[1]Discipline of Genetics, School of Life Sciences, University of Kwazulu-Natal, Durban, South Africa, [2]National Institute of Parasitic Diseases, Chinese Center for Diseases Control and Prevention (Chinese Center for Tropical Diseases Research), Shanghai, China, [3]National Health Commission of the People's Republic of China (NHC) Key Laboratory of Parasite and Vector Biology, Shanghai, China, [4]World Health Organization (WHO) Collaborating Centre for Tropical Diseases, Shanghai, China, [5]National Center for International Research on Tropical Diseases, Shanghai, China

Keywords: apicomplexan-related parasitic diseases, genetics, malaria, theileriosis, tick, anopheles

Editorial on the Research Topic

Genetics of Apicomplexans and Apicomplexan-Related Parasitic Diseases

Apicomplexa are a large phylum of parasitic organisms. They are mostly unicellular and spore-forming in nature. Majority are parasites of humans, livestock, wild and domesticated animals, birds, and aquatic organisms; responsible for some parasitic diseases of serious economic importance like babesiosiosis, cryptosporidiosis, cystoisosporiasis, malaria and toxoplasmosis. Although great efforts in epidemiology, vector control, and use of medicinal drugs have been very helpful in reducing disease burden of Apicomplexa-based disease: control, prevention, and elimination of these diseases is dependent on a better understanding of their genetics, mechanisms of infection, and immunity control methods (such as drugs/vaccines and vector control methods). Furthermore, knowledge of host-vector-parasite interactions is vital for stimulating new concepts and tools for control, management, and policy formulation. This issue puts together some current researches involving apicomplexan.

The phylum Apicomplexa is entirely composed of obligate intracellular parasites, causing a plethora of severe diseases in humans, wild and domestic animals. Mitochondria are vital organelles of eukaryotic cells, participating in key metabolic pathways. The mitochondria in Apicomplexa is as a promising source of undiscovered drug targets, and it has been validated as the target of atovaquone, a drug currently used in the clinic to counter malaria, but highly diverse (Berná et al.). In this issue, Berná et al. extensively reviewed mitochondrial genomic and mechanistic features found in evolutionarily related alveolates, as well as common and distinct origins of the apicomplexan mitochondria peculiarities, with a focus on genomic insight gained from second and third generation sequencing technologies.

Malaria accounts to be responsible for large morbidity and mortality, particularly in sub-Saharan Africa where the greatest burden of the disease is borne. Chloroquine (CQ) and sulfadoxine–pyrimethamine (SP) are core drugs used in malaria control. However, a core challenge is the emergence and spread of drug resistance. Roux et al. traced the spread of CQ and SP resistance in sub-Saharan Africa, catalogued different control strategies adopted to control the spread of resistant parasites. This review suggests that close monitoring is required to prevent establishment and spread of resistance and to detection strategies for tracking possible reintroduction of drug resistance to specific anti-malarial drugs, especially as the region works towards malaria elimination. In another contribution, Yang et al. focused on *Plasmodium falciparum* surface-related antigen (SRA) located on the surfaces of gametocyte and merozoite. The SRA possess structural and functional characteristics that make them

OPEN ACCESS

Edited and reviewed by:
John R. Battista, Louisiana State University, United States

***Correspondence:**
Moses Okpeku
okpekum@ukzn.ac.za
Matthew A. Adeleke
AdelekeM@ukzn.ac.za
Jun-Hu Chen
chenjh@nipd.chinacdc.cn

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: *14 February 2022*
Accepted: *23 February 2022*
Published: *10 March 2022*

Citation:
Okpeku M, Adeleke MA and Chen J-H (2022) Editorial: Genetics of Apicomplexans and Apicomplexan-Related Parasitic Diseases. Front. Genet. 13:875778. doi: 10.3389/fgene.2022.875778

potential targets for multistage vaccine development. However, little information is available regarding the genetic polymorphism of *pfsra*. The study determined genetic variation in *pfsra* sequences generated from 74 samples collected from migrant workers who returned to China from 12 malaria endemic countries in Africa. Average pairwise nucleotide diversities (π), haplotype diversity (*Hd*), average number of nucleotide differences (*k*) of *P. falciparum sra* gene were 0.00132, 0.770, and 3.049, respectively. The ratio of non-synonymous (*dN*) to synonymous (*dS*) substitutions across sites (*dN*/*dS*) was 1. 365. Neutrality tests showed polymorphism of *pfsra* genes maintained by positive diversifying selection, implicating this gene as a potential target of protective immune responses and a vaccine candidate (Yang et al.). In a separate study from Pakistan, Khan et al. investigated the molecular diversity of *P. falciparum* based on msp-1 and msp-2 genes. The study concluded that the *P. falciparum* populations and allelic variants of msp-1 and msp-2 in the study region are highly polymorphic and diverse. HSP101 is related to host cell remodelling of malaria parasites. Kreutzfeld et al. generated transgenic *P. berghei* parasite lines that restore liver stage expression of HSP101, which were able to complete the life cycle, but failed to export PEXEL-proteins in early liver stages. The results suggest that PTEX-dependent early liver stage export cannot be restored by addition of HSP101, indicative of alternative export complexes or other functions of the PTEX core complex during liver infection.

Theileria parva is a protozoan parasite transmitted by the brown-eared ticks, *Rhipicephalus appendiculatus* and *R. zambeziensis*. The parasite has two transmission modes; cattle–cattle and buffalo–cattle transmission. Maboko et al., using Illumina HiSeq whole genome sequence data, compared genetic diversity in *T. parva* sampled from East and Southern Africa. Buffalo-derived strains had higher genetic diversity, with twice the number of variants compared to cattle-derived strains, with possible geographic sub-structuring. They conclude that *T. parva* strains circulating in South Africa are buffalo derived, and associated with corridor disease, but further work is needful for identifying host specificity (Maboko et al.), but suggested that East Coast fever (ECF) was not circulating in South African buffaloes. In a separate study, Allan et al., compared the complex interplay between *T. parva* populations in buffalo and cattle, revealing the extent of the significant genetic diversity in the buffalo *T. parva* population, the limited sharing of parasite genotypes between the host species, and highlighting that a subpopulation of *T. parva* is maintained by transmission within cattle. The results emphasize the importance of obtaining a fuller understanding of buffalo *T. parva* population dynamics in particular, for holistic understanding of *T. parva* genetics to enable a realistic assessment of the extent of buffalo-derived infection risk in cattle. *T. annulata* which causes tropical theileriosis, is a major impediment to improving cattle production in Sudan (Salih et al.). Salih et al., investigated the genetic structure of four *T. annulata* populations in Sudan revealed substantial intermixing, with only two groups exhibiting regional origin independence. The results show that the live schizont attenuated vaccine, Atbara strain may be acceptable for use in all Sudanese regions where tropical theileriosis occurs.

The distribution, genetic diversity, and zoonotic potential of *Cryptosporidium*, *Enterocytozoon bieneusi*, and *Giardia duodenalis* in wildlife are poorly understood. Zhang et al. reported the first molecular epidemiological investigation of three pathogens in wildlife from Zhejiang province and Shanghai city, China. Sequence analyses revealed five known (BEB6, D, MJ13, SC02, and type IV) and two novel (designated SH_ch1 and SH_deer1) genotypes of *E. bieneusi*. The overall results suggest that wildlife act as host reservoirs carrying zoonotic *E. bieneusi* and *G. duodenalis*, potentially enabling transmission from wildlife to humans and other animals (Zhang et al.). Cabarcas et al. performed a thorough comparison of 100+ *de novo* assembled genomes of *C. hominis*. After quality genome filtering, a comprehensive phylogenomic analysis allowed them to discover that *C. hominis* encompasses two lineages with continental segregation. These lineages were named as *C. hominis* Euro-American (EA) and the *C. hominis* Afro-Asian (AA) lineages based on the observed continental distribution bias.

Ticks are ectoparasites of humans and animals, and important vectors of apicomplenxan parasites. In this issue, Cao et al. identified 189 microbial genera and 284 species from different tick populations, using mNGS Illumina HiSeq platform, identified at taxa level (*Anaplasma*, *Rickettsia*, *Ehrlichia* species), genus (*Wolbachia*), as well as species (*Candidatus* Entotheonella) levels. The results of this study provide insights into possible profiling of known and novel tick-borne pathogens in the study region. Metagenomics next-generation sequencing (mNGS) has been suggested to be useful for rapid and accurate characterization of microorganism abundance and diversity. The promotion of NGS usage in apicomplexan research holds promises for advancement in apicomplexan profiling and characterization of known and novel types. Understanding genetic architecture of such is key in combating diseases transmitted by these myriads of microbes that co-habit the planet with us. *Anopheles coluzzii* is a major African malaria vector in the Gambiae species complex. Holm et al. used a massively parallel reporter assay, Self-Transcribing Active Regulatory Region sequencing (STARR-seq), to generate the first comprehensive genome-wide map of transcriptional enhancers in *A. coluzzii* and reported a catalog of 3,288 active genomic enhancers. This enhancer catalogue will be valuable in identifying regulatory elements that could be employed in vector manipulation.

Apicomplexan-related parasitic diseases are ancient diseases, great deal of effort invested into control and prevention of the diseases from the globe has seen measurable success, but the diseases persist. The unique multifaceted nature of the apicomplexan parasites, the potential to infect and cross infect more than one host has made viable vaccine intervention a difficult task. The

deployment of genomic techniques is yielding useful information that can be explore for predicted allele-based vaccines, concocted into multi-target vaccine coupled with potent drugs is promising.

AUTHOR CONTRIBUTIONS

MO, MA, and J-HC drafted the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the National Research and Development Plan of China (Grant No. 2020YFE0205700 and 2018YFE0121600), the National Sharing Service Platform for Parasite Resources (Grant No. TDRC-2019-194-30) and National Research Foundation (NRF) South Africa (Grant No 120368). The funding bodies had no role in the design of the study, in collection, analysis, and interpretation of data, or in writing the manuscript.

ACKNOWLEDGMENTS

We appreciate all hard working scientist who received and responded to our call for submission to this collection. We are also indebted to all the selfless reviewers who contributed of their valuable time and energy to ensure timeous review and feedbacks that went into improving every single manuscript published in this collection.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 June 2021
doi: 10.3389/fgene.2021.666096

South African Buffalo-Derived *Theileria parva* Is Distinct From Other Buffalo and Cattle-Derived *T. parva*

Boitumelo B. Maboko[1,2]*, Kgomotso P. Sibeko-Matjila[2], Rian Pierneef[3], Wai Y. Chan[3], Antoinette Josemans[1], Ratselane D. Marumo[1], Sikhumbuzo Mbizeni[1,4], Abdalla A. Latif[5] and Ben J. Mans[1,2,6]*

[1] Agricultural Research Council, Onderstepoort Veterinary Research, Pretoria, South Africa, [2] Department of Veterinary Tropical Diseases, University of Pretoria, Pretoria, South Africa, [3] Agricultural Research Council, Biotechnology Platform, Pretoria, South Africa, [4] Department of Agriculture and Animal Health, University of South Africa, Pretoria, South Africa, [5] School of Life Sciences, University of KwaZulu Natal, Durban, South Africa, [6] Department of Life and Consumer Sciences, University of South Africa, Pretoria, South Africa

OPEN ACCESS

Edited by:
Jun-Hu Chen,
National Institute of Parasitic Diseases, China

Reviewed by:
Isaiah Obara,
Freie Universität Berlin, Germany
Erik Bongcam-Rudloff,
Swedish University of Agricultural Sciences, Sweden

***Correspondence:**
Boitumelo B. Maboko
bpule@arc.agric.za
Ben J. Mans
mansb@arc.agric.za

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: 09 February 2021
Accepted: 27 May 2021
Published: 25 June 2021

Citation:
Maboko BB, Sibeko-Matjila KP, Pierneef R, Chan WY, Josemans A, Marumo RD, Mbizeni S, Latif AA and Mans BJ (2021) South African Buffalo-Derived Theileria parva Is Distinct From Other Buffalo and Cattle-Derived T. parva. Front. Genet. 12:666096. doi: 10.3389/fgene.2021.666096

Theileria parva is a protozoan parasite transmitted by the brown-eared ticks, *Rhipicephalus appendiculatus* and *Rhipicephalus zambeziensis*. Buffaloes are the parasite's ancestral host, with cattle being the most recent host. The parasite has two transmission modes namely, cattle–cattle and buffalo–cattle transmission. Cattle–cattle *T. parva* transmission causes East Coast fever (ECF) and January disease syndromes. Buffalo to cattle transmission causes Corridor disease. Knowledge on the genetic diversity of South African *T. parva* populations will assist in determining its origin, evolution and identify any cattle–cattle transmitted strains. To achieve this, genomic DNA of blood and *in vitro* culture material infected with South African isolates (8160, 8301, 8200, 9620, 9656, 9679, Johnston, KNP2, HL3, KNP102, 9574, and 9581) were extracted and paired-end whole genome sequencing using Illumina HiSeq 2500 was performed. East and southern African sample data (Chitongo Z2, Katete B2, Kiambu Z464/C12, Mandali Z22H10, Entebbe, Nyakizu, Katumba, Buffalo LAWR, and Buffalo Z5E5) was also added for comparative purposes. Data was analyzed using BWA and SAMtools variant calling with the *T. parva* Muguga genome sequence used as a reference. Buffalo-derived strains had higher genetic diversity, with twice the number of variants compared to cattle-derived strains, confirming that buffaloes are ancestral reservoir hosts of *T. parva*. Host specific SNPs, however, could not be identified among the selected 74 gene sequences. Phylogenetically, strains tended to cluster by host with South African buffalo-derived strains clustering with buffalo-derived strains. Among the buffalo-derived strains, South African strains were genetically divergent from other buffalo-derived strains indicating possible geographic sub-structuring. Geographic sub-structuring was also observed within South Africa strains. The knowledge generated from this study indicates that to date, ECF is not circulating in buffalo from South Africa. It also shows that *T. parva* has historically been present in buffalo from South Africa before the introduction of ECF and was not introduced into buffalo during the ECF epidemic.

Keywords: *Theileria parva*, whole genome sequencing, cattle, buffalo, genetic diversity, South Africa

INTRODUCTION

Theileria parva is a protozoan parasite transmitted by brown-eared ticks, *Rhipicephalus appendiculatus* and *Rhipicephalus zambeziensis* (Blouin and Stoltsz, 1989). Cape buffaloes are considered the parasite's ancestral host, with cattle being a more recent host (Norval et al., 1991). The parasite has two transmission modes namely, cattle–cattle and buffalo–cattle transmission. Cattle–cattle *T. parva* transmission by ticks causes East Coast fever (ECF) and January disease syndromes (Theiler, 1904; Lawrence, 1979). East Coast fever is endemic in different East and southern African countries, notably Kenya, Rwanda, Tanzania, Uganda, Malawi, Mozambique, and Zambia (Lawrence et al., 1992, 2017). The name East Coast fever is derived from the historical origin of the disease in the East Coast of Africa (Theiler, 1904). South Africa had ECF outbreaks in the early 1900s that caused extensive mortalities before its eradication in the country in the 1950s (Potgieter et al., 1988; Norval et al., 1992). January disease is only reported in Zimbabwe (Lawrence, 1992; Geysen et al., 1999). These diseases cause major cattle and financial losses with January disease's seasonality making it the milder of the two diseases due to lesser mortalities (Matson, 1967; Lawrence, 1979; Norval et al., 1992).

Buffalo–cattle transmission causes a disease known as Corridor disease. The name is derived from an outbreak reported between the corridor of the then Hluhluwe and Imfolozi nature reserves (Neitz et al., 1955). Currently, Corridor disease is endemic in certain regions of South Africa, Mpumalanga, Limpopo, and KwaZulu-Natal provinces where cattle and game share grazing land (Potgieter et al., 1988; Stoltsz, 1989; Mbizeni et al., 2013). It is very lethal to cattle to an extent that parasites rarely reach the tick-infective piroplasm stage in this host. If piroplasms are present, they are not regularly transmitted to cattle (Young et al., 1973; Uilenberg, 1999; Morrison, 2015). Barnett and Brocklesby (1966) showed that passage of buffalo-derived *T. parva* in cattle transforms the parasite to become cattle–cattle transmitted *T. parva* (Lawrence, 1992). Conversely, Potgieter et al. (1988) could not reproduce the transformation of South African buffalo-derived *T. parva* to resemble ECF strains. The latter study suggested that possible selection of cattle infective strains instead of behavioral change of the parasite may have been responsible for what was reported as "transformation."

Recovery from *T. parva* infection can result in cattle being immune to specific parasite strains, but remain piroplasm carriers that are able to transmit the parasite to other susceptible hosts, a phenomenon known as carrier status (Young et al., 1981; Kariuki et al., 1995). The presence of carrier animals means year-round sources of parasites that can be picked up by different life stages of the vector tick, resulting in year round outbreaks. The carrier state is therefore a mode of the parasite to maintain itself in the field (Young et al., 1986; Kariuki et al., 1995; Latif et al., 2001). Naturally recovered life-long carriers of *T. parva* are common in African countries where cattle–cattle transmission is common (Kariuki et al., 1995). Asymptomatic *T. parva* positive cattle (Yusufmia et al., 2010) as well as strains with antigenic gene sequences identical to the Muguga isolate (Sibeko et al., 2010, 2011) have been reported in KwaZulu-Natal, raising concerns of the possible circulation of ECF strains in South Africa. The animals, however, lost their parasite load approximately 120 days post infection and could not transmit the parasite (Mbizeni et al., 2013).

There is genetic evidence that phenotypic differences between cattle-derived and buffalo-derived parasite strains have an underlying genetic component. A deletion in the p67 gene (allele 1) has been associated with strains of cattle origin, whereas lack of that deletion indicated buffalo origin (allele 2) (Nene et al., 1996). The observation that cattle-derived *T. parva* showed genotypes found in buffalo was first observed in South Africa (Sibeko et al., 2010) later confirmed by Mukolwe et al. (2020) and Mwamuye et al. (2020) using strains from various countries. Sibeko et al. (2010) found both alleles and additional ones in South African strains, irrespective of host origin indicating that South African populations are different from East African populations. Thus, the p67 gene could not differentiate host-derived *T. parva* in South Africa. On the other hand, 200 random genes selected by Hayashida et al. (2013) showed phylogenetic differentiation of the different host strains.

Characterization of *T. parva* has been extensively done using various loci specific molecular tools. Bishop et al. (2001) used PIM and p104 to identify different cattle-derived *T. parva* isolates, which was supported by Norling et al. (2015) using whole genome sequencing. In addition, Pelle et al. (2011) found that a less diverse population subset of *T. parva* is maintained in cattle for cattle–cattle transmission. Microsatellites have found buffalo-derived strains to be geographically diverse and the strains to be distinct from cattle-derived strains. They have also shown evidence of allele sharing among hosts in different countries (Oura et al., 2011; Elisa et al., 2014; Hemmink et al., 2018).

Next generation sequencing (NGS) has opened various avenues for research into genomic diversity. In *T. parva* for example, NGS allowed access to additional markers to characterize the parasite and discriminate ECF and Corridor disease on a genome-wide scale (Gardner et al., 2005; Hayashida et al., 2013; Norling et al., 2015). In spite of the distinction between cattle-derived and buffalo-derived strains, regions responsible for this are yet to be identified. It is thus important to identify genetic loci linked to buffalo-transmitted *T. parva* (Corridor disease) only and not cattle–cattle transmitted *T. parva* (January disease and ECF) in a region like South Africa. This may also allow the discovery of genes responsible for host adaptation.

Currently, *T. parva* whole genome sequences have been generated for eastern and southern African (ESA) strains only. While this is an important region for ECF where cattle–cattle transmission is predominant, these genomic sequences only capture genetic diversity in a subset of the geographical range of *T. parva*. To expand our understanding of the genetic diversity of *T. parva*, we have sequenced the genomes of South African strains where the primary transmission (buffalo–cattle) is different.

MATERIALS AND METHODS

Sample History and Background

Samples used in this study were from ARC-OVR *in vitro* biobank, field collections and on-going research (**Table 1**).

Hluhluwe 3 (HL3)

Engorged and unengorged *Rhipicephalus appendiculatus* adults, larvae and nymphs were collected from buffaloes at Hluhluwe Game Reserve and divided into batches from which tick stabilates were made. Theilerial isolations from engorged larvae and nymphs were put in batch 3 and called isolate 3. Initially, HL3 was maintained in the laboratory by cattle–cattle passages (De Vos, 1982; Potgieter et al., 1988) and later maintained through *in vitro* culturing. The *in vitro* biobank reference for HL3 used in this study was 1/1/6/607.

8160

Bovine 9433 was infected with the buffalo-derived HL3 *T. parva* strain through a blood transfusion from bovine 9266. In 2006, *Rhipicephalus zambeziensis* nymphal ticks fed on bovine 9433, picking up the parasite and transmitting it to bovine 8160. *In vitro* cultures of 8160 were subsequently initiated and used in this study. It is a third generation cattle-derived strain originating from HL3. The *in vitro* biobank reference for 8160 used in this study was 1/12/2/826.

TABLE 1 | Source of samples used in the genomic characterization of *Theileria parva*.

Strain	Accession	Host	Location/ Province[a]	Origin
8160	SAMN18117497	Bovine	Hluhluwe Nature Reserve/KZN	Game reserve buffalo (CD endemic zone)
8301	SAMN18117498	Buffalo	Welgevonden Private Game Reserve/LP	Pickup and transmission from buffalo to bovine
8200	SAMN16622817	Bovine	Ithala Game Reserve/KZN	Pickup and transmission from buffalo to bovine
9620	SAMN16622818	Bovine pastures	Pongola/KZN	Corridor disease outbreak in bovines
9656	SAMN18117499	Bovine pastures	Nyalisa/KZN	Corridor disease outbreak in bovines
9679	SAMN16622819	Bovine	Pongola/KZN Private Game Reserve	Game reserve buffalo (CD endemic zone)
Johnston	SAMN18117500	Bovine	Ladysmith/KZN	Corridor disease outbreak in bovines
KNP2	SAMN18117501	Buffalo	Kruger National Park/LP_MP	Game reserve buffalo (CD endemic zone)
HL3	SAMN18117502	Buffalo	Hluhluwe Nature Reserve/KZN	Game reserve buffalo (CD endemic zone)
KNP102	SAMN16622805	Buffalo	Kruger National Park/LP_MP	Game reserve buffalo (CD endemic zone)
9574	SAMN16622808	Bovine	Welgevonden Private Game Reserve/LP	Corridor disease outbreak in bovines
9581	SAMN18117503	Buffalo	Belvedere Game Reserve/KZN	Game reserve buffalo (CD endemic zone)

[a]*KZN, KwaZulu-Natal; LP, Limpopo; MP, Mpumalanga.*

8301

Rhipicephalus appendiculatus ticks that had previously fed on two buffaloes from Welgevonden Private Game Reserve in 2004 were fed on bovine 9288 (Sibeko et al., 2008). *Rhipicephalus zambeziensis* ticks then picked up *T. parva* from 9288 in 2006 and transmitted it to bovine 8301. *In vitro* cultures of 8301 were subsequently initiated and used in this study. It is a second generation cattle-derived strain originating from the Welgevonden buffalo. The *in vitro* biobank reference for 8301 used in this study was 1/4/4/818.

9574

Rhipicephalus appendiculatus ticks that had picked up *T. parva* from bovine carrier 8301 fed and transmitted the parasite to bovine 9574 at ARC-OVR quarantine facility. It is a third generation cattle-derived strain originating from the Welgevonden buffalo. Whole blood for the current work was drawn in 2017.

8200

Bovine 8200 received adult *Rhipicephalus appendiculatus* collected as engorged nymphs from a buffalo in Ithala Game reserve in 2006. It had severe body reactions but recovered spontaneously without treatment to become a *T. parva* parasite carrier. *In vitro* cultures were subsequently initiated and used in this study. 8200 is considered a first generation cattle derived *T. parva* strain from buffalo strains. The *in vitro* biobank reference for 8200 used in this study was 1/12/2/844.

9620 and 9656

Rhipicephalus appendiculatus ticks were collected from bovine pastures with a history of Corridor disease outbreaks in Pongola in 2017 and not from a game reserve as erroneously reported by Zweygarth et al. (2020). They fed on bovines 9620 and 9656, erroneously called 9596 by Zweygarth et al. (2020) at ARC-OVR quarantine facility. *In vitro* cultures were initiated upon presentation of *T. parva* clinical symptoms. 9620 and 9656 are considered first generation cattle-derived *T. parva* strains with an unknown prior host history. Biobank references 7/14/8 and 7/14/9 were allocated to 9620 and 9656, respectively.

9581 and 9679

Rhipicephalus appendiculatus ticks collected from buffalo pastures in Pongola in 2017 and 2018, respectively, were fed on bovines 9581 and 9679 at ARC-OVR quarantine facility. *In vitro* cultures were initiated and maintained from lymph node aspirates after the bovines showed *T. parva* clinical signs. 9581 and 9679 *in vitro* cultures were subsequently initiated for the current study. 9581 and 9679 are first generation *T. parva* strains originating from buffalo pastures. Biobank references 7/14/10 and 7/18/3 were allocated to 9581 and 9679, respectively.

Johnston

A Corridor outbreak was suspected on a cattle farm near Pongola in 2010, based on observed clinical reactions. Lymph node biopsies were collected and *in vitro* cultures were subsequently initiated and used in this study (Latif and Mans, 2010). Johnston is considered a first generation cattle-derived *T. parva* strain with an unknown prior host history. The *in vitro* biobank reference for Johnston used in this study was 8/48/7/632.

KNP2

Theileria parva negative, laboratory reared *Rhipicephalus zambeziensis* fed on a naturally *T. parva* infected captive buffalo at the Kruger National Park (KNP) (Collins and Allsopp, 1999), which borders two provinces Mpumalanga and Limpopo (LP_MP). The ticks fed and transmitted the parasite to a bovine at ARC-OVR quarantine facility. *In vitro* cultures of KNP2 were then initiated upon presentation of *T. parva* clinical symptoms. KNP2 is a first generation cattle-derived *T. parva* strain originating from a buffalo carrier. The *in vitro* biobank reference for KNP2 used in this study was 7/14/4.

KNP102

Buffalo KNP102 originating from Kruger National Park in LP_MP, was donated by the South African National Parks to the University of Pretoria/Agricultural Research Council in 2004 as part of the UP/ARC BioPad consortium (Sibeko et al., 2008). It was later handed over to the ARC and has been in a tick free environment at ARC-OVR quarantine facility ever since. Whole blood for the current work was drawn in 2017. KNP102 is a zero generation buffalo-derived strain assumed to have been infected trans-placental.

Transmission experiments were done within the ARC-OVR-East Coast fever quarantine facility with ethical clearance under project 000760-Y5 and Department of Agriculture, Land reform and Rural development (DALRRD) (Section 20 12/11/1/1). *In vitro* cell cultures were established using blood and lymph aspirates from clinical animals as described with modifications (Brown, 1987). Briefly, blood and lymph aspirates were collected in heparin tubes. Lymphocytes were separated and collected using equal volume Percoll pH 8.5–9.5 (Sigma-Aldrich, United States) and washed with Dulbelcco's phosphate buffered saline (PBS) (Sigma-Aldrich, United States). Prior to *T. parva* induced cell transformation, bovine endothelial cell lines (BA 886) (Yunker et al., 1988; Zweygarth et al., 1997) were used as feeder cells to assist in parasite multiplication. HL-1 medium was supplemented with 10% heat inactivated bovine serum or sterile fetal calf serum, 2 mM L-glutamine, 100 IU/ml penicillin, 100 μg/ml streptomycin and 2.5 mg/l amphotericin B. Cultures were incubated at 37°C. CO_2 was only supplemented when cultures struggled to grow.

Theileria parva Detection

Genomic DNA was extracted from *in vitro* cell cultures and blood material using MPLC or MP96 instruments with their respective MagNa Pure LC DNA Isolation Kit – Large Volume or MagNA Pure 96 DNA and Viral Nucleic Acid Small Volume Kit (Roche Diagnostics, Mannheim, Germany), according to manufacturer's instructions. Eluted DNA was tested for *T. parva* using the Hybrid II assay (Pienaar et al., 2011). Molecular biology work was done in a South African National Accreditation System (SANAS) accredited (V0017) and Department of Agriculture, Land Reform and Rural development (DALRRD) (DAFF-30) approved laboratory.

Genomic Library Preparation

Genomic library preparations for the various samples were as described by Maboko et al. (2021).

Library preparation, hybridization capture, and sequencing were conducted at ARC-Biotechnology Platform using Agilent SureSelect^{XT} Target Enrichment System (Agilent Technologies, Santa Clara, CA, United States) for Illumina paired-end Multiplexed Sequencing Library protocol version C1 according to the manufacturer's instructions. Briefly, the sample enrichment process involved fragmentation of the DNA using Covaris E220 sonication (Covaris Inc., United States) set to duty factor 10, Peak Power 175 and 200 cycles per burst, resulting in a fragment size distribution of 125 bp. This was followed by purification, end-repair, addition of a 3′ -A and ligation using SureSelect adapters. Adapter-mediated PCR was done on the fragmented DNA prior to in-solution hybridization. Hybridization was conducted at 65°C for 16–24 h. Streptavidin Dynabeads (Invitrogen Carlsbad, CA, United States) were used to isolate the biotinylated baits with hybridized DNA fragments. Host DNA was washed away and captured DNA was once again amplified and indexed prior to sequencing. Library concentrations were determined using Qubit Flourometric Quantitation (Thermo Fisher Scientific, Life Technologies Holdings, Singapore) and LabChip GX Touch Nucleic Acid Analyzer (PerkinElmer Inc., United States) was used to estimate the library size. Datasets generated in the current study were submitted to GenBank under Bioproject PRJNA673089 with SRA database numbers SAMN18117497, SAMN18117498, SAMN16622817, SAMN16622818, SAMN18117499, SAMN16622819, SAMN18117500, SAMN18117501, SAMN18117502, SAMN16622805, SAMN16622808, SAMN18117503.

Processing, Filtering and Mapping to Reference

For our genomic data to have host and regional context and for comparative consistency, single-end Illumina reads from nine East and southern Africa (ESA) *T. parva* isolates were re-analyzed (Hayashida et al., 2013) (**Table 2**). Fastq files were downloaded from Sequence Read Archive (SRA) of the National Center for Biotechnology Information (United States of America) and processed in the same way as our sequences except for the quality trimming and mapping which were defined as single-end reads.

Data quality trimming was done by Trimmomatic (v 0.36) using a combination of SureSelect and trimmomatic-derived adapter sequences. Trimming parameters were as follows: Removed leading and trailing base quality: 4-base wide sliding window cutting when the average quality per base was below 15 (4:15). Reads below 36 bp were removed (MINLEN: 36). Sequence data quality was assessed before and after trimming

TABLE 2 | Eastern and southern African additional genomes used for the genetic diversity studies of *Theileria parva**.

Strain name	Host	Year of isolation	Place isolated	SRA accession
Chitongo Z2	Cattle	1982	Zambia	DRR002438
Katete B2	Cattle	1989	Zambia	DRR002439
Kiambu Z464/C12	Cattle	1972	Kenya	DRR002440
Mandali Z22H10	Cattle	1985	Zambia	DRR002441
Entebbe	Cattle	1980	Uganda	DRR002442
Nyakizu	Cattle	1979	Rwanda	DRR002443
Katumba	Cattle	1981	Tanzania	DRR002444
Buffalo LAWR	Buffalo	1990	Kenya	DRR002445
Buffalo Z5E5	Buffalo	1982	Zambia	DRR002446

**Data derived from Hayashida et al. (2013).*

using the FastQC (v 0.11.5) (Andrews, 2010). Trimmed fastq reads were first mapped to the bovine host *Bos taurus* (Hereford) genome (Accession DAAA00000000.2) using Burrows-Wheeler Aligner-Minimum Exact Match (BWA-MEM) (v 0.7.17). The default parameters used were: matching score of 1, mismatch penalty of 4, gap extension penalty of 1, gap open penalty of 6 and an unpaired read pair penalty of 9 were used except for minimum seed length (-k) of 23, which is the minimum base pair length of exact match. Bedtools (v 2.27.1) (Quinlan and Hall, 2010) was used to convert the unmapped *Bos taurus* bam files into paired fastq files to be mapped to the Muguga reference genome (Accession AAGK00000000) (Gardner et al., 2005) using BWA-MEM (v 0.7.17) also at a minimum seed length (-k) of 23. Sequences that had mapped multiple times were removed using Sequence Alignment/Map (SAMtools) (v 1.9) –F 2048 as they may represent repeat regions. PCR duplicate reads were identified using Picard MarkDuplicates (v 2.2.1) (Picard, 2019) and played no part in the subsequent downstream analysis. The general workflow for downstream analysis is indicated in **Figure 1**.

Variant Calling and Phylogenetic Analysis

BCFtools mpileup was used for individual and joint variant calling. Joint variant calling was for host (buffalo and cattle), regional (South Africa and ESA) origin and the whole *T. parva* populations. Default parameters of maximum base depth of 250, minimum base quality of 13 were used. SNPs were selected from other variants using GATK SelectVariants. SNP filtration with a minimum variant quality equal and above 20 phred (QUAL), read depth (DP) equal and above 10 were called using BCFtools. VCFtools (v 0.1.15) (Danecek et al., 2011) was used to determine variant and SNP counts. Strains were analyzed as haploid organisms because the parasite is haploid in the host.

SNP-derived maximum likelihood phylogenetic tree was constructed using PAUP* with 1000 bootstraps. A reticulated tree was constructed using SplitsTree 4 (v 4.16.1) (Huson, 1998) employing the following parameters: all characters with uncorrected P parameter, neighbor-joining network at equal angles (Huson and Bryant, 2006). Inkscape v 1.0.1[1] was used to visualise the tree.

[1]Harrington, B. et al (2004-2005). Inkscape. http://www.inkscape.org/.

Different strains were grouped according to host origin (buffalo and cattle). Pairwise comparison of each SNP position in individual strains was done to remove cumulative SNP counting. Shared and unique SNP positions were identified using VCFtools while SNP annotation was done using Ensembl Variant Effect Predictor (VEP) (Mclaren et al., 2016). SNP classification was based on effect of SNP on a gene, ranging from high to low. If there was no evidence of impact or it was difficult to determine whether there is an impact, it was classified as a modifier SNP. Moderate SNPs have a non-disruptive impact that might change protein effectiveness. High impact SNPs have a disruptive effect on protein function.

A total number of 74 genes of interest were selected for determining host specificity based on the following criteria (**Supplementary Table 1**):

a. Predicted impact of SNPs on the protein (high).
b. Subtelomeric variable secreted protein (SVSP) gene sequence associated with host–pathogen interaction (Schmuckli-Maurer et al., 2009).
c. Signal protein gene sequence involved in protein translocation (Hayashida et al., 2013).
d. Differential protein expression levels at different stages of the parasite life cycle (Tonui et al., 2018).

Our study's strain gene sequences were obtained by searching against the Muguga sequences using BLASTN (Altschul et al., 1990). In this study, host specific SNP genes were defined as occurring in one host and not the other and are not shared among the hosts.

Mixed Populations

To assess whether there were multiple strains in individual samples (mixed infections); a *de novo* assembly was done. Input files (fastq files) were mapped to the full-length assembled consensus sequences in order to identify the extent of intra-strain variation, which would indicate mixed infections (Mans et al., 2019). Paired-end, trimmed and host *Bos taurus* DNA filtered reads as well as ESA acquired reads were assembled using CLC Genomics Workbench (v 9.5.2) (Qiagen, Hilden, Germany) using the standard *de novo* de Bruijn graph assembly algorithm implemented in this version. The parameters used were as follows: minimum contig length: 200, mismatch cost: 2, insertion cost: 3, deletion cost: 3, length fraction: 0.5, similarity fraction: 0.8. Contig regions with low coverage were removed and the remaining ones were joined to generate final consensus sequences. Mapping parameters of reads to the assembled consensus sequences were as follows: match score: 1, mismatch cost: 2, cost of insertions and deletions: linear gap cost, insertion cost: 3, deletion cost; 3, length fraction: 0.5, similarity fraction: 0.8. Variant calling parameters were as follows: ploidy: 1, minimum coverage: 20, minimum count: 2, % minimum frequency: 35, neighborhood radius: 5, minimum central quality: 20, minimum neighborhood quality: 15.

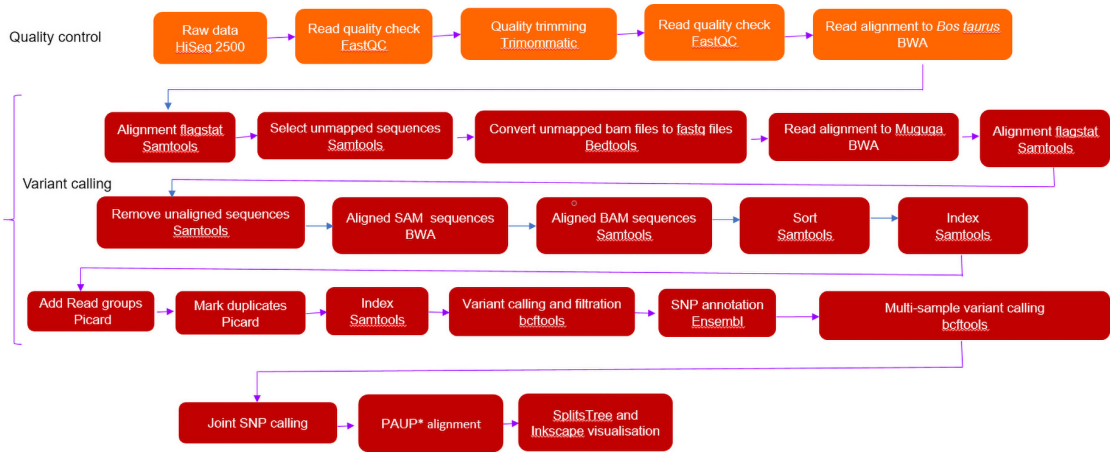


FIGURE 1 | Sequential analysis workflow used for whole genome studies of *Theileria parva*. SplitsTree and Inkscape visualisation.

Population Genetic Diversity

To explore the levels of regional and host population genetic diversity, strains were separated according to host and region origin. DNASP (v 6.12.03) (Rozas et al., 2017) was used to determine the average pairwise nucleotide diversity (Pi) defined as the average number of nucleotide differences per site between pairs of DNA sequences and the fixation index (F_{ST}) which measures the extent of DNA divergence between populations. F_{ST} values were interpreted as follows: low (0–0.05), intermediate (0.05–0.15), high (0.15–0.25) and very high (greater than 0.25) (Amzati et al., 2019).

RESULTS

Mapping and Alignment Efficiency

Removal of contaminating host-derived reads by mapping to the *Bos taurus* genome led to unacceptable levels of incorrect mapping (loss of *T. parva* reads) when using default parameters (minimum seed length of 19). Adjustment of the BWA's minimum seed length to 23 resulted in a reduced number of reads that incorrectly mapped to the host (*Bos taurus)*. Blood extracted strains (KNP102 and 9574) had the highest bovine sequence contamination 69% and 67%, respectively (**Table 3**). Removal of host-derived reads led to increased mapping percentages of reads to the parasite reference genome. Blood extracted strains also had the lowest number of mapped reads as well as average mapping coverage to the reference compared to *in vitro* culture samples. Active removal of contaminating bovine reads led to an increased mapping efficiency of approximately 20% for LAWR and Z5E5 whilst 8200 had a 50% increase. Cattle-derived strains had an improvement of 40%–66% increase with the exception of Chitongo, which had an increase of only 6%. The extent of mixed populations among South African strains was at 6% or below except for two field samples 9679 and 9581 at 86% and 77%, respectively. KNP102, kept in a tick-free environment, did not show any mixed infection (**Table 4**) compared to ESA strains, which had consistent extensive mixed infections except for Entebbe at 1% (**Table 5**).

South African strains had higher mapped reads percentage due to larger read length and paired-end sequencing with genome coverage at over 88%, allowing confirmation of the identity of strains as *T. parva*. The lowest average mapping coverage was at 23×, within range of good variant calling (**Table 3**).

The SNP alignment was 541,524 bp with 541,524 informative sites.

Population Structure of South Africa *T. parva*

The South African *T. parva* strains had SNP numbers ranging from 58,684 to 192,411 compared to the Muguga reference genome (**Table 3**). South African strains clustered together with no differentiation between those collected from buffalo and cattle pastures (**Figure 2**). Most of the strains used in our study are from the KwaZulu-Natal province and these strains clustered together, with the top branch being a mixture of KwaZulu-Natal and a Limpopo/Mpumalanga strains (**Figure 2**). Strains isolated from bovine carrier animals (8200 and 9574), also clustered with other buffalo strains. 8160 is a Hluhluwe-Imfolozi (HL3) buffalo strain isolated from a bovine and the two strains (8160 and HL3) clustered together; similarly 8301 and 9574 that clustered together, which are second and third generations of the Welgevonden strain.

Eastern and southern African and South African strains mapped to 88.46%–97.54% of the reference genome. However, the South African strains had close to 3× more variants than ESA *T. parva* strains from the 97% of RSA and 99% of the ESA strains that were variable (segregating sites) (**Table 6**). Nucleotide diversity (Pi) for all strains were low across all analysis areas (geographic and host origin) but South Africa and buffalo strains tended to have higher values. Genome-wide F_{ST} values of 0.3 among South Africa and ESA supports differentiation between the two groups. All strains were distinct to an extent that haplotypes could not be determined.

TABLE 3 | Mapping and sequencing statistics of different *Theileria parva* strains used in this study.

Strain	Total paired reads before quality trimming	Total paired reads after quality trimming	Percentage mapped reads to *Bos taurus*	Total paired reads used for reference mapping (Muguga)	Number of reads mapped to reference (Muguga)	Percentage mapped reads to reference (Muguga)	Average mapping coverage (×)	Genome coverage (%)	Total variant number	Total SNPs
8200	8,938,734	8,742,536	22.06	6,669,252	6,482,174	97.19	92.366	96.61	175,224	171,748
9620	7,279,672	6,721,550	6.82	6,225,554	6,079,644	97.66	84.967	93.09	138,470	136,186
9679	7,874,554	7,062,084	8.29	6,303,390	6,111,848	96.96	85.896	88.46	66,930	65,947
KNP102	7,455,982	7,206,184	68.85	1,685,168	1,564,764	92.86	22.593	95.31	59,233	58,564
9574	8,232,602	8,028,762	67.42	2,545,358	2,419,424	95.05	34.808	93.80	120,401	118,244
8160	168,501,398	158,906,702	12.18	139,139,772	136,202,234	97.89	1917.728	97.22	196,548	192,411
KNP2	7,833,900	7,106,360	13.99	5,861,456	8,726,194	97.69	80.011	92.99	154,524	151,851
HL3	22,394,860	22,015,858	3.76	21,107,322	18,568,916	87.97	268.034	97.52	184,018	180,420
9581	8,060,570	7,870,172	11.23	6,873,992	6,651,394	96.76	94.762	97.54	94,898	93,343
9656	7,325,764	6,772,496	4.62	6,424,462	6,291,954	97.94	87.733	92.11	146,306	143,909
8301	11,114,236	10,223,658	4.87	9,587,565	9,457,338	97.85	131.847	94.44	159,577	156,709
Johnston	8,990,948	8,794,236	14.11	7,199,810	7,044,192	97.84	100.361	94.53	160,811	158,005
Chitongo	Unknown	14,405,285	10.80	12,849,208	10,179,936	79.23	43.871	94.63	23,786	23,749
Entebbe	Unknown	10,171,312	54.32	4,645,961	3,164,429	68.11	13.637	91.75	7,324	7,323
Katete	Unknown	16,558,765	62.31	6,241,639	4,528,707	72.56	19.517	94.65	12,951	12,946
Katumba	Unknown	35,406,725	84.48	5,495,080	3,727,410	67.83	16.064	94.41	10,759	10,745
Kiambu	Unknown	15,848,447	52.69	7,497,851	5,713,477	76.20	18.948	95.16	16,419	16,409
Mandali	Unknown	16,362,287	70.00	4,909,394	3,539,263	72.09	15.253	93.77	8,976	8,971
Nyakizu	Unknown	7,212,228	1.73	7,087,364	5,542,766	78.21	23.887	94.55	18,543	18,506
LAWR	Unknown	17,072,360	38.72	10,461,767	4,803,551	45.92	20.701	89.29	45,919	45,880
Z5E5	Unknown	14,821,054	39.59	8,953,444	3,968,924	44.33	17.104	90.08	36,856	36,819

TABLE 4 | Genome features of South African *T. parva* strains.

	8200	9620	9679	KNP102	9574	8160	KNP2	HL3	9581	9656	8301	Johnston
Contig count	1,445	6,517	9,660	7,541	3946	1,299	3,935	4,939	12,482	5,806	3,676	3,410
N50 (mean length, bp)	14,679	1,885	918	1,355	3762	63,285	3,276	13,127	1,085	1,989	3,626	4,159
GC content (%)	34.24	35.03	35.43	34.71	34.83	34.28	34.82	39.10	34.89	35.07	34.69	34.63
Max. contig length	74,211	20,398	12,283	10,572	26,418	252,906	21,180	125,690	14,559	18,989	20,552	31,931
Min. contig length	200	150	151	118	188	176	184	150	118	190	126	195
Total contig length >1000 bp	7,913,602	5,668,460	3,050,776	4,081,617	6,779,720	8,260,230	6,792,895	17,845,428	4,571,833	5,586,198	7,032,359	7,099,677
Total contig length of all contigs	8,172,769	7,633,869	6,474,760	6,490,592	7,741,301	8,609,709	7,666,245	19,053,318	8,765,810	7,293,380	7,830,966	7,841,276
Total variants	1712 (1%)	8918 (6%)	57628 (86%)	0	2757 (2%)	3166 (2%)	2776 (2%)	3853 (2%)	73094 (77%)	3079 (2%)	2597 (2%)	2310 (1%)
SNPs	1,341	7,678	54,200	0	2,184	2,572	2,113	3,101	68,010	2,305	1,953	1,794

Values in brackets represent extend of mixed variants (%) to the total number of Muguga mapped variant number in ***Table 3****.*

Host Differentiation

Buffalo-derived strains (mostly from South Africa) had 6× more SNPs than the cattle-derived (**Table 5**) with different strains clustering according to host origin. ESA buffalo-derived strains clustered with South African buffalo-derived strains (**Figure 2**).

A total of 679 SNPs were specific to cattle, 65 SNPs were shared by the two hosts and 2499 were specific to buffaloes (**Figure 3**). These values include strains with mixed infections. Only two high impact SNPs were found among all identified SNPs, poly (A) polymerase and a hypothetical protein (TP02_0956), both from buffalo-derived *T. parva*. However, gene and protein multiple sequence alignments did not support the high impact classification due to limited variation between cattle- and buffalo-derived sequences.

TABLE 5 | Genome features of eastern and southern African *T. parva* strains.

	Chitongo	Entebbe	Katete	Katumba	Kiambu	Mandali	Nyakizu	LAWR	Z5E5
Contig count	4,260	9,402	7,553	8,473	6,070	8,933	4,159	5,028	7,113
N50 (mean length, bp)	4,670	608	1,524	2,824.11	2,151	1,004	4,840	3,358	1,686
GC content (%)	34.40	37.22	34.95	34.50	34.83	35.72	34.19	34.47	34.95
Maximum contig length	42,436	5,469	21,314	17,455	21,136	14,971	48,207	43,529	18,599
Minimum contig length	199	200	200	1,367	200	200	200	199	199
Total contig length >1000 bp	6,787,839	1,355,599	4,780,159	4,380,612	5,504,375	3,182,192	6,663,018	6,129,843	4,885,893
Total contig length of all contigs	7,886,524	4,931,777	7,274,837	7,306,660	7,356,218	6,350,803	7,802,537	7,582,538	7,185,596
Total variants	7846 (1%)	7850 (33%)	8533 (66%)	9081 (84%)	8688 (53%)	9107 (101%)	8627 (47%)	6505 (14%)	8141 (22%)
SNPs	6,471	6,253	7,030	7,327	7,054	7,355	7,141	5,453	6,705

Values in brackets represent extend of mixed variants (%) to the total number of Muguga mapped variant number in ***Table 3****.*

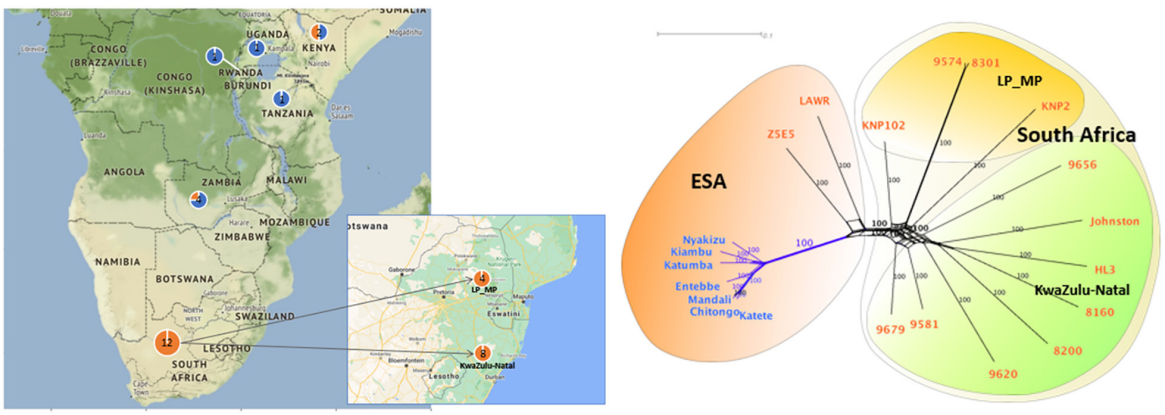


FIGURE 2 | SNP phylogenetic and geographical relationships of *T. parva* strains across selected South and East African countries. Blue font represents cattle-derived strains. Orange font represents buffalo-derived strains. LP stands for Limpopo. MP stands for Mpumalanga. ESA stands for eastern and southern Africa. The phylogenetic tree was calculated by maximum likelihood analysis using PAUP*. Nodal support was assessed with 1000 bootstrap replicates. Number of nucleotide substitutions estimated at 0.01. 9574, 9656, Johnston, 8200 and 9620 were collected from a cattle host.

Genome-wide F_{ST} values of 0.4 among cattle and buffalo supports differentiation between the two groups, further supported by the high nucleotide diversity observed in buffaloes (**Table 6**).

DISCUSSION

In the present study, we describe the use of whole genome sequencing for identification of genetic variants in the genomes of South African and ESA buffalo-derived *T. parva* strains sequenced by next generation sequencing. Variant calling was also done for the first time from blood extracted *T. parva* samples. This allowed a multi-locus phylogenetic analysis to determine phylogenetic relationships with greater confidence.

The study indicated that even with bait capture, host-derived DNA remains a potential contamination problem. Seed length optimization was important in getting rid of host DNA. Failure to do so leads to a high number of reads mapping to the host, leading to wrong interpretation of the purification method. Robinson et al. (2017) also found that seed length adjustment worked well in cases where parasite DNA concentration is low compared to host's, as was the case for *Plasmodium*.

Although sample numbers were low, high SNP numbers allowed for meaningful analysis of data to determine *T. parva* diversity across host and geographical area. Available data so far supports the arguments that buffalo-derived *T. parva* is distinct from cattle-derived strains (Oura et al., 2011; Elisa et al., 2014) with cattle strains being less diverse, subset of the buffalo strains (Oura et al., 2003; Pelle et al., 2011). A similar study done by Palmateer et al. (2020), also using bait capture found buffalo-derived *T. parva* to be highly divergent from cattle-derived *T. parva*. While this latter genome would be of interest to use as an alternative reference genome to the Muguga genome, it still needs refinement with regard to assignment of contigs to the correct homologous chromosomes of the Muguga genome. Whole genome alignment indicated that large regions of various contigs do not align to the Muguga reference genome (results not shown). This suggest significant differences between the cattle- and buffalo-derived *T. parva* genomes, or that the current draft genome is not yet at reference genome status.

TABLE 6 | Genetic diversity of all joint *T. parva* strains.

	Regions		Host	
	South Africa	Eastern and Southern Africa	Buffalo	Cattle
Sample number	12	9	14	7
Total variant number	501,479	188,999	536,603	91,849
SNP number	485,520	187,620	520,362	91,115
Segregating/variable SNP sites	485,520	187,620	520,362	91,115
Nucleotide diversity (Pi)	0.008275	0.003034	0.005959	0.001640
Tajima's *D*	−0.02274	−0.36771	−0.16821	0.3507380
F_{ST}	0.30334		0.41099	

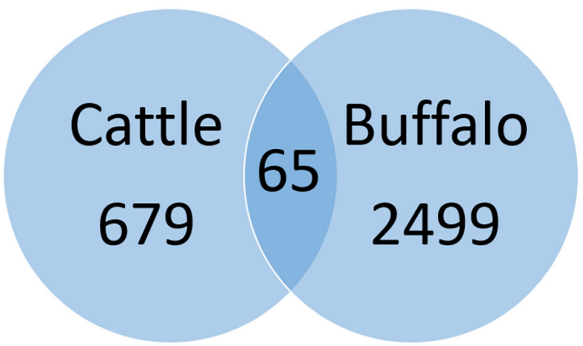


FIGURE 3 | Pairwise comparison of cattle and buffalo-derived SNPs. Sixty-five SNPs were shared between cattle and buffalo strains whereas 679 and 2499 SNPs were specific to cattle and buffalo, respectively.

The clustering of all buffalo-derived strains regardless of geographic origin suggests that they have a common ancestor (**Figure 2**). The Cape buffalo evolved from an ancestor similar to the western African buffalo (*Syncerus brachyceros*) and is distributed over eastern and southern Africa (Smitz et al., 2013). It is believed that the Cape buffalo has always been infected with *T. parva* (Young et al., 1978; Mans, 2020), which explains the clustering of buffalo-derived *T. parva* while being distinct due to lack of gene flow between the two regionally distinct strains.

Nucleotide diversity and F_{ST} values were skewed toward higher numbers in buffalo-derived *T. parva* as indicated by South African strains that were exclusively buffalo-derived. Similarly, clustering of cattle-derived ECF strains suggest that they also share a common ancestor that may have resulted from a single origin from the more diverse East African buffalo-derived stock (Obara et al., 2015; Nene and Morrison, 2016). Previous suggestions that *T. parva* originated in East Africa (Norval et al., 1991) may need to be reassessed, since the high diversity observed in southern African strains would suggest that *T. parva* rather originated in this latter region. However, more data from ESA buffalo- and cattle-derived *T. parva* is necessary to confirm the higher diversity observed in South Africa that would support this hypothesis.

Buffalo-derived strains had more substitutions (longer branch lengths) suggesting that cattle-derived strains separated from buffalo-derived strains at a later stage. 8200, 9581, 9620, 9679 and 9656 are first generation cattle-associated strains of buffalo origin that show no sign of population selection. Similarly, 8160 is a second-generation buffalo-derived strain with at least two cattle-dependent bottlenecks (transmission from buffalo HL3 to bovine 9433 to bovine 8160), but still clusters with HL3 with no sign of population selection. Correspondingly, 9574, a third generation buffalo-derived strain that was tick passaged through three cattle-dependent bottlenecks (transmission from buffalo Welgevonden to bovine 9288 to bovine 8301 to bovine 9574), still clusters with 8301 (**Figure 2**). Parasites that were derived from cattle in their latest passage have not transformed or been strain selected for cattle–cattle transmission. In the latter case, it is expected that they would have clustered with other cattle *T. parva* strains (**Figure 2**). In South Africa, the seasonality observed for the various tick life stages (Mbizeni et al., 2013), may decrease the likelihood of strains going through several cattle passages for transformation like in Barnett and Brocklesby (1966). Selection of cattle transmissible strains as suggested by Potgieter et al. (1988) is also unlikely because buffaloes and cattle are not permitted to come into contact (Department of Agriculture, 1984). Therefore, these parasites are cattle associated strains of buffalo origin that are infective to cattle. Buffalo movement is restricted in South Africa and therefore, *T. parva* positive buffaloes can only be moved to *T. parva* endemic areas and thus clustering of KwaZulu-Natal and Limpopo/Mpumalanga strains is not surprising.

Mixed populations were observed in two field strains (9679 and 9581) which had over 70% intra-strain variation (**Table 4**). These field strains had recently been introduced into *in vitro* culturing and strain selection may have not occurred as reported for cell-cultured *T. parva* schizonts (Kasai et al., 2016).

Theileria subtelomeric variable proteins and signal proteins expressed in the schizont phase are believed to be involved in host cell manipulation (Shiels et al., 2006; Schmuckli-Maurer et al., 2009) and thus may be involved in host specificity. The inability to find specific regions or genes responsible for host specificity from the selected genes may indicate that other unidentified and/or unselected gene regions are responsible. It is therefore necessary to explore additional genes including PIN1, a gene previously unknown and identified during the recent re-annotation of *T. parva* (Tretina et al., 2020) as well as the other 127 previously unknown genes. It is also possible that buffalo diversity is so great that no unique SNP markers shared by all buffalo could be found. This study also did not find the p67 diversity reported by Mukolwe et al. (2020). A likely reason for this is that their study used blood extracted field samples, which may have had multiple *T. parva* strains whereas the current work used clonal *in vitro* samples.

In South Africa, buffaloes are only allowed to move to endemic regions, this might lead to increased parasite diversity as a result of sexual reproduction of the parasite in ticks, which will explain the lack of haplotypes and all distinct strains

(Mehlhorn and Shein, 1984). Differentiation of South African strains from ESA strains may be as a result of geographic separation. Genetic drift is most likely due to the strict control measures in South Africa regarding buffalo movement and wildlife-livestock interaction as well as theileriosis control measures. This lack of interaction prevents the parasite from being selected for cattle–cattle transmission and a possible reason why South Africa does not have cattle-adapted *T. parva* populations that would allow cattle–cattle transmission that would resemble ECF.

Based on heavy mortalities of domestic livestock during the ECF epidemic in South Africa and the discovery of Corridor disease after eradication of ECF from South Africa, it was proposed that *T. parva* did not circulate in African buffalo in southern Africa before the introduction of ECF, but was transmitted to and established in buffalo during the ECF epidemic (Norval et al., 1991, 1992). Recently it was suggested that infected cattle introduced into South Africa did not carry buffalo-derived *T. parva*, but that extensive genetic diversity of buffalo-derived *T. parva* from South Africa suggest that this parasite was present in South Africa before introduction of ECF (Mans, 2020; Morrison et al., 2020). For the alternative scenario where cattle-derived *T. parva* established in buffalo during the ECF epidemic, a low genetically diverse population would have been expected, since at least two genetic bottlenecks occurred for cattle-derived *T. parva* introduced into South Africa. The first bottleneck occurred as adaptation of buffalo-derived *T. parva* to cattle in East Africa that resulted in the genetically restricted low diversity populations currently observed for cattle-derived *T. parva* throughout East and southern Africa. The second bottleneck occurred when cattle-derived *T. parva* was introduced into South Africa from a single or restricted number of *T. parva* populations (Norval et al., 1991). However, the current study provides concrete evidence that buffalo-derived *T. parva* from South Africa shows very high genetic diversity and is distinct from ECF strains. This genetic diversity would suggest an ancient association of *T. parva* with African buffalo from southern Africa.

CONCLUSION

Our data indicate that *T. parva* strains circulating in South Africa are buffalo derived, associated with Corridor disease and that there are no ECF circulating South African strains discovered to date in buffalo. Further work still needs to be done to identify regions responsible for host specificity.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://dataview.ncbi.nlm.nih.gov/object/PRJNA673089?reviewer=42193fcf2rb2g9jk5k9o10u7rj, SAMN18117497, SAMN18117498, SAMN16622817, SAMN16622818, SAMN18117499, SAMN16622819, SAMN18117500, SAMN18117501, SAMN18117502, SAMN16622805, SAMN16622808, and SAMN18117503.

ETHICS STATEMENT

Ethics approval was obtained from the ARC-OVR (project 000760-Y5) and University of Pretoria (V033-17). Written permission from national authorities was obtained.

AUTHOR CONTRIBUTIONS

BBM developed the method, designed the experiments, performed the data analysis, and wrote the manuscript. KS-M co-supervised, performed the data analysis and critically reviewed the manuscript. RP assisted in data analysis and critically reviewed the manuscript. AJ assisted in experiments and critically reviewed the manuscript. RM and SM assisted in all animal related experiments and critically reviewed the manuscript. AL provided intellectual assistance and critically reviewed the manuscript. BJM supervised, conceived the idea, and critically reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the Department of Agriculture, Land Reform and Rural Development under project P10000062.

ACKNOWLEDGMENTS

The authors would like to thank the National Department of Agriculture, Land Reform and Rural Development for funding and Provincial Department of Agriculture, Land Reform and Rural Development in sample collection. The authors also thank the Agricultural Research Council for providing facilities and staff for assistance with *in vitro* cultures and animal experiments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.666096/full#supplementary-material

Supplementary Table 1 | Genes of interest studied to determine host specificity.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Amzati, G. S., Djikeng, A., Odongo, D. O., Nimpaye, H., Sibeko, K. P., Muhigwa, J. B., et al. (2019). Genetic and antigenic variation of the bovine tick-borne pathogen *Theileria parva* in the Great Lakes region of Central Africa. *Parasit. Vect.* 12:588.

Andrews, S. (2010). *FASTQC. A Quality Control Tool for High Throughput Sequence Data*. Babraham: The Babraham Institute.

Barnett, S. F., and Brocklesby, D. W. (1966). The passage of *Theileria lawrencei* (Kenya) through cattle. *Br. Vet. J.* 122, 396–409. doi: 10.1016/s0007-1935(17)40406-4

Bishop, R., Geysen, D., Spooner, P., Skilton, R., Nene, V., Dolan, T., et al. (2001). Molecular and immunological characterisation of *Theileria parva* stocks which are components of the 'Muguga cocktail' used for vaccination against East Coast fever in cattle. *Vet. Parasitol.* 94, 227–237. doi: 10.1016/s0304-4017(00)00404-0

Blouin, E. F., and Stoltsz, W. H. (1989). Comparative infection rates of *Theileria parva* lawrencei in salivary glands of *Rhipicephalus appendiculatus* and *Rhipicephalus zambeziensis*. *Onderstepoort J. Vet. Res.* 56, 211–213.

Brown, C. G. D. (1987). "Theileriidae," in *In Vitro Methods for Parasite Cultivation*, eds A. E. R. Taylor and J. R. Baker (New York, NY: Academic Press).

Collins, N. E., and Allsopp, B. A. (1999). *Theileria parva* ribosomal internal transcribed spacer sequences exhibit extensive polymorphism and mosaic evolution: application to the characterization of parasites from cattle and buffalo. *Parasitology* 118(Pt 6), 541–551. doi: 10.1017/s0031182099004321

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

De Vos, A. J. (1982). *The Identity of Bovine Theileria spp. in South Africa*. Pretoria: University of Pretoria.

Department of Agriculture (1984). *Animal Diseases Act 1984, Act No. 35*. Republic of South Africa: Department of Agriculture.

Elisa, M., Gwakisa, P., Sibeko, K. P., Oosthuizen, M. C., and Geysen, D. (2014). Molecular characterization of *Theileria parva* field strains derived from Cattle and Buffalo sympatric populations of Northern Tanzania. *Am. J. Res. Commun.* 2, 10–22.

Gardner, M. J., Bishop, R., Shah, T., De Villiers, E. P., Carlton, J. M., Hall, N., et al. (2005). Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science* 309, 134–137. doi: 10.1126/science.1110439

Geysen, D., Bishop, R., Skilton, R., Dolan, T. T., and Morzaria, S. (1999). Molecular epidemiology of *Theileria parva* in the field. *Trop. Med. Int. Health* 4, A21–A27.

Hayashida, K., Abe, T., Weir, W., Nakao, R., Ito, K., Kajino, K., et al. (2013). Whole-genome sequencing of *Theileria parva* strains provides insight into parasite migration and diversification in the African continent. *DNA Res.* 20, 209–220. doi: 10.1093/dnares/dst003

Hemmink, J. D., Sitt, T., Pelle, R., De Klerk-Lorist, L. M., Shiels, B., Toye, P. G., et al. (2018). Ancient diversity and geographical sub-structuring in African buffalo *Theileria parva* populations revealed through metagenetic analysis of antigen-encoding loci. *Int. J. Parasitol.* 48, 287–296. doi: 10.1016/j.ijpara.2017.10.006

Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14, 68–73. doi: 10.1093/bioinformatics/14.1.68

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Kariuki, D. P., Young, A. S., Morzaria, S. P., Lesan, A. C., Mining, S. K., Omwoyo, P., et al. (1995). *Theileria parva* carrier state in naturally infected and artificially immunised cattle. *Trop. Anim. Health Prod.* 27, 15–25. doi: 10.1007/bf02236328

Kasai, F., Hirayama, N., Ozawa, M., Iemura, M., and Kohara, A. (2016). Changes of heterogeneous cell populations in the Ishikawa cell line during long-term culture: proposal for an *in vitro* clonal evolution model of tumor cells. *Genomics* 107, 259–266. doi: 10.1016/j.ygeno.2016.04.003

Latif, A., and Mans, B. J. (2010). *Corridor Disease Outbreak: Johnston Farm Visit*. Report to the Director Onderstepoort Veterinary Institute

Latif, A. A., Hove, T., Kanhai, G. K., Masaka, S., and Pegram, R. G. (2001). Epidemiological observations of *Zimbabwean theileriosis*: disease incidence and pathogenicity in susceptible cattle during *Rhipicephalus appendiculatus* nymphal and adult seasonal activity. *Onderstepoort J. Vet. Res.* 68, 187–195.

Lawrence, J. A. (1979). The differential diagnosis of the bovine theilerias of Southern Africa. *J. South Afr. Vet. Assoc.* 50, 311–313.

Lawrence, J. A. (1992). "History of bovine theileriosis in southern Africa," in *The Epidemiology of Theileriosis in Southern Africa*, eds R. A. I. Norval, B. D. Perry, and A. S. Young (London: Academic Press).

Lawrence, J. A., Perry, B. D., and Williamson, M. S. (1992). "East Coast fever," in *Infectious Diseases of Livestock*, eds J. A. W. Coetzer and R. C. Tustin (Cape Town: Oxford University Press).

Lawrence, J. A., Sibeko-Matjila, K. P., and Mans, B. J. (2017). "East Coast fever," in *Infectious Diseases of Livestock*, eds J. A. W. Coetzer, M. L. Penrith, N. J. Maclachlan, and G. R. Thomson (Pretoria: Anipedia).

Maboko, B. B., Featherston, J., Sibeko-Matjila, K. P., and Mans, B. J. (2021). Whole genome sequencing of *Theileria parva* using target capture. *Genomics* 113, 429–438. doi: 10.1016/j.ygeno.2020.12.033

Mans, B. J. (2020). "The history of East Coast fever in Southern Africa," in *Proceedings of the 44th International Congress of the World Assosciation for the History of Veterinary Medicine*, (Pretoria: The Farm Inn Hotel & Conference Centre), 52–54.

Mans, B. J., Pienaar, R., Christo Troskie, P., and Combrink, M. P. (2019). Investigation into limiting dilution and tick transmissibility phenotypes associated with attenuation of the S24 vaccine strain. *Parasit. Vect.* 12:419.

Matson, B. A. (1967). Theileriosis in Rhodesia. I. A study of diagnostic specimens over two seasons. *J. South Afr. Vet. Med. Assoc.* 38, 93–102.

Mbizeni, S., Potgieter, F. T., Troskie, C., Mans, B. J., Penzhorn, B. L., and Latif, A. A. (2013). Field and laboratory studies on Corridor disease (*Theileria parva* infection) in cattle population at the livestock/game interface of uPhongolo-Mkuze area, South Africa. *Ticks Tick Borne Dis.* 4, 227–234. doi: 10.1016/j.ttbdis.2012.11.005

Mclaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17:122.

Mehlhorn, H., and Shein, E. (1984). The piroplasms: life cycle and sexual stages. *Adv. Parasitol.* 23, 37–103. doi: 10.1016/s0065-308x(08)60285-7

Morrison, W. I. (2015). The aetiology, pathogenesis and control of theileriosis in domestic animals. *Rev. Sci. Technol.* 34, 599–611. doi: 10.20506/rst.34.2.2383

Morrison, W. I., Hemmink, J. D., and Toye, P. G. (2020). *Theileria parva*: a parasite of African buffalo, which has adapted to infect and undergo transmission in cattle. *Int. J. Parasitol.* 50, 403–412. doi: 10.1016/j.ijpara.2019.12.006

Mukolwe, L. D., Odongo, D. O., Byaruhanga, C., Snyman, L. P., and Sibeko-Matjila, K. P. (2020). Analysis of p67 allelic sequences reveals a subtype of allele type 1 unique to buffalo-derived *Theileria parva* parasites from southern Africa. *PLoS One* 15:e0231434. doi: 10.1371/journal.pone.0231434

Mwamuye, M. M., Odongo, D., Kazungu, Y., Kindoro, F., Gwakisa, P., Bishop, R. P., et al. (2020). Variant analysis of the sporozoite surface antigen gene reveals that asymptomatic cattle from wildlife-livestock interface areas in northern Tanzania harbour buffalo-derived *T. parva*. *Parasitol. Res.* 119, 3817–3828. doi: 10.1007/s00436-020-06902-1

Neitz, W. O., Canhan, A. S., and Kluge, E. B. (1955). Corridor disease: a fatal form of bovine theileriosis encountered in Zululand. *J. South Afr. Vet. Assoc.* 26, 79–88.

Nene, V., and Morrison, W. I. (2016). Approaches to vaccination against *Theileria parva* and *Theileria annulata*. *Parasite Immunol.* 38, 724–734. doi: 10.1111/pim.12388

Nene, V., Musoke, A., Gobright, E., and Morzaria, S. (1996). Conservation of the sporozoite p67 vaccine antigen in cattle-derived *Theileria parva* stocks with different cross-immunity profiles. *Infect. Immun.* 64, 2056–2061. doi: 10.1128/iai.64.6.2056-2061.1996

Norling, M., Bishop, R. P., Pelle, R., Qi, W., Henson, S., Drabek, E. F., et al. (2015). The genomes of three stocks comprising the most widely utilized live sporozoite *Theileria parva* vaccine exhibit very different degrees and patterns of sequence divergence. *BMC Genomics* 16:729. doi: 10.1186/s12864-015-1910-9

Norval, R. A., Lawrence, J. A., Young, A. S., Perry, B. D., Dolan, T. T., and Scott, J. (1991). *Theileria parva*: influence of vector, parasite and host relationships on the epidemiology of theileriosis in southern Africa. *Parasitology* 102(Pt 3), 347–356. doi: 10.1017/s0031182000064295

Norval, R. A. I., Perry, B. D., and Young, A. S. (1992). *The Epidemiology of Theileriosis in Africa*. London: Academic Press.

Obara, I., Ulrike, S., Musoke, T., Spooner, P. R., Jabbar, A., Odongo, D., et al. (2015). Molecular evolution of a central region containing B cell epitopes in the gene encoding the p67 sporozoite antigen within a field population of *Theileria parva*. *Parasitol. Res.* 114, 1729–1737. doi: 10.1007/s00436-015-4358-6

Oura, C. A., Tait, A., Asiimwe, B., Lubega, G. W., and Weir, W. (2011). *Theileria parva* genetic diversity and haemoparasite prevalence in cattle and wildlife in and around Lake Mburo National Park in Uganda. *Parasitol. Res.* 108, 1365–1374. doi: 10.1007/s00436-010-2030-8

Oura, C. A. L., Odongo, D. O., Lubega, G. W., Spooner, P. R., Tait, A., and Bishop, R. P. (2003). A panel of microsatellite and minisatellite markers for the characterisation of field isolates of *Theileria parva*. *Int. J. Parasitol.* 33, 1641–1653. doi: 10.1016/s0020-7519(03)00280-7

Palmateer, N. C., Tretina, K., Orvis, J., Ifeonu, O. O., Crabtree, J., Drabek, E., et al. (2020). Capture-based enrichment of *Theileria parva* DNA enables full genome assembly of first buffalo-derived strain and reveals exceptional intra-specific genetic diversity. *PLoS Negl. Trop. Dis.* 14:e0008781. doi: 10.1371/journal.pntd.0008781

Pelle, R., Graham, S. P., Njahira, M. N., Osaso, J., Saya, R. M., Odongo, D. O., et al. (2011). Two *Theileria parva* CD8 T cell antigen genes are more variable in buffalo than cattle parasites, but differ in pattern of sequence diversity. *PLoS One* 6:e19015. doi: 10.1371/journal.pone.0019015

Picard (2019). *Mark Duplicates [Online]*. Available online at: http://broadinstitute.github.io/picard/ (accessed 2019)

Pienaar, R., Potgieter, F. T., Latif, A. A., Thekisoe, O. M., and Mans, B. J. (2011). The Hybrid II assay: a sensitive and specific real-time hybridization assay for the diagnosis of *Theileria parva* infection in Cape buffalo (Syncerus caffer) and cattle. *Parasitology* 138, 1935–1944. doi: 10.1017/s0031182011001454

Potgieter, F. T., Stoltsz, W. H., Blouin, E. F., and Roos, J. A. (1988). Corridor disease in South Africa: a review of the current status. *J. South Afr. Vet. Assoc.* 59, 155–160.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Robinson, K. M., Hawkins, A. S., Santana-Cruz, I., Adkins, R. S., Shetty, A. C., Nagaraj, S., et al. (2017). Aligner optimization increases accuracy and decreases compute times in multi-species sequence data. *Microb. Genome* 3:e000122.

Rozas, J., Ferrer-Mata, A., Sanchez-Delbarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 34, 3299–3302. doi: 10.1093/molbev/msx248

Schmuckli-Maurer, J., Casanova, C., Schmied, S., Affentranger, S., Parvanova, I., Kang'a, S., et al. (2009). Expression analysis of the *Theileria parva* subtelomere-encoded variable secreted protein gene family. *PLoS One* 4:e4839. doi: 10.1371/journal.pone.0004839

Shiels, B., Langsley, G., Weir, W., Pain, A., Mckellar, S., and Dobbelaere, D. (2006). Alteration of host cell phenotype by *Theileria annulata* and *Theileria parva*: mining for manipulators in the parasite genomes. *Int. J. Parasitol.* 36, 9–21. doi: 10.1016/j.ijpara.2005.09.002

Sibeko, K. P., Collins, N. E., Oosthuizen, M. C., Troskie, M., Potgieter, F. T., Coetzer, J. A., et al. (2011). Analyses of genes encoding *Theileria parva* p104 and polymorphic immunodominant molecule (PIM) reveal evidence of the presence of cattle-type alleles in the South African *T. parva* population. *Vet. Parasitol.* 181, 120–130. doi: 10.1016/j.vetpar.2011.04.035

Sibeko, K. P., Geysen, D., Oosthuizen, M. C., Matthee, C. A., Troskie, M., Potgieter, F. T., et al. (2010). Four p67 alleles identified in South African *Theileria parva* field samples. *Vet. Parasitol.* 167, 244–254. doi: 10.1016/j.vetpar.2009.09.026

Sibeko, K. P., Oosthuizen, M. C., Collins, N. E., Geysen, D., Rambritch, N. E., Latif, A. A., et al. (2008). Development and evaluation of a real-time polymerase chain reaction test for the detection of *Theileria parva* infections in Cape buffalo (Syncerus caffer) and cattle. *Vet. Parasitol.* 155, 37–48. doi: 10.1016/j.vetpar.2008.03.033

Smitz, N., Berthouly, C., Cornelis, D., Heller, R., Van Hooft, P., Chardonnet, P., et al. (2013). Pan-African genetic structure in the African buffalo (Syncerus caffer): investigating intraspecific divergence. *PLoS One* 8:e56235. doi: 10.1371/journal.pone.0056235

Stoltsz, W. H. (1989). Theileriosis in South Africa: a brief review. *Sci. Techn. Rev. Office Int. Epizoot.* 8, 93–102. doi: 10.20506/rst.8.1.395

Theiler, A. (1904). Rhodesian tick fever. *Transvaal Agric. J.* 2, 421–438.

Tonui, T., Corredor-Moreno, P., Kanduma, E., Njuguna, J., Njahira, M. N., Nyanjom, S. G., et al. (2018). Transcriptomics reveal potential vaccine antigens and a drastic increase of upregulated genes during *Theileria parva* development from arthropod to bovine infective stages. *PLoS One* 13:e0204047. doi: 10.1371/journal.pone.0204047

Tretina, K., Pelle, R., Orvis, J., Gotia, H. T., Ifeonu, O. O., Kumari, P., et al. (2020). Re-annotation of the *Theileria parva* genome refines 53% of the proteome and uncovers essential components of N-glycosylation, a conserved pathway in many organisms. *BMC Genomics* 21:279. doi: 10.1186/s12864-020-6683-0

Uilenberg, G. (1999). Immunization against diseases caused by *Theileria parva*: a review. *Trop. Med. Int. Health* 4, A12–A20.

Young, A. S., Branagan, D., Brown, C. G., Burridge, M. J., Cunningham, M. P., and Purnell, R. E. (1973). Preliminary observations on a theilerial species pathogenic to cattle isolated from buffalo (Syncerus caffer) in Tanzania. *Br. Vet. J.* 129, 382–389. doi: 10.1016/s0007-1935(17)36442-4

Young, A. S., Brown, C. G., Burridge, M. J., Grootenhuis, J. G., Kanhai, G. K., Purnell, R. E., et al. (1978). The incidence of theilerial parasites in East African buffalo (Syncerus caffer). *Tropenmed. Parasitol.* 29, 281–288.

Young, A. S., Leitch, B. L., and Newson, R. M. (1981). "The occurrence of a *Theileria parva* carrier state in cattle from an East Coast fever endemic area of Kenya," in *Advances in the Control of Theileriosis. Current Topics in Veterinary Medicine and Animal Science*, eds A. D. Irvin, M. P. Cunningham, and A. S. Young (Dordrecht: Springer).

Young, A. S., Leitch, B. L., Newson, R. M., and Cunningham, M. P. (1986). Maintenance of *Theileria parva* parva infection in an endemic area of Kenya. *Parasitology* 93(Pt 1), 9–16. doi: 10.1017/s0031182000049787

Yunker, C. E., Byrom, B., and Semu, S. (1988). Cultivation of Cowdria ruminantium bovine vascular endothelial cells. *Kenya Vet.* 12, 12–16.

Yusufmia, S. B., Collins, N. E., Nkuna, R., Troskie, M., Van Den Bossche, P., and Penzhorn, B. L. (2010). Occurrence of *Theileria parva* and other haemoprotozoa in cattle at the edge of Hluhluwe-iMfolozi Park, KwaZulu-Natal, South Africa. *J. South Afr. Vet. Assoc.* 81, 45–49.

Zweygarth, E., Nijhof, A. M., Knorr, S., Ahmed, J. S., Al-Hosary, A. T. A., Obara, I., et al. (2020). Serum-free *in vitro* cultivation of *Theileria annulata* and *Theileria parva* schizont-infected lymphocytes. *Transbound Emerg. Dis.* 67, 35–39. doi: 10.1111/tbed.13348

Zweygarth, E., Vogel, S. W., Josemans, A. I., and Horn, E. (1997). *In vitro* isolation and cultivation of *Cowdria ruminantium* under serum-free culture conditions. *Res. Vet. Sci.* 63, 161–164. doi: 10.1016/s0034-5288(97)90011-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 25 June 2021
doi: 10.3389/fgene.2021.668574

Chloroquine and Sulfadoxine–Pyrimethamine Resistance in Sub-Saharan Africa—A Review

*Alexandra T. Roux[1], Leah Maharaj[1], Olukunle Oyegoke[1], Oluwasegun P. Akoniyon[1], Matthew Adekunle Adeleke[1], Rajendra Maharaj[2] and Moses Okpeku[1]**

[1] Discipline of Genetics, School of Life Sciences, University of KwaZulu-Natal, Westville, South Africa, [2] Office of Malaria Research, South African Medical Research Council, Cape Town, South Africa

Malaria is a great concern for global health and accounts for a large amount of morbidity and mortality, particularly in Africa, with sub-Saharan Africa carrying the greatest burden of the disease. Malaria control tools such as insecticide-treated bed nets, indoor residual spraying, and antimalarial drugs have been relatively successful in reducing the burden of malaria; however, sub-Saharan African countries encounter great challenges, the greatest being antimalarial drug resistance. Chloroquine (CQ) was the first-line drug in the 20th century until it was replaced by sulfadoxine–pyrimethamine (SP) as a consequence of resistance. The extensive use of these antimalarials intensified the spread of resistance throughout sub-Saharan Africa, thus resulting in a loss of efficacy for the treatment of malaria. SP was replaced by artemisinin-based combination therapy (ACT) after the emergence of resistance toward SP; however, the use of ACTs is now threatened by the emergence of resistant parasites. The decreased selective pressure on CQ and SP allowed for the reintroduction of sensitivity toward those antimalarials in regions of sub-Saharan Africa where they were not the primary drug for treatment. Therefore, the emergence and spread of antimalarial drug resistance should be tracked to prevent further spread of the resistant parasites, and the re-emergence of sensitivity should be monitored to detect the possible reappearance of sensitivity in sub-Saharan Africa.

Keywords: antimalarial drug resistance, chloroquine, malaria, malaria control, sulfadoxine–pyrimethamine

OPEN ACCESS

Edited by:
Apichai Tuanyok, University of Florida, United States

Reviewed by:
Rajeev Kumar Mehlotra, Case Western Reserve University, United States
Chenqi Wang, University of South Florida, United States

****Correspondence:***
Moses Okpeku
okpekum@ukzn.ac.za

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: *16 February 2021*
Accepted: *20 April 2021*
Published: *25 June 2021*

Citation:
Roux AT, Maharaj L, Oyegoke O, Akoniyon OP, Adeleke MA, Maharaj R and Okpeku M (2021) Chloroquine and Sulfadoxine–Pyrimethamine Resistance in Sub-Saharan Africa—A Review. Front. Genet. 12:668574. doi: 10.3389/fgene.2021.668574

INTRODUCTION

Malaria is a global health concern regarding morbidity and mortality, with approximately 228 million worldwide cases and an estimated 405,000 deaths in 2018 (World Health Organisation (WHO), 2019). Underprivileged, rural populations consisting of young children and pregnant women are disproportionately affected by malaria. The cornerstone of malaria control efforts for the past decade has been to address the inequalities, including the availability of commodities (Taylor et al., 2017). Sub-Saharan Africa experiences the greatest burden of the malaria disease, accounting for approximately 90% of the world's *Plasmodium falciparum* infections and deaths with almost all malaria caused by *P. falciparum* in this area (Yeka et al., 2012; Maitland, 2016).

P. falciparum is the most virulent of the malaria parasites that infect humans (Recker et al., 2018). Therefore, the utilization of effective tools for malaria control has increased to lessen the burden of malaria (Greenwood et al., 2005). In spite of increased efforts, many countries in Africa are still confronted with challenges with malaria control, partially as a consequence of the identified limitations in public health structures as well as the infrastructure of primary health care (Xia et al., 2014). The development of resistance toward antimalarial drugs is one of the main challenges when dealing with malaria control and elimination, as such antimalarial drug resistance in a setting where access to health care is limited has severe consequences (Takala-Harrison and Laufer, 2015). Drug pressure, which refers to the extensive and/or misuse of antimalarial drugs, is an identified factor associated with the emergence of resistance; additionally, the misuse or sole use of a drug can encourage selection of resistant strains (Hanboonkunupakarn and White, 2016).

In Africa, malaria control programs rest majorly on vector control and the use of antimalarial drugs (Conrad and Rosenthal, 2019). There are different types of malaria, namely, asymptomatic, uncomplicated, and severe malaria. Asymptomatic malaria refers to malaria parasites being present in the blood, providing a reservoir for transmission, without the individual displaying symptoms (Chourasia et al., 2017). Uncomplicated malaria refers to malaria symptoms presented by a patient together with a positive parasitological test (Grobusch and Kremsner, 2005). Severe malaria is defined by positive parasitological test (microscopy or rapid diagnostic test) detecting *P. falciparum* and at least one condition for severe disease such as severe anemia or respiratory distress (Conroy et al., 2019).

Sub-Saharan African countries used chloroquine (CQ) as the first-line drug for malaria up to the start of this millennium. However, it was replaced with sulfadoxine–pyrimethamine (SP) after CQ was declared ineffective as a result of resistant parasites (Lusingu and Von Seidlein, 2008) but soon lost its efficacy when parasites began to develop resistance toward it. Mutations in the *P. falciparum* CQ-resistant transporter (*pfcrt*) gene and the *P. falciparum* multidrug-resistance gene 1 (*pfmdr-1*) have been implicated in CQ resistance (CQR) (Bin Dajem and Al-Qahtani, 2010). CQR was observed on the Thai-Cambodian border and concurrently in South America in the late 1950s (Young and Moore, 1961; Harinasuta et al., 1965). In Southeast Asia, specifically Thailand and Myanmar, drug resistance in local parasite populations is a serious concern, with *P. falciparum* in the region tending to develop resistance (Parker et al., 2015). Key factors associated with drug resistance in these regions is drug pressure, which results in selection of resistant parasites and their propagation by local transmission and reservoir migration (Wernsdorfer, 1994). In 1978, resistance spread to Africa, with confirmed treatment failures in both Kenya and Tanzania, later reaching West Africa in 1980. CQ continued to be the first-line treatment for uncomplicated *P. falciparum* malaria in the most sub-Saharan countries until after 2,000 despite its declining use (Flegg et al., 2013). However, some countries, including South Africa, Zambia, Tanzania, and Kenya (D'Alessandro and Buttiëns, 2001; Mwendera et al., 2016), were only forced to switch to SP as the new first-line treatment for malaria, as levels of CQR increased (Gatton et al., 2004). Shortly after its introduction as the new first-line treatment for malaria, polymorphisms in the parasite's genes—*pfdhps* [*P. falciparum* dihydropteroate synthase (DHFR)] and *pfdhfr* [*P. falciparum* dihydrofolate reductase (DHPS)]—became widespread throughout Africa (Winstanley et al., 2004). The association of the mutations with SP treatment failure in children made the antimalarial unsuitable for therapy (Desai et al., 2015). In Africa, the use of SP for clinical malaria was stopped as a consequence of its extensive resistance. However, in malaria endemic areas, it is still used for intermittent preventative treatment during pregnancy (Zhao et al., 2020). By 2007, 90% of sub-Saharan Africa had implemented policies of artemisinin-based combination therapy (ACT) for the treatment of uncomplicated malaria as a consequence of widespread CQ and SP resistance (Frosch et al., 2011). The efficacy of ACTs leads to substantial declines in morbidity and mortality in areas with high malaria endemicity. But the success of such therapies is threatened by the appearance of artemisinin-resistant strains of *P. falciparum* from Thai-Cambodian and Thai-Myanmar borders (Ajayi and Ukwaja, 2013). There have been isolated reports of artemisinin-resistant parasites in sub-Saharan Africa, but they have not become established on the continent yet (Lu et al., 2017a). The emergence and spread of artemisinin-resistant parasites would have an immense impact on Africa's control efforts (Raman et al., 2019), especially as the world is anticipating a global elimination strategy (World Health Organisation (WHO), 2020). *P. falciparum* contains Kelch13 (*K13*), which is needed in the asexual erythrocytic development stage. Although details of its function are not fully known (Siddiqui et al., 2020), they are highly conserved, and single-point mutations are associated with artemisinin resistance (Birnbaum et al., 2020). According to Meshnick (1998), the mechanism of action of artemisinin is in two steps: first, an initial activation that catalyzes the cleavage of endoperoxide produces free radicals. In the second step, these free radicals are responsible for killing the parasites. However, a single-nucleotide polymorphism (SNP) in the gene associated with up-regulated pathway antagonizes the artemisinin oxidation activity (Fairhurst and Dondorp, 2016). A recent study revealed that parasites with inactivated *K13* or its mutated form displayed reduced hemoglobin endocytosis (Birnbaum et al., 2020). *K13* is important in the uptake and degradation of hemoglobin, which is vital for parasite survival. The mutations and mislocalization of *K13* induce artemisinin resistance (Xie et al., 2020); therefore, a close observation of ACT resistance is necessary, as they are the current first-line treatment for *P. falciparum* malaria (Siddiqui et al., 2020).

A key challenge for malaria control is the emergence and spread of antimalarial drug resistance, particularly in the malaria endemic areas of sub-Saharan Africa where disastrous consequences are observed as a result of the spread of CQ- and SP-resistant *P. falciparum* strains (von Seidlein and Dondorp, 2015). Furthermore, the development of ACT resistance threatens the control efforts in malaria elimination. This review aims to track the spread of CQ and SP resistance in sub-Saharan Africa. It identifies malaria control strategies

employed and ways to combat the challenges being faced to further prevent the emergence and spread of resistant parasites.

MALARIA CONTROL IN SUB-SAHARAN AFRICA

A global public health concern of this century includes controlling vector-borne diseases such as malaria; therefore, much effort has focused on the development of vector control approaches (Tizifa et al., 2018). These approaches encompass methods directed toward the malaria vector by restricting its ability to transmit malaria by shielding areas that are recognized as receptive areas for transmission. Receptivity of the disease is dependent on the vectorial capacity of local vector populations, besides the presence of the vector; this includes the vector population size, biting habits, and the longevity of the sporogony period. These parameters are greatly affected by local ecology, climate, and human and vector behaviors (Smith Gueye et al., 2016). Examples of vector control strategies include insecticide-treated nets (ITNs) and indoor residual spraying (IRS).

In Africa, ITNs are the most extensively used intervention for malaria control, signifying the main tool for vector control in almost all malaria endemic African countries (World Health Organisation (WHO), 2017). ITNs act as a direct barrier to mosquito biting, hence proving effective in the reduction of malaria-related morbidity and mortality. Additionally, they reduce vector density and the average life span by providing community-wide protection through the killing of mosquitoes (Scates et al., 2020). They exploit the indoor feeding and resting behavioral patterns displayed by some *Anopheles* mosquitoes (Steinhardt et al., 2017). The universal coverage of ITN distribution, with mass distribution campaigns conducted in intervals, is implemented by most national malaria control programs (World Health Organisation (WHO), 2017). More than 800 million ITNs have been distributed in sub-Saharan Africa between 2011 and 2016, to ultimately achieve universal coverage (Olapeju et al., 2018). Through this initiative, a greater proportion (30% in 2010 to 54% in 2016) of Africans in malaria endemic areas slept under ITNs (Olapeju et al., 2018). In sub-Saharan Africa, 67–73% of the total 663 million prevented malaria cases in the past 15 years have been credited to the widespread distribution and use of ITNs (World Health Organisation (WHO), 2017). Another main method used for malaria control on a large scale includes the spraying of houses with insecticides, referred to as IRS. This method has aided the elimination of malaria from large parts of Latin America, Europe, Russia, and Asia. There are also successful IRS programs implemented in parts of Africa (Pluess et al., 2010). It is believed to function by repelling mosquitoes from entering houses as well as through killing female mosquitoes that rest inside houses after taking up a blood meal, thus suggesting that IRS is largely effective against endophilic mosquitoes (those species resting indoors). Additionally, this method is reliant on the vectorial mass effect, which refers to the reduction in transmission as a result of increased mortality of adult vectors typically after feeding (Najera and Zaim, 2001). Considering the slow development and implementation of alternative interventions, ITNs and IRS continue to be the foundation of the malaria control agenda. Consequently, a great challenge is the optimization of the continued use and sustained success of existing ITNs and IRS, while new vector control tools are being studied (Okumu and Moore, 2011).

Antimalarial drugs are used as a malaria control strategy to essentially reduce transmission. During the 20th century, CQ, a safe and inexpensive antimalarial, was a pillar of malaria control and eradication; however, it has lost its effectiveness. Of late, SP was the only alternative extensively available, but the increase and spread of drug resistance have compromised its effectiveness. Resistance to CQ and SP appeared in Southeast Asia, thereafter, spread to Africa (Naidoo and Roper, 2010). Therefore, the World Health Organization (WHO) recommended the implementation of policies approving ACTs as the primary treatment strategy for uncomplicated malaria in the majority of malaria endemic countries (Frosch et al., 2011). In spite of this, according to the 2008 World Malaria Report, ACTs were used to treat only 3% of children with suspected malaria (World Health Organisation (WHO), 2008), implying that CQ and SP were still used to treat malaria in children. Although ACTs are the recommended first-line treatment for malaria in both Asia and Africa, artemisinin-resistant *P. falciparum* strains have appeared and spread within Southeast Asia, subsequently leading to a reduction in treatment efficacy (Dondorp and Ringwald, 2013). Since Africa holds the greatest burden of malaria, there is a concern regarding the impact of artemisinin resistance on malaria morbidity and mortality (Slater et al., 2016).

THE MECHANISMS AND CONTRIBUTING FACTORS ASSOCIATED WITH ANTIMALARIAL DRUG RESISTANCE IN *Plasmodium falciparum*

Drug resistance is a growing problem in the fight to control malaria, as pathogens regularly develop mechanisms that allow them to survive the use of drugs. These mechanisms are generally the result of mutations that affect the drug's target site, thereby hindering or completely preventing binding between the drug and its target (**Table 1**). Another way that drug resistance may occur is by the increased levels of the target—this means that more drug is needed to reach inhibition of the parasite.

Chloroquine functions by accumulating in the parasite's acidic food vacuole (Lehane et al., 2012; Lawrenson et al., 2018). The xenobiotic inhibits heme catabolism (i.e., the mechanism by which the malaria parasite "feeds"). Upon infection, host hemoglobin is degraded by the parasite; this releases heme and leads to the ultimate development of an acidic lysosome-like digestive vacuole (Coban, 2020). Downstream, there is accumulation of heme–CQ complexes, which negatively impact parasite survival (Loria et al., 1999; Lehane et al., 2012).

CQ is able to move through biological membranes and accumulate in the acidic digestive vacuole (Ehlgen et al., 2012). This digestive vacuole functions to conduct

TABLE 1 | Summary of chloroquine- and sulfadoxine–pyrimethamine-resistant genes, their mutation sites, site, and mode of action.

Antimalarial and the stage it targets	Resistance gene(s)	Mutation site	Site of action	Mode of action
Chloroquine Stage of lifecycle—erythrocytic asexual stages (Lehane et al., 2012)	*Pfcrt* [located on chromosome 7 (Ikegbunam et al., 2019)]	Polymorphism at position 76 (K76T) in the first transmembrane domain (Pulcini et al., 2015)	Digestive vacuole (Shafik et al., 2020)	Mutant *Pfcrt*-mediated CQ efflux lessens access of CQ to its heme target (Ecker et al., 2012)
	pfmdr-1 [located on chromosome 5 (Ikegbunam et al., 2019)]	The amino-terminal mutations N86Y and F184Y—common to Asian and African parasites (Veiga et al., 2016). The 3 carboxy-terminal mutations S1034C, N1042D, and D1246Y—common to South American isolates (Veiga et al., 2016).	Digestive vacuole (Reiling et al., 2018)	Acts as an auxiliary mechanism alongside diffusion for drug entry into the digestive vacuole (Ibraheem et al., 2014)
Sulfadoxine–pyrimethamine **Stage of life cycle—liver and blood stages (**Raphemot et al., 2016)	*Pfdhfr* [located on chromosome 4 (Fortes et al., 2011)]	Amino acid point mutations at codons N51I, C59R, S108N, and I164L (Jiang et al., 2019)	Folate metabolic pathway (Sharma et al., 2015)	*Pfdhfr*—associated with pyrimethamine resistance (Sharma et al., 2015) *Pfdhps*—associated with sulfadoxine resistance (Sharma et al., 2015). These mutations reduce the binding affinity of SP to the targeted enzymes (Kümpornsin et al., 2014)
	Pfdhps [located on chromosome 8 (Fortes et al., 2011)]	Amino acid point mutations at codons S436A, A437G, K540E, A581G, and A613S (Jiang et al., 2019)	Folate metabolic pathway (Sharma et al., 2015)	

CQ, chloroquine; SP, sulfadoxine–pyrimethamine.

proteolysis of hemoglobin, which produces dipeptides and ferriprotoporphyrin IX (a substance that is toxic to the parasite at high concentrations). The parasite biocrystallizes ferriprotoporphyrin IX to hemozoin—this is inhibited by use of CQ (Reiling et al., 2018). CQR arises when CQ cannot accumulate at this active site to break the parasite's hemoglobin degradation cycle that *Plasmodium* needs (Loria et al., 1999).

CQR is often associated with genes *pfcrt* and/or *pfmdr-1*. The *P. pfcrt* gene is 424-amino acid long and is made up of 10 predicted transmembrane domains (Griffin et al., 2012). It is expressed at all infected erythrocyte stages with maximal expression at the trophozoite stage (Roepe, 2011). In its natural state, CQ is able to permeate the membrane of the food vacuole but becomes protonated in the vacuole and cannot pass the membrane to exit the vacuole (Homewood et al., 1972; Chinappi et al., 2010). As a result, there is a collection of CQ in the vacuole that binds to heme.

However, mutation on the gene increases export of CQ molecules. It has been reported that mutations in *pfcrt* lead to CQ and hydrogen ions being transported out of the food vacuole, which is where CQ exerts its effects from (Lehane and Kirk, 2008; Ecker et al., 2012). As a result, CQ is not activated and does not perform as it should. This is seen phenotypically as CQR.

Gene *pfmdr-1* has also been implicated in CQR. It mediates the production of P-glycoprotein homolog 1. The protein is localized to the food vacuole membrane (Cowman et al., 1991; Njokah et al., 2016). It is a member of a family of proteins that couple ATP hydrolysis to translocation of solutes across cell membranes (Cowman et al., 1991). Structural modeling of *pfmdr-1* assesses biophysical mechanisms of this gene and its protein in conferring resistance with respect to aminoquinoline (Ferreira et al., 2011). It has been established that *pfmdr-1* is involved in transporting xenobiotics (such as CQ) to the food vacuole. To do this, it is located along the food vacuole membrane and pushes CQ away off the cytosol (Dorsey et al., 2001; Ibraheem et al., 2014). The binding domain is on the cytosol-facing side of the food vacuole, allowing it to first encounter antimalarials—this side is where most resistance-causing mutations occur (Ibraheem et al., 2014). Mutations in *pfmdr-1* prevent movement of antimalarials from cytosol into the food vacuole—this reduces potency of some drugs such as CQ, which act in the food vacuole, but drugs that inhibit targets outside the food vacuole can become more potent (Wicht et al., 2020).

SP is an antifolate that is routinely used to treat uncomplicated malaria (Terlouw et al., 2003). Sulfadoxine (SDX) inhibits the activity of DHFR, and pyrimethamine (PYR) inhibits DHPS; these constituents are active against asexual erythrocytic stages of *P. falciparum* (Sandefur et al., 2007). Both DHFR and DHPS are central in parasite metabolism. Specifically, these enzymes are in the folate metabolic pathway. At the end of the pathway, reduced folate cofactors are produced, which are needed for DNA synthesis and metabolism of particular amino acids (Hyde, 2009).

Dihydrofolate reductase is the third enzyme involved in the folate-synthesis pathway in which it combines pteridine with *para*-amino benzoic acid to form dihydropteroate (Heinberg and Kirkman, 2015). SDX is a structural analog of *para*-amino benzoic acid, allowing it to inhibit the DHPS enzyme through competitive inhibition (Nzila, 2006). PYR is a competitive inhibitor of DHFR (Sandefur et al., 2007); thus, in its presence,

the folate-metabolism pathway is halted or made less effective. Both SDX and PYR function through competitive inhibition of different targets. In combination, two different enzymes of the same pathway are disrupted and present as resistance to SP.

THE SPREAD OF CHLOROQUINE RESISTANCE IN SUB-SAHARAN AFRICA

Introduced in the mid-1940s, CQ became the most extensively utilized therapeutic antimalarial drug by 1950. Quinine was essentially swapped out for CQ in areas such as tropical Africa, which lacked systematic malaria eradication programs, therefore making it a pillar of presumptive and mass treatment in eradication campaigns (Nuwaha, 2001). CQ was used as the main malaria treatment therapy up to 1990, owing to its safety, efficacy, and low cost. However, CQR emerged in various parts of the world and quickly spread to West Africa in the 1980s and 1990s (Dagnogo et al., 2018) (**Figure 1**). Treatment failure, particularly in children who are too young to have acquired immunity, in malaria endemic regions is attributed to the presence of CQR (Trape et al., 1998). CQR is primarily related to an SNP in *pfcrt*, thus leading to an amino acid mutation from threonine to lysine at codon 76 (wild-type K to the mutant T-K76T) (Fidock et al., 2000); additionally, SNPs in exons 2, 3, 4, 6, 9, 10, and 11 of the *pfcrt* gene display possible association with CQR (Awasthi and Das, 2013). Three main haplotypes occur in codons 72–76 of *pfcrt*, thus resulting in the wild-type CVMNK and CQR haplotypes CVIET and SVMNT. In Africa, the most dominant mutant haplotype is CVIET (Thomsen et al., 2013). Another mutation conferring resistance to CQ is the N86Y allele of *pfmdr-1* (Ebel et al., 2020). The role of *pfmdr-1* mutations (N86Y, Y184F, S1034C, and D1246Y) in facilitating *in vitro* and *in vivo* CQR has received a lot of interest in research (Oladipo et al., 2015). CQR advances in three particular ways in each newly affected country, including its increasing spread over a number of locations and regions, the increased prevalence of resistant strains in each area, and the increase in the intensity of resistance (Trape, 2001).

The first reported cases of CQ-resistant infections were confirmed in non-immune tourists in Kenya and Tanzania in 1978, followed by semi−immune Kenyans in 1982 (Shretta et al., 2000). According to Annual Report: Diagnosis, Treatment and Prevention of Malaria: Nairobi, Kenyan Ministry of Health, 61–80% of isolated parasites were resistant to CQ, and 30% of cases treated with CQ experienced clinical treatment failure (Shretta et al., 2000). According to a study, 92% (22 of 24) of isolates assessed for CQR carried the point mutation, Asn to Tyr in *pfmdr-1* codon 86, which was the most frequently reported sequence variation related to CQR in Africa. Furthermore, 83% (20 of 24) had an Asp to Tyr mutation, and both mutations were observed in 82% (18 of 24) of the isolates. Both alleles have reported association with CQR. In addition, 75% (18 of 24) of samples each had polymorphisms on both *cg2* and *pfmdr-l* (Omar et al., 2001). Polymorphisms in the *cg2* gene, located in the 36-kb region on chromosome 7, are also related to CQR. This gene is characterized by 12 point mutations and three length polymorphisms, namely, kappa, gamma, and omega, and are associated with clones encompassing CQR phenotypes (Viana et al., 2006). Durand et al. (1999) used DNA sequencing to confirm the association of *cg2* with CQR. It was found that the *cg2* genotype including identical κ14 repeats and particular ω16 repeats displayed a strong association with CQR in all *P. falciparum* isolates from the African countries included in the study. CQR was heightened during the 1980s but continued to be the first-line treatment for uncomplicated malaria. In 1998, revised guideline for malaria treatment officially replaced CQ with a combination of SDX and PYR (Shretta et al., 2000).

In Tanzania, CQR was first demonstrated in semi-immune citizens in 1982 and in 1983 resistance spread and was reported at 34% among a Zanzibar school population. Furthermore, studies between 1982 and 1985 in various Tanzanian revealed a 20% average *in vivo*-resistant rate in schoolchildren (Trape, 2001). Consequently, CQ was the official first-line antimalarial drug for uncomplicated malaria until the end of July 2001. The parliament of Tanzania was notified by the Minister of Health, during the 2002–2003 budgetary session, that a decision was made to suspend CQ as the first-line antimalarial drug. The grounds for that decision were based on researched evidence of high cure-rate failure observed for CQ, of approximately 60% (Mubyazi and Gonzalez-Block, 2005).

By 1983, CQR had been reported in several sub-Saharan countries such as Madagascar, Burundi, Sudan, Uganda, Zambia, Comoros, Burkina Faso, Malawi, and Mozambique (Nuwaha, 2001).

The prevalence of CQR in Malawi by 1992 was an estimated 85% (Kublin et al., 2003), therefore prompting Malawi to be the first country in sub-Saharan Africa to terminate the routine use of CQ in 1993. This was supported by the increased failure of CQ treatment and its inability to produce sufficient clinical and hematological recovery. In 1990, more than 80% of Malawian children treated with CQ presented high-level parasitological resistance (Bloland et al., 1993). The removal for CQ in Malawi was supplemented by reduced prevalence of the CQR molecular marker, *pfcrt* T76, from 85% to 13% between 1992 and 2000 (Kublin et al., 2003). Widespread sensitivity to CQ was observed in Malawi after the removal of CQ as the first-line drug for malaria treatment. In 2005, a clinical trial estimated CQ efficacy at 99%, 12 years after removal. Furthermore, CQ susceptibility was maintained when used for recurring malaria episodes (Takala-Harrison and Laufer, 2015).

In 1999, a 90% prevalence of the *pfcrt* K76T mutation was identified in infected Mozambican children. A trial in southern Mozambique estimated an average of 47% for CQ's clinical efficacy between 2001 and 2002. Furthermore, another study revealed more than 90% occurrence of the mutant CVIET haplotype in the same area (Thomsen et al., 2013). This led to the abandonment of CQ treatment for malaria in 2003 (Galatas et al., 2017).

By 1984, CQR was observed in countries including Gabon, Angola, Namibia, Senegal, Zimbabwe, and South Africa (Nuwaha, 2001). A study in Senegal examined the prevalence of *pfcrt* to determine the molecular level of CQR. A high prevalence of single *pfcrt* CVMNK wild-type haplotype was observed in the study, and the data suggested that increased prevalence of *pfcrt* wild types is occurring country-wide (Ndiaye et al., 2012). The use of CQ was abandoned by health authorities in 2003

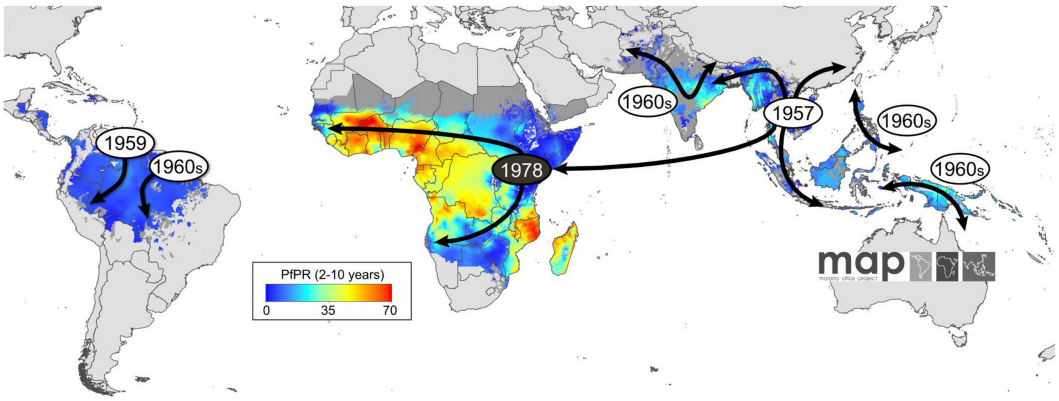


FIGURE 1 | The appearance and global spread of chloroquine resistance (CQR) in *P. falciparum*. Resistance is thought to have arisen in at least six independent origin (gray circles) and moved progressively as a CQ-driven selective sweep, including from Asia to Africa, where it established itself on the East coast in the late 1970s (black circle). The geographic spread of CQR is overlaid onto a current map of *P. falciparum* endemicity modeled for 2010. This map was derived from *P. falciparum* parasite rate (PfPR) surveys, age standardized to the 2- to 10-year age range, using model-based geostatistics (Ecker et al., 2012).

(Trape et al., 1998). In Zimbabwe, 50% of the population reside in malaria-risk areas, and malaria transmission occurs in 51 out of 59 administrative regions (Mlambo et al., 2007). Until 1983, there were no cases of CQR in Zimbabwe; however, in 1984, the first seven cases were reported in Zambezi Valley. CQR was initially confined to the Zambezi Valley, but by 1990, it was clear that the problem had spread throughout the country (Makono and Sibanda, 1999). According to a study facilitated in the Njelele area of Gokwe following a malaria outbreak, 83% of infections were CQ resistant, with only 3% of cases responding to CQ treatment (Mharakurwa and Mugochi, 1994). The establishment of maintainable national surveillance approaches was necessary to counteract the increasing CQR and monitor its spread throughout the country (Makono and Sibanda, 1999). The clinical failure of CQ was exceptionally high and, thus, only remained the drug of choice until 2000 (Mlambo et al., 2007).

Since 1932, South Africa had faced sporadic malaria epidemics, with KwaZulu Natal, Mpumalanga, and Limpopo bearing the greatest burden within the country (Knight et al., 2009). In 1985, KwaZulu Natal had its first report of *in vitro* CQR, which spread under persistent CQ pressure, thereby leading to an escalation in cases and treatment failures. In spite of being treated with CQ four times, 3% of malaria-treated patients remained positive (Ukpe et al., 2013). CQ was replaced and no longer used as the first-line drug for malaria treatment by February 1988 (Knight et al., 2009). Malaria parasites stayed susceptible to CQ in Mpumalanga and Limpopo until the mid-to-late 1990s. The number of CQ-resistant parasites escalated, thus prompting the replacement of CQ as the drug of choice in Mpumalanga and Limpopo in 1997 and 1998, respectively (Maharaj et al., 2013).

THE SPREAD OF SULFADOXINE–PYRIMETHAMINE RESISTANCE IN SUB-SAHARAN AFRICA

Africa, most predominantly, sub-Saharan Africa, still remains a home to malaria, bearing high morbidity, mortality, and risk of transmission, in spite of significant reduction in malaria incidence in the region, with children and pregnant women being the most vulnerable (Zareen et al., 2016). Malaria in pregnancy is estimated to cause hundreds of thousand infant deaths every year (Kajubi et al., 2019). Therefore, in 2012, WHO recommended SP as an intermittent preventive treatment (IPT) for fetal malaria prevention in pregnant women across malarious regions (Capan et al., 2010). However, efficacy of chemoprophylaxis with SP was halted by poor compliance of patients, which has led to emergence of drug-resistant strains of *P. falciparum* (Oyibo and Agomo, 2011).

The SP is a combination comprising SDX, which inhibits DHPS, a bifunctional protein that interacts with hydroxymethylpterin pyrophosphokinase, and PYR, which inhibits DHFR bifunctional protein (Nzila et al., 2000); it, therefore, synergistically inhibits the folate biosynthesis pathway of *P. falciparum* (Ahmed et al., 2006).

The molecular basis of *Plasmodium* resistance to SP is shown to be point mutations (Gatton et al., 2004) at seven sites in the DHPS gene (*dhfr*) that confer resistance to PYR and five sites in the DHFR (*dhps*) gene that confer resistance to SDX. Different combinations of mutations in each gene lead to varied resistance levels of SP (Pearce et al., 2003). For *pfdhps*, several point mutations—Ser → Ala at codon 436, Ala → Gly at codon 437, Lys → Glu at codon 540, Ala → Gly at codon 581, and Ser → Phe at codon 436—coupled with either Ala → Thr or Ala → Ser at codon 613 confer resistance to SDX. For *pfdhfr*, a point mutation of Ser → Asn changes at position 108 (S108N); resistance mutations Asn → Ile at codon 51 and/or Cys → Arg at codon 59 (C59R) and a Ser → Thr mutation at position 108 (S108N) with an Ala → Val change at position 16, synergistically confering PYR resistance in *Plasmodium* species (Kublin et al., 2003). The DHPS domain of the bifunctional 7,8-dihydro-6-hydroxymethylpterin pyrophosphokinase (PPPK)-dhps enzyme for SDX and the DHFR domain of the bifunctional DHFR-thymidylate synthase (DHFRTS) enzyme for PYR are key targets of SP drug, as these enzymes are responsible for the folic acid biosynthetic pathway (Basco et al., 1998). After approximately

50 years of SP's introduction, kinetics in children is still poorly understood; therefore, dosing is predominantly based on age for practical reasons, which may result in the administration of inaccurate quantities (De Kock et al., 2018). Increasing resistance to SP is linked to the stepwise acquisition of specific point mutations at specific codons in the *dhfrdhps* genes, which alter the drug-binding sites of SP (Oyibo and Agomo, 2011). However, the prophylaxis treatment of malaria was selected for double and triple mutations in *dhps* and *dhfr* genotypes, respectively, which are implicated in *Plasmodium* resistance to SP across sub-Saharan Africa (Capan et al., 2010; Sridaran et al., 2010). Accumulation of mutations within the genes (N51I, C59R, and S108N) in *pfdhfr* and mutations in codons A437G, K540E, and A581G within *pfdhps*, which confer resistance of *Plasmodium* to SP (Takala-Harrison and Laufer, 2015), has spread across sub-Saharan Africa (**Figure 2**).

In Benin, *pfdhps*, which confers resistance to SDX, contained a mixture of wild-type and mutant alleles (A437 and G437). High mutated alleles of *pfdhfr* were reported for codons N51I, C59R, and codon S108N. The *pfdhfr* (N51I, C59R, and S108N) proportion of single, triple, and quadruple mutations was very high in the study population (Ogouyèmi-Hounto et al., 2013). In Nigeria, parasite clearance was observed in pregnant women using SP (Kalanda et al., 2006), with mutations, N51I, C59R, S108N, and the triple mutation conferring a high level of *Plasmodium* resistance to SP (Iriemenam et al., 2012). A study conducted in Mali between August 2004 and January 2006 and in Ghana in 2003 supported the use of SP + AS (SP and amodiaquine) to treat young children with uncomplicated *P. falciparum* malaria and IPT in pregnant women (Tagbor et al., 2006; Tekete et al., 2011; Maiga et al., 2015). Additionally, in Mali, prevalence of each of the three *pfdhfr* mutations at codons N51I, C59R, and S108N was observed in 2007. Notably, *pfdhps* mutations at codons A437G and K540E as well as *pfdhfr* triple and *pfdhps* A437G quadruple mutations were present (Dicko et al., 2010). SP resistance was still quite low (Figueiredo et al., 2008; Dicko et al., 2012) in 2008 but rapidly increased in Mali in 2010 and Burkina Faso in 2012 (Coulibaly et al., 2014). The triple mutant *pfdhfr* in Africa evolved from the southeast Asian lineage (Coulibaly et al., 2014), while the N51I + S108N double-mutant *pfdhfr* alleles are a local origin. Both the *pfdhps* double mutants (Gly-437 and Glu-540) and the *pfdhfr* triple mutants were individually associated with SP treatment failure in children aged less than 5 years, with *pfdhps* and *pfdhfr* quintuple mutations being highly associated with SP treatment failure in Ibadan, Nigeria (Happi et al., 2005). The *pfdhfr* N51I, C59R, S108N triple mutation had a prevalence close to 100% in Cameroon, western Africa. The most frequent *pfdhps* mutation there was A437G, with a prevalence of 76.5% and had a higher significant prevalence in pregnant women with SP uptake (Chauvin et al., 2015).

Ghana, a western African country, in 2003 also recorded the quadruple mutation in *pfdhps* A437G and *pfdhfr*—N51I, C59R, and S108N (Marks et al., 2005)—but SP was still effective in 2001 for treating uncomplicated *P. falciparum* malaria whether as monotherapy or combined therapy in Cameroon (Basco et al., 2002). Meanwhile, emergence of SP combination resistance was sustained for many years before introduction of ACTs as a frontline antimalarial in Africa (Marks et al., 2005). There is a strong relationship between protective efficacy and the frequency of resistance mutations, as negative correlation exists between SP prophylaxis efficacy and parasite mutations, indicating that as resistance increases, protective efficacy decreases. This decreased efficacy of SP on IPTi was reported across Ghana and Gabon where decreased efficacy was linked to *pfdhfr* triple mutation (Griffin et al., 2010). A study in Niger showed a high proportion of *pfdhfr* N51I, C59R, and S108N haplotypes associated with resistance to PYR and *pfdhps* S436AFH and A437G mutations associated with reduced susceptibility to SDX (Grais et al., 2018). In 2004, a molecular genotyping study conducted in Laine, Guinea, found three *pfdhfr* mutations in 85.6% patients and quintuple *pfdhfr*/*pfdhps* mutations in 9.6% showing a progressive increase in resistance to SP (Bonnet et al., 2007). No *pfdhfr* I164L or *pfdhps* A581G mutations were found, but *pfdhps* K540E mutation was found in Mali, with a very low prevalence in Mali and Burkina-Faso (Coulibaly et al., 2014).

In Central Africa, between 2002 and 2003, there was adequate clearance of parasitemia in children treated with SP with low presence of *pfdhfr* S108N, *pfdhfr* triple, and *pfdhfr*/*pfdhps* quadruple mutation genotypes. This shows that studies conducted a decade ago still showed some SP efficacy despite accompanying resistance genotypes. The parasite population of Uige, Angola, revealed high frequency mutations in *pfcrt*, *pfdhps*, and *pfdhfr*, conferring resistance to CQ and SP (Coulibaly et al., 2014).

The frequency of the double A437G, K540E mutant *pfdhps* allele (conferring SDX resistance) increased from 200% to 300%, and the triple mutant *pfdhfr* (N51I, C59R, and S108N) allele (conferring PYR resistance) increased by 37–63% in Tanzanian population when SP was used as a first-line antimalarial (Malisa et al., 2010). A study conducted in 2002 in Tanzania confirmed the presence of *pfdhfr*/*pfdhps* mutations even when SP was efficient (Mbugi et al., 2006). SP resistance was reported across regions of Tanzania where the triple mutations (N51I, C59R, and S108N) were constantly predominant across all the regions, implying a high level of SP resistance (Schönfeld et al., 2007; Bertin et al., 2011; Baraka et al., 2015), which was corroborated by same result in Benin, a western African country (Bertin et al., 2011). No relationship exists between the number of mutations and the degree of parasitological resistance (Eriksen et al., 2004). Additionally, Tanzania not only reported SP failure, but its prophylaxis was of no significant effect in northern Tanzania in 2006, as triple mutations previously reported in other parts of Africa were also detected (Gesase et al., 2009). The highest mutations of *pfdhfr* and *pfdhps* genes were predominantly distributed across eastern and southern Africa where SP use has been the highest (Enosse et al., 2008; Griffin et al., 2010). In 2005, northwest Ethiopia displayed a gradual reduction of individual *pfdhps*/*pfdhfr* mutations; triple, quadruple, and quintuple mutations were observed after 5 years of SP withdrawal as a frontline antimalarial (Hailemeskel et al., 2013), but quintupled in western Kenya between 2008 and 2009 (Iriemenam et al., 2012). Moreover, by 2000, the East African Network for Monitoring Antimalarial Treatment (EANMAT)

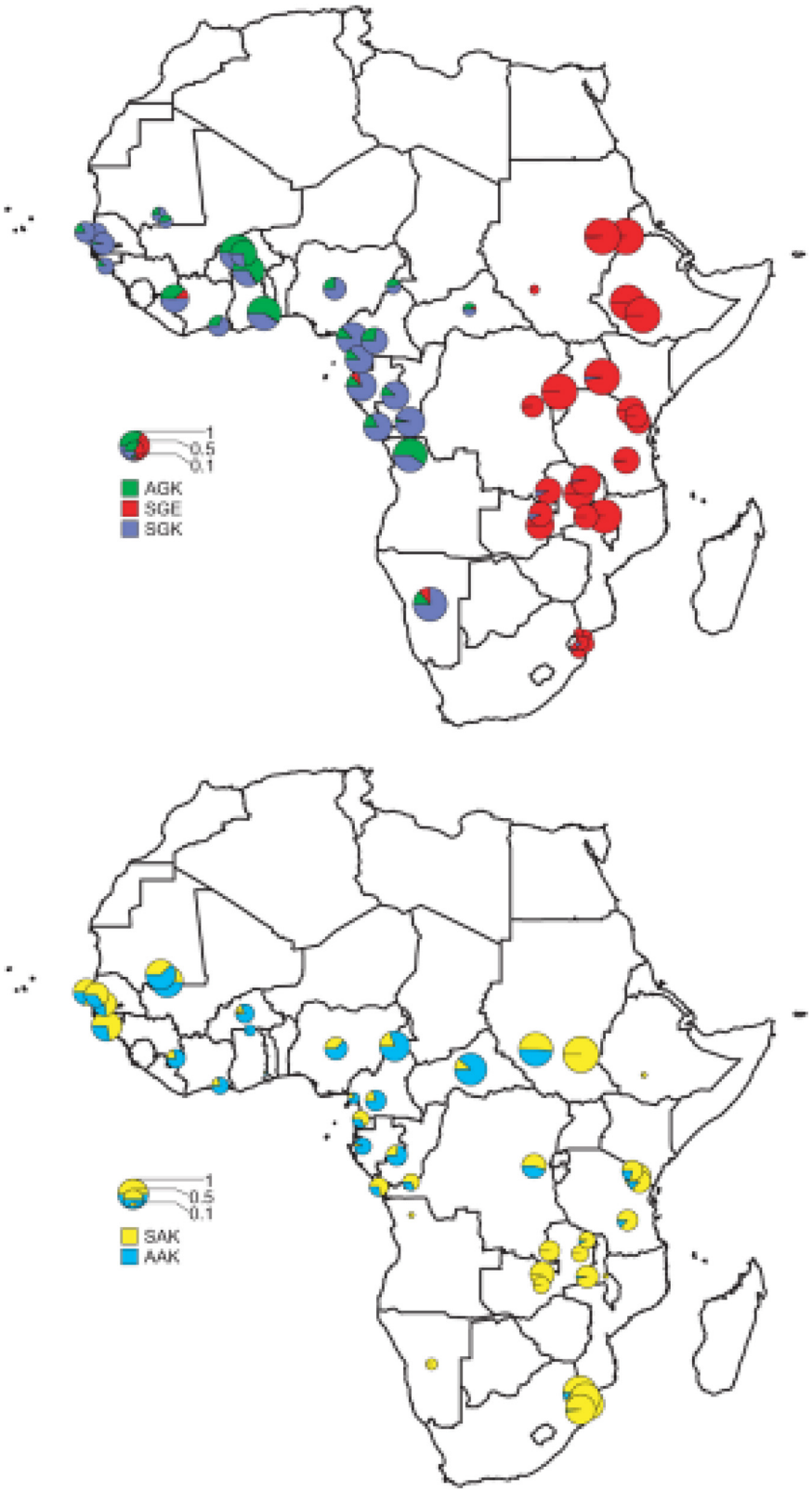


FIGURE 2 | The distribution of the major *pfdhps* alleles across sub-Saharan Africa. Resistant alleles; the upper map shows the relative proportions of the three major resistance alleles, SGK, AGK, and SGE. Wild-type alleles; the lower map shows the ratio of SAK and AAK alleles among wild-type *pfdhps* alleles (Pearce et al., 2009).

reported SP failure across East Africa, with Rwanda and Burundi exceeding the critical 25% value SP failure rates (Rapuoda et al., 2003); this could be due to a low uptake of SP by pregnant women due to weak health system and sub-optimal implementation policy, among other factors (Martin et al., 2020). The *pfdhps* A581G and *pfdhfr* I164L mutations, which confer SP resistance, have been found to be unevenly distributed across East Africa, suggesting a high level of SP resistance, as indicated by the prevalence of the K540E mutation, found across East Africa (Naidoo and Roper, 2011).

Zimbabwe reported high prevalence of *pfcrt* (K76T) to be 64%, 82%, and 92% in Chiredzi, Kariba, and Bindura, respectively. On the *pfdhfr* locus, the presence of triple mutations at codons N51I, C59R, and S108N was approximately 50% across the three locations, therefore indicative of widespread prevalence of molecular markers associated with CQ and PYR resistance in Zimbabwe (Schleicher et al., 2018). KwaZulu Natal in South Africa had reported SP resistance in 2000 and thus changed its malaria policy to the use of ACTs for malaria treatment (Knight et al., 2009).

In 2010, the quadruple mutation (A437G, N51I, C59R, and S108N) haplotypes were widespread throughout sub-Saharan Africa, particularly in Uganda, Mali, Kenya, and Malawi (Takechi et al., 2001; Osman et al., 2007), but the quintuple and sextuple mutants were found only in specific regions of Africa (Walker et al., 2017). The quintuple mutants, which plateaued in East and Southern Africa, were relatively rare in west Africa and parts of central Africa. Estimates of the prevalence of the *pfdhps*-A581G mutation suggest that the sextuple haplotype had become established in Rwanda, Burundi, D R Congo, southwestern Uganda, northwestern Tanzania, southeastern Kenyan, and northeastern Tanzanian borders.

Genotyping of *pfdhfr* responsible for PYR resistance showed no mutation in codons A16V, C59R, or I164L, while there was 84% mutation at codons N51I and S108N (Osman et al., 2007). Additionally, linkage exists between CQ mutation, *pfcrt* K76T, SDX mutation, *pfdhps* K540E, and PYR mutation *pfdhfr* (N51I and S108N), thus conferring parasite resistance against both CQ and SP. The C59R mutation observed in Mali, Kenya, and Malawi was, however, absent in Sudan in spite of 80% prevalence of *pfdhps* K540E. *Pfdhfr* N51I and S108N occurred in more than 80% (Osman et al., 2007). The A437G mutation at *pfdhps* and the triple mutation (N51I, C59R, and S108N) at *pfdhfr* associated with SP resistance is concurrent with those found even outside sub-Saharan Africa (Checchi et al., 2002). In six countries of Eastern and Southern Africa, more than 90% prevalence of *pfdhps*-K540E was implicated in SP resistance (Iriemenam et al., 2012). However, countries like Cameroon, Ghana, Zambia, Mozambique, Mali, and Zimbabwe had three doses of SP in their policy for all pregnant women, which seemed efficacious with resistant genotypes concurrently found in these countries (Mayor et al., 2008; Maiga et al., 2011; Kayentao et al., 2013). The quadruple mutation was uncommon compared with triple mutation, which confers a major resistance to SP in Africa (Peters et al., 2007). Conclusively, a systematic analysis conducted in African countries between 1996 and 2006 reported delayed clearance alluding to SP treatment failure (ter Kuile et al., 2015).

COMBATING CHLOROQUINE AND SULFADOXINE–PYRIMETHAMINE RESISTANCE

Several publications confirm the widespread resistance of CQ and SP, with substantial occurrence in Asian and African countries. The reasons for this hinged on various factors, which include, but is not limited to, the overuse of these antimalarial medications, underdosage of the medications in the treatment of active malaria infection, and high susceptibility of the parasite to adapt at the genetic metabolic levels rapidly (Hyde, 2007). General preventive as well as some of the specific measures have been considered as ways to combat further spread of antimalarial resistance.

Vector Control

Vector control makes use of ITNs/long-lasting insecticidal nets (LLINs) and IRS, which are the mainstay of malaria control. These two strategies have proven useful in the reduction of local malaria transmission by protecting susceptible individuals against infective mosquito bites, thereby leading to a reduction in the level of malaria intensity. When sleeping under the net, the net serves as a means of preventing mosquito contact or bite. Reducing the vector population through this means can indirectly serve as a measure to mitigate the spread of CQ- and SP-resistant *Plasmodium*. In fact, since 2007, the WHO has recommended universal coverage with ITN (Macintyre et al., 2011; World Health Organisation (WHO), 2011). IRS has also been reportedly effective in reducing the challenge of transmission. It serves as a means to terminating these vectors, thereby reducing the possible malaria morbidity and mortality (World Health Organisation (WHO), 2011). When vector control measures are well applied in combination, it is evident that the outcome of malaria transmission (CQ/SP resistant in this case) will be mitigated (World Health Organisation (WHO), 2010; Thu et al., 2017).

The Introduction of Artemisinin-Based Combination Therapies

An important strategy continues to be the use of effective antimalarials. However, this is threatened by the extensive resistance observed in the parasite toward the most affordable classes of drugs, thus highlighting the need for novel antimalarials (Winstanley et al., 2004) and ultimately a vaccine. The widespread resistance of the known frontline antimalarial drugs—CQ and SP—in April 2002 led to the introduction of ACTs as the first-line treatment for malaria by the WHO. Notably in the treatment of *P. falciparum*, it is recommended that two or more drugs must be combined. The combined drugs must have different modes of action. Since artemisinin-based compounds (dihydroartemisinin, artemether, and artesunate) are fast acting, they are usually combined with other classes of antimalarial with long half-life such as mefloquine, amodiaquine, lumefantrine, and piperaquine (Bell and Winstanley, 2005). The ACTs, which

come in various combinations, is used to replace CQ and SP in order to prevent further increasing morbidity and mortality resulting from the observed resistance in the initial use of first-line medications (World Health Organisation (WHO), 2003; Dalrymple, 2009). The high efficacy and ability to interrupt the development of resistance supported the use of ACTs as the core of malaria treatment (Laxminarayan et al., 2006). The artemisinin component of ACTs has the ability to reduce parasitemia; thus, the progression from uncomplicated malaria to a severe form of the disease can be prevented by early treatment with ACTs, consequently reducing the amount of severe malaria cases and the rate of malaria mortality (Ndeffo Mbah et al., 2015). By June 2008, most malaria endemic nations have adopted the ACTs as its front-line antimalarial medication (World Health Organisation (WHO), 2008). Part of the measures put in place to avoid similar medication resistance as seen in CQ and SP is the disallowance of artemisinin monotherapy medications. This was part of the World Health Assembly resolution of 2007 (World Health Organisation (WHO), 2008).

Medication Adherence and Avoiding Self-Medication

Among humans, there are a variety of beliefs that can impact our behavior significantly. These beliefs are scarcely changed and can therefore affect adherence to antimalarials such as ACTs, thus highlighting the need for patients' beliefs to be considered when care is offered (Amponsah et al., 2015). Numerous interventions have been developed and employed in malaria endemic areas to promote adherence to medications. Clear guidelines to better adherence can be achieved through deeper understanding of the relative efficacy of each intervention (Fuangchan et al., 2014). The understanding of malaria treatment could be improved by community education, such as training of dispensers and promoting public awareness, thus improving adherence to antimalarials (Denis, 1998). To ensure malaria treatment approaches continue to be effective and improve the possibility of a positive outcome, medication adherence should be evaluated regularly, thereafter used as tools to encourage appropriate action, if necessary (Gerstl et al., 2015).

The CQ/SP resistance developed partly due to widespread presumptuous use of these medications without prior objective confirmation of the diagnosis. Due to the poverty and economic setting of many malaria endemic countries especially in sub-Saharan Africa, most people self-medicate and thereby inappropriately use CQ/SP (Idowu et al., 2006). To this end, the WHO essentially recommended and introduced affordable point-of-care rapid diagnostic tests for definitive malaria testing and diagnosing before the usage of an antimalarial even in under-resourced settings. This is expected to be accompanied by correct and complete usage of the antimalarial especially in this era of ACTs (White et al., 2009; Thu et al., 2017).

Combating the Counterfeit Medication Market

Generally, counterfeit medication presents a great challenge globally, thus contributing to the burden presented by the development of antimalarial drug resistance. Studies done in Eastern DR Congo showed that majority of the CQ has its active ingredient to be significantly underdosed while the SP was overdosed in terms of the active ingredients (Atemnkeng et al., 2007). A similar study in Nigeria corroborated the fact that the menace of counterfeit medication is huge, with 50% of CQ dispensed by street vendors underdosed of the active ingredients (Idowu et al., 2006). In light of this, the development of CQ and SP resistance is substantiated. However, going forward, this can be combatted by action at the policy level. Enacting the laws guiding counterfeit medications and strengthening of the institutions responsible will go a long way to reduce, if not totally remove, counterfeit medications in the society. This can mitigate the CQ/SP resistance in the long run (Yadav and Rawal, 2016). Another way to combat the problem of counterfeit medication is the provision of novel, effective, and affordable antimalarial drugs. This issue is addressed by the Medicines for Malaria Venture, whose aim consists of the discovery, development, and the facilitation of those antimalarials (Medicines for Malaria Venture [MMV], 2020). MMV is a non-profit Swiss foundation, officially introduced in 1999. It was one of the first public–private partnerships created to address the drug innovation gap of malaria (Bathurst and Hentschel, 2006). Developing medicines for children addresses the population most at risk of dying from malaria, thus serving as motivation for MMV to develop child-friendly formulations of current antimalarials and to develop next-generation medicines (Medicines for Malaria Venture [MMV], 2020). In 2009, Novartis and MMV developed an artemether–lumefantrine formulation, called *Coartem® Dispersible*, adapted to cater to the needs of children affected by *P. falciparum* malaria and was the first successfully co-developed product launched by the MMV, in which 400 million products were distributed (Premji, 2009; Medicines for Malaria Venture [MMV], 2020). MMV continues to develop and improve access to affordable medicines to lessen the global burden of malaria.

Reintroduction of Chloroquine in the Chloroquine-Sensitive Environment

Various studies have shown that with replacement of CQ/SP as the first line of treatment and subsequent replacement with ACTs, there was notable decrease in the prevalence of pfcrt K76T present in such communities. This was first demonstrated in Malawi in 2004, with the prevalence of pfcrt K76T reaching an undetectable level (Plowe, 2014).

Similar studies have validated this point in China and Zambia, with the outcome of both studies showing a resurgence of CQ sensitivity (Mwanza et al., 2016; Lu et al., 2017b). Considering the role of CQ/SP safety profile and affordability, this looks promising, and reintroduction of CQ/SP may be the way to go in the near future.

CONCLUSION

In sub-Saharan Africa, where the burden of malaria infection is great, one of the core challenges for malaria control is the emergence and spread of antimalarial drug resistance. Although other strategies are employed for malaria control such as ITNs and IRS, which are the foundation of the malaria control agenda, antimalarial drugs are essential to reduce transmission. However, because of their accessibility and affordability, CQ and SP—both first-line drugs in the past—were used excessively, which resulted in the establishment of resistance, which in turn hindered malaria control in sub-Saharan Africa. This prompted the introduction of ACTs, which decreased the pressure on CQ and SP in some regions of sub-Saharan Africa, thus encouraging their reintroduction in those regions. Ongoing antimalarial drug resistance should be closely monitored in sub-Saharan Africa to prevent the establishment and spread of resistance and to detect the return of sensitivity, which could result in the possible reintroduction of specific antimalarials.

AUTHOR CONTRIBUTIONS

AR: writing, editing, and collation. LM, OO, and OA: writing and editing. RM: supervision and editing. MA and MO: conceptualization, supervision, editing, and funding.

FUNDING

This work was supported by the National Research Foundation, (NRF) South Africa (Grant No. 120368).

REFERENCES

Ahmed, A., Lumb, V., Das, M. K., Dev Wajihullah, V., and Sharma, Y. D. (2006). Prevalence of mutations associated with higher levels of sulfadoxine-pyrimethamine resistance in *plasmodium falciparum* isolates from car nicobar Island and Assam, India. *Antimicrob. Agents Chemother.* 50, 3934–3938. doi: 10.1128/aac.00732-06

Ajayi, N. A., and Ukwaja, K. N. (2013). Possible artemisinin-based combination therapy-resistant malaria in Nigeria: a report of three cases. *Rev. Soc. Bras. Med. Trop.* 46, 525–527. doi: 10.1590/0037-8682-0098-2013

Amponsah, A. O., Vosper, H., and Marfo, A. F. A. (2015). Patient related factors affecting adherence to antimalarial medication in an urban estate in ghana. *Malar. Res. Treat.* 2015, 452539–452539.

Atemnkeng, M. A., Chimanuka, B., and Plaizier-Vercammen, J. (2007). Quality evaluation of chloroquine, quinine, sulfadoxine–pyrimethamine and proguanil formulations sold on the market in East Congo DR. *J. Clin. Pharm. Ther.* 32, 123–132. doi: 10.1111/j.1365-2710.2007.00797.x

Awasthi, G., and Das, A. (2013). Genetics of chloroquine-resistant malaria: a haplotypic view. *Memor. Instit. Oswal. Cruz* 108, 947–961. doi: 10.1590/0074-0276130274

Baraka, V., Ishengoma, D. S., Fransis, F., Minja, D. T. R., Madebe, R. A., Ngatunga, D., et al. (2015). High-level *Plasmodium falciparum* sulfadoxine-pyrimethamine resistance with the concomitant occurrence of septuple haplotype in Tanzania. *Mal. J.* 14:439.

Basco, L. K., Tahar, R., and Ringwald, P. (1998). Molecular basis of in vivo resistance to sulfadoxine-pyrimethamine in African adult patients infected with *Plasmodium falciparum* malaria parasites. *Antimicr. Agents Chemother.* 42, 1811–1814. doi: 10.1128/aac.42.7.1811

Basco, L., Same-Ekobo, A., Ngane, V., Ndounga, M., Metoh, T., Ringwald, P., et al. (2002). Therapeutic efficacy of sulfadoxine-pyrimethamine, amodiaquine and the sulfadoxine-pyrimethamine-amodiaquine combination against uncomplicated *Plasmodium falciparum* malaria in young children in Cameroon. *Bull. World Health Organiz.* 80, 538–545.

Bathurst, I., and Hentschel, C. (2006). Medicines for malaria venture: sustaining antimalarial drug development. *Trends Parasitol.* 22, 301–307. doi: 10.1016/j.pt.2006.05.011

Bell, D., and Winstanley, P. (2005). Current issues in the treatment of uncomplicated malaria in Africa. *Br. Med. Bull.* 71, 29–43. doi: 10.1093/bmb/ldh031

Bertin, G., Briand, V., Bonaventure, D., Carrieu, A., Massougbodji, A., Cot, M., et al. (2011). Molecular markers of resistance to sulphadoxine-pyrimethamine during intermittent preventive treatment of pregnant women in Benin. *Mal. J.* 10:196. doi: 10.1186/1475-2875-10-196

Bin Dajem, S. M., and Al-Qahtani, A. (2010). Analysis of gene mutations involved in chloroquine resistance in *Plasmodium falciparum* parasites isolated from patients in the southwest of Saudi Arabia. *Ann. Saudi Med.* 30, 187–192. doi: 10.4103/0256-4947.62826

Birnbaum, J., Scharf, S., Schmidt, S., Jonscher, E., Hoeijmakers, W. A. M., Flemming, S., et al. (2020). A Kelch13-defined endocytosis pathway mediates artemisinin resistance in malaria parasites. *Science* 367:51. doi: 10.1126/science.aax4735

Bloland, P. B., Lackritz, E. M., Kazembe, P. N., Were, J. B., Steketee, R., and Campbell, C. C. (1993). Beyond chloroquine: implications of drug resistance for evaluating malaria therapy efficacy and treatment policy in Africa. *J. Infect. Dis.* 167, 932–937. doi: 10.1093/infdis/167.4.932

Bonnet, M., Roper, C., Félix, M., Coulibaly, L., Kankolongo, G. M., and Guthmann, J. P. (2007). Efficacy of antimalarial treatment in Guinea: in vivo study of two artemisinin combination therapies in Dabola and molecular markers of resistance to sulphadoxine-pyrimethamine in N'Zérékoré. *Malar. J.* 6:54.

Capan, M., Mombo-Ngoma, G., Makristathis, A., and Ramharter, M. (2010). Anti-bacterial activity of intermittent preventive treatment of malaria in pregnancy: comparative in vitro study of sulphadoxine-pyrimethamine, mefloquine, and azithromycin. *Mal J.* 9:303. doi: 10.1186/1475-2875-9-303

Chauvin, P., Menard, S., Iriart, X., Nsango, S. E., Tchioffo, M. T., Abate, L., et al. (2015). Prevalence of *Plasmodium falciparum* parasites resistant to sulfadoxine/pyrimethamine in pregnant women in Yaoundé, Cameroon: emergence of highly resistant *pfdhfr/pfdhps* alleles. *J. Antimicrob. Chemother.* 70, 2566–2571. doi: 10.1093/jac/dkv160

Checchi, F., Durand, R., Balkan, S., Vonhm, B. T., Kollie, J. Z., Biberson, P., et al. (2002). High *Plasmodium falciparum* resistance to chloroquine and sulfadoxine-pyrimethamine in Harper, Liberia: results in vivo and analysis of point mutations. *Transact. R. Soc. Trop. Med. Hygiene* 96, 664–669. doi: 10.1016/s0035-9203(02)90346-9

Chinappi, M., Via, A., Marcatili, P., and Tramontano, A. (2010). On the mechanism of chloroquine resistance in *Plasmodium falciparum*. *PLoS One* 5:e14064.

Chourasia, M. K., Raghavendra, K., Bhatt, R. M., Swain, D. K., Valecha, N., and Kleinschmidt, I. (2017). Burden of asymptomatic malaria among a tribal population in a forested village of central India: a hidden challenge for malaria control in India. *Public Health* 147, 92–97. doi: 10.1016/j.puhe.2017.02.010

Coban, C. (2020). The host targeting effect of chloroquine in malaria. *Curr. Opin. Immunol.* 66, 98–107. doi: 10.1016/j.coi.2020.07.005

Conrad, M. D., and Rosenthal, P. J. (2019). Antimalarial drug resistance in Africa: the calm before the storm? *Lancet Infect. Dis.* 19, e338–e351.

Conroy, A. L., Datta, D., and John, C. C. (2019). What causes severe malaria and its complications in children? Lessons learned over the past 15 years. *BMC Med.* 17:52.

Coulibaly, S. O., Kayentao, K., Taylor, S., Guirou, E. A., Khairallah, C., Guindo, N., et al. (2014). Parasite clearance following treatment with sulphadoxine-pyrimethamine for intermittent preventive treatment in Burkina-Faso and Mali: 42-day in vivo follow-up study. *Mal. J.* 13:41. doi: 10.1186/1475-2875-13-41

Cowman, A. F., Karcz, S., Galatis, D., and Culvenor, J. G. (1991). A P-glycoprotein homologue of *Plasmodium falciparum* is localized on the digestive vacuole. *J. Cell Biol.* 113, 1033–1042. doi: 10.1083/jcb.113.5.1033

Dagnogo, O., Ako, A. B., Ouattara, L., Dago, N. D., Coulibaly, D. N. g, Touré, A. O., et al. (2018). Towards a re-emergence of chloroquine sensitivity in Côte d'Ivoire? *Mal. J.* 17, 413–413.

D'Alessandro, U., and Buttiëns, H. (2001). History and importance of antimalarial drug resistance. *Trop Med. Int. Health* 6, 845–848. doi: 10.1046/j.1365-3156.2001.00819.x

Dalrymple, D. G. (2009). *Artemisia Annua, Artemisinin, ACTs and Malaria Control in Africa: The Interplay of Tradition, Science and Public Policy, Working Paper*. Geneva: Medicines for Malaria Venture.

De Kock, M., Tarning, J., Workman, L., Allen, E. N., Tekete, M. M., Djimde, A. A., et al. (2018). Population pharmacokinetic properties of sulfadoxine and pyrimethamine: a pooled analysis to inform optimal dosing in african children with uncomplicated Malaria. *Antimicr. Agents Chemother.* 62:e01370-17.

Denis, M. B. (1998). Improving compliance with quinine + tetracycline for treatment of malaria: evaluation of health education interventions in Cambodian villages. *Bull. World Health Organ.* 76(Suppl. 1), 43–49.

Desai, M., Gutman, J., Taylor, S. M., Wiegand, R. E., Khairallah, C., Kayentao, K., et al. (2015). Impact of sulfadoxine-pyrimethamine resistance on effectiveness of intermittent preventive therapy for malaria in pregnancy at clearing infections and preventing low birth weight. *Clin. Infect. Dis.* 62, 323–333. doi: 10.1093/cid/civ881

Dicko, A., Konare, M., Traore, D., Testa, J., Salamon, R., Doumbo, O., et al. (2012). The implementation of malaria intermittent preventive trialtreatment with sulphadoxine-pyrimethamine in infants reduced all-cause mortality in the district of Kolokani, Mali: results from a cluster randomized control. *Mal. J.* 11:73.

Dicko, A., Sagara, I., Djimdé, A. A., Touré, S. O., Traore, M., Dama, S., et al. (2010). Molecular markers of resistance to sulphadoxine-pyrimethamine one year after implementation of intermittent preventive treatment of malaria in infants in Mali. *Mal. J.* 9:9.

Dondorp, A. M., and Ringwald, P. (2013). Artemisinin resistance is a clear and present danger. *Trends Parasitol* 29, 359–360. doi: 10.1016/j.pt.2013.05.005

Dorsey, G., Kamya, M. R., Singh, A., and Rosenthal, P. J. (2001). Polymorphisms in the *Plasmodium falciparum pfcrt* and *pfmdr*-1 genes and clinical response to chloroquine in Kampala, Uganda. *J. Infect Dis.* 183, 1417–1420.

Durand, R., Gabbett, E., Di Piazza, J-P., Delabre, J-F., and Le Bras, J. (1999). Analysis of κ and ω repeats of the *cg*2 gene and chloroquine susceptibility in isolates of *Plasmodium falciparum* from sub-Saharan Africa. *Mol. Biochem. Parasitol.* 101, 185–197. doi: 10.1016/s0166-6851(99)00073-0

Ebel, E. R., Reis, F., Petrov, D. A., and Beleza, S. (2020). Shifts in antimalarial drug policy since 2006 have rapidly selected *P. falciparum* resistance alleles in Angola. doi: 10.1101/2020.09.29.310706

Ecker, A., Lehane, A. M., Clain, J., and Fidock, D. A. (2012). *pfcrt* and its role in antimalarial drug resistance. *Trends Parasitol.* 28, 504–514. doi: 10.1016/j.pt.2012.08.002

Ehlgen, F., Pham, J. S., de Koning-Ward, T., Cowman, A. F., and Ralph, S. A. (2012). Investigation of the *Plasmodium falciparum* food vacuole through inducible expression of the chloroquine resistance transporter (PfCRT). *PLoS One* 7:e38781. doi: 10.1371/journal.pone.0038781

Enosse, S., Magnussen, P., Abacassamo, F., ómez-Olivé, X. G., Rønn, A. M., Thompson, R., et al. (2008). Rapid increase of *Plasmodium falciparum dhfr/dhps* resistant haplotypes, after the adoption of sulphadoxine-pyrimethamine as first line treatment in 2002, in southern Mozambique. *Mal. J.* 7:115.

Eriksen, J., Mwankusye, S., Mduma, S., Kitua, A., Swedberg, G., Tomson, G., et al. (2004). Patterns of resistance and DHFR/DHPS genotypes of *Plasmodium falciparum* in rural Tanzania prior to the adoption of sulfadoxine—pyrimethamine as first-line treatment. *Transact. R. Soc. Trop. Med. Hyg.* 98, 347–353. doi: 10.1016/j.trstmh.2003.10.010

Fairhurst, R. M., and Dondorp, A. M. (2016). Artemisinin-resistant *Plasmodium falciparum* malaria. *Microbiol. Spect.* 4:13. doi: 10.1128/microbiolspec.EI1110-0013-2016

Ferreira, P. E., Holmgren, G., Veiga, M. I., Uhlén, P., Kaneko, A., and Gil, J. P. (2011). Pfmdr1: Mechanisms of Transport Modulation by Functional Polymorphisms. *PLoS One* 6:e23875. doi: 10.1371/journal.pone.0023875

Fidock, D. A., Nomura, T., Talley, A. K., Cooper, R. A., Dzekunov, S. M., Ferdig, M. T., et al. (2000). Mutations in the *P. falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol. Cell* 6, 861–871. doi: 10.1016/s1097-2765(05)00077-8

Figueiredo, P., Benchimol, C., Lopes, D., Bernardino, L., do Rosário, V. E., Varandas, L., et al. (2008). Prevalence of *pfmdr1*, *pfcrt, pfdhfr* and *pfdhps* mutations associated with drug resistance, in Luanda, Angola. *Mal. J.* 7:236.

Flegg, J. A., Metcalf, C. J. E., Gharbi, M., Venkatesan, M., Shewchuk, T., Hopkins Sibley, C., et al. (2013). Trends in antimalarial drug use in Africa. *Am. J. Trop. Med. Hyg.* 89, 857–865.

Fortes, F., Dimbu, R., Figueiredo, P., Neto, Z., do Rosário, V. E., and Lopes, D. (2011). Evaluation of prevalence's of *pfdhfr* and *pfdhps* mutations in Angola. *Malaria J.* 10:22. doi: 10.1186/1475-2875-10-22

Frosch, A. E. P., Venkatesan, M., and Laufer, M. K. (2011). Patterns of chloroquine use and resistance in sub-Saharan Africa: a systematic review of household survey and molecular data. *Mal. J.* 10:116. doi: 10.1186/1475-2875-10-116

Fuangchan, A., Dhippayom, T., and Kongkaew, C. (2014). Intervention to promote patients' adherence to antimalarial medication: a systematic review. *Am. J. Trop. Med. Hyg.* 90, 11–19. doi: 10.4269/ajtmh.12-0598

Galatas, B., Nhamussua, L., Candrinho, B., Mabote, L., Cisteró, P., Gupta, H., et al. (2017). In-Vivo efficacy of chloroquine to clear asymptomatic infections in mozambican adults: a randomized, placebo-controlled trial with implications for elimination strategies. *Sci. Rep.* 7, 1356–1356.

Gatton, M. L., Martin, L. B., and Cheng, Q. (2004). Evolution of resistance to sulfadoxine-pyrimethamine in *Plasmodium falciparum*. *Antimicrob Agents Chemother.* 48, 2116–2123. doi: 10.1128/aac.48.6.2116-2123.2004

Gerstl, S., Namagana, A., Palacios, L., Mweshi, F., Aprile, S., and Lima, A. (2015). High adherence to malaria treatment: promising results of an adherence study in South Kivu, Democratic Republic of the Congo. *Mal. J.* 14:414.

Gesase, S., Gosling, R. D., Hashim, R., Ord, R., Naidoo, I., Madebe, R., et al. (2009). High resistance of *Plasmodium falciparum* to sulphadoxine/pyrimethamine in northern Tanzania and the emergence of dhps resistance mutation at Codon 581. *PLoS One* 4:e4569. doi: 10.1371/journal.pone.0004569

Grais, R. F. I., Laminou, M., Woi-Messe, L., Makarimi, R., Bouriema, S. H., Langendorf, C., et al. (2018). Molecular markers of resistance to amodiaquine plus sulfadoxine–pyrimethamine in an area with seasonal malaria chemoprevention in south central Niger. *Mal. J.* 17:98.

Greenwood, B. M., Bojang, K., Whitty, C. J., and Targett, G. A. (2005). Malaria. *Lancet* 365, 1487–1498.

Griffin, C. E., Hoke, J. M., Samarakoon, U., Duan, J., Mu, J., Ferdig, M. T., et al. (2012). Mutation in the *Plasmodium falciparum* CRT protein determines the stereospecific activity of antimalarial cinchona alkaloids. *Antimicrob Agents Chemother.* 56, 5356–5364. doi: 10.1128/aac.05667-11

Griffin, J. T., Cairns, M., Ghani, A. C., Roper, C., Schellenberg, D., Carneiro, I., et al. (2010). Protective efficacy of intermittent preventive treatment of malaria in infants (IPTi) using sulfadoxine-pyrimethamine and parasite resistance. *PLoS One* 5:e12618. doi: 10.1371/journal.pone.0012618

Grobusch, M. P., and Kremsner, P. G. (2005). Uncomplicated malaria. *Curr. Top. Microbiol. Immunol.* 295, 83–104.

Hailemeskel, E., Kassa, M., Taddesse, G., Mohammed, H., Woyessa, A., Tasew, G., et al. (2013). Prevalence of sulfadoxine-pyrimethamine resistance-associated mutations in *dhfr* and *dhps* genes of *Plasmodium falciparum* three years after SP withdrawal in Bahir Dar, Northwest Ethiopia. *Acta Trop.* 128, 636–641. doi: 10.1016/j.actatropica.2013.09.010

Hanboonkunupakarn, B., and White, N. J. (2016). The threat of antimalarial drug resistance. *Trop. Dis. Travel Med. Vacc.* 2:10.

Happi, C. T., Gbotosho, G. O., Folarin, O. A., Akinboye, D. O., Yusuf, B. O., Ebong, O. O., et al. (2005). Polymorphisms in *Plasmodium falciparum dhfr* and *dhps* genes and age related in vivo sulfadoxine–pyrimethamine resistance in malaria-infected patients from Nigeria. *Acta Trop.* 95, 183–193. doi: 10.1016/j.actatropica.2005.06.015

Harinasuta, T., Suntharasamai, P., and Viravan, C. (1965). Chloroquine-resistant *falciparum* malaria in Thailand. *Lancet* 2, 657–660. doi: 10.1016/s0140-6736(65)90395-8

Heinberg, A., and Kirkman, L. (2015). The molecular basis of antifolate resistance in *Plasmodium falciparum*: looking beyond point mutations. *Annal. N. Y. Acad. Sci.* 1342, 10–18. doi: 10.1111/nyas.12662

Homewood, C. A., Warhurst, D. C., Peters, W., and Baggaley, V. C. (1972). Lysosomes, pH and the Anti-malarial action of chloroquine. *Nature* 235, 50–52. doi: 10.1038/235050a0

Hyde, J. E. (2009). Exploring the folate pathway in *Plasmodium falciparum*. *Acta Trop.* 94, 191–206. doi: 10.1016/j.actatropica.2005.04.002

Hyde, J. E. (2007). Drug-resistant malaria - an insight. *FEBS J.* 274, 4688–4698. doi: 10.1111/j.1742-4658.2007.05999.x

Ibraheem, Z. O., Abd Majid, R., Noor, S. M., Sedik, H. M., and Basir, R. (2014). Role of different *pfcrt* and *pfmdr-1* mutations in conferring resistance to antimalaria drugs in *Plasmodium falciparum*. *Mal. Res. Treat.* 2014:950424.

Idowu, O. A., Apalara, S. B., and Lasisi, A. A. (2006). Assessment of quality of chloroquine tablets sold by drug vendors in Abeokuta, Nigeria. *Tanzan Health Res. Bull.* 8, 45–46.

Ikegbunam, M. N., Nkonganyi, C. N., Thomas, B. N., Esimone, C. O., Velavan, T. P., and Ojurongbe, O. (2019). Analysis of *Plasmodium falciparumpfcrt* and *pfmdr1* genes in parasite isolates from asymptomatic individuals in Southeast Nigeria 11 years after withdrawal of chloroquine. Malaria J. 18:343. doi: 10.1186/s12936-019-2977-6

Iriemenam, N. C., Shah, M., Gatei, W., van Eijk, A. M., Ayisi, J., Kariuki, S., et al. (2012). Temporal trends of sulphadoxine-pyrimethamine (SP) drug-resistance molecular markers in *Plasmodium falciparum* parasites from pregnant women in western Kenya. *Mal. J.* 11:134. doi: 10.1186/1475-2875-11-134

Jiang, T., Chen, J., Fu, H., Wu, K., Yao, Y., Eyi, J. U. M., et al. (2019). High prevalence of *pfdhfr-pfdhps* quadruple mutations associated with sulfadoxine-pyrimethamine resistance in *Plasmodium falciparum* isolates from Bioko Island, Equatorial Guinea. *Malaria J.* 18:101. doi: 10.1186/s12936-019-2734-x

Kajubi, R., Ochieng, T., Kakuru, A., Jagannathan, P., Nakalembe, M., Ruel, T., et al. (2019). Monthly sulfadoxine-pyrimethamine versus dihydroartemisinin-piperaquine for intermittent preventive treatment of malaria in pregnancy: a double-blind, randomised, controlled, superiority trial. *Lancet* 393, 1428–1439. doi: 10.1016/s0140-6736(18)32224-4

Kalanda, G. C., Hill, J., Verhoeff, F. H., and Brabin, B. J. (2006). Comparative efficacy of chloroquine and sulphadoxine–pyrimethamine in pregnant women and children: a meta-analysis. *Trop Med. Int. Health* 11, 569–577. doi: 10.1111/j.1365-3156.2006.01608.x

Kayentao, K., Garner, P., Maria van Eijk, A., Naidoo, I., Roper, C., Mulokozi, A., et al. (2013). Intermittent preventive therapy for malaria during pregnancy using 2 vs 3 or more doses of sulfadoxine-pyrimethamine and risk of low birth weight in africa: systematic review and meta-analysis. *JAMA* 309, 594–604. doi: 10.1001/jama.2012.216231

Knight, S. E., Anyachebelu, E. J., Geddes, R., and Maharaj, R. (2009). Impact of delayed introduction of sulfadoxine–pyrimethamine and arthemeter–lumefantrine on malaria epidemiology in KwaZulu-Natal, South Africa. *Trop Med. Int. Health* 14, 1086–1092. doi: 10.1111/j.1365-3156.2009.02333.x

Kublin, J. G., Cortese, J. F., Njunju, E. M., Mukadam, R. A., Wirima, J. J., Kazembe, P. N., et al. (2003). Reemergence of chloroquine-sensitive *Plasmodium falciparum* malaria after cessation of chloroquine use in Malawi. *J. Infect. Dis.* 187, 1870–1875. doi: 10.1086/375419

Kümpornsin, K., Modchang, C., Heinberg, A., Ekland, E. H., Jirawatcharadech, P., Chobson, P., et al. (2014). Origin of robustness in generating drug-resistant malaria parasites. *Mol. Biol. Evol.* 31, 1649–1660. doi: 10.1093/molbev/msu140

Lawrenson, A., Cooper, D., O'Neill, P., and Berry, N. (2018). Study of the antimalarial activity of 4-aminoquinoline compounds against chloroquine-sensitive and chloroquine-resistant parasite strains. *J. Mol. Model.* 24:237. doi: 10.1007/s00894-018-3755-z

Laxminarayan, R., Over, M., and Smith, D. L. (2006). Will a global subsidy of new antimalarials delay the emergence of resistance and save lives? *Health Aff. (Millwood)* 25, 325–336. doi: 10.1377/hlthaff.25.2.325

Lehane, A. M., and Kirk, K. (2008). Chloroquine resistance-conferring mutations in pfcrt give rise to a chloroquine-associated H+ leak from the malaria parasite's digestive vacuole. *Antimicrob Agents Chemothe.r* 52, 4374–4380. doi: 10.1128/aac.00666-08

Lehane, A. M., McDevitt, C. A., Kirk, K., and Fidock, D. A. (2012). Degrees of chloroquine resistance in *Plasmodium* - is the redox system involved? *Int. J. Parasitol. Drugs Drug Resist.* 2, 47–57. doi: 10.1016/j.ijpddr.2011.11.001

Loria, P., Miller, S., Foley, M., and Tilley, L. (1999). Inhibition of the peroxidative degradation of haem as the basis of action of chloroquine and other quinoline antimalarials. *Biochem. J.* 339(Pt 2), 363–370. doi: 10.1042/0264-6021:3390363

Lu, F., Culleton, R., Zhang, M., Ramaprasad, A., von Seidlein, L., Zhou, H., et al. (2017b). Emergence of indigenous artemisinin-resistant *Plasmodium falciparum* in Africa. *N. Engl. J. Med.* 376, 991–993.

Lu, F., Zhang, M., Culleton, R. L., Xu, S., Tang, J., Zhou, H., et al. (2017a). Return of chloroquine sensitivity to Africa? Surveillance of African *Plasmodium falciparum* chloroquine resistance through malaria imported to China. *Paras. Vect.* 10, 355.

Lusingu, J. P., and Von Seidlein, L. (2008). Challenges in malaria control in sub-Saharan Africa: the vaccine perspective. *Tanzan J. Health Res.* 10, 253–266.

Macintyre, K., Littrell, M., Keating, J., Hamainza, B., Miller, J., and Eisele, T. P. (2011). Determinants of hanging and use of ITNs in the context of near universal coverage in Zambia. *Health Policy Plann.* 27, 316–325. doi: 10.1093/heapol/czr042

Maharaj, R., Raman, J., Morris, N., Moonasar, D., Durrheim, D. N., Seocharan, I., et al. (2013). Epidemiology of malaria in South Africa: From control to elimination. *SAMJ South Afr. Med. J.* 103, 779–783. doi: 10.7196/samj.7441

Maiga, H., Djimde, A. A., Beavogui, A. H., Toure, O., Tekete, M., Sangare, C. P. O., et al. (2015). Efficacy of sulphadoxine-pyrimethamine + artesunate, sulphadoxine-pyrimethamine + amodiaquine, and sulphadoxine-pyrimethamine alone in uncomplicated *falciparum* malaria in Mali. *Mal. J.* 14:64. doi: 10.1186/s12936-015-0557-y

Maiga, O. M., Kayentao, K., Traoré, B. T., Djimde, A., Traoré, B., Diallo, M., et al. (2011). Superiority of 3 Over 2 doses of intermittent preventive treatment with sulfadoxine-pyrimethamine for the prevention of malaria during pregnancy in mali: a randomized controlled trial. *Clin. Infect. Dis.* 53, 215–223. doi: 10.1093/cid/cir374

Maitland, K. (2016). Severe malaria in african children – the need for continuing investment. *N. Engl. J. Med.* 375, 2416–2417. doi: 10.1056/nejmp1613528

Makono, R., and Sibanda, S. (1999). Review of the prevalence of malaria in Zimbabwe with specific reference to parasite drug resistance (1984-96). *Trans. R. Soc. Trop. Med. Hyg.* 93, 449–452. doi: 10.1016/s0035-9203(99)90331-0

Malisa, A. L., Pearce, R. J., Abdulla, S., Mshinda, H., Kachur, P. S., Bloland, P., et al. (2010). Drug coverage in treatment of malaria and the consequences for resistance evolution – evidence from the use of sulphadoxine/pyrimethamine. *Mal. J.* 9:190.

Marks, F., Evans, J., Meyer, C. G., Browne, E. N., Flessner, C., von Kalckreuth, V., et al. (2005). High prevalence of markers for sulfadoxine and pyrimethamine resistance in *Plasmodium falciparum* in the absence of drug pressure in the Ashanti region of Ghana. *Antimicrob Agents Chemother.* 49, 1101–1105. doi: 10.1128/aac.49.3.1101-1105.2005

Martin, M. K., Venantius, K. B., Patricia, N., Bernard, K., Keith, B., Allen, K., et al. (2020). Correlates of uptake of optimal doses of sulfadoxine-pyrimethamine for prevention of malaria during pregnancy in East-Central Uganda. *Mal. J.* 19:153.

Mayor, A., Serra-Casas, E., Sanz, S., Aponte, J. J., Macete, E., Mandomando, I., et al. (2008). Molecular Markers of resistance to sulfadoxine-pyrimethamine during intermittent preventive treatment for malaria in mozambican infants. *J. Infect. Dis.* 197, 1737–1742. doi: 10.1086/588144

Mbugi, E. V., Mutayoba, B. M., Malisa, A. L., Balthazary, S. T., Nyambo, T. B., and Mshinda, H. (2006). Drug resistance to sulphadoxine-pyrimethamine in *Plasmodium falciparum* malaria in Mlimba, Tanzania. *Mal. J.* 5:94.

Medicines for Malaria Venture (MMV) (2020). Available online at: www.mmv.org (accessed October 27, 2020).

Meshnick, S. R. (1998). Artemisinin antimalarials: mechanisms of action and resistance. *Med. Trop. (Mars)* 58(Suppl. 3), 13–17.

Mharakurwa, S., and Mugochi, T. (1994). Chloroquine-resistant *falciparum* malaria in an area of rising endemicity in Zimbabwe. *J. Trop. Med. Hyg.* 97, 39–45.

Mlambo, G., Sullivan, D., Mutambu, S. L., Soko, W., Mbedzi, J., Chivenga, J., et al. (2007). High prevalence of molecular markers for resistance to chloroquine and pyrimethamine in *Plasmodium falciparum* from Zimbabwe. *Parasitol. Res.* 101, 1147–1151. doi: 10.1007/s00436-007-0597-5

Mubyazi, G. M., and Gonzalez-Block, M. A. (2005). Research influence on antimalarial drug policy change in Tanzania: case study of replacing chloroquine with sulfadoxine-pyrimethamine as the first-line drug. *Mal. J.* 4, 51–51.

Mwanza, S., Joshi, S., Nambozi, M., Chileshe, J., Malunga, P., Kabuya, J-B., et al. (2016). The return of chloroquine-susceptible *Plasmodium falciparum* malaria in Zambia. *Malaria J.* 15, 584–584. doi: 10.1186/s12936-016-1637-3

Mwendera, C., de Jager, C., Longwe, H., Phiri, K., Hongoro, C., and Mutero, C. M. (2016). Malaria research and its influence on anti-malarial drug policy in Malawi: a case study. *Health Res. Policy Syst.* 14:41. doi: 10.1186/s12961-016-0108-1

Naidoo, I., and Roper, C. (2010). Following the path of most resistance: *dhps* K540E dispersal in African *Plasmodium falciparum*. *Trends Parasitol.* 26, 447–456. doi: 10.1016/j.pt.2010.05.001

Naidoo, I., and Roper, C. (2011). Drug resistance maps to guide intermittent preventive treatment of malaria in African infants. *Parasitology* 138, 1469–1479. doi: 10.1017/s0031182011000746

Najera, J. A., and Zaim, M. (2001). *Malaria Vector Control – Insecticides for Indoor Residual Spraying*. Geneva: World Health Organization Communicable Disease Control, Prevention and Eradication.

Ndeffo Mbah, M. L., Parikh, S., and Galvani, A. P. (2015). Comparing the impact of artemisinin-based combination therapies on malaria transmission in sub-Saharan Africa. *Am. J. Trop. Med. Hyg.* 92, 555–560. doi: 10.4269/ajtmh.14-0490

Ndiaye, M., Faye, B., Tine, R., Ndiaye, J. L., Lo, A., Abiola, A., et al. (2012). Assessment of the molecular marker of Plasmodium falciparum chloroquine resistance (Pfcrt) in Senegal after several years of chloroquine withdrawal. *Am. J. Trop. Med. Hyg.* 87, 640–645. doi: 10.4269/ajtmh.2012.11-0709

Njokah, M. J., Kang'ethe, J. N., Kinyua, J., Kariuki, D., and Kimani, F. T. (2016). In vitro selection of *Plasmodium falciparum pfcrt* and *pfmdr1* variants by artemisinin. *Malar. J.* 15:381.

Nuwaha, F. (2001). The challenge of chloroquine-resistant malaria in sub-Saharan Africa. *Health Policy Plann.* 16, 1–12. doi: 10.1093/heapol/16.1.1

Nzila, A. (2006). The past, present and future of antifolates in the treatment of *Plasmodium falciparum* infection. *J. Antimicrob. Chemother.* 57, 1043–1054. doi: 10.1093/jac/dkl104

Nzila, A. M., Mberu, E. K., Sulo, J., Dayo, H., Winstanley, P. A., Sibley, C. H., et al. (2000). Towards an understanding of the mechanism of pyrimethamine-sulfadoxine resistance in *Plasmodium falciparum*: genotyping of dihydrofolate reductase and dihydropteroate synthase of Kenyan parasites. *Antimicrob. Agents Chemother.* 44, 991–996. doi: 10.1128/aac.44.4.991-996.2000

Ogouyèmi-Hounto, A., Ndam, N. T., Kinde Gazard, D., d'Almeida, S., Koussihoude, L., Ollo, E., et al. (2013). Prevalence of the molecular marker of *Plasmodium falciparum* resistance to chloroquine and sulphadoxine/pyrimethamine in Benin seven years after the change of malaria treatment policy. *Mal. J.* 12:147.

Okumu, F. O., and Moore, S. J. (2011). Combining indoor residual spraying and insecticide-treated nets for malaria control in Africa: a review of possible outcomes and an outline of suggestions for the future. *Malar. J.* 10:208.

Oladipo, O. O., Wellington, O. A., and Sutherland, C. J. (2015). Persistence of chloroquine-resistant haplotypes of *Plasmodium falciparum* in children with uncomplicated Malaria in Lagos, Nigeria, four years after change of chloroquine as first-line antimalarial medicine. *Diagn. Pathol.* 10:41. doi: 10.1186/s13000-015-0276-2

Olapeju, B., Choiriyyah, I., Lynch, M., Acosta, A., Blaufuss, S., Filemyr, E., et al. (2018). Age and gender trends in insecticide-treated net use in sub-Saharan Africa: a multi-country analysis. *Mal. J.* 17:423.

Omar, S. A. I, Adagu, S., Gump, D. W., Ndaru, N. P., and Warhurst, D. C. (2001). *Plasmodium falciparum* in Kenya: high prevalence of drug-resistance-associated polymorphisms in hospital admissions with severe malaria in an epidemic area. *Annal. Trop. Med. Parasitol.* 95, 661–669. doi: 10.1080/00034980120103234

Osman, M. E., Mockenhaupt, F. P., Bienzle, U., Elbashir, M. I., and Giha, H. A. (2007). Field-based evidence for linkage of mutations associated with chloroquine (*pfcrt/pfmdr1*) and sulfadoxine-pyrimethamine (pfdhfr/pfdhps) resistance and for the fitness cost of multiple mutations in *P. falciparum*. *Infect. Genet. Evol.* 7, 52–59. doi: 10.1016/j.meegid.2006.03.008

Oyibo, W. A., and Agomo, C. O. (2011). Scaling up of intermittent preventive treatment of malaria in pregnancy using sulphadoxine-pyrimethamine: prospects and challenges. *Mat. Child Health J.* 15, 542–552. doi: 10.1007/s10995-010-0608-5

Parker, D., Carrara, V., Pukrittayakamee, S., McGready, R., and Nosten, F. (2015). Malaria ecology along the Thailand– Myanmar border. *Mal. J.* 14:921.

Pearce, R. J., Drakeley, C., Chandramohan, D., Mosha, F., and Roper, C. (2003). Molecular determination of point mutation haplotypes in the dihydrofolate reductase and dihydropteroate synthase of *Plasmodium falciparum* in three districts of northern Tanzania. *Antimicrob Agents Chemother.* 47, 1347–1354. doi: 10.1128/aac.47.4.1347-1354.2003

Pearce, R. J., Pota, H., Evehe, M. S., El-Hadj, B., Mombo-Ngoma, G., Malisa, A. L., et al. (2009). Multiple origins and regional dispersal of resistant dhps in African *Plasmodium falciparum* malaria. *PLoS Med.* 6:e1000055. doi: 10.1371/journal.pmed.1000055

Peters, P. J., Thigpen, M. C., Parise, M. E., and Newman, R. D. (2007). Safety and toxicity of sulfadoxine/pyrimethamine: implications for malaria prevention in pregnancy using intermittent preventive treatment. *Drug Saf.* 30, 481–501. doi: 10.2165/00002018-200730060-00003

Plowe, C. V. (2014). Resistance nailed. *Nature* 505, 30–31. doi: 10.1038/nature12845

Pluess, B., Tanser, F. C., Lengeler, C., and Sharp, B. L. (2010). Indoor residual spraying for preventing malaria. *Cochr. Database Syst. Rev.* 2010:CD006657.

Premji, Z. G. (2009). Coartem® : the journey to the clinic. *Mal. J.* 8:S3.

Pulcini, S., Staines, H. M., Lee, A. H., Shafik, S. H., Bouyer, G., Moore, C. M., et al. (2015). Mutations in the *Plasmodium falciparum* chloroquine resistance transporter, PfCRT, enlarge the parasite's food vacuole and alter drug sensitivities. *Sci. Rep.* 5:14552. doi: 10.1038/srep14552

Raman, J., Kagoro, F. M., Mabuza, A., Malatje, G., Reid, A., Frean, J., and Barnes, K. I. (2019). Absence of *kelch13* artemisinin resistance markers but strong selection for lumefantrine-tolerance molecular markers following 18 years of artemisinin-based combination therapy use in Mpumalanga Province, South Africa (2001–2018). *Malaria J.* 18:280. doi: 10.1186/s12936-019-2911-y

Raphemot, R., Posfai, D., and Derbyshire, E. R. (2016). Current therapies and future possibilities for drug development against liver-stage malaria. *J. Clin. Investigat.* 126, 2013–2020. doi: 10.1172/jci82981

Rapuoda, B., Ali, A., Bakyaita, N., Mutabingwa, T., Mwita, A., Rwagacondo, C., et al. (2003). The efficacy of antimalarial monotherapies, sulfadoxine-pyrimethamine and amodiaquine in East Africa: implication for sub-regional policy. The East African Network for Monitoring Antimalarial Treatment (EANMAT). *Trop. Med. Int. Health* 8, 860–867. doi: 10.1046/j.1360-2276.2003.01114.x

Recker, M., Bull, P. C., and Buckee, C. O. (2018). Recent advances in the molecular epidemiology of clinical malaria. *F1000Research* 7, F1000FacultyRev-1159.

Reiling, S. J., Krohne, G., Friedrich, O., Geary, T. G., and Rohrbach, P. (2018). Chloroquine exposure triggers distinct cellular responses in sensitive versus resistant Plasmodium falciparum parasites. *Sci. Rep.* 8:11137.

Roepe, P. D. (2011). PfCRT-mediated drug transport in malarial parasites. *Biochemistry* 50, 163–171. doi: 10.1021/bi101638n

Sandefur, C. I., Wooden, J. M. I, Quaye, K., Sirawaraporn, W., and Sibley, C. H. (2007). Pyrimethamine-resistant dihydrofolate reductase enzymes of *Plasmodium falciparum* are not enzymatically compromised in vitro. *Mol. Biochem. Parasitol.* 154, 1–5. doi: 10.1016/j.molbiopara.2007.03.009

Scates, S. S., Finn, T. P., Wisniewski, J., Dadi, D., Mandike, R., Khamis, M., et al. (2020). Costs of insecticide-treated bed net distribution systems in sub-Saharan Africa. *Mal. J.* 19, 105–105.

Schleicher, T. R., Yang, J., Freudzon, M., Rembisz, A., Craft, S., Hamilton, M., et al. (2018). A mosquito salivary gland protein partially inhibits *Plasmodium* sporozoite cell traversal and transmission. *Nat. Commun.* 9:2908.

Schönfeld, M., Barreto Miranda, I., Schunk, M., Maduhu, I., Maboko, L., Hoelscher, M., et al. (2007). Molecular surveillance of drug-resistance associated mutations of *Plasmodium falciparum* in south-west Tanzania. *Mal. J.* 6:2.

Shafik, S. H., Cobbold, S. A., Barkat, K., Richards, S. N., Lancaster, N. S., Llinás, M., et al. (2020). The natural function of the malaria parasite's chloroquine resistance transporter. *Nat. Commun.* 11:3922. doi: 10.1038/s41467-020-17781-6

Sharma, J., Khan, S., Dutta, P., and Mahanta, J. (2015). Molecular determination of antifolate resistance associated point mutations in *Plasmodium falciparum* dihydrofolate reductase (dhfr) and dihydropteroate synthetase (dhps) genes among the field samples in Arunachal Pradesh. *J. Vector Borne Dis.* 52, 116–121.

Shretta, R., Omumbo, J., Rapuoda, B., and Snow, R. W. (2000). Using evidence to change antimalarial drug policy in Kenya. *Trop. Med. Int. Health* 5, 755–764. doi: 10.1046/j.1365-3156.2000.00643.x

Siddiqui, F. A., Boonhok, R., Cabrera, M., Mbenda, H. G. N., Wang, M., Min, H., et al. (2020). Role of *Plasmodium falciparum* Kelch 13 protein mutations in *P. falciparum* populations from northeastern myanmar in mediating artemisinin resistance. *mBio* 11:e001134-19.

Slater, H. C., Griffin, J. T., Ghani, A. C., and Okell, L. C. (2016). Assessing the potential impact of artemisinin and partner drug resistance in sub-Saharan Africa. *Mal. J.* 15, 10–10.

Smith Gueye, C., Newby, G., Gosling, R. D., Whittaker, M. A., Chandramohan, D., Slutsker, L., et al. (2016). Strategies and approaches to vector control in nine malaria-eliminating countries: a cross-case study analysis. *Mal. J.* 15:2.

Sridaran, S., McClintock, S. K., Syphard, L. M., Herman, K. M., Barnwell, J. W., and Udhayakumar, V. (2010). Anti-folate drug resistance in Africa: meta-analysis of reported dihydrofolate reductase (*dhfr*) and dihydropteroate synthase (*dhps*) mutant genotype frequencies in African Plasmodium falciparum parasite populations. *Mal. J.* 9:247. doi: 10.1186/1475-2875-9-247

Steinhardt, L. C., Jean, Y. S., Impoinvil, D., Mace, K. E., Wiegand, R., Huber, C. S., et al. (2017). Effectiveness of insecticide-treated bednets in malaria prevention in Haiti: a case-control study. *Lancet Global Health* 5, e96–e103.

Tagbor, H., Bruce, J., Browne, E., Randal, A., Greenwood, B., and Chandramohan, D. (2006). Efficacy, safety, and tolerability of amodiaquine plus sulphadoxine-pyrimethamine used alone or in combination for malaria treatment in pregnancy: a randomised trial. *Lancet* 368, 1349–1356. doi: 10.1016/s0140-6736(06)69559-7

Takala-Harrison, S., and Laufer, M. K. (2015). Antimalarial drug resistance in Africa: key lessons for the future. *Annal. N. Y. Acad. Sci.* 1342, 62–67. doi: 10.1111/nyas.12766

Takechi, M., Matsuo, M., Ziba, C., Macheso, A., Butao, D. I, Zungu, L., et al. (2001). Therapeutic efficacy of sulphadoxine/pyrimethamine and susceptibility in vitro of P. falciparum isolates to sulphadoxine-pyremethamine and other antimalarial drugs in Malawian children. *Trop. Med. Int. Health* 6, 429–434. doi: 10.1046/j.1365-3156.2001.00735.x

Taylor, C., Florey, L., and Ye, Y. (2017). Equity trends in ownership of insecticide-treated nets in 19 sub-Saharan African countries. *Bull. World Health Organ.* 95, 322–332. doi: 10.2471/blt.16.172924

Tekete, M. M., Toure, S., Fredericks, A., Beavogui, A. H., Sangare, C. P. O., Evans, A., et al. (2011). Effects of amodiaquine and artesunate on sulphadoxine-pyrimethamine pharmacokinetic parameters in children under five in Mali. *Mal. J.* 10:275. doi: 10.1186/1475-2875-10-275

ter Kuile, F. O., van Eijk, A. M., and Filler, S. J. (2015). Effect of sulfadoxine-pyrimethamine resistance on the efficacy of intermittent preventive therapy for malaria control during pregnancy: a systematic review. *JAMA* 297, 2603–2616. doi: 10.1001/jama.297.23.2603

Terlouw, D. J., Nahlen, B. L., Courval, J. M., Kariuki, S. K., Rosenberg, O. S., Oloo, A. J., et al. (2003). Sulfadoxine-pyrimethamine in treatment of malaria in Western Kenya: increasing resistance and underdosing. *Antimicrob Agents Chemother.* 47, 2929–2932. doi: 10.1128/aac.47.9.2929-2932.2003

Thomsen, T. T., Madsen, L. B., Hansson, H. H., Tomás, E. V., Charlwood, D. I, Bygbjerg, C., et al. (2013). Rapid selection of *Plasmodium falciparum* chloroquine resistance transporter gene and multidrug resistance gene-1 haplotypes associated with past chloroquine and present artemether-lumefantrine use in Inhambane District, southern Mozambique. *Am. J. Trop. Med. Hyg.* 88, 536–541. doi: 10.4269/ajtmh.12-0525

Thu, A. M., Phyo, A. P., Landier, J., Parker, D. M., and Nosten, F. H. (2017). Combating multidrug-resistant *Plasmodium falciparum* malaria. *FEBS J.* 284, 2569–2578.

Tizifa, T. A., Kabaghe, A. N., McCann, R. S., van den Berg, H., Van Vugt, M., and Phiri, K. S. (2018). Prevention efforts for Malaria. *Curr. Trop. Med. Rep.* 5, 41–50.

Trape, J. F. (2001). The public health impact of chloroquine resistance in Africa. *Am. J. Trop. Med. Hyg.* 64, 12–17. doi: 10.4269/ajtmh.2001.64.12

Trape, J. F., Pison, G., Preziosi, M. P., Enel, C., Desgrées du Loû, A., Delaunay, V., et al. (1998). Impact of chloroquine resistance on malaria mortality. *C R Acad. Sci. III* 321, 689–697.

Ukpe, I. S., Moonasar, D., Raman, J., Barnes, K. I., Baker, L., and Blumberg, L. (2013). Case management of malaria: Treatment and chemoprophylaxis. *S. Afr. Med. J.* 103(10 Pt 2), 793–798. doi: 10.7196/samj.7443

Veiga, M. I., Dhingra, S. K., Henrich, P. P., Straimer, J., Gnädig, N., Uhlemann, A-C., et al. (2016). Globally prevalent *pfmdr1* mutations modulate *Plasmodium falciparum* susceptibility to artemisinin-based combination therapies. *Nat. Commun.* 7:11553. doi: 10.1038/ncomms11553

Viana, G. M. R., Machado, R. L. D., Calvosa, V. S. P., and Póvoa, M. M. (2006). Mutations in the *pfmdr1, cg2*, and *pfcrt* genes in *Plasmodium falciparum* samples from endemic malaria areas in Rondonia and Pará State, Brazilian Amazon Region. *Cad Saude Publica* 22, 2703–2711. doi: 10.1590/s0102-311x2006001200019

von Seidlein, L., and Dondorp, A. (2015). Fighting fire with fire: mass antimalarial drug administrations in an era of antimalarial resistance. *Exp. Rev. Anti. Infect. Ther.* 13, 715–730. doi: 10.1586/14787210.2015.1031744

Walker, P. G. T., Floyd, J., ter Kuile, F., and Cairns, M. (2017). Estimated impact on birth weight of scaling up intermittent preventive treatment of malaria in pregnancy given sulphadoxine-pyrimethamine resistance in Africa: a mathematical model. *PLoS Med.* 14:e1002243. doi: 10.1371/journal.pmed.1002243

Wernsdorfer, W. H. (1994). Epidemiology of drug resistance in malaria. *Acta Tropica* 56, 143–156. doi: 10.1016/0001-706x(94)90060-4

White, N. J., Pongtavornpinyo, W., Maude, R. J., Saralamba, S., Aguas, R., Stepniewska, K., et al. (2009). Hyperparasitaemia and low dosing are an important source of anti-malarial drug resistance. *Mal. J.* 8:253.

Wicht, K. J., Mok, S., and Fidock, D. A. (2020). Molecular Mechanisms of drug resistance in *Plasmodium falciparum* malaria. *Annu. Rev. Microbiol.* 74, 431–454. doi: 10.1146/annurev-micro-020518-115546

Winstanley, P., Ward, S., Snow, R., and Breckenridge, A. (2004). Therapy of falciparum malaria in sub-saharan Africa: from molecule to policy. *Clin. Microbiol. Rev.* 17, 612–637. doi: 10.1128/cmr.17.3.612-637.2004

World Health Organisation (WHO) (2008). *World Malaria Report 2007*. Geneva: World Health Organization.

World Health Organisation (WHO) (2003). *World Malaria Report 2002*. Geneva: World Health Organization.

World Health Organisation (WHO) (2010). *WHO Technical Report Series 957*. Geneva: World Health Organization.

World Health Organisation (WHO) (2011). *World Malaria Report 2010*. Geneva: World Health Organization.

World Health Organisation (WHO) (2017). *World Malaria Report 2016*. Geneva: World Health Organization.

World Health Organisation (WHO) (2019). *World Malaria Report 2018*. Geneva: World Health Organization.

World Health Organisation (WHO) (2020). *World Malaria Report 2018*. Geneva: World Health Organization.

Xia, Z.-G., Wang, R.-B., Wang, D.-Q., Feng, J., Zheng, Q., Deng, C.-S., et al. (2014). China-Africa cooperation initiatives in malaria control and elimination. *Adv. Parasitol.* 86, 319–337. doi: 10.1016/b978-0-12-800869-0.00012-3

Xie, S. C., Ralph, S. A., and Tilley, L. (2020). K13, the cytostome, and artemisinin resistance. *Trends Parasitol.* 36, 533–544. doi: 10.1016/j.pt.2020.03.006

Yadav, S., and Rawal, G. (2016). The menace due to fake antimalarial drugs. *Int. J. Pharmaceut. Chem. Anal.* 3, 53–55. doi: 10.5958/2394-2797.2016.00007.1

Yeka, A., Gasasira, A., Mpimbaza, A., Achan, J., Nankabirwa, J., Nsobya, S., et al. (2012). Malaria in Uganda: challenges to control on the long road to elimination: I. Epidemiology and current control efforts. *Acta Trop.* 121, 184–195. doi: 10.1016/j.actatropica.2011.03.004

Young, M. D., and Moore, D. V. (1961). Chloroquine resistance in *Plasmodium falciparum*. *Am. J. Trop. Med. Hyg.* 10, 317–320.

Zareen, S., Ur Rehman, H., Gul, N., Ur Rehman, M., Bibi, S., Bakht, A., et al. (2016). Malaria is still a life-threatening disease review. *J. Entomol. Zool. Stud. JEZS* 105, 105–112.

Zhao, L., Pi, L., Qin, Y., Lu, Y., Zeng, W., Xiang, Z., et al. (2020). Widespread resistance mutations to sulfadoxine-pyrimethamine in malaria parasites imported to China from Central and Western Africa. *Int. J. Parasitol. Drugs Drug Resist.* 12, 1–6. doi: 10.1016/j.ijpddr.2019.11.002

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 July 2021
doi: 10.3389/fgene.2021.684127

Antigenic Diversity in *Theileria parva* Populations From Sympatric Cattle and African Buffalo Analyzed Using Long Read Sequencing

*Fiona K. Allan[1†], Siddharth Jayaraman[1†], Edith Paxton[1], Emmanuel Sindoya[2], Tito Kibona[3], Robert Fyumagwa[4], Furaha Mramba[5], Stephen J. Torr[6], Johanneke D. Hemmink[1,7], Philip Toye[7], Tiziana Lembo[8], Ian Handel[1], Harriet K. Auty[8], W. Ivan Morrison[1] and Liam J. Morrison[1]**

[1] Royal (Dick) School of Veterinary Studies, Roslin Institute, University of Edinburgh, Edinburgh, United Kingdom, [2] Ministry of Livestock and Fisheries, Serengeti District Livestock Office, Mugumu, Tanzania, [3] Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania, [4] Tanzania Wildlife Research Institute, Arusha, Tanzania, [5] Vector and Vector-Borne Diseases Research Institute, Tanga, Tanzania, [6] Liverpool School of Tropical Medicine, Liverpool, United Kingdom, [7] International Livestock Research Institute, Nairobi, Kenya, [8] Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom

OPEN ACCESS

Edited by:
Matthew Adekunle Adeleke, University of KwaZulu-Natal, South Africa

Reviewed by:
Stefano D'Amelio, Sapienza University of Rome, Italy
Jun-Hu Chen, National Institute of Parasitic Diseases, China

***Correspondence:**
Liam J. Morrison
Liam.Morrison@roslin.ed.ac.uk

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: *22 March 2021*
Accepted: *24 May 2021*
Published: *15 July 2021*

Citation:
Allan FK, Jayaraman S, Paxton E, Sindoya E, Kibona T, Fyumagwa R, Mramba F, Torr SJ, Hemmink JD, Toye P, Lembo T, Handel I, Auty HK, Morrison WI and Morrison LJ (2021) Antigenic Diversity in Theileria parva Populations From Sympatric Cattle and African Buffalo Analyzed Using Long Read Sequencing. Front. Genet. 12:684127. doi: 10.3389/fgene.2021.684127

East Coast fever (ECF) in cattle is caused by the Apicomplexan protozoan parasite *Theileria parva*, transmitted by the three-host tick *Rhipicephalus appendiculatus*. The African buffalo (*Syncerus caffer*) is the natural host for *T. parva* but does not suffer disease, whereas ECF is often fatal in cattle. The genetic relationship between *T. parva* populations circulating in cattle and buffalo is poorly understood, and has not been studied in sympatric buffalo and cattle. This study aimed to determine the genetic diversity of *T. parva* populations in cattle and buffalo, in an area where livestock co-exist with buffalo adjacent to the Serengeti National Park, Tanzania. Three *T. parva* antigens (Tp1, Tp4, and Tp16), known to be recognized by $CD8^+$ and $CD4^+$ T cells in immunized cattle, were used to characterize genetic diversity of *T. parva* in cattle (n = 126) and buffalo samples (n = 22). Long read (PacBio) sequencing was used to generate full or near-full length allelic sequences. Patterns of diversity were similar across all three antigens, with allelic diversity being significantly greater in buffalo-derived parasites compared to cattle-derived (e.g., for Tp1 median cattle allele count was 9, and 81.5 for buffalo), with very few alleles shared between species (8 of 651 alleles were shared for Tp1). Most alleles were unique to buffalo with a smaller proportion unique to cattle (412 buffalo unique vs. 231 cattle-unique for Tp1). There were indications of population substructuring, with one allelic cluster of Tp1 representing alleles found in both cattle and buffalo (including the TpM reference genome allele), and another containing predominantly only alleles deriving from buffalo. These data illustrate the complex interplay between *T. parva* populations in buffalo and cattle, revealing the significant genetic diversity in the buffalo *T. parva* population, the limited sharing of parasite genotypes between the host species, and highlight that a subpopulation of *T. parva* is maintained by transmission within cattle. The data indicate that fuller understanding of

buffalo *T. parva* population dynamics is needed, as only a comprehensive appreciation of the population genetics of *T. parva* populations will enable assessment of buffalo-derived infection risk in cattle, and how this may impact upon control measures such as vaccination.

Keywords: *Theileria parva*, East Coast fever, African buffalo (*Syncerus caffer)*, molecular epidemiology, cattle

INTRODUCTION

The protozoan parasite *Theileria parva*, transmitted by the three host tick *Rhipicephalus appendiculatus,* infects cattle and African buffalo (*Syncerus caffer*) across Eastern, Central and Southern Africa (McKeever and Morrison, 1990; Morrison and McKeever, 2006). *T. parva* causes East Coast fever (ECF) in cattle, an acute and often fatal disease, which is responsible for significant economic impact on the livestock industry, estimated at US $596 million annually (GALVmed, 2019). The African buffalo is considered the natural host for *T. parva*, but in contrast to cattle, where mortality can be as high as 90% (Brocklesby, 1961), infected buffalo do not suffer clinical disease. The parasite infects and multiplies within lymphocytes (Baldwin et al., 1988) causing an acute lymphoproliferative disease that can result in death within 3–4 weeks. A proportion of the intra-lymphocytic parasites (schizonts) differentiate to produce merozoites, which upon release, infect erythrocytes giving rise to the tick-infective piroplasm stage. In both host species, a low-level persistent infection, referred to as the carrier state, is established following recovery from the acute phase of infection, facilitating transmission by feeding ticks (Young et al., 1986). The transmission dynamics of *T. parva* between buffalo and cattle are complex (Morrison et al., 2020), with evidence that only a subset of parasites from buffalo are able to differentiate to piroplasms and establish tick-transmissible infections in cattle. However, the nature of this putative population substructuring within *T. parva* is currently unclear.

Cattle have only relatively recently arrived in Africa, having migrated in two broad waves, the first from the fertile crescent approximately 10,000 years ago that established African taurine (*Bos taurus*) breeds, and the second approximately 5,000 years ago involving *Bos indicus* cattle originating from Asia. European *B. taurus* breeds are a more recent introduction into Africa, only arriving in the last 150 years (Hanotte et al., 2002; Kim et al., 2017). This relatively short period of co-existence of cattle with *T. parva* has limited their ability to adapt to the parasite. European *B. taurus* breeds and improved *B. indicus* breeds are highly susceptible, suffering severe disease and high levels of mortality. However, some East African zebu (*B. indicus*) cattle residing in tick-infested areas are more tolerant of *T. parva* infections, with mortality usually <10% in the absence of control measures (Barnett, 1957; Paling et al., 1991). In contrast, *S. caffer* and *T. parva* have co-existed for millennia, allowing co-evolution that has resulted in continued susceptibility to infection but no apparent clinical disease.

Infection of cattle with buffalo-derived *T. parva* results in an acute, usually fatal disease clinically similar to classical ECF (referred to as corridor disease in southern Africa), but with lower levels of parasitized cells in peripheral lymphoid tissues and usually no detectable piroplasms (Young et al., 1977; Sitt et al., 2015). Importantly, such infections are usually not transmissible to ticks (Schreuder et al., 1977). In a few instances, experimental feeding of large numbers of ticks on recovered animals has resulted in transmission and passage of the parasites by ticks in cattle (Barnett and Brocklesby, 1966; Young and Purnell, 1973; Maritim et al., 1992), but it is unclear whether this enables the buffalo *T. parva* parasites to establish in cattle populations in the field.

Population analyses of *T. parva* have demonstrated that in both cattle and buffalo the parasite exists in freely mating (panmictic) populations. However, greater diversity in buffalo-derived *T. parva* has been consistently observed, whether by monoclonal antibody, RFLP, microsatellite or gene sequencing technologies, with analyses indicating that cattle-derived populations comprise a subset of the far greater diversity observed in buffalo (Conrad et al., 1989; Oura et al., 2011b; Hemmink et al., 2018). Additionally, microsatellite and gene sequence analyses of samples from animals naturally infected with *T. parva* have indicated that multiplicity of infection is the norm in both cattle and buffalo, but again buffalo samples having greater diversity, in some instances more than 20 alleles of individual satellite or gene loci detected in single animals (Oura et al., 2011b; Hemmink et al., 2018). This is consistent with the fact that development of immunity does not prevent establishment of infection following subsequent parasite challenge, allowing accumulation of parasite genotypes following repeated parasite challenge.

Experimental studies of immune responses of cattle to *T. parva* infection have provided evidence that immunity is mediated by CD8 T cells specific for parasitized lymphocytes (Morrison and McKeever, 1998). This immunity has been shown to be strain-specific, and the parasite strain-restriction is reflected by strain specificity of CD8 T cell responses (Taracha et al., 1995a,b). Use of parasite-specific CD8 T cell lines for antigen screening has resulted in the identification of a series of CD8 T cell target antigens (Tp1- Tp10) (Graham et al., 2006) and the epitopes within them (Graham et al., 2008) recognized by cells from immune animals. Given the evidence of the importance of CD8 T cell responses in immunity, these antigens have been investigated not only for vaccine development but also to examine antigenic diversity in field populations of *T. parva.*

A study of the sequences of the Tp1 and Tp2 antigen genes in a series of *T. parva*-infected cell lines isolated from cattle or buffalo in different locations in east Africa demonstrated that both of these antigens are highly polymorphic and, by comparing parasites isolated from cattle with those obtained from buffalo or cattle co-grazed with buffalo, the latter were found to contain

much greater sequence diversity (Pelle et al., 2011). A further study, utilizing high throughput sequencing of PCR amplicons of 6 *T. parva* antigen genes (Tp1, Tp2, Tp4, Tp5, Tp6, and Tp10), performed on blood samples from buffalo, showed that the genes varied in the extent of polymorphism but confirmed extensive sequence diversity and the presence of multiple alleles in individual animals (Hemmink et al., 2018).

The advent of genomic analysis has added further insight, with the initial sequencing of nine isolates indicating that the two genomes of buffalo-derived *T. parva* analyzed were divergent, with twice the number of single nucleotide polymorphisms (SNPs) compared to the *T. parva* reference genome (*T. parva* Muguga—originally isolated from a cow) compared to seven genomes of cattle-derived *T. parva* (Hayashida et al., 2013). This putative divergence has been recently supported by the first *de novo* genome assembly of a buffalo-derived *T. parva* isolate, which indicated a slightly larger genome, significantly high non-synonymous nucleotide diversity and genome-wide F_{ST} values compared to the *T. parva* Muguga genome, at levels compatible with those observed between species rather than within species (Palmateer et al., 2020).

The diversity and transmissibility (to cattle) of buffalo-derived *T. parva* is of obvious relevance with respect to immunity, whether through natural infected tick exposure or via the only currently available vaccine, the "Infection and Treatment Method" (ITM), which involves administration of live *T. parva* (three parasite isolates, including *T. parva* Muguga) and simultaneous treatment with oxytetracycline (Radley et al., 1975a,b,c). This vaccination provides immunity against challenge with cattle *T. parva* (Burridge et al., 1972; Morzaria et al., 1987) which is boosted by challenge. The vaccine has been used successfully to vaccinate cattle in the field, including some areas where buffalo are present (Radley et al., 1979; Radley, 1981; Di Giulio et al., 2009; Martins et al., 2010; Bishop et al., 2015; Sitt et al., 2015), but has been ineffective in protecting cattle introduced into areas of heavy tick challenge grazed only or predominantly by buffalo (Bishop et al., 2015; Sitt et al., 2015). The latter has been interpreted to reflect the greater antigenic diversity in the buffalo-derived *T. parva*.

Areas where buffalo populations interact with cattle are important across sub-Saharan Africa for transmission of multiple infectious diseases (Casey et al., 2014). This is particularly so for *T. parva*, where furthering understanding of the relationship between the parasite populations circulating in cattle and buffalo is a critical factor in resolving both the epidemiology of *T. parva* and vaccine utility. However, few studies have examined *T. parva* genetic diversity in co-circulating populations deriving from sympatric cattle and buffalo. This study aimed to dissect the genetic relationship between *T. parva* identified in cattle and African buffalo sampled in and around the Serengeti National Park, Tanzania, an area endemic for *T. parva*. Long read (PacBio) sequencing was used to obtain sequences from full or near-full length amplicons of genes encoding three antigens recognized by parasite-specific T cells; Tp1, Tp4, and Tp16 (Graham et al., 2006; Tretina et al., 2020; Morrison et al., 2021, submitted), which were selected on the basis of amplifying from *T. parva*, but not from the closely related *T.* sp. buffalo. These data indicated significantly greater allelic diversity in buffalo-derived *T. parva*, with substantial multiplicity of infection in buffalo, and very little sharing of antigen alleles between buffalo and cattle *T. parva* populations. While there was limited evidence for genetic sub-structuring of the two populations, one clade of Tp1 alleles was almost entirely only found in buffalo, suggesting subpopulation host restriction.

MATERIALS AND METHODS

Ethics Statement

Ethical clearance for cattle sampling was gained from the Animal Experimentation Committee of Scotland's Rural College (SRUC), and the Commission for Science and Technology (COSTECH), Tanzania (Research Permit Number 2016-32-NA-2016-19). Buffalo sampling was carried out with the approval of and in collaboration with the Tanzania Wildlife Research Institute (TAWIRI), as previously described (Casey-Bryars et al., 2018).

Study Area

The study was undertaken in the Serengeti National Park, Tanzania and adjacent livestock-keeping areas (**Figure 1**). The protected areas have unfenced boundaries where livestock can interact with wildlife, including buffalo. Communities in the study area practice livestock keeping as well as mixed crop-livestock farming (Estes et al., 2012). Cattle breeds farmed in this area are predominantly indigenous zebu × Tarime and Zebu × Maswa (Sahiwal, Boran and Mpwapwa zebu cross breeds). Livestock density is highest along the game reserve boundaries north-west and south-west of the National Park (TAWIRI, 2016).

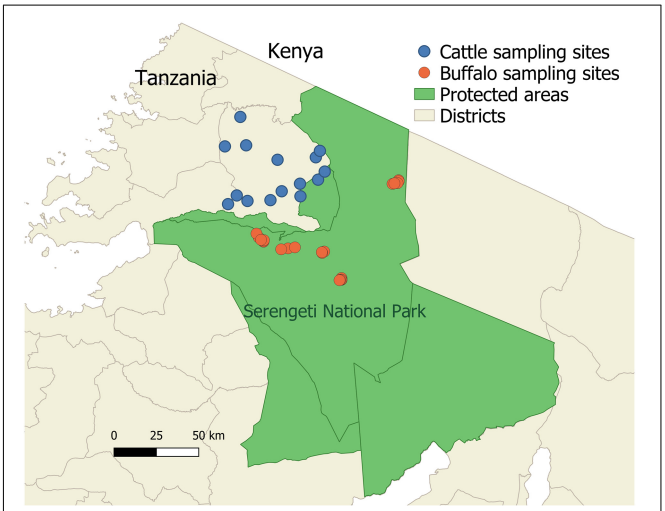


FIGURE 1 | The study area showing locations of cattle and buffalo sampling in the Serengeti-Mara ecosystem. Cattle sampling sites are shown in blue and buffalo sampling sites are shown in orange. Cattle sampling sites (villages) were selected from those close to the protected area boundaries in Serengeti District. Protected wildlife areas are shown in green.

Sample Collection

Cattle samples were from agropastoral farms, where cattle are herded daily for grazing and water, with herd sizes ranging from 4 to 1,000 cattle. Samples derived from several studies within the study area (**Figure 1**): In 2016 a randomized cross-sectional survey was designed and carried out in order to gain an estimate of the prevalence of *T. parva* in cattle farmed at the boundary of the Serengeti National Park, using a multistage stratified strategy to select herds, which resulted in 48 herds being sampled (n = 770). In 2011 a cross-sectional livestock survey was carried out in a similar study area, which sampled cattle from six herds (n = 199). Samples were also analyzed from herds recruited for longitudinal studies on Animal African Trypanosomiasis; samples from four herds were used from 2013, 2015, and 2017 (n = 432), selected on the basis of providing resolution for potential analysis of genotypes over space and time. A further subset of the longitudinal cattle samples was used as they derived from cattle described to be clinically ill at the time of sampling (n = 201)—while not specifically diagnosed with ECF, these were included in the expectation that some may have been clinical ECF cases (collected in 2013, 2014, and 2015). Buffalo samples (n = 22; **Figure 1**) were collected from the Serengeti National Park in 2011 as previously described (Casey-Bryars et al., 2018). The study groups are referred to as ST1 (cross-sectional survey 2011), ST2 (cross-sectional survey 2016), ST3 (longitudinal 2013), ST5 (longitudinal 2017), ST6 (clinically ill), ST7 (buffalo), and ST8 (reference strain TpM). Sample names used in the manuscript include a designation of study group (ST), village/location of sampling (number) and host (C or B for cattle or buffalo, respectively).

Sampling and handling of cattle was carried out by local Serengeti District livestock officers and a Tanzanian veterinarian. A 10 ml blood sample was collected from the jugular vein into a Paxgene DNA tube (Qiagen). Buffalo samples were collected in the Serengeti National Park by Tanzanian Wildlife Research Institute (TAWIRI) veterinary teams, under strict guidelines regulating wildlife immobilizations. Buffalo were anaesthetized for a short period for jugular venepuncture to collect blood samples, as previously described (Casey-Bryars et al., 2018). Global Positioning System (GPS) location was recorded for all sampled cattle and buffalo (**Figure 1**).

DNA Extraction and PCR Amplification

The PAXgene Blood DNA Kit (Qiagen) was used to isolate genomic DNA from whole blood, according to manufacturer's instructions. DNA was stored at −20 or −80°C for longer term.

Diagnostic PCR

A nested p104 PCR (nPCR) was used to screen all field samples for the presence of *T. parva* (Skilton et al., 2002; Odongo et al., 2010). Each PCR reaction (25 μl) consisted of 12.5 μl Quick-Load Taq 2X Master Mix (New England Biolabs), 1 μl of each primer (10 μM), 10 μl nuclease-free water and 1 μl of DNA template. For the second round PCR, first round product was diluted 1:100 in dH_2O for use as template. The nPCR reactions were carried out in a thermal cycler (MJ Research PTC-200 Engine). First round (primer sequences: 5′ATT TAA GGA ACC TGA CGT GAC TGC 3′ and 5′TAA GAT GCC GAC TAT TAA TGA CAC C 3′) and second round (primer sequences: 5′GGC CAA GGT CTC CTT CAG ATT ACG 3′ and 5′TGG GTG TGT TTC CTC GTC ATC TGC 3′) PCR conditions were as previously described (Skilton et al., 2002; Odongo et al., 2009). The nPCR products were visualized by UV trans-illumination in a 1.5% agarose gel containing GelRed (Biotium) following electrophoresis.

Antigen Gene PCR

Samples positive by p104 nPCR were selected for amplification and characterization of *T. parva* antigen genes Tp1, Tp4, and Tp16. Specific primers were designed for nested PCR amplification of full or near full-length gene sequences, with the exception of second round primers for Tp1, previously published by Pelle et al. (2011). Primers were predicted to amplify a 1,618 bp product from Tp1 (TP03_0849: annotated gene length 1,771 bp), 1,473 bp from Tp4 (TP03_0210: 1,740 bp) and 983 bp product from Tp16 (TP01_0726: 1,347 bp). Primer specificity was validated using a panel of DNA (**Table 1**) from multiple *T. parva* isolates, as well as DNA samples from *T. annulata*, *T. buffeli*, *T. taurotragi*, *T.* sp. (buffalo) and DNA from uninfected *R. appendiculatus* ticks as a negative control. Primers were tested against 22 previously obtained *T. parva* isolates (including cattle and buffalo-derived isolates) to ensure they amplified across the range of strains expected, and were also tested against DNA from 5 *T.* sp. (buffalo) isolates to ensure that primers were specific to *T. parva* and did not amplify product from this closely related species (Bishop et al., 2015; Hemmink et al., 2018), expected to be present in buffalo samples. Each PCR reaction (25 μl) consisted of 5 μl Q5 High-Fidelity Reaction Buffer (New England Biolabs), 0.5 μl dNTPs (Bioline), 1.25 μl of each primer (10μM), 0.25 μl q5 High-Fidelity DNA Polymerase (New England Biolabs), 15.75 μl nuclease-free water and 1 μl of DNA template. For the second round template the first round product was diluted 1:50 in DH_2O, and a unique barcode sequence was added to the 5′ end of second round primers for each sample (Eurofins Scientific) in order to combine individual amplicons into a single pool, per antigen gene, for multiplexed sequencing[1]. Primer sequences and PCR conditions are shown in **Table 2**. All PCR reactions were carried out in a MJ Research PTC-200 DNA Engine thermal cycler, and the nPCR products were visualized by UV trans-illumination in a 1.5% agarose gel containing GelRed (Biotium) after electrophoresis. Positive amplicons were subsequently purified using the QIAquick PCR Purification Kit Protocol (Qiagen), according to manufacturer's instructions.

Sample Preparation for Long Read (PacBio) Sequencing

In addition to field samples, amplicons derived from the reference genome stock, *T. parva* Muguga (TpM) were included in each pool as a positive single clone (and allele) control. As a technical control and in order to be able to assess whether the depth of sequencing used captured all diversity present, every sample for Tp1 and Tp4 was submitted for sequencing twice, with a further barcode differentiating these batches (referred to as barcode A

[1]https://www.pacb.com/products-and-services/analytical-software/multiplexing/

TABLE 1 | DNA panel used to test primer specificity.

DNA Sample ID	Country of origin	Sample type
T. parva (Muguga)	Kenya	Reference genome
T. annulata	Turkey	Non-pathogenic/alternative *Theileria* species
T. buffeli	Kenya	
T. taurotragi	Kenya	
T. sp. (buffalo) 6834 clone 10	Kenya	
T. sp. (buffalo) 6998 clone 10	Kenya	
T. sp. (buffalo) 6834 clone 5	Kenya	
T. sp. (buffalo) 6998 clone 2	Kenya	
T. sp. (buffalo) 6998 clone 4	Kenya	
Rhipicephalus appendiculatus DNA	Kenya (ILRI)	Negative control
Buffalo clone M3.3	Kenya	Buffalo-derived *T. parva* clones
Buffalo clone M3.6	Kenya	
Buffalo clone M3.7	Kenya	
Buffalo clone M3.9	Kenya	
Buffalo clone M30.2	Kenya	
Buffalo clone M30.5	Kenya	
Buffalo clone M30.8	Kenya	
Buffalo clone M30.11	Kenya	
Buffalo clone M42.2	Kenya	
Buffalo clone M42.5	Kenya	
Buffalo clone M42.8	Kenya	
Buffalo clone M42.12	Kenya	
Buffalo clone 6998.9	Kenya	
Buffalo clone 6998.11	Kenya	
Buffalo-assoc. clone N33.1	Kenya	Buffalo-associated cattle *T. parva* clones
Buffalo-assoc. clone N33.3	Kenya	
Buffalo-assoc. clone N33.4	Kenya	
Buffalo-assoc. clone N33.5	Kenya	
Buffalo-assoc. clone N43.1	Kenya	
Buffalo-assoc. clone N43.3	Kenya	
Buffalo-assoc. clone N43.5	Kenya	
Buffalo-assoc. clone N43.6	Kenya	

or barcode B). Qubit fluorimetry (Thermo Fisher Scientific) was used to calculate equimolar quantities of PCR products for each pool and to verify final DNA concentration of each pool. DNA purity was assessed with NanoDrop spectrophotometry (Thermo Fisher Scientific) and Agilent 2200 TapeStation System (Agilent Technologies) was used to assess DNA integrity and size. Ethanol precipitation was carried out to further purify and concentrate the pooled sample, resulting in 3–5 μg DNA at 50 ng/μl per amplicon sample for library preparation.

The pooled amplicons (96 samples for Tp16; 48 samples for Tp1 and Tp4 due to running A and B replicates) comprised 46 samples that had generated amplicons for Tp1 (34 cattle, 12 buffalo; 3 μg at 61 ng/μl), 47 samples for Tp4 (32 cattle, 15 buffalo; 3.8 μg at 75.4 ng/μl), 94 samples for Tp16 (73 cattle, 21 buffalo; 3.3 μg at 65.6 ng/μl of Tp16), as well as amplicons derived from TpM in each antigen pool (**Table 3** for sample details). Each pool was sequenced on a single SMRT cell on a PacBio Sequel machine (Edinburgh Genomics).

PacBio Sequencing Analysis

Raw Data Processing

PacBio raw unfiltered multiplexed sub-reads were received in bam format from Edinburgh Genomics for each of the three separate PacBio runs containing reads from each antigen—Tp1, Tp4, and Tp16, respectively. PacBio circular consensus caller (PBCCS v3[2]) was used to generate CCS reads from sub-reads. PacBio barcode demultiplexer (lima v1[3]) was used to demultiplex the CCS reads into their respective biological replicate samples using the primer barcode information used in the PCR amplification step. PacBio long read aligner (Blasr[4]) was used to align CCS reads from each sample to their respective reference amplicon sequence and outputted in the machine parsable format using the –placeGapConsistently option for better variant calling downstream using downstream data analysis scripts. Data and scripts are available via the following link: https://doi.org/10.7488/ds/3003. The main script files for each data set are titled Tp1_Variant_Analysis.m, Tp4_Variant_Analysis.m and

[2]https://github.com/PacificBiosciences/CCS
[3]https://github.com/PacificBiosciences/barcoding
[4]https://github.com/PacificBiosciences/blasr

TABLE 2 | Primers and PCR cycling conditions for antigens.

Gene	Primers	Cycling conditions
Tp1		
Round 1	F: GCTACGCGGAAATCTAGGCT R: CATCGTTTGCCAGCACTATGA	98°C for 30 s, 40 cycles (98°C for 10 s, 56°C for 20 s, 72°C for 2.5 min), 72°C for 2 min
Round 2*	F: AGGGTCAAAAAAGTTTTATTA R: TTAATTTTTGAGGTAAATTTTG	98°C for 30 s, 37 cycles (98°C for 10 s, 54°C for 20 s, 72°C for 1.5 min), 72°C for 2 min
Tp4		
Round 1	F: ATACATCCCAAGGCCAAGCT R: GGAAGGGGTTGGATAGTGCT	98°C for 30 s, 40 cycles (98°C for 10 s, 58°C for 20 s, 72°C for 2.5 min), 72°C for 2 min
Round 2	F: TTACTCATCCTGCCGCTTCT R: TGACCTCCACCTCTCAACAC	98°C for 30 s, 30 cycles (98°C for 10 s, 64°C for 20 s, 72°C for 2 min), 72°C for 2 min
Tp16		
Round 1	F: TGATCTACAAGCTCGGTGGA R: GCGGGTATTCTGTGAAGGTC	98°C for 30 s, 35 cycles (98°C for 10 s, 68 C for 20 s, 72°C for 1.5 min), 72°C for 2 min
Round 2	R: AGACATGGGAAAGGGAAGCT F: CCTCCAGTGTCTTTCCGGTA	98°C for 30 s, 32 cycles (98°C for 10 s, 56°C for 20 s, 72°C for 1.5 min), 72°C for 2 min

*Pelle et al. (2011).

TABLE 3 | Summary of data per sample.

Sample ID	Date sampled	Sex*	Study**	Antigen†		
				Tp1	Tp4	Tp16
ST1_01_C01	31/10/2011	F	CS 2011	+	+	-
ST1_01_C02	31/10/2011	F	CS 2011	-	+	-
ST1_01_C03	01/11/2011	F	CS 2011	+	+	-
ST1_01_C04	01/11/2011	F	CS 2011	+	-	-
ST1_02_C05	26/10/2011	F	CS 2011	-	-	+
ST1_02_C06	26/10/2011	M	CS 2011	-	-	+
ST1_02_C07	26/10/2011	F	CS 2011	-	-	+
ST1_02_C08	26/10/2011	F	CS 2011	-	-	+
ST1_02_C09	27/10/2011	M	CS 2011	-	-	+
ST1_02_C10	27/10/2011	M	CS 2011	-	-	+
ST1_02_C11	27/10/2011	F	CS 2011	+	+	+
ST1_03_C12	01/08/2011	F	CS 2011	+	-	+
ST1_04_C13	27/07/2011	F	CS 2011	-	-	+
ST1_04_C14	27/07/2011	F	CS 2011	+	-	+
ST1_04_C15	27/07/2011	F	CS 2011	-	-	+
ST1_04_C16	27/07/2011	F	CS 2011	-	+	+
ST1_04_C17	27/07/2011	F	CS 2011	+	-	+
ST1_05_C18	24/10/2011	M	CS 2011	+	-	+
ST1_05_C19	24/10/2011	M	CS 2011	+	-	+
ST1_05_C20	24/10/2011	M	CS 2011	-	-	+
ST1_05_C21	24/10/2011	M	CS 2011	+	+	+
ST1_05_C22	25/10/2011	M	CS 2011	-	-	+
ST1_05_C23	25/10/2011	M	CS 2011	-	-	+
ST1_05_C24	25/10/2011	M	CS 2011	+	+	+
ST1_06_C25	25/07/2011	F	CS 2011	-	-	+
ST1_06_C26	25/07/2011	F	CS 2011	-	+	+
ST1_06_C27	25/07/2011	M	CS 2011	-	-	+
ST1_06_C28	25/07/2011	F	CS 2011	-	-	+
ST1_06_C29	25/07/2011	F	CS 2011	-	-	+
ST1_06_C30	25/07/2011	F	CS 2011	-	-	+
ST1_06_C31	25/07/2011	F	CS 2011	-	-	+
ST1_06_C32	25/07/2011	F	CS 2011	-	-	+
ST1_06_C33	26/07/2011	M	CS 2011	-	-	+
ST1_06_C34	26/07/2011	F	CS 2011	+	-	+
ST1_06_C35	26/07/2011	F	CS 2011	-	-	+
ST1_06_C36	26/07/2011	F	CS 2011	-	-	+
ST2_01_C37	21/07/2016	F	CS 2016	-	+	+
ST2_02_C38	18/07/2016	M	CS 2016	+	+	+
ST2_02_C39	18/07/2016	F	CS 2016	+	+	+
ST2_02_C40	20/07/2016	F	CS 2016	+	+	+
ST2_02_C41	19/07/2016	F	CS 2016	-	-	+
ST2_02_C42	19/07/2016	F	CS 2016	+	+	+
ST2_02_C43	19/07/2016	M	CS 2016	+	+	+
ST2_03_C44	04/08/2016	M	CS 2016	-	-	+
ST2_03_C45	04/08/2016	F	CS 2016	+	+	+
ST2_03_C46	05/08/2016	F	CS 2016	-	-	+
ST2_03_C47	05/08/2016	F	CS 2016	+	+	+
ST2_03_C48	05/08/2016	F	CS 2016	-	-	+
ST2_04_C49	10/08/2016	F	CS 2016	-	-	+
ST2_04_C50	11/08/2016	M	CS 2016	-	+	+
ST2_04_C51	11/08/2016	F	CS 2016	+	+	+
ST2_04_C52	11/08/2016	F	CS 2016	-	-	+
ST2_05_C53	28/07/2016	F	CS 2016	+	+	+
ST2_05_C54	28/07/2016	M	CS 2016	-	-	+
ST2_06_C55	08/08/2016	F	CS 2016	-	+	+
ST2_06_C56	08/08/2016	F	CS 2016	+	+	+
ST2_06_C57	09/08/2016	M	CS 2016	+	-	+
ST2_06_C58	09/08/2016	F	CS 2016	+	-	+
ST2_06_C59	09/08/2016	M	CS 2016	-	+	+
ST2_07_C60	25/07/2016	F	CS 2016	+	-	+
ST2_07_C61	25/07/2016	F	CS 2016	+	-	+
ST3_01_C62	15/04/2013	F	Long 2013	+	+	+
ST3_02_C63	15/05/2013	F	Long 2013	-	-	+
ST3_02_C64	15/05/2013	M	Long 2013	-	-	+
ST3_02_C65	15/05/2013	F	Long 2013	+	-	+
ST3_02_C66	15/05/2013	M	Long 2013	-	-	+
ST4_01_C67	20/01/2015	?	Long 2015	-	+	+
ST4_01_C68	28/01/2015	M	Long 2015	-	-	+
ST4_02_C69	14/03/2015	F	Long 2015	-	-	+
ST4_02_C70	14/03/2015	F	Long 2015	-	-	+
ST5_01_C71	18/05/2017	F	Long 2017	+	+	+
ST5_01_C72	18/05/2017	F	Long 2017	+	+	+
ST5_02_C73	30/05/2017	M	Long 2017	+	+	+
ST5_02_C74	30/05/2017	M	Long 2017	-	-	+
ST5_02_C75	30/05/2017	F	Long 2017	-	+	-
ST5_03_C76	10/05/2017	F	Long 2017	+	+	+
ST6_01_C77	09/09/2014	F	Sick 2014	+	+	+
ST6_01_C78	09/09/2014	F	Sick 2014	+	+	-
ST7_01_B01	01/07/2011	F	Buffalo 2011	-	-	+
ST7_01_B02	01/07/2011	F	Buffalo 2011	+	+	+
ST7_01_B03	01/07/2011	M	Buffalo 2011	+	+	+
ST7_01_B04	01/07/2011	F	Buffalo 2011	-	+	+
ST7_01_B05	01/07/2011	F	Buffalo 2011	-	+	+
ST7_01_B06	02/07/2011	M	Buffalo 2011	+	+	+
ST7_01_B07	02/07/2011	M	Buffalo 2011	-	+	+
ST7_01_B08	02/07/2011	M	Buffalo 2011	-	+	+
ST7_01_B09	02/07/2011	M	Buffalo 2011	+	+	+
ST7_01_B10	02/07/2011	M	Buffalo 2011	+	+	+
ST7_01_B11	03/07/2011	M	Buffalo 2011		+	+
ST7_01_B12	03/07/2011	F	Buffalo 2011	+	+	+
ST7_01_B13	03/07/2011	F	Buffalo 2011	-	-	+
ST7_01_B14	05/07/2011	?	Buffalo 2011	-	-	+
ST7_01_B15	05/07/2011	M	Buffalo 2011	+	+	+
ST7_01_B16	05/07/2011	M	Buffalo 2011	+	-	+
ST7_01_B17	06/07/2011	M	Buffalo 2011	+	+	+
ST7_01_B18	06/07/2011	M	Buffalo 2011	+	-	+
ST7_01_B19	06/07/2011	F	Buffalo 2011	-	+	+
ST7_01_B20	06/07/2011	F	Buffalo 2011	+	-	+
ST7_01_B21	06/07/2011	F	Buffalo 2011	+	+	+

**M = Male; F = Female; ? = sex unrecorded.*

***Study groups are: CS 2011 (cross-sectional survey 2011; ST1), CS 2016 (cross-sectional survey 2016; ST2), Long 2013 (longitudinal 2013; ST3), Long 2015 (longitudinal 2015; ST4), Long 2017 (longitudinal 2017; ST5), Sick 2014 (clinically ill; ST6) and Buffalo 2011 (buffalo; ST7).*

†Symbols represent successful (+) or unsuccessful (-) amplification of antigen gene by PCR.

Tp16_Variant_Analysis.m, respectively (please note that within the data files Tp16 is referred to by the alias N60).

Read QC and Filtering

CCS reads were filtered to retain only high-quality reads for variant calling and allelic distribution using the following criteria (PCR replicates accounted as separate samples at this stage); (a) sequenced length of a CCS read length should fall within the mean ± SD sequence length distribution window for reads coming from all samples of a given antigen, to remove short fragments and long artifacts, (b) using the PCR primers as a guide, each CCS read was tested to make sure they were full length amplicons and remove all fragmented reads, (c) the PacBio read quality score was used to retain only high confidence reads with score RQ > 0.99 from the consensus caller, (d) for Tp1 and Tp16, genomic amplicons without introns translated in frame and reads with non-sense codons in open reading frame (ORF) were filtered out.

Variant Calling and Allelic Distribution

CCS reads passing the above filters were analyzed for their Blasr alignment to their respective reference amplicon and variant calling was done using an in-house pipeline[5]. The CCS reads were grouped based on whether the reads originated from cattle, buffalo or TpM samples. For each of the groups, variant frequency at each base pair locus was calculated for SNPs and insertions/deletions (INDELs) based on their alignments against the reference sequence (**Figure 2A**).

SNP and INDEL sites identified were filtered for variant frequency greater than 1% for Tp1 and Tp16 samples and greater than 5% for Tp4 samples to account for background error introduced by PacBio sequencing, such as homopolymer errors, for each of the source group—cattle, buffalo and TpM—samples. For each source group, the SNP/INDEL which passed the variant frequency threshold from each sequenced read was substituted into the reference amplicon sequence, in order to remove any background noise from the sequenced reads. For Tp1 and Tp16 an additional filtering step was applied to the modified reads to remove any alleles which failed to produce a complete ORF, by removing any alleles which were out of frame or introduced non-sense mutation during SNP/INDEL substitution. Unique alleles were identified by clustering the modified reads originating from cattle and buffalo at 100% identity. The sequenced read count evidence supporting each identified allele was summarized across the total sequenced pool (cattle, buffalo, TpM) and each individual sample replicate (**Supplementary Data 1–3**). For Tp1 and Tp16, each allele was tested against the reference to check if the mutation was synonymous or non-synonymous. For Tp1, the known epitope region (Nanteza et al., 2020) was selected from each allele sequence to list all unique epitope residues. Ts/Tv ratio was calculated for each source group Tp1 samples—cattle, buffalo, TpM. Identified alleles were summarized for each animal (replicates combined) into total alleles, alleles shared between two replicates, and alleles in replicate A and replicate B separately. To assess the allelic diversity and potential genetic relationships between cattle- and buffalo-derived *T. parva* populations, a phylogenetic (neighbor-joining) tree was constructed using MEGA v.7.

[5]https://doi.org/10.7488/ds/3003

RESULTS

Detection of *Theileria parva* in Field Samples and Amplification of Antigen Genes

A nested p104 PCR assay was used to screen 1,602 cattle and 22 buffalo samples, of which 126 cattle (7.87%) and 22 buffalo (100%) were positive for *T. parva*. Primers designed to amplify full or near-full length gene sequences of the antigens Tp1, Tp4, and Tp16 were applied to the 126 cattle and 22 buffalo *T. parva* positive samples. A PCR product of the predicted size (1,618 bp) was obtained from 34 cattle and 12 buffalo samples for Tp1, 32 cattle samples and 15 buffalo samples for Tp4 (1,473 bp), and 80 cattle and 21 buffalo samples for Tp16 (983 bp). All 46 samples for Tp1 (34 cattle, 12 buffalo) and 47 samples for Tp4 (32 cattle, 15 buffalo) that generated PCR products, as well as amplicons deriving from TpM, were submitted in duplicate for sequencing (**Table 3**). For Tp16, 94 samples that had generated amplicons (73 randomly selected cattle and 21 buffalo), along with TpM amplicons, were submitted for sequencing (no duplication; **Table 3**).

Tp1 PacBio Sequencing

Using PacBio long read sequencing, a total of 404,723 raw CCS reads was obtained; 282,998 for cattle-derived samples (34 samples, submitted in duplicate), 113,149 for buffalo-derived samples (12 samples, submitted in duplicate) and 8,576 for TpM ($n = 2$; **Supplementary Data 1**).

After filtering steps based upon read quality (RQ ≥ 0.99; 356,440 reads passed this threshold; 250,139 for cattle, 98,710 for buffalo and 7,591 for TpM), read length (mean ± 1 SD; 311,618 full length reads; 222,726 for cattle, 82,130 for buffalo, and 6,762 for TpM), codon length (checking in-frame; 224,513 reads; 166,801 for cattle, 52,517 for buffalo, and 5,195 for TpM) and ORF (checking for stop codons; 160,326 reads; 128,200 for cattle, 28,013 for buffalo, and 4,113 for TpM), there was a total of 155,159 reads remaining (38.3% of starting total); 125,851 from 68 cattle samples (average of 1,850 reads per sample), 25,359 for 24 buffalo samples (average of 1,056 reads per sample), and 3,949 for TpM (average of 1,974 per TpM sample).

Tp1 Alleles

At the SNP/INDEL frequency threshold of 1%, mismatches, deletions and insertions were identified at 110, 48 and 13 nucleotide positions, respectively, i.e., there were a total of 171 variable nucleotide positions across the 1,618 bp amplicon (**Figure 2A**). This resulted in a total of 3,876 alleles being initially identified. However, a filter was applied to avoid potential errors deriving from single reads (i.e., to be considered in our dataset

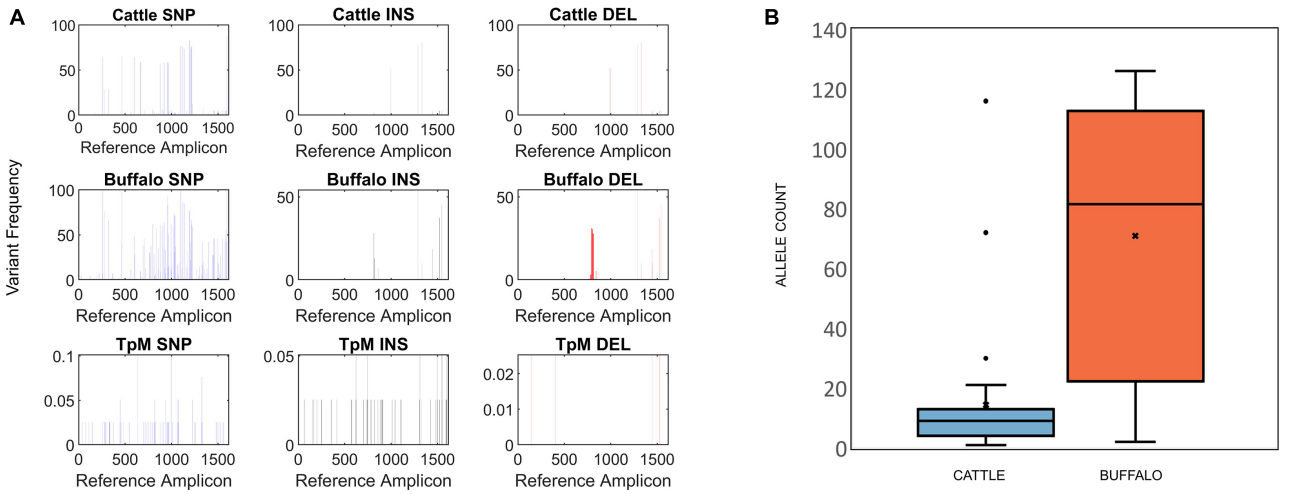


FIGURE 2 | (A) Single Nucleotide Polymorphism (SNP), insertion (INS) and deletion (DEL) frequency for amplicon data across the length of the Tp1 amplicon, with respect to the TpM reference sequence. The mean calculated SNP/INDEL frequency is shown at each base location for the Tp1 gene. Blue bars represent SNPs, black bars represent insertions and red bars represent deletions. **(B)** Tp1 allele counts in cattle- (blue) and buffalo-derived (orange) samples. Median (line), mean (x) and standard deviation (whiskers) are shown, dots represent outlier samples with extreme allele counts.

a sequence must be represented by two independent reads), which resulted in a final allele count of 651, which included the reference TpM allele. The read count of this final filtered dataset was 150,198.

Importantly, the data filtering steps resulted in only a single allele being detected in the TpM control dataset (3,949 reads for TpM; 2,165 reads in replicate A and 1,784 reads in replicate B), consistent with expectations as this DNA derived from a clonal *in vitro* cultured *T. parva* cell line, and suggesting filtering steps were appropriately stringent. Notably, the TpM sequence was also identified in one buffalo (294 reads) and eight cattle (7,715 reads).

Buffalo samples on average had much higher allelic diversity than cattle, with a median of 81.5 alleles (S.D. = 45.2) being identified in buffalo compared to a median of nine alleles (S.D. = 21.8) in cattle samples (**Figure 2B**). Buffalo samples ranged from 126 alleles identified in ST7_01_B06 to 22 alleles in ST7_01_B02, with two buffalo samples, ST7_01_B18 and ST7_01_B20, having relatively low numbers of alleles detected, 10 and 2, respectively. Cattle sample-derived allele numbers ranged from 30 in ST1_04_C14 to a single allele in ST2_04_C51, with the exception of samples ST1_05_C21 and ST2_03_C45, in which 116 and 72 were detected, respectively. After data filtering, cattle sample ST1_01_C03 had no detectable alleles, and so this sample was removed from further analysis. The number of alleles identified did not necessarily correlate directly with number of reads—for example, for the cattle-derived samples ST1_05_C21 and ST2_03_C45, which were outliers in terms of high numbers of alleles (116 and 72), there were 3,281 and 6,067 reads, compared to the other cattle from same study groups; 1,443 (ST1_05_C18—2 alleles), 1,720 (ST1_05_C19—9 alleles), 4,400 reads (ST1_24_C24—13 alleles) and 6,785 reads (ST3_05_C47—3 alleles). Similarly, the buffalo-derived samples ST7_01_B18 and ST7_01_B20 which had unusually low allele counts (10 and 2, respectively) did not have significantly lower read counts compared to other samples in the buffalo-derived sample cohort. These data suggest that it is not simply differential sequencing coverage that is responsible for the high or low allele detection in these samples.

Each individual sample was sequenced twice independently, resulting in two replicate datasets (A and B) from the same amplicon (note that this pertains to 36 samples; for samples ST7_01_B21, ST7_01_B20, ST5_01_C71, ST1_01_C03, ST7_01_B02, ST6_01_C78, ST2_05_C53, ST2_06_C57, ST2_04_C51 and ST1_01_C03 data was only successfully generated for one replicate). This approach provided the ability to assess the coverage of sequencing with respect to allele detection (summarized in **Figure 3**). The alleles detected in buffalo-derived replicates A and B for the same samples, while showing a degree of overlap were very often different, indicating that the depth of sequencing coverage was not sufficient to detect all alleles in a sample. In contrast, sequencing replicates of cattle-derived samples indicated much greater sharing of alleles, with the exception of outlying cattle C21 and C45 (**Figure 3**) between the replicates of individual samples, suggesting that for cattle-derived samples the sequencing coverage was sufficient to identify all or most alleles present. TpM replicates were identical, as expected with a single allele control. These data confirm that the diversity in buffalo-derived samples is much greater than in cattle-derived samples, and indeed indicate that the approach taken was not sufficient to identify all alleles potentially present in buffalo-derived samples.

Tp1 Allele Sharing

When allele sharing within and between species groups was analyzed, sharing clearly commonly occurs within cattle or within buffalo (blue or orange ribbons, respectively, in **Figure 4**), whereas sharing between species (black ribbons

in **Figure 4**) was much less frequent. Of the 420 alleles detected in buffalo, 412 alleles were unique to buffalo and not detected in cattle samples. 239 alleles were identified in cattle samples, and of those 231 alleles were only found in cattle samples. Therefore, only eight of the 651 alleles were present in both cattle and buffalo, and this included the reference TpM allele (green ribbon in **Figure 4**). Most of the cattle-unique alleles derive from the two samples with high allele counts (ST1_05_C21 and ST2_03_C45), with 118/231 cattle-derived alleles (51.1%) being detected only in these samples.

Of the eight alleles shared between species (**Figure 4**), four alleles were found in multiple buffalo and a single cow (alleles #0304, #0461, #0501, and #0300). One allele was found in one buffalo and one cow (#0120), and two alleles (#0055 and TpM) were found in multiple cattle and a single buffalo, TpM being present in eight and #0055 in seven cattle, respectively. It is difficult to discern patterns with such low numbers of shared alleles.

Of the 34 cattle, 12 did not share any alleles with buffalo (ST1_01_C01, ST1_02_C11, ST1_05_C18, ST1_06_C34, ST2_02_C43, ST2_03_C45, ST2_03_C47, ST2_06_C57, ST3_01_C62, ST3_02_C65, ST5_01_C71, ST5_03_C76; **Figure 4**). For the two cattle samples with unusually high allele numbers, cow ST2_03_C45 (72 alleles) shared no alleles with buffalo and ST1_05_C21 (116 alleles) shared four alleles with four buffalo (B10, B12, B15, B16), including the TpM allele. For the 12 buffalo samples that gave data, only two buffalo, the samples with low allele counts (ST7-01-B20 and ST7-01-B18), shared no alleles at all with cattle. The buffalo samples with the highest allele counts, B12 (120), B15 (113), and B16 (98) each had four alleles shared with cattle.

These data further indicate that for cattle and buffalo the degree of sharing between species was low, although for buffalo most animals had at least some alleles that are detected in cattle, whereas for cattle (where the data indicate that we have the resolution to detect all alleles in the samples), a substantial proportion did not have any alleles that are also present in buffalo.

Clearly most alleles detected were unique to each species in the dataset. Of the 231 cattle-unique alleles, 90.9% were shared between two or more cattle, and this ranged from being shared between two animals to being shared between 19 animals. For example, 93 alleles (40.3%) were present in two cattle, 43 alleles (18.6%) in three, 22 alleles (9.5%) in four, 12 alleles (5.2%) in five cattle, and 8 alleles (3.5%) in six cattle. The remaining 32 alleles were shared between seven and 19 cattle; of the most frequently occurring, allele #0002 was present in 19 cattle (read

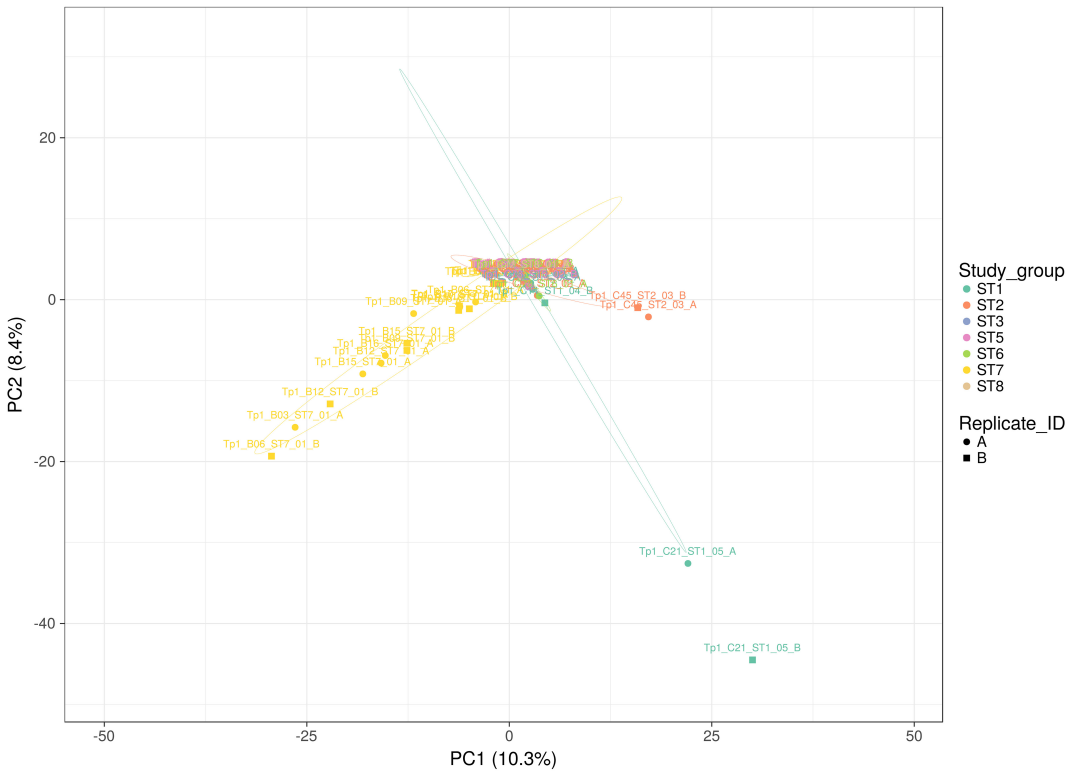


FIGURE 3 | Principal Component analysis applied to Tp1 alleles present in all individual replicates (A or B representing paired independent sequencing data from the same sample). Study group ST1 (aqua) represents cross-sectional cattle from 2011, ST2 (pink) represents cross-sectional cattle from 2016, ST3 (blue) represents longitudinal cattle from 2013, ST5 (lilac) represents longitudinal cattle from 2017, ST6 (green) represents clinically ill cattle, ST7 (yellow) represents buffalo, and ST8 (beige) represents reference strain TpM. Replicates A and B are represented by a circle and a square respectively. The proportion of variation in the dataset, explained by the 1st and 2nd principal components, is indicated in parenthesis on each axis.

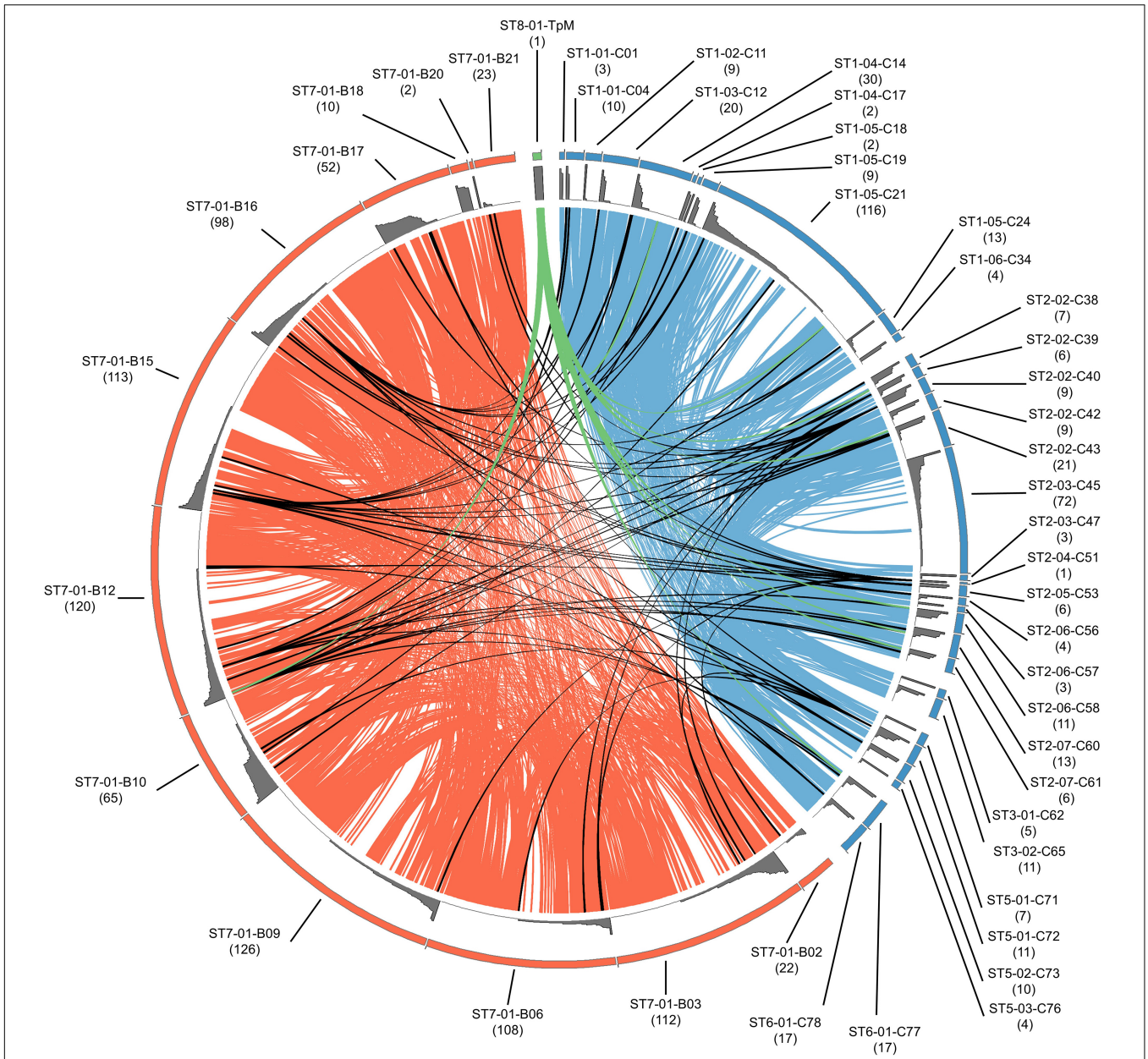


FIGURE 4 | Relationship of alleles within and between buffalo- and cattle-derived samples. Individual samples are arranged around the outer ring, as indicated by labels with sample name (blue indicating cattle and orange indicating buffalo); the size of the segment in the outer ring corresponds with the allele count per individual (also noted in brackets under sample name), with the relative abundance of each allele illustrated by the gray histogram. Alleles shared between buffalo samples are joined by orange ribbons, alleles shared between cattle samples are joined by blue ribbons, and alleles shared between buffalo and cattle are joined by black ribbons. TpM is shown in green, and is joined to animals in which TpM was detected by a green ribbon.

counts ranging from 1 to 6,781), allele #0021 was in 17 cattle (1–6,867) and allele #0031 was in 10 cattle (1-2). These data indicate that there were frequently shared alleles between cattle, but the most frequently shared were not necessarily the most abundant allele in all cattle.

Of the 412 buffalo-unique alleles, almost all were shared across at least two animals, with only three alleles being found in just one animal. 220 (53.4%) alleles were present in two buffalo, 90 alleles (21.8%) in 3 buffalo, 46 (11.2%) in four buffalo, 25 (6.1%) in five buffalo, and 18 (4.4%) in six buffalo. The remaining five most frequently encountered alleles were found in seven (alleles #611 and #771; read counts ranging from 1 to 4 and 5 to 524, respectively), eight (alleles #601 and #667; read counts 1–3 and 21–378, respectively) and ten buffalo (allele #0590; read counts ranging from 1 to 247). Again, these data indicate that there is not a correlation between the degree of sharing and allele dominance in individual buffalo.

Tp1 Allele Phylogeny

The lack of sharing between species could be due to many reasons, including potential genetic substructuring reflecting either adaptation to the respective host species or different transmission cycles. In order to investigate potential substructuring by host species, the genetic relationship between alleles found in cattle and buffalo was analyzed using a neighbor-joining clustering approach. The data indicated that there were two main clusters of alleles, one mainly buffalo-derived cluster, and one that broadly split into cattle-derived alleles and buffalo-derived alleles (**Figure 5**), with TpM being in the mainly cattle-derived sub-cluster. There were only six cattle alleles, and one allele found in both species, that group with the main buffalo-derived cluster, with all other alleles found in cattle or both species being present in the second cattle/buffalo cluster. While the bootstrap values indicated little support overall for strong signals of substructuring (perhaps not surprising given the relatively few variable sites across the dataset), and the small number of alleles shared between species limited what can be concluded, the species-specific clustering of alleles was very marked.

Tp1 CTL Epitopes

A Tp1 epitope recognized by the bovine immune response has been previously described (Graham et al., 2006, 2008; MacHugh et al., 2009; Pelle et al., 2011). This epitope sequence was analyzed to understand the diversity of a sequence that has been defined as interacting with the host immune response, to examine if there was any relationship between epitope identity and host species. Across the 651 Tp1 alleles, five epitope variants (i.e., nucleotide sequences that result in an amino acid change) were identified in the previously described sequence (VGYPKVKEEML in TpM) that demarcates an epitope recognized by protective cytotoxic CD8 T cells (**Table 4**). The predominant variant circulating in cattle was VGYPKVKEEII, found in 31 cattle (73,202 reads), followed by VGYPKVKEEML in 24 cattle (50,791 reads), VGYPKVKEEMI in 12 cattle (196 reads) and VGYPKVKEEIL in six cattle (56 reads). In buffalo, variant VGYPKVKEEML was predominant, found in all 12 buffalo (19,896 reads), followed by VGYPKVKEEII in nine buffalo (1,805 reads), VGYPKVKEEMV in four buffalo (301 reads) and VGYPKVKEEIL in a single buffalo (2 reads). The variant with MI at positions 10 and 11 was therefore only found in cattle, and that with MV was only found in buffalo.

TABLE 4 | Tp1 epitope variants found in cattle and buffalo samples.

Epitope	Number cattle	Cattle reads	Number buffalo	Buffalo reads
VGYPKVKEEML	24	50,791	12	19,896
VGYPKVKEEMV	0	0	4	301
VGYPKVKEEMI	12	196	0	0
VGYPKVKEEIL	6	56	1	2
VGYPKVKEEII	31	73,202	9	1,805

Tp4 and Tp16 Alleles

In addition to Tp1, analysis of two additional antigens recognized by CD8 or CD4 T cells, Tp4 and Tp16, was also carried out (**Supplementary Data 2** and **3**). For Tp4, amplicons from 48 samples (32 cattle, 15 buffalo and TpM) were sequenced in duplicate (total $n = 96$). A total of 74,274 alleles were initially identified, but after the removal of intronic regions (**Figure 6A**) and filtering (a sequence must be represented by two independent reads), the final allele count was 4,449 (286,100 reads). A single TpM allele was identified (11,422 reads) at the SNP/INDEL variant frequency threshold of 5% used (**Figure 6B**). The median allele count in cattle samples was 32.5 (SD = 116.4), with a range of 5–84 alleles per cow (reads per allele ranged from 410 to 12,388; **Figure 7A**), excluding seven outlying cattle samples with significantly higher allele counts (C42—570 alleles, 6,399 reads; C40—298 alleles, 7,922 reads; C21—242 alleles, 7,994 reads; C45—210 alleles, 7,465 reads; C75—207 alleles, 9,828 reads; C26—204 alleles, 9,229 reads; C55—192 alleles, 7,708 reads). In contrast, the median allele count in buffalo was 484 (SD = 111.1), with a range of 298–684 alleles per buffalo (reads per allele 319–524; **Figure 7A**). When allele sharing within and between species was analyzed, 1,248 Tp4 alleles were unique to cattle (including TpM, #0001), 3,201 were unique to buffalo and no alleles were shared.

For Tp16, amplicons from 95 samples (73 cattle, 21 buffalo, and TpM) were sequenced (no duplicates, total $n = 95$). Sequence data were not generated for three cattle samples and after filtering, a further 12 cattle samples were lost, resulting in post-filtering data for 58 cattle samples, as well as all buffalo samples and TpM. A total of 4,215 alleles were initially identified, but after filtering (a sequence must be represented by two independent reads) the final allele count was 1,451 (175,395 reads). A single TpM allele was identified (3,948 reads) at the 1% SNP/INDEL variant frequency threshold used (**Figure 6C**). The median allele count in cattle samples was 3 (SD = 6.9), with a range of 1–9 alleles per cow (reads per allele 1–7,252; **Figure 7B**), excluding four outlying cattle samples with significantly higher reads (C25—42 alleles, 3,636 reads; C14—26 alleles, 1,868 reads; C60—24 alleles, 987 reads; C24—17 alleles, 6,139 reads). The median allele count in buffalo was 275.5 (SD = 80.8), with a range of 153–369 (reads per allele 438–1,822; **Figure 7B**), excluding one outlying buffalo sample with a significantly lower allele count (B20—3 alleles, 5,283 reads). Analysis of allele sharing identified 58 Tp16 alleles unique to cattle (including TpM, #0080), 1,389 alleles unique to buffalo, and four shared alleles (#0002 present in 16 cattle and 2 buffalo; #0014 in 12 cattle and 2 buffalo; #0016 in 2 cattle and 2 buffalo; #0052 in 1 cow and 3 buffalo).

Therefore, the Tp4 and Tp16 data showed the same trends as seen in Tp1, with significantly higher allele numbers observed in the buffalo samples, and limited evidence of sharing alleles between species. When data for all three antigens was integrated, for the 18 cattle and 9 buffalo samples that had data for all three antigens, it was evident that the cattle and buffalo samples represented largely non-overlapping populations (**Figure 8**), with greater distances between the buffalo samples, and the cattle samples mostly clustering very tightly together.

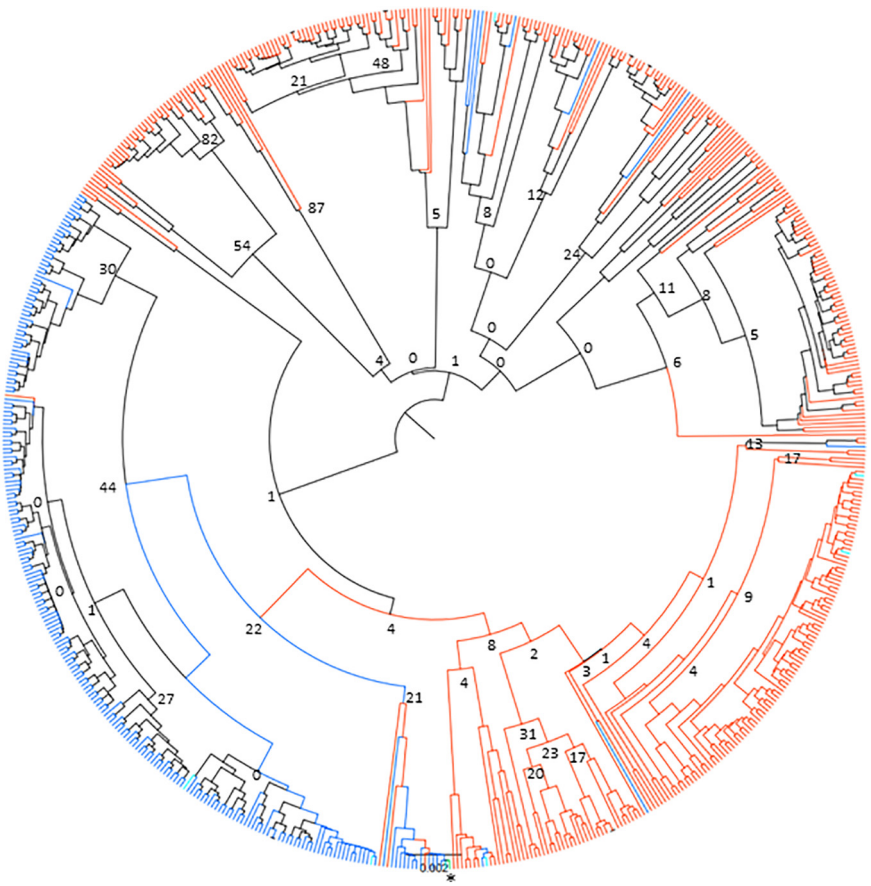


FIGURE 5 | Neighbor-joining phylogenetic tree of 651 Tp1 alleles found in cattle and buffalo samples. Alleles only found in cattle are shown in blue, alleles only found in buffalo are shown in orange and alleles present in both cattle and buffalo are shown in turquoise. The TpM reference sequence allele is highlighted in green and marked with an asterisk. Bootstrap values are shown at selected tree nodes.

DISCUSSION

This study aimed to assess *T. parva* genetic and antigenic diversity at the population level, in sympatric cattle and buffalo, by analyzing the sequence of near full-length amplicons of *T. parva* antigen genes Tp1, and Tp4 and Tp16. Samples deriving from cattle and buffalo in the same time and space enable the analysis of key unresolved questions around the epidemiology of *T. parva*—to what degree are parasites shared between buffalo and cattle, and can this inform on disease risk as well as potential utility of tools such as vaccination? A risk in analyzing *T. parva* from buffalo-endemic areas is the circulation of closely related *Theileria* species, which can confound results through cross-amplification. Significant care was taken to ensure that the PCR primers designed did not cross-amplify with other species, and particularly with the closely related *T.* sp. (buffalo) (Bishop et al., 2015), and so we are confident that the amplicons are *T. parva*-specific. For Tp1, at 1% threshold, a high level of diversity at the nucleotide level was observed, with 651 allele variants being identified across the 46 samples (34 cattle, 12 buffalo), and importantly only a single allele in the control clonal TpM sample. There was greater allelic diversity seen in buffalo, with a median of 81.5 alleles (SD = 45.2) being identified in buffalo compared to a median of nine alleles (SD = 21.8) in cattle samples. Additionally, there was overall a higher number of alleles detected in buffalo (n = 420) than in cattle (n = 239), despite there being much fewer buffalo samples. Only a small number of alleles (n = 8) were found in both species. These data suggest that the method described provides an increased resolution for detecting *T. parva* antigen alleles over those in previous studies. The results are in agreement with the findings of several previous studies that there is greater *T. parva* parasite heterogeneity in buffalo-derived populations compared to cattle-derived populations (Bishop et al., 1994; Oura et al., 2011a; Pelle et al., 2011; Kerario et al., 2019). The Tp1 alleles in the study population showed a high degree of relatedness, consistent with the samples all deriving from a largely co-circulating population. While there was limited support for genetic substructuring by host, time or location, there was evidence that there may be some incipient divergence between parasites deriving from cattle and buffalo, with one large group of buffalo-derived samples clustering together and being rarely found in cattle—in contrast to a second cluster in which multiple of both cattle- and buffalo-derived alleles were found.

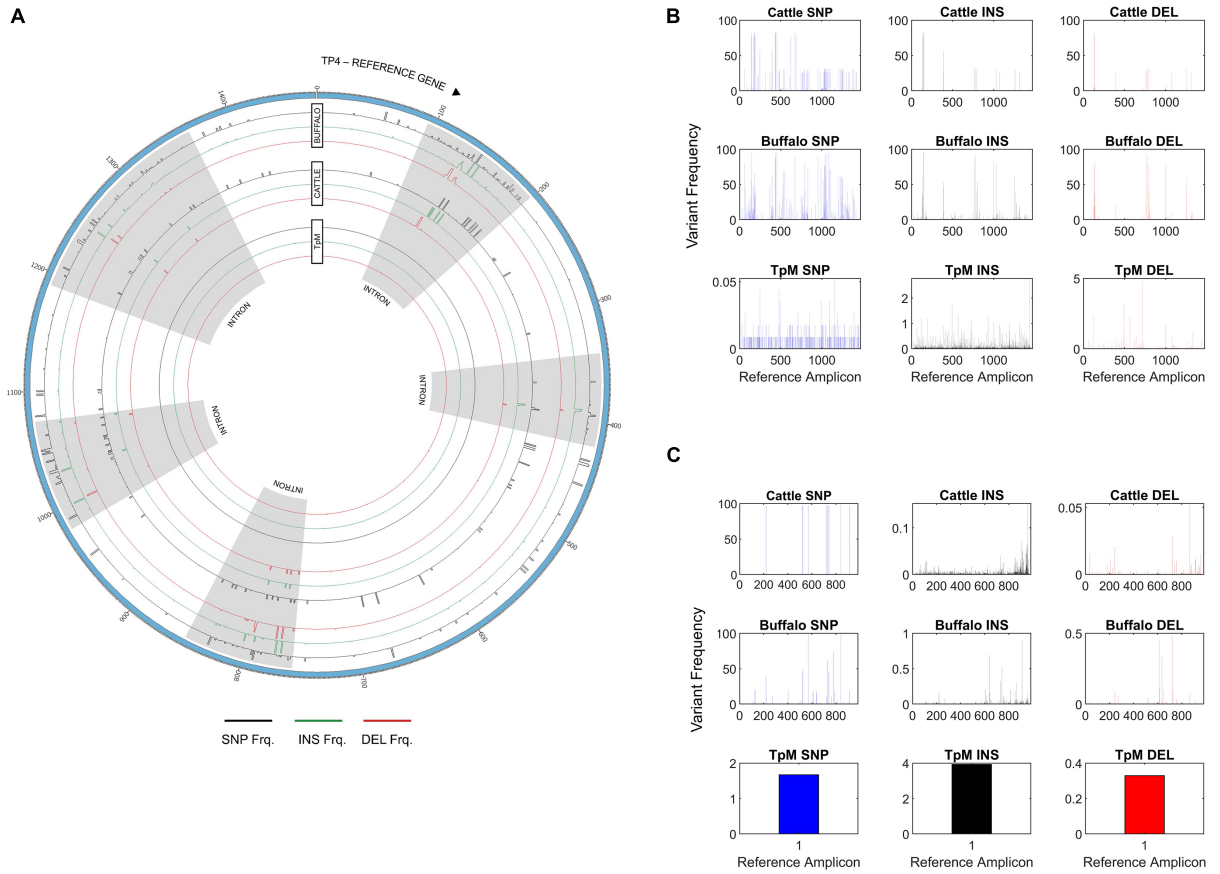


FIGURE 6 | (A) Single Nucleotide Polymorphism (SNP), insertion and deletion (INS/DEL) frequencies across the amplicon length of Tp4, with respect to the TpM reference sequence, highlighting increased mutation rates in intronic regions. SNPs are represented in black, insertions in green, and deletions in red. The outermost histogram represents buffalo, the middle represents cattle and the inner represents TpM. Intronic regions are highlighted in gray. **(B)** SNP/INDEL frequency for amplicon data across the length of the Tp4 amplicon, with respect to the TpM reference sequence. The mean calculated SNP/INDEL frequency is shown at each base location for the Tp4 gene. Blue bars represent SNPs, black bars represent insertions and red bars represent deletions. **(C)** SNP/INDEL frequency across the length of the Tp16 amplicon, with respect to the TpM reference sequence. The mean calculated SNP/INDEL frequency is shown at each base location for the Tp16 gene. Blue bars represent SNPs, black bars represent insertions and red bars represent deletions.

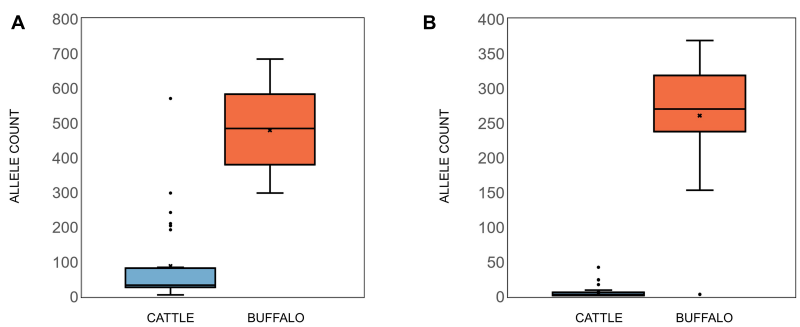


FIGURE 7 | (A) Tp4 and **(B)** Tp16 allele counts in cattle (blue) and buffalo (orange). Median (line), mean (x), and standard deviation (whiskers) are shown.

The analysis of Tp4 and Tp16 showed the same pattern of greater allelic diversity in buffalo compared to cattle, and with relatively few alleles shared between species. The difference in diversity was particularly marked for Tp16, which had very low allele counts in cattle-derived samples compared to buffalo (3 vs. 275 median number of alleles, respectively). This may reflect functional restriction on diversity in Tp16 when *T. parva* is in cattle compared to buffalo, but also may be due to the lower overall levels of polymorphism in Tp16, resulting in a cleaner

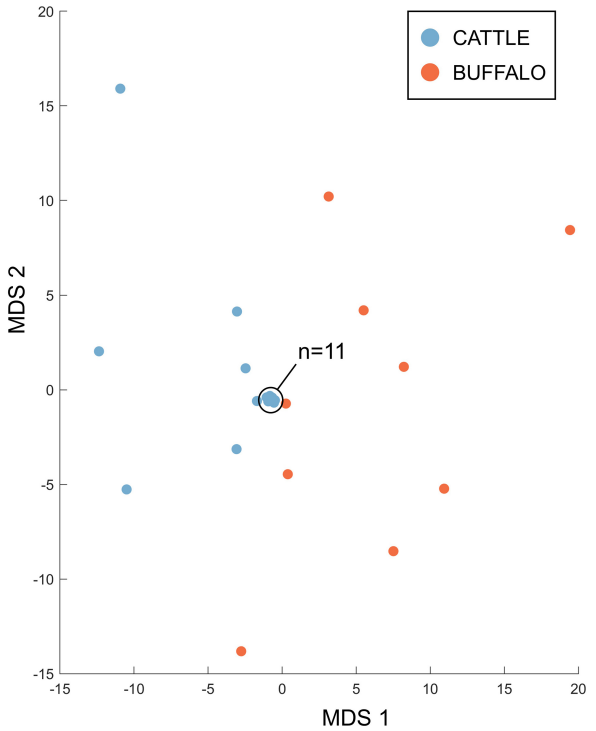


FIGURE 8 | Multidimensional scaling analysis incorporating counts for non-singleton alleles from samples for which data was generated for all three antigens. Data across the three antigens were integrated; the first two dimensions are shown. Cattle samples (n = 18) are represented in blue and buffalo samples (n = 9) in orange.

signature of differential polymorphism between cattle- and buffalo-derived parasites.

Of note, when a SNP/INDEL variant frequency threshold of 1% was used for Tp4, there were thousands of alleles detected, including multiple alleles for the TpM control sample derived from a clonal laboratory line. However, on closer examination this was as a result of multiple INDELS, which were found to correlate with homopolymer runs (more frequent in Tp4 than in Tp1 or Tp16), and thus likely to derive from PacBio error, as homopolymer stretches are an acknowledged issue with PacBio (Weirather et al., 2017). The SNP/INDEL frequency threshold was adjusted to the level (5%) at which only a single TpM allele was detected (**Figure 6B**), but this threshold will likely result in a conservative estimate of allele numbers. Additionally, analysis of Tp4 was complicated by the presence of five introns. The allele counts reported above were calculated based on exonic sequence only. Notably, the INDEL and SNP incidence was significantly higher in the intronic regions (**Figure 6A**). This provides indirect validation of the PacBio approach and the analytical pipeline used, as higher mutation rate would be expected in non-coding sections of DNA not under obvious functional restriction. It also provides some indication of the background mutation rate in these populations, with the average number of insertions, deletions and SNPs being 37, 24, and 75 and 13, 12, and 35 in buffalo-derived and cattle-derived samples (introns, 5% threshold), respectively, further emphasizing the reduced diversity in the cattle *T. parva* samples.

As outlined above, multiple alleles were detected in every animal, indicating a high level of multiplicity of infection. Hemmink et al. (2018) had previously observed an average of 16 Tp1 alleles in individual buffalo from Kruger, South Africa, and 14 in Ol Pejeta, Kenya—a total of 72 alleles from 14 animals, also using an amplicon sequencing approach, albeit amplifying a smaller (344 bp) region of Tp1. The number of alleles identified in buffalo in our study was much higher (median of 81.5), which may reflect the increased amplicon size of 1,618 bp, but also suggests that the PacBio approach and analytical pipeline results in greater resolution in terms of allele detection. While PacBio can introduce errors into the sequencing data, we included TpM amplified from *in vitro* cells as a control, and with our analysis only one allele was detected in TpM, as expected, giving confidence to the alleles detected in the field samples. Mixed genotype infections of *T. parva* are common in field samples, and have been suggested to derive from parasite diversity within the tick vector population (Elisa et al., 2015), which can be amplified through sexual recombination in the tick (Katzer et al., 2011; Henson et al., 2012). However, ticks usually have only very few infected salivary gland acini, and the sporozoites in each acinus normally originate from a single parasite (Gitau et al., 2000), which in turn suggests that extensive allelic diversity within individual animals is likely to be the result of cumulative infections through multiple and sequential tick bites. This is supported by previous studies of cattle in Uganda using satellite DNA markers, which showed an increase in number of alleles with age (Oura et al., 2005).

The allelic polymorphism observed in Tp1 was predominantly composed of synonymous SNPs, but did include non-synonymous SNPs, including five variants observed in the previously defined epitope that is recognized by protective CTL responses against Tp1 (MacHugh et al., 2009; Connelley et al., 2011); the epitope variant in the reference stock, TpM, was VGYPKVKEEML, as expected (Graham et al., 2008; Pelle et al., 2011; Elisa et al., 2015; Hemmink et al., 2018). Epitope variant VGYPKVKEEII was most frequently observed in cattle (31) and VGYPKVKEEML was most common in buffalo (all 12). Several other studies have demonstrated that -ML is the dominant epitope across East Africa, being the predominant allele detected in cattle in Kenya, Tanzania, Burundi, Democratic Republic of Congo, Malawi, Rwanda and Uganda (Pelle et al., 2011; Elisa et al., 2015; Amzati et al., 2019; Kerario et al., 2019; Atuhaire et al., 2020; Chatanga et al., 2020; Nanteza et al., 2020). Pelle et al. (2011) is the most comparable study to ours, having examined Tp1 diversity in parasites isolated from buffalo or from cattle with varying exposure to buffalo-derived *T. parva*. That study found variant –ML in the majority of their isolates in Kenya (6 of 17 cattle-derived laboratory isolates, 16/27 cattle-derived field isolates, 22/25 buffalo-associated cattle isolates [(i.e., from cattle co-grazed with buffalo) and 13/16 buffalo-derived isolates]. –II, the next most common variant in Pelle's dataset, suggested possible association with cattle (38% of cattle samples compared to 13% of buffalo-derived or associated samples), but our data indicated that –II was frequently found in both cattle and buffalo.

Of the rare variants in the Pelle dataset (-IL in one cattle-derived cell line and -MI in one buffalo-associated cattle cell line), -IL was found in both cattle (6) and buffalo (1), and -MI was only found in one cow in our samples. Kerario et al. (2019) also examined Tp1 alleles in cattle associated with buffalo in five areas of Tanzania, and found variant –ML in the majority of their samples (48/98 cattle, 17/19 buffalo-associated cattle and 6/13 buffalo-derived), and similar to Pelle et al. (2011), found the next most common variant to be –II (38/98 cattle, 1/19 buffalo-associated cattle and 2/13 buffalo-derived samples). Although in our dataset epitope variant –MV was only observed in buffalo, it was recently reported in a single cow from Mara region in northern Tanzania—the cow having had no obvious history of co-grazing with buffalo (Kerario et al., 2019). Therefore, our data add to previous reports and agree that epitope -ML is common in *T. parva* cattle samples. In our study, epitope -II is the most frequent in cattle, but this is also commonly found in other populations, and without further work it is difficult to conclude if this is meaningful in terms of relating to epidemiological differences. We did detect all five known Tp1 epitopes in the sample set, suggesting that this is a genetically diverse *T. parva* population—probably reflecting the large resident population of both buffalo and cattle, and the lack of tick control impact on the tick population in protected areas. However, it may also reflect the improved resolution of the protocol employed, suggesting this may prove advantageous in future studies aiming to define *T. parva* genetic diversity.

While there was limited evidence for genetic substructuring in our data, and the presence of alleles shared between cattle and buffalo does indicate that there is (probably infrequent) mixing of parasite populations, it may be that the occasional transmission between cattle and buffalo, and genetic exchange in ticks, is sufficient to blur the signature of any genetic drift that may be happening between subpopulations. It may also simply reflect that the diverse *T. parva* population evolved in buffalo before the arrival of cattle in Africa, and that the cattle-maintained *T. parva* parasite population is a subset of this ancient buffalo-derived population. However, the indications of separate clusters of genotypes that derive from either only buffalo or from both cattle and buffalo, alternatively suggest that this may be an example of incipient speciation, with divergence between cattle and buffalo-derived samples in process, i.e., comparable to but at an earlier stage of divergence than the very closely related *T.* sp. (buffalo) (Palmateer et al., 2020)—albeit that in *T.* sp. (buffalo) this is suggested to be due to adaptation to a different tick species rather than a different mammalian host (Bishop et al., 2015).

There were a variety of potential factors that could have influenced potential substructuring of the *T. parva* population, and we did collect sample metadata that included host species, date of sampling, place of sampling and management practice. However, for each antigen the number of samples resulting in successful generation of sequence data were small for each category, meaning the data was under-powered for analysis of any genetic association. Although spatiotemporal analysis was not formally carried out, given the close relatedness of the sequences overall there was not an evident trend in allele distribution across space or time in the data.

It is well established that buffalo-derived *T. parva* do not differentiate well into the piroplasm state in cattle (Young et al., 1977; Sitt et al., 2015), and most attempts to transmit infection from cattle infected with buffalo-derived *T. parva* have been unsuccessful (reviewed in Morrison et al., 2020). Since the cattle samples in this study were almost all from healthy animals, the *T. parva* detected predominantly represent the carrier state cattle-maintained parasite population (i.e., *T. parva* that is able to differentiate to transmissible infections in cattle). The inability of buffalo parasites to undergo differentiation to the piroplasm stage in cattle is believed to result in a barrier to maintenance of buffalo-derived parasites in the cattle population, thus accounting for less genotypic diversity in the cattle *T. parva*. However, in a few instances experimental feeding of large numbers of ticks on cattle experimentally infected with buffalo parasites has resulted in low level transmission and further passage by ticks in cattle, yielding parasites that behave similarly to cattle *T. parva* (Barnett and Brocklesby, 1966; Young and Purnell, 1973; Maritim et al., 1992). Whether or not such adaptation occurs naturally at a sufficient level to allow the parasites to establish in cattle is unclear. Previous comparisons of the genotypes of buffalo and cattle parasites have focused mainly on cattle parasites from buffalo-free areas. Therefore, it was of interest in the current study, where cattle are exposed to buffalo parasites, to determine whether there is greater diversity in the cattle *T. parva* population, possibly representing introgression of buffalo *T. parva* genotypes into the cattle parasite population. However, the cattle parasite population detected showed less diversity compared to the buffalo population and the genotypes were largely different in the two host species, with most antigen alleles unique to one or other host species and only comparatively few shared. While there has been considerable focus on the transmission of buffalo-derived *T. parva* to cattle, there have been very few studies to examine infection of buffalo with cattle-derived *T. parva*. Our data suggest that, although very few alleles are directly shared, one clade of Tp1 alleles is found regularly in both cattle and buffalo, indicating that there is some degree of transmission between buffalo and cattle within this clade.

The level of interaction between cattle and buffalo, and the ticks that feed on them, in the study area is currently not quantifiable. The limited number of alleles found in both hosts likely reflects that acute infections of cattle with buffalo-derived parasites could not be examined in this study, and this highlights the need for further studies to examine parasites in clinical cases of *T. parva*, in order to assess the impact of buffalo-derived parasites upon clinical disease burden in this epidemiological setting. The national park boundary is unfenced, and buffalo (and ticks) are free to move into suitable habitat outside the national park, although fragmentation of habitat and increasing human population may mean that this is occurring less frequently. Discussion with farmers in the study area did reveal that grazing of cattle within the protected areas does occur (particularly seasonally when water supplies are limited) providing opportunity for ticks to feed on both cattle and buffalo. In order to fully understand the parasite dynamics across this wildlife-livestock interface, it would also therefore be necessary to examine the parasites circulating in the tick population, to

establish the parasite genotypes they carry, as well as which hosts they are feeding on.

Only a relatively low proportion (7.87%) of the 1,602 cattle samples tested were positive for *T. parva*, and only some of these resulted in successful amplification of Tp1, Tp4, and Tp16 (likely due to the samples representing low parasitaemic carrier state, as well as the relative efficiency of the PCR assays). The limited number of sympatric buffalo samples reflected the difficulty in obtaining such samples. A further limitation of the current study is that by focusing on antigens recognized by the bovine immune system, we may be analyzing genes that are under particular selection pressures, which has the potential to confound other genetic signals, such as those associated with population substructuring by host. In order to more fully understand the population structure, ideally a larger sample set of both *T. parva*-positive cattle and buffalo would be required, and an approach such as sequence capture used to enable analysis of polymorphism at the whole genome level (Palmateer et al., 2020). Analysis of whole genome data would facilitate unbiased and thorough analysis of genetic diversity in cattle and buffalo-derived *T. parva*, and enable identification of signatures of selection that may underpin adaptation to differentiation to piroplasms in cattle, for example. A greater understanding of the pattern of such diversity across the geographic range of *T. parva*, in particular focusing on the remaining areas where cattle and buffalo interact, would be particularly enlightening in terms of the overall population structure, and potentially how much a role buffalo-derived *T. parva* play in disease epidemiology across the parasite's range.

The shape of the parasite population in cattle and buffalo has important implications for use of the ITM vaccine. As has been observed previously, cattle vaccinated with ITM are unlikely to be protected against heavy challenge with buffalo-derived parasites (Sitt et al., 2015). Therefore, the increased allelic diversity in buffalo in our study site, and in particular the clade of Tp1 alleles only found in buffalo, presumably reflects a substantial pathogen diversity that cattle may be exposed to if fed on by infected ticks. How effective the ITM vaccine would be in this area may depend on the extent of exposure to buffalo-derived parasites, which as outlined above could not be captured with the current sample set. Additionally, frequent tick challenge resulting in exposure to multiple cattle-derived *T. parva* genotypes may be sufficient to confer broad protection prior to encountering buffalo-derived infection, and this also is an aspect that could not be inferred from our data.

T. parva is an interesting example where there has been a relatively recent species jump by the pathogen, and provides a prime case study for adaptation of a complex eukaryote pathogen to a new host. This study established and validated a novel genotyping pipeline to analyze genetic and antigenic diversity of *T. parva* by long read sequencing, allowing analysis of near full-length sequences from three polymorphic antigen genes, importantly in samples derived from sympatric cattle and African buffalo. These data shed light on the complex interplay between *T. parva* populations in buffalo and cattle, revealing the extent of the significant genetic diversity in the buffalo *T. parva* population, the limited sharing of parasite genotypes between the host species, and highlighting that a subpopulation of *T. parva* is maintained by transmission within cattle. The results emphasize the importance of obtaining a fuller understanding of buffalo *T. parva* population dynamics in particular, as ultimately a more holistic understanding of the population genetics of *T. parva* populations will enable a realistic assessment of the extent of buffalo-derived infection risk in cattle, and how this may impact upon current control measures such as vaccination.

DATA AVAILABILITY STATEMENT

Data presented in this study and scripts used to analyse the data are available via the following link: https:/doi.org/10.7488/ds/3003.

AUTHOR CONTRIBUTIONS

FA, SJ, WM, HA, and LM designed the study. FA, EP, ES, TK, RF, FM, ST, TL, JH, PT, and HA carried out the field work and generated the samples and data. FA, SJ, IH, HA, WM, and LM undertook analysis and interpretation. All authors wrote the manuscript.

FUNDING

This work was funded by a Biotechnology and Biological Sciences Research Council (BBSRC, United Kingdom) KTN/iCASE Ph.D. studentship awarded to FA, in partnership with the Global Alliance for Livestock Veterinary Medicines (GALVmed) with funding from the Bill and Melinda Gates Foundation and UK Government. ST, LM, HA, FM, and EP were supported by a grant from the Zoonoses and Emerging Livestock Systems (ZELS) programme (BB/L019035/1). LM, WM, EP, FA, SJ, and the Roslin Institute were also supported by a core grant from the BBSRC (BBS/E/D/20231762; BBS/E/D/ 20002173).

ACKNOWLEDGMENTS

We would like to thank Geofrey Mbata, Ahmed Lugelo, Deogratius Mshanga, and Raphael Mahemba Shabani for cattle sampling, and staff at the Nelson Mandela African Institution of Science and Technology (NM-AIST) for providing access to laboratory facilities for some of the sample processing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.684127/full#supplementary-material

REFERENCES

Amzati, G. S., Djikeng, A., Odongo, D. O., Nimpaye, H., Sibeko, K. P., Muhigwa, J. B., et al. (2019). Genetic and antigenic variation of the bovine tick-borne pathogen *Theileria parva* in the Great Lakes region of Central Africa. *Parasit. Vectors* 12:588.

Atuhaire, D. K., Muleya, W., Mbao, V., Bazarusanga, T., Gafarasi, I., Salt, J., et al. (2020). Sequence diversity of cytotoxic T cell antigens and satellite marker analysis of *Theileria parva* informs the immunization against East Coast fever in Rwanda. *Parasit. Vectors* 13:452.

Baldwin, C. L., Malu, M. N., and Grootenhuis, J. G. (1988). Evaluation of cytotoxic lymphocytes and their parasite strain specificity from African buffalo infected with *Theileria parva*. *Parasite Immunol.* 10, 393–403. doi: 10.1111/j.1365-3024.1988.tb00229.x

Barnett, S. F. (1957). Theileriasis control. *Bull. Epizoot. Dis. Afr.* 5, 343–357.

Barnett, S. F., and Brocklesby, D. W. (1966). The passage of *Theileria* Lawrencei (Kenya) through cattle. *Br. Vet. J.* 122, 396–409. doi: 10.1016/s0007-1935(17)40406-4

Bishop, R. P., Hemmink, J. D., Morrison, W. I., Weir, W., Toye, P. G., Sitt, T., et al. (2015). The African buffalo parasite *Theileria*. sp. (buffalo) can infect and immortalize cattle leukocytes and encodes divergent orthologues of *Theileria parva* antigen genes. *Int. J. Parasitol. Parasites Wildl.* 4, 333–342. doi: 10.1016/j.ijppaw.2015.08.006

Bishop, R. P., Spooner, P. R., Kanhai, G. K., Kiarie, J., Latif, A. A., Hove, T., et al. (1994). Molecular characterization of *Theileria* parasites: application to the epidemiology of theileriosis in Zimbabwe. *Parasitology* 109, 573–581. doi: 10.1017/s0031182000076459

Brocklesby, D. W. (1961). Morbidity and mortality rates in East Coast Fever (*Theileria parva* infection) and their application to drug screening procedures. *Br. Vet. J.* 117, 529–532. doi: 10.1016/s0007-1935(17)43303-3

Burridge, M. J., Morzaria, S. P., Cunningham, M. P., and Brown, C. G. (1972). Duration of immunity to East Coast fever (*Theileria parva* infection of cattle). *Parasitology* 64, 511–515. doi: 10.1017/s0031182000045571

Casey, M. B., Lembo, T., Knowles, N. J., Fyumagwa, R., Kivaria, F., Maliti, H., et al. (2014). "Patterns of Foot-and-Mouth Disease Virus Distribution in Africa: the Role of Livestock and Wildlife in Virus Emergence," in *The Role of Animals in Emerging Viral Diseases*, ed. N. Johnson (Cambridge: Academic Press), 21–38.

Casey-Bryars, M., Reeve, R., Bastola, U., Knowles, N. J., Auty, H., Bachanek-Bankowska, K., et al. (2018). Waves of endemic foot-and-mouth disease in eastern Africa suggest feasibility of proactive vaccination approaches. *Nat. Ecol. Evol.* 2, 1449–1457. doi: 10.1038/s41559-018-0636-x

Chatanga, E., Hayashida, K., Muleya, W., Kusakisako, K., Moustafa, M. A. M., Salim, B., et al. (2020). Genetic Diversity and Sequence Polymorphism of Two Genes Encoding *Theileria parva* Antigens Recognized by CD8(+) T Cells among Vaccinated and Unvaccinated Cattle in Malawi. *Pathogens* 9:334. doi: 10.3390/pathogens9050334

Connelley, T. K., MacHugh, N. D., Pelle, R., Weir, W., and Morrison, W. I. (2011). Escape from CD8+ T cell response by natural variants of an immunodominant epitope from *Theileria parva* is predominantly due to loss of TCR recognition. *J. Immunol.* 187, 5910–5920. doi: 10.4049/jimmunol.1102009

Conrad, P. A., ole-MoiYoi, O. K., Baldwin, C. L., Dolan, T. T., O'Callaghan, C. J., Njamunggeh, R. E., et al. (1989). Characterization of buffalo-derived theilerial parasites with monoclonal antibodies and DNA probes. *Parasitology* 98, 179–188. doi: 10.1017/s0031182000062089

Di Giulio, G., Lynen, G., Morzaria, S., Oura, C., and Bishop, R. (2009). Live immunization against East Coast fever–current status. *Trends Parasitol.* 25, 85–92. doi: 10.1016/j.pt.2008.11.007

Elisa, M., Hasan, S. D., Moses, N., Elpidius, R., Skilton, R., and Gwakisa, P. (2015). Genetic and antigenic diversity of *Theileria parva* in cattle in Eastern and Southern zones of Tanzania. A study to support control of East Coast fever. *Parasitology* 142, 698–705. doi: 10.1017/s0031182014001784

Estes, A. B., Kuemmerle, T., Kushnir, H., Radeloff, V. C., and Shugart, H. H. (2012). Land-cover change and human population trends in the greater Serengeti ecosystem from 1984-2003. *Biol. Conserv.* 14, 255–263. doi: 10.1016/j.biocon.2012.01.010

GALVmed. (2019). *East Coast Fever*. https://www.galvmed.org/livestock-and-diseases/livestock-diseases/east-coast-fever/ (accessed March, 2021).

Gitau, G. K., McDermott, J. J., Katende, J. M., O'Callaghan, C. J., Brown, R. N., and Perry, B. D. (2000). Differences in the epidemiology of theileriosis on smallholder dairy farms in contrasting agro-ecological and grazing strata of highland Kenya. *Epidemiol. Infect.* 124, 325–335. doi: 10.1017/s0950268800003526

Graham, S. P., Pelle, R., Honda, Y., Mwangi, D. M., Tonukari, N. J., Yamage, M., et al. (2006). *Theileria parva* candidate vaccine antigens recognized by immune bovine cytotoxic T lymphocytes. *Proc. Natl. Acad. Sci. U. S. A.* 103, 3286–3291.

Graham, S. P., Pelle, R., Yamage, M., Mwangi, D. M., Honda, Y., Mwakubambanya, R. S., et al. (2008). Characterization of the fine specificity of bovine CD8 T-cell responses to defined antigens from the protozoan parasite *Theileria parva*. *Infect. Immun.* 76, 685–694. doi: 10.1128/iai.01244-07

Hanotte, O., Bradley, D. G., Ochieng, J. W., Verjee, Y., Hill, E. W., and Rege, J. E. (2002). African pastoralism: genetic imprints of origins and migrations. *Science* 296, 336–339. doi: 10.1126/science.1069878

Hayashida, K., Abe, T., Weir, W., Nakao, R., Ito, K., Kajino, K., et al. (2013). Whole-genome sequencing of *Theileria parva* strains provides insight into parasite migration and diversification in the African continent. *DNA Res.* 20, 209–220. doi: 10.1093/dnares/dst003

Hemmink, J. D., Sitt, T., Pelle, R., de Klerk-Lorist, L. M., Shiels, B., Toye, P. G., et al. (2018). Ancient diversity and geographical sub-structuring in African buffalo *Theileria parva* populations revealed through metagenetic analysis of antigen-encoding loci. *Int. J. Parasitol.* 48, 287–296. doi: 10.1016/j.ijpara.2017.10.006

Henson, S., Bishop, R. P., Morzaria, S., Spooner, P. R., Pelle, R., Poveda, L., et al. (2012). High-resolution genotyping and mapping of recombination and gene conversion in the protozoan *Theileria parva* using whole genome sequencing. *BMC Genomics* 13:503. doi: 10.1186/1471-2164-13-503

Katzer, F., Lizundia, R., Ngugi, D., Blake, D., and McKeever, D. (2011). Construction of a genetic map for *Theileria parva*: identification of hotspots of recombination. *Int. J. Parasitol.* 41, 669–675. doi: 10.1016/j.ijpara.2011.01.001

Kerario, I. I., Chenyambuga, S. W., Mwega, E. D., Rukambile, E., Simulundu, E., and Simuunza, M. C. (2019). Diversity of two *Theileria parva* CD8+ antigens in cattle and buffalo-derived parasites in Tanzania. *Ticks Tick Borne Dis.* 10, 1003–1017. doi: 10.1016/j.ttbdis.2019.05.007

Kim, J., Hanotte, O., Mwai, O. A., Dessie, T., Bashir, S., Diallo, B., et al. (2017). The genome landscape of indigenous African cattle. *Genome Biol.* 18:34.

MacHugh, N. D., Connelley, T., Graham, S. P., Pelle, R., Formisano, P., Taracha, E. L., et al. (2009). CD8+ T-cell responses to *Theileria parva* are preferentially directed to a single dominant antigen: implications for parasite strain-specific immunity. *Eur. J. Immunol.* 39, 2459–2469. doi: 10.1002/eji.200939227

Maritim, A. C., Young, A. S., Lesan, A. C., Ndungu, S. G., Stagg, D. A., and Ngumi, P. N. (1992). Transformation of *Theileria parva* derived from African buffalo (*Syncerus caffer*) by tick passage in cattle and its use in infection and treatment immunization. *Vet. Parasitol.* 43, 1–14. doi: 10.1016/0304-4017(92)90043-9

Martins, S. B., Di Giulio, G., Lynen, G., Peters, A., and Rushton, J. (2010). Assessing the impact of East Coast Fever immunisation by the infection and treatment method in Tanzanian pastoralist systems. *Prev. Vet. Med.* 97, 175–182. doi: 10.1016/j.prevetmed.2010.09.018

McKeever, D. J., and Morrison, W. I. (1990). *Theileria parva*: the nature of the immune response and its significance for immunoprophylaxis. *Rev. Sci. Tech.* 9, 405–421. doi: 10.20506/rst.9.2.504

Morrison, W. I., Hemmink, J. D., and Toye, P. G. (2020). *Theileria parva*: a parasite of African buffalo, which has adapted to infect and undergo transmission in cattle. *Int. J. Parasitol.* 50, 403–412. doi: 10.1016/j.ijpara.2019.12.006

Morrison, W. I., and McKeever, D. J. (1998). Immunology of infections with *Theileria parva* in cattle. *Chem. Immunol.* 70, 163–185. doi: 10.1159/000058705

Morrison, W. I., and McKeever, D. J. (2006). Current status of vaccine development against *Theileria* parasites. *Parasitology* 133, S169–S187.

Morzaria, S. P., Irvin, A. D., Voigt, W. P., and Taracha, E. L. (1987). Effect of timing and intensity of challenge following immunization against East Coast fever. *Vet. Parasitol.* 26, 29–41. doi: 10.1016/0304-4017(87)90074-4

Nanteza, A., Obara, I., Kasaija, P., Mwega, E., Kabi, F., Salih, D. A., et al. (2020). Antigen gene and variable number tandem repeat (VNTR) diversity in *Theileria parva* parasites from Ankole cattle in south-western Uganda:

evidence for conservation in antigen gene sequences combined with extensive polymorphism at VNTR loci. *Transbound. Emerg. Dis.* 67, 99–107. doi: 10.1111/tbed.13311

Odongo, D. O., Sunter, J. D., Kiara, H. K., Skilton, R. A., and Bishop, R. P. (2010). A nested PCR assay exhibits enhanced sensitivity for detection of *Theileria parva* infections in bovine blood samples from carrier animals. *Parasitol. Res.* 106, 357–365. doi: 10.1007/s00436-009-1670-z

Odongo, D. O., Ueti, M. W., Mwaura, S. N., Knowles, D. P., Bishop, R. P., and Scoles, G. A. (2009). Quantification of *Theileria parva* in *Rhipicephalus appendiculatus* (*Acari: ixodidae*) confirms differences in infection between selected tick strains. *J. Med. Entomol.* 46, 888–894. doi: 10.1603/033.046.0422

Oura, C. A., Asiimwe, B. B., Weir, W., Lubega, G. W., and Tait, A. (2005). Population genetic analysis and sub-structuring of *Theileria parva* in Uganda. *Mol. Biochem. Parasitol.* 140, 229–239. doi: 10.1016/j.molbiopara.2004.12.015

Oura, C. A., Tait, A., Asiimwe, B., Lubega, G. W., and Weir, W. (2011a). Haemoparasite prevalence and *Theileria parva* strain diversity in Cape buffalo (*Syncerus caffer*) in Uganda. *Vet. Parasitol.* 175, 212–219. doi: 10.1016/j.vetpar.2010.10.032

Oura, C. A., Tait, A., Asiimwe, B., Lubega, G. W., and Weir, W. (2011b). *Theileria parva* genetic diversity and haemoparasite prevalence in cattle and wildlife in and around Lake Mburo National Park in Uganda. *Parasitol. Res.* 108, 1365–1374. doi: 10.1007/s00436-010-2030-8

Paling, R. W., Mpangala, C., Luttikhuizen, B., and Sibomana, G. (1991). Exposure of Ankole and crossbred cattle to theileriosis in Rwanda. *Trop. Anim. Health Prod.* 23, 203–214. doi: 10.1007/bf02357101

Palmateer, N. C., Tretina, K., Orvis, J., Ifeonu, O. O., Crabtree, J., Drabék, E., et al. (2020). Capture-based enrichment of *Theileria parva* DNA enables full genome assembly of first buffalo-derived strain and reveals exceptional intra-specific genetic diversity. *PLoS Negl. Trop. Dis.* 14:e0008781. doi: 10.1371/journal.pntd.0008781

Pelle, R., Graham, S. P., Njahira, M. N., Osaso, J., Saya, R. M., Odongo, D. O., et al. (2011). Two *Theileria parva* CD8 T cell antigen genes are more variable in buffalo than cattle parasites, but differ in pattern of sequence diversity. *PLoS One* 6:e19015. doi: 10.1371/journal.pone.0019015

Radley, D. E. (1981). "Infection and treatment method of immunisation against theileriosis," in *Advances in the Control of Theileriosis: Proceedings of an International Conference Held at the International Laboratory for Research on Animal Diseases, Nairobi, 9-13 February 1981*, eds A.D. Irvin, M.P. Cunningham, A.S. Young. (Leiden: Martinus Nijhoff Publishers), 227–237. doi: 10.1007/978-94-009-8346-5_41

Radley, D. E., Brown, C. G. D., Burridge, M. J., Cunningham, M. P., Kirimi, I. M., Purnell, R. E., et al. (1975a). East Coast Fever: 1. Chemoprophylactic Immunisation of cattle against *Theileria parva* (Muguga) and five theilerial strains. *Vet. Parasitol.* 1, 35–41. doi: 10.1016/0304-4017(75)90005-9

Radley, D. E., Brown, C. G. D., Cunningham, M. P., Kimber, C. D., Musisi, F. L., Payne, R. C., et al. (1975b). East Coast Fever: 3. Chemoprophylactic immunization of cattle using oxytetracycline and a combination of theilerial strains. *Vet. Parasitol.* 1, 51–60. doi: 10.1016/0304-4017(75)90007-2

Radley, D. E., Young, A. S., Brown, C. G. D., Burridge, M. J., Cunningham, M. P., Musisi, F. L., et al. (1975c). East Coast Fever: 2. Cross-immunity trials with a Kenya strain of *Theileria* Lawrencei. *Vet. Parasitol.* 1, 43–50. doi: 10.1016/0304-4017(75)90006-0

Radley, D. E., Young, A. S., Grootenhuis, J. G., Cunningham, M. P., Dolan, T. T., and Morzaria, S. P. (1979). Further-Studies on the Immunization of Cattle against Theileria-Lawrencei by Infection and Chemoprophylaxis. *Vet. Parasitol.* 5, 117–128. doi: 10.1016/0304-4017(79)90003-7

Schreuder, B. E., Uilenberg, G., and Tondeur, W. (1977). Studies on Theileriidae (Sporozoa) in Tanzania. VIII. Experiments with African buffalo (*Syncerus caffer*). *Tropenmed. Parasitol.* 28, 367–371.

Sitt, T., Poole, E. J., Ndambuki, G., Mwaura, S., Njoroge, T., Omondi, G. P., et al. (2015). Exposure of vaccinated and naive cattle to natural challenge from buffalo-derived *Theileria parva*. *Int. J. Parasitol. Parasites Wildl.* 4, 244–251. doi: 10.1016/j.ijppaw.2015.04.006

Skilton, R. A., Bishop, R. P., Katende, J. M., Mwaura, S., and Morzaria, S. P. (2002). The persistence of *Theileria parva* infection in cattle immunized using two stocks which differ in their ability to induce a carrier state: analysis using a novel blood spot PCR assay. *Parasitology* 124, 265–276. doi: 10.1017/s0031182001001196

Taracha, E. L., Goddeeris, B. M., Morzaria, S. P., and Morrison, W. I. (1995a). Parasite strain specificity of precursor cytotoxic T cells in individual animals correlates with cross-protection in cattle challenged with *Theileria parva*. *Infect. Immun.* 63, 1258–1262. doi: 10.1128/iai.63.4.1258-1262.1995

Taracha, E. L., Goddeeris, B. M., Teale, A. J., Kemp, S. J., and Morrison, W. I. (1995b). Parasite strain specificity of bovine cytotoxic T cell responses to *Theileria parva* is determined primarily by immunodominance. *J. Immunol.* 155, 4854–4860.

TAWIRI. (2016). *Tanzania Wildlife Research Institute, 2016, Wildlife, Livestock and Bomas census in the Serengeti Ecosystem, Dry Season, 2016. TAWIRI Aerial Survey Report.: 35*. Tanzania: Tanzania Wildlife Research Institute.

Tretina, K., Pelle, R., Orvis, J., Gotia, H. T., Ifeonu, O. O., Kumari, P., et al. (2020). Re-annotation of the *Theileria parva* genome refines 53% of the proteome and uncovers essential components of N-glycosylation, a conserved pathway in many organisms. *BMC Genomics* 21:279.

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X. J., et al. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res* 6:100. doi: 10.12688/f1000research.10571.1

Young, A. S., Leitch, B. L., Newson, R. M., and Cunningham, M. P. (1986). Maintenance of *Theileria parva* parva infection in an endemic area of Kenya. *Parasitology* 93, 9–16. doi: 10.1017/s0031182000049787

Young, A. S., and Purnell, R. E. (1973). Transmission of Theileria lawrencei (Serengeti) by the ixodid tick, *Rhipicephalus appendiculatus*. *Trop. Anim. Health Prod.* 5, 146–152. doi: 10.1007/bf02251383

Young, A. S., Radley, D. E., Cunningham, M. P., Musisi, F. L., Payne, R. C., and Purnell, R. E. (1977). Exposure of immunised cattle to prolonged natural challenge of Theileria lawrencei derived from buffalo (*Syncerus caffer*). *Vet. Parasitol.* 3, 283–290. doi: 10.1016/0304-4017(77)90015-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 August 2021
doi: 10.3389/fgene.2021.688606

Genetic Diversity Analysis of Surface-Related Antigen (SRA) in *Plasmodium falciparum* Imported From Africa to China

Bo Yang[1], Hong Liu[2], Qin-Wen Xu[1], Yi-Fan Sun[1], Sui Xu[3], Hao Zhang[1], Jian-Xia Tang[2,3], Guo-Ding Zhu[2,3], Yao-Bao Liu[2,3], Jun Cao[1,2,3] and Yang Cheng[1]**

[1] *Laboratory of Pathogen Infection and Immunity, Department of Public Health and Preventive Medicine, Wuxi School of Medicine, Jiangnan University, Wuxi, China,* [2] *Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China,* [3] *Key Laboratory of National Health and Family Planning Commission on Parasitic Disease Control and Prevention, Jiangsu Provincial Key Laboratory on Parasite and Vector Control Technology, Jiangsu Institute of Parasite Diseases, Wuxi, China*

OPEN ACCESS

Edited by:
Matthew Adekunle Adeleke, University of KwaZulu-Natal, South Africa

Reviewed by:
Francis Ntumngia, University of South Florida, United States
Ikhide G. Imumorin, Georgia Institute of Technology, United States

***Correspondence:**
Jun Cao
caojuncn@hotmail.com
Yang Cheng
woerseng@126.com

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: *31 March 2021*
Accepted: *01 July 2021*
Published: *05 August 2021*

Citation:
Yang B, Liu H, Xu Q-W, Sun Y-F, Xu S, Zhang H, Tang J-X, Zhu G-D, Liu Y-B, Cao J and Cheng Y (2021) Genetic Diversity Analysis of Surface-Related Antigen (SRA) in Plasmodium falciparum Imported From Africa to China. Front. Genet. 12:688606. doi: 10.3389/fgene.2021.688606

Plasmodium falciparum surface-related antigen (SRA) is located on the surfaces of gametocyte and merozoite and has the structural and functional characteristics of potential targets for multistage vaccine development. However, little information is available regarding the genetic polymorphism of *pfsra*. To determine the extent of genetic variation about *P. falciparum* by characterizing the *sra* sequence, 74 *P. falciparum* samples were collected from migrant workers who returned to China from 12 countries of Africa between 2015 and 2019. The full length of the *sra* gene was amplified and sequenced. The average pairwise nucleotide diversities (π) of *P. falciparum sra* gene was 0.00132, and the haplotype diversity (*Hd*) was 0.770. The average number of nucleotide differences (*k*) for *pfsra* was 3.049. The ratio of non-synonymous (*dN*) to synonymous (*dS*) substitutions across sites (*dN*/*dS*) was 1.365. Amino acid substitutions of *P. falciparum* SRA could be categorized into 35 unique amino acid variants. Neutrality tests showed that the polymorphism of PfSRA was maintained by positive diversifying selection, which indicated its role as a potential target of protective immune responses and a vaccine candidate. Overall, the ability of the N-terminal of PfSRA antibodies to evoke inhibition of merozoite invasion of erythrocytes and conserved amino acid at low genetic diversity suggest that the N-terminal of PfSRA could be evaluated as a vaccine candidate against *P. falciparum* infection.

Keywords: *Plasmodium falciparum*, PfSRA, genetic diversity, imported malaria cases, DNA sequencing, Jiangsu Province

INTRODUCTION

Malaria has been a major global health concern of humans throughout history and is a leading cause of disease and death across many tropical and subtropical countries. In 2019, an estimated 229 million malaria cases and 409,000 malaria-caused deaths globally were reported (WHO, 2020). Among the five species of *Plasmodium* that infect humans, *P. falciparum* infection causes the highest mortality and morbidity and the most serious clinical symptoms (Buffet et al., 2011).

The resurgence and spread of antimalarial drug resistance (Ménard et al., 2015; WWARN, 2015) along with vector resistance to insecticides (Dhiman and Veer, 2014; Strode et al., 2014) have the potential to reduce the impact of existing malaria control strategies and make vaccines a public health priority. Although extensive studies have been conducted on several blood-stage antigens, few have shown the quality required for a candidate vaccine. In a systematic screen of uncharacterized *P. falciparum* proteins for potential blood-stage vaccine candidates, using data from transcriptome studies of *P. falciparum*, data-mining analysis of the genes with peak mRNA expression levels in late schizogony was performed (Bozdech et al., 2003; Le Roch et al., 2003) and another study on the prediction of PfSUB-1 protease specificity (Gilson et al., 2006). The results showed that *P. falciparum* surface-related antigen (PfSRA) emerged as the top hit with both signal peptide and a predicted glycosylphosphatidylinositol (GPI) attachment site. PfSRA is localized on the surfaces of both gametocytes and merozoites. The processed 32-kDa PfSRA protein fragment binds normal human erythrocytes. Immunoepidemiological studies in malaria-infected populations suggest the presence of naturally acquired protective antibodies against PfSRA. Parasite growth inhibition assays indicated that the antibodies against PfSRA could potently inhibit the invasion of merozoite on erythrocytes. Overall, the structural and functional characteristics of PfSRA indicate that it would be a promising vaccine target (Amlabu et al., 2018).

The low protective efficacy of vaccines against clinical malaria has been in part limited by extensive genetic diversity, which enables parasites to evade human immune responses and may lead to vaccine failure (Takala et al., 2002). However, the evidence for *pfsra* genetic diversity is limited. Therefore, it is necessary to study the characteristics of *pfsra* toward finding suitable vaccine candidates and understanding its population genetic structure. Accordingly, this study analyzed the full-length sequence of *sra* from the *P. falciparum* collected from infected migrant workers returning to the Jiangsu Province from Africa. We determined the nucleotide divergence and polymorphisms level of *sra* sequences to trace signatures of selection and to determine the extent of genetic variation in *P. falciparum* by characterizing the *sra* sequence at the nucleotide and protein levels.

MATERIALS AND METHODS

Study Areas and Blood Samples Collection

The samples of *P. falciparum* were obtained from febrile patients in Jiangsu Province, China, from 2015 to 2019, who had returned from working in tropical regions of sub-Saharan Africa endemic for malaria (Chu et al., 2018). A total of 74 *P. falciparum*-infected blood samples were collected from 12 countries. The subjects were identified for mono-infection of *P. falciparum* by microscopic examination of blood smears stained with Giemsa. The isolates were identified by specific polymerase chain reaction (PCR).

Amplification and Sequencing Analysis of *pfsra*

The full-length nucleotide sequences of *sra* from *P. falciparum* were divided into four fragments and amplified by PCR with primers designed as *pfsra*-1-Forward (5′-ATG TTT CTA AGT TCT AAG AAA AGA A-3′) and *pfsra*-1-Reverse (5′-AAA GGA ATC TGT CTC ATT ATT TGT T-3′), *pfsra*-2-Forward (5′-GAT AAT GAA GAA ACA GAA GAT ATT G-3′), and *pfsra*-2-Reverse (5′-ATC TAA TAG TTG TAT ATA AGC ATA TTT ATT AAC-3′), *pfsra*-3-Forward (5′-AAT AAG AAT TCA AAT CAA TCA TAT AAT T-3′) and *pfsra*-3-Reverse (5′-ATA ATA TTT CCT CAC AAT TTT TAC ATG-3′), and *pfsra*-4-Forward (5′-GTA CCT GCC AAA ATT AAA TAT ATA GAA-3′) and *pfsra*-4-Reverse (5′-TTA ATA TAT CGA AAT AAA TAT CAT AAG-3′), respectively. The *pfsra* (PlasmoDB, PF3D7_1431400) sequence from the *Plasmodium* Genomics Resource database was used as the reference gene sequence. The PCR amplification reactions were performed in a volume of 50 μl including 100 ng of genomic DNA, 0.2 μM each of the forward and reverse primers, 0.2 mM deoxynucleoside triphosphate, 2.5 units of DNA polymerase in 1 × *FastPfu* buffer (*TransStart® FastPfu* DNA polymerase, Beijing, China), and nuclease-free water up to 50 μl. The PCR amplification of *pfsra* genes was carried out in Mastercycler (Eppendorf, Hamburg, Germany). Amplification was performed as follows: denaturation at 95°C for 2 min, 35 cycles of 95°C for 20 s, 50°C for 20 s, and 65°C for 1 min, and final extension at 65°C for 5 min. The PCR products were analyzed using 1% agarose gel electrophoresis, stained with SuperStain (CWBIO, Jiangsu, China), and visualized by ultraviolet transilluminator (Bio-Rad ChemiDoc MP, Hercules, United States). The lengths of the PCR products were estimated based on their mobility relative to a standard DNA marker (TransGen Biotech, Beijing, China). Sequencing reactions were performed using GENEWIZ (Suzhou, China) with an ABI 3730xl DNA Analyzer (Thermo Fisher Scientific, Waltham, United States). All 74 samples generated a single amplification fragment of the expected size, and direct Sanger DNA sequencing of the forward and reverse directions was conducted to ensure the accuracy of the obtained sequences.

Sequence Alignment and Genetic Data Analysis

The geographical distribution map of *P. falciparum* samples was constructed by Arcgis10.2 software (ESRI, 2011). In order to evaluate diversity, *pfsra* sequence was used as template and aligned using GeneDoc2.7.0.[1] The primary structure of the PfSRA protein was demonstrated by UniProt.[2] The nucleotide sequences of *pfsra* were translated into the deduced amino acid (aa) sequences by DNASTAR (Burland, 2000). The predicted amino acid sequences of PfSRA from the PCR sequenced genomic fragments were aligned with the sequence of *P. falciparum* genome strain 3D7 by the MUSCLE algorithms in the MEGA 7.0 program (Kumar et al., 2016). A logo plot for each *pfsra* population was constructed to analyze the polymorphic

[1]https://github.com/karlnicholas/GeneDoc

[2]https://www.uniprot.org/uniprot/Q8ILF1

characteristics of PfSRA by the WebLogo program.[3] In addition, a codon-based test of purifying selection was analyzed by MEGA 7.0 program (Kumar et al., 2016). The non-synonymous mutations (*dN*), synonymous mutations (*dS*), and the *dN*/*dS* ratio from MEGA 7.0 were tested and compared by the *Z*-test ($p < 0.05$) using Nei and Gojobori's method, corrected by Jukes and Cantor and 1,000 bootstrap replications (Nei and Gojobori, 1986). Under purifying selection, *dN* will be less than *dS* (*dN*/*dS* < 1), while when the positive selection is more advantageous, *dN* will exceed *dS* (*dN*/*dS* > 1).

The average pairwise nucleotide diversity (π), number of haplotypes (*H*), and haplotype diversity (*Hd*) were calculated by DnaSP v6 (Rozas et al., 2017). The nucleotide diversity was analyzed by DnaSP v6 with a window length of 100 base pairs (bp) and a step size of 25 bp. In addition, the neutrality tests (Tajima's *D*, Fu and Li's D^*, and Fu and Li's F^*) implemented in DnaSP v6 software were utilized to measure the departure of the neutral mode prediction of molecular evolution (Fu and Li, 1993; Tajima, 1993). In order to determine the evolutionary relationship of the aligned sequences, based on nucleotide sequences, the phylogenetic tree of *sra* was constructed with the neighbor-joining method in MEGA 7.0. The *sra* sequences of diverse malaria parasites species included the SRA haplotypes of *plasmodium* from humans, non-human primate, avian, and murine malaria, which were obtained from the PlasmoDB and NCBI databases.

RESULTS

Geographical Origin of *P. falciparum*

A total of 74 clinical isolates of *P. falciparum* showed the geographical distribution in 12 sub-Saharan Africa countries. These isolates were mainly from the west coast of Africa, including Angola ($n = 16$, 21.6%), Nigeria and Equatorial Guinea ($n = 13$, 17.6%), and the Republic of the Congo ($n = 9$, 12.2%) (**Figure 1**). Of the 74 sequencing samples, 3 were from Eastern Africa (Uganda), 22 were from Western Africa (Sierra Leone, Côte d'Ivoire, Ghana, and Nigeria), 31 were from Central Africa (Cameroon, Equatorial Guinea, Gabon, Republic of the Congo, and Democratic Republic of the Congo), and 18 were from Southern Africa (Zambia and Angola). Overall, 74 cases of *P. falciparum* infection were identified in our study (**Table 1**). Full details for the isolates were provided (**Supplementary Table 1**).

Characterization of PfSRA

The length of SRA encoded by the *pfsra* full-length was 990 (aa), beginning with the predicted 24-aa signal peptide sequence (aa 1–24) and ending with a GPI-anchor (aa 969–990). Other specific regions, such as two coiled-coil regions, were also identified in the predicted protein primary structure of *P. falciparum* (aa 413–433 and 437–457) (**Figure 2A**). The full-length nucleotide sequences of *sra* from *P. falciparum* was 3,149 bp and was amplified by PCR using primer 1 (1–25 bp), primer 2 (924–948 bp), primer 3 (801–826 bp), primer 4 (1,768–1,800 bp), primer 5 (1,702–1,729 bp), primer 6 (2,675–2,701 bp), primer 7 (2,602–2,628 bp), and primer 8 (3,123–3,149 bp) (**Figure 2B**).

[3]https://weblogo.berkeley.edu/logo.cgi

Nucleotide Polymorphism of *pfsra*

The *sra* genes of 74 *P. falciparum* isolates were successfully amplified by PCR, corresponding to nucleotides 1–948, 801–1,800, 1,702–2,701, and 2,602–3,149, respectively, and a single PCR product with an expected size of 948 bp (*pfsra1*), 1 kb (*pfsra2* and *pfsra3*), and 548 bp (*pfsra4*) (**Figure 2C**). The direct sequencing of the purified PCR fragments indicated that there were no superimposed signals on the electropherograms of *pfsra*. Compared with the reference 3D7 strain, 74 isolates (100%) showed non-synonymous mutation. Overall, 63 single nucleotide polymorphisms (SNPs) were found in 74 isolates with the average π value of 0.00132 for *pfsra*. The sliding method plot using DnaSP v6 with a window length of 100 bp and a step size of 25 bp showed that the π value of *pfsra* is in the range of 0–0.01023. The conservative regions of 0–0.6 and 0.8–1.6 kb were observed in *pfsra* with π values of 0 approximately (**Figure 3**). The average number of nucleotide differences (*k*) of *pfsra* was 3.049. Nucleotide diversity of *pfsra* was categorized into 35 distinct haplotypes, and the estimated *Hd* was 0.770 (**Table 2**). For amino acid, the frequencies and types of mutation in the full-length of PfSRA (aa 1–990) were briefly presented in **Figure 4**. The C-terminal fragments of the PfSRA (aa 585–591, aa 879–900) showed relatively high polymorphism. More detailed amino acid comparison results are reported in **Supplementary Figure 1**.

Genetic Population Structure of *pfsra*

Based on the average values of *dS* and *dN*, the population genetic structure of the *P. falciparum* samples was analyzed using the *sra* gene polymorphisms in the codon-based purifying selection test. Results showed that there was diversifying selection or positive selection in *P. falciparum sra* population ($dS - dN = -0.00075$). In addition, the mean ratio of across sites non-synonymous (*dN*) to synonymous (*dS*) substitutions (*dN*/*dS*) was 1.365, and most of the nucleotide substitutions detected were non-synonymous, which also showed that the genetic variations of *pfsra* were maintained by positive selection. Tajima's *D* and Fu and Li's D^* and F^* tests rejected a neutral polymorphism occurrence model with values of *pfsra* (Tajima's $D = -2.75387$, $p < 0.05$, Fu and Li's $D^* = -6.85882$, $p < 0.05$, and Fu and Li's $F^* = -6.25597$, $p < 0.05$) (**Table 2**). Full details for all study countries were provided (**Supplementary Table 2**).

Phylogenetic Analysis of *sra*

As predicted based on the signature of positive selection and the level of genetic diversity described above, the phylogenetic relationship among 35 distinct haplotypes was detected in the *pfsra* sequences (1 was from Eastern Africa; 9 were from Western Africa; 14 were from Central Africa; and 11 were from Southern Africa) (**Figure 5**). The phylogenetic tree of 11 alleles of *sra* gene of 11 species of human and non-human *Plasmodium* primates was constructed by the neighbor-joining method (**Supplementary Figure 2**). **Supplementary Table 3** provides the *sra* gene ID number and gene length of other malaria parasite species.

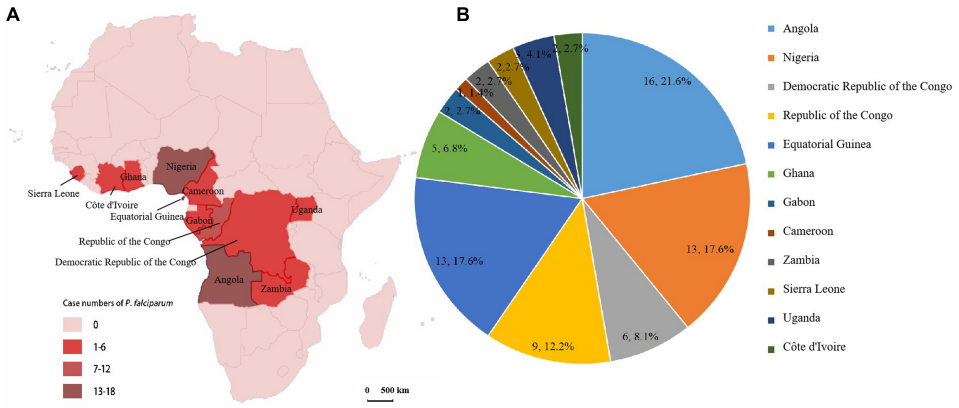


FIGURE 1 | Origins of malaria cases were imported from 12 countries of sub-Saharan Africa to the Jiangsu Province of China. **(A)** A map of Africa showing the countries of origin of *P. falciparum* isolates. **(B)** The total number of genotyped samples per region and percentage of samples.

TABLE 1 | Origin of imported *Plasmodium falciparum* in 2015–2019.

Country	2015 number	2016 number	2017 number	2018 number	2019 number	Total number (%)
Angola	3	3	2	4	4	16 (21.6%)
Nigeria	2	3	4	3	1	13 (17.6%)
Democratic Republic of the Congo	3	0	0	2	1	6 (8.1%)
Republic of the Congo	3	2	0	1	3	9 (12.2%)
Equatorial Guinea	7	1	2	2	1	13 (17.6%)
Ghana	2	0	1	2	0	5 (6.8%)
Gabon	1	0	0	1	0	2 (2.7%)
Cameroon	0	1	0	0	0	1 (1.4%)
Zambia	0	0	0	1	1	2 (2.7%)
Sierra Leone	0	1	0	0	1	2 (2.7%)
Uganda	0	1	0	0	2	3 (4.1%)
Côte d'Ivoire	0	0	0	0	2	2 (2.7%)
Total	21	12	9	16	16	74 (100%)

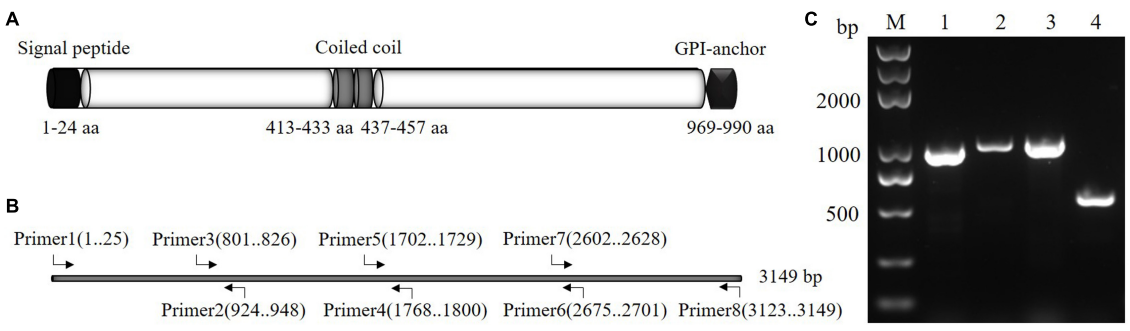


FIGURE 2 | Predicted *P. falciparum* SRA protein primary structure and primer design diagram. **(A)** Diagram of PfSRA protein primary structure. **(B)** Diagram of *pfsra* primers design. **(C)** Amplification of four fragments of *pfsra* by PCR. Abbreviations: M, DNA marker; 1, the first fragment of *pfsra* (948 bp); 2, the second fragment of *pfsra* (1 kb); 3, the third fragment of *pfsra* (1 kb); 4, the fourth fragment of *pfsra* (548 bp).

DISCUSSION

Apart from the complex life cycle of the malaria parasite involving the mosquito vector and human host, the malaria parasite exhibits extensive antigenicity and genetically diverse stages that may pose an adverse obstacle to malarial control strategies. Thus, a deeper understanding of patterns and mechanisms of sequence variation and genetic recombination may contribute to the design of a vaccine that represents the global repertoire of polymorphic malaria surface antigens

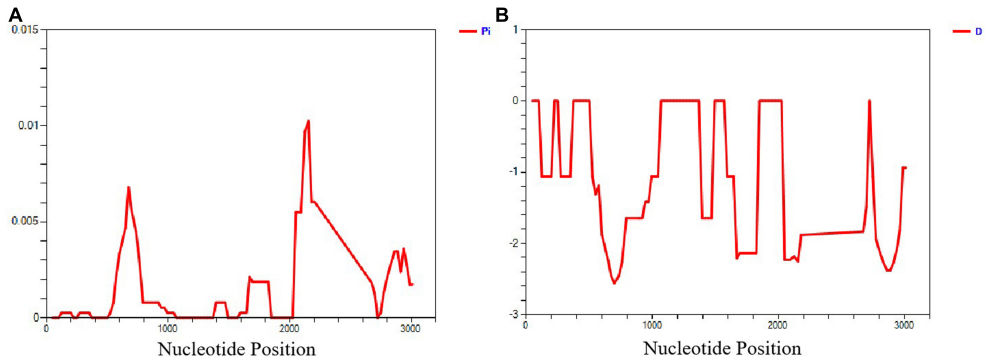


FIGURE 3 | Sliding window plot analyses show the sequence diversity (π) and Tajima's *D*. **(A)** Sequence diversity of PfSRA. **(B)** Tajima's *D* of PfSRA.

TABLE 2 | Estimates of number of haplotypes, haplotype diversity, nucleotide diversity, and neutrality indices of *pfsra*.

Region	No. of samples	No. of haplotypes	*Hd*	*dN*/*dS*	Diversity ± SD		Tajima's *D*	FU and Li's *D**	FU and Li's *F**
					Nucleotide	Haplotype			
Eastern Africa	3	3	1	1.579	0.00437 ± 0.00092	1 ± 0.177	−0.73807	−0.73807	−0.77178
Western Africa	21	21	0.996	1.313	0.00252 ± 0.00072	0.996 ± 0.014	−2.10441	−3.02686	−3.21120
Central Africa	30	25	0.968	1.267	0.00208 ± 0.00073	0.968 ± 0.024	−2.34218	−3.94714	−4.03442
Southern Africa	18	15	0.977	1.252	0.00186 ± 0.00086	0.977 ± 0.027	−2.42786	−3.45933	−3.66946
Total	74	35	0.770	1.365	0.00132 ± 0.00078	0.770 ± 0.054	−2.75387	−6.85882	−6.25597

In DnaSP software, D and F* tests are based on the neutral model prediction. This command calculates the statistical tests D* and F* proposed by Fu and Li (1993) for testing the hypothesis that all mutations are selectively neutral (Kimura, 1983).*

(Bharti et al., 2012). Systematic screens for uncharacterized *P. falciparum* invasion-related proteins evaluated PfSRA as one of the top hits that emerged; it contains coiled-coil domains known to be less polymorphic (Villard et al., 2007; Kulangara et al., 2009; Amlabu et al., 2018). Our investigation into the extent of sequence variation is consistent with this. In addition, coiled-coil domains form a stable structure, which elicit functional antibodies, thus blocking the related domains in many organisms and were considered to be the basis for the chemical synthesis of three PfSRA peptides designed to generate antibodies (Tripet et al., 2006; Gustchina et al., 2013; Jiang et al., 2016; Amlabu et al., 2018). Furthermore, these domains have been evaluated as potential targets for immunotherapy such as peptide-based vaccine strategies (Stoute et al., 1995; Demangel et al., 1998; Adda et al., 1999). We analyzed the full-length of *pfsra* (74 isolates) and found that the C-terminal fragments of the *pfsra* (π = 0.00198) show polymorphism probably due to selection pressure. Comparatively, the N-terminal of *sra* is a relatively conserved sequence (π = 0.00083).

A previous study suggested that people infected with malaria have naturally acquired antibodies against PfSRA and PfSRA N-terminal antibodies could partially inhibit merozoite invasion of erythrocytes by parasite growth inhibition assays (Amlabu et al., 2018). Now, the evidence of relatively conservative N-terminus might raise the possibility that it has the potential to be a candidate for anti-malarial vaccine. The ratio of non-synonymous (*dN*) to synonymous (*dS*) substitutions across sites was used as an index to evaluate selection pressure; $dN/dS > 1$ indicates diversifying positive selection. Further neutrality tests were carried out to determine the types and characteristics of natural selection on the *pfsra*. Statistically significant negative values of neutrality tests suggest an excess of rare polymorphisms in the population and provide evidence of purifying or directional (positive) selection (Fu and Li, 1993; Akey et al., 2004). The phylogenetic tree of 11 alleles of *sra* gene of 11 species showed that *pfsra* and other species occupied distinct bifurcating branches, supporting an ancient divergence times of the malarial parasite lineage.

The nucleotide diversity of *pfsra* in Southern Africa (π = 0.00186 ± SD 0.00086), Central Africa (π = 0.00208 ± SD 0.00073), and Western Africa (π = 0.00252 ± SD 0.00072) was lower than that in Eastern Africa (π = 0.00437 ± SD 0.00092), which may be related to the higher transmission rate of *P. falciparum* in Eastern Africa. Furthermore, more samples are needed in future research to support our findings and to control the limitations of small sample size (large confidence interval) in a single area. A previous study had also shown that *P. falciparum* has a spectrum of population structure: linkage "equilibrium," low levels of differentiation and high diversity in regions with high levels of transmission (Anderson et al., 2000; Huang et al., 2020). Mutation, recombination, gene flow, and natural selection may contribute to the genetic diversity of malaria parasites (Cole-Tobian and King, 2003).

In the analysis of *pfsra* full-length, there were abundant polymorphisms found. Samples from the four Africa regions showed their own distinct diversity patterns. Interestingly, two

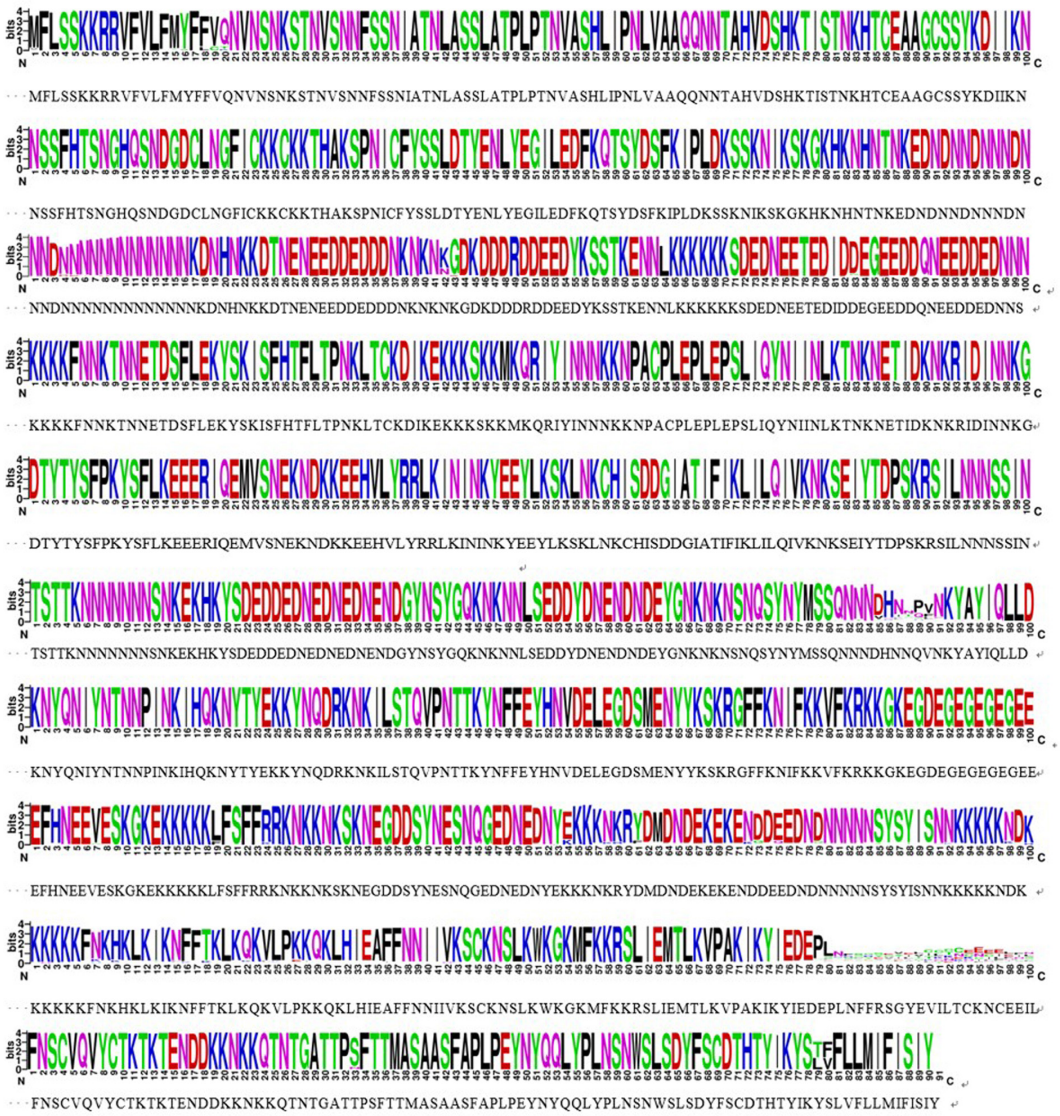


FIGURE 4 | Conservative locus analysis of the PfSRA amino acid sequences defined by WEBLOGO. Each logo consists of stacks of symbols, and each position in the sequence corresponds to stacks of symbols. The height within the stack of each individual amino acid abbreviation indicates its relative frequency at that specific position.

larger-size parasite population (Western Africa and Central Africa) showed more polymorphisms compared to those in Eastern Africa and Southern Africa. Some mutations showed the regional differences based on the geographical isolation effect; for example, the 15th amino acid mutant (M15Y) only occurred in Central Africa; K333E was only found in Western Africa. These phenomena indicate that it is necessary to continuously monitor these regional characteristic mutations in order to explore their association with regional malaria epidemics. Overall, apart from the conserved N-terminus, the composition of PfSRA vaccine should consider the high-frequency alleles instead of the C-terminus of wild-type ones (Conway, 2007).

Epidemiological studies have indicated that the level of heterologous mating in malaria populations is positively correlated with the prevalence of mixed allele infections and transmission rates (Chenet et al., 2008). The generation of relevant genetic, immunologic, and epidemiologic data for the *sra* gene is necessary, especially in areas with low malaria endemicity. Even in geographical areas with low transmission, the development of vaccine strategies should include results of diversity analysis. The uneven geographical distribution of alleles may jeopardize the development and use of vaccines targeting specific variable site, as local variation may not be taken into account in vaccine design (Cole-Tobian and King, 2003). The study of different genes and their alleles is helpful for us to understand the trends of genetic variation and if alleles could render vaccine ineffective. Given the genetic diversity found in the region, an alternative to improve the vaccine effectiveness is to

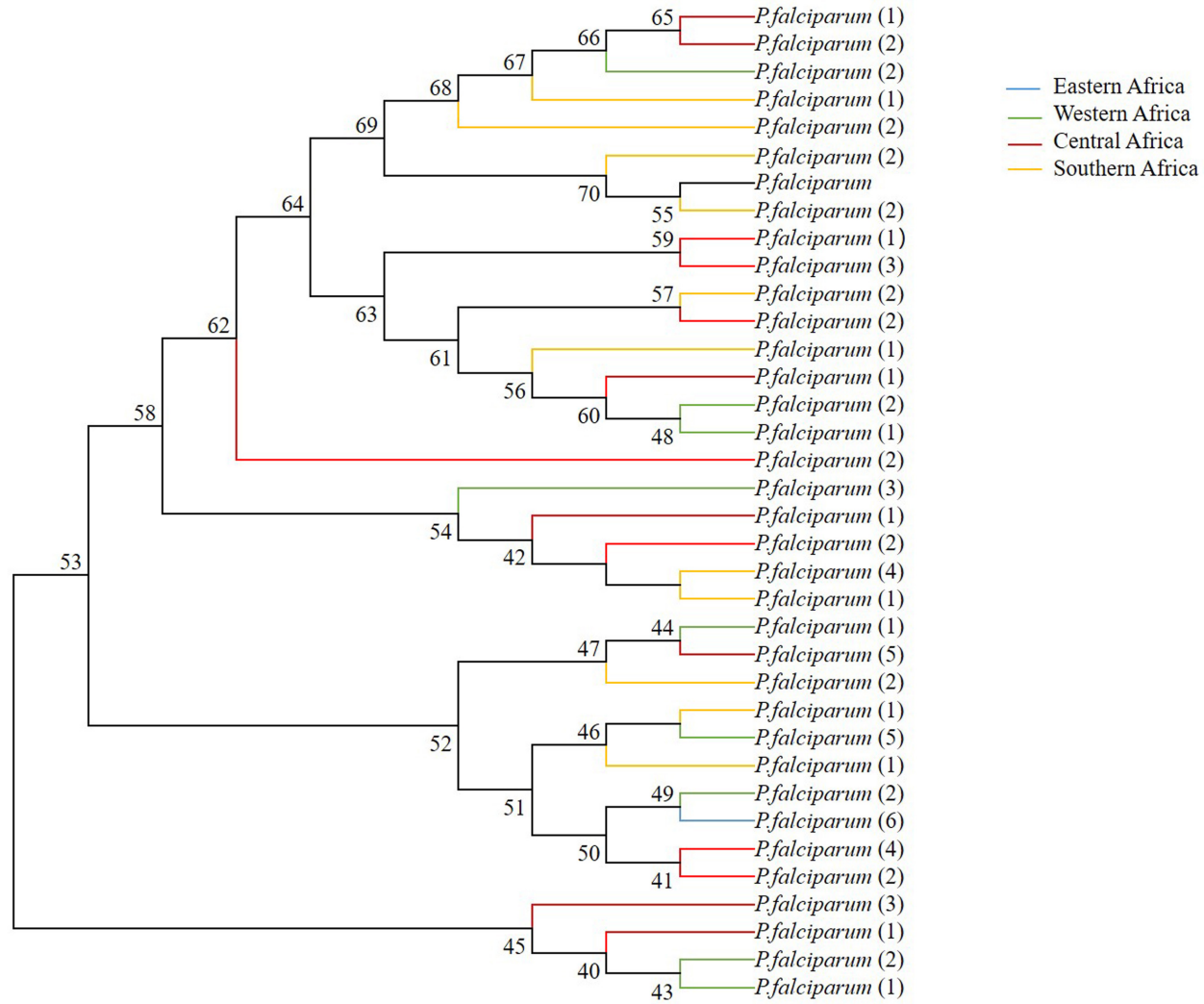


FIGURE 5 | Phylogenetic relationship of *sra* full-length genes within *pfsra* sequences based on the neighbor-joining method. Numbers at nodes show bootstrap values.

create a construct with the most common region-specific alleles (Cole-Tobian and King, 2003; Chenet et al., 2008).

CONCLUSION

The C-terminal fragments of the *sra* gene of *P. falciparum* showed polymorphism due to positive diversifying selection, which would hinder SRA-based vaccine development. Comparatively, in addition to the coiled-coil domains that have been evaluated as potential targets of peptide-based vaccines previously, the conserved N-terminal of *pfsra* is also a promising vaccine candidate against *P. falciparum* infection.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

ETHICS STATEMENT

This study was approved by the Ethics Committee of Jiangsu Institute of Parasitic Diseases (JIPD) (IRB00004221), Wuxi, China. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

BY and YC conceptualized the study, wrote the manuscript and contributed to the interpretation of the data. BY, HL, Q-WX, SX, HZ, J-XT, and G-DZ collected and analyzed the samples. Y-FS performed statistical and bioinformatics analysis. BY, YC, Y-BL, and JC revised the manuscript critically for important intellectual

content. All the authors contributed to this article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (81871681 and 81971967), the Jiangsu Provincial Department of Science and Technology (BM2018020), the Jiangnan University Student Innovation Training Program (1285210232175260), the National First-class Discipline Program of Food Science and Technology (JUFSTR20180101), and the Jiangsu Provincial Project of Invigorating Health Care through Science, Technology and Education.

ACKNOWLEDGMENTS

We thank all participants in this study, doctors, and local health departments for their support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.688606/full#supplementary-material

Supplementary Figure 1 | Amino acid sequence alignment of *P. falciparum* SRA. Gray area indicates that amino acid is identical across all aligned sequences.

Supplementary Figure 2 | Neighbor-joining tree of 11 unique alleles encoded by *sra* from 11 *Plasmodium* parasite species. Numbers at nodes show bootstrap values.

Supplementary Table 1 | Country of origin and parasitemia of *P. falciparum* samples.

Supplementary Table 2 | Neutrality tests of *pfsra* among 74 *P. falciparum* isolates.

Supplementary Table 3 | The *sra* Gene ID number and gene length of other *Plasmodium* species.

REFERENCES

Adda, C. G., Tilley, L., Anders, R. F., and Foley, M. (1999). Isolation of peptides that mimic epitopes on a malarial antigen from random peptide libraries displayed on phage. *Infect. Immun*. 67, 4679–4688. doi: 10.1128/Iai.67.9.4679-4688.1999

Akey, J. M., Eberle, M. A., Rieder, M. J., Carlson, C. S., Shriver, M. D., Nickerson, D. A., et al. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol*. 2:e286. doi: 10.1371/journal.pbio.0020286

Amlabu, E., Mensah-Brown, H., Nyarko, P. B., Akuh, O. A., Opoku, G., Ilani, P., et al. (2018). Functional characterization of *Plasmodium falciparum* surface-related antigen as a potential blood-stage vaccine target. *J. Infect. Dis*. 218, 778–790. doi: 10.1093/infdis/jiy222

Anderson, T. J., Haubold, B., Williams, J. T., Estrada-Franco, J. G., Richardson, L., Mollinedo, R., et al. (2000). Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol. Biol. Evol*. 17, 1467–1482. doi: 10.1093/oxfordjournals.molbev.a026247

Bharti, P. K., Shukla, M. M., Sharma, Y. D., and Singh, N. (2012). Genetic diversity in the block 2 region of the merozoite surface protein-1 of *Plasmodium falciparum* in central India. *Malar. J*. 11:78. doi: 10.1186/1475-2875-11-78

Bozdech, Z., Llinás, M., Pulliam, B. L., Wong, E. D., Zhu, J., and DeRisi, J. L. (2003). The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol*. 1:E5. doi: 10.1371/journal.pbio.0000005

Buffet, P. A., Safeukui, I., Deplaine, G., Brousse, V., Prendki, V., Thellier, M., et al. (2011). The pathogenesis of *Plasmodium falciparum* malaria in humans: insights from splenic physiology. *Blood* 117, 381–392. doi: 10.1182/blood-2010-04-202911

Burland, T. G. (2000). DNASTAR's lasergene sequence analysis software. *Methods Mol. Biol*. 132, 71–91. doi: 10.1385/1-59259-192-2:71

Chenet, S. M., Branch, O. H., Escalante, A. A., Lucas, C. M., and Bacon, D. J. (2008). Genetic diversity of vaccine candidate antigens in *Plasmodium falciparum* isolates from the Amazon basin of Peru. *Malar. J*. 7:93. doi: 10.1186/1475-2875-7-93

Chu, R., Zhang, X., Xu, S., Chen, L., Tang, J., Li, Y., et al. (2018). Limited genetic diversity of N-terminal of merozoite surface protein-1 (MSP-1) in *Plasmodium ovale curtisi* and *P. ovale wallikeri* imported from Africa to China. *Parasit. Vectors* 11:596. doi: 10.1186/s13071-018-3174-0

Cole-Tobian, J., and King, C. L. (2003). Diversity and natural selection in *Plasmodium vivax* Duffy binding protein gene. *Mol. Biochem. Parasitol*. 127, 121–132. doi: 10.1016/s0166-6851(02)00327-4

Conway, D. J. (2007). Molecular epidemiology of malaria. *Clin. Microbiol. Rev*. 20, 188–204. doi: 10.1128/CMR.00021-06

Demangel, C., Rouyre, S., Alzari, P. M., Nato, F., Longacre, S., Lafaye, P., et al. (1998). Phage-displayed mimotopes elicit monoclonal antibodies specific for a malaria vaccine candidate. *Biol. Chem*. 379, 65–70. doi: 10.1515/bchm.1998.379.1.65

Dhiman, S., and Veer, V. (2014). Culminating anti-malaria efforts at long lasting insecticidal net? *J. Infect. Public Health* 7, 457–464. doi: 10.1016/j.jiph.2014.06.002

ESRI (2011). *ArcGIS Desktop: Release 10*. Redlands, CA: Environmental Systems Research Institute.

Fu, Y. X., and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* 133, 693–709. doi: 10.1093/genetics/133.3.693

Gilson, P. R., Nebl, T., Vukcevic, D., Moritz, R. L., Sargeant, T., Speed, T. P., et al. (2006). Identification and stoichiometry of glycosylphosphatidylinositol-anchored membrane proteins of the human malaria parasite *Plasmodium falciparum*. *Mol. Cell Proteomics* 5, 1286–1299. doi: 10.1074/mcp.M600035-MCP200

Gustchina, E., Li, M., Ghirlando, R., Schuck, P., Louis, J. M., Pierson, J., et al. (2013). Complexes of neutralizing and non-neutralizing affinity matured Fabs with a mimetic of the internal trimeric coiled-coil of HIV-1 gp41. *PLoS One* 8:e78187. doi: 10.1371/journal.pone.0078187

Huang, H. Y., Liang, X. Y., Lin, L. Y., Chen, J. T., Ehapo, C. S., Eyi, U. M., et al. (2020). Genetic polymorphism of *Plasmodium falciparum* circumsporozoite protein on Bioko Island, equatorial guinea and global comparative analysis. *Malar. J*. 19:245. doi: 10.1186/s12936-020-03315-4

Jiang, Z., Gera, L., Mant, C. T., Hirsch, B., Yan, Z., Shortt, J. A., et al. (2016). Platform technology to generate broadly cross-reactive antibodies to alpha-helical epitopes in hemagglutinin proteins from influenza A viruses. *Biopolymers* 106, 144–159. doi: 10.1002/bip.22808

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge, MA: Cambridge University Press.

Kulangara, C., Kajava, A. V., Corradin, G., and Felger, I. (2009). Sequence conservation in *Plasmodium falciparum* alpha-helical coiled coil domains proposed for vaccine development. *PLoS One* 4:e5419. doi: 10.1371/journal.pone.0005419

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol*. 33, 1870–1874. doi: 10.1093/molbev/msw054

Le Roch, K. G., Zhou, Y., Blair, P. L., Grainger, M., Moch, J. K., Haynes, J. D., et al. (2003). Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* 301, 1503–1508. doi: 10.1126/science.1087025

Ménard, S., Ben Haddou, T., Ramadani, A. P., Ariey, F., Iriart, X., Beghain, J., et al. (2015). Induction of multidrug tolerance in *Plasmodium falciparum* by

extended artemisinin pressure. *Emerg. Infect. Dis*. 21, 1733–1741. doi: 10.3201/eid2110.150682

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol*. 3, 418–426. doi: 10.1093/oxfordjournals.molbev.a040410

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol*. 34, 3299–3302. doi: 10.1093/molbev/msx248

Stoute, J. A., Ballou, W. R., Kolodny, N., Deal, C. D., Wirtz, R. A., and Lindler, L. E. (1995). Induction of humoral immune-response against *Plasmodium-Falciparum* sporozoites by immunization with a synthetic peptide mimotope whose sequence was derived from screening a filamentous phage epitope library. *Infect. Immun*. 63, 934–939. doi: 10.1128/Iai.63.3.934-939.1995

Strode, C., Donegan, S., Garner, P., Enayati, A. A., and Hemingway, J. (2014). The impact of pyrethroid resistance on the efficacy of insecticide-treated bed nets against African anopheline mosquitoes: systematic review and meta-analysis. *PLoS Med*. 11:e1001619. doi: 10.1371/journal.pmed.1001619

Tajima, F. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* 135, 599–607. doi: 10.1093/genetics/135.2.599

Takala, S., Branch, O., Escalante, A. A., Kariuki, S., Wootton, J., and Lal, A. A. (2002). Evidence for intragenic recombination in *Plasmodium falciparum*: identification of a novel allele family in block 2 of merozoite surface protein-1: Asembo Bay area cohort project XIV. *Mol. Biochem. Parasitol*. 125, 163–171. doi: 10.1016/S0166-6851(02)00237-2

Tripet, B., Kao, D. J., Jeffers, S. A., Holmes, K. V., and Hodges, R. S. (2006). Template-based coiled-coil antigens elicit neutralizing antibodies to the SARS-coronavirus. *J. Struct. Biol*. 155, 176–194. doi: 10.1016/j.jsb.2006.03.019

Villard, V., Agak, G. W., Frank, G., Jafarshad, A., Servis, C., Nébié, I., et al. (2007). Rapid identification of malaria vaccine candidates based on alpha-helical coiled coil protein motif. *PLoS One* 2:e645. doi: 10.1371/journal.pone.0000645

WHO (2020). *World Malaria Report*. Available online at: http://www.who.int/publications/i/item/9789240015791 (accessed November 30, 2020).

WWARN (2015). Artemether-lumefantrine treatment of uncomplicated *Plasmodium falciparum* malaria: a systematic review and meta-analysis of day 7 lumefantrine concentrations and therapeutic response using individual patient data. *BMC Med*. 13:227. doi: 10.1186/s12916-015-0456-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 10 August 2021
doi: 10.3389/fmicb.2021.692837

Wildlife Is a Potential Source of Human Infections of *Enterocytozoon bieneusi* and *Giardia duodenalis* in Southeastern China

*Yan Zhang[1], Rongsheng Mi[1], Lijuan Yang[1], Haiyan Gong[1], Chunzhong Xu[2], Yongqi Feng[2], Xinsheng Chen[2], Yan Huang[1], Xiangan Han[1] and Zhaoguo Chen[1]**

[1] Key Laboratory of Animal Parasitology of Ministry of Agriculture, Laboratory of Quality and Safety Risk Assessment for Animal Products on Biohazards (Shanghai) of Ministry of Agriculture, Shanghai Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Shanghai, China, [2] Shanghai Wild Animal Park, Shanghai, China

Wildlife is known to be a source of high-impact pathogens affecting people. However, the distribution, genetic diversity, and zoonotic potential of *Cryptosporidium*, *Enterocytozoon bieneusi*, and *Giardia duodenalis* in wildlife are poorly understood. Here, we conducted the first molecular epidemiological investigation of these three pathogens in wildlife in Zhejiang and Shanghai, China. Genomic DNAs were derived from 182 individual fecal samples from wildlife and then subjected to a nested polymerase chain reaction–based sequencing approach for detection and characterization. Altogether, 3 (1.6%), 21 (11.5%), and 48 (26.4%) specimens tested positive for *Cryptosporidium* species, *E. bieneusi*, and *G. duodenalis*, respectively. Sequence analyses revealed five known (BEB6, D, MJ13, SC02, and type IV) and two novel (designated SH_ch1 and SH_deer1) genotypes of *E. bieneusi*. Phylogenetically, novel *E. bieneusi* genotype SH_deer1 fell into group 6, and the other genotypes were assigned to group 1 with zoonotic potential. Three novel *Cryptosporidium* genotypes (*Cryptosporidium* avian genotype V-like and *C. galli*-like 1 and 2) were identified, *C. galli*-like 1 and 2 formed a clade that was distinct from *Cryptosporidium* species. The genetic distinctiveness of these two novel genotypes suggests that they represent a new species of *Cryptosporidium.* Zoonotic assemblage A (n = 36) and host-adapted assemblages C (n = 1) and E (n = 7) of *G. duodenalis* were characterized. The overall results suggest that wildlife act as host reservoirs carrying zoonotic *E. bieneusi* and *G. duodenalis*, potentially enabling transmission from wildlife to humans and other animals.

Keywords: *Cryptosporidium*, *Enterocytozoon bieneusi*, *Giardia duodenalis*, genotypes, wildlife, prevalence, zoonotic potential

OPEN ACCESS

Edited by:
Matthew Adekunle Adeleke, University of KwaZulu-Natal, South Africa

Reviewed by:
Haileeyesus Adamu, Addis Ababa University, Ethiopia
Meng Qi, Tarim University, China

****Correspondence:***
Zhaoguo Chen
zhaoguochen@shvri.ac.cn

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *09 April 2021*
Accepted: *14 July 2021*
Published: *10 August 2021*

Citation:
Zhang Y, Mi R, Yang L, Gong H, Xu C, Feng Y, Chen X, Huang Y, Han X and Chen Z (2021) Wildlife Is a Potential Source of Human Infections of Enterocytozoon bieneusi and Giardia duodenalis in Southeastern China. Front. Microbiol. 12:692837. doi: 10.3389/fmicb.2021.692837

INTRODUCTION

Wildlife has been an important source of various high-impact pathogens affecting people, and zoonoses originated in wildlife remain a major public health issue around the world (Kruse et al., 2005). Novel diseases continue to emerge, and the responsible pathogens are often from unexpected wildlife, such as Ebola and Marburg virus (Bats; Amman et al., 2015; Jones et al., 2015); HIV-1 and

HIV-2 (Primates; Gao et al., 1999); Nipah, Hendra, and Menangle virus (Bats; Field et al., 2007); West Nile virus (Mosquitoes; CDC, 1999, 2000); SARS (severe acute respiratory syndrome)–like virus (Bats; Li et al., 2005); and 2019 novel coronavirus (COVID-19) (Wildlife; Liu et al., 2020; Xiao et al., 2020), highlighting the important role of wildlife in the transmission of zoonotic pathogens. Currently, the ongoing 2019 novel coronavirus pandemic has resounded the alarm on pathogens in wildlife. Thus, it is crucial to screen and identify potentially zoonotic pathogens in wildlife from different geographical regions for prediction, prevention, and control of zoonotic diseases outbreaks in humans (Cunningham et al., 2017).

Infectious diarrhea remains a major public health concern worldwide (Walker et al., 2012; Liu et al., 2015; GBD, 2016). It kills more than 2,000 children every day, more than AIDS, malaria, and measles (Liu et al., 2015; Chingwaru and Vidmar, 2018; Lemos et al., 2018; Maniga et al., 2018). Multiple pathogens including viruses, bacteria (Blander et al., 2017), fungi (Liguori et al., 2015; Hallen-Adams and Suhr, 2017), and protists (Lanata et al., 2013; Blander et al., 2017) are responsible for diarrhea. Among protists, *Cryptosporidium* species, *Giardia duodenalis*, and *Enterocytozoon bieneusi* are the most common etiological pathogens of the intestinal disease and are known to cause large disease outbreaks in humans, especially for *Cryptosporidium* and *G. duodenalis* (Karanis et al., 2007; Baldursson and Karanis, 2011; Decraene et al., 2012; Ryan and Cacciò, 2013; Checkley et al., 2015).

Currently, ∼40 named *Cryptosporidium* species and close to 50 genotypes have been reported (Feng et al., 2018; Haghi et al., 2020). There are ∼20 species of *Cryptosporidium* identified in humans (Xiao, 2010), of which *Cryptosporidium parvum* and *Cryptosporidium hominis* are the most common species infecting humans (Feng et al., 2018). *C. parvum* has a broad host range that includes humans and various animal species. By contrast, *C. hominis* is mainly restricted to humans, non-human primates, and equine animals (Feng et al., 2018). *G. duodenalis* is recognized as a species complex consisting of eight assemblages (A–H). Assemblages A and B can infect humans and other mammals, assemblages C and D are frequently found in dogs and other canids, assemblage E in hoofed animals, assemblage F in cats, assemblage G in rodents, and assemblage H in pinnipeds (Ryan and Zahedi, 2019). Until recently, assemblages C to H were considered host-specific, except that assemblages C, D, E, and F are occasionally found in humans (i.e., assemblage E has been found in human samples more frequently than F) (Gelanew et al., 2007; Broglia et al., 2013; Liu et al., 2014; Štrkolcová et al., 2015; Scalia et al., 2016; Zahedi et al., 2017). Among 14 species of microsporidia infecting humans, *E. bieneusi* is the most common microbe causing diarrhea. *E. bieneusi* can infect a broad host range, including mammals, birds, reptiles (Squamata), and insects (Diptera). Currently, there are more than 600 genotypes, and most genotypes can be found in both humans and animals, showing zoonotic potential (Li et al., 2019a,b; Zhang, 2019; Zhang et al., 2021). The three enteric eukaryotic agents can infect humans through the fecal–oral route, via direct contact with infected individuals or ingestion of contaminated water or food (Yu et al., 2020).

The three microbes can be identified or characterized at species, subspecies, and/or genotypic level using molecular techniques. Currently, small subunit ribosomal DNA (*SSU* rDNA) has been wildly used for *Cryptosporidium* species identification, whereas a genetic marker in the 60-kDa glycoprotein (*gp60*) gene has been commonly used for differentiating *Cryptosporidium* at the genotypic and subgenotypic levels (Abeywardena et al., 2015). For *G. duodenalis*, triose-phosphate isomerase (*tpi*), β-giardin (*bg*), glutamate dehydrogenase (*gdh*), elongation factor 1-alfa (*ef1*-α), and *SSU* rDNA are commonly used for genotypic identification (Ryan and Cacciò, 2013). As *ef1*-α and *SSU* rDNA are relatively problematic, and they cannot discriminate *G. duodenalis* subtypes within assemblages accurately and are thus not useful for transmission analyses (Traub et al., 2004). Internal transcribed spacer (ITS) of nuclear ribosomal DNA is sufficiently variable for the identification and genotypic characterization of *E. bieneusi* (Santín et al., 2009).

Using the approach above, we have explored the microbes from various animals including wild deer (Zhang et al., 2018b; Koehler et al., 2020), marsupials (Zhang et al., 2018c), domestic alpacas (Koehler et al., 2018), cattle (Zhang et al., 2018a), goats and sheep (Zhang et al., 2020), companion cats and dogs (Zhang et al., 2019), and humans (Zhang et al., 2018e). We also established a new phylogenetic classification system of overall 600 *E. bieneusi* genotypes (Zhang et al., 2021). The present study aims to identify three pathogens (*Cryptosporidium* species, *E. bieneusi*, and *G. duodenalis*) in wildlife in Zhejiang and Shanghai, characterize their genotypes and analyze their zoonotic potential. The findings in this study would help to understand the genetic diversity of the three agents and provide critical information for future global strategies to prevent outbreaks of their zoonoses.

MATERIALS AND METHODS

Samples and DNA Isolation

In total, 182 fecal samples were collected from 48 species of zoo animals from Zhejiang zoo (n = 52) and Shanghai Wild Animal Park (n = 130) from May 2018 to August 2020 (**Supplementary Table 1**). Some fecal samples were collected from wildlife rectum directly, whereas others were fresh deposited fecal samples. Genomic DNA was extracted directly from 0.1 to 0.4 g of each of the 182 fecal samples using the FastDNA SPIN Kit for Soil (MP Biomedicals, Santa Ana, CA, United States) according to the manufacturer's recommendations. The extracted DNA was stored at −20°C for further polymerase chain reaction (PCR) assay.

Detection of *Cryptosporidium* Species, *E. bieneusi* and *G. duodenalis*

Nested PCR-Based Sequencing of *Cryptosporidium* Species *SSU* rDNA

The small subunit of ribosomal nuclear DNA locus (target length 830 bp) of each sample was screened for identification of *Cryptosporidium* species (i.e., primers are listed in **Table 1**). In

the first run, PCR contained 25 μL of 2 × PCR buffer for KOD FX (Mg^{2+} plus) (Toyobo, Japan), 2 mM dNTPs, 100 nM (each) primer, 1.0 U KOD FX, and 1 μL of DNA template in a total 50 μL reaction mixture. A total of 35 cycles were carried out, each consisting of 94°C for 45 s, 55°C for 45 s, and 72°C for 1 min, with an initial hot start at 94°C for 3 min, and a final extension at 72°C for 7 min. A secondary PCR product was then amplified from 2 μL of the primary PCR products with the same cycling conditions as the first run, except for 60°C annealing temperature (Xiao et al., 1999, 2001; Jiang et al., 2005).

Nested PCR-Based Sequencing of *E. bieneusi* ITS

Individual genomic DNA samples were subjected to nested PCR-coupled sequencing of the *ITS* (243–245 bp) region (i.e., only 243–245-bp fragment of the ITS was used for further phylogenetic analyses) using an established technique (Katzwinkel-Wladarsch et al., 1996). Nested PCR (in 50 μL) was conducted in a standard buffer containing 3.0 μM $MgCl_2$, 0.4 mM dNTPs, 50 pmol of each primer, 1.25 U of Ex Taq DNA (TaKaRa Bio Inc., Beijing, China), and DNA template—except for the negative (no-template) control. The cycling conditions for both primary and secondary (nested) PCRs were as follows: 94°C for 5 min (initial denaturation), followed by 35 cycles of 94°C for 45 s (denaturation), 54°C for 45 s (annealing), and 72°C for 1 min (extension), followed by 72°C for 10 min (final extension).

Nested PCR-Based Sequencing of *G. duodenalis TPI*

G. duodenalis assemblages were identified and characterized by nested PCR-based sequencing of the *tpi* gene (∼530 bp) using the established methods (Sulaiman et al., 2003). PCR was carried out in a volume of 50 μL containing 3.0 μM $MgCl_2$, 0.4 mM dNTPs, 50 pmol of each primer, 1.25 U of Ex Taq DNA (TaKaRa Bio Inc., Beijing, China), and DNA template. A cycling protocol of 94°C for 5 min (initial denaturation), followed by 35 cycles of 94°C for 45 s (denaturation), 50°C for 45 s (annealing), 72°C for 1 min (extension), and a final extension of 72°C for 10 min. The secondary amplification was achieved using the same cycling conditions, except for the annealing temperature of 55°C for 30 s.

Known test-positive, test-negative, and no-template controls were included in each PCR run. The secondary PCR products were examined by gel electrophoresis on a 1.5% agarose gel containing 4S Green Plus Nucleic Acid Stain (Sangon Biotech, Shanghai, China) and directly sequenced using second-round PCR primers in both directions. All sequences obtained (GenBank accession nos. *Cryptosporidium*: MW168840-MW168842; *E. bieneusi*: MT895455-MT895461 and *G. duodenalis*: MW048593-MW048601) were inspected for quality and compared with reference sequences acquired from the GenBank database.

Phylogenetic Analysis

Obtained sequences from this and previous studies were aligned over a consensus length of 735 (*Cryptosporidium*), 459 (*G. duodenalis*), and 270 (*E. bieneusi;* after trimming, approximately 243-bp fragment of the ITS was analyzed) positions using previously established methods (Zhang et al., 2018d) and then subjected to phylogenetic analyses using the Bayesian inference (BI) and Monte Carlo Markov Chain methods in MrBayes v.3.2.3 (Huelsenbeck and Ronquist, 2001). The Akaike Information Criteria test in jModeltest v.2.1.7 (Darriba et al., 2012) was used to evaluate the likelihood parameters set for BI analysis. Posterior probability (*pp*) values were calculated by running 2,000,000 generations with four simultaneous tree-building chains, with trees saved every one-hundredth generation. A 50% majority-rule consensus tree for each analysis was constructed based on the final 75% of trees generated by BI. The clades and subclades were assigned and named using an established classification system (Santín and Fayer, 2009, 2011; Feng and Xiao, 2011; Karim et al., 2015; Li W. et al., 2015; Koehler et al., 2016; Li et al., 2019a,b; Ryan and Zahedi, 2019).

RESULTS

Molecular Detection of *Cryptosporidium* Species Based on *SSU* rDNA Gene

In total, three fecal DNA samples were identified *Cryptosporidium* species with the prevalence of 1.6% (3/182) (**Table 2**). They were all novel *SSU* rDNA sequences (i.e., < 100% identity with a sequence on GenBank) uniquely form the zoo in Zhejiang (**Table 3**). The three novel *SSU* rDNA sequences were assigned to the most closely related species or genotypes of *Cryptosporidium* based on sequence identity, representing *Cryptosporidium galli*-like 1 (from a species of Psittacidae) and *C. galli*-like 2 (channel-billed toucan) and *Cryptosporidium* avian genotype V-like (green aracari). *C. galli*-like 1 and 2 differed by 18 bp (763/781; 97.7%) and 17 bp (765/778; 98.3%) from the sequences representing *C. galli* (GenBank accession no. MG516766), and *Cryptosporidium* avian genotype V-like differed by 7 bp (780/787; 99.1%) from a sequence with GenBank accession no. JX548292 (*Cryptosporidium* avian genotype V).

The three *SSU* rDNA sequences were aligned with selected representative sequences in particular clades and subjected to the phylogenetic analysis (**Figure 1**). Genotypes *Cryptosporidium* avian genotype V-like clustered with genotype V with strong statistical support (*pp* = 1). *Cryptosporidium galli*-like 1 and 2 fell in one group and clustered with a clade of *C. galli* (*pp* = 0.99).

E. bieneusi Genotype Characterizations Based on ITS Region

Enterocytozoon DNA was specifically detected by nested PCR of ITS in 21 of 182 (11.5%) fecal samples from zoo animals in Zhejiang (3.8%; 2/52) and Shanghai (14.6%; 19/130) (**Table 2**), including 10 mammal species: Alpaca (*Vicugna pacos*), amur tiger (*Panthera tigris altaica*), brown bear (*Ursus arctos pruinosus*), cheetah (*Acinonyx jubatus*), fallow deer (*Dama dama*), lion (*Panthera leo*), red deer (*Cervus elaphus*), sika deer (*Cervus Nippon*), snub-nosed monkey (*Rhinopithecus* species), tiger (*Panthera tigris tigris*), and three species of birds: Chestnut-fronted macaw (*Ara severa*), great pied hornbill (*Buceros bicomis*), and red-and-green macaw (*Ara chloropterus*) (**Table 3**).

TABLE 1 | PCR primers (forward and reserve) used for the amplification of *Cryptosporidium*, *Enterocytozoon bieneusi*, and *Giardia duodenalis* in this study.

Species (genetic marker)	Primers (5′-3′)	Length (~bp)	References
Cryptosporidium	TTC TAG AGC TAA TAC ATG CG	1,325	Xiao et al., 1999
(*SSU* rDNA)	CCC ATT TCC TTC GAA ACA GGA		Xiao et al., 2001
	GGA AGG GTT GTA TTT ATT AGA TAA AG	830	Jiang et al., 2005
	CTC ATA AGG TGC TGA AGG AGT A		
E. bieneusi	MSP-1 (TGA ATG KGT CCC TGT)	590	Katzwinkel-Wladarsch et al., 1996
(ITS)	MSP-2B (GTT CAT TCG CAC TAC T)		
	MSP-3 (GGA ATT CAC ACC GCC CGT CRY TAT)	508	
	MSP-4B (CCA AGC TTA TGC TTA AGT CCA GGG AG)		
G. duodenalis	AL3543 (AAA TTA TGC CTG CTC GTC G)	605	Sulaiman et al., 2003
(*tpi*)	AL3546 (CAA ACC TTT TCC GCA AAC C)		
	AL3544 (CCC TTC ATC GGT GGT AAC TT)	532	
	AL3545 (GTG GCC ACC ACT CCC GTG CC)		

TABLE 2 | Prevalence of *Cryptosporidium* species, *Enterocytozoon bieneusi*, and *Giardia duodenalis* in Shanghai Wild Animal Park and Zhejiang zoo of China.

Species	Prevalence of each species (%)	Total no. of positive/total no. of samples	Prevalence in each region (%) (no. of positive/no. of samples)
		Shanghai	Zhejiang
Cryptosporidium	1.6 3/182	0	1.6 (3/52)
E. bieneusi	11.5 21/182	14.6 (19/130)	3.8 (2/52)
G. duodenalis	26.4 48/182	30.8 (40/130)	15.4 (8/52)

The 21 ITS amplicons (243 bp) were aligned to reference sequences in the GenBank database, and seven distinct genotypes were identified, including five known (BEB6, D, MJ13, SC02, and type IV) and two novel genotypes (designated SH_ch1 and SH_deer1) (**Table 3**). Novel genotype SH_ch1 (n = 2; from cheetahs) differed by 1 bp (242/243; 99.6%) from the sequence representing genotypes and KIN-1 (GenBank number MT231508). Novel genotype SH_deer1 (n = 1; from a sika deer) showed 8-bp (234/242; 99.7%) differences from the sequence with GenBank accession number KF261802. Two ambiguous sequences were derived from two amplicons, each containing multiple genotypes.

The eight ITS sequences representing seven distinct genotypes were aligned with sequences representing 10 groups of *E. bieneusi* and subjected to phylogenetic analysis (**Figure 2**). Genotypes BEB6, D, MJ13, SC02, SH_ch1, and type IV could be assigned to group 1 (pp = 0.96). Novel genotype SH_deer1 clustered with genotypes in group 6 with strong statistical support (pp = 0.95).

G. duodenalis Assemblages Identification Based on *tpi* Gene

Sequencing of all *tpi* amplicons identified 48 of 182 (26.4%) individual fecal samples to contain *Giardia* based on direct sequence comparisons, including 8 (15.4%; 8/52) in Zhejiang zoo and 40 (30.8%; 40/130) in Shanghai Wild Animal Park (**Table 2**). Genetic assemblages A (n = 36), C (n = 1), and E (n = 7) of *G. duodenalis* were characterized, and four amplicons contained mixed indeterminate genotypes. In total, eight distinct sequence types for *tpi* were defined (**Table 3**), including four representing *Giardia* sub-assemblage A (i.e., one known type from 16 species of wildlife and three novel sequence types from cheetah, fennec fox, lion, and snub-nosed monkey), one novel sequence type from a spotted hyena defined as assemblage C, and two novel distinct sequence types all representing assemblage E from giraffes.

The eight distinct *tpi* sequences representing four distinct assemblages or sub-assemblages were aligned with sequences representing *Giardia* assemblages A–G and subjected to the phylogenetic analysis (**Figure 3**). A novel sequence type (GenBank accession no. MW048604) clustered with assemblage C with strong statistical support (pp = 1.00).

DISCUSSION

The zoonotic enteric pathogens *Cryptosporidium*, *Enterocytozoon*, and *Giardia* have been reported in captive, wild, and zoo animals around the world (Leśniańska et al., 2016; Li N. et al., 2018; Amer et al., 2019). Their ability to spread via contaminated food, water, or direct contact with humans (e.g., zookeeper) poses a risk to public health.

Cryptosporidium

PCR-based sequencing of all three amplicons from 182 fecal DNA samples (1.6%; 3/182) revealed three operational taxonomic units (OTUs) of *Cryptosporidium* from three birds (channel-billed toucan, green aracari, and an unknown species of Psittacidae). Their *SSU* rDNA sequences were aligned (over a consensus length of 735 positions) with publicly available sequences, representing 14 species and an outgroup *C. muris* (**Figure 1**). Phylogenetic analyses of *SSU* rDNA data revealed that *Cryptosporidium* avian genotype V-like clustered with the genotypes *C. galli* and *Cryptosporidium* avian genotype V, which

TABLE 3 | Summary of all pathogen species, genotypes, and/or assemblages identified in wildlife in Zhenjiang and Shanghai, China, using PCR-based sequencing of particular genetic markers.

Species/genotype/ assemblage identified by PCR based on sequencing (positivity no.)	Genetic marker used	GenBank accession no.	Host (Latin name)	Positivity no. for each wild animal species
Cryptosporidium species V-like	(1) *SSU*	MW168842*	Green aracari (*Pteroglossus viridis*)	(1)
C. galli-like 1	(2) *SSU*	MW168841*	Psittacidae (species unknown)	(1)
C. galli-like *2*		MW168840*	Channel-billed toucan (*Ramphastos vitellinus*)	(1)
Giardia duodenalis A	(40) *tpi*	MW048593	Alpaca (*Vicugna pacos*)	(2)
		MW048598*	Siberian tiger (*Panthera tigris altaica*)	(2)
		MW048599*	Black-necked Crane (*Grus nigricollis*)	(2)
		MW048600*	Blue-headed macaw (*Propyrrhura couloni*)	(3)
		MW048601*	Cheetah (*Acinonyx jubatus*)	(3)
		MW048594[a]	Fennec fox (*Vulpes zerda*)	(2)
		MW048595[a]	Giant Eland (*Tragelaphus derbianus*)	(1)
		MW048596[a]	Giraffe (*Giraffa camelopardalis*)	(3)
		MW048597[a]	Golden takin (*Budorcas taxicolor bedfordi*)	(1)
			Great pied hornbill (*Buceros bicomis*)	(1)
			Hippopotamus (*Hippopotamus amphibious*)	(1)
			Lion (*Panthera leo*)	(2)
			Malabar pied hornbill (*Anthracoceros coronatus*)	(1)
			Snub-nosed monkey (*Rhinopithecus roxellana*)	(8)
			Ostrich (*Struthio camelus*)	(2)
			Peafowl (*Pavo cristatus*)	(2)
			Scarlet macaw (*Ara macao*)	(1)
			Sika deer (*Cervus Nippon*)	(1)
			Sun parakeet (*Aratinga solstitialis*)	(1)
			Tiger (*Panthera tigris tigris*)	(1)
G. duodenalis C	(1) *tpi*	MW048604*	Spotted hyena (*Crocuta crocuta*)	(1)
G. duodenalis E	(5) *tpi*	MW048602	Giraffe (*Giraffa camelopardalis*)	(4)
			Kangaroo (*Macropus* species)	(1)
	(2) *tpi*	MW048603*	Giraffe (*Giraffa camelopardalis*)	(2)
Enterocytozoon bieneusi BEB6	(3) ITS	MT895455	Alpaca (*Vicugna pacos*)	(1)
			Fallow deer (*Dama dama*)	(1)
			Red deer (*Cervus elaphus*)	(1)
E. bieneusi D	(8) ITS	MT895457	Siberian tiger (*Panthera tigris altaica*)	(2)
			Lion (*Panthera leo*)	(2)
			Snub-nosed monkey (*Rhinopithecus roxellana*)	(2)
			Tiger (*Panthera tigris tigris*)	(2)
E. bieneusi MJ13	(1) ITS	MT895460	Red-and-green macaw (*Ara chloropterus*)	
E. bieneusi SC02	(3) ITS	MT895459	Great pied hornbill (*Buceros bicomis*)	(2)
			Red-and-green macaw (*Ara chloropterus*)	(1)
E. bieneusi SH_ch1	(2) ITS	MT895458*	Cheetah (*Acinonyx jubatus*)	(2)
E. bieneusi SH_deer1	(1) ITS	MT895456*	Sika deer (*Cervus Nippon*)	(1)
E. bieneusi type IV	(1) ITS	MT895461	Chestnut-fronted macaw (*Ara severa*)	(1)
E. bieneusi BEB6-like	(1) ITS	MT895462[a]	Red deer (*Cervus elaphus*)	(1)
E. bieneusi MJ17-like	(1) ITS	MT895463[a]	Brown bear (*Ursus arctos pruinosus*)	(1)

**Novel genotypes.*
[a]*Mixed (indeterminate) genotypes.*

are typically found in birds (Xiao et al., 2004), and novel OTUs (genotypes *C. galli*-like 1 and 2 grouped, with strong nodal support (*pp* = 0.99). This analysis clearly showed that *C. galli*-like 1 and 2 represent a new and distinct clade. As the sequence variation (0–1.2%) within novel *C. galli*-like group was substantially less than differences (2.7–3.7%) between *C. galli* group and *C. galli*-like 1 and 2 upon pairwise comparison (**Figure 1** and **Supplementary Table 2**), we propose that the latter two genotypes may represent a novel species of *Cryptosporidium*. However, it should be cautious to draw this conclusion. Definitely, further histological and morphological studies are needed. Sequencing *SSU* rDNA from many more

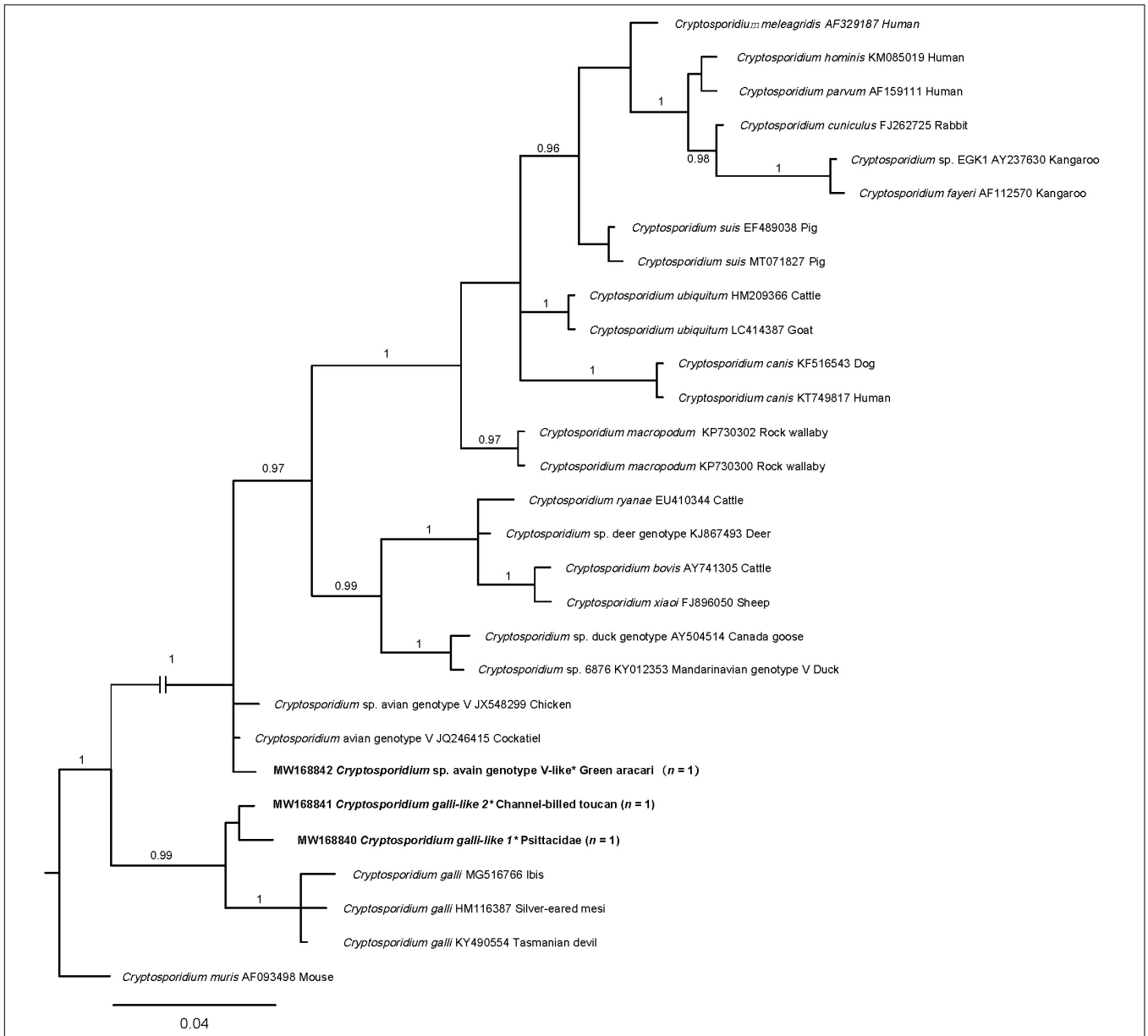


FIGURE 1 | Relationships among *Cryptosporidium* taxa inferred from the phylogenetic analysis of partial small subunit ribosomal rDNA gene (*SSU* rDNA) sequence data by Bayesian inference (BI). Posterior probabilities are indicated at all major nodes. Bold font indicates *Cryptosporidium* species or genotypes characterized from fecal DNA samples in this study. In parentheses are the numbers of samples representing a particular species, genotype, and sequence (GenBank accession numbers indicated). Novel genotypes (*). Scale bar represents the number of substitutions per site. Most clades were strongly supported (*pp* = 0.96–1.00). *pp* < 0.95 was not shown.

representatives of *Cryptosporidium* to conduct a comprehensive phylogenetic analysis is also required.

E. bieneusi

E. bieneusi was identified in three wildlife fecal DNA samples in Zhejiang zoo (3.8%; 2/52) and 19 in Shanghai Wild Animal Park (14.6%; 19/130), with a total prevalence of 11.5% (21/182). Similarly, Li J. et al. (2015) and Yu et al. (2017) studied the prevalence of *E. bieneusi* in Shanghai wildlife animal park and reported 44.8% (30/67) and 69.1% (38/55), respectively. These cited prevalences are all higher than that in our study; however, they uniquely focused on the populations of non-human primates. By contrast, Li et al. (2016) studied 70 different wildlife species (272 fecal samples) in Chengdu zoo and Bifengxia zoo with prevalences of 10.6% (21/198) and 29.7% (22/74), respectively, both of which are higher than that in Zhejiang zoo, but *E. bieneusi* positivity in Chengdu zoo was lower than that in Shanghai, indicating that *E. bieneusi* might be widespread in Shanghai wild animal park. Internationally, the overall prevalences of *E. bieneusi* in farmed and/or captive wildlife and zoo animals globally ranged from 1.4% in Australia (Zhang et al., 2018c) to 53.3% in China (Yu et al., 2020). The

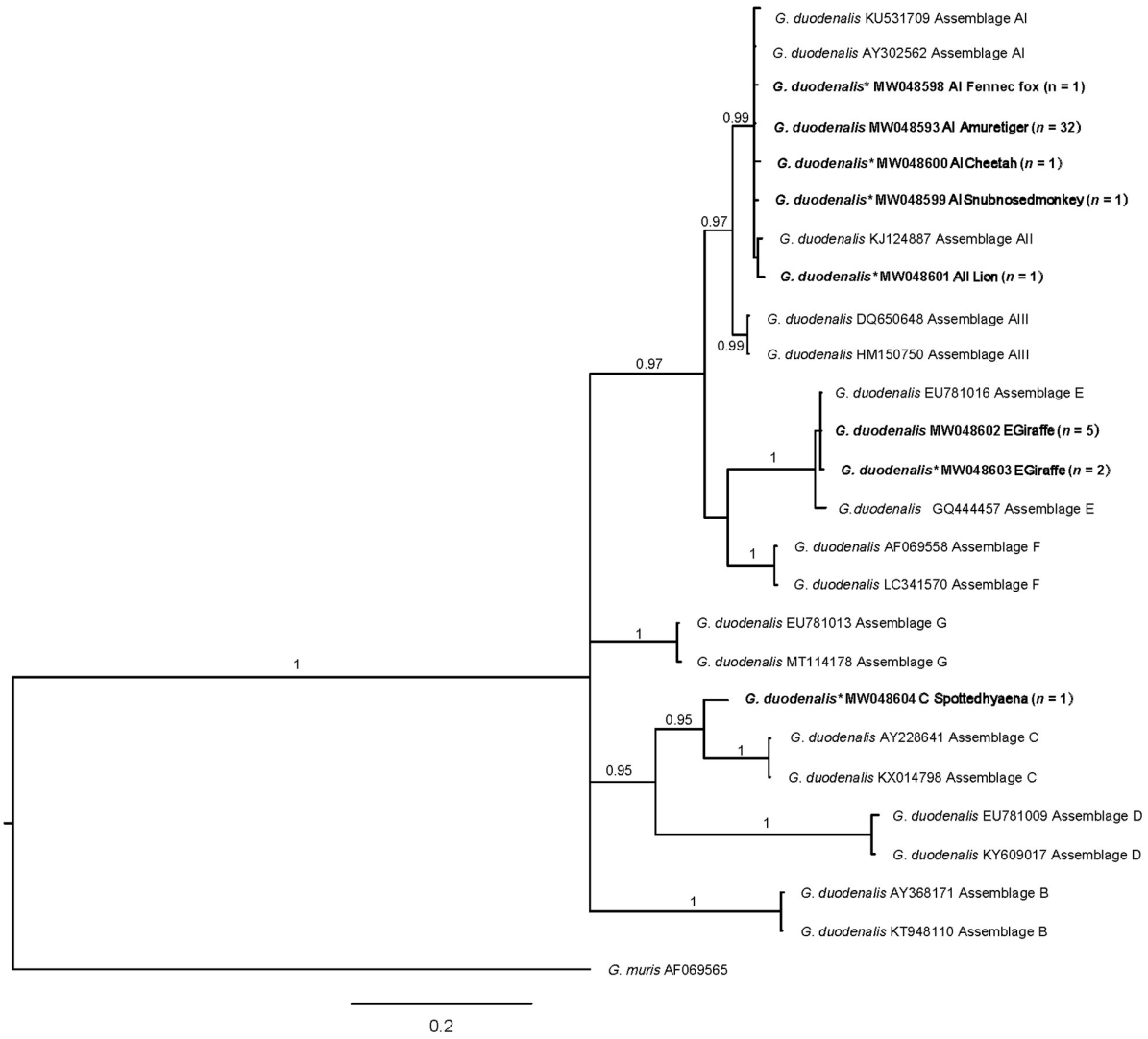


FIGURE 2 | Relationships among genotypes of *Enterocytozoon bieneusi* recorded in wildlife in this study inferred from phylogenetic analysis of sequence data for the internal transcribed spacer (ITS) of nuclear ribosomal DNA by Bayesian inference (BI). Statistically significant posterior probabilities (*pp*) are indicated on branches. Individual GenBank accession numbers precede genotype designation (in italics), followed by sample and locality descriptions. *Enterocytozoon bieneusi* genotypes identified and characterized from fecal DNA samples in the present study are indicated in bold font. Clades were assigned group names based on the established classification system established by Karim et al. (2015) and Li et al. (2019a). The scale bar represents the number of substitutions per site. The *E. bieneusi* genotypes PtEbIX (DQ885585) and CD8 (KJ668735) from dogs were used as outgroups. All groups were strongly supported (*pp* = 0.96–1.00). *pp* < 0.95 were not shown. Novel genotypes (*).

variety of *E. bieneusi* prevalences might be due to host species, health status, and immunity of animals; management; locations; sample size; and environmental factors—season, temperature, sunlight, and humidity.

In total, five known (BEB6, D, MJ13, SC02, and type IV) and two novel genotypes (designated SH_ch1 and SH_deer1) were identified in this study. The predominant genotype here was genotype D (38.1%; 8/21), followed by BEB6 and SC02 (each 14.3%; 3/21), SH_ch1 (9.5%; 2/21), and four other genotypes (each 4.8%; 1/21). Genotype D is frequently identified in humans and nearly 70 species of animals, including birds (Anseriformes, Columbiformes, Falconiformes, Galliformes, Gruiformes, and Passeriformes) and mammals (Artiodactyla, Carnivora, Lagomorpha, Perissodactyla, Primates, and Rodentia) (Zhang, 2019; Zhang et al., 2021), indicating that genotype D has the capability of intra-species transmission. Similarly, genotype BEB6 has also been found in humans and 23 animal species (Zhang, 2019; Zhang et al., 2021), and fallow deer (reported here) is the first record of this genotype. Genotype SC02 was found in human and bear (Wu et al., 2018), giant panda (Li W. et al., 2018), horse (Deng et al., 2016b), Pallas's squirrel, raccoon (Li et al., 2016), red-bellied tree squirrel (Deng et al., 2016a), rhesus macaque (Zhong et al., 2017), and wild boar (Li et al., 2017); great pied hornbill (*Buceros bicomis*) identified

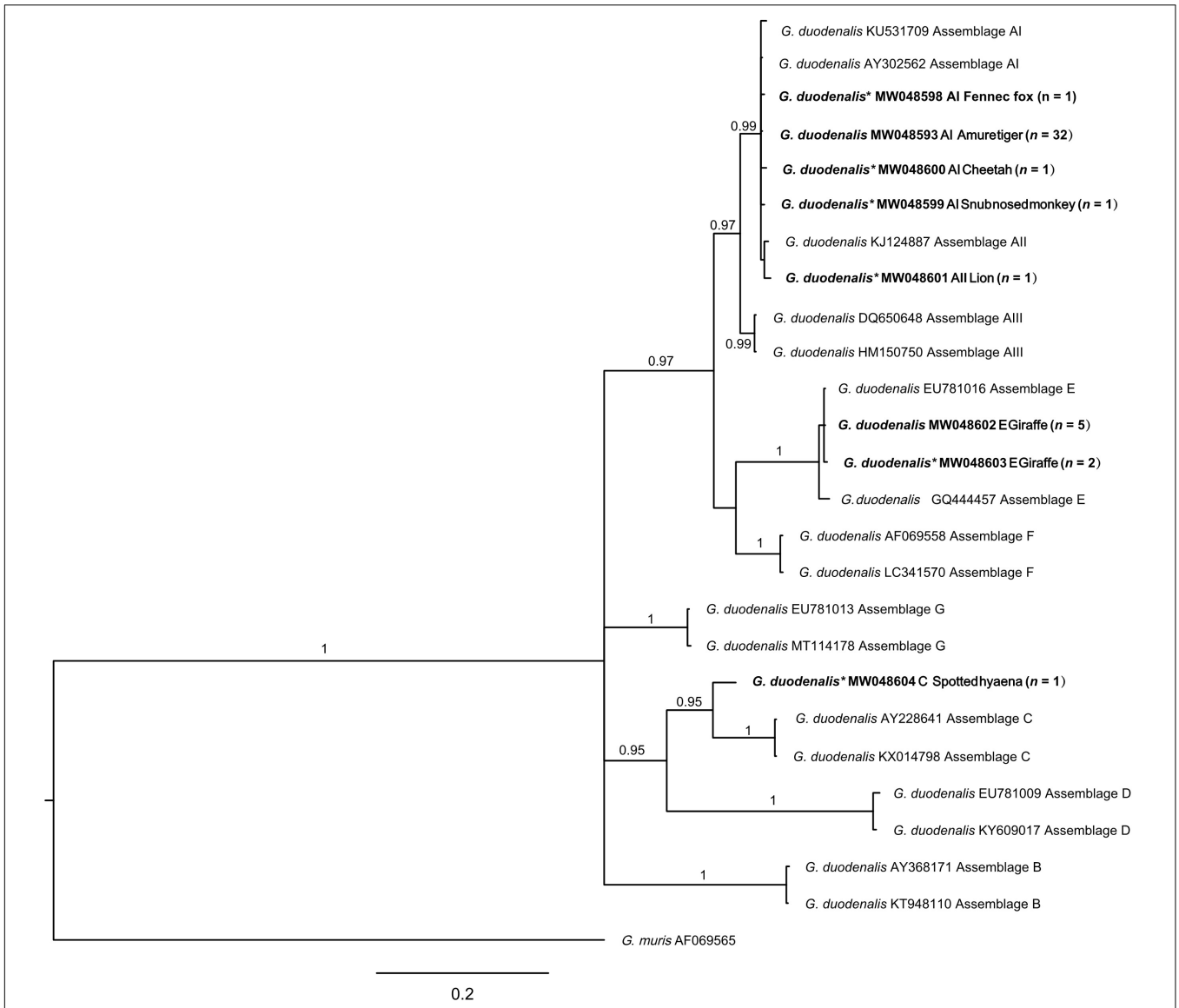


FIGURE 3 | Relationships among *Giardia* taxa inferred from the phylogenetic analysis of partial triose-phosphate isomerase gene (*tpi*) sequence data by Bayesian inference (BI). Posterior probabilities are indicated at all major nodes. Bold font indicates *Giardia* species or genotypes characterized from fecal DNA samples in this study. In parentheses are the numbers of samples representing a particular species, genotype, and sequence (GenBank accession numbers indicated). Novel genotypes (*). Scale bar represents the number of substitutions per site. All groups were strongly supported (*pp* = 0.96–1.00). *pp* < 0.95 were not shown.

in this study is the first such published record. Similarly, red-and-green macaw (*Ara chloropterus*) is the first host record of genotype MJ13. Predominant genotypes BEB6, D, and SC02 were also found in water samples (Ayed et al., 2012; Li et al., 2012; Huang et al., 2017; Li W. et al., 2018), indicating that they might spread via *E. bieneusi* spores–contaminated water.

Phylogenetic analyses revealed that novel genotype SH_deer1 clustered with genotypes CAM1 (camel), horse 2 (horse), MAY 1 (human), and Nig3 (human), falling into group 6. Previously, genotypes in this group were predominantly found in animals. Thus, group 6 was typically considered as the host-adapted group. However, with more genotypes from this group identified in humans (Akinbo et al., 2012; Qi et al., 2018), demonstrating that group 6 revealed zoonotic potential. Additionally, we have also created a phylogeny using all nearly 600 unique genotypes from all published studies employing complete *ITS* sequences, with the aim of assessing the relationships of the genotypes and the validity of groups (Zhang et al., 2021), proving the zoonotic potential of group 6. The overall results indicate that wildlife carrying zoonotic genotypes have the capacity to transmit from them to humans.

G. duodenalis

In the present study, 48 wildlife tested positive for *G. duodenalis* with a total prevalence of 26.37% (48/182), which was higher than that of *Cryptosporidium* (1.6%; 3/182) and *E. bieneusi* (11.5%;

21/182), indicating that *G. duodenalis* is more widely spread than the other two microbes. The prevalences of *G. duodenalis* in Shanghai Wild Animal Park and in Zhejiang were 30.8% (40/130) and 15.4% (8/52), respectively, both of which were higher than that in a number of studies of *G. duodenalis* globally (Matsubayashi et al., 2005; Lallo et al., 2009; Beck et al., 2011b; Majewska et al., 2012; Oates et al., 2012; Aghazadeh et al., 2015; Reboredo-Fernández et al., 2015; Adriana et al., 2016; Mynarova et al., 2016; Mateo et al., 2017; Helmy et al., 2018). Additionally, the prevalence of *G. duodenalis* in wild animals worldwide ranged from 1.1% in zoo in Japan (Matsubayashi et al., 2005) to 29.0% in Zagreb zoo in Croatia (Beck et al., 2011a); 30.8% (40/130) here in wildlife in Shanghai is the highest prevalence around the world. The overall results indicate relatively high *G. duodenalis* infections in zoo animals in this study. However, it cannot be entirely excluded that *G. duodenalis* cysts might only pass through the gastrointestinal tract (pseudoparasitism), as identification of *G. duodenalis* DNA from fecal samples is not a direct evidence of infection.

In total, three assemblages A, C, and E of *G. duodenalis* were characterized. Zoonotic assemblage A is predominant (75%; 36/48) in this study, followed by genotype E (14.58%; 7/48) and C (2.08%; 1/48). Genotype A has been reported in humans and a large number of animal species with the capacity of cross-species transmission (Ryan and Zahedi, 2019). In this study, assemblage E was mostly identified in giraffe, except for one positivity in kangaroo. This is the first time that kangaroo was recorded in the *G. duodenalis* assemblage E. This assemblage has been mainly reported in hoofed animals, but it was also detected in human specimens in Brazil (Fantinatti et al., 2016), Egypt (Foronda et al., 2008), and Australia (Zahedi et al., 2017), posing less risk to public health. Phylogenetically, the novel *tpi* sequence found in spotted hyena (*Crocuta crocuta*) clustered with assemblage C (**Figure 3**), which has been frequently reported in canids and occasionally reported in humans (Hopkins et al., 1997; Monis et al., 1998). The overall results indicate that zoo animals can harbor zoonotic *G. duodenalis* and potentially act as a host reservoir for human infections of giardiasis.

CONCLUSION

Exploring the genetic composition of *Cryptosporidium* species *E. bieneusi* and *G. duodenalis* populations in animals and humans is important for understanding transmission patterns of enteric disease and for its prevention and control. By conducting the present molecular-phylogenetic investigation of three pathogens target sequences derived from fecal samples ($n = 182$) from zoo animals in China, we found (phylogenetically) a novel species of *Cryptosporidium*. We also identified genotypes or assemblages (*E. bieneusi*: BEB6, D, MJ13, SC02, SH_ch1, SH_deer1, and type IV; *G. duodenalis*: A, C, and E), all of which have zoonotic potential. The overall results indicate that wildlife carrying zoonotic *E. bieneusi* and *G. duodenalis* can potentially transmit the pathogens to humans, thus posting a public health risk.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in GenBank under the accession numbers, *Cryptosporidium*: MW168840-MW168842; *E. bieneusi*: MT895455- MT895461 *and G. duodenalis*: MW048593-MW048601.

ETHICS STATEMENT

Sample collections were carried out by colleges from the Shanghai Wild Animal Park and Zhejiang Zoo. Animals were handled in accordance with the Animal Ethics Procedures and Guidelines of the People's Republic of China.

AUTHOR CONTRIBUTIONS

YZ, RM, LY, ZC, YF, XC, YH, and HG: sample collection. YZ and LY: designed the study and performed the experiments. YZ: analysis and interpretation and wrote the manuscript. ZC: review the draft and supervision of project. All authors read and approved the final version of the manuscript.

FUNDING

This study was supported in part by the National Key Research and Development Program of China (Grant No. 2017YFD0500401), Shanghai Science and Technology Commission Scientific Research Project (Grant No. 20140900400), National Risk Assessment Project for Quality and Safety of Agricultural Products (Grant No. GJFP2019027), International Postdoctoral Exchange Fellowship Program (Talent-Introduction Program), Scientific Research Project of Special Training Program for Scientific and Technological Talents of Ethnic Minorities in Xinjiang (Grant No. 2020D03030), China Postdoctoral Science Foundation (2021M693455), and Shanghai Agriculture Applied Technology Development Program, China (Grant No. G20050304).

ACKNOWLEDGMENTS

We thank colleagues from zoos in Zhejiang and Shanghai for donating fecal samples.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.692837/full#supplementary-material

Supplementary Table 1 | The information regarding fecal samples (sample codes are given) collected from various species of wildlife (host and Latin name are given) located in Zhejiang and Shanghai from May 2018 to August 2020.

Supplementary Table 2 | Pairwise comparison of sequence differences in the small subunit of nuclear ribosomal RNA gene (*SSU* rDNA; 735 bp) among *Cryptosporidium* species from samples tested in the present study (cf. **Figure 1**).

REFERENCES

Abeywardena, H., Jex, A. R., and Gasser, R. B. (2015). A perspective on *Cryptosporidium* and *Giardia*, with an emphasis on bovines and recent epidemiological findings. *Adv. Parasitol.* 88, 243–301. doi: 10.1016/bs.apar.2015.02.001

Adriana, G., Zsuzsa, K., Oana, D. M., Gherman, C. M., and Viorica, M. (2016). *Giardia duodenalis* genotypes in domestic and wild animals from Romania identified by PCR-RFLP targeting the *gdh* gene. *Vet. Parasitol.* 217, 71–75. doi: 10.1016/j.vetpar.2015.10.017

Aghazadeh, M., Elson-Riggins, J., Reljic, S., De Ambrogi, M., Huber, D., Majnaric, D., et al. (2015). Gastrointestinal parasites and the first report of *Giardia* spp. in a wild population of European brown bears (*Ursus arctos*) in Croatia. *Vet. Arhiv.* 85, 201–210.

Akinbo, F. O., Okaka, C. E., Omoregie, R., Dearen, T., Leon, E. T., and Xiao, L. H. (2012). Molecular epidemiologic characterization of *Enterocytozoon bieneusi* in HIV-infected persons in Benin city, Nigeria. *Am. J. Trop. Med. Hyg.* 86, 441–445. doi: 10.4269/ajtmh.2012.11-0548

Amer, S., Kim, S., Han, J. I., and Na, K. J. (2019). Prevalence and genotypes of *Enterocytozoon bieneusi* in wildlife in Korea: a public health concern. *Parasit. Vectors* 12:160. doi: 10.1186/s13071-019-3427-6

Amman, B. R., Jones, M. E., Sealy, T. K., Uebelhoer, L. S., Schuh, A. J., Bird, B. H., et al. (2015). Oral shedding of Marburg virus in experimentally infected Egyptian fruit bats (*Rousettus aegyptiacus*). *J. Wildl. Dis.* 51, 113–124. doi: 10.7589/2014-08-198

Ayed, L. B., Yang, W., Widmer, G., Cama, V., Ortega, Y., and Xiao, L. H. (2012). Survey and genetic characterization of wastewater in Tunisia for *Cryptosporidium* spp., *Giardia duodenalis*, *Enterocytozoon bieneusi*, *Cyclospora cayetanensis* and *Eimeria* spp. *J. Water Health* 10, 431–444. doi: 10.2166/wh.2012.204

Baldursson, S., and Karanis, P. (2011). Waterborne transmission of protozoan parasites: review of worldwide outbreaks-an update 2004-2010. *Water Res.* 45, 6603–6614. doi: 10.1016/j.watres.2011.10.013

Beck, R., Sprong, H., Bata, I., Lucinger, S., Pozio, E., and Cacciò, S. M. (2011a). Prevalence and molecular typing of *Giardia* spp. in captive mammals at the zoo of Zagreb, Croatia. *Vet. Parasitol.* 175, 40–46. doi: 10.1016/j.vetpar.2010.09.026

Beck, R., Sprong, H., Lucinger, S., Pozio, E., and Caccio, S. M. (2011b). A large survey of Croatian wild mammals for *Giardia duodenalis* reveals a low prevalence and limited zoonotic potential. *Vector Borne Zoonot.* 11, 1049–1055. doi: 10.1089/vbz.2010.0113

Blander, J. M., Longman, R. S., Iliev, I. D., Sonnenberg, G. F., and Artis, D. (2017). Regulation of inflammation by microbiota interactions with the host. *Nat. Immunol.* 18:851. doi: 10.1038/ni.3780

Broglia, A., Weitzel, T., Harms, G., Cacció, S. M., and Nöckler, K. (2013). Molecular typing of *Giardia duodenalis* isolates from German travellers. *Parasitol. Res.* 112, 3449–3456. doi: 10.1007/s00436-013-3524-y

CDC (1999). Outbreak of West Nile-like viral encephalitis-New York, 1999. *MMWR Morb. Mortal. Wkly. Rep.* 48, 845–849.

CDC (2000). Update: West Nile Virus activity-Eastern United States, 2000. *MMWR Morb. Mortal. Wkly. Rep.* 49, 1044–1047.

Checkley, W., White, A. C. J., Jaganath, D., Arrowood, M. J., Chalmers, R. M., Chen, X. M., et al. (2015). A review of the global burden, novel diagnostics, therapeutics, and vaccine targets for *Cryptosporidium*. *Lancet Infect. Dis.* 15, 85–94. doi: 10.1016/S1473-3099(14)70772-8

Chingwaru, W., and Vidmar, J. (2018). Culture, myths and panic: three decades and beyond with an HIV/AIDS epidemic in Zimbabwe. *Glob. Public Health* 13, 249–264. doi: 10.1080/17441692.2016.1215485

Cunningham, A. A., Daszak, P., and Wood, J. L. N. (2017). One health, emerging infectious diseases and wildlife: two decades of progress? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 372:20160167. doi: 10.1098/rstb.2016.0167

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9:772. doi: 10.1038/nmeth.2109

Decraene, V., Lebbad, M., Botero-Kleiven, S., Gustavsson, A. M., and Löfdahl, M. (2012). First reported foodborne outbreak associated with microsporidia, Sweden, October 2009. *Epidemiol. Infect.* 140, 519–527. doi: 10.1017/S095026881100077X

Deng, L., Li, W., Yu, X., Gong, C., Liu, X., Zhong, Z., et al. (2016a). First report of the human-pathogenic *Enterocytozoon bieneusi* from red-bellied tree squirrels (*Callosciurus erythraeus*) in Sichuan, China. *PLoS One* 11:e0163605. doi: 10.1371/journal.pone.0163605

Deng, L., Li, W., Zhong, Z., Gong, C., Liu, X., Huang, X., et al. (2016b). Molecular characterization and multilocus genotypes of *Enterocytozoon bieneusi* among horses in southwestern China. *Parasit. Vectors* 9:561. doi: 10.1186/s13071-016-1844-3

Fantinatti, M., Bello, A. R., Fernandes, O., and Da-Cruz, A. M. (2016). Identification of *Giardia lamblia* assemblage E in humans points to a new anthropozoonotic cycle. *J. Infect. Dis.* 214, 1256–1259. doi: 10.1093/infdis/jiw361

Feng, Y. Y., Ryan, U. M., and Xiao, L. H. (2018). Genetic diversity and population structure of *Cryptosporidium*. *Trends Parasitol.* 34, 997–1011. doi: 10.1016/j.pt.2018.07.009

Feng, Y., and Xiao, L. (2011). Zoonotic potential and molecular epidemiology of *Giardia* species and giardiasis. *Clin. Microbiol. Rev.* 24, 110–140. doi: 10.1128/CMR.00033-10

Field, H. E., Mackenzie, J. S., and Daszak, P. (2007). Henipaviruses: emerging paramyxoviruses associated with fruit bats. *Curr. Top. Microbiol. Immunol.* 315, 133–159. doi: 10.1007/978-3-540-70962-6_7

Foronda, P., Bargues, M. D., Abreu-Acosta, N., Periago, M. V., Valero, M. A., Valladares, B., et al. (2008). Identification of genotypes of *Giardia intestinalis* of human isolates in Egypt. *Parasitol. Res.* 103, 1177–1181. doi: 10.1007/s00436-008-1113-2

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., et al. (1999). Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* 397, 436–441. doi: 10.1038/17130

GBD (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015. *Lancet* 388, 1545–1602.

Gelanew, T., Lalle, M., Hailu, A., Pozio, E., and Cacciò, S. M. (2007). Molecular characterization of human isolates of *Giardia duodenalis* from Ethiopia. *Acta Trop.* 102, 92–99. doi: 10.1016/j.actatropica.2007.04.003

Haghi, M. M., Khorshidvand, Z., Khazaei, S., Foroughi-Parvar, F., and Ghasemikhah, R. (2020). *Cryptosporidium* animal species in Iran: a systematic review and meta-analysis. *Trop. Med. Health* 48:97. doi: 10.1186/s41182-020-00278-9

Hallen-Adams, H. E., and Suhr, M. J. (2017). Fungi in the healthy human gastrointestinal tract. *Virulence* 8, 352–358. doi: 10.1080/21505594.2016.1247140

Helmy, Y. A., Spierling, N. G., Schmidt, S., Rosenfeld, U. M., Reil, D., Imholt, C., et al. (2018). Occurrence and distribution of *Giardia* species in wild rodents in Germany. *Parasit. Vectors* 11:213. doi: 10.1186/s13071-018-2802-z

Hopkins, R. M., Meloni, B. P., Groth, D. M., Wetherall, J. D., Reynoldson, J. A., and Thompson, R. C. (1997). Ribosomal RNA sequencing reveals differences between the genotypes of *Giardia* isolates recovered from humans and dogs living in the same locality. *J. Parasitol.* 83, 44–51. doi: 10.2307/3284315

Huang, C., Hu, Y., Wang, L., Wang, Y., Li, N., Guo, Y., et al. (2017). Environmental transport of emerging human-pathogenic *Cryptosporidium* species and subtypes through combined sewer overflow and wastewater. *Appl. Environ. Microbiol.* 83:e00682-17. doi: 10.1128/AEM.00682-17

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755. doi: 10.1093/bioinformatics/17.8.754

Jiang, J., Alderisio, K. A., and Xiao, L. (2005). Distribution of *Cryptosporidium* genotypes in storm event water samples from three watersheds in New York. *Appl. Environ. Microbiol.* 71, 4446–4454. doi: 10.1128/AEM.71.8.4446-4454.2005

Jones, M. E. B., Schuh, A. J., Amman, B. R., Sealy, T. K., Zaki, S. R., Nichol, S. T., et al. (2015). Experimental inoculation of Egyptian rousette bats (*Rousettus aegyptiacus*) with viruses of the *Ebolavirus* and *Marburgvirus* genera. *Viruses* 7, 3420–3442. doi: 10.3390/v7072779

Karanis, P., Kourenti, C., and Smith, H. (2007). Waterborne transmission of protozoan parasites: a worldwide review of outbreaks and lessons learnt. *J. Water Health* 5, 1–38. doi: 10.2166/wh.2006.002

Karim, M. R., Dong, H., Li, T., Yu, F., Li, D., Zhang, L., et al. (2015). Predomination and new genotypes of *Enterocytozoon bieneusi* in captive nonhuman primates

in zoos in China: high genetic diversity and zoonotic significance. *PLoS One* 10:e0117991. doi: 10.1371/journal.pone.0117991

Katzwinkel-Wladarsch, S., Lieb, M., Helse, W., Löscher, T., and Rinder, H. (1996). Direct amplification and species determination of microsporidian DNA from stool specimens. *Trop. Med. Int. Health* 1, 373–378. doi: 10.1046/j.1365-3156.1996.d01-51.x

Koehler, A. V., Haydon, S. R., Jex, A. R., and Gasser, R. B. (2016). *Cryptosporidium* and *Giardia* taxa in faecal samples from animals in catchments supplying the city of Melbourne with drinking water (2011 to 2015). *Parasit. Vectors* 9:315. doi: 10.1186/s13071-016-1607-1

Koehler, A. V., Rashid, M. H., Zhang, Y., Vaughan, J. L., Gasser, R. B., and Jabbar, A. (2018). First cross-sectional, molecular epidemiological survey of *Cryptosporidium*, *Giardia* and *Enterocytozoon* in alpaca (*Vicugna pacos*) in Australia. *Parasit. Vectors* 11:498. doi: 10.1186/s13071-018-3055-6

Koehler, A. V., Zhang, Y., Wang, T., Haydon, S. R., and Gasser, R. B. (2020). Multiplex PCRs for the specific identification of marsupial and deer species from scats as a basis for non-invasive epidemiological studies of parasites. *Parasit. Vectors* 13:144. doi: 10.1186/s13071-020-04009-1

Kruse, H., Kirkemo, A. M., and Handeland, K. (2005). Wildlife as source of zoonotic infections. *Emerg. Infect. Dis.* 10, 2067–2072. doi: 10.3201/eid1012.040707

Lallo, M. A., Pereira, A., Araujo, R., Favorito, S. E., Bertolla, P., and Bondan, E. F. (2009). Occurrence of *Giardia*, *Cryptosporidium* and microsporidia in wild animals from a deforestation area in the state of Sao Paulo, Brazil. *Cienc. Rural* 39, 1465–1470. doi: 10.1590/S0103-84782009005000085

Lanata, C. F., Fischer-Walker, C. L., Olascoaga, A. C., Torres, C. X., Aryee, M. J., and Black, R. E. (2013). Global causes of diarrheal disease mortality in children $<$ 5 years of age: a systematic review. *PLoS One* 8:e72788. doi: 10.1371/journal.pone.0072788

Lemos, D. R. Q., Franco, A. R., de Oliveir Garcia, M. H., Pastor, D., Bravo-Alcântara, P., de Moraes, J. C., et al. (2018). Risk analysis for the reintroduction and transmission of measles in the post-elimination period in the Americas. *Rev. Panam. Salud Pública* 41:e157. doi: 10.26633/RPSP.2017.157

Leśniańska, K., Perec-Matysiak, A., Hildebrand, J., Bunkowska-Gawlik, K., Pirog, A., and Popiolek, M. (2016). *Cryptosporidium* spp. and *Enterocytozoon bieneusi* in introduced raccoons (*Procyon lotor*)-first evidence from Poland and Germany. *Parasitol. Res.* 115, 4535–4541. doi: 10.1007/s00436-016-5245-5

Li, J., Qi, M., Chang, Y., Wang, R., Li, T., Dong, H., et al. (2015). Molecular characterization of *Cryptosporidium* spp., *Giardia duodenalis*, and *Enterocytozoon bieneusi* in captive wildlife at Zhengzhou Zoo, China. *J. Eukaryot. Microbiol.* 62, 833–839. doi: 10.1111/jeu.12269

Li, N., Ayinmode, A. B., Zhang, H. W., Feng, Y. Y., and Xiao, L. H. (2018). Host-adapted *Cryptosporidium* and *Enterocytozoon bieneusi* genotypes in straw-colored fruit bats in Nigeria. *Int. J. Parasitol. Parasites Wildl.* 8, 19–24. doi: 10.1016/j.ijppaw.2018.12.001

Li, N., Xiao, L. H., Wang, L., Zhao, S., Zhao, X., Duan, L., et al. (2012). Molecular surveillance of *Cryptosporidium* spp., *Giardia duodenalis*, and *Enterocytozoon bieneusi* by genotyping and subtyping parasites in wastewater. *PLoS Negl. Trop. Dis.* 6:e1809. doi: 10.1371/journal.pntd.0001809

Li, W., Deng, L., Wu, K., Huang, X., Song, Y., Su, H., et al. (2017). Presence of zoonotic *Cryptosporidium scrofarum*, *Giardia duodenalis* assemblage A and *Enterocytozoon bieneusi* genotypes in captive Eurasian wild boars (*Sus scrofa*) in China: potential for zoonotic transmission. *Parasit. Vectors* 10:10. doi: 10.1186/s13071-016-1942-2

Li, W., Deng, L., Yu, X., Zhong, Z., Wang, Q., Liu, X., et al. (2016). Multilocus genotypes and broad host-range of *Enterocytozoon bieneusi* in captive wildlife at zoological gardens in China. *Parasit. Vectors* 9:395. doi: 10.1186/s13071-016-1668-1

Li, W., Feng, Y. Y., and Santín, M. (2019a). Host specificity of *Enterocytozoon bieneusi* and public health implications. *Trends Parasitol.* 35, 436–451. doi: 10.1016/j.pt.2019.04.004

Li, W., Feng, Y. Y., Zhang, L. X., and Xiao, L. H. (2019b). Potential impacts of host specificity on zoonotic or interspecies transmission of *Enterocytozoon bieneusi*. *Infect. Genet. Evol.* 75:104033. doi: 10.1016/j.meegid.2019.104033

Li, W., Li, Y., Song, M., Lu, Y., Yang, J., Tao, W., et al. (2015). Prevalence and genetic characteristics of *Cryptosporidium*, *Enterocytozoon bieneusi* and *Giardia duodenalis* in cats and dogs in Heilongjiang province, China. *Vet. Parasitol.* 208, 125–134. doi: 10.1016/j.vetpar.2015.01.014

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J. H., et al. (2005). Bats are natural reservoirs of SARS-like coronaviruses. *Science* 310, 676–679. doi: 10.1126/science.1118391

Li, W., Zhong, Z., Song, Y., Gong, C., Deng, L., Cao, Y., et al. (2018). Human-pathogenic *Enterocytozoon bieneusi* in captive giant pandas (*Ailuropoda melanoleuca*) in China. *Sci. Rep.* 8:6590. doi: 10.1038/s41598-018-25096-2

Liguori, G., Lamas, B., Richard, M. L., Brandi, G., da Costa, G., Hoffmann, T. W., et al. (2015). Fungal dysbiosis in mucosa-associated microbiota of Crohn's disease patients. *J. Crohns Colitis* 10, 296–305. doi: 10.1093/ecco-jcc/jjv209

Liu, A., Yang, F., Shen, Y., Zhang, W., Wang, R., Zhao, W., et al. (2014). Genetic analysis of the *gdh* and *bg* genes of animal-derived *Giardia duodenalis* isolates in northeastern China and evaluation of zoonotic transmission potential. *PLoS One* 9:e95291. doi: 10.1371/journal.pone.0095291

Liu, L., Oza, S., Hogan, D., Perin, J., Rudan, I., Lawn, J. E., et al. (2015). Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. *Lancet* 385, 430–440. doi: 10.1016/S0140-6736(14)61698-6

Liu, P., Jiang, J. Z., Wan, X. F., Hua, Y., Li, L., Zhou, J., et al. (2020). Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? *PLoS Pathog.* 16:e1008421. doi: 10.1371/journal.ppat.1008421

Majewska, A. C., Solarczyk, P., Moskwa, B., Cabaj, W., Jankowska, W., and Nowosad, P. (2012). *Giardia* prevalence in wild cervids in Poland. *Ann. Parasitol.* 58, 207–209.

Maniga, J. N., Aliero, A. A., Ibrahim, N., Okech, M. A., and Mack, M. C. (2018). *Plasmodium falciparum* malaria clinical and parasitological outcomes after in-vivo artemether-lumefantrine (AL) treatment at Bushenyi district Uganda. *Int. J. Trop. Dis. Health* 29, 1–17. doi: 10.9734/IJTDH/2018/39642

Mateo, M., de Mingo, M. H., de Lucio, A., Morales, L., Balseiro, A., Espí, A., et al. (2017). Occurrence and molecular genotyping of *Giardia duodenalis* and *Cryptosporidium* spp. in wild mesocarnivores in Spain. *Vet. Parasitol.* 235, 86–93. doi: 10.1016/j.vetpar.2017.01.016

Matsubayashi, M., Takami, K., Kimata, I., Nakanishi, T., Tani, H., Sasai, K., et al. (2005). Survey of *Cryptosporidium* spp. and *Giardia* spp. infections in various animals at a zoo in Japan. *J. Zoo Wildl. Med.* 36, 331–335. doi: 10.1638/04-032.1

Monis, P. T., Andrews, R. H., Mayrhofer, G., Mackrill, J., Kulda, J., Isaac-Renton, J. L., et al. (1998). Novel lineages of *Giardia intestinalis* identified by genetic analysis of organisms isolated from dogs in Australia. *Parasitology* 116, 7–19. doi: 10.1017/S0031182097002011

Mynarova, A., Foitova, I., Kvac, M., Kvetonova, D., Rost, M., Morrogh-Bernard, H., et al. (2016). Prevalence of *Cryptosporidium* spp., *Enterocytozoon bieneusi*, *Encephalitozoon* spp. and *Giardia intestinalis* in wild, semi-wild and captive orangutans (*Pongo abelii* and *Pongo pygmaeus*) on Sumatra and Borneo, Indonesia. *PLos One* 11:e0152771. doi: 10.1371/journal.pone.0152771

Oates, S. C., Miller, M. A., Hardin, D., Conrad, P. A., Melli, A., Jessup, D. A., et al. (2012). Prevalence, environmental loading, and molecular characterization of *Cryptosporidium* and *Giardia* isolates from domestic and wild animals along the central California coast. *Appl. Environ. Microbiol.* 78, 8762–8772. doi: 10.1128/AEM.02422-12

Qi, M., Li, J., Zhao, A., Cui, Z., Wei, Z., Jing, B., et al. (2018). Host specificity of *Enterocytozoon bieneusi* genotypes in bactrian camels (*Camelus bactrianus*) in China. *Parasit. Vectors* 11:219. doi: 10.1186/s13071-018-2793-9

Reboredo-Fernández, A., Ares-Mazás, E., Cacciò, S. M., and Gómez-Couso, H. (2015). Occurrence of *Giardia* and *Cryptosporidium* in wild birds in Galicia (northwest Spain). *Parasitology* 142, 917–925. doi: 10.1017/S0031182015000049

Ryan, U. M., and Cacciò, S. M. (2013). Zoonotic potential of *Giardia*. *Int. J. Parasitol.* 43, 943–956. doi: 10.1016/j.ijpara.2013.06.001

Ryan, U., and Zahedi, A. (2019). Molecular epidemiology of giardiasis from a veterinary perspective. *Adv. Parasitol.* 106, 209–254. doi: 10.1016/bs.apar.2019.07.002

Santín, M., and Fayer, R. (2009). *Enterocytozoon bieneusi* genotype nomenclature based on the internal transcribed spacer sequence: a consensus. *J. Eukaryot. Microbiol.* 56, 34–38. doi: 10.1111/j.1550-7408.2008.00380.x

Santín, M., and Fayer, R. (2011). Microsporidiosis: *Enterocytozoon bieneusi* in domesticated and wild animals. *Res. Vet. Sci.* 90, 363–371. doi: 10.1016/j.rvsc.2010.07.014

Santín, M., Trout, J. M., and Fayer, R. (2009). A longitudinal study of *Giardia duodenalis* genotypes in dairy cows from birth to 2 years of age. *Vet. Parasitol.* 162, 40–45. doi: 10.1016/j.vetpar.2009.02.008

Scalia, L. A. M., Fava, N. M. N., Soares, R. M., Limongi, J. E., da Cunha, M. J. R., Pena, I. F., et al. (2016). Multilocus genotyping of *Giardia duodenalis* in Brazilian children. *Trans. R. Soc. Trop. Med. Hyg.* 110, 343–349. doi: 10.1093/trstmh/trw036

Štrkolcová, G., Mad'án, M., Hinney, B., Goldová, M., Mojžišová, J., and Halánová, M. (2015). Dog's genotype of *Giardia duodenalis* in human: first evidence in Europe. *Acta Parasitol.* 60, 769–799. doi: 10.1515/ap-2015-0113

Sulaiman, I. M., Fayer, R., Bern, C., Gilman, R. H., and Xiao, L. (2003). Triosephosphate isomerase gene characterization and potential zoonotic transmission of *Giardia duodenalis*. *Emerg. Infect. Dis.* 9, 1444–1452. doi: 10.3201/eid0911.030084

Traub, R. J., Monis, P. T., Robertson, I., Irwin, P., Mencke, N., and Thompson, R. C. (2004). Epidemiological and molecular evidence supports the zoonotic transmission of *Giardia* among humans and dogs living in the same community. *Parasitology* 128, 253–262. doi: 10.1017/S0031182003004505

Walker, C. L. F., Aryee, M. J., Boschi-Pinto, C., and Black, R. E. (2012). Estimating diarrhea mortality among young children in low and middle income countries. *PLoS One* 7:e29151. doi: 10.1371/journal.pone.0029151

Wu, J., Han, J. Q., Shi, L. Q., Zou, Y., Li, Z., Yang, J. F., et al. (2018). Prevalence, genotypes, and risk factors of *Enterocytozoon bieneusi* in Asiatic black bear (*Ursus thibetanus*) in Yunnan province, southwestern China. *Parasitol. Res.* 117, 1139–1145. doi: 10.1007/s00436-018-5791-0

Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J.-J., et al. (2020). Isolation of SARS-CoV-2-related coronavirus from *Malayan pangolins*. *Nature* 583, 286–289. doi: 10.1038/s41586-020-2313-x

Xiao, L. H. (2010). Molecular epidemiology of cryptosporidiosis: an update. *Exp. Parasitol.* 124, 80–89. doi: 10.1016/j.exppara.2009.03.018

Xiao, L. H., Bern, C., Limor, J., Sulaiman, I., Roberts, J., Checkley, W., et al. (2001). Identification of 5 types of *Cryptosporidium* parasites in children in Lima, Peru. *J. Infect. Dis.* 183, 492–497. doi: 10.1086/318090

Xiao, L. H., Fayer, R., Ryan, U., and Upton, S. J. (2004). *Cryptosporidium* taxonomy: recent advances and implications for public health. *Clin. Microbiol. Rev.* 17, 72–97. doi: 10.1128/CMR.17.1.72-97.2004

Xiao, L. H., Morgan, U. M., Limor, J., Escalante, A., Arrowood, M., Shulaw, W., et al. (1999). Genetic diversity within *Cryptosporidium parvum* and related *Cryptosporidium* species. *Appl. Environ. Microbiol.* 65, 3386–3391. doi: 10.1128/AEM.65.8.3386-3391.1999

Yu, F., Wu, Y., Li, T., Cao, J., Wang, J., Hu, S., et al. (2017). High prevalence of *Enterocytozoon bieneusi* zoonotic genotype D in captive golden snub-nosed monkey (*Rhinopithecus roxellanae*) in zoos in China. *BMC Vet. Res.* 13:158. doi: 10.1186/s12917-017-1084-6

Yu, Z., Wen, X., Huang, X., Yang, R., Guo, Y., Feng, Y. Y., et al. (2020). Molecular characterization and zoonotic potential of *Enterocytozoon bieneusi*, *Giardia duodenalis* and *Cryptosporidium* sp. in farmed masked palm civets (*Paguma larvata*) in southern China. *Parasit. Vectors* 13:403. doi: 10.1186/s13071-020-04274-0

Zahedi, A., Field, D., and Ryan, U. (2017). Molecular typing of *Giardia duodenalis* in humans in Queensland – first report of assemblage E. *Parasitology* 144, 1154–1161. doi: 10.1017/S0031182017000439

Zhang, Y. (2019). *Molecular Epidemiological Investigations of Enterocytozoon Bieneusi*. Ph.D. thesis. Melbourne, VIC: The University of Melbourne.

Zhang, Y., Koehler, A. V., Wang, T., and Gasser, R. B. (2021). *Enterocytozoon bieneusi* of animals—with an 'Australian twist'. *Adv. Parasitol.* 111, 1–73. doi: 10.1016/bs.apar.2020.10.001

Zhang, Y., Koehler, A. V., Wang, T., Cunliffe, D., and Gasser, R. B. (2019). *Enterocytozoon bieneusi* genotypes in cats and dogs in Victoria, Australia. *BMC Microbiol.* 19:183. doi: 10.1186/s12866-019-1563-y

Zhang, Y., Koehler, A. V., Wang, T., Haydon, S. R., and Gasser, R. B. (2018a). *Enterocytozoon bieneusi* genotypes in cattle on farms located within a water catchment area. *J. Eukaryot. Microbiol.* 66, 553–559. doi: 10.1111/jeu.12696

Zhang, Y., Koehler, A. V., Wang, T., Haydon, S. R., and Gasser, R. B. (2018b). First detection and genetic characterisation of *Enterocytozoon bieneusi* in wild deer in Melbourne's water catchments in Australia. *Parasit. Vectors* 11:2. doi: 10.1186/s13071-017-2577-7

Zhang, Y., Koehler, A. V., Wang, T., Haydon, S. R., and Gasser, R. B. (2018c). New operational taxonomic units of *Enterocytozoon* in three marsupial species. *Parasit. Vectors* 11:371. doi: 10.1186/s13071-018-2954-x

Zhang, Y., Koehler, A. V., Wang, T., Robertson, G. J., Bradbury, R. S., and Gasser, R. B. (2018d). *Enterocytozoon bieneusi* genotypes in people with gastrointestinal disorders in Queensland and Western Australia. *Infect. Genet. Evol.* 65, 293–299. doi: 10.1016/j.meegid.2018.08.006

Zhang, Y., Koehler, A. V., Wang, T., Robertson, G. J., Bradbury, R. S., and Gasser, R. B. (2018e). *Enterocytozoon bieneusi* genotypes in people with gastrointestinal disorders in Queensland and Western Australia. *Infect. Genet. Evol.* 65, 293–299. doi: 10.1016/j.meegid.2018.08.006

Zhang, Y., Mi, R. S., Yang, J. B., Wang, J. X., and Chen, Z. G. (2020). *Enterocytozoon bieneusi* genotypes in farmed goats and sheep in Ningxia, China. *Infect. Genet. Evol.* 89:104559. doi: 10.1016/j.meegid.2020.104559

Zhong, Z., Li, W., Deng, L., Song, Y., Wu, K., Tian, Y., et al. (2017). Multilocus genotyping of *Enterocytozoon bieneusi* derived from nonhuman primates in southwest China. *PLoS One* 12:e0176926. doi: 10.1371/journal.pone.0176926

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 September 2021
doi: 10.3389/fgene.2021.723773

Analysis of Microorganism Diversity in *Haemaphysalis longicornis* From Shaanxi, China, Based on Metagenomic Sequencing

Runlai Cao[1], Qiaoyun Ren[1], Jin Luo[1], Zhancheng Tian[1], Wenge Liu[1], Bo Zhao[2], Jing Li[3], Peiwen Diao[1], Yangchun Tan[1], Xiaofei Qiu[1], Gaofeng Zhang[1], Qilin Wang[1], Guiquan Guan[1], Jianxun Luo[1], Hong Yin[1,4] and Guangyuan Liu[1]**

[1] State Key Laboratory of Veterinary Etiological Biology, Key Laboratory of Veterinary Parasitology of Gansu Province, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Lanzhou, China, [2] Gansu Agriculture Technology College, Lanzhou, China, [3] Animal Disease Prevention and Control Center of Qinghai Province, Xining, China, [4] Jiangsu Co-innovation Center for the Prevention and Control of Important Animal Infectious Disease and Zoonose, Yangzhou University, Yangzhou, China

OPEN ACCESS

Edited by:
Jun-Hu Chen,
National Institute of Parasitic Diseases, China

Reviewed by:
Zhijun Yu,
Hebei Normal University, China
Luca Ermini,
King Abdullah University of Science and Technology, Saudi Arabia

***Correspondence:**
Qiaoyun Ren
renqiaoyun@gmail.com
Guangyuan Liu
liuguangyuan@caas.cn

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: 11 June 2021
Accepted: 13 August 2021
Published: 09 September 2021

Citation:
Cao R, Ren Q, Luo J, Tian Z, Liu W, Zhao B, Li J, Diao P, Tan Y, Qiu X, Zhang G, Wang Q, Guan G, Luo J, Yin H and Liu G (2021) Analysis of Microorganism Diversity in Haemaphysalis longicornis From Shaanxi, China, Based on Metagenomic Sequencing. Front. Genet. 12:723773. doi: 10.3389/fgene.2021.723773

Ticks are dangerous ectoparasites of humans and animals, as they are important disease vectors and serve as hosts for various microorganisms (including a variety of pathogenic microorganisms). Diverse microbial populations coexist within the tick body. Metagenomic next-generation sequencing (mNGS) has been suggested to be useful for rapidly and accurately obtaining microorganism abundance and diversity data. In this study, we performed mNGS to analyze the microbial diversity of *Haemaphysalis longicornis* from Baoji, Shaanxi, China, with the Illumina HiSeq platform. We identified 189 microbial genera (and 284 species) from ticks in the region; the identified taxa included *Anaplasma* spp., *Rickettsia* spp., *Ehrlichia* spp., and other important tick-borne pathogens at the genus level as well as symbiotic microorganisms such as *Wolbachia* spp., and *Candidatus* Entotheonella. The results of this study provide insights into possible tick-borne diseases and reveal new tick-borne pathogens in this region. Additionally, valuable information for the biological control of ticks is provided. In conclusion, this study provides reference data for guiding the development of prevention and control strategies targeting ticks and tick-borne diseases in the region, which can improve the effectiveness of tick and tick-borne disease control.

Keywords: metagenomic, *Haemaphysalis longicornis*, microorganism diversity, tick-borne pathogens, symbiotic microorganisms

INTRODUCTION

Ticks are important disease vectors that serve as hosts for various microorganisms; they can transmit etiological agents, including bacteria, viruses, and parasitic protozoa (Sanchez-Vicente et al., 2019; Velay et al., 2019), which cause a variety of animal diseases, including zoonoses transferred from animal to animal or human (Sharifah et al., 2020). In several countries, tick-borne diseases have been reported to cause immeasurable economic losses and negatively impact

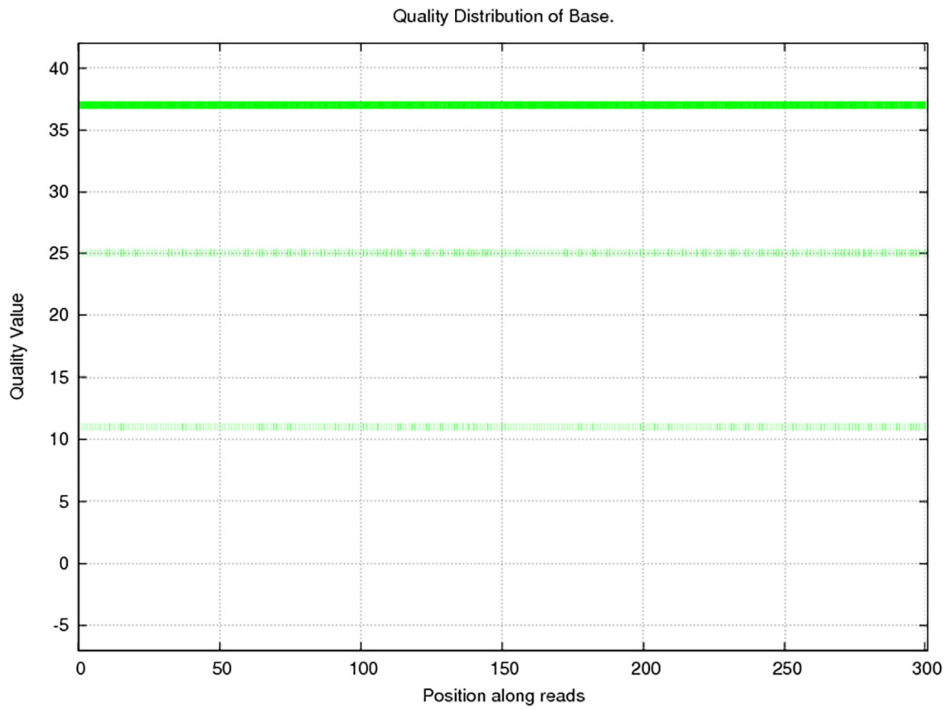


FIGURE 1 | Sequencing quality distribution of bases. The quality of the sequencing data was mainly distributed above Q20, which guarantees the normal conduct of subsequent high-level analyses. The abscissa presents the position on the reads, and the ordinate presents the base mass distribution over the position on the reads.

livestock development (Yang et al., 2018; Sahara et al., 2019; Zeb et al., 2020). Ticks are blood-sucking parasites that can cause host mortality *via* mechanisms (Edlow and McGillicuddy, 2008) such as anemia and tick paralysis. Therefore, analysis of tick microbial diversity is very important for identifying unknown pathogens or detecting known pathogens early and is beneficial for determining new biological control methods based on symbiotic microorganisms.

Haemaphysalis longicornis is a dominant tick species in China (Jia et al., 2020) that is prevalent in multifarious climatic environments in the northern and southern regions of China. It is an important pathogen vector and has one of the largest pathogen loads in China (Zhao et al., 2021); it can harbor pathogens such as *Anaplasma* spp., *Rickettsia* spp., and severe fever with thrombocytopenia syndrome virus (Luo et al., 2015; Jiang et al., 2018; Qin et al., 2018). Moreover, multiple human infections of tick-borne pathogens have been reported to be due to bites by *H. longicornis* (Kondo et al., 2017; Li et al., 2018, 2020), so this tick has become a focus of public health. Baoji City is an animal husbandry region in China with abundant wildlife resources (He, 2012). *H. longicornis*, as a pathogen vector, may cause epidemics due to transmission between livestock and wildlife in the region; therefore, understanding tick-borne pathogens transmitted by *H. longicornis* is significant for the protection of wildlife and livestock in the region.

Metagenomic next-generation sequencing (mNGS) has been suggested to be useful for identifying total microbial species from a single sample. An advantage of the technique is that it can obtain abundant microorganism data rapidly and accurately, allowing the analysis of microorganism diversity. This technique has been widely applied in veterinary science and animal husbandry production (Minamoto et al., 2015; Patel et al., 2018; Mora-Diaz et al., 2020) because of this advantage. mNGS has also been widely used in the analysis of microorganism diversity in vectors (Lambert et al., 2019; Ravi et al., 2019), and mNGS is helpful for understanding insect-borne pathogens and guiding the development of control vectors by identifying new biocontrol agents among symbiotic microorganisms.

In this study, we used the mNGS technique to obtain data for DNA-seq and bioinformatic analyses and obtained microbial species information for *H. longicornis* collected from Baoji, Shaanxi Province. mNGS detected some important tick-borne zoonotic pathogens in the samples. Therefore, this study will be a useful resource for efforts aimed at tick-borne disease control and the biological control of ticks in this region.

MATERIALS AND METHODS

Tick Collection and Preparation

Ticks were collected as adults from host animals (cattle) in Baoji, Shaanxi, China, and were partially fed. The ticks were collected on July 3, 2020. Tick collection was performed after obtaining permission from the farmer. The tick species were identified by morphological features according to the descriptions in the *Economic Insect Fauna of China* (Deng and Jiang, 1991). The ticks

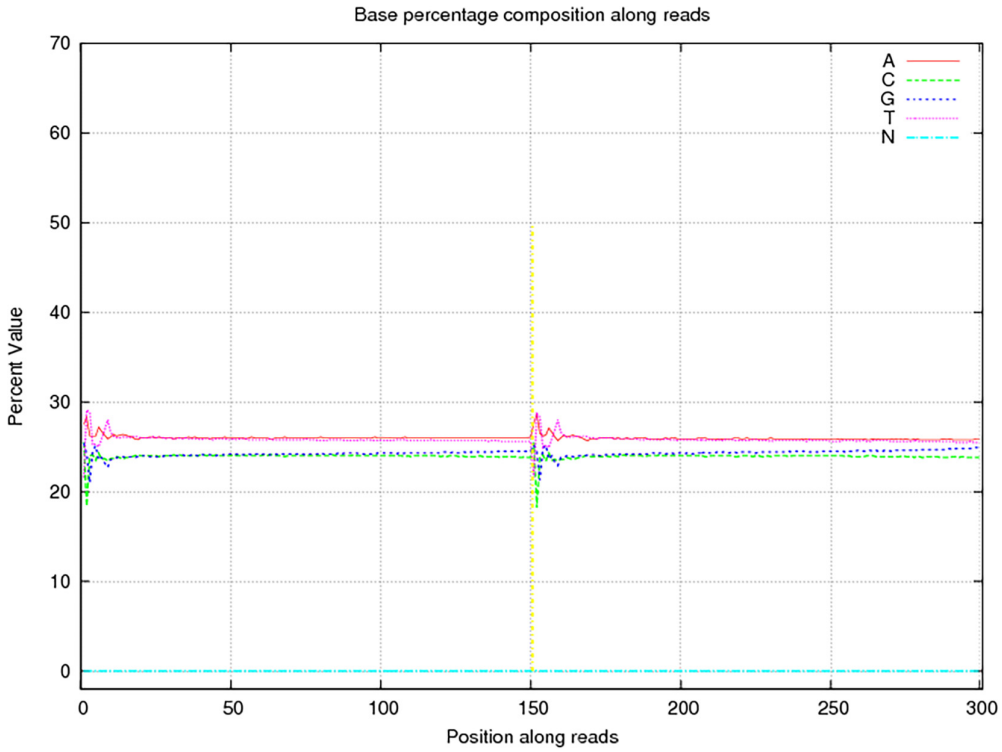


FIGURE 2 | Sequencing base percentage composition among reads. The abscissa presents the position on the reads, the ordinate presents the content of a certain base at that position, and the five colors indicate the proportions of content of the four bases of ATGC and undetected bases (Ns).

were stored at −20°C before DNA analysis. A total of 131 ticks were analyzed in this study.

DNA Extraction

The ticks were placed into new 50-ml sterile centrifuge tubes. Then, after being washed once with 75% ethanol, the ticks were rinsed with normal saline until the liquid was clear. QIAamp® DNA Mini Kit (Germany) was used to extract the total DNA from each tick according to the protocol of the manufacturer. Each DNA sample was stored at −20°C.

Library Construction and DNA Sequencing

To ensure quality, reduce sequencing costs, and be as comprehensive as possible, we selected the best 50 samples randomly from all the DNA samples for sequencing and combined them in a 100-μl pooled sample that comprised 2 μl of DNA solution from each of the 50 samples. The pooled sample was transported in Drikold (at below 0°C) to Novogene in China. The pooled DNA sample was randomly fragmented into 350-bp fragments with a Covaris ultrasonic disruptor, and then the entire library was end-repaired, A-tailed, ligated with a full-length adaptor, and subjected to purification and PCR amplification at Novogene.

Clustering of the index-coded samples was performed on a cBot Cluster Generation System according to the instructions of the manufacturer. After cluster generation, DNA sequencing of 150-bp paired-end reads was conducted with the Illumina HiSeq platform at Novogene.

Data Analysis

Low-quality data were excluded from the raw data to acquire clean data for subsequent analysis using Readfq (version 8)[1]; the obtained data were compared against tick DNA data using Bowtie2.2.4 software to filter out host reads (Karlsson et al., 2012, 2013), and the parameters were as follows: --end-to-end, --sensitive, -I 200, and -X 400. Then, metagenome assembly, gene prediction, abundance analysis, and taxonomy prediction were completed according to the information analysis of the Metagenomic Project of Novogene Content. The detailed methods were as follows: For metagenome assembly, the clean data were assembled and analyzed (Luo et al., 2012) with SOAPdenovo software (V2.04),[2] and the parameters (Scher et al., 2013; Qin et al., 2014; Brum et al., 2015; Feng et al., 2015) were as follows: -d 1, -M 3, -R, -u, -F, and -K 55. Then, the assembled scaffolds were interrupted at N positions to produce scaffolds without Ns (Mende et al., 2012; Nielsen et al., 2014; Qin et al., 2014), called scaftigs (i.e., continuous sequences within scaffolds), using SOAPdenovo (V2.04). Fragments shorter than 500 bp were filtered from all scaftigs for statistical analysis (Li et al., 2014; Qin et al., 2014; Zeller et al., 2014; Sunagawa et al., 2015).

[1]https://github.com/cjfields/readfq

[2]http://soapdenovo2.sourceforge.net/

For gene prediction and abundance analysis, scaftigs (≥ 500 bp) assembled from the sample were used for the open reading frame (ORF) (Zhu et al., 2010; Karlsson et al., 2012, 2013; Mende et al., 2012; Nielsen et al., 2014; Oh et al., 2014) prediction by MetaGeneMark software (V2.10),[3] and sequences shorter than 100 nt (Qin et al., 2010, 2014; Li et al., 2014; Nielsen et al., 2014; Zeller et al., 2014) were filtered from the prediction results with the default parameters. For ORF prediction, CD-HIT (Li and Godzik, 2006; Fu et al., 2012) software (V4.5.8)[4] was adopted to eliminate redundancy and obtain the unique initial gene catalog with the parameter options (Zeller et al., 2014; Sunagawa et al., 2015) -c 0.95, -G 0, -aS 0.9, -g 1, and -d 0. The clean data of the sample were mapped to the initial gene catalog using Bowtie 2.2.4 with the parameter settings (Li et al., 2014; Qin et al., 2014) -end-to-end, -sensitive, -I 200, and -X 400, and the number of reads in the sample mapping to individual genes was determined. The genes with less than or equal to two reads (Qin et al., 2012; Li et al., 2014) in the sample were filtered out to obtain the gene catalog (unigenes) used for subsequent analysis.

For taxonomy prediction, DIAMOND (Buchfink et al., 2015) software (V0.9.9)[5] was used to blast the unigenes to the sequences of bacteria, eukaryote, archaea, and viruses, which were all extracted from the nucleotide (NR) database (version 2018-01-02)[6] of NCBI with the parameter settings blastp and -e 1e-5. To obtain the final alignment results of each sequence, as each sequence may have multiple alignment results, the results with *e*-values less than or equal to the smallest *e*-value × 10 (Oh et al., 2014) were selected. Taxon classification with MEGAN (Huson et al., 2011) software using the last common ancestor (LCA) algorithm (Lowest_common_ancestor)[7] was performed to confirm the species annotation information of the sequences. A table of the number of genes and the abundance information of the sample for each taxonomic level (kingdom, phylum, class, order, family, genus, and species) was produced based on the LCA annotation results and the gene abundance table. Krona analysis (Ondov et al., 2011) was used to visualize the results of the species annotation.

[3]http://topaz.gatech.edu/GeneMark

[4]http://www.bioinformatics.org/cd-hit

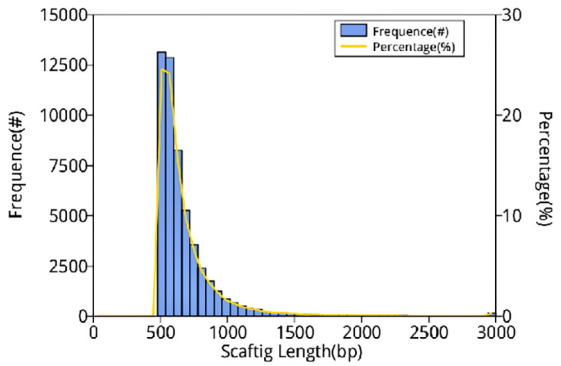


FIGURE 3 | Distribution of scaftig lengths across samples. The first vertical axis (frequency, #) indicates the number of scaftigs, the second vertical axis (percentage, %) indicates the percentage of the number of scaftigs, and the horizontal axis indicates the scaftig length.

TABLE 1 | Gene catalog and basic information.

Catalog	Amount
ORFs NO.	30,255
Integrity:all	9,458 (31.26%)
Integrity:end	9,558 (31.59%)
Integrity:none	2,601 (8.6%)
Integrity:start	8,638 (28.55%)
Total Len. (Mbp)	11.42
Average Len. (bp)	377.51
GC percent	50.74

ORF No. indicates the number of genes in the gene catalog; integrity:start indicates the number and percentage of genes containing only the start codon; integrity:end indicates the number and percentage of genes containing only stop codons; integrity:none indicates the number and percentage of genes with no start or stop codons; integrity:all indicates the percentage of the number of complete genes (both start and stop codons); Total len. (Mbp) represents the total length of a gene in a gene catalog in millions; Average len. indicates the average length of the gene in the gene catalog; GC percent represents the overall GC content value of a gene in the predicted gene catalog.

RESULTS

Tick Identification

In total, 131 tick samples were collected from Baoji, Shaanxi Province, in July 2020. The samples were identified as *H. longicornis* (Acari: Ixodidae).

[5]https://github.com/bbuchfink/diamond/

[6]https://www.ncbi.nlm.nih.gov/

[7]https://en.wikipedia.org/wiki/

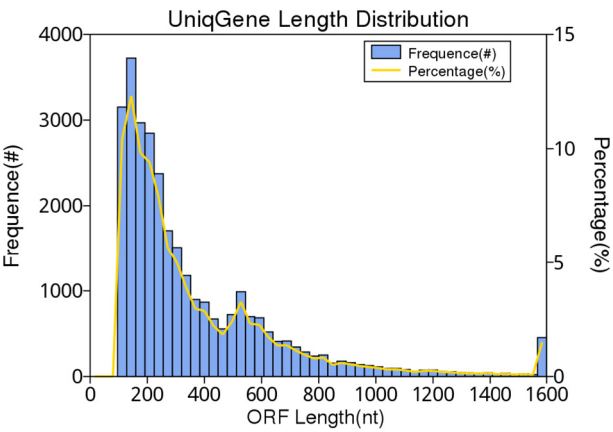


FIGURE 4 | Gene catalog length distribution. The first vertical axis frequency (#) indicates the number of genes in the catalog, the second vertical axis percentage (%) represents the percentage of the number of genes in the catalog, and the horizontal axis indicates the length of the gene in the gene catalog.

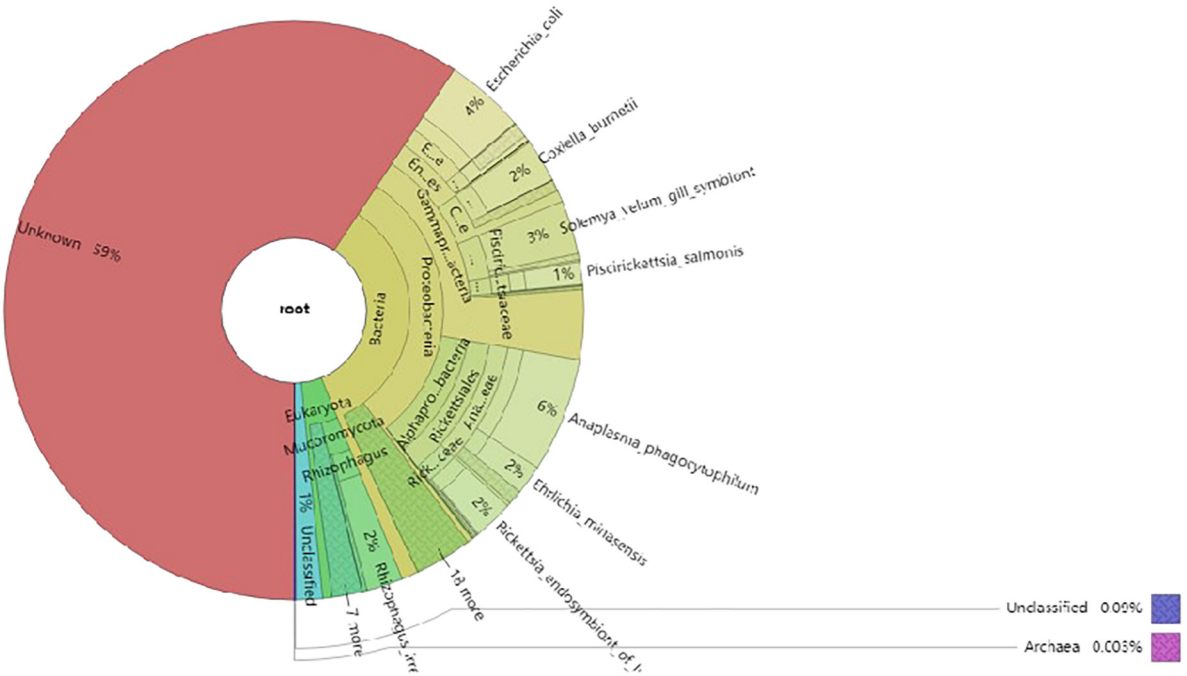


FIGURE 5 | The Krona diagram of taxonomy prediction. In the figure, circles represent different classification levels (kingdom, phylum, class, order, family, and genus) from inside to outside. The size of the sector represents the relative abundance of different species.

Overview of DNA Sequencing

In this study, 10,117.72 Mbp of clean data were generated by sequencing with the Illumina HiSeq platform; the effective data rate was 99.64%, and the results of the quality control are shown in **Figures 1**, **2**. After a single sample assembly, 58,543,947-bp scaffolds were obtained. Then, a total of 37,959,124-bp scaftigs were obtained by interrupting scaffolds at the N-site; the distribution of scaftig lengths across samples is shown in **Figure 3**. After obtaining the assembly results, MetaGeneMark software was used for gene prediction, basic gene catalog information statistics were obtained (**Table 1** and **Figure 4**), and a total of 30,341 ORFs were identified. After excluding redundant sequences, 30,255 ORFs were identified, and among them, the number of complete genes was 9,458, accounting for 31.26% of the ORFs.

Taxonomy Prediction and Diversity Analysis

The non-redundant gene catalog (unigenes) was BLASTP-aligned with the NR database of NCBI (see text footnote 6; version: 2018-01-02) using DIAMOND software. Based on the LCA algorithm, species annotation was performed. There were 30,255 genes after the initial redundancy removal, and the number of ORFs that were annotated to the NR database was 7,754 (25.63%). The proportion of ORFs annotated to the genus level was 67.08% (189 genera) and that annotated to the species level was 58.72% (284 species). According to the results, the most common species of microorganisms was bacteria; a total of 145 species were identified. The next was eukaryote and had 116 species. The last was virus and archaea; there were 21 species and two species, respectively. The result of the species annotation is shown in the Krona figure (taxonomy.krona.html, **Figure 5**).

Anaplasma, *Escherichia*, *Coxiella*, *Rickettsia*, *Rhizophagus*, *Ehrlichia*, *Piscirickettsia*, *Serratia*, and *Wolbachia* were the predominant genera based on relative abundance in all microorganisms. *Anaplasma phagocytophilum*, *Escherichia coli*, *Solemya velum* gill symbiont, *Coxiella burnetiid*, *Rickettsia endosymbiont* of *Ixodes scapularis*, *Rhizophagus irregularis*, *Ehrlichia minasensis*, *Piscirickettsia salmonis*, and bacterium 2013Ark19i were the predominant species based on relative abundance in all microorganisms. The predominant genera and species are shown in **Figures 6**, **7**, respectively.

The predominant eukaryotes were *R. irregularis*, *Metarhizium anisopliae*, *Enterospora canceri*, *Rhizopus delemar*, *Rhizoctonia solani*, *Absidia glauca*, *Puccinia striiformis*, *Smittium culicis*, and *Nosema apis* on relative abundance in eukaryote. In addition, *Lymphocystis disease virus Sa*, *Cotesia sesamiae bracovirus*, and *Autographa californica multiple nucleopolyhedrovirus* were the predominant viruses on relative abundance in virus, and the only two archaea were *Thermococcus* sp. 2319 × 1 and *Lokiarchaeum* sp. GC14_75. Those results are shown in the Krona figure (taxonomy.krona.html, **Figure 5**).

DISCUSSION

Ticks are important pathogen vectors (Sharifah et al., 2020). In this study, *H. longicornis* was collected from the Qinling area in Baoji, Shaanxi, China. Then, tick DNA was extracted

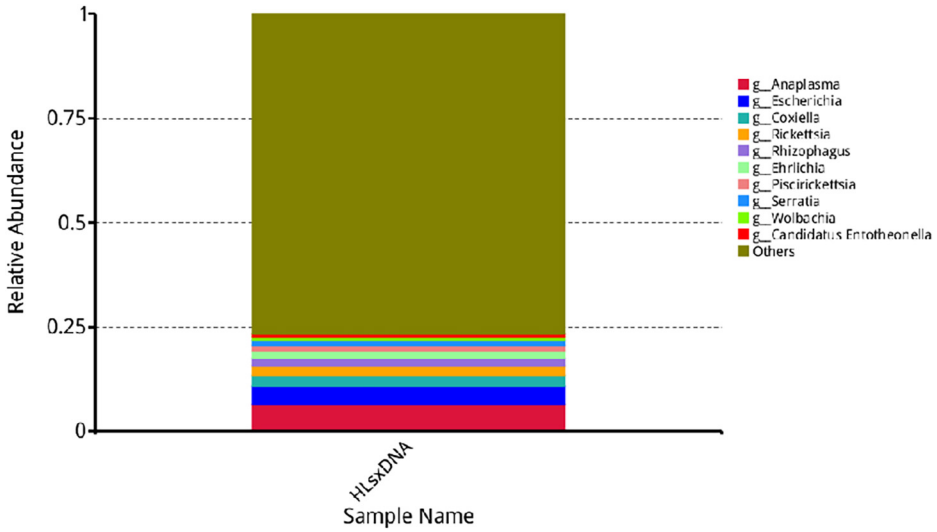


FIGURE 6 | Relative abundance of the top 10 genera. The horizontal axis presents the sample names, the vertical axis presents the relative proportion of species annotated to a certain type, and the genera corresponding to individual color blocks are indicated in the legend on the right.

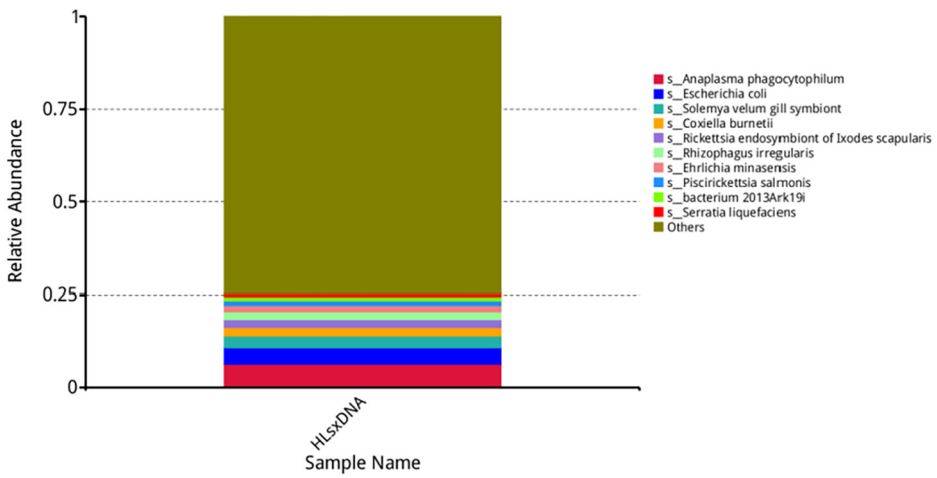


FIGURE 7 | Top 10 most abundant species. The horizontal axis presents the sample names, the vertical axis presents the relative proportion of species annotated to a certain type, and the species corresponding to individual color blocks are indicated in the legend on the right.

and analyzed with mNGS technology. The primary objective was to understand the microbial diversity in *H. longicornis* in the region. Prior to sequencing, 50 samples of the 131 collected samples were pooled. Finally, a total of 284 microorganisms were annotated and classified as bacteria, eukaryote, viruses, or archaea. Common pathogens transmitted by *H. longicornis*, such as *Rickettsia* spp., *Anaplasma* spp., and *Ehrlichia* spp., were detected. Most known *Rickettsia* species belong to the spotted fever group (Abdad et al., 2018) and cause spotted fever. *Anaplasma* and *Ehrlichia* belong to Anaplasmataceae and are prevalent and potentially fatal arthropod-borne pathogens (Dumler, 2005). Previous studies on *H. longicornis*-borne pathogens have revealed that these pathogens are commonly detected and widely distributed in *H. longicornis* in all regions of China. The *Rickettsia* positivity rate was 7.36% in Liaoning, China (Xu et al., 2019), and the *Anaplasma* and *Ehrlichia* positivity rates were 2.2 and 0.8%, respectively, in northeastern China (Wei et al., 2016). The rate of *Ehrlichia* positivity was 1.82%, and the rate of *Anaplasma* positivity was 11.82% in Zhejiang, China (Hou et al., 2019). Moreover, *Rickettsia* (0.67%, 2/298), *Anaplasma* (3.02%, 9/298), and *Ehrlichia* (1.01%, 3/298) were detected in central China (Chen et al., 2014). The results of this study show that these pathogens may be common in the region and that the incidence of diseases caused by these pathogens could be very high; therefore, these diseases need to be monitored to ensure successful animal husbandry and healthy production in the region.

In addition to the abovementioned pathogens, *Coxiella* spp. (including *Coxiella burnetiid* and *Coxiella*-like endosymbiont) were detected, with a high relative abundance. *Francisella tularensis* was detected, but it had a low relative abundance in this study. Both *C. burnetii* and *F. tularensis* are important zoonotic pathogens (Maurin and Raoult, 1999; Proksova et al., 2019). The different abundances may be related to the carrying capacity of *H. longicornis*. Previous studies have found that *F. tularensis* did not persist in *H. longicornis* after artificial infection (Tully and Huntley, 2020). However, *Coxiella* has a symbiotic relationship with *H. longicornis*. Phylogenetic analysis has revealed that *C. burnetii* is closely related to *Coxiella*-like endosymbiont (Elsa et al., 2015; Trinachartvanit et al., 2018). As expected, these pathogens were detected in *H. longicornis* in this study, which suggests that epidemics of *Coxiella*-related diseases could occur in this area.

Wolbachia spp. was also detected in *H. longicornis*, with a high relative abundance in this study. Studies on *Wolbachia* in arthropods (Carvajal et al., 2019; Landmann, 2019) have confirmed that it is prevalent in mosquitoes and important for insect control (Beckmann et al., 2017; Madhav et al., 2020). However, only a few studies have reported the detection of *Wolbachia* genes in individual tick species, and none has reported the presence of *Wolbachia* in *H. longicornis* (Madhav et al., 2020). The results of this study suggest that *Wolbachia* is present in *H. longicornis*, but further investigation is required to determine whether it has the same cytoplasmic incompatibility function in *H. longicornis* as it does in other taxa.

In summary, this study using mNGS technology revealed the microbial species composition in *H. longicornis* in the Baoji Qinling region. Pathogenic infection *via* ticks is complex and diverse, and tick bites are a potential threat to wildlife, domestic animals, and people involved in forestry, livestock, and tourism in the region. Therefore, workers in relevant industries in the region need to adhere to surveillance measures and implement control strategies. It is very important to monitor for possible emerging tick-borne diseases by analyzing microbial species; monitoring may have profound and great implications for public health and the safety of animal husbandry in the region.

CONCLUSION

In this study, microorganism diversity data were obtained from *H. longicornis* in Baoji, Shaanxi, China. The data indicated that pathogens harbored by *H. longicornis* may pose a great threat to animals and humans in this region. Therefore, it is important to control ticks in this region. In addition, the results provide evidence of the presence of *Wolbachia* spp., in *H. longicornis*. These results are of great significance for the prevention and control of ticks and tick-borne diseases in the region, as these data are necessary for guiding human and animal safety measures in the region.

DATA AVAILABILITY STATEMENT

The raw sequence data are available in the National Genomics Data Center (NGDC), the China National Center for Bioinformation (CNCB), using accession number CRA004680 (https://bigd.big.ac.cn/gsa/browse/CRA004680).

AUTHOR CONTRIBUTIONS

RC contributed to the methodology, formal analysis, and writing of the original draft. QR contributed to the methodology and validation. JLu, ZT, GG, JLu, and HY contributed to writing—review and editing. WL, BZ, and JLi contributed to formal analysis and writing of the original draft. PD, YT, and XQ contributed to formal analysis and resources. GZ and QW contributed to formal analysis. GL conceived and designed the experiment and contributed to writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This study was financially supported by the National Key Research and Development Program of China (Nos. 2019YFC1200502, 2019YFC1200504, 2019YFC1200500, and 2017YFD0501200), the National Parasitic Resources Center (No. NPRC-2019-194-30), the National Natural Science Foundation of China (31572511), the Central Public-interest Scientific Institution Basal Research Fund (Nos. Y2019YJ07-04 and Y2018PT76), and NBCIS (CARS-37).

REFERENCES

Abdad, M. Y., Abou, A. R., Fournier, P. E., Stenos, J., and Vasoo, S. (2018). A concise review of the epidemiology and diagnostics of rickettsioses: *Rickettsia* and *Orientia* spp. *J. Clin. Microbiol.* 56:e01728-17. doi: 10.1128/JCM.01728-17

Beckmann, J. F., Ronau, J. A., and Hochstrasser, M. (2017). A *Wolbachia* deubiquitylating enzyme induces cytoplasmic incompatibility. *Nat. Microbiol.* 2:17007. doi: 10.1038/nmicrobiol.2017.7

Brum, J. R., Ignacio-Espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., et al. (2015). Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* 348:1261498. doi: 10.1126/science.1261498

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Carvajal, T. M., Hashimoto, K., Harnandika, R. K., Amalin, D. M., and Watanabe, K. (2019). Detection of *Wolbachia* in field-collected *Aedes aegypti* mosquitoes in metropolitan Manila, Philippines. *Parasit. Vectors* 12:361. doi: 10.1186/s13071-019-3629-y

Chen, Z., Liu, Q., Liu, J. Q., Xu, B. L., Lv, S., Xia, S., et al. (2014). Tick-borne pathogens and associated co-infections in ticks collected from domestic animals in central China. *Parasit. Vectors* 7:237. doi: 10.1186/1756-3305-7-237

Deng, G. F., and Jiang, Z. J. (1991). *Economic Insect Fauna of China. Fasc 39 Acari:Ixodidae. Thirty-nine volumes. [In Chinese]*. Beijing: Science Press.

Dumler, J. S. (2005). *Anaplasma* and *Ehrlichia* infection. *Ann. N. Y. Acad. Sci.* 1063, 361–373. doi: 10.1196/annals.1355.069

Edlow, J. A., and McGillicuddy, D. C. (2008). Tick paralysis. *Infect. Dis. Clin. North Am.* 22, 397–413. doi: 10.1016/j.idc.2008.03.005

Elsa, J., Duron, O., Severine, B., Gonzalez-Acuna, D., and Sidi-Boumedine, K. (2015). Molecular methods routinely used to detect *Coxiella burnetii* in ticks cross-react with *Coxiella*-like bacteria. *Infect. Ecol. Epidemiol.* 5:29230. doi: 10.3402/iee.v5.29230

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., et al. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun* 6:6528. doi: 10.1038/ncomms7528

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

He, C. (2012). *Baoji City Development Model of Animal Husbandry and Construction[D]*, Xi'an: Northwest A&F University.

Hou, J., Ling, F., Liu, Y., Zhang, R., Song, X., Huang, R., et al. (2019). A molecular survey of *Anaplasma*. *Ehrlichia*, *Bartonella* and *Theileria* in ticks collected from southeastern China. *Exp. Appl. Acarol.* 79, 125–135. doi: 10.1007/s10493-019-00411-2

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 21, 1552–1560. doi: 10.1101/gr.120618.111

Jia, N., Wang, J., Shi, W., Du, L., Sun, Y., Zhan, W., et al. (2020). Large-scale comparative analyses of tick genomes elucidate their genetic diversity and vector capacities. *Cell* 182, 1328–1340. doi: 10.1016/j.cell.2020.07.023

Jiang, J., An, H., Lee, J. S., O'Guinn, M. L., Kim, H. C., Chong, S. T., et al. (2018). Molecular characterization of *Haemaphysalis longicornis*-borne rickettsiae, Republic of Korea and China. *Ticks Tick Borne Dis.* 9, 1606–1613. doi: 10.1016/j.ttbdis.2018.07.013

Karlsson, F. H., Fak, F., Nookaew, I., Tremaroli, V., Fagerberg, B., Petranovic, D., et al. (2012). Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nat. Commun.* 3:1245. doi: 10.1038/ncomms2266

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergstrom, G., Behre, C. J., Fagerberg, B., et al. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498, 99–103. doi: 10.1038/nature12198

Kondo, M., Nakagawa, T., Yamanaka, K., and Mizutani, H. (2017). Case with acute urticaria by red meat after *Haemaphysalis longicornis* bite. *J. Dermatol.* 44, e168–e169. doi: 10.1111/1346-8138.13865

Lambert, J. S., Cook, M. J., Healy, J. E., Murtagh, R., Avramovic, G., and Lee, S. H. (2019). Metagenomic 16s rRNA gene sequencing survey of *Borrelia* species in Irish samples of *Ixodes ricinus* ticks. *PLoS One* 14:e209881. doi: 10.1371/journal.pone.0209881

Landmann, F. (2019). The *Wolbachia* endosymbionts. *Microbiol. Spectr.* 7:BAI-0018-2019. doi: 10.1128/microbiolspec.BAI-0018-2019

Li, H., Li, X. M., Du, J., Zhang, X. A., Cui, N., Yang, Z. D., et al. (2020). *Candidatus Rickettsia* xinyangensis as cause of spotted fever group rickettsiosis, Xinyang, China, 2015. *Emerg. Infect. Dis.* 26, 985–988. doi: 10.3201/eid2605.170294

Li, J., Hu, W., Wu, T., Li, H. B., Hu, W., Song, Y., et al. (2018). Japanese spotted fever in eastern China, 2013. *Emerg. Infect. Dis.* 24, 2107–2109. doi: 10.3201/eid2411.170264

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* 32, 834–841. doi: 10.1038/nbt.2942

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Luo, L. M., Zhao, L., Wen, H. L., Zhang, Z. T., Liu, J. W., Fang, L. Z., et al. (2015). *Haemaphysalis longicornis* ticks as reservoir and vector of severe fever with thrombocytopenia syndrome virus in China. *Emerg. Infect. Dis.* 21, 1770–1776. doi: 10.3201/eid2110.150126

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18. doi: 10.1186/2047-217X-1-18

Madhav, M., Baker, D., Morgan, J., Asgari, S., and James, P. (2020). *Wolbachia*: a tool for livestock ectoparasite control. *Vet. Parasitol.* 288:109297. doi: 10.1016/j.vetpar.2020.109297

Maurin, M., and Raoult, D. (1999). Q fever. *Clin. Microbiol. Rev.* 12, 518–553. doi: 10.1128/CMR.12.4.518

Mende, D. R., Waller, A. S., Sunagawa, S., Jarvelin, A. I., Chan, M. M., Arumugam, M., et al. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS One* 7:e31386. doi: 10.1371/journal.pone.0031386

Minamoto, Y., Otoni, C. C., Steelman, S. M., Buyukleblebici, O., Steiner, J. M., Jergens, A. E., et al. (2015). Alteration of the fecal microbiota and serum metabolite profiles in dogs with idiopathic inflammatory bowel disease. *Gut Microbes* 6, 33–47. doi: 10.1080/19490976.2014.997612

Mora-Diaz, J., Pineyro, P., Shen, H., Schwartz, K., Vannucci, F., Li, G., et al. (2020). Isolation of PCV3 from perinatal and reproductive cases of pcv3-associated disease and in vivo characterization of PCV3 replication in CD/CD growing pigs. *Viruses* 12:219. doi: 10.3390/v12020219

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* 32, 822–828. doi: 10.1038/nbt.2939

Oh, J., Byrd, A. L., Deming, C., Conlan, S., Kong, H. H., Segre, J. A., et al. (2014). Biogeography and individuality shape function in the human skin metagenome. *Nature* 514, 59–64. doi: 10.1038/nature13786

Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a web browser. *BMC Bioinformatics* 12:385. doi: 10.1186/1471-2105-12-385

Patel, J. G., Patel, B. J., Patel, S. S., Raval, S. H., Parmar, R. S., Joshi, D. V., et al. (2018). Metagenomic of clinically diseased and healthy broiler affected with respiratory disease complex. *Data Brief* 19, 82–85. doi: 10.1016/j.dib.2018.05.010

Proksova, M., Bavlovic, J., Klimentova, J., Pejchal, J., and Stulik, J. (2019). Tularemia-zoonosis carrying a potential risk of bioterrorism. *Epidemiol. Mikrobiol. Imunol.* 68, 82–89.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60. doi: 10.1038/nature11450

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513, 59–64. doi: 10.1038/nature13568

Qin, X. R., Han, F. J., Luo, L. M., Zhao, F. M., Han, H. J., Zhang, Z. T., et al. (2018). *Anaplasma* species detected in *Haemaphysalis longicornis* tick from China. *Ticks Tick Borne Dis.* 9, 840–843. doi: 10.1016/j.ttbdis.2018.03.014

Ravi, A., Ereqat, S., Al-Jawabreh, A., Abdeen, Z., Abu Shamma, O., Hall, H., et al. (2019). Metagenomic profiling of ticks: identification of novel *Rickettsial* genomes and detection of tick-borne canine parvovirus. *PLoS Negl. Trop. Dis.* 13:e0006805. doi: 10.1371/journal.pntd.0006805

Sahara, A., Nugraheni, Y. R., Patra, G., Prastowo, J., and Priyowidodo, D. (2019). Ticks (*Acari:Ixodidae*) infestation on cattle in various regions in Indonesia. *Vet. World* 12, 1755–1759. doi: 10.14202/vetworld.2019.1755-1759

Sanchez-Vicente, S., Tagliafierro, T., Coleman, J. L., Benach, J. L., and Tokarz, R. (2019). Polymicrobial nature of tick-borne diseases. *mBio* 10:e02055-19. doi: 10.1128/mBio.02055-19

Scher, J. U., Sczesnak, A., Longman, R. S., Segata, N., Ubeda, C., Bielski, C., et al. (2013). Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *eLife* 2:e1202. doi: 10.7554/eLife.01202

Sharifah, N., Heo, C. C., Ehlers, J., Houssaini, J., and Tappe, D. (2020). Ticks and tick-borne pathogens in animals and humans in the island nations of southeast Asia: a review. *Acta Trop.* 209:105527. doi: 10.1016/j.actatropica.2020.105527

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., et al. (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science* 348:1261359. doi: 10.1126/science.1261359

Trinachartvanit, W., Maneewong, S., Kaenkan, W., Usananan, P., Baimai, V., Ahantarig, A., et al. (2018). *Coxiella*-like bacteria in fowl ticks from Thailand. *Parasit. Vectors* 11:670. doi: 10.1186/s13071-018-3259-9

Tully, B. G., and Huntley, J. F. (2020). A *Francisella tularensis* chitinase contributes to bacterial persistence and replication in two major U.S. tick vectors. *Pathogens* 9:1037. doi: 10.3390/pathogens9121037

Velay, A., Paz, M., Cesbron, M., Gantner, P., Solis, M., Soulier, E., et al. (2019). Tick-borne encephalitis virus: molecular determinants of neuropathogenesis of an emerging pathogen. *Crit. Rev. Microbiol.* 45, 472–493. doi: 10.1080/1040841X.2019.1629872

Wei, F., Song, M., Liu, H., Wang, B., Wang, S., Wang, Z., et al. (2016). Molecular detection and characterization of zoonotic and veterinary pathogens in ticks from northeastern China. *Front. Microbiol.* 7:1913. doi: 10.3389/fmicb.2016.01913

Xu, H., Zhang, Q., Guan, H., Zhong, Y., Jiang, F., Chen, Z., et al. (2019). High incidence of a novel *Rickettsia* genotype in parasitic *Haemaphysalis longicornis* from China-north Korea border. *Sci. Rep.* 9:5373. doi: 10.1038/s41598-019-41879-7

Yang, J., Han, R., Niu, Q., Liu, Z., Guan, G., Liu, G., et al. (2018). Occurrence of four *Anaplasma* species with veterinary and public health significance in sheep, northwestern China. *Ticks Tick Borne Dis.* 9, 82–85. doi: 10.1016/j.ttbdis.2017.10.005

Zeb, J., Shams, S., Din, I. U., Ayaz, S., Khan, A., Nasreen, N., et al. (2020). Molecular epidemiology and associated risk factors of *Anaplasma marginale* and *Theileria annulata* in cattle from north-western Pakistan. *Vet. Parasitol.* 279:109044. doi: 10.1016/j.vetpar.2020.109044

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* 10:766. doi: 10.15252/msb.20145645

Zhao, G. P., Wang, Y. X., Fan, Z. W., Ji, Y., Liu, M. J., Zhang, W. H., et al. (2021). Mapping ticks and tick-borne pathogens in China. *Nat. Commun.* 12:1075. doi: 10.1038/s41467-021-21375-1

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res.* 38:e132. doi: 10.1093/nar/gkq275

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 October 2021
doi: 10.3389/fgene.2021.740940

Cryptosporidium hominis Phylogenomic Analysis Reveals Separate Lineages With Continental Segregation

*Felipe Cabarcas[1,3], Ana Luz Galvan-Diaz[4], Laura M. Arias-Agudelo[1], Gisela María García-Montoya[1,2,5], Juan M. Daza[6] and Juan F. Alzate[1,2,5]**

[1]Centro Nacional de Secuenciación Genómica CNSG, Sede de Investigación Universitaria-SIU, Medellín, Colombia, [2]Grupo SISTEMIC, Departamento de Ingeniería Electrónica, Facultad de Ingeniería, Universidad de Antioquia, Medellín, Colombia, [3]Environmental Microbiology Group, School of Microbiology, Universidad de Antioquia, Medellín, Colombia, [4]Departamento de Microbiología y Parasitología, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia, [5]Grupo Pediaciencias, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia, [6]Grupo Herpetológico de Antioquia, Institute of Biology, Universidad de Antioquia, Medellín, Colombia

OPEN ACCESS

Edited by:
Jun-Hu Chen,
National Institute of Parasitic Diseases, China

Reviewed by:
Karen Luisa Haag,
Federal University of Rio Grande do Sul, Brazil
Adam Sateriale,
Francis Crick Institute, United Kingdom
Alejandro Castellanos-Gonzalez,
University of Texas Medical Branch at Galveston, United States

***Correspondence:**
Juan F. Alzate
jfernando.alzate@udea.edu.co

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: 13 July 2021
Accepted: 27 September 2021
Published: 14 October 2021

Citation:
Cabarcas F, Galvan-Diaz AL, Arias-Agudelo LM, García-Montoya GM, Daza JM and Alzate JF (2021) Cryptosporidium hominis *Phylogenomic Analysis Reveals Separate Lineages With Continental Segregation. Front. Genet. 12:740940. doi: 10.3389/fgene.2021.740940*

Cryptosporidium is a leading cause of waterborne outbreaks globally, and *Cryptosporidium hominis* and *C. parvum* are the principal cause of human cryptosporidiosis on the planet. Thanks to the advances in Next-Generation Sequencing (NGS) sequencing and bioinformatic software development, more than 100 genomes have been generated in the last decade using a metagenomic-like strategy. This procedure involves the parasite oocyst enrichment from stool samples of infected individuals, NGS sequencing, metagenomic assembly, parasite genome computational filtering, and comparative genomic analysis. Following this approach, genomes of infected individuals of all continents have been generated, although with striking different quality results. In this study, we performed a thorough comparison, in terms of assembly quality and purity, of 100+ *de novo* assembled genomes of *C. hominis*. Remarkably, after quality genome filtering, a comprehensive phylogenomic analysis allowed us to discover that *C. hominis* encompasses two lineages with continental segregation. These lineages were named based on the observed continental distribution bias as *C. hominis* Euro-American (EA) and the *C. hominis* Afro-Asian (AA) lineages.

Keywords: *Cryptosporidium hominis*, comparative genomics, *de novo* genome assembly, evolution, lineage

INTRODUCTION

Cryptosporidium is a ubiquitous apicomplexan parasite with gastrointestinal habitat and a broad range of vertebrate hosts, including humans, other mammals, birds, fish, and reptiles (Zahedi and Ryan, 2020). *Cryptosporidium hominis* and *C. parvum* are the preponderant cause of human cryptosporidiosis around the world, with the former being more frequently found in developing nations (Gilchrist et al., 2018; Tichkule et al., 2021). Cryptosporidiosis is usually a self-limiting infection in immunocompetent individuals. However, vulnerable populations, like immunocompromised individuals (especially those with T cell impairment) and children (particularly those below 5 years old), develop persistent to chronic syndromes, with diarrhea as

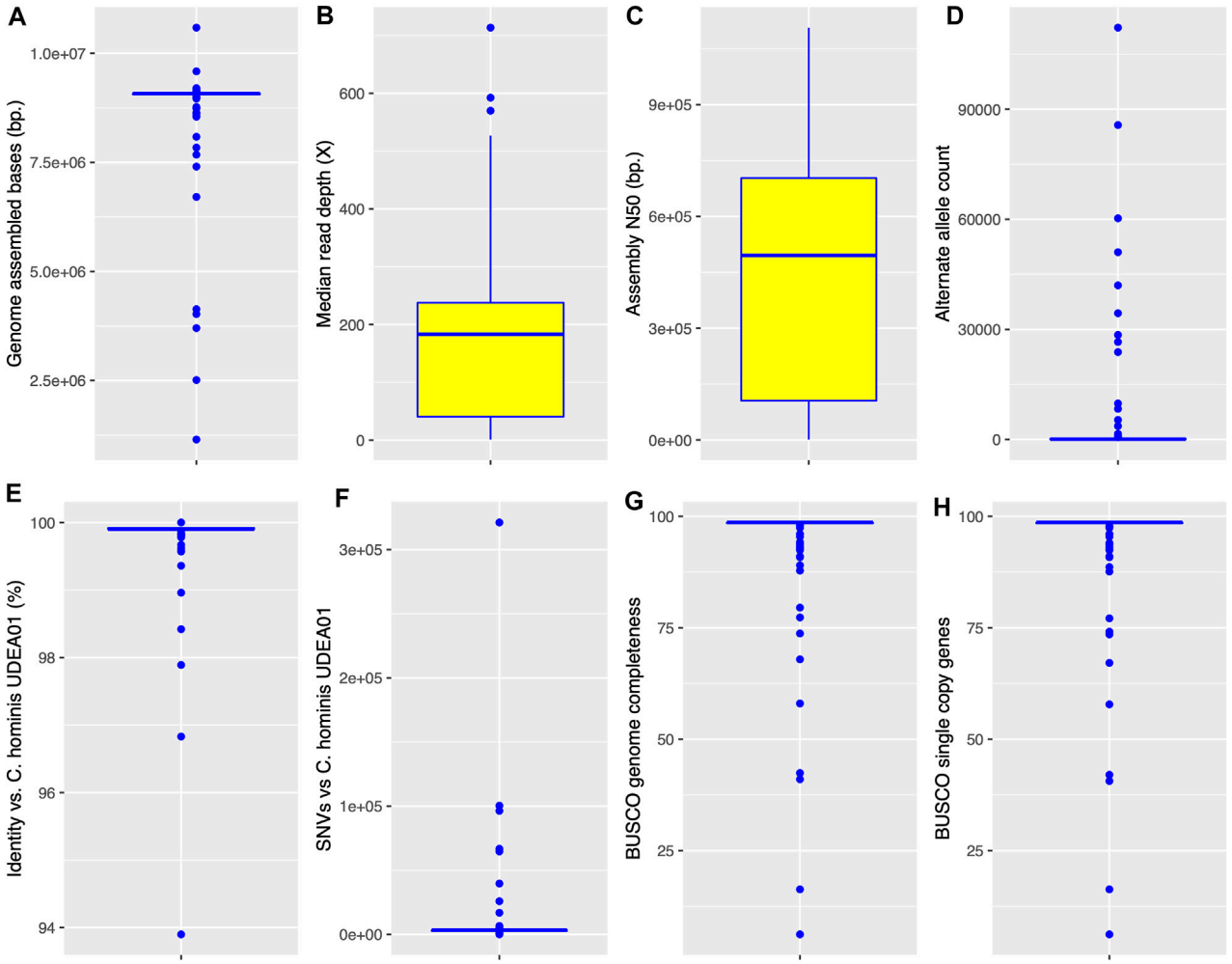


FIGURE 1 | Boxplots representing the descriptive statistics of the *de novo* assembled *C. hominis* genomes. The box represents the interquartile range, IQR 25th -75th percentile. The line represents the 50th percentile. Whiskers denote the 0th and 100th percentile. **(A)** *Cryptosporidium* genome assembled bases. **(B)** Scaffold median read depth. **(C)** *Cryptosporidium* assembly N50 value. **D**. Scaffold positions with an alternate allele. **(E)** Assembly genomic nucleotide identity against *C. hominis* UDEA01 reference. **(F)** Single nucleotide variants (SNVs) detected in each *C. hominis* assembly against *C. hominis* UDEA01 reference. **(G)** BUSCO genome completeness. **(H)** BUSCO single-copy genes.

the principal symptom (Innes et al., 2020). According to the Global Enteric Multicenter Study (GEMS), *Cryptosporidium* was the second cause of moderate-to-severe diarrhea in children from sub-Saharan Africa and South Asia countries; and this infection increased the risk of death in children aged 12–23 months (Kotloff et al., 2013).

Cryptosporidium oocysts are excreted in feces by symptomatic hosts. In this stage, the parasite is highly resistant to common disinfection methods like chlorination (Cacciò and Chalmers, 2016). These characteristics might explain why *Cryptosporidium* is a leading cause of waterborne outbreaks globally (Efstratiou et al., 2017). Furthermore, depending on the species, the parasite can be transmitted via direct person-to-person contact, indirect (food, water, or fomites), or zoonotic routes (Cacciò and Chalmers, 2016).

Currently, 42 species of *Cryptosporidium* are recognized (Zahedi and Ryan, 2020), with *C. parvum* and *C. hominis* being responsible for greater than 90% of human infections (Feng et al., 2018). Geographic differences in species distribution have been reported, with *C. hominis* as the leading species in human cases in developing countries (Xiao and Feng, 2008; Ryan et al., 2014; Xiao and Feng, 2017). While *C. hominis* is associated with a predominant anthroponotic transmission, *C. parvum* presents a zoonotic transmission route with livestock as the primary source of infection (Nader et al., 2019). In developed European nations, *C. parvum* infections are more common in rural areas with low human population density and activities related to livestock and agriculture (Lake et al., 2007; Pollock et al., 2010). By contrast, *C. hominis* has a narrow host range showing a specialization trend toward human hosts. Although it can successfully infect other mammals, it produces mild and asymptomatic infections (Widmer et al., 2020). Subtyping characterization through gp 60 gene analysis reveals more than ten subtype families, with six of them as the

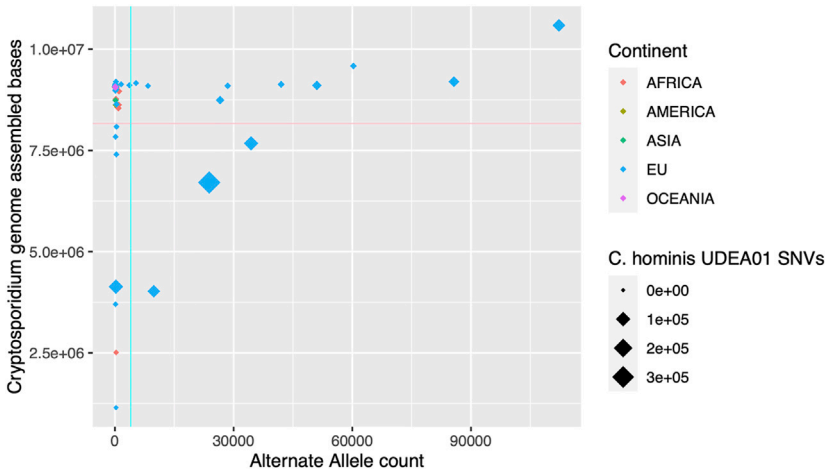


FIGURE 2 | Graphical representation of the *de novo* assembled *C. hominis* genome size and alternate allele detection. In the x-axis, alternate allele count for each genome. In the y-axis, *Cryptosporidium* assembly size in bases pairs. The continent where the parasite was isolated is represented in colors. The size of the rhombus depicts the single nucleotide variants (SNVs) counts against the *C. hominis* UDEA01 genome reference. The cyan and pink lines indicate the thresholds established for low-quality genomes, alternate allele count = 3,992, and assembled genome size = 8,166,447 bases, respectively.

most common in humans (Ia, Ib, Id, Ie, If, and Ig), being the IbA10G2 subtype the predominant and most virulent, widely distributed in both developing and developed countries, and frequently associated with outbreaks worldwide (Feng et al., 2018).

Cryptosporidium hominis is an unculturable parasite. This condition implies that the only possibility to obtain genomic DNA of the parasite is to extract it from stool samples of infected hosts. *Cryptosporidium* oocysts in human feces are usually present at low numbers, requiring specific procedures to capture them. Despite the purification step, DNA of the intestinal microbiota is often abundant, requiring a metagenomic bioinformatics strategy to study the *Cryptosporidium* genome. A metagenome assembly is performed to start the genome reconstruction. Then, *C. hominis* genome scaffolds should be selected using informatics tools. Due to the risk of *Cryptosporidium* coinfections, different species, or isolates, additional quality control steps are needed to avoid contaminated/chimeric genome reconstructions (Isaza et al., 2015).

In this study, we conducted a comparative analysis of the publicly available genomic data of *C. hominis* under normalized conditions with the aim to have a better understanding of the quality of the generated data in the last decade. Since all these genomes come from a methodology more like a metagenomic approach, additional analyses were performed to detect *Cryptosporidium* genome mixtures present in one individual.

We also used phylogenetic tools to reconstruct the evolutionary history of ninety-nine *C. hominis* genomes collected from human individuals in five continents: America, Europe, Asia, Africa, and Oceania. Our phylogenomic analysis showed that *C. hominis* species encompasses two lineages with phylogeographic structure, with one clade composed mainly of European and American isolates, while the other mainly with African and Asian isolates. The two Oceania representatives were grouped into the Euro-American lineage.

MATERIALS AND METHODS

Genome Data From Sequence Read Archive-SRA Public Database

One hundred nineteen *C. hominis* Next-Generation Sequencing (NGS) genome projects with shotgun reads were selected and downloaded (**Supplementary Table S1**) from the Sequence Read Archive-SRA database. Genome projects based only in 454 Technology were excluded. The raw read data were directly download from the SRA database using the fastq-dump tool with the split option activated.

As outgroups, another 15 genomes were used: *C. cuniculus* (UKCU5: PRJNA492839, UKCU2: PRJNA315496), *C. parvum* (UKP2:PRJNA253836, UKP3:PRJNA253840, UKP4: PRJNA253843, UKP5:PRJNA253845, UKP6:PRJNA253846, UKP7:PRJNA253847, UKP8:PRJNA253848, UKP14: PRJNA315506, UKP15:PRJNA315507), and *C. meleagridis* (UKMEL1:PRJNA222838, UKMEL3:PRJNA315502, UKMEL4: PRJNA315503).

Generation of the New Colombian *Cryptosporidium hominis* Genome Reference UDEAa567

C. hominis oocysts were purified from a fecal sample collected from an HIV Colombian female patient by flotation in a saturated sodium chloride solution (Kar et al., 2011). Parasite diagnosis was previously confirmed by Kinyoun stain. Species and subtype identification was done by a nested PCR and sequence analysis of the small-subunit (SSU) rDNA gene and 60 kDa glycoprotein gene (IbA10G2), respectively. Purified oocysts were resuspended

in PBS and quantified using a Neubauer chamber slide and light microscopy. The sample was stored at 4°C until DNA was extracted. DNA extraction was performed using the kit NORGEN Stool DNA Isolation (CAT 27600), following the manufacturer's instructions, with previous freeze-thaw cycles (each for 2 min) in liquid nitrogen and a water bath at 37°C. A solution with approximately 1.85×10^8 oocyst was used for DNA purification. DNA obtained was quantified by light absorption at 260 nm (NanoDropTM, Thermo Scientific) and with PicoGreen® reagent. The genome was sequenced using Illumina NOVASEQ 6000 instrument at Macrogen (Seoul, Korea). Paired-end reads of 150 bases were generated and deposited at the SRA database under the accession number SRR14522748.

Cryptosporidium hominis De Novo Genome Assembly

Genomic reads from each experiment were independently assembled using SPADES v3.14.1. First, reads ends were cleaned with rapifilt (homebrew program) filtering at Q30 and only keeping reads with a minimum of 50 bases. Then, assembly was performed using SPADES parameters: -careful -t 40 -m 160 -k 33,55,77,99.

Each assembly was filtered to excluded contaminating sequences using BLASTN (v2.10.1+). Only those longer than 1,000 bases and with a bit score greater than 300 were kept in each *C. hominis* assembled genome dataset. The bit score was obtained mapping the scaffolds to the *C. parvum* IOWAII genome reference (CryptoDB v 52), using BLASTN with parameters -evalue 1e-30 -num_alignments 5.

The genomes stats (Total length of sequence, Largest contig, N50 stats, Total number of sequences) were obtained using an in-house python script (see **Supplementary Material**). The "AltSNPs" (alternate single nucleotide polymorphisms) and "Coverage" of the assembly, were obtained by mapping the cleaned reads of each experiment against the assembled genome, using bowtie2 (version 2.4.1) with default parameters and using samtools (v1.10) view -F 3584 to keep only mapped reads. The coverage was obtained using samtools coverage and obtaining the mean and median of the mean depth of each scaffold. While the SNVs were obtained by counting the number of variants from the Variant Call Format (VCF) created as bcftools mpileup--redo-BAQ--in-BQ 30--per-sample-mF--skip-indels, then bcftools call--multiallelic-caller--variants-only-Ov and then bcftools view -i "%QUAL ≥ 30."

The SNVs and Identity vs. *C. hominis* UDEA01 were obtained comparing each genome with the *C. hominis* UDEA01 using DNAdiff (version 1.3) and getting the TotalSNPs and AvgIdentity from the.report file. The *C. hominis* genomes that had less than 8,167,000 assembled bases and with more than 3,825 AltSNPs were dropped as they did not meet basic quality parameters, to obtain the selected genomes that are further analyzed. We run BUSCO v5.2.2 (Simão et al., 2015) with parameters "busco-m geno-l coccidia_odb10 -i" for each selected genomes, including outgroups.

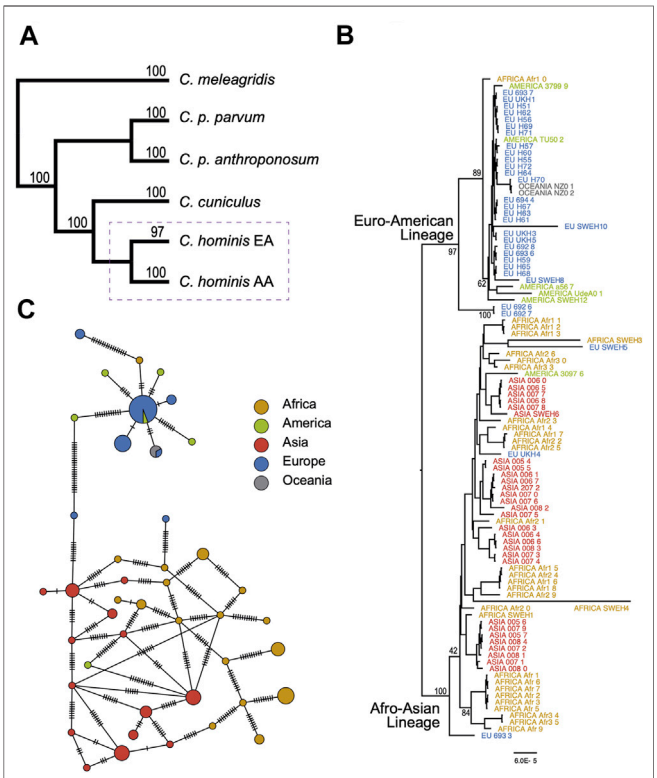


FIGURE 3 | Evolutionary relationships among species and populations of *Cryptosporidium hominis*. **(A)** Summary of main phylogenetic relationships among the main human-infecting *Cryptosporidium* species. The complete tree is provided in the **Supplementary Material**. The dashed rectangle is expanded in **(B,C)**. **(B)** Subtree depicting the phylogenetic relationships among *C. hominis* isolates from five continents. Tree inference based on 61 coding genes and 140,425 sites. **(C)** Haplotype network based on 116 segregating sites across 61 coding genes (see methods for details). The size of the circles is proportional to the number of individuals with each haplotype. The cross lines represent the number of expected mutations among haplotypes.

The indels vs *C. parvum* IOWAII were obtained using nucdiff (v2.0.3), with default parameters for each selected genome and the additional *C. parvum*, *C. meleagridis*, and *C. cuniculus* genomes. Using an in-house python script the indels were extracted from "_query_snps.gff" files. The principal component analysis (PCA) analysis was created with python's sklearn package and plotted with matplotlib.

Phylogenomic Analysis

The selected *C. hominis* genomes, together with the *C. parvum, C. meleagridis* and *C. cuniculus* genomes were used to infer a phylogenetic tree. The sixty one neutrally evolving genes of *C. parvum* described by Nader et al. (2019) were extracted from each genome. Each gene from all genomes were aligned using MAFFT (version v7.475) with parameters--inputorder--adjustdirection--anysymbol--auto. All the aligned genes were concatenated using catsequences. These aligned sequences were used to infer a

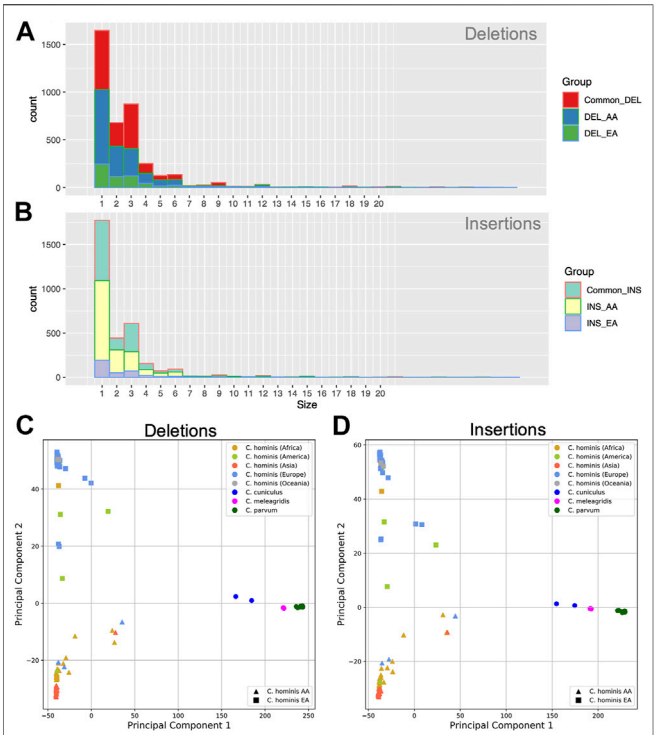


FIGURE 4 | Genomic insertion and deletion events in *C. hominis* lineages. **(A)** Genomic deletion event length (base pair) frequency common to both lineages (Common_DEL), EA and AA. DEL_AA: Specific to the AA lineage. DEL_EA: specific for the EA lineage. **(B)** Genomic Insertion events length (base pair) frequency common to both lineages (Common_INS), EA and AA. INS_AA: Specific to the AA lineage. INS_EA: specific for the EA lineage. **(C)** Principal component analysis (PCA) of the genomic deletion events detected in *C. hominis*. **(D)** Principal component analysis (PCA) of the genomic insertion events detected in *C. hominis*.

maximum likelihood tree using IQTREE2 (2.1.2 COVID-edition for Linux 64-bit) with parameters -B 5000-T AUTO -m MFP+MERGE -rcluster 10 (Lanfear et al., 2012; Lanfear et al., 2014).

Best-fit models selected according to BIC for each coding gene were: K3Pu + F + G4:cgd1_1450 + cgd4_4440 + cgd6_2560 + cgd7_1810 + cgd7_2600 + cgd8_3560,TPM3u + F +G4:cgd1_1730 + cgd1_3650 + cgd1_640 + cgd6_2720 + cgd6_5300 + cgd7_1270 + cgd7_340 + cgd7_890,TIM2 + F + G4:cgd1_2000 + cgd1_3790 + cgd2_2470 + cgd3_4230 + cgd4_2820 + cgd5_2250 + cgd5_4240 + cgd7_2340 + cgd8_2850,TN + F + G4:cgd1_3780 + cgd3_1720 + cgd4_3800 + cgd4_4360 + cgd6_4090 + cgd6_4280 + cgd6_5370 + cgd8_2080 + cgd8_830,TN + F + G4:cgd2_180 + cgd2_3630 + cgd8_1960 + cgd8_3030 + cgd8_5310,HKY + F:cgd2_2060 + cgd5_1340,TN + F + G4:cgd2_3110 + cgd3_1010,K3Pu + F + G4:cgd2_3810 + cgd2_940 + cgd3_380 + cgd4_2620 + cgd5_2890 + cgd7_1330 + cgd7_3550,TPM3 + F + G4:cgd3_2600 + cgd5_3600 + cgd6_2100,HKY + F + I:cgd3_3070 + cgd3_3650 + cgd4_2210 + cgd5_2730,TPM3 + F + G4:cgd3_3310 + cgd5_2700,HKY + F + G4:cgd4_2180 + cgd4_370 + cgd5_1860,TPM2 + F + G4:cgd8_140

We built a haplotype network using the minimum spanning network algorithm (Bandelt et al., 1999) and implemented in PopART (Leigh and Bryant, 2015). We included 95 genomes and excluded four genomes (Chom_EU_SWEH8, Chom_EU_SWEH5, Chom_AFRICA_SWEH3, Chom_AFRICA_SWEH4) because their long branches inferred in the phylogeny, resulting most likely from sequencing errors instead of true genetic variation. The dataset included 118 segregating sites across the 61 genes. The complete step by step analysis pipeline can be found in **Supplementary Material**.

Statistical and Graphical Analysis

Statistical analysis and graphics were done in R [R version 4.0.4 64, x86_64-apple-darwin17.0 (64-bit)] and R studio (Version 1.2.1335, Macintosh; Intel Mac OS X 10_16_0). Boxplots were generated in R with the function "boxplot." Central tendency measures and inter-quartile ranges (IQR) were calculated with R functions "mean," "median," "quantile," and "IQR."

RESULTS

Cryptosporidium hominis Genome Quality Analysis

One hundred and nineteen *Cryptosporidium hominis* genomes were included in the present study. A new Colombian *C. hominis* genome with the code UDEAa567 was also generated following *C. hominis* UdeA01 strategy (Isaza et al., 2015): the oocysts were enriched using a flotation protocol from a stool sample from an HIV + infected individual and then the NGS sequencing was performed using an Illumina Novaseq 6,000 instrument. *C. hominis* UDEAa567 assembly genome size was 9,052,438 bp. with an N50 value of 93,658 bp. The mean and median read depths were 10.2X and 10.5X, respectively. The UDEAa567 has 99.93% genome identity and 2,655 Single nucleotide variants (SNVs) respect *C. hominis* UDEA01.

To normalize the analytical conditions for all *C. hominis* genomes, the processing started with the original raw read data, and then they were all assembled using the same strategy. The *C. parvum* IOWA II genome (version 52) was downloaded from CryptoDB. All the read sequences used (*C. hominis*, *C. parvum*, *C. cuniculus*, *C. meleagridis*) are publicly available at the NCBI/SRA website (see *Materials and Methods*).

The *C. hominis* genome projects were executed in the last 10 years with different Illumina instruments, different library preparation kits, and were conducted in different laboratories around the world. A small set of these genomes were generated on the ION TORRENT platform. To assess the quality of the assemblies, their general metrics were compared in **Figure 1**: total assembled bases, median sequencing depth, assembly N50 values, genome nucleotide identity with *C. hominis* UDEA01, alternate allele count, BUSCO genome completeness, and BUSCO single-copy genes (**Supplementary Table S1**). The *Cryptosporidium* assembled genome sizes ranged between 1,148,020 and 10,587,153 Mbp. Despite the broad range of the assembled bases metric, a very narrow dispersion range (an IQR of 21,974 bases) was observed around the median, 9,073,830 bp

(**Figure 1A**). By contrast, the observed median sequencing depth and its IQR were both very variable. The depth ranged from 1.1× to 714×, with a median value of 183× and an IQR of 197 (**Figure 1B**).

The assembly N50 value also showed a very broad range, from 1,302 to 1,106,561 bp. The median N50 value was 495,148 bases with an IQR of 597,468 (**Figure 1C**). Furthermore, the global nucleotide identity of all genomes against the *C. hominis* UDEA01 reference was calculated. They have a median identity of 99.9% and an IQR of 0.02% (**Figure 1E**). Only 10 genomes had a global identity value below 99.7% (Chom_EU_4127, Chom_AFRICA_SWEH7, Chom_EU_SWEH9, Chom_EU_6940, Chom_EU_6946, Chom_EU_6925, Chom_EU_6934, Chom_EU_4120, Chom_EU_4128, Chom_EU_4118). These genomes are poor-quality references since they are outliers in either assembled genome size or alternative allele count metrics (**Supplementary Table S1**).

To assess the possible *Cryptosporidium* coinfections, the reads of each isolate were mapped to its respective assembly and alternate alleles were counted. The observed median number of genomic positions with an alternate allele was 83, with an IQR of 196 (**Figure 1D**). Eighty-five percent of the *C. hominis* isolates showed alternate allele counts below 470. Then, to estimate the genetic variations among the *C. hominis* genomes, we counted the number of SNVs against the *C. hominis* UDEA01 reference. This analysis showed a median value of 3,324 SNVs with an IQR of 831 (**Figure 1F**). Ninety percent of the *C. hominis* genomes had less than 4,000 SNVs regardless its continental origin.

We analyzed the *C. hominis* genomes with BUSCO using the coccidia ODB10 database. The BUSCO analysis shows the same median and IQR values for both completeness and single-copy metrics, 98.6 and 0.2%, respectively (**Figures 1G,H**). The duplicated BUSCO median value was 0 with an IQR of 0. All *C. hominis* genomes, except 1 (Chom_EU_6925), showed duplicated values below 1%. Finally, the BUSCO fragmented metrics showed a median value of 0.4% with an IQR of 0.2% (**Supplementary Table S1**).

The genomic descriptive measurements showed outliers in several statistic indices, indicating the presence of poor-quality genomes. To filter out them, we set two thresholds: assembled genome size and alternate allele count. The first parameter is directly related to the coverage achieved for each genome and failing to reach a minimum threshold will affect the assembly size and also the accuracy of the assembly. The second parameter, alternate allele count, might indicate the presence of more than one *Cryptosporidium* genome in the human stool sample. The proposed thresholds rationale are: 1) Assembly size > 8,166,447 bp., which corresponds to 90% of the median value of *C. hominis* assembled genome size. We think that a representation of 90% of the expected assembly genome size is enough for subsequent comparative or phylogenetic analysis and is equivalent to a BUSCO genome completeness score of >90% (**Supplementary Table S1**); and 2) AltSNVs threshold of 3,992 SNVs, which corresponds to the percentile 90th of the observed HQ SNVs of the analyzed *C. hominis* genomes (**Figure 2**). Applying these two thresholds, 20 genomes were excluded for further analysis, 18 with European origin and 2 with African origin (Chom_EU_SWEH9, Chom_EU_UKH6, Chom_EU_SWEH13, Chom_EU_SWEH11, Chom_EU_H58, Chom_EU_6946, Chom_EU_6945, Chom_EU_6942, Chom_EU_6940, Chom_EU_6939, Chom_EU_6935, Chom_EU_6934, Chom_EU_6925, Chom_EU_6919, Chom_EU_4128, Chom_EU_4127, Chom_EU_4120, Chom_EU_4118, Chom_AFRICA_SWEH7, and Chom_AFRICA_SWEH2). After filtering, the number of genomes per continent was Africa, 33; America, 6; Asia, 29; European Union, 49; Oceania, 2.

Phylogenomic Analysis of the *Cryptosporidium hominis* Species

Following the previous work published by Nader et al. (2019), 61 genes identified as neutrally evolving in *C. parvum* and *C. hominis* were used for the phylogenomic analysis. After filtering out the low-quality *C. hominis* genomes, ninety-nine genomes were kept for the phylogenetic reconstruction. As outgroups, 15 genomes were used: two *C. cuniculus* (UKCU5 and UKCU2), 8 *C. parvum parvum* (UKP2, UKP3, UKP4, UKP5, UKP6, UKP7, and UKP8), three *C. parvum anthroponosum* (UKP14 and UKP15), and three *C. meleagridis* (UKMEL1, UKMEL3, and UKMEL4).

A maximum-likelihood tree was constructed encompassing 114 tips. The total matrix of the 61 aligned genes encompasses 140,425 sites. BIC partitions were reduced to 13 (see *Materials and Methods*). The consensus tree has a Log-likelihood of −266,302.24. The basal branches recreate the previously described topology for the analyzed species, with *C. meleagridis* defined as the outgroup. The *C. parvum* subspecies clades *C. parvum* parvum and *C. parvum anthroponosum* are reciprocally monophyletic with 100% bootstrap support. The clade *C. hominis* comprises two well-supported lineages with strong continental bias: One lineage representing isolates from Asia and Africa (100% nodal support), and the other lineage representing genomes isolated from European or American infected humans (97% nodal support). The two isolates from Oceania are part of the European-American clade (**Figure 3A**). From now on, these two lineages will be referred to as the *C. hominis* Euro-American lineage (EA) and the Afro-Asian lineage (AA). The consensus tree shows very well-supported basal nodes but low support at the terminal branches due to low divergence among closely related genomes (**Figure 3B**). It is also noteworthy that the *C. hominis* EA lineage is populated by Ib genotypes, mostly IbA10G2/IbA12G2, except for *C. hominis* UdeA01, which has been classified with the genotype Ie.

Four terminal branches showed exceptional longer branch lengths, one in the AA lineage, and three in the EA lineage. It is also particular to see that these four samples come from the same study and were sequenced with the same instrument, the ION TORRENT.

To further confirm the lineage segregation in *C. hominis*, haplotype analysis was performed based on the 61 neutrally evolving genes. Haplotype network is congruent with the two main lineages (Euro-American and Afro-Asian lineages,

Figure 3C). However, we did not find structure within these two lineages, and haplotype diversity appears to be higher in the Afro-Asian lineage.

Genomic Deletions Analysis

Indel events are common in *Cryptosporidium* parasites (Arias-Agudelo et al., 2020). To assess the genomic deletion and insertion profiles within *C. hominis* species, genomic alignment of the assemblies against the *C. parvum* IOWA II reference was performed and analyzed. Insertion and deletion events were treated independently to give more clarity.

Deletion events were more common than insertions and spanned from 1 to 78 bases. The most frequent loss harbored a single base and was followed in frequency by the loss of three bases (**Figure 4A**). All the *C. hominis* genomes displayed a set of 3,958 shared deletion events. Furthermore, there are 609 specific deletion events in *C. hominis* EA lineage, and 2,338 specific deletion events in AA lineages.

Like the deletions, the insertion events spanned from 1 to 68 bases, being, in order, more frequent those that involve one and three bases (**Figure 4B**). Most of the insertion events were shared between both lineages (n = 3,272). Nonetheless, each lineage displayed specific insertion events. The AA lineage has 1,996 common insertions, while the EA lineage has 380.

Indel events may help to understand the relationship among the different parasite clades. With the aim to test the informative signal of the indel event profile of each genome, a Principal component analysis (PCA) analysis of the studied genomes was performed. As can be seen in **Figures 4C,D**, deletions or insertion event profiles segregate the parasites according to species and the lineage of *C. hominis*.

DISCUSSION

We are at the dawn of a new era in microbiology. Nowadays, genomes of pathogens are analyzed directly from infected tissues or other human-derived samples like feces. Next-generation sequencing technologies not only fostered, in general, the high-resolution analysis of life on Earth but also prompted the genomic analysis of non-culturable pathogens like *C. hominis*. In this apicomplexan, we can observe how metagenomic analyses nurtured the advance in comparative and evolutionary genomics. Hundreds of reference genomes were generated in the last decade using the oocysts enrichment and metagenomic shotgun sequencing strategy. (Guo et al., 2015; Hadfield et al., 2015; Isaza et al., 2015; Ifeonu et al., 2016; Sikora et al., 2017; Gilchrist et al., 2018; Morris et al., 2019; Nader et al., 2019; Xu et al., 2019; Arias-Agudelo et al., 2020; Tichkule et al., 2021).

This metagenomic strategy poses challenges like the need for new standardized informatics methods to validate genome quality and purity since the infected individual could harbor a mixture of *Cryptosporidium* parasites: either different lineages of the same species or even multispecies infections. Furthermore, this strategy offers a great opportunity to study pathogens without the bias of artificial selection *in vitro* cultures and observe their genomes in natural settings under the complex conditions where they must co-exist and compete with other microorganisms.

The *C. hominis* genomes, sequenced so far, were generated in very dissimilar conditions, with different oocysts isolation protocols, and sequenced with different instruments in the last decade (Hadfield et al., 2015; Isaza et al., 2015; Ifeonu et al., 2016; Sikora et al., 2017; Gilchrist et al., 2018; Morris et al., 2019; Nader et al., 2019; Tichkule et al., 2021). The results presented in this work show that descriptive genome assembly statistics differ significantly, although some metrics have very narrow dispersion values enabling a rapid way to detect outliers. For instance, *C. hominis* assembled genome size has a narrow dispersion around the median value of 9,074,262 bp (IQR 20688 bp.), regardless of its continental origin or the NGS platform used. Additionally, the global nucleotide identity and SNVs count against the *C. hominis* UDEA01 also have a narrow dispersion close to their respective median values. This might be used as a marker for intraspecies boundaries for *C. hominis* genome around 99.7% identity and 4000 SNVs.

The median sequencing depth and the N50 values show the highest dispersion ranges. Median sequencing depth depends mainly on two main factors: The efficiency of the *Cryptosporidium* oocysts enrichment process, and the sequencing scale (number of reads generated). Even after an efficient enrichment process, *C. hominis* oocysts will be accompanied by intestinal microbiota, mainly bacteria. The more bacteria present in the sample, the fewer reads that will be available to support the target *Cryptosporidium* genome.

Insufficient sequencing depth might generate smaller assembled genome sizes, less accurate scaffolds, lower N50 values, and more fragmented genome models. Another situation that should be considered is that mixed parasite populations in one infected individual might lead, even under sufficient sequencing depth conditions, to more fragmented assemblies. In the case of samples with mixed *Cryptosporidium* species, it might be possible to observe outlier assemblies with larger-than-expected genome sizes and higher counts of alternate alleles.

It is noteworthy to mention that the phylogenomic analysis also enables to detection of outlier genome models since suspicious genomes with long branches can be spotted. In this work, four genomes showed extreme branch lengths, albeit they passed the genome quality thresholds. These four genomes come from different continents but have in common the instrument used for NGS data generation, the ION TORRENT. This instrument has shown to have higher mutation rates than Illumina instruments, supporting the hypothesis that these longer branches are more likely associated with sequencing artifacts rather than real higher evolution rates (Quail et al., 2012). This observation implies that genome assembly statistics alone are not enough for quality control. Complementary phylogenomic analyses help to detect genome assemblies that could accumulate higher error rates.

In general, the analyzed *C. hominis* genomes have low alternate allele counts, 85% with less than 470 events,

suggesting that most of the infected individuals might harbor one clonal *Cryptosporidium* genome with minor nucleotide variants (0.005%).

The highest number of alternate alleles went as high as 112,225, nearly 34 times higher than the observed median value. In this specific patient, we might be dealing with a case of *Cryptosporidium* mixed species infection. This observation is supported by the facts that the global genome identity went down to 98.96% and that the *Cryptosporidium* assembled genome size was exceptionally large, exceeding 11 Mbp.

Several research works have described the presence of mixed *C. hominis* genotypes or even species in one infected individual (Cama et al., 2006; Gilchrist et al., 2018; Korpe et al., 2019; Sannella et al., 2019). These observations can be reconciled with our results considering that coinfecting *C. hominis* genotypes of other *Cryptosporidium* species could be below the detection threshold of the actual metagenomic approaches.

The phylogenomic analysis allowed us to discover two *C. hominis* lineages with continental segregation. These lineages were named based on the observed continental distribution bias as *C. hominis* Euro-American (EA) lineage and the *C. hominis* Afro-Asian (AA) lineage.

The reference isolate UDEA01 is part of the EA lineage, as well as the other Latin-American isolates UdeAa567 and SWEH12. This clade also encompasses isolates of the United States, the United Kingdom, Spain, and Greece. Only one genome belongs to a different continental region, Afr10, which was isolated from a human in Madagascar, Africa. All the genomes in the EA clade are classified as genotypes IbA10G2/IbA12G2 except for *C. hominis* UDEA01, which has been classified as genotype IeA11G3T3. Using an SNVs phylogenomic approach, Sikora et al. (2017) made a previous observation that supports our findings. In their study, *C. hominis* genomes of the genotype IbA10G2 separated in a monophyletic clade independent of other studied *C. hominis* genotypes. Another line of evidence that supports our finding was published recently by Tichkule et al. In their work, the researchers observed a geographic structuring of the *C. hominis* isolates in African countries (Tichkule et al., 2021).

Haplotype network analysis supported the geographical isolation of the EA and AA *C. hominis* lineages. Additionally, the AA lineage showed a higher haplotype diversity. Nevertheless, it must be highlighted that the sampling was higher in the AA lineage.

INDEL events are frequent in *Cryptosporidium* genomes (Arias-Agudelo et al., 2020). PCA analysis of the INDEL profiles showed that they can segregate *C. hominis* of the other studied *Cryptosporidium* species: *C. parvum*, *C. meleagridis*, and *C cuniculus*. Additionally, INDEL events profile showed to be taxonomically informative in *C. hominis* since they were able to segregate the two lineages, adding support to the previous observe results.

Summarizing, three independent lines of evidence support our observation that *C. hominis* encompasses two separate lineages. These two monophyletic lineages have a differential continental distribution pattern with a clear dominance of one lineage in Eastern Europe and the Americas, while the other is the main lineage in Africa and Asia.

Previous research works on infectious diseases have demonstrated a phylogeographic relationship among European and American human populations. This is the case for the etiological agent of human tuberculosis, *Mycobacterium tuberculosis*. This pathogen presents a specific lineage, called the LAM family (L4.3), which is more common in the Mediterranean and Latin-American countries, and showed the spread of the bacteria associated with human migrations from Europe to America (Gagneux et al., 2006; Brynildsrud et al., 2018).

By the beginning of the 2000s *C. parvum* and *C. hominis* were considered a single species (Sulaiman et al., 2000), with the name of the former as the type species. Two years later, Morgan-Ryan et al., using complementary molecular analysis showed that, in fact, two closely related species could be discriminated and *C. hominis* and *C. parvum* were separated as independent species (Morgan-Ryan et al., 2002).

Ending the next decade, in 2019, within the *C. parvum* clade, a new subspecies was discovered, *C. parvum antroponosum* (Nader et al., 2019). *Cryptosporidium parvum* parvum is described as a zoonotic parasite while the new subspecies *C. parvum* antroponosum has an anthroponotic transmission scheme. These discoveries were achieved thanks to the incorporation of more resolutive molecular analytical methods, like phylogenomic tools.

In this work, we present a similar observation for *C. hominis*, this new lineage may represent a new subspecies, but its confirmation and biological implications should be further investigated. Previous evolutionary works have shown an evolution model in *Cryptosporidium* parasites where speciation events are related to new host adaptations. In almost every major vertebrate lineage, an adapted *Cryptosporidium* species have been described (Garcia-R and Hayman, 2016). In the case of *C. hominis*, a particularly interesting question raises since *C. hominis* has been already adapted to our human ancestors around 6 million years ago (Garcia-R and Hayman, 2016). Additional work should be performed in order to establish the origin of the separation of the *C. hominis* lineages.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories (NCBI SRA). See **Supplementary Table S1**. New Colombian C hominis genome read data were deposited in the NCBI SRA database accession number SRR14522748 (https://dataview.ncbi.nlm.nih.gov/object/SRR14522748).

AUTHOR CONTRIBUTIONS

FC: Bioinformatic analysis design, Data processing, and statistical analysis. Article writing and review. AG-D: Parasite sample collection and enrichment. Patient enrollment and bioethics follow-up. Literature review. Articles writing and review. LA: Parasite wet-lab

experiments. Parasite enrichment and DNA extraction and QC. Literature review. GG-M: Parasite sample collection and enrichment. Patient enrollment and bioethics follow-up. Articles review. JD: Phylogenetic and haplotype analysis. Articles review. JA: Study design. Bioinformatic analysis design, Data processing, and statistical analysis. Article writing and review.

FUNDING

COLCIENCIAS Convocatoria 777-2017 Para Proyectos de Ciencia, Tecnología e Innovación en Salud 2017, project name: "Estudio de prevalencia, diversidad genética y genomica de Cryptosporidium en población VIH positiva de Antioquia'." 111577757608. Vicerrectoría de Investigación (CODI), Universidad de Antioquia, grant "Finalización del genoma de referencia de C. hominis aislado UdeA01'": 2017-16-171.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.740940/full#supplementary-material

REFERENCES

Arias-Agudelo, L. M., Garcia-Montoya, G., Cabarcas, F., Galvan-Diaz, A. L., and Alzate, J. F. (2020). Comparative Genomic Analysis of the Principal Cryptosporidium Species that Infect Humans. *PeerJ* 8, e10478. doi:10.7717/peerj.10478

Bandelt, H. J., Forster, P., and Röhl, A. (1999). Median-Joining Networks for Inferring Intraspecific Phylogenies. *Mol. Biol. Evol.* 16 (1), 37–48. doi:10.1093/oxfordjournals.molbev.a026036

Brynildsrud, O. B., Pepperell, C. S., Suffys, P., Grandjean, L., Monteserin, J., Debech, N., et al. (2018). Global Expansion of Mycobacterium Tuberculosis Lineage 4 Shaped by Colonial Migration and Local Adaptation. *Sci. Adv.* 4 (10), eaat5869. doi:10.1126/sciadv.aat5869

Cacciò, S. M., and Chalmers, R. M. (2016). Human Cryptosporidiosis in Europe. *Clin. Microbiol. Infect.* 22 (6), 471–480. doi:10.1016/j.cmi.2016.04.021

Cama, V., Gilman, R. H., Vivar, A., Ticona, E., Ortega, Y., Bern, C., et al. (2006). MixedCryptosporidiumInfections and HIV. *Emerg. Infect. Dis.* 12 (6), 1025–1028. doi:10.3201/eid1206.060015

Efstratiou, A., Ongerth, J. E., and Karanis, P. (2017). Waterborne Transmission of Protozoan Parasites: Review of Worldwide Outbreaks - an Update 2011-2016. *Water Res.* 114, 14–22. doi:10.1016/j.watres.2017.01.036

Feng, Y., Ryan, U. M., and Xiao, L. (2018). Genetic Diversity and Population Structure of Cryptosporidium. *Trends Parasitol.* 34, 997–1011. doi:10.1016/j.pt.2018.07.009

Gagneux, S., Deriemer, K., Van, T., Kato-Maeda, M., de JongDe Jong, B. C., Narayanan, S., et al. (2006). Variable Host-Pathogen Compatibility in Mycobacterium Tuberculosis. *Proc. Natl. Acad. Sci. U S A.* 103 (8), 2869–2873. Available at: www.pnas.org. doi:10.1073/pnas.0511240103

Garcia-R, J. C., and Hayman, D. T. S. (2016). Origin of a Major Infectious Disease in Vertebrates: The Timing of Cryptosporidium Evolution and its Hosts. *Parasitology* 143 (13), 1683–1690. doi:10.1017/S0031182016001323

Gilchrist, C. A., Cotton, J. A., Burkey, C., Arju, T., Gilmartin, A., Lin, Y., et al. (2018). Genetic Diversity of Cryptosporidium Hominis in a Bangladeshi Community as Revealed by Whole-Genome Sequencing. *J. Infect. Dis.* 218 (2), 259–264. doi:10.1093/infdis/jiy121

Guo, Y., Tang, K., Rowe, L. A., Li, N., Roellig, D. M., Knipe, K., et al. (2015). Comparative Genomic Analysis Reveals Occurrence of Genetic Recombination in Virulent Cryptosporidium Hominis Subtypes and Telomeric Gene Duplications in Cryptosporidium Parvum. *BMC Genomics* 16 (1), 320. doi:10.1186/s12864-015-1517-1

Hadfield, S. J., Pachebat, J. A., Swain, M. T., Robinson, G., Cameron, S. J., Alexander, J., et al. (2015). Generation of Whole Genome Sequences of New Cryptosporidium Hominis and Cryptosporidium Parvum Isolates Directly from Stool Samples. *BMC Genomics* 16 (1), 650. doi:10.1186/s12864-015-1805-9

Ifeonu, O. O., Chibucos, M. C., Orvis, J., Qi, S., Elwin, K., Guo, F., et al. (2016). Annotated Draft Genome Sequences of Three Species of Cryptosporidium: Cryptosporidium Meleagridis Isolate UKMEL1, C. Baileyi Isolate TAMU-09Q1 and C. Hominis Isolates TU502 2012 and UKH1. *Pathog. Dis.* 74 (7), 1–5. doi:10.1093/femspd/ftw080

Innes, E. A., Chalmers, R. M., Wells, B., and Pawlowic, M. C. (2020). A One Health Approach to Tackle Cryptosporidiosis. *Trends Parasitol.* 36, 290–303. doi:10.1016/j.pt.2019.12.016

Isaza, J. P., Galván, A. L., Polanco, V., Huang, B., Matveyev, A. V., Serrano, M. G., et al. (2015). Revisiting the Reference Genomes of Human Pathogenic Cryptosporidium Species: Reannotation of C. Parvum Iowa and a New C. Hominis Reference. *Sci. Rep.* 5, 16324. doi:10.1038/srep16324

Kar, S., Gawlowska, S., Daugschies, A., and Bangoura, B. (2011). Quantitative Comparison of Different Purification and Detection Methods for Cryptosporidium Parvum Oocysts. *Vet. Parasitol.* 177 (3), 366–370. doi:10.1016/j.vetpar.2010.12.005

Korpe, P. S., Gilchrist, C., Burkey, C., Taniuchi, M., Ahmed, E., Madan, V., et al. (2019). Case-Control Study of Cryptosporidium Transmission in Bangladeshi Households. *Clin. Infect. Dis.* 68 (7), 1073–1079. doi:10.1093/cid/ciy593

Kotloff, K. L., Nataro, J. P., Blackwelder, W. C., Nasrin, D., Farag, T. H., Panchalingam, S., et al. (2013). Burden and Aetiology of Diarrhoeal Disease in Infants and Young Children in Developing Countries (The Global Enteric Multicenter Study, GEMS): A Prospective, Case-Control Study. *Lancet* 382 (9888), 209–222. doi:10.1016/S0140-6736(13)60844-2

Lake, I. R., Harrison, F. C. D., Chalmers, R. M., Bentham, G., Nichols, G., Hunter, P. R., et al. (2007). Case-Control Study of Environmental and Social Factors Influencing Cryptosporidiosis. *Eur. J. Epidemiol.* 22 (11), 805–811. doi:10.1007/s10654-007-9179-1

Lanfear, R., Calcott, B., Ho, S. Y. W., and Guindon, S. (2012). PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. *Mol. Biol. Evol.* 29 (6), 1695–1701. doi:10.1093/molbev/mss020

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., and Stamatakis, A. (2014). Selecting Optimal Partitioning Schemes for Phylogenomic Datasets. *BMC Evol. Biol.* 14 (1), 82–14. doi:10.1186/1471-2148-14-82

Leigh, J. W., and Bryant, D. (2015). Popart : Full-Feature Software for Haplotype Network Construction. *Methods Ecol. Evol.* 6 (9), 1110–1116. doi:10.1111/2041-210X.12410

Morgan-Ryan, U. M., Fall, A., Ward, L. A., Hijjawi, N., Sulaiman, I., Payer, R., et al. (2002). Cryptosporidium Hominis N. Sp. (Apicomplexa: Cryptosporidiidae) from Homo Sapiens. *J. Eukaryot. Microbiol.* 49 (6), 433–440. doi:10.1111/j.1550-7408.2002.tb00224.x

Morris, A., Robinson, G., Swain, M. T., and Chalmers, R. M. (2019). Direct Sequencing of Cryptosporidium in Stool Samples for Public Health. *Front. Public Health* 7, 360. doi:10.3389/fpubh.2019.00360

Nader, J. L., Mathers, T. C., Ward, B. J., Pachebat, J. A., Swain, M. T., Robinson, G., et al. (2019). Evolutionary Genomics of Anthroponosis in Cryptosporidium. *Nat. Microbiol.* 4 (5), 826–836. doi:10.1038/s41564-019-0377-x

Pollock, K. G. J., Ternent, H. E., Mellor, D. J., Chalmers, R. M., Smith, H. V., Ramsay, C. N., et al. (2010). Spatial and Temporal Epidemiology of Sporadic Human Cryptosporidiosis in Scotland. *Zoonoses and Public Health* 57 (7–8), 487–492. doi:10.1111/j.1863-2378.2009.01247.x

Quail, M., Smith, M. E., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A Tale of Three Next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers. *BMC Genomics* 13, 341. doi:10.1186/1471-2164-13-341

Ryan, U., Fayer, R., and Xiao, L. (2014). Cryptosporidiumspecies in Humans and Animals: Current Understanding and Research Needs. *Parasitology* 141 (13), 1667–1685. doi:10.1017/S0031182014001085

Sannella, A. R., Suputtamongkol, Y., Wongsawat, E., and Cacciò, S. M. (2019). A Retrospective Molecular Study of Cryptosporidium Species and Genotypes in HIV-Infected Patients from Thailand. *Parasites Vectors* 12 (1), 91. doi:10.1186/s13071-019-3348-4

Sikora, P., Andersson, S., Winiecka-Krusnell, J., Hallström, B., Alsmark, C., Troell, K., et al. (2017). Genomic Variation in IbA10G2 and Other Patient-Derived Cryptosporidium Hominis Subtypes. *J. Clin. Microbiol.* 55 (3), 844–858. doi:10.1128/JCM.01798-16

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* 31 (19), 3210–3212. doi:10.1093/bioinformatics/btv351

Sulaiman, I. M., Morgan, U. M., Thompson, R. C. A., Lal, A. A., and Xiao, L. (2000). Phylogenetic Relationships of Cryptosporidium Parasites Based on the 70-Kilodalton Heat Shock Protein (HSP70) Gene. *Appl. Environ. Microbiol.* 66 (6), 2385–2391. doi:10.1128/AEM.66.6.2385-2391.2000

Tichkule, S., Jex, A. R., Oosterhout, C. V., Rosa Sannella, A., Krumkamp, R., Aldrich, C., et al. (2021). Comparative Genomics Revealed Adaptive Admixture in Cryptosporidium Hominis in Africa. *Microb. Genomics* 7 (1), 1–14. doi:10.1099/mgen.0.000493

Widmer, G., Köster, P. C., and Carmena, D. (2020). Cryptosporidium Hominis Infections in Non-human Animal Species: Revisiting the Concept of Host Specificity. *Int. J. Parasitol.* 50, 253–262. doi:10.1016/j.ijpara.2020.01.005

Xiao, L., and Feng, Y. (2017). Molecular Epidemiologic Tools for Waterborne Pathogens Cryptosporidium Spp. And Giardia Duodenalis. *Food Waterborne Parasitol.* 8-9, 14–32. doi:10.1016/j.fawpar.2017.09.002

Xiao, L., and Feng, Y. (2008). Zoonotic Cryptosporidiosis. *FEMS Immunol. Med. Microbiol.* 52, 309–323. doi:10.1111/j.1574-695X.2008.00377.x

Xu, Z., Guo, Y., Roellig, D. M., Feng, Y., and Xiao, L. (2019). Comparative Analysis Reveals Conservation in Genome Organization Among Intestinal Cryptosporidium Species and Sequence Divergence in Potential Secreted Pathogenesis Determinants Among Major Human-Infecting Species. *BMC Genomics* 20 (1), 1–15. doi:10.1186/s12864-019-5788-9

Zahedi, A., and Ryan, U. (2020). Cryptosporidium – an Update with an Emphasis on Foodborne and Waterborne Transmission. *Res. Vet. Sci.* 132, 500–512. doi:10.1016/j.rvsc.2020.08.002

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Microbiology

REVIEW
published: 15 October 2021
doi: 10.3389/fmicb.2021.751775

The Elusive Mitochondrial Genomes of Apicomplexa: Where Are We Now?

Luisa Berná[1,2,3,4†], *Natalia Rego*[3†] *and María E. Francia*[1,5*]

[1] Laboratory of Apicomplexan Biology, Institut Pasteur de Montevideo, Montevideo, Uruguay, [2] Molecular Biology Unit, Institut Pasteur de Montevideo, Montevideo, Uruguay, [3] Bioinformatics Unit, Institut Pasteur de Montevideo, Montevideo, Uruguay, [4] Sección Biomatemática-Laboratorio de Genómica Evolutiva, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay, [5] Departamento de Parasitología y Micología, Facultad de Medicina, Universidad de la República, Montevideo, Uruguay

Mitochondria are vital organelles of eukaryotic cells, participating in key metabolic pathways such as cellular respiration, thermogenesis, maintenance of cellular redox potential, calcium homeostasis, cell signaling, and cell death. The phylum Apicomplexa is entirely composed of obligate intracellular parasites, causing a plethora of severe diseases in humans, wild and domestic animals. These pathogens include the causative agents of malaria, cryptosporidiosis, neosporosis, East Coast fever and toxoplasmosis, among others. The mitochondria in Apicomplexa has been put forward as a promising source of undiscovered drug targets, and it has been validated as the target of atovaquone, a drug currently used in the clinic to counter malaria. Apicomplexans present a single tubular mitochondria that varies widely both in structure and in genomic content across the phylum. The organelle is characterized by massive gene migrations to the nucleus, sequence rearrangements and drastic functional reductions in some species. Recent third generation sequencing studies have reignited an interest for elucidating the extensive diversity displayed by the mitochondrial genomes of apicomplexans and their intriguing genomic features. The underlying mechanisms of gene transcription and translation are also ill-understood. In this review, we present the state of the art on mitochondrial genome structure, composition and organization in the apicomplexan phylum revisiting topological and biochemical information gathered through classical techniques. We contextualize this in light of the genomic insight gained by second and, more recently, third generation sequencing technologies. We discuss the mitochondrial genomic and mechanistic features found in evolutionarily related alveolates, and discuss the common and distinct origins of the apicomplexan mitochondria peculiarities.

Keywords: Apicomplexa, mitochondria, mitogenome, genome, *Toxoplasma*, malaria, PacBio and ONT

OPEN ACCESS

Edited by:
Moses Okpeku, University of KwaZulu-Natal, South Africa

Reviewed by:
Lilach Sheiner, University of Glasgow, United Kingdom
Jessica C. Kissinger, University of Georgia, United States

***Correspondence:**
María E. Francia
mfrancia@pasteur.edu.uy

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *02 August 2021*
Accepted: *16 September 2021*
Published: *15 October 2021*

Citation:
Berná L, Rego N and Francia ME (2021) The Elusive Mitochondrial Genomes of Apicomplexa: Where Are We Now? Front. Microbiol. 12:751775. doi: 10.3389/fmicb.2021.751775

INTRODUCTION

Mitochondria are distinctive double-membrane organelles of central importance to the biology of eukaryotic cells. Mitochondria are ubiquitous through the eukaryotic domain. It is accepted that the last common ancestor of extant eukaryotes was a mitochondrion-containing organism generated by an event of endosymbiosis following the integration of an alpha-proteobacterium

into a cell related to Asgard archaea (Roger et al., 2017). Classically, mitochondria are regarded as the prime energy conversion hubs in aerobic organisms, as a product of the electrochemical potential generated along the respiratory chain. However, it is well established that mitochondrial functions extend well beyond respiration and energy conversion, impacting other biological processes central to life. These include the participation of mitochondria in apoptosis, amino acid metabolism, calcium homeostasis, pyruvate decarboxylation, folate, phospholipids, heme, and iron-sulfur clusters biosynthesis, among others (Maguire and Richards, 2014; Santos et al., 2018). Many of these functions are performed either by proteins coded for in the nucleus and imported into the mitochondrion, or by proteins coded for in the mitochondrial genome. The former arose mainly by past lateral gene transfer from the mitochondrion to the nucleus and repurposing and/or redirecting pre-existing nuclear genes. Anaerobic and facultative aerobes harbor functionally reduced and structurally simplified mitochondrion-related organelles, generally referred to as MROs. These include the hydrogenosomes and the mitosomes. In some cases mitochondria are no longer observable as a product of a secondary loss. In these cases, the lack of mitochondria can be tolerated by the novel acquisition of a functionally compensating cytosolic metabolic pathway. This is possible, in some instances, by lateral gene transfer from bacteria. Mitochondrial reduction accompanied by specific alterations in metabolic capabilities is often connected to adaptation to specific ecological niches (Seeber et al., 2008; Karnkowska et al., 2016).

Unlike other membrane bound organelles, with the exception of plastids, mitochondria bear their own extranuclear genome. The mitochondrial genome, from hereon referred to as "mitogenomes" can encode for tens of proteins, including those required for its maintenance (i.e., its replication). However, the coding information varies widely in nature. It can encompass part of the translation apparatus, including mitochondrial-specific tRNAs, large and small ribosomal RNAs (*rns* and *rnl)*; membrane associated proteins which catalyze oxidative phosphorylation: cytochrome b (*cob*), subunits of cytochrome c oxidase (*cox*), ATP synthase subunits 6, 8, and 9, subunits of the NADH dehydrogenase complex; and a few additional ORFs of unknown function. However, examples of mitogenomes displaying coding reduction are readily found in nature. Mitogenome sizes can vary widely in nature, ranging from 6 kb in protists to 2,400 kb in angiosperms.

Animal mitochondrial genomes exhibit an exceptional circular-supercoiled DNA topology (Lecrenier and Foury, 2000). This topology can also be found in kinetoplastids (Simpson, 1986). However, classical pulsed-field electrophoresis experiments from the 90's showed that most commonly, mitogenomes of plants and fungi exist in linear topologies (Bendich, 1993, 1996). Linear configurations are found as concatemers of monomeric subunits (Burger et al., 2003), arranged in head to head or head to tail configuration (Oldenburg and Bendich, 2001). Linear combinations of distinct subunits and *bona fide* linear monomeric mitogenomes have been documented. Illustrative examples of the former include the mitogenomes of the ciliate *Tetrahymena pyriformis* (Burger et al., 2000) and the algae *Chlamydomonas reinhardtii* (Gray and Boer, 1988).

The phylum Apicomplexa comprises over six thousand protozoan species of unicellular obligate intracellular parasites. Infamous pathogens causing an immense burden of mortality and morbidity, in both humans and animals, belong to this phylum. These include the causative agents of malaria, cryptosporidiosis, East Coast fever, tropical theileriosis, babesiosis and toxoplasmosis, among others. Though the pathobiology of apicomplexan diseases are notably distinct, parasites belonging to the phylum share a number of conserved structural features which underlie their pathogenesis. Namely, apicomplexans owe their name to a complex of apical secretory organelles equipped with effectors, adhesins and virulence factors that are hierarchically and sequentially deployed to allow recognition, invasion and subversion of their preferred host cell type (Dubremetz, 1973). In most species, each individual apicomplexan cell is equipped with a non-photosynthetic plastid remnant and a single, ramified mitochondrion, morphologically characterized by a dense matrix and inner mitochondrial membrane folds (cristae) with a circular profile (de Souza et al., 2009). Beyond these shared features, apicomplexans constitute a phylum of diverse organisms, showcasing an array of fascinating biological adaptations to the parasitism of specific niches. In this context, the presence, morphology and complexity of endosymbiotic organelles -the mitochondrion and the apicoplast- vary widely within lineages and between species.

The detailed study of the Api-mitochondrion and its complexity has been driven by the demonstration that Complex III of the mitochondrial electron transport chain (mETC) is a major drug target of the antimalarial atovaquone (Korsinczky et al., 2000; Mather et al., 2007; Vaidya et al., 2008; Biagini et al., 2012; Goodman et al., 2017; Pan et al., 2017). More recently, genome-wide genetic screens have deepened the interest. Genes coding for mitochondrial functions rank amongst the most important in conferring fitness, demonstrating the organelle's vital role in the survival of the causative agent of human malaria, *Plasmodium falciparum*, and of toxoplasmosis, *Toxoplasma gondii* (Sidik et al., 2016; Bushell et al., 2017). Complexome analyses have recently shown that the major mitochondrial complexes in *P. falciparum* and *T. gondii* are formed by a combination of conserved and divergent protein components further putting the organelle forward as a promising source of much sought-after novel drug targets (Huet et al., 2018; Salunke et al., 2018; Seidi et al., 2018; Hayward and van Dooren, 2019; Evers et al., 2021; Maclean et al., 2021).

Recent sequencing analyses using third generation sequencing, such as Pacific Bioscience (PacBio) and Oxford Nanopore Technologies (ONT) have revealed a highly variable, and unusual mitochondrial genome organization in a number of medically relevant apicomplexans. However, the contribution of the mitochondrial genome to the biology of Apicomplexa remains unclear. Apicomplexan parasites constitute a heterogenous group of fascinating model organisms ideally suited to explore and understand the diversity and functional significance of distinct configurations of the mitochondrial genome. Apicomplexan mitochondria also

represent extraordinary examples of selective retention, loss and gain, depicting unique features of evolution within the alveolates. This review summarizes the structure of apicomplexan mitochondrial genomes as currently understood. We highlight the unique and conserved apicomplexan mitogenome features, in the light of classical ultrastructural and biochemical insight, and emerging understanding brought by third generation sequencing technologies and single cell genomics.

PHYLOGENY OF THE PHYLUM APICOMPLEXA

The currently accepted phylogenetic classification of Apicomplexa was originally proposed by Levin in 1970. Based on electron microscopy studies from the 50's, this phylogeny contributed to reclassifying organisms originally grouped together under the term "Sporozoa." The apicomplexans were grouped based on the presence of a visually detectable apical complex. Under Levin's classification, groups of organisms such as Microspora, Myxozoa, and Ascetospora, were separated from Apicomplexa based on their distinct morphologies (Levine, 1988). The advent of molecular techniques in the 90's allowed insight into genomes and further refinement of this classification. Apicomplexans belong to the superphylum Alveolata, which they share with Chrompodellids, Perkinsozoa, Dinoflagellata, and Cilliophora.

Emerging data based on single-cell genomics of apicomplexan groups suggests that the currently accepted phylum, encompassing over 6,000 species, could actually be polyphyletic (Janouškovec et al., 2019; Mathur et al., 2021). In light of these results, it has been put forward that morphological similarities could have evolved convergently at least three times. Nonetheless, taxa important from a medical and veterinary perspective are all classified and grouped within Apicomplexa *sensu stricto* (Janouškovec et al., 2019), where species are further classified as coccidiomorphs, gregarines and cryptosporidians. Coccidiomorphs include species classified as agamococcidians, Coccidia and Hematozoa; arguably the most intensely studied and best understood species within the phylum belong to the latter two groups. Coccidians can be further classified as eucoccidians and protococcidians. Eucoccidians can be further classified as either "cyst forming" or monoxenic. The former includes *T. gondii*, *Neospora caninum, Hammondia hammondi, and Sarcocystis,* while the latter encompasses the family Eimeriidae. Eimeriidae is composed of multiple important genera, including *Eimeria*, *Isospora,* and *Cyclospora*. Species within these genera cause gastrointestinal disease, and characteristically complete their life cycle within a single host's intestinal tract.

Hematozoa can be further classified into Haemosporida; the malaria causing species of *Plasmodium* belong to this group, and Piroplasmida which includes *Theileria* and *Babesia* spp. Finally, the cyrptosporidians include all *Cryptosporidium* species; *C. hominis, C. parvum, C. muris,* among others. Recently, transcriptomic data allowed the assessment of the divergence pattern involving coccidiomorphs, gregarines and cryptosporidians and results support the placement of *Cryptosporidium* as the earliest diverging lineage of apicomplexans (Mathur et al., 2021; Salomaki et al., 2021). However, the history of early apicomplexan radiation remains highly contentious (Janouškovec et al., 2019; Mathur et al., 2019, 2021; Salomaki et al., 2021), which impacts our interpretation of evolutionary events within the phylum. **Figure 1** summarizes our current understanding of the phylogeny of alveolates.

THE MITOCHONDRIA OF APICOMPLEXANS

Electron microscopy from the 1960s clearly identified double membrane-bound mitochondrial organelles in many apicomplexans including *Plasmodium, Toxoplasma,* and *Eimeria* (Aikawa et al., 1966; Hepler et al., 1966; Rudzinska, 1969; Divo et al., 1985; Slomianny and Prensier, 1986; Ferguson and Dubremetz, 2014). However, it was noted that the cristae morphology differed from that found in mammals, and that there is a single mitochondrion per cell. Both in *Toxoplasma* and in *Plasmodium,* cristae are bulbous in shape and not lamellar (as are in mammals) (Aikawa et al., 1969; Ferguson and Dubremetz, 2014; Muhleip et al., 2021). The proper formation of hexamers of the F1F0 ATP synthase residing in the cristae membrane of the *T. gondii* mitochondria was shown to be critical for the maintenance of the native cristae morphology (Muhleip et al., 2021). However, tubular cristae morphology is conserved in gregarines whose expression of the ATP synthase genes is not detectable. This suggests that alternative cristae formation mechanisms could operate in Apicomplexa (Mathur et al., 2021).

The localization and morphology of the organelle vary as the cell cycle progresses (Melo et al., 2000; Dooren et al., 2006; Ovciarikova et al., 2017). The non-dividing parasite mitochondrion is generally elongated or branched (Melo et al., 2000; Dooren et al., 2006; Ovciarikova et al., 2017). However, as cell division onsets, mitochondria undergo marked morphological changes to accommodate the specific requirements of distinct cell division mechanisms used by each species. For instance, the *Plasmodium* mitochondrion replicates its DNA coincidentally with nuclear DNA replication (Preiser et al., 1996). In the process, it becomes highly branched during schizogony; a cell division mechanism which generates tens of emerging daughter cells simultaneously. The mitochondrion divides into multiple organelles late during schizogony, prior to cytokinesis, at the time in which other organelles such as the nucleus are packed (Dooren et al., 2006). Mitochondrial duplication starts at the early stages of daughter formation in *Toxoplasma*'s endodyogeny. The mitochondrion branches at multiple locations, entering the developing daughter cells. In the cell division scheme used by *T. gondii* two daughter cells are formed at a time; this mechanism is known as endodyogeny. Here, again packing of the mitochondria into the new parasites happens at the very end of the division cycle (Nishi et al., 2008; Ovciarikova et al., 2017; Melatti et al., 2019).

In *T. gondii*, mitochondrial segregation has been further elucidated. Two proteins, Fis1, a homolog of Fis1p protein of

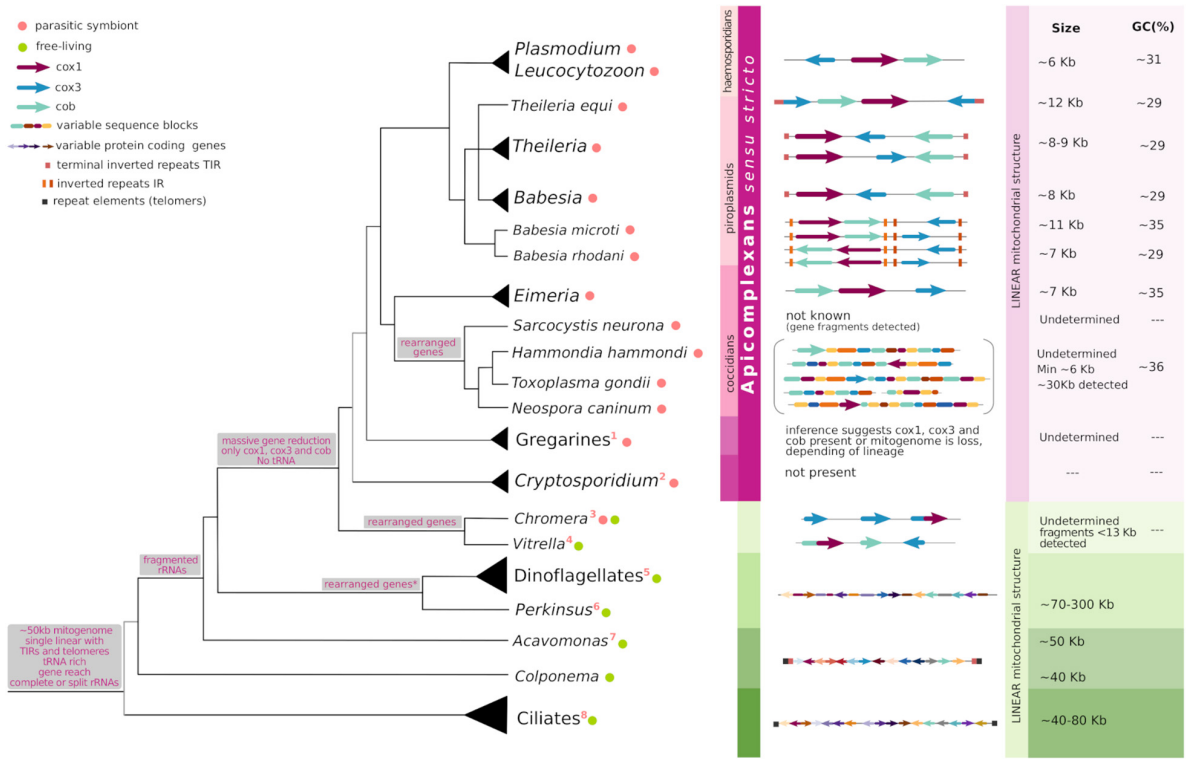


FIGURE 1 | Commonalities and peculiarities of apicomplexan mitogenomes in light of their evolution. The tree shows evolutionary relationships among taxa in Alveolata, the gray branches at the base of Alveolata and Apicomplexa depict less confidence for the topology shown. Gray boxes associated with some tree nodes describe the characters present or lost at the hypothetical common ancestors. To the right of taxon names, colored circles indicate the organism lifestyle following the legend and numbers enumerate peculiar features of each lineage (see below). At the middle, the horizontal figures represent the size, topology and gene content of known mitogenomes. The different elements follow the shape and color depicted in the legend. To the right, when the data is available, approximated mitogenome size and GC content (%) are indicated. This figure results from an adaptation of Figure 7 in Namasivayam et al. (2021), Figure 5 in Hikosaka et al. (2010), Figure 5 in Schreeg et al. (2016), and Figure 1 in Janouškovec et al. (2019). 1: mitogenome is lost in some gregarines; 2: mitogenome loss; 3: *cox1–cox3* fusion, *cob* loss; 4: *cob-cox1* fusion; 5: fragmented *cox3* and *cob-cox3* fusion, extensive RNA editing; 6: *cox3* loss and translational frameshifting in *cox1*; 7: *cox2* encoded for in the mitogenome; 8: bipartite rRNAs, few tRNAs, *cox3* loss, bipartite *nad1*.

yeast, and LMF1, an outer mitochondrial membrane protein, were shown to interact with each other. Fis1 homologs play a role in recruiting fission-mediating dynamins in other systems, such as bacteria. Disruption of LMF1 in *T. gondii* severely impacts mitochondrial morphology, and impairs proper organellar segregation (Jacobs et al., 2020). In all cases, however, little is known about the mechanism of mitochondrial genome duplication and maintenance and its coordination with cell cycle progression.

A remarkable exception regarding the biology of mitochondria among Apicomplexa is *Cryptosporidium*. While all others in the phylum retain the mitochondria, cryptosporidians possess a double membrane remnant organelle which has lost its metabolic capability, generally referred to as the mitosome (Putignani et al., 2004). In contrast to other mitochondrial reductions found in nature, as for example hydrogenosomes, the mitosome represents an additional level of reduction as it does not retain any genetic material (Tachezy, 2008; Hjort et al., 2010). The mitosome evolutionary history can only be traced thanks to the localization of resident mitochondrial proteins, such as the mitochondrial chaperonin 60 (Cpn60) (Riordan et al., 2003) and mitochondrial heat shock protein 70 (HSP70) (Šlapeta and Keithly, 2004) to the organelle. The loss of the mitochondrion is accompanied by the loss of the plastid remnant and a reduction of the nuclear genome. It is thought that the abundance of host-derived nutrients combined with anaerobic/hypoxic growth conditions support the massive reduction in the mitochondrial metabolic capability in cryptosporidians (Gagat et al., 2017). Two additional independent mitochondrial functional reductions and electron chain losses within Apicomplexa have been recently shown to have occurred amongst gregarines (Mathur et al., 2021; Salomaki et al., 2021). While these mitochondrial reductions might be related to the hypoxic conditions found within the digestive tract of their respective hosts, further studies are needed to proper characterize these organelles.

API-MITOCHONDRIAL GENOME SIZES AND TOPOLOGICAL ORGANIZATION

A common theme to Myzozoa, including chromerids, dinoflagellates, perkinsids, and Apicomplexa *sensu stricto*,

is the extreme reductive evolution of their mitogenomes. A hallmark of these genomes is that they are stripped down to the bare minimum, retaining only the coding information required to continue existing as independent entities. Most myzozoans retain at most three protein coding genes; apocytochrome b (*cob*) and two subunits of cytochrome c oxidase (*cox1* and *cox3*). Mitogenomes also include fragmented genes for ribosomal RNAs of the Large and Small subunits (LSU and SSU fragments, respectively), and other fragments of unknown function. No mitochondrial tRNA genes are present and, as it is understood, their loss is compensated by functional replacement with cytosol-imported tRNAs (Esseiva et al., 2004; Pino et al., 2010). In stark contrast, sister lineages to myzozoans, such as Acavonomidia and Colponemidia, include over 45 protein coding genes in their mitochondrial genomes. Of these, nine were transferred to the nucleus in Apicomplexa (including the cytochrome c oxidase 2 gene), while 30 were presumably lost. One hypothesis for the selective retention of the three protein-coding genes in the myzozoan mitogenomes argues that given the key roles played by these proteins in electron transport and energy coupling, their expression must be quickly and directly regulated by the redox state of the mitochondrion (Allen, 1993; Adams and Palmer, 2003; Martin, 2003). Additionally, GC content and protein hydrophobicity have been proposed as factors favoring mitochondrial retention (Johnston and Williams, 2016). In fact, the hydrophobicity-related hypothesis was the first to be proposed based on the rationale that hydrophobic segments present in mitochondrial proteins could potentially target them for transport to the endoplasmic reticulum, thus preventing import across the mitochondrial membrane (von Heijne, 1986). Experimental evidence and additional observations have contributed to further support this mitochondrial-targeting hypothesis (Björkholm et al., 2015; Muhleip et al., 2021). Concordantly, the three retained genes, which are almost universal among currently examined mitochondrial genomes, exhibit highly hydrophobic profiles. Importantly, three other respiratory genes, *nad1*, *nad4,* and *nad5*, are present in all mitogenomes except for a number of alveolates, including all apicomplexans, and some yeasts. In fact, what is missing is the entire NADH dehydrogenase complex (composed of over 20 subunits) known as Complex I of the electron transfer chain (Adams and Palmer, 2003; Vaidya and Mather, 2009; Flegontov et al., 2015). In such a context, organisms rely on alternative NADH dehydrogenases (NDH2) to transfer electrons to ubiquinone (Biagini et al., 2006; Flegontov et al., 2015; Hayward and van Dooren, 2019). In absence of Complex I, Complex III is the first proton pumping complex in the apicomplexan mETC.

Dinoflagellates have increased their mitogenome size by accumulating a surprisingly high (99%) percentage of non-coding and repetitive sequence elements, reaching mitogenome sizes of up to 300 kb. It has been proposed that these sequence additions might be functional and positively selected in free-living organisms subject to more stringent metabolic conditions than those which have remained exclusively parasitic and under a scenery of relaxed selection (Gagat et al., 2017). Comparatively, mitogenomes in Apicomplexa are generally small, with the largest one characterized to date being of less than 30 kb, and the majority being smaller than 12 kb (**Figure 1**). The mitochondrial genomes of haemosporidians (*Plasmodium* and *Leucocytozoon*), piroplasmids (*Babesia, Theileria* and archaeopiroplasmids) together with Eimeriidae coccidians (*Eimeria, C. cayetanensis,* and *Isospora),* are amongst the smallest characterized in nature, ranging from 5 to 12 kb in size (Hikosaka et al., 2010; Cinar et al., 2015; Hafeez and Barta, 2019). Nonetheless, no correlation *per se* has been established between the mitochondrial genome size and its gene content. This is clearly illustrated by the fact that ciliates, whose mitogenomes encode for tens of proteins usually range from 50 to 70 kb in size while dinoflagellates bear mitochondrial genomes of about 300 kb, but conserve only three protein coding genes (Burger et al., 2003; Gagat et al., 2017; Park et al., 2019). Variations in genome sizes are mostly caused by repetitive elements, and differences in the length and organization of intergenic regions.

One of the best characterized mitochondrial genomes in the phylum Apicomplexa belongs to *Plasmodium.* As for many other aspects of their cell biology, our disproportionate insight into the -omics of *Plasmodium* species is likely a consequence of their tremendous impact to human and animal health. Malaria remains one of the most deadly infections affecting millions of people worldwide, yearly. This interest is exacerbated for the *Plasmodium* mitogenome, as the organelle has been identified as the target of the antimalarial drug atovaquone, and of other drugs showing antimalarial effects *in vitro* (Vaidya et al., 1993). The first molecular insights into the mitogenome in *Plasmodium* inadvertently emerged in the late 80's when "the 6 kb element" was first identified in the murine and human infecting species *P. yoelii* and *P. falciparum* (Vaidya and Arasu, 1987; Vaidya et al., 1989), respectively.

The "6 kb element" initially attracted interest due to its repetitive nature (Suplick et al., 1988). However, the presence of an additional extrachromosomal DNA element -which we know now corresponds to the remnant plastid genome- confounded its conclusive identification (Feagin, 1994). Definitive proof came from sequencing experiments which demonstrated that this element coded for three classical mitochondrial proteins (Suplick et al., 1990; Feagin, 1992). Early sequencing experiments of the 6 kb element in different *Plasmodium* species showed that both the sequence and the physical arrangement of the mitogenome are highly conserved. The fast paced advances in sequencing technologies exponentially increased the number of sequences available: by 2011, complete or near-complete mitochondrial genome sequences were already available for 25 species of *Plasmodium* (Hikosaka et al., 2011b), consolidating the notion that the organizational principle of the 6 kb element is universal to the mitogenomes of *Plasmodium* species. The 6 kb element was estimated at the time to exist in about 20 copies per cell in *P. falciparum*. This coincides with the copy number reported for *P. gallinaceum* (15–20/cell) (Joseph et al., 1989), but is markedly lower than the number of copies reported for *P. yoelii* (150/cell) (Vaidya and Arasu, 1987). The biological significance of these differences remains unclear.

The topology of the mitogenome in *Plasmodium* was initially proposed to be a combination of circular and linear molecules. Still, this combination is actively reported in the literature.

However, sequencing and biochemical assays corroborated the identity, topology and size of *Plasmodium* mitogenome. Hybridization experiments elegantly demonstrated that the 6 kb element did not exist as a linear monomer, but rather exists within tandem arrays of varying sizes. These tandems range from 6 to 30 kb, with a marked preponderance of dimers of 12 kb (Joseph et al., 1989). Preiser and colleagues showed in 1996, through fine analyses of the mitogenome configuration of *P. falciparum,* using 2D gels, that circular topologies observed by electron microscopy of isolated mitochondrial DNA constitute a minor fraction of the molecules. These data reinforced the notion that the vast majority of DNA molecules of the mitogenome exist in linear configuration (Preiser et al., 1996).

Mitochondrial genomes have been looked into in a number of other haemosporidians belonging to three genera distinct from *Plasmodium*: *Leucocytozoon*, *Haemoproteus* and *Hepatocystis*. The mitogenomes of *L. caulleryi*, *L. fringillinarum,* and *L. majoris* share both size and organization with *Plasmodium*. Similarly, analyses of mitogenomes of seven *Haemoproteus* spp., two *Hepatocystis* spp., and 24 *Parahemoproteus* spp. support the notion that the mitogenome conservation extends throughout the order Haemosporida (Seeber et al., 2014; Pacheco et al., 2018). However, contradicting reports exist when it comes to the mitogenome sequences of bat parasites of the *Nycteria* genus; also haemosporidians. Karadjian et al. (2016) showed that in these species mitogenomes exhibit two distinct organizations; while *N. gabonensis* and *N. heischi,* display the typical haemosporidian sequence and organization found in *Plasmodium*, *N. medusiformis,* and *Nycteria* sp. (isolated as mixed infection), differ in that the *cob*, *cox1*, and *cox3* genes appear in a different order and transcriptional direction. In addition, the pattern of fragmentations of the large- and small-subunit rRNA genes is distinct. Though no evidence of such genome rearrangements were found for other species in over 100 mitogenome analyses, these results suggest that subtle differences may exist. Though the actual diversity in organization remains slightly controversial in Haemosporida, data disproportionally supports a highly conserved topology, and in all cases it is clear that all mitogenomes abide by the linear "6 kb element" monomeric subunit rule.

The mitogenomes of several Babesia and Theileria are composed of linear molecules characterized by the presence of large terminal inverted repeats on both ends (Hikosaka et al., 2010). However, differences in topology and organization exist between species and within lineages. The first sequenced mitochondrial genomes among the piroplasmids showed conservation in the structure and size within the order.

Within the genus *Babesia* the mitogenome structure is highly conserved. The mitogenomes of *B. bovis, B. bigemina, B. caballi, B. orientalis,* and *B. gibsoni* are made up of monomeric linear molecules of 6.6 kb characterized by the presence of flanking terminal inverted repeats (TIRs) of 440 bp (Hikosaka et al., 2010; He et al., 2014). Recently, mitogenomes for *B. canis*, *B. rossi*, *B. vogeli*, *B. gibsoni*, *Babesia sp. Coco*, and *Cytauxzoon felis* were characterized, revealing that they share the organization reported in *Babesia sensu stricto* and *Theileria* species (Schreeg et al., 2016; Guo et al., 2019). Nonetheless, differences are also identifiable; the presence of the repeated terminal TIRs and a *cox1* segment for *B. vogeli* and *Babesia sp. Coco* could not be confirmed. Likewise, the mitochondrial genome of *B. conradae* presents a novel arrangement of its genes, including the loss of *cox3* and a duplicated inversion that includes the 3′ end of *cox1* and ribosomal gene fragments (Schreeg et al., 2016).

Whole genome sequencing of *B. microti* R1 and Gray strains supported a previous notion based on 18S rRNA sequence, that it defines a basal branching new lineage within Piroplasmida, different from *Babesia* and *Theileria* (Cornillot et al., 2012). Genome mapping of these *B. microti* strains revealed that the mitochondrial genome is made up of four distinct linear molecule configurations in this species. Similar findings were reported for *B. microti* Munich strain, and for *B. rodhaini* (Hikosaka et al., 2012). These archeapiroplasmid mitogenomes present two pairs of novel inverted repeats -named IRa and IRb- and a flip-flop inversion in between. Combination of distinct pairs generates four distinct mitogenome structures present at an equi-molar ratio (Hikosaka et al., 2012). All possible configurations are similarly organized in linear molecules, however, mitogenome sizes differ amongst archeapiroplasmids; *B. microti* mitogenome is 11.1 kb while that of *B. rodhaini* is 6.9 kb. For both species, it was shown that multiple forms of the mitogenome coexist within a mitochondrion, displaying a total of about 20 copies per cell (Hikosaka et al., 2012).

While a number of *Theileria* species, such as *T. parva* and *T. annulata,* display mitogenomes very similar to those found in *Babesia* (Kairo et al., 1994; Hikosaka et al., 2010; Cornillot et al., 2013), sequence variations are more frequently observed. For instance, *T. orientalis* has a 3 kb central inversion including *cox3* and LSU fragments flanked by distinctive insertions (Hikosaka et al., 2010). *T. ducani* presents inverted repeats of only 48 bp at both ends (Virji et al., 2019). *T. equi* presents the most divergent mitochondrial genome of the theilerids displaying evidence of gene duplication and rearrangements, larger and more divergent TIR sequences which include *cox3* and rRNA gene fragments, resulting in a longer mitogenome molecule of 9 kb in size (Hikosaka et al., 2010; Kappmeyer et al., 2012). Similarly to *B. microti*, the phylogenetic position of *T. equi* is not clear and, while these features are consistent with an early branching of these species within Piroplasmida, a better understanding of the evolutionary patterns within this group requires further and better understanding of the group's systematics.

Regarding Eimeriidae, varying size mitochondrial DNA fragments of 4–20 kb were observed by Southern blots when probing for the mitochondrial *cox3* gene of *E. tenella* (Hikosaka et al., 2011a). Consistently, the *E. tenella* mitogenome was shown to be arranged in tandemly repeated linear 6.2 kb elements (Hikosaka et al., 2011a). This structure and the mitogenome organization was later shown to be highly conserved amongst species of *Eimeria*. In fact, the sequencing of *E. mitis* (Liu et al., 2012), *E. magna* (Tian et al., 2015) and seven other species responsible for coccidiosis in turkey (Ogedengbe et al., 2014; Hafeez and Barta, 2019), revealed that mitogenomes are more stringently conserved among these species (displaying 0–0.5% variation) than their nuclear genomes are (1.6–3.2%)

(Ogedengbe et al., 2014), suggesting they are under stricter constrain to change.

Cyclospora cayetanensis is another human pathogen of emerging importance within Eimeriidae. Its mitogenome has been shown to be formed by concatemers of a monomeric 6.3 kb subunit (Cinar et al., 2015; Ogedengbe et al., 2015; Tang et al., 2015). Parasites of the genus *Isospora,* composed of mostly avian and reptile parasites, display an akin mitogenome organization to that of *C. cayetanensis* and others within Eimeriidae. The mitogenomes of *I. serinuse, I. manorinae, I. amphiboluri* and an unnamed *Isospora* sp. have been shown to be of about 6 kb in size. The latter two were reported to be circular-mapping but their physical structure was not determined; circular mapping molecules plausibly suggest a topology of linear concatemers. Gene arrangement and directions are identical to that of other mitogenomes from *Eimeria* spp. and others in the family (Yang et al., 2017; Hafeez and Barta, 2019).

In contrast to the abundance of detailed information available for *Plasmodium* and its close relatives, and the relatively good understanding of the mitogenomes of monoxenic coccidans, cyst forming coccidians are still in the works. Early attempts to assemble these mitogenomes were presumably hampered due to the presence of fragmented copies of mitochondrial genes, especially *cob* and *cox1*, in the nuclear genome. This, has been argued, precluded the definitive identification of *bona fide* mitochondrial sequences. Using PacBio and ONT sequencing technologies, we and others recently attempted to assemble the mitochondrial genomes of both *T. gondii* and *N. caninum*. Contigs belonging to the mitochondrion could be readily identified by their markedly lower GC content (approximately 36%). Unexpectedly, however, neither we nor others embarking in similar efforts could assemble the tens of different mitochondrial contigs and reads into a single mitogenome. Contigs corresponding to the mitogenome vary widely in size. We found the mitogenome of *N. caninum* to be distributed in 20 distinct contigs, ranging from 1.4 to 86 kb with a mean of 12 kb in size (Berná et al., 2021). Twenty nine contigs ranging from 1.1 to 39 kb in size, with a median of 8.0 kb, were identified as the mitochondrial genome of *T. gondii* (Berná et al., 2021).

Consistently, Namasivayam et al. (2021) reported ONT reads of variable size ranging from 320 to 23,619 bp for *T. gondii*. This study carefully analyzed individual sequencing reads, revealing that though the composition of each read is not equal, nor is it made up of linear arrangements of a unique monomeric subunit, discrete common short sequence elements can be identified. Twenty one of these elements, referred to as "sequence blocks" named after the alphabet (A to V), appear in a plethora of arrangements in molecules of varying total size. These variable combinations give rise to an extraordinarily complex genome sequence. However complex, these authors discovered that the sequence blocks do not assemble randomly (Namasivayam et al., 2021). Rather, blocks appear in a finite number of combinations. As an illustrative example, *cox1* fragments are only present in sequence blocks V, S, C, and Q. The correct order combination of these results in a nearly full length *cox1*. Astoundingly, the reads representing the *T. gondii* mitogenome identified by Namasivayam et al. (2021), and the contigs assembled by us, differ significantly (though sequence blocks are clearly identifiable in both), suggesting the saturating sampling of possible combinations has not been reached even by the sum of our sequencing efforts. The mitogenome of *H. hammondi* seems to abide by the same principles of complexity both in size and organization as those of *T. gondii* and *N. caninum,* suggesting this composition complexity is a common feature of cyst-forming coccidia mitogenomes (Namasivayam et al., 2021).

A partially assembled genome of *S. neurona* has been produced (Blazejewski et al., 2015), showing an unusually large nuclear genome size, twice as large as other coccidian genomes, and with a repeat content approximately five times that of *T. gondii*. A definitive mitogenome has not been identified. However, analyses of publicly available *Sarcocystis* sequence reads (Blazejewski et al., 2015) by us (unpublished) and Namasivayam et al. (2021), clearly identify variable size contigs, ranging from 500 to 1,300 bp, whereby mitochondrial gene fragments can be identified. This suggests that the complexity of the mitogenome in these species might be akin to that of other cyst forming coccidians.

Though recent studies using third generation sequencing have allowed unequivocal identification of *bona fide* mitogenome sequences, overcoming the limitations of earlier technologies, the definitive size and topology of the elusive mitogenomes of cyst forming coccidians continues to be an enigma. Nonetheless, it should be noted that generally speaking, mitogenome sequencing can be influenced by low sequencing depths caused by low abundance of starting material (derived from difficult to culture parasites bearing a single copy of the organelle). In addition, protocols to obtain pure mitochondrial fractions from many species are lacking, and have only recently been optimized and used for organelle proteome profiling in a few species (Evers et al., 2021; Maclean et al., 2021). These fractions, however, are usually contaminated with plastid DNA. These limitations, combined with the use of analysis strategies based on mapping short-read sequences to highly repetitive DNAs, pose specific technological limitations which may influence the outcomes, either generating *de novo* artifactual assemblies, or missing subtle differences by inadvertently forcing synteny.

API-MITOCHONDRIAL GENOME TRANSCRIPTION AND TRANSLATION

Mitochondrial components are generally encoded by the nucleus while the organelle genome encodes a small number of proteins. The mitogenome encoded mRNAs are transcribed and translated by a distinctive mitochondrial protein-synthesizing system, some of whose components are specified by the mitochondrial genome. Characteristically, mitogenomes encode for rRNAs, while tRNAs and ribosomal proteins can be encoded for in the nucleus (Lang et al., 1999). The mitogenomes of Apicomplexa are characterized by their extreme coding reductionism, harboring only highly fragmented rRNAs, and *cox1*, *cox3,* and *cob*

(Mathur et al., 2021). Protein coding genes are also highly fragmented in cyst-forming coccidians. In this section we succinctly (and certainly not exhaustively) describe a number of known molecular mechanisms underlying transcription and translation of apicomplexan mitogenomes genes.

We note that the complexity and breadth of nuclearly encoded proteins that significantly contribute to mitochondrial function -including protein import, organelle fission, other metabolic pathways, etc.- is by no means sufficiently reviewed in this section. Interested readership in the subject can find excellent recent reviews on the subject which are beyond the scope of the present work. Here, we highlight feature and mechanisms occurring in apicomplexans and other alveolates to showcase the diversity of processes that have been selected through thier evolution (**Figure 1**), and propose plausible mechanisms underlying the apicomplexan peculiarities.

Mitochondrial Electron Transport Chain Complexes and Mitoribosomes

Recent "complexome" profiling has shown that the respiratory chain complex IV in *T. gondii*, which in humans and yeast houses *cox1* and *cox3*, contains peptides of these proteins, and 17 other divergent Apicomplexa-specific proteins (Maclean et al., 2021). Likewise, complexome profiling in *Plasmodium* revealed both divergent and clade-specific additions to all respiratory chain complexes, and expression of Cox1 and Cox3 peptides (Evers et al., 2021).

While the majority of mitochondrion-resident proteins are encoded for in the nucleus and imported, genes encoded for in the mitogenome are necessarily translated by mitochondrial ribosomes. Mitochondrial translation has been shown essential for survival in *P. falciparum* (Ke et al., 2018). Cross-species comparisons amongst Apicomplexa, alga and diatoms revealed that organellar ribosomes (i.e. of the mitochondrion and the apicoplast) have species-specific composition. In addition, they have undergone reduction, losing several small and large subunit proteins, and divergence both in sequence and ortholog length, with respect to bacterial ribosomes (Ke et al., 2018). In fact, several subunits of conserved mitochondrial ribosomes, which are present in diatoms, are missing in Apicomplexa. It has been proposed that these reductions in the mitochondria could be partially rescued by dually targeted proteins of plastid origin to both the mitochondrion and the apicoplast.

Noteworthy, the mitoribosomes of *T. gondii* and *P. falciparum* were recently identified and characterized by genetic ablation of genes encoding for putative protein components (Lacombe et al., 2019). In *T. gondii* it was shown that three distinct proteins, encoded by the nucleus, belong to two distinct macromolecular complexes, possibly corresponding to the small and large subunits of the ribosome. Meanwhile, in *Plasmodium*, immuno-EM experiments showed that three putative protein components of mitoribosomes are closely associated with the cristae. Knockdown of these proteins hypersensitized parasites to inhibitors targeting Complex III and blocked the pyrimidine biosynthesis pathway (Ling et al., 2020).

tRNAs

In addition to the presence of a mitochondrial ribosome, translation of mito-encoded proteins implies the need for an organelle resident protein translation system, which includes tRNAs. Characteristically, apicomplexan mitogenomes do not encode for tRNAs. In fact, this is a character common to all alveolates, which we discuss in further detail below. Northern blot experiments probing for nuclear-encoded tRNAs on purified mitochondrial fractions, showed that in *T. gondii*, tRNAs are indeed imported. All but the cytosol-specific initiator tRNAMet were shown to be present both in the cytosol and in the mitochondria. Absence of formylated tRNAMet suggests that this amino acid modification is not required for translation initiation in the *T. gondii* mitochondrion (Esseiva et al., 2004; Pino et al., 2010). In addition to importing tRNAs, *P. falciparum* has been shown to be equipped with a mitochondrion-dedicated aminoacyl-tRNA synthetase (aaRS), known as mFRS, responsible for charging tRNAs imported into the mitochondria with phenylalanine on site. Intriguingly, mFRS is exclusive to the malaria parasites within the phylum (Sharma and Sharma, 2015). In contrast, no mitochondrial aaRS has been identified in *T. gondii*, suggesting that all tRNAs are imported in their charged form (Pino et al., 2010).

rRNAs

rRNA fragmentation is widespread in nature, and by no means an exclusive feature of the Apicomplexa mitochondrial genomes. In Alveolata, extensive rRNA fragmentation might have appeared before divergence of colponemids and myzozoans (see below for a more detailed discussion of rRNA fragmentation in alveolates). However, the fragmentation observed in *Plasmodium* is the most extreme reported, and we therefore highlight it as an illustrative -extreme- example of rRNA fragmentation in Apicomplexa.

In *Plasmodium* out of a total of 34 identified small RNAs, 20 were initially assigned to both rRNA genes (Feagin et al., 1997); seven more were described later. These fragments range from 23 to 200 nucleotides in size whereby 12 correspond to SSU RNAs (totaling 804 nt) and 15 to LSU RNAs (totaling 1,233 nt) (Feagin et al., 2012). Fragments are dispersed throughout the mitogenome, and are coded for in either strand. It is unclear how these small RNA fragments come together to form a ribosome (Hillebrand et al., 2018). However, it has been proposed, based on 3D structure simulations, that fragments could cluster, by secondary-pairing interactions, on the interfaces of the two ribosomal subunits as separate entities to form a functional ribosome (Feagin et al., 2012). In addition, clustered organellar short RNA fragments (cosRNAs) complementary to the 5′ ends of rRNA fragments have been identified. These cosRNAs could be a recognizable footprint by RNA-binding proteins (RBPs) which might aid in the assembling of the rRNA fragments into a functional ribosome (Hillebrand et al., 2018).

The size, nucleotide sequence and arrangements are almost completely conserved among *Plasmodium* and *Leucocytozoon*. Despite complete apicomplexan mitogenomes, not all sequences presumed necessary for functional rRNAs have been identified (Feagin et al., 1997). Such absences might be accounted for by

functional relocation to the nucleus, and targeting of small rRNA fragments from the cytoplasm. As a precedent, mitochondrial import of cytosolic 5S rRNA has been demonstrated in mammals (Entelis et al., 2001).

rRNA fragmentation has also been well described in dinoflagellates (Jackson et al., 2007; Kamikawa et al., 2007; Nash et al., 2007; Slamovits et al., 2007). Common with apicomplexans, these small rRNAs are oligo-adenylated, and presumably assemble into complete ribosomes carrying these short extensions (Feagin et al., 1997; Jackson et al., 2007).

Cox1, Cox3, and Cob

The precise mechanisms of translation of resident mitochondrial protein are still to be teased out in Apicomplexa. As mentioned above, this is likely carried out by a ribosome resident to the organelle, aided by fragments of rRNA possibly brought together by specialized nuclearly encoded RBPs, and a combination of charged imported cytosolic and *in situ* charged tRNAs. Transcription of the mitogenome is mediated by a resident mitochondrial RNA polymerase, which most likely is also involved in mitogenome maintenance (Suplick et al., 1990; Ke et al., 2012). Transcripts are generated as polycistronic units, followed by cleavage and individual transcript polyadenylation (Rehkopf et al., 2000) to yield the shorter RNAs which correspond to the protein-coding and rRNA genes (Ji et al., 1996). In *Plasmodium*, for example, genome length transcripts have been detected from both DNA strands (Suplick et al., 1990).

*Cox*1 is the best conserved of all three protein coding genes amongst Apicomplexa (Feagin, 1994). The *Plasmodium* gene codes for the shortest *cox1* known, lacking the last 45 amino acids of the mammalian *cox1*. *Cob* is also highly conserved amongst species (Feagin, 1994) and, together with *cox1,* they have been used as molecular markers for phylogenetic analysis at different taxonomic levels. On the contrary, *cox3* is not well conserved (Feagin, 1994) and it is not present in all alveolates -it has been lost in ciliates (Burger et al., 2000). Nucleotide sequence identity has been shown for Hematozoa (He et al., 2014). Protein identity values for these three genes in different taxa representative of Alveolata are shown in **Figure 2**. It consistently shows that *cox1* is the most conserved gene, followed by *cob*. On the other hand, *cox3* shows overall low identity values. In addition, its coding region has not yet been identified for some taxa (in which they are expectedly present). For these proteins the minimum percentage of identity within Apicomplexa *sensu stricto* is 65, 59, and 42 (for Cox1, Cob, and Cox3, respectively), and these minimum percentages fall to 53, 35, and 42 within Alveolata. Of interest, **Figure 2** also depicts the extension of the pairwise local alignments. It can be seen that, for Cob and Cox1, the conservation extends over almost the entire protein length and it is not limited to specific domains. Contrarily, for *cox3*, conservation is markedly reduced and in some cases, homology can be established only for specific fragments or domains (less than 20% of the length).

Initiation codons are still uncertain in *cox1* and *cox3*. Only the *cob* in *Leucocytozoon*, *H.* (*Haemoproteus*), and *Plasmodium* has an ATG triplet at the 5′ terminus that could be its initiation codon. The presence of alternative initiation codons has been determined for the *cob* genes in *H.* (*Parahaemoproteus*) and *Hepatocystis*. In these, ATT (Ile) or ACT (Thr) and AGT (Ser) substitute the canonical ATG, respectively (Pacheco et al., 2018). Similarly, the presence of canonical stop codons is seemingly reduced in apicomplexan and alveolate mitogenomes, with frequent loss of the TAA stop codon. A common occurrence is the rescue system using oligo-adenylation following a U nucleotide which creates an in-frame standard stop codon (TAA). This has been shown to happen in the *cox3* transcripts of all dinoflagellates analyzed (Waller and Jackson, 2009).

The redefinition of ORF boundaries, with apparent abandonment of standard start and stop codons is presumably observed in all alveolates (Edqvist et al., 2000; Waller and Jackson, 2009). In absence of RNA editing mechanisms, novel mechanisms are expected to either allow alternative start codons or the recognition of tRNA Met of non-standard initiation codons as Met codons. In either case, an unusual degree of flexibility during the initiation phase of translation is expected. It is interesting that, at least in ciliates, where tRNA Met is encoded in the mitogenome, it shows a truncated D stem and loop and unusual structural features in the anticodon loop, which might allow a more flexible codon-anticodon pairing to accommodate initiation codons that are variants of AUG (discussed in Edqvist et al., 2000). Conspicuously, as mentioned above, cytosol-specific initiator tRNAMet is undetectable in the *T. gondii* mitochondria.

When stop codons are absent or cannot be formed by oligo-adenylation, active mechanisms are needed to avoid translation arrest when the ribosome arrives at the undefined transcript termini. One such mechanism in eukaryotes involves the protein Ski7p and an equivalent system in prokaryotes employs a transfer messenger RNA. It is conceivable that alveolates have adopted components of either of these two rescue systems for regular translation termination in the mitochondrion, in order to avoid transcripts and proteins being targeted for degradation (Waller and Jackson, 2009).

Out of all apicomplexan mitogenomes, the cyst forming coccidians pose the most challenging scenario to understanding transcriptional regulation. RT-PCR experiments in *T. gondii* by Namasivayam et al. (2021) detect nearly full-length *cob* and *cox3,* and a partial *cox1*. Whether these are exclusively generated from the low abundance full length coding sequences detected by ONT (Namasivayam et al., 2021), or are alternatively or additionally generated by an underlying mechanism of recombination of the highly fragmented ORFs coding for Cox1, Cox3, and Cob, remains to be deciphered. An interesting finding of Namasivayam et al. (2021), is that many ORFs were shown to be followed by variant polyadenylation signals which may affect the detection of full length transcripts by Illumina based RNA-seq as sequencing libraries are typically generated following oligo(dT) purification (Namasivayam et al., 2021).

Long reads have been widely used to resolve or enhance nuclear genomes of various parasites. Since they are orders of magnitude smaller and simpler, complete mitochondrial genomes have surged from these same sequencing projects. However, there are notable exceptions; the mitochondrial genome of coccidians being a beautifully illustrative example. From PacBio and ONT sequencing the number of nuclear

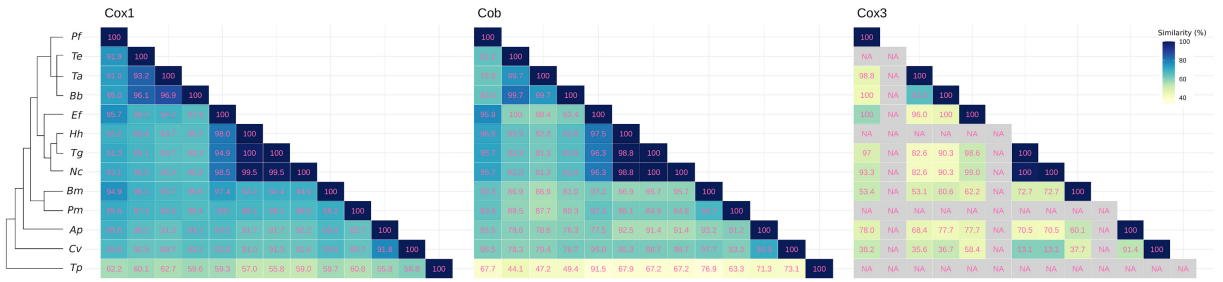


FIGURE 2 | Protein sequence conservation of the three retained mitochondrial genes within Apicomplexa and evolutionarily related organisms. Heatmaps show similarity in local blastp pairwise protein alignments. Percentages of positive-scoring matches for Cox1, Cob, and Cox3, are shown according to the color scale, as indicated. Percentages of alignment coverage (alignment length normalized by the average protein length) are indicated as numbers. Abbreviated labels and NCBI accession identifiers for protein sequences used for Cox1, Cob, and Cox3, are as follows: Pf: *P. falciparum;* YP_009480320.1, YP_009480321.1, AAC06268.1; Te: *T. equi,* XP_025033546.1, XP_025033545.1, NA; Ta: *T. annulata,* XP_039113697.1, XP_039113699.1, XP_039113698.1; Bb: *B. bovis*, YP_001504106.1, YP_001504108.1, YP_001504107.1; Ef: *E. fuonis,* NC_039745.1, NC_039745, YP_009531492.1; Hh: *H. heydorni,* AFQ31671.1, AFQ31670.1, NA; Tg: *T. gondii,* AFQ31675.1, XP_018634734.1, AFQ31675.1; Nc: *N. caninum*, AFQ31673.1, QLF97291.1, QLF97290.1; Bm: *B. minutum,* BAS21362.1, BAS21364.1, BAS21363.1; Pm: *P. marinus*, ADZ98701.1, ADZ98702.1, NA; Ap: *A. peruviana,* AHA41651.1, AHA41650.1, AHA41680.1; Cv: *C. vietnamica*, AHA41634.1, AHA41615.1, AHA41635.1; Tp: *T. pyriformis,* AAA32102.2, AAD41943.1, NA. NA corresponds to missing or unavailable data. To the left, tree topology corresponds to that used for **Figure 1**.

chromosomes of both *T. gondii* and *N. caninum* was reduced from 14 to 13, and important chromosomal rearrangements between both species were described (Berná et al., 2021; Xia et al., 2021). However, none of the mitochondrial genomes could be completely resolved; neither have been conclusively resolved the level at which the diversity exists nor the mechanisms underlying this diversity. For these complex mitogenomes, as can also be the case for *Sarcocystis*, additional strategies must be adopted. In particular, better coverage and more robust results are necessary. Direct RNA long reads emerge as a promising alternative to shed light onto the undescribed diversity and perhaps new transcriptional mechanisms in these parasites.

With this in mind, we explored the recently published direct ONT mRNA data from *T. gondii* (Lee et al., 2021). We searched for clues of the underlying transcription mechanism of gene fragments in the *T. gondii* mitogenome. Although this sequencing was performed to identify possible isoforms in the nuclear genes (Lee et al., 2021), a total of 355 ONT reads (from 112 to 2,962 bp long) mapped in the mitogenome of *T. gondii*. The longest mRNAs correspond almost exactly to the series of discrete sequence blocks described by Namasivayam et al. (2021). We note that this dataset had been analyzed by Namasivayam et al. (2021) prior to publication. While the authors reached an akin conclusion, a markedly lower number of reads mapped to the individual sequence blocks and their combinations. Thus, the significantly higher number of reads mapped here further support the mechanism suggested by Namasivayam; ordered blocks are indeed transcribed in the required order so as to produce a full length mRNA (Namasivayam et al., 2021).

However, the three protein-coding genes encoded for the *T. gondii* mitogenome are formed by a minimum of three sequence blocks. Interestingly, we and Namasivayam and colleagues, found reads supporting the existence of transcripts corresponding to single blocks, or the combination of only two blocks for *cob*. Several of these mRNA reads, as shown in **Figure 3**, do not extend beyond, nor do they include other sequence blocks. Instead, they are truncated in the exact same position, coinciding with the block's boundary. To us, this suggests that an unidentified post-transcriptional mechanism may exist to precisely cleave the individual block sequences from their primary polycistronic mRNA. One could also envision the possibility that these expressed fragments are somehow additionally combined *a posteriori* to form a complete messenger, thus originating a mature and complete mRNA ready for translation. In fact, we propose that the existence of such mechanisms could even be needed to compensate for the low abundance of full-length transcripts originating from a only handful of sequence block combinations coding for the complete ORFs in the mitogenome.

Post-transcriptional Processing in Alveolata

Mitochondrial genomes in *Plasmodium* and piroplasmids seem to produce mRNA transcripts that are more or less ready for translation. In *Toxoplasma* and *Neospora*, one complete gene copy could produce all full length mRNAs and/or transcript fragments might be further processed to obtain full length messenger RNAs. While this is unknown, some clues can be obtained from what has been described in other alveolate lineages. Thus, we summarize some post-transcriptional oddities to highlight the diversity of molecular processes that could be at hand for these protists.

First, apart from the rRNAs fragments, additional genes have been found split. In *Paramecium* and *Tetrahymena*, the *nad1* gene is split in two portions that are encoded for and transcribed from different DNA strands. Edqvist et al. (2000) determined that the corresponding transcripts are not trans-spliced, thus separate N- and C-terminal portions of Nad1 are independently produced in these systems. Whether the separate polypeptides can associate to generate a bipartite, functional Nad1 protein - analogous to what occurs with the rRNA fragments- or some

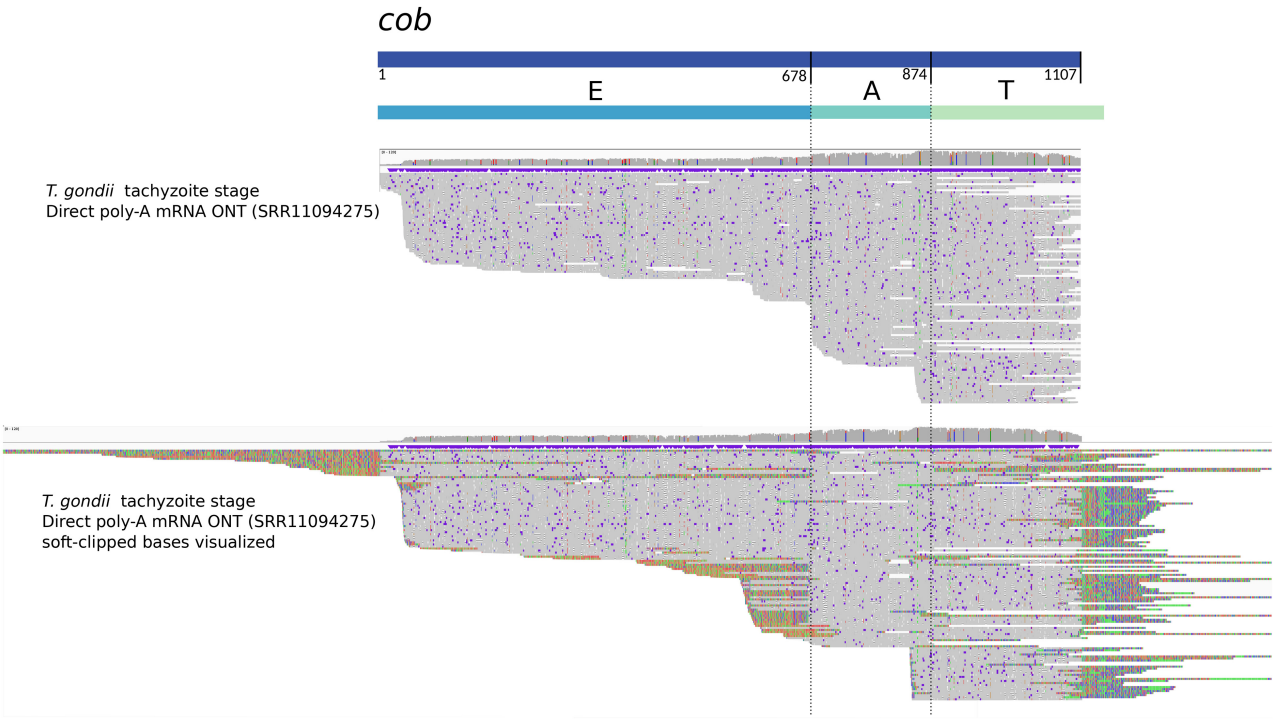


FIGURE 3 | Mapping of Oxford Nanopore mRNA reads with respect to the *cob* sequence in the mitogenome of *T. gondii*. The figure shows an adapted visualization obtained with IGV (Thorvaldsdóttir et al., 2013) of ONT direct RNA reads (Lee et al., 2021) aligned with minimap2 (Li, 2018) to the *cob* sequence inferred for *T. gondii*. Above, the rectangles represent the *cob* gene sequence and the proposed constituent sequence blocks "EAT" (Namasivayam et al., 2021), respectively. Vertical lines highlight the limits between E and A, and A and T blocks. In the alignment section, the upper part shows ONT reads aligned with hard clipping, and the bottom shows the same ONT reads with soft clipping visualization. Notoriously, there are reads covering the complete *cob* gene, but others align only to one (T) or two blocks (AT). In the case of the visualization with soft-clipping, the aligning part of the reads extend to sequence contexts not contiguous but otherwise present in other parts of the mitogenome. The latter correspond to other sequence blocks.

other type of covalent joining occurs at the protein level, remains unclear. In dinoflagellates, *cox3* is split in two fragments, each producing a polyadenylated transcript. To achieve a mature mRNA, the fragments are joined when a certain number (taxon specific), of A nucleotides of the oligo(A) tail from the 5′ major transcript are incorporated (Jackson et al., 2007), implying a novel mechanism for splicing in dinoflagellate mitochondria. The inserted adenosine nucleotides in turn result in one or more lysines in the Cox3 protein, representing an hydrophilic insertion which is tolerated due to its position between the transmembrane domains (Jackson and Waller, 2013). *Trans*-splicing is present in the mitochondrial genome of plants. Autocatalytic molecular events are mediated by sequence flanking the precursor elements, which resemble group II introns (Bonen, 1993; Edqvist et al., 2000). To our knowledge, no such elements have been identified flanking genes in alveolates. On the other hand, in mycorrhizal fungi, there is evidence that the RNA from a fragmented *cox1* gene is joined by flanking exon sequences that form a group I intron structure, potentially in conjunction with the *nad5* intron sequences. The latter acts as an additional helper (Nadimi et al., 2012). Additional third helpers are also used in *Chlamydomonas* (Goldschmidt-Clermont et al., 1991). Though purely speculative, the possibility of small RNAs produced in *trans* being involved in the resolution of trans-splicing events is appealing to propose as a putative mechanism operating in Apicomplexa. These could be coded for by the nucleus and imported into the mitochondrion to aid in post-transcriptional processing of mitochondrial transcripts.

In addition to split of mitogenome genes, multiple mitochondrial gene fusions have been reported in early branching dinoflagellates (*cob-cox3*) and chromerids (*Chromera*: *cox3-cox1*; *Vitrella*: *cob-cox1*) (Slamovits et al., 2007; Oborník and Lukeš, 2015). As proteins are part of different electron transport complexes, they must function separately. It is unknown if the processing mechanism -shared or not in both lineages- involves cleavage of the initial transcript, independent translation from the bi-cistronic transcript or post-translational cleavage.

Third, the most extensive case of translational frameshifting was determined for *cox1* mRNA in *Perkinsus*, a sister taxon to Dinoflagellata. From the DNA sequence it can be inferred that the reading frame must be shifted ten times, at every AGG and CCC codon, to yield a consensus Cox1 protein (Masuda et al., 2010). Whether the mechanism involves ribosomal frameshifts and/or specialized tRNAs, is unknown. No evidence of an akin mechanism operating in Apicomplexa has been reported.

Finally, RNA editing can explain differences between nucleotide sequences at the DNA and mRNA levels. All mitochondrial protein-coding genes and some rRNA fragments

are edited in dinoflagellates, where nine of twelve possible substitutions have been observed and concern up to 6% of the nucleotides in transcripts (Lin et al., 2002; Waller and Jackson, 2009). The mechanisms underlying this unprecedentedly versatile editing are unknown. However, with the sole exception of core dinoflagellates, RNA editing is absent from most alveolates making it unlikely for this to be the adopted mechanism used by cyst forming coccidians to make sense of their fragmented genomes.

HOW PECULIAR ARE MITOGENOMES IN APICOMPLEXA?

The progression of a free-living alpha-proteobacterium into a symbiotic mitochondrion necessarily involves a history of decayment. To fully grasp how peculiarly reductive the apicomplexan mitogenomes are we analyze the general features of mitochondrial genomes in additional Alveolata lineages, providing an expanded comparative evolutionary framework.

Apart from apicomplexans, alveolates comprise the well-studied dinoflagellates and ciliates - and *Perkinsus*. These major lineages have evolved distinct and peculiar characteristics. However, the reconstruction of ancient transitions cannot be fully understood unless deep-branching lineages, which retain ancestral characteristics, are incorporated into the discussion. As such, it becomes relevant to highlight inferences drawn from the recently described *Colponema* and *Acavomonas* alveolate lineages. Lastly, we revisit the mitogenome features of chromerids, a sister lineage to Apicomplexa, consisting of coral-associated photosynthetic protists.

To date, about a dozen mitogenomes have been described for ciliates (Burger et al., 2000; Edqvist et al., 2000; Park et al., 2019). They are linear and relatively large (up to 80 kb), with repeat elements at the ends of the linear DNA (telomeres). They present a variable number of protein coding genes involved in aerobic respiration, an unusually high number of unassigned ORFs -many of them being ciliate-specific, two bipartite rRNAs (LSU and SSU) and a few tRNAs. According to this, they contain more than the standard set of genes encoded by mitochondrial genomes in other organisms but less tRNAs. Interestingly, as mentioned before, ciliates do not encode for *cox3* but harbor additional NAD dehydrogenase subunit genes (*nad7*, *nad9,* and *nad10*). No intron sequences have been reported. In addition to the standard ATG, start codons include: ATA, ATT, GTG, and TTG. The latter makes it challenging to assign initiation codons, raising questions about the underlying mechanisms of gene expression. Only TAA is used to terminate translation, while TGA, as in many other mitochondrial translation systems, specifies tryptophan. To note, ciliate mitogenomes harbor, in addition to the rRNA genes, a *nad1* gene which is split and rearranged, whereby N-terminal and C-terminal portions are transcribed from different strands, and translated separately (Burger et al., 2000; Edqvist et al., 2000).

Curiously, dinoflagellates exhibit an evolutionary trend toward mitogenome simplicity, all the while increasing their complexity (Gagat et al., 2017). As in apicomplexans, mitogenomes of several dinoflagellates code for *cox1*, *cox3,* and *cob*, and display the two fragmented rRNAs. No tRNAs have been identified implying, as in Apicomplexa, complete reliance on imported tRNAs for protein translation. None of the NADH dehydrogenase genes (*nad1*, *nad4,* and *nad5*) have been observed in dinoflagellate suggesting that the loss of complex I has occurred before lineage divergence (Waller and Jackson, 2009; Flegontov and Lukeš, 2012). However, mitogenomes consist of multiple linear chromosomes with sizes ranging 6–10 kb, and even longer. The mitogenome of *Symbiodinium minutum*, a reef-building coral endosymbiont, for example, totals 326 kb; 99% of which is composed of non-coding sequences (Shoguchi et al., 2015). The large size and complexity of the dinoflagellate mitogenome is owed to numerous amplification and recombination events; not only does this impact genome structure, but also its gene content. Similarly to what is observed in coccidians, dinoflagellates display multiple copies of each gene, and gene fragments linked in numerous configurations (Nash et al., 2008). Some of the dinoflagellates gene features are shared with other alveolates; for instance, the usage of unconventional start codons or use of polyadenylation to generate stop codons is an alveolate-wide observation.

Other features, however, are peculiar to the lineage. Specifically, dinoflagellates exhibit RNA editing and *cox3* processing. RNA editing occurs in the transcripts of all protein-coding genes and in some rRNA fragments. Editing is absent in *Oxyrrhis marina*, an early branching dinoflagellate. However, the younger the genus, the greater the versatility and scale of changes that occur through RNA editing (Waller and Jackson, 2009). The *cox3* gene is broken between regions coding for the sixth and seventh transmembrane helices, in all core dinoflagellates. In order to create the mature mRNA, a certain number of adenosine nucleotides are added between both primary transcripts. This in turn results in one or more lysines in the Cox3 protein (Jackson and Waller, 2013). Finally, another peculiarity of early *O. marina*, is the *cob-cox3* gene fusion (Slamovits et al., 2007). The mechanisms to manage this gene fusion and its products are currently unknown.

Perkinsus comprises species that affect many bivalve mollusk host species globally. Current evidence supports that, together with other taxa, they constitute the separate phylum Perkinsozoa, sister to Dinoflagellata (Mangot et al., 2011; Janouškovec et al., 2017). Mitogenomes have been assembled for some species; their size, topology and gene content resemble those described in dinoflagellates (Bogema et al., 2021). However, a few differences are attention worthy. First, the *cox3* gene is undetectable. In this sense, the phylum is more similar to ciliates. Secondly, there exists a peculiar *cox1* gene; the reading frame must be shifted ten times, at every conserved AGG and ACC codon, to yield the consensus Cox1 protein (Masuda et al., 2010; Bogema et al., 2021). The mechanisms needed to couple with this translational frameshifts are unknown but an active and efficient machinery is proposed to be required (Masuda et al., 2010).

Colponema and *Acavomonas* belong to two predatory alveolate phyla that lack secondarily derived morphological characteristics and retain cytoskeletal characteristics of the ancestral aveolate.

Thus these phyla are pivotal to our understanding of alveolate evolution (Janouškovec et al., 2013; Tikhonenkov et al., 2014). Inferences strongly support *Acavomonas* as sister to myzozoans and *Colponema* branching deeper, near the base of alveolates or sister to ciliates (Janouškovec et al., 2013). In these organisms, mitochondria are oval and have tubular cristae. The *Acavomonas* mitogenome is a 50.4 kb single linear chromosome with terminal inverted repeats (TIRs), and 38 bp tandem telomeric repeats at the ends. This linear monomeric organization ending in telomeres resembles features described in ciliates, suggesting this represents the ancestral state in Alveolata. While this organization has changed in dinoflagellates and in most apicomplexans, piroplasmids may retain the ancestral state (Janouškovec et al., 2013).

The *Acavomonas* mitogenome has retained the largest gene set among all alveolates, including 14 genes previously unknown in the group. While ciliates are somewhat tRNA gene poor, *Acavomonas* has retained 21 mitochondrial-encoded tRNAs, and requires import of only three tRNA species. In this sense, it appears that independent loss of tRNA genes with concomitant import of aminoacylated tRNAs from the cytosol and use of alternative NADH dehydrogenase, has driven the main gene content reduction in myzozoans. The ancestral alveolate mitogenome was therefore tRNA rich, and reduction in ciliates and myzozoans occurred independently (Janouškovec et al., 2013). Interestingly, large and small rRNA genes are fragmented into multiple pieces, but the *nad1* gene is complete. The bipartite rRNAs and *nad1* are particularities of ciliates and fragmented genes appeared independently in myzozoa (instead, *nad5* is split into two fragments that are widely separated in the genome). *Acavomonas cox2* gene is intact and encoded for in the mitogenome, thus its split and migration to the nucleus is an event that occurred in the ancestor of myzozoans. Other gene peculiarities in ciliates, such as *cox3* absence, are not present in this lineage, suggesting that ciliates harbor many peculiarities that were not present in the alveolate ancestor.

Given their phylogenetic position within Alveolata, the deciphering of the chromerids *C. velia* and *V. brassicaformis* genomes have been instrumental in shedding light onto some conundrums about the evolution of obligate apicomplexan parasites from the free-living phototrophic algae (Woo et al., 2015). Their mitochondrial genomes have been reported (Flegontov et al., 2015) and, together with the analysis of mitochondrial pathways reconstructed from their nuclear genomes, they point to an unprecedented reduction of the respiratory chain in *C. velia*, an aerobic eukaryote. *Chromera* lacks respiratory complexes I and III, whereas *Vitrella* and apicomplexans are missing only complex I (Flegontov et al., 2015; Oborník and Lukeš, 2015). *Chromera* mitochondrial DNA is composed of short linear molecules and similarity searches identified a single conserved gene *cox1* and a highly divergent *cox3* gene. However, these genes appear multiple times in different mitogenome contexts. What is more, *cox3* seems to be fused with an upstream *cox1* fragment, and a fusion transcript is produced which apparently undergoes a subsequent cleavage. Unexpectedly, *cob* gene was absent from the assembly and original reads. Fragments of rRNA genes were also found. No RNA editing was determined. *Vitrella* mitogenome showed similar characteristics but, regarding its protein-coding gene content, it contains a fused *cob-cox1* gene and a divergent *cox3*.

The common features and oddities of the mitogenomes in Alveolata are summarized in **Figure 1**. Based on the mitochondrial genome features of ciliates (with exclusion of particular oddities) and *Acavomonas*, it can be hypothesized that the common ancestor of alveolates carried a mitogenome of about 50 kb, in a single linear chromosome, ending in TIRs and telomeres. The ancestor genome was protein-coding and -relatively- tRNA gene rich. Canonical start and stop codons were not always present. The two rRNA genes may have been complete, or not, but the direct ancestor of *Acavomonas* likely acquired fragmented rRNA genes. At the common ancestor of Myzozoa, a spectacular genome and gene content reduction occurred, arguably representing the largest and most profound reduction discovered in any aerobic mitochondrion thus far. Only three protein coding genes -*cox1*, *cox3,* and *cob* were retained; all tRNA genes were lost with the concomitant gain of mechanisms for cytosolic tRNAs import. NADH dehydrogenase genes were irrecoverably lost. With certain differences and exceptions (further reductions in *Chromera*, some gregarines and *Cryptosporidium*), gene content has been maintained more or less constant throughout myzozoans. In contrast, the evolutionary history of the group is marked by changes in the mitogenome size, organization and topology which seems to have evolved bimodally, with two opposite possible configurations. The first configuration is represented by the smallest known mitogenomes, those of haemosporidians and piroplasmids. They are small linear monomers, with telomeric ends and can be found forming linear concatemers or alternative conformations. This configuration seems to be more parsimonious as the default state in the common ancestor of Myzozoa thus it might be retained in apicomplexans such as *Plasmodium* and *Theileria*. The second configuration is present in dinoflagellates, *Chromera*, *Vitrella,* and also coccidians, and involves an expanded, fragmented and scrambled state of the mitogenome, possibly due to extensive recombination. While this configuration was already known for dinoflagellates and chromerids, it has only been recently described for *Toxoplasma* and *Neospora* (Berná et al., 2021; Namasivayam et al., 2021). In fact, for coccidians, additional details on the "21 building sequence blocks" are only beginning to emerge. While it is inferred that these seemingly messy mitogenomes have arisen recurrently (at least three times), it is unknown whether they share the same underlying molecular mechanisms of emergence. What is more, whether coccidian sequence blocks are mirrored in the other myzozoans is unknown. Finally, what the selective forces operating in these extraordinary complex mitogenome architectures are, remains a mystery.

In this sense, new long read technologies will be instrumental in shedding light onto the unknown aspects of structure and function of the mitogenomes in Apicomplexa, particularly illuminating our understanding of the possible transcriptional and post-transcriptional mechanisms underlying the survival of apicomplexans, in light of their unconventional mitogenomes and parasitic lifestyles.

AUTHOR CONTRIBUTIONS

LB, NR, and MF contributed to conception and design of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

MF is funded by an installment grant awarded by Banco de Seguros del Estado (BSE). Work at the Institut Pasteur de Montevideo is supported by an Institutional grant by FOCEM - Fondo para la Convergencia Estructural del Mercosur (COF 03/11). LB and MF are PEDECIBA and Sistema Nacional de Investigadores researchers.

ACKNOWLEDGMENTS

We are thankful to Drs. Fernando Alvarez and Martin Graña for helpful advice and insightful discussion during the conception and writing of this manuscript. We apologize to all authors whose work we were unable to cite due to space constraints.

REFERENCES

Adams, K., and Palmer, J. (2003). Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Mol. Phylogenet. Evol.* 29, 380–395. doi: 10.1016/s1055-7903(03)00194-5

Aikawa, M., Hepler, P. K., Huff, C. G., and Sprinz, H. (1966). The feeding mechanism of avian malarial parasites. *J. Cell Biol.* 28, 355–373. doi: 10.1083/jcb.28.2.355

Aikawa, M., Huff, C. G., and Sprinz, H. (1969). Comparative fine structure study of the gametocytes of avian, reptilian, and mammalian malarial parasites. *J. Ultrastruct. Res.* 26, 316–331. doi: 10.1016/s0022-5320(69)80010-9

Allen, J. F. (1993). Redox control of transcription: sensors, response regulators, activators and repressors. *FEBS Lett.* 332, 203–207. doi: 10.1016/0014-5793(93)80631-4

Bendich, A. J. (1993). Reaching for the ring: the study of mitochondrial genome structure. *Curr. Genet.* 24, 279–290. doi: 10.1007/BF00336777

Bendich, A. J. (1996). Structural analysis of mitochondrial DNA molecules from fungi and plants using moving pictures and pulsed-field gel electrophoresis. *J. Mol. Biol.* 255, 564–588. doi: 10.1006/jmbi.1996.0048

Berná, L., Márquez, P., Cabrera, A., Greif, G., Francia, M. E., and Robello, C. (2021). Reevaluation of the *Toxoplasma gondii* and *Neospora caninum* genomes reveals misassembly, karyotype differences, and chromosomal rearrangements. *Genome Res.* 31, 823–833. doi: 10.1101/gr.262832.120

Biagini, G. A., Fisher, N., Shone, A. E., Mubaraki, M. A., Srivastava, A., Hill, A., et al. (2012). Generation of quinolone antimalarials targeting the *Plasmodium falciparum* mitochondrial respiratory chain for the treatment and prophylaxis of malaria. *Proc. Natl. Acad. Sci. U.S.A.* 109, 8298–8303. doi: 10.1073/pnas.1205651109

Biagini, G., Viriyavejakul, P., O'Neill, P., Bray, P., and Ward, S. (2006). Functional characterization and target validation of alternative complex I of *Plasmodium Falciparum* mitochondria. *Antimicrob. Agents Chemother.* 50, 1841–1851. doi: 10.1128/AAC.50.5.1841-1851.2006

Björkholm, P., Harish, A., Hagstrom, E., Ernst, A. M., and Andersson, S. G. (2015). Mitochondrial genomes are retained by selective constraints on protein targeting. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10154–10161. doi: 10.1073/pnas.1421372112

Blazejewski, T., Nursimulu, N., Pszenny, V., Dangoudoubiyam, S., Namasivayam, S., Chiasson, M. A., et al. (2015). Systems-Based analysis of the *Sarcocystis* neurona genome identifies pathways that contribute to a heteroxenous life cycle. *MBio* 6:e02445-14. doi: 10.1128/mBio.02445-14

Bogema, D., Yam, J., Micallef, M., Kanani, H. G., Go, J., Jenkins, C., et al. (2021). Draft genomes of *Perkinsus olseni* and *Perkinsus chesapeaki* reveal polyploidy and regional differences in heterozygosity. *Genomics* 113(1 Pt 2), 677–688. doi: 10.1016/j.ygeno.2020.09.064

Bonen, L. (1993). Trans-Splicing of Pre-MRNA in plants, animals, and protists. *FASEB J.* 7, 40–46. doi: 10.1096/fasebj.7.1.8422973

Burger, G., Gray, M. W., and Lang, B. (2003). Mitochondrial genomes: anything goes. *Trends Genet.* 19, 709–716. doi: 10.1016/j.tig.2003.10.012

Burger, G., Zhu, Y., Littlejohn, T. G., Greenwood, S. J., Schnare, M. N., Lang, B. F., et al. (2000). Complete sequence of the mitochondrial genome of *Tetrahymena pyriformis* and comparison with *Paramecium aurelia* mitochondrial DNA. *J. Mol. Biol.* 297, 365–380. doi: 10.1006/jmbi.2000.3529

Bushell, E., Gomes, A. R., Sanderson, T., Anar, B., Girling, G., Herd, C., et al. (2017). Functional profiling of a *Plasmodium* genome reveals an abundance of essential genes. *Cell* 170, 260–272.e8. doi: 10.1016/j.cell.2017.06.030

Cinar, H., Gopinath, G., Jarvis, K., and Murphy, H. (2015). The complete mitochondrial genome of the foodborne parasitic pathogen *Cyclospora cayetanensis*. *PLoS One* 10:e0128645. doi: 10.1371/journal.pone.0128645

Cornillot, E., Dassouli, A., Garg, A., Pachikara, N. D., Randazzo, S., Depoix, D., et al. (2013). Whole genome mapping and re-organization of the nuclear and mitochondrial genomes of *Babesia microti* isolates. *PLoS One* 8:e72657. doi: 10.1371/journal.pone.0072657

Cornillot, E., Hadj-Kaddour, K., Dassouli, A., Noel, B., Ranwez, V., Vacherie, B., et al. (2012). Sequencing of the smallest apicomplexan genome from the human pathogen babesia microti. *Nucleic Acids Res.* 40, 9102–9114. doi: 10.1093/nar/gks700

de Souza, W., Attias, M., and Rodrigues, J. C. F. (2009). Particularities of mitochondrial structure in parasitic protists (*Apicomplexa* and *Kinetoplastida*). *Int. J. Biochem. Cell Biol.* 41, 2069–2080. doi: 10.1016/j.biocel.2009.04.007

Divo, A. A., Geary, T. G., Jensen, J. B., and Ginsburg, H. (1985). The mitochondrion of *Plasmodium falciparum* visualized by rhodamine 123 fluorescence. *J. Protozool.* 32, 442–446. doi: 10.1111/j.1550-7408.1985.tb04041.x

Dooren, G. V., Stimmler, L. M., and McFadden, G. (2006). Metabolic maps and functions of the *Plasmodium* mitochondrion. *FEMS Microbiol. Rev.* 30, 596–630. doi: 10.1111/j.1574-6976.2006.00027.x

Dubremetz, J. F. (1973). Etude ultrastructurale de la mitose schizogonique chez la coccidie *Eimeria necatrix* (Johnson 1930). *J. Ultrastruct. Res.* 42, 354–376. doi: 10.1016/S0022-5320(73)90063-4

Edqvist, J., Burger, G., and Gray, M. W. (2000). Expression of mitochondrial protein-coding genes in *Tetrahymena pyriformis*. *J. Mol. Biol.* 297, 381–393. doi: 10.1006/jmbi.2000.3530

Entelis, N. S., Kolesnikova, O. A., Dogan, S., Martin, R. P., and Tarassov, I. A. (2001). 5 S RRNA and TRNA import into human mitochondria. comparison of in vitro requirements. *J. Biol. Chem.* 276, 45642–45653. doi: 10.1074/jbc.M103906200

Esseiva, A. C., Naguleswaran, A., Hemphill, A., and Schneider, A. (2004). Mitochondrial TRNA Import in *Toxoplasma gondii*. *J. Biol. Chem.* 279, 42363–42368. doi: 10.1074/jbc.M404519200

Evers, F., Cabrera-Orefice, A., Elurbe, D. M., Kea-Te Lindert, M., Boltryk, S. D., Voss, T. S., et al. (2021). Composition and stage dynamics of mitochondrial complexes in *Plasmodium falciparum*. *Nat. Commun.* 12:3820. doi: 10.1038/s41467-021-23919-x

Feagin, J. E. (1992). The 6-Kb element of *Plasmodium* falciparum encodes mitochondrial cytochrome genes. *Mol. Biochem. Parasitol.* 52, 145–148. doi: 10.1016/0166-6851(92)90046-m

Feagin, J. E. (1994). The extrachromosomal DNAs of apicomplexan parasites. *Annu. Rev. Microbiol.* 48, 81–104. doi: 10.1146/annurev.mi.48.100194.000501

Feagin, J. E., Mericle, B. L., Werner, E., and Morris, M. (1997). Identification of additional RRNA fragments encoded by the *Plasmodium falciparum* 6 Kb element. *Nucleic Acids Res.* 25, 438–446. doi: 10.1093/nar/25.2.438

Feagin, J., Harrell, M., Lee, J., Coe, K., Sands, B., Cannone, J., et al. (2012). The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum*. *PLoS One* 7:e38320. doi: 10.1371/journal.pone.0038320

Ferguson, D. J. P., and Dubremetz, J. F. (2014). "Chapter 2 – the ultrastructure of *Toxoplasma gondii*," in *Toxoplasma Gondii*, 2nd Edn, eds L. M. Weiss and K. Kim (Boston, MA: Academic Press), 19–59. doi: 10.1016/B978-0-12-396481-6.00002-7

Flegontov, P., and Lukeš, J. (2012). "Chapter Six – mitochondrial genomes of photosynthetic *Euglenids* and *Alveolates*," in *Advances in Botanical Research: Mitochondrial Genome Evolution*, Vol. 63, ed. L. Maréchal-Drouard (Boston, MA: Academic Press), 127–153. doi: 10.1016/B978-0-12-394279-1.00006-5

Flegontov, P., Michálek, J., Janouškovec, J., Lai, D., Jirkù, M., Hajdušková, E., et al. (2015). Divergent mitochondrial respiratory chains in phototrophic relatives of apicomplexan parasites. *Mol. Biol. Evol.* 32, 1115–1131. doi: 10.1093/molbev/msv021

Gagat, P., Mackiewicz, D., and Mackiewicz, P. (2017). Peculiarities within peculiarities – dinoflagellates and their mitochondrial genomes. *Mitochondrial DNA Part B Resour.* 2, 191–195. doi: 10.1080/23802359.2017.1307699

Goldschmidt-Clermont, M., Choquet, Y., Girard-Bascou, J., Michel, F., Schirmer-Rahire, M., and Rochaix, J.-D. (1991). A small chloroplast RNA may be required for trans-splicing in *Chlamydomonas reinhardtii*. *Cell* 65, 135–143. doi: 10.1016/0092-8674(91)90415-U

Goodman, C. D., Buchanan, H. D., and McFadden, G. I. (2017). Is the mitochondrion a good malaria drug target? *Trends Parasitol.* 33, 185–193. doi: 10.1016/j.pt.2016.10.002

Gray, M. W., and Boer, P. H. (1988). Organization and expression of algal (*Chlamydomonas reinhardtii*) mitochondrial DNA. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 319, 135–147. doi: 10.1098/rstb.1988.0038

Guo, J., Miao, X., He, P., Li, M., Wang, S., Cui, J., et al. (2019). Babesia gibsoni endemic to wuhan, china: mitochondrial genome sequencing, annotation, and comparison with *Apicomplexan Parasites*. *Parasitol. Res.* 118, 235–243. doi: 10.1007/s00436-018-6158-2

Hafeez, M. A., and Barta, J. R. (2019). Conserved mitochondrial genome organization of *Isospora* species (eimeriidae, coccidia, apicomplexa) infecting reptiles (*Pogona vitticeps* (sauria: agamidae) and birds (garrulax chinensis aves: passeriformes). *Mitochondrial DNA Part B* 4, 1133–1135. doi: 10.1080/23802359.2018.1564388

Hayward, J. A., and van Dooren, G. G. (2019). Same same, but different: uncovering unique features of the mitochondrial respiratory chain of apicomplexans. *Mol. Biochem. Parasitol.* 232:111204. doi: 10.1016/j.molbiopara.2019.111204

He, L., Zhang, Y., Zhang, Q., Zhang, W., Feng, H., Khan, M. K., et al. (2014). Mitochondrial Genome of *Babesia orientalis*, apicomplexan parasite of water buffalo (*Bubalus babalis*, Linnaeus, 1758) endemic in China. *Parasit. Vect.* 7:82. doi: 10.1186/1756-3305-7-82

Hepler, P. K., Huff, C. G., and Sprinz, H. (1966). The fine structure of the exoerythrocytic stages of *Plasmodium fallax*. *J. Cell Biol.* 30, 333–358. doi: 10.1083/jcb.30.2.333

Hikosaka, K., Watanabe, Y., Kobayashi, F., Waki, S., Kita, K., and Tanabe, K. (2011b). Highly conserved gene arrangement of the mitochondrial genomes of 23 *Plasmodium* species. *Parasitol. Int.* 60, 175–180. doi: 10.1016/j.parint.2011.02.001

Hikosaka, K., Nakai, Y., Watanabe, Y., Tachibana, S., Arisue, N., Palacpac, N., et al. (2011a). Concatenated mitochondrial DNA of the coccidian parasite *Eimeria tenella*. *Mitochondrion* 11, 273–278. doi: 10.1016/j.mito.2010.10.003

Hikosaka, K., Tsuji, N., Watanabe, Y., Kishine, H., Horii, T., Igarashi, I., et al. (2012). Novel type of linear mitochondrial genomes with dual flip-flop inversion system in apicomplexan parasites, *Babesia microti* and *Babesia rodhaini*. *BMC Genomics* 13:622. doi: 10.1186/1471-2164-13-622

Hikosaka, K., Watanabe, Y., Tsuji, N., Kita, K., Kishine, H., Arisue, N., et al. (2010). Divergence of the mitochondrial genome structure in the apicomplexan parasites, *Babesia* and *Theileria*. *Mol. Biol. Evol.* 27, 1107–1116. doi: 10.1093/molbev/msp320

Hillebrand, A., Matz, J. M., Almendinger, M., Müller, K., Matuschewski, K., and Schmitz-Linneweber, C. (2018). Identification of clustered organellar short (Cos) RNAs and of a conserved family of organellar RNA-binding proteins, the heptatricopeptide repeat proteins, in the malaria parasite. *Nucleic Acids Res.* 46, 10417–10431. doi: 10.1093/nar/gky710

Hjort, K., Goldberg, A. V., Tsaousis, A., Hirt, R., and Embley, T. (2010). Diversity and reductive evolution of mitochondria among microbial eukaryotes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 713–727. doi: 10.1098/rstb.2009.0224

Huet, D., Rajendran, E., Dooren, G. V., and Lourido, S. (2018). Identification of cryptic subunits from an apicomplexan ATP synthase. *ELife* 7:e38097. doi: 10.7554/eLife.38097

Jackson, C., and Waller, R. (2013). A widespread and unusual RNA trans-splicing type in dinoflagellate mitochondria. *PLoS One* 8:e56777. doi: 10.1371/journal.pone.0056777

Jackson, C., Norman, J. E., Schnare, M. N., Gray, M. W., Keeling, P., and Waller, R. (2007). Broad genomic and transcriptional analysis reveals a highly derived genome in *Dinoflagellate mitochondria*. *BMC Biol.* 5:41. doi: 10.1186/1741-7007-5-41

Jacobs, K., Charvat, R., and Arrizabalaga, G. (2020). Identification of Fis1 interactors in *Toxoplasma gondii* reveals a novel protein required for peripheral distribution of the mitochondrion. *MBio* 11:e02732-19. doi: 10.1128/mBio.02732-19

Janouškovec, J., Gavelis, G. S., Burki, F., Dinh, D., Bachvaroff, T., Gornik, S., et al. (2017). Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. *Proc. Natl. Acad. Sci. U.S.A.* 114, E171–E180. doi: 10.1073/pnas.1614842114

Janouškovec, J., Paskerova, G., Miroliubova, T. S., Mikhailov, K., Birley, T., Aleoshin, V., et al. (2019). Apicomplexan-like parasites are polyphyletic and widely but selectively dependent on cryptic plastid organelles. *ELife* 8:e49662. doi: 10.7554/eLife.49662

Janouškovec, J., Tikhonenkov, D. V., Mikhailov, K. V., Simdyanov, T. G., Aleoshin, V. V., Mylnikov, A. P., et al. (2013). Colponemids represent multiple ancient alveolate lineages. *Curr. Biol.* 23, 2546–2552. doi: 10.1016/j.cub.2013.10.062

Ji, Y. E., Mericle, B. L., Rehkopf, D. H., Anderson, J. D., and Feagin, J. E. (1996). The *Plasmodium falciparum* 6 Kb element is polycistronically transcribed. *Mol. Biochem. Parasitol.* 81, 211–223. doi: 10.1016/0166-6851(96)02712-0

Johnston, I., and Williams, B. P. (2016). Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. *Cell Syst.* 2, 101–111. doi: 10.1016/j.cels.2016.01.013

Joseph, J. T., Aldritt, S. M., Unnasch, T., Puijalon, O., and Wirth, D. F. (1989). Characterization of a conserved extrachromosomal element isolated from the avian malarial parasite *Plasmodium gallinaceum*. *Mol. Cell. Biol.* 9, 3621–3629. doi: 10.1128/mcb.9.9.3621-3629.1989

Kairo, A., Fairlamb, A. H., Gobright, E., and Nene, V. (1994). A 7.1 Kb linear DNA molecule of *Theileria parva* has scrambled RDNA sequences and open reading frames for mitochondrially encoded proteins. *EMBO J.* 13, 898–905.

Kamikawa, R., Inagaki, Y., and Sako, Y. (2007). Fragmentation of mitochondrial large subunit RRNA in the dinoflagellate *Alexandrium catenella* and the evolution of RRNA structure in alveolate mitochondria. *Protist* 158, 239–245. doi: 10.1016/j.protis.2006.12.002

Kappmeyer, L. S., Thiagarajan, M., Herndon, D. R., Ramsay, J. D., Caler, E., Djikeng, A., et al. (2012). Comparative genomic analysis and phylogenetic position of *Theileria equi*. *BMC Genomics* 13:603. doi: 10.1186/1471-2164-13-603

Karadjian, G., Hassanin, A., Saintpierre, B., Tungaluna, G. G., Ariey, F., Ayala, F., et al. (2016). Highly rearranged mitochondrial genome in *Nycteria parasites* (Haemosporidia) from bats. *Proc. Natl. Acad. Sci. U.S.A.* 113, 9834–9839. doi: 10.1073/pnas.1610643113

Karnkowska, A., Vacek, V., Zubáčová, Z., Treitli, S., Petrželková, R., Eme, L., et al. (2016). A eukaryote without a mitochondrial organelle. *Curr. Biol.* 26, 1274–1284. doi: 10.1016/j.cub.2016.03.053

Ke, H., Dass, S., Morrisey, J., Mather, M., and Vaidya, A. (2018). The mitochondrial ribosomal protein L13 is critical for the structural and functional integrity of the mitochondrion in *Plasmodium falciparum*. *J. Biol. Chem.* 293, 8128–8137. doi: 10.1074/jbc.RA118.002552

Ke, H., Morrisey, J., Ganesan, S. M., Mather, M., and Vaidya, A. (2012). Mitochondrial RNA polymerase is an essential enzyme in erythrocytic stages of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* 185, 48–51. doi: 10.1016/j.molbiopara.2012.05.001

Korsinczky, M., Chen, N., Kotecka, B., Saul, A., Rieckmann, K., and Cheng, Q. (2000). Mutations in *Plasmodium falciparum* cytochrome b that are associated with atovaquone resistance are located at a putative drug-binding

site. *Antimicrob. Agents Chemother.* 44, 2100–2108. doi: 10.1128/AAC.44.8.2100-2108.2000

Lacombe, A., Maclean, A. E., Ovciarikova, J., Tottey, J., Mühleip, A., Fernandes, P., et al. (2019). Identification of the *Toxoplasma gondii* mitochondrial ribosome, and characterisation of a protein essential for mitochondrial translation. *Mol. Microbiol.* 112, 1235–1252. doi: 10.1111/mmi.14357

Lang, B. F., Gray, M. W., and Burger, G. (1999). Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.* 33, 351–397. doi: 10.1146/annurev.genet.33.1.351

Lecrenier, N., and Foury, F. (2000). New features of mitochondrial DNA replication system in yeast and man. *Gene* 246, 37–48. doi: 10.1016/s0378-1119(00)00107-4

Lee, V. V., Judd, L. M., Jex, A. R., Holt, K. E., Tonkin, C. J., and Ralph, S. A. (2021). Direct nanopore sequencing of MRNA reveals landscape of transcript isoforms in apicomplexan parasites. *MSystems* 6:e01081-20. doi: 10.1128/mSystems.01081-20

Levine, N. D. (1988). *The Protozoan Phylum Apicomplexa*, Vol. 1–2. Boca Raton, FL: CRC Press, Inc.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Lin, S., Zhang, H., Spencer, D., Norman, J. E., and Gray, M. W. (2002). Widespread and extensive editing of mitochondrial MRNAS in dinoflagellates. *J. Mol. Biol.* 320, 727–739. doi: 10.1016/s0022-2836(02)00468-0

Ling, L., Mulaka, M., Munro, J., Dass, S., Mather, M., Riscoe, M., et al. (2020). Genetic ablation of the mitoribosome in the malaria parasite *Plasmodium falciparum* sensitizes it to antimalarials that target mitochondrial functions. *J. Biol. Chem.* 295, 7235–7248. doi: 10.1074/jbc.RA120.012646

Liu, G.-H., Hou, J., Weng, Y.-B., Song, H.-Q., Li, S., Yuan, Z.-G., et al. (2012). The complete mitochondrial genome sequence of *Eimeria mitis* (Apicomplexa: coccidia). *Mitochondrial DNA* 23, 341–343. doi: 10.3109/19401736.2012.690750

Maclean, A. E., Bridges, H. R., Silva, M. F., Ding, S., Ovciarikova, J., Hirst, J., et al. (2021). Complexome profile of *Toxoplasma gondii* mitochondria identifies divergent subunits of respiratory chain complexes including new subunits of cytochrome Bc1 complex. *PLoS Pathog.* 17:e1009301. doi: 10.1371/journal.ppat.1009301

Maguire, F., and Richards, T. A. (2014). "Organelle evolution: a mosaic of 'mitochondrial', Functions." *Curr. Biol.* 24, R518–R520. doi: 10.1016/j.cub.2014.03.075

Mangot, J., Debroas, D., and Domaizon, I. (2011). Perkinsozoa, a well-known marine protozoan flagellate parasite group, newly identified in lacustrine systems: a review. *Hydrobiologia* 659, 37–48. doi: 10.1007/s10750-010-0268-x

Martin, W. (2003). Gene transfer from organelles to the nucleus: frequent and in big chunks. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8612–8614. doi: 10.1073/pnas.1633606100

Masuda, I., Matsuzaki, M., and Kita, K. (2010). Extensive frameshift at all AGG and CCC codons in the mitochondrial cytochrome c oxidase subunit 1 gene of *Perkinsus marinus* (Alveolata; Dinoflagellata). *Nucleic Acids Res.* 38, 6186–6194. doi: 10.1093/nar/gkq449

Mather, M., Henry, K., and Vaidya, A. (2007). Mitochondrial drug targets in apicomplexan parasites. *Curr. Drug Targets* 8, 49–60. doi: 10.2174/138945007779315632

Mathur, V., Kolisko, M., Hehenberger, E., Irwin, N. A., Leander, B., Kristmundsson, A., et al. (2019). Multiple independent origins of apicomplexan-like parasites. *Curr. Biol.* 29, 2936–2941.e5. doi: 10.1016/j.cub.2019.07.019

Mathur, V., Wakeman, K. C., and Keeling, P. J. (2021). Parallel functional reduction in the mitochondria of apicomplexan parasites. *Curr. Biol.* 31, 2920–2928.e4. doi: 10.1016/j.cub.2021.04.028

Melatti, C., Pieperhoff, M., Lemgruber, L., Pohl, E., Sheiner, L., and Meissner, M. (2019). A unique dynamin-related protein is essential for mitochondrial fission in *Toxoplasma gondii*. *PLoS Pathog.* 15:e1007512. doi: 10.1371/journal.ppat.1007512

Melo, E. J., Attias, M., and De Souza, W. (2000). The single mitochondrion of tachyzoites of *Toxoplasma gondii*. *J. Struct. Biol.* 130, 27–33. doi: 10.1006/jsbi.2000.4228

Muhleip, A., Kock Flygaard, R., Ovciarikova, J., Lacombe, A., Fernandes, P., Sheiner, L., et al. (2021). ATP synthase hexamer assemblies shape cristae of toxoplasma mitochondria. *Nat. Commun.* 12:120. doi: 10.1038/s41467-020-20381-z

Nadimi, M., Beaudet, D., Forget, L., Hijri, M., and Lang, B. (2012). Group I intron-mediated trans-splicing in mitochondria of *Gigaspora rosea* and a robust phylogenetic affiliation of Arbuscular mycorrhizal fungi with Mortierellales. *Mol. Biol. Evol.* 29, 2199–2210. doi: 10.1093/molbev/mss088

Namasivayam, S., Baptista, R. P., Xiao, W., Hall, E. M., Doggett, J. S., Troell, K., et al. (2021). A novel fragmented mitochondrial genome in the Protist pathogen *Toxoplasma gondii* and related tissue *Coccidia*. *Genome Res.* 31, 852–865. doi: 10.1101/gr.266403.120

Nash, E. A., Barbrook, A. C., dwards-Stuart, R. K. E., Bernhardt, K., Howe, C. J., and Nisbet, R. E. (2007). Organization of the mitochondrial genome in the dinoflagellate *Amphidinium* Carterae. *Mol. Biol. Evol.* 24, 1528–1536. doi: 10.1093/molbev/msm074

Nash, E. A., Nisbet, R. E. R., Barbrook, A. C., and Howe, C. J. (2008). Dinoflagellates: a mitochondrial genome all at sea. *Trends Genet. TIG* 24, 328–335. doi: 10.1016/j.tig.2008.04.001

Nishi, M., Hu, K., Murray, J. M., and Roos, D. S. (2008). Organellar dynamics during the cell cycle of *Toxoplasma gondii*. *J. Cell Sci.* 121(Pt 9), 1559–1568. doi: 10.1242/jcs.021089

Oborník, M., and Lukeš, J. (2015). The organellar genomes of *Chromera* and *Vitrella*, the phototrophic relatives of apicomplexan parasites. *Annu. Rev. Microbiol.* 69, 129–144. doi: 10.1146/annurev-micro-091014-104449

Ogedengbe, M. E., El-Sherry, S., Whale, J., and Barta, J. R. (2014). Complete mitochondrial genome sequences from five *Eimeria* species (apicomplexa; coccidia; eimeriidae) infecting domestic Turkeys. *Parasit. Vect.* 7:335. doi: 10.1186/1756-3305-7-335

Ogedengbe, M. E., Qvarnstrom, Y., da Silva, A. J., Arrowood, M. J., and Barta, J. R. (2015). A linear mitochondrial genome of *Cyclospora cayetanensis* (eimeriidae, eucoccidiorida, coccidiasina, apicomplexa) suggests the ancestral start position within mitochondrial genomes of *Eimeriid coccidia*. *Int. J. Parasitol.* 45, 361–365. doi: 10.1016/j.ijpara.2015.02.006

Oldenburg, D. J., and Bendich, A. J. (2001). "Mitochondrial DNA from the liverwort marchantia polymorpha: circularly permuted linear molecules, head-to-tail concatemers, and a 5′, Protein. *J. Mol. Biol.* 310, 549–562. doi: 10.1006/jmbi.2001.4783

Ovciarikova, J., Lemgruber, L., Stilger, K. L., Sullivan, W. J., and Sheiner, L. (2017). Mitochondrial behaviour throughout the lytic cycle of *Toxoplasma gondii*. *Sci. Rep.* 7:42746. doi: 10.1038/srep42746

Pacheco, M. A., Matta, N. E., Valkiunas, G., Parker, P. G., Mello, B., Stanley, C. E., et al. (2018). Mode and rate of evolution of *Haemosporidian* mitochondrial genomes: timing the radiation of avian parasites. *Mol. Biol. Evol.* 35, 383–403. doi: 10.1093/molbev/msx285

Pan, Y., Cao, M., Liu, J., Yang, Q., Miao, X., Go, V. L. W., et al. (2017). Metabolic regulation in mitochondria and drug resistance. *Adv. Exp. Med. Biol.* 1038, 149–171. doi: 10.1007/978-981-10-6674-0_11

Park, K.-M., Min, G.-S., and Kim, S. (2019). The mitochondrial genome of the ciliate *Pseudourostyla cristata* (Ciliophora, Urostylida). *Mitochondrial DNA Part B* 4, 66–67. doi: 10.1080/23802359.2018.1536458

Pino, P., Aeby, E., Foth, B. J., Sheiner, L., Soldati, T., Schneider, A., et al. (2010). Mitochondrial translation in absence of local TRNA aminoacylation and methionyl TRNA met formylation in apicomplexa. *Mol. Microbiol.* 76, 706–718. doi: 10.1111/j.1365-2958.2010.07128.x

Preiser, P. R., Wilson, R. J., Moore, P. W., McCready, S., Hajibagheri, M. A., Blight, K. J., et al. (1996). Recombination associated with replication of malarial mitochondrial DNA. *EMBO J.* 15, 684–693.

Putignani, L., Tait, A., Smith, H. V., Horner, D., Tovar, J., Tetley, L., et al. (2004). Characterization of a mitochondrion-like organelle in *Cryptosporidium parvum*. *Parasitology* 129(Pt 1), 1–18. doi: 10.1017/s003118200400527x

Rehkopf, D. H., Gillespie, D. E., Harrell, M. I., and Feagin, J. E. (2000). Transcriptional mapping and RNA processing of the *Plasmodium falciparum* mitochondrial MRNAs. *Mol. Biochem. Parasitol.* 105, 91–103. doi: 10.1016/s0166-6851(99)00170-x

Riordan, C. E., Ault, J. G., Langreth, S. G., and Keithly, J. S. (2003). *Cryptosporidium parvum* Cpn60 targets a relict organelle. *Curr. Genet.* 44, 138–147. doi: 10.1007/s00294-003-0432-1

Roger, A. J., Muñoz-Gómez, S. A., and Kamikawa, R. (2017). The origin and diversification of mitochondria. *Curr. Biol.* 27, R1177–R1192. doi: 10.1016/j.cub.2017.09.015

Rudzinska, M. A. (1969). The fine structure of malaria parasites. *Int. Rev. Cytol.* 25, 161–199. doi: 10.1016/s0074-7696(08)60203-x

Salomaki, E. D., Terpis, K. X., Rueckert, S., Kotyk, M., Varadínová, Z. K., Čepička, I., et al. (2021). Gregarine single-cell transcriptomics reveals differential mitochondrial remodeling and adaptation in apicomplexans. *BMC Biol.* 19:77. doi: 10.1186/s12915-021-01007-2

Salunke, R., Mourier, T., Banerjee, M., Pain, A., and Shanmugam, D. (2018). Highly diverged novel subunit composition of apicomplexan F-Type ATP synthase identified from *Toxoplasma gondii*. *PLoS Biol.* 16:e2006128. doi: 10.1371/journal.pbio.2006128

Santos, H. J., Makiuchi, T., and Nozaki, T. (2018). Reinventing an organelle: the reduced mitochondrion in parasitic protists. *Trends Parasitol.* 34, 1038–1055. doi: 10.1016/j.pt.2018.08.008

Schreeg, M. E., Marr, H. S., Tarigo, J. L., Cohn, L. A., Bird, D. M., Scholl, E. H., et al. (2016). Mitochondrial genome sequences and structures aid in the resolution of piroplasmida phylogeny. *PLoS One* 11:e0165702. doi: 10.1371/journal.pone.0165702

Seeber, F., Feagin, J. E., and Parsons, M. (2014). "Chapter 9 – the apicoplast and mitochondrion of *Toxoplasma gondii*," in *Toxoplasma Gondii*, 2nd Edn, eds L. M. Weiss and K. Kim (Boston, MA: Academic Press), 297–350. doi: 10.1016/B978-0-12-396481-6.00009-X

Seeber, F., Limenitakis, J., and Soldati-Favre, D. (2008). Apicomplexan mitochondrial metabolism: a story of gains, losses and retentions. *Trends Parasitol.* 24, 468–478. doi: 10.1016/j.pt.2008.07.004

Seidi, A., Muellner-Wong, L. S., Rajendran, E., Tjhin, E. T., Dagley, L. F., Aw, V. Y., et al. (2018). Elucidating the mitochondrial proteome of *Toxoplasma gondii* reveals the presence of a divergent cytochrome c oxidase. *ELife* 7:e38131. doi: 10.7554/eLife.38131

Sharma, A., and Sharma, A. (2015). *Plasmodium falciparum* mitochondria import TRNAs along with an active phenylalanyl-TRNA synthetase. *Biochem. J.* 465, 459–469. doi: 10.1042/BJ20140998

Shoguchi, E., Shinzato, C., Hisata, K., Satoh, N., and Mungpakdee, S. (2015). The large mitochondrial genome of *Symbiodinium minutum* reveals conserved noncoding sequences between dinoflagellates and apicomplexans. *Genome Biol. Evol.* 7, 2237–2244. doi: 10.1093/gbe/evv137

Sidik, S. M., Huet, D., Ganesan, S. M., Huynh, M.-H., Wang, T., Nasamu, A. S., et al. (2016). a genome-wide CRISPR screen in toxoplasma identifies essential *Apicomplexan* genes. *Cell* 166, 1423–1435.e12. doi: 10.1016/j.cell.2016.08.019

Simpson, L. (1986). Kinetoplast DNA in trypanosomid flagellates. *Int. Rev. Cytol.* 99, 119–179. doi: 10.1016/s0074-7696(08)61426-6

Slamovits, C. H., Saldarriaga, J. F., Larocque, A., and Keeling, P. J. (2007). The highly reduced and fragmented mitochondrial genome of the early-branching dinoflagellate oxyrrhis marina shares characteristics with both apicomplexan and dinoflagellate mitochondrial genomes. *J. Mol. Biol.* 372, 356–368. doi: 10.1016/j.jmb.2007.06.085

Šlapeta, J., and Keithly, J. S. (2004). *Cryptosporidium parvum* mitochondrial-type HSP70 targets homologous and heterologous mitochondria. *Eukaryot. Cell* 3, 483–494. doi: 10.1128/EC.3.2.483-494.2004

Slomianny, C., and Prensier, G. (1986). Application of the serial sectioning and tridimensional reconstruction techniques to the morphological study of the *Plasmodium falciparum* mitochondrion. *J. Parasitol.* 72, 595–598.

Suplick, K., Akella, R., Saul, A., and Vaidya, A. B. (1988). Molecular cloning and partial sequence of a 5.8 kilobase pair repetitive DNA from *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* 30, 289–290. doi: 10.1016/0166-6851(88)90098-9

Suplick, K., Morrisey, J., and Vaidya, A. B. (1990). Complex transcription from the extrachromosomal DNA encoding mitochondrial functions of *Plasmodium yoelii*. *Mol. Cell. Biol.* 10, 6381–6388. doi: 10.1128/mcb.10.12.6381-6388.1990

Tachezy, J. ed. (2008). *Hydrogenosomes and Mitosomes: Mitochondria of Anaerobic Eukaryotes. Microbiology Monographs*. Berlin: Springer-Verlag. doi: 10.1007/978-3-540-76733-6

Tang, K., Guo, Y., Zhang, L., Rowe, L. A., Roellig, D. M., Frace, M. A., et al. (2015). Genetic similarities between *Cyclospora cayetanensis* and cecum-infecting avian *Eimeria* Spp. in apicoplast and mitochondrial genomes. *Parasit. Vect.* 8:358. doi: 10.1186/s13071-015-0966-3

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics* 14, 178–192. doi: 10.1093/bib/bbs017

Tian, S.-Q., Cui, P., Fang, S.-F., Liu, G.-H., Wang, C.-R., and Zhu, X.-Q. (2015). The complete mitochondrial genome sequence of *Eimeria magna* (apicomplexa: coccidia). *Mitochondrial DNA* 26, 714–715. doi: 10.3109/19401736.2013.843088

Tikhonenkov, D. V., Janouškovec, J., Mylnikov, A. P., Mikhailov, K. V., Simdyanov, T. G., Aleoshin, V. V., et al. (2014). Description of *Colponema vietnamica* Sp.n. and *Acavomonas peruviana* n. Gen. n. Sp., two new *Alveolate phyla* (*Colponemidia* Nom. Nov. and *Acavomonidia* Nom. Nov.) and their contributions to reconstructing the ancestral state of *Alveolates* and *Eukaryotes*. *PLoS One* 9:e95467. doi: 10.1371/journal.pone.0095467

Vaidya, A. B., Akella, R., and Suplick, K. (1989). Sequences similar to genes for two mitochondrial proteins and portions of ribosomal RNA in tandemly arrayed 6-kilobase-pair DNA of a malarial parasite. *Mol. Biochem. Parasitol.* 35, 97–107. doi: 10.1016/0166-6851(89)90112-6

Vaidya, A. B., and Arasu, P. (1987). Tandemly arranged gene clusters of malarial parasites that are highly conserved and transcribed. *Mol. Biochem. Parasitol.* 22, 249–257. doi: 10.1016/0166-6851(87)90056-9

Vaidya, A. B., and Mather, M. W. (2009). Mitochondrial evolution and functions in malaria parasites. *Annu. Rev. Microbiol.* 63, 249–267. doi: 10.1146/annurev.micro.091208.073424

Vaidya, A. B., Lashgari, M. S., Pologe, L. G., and Morrisey, J. (1993). Structural features of *Plasmodium* cytochrome b that may underlie susceptibility to 8-aminoquinolines and hydroxynaphthoquinones. *Mol. Biochem. Parasitol.* 58, 33–42. doi: 10.1016/0166-6851(93)90088-f

Vaidya, A. B., Painter, H. J., Morrisey, J. M., and Mather, M. W. (2008). The validity of mitochondrial dehydrogenases as antimalarial drug targets. *Trends Parasitol.* 24, 8–9. doi: 10.1016/j.pt.2007.10.005

Virji, A. Z., Thekkiniath, J., Ma, W., Lawres, L., Knight, J., Swei, A., et al. (2019). Insights into the evolution and drug susceptibility of *Babesia duncani* from the sequence of its mitochondrial and apicoplast genomes. *Int. J. Parasitol.* 49, 105–113. doi: 10.1016/j.ijpara.2018.05.008

von Heijne, G. (1986). Why mitochondria need a genome. *FEBS Lett.* 198, 1–4. doi: 10.1016/0014-5793(86)81172-3

Waller, R. F., and Jackson, C. J. (2009). Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. *BioEssays* 31, 237–245. doi: 10.1002/bies.200800164

Woo, Y. H., Ansari, H., Otto, T. D., Klinger, C. M., Kolisko, M., Michálek, J., et al. (2015). Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *ELife* 4:e06974. doi: 10.7554/eLife.06974

Xia, J., Venkat, A., Bainbridge, R. E., Reese, M. L., Le Roch, K. G., Ay, F., et al. (2021). Third-Generation sequencing revises the molecular karyotype for *Toxoplasma gondii* and identifies emerging copy number variants in sexual recombinants. *Genome Res.* 31, 834–851. doi: 10.1101/gr.262816.120

Yang, R., Brice, B., Oskam, C., Zhang, Y., Brigg, F., Berryman, D., et al. (2017). Characterization of two complete isospora mitochondrial genomes from passerine birds*: Isospora serinuse* in a domestic canary and *Isospora manorinae* in a yellow-throated miner. *Vet. Parasitol.* 237, 137–142. doi: 10.1016/j.vetpar.2017.01.027

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 November 2021
doi: 10.3389/fgene.2021.751552

Genetic Diversity of Polymorphic Marker Merozoite Surface Protein 1 (*Msp-1*) and 2 (*Msp-2*) Genes of *Plasmodium falciparum* Isolates From Malaria Endemic Region of Pakistan

Shahid Niaz Khan[1], Rehman Ali[1]*, Sanaullah Khan[2], Muhammad Rooman[3], Sadia Norin[1], Shehzad Zareen[1], Ijaz Ali[4] and Sultan Ayaz[5]*

[1]Department of Zoology, Faculty of Biological Sciences, Kohat University of Science and Technology, Kohat, Pakistan, [2]Department of Zoology, University of Peshawar, Peshawar, Pakistan, [3]Department of Zoology, Hazara University, Mansehra, Pakistan, [4]Department of Biosciences, COMSATS University Islamabad, Islamabad, Pakistan, [5]College of Veterinary Sciences and Animal Husbandry, Abdul Wali Khan University Garden Campus Mardan, Mardan, Pakistan

OPEN ACCESS

Edited by:
Moses Okpeku, University of KwaZulu-Natal, South Africa

Reviewed by:
Diego Garzón-Ospina, Universidad Pedagógica y Tecnológica de Colombia, Colombia
Olusola Ojurongbe, Ladoke Akintola University of Technology, Nigeria

***Correspondence:**
Shahid Niaz Khan
shahid@kust.edu.pk
Rehman Ali
rehmanali7680@gmail.com

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: *01 August 2021*
Accepted: *04 October 2021*
Published: *17 November 2021*

Citation:
Khan SN, Ali R, Khan S, Rooman M, Norin S, Zareen S, Ali I and Ayaz S (2021) Genetic Diversity of Polymorphic Marker Merozoite Surface Protein 1 (Msp-1) and 2 (Msp-2) Genes of Plasmodium falciparum Isolates From Malaria Endemic Region of Pakistan. Front. Genet. 12:751552. doi: 10.3389/fgene.2021.751552

Background: Understanding the genetic diversity of *Plasmodium* species through polymorphic studies can assist in designing more effective control strategies of malaria like new drug formulation and development of a vaccine. Pakistan is moderate endemic for *Plasmodium falciparum*, but little is known about the genetic diversity of this parasite. This study aimed to investigate the molecular diversity of *P. falciparum* based on *msp-1* and *msp-2* genes in the malaria-endemic regions of Khyber Pakhtunkhwa, Pakistan.

Methods: A total of 199/723 blood samples, tested positive by microscopy for *falciparum* malaria, were collected from four districts (Dera Ismail Khan, Karak, Mardan, and Peshawar) of Khyber Pakhtunkhwa. Nested PCR amplification technique was employed to target block 2 of *msp-1* and the central domain of *msp-2* genes, including their respective allelic families K1, MAD20, RO33, FC27, and 3D7/IC, and to detect the extent of genetic diversity of *P. falciparum* clinical isolates.

Results: Among the 199 microscopy-positive *P. falciparum* samples, a total of 192 were confirmed using PCR. Ninety-seven amplicons were observed for *msp-1* and 95 for *msp-2*. A total of 33 genotypes, 17 for *msp-1* (eight K1, six MAD20, and three RO33) and 16 for *msp-2* (nine FC27 and seven 3D7/IC), were identified. The specific allelic frequency of the K1 family was higher (44.3%) than that of MAD20 (33.0%) and RO33 (23.0%) for *msp-1*, while the FC27 allelic family was dominant (60.0%) compared with 3D7/IC (40.0%) for *msp-2*. No polyclonal infection was observed in *msp-1* and *msp-2*. The expected heterozygosity was 0.98 and 0.97 for *msp-1* and *msp-2*, respectively.

Conclusion: It was concluded that the *P. falciparum* populations are highly polymorphic, and diverse allelic variants of *msp-1* and *msp-2* are present in Khyber Pakhtunkhwa, Pakistan.

Keywords: *Plasmodium falciparum*, genetic diversity, polymorphism, msp-1 and msp-2 genes, nested PCR, Khyber Pakhtunkhwa, Pakistan

INTRODUCTION

Malaria causes 300–500 million cases worldwide and approximately 0.5–3 million deaths annually, the majority of which are caused by *Plasmodium falciparum* (Phillips, 2001; Ferreira et al., 2004). Pakistan is a moderate malaria-endemic region, and approximately 60% of its population is living in the endemic areas, whereas 177 million individuals are at risk of malaria (Qureshi et al., 2019). Annually, the estimated number of suspected and confirmed individuals is 3.5 million in Pakistan. The WHO included Pakistan as one of the six Eastern Mediterranean region countries with almost 100% population at risk of malaria (WHO, 2017; 2018). The endemicity of malaria varies in different provinces and even in different cities having variable climates. In 2017, about 30% of total malaria cases were reported from Khyber Pakhtunkhwa only, and the province has the highest reported cases of malaria (WHO, 2017).

The National Malaria Control Programmes (NMCP) has reported a record sixfold increase in *P. falciparum* during the last decade. The *falciparum* malaria has acknowledged sparse scientific attention, particularly concerning the molecular characterization of the local parasite population. The rise of *P. falciparum* in various districts of Pakistan may be attributable to the failed treatment of chloroquine resistance (Nizamani et al., 2006). Besides that, the heavy influx and continued presence of refugees from Afghanistan, where malaria is most prevalent, may contribute not only to the increasing number of cases of malaria but also to its genetic variations in Khyber Pakhtunkhwa province (Murtaza et al., 2004; Sheikh et al., 2005; Howard et al., 2011).

Analyses of the genotypes of *Plasmodium* spp. by PCR have remarkably improved our understanding of the biology of these parasites. In this regard, genetically distinct *P. falciparum* has been identified and extensively studied to further unveil its molecular epidemiology, parasite resistance, and potential vaccine candidates (Diggs et al., 1993; Greenhouse et al., 2006). The mainly used genetic markers of *P. falciparum* are the merozoite surface protein 1 (*msp-1*), merozoite surface protein 2 (*msp-2*), and glutamate-rich protein (*glurp*) (Smythe et al., 1991; Snounou and Beck, 1998). Allelic forms of these polymorphic markers have been reported in various parts of the world (Babiker et al., 1997; Jordan et al., 2001).

The *msp-1* and *msp-2* are antigenic proteins responsible for immunological responses in humans (Taylor et al., 1995; Aubouy et al., 2003). Block 2 of *msp-1* has three polymorphic allelic families identified as MAD20, K1, and RO33 (Contamin et al., 1996). Similarly, the central domain of the *msp-2* has two distinct families, i.e., 3D/IC and FC27 (Sallenave-Sales et al., 2000). These markers are unlinked and located on different chromosomes (Färnert et al., 2001). These features make them attractive candidates for studies where identification and enumeration of genetically distinct *P. falciparum* parasite subpopulations are of interest. As such, they have proven to be useful tools in molecular epidemiology studies in different epidemiological settings as well as to distinguish treatment failures from new infections in antimalarial drug trials (Cattamanchi et al., 2003; Collins et al., 2006).

The genetic diversity of *P. falciparum* population is an important indicator of the malaria transmission intensity in an area (Babiker et al., 1995; Paul et al., 1998). A high endemic area is generally characterized by extensive parasite diversity, and infected humans often carry multiple genotypes. Conversely, the parasite population in a low transmission area has a limited genetic diversity, and most infections are monoclonal (Babiker et al., 1997; Haddad et al., 1999; Peyerl-Hoffmann et al., 2001; Gómez et al., 2002). The *P. falciparum* field isolates have been characterized in Afghanistan, Iran, and India and previously from Sindh and Baluchistan in Pakistan, using the above-mentioned molecular markers. Therefore, we investigated the genetic diversity and polymorphic nature of *P. falciparum* isolates in selective districts of the malaria endemic province Khyber Pakhtunkhwa of Pakistan.

MATERIALS AND METHODS

Ethics and Consent for Participation

The study protocol was approved by the Institutional Ethical Review Committee of Kohat University of Science and Technology (KUST), Kohat-26000, Pakistan. Signed and written informed consent was obtained from the participants/legal guardians before sample collection.

Sample Collection and Analysis

Blood samples were randomly collected from suspected individuals ≥1 year at the Malaria Control Laboratories of District Headquarters Hospitals (DHQs) in four districts of Khyber Pakhtunkhwa (Dera Ismail Khan, Karak, Peshawar, and Mardan). A total of 723 suspected individuals with fever or history of fever were screened. Finger-pricked blood was collected on a glass-slide to prepare thick and thin blood smears, air-dried, and stained with Giemsa's stain (10%) for 15 min. The slides were examined under a microscope (Olympus CX31, Tokyo, Japan) by experienced laboratory technicians for *Plasmodium* species-specific identification. After careful examination, blood smears were considered negative when no parasite was detected and vice versa. Among the screened individuals, a total of 199 were confirmed positive for *P. falciparum* infection by microscopy.

Additionally, 200 µl of blood was collected into EDTA tubes, labelled, and transferred to the Molecular Parasitology and Virology Laboratory, Department of Zoology, KUST, Kohat-26000, Khyber Pakhtunkhwa, Pakistan, and stored at −80°C in a low deep freezer until genomic DNA extraction. Furthermore, a brief epidemiological/demographic history was also recorded using a structured questionnaire. The demographic data will be published elsewhere.

DNA Extraction and PCR Amplification

Genomic DNA was extracted from blood using the GF1 DNA extraction kit (Vivantis, Shah Alam, Selangor). Nested PCR was recruited to determine *Plasmodium* species with the help of 18S rRNA amplification using primers rPLUf: 5′-TTAAAATTGTTG CAGTTAAAACG-3′ and rPLUr: 5′-CCTGTTGTTGCCTTA AACTTC-3′ as previously described (Snounou et al., 1993; Singh et al., 1999). The amplicons were used as a template in

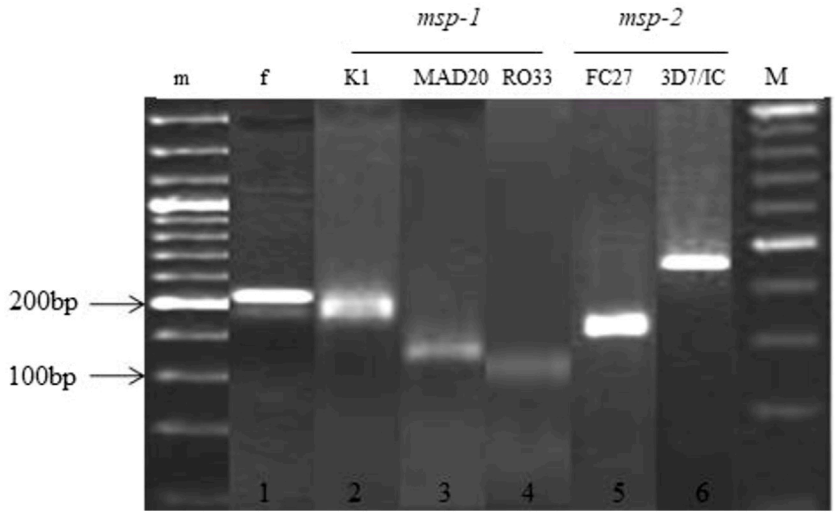


FIGURE 1 | Allelic families of *msp-1* and *msp-2* genes of *Plasmodium falciparum* 1 show the positive infection of *Plasmodium falciparum*; 2, 3, and 4 represent the corresponding alleles of *msp-1*; 5 and 6 show the alleles of *msp-2*; m is a 50-bp DNA ladder marker, and M is a 100-bp DNA ladder marker.

the nested PCR; and species-specific primers rFALf: 5′-TTAAAC TGGTTTGGGAAAACCAAATATATT-3′ and rFALr: 5′-ACA CAATGAACTCAATCATGACTACCCGTC-3′ were employed for *P. falciparum* detection (Snounou et al., 1993). Deionized water and genomic DNA from laboratory strains were used as negative and positive control, respectively. Both primary and secondary PCRs were carried out in a final volume of 20 µl containing 5.0 µl of DNA, 0.4 µl of 10 mM of each dNTP (Fermentas, Hanover, MD, USA), 2.4 µl of 10× PCR buffer, 1.5 µl of 25 mM $MgCl_2$, 0.5 µl of 5 U/µl *Taq* DNA polymerase (Fermentas, Hanover, MD, USA), paired primers of 1.0 µl each (20 µM), and 8.2 µl of sterile water. Thermocycler (Nyxtech, Boston, MA, USA) was used to complete all the reactions under the following conditions for primary (35 cycles) and secondary PCR (30 cycles): initial denaturation at 94°C for 1 min, extension at 94°C for 1 min, 58°C for 2 min and 72°C for 5 min, and final elongation at 72°C for 5 min.

The *msp-1* and *msp-2* genes were amplified using specific primers as per standard protocol previously described (Snounou et al., 1999). The primary reaction used a set of primers corresponding to the conserved regions of block 2 for *msp-1* and block 3 for *msp-2*. The second reaction primer set targets specific allelic families of *msp-1* (KI, MAD20, and RO33) or *msp-2* (3D7/IC and FC27). The cycling conditions for both *msp-1* and *msp-2* as well as primers were previously described (Somé et al., 2018).

Agarose gel (2%) stained with ethidium bromide was used to evaluate the PCR products under UV illumination. A 50 and 100-bp DNA ladders (Promega, Madison, WI, USA) were used; and alleles of *msp-1* and *msp-2* were categorized according to their molecular weights.

Statistical Analysis

Primarily, Microsoft Excel was used to manage the data. Statistical analyses were performed using SPSS (Version 20). The allelic frequencies for *msp-1* and *msp-2* were calculated and expressed in percentages. The proportion of alleles observed at each locus was compared using a chi-square test. The expected heterozygosity (H_e) was calculated using the following formula: $H_e = [n/(n-1)][(1-\Sigma p_i^2)]$, where n is the number of isolates sampled and p_i is the allele frequency at a given locus (Nei, 1978). The p-value (0.05) was assumed to be statistically significant.

TABLE 1 | Genotyping of *Plasmodium falciparum msp-1* in Khyber Pakhtunkhwa.

MSP-1 (n = 97)	Size (bp)	Frequency (%)	No. of alleles	*p*-Value[a]
K1	180–250	43 (44.3)	8	0.994
MAD20	110–250	32 (33.0)	6	
RO33	130–190	22 (23.0)	3	

Note. bp, base pair.
[a]The p-value represents significance level of allelic variants among different districts.

TABLE 2 | Genotyping of *Plasmodium falciparum msp-2* in Khyber Pakhtunkhwa.

MSP-2 (n = 95)	Size (bp)	Frequency (%)	No. of alleles	*p*-Value[a]
3D7/IC	400–580	38 (40.0)	9	0.963
FC27	300–430	57 (60.0)	7	

Note. bp, base pair.
[a]The p-value represents significance level of allelic variants among different districts.

RESULTS

Out of 199 microscopy-positive samples, 192 were further confirmed for *P. falciparum* using PCR, whereas seven samples (3.2%) were excluded due to negative PCR outcome (**Figure 1**).

Out of total, 97 amplicons were observed for *msp-1* and 95 for *msp-2*. In *msp-1*, the highest frequency of K1, MAD20, and RO33 was 44.3% (43/97), 33.0% (32/97), and 23.0% (22/97), respectively (**Table 1**). In *msp-2*, the FC27 has the highest frequency of 60.0% (57/95), followed by 3D7/IC with 40.0% (38/95) (**Table 2**). However, no polyclonal infection was observed in *msp-1* and *msp-2*.

Seventeen alleles were detected in *msp*-1 comprising eight alleles (180–350 bp) for K1, six alleles (110–250 bp) for MAD20, and three alleles (130–190 bp) for RO33. However, 16 alleles were attributed to *msp-2*, with nine alleles (400–580 bp) for FC27 and seven alleles (300–430 bp) for 3D7/IC. The district-wise details of allelic frequencies are presented in the supplementary materials (**Supplementary Tables S1, S2**). The expected heterozygosity was 0.98 and 0.97 for *msp-1* and *msp-2*, respectively.

DISCUSSION

Molecular studies provide an insight into the transmission intensity and genetic variation of parasite population within a region. The genetic diversity may be linked with the cross-border movement of populations living in the Frontier Regions of Pakistan (Ghanchi et al., 2010; Zakeri et al., 2010). Usually, areas with high malaria transmission are observed to have an extensive genetic diversity (Paul et al., 1998; Peyerl-Hoffmann et al., 2001). This study provides the basis to explore the genetic diversity of *P. falciparum* in the endemic regions of Khyber Pakhtunkhwa.

In the present study, the number of successfully genotyped samples for *msp-1* was higher as compared with that for *msp-2*. This result is consistent with studies reported from Cote d'Ivoire and Gabon (Yavo et al., 2016) and Burkina Faso (Soulama et al., 2009; Somé et al., 2018). However, in contrast, Mohammad et al. (2019) from Ethiopia, Soe et al. (2017) from Myanmar, and A-Elbasit et al. (2007) from Sudan reported a relatively high frequency of *msp-2* genotyped samples. Furthermore, of 192 positive samples for *P. falciparum*, less than half showed an amplicon for *msp-1* (n = 97) or *msp-2* (n = 95). In previous studies, relatively higher numbers of amplicons were observed for *msp-1* or *msp-2* (Somé et al., 2018; Mohammed et al., 2019; Eltayeb et al., 2020; Papa Mze et al., 2020). The nonspecific binding during *P. falciparum* identification by PCR may justify the smaller number of amplicons for *msp-1* and *msp-2* in the current study.

It was observed that 17 allelic variants of *msp-1* and 16 of *msp-2* were present in the studied areas. This result is in comparison with a similar study from the south of Pakistan (Ghanchi et al., 2010), Iran, South Africa, Myanmar, Sudan, Senegal, and Thailand (Heidari et al., 2007; Soe et al., 2017; Somé et al., 2018; Ndiaye et al., 2019; Eltayeb et al., 2020). However, in two southern districts Bannu and Kohat of Khyber Pakhtunkhwa, less allelic variants of *P. falciparum* were reported (Khatoon et al., 2010; Khatoon et al., 2012). Similarly, a study from the hypo-endemic area of Colombia reported only one allele of *msp-1* and three alleles of *msp-2* (Montoya et al., 2003). This difference might be due to low malaria endemicity since higher allele frequencies have been reported with high malaria transmission (Konaté et al., 1999; Soulama et al., 2009), suggesting that malaria endemicity affects the circulating strain number. However, high genetic diversity was observed in the Kingdom of Eswatini, which is regarded as a low transmission area for *P. falciparum* (Roh et al., 2019). Therefore, the genetic diversity may also be attributed to the effects of several factors such as indiscriminate use of long-lasting insecticide-treated nets (LLINs), indoor insecticide spraying, and antimalarial pressure (Soulama et al., 2009).

The K1 and FC27 alleles of *msp-1* and *msp-2* were predominant, respectively. The K1 and FC27 allelic families were previously reported from Kohat district (Khatoon et al., 2012). Nevertheless, the present findings also showed slight discrepancy with the previous studies (Zakeri et al., 2005; Ghanchi et al., 2010). The current study and previous findings suggest that K1 and FC27 allelic families might have prominent roles in clinical malaria at least in southern Khyber Pakhtunkhwa. However, further in-depth investigation with larger dataset should be carried out to unveil the genetic diversity and prevailing genotypes.

Interestingly, no polyclonal infection was detected in the present study. Malaria treatment policies, geographic isolation, and transmission intensities may result in spatial heterogeneity of *P. falciparum* (Khatoon et al., 2010). Most importantly, the use of highly potent antiplasmodial drugs that kill the asexual blood stage parasites and gametocytes are more likely to decrease parasite transmission and clonal diversity (Targett et al., 2001; Greenwood et al., 2008). However, in the adjacent districts (Bannu and Kohat) of Khyber Pakhtunkhwa, polyclonal infections were reported previously (Khatoon et al., 2010; Khatoon et al., 2012). It is worth mentioning that these two districts (Bannu and Kohat) accommodated one million internally displaced persons (IDPs) from tribal areas of North Waziristan Agency (NWA) sharing its border with Afghanistan, which may have introduced new genetically distinct variants of *P. falciparum*. It shows that the huge influx and migration of people between the study area and the neighboring countries like Iran and especially Afghanistan may introduce different alleles of *P. falciparum* into Khyber Pakhtunkhwa province of Pakistan.

Malaria transmission in Pakistan is markedly seasonal and prone to outbreaks, in particular geographical areas, especially Khyber Pakhtunkhwa, Baluchistan, and Sindh province (Yasinzai and Kakarsulemankhel, 2009). Pakistan is considered to be endemic for malaria, but the precise data on the genetic diversity of malaria in Pakistan are still lacking (Khan et al., 2005). As limitations, a small number of samples were amplified for *msp-1* and *msp-2*, and the use of nested PCR instead of DNA sequencing could possibly underestimate the genetic diversity. Therefore, studies with larger datasets and more robust techniques should be used to explore the genetic diversity in the future. Furthermore, only microscopy-positive samples for *P. falciparum* were further subjected to nested PCR, which is another limitation of the present study.

CONCLUSION

It was concluded that extensively diverse and polymorphic *P. falciparum* populations of merozoite surface protein 1 (*msp-1*) and 2 (*msp-2*) are present in Khyber Pakhtunkhwa, Pakistan.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the recommendation of the Institutional Ethical Review Committee of Kohat University of Science and Technology (KUST), Kohat-26000, Pakistan. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

SNK, IA, and SA designed the research study. IA and SA supervised the study. SNK collected and analyzed the data. SNK and RA drafted the manuscript. MR and SN helped during literature search, text incorporation, and table and graph designing. SK, SZ, and RA provided critical comments for the improvement of the manuscript. Finally, all the authors have read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.751552/full#supplementary-material

REFERENCES

A-Elbasit, I. E., ElGhazali, G., A-Elgadir, T. M. E., Hamad, A. A., Babiker, H. A., Elbashir, M. I., et al. (2007). Allelic Polymorphism of MSP2 Gene in Severe *P. Falciparum* Malaria in an Area of Low and Seasonal Transmission. *Parasitol. Res.* 102 (1), 29–34. doi:10.1007/s00436-007-0716-3

Aubouy, A., Migot-Nabias, F., and Deloron, P. (2003). Polymorphism in Two Merozoite Surface Proteins of *Plasmodium Falciparum* Isolates from Gabon. *Malar. J.* 2, 12. doi:10.1186/1475-2875-2-12

Babiker, H. A., Charlwood, J. D., Smith, T., and Walliker, D. (1995). Gene Flow and Cross-Mating in *Plasmodium Falciparum* in Households in a Tanzanian Village. *Parasitology* 111 (Pt 4), 433–442. doi:10.1017/s0031182000065938

Babiker, H. A., Hill, W. G., Lines, J., and Walliker, D. (1997). Population Structure of *Plasmodium Falciparum* in Villages with Different Malaria Endemicity in East Africa. *Am. J. Trop. Med. Hyg.* 56 (2), 141–147. doi:10.4269/ajtmh.1997.56.141

Cattamanchi, A., Hubbard, A., Kyabayinze, D., Dorsey, G., and Rosenthal, P. J. (2003). Distinguishing Recrudescence from Reinfection in a Longitudinal Antimalarial Drug Efficacy Study: Comparison of Results Based on Genotyping of Msp-1, Msp-2, and Glurp. *Am. J. Trop. Med. Hyg.* 68 (2), 133–139. doi:10.4269/ajtmh.2003.68.133

Collins, W. J., Greenhouse, B., Rosenthal, P. J., and Dorsey, G. (2006). The Use of Genotyping in Antimalarial Clinical Trials: a Systematic Review of Published Studies from 1995-2005. *Malar. J.* 5, 122. doi:10.1186/1475-2875-5-122

Contamin, H., Konate, L., Trape, J.-F., Mercereau-Puijalon, O., Rogier, C., Fandeur, T., et al. (1996). Different Genetic Characteristics of *Plasmodium Falciparum* Isolates Collected during Successive Clinical Malaria Episodes in Senegalese Children. *Am. J. Trop. Med. Hyg.* 54 (6), 632–643. doi:10.4269/ajtmh.1996.54.632

Diggs, C. L., Ballou, W. R., and Miller, L. H. (1993). The Major Merozoite Surface Protein as a Malaria Vaccine Target. *Parasitol. Today* 9 (8), 300–302. doi:10.1016/0169-4758(93)90130-8

Eltayeb, L. B., Abdelghani, S., Madani, M., and Waggiallah, H. A. (2020). Molecular Characterization of Msp-1 and Msp-2 Among *Plasmodium Falciparum*-Infected Children: A Gene Polymorphisms Analysis. *J. Biochem. Tech.* 11 (2), 102–107.

Färnert, A., Arez, A. P., Babiker, H. A., Beck, H. P., Benito, A., Björkman, A., et al. (2001). Genotyping of *Plasmodium Falciparum* Infections by PCR: a Comparative Multicentre Study. *Trans. R. Soc. Trop. Med. Hyg.* 95 (2), 225–232. doi:10.1016/s0035-9203(01)90175-0

Ferreira, M. U., da Silva Nunes, M., and Wunderlich, G. (2004). Antigenic Diversity and Immune Evasion by Malaria Parasites. *Clin. Vaccin. Immunol* 11 (6), 987–995. doi:10.1128/cdli.11.6.987-995.2004

Ghanchi, N. K., Mårtensson, A., Ursing, J., Jafri, S., Bereczky, S., Hussain, R., et al. (2010). Genetic Diversity Among *Plasmodium Falciparum* Field Isolates in Pakistan Measured with PCR Genotyping of the Merozoite Surface Protein 1 and 2. *Malar. J.* 9, 1. doi:10.1186/1475-2875-9-1

Gómez, D., Rojas, M. O., Rubiano, C., Chaparro, J., and Wasserman, M. (2002). Genetic Diversity of *Plasmodium Falciparum* Field Samples from an Isolated Colombian Village. *Am. J. Trop. Med. Hyg.* 67 (6), 611–616. doi:10.4269/ajtmh.2002.67.611

Greenhouse, B., Dorsey, G., Myrick, A., Carlson, E. J., Dokomajilar, C., Rosenthal, P. J., et al. (2006). Validation of Microsatellite Markers for Use in Genotyping Polyclonal *Plasmodium Falciparum* Infections. *Am. J. Trop. Med. Hyg.* 75 (5), 836–842. doi:10.4269/ajtmh.2006.75.836

Greenwood, B. M., Fidock, D. A., Kyle, D. E., Kappe, S. H. I., Alonso, P. L., Collins, F. H., et al. (2008). Malaria: Progress, Perils, and Prospects for Eradication. *J. Clin. Invest.* 118 (4), 1266–1276. doi:10.1172/jci33996

Haddad, D., Enamorado, I. G., Berzins, K., Snounou, G., Ståhl, S., Mattei, D., et al. (1999). Limited Genetic Diversity of *Plasmodium Falciparum* in Field Isolates from Honduras. *Am. J. Trop. Med. Hyg.* 60 (1), 30–34. doi:10.4269/ajtmh.1999.60.30

Heidari, A., Keshavarz, H., Rokni, M. B., and Jelinek, T. (2007). Genetic Diversity in Merozoite Surface Protein (MSP)-1 and MSP-2 Genes of *Plasmodium Falciparum* in a Major Endemic Region of Iran. *Korean J. Parasitol.* 45 (1), 59–63. doi:10.3347/kjp.2007.45.1.59

Howard, N., Durrani, N., Sanda, S., Beshir, K., Hallett, R., and Rowland, M. (2011). Clinical Trial of Extended-Dose Chloroquine for Treatment of Resistant *Falciparum* Malaria Among Afghan Refugees in Pakistan. *Malar. J.* 10, 171. doi:10.1186/1475-2875-10-171

Jordan, S., Jelinek, T., Aida, A. O., Peyerl-Hoffmann, G., Heuschkel, C., el Valy, A. O., et al. (2001). Population Structure of *Plasmodium Falciparum* Isolates during an Epidemic in Southern Mauritania. *Trop. Med. Int. Health* 6 (10), 761–766. doi:10.1046/j.1365-3156.2001.00802.x

Khan, M. A., Mekan, S. F., Abbas, Z., and Smego, R. A., Jr. (2005). Concurrent Malaria and Enteric Fever in Pakistan. *Singapore Med. J.* 46 (11), 635–638.

Khatoon, L., Baliraine, F. N., Bonizzoni, M., Malik, S. A., and Yan, G. (2010). Genetic Structure of *Plasmodium Vivax* and *Plasmodium Falciparum* in the Bannu District of Pakistan. *Malar. J.* 9 (1), 112. doi:10.1186/1475-2875-9-112

Khatoon, L., Khan, I. U., Shah, S. A., Jan, M. I., Ullah, F., and Malik, S. A. (2012). Genetic Diversity of *Plasmodium Vivax* and *Plasmodium Falciparum* in Kohat

District, Pakistan. *Braz. J. Infect. Dis.* 16 (2), 184–187. doi:10.1590/s1413-86702012000200014

Konaté, L., Zwetyenga, J., Rogier, C., Bischoff, E., Fontenille, D., Tall, A., et al. (1999). 5. Variation of *Plasmodium Falciparum* Msp1 Block 2 and Msp2 Allele Prevalence and of Infection Complexity in Two Neighbouring Senegalese Villages with Different Transmission Conditions. *Trans. R. Soc. Trop. Med. Hyg.* 93 (Suppl. ment_1), 21–28. doi:10.1016/s0035-9203(99)90323-1

Mohammed, H., Hassen, K., Assefa, A., Mekete, K., Tadesse, G., Taye, G., et al. (2019). Genetic Diversity of *Plasmodium Falciparum* Isolates from Patients with Uncomplicated and Severe Malaria Based on Msp-1 and Msp-2 Genes in Gublak, North West Ethiopia. *Malar. J.* 18 (1), 413. doi:10.1186/s12936-019-3039-9

Montoya, L., Maestre, A., Carmona, J., Lopes, D., Do Rosario, V., and Blair, S. (2003). *Plasmodium Falciparum*: Diversity Studies of Isolates from Two Colombian Regions with Different Endemicity. *Exp. Parasitol.* 104 (1-2), 14–19. doi:10.1016/s0014-4894(03)00112-7

Murtaza, G., Memon, I. A., and Noorani, A. K. (2004). Academic Works and Political Activities. *Med. Channel* 10 (2), 41–95. doi:10.4324/9780203335239-7

Ndiaye, T., Sy, M., Gaye, A., and Ndiaye, D. (2019). Genetic Polymorphism of Merozoite Surface Protein 1 (Msp1) and 2 (Msp2) Genes and Multiplicity of *Plasmodium Falciparum* Infection across Various Endemic Areas in Senegal. *Afr. H. Sci.* 19 (3), 2446–2456. doi:10.4314/ahs.v19i3.19

Nei, M. (1978). Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. *Genetics* 89 (3), 583–590. doi:10.1093/genetics/89.3.583

Nizamani, A., Kalar, N., and Khushk, I. (2006). Burden of Malaria in Sindh,Pakistan: A Two Years Surveillance Report. *J. Liaquat Uni Med. Health Sci.* 05, 76–83. doi:10.22442/jlumhs.06520092

Papa Mze, N., Bogreau, H., Diedhiou, C. K., Herdell, V., Rahamatou, S., Bei, A. K., et al. (2020). Genetic Diversity of *Plasmodium Falciparum* in Grande Comore Island. *Malar. J.* 19 (320). doi:10.1186/s12936-020-03384-5

Paul, R. E., Nosten, F., White, N. J., Luxemburger, C., Price, R., Hackford, I., et al. (1998). Transmission Intensity and *Plasmodium Falciparum* Diversity on the Northwestern Border of Thailand. *Am. J. Trop. Med. Hyg.* 58 (2), 195–203. doi:10.4269/ajtmh.1998.58.195

Peyerl-Hoffmann, G., Jelinek, T., Kilian, A., Kabagambe, G., Metzger, W. G., and von Sonnenburg, F. (2001). Genetic Diversity of *Plasmodium Falciparum* and its Relationship to Parasite Density in an Area with Different Malaria Endemicities in West Uganda. *Trop. Med. Int. Health* 6 (8), 607–613. doi:10.1046/j.1365-3156.2001.00761.x

Phillips, R. S. (2001). Current Status of Malaria and Potential for Control. *Clin. Microbiol. Rev.* 14 (1), 208–226. doi:10.1128/cmr.14.1.208-226.2001

Qureshi, N. A., Fatima, H., Afzal, M., Khattak, A. A., and Nawaz, M. A. (2019). Occurrence and Seasonal Variation of Human *Plasmodium* Infection in Punjab Province, Pakistan. *BMC Infect. Dis.* 19 (1), 935. doi:10.1186/s12879-019-4590-2

Roh, M. E., Tessema, S. K., Murphy, M., Nhlabathi, N., Mkhonta, N., Vilakati, S., et al. (2019). High Genetic Diversity of *Plasmodium Falciparum* in the Low-Transmission Setting of the Kingdom of Eswatini. *J. Infect. Dis.* 220 (8), 1346–1354. doi:10.1093/infdis/jiz305

Sallenave-Sales, S., Daubersies, P., Mercereau-Puijalon, O., Rahimalala, L., Contamin, H., Druilhe, P., et al. (2000). *Plasmodium Falciparum*: a Comparative Analysis of the Genetic Diversity in Malaria-Mesoendemic Areas of Brazil and Madagascar. *Parasitol. Res.* 86 (8), 692–698. doi:10.1007/pl00008554

Sheikh, A. S., Sheikh, A. A., Sheikh, N. S., and Paracha, S. M. (2005). Endemicity of Malaria in Quetta. *Pak J. Med. Res.* 44 (1), 41–50.

Singh, B., Snounou, G., Abdullah, M. S., Rahman, H. A., Bobogare, A., and Cox-Singh, J. (1999). A Genus- and Species-specific Nested Polymerase Chain Reaction Malaria Detection Assay for Epidemiologic Studies. *Am. J. Trop. Med. Hyg.* 60 (4), 687–692. doi:10.4269/ajtmh.1999.60.687

Smythe, J. A., Coppel, R. L., Day, K. P., Martin, R. K., Oduola, A. M., Kemp, D. J., et al. (1991). Structural Diversity in the *Plasmodium Falciparum* Merozoite Surface Antigen 2. *Proc. Natl. Acad. Sci.* 88 (5), 1751–1755. doi:10.1073/pnas.88.5.1751

Snounou, G., and Beck, H.-P. (1998). The Use of PCR Genotyping in the Assessment of Recrudescence or Reinfection after Antimalarial Drug Treatment. *Parasitol. Today* 14 (11), 462–467. doi:10.1016/s0169-4758(98)01340-4

Snounou, G., Viriyakosol, S., Xin Ping Zhu, Z. P., Jarra, W., Pinheiro, L., do Rosario, V. E., et al. (1993). High Sensitivity of Detection of Human Malaria Parasites by the Use of Nested Polymerase Chain Reaction. *Mol. Biochem. Parasitol.* 61, 315–320. doi:10.1016/0166-6851(93)90077-b

Snounou, G., Zhu, X., Siripoon, N., Jarra, W., Thaithong, S., Brown, K. N., et al. (1999). Biased Distribution of Msp1 and Msp2 Allelic Variants in *Plasmodium Falciparum* Populations in Thailand. *Trans. R. Soc. Trop. Med. Hyg.* 93 (4), 369–374. doi:10.1016/s0035-9203(99)90120-7

Soe, T. N., Wu, Y., Tun, M. W., Xu, X., Hu, Y., Ruan, Y., et al. (2017). Genetic Diversity of *Plasmodium Falciparum* Populations in Southeast and Western Myanmar. *Parasites Vectors* 10 (1), 322. doi:10.1186/s13071-017-2254-x

Somé, A. F., Bazié, T., Zongo, I., Yerbanga, R. S., Nikiéma, F., Neya, C., et al. (2018). *Plasmodium Falciparum* Msp1 and Msp2 Genetic Diversity and Allele Frequencies in Parasites Isolated from Symptomatic Malaria Patients in Bobo-Dioulasso, Burkina Faso. *Parasites Vectors* 11 (1), 323. doi:10.1186/s13071-018-2895-4

Soulama, I., Nébié, I., Ouédraogo, A., Gansane, A., Diarra, A., Tiono, A. B., et al. (2009). *Plasmodium Falciparum* Genotypes Diversity in Symptomatic Malaria of Children Living in an Urban and a Rural Setting in Burkina Faso. *Malar. J.* 8, 135. doi:10.1186/1475-2875-8-135

Targett, G., Drakeley, C., Jawara, M., von Seidlein, L., Coleman, R., Deen, J., et al. (2001). Artesunate Reduces but Does Not Prevent Posttreatment Transmission of *Plasmodium Falciparum* to Anopheles Gambiae. *J. Infect. Dis.* 183 (8), 1254–1259. doi:10.1086/319689

Taylor, R. R., Smith, D. B., Robinson, V. J., McBride, J. S., and Riley, E. M. (1995). Human Antibody Response to *Plasmodium Falciparum* Merozoite Surface Protein 2 Is Serogroup Specific and Predominantly of the Immunoglobulin G3 Subclass. *Infect. Immun.* 63 (11), 4382–4388. doi:10.1128/iai.63.11.4382-4388.1995

Who (2017). World malaria report. World Health Organization, Available at: https://www.who.int/publications-detail-redirect/9789241565523.

Who (2018). World malaria report. World Health Organization, Available at: https://apps.who.int/iris/handle/10665/275867.

Yasinzai, M. I., and Kakarsulemankhel, J. K. (2009). Prevalence of Human Malaria Infection in Bordering Areas of East Balochistan, Adjoining with Punjab: Loralai and Musakhel. *J. Pak Med. Assoc.* 59 (3), 132–135.

Yavo, W., Konaté, A., Mawili-Mboumba, D. P., Kassi, F. K., Tshibola Mbuyi, M. L., Angora, E. K., et al. (2016). Genetic Polymorphism of msp1 and msp2 in *Plasmodium falciparum* isolates from Côte d'Ivoire versus Gabon. *J. Parasitol. Res.* 2016, 1–7. doi:10.1155/2016/3074803

Zakeri, S., Bereczky, S., Naimi, P., Pedro Gil, J., Djadid, N. D., Färnert, A., et al. (2005). Multiple Genotypes of the Merozoite Surface Proteins 1 and 2 in *Plasmodium Falciparum* Infections in a Hypoendemic Area in Iran. *Trop. Med. Int. Health* 10 (10), 1060–1064. doi:10.1111/j.1365-3156.2005.01477.x

Zakeri, S., Raeisi, A., Afsharpad, M., Kakar, Q., Ghasemi, F., Atta, H., et al. (2010). Molecular Characterization of *Plasmodium Vivax* Clinical Isolates in Pakistan and Iran Using Pvmsp-1, Pvmsp-3α and Pvcsp Genes as Molecular Markers. *Parasitol. Int.* 59 (1), 15–21. doi:10.1016/j.parint.2009.06.006

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 November 2021
doi: 10.3389/fgene.2021.742808

OPEN ACCESS

Edited by:
Moses Okpeku, University of KwaZulu-Natal, South Africa

Reviewed by:
Liang Wang, Institut Pasteur of Shanghai (CAS), China
Aleksandr G. Bulaev, Federal Center Research Fundamentals of Biotechnology (RAS), Russia

***Correspondence:**
Diaeldin A. Salih
diaeldin2000@hotmail.com

†Present address:
Moses Njahira, International Centre of Insect Physiology and Ecology (icipe), Nairobi, Kenya
Joram M. Mwacharo, International Centre for Agricultural Research in the Dry Areas (ICARDA), Addis Ababa, Ethiopia
Richard Bishop, Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, WA, United States
Robert Skilton, International Centre of Insect Physiology and Ecology (icipe), Nairobi, Kenya

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: *16 July 2021*
Accepted: *25 October 2021*
Published: *19 November 2021*

Citation:
Salih DA, Ali AM, Njahira M, Taha KM, Mohammed MS, Mwacharo JM, Mbole-Kariuki N, El Hussein AM, Bishop R and Skilton R (2021) Population Genetic Analysis and Sub-Structuring of Theileria annulata *in Sudan. Front. Genet. 12:742808. doi: 10.3389/fgene.2021.742808*

Population Genetic Analysis and Sub-Structuring of *Theileria annulata* in Sudan

Diaeldin A. Salih [1,2]*, ***Awadia M. Ali*** [3], ***Moses Njahira*** [1†], ***Khalid M. Taha*** [4], ***Mohammed S. Mohammed*** [5], ***Joram M. Mwacharo*** [6†], ***Ndila Mbole-Kariuki*** [7], ***Abdelrhim M. El Hussein*** [8], ***Richard Bishop*** [7†] ***and Robert Skilton*** [1†]

[1]*Biosciences Eastern and Central Africa-International Livestock Research Institute Hub (BecA-ILRI Hub), Nairobi, Kenya,* [2]*Central Veterinary Research Laboratory, Khartoum, Sudan,* [3]*Faculty of Veterinary Medicine, University of Khartoum, Khartoum, Sudan,* [4]*Atbara Veterinary Research Laboratory, Atbara, Sudan,* [5]*Faculty of Veterinary Medicine, University of Al-Butana, Tamboul, Sudan,* [6]*School of Life Sciences, Centre for Genetics and Genomics, University of Nottingham, Nottingham, United Kingdom,* [7]*International Livestock Research Institute, Nairobi, Kenya,* [8]*Central Laboratory, Khartoum, Sudan*

Theileria annulata, which causes tropical theileriosis, is a major impediment to improving cattle production in Sudan. Tropical theileriosis disease is prevalent in the north and central regions of Sudan. Outbreaks of the disease have been observed outside the known endemic areas, in east and west regions of the country, due to changes in tick vector distribution and animal movement. A live schizont attenuated vaccination based on tissue culture technology has been developed to control the disease. The parasite in the field as well as the vaccine strain need to be genotyped before the vaccinations are practiced, in order to be able to monitor any breakthrough or breakdown, if any, after the deployment of the vaccine in the field. Nine microsatellite markers were used to genotype 246 field samples positive for *T. annulata* DNA and the vaccine strain. North and central populations have a higher multiplicity of infection than east and west populations. The examination of principal components showed two sub-structures with a mix of all four populations in both clusters and the vaccine strain used being aligned with left-lower cluster. Only the north population was in linkage equilibrium, while the other populations were in linkage disequilibrium, and linkage equilibrium was found when all samples were regarded as single population. The genetic identity of the vaccine and field samples was 0.62 with the north population and 0.39 with west population. Overall, genetic investigations of four *T. annulata* populations in Sudan revealed substantial intermixing, with only two groups exhibiting regional origin independence. In the four geographically distant regions analyzed, there was a high level of genetic variation within each population. The findings show that the live schizont attenuated vaccine, Atbara strain may be acceptable for use in all Sudanese regions where tropical theileriosis occurs.

Keywords: cattle, cell culture vaccine, *Theileria annulata*, genotyping, population genetics, sub-structure, Sudan

INTRODUCTION

Tropical theileriosis is a tick-borne disease, caused by *Theileria annulata* that continues to be a major concern for livestock in tropical countries affecting millions of animals, particularly crossbreed and exotic cattle, and resulting in significant economic loss and mortality (Dolan, 1989). *Hyalomma anatolicum* transmits *T. annulata* sporozoites causing a lymphoproliferative disease (Ghosh and Azhahianambi, 2007). The sporozoites develop into schizonts which reside inside the host's lymphocyte and macrophages (Sager et al., 1998). The schizont stage is the only symptomatic stage among different parasite stages in the host, on which attenuated schizont vaccine was designed (Pipano and Shkap, 2006). Many countries, including, Israel (Pipano et al., 1981), Iran (Hashemi-Fesharki, 1988), Russia (Stepanova and Zablotskii, 1989), India (Beniwal et al., 1997), China (Zhang, 1997), Turkey (Sayin et al., 1997), Spain (Viseras et al., 1997), Morocco (Ouhelli et al., 1997), Tunisia (Darghouth, 2008), and India (Roy et al., 2019), have employed the attenuated schizont vaccines to control *T. annulata* infection.

The importance of genetic diversity research in providing information on protozoan parasites, such as epidemiology, control, evolution, virulence, antigenicity, infectivity, treatment sensitivity, and host preference, has been demonstrated (Weir, et al., 2011; Sivakumar et al., 2014). *T. annulata* from other endemic regions, such as China, Oman, Turkey, Tunisia, and Portugal have all been researched utilizing a multilocus genotyping technique for assessing genetic diversity, population structure, and transmission patterns (Weir et al., 2007, 2011; Al-Hamidhi et al., 2015; Gomes et al., 2016; Yin et al., 2018; Roy et al., 2021). *T. annulata* genetic populations studies were notable for its genetic variation, the availability of many genotypes per sample, and sub-structuring by geography (Weir et al., 2007, 2011; Al-Hamidhi et al., 2015; Gomes et al., 2016; Yin et al., 2018; Roy et al., 2019).

Microsatellite-based genotyping was utilized in this work to better understand genetic diversity, population structure, and geographical sub structuring of the *T. annulata* vaccine and parasite samples collected from four different regions in Sudan. The diversity of the markers used would reflect the genetic makeup of the samples as well as the vaccine genetic makeup. The recognition of the vaccine strain would be as fast and efficient if the genetic makeup of the vaccine is the same as the field strain. The findings offer the first glimpse of the *T. annulata* parasites population genetics and diversity in Sudan.

MATERIALS AND METHODS

Cattle Blood Samples

A total of 530 blood samples were collected from cattle in four Sudanese regions using FTA™ cards (Whatman Biosciences, United Kingdom). The four regions were north (n = 69), central (n = 195), east (n = 158), and west (n = 108). North and central regions were designated endemic region, while east and west were designated new extension regions. Information on the sampling locations, whether from endemic or new extension regions, total number examined and *T. annulata* positive samples by PCR are provided in **Table 1**.

Extraction of DNA and Small Subunit "SSU" rRNA PCR

Extraction of DNA from cattle blood samples and Atbara vaccine strain was carried out using the PureLink™ Genomic DNA Mini extraction kit (Invitrogen, Germany). For diagnosis of *T. annulata*, the primer used was SSU rRNA gene 989 5′AGT TTCTGACCTATCAG3′ and the reverse primer was 1,347 5′TGCACAGACCCCAGAG G 3′ giving an amplicon of 370 bp (Allsopp et al., 1993; Taha et al., 2013).

Microsatellite PCR Assay

In this study, the primers used were designed by Weir et al. (2007). For detection in capillary electrophoresis, the forward primer was labeled with standard labeling dyes at the 5′ end (**Supplementary Table S1**). The PCR amplification was carried out as described in Salih et al. (2018). For negative control, the nuclease free water was used, while DNA extracted from a schizont-infected lymphocyte culture derived from *T. annulata* Ankara strain was used as a positive control.

Capillary Electrophoresis and Genotyping

The ABI 3730 Genetic Analyzer (Applied Biosystems-USA) was used to analyze the PCR amplicons at the BecA-ILRI Hub, SegoliP sequencing unit, Nairobi, Kenya. For size fractionation, the Gene Scan 500 LIZ internal lane size standard (Applied Biosystems-USA) was employed. The Gene Mapper tool (Applied Biosystems-USA) was used to score the results, which allowed for the resolution of 1 base pair (bp) changes with many products from a single PCR reaction. The predominant allele was determined as the one with the biggest area under the curve, and amplicons with highest peak height were scored. Allelobin software (Idury and Cardon, 1997) was used to re-sized all Gene Mapper data based on consensus sequence repeats of each marker (**Table 2**). The inaugural form of file, designated multi locus genotype (MLG) consisted of genotypes created from only the predominant allele at each locus (Weir et al., 2007). The allelic profile dataset, on the other hand, contained genotypic profiles derived from all alleles observed at each locus (where minor peaks were greater than 33% the height of the predominant allele present). The MLG file was used to determine population genetic diversity and structure, while the allelic profile file was utilized to calculate the multiplicity of infection (MOI), as well as to rule out linkage disequilibrium as a null hypothesis.

Analyses of Population Genetic

Arlequin v. 3.5 http://cmpg.unibe.ch/software/arlequin 35/ (Excoffier, and Lischer, 2010) was used to calculate the expected heterozygosity, as *Theileria* is haploid and heterozygosity cannot be observed directly. To investigate the genetic relationships between population, principal component analysis (PCA) was calculated in GenAlEx6.5 (Peakall and Smouse, 2006; 2012). Analysis of molecular variance (AMOVA) was performed using ARLEQUIN to test for hierarchical population structure. Nei's genetic distance (D) (Nei, 1978) was calculate between each group of samples from

TABLE 1 | Information on sample numbers and location.

Status of the disease	Region	Town	Total number examined	*T. annulata* positive (PCR)
Endemic areas	North Sudan	Atbara field samples	25	14
		Research station	44	22
	Sub total		69	36
	Central Sudan	Khartoum	100	59
		Omderman	79	56
		Madani	4	1
		Singa	9	2
		Kuku	3	2
Sub total			195	120
Grand total			264	156
New extended areas	East Sudan	Kassala	25	22
		Halfa	133	25
	Sub total		158	47
	West Sudan	Nyala	5	4
		Nihod	7	2
		Fashir	6	0
		Obied	90	37
	Sub total		108	43
Grand total			266	90
Grand total			530	246

Out group controls: Theileria annulata *Ankara strain (Turkey), Tissue culture vaccine strain.*

TABLE 2 | Major allele frequency, gene diversity, number of alleles and polymorphic information content (PIC) of the nine microsatellite markers used in this study.

Marker	Major allele frequency	Number of allele	Gene diversity (*He*)	PIC
TS5	0.62	10.00	0.58	0.55
TS6	0.46	14.00	0.75	0.73
TS8	0.29	22.00	0.87	0.87
TS9	0.65	4.00	0.46	0.36
TS12	0.48	9.00	0.62	0.54
TS15	0.34	10.00	0.80	0.78
TS20	0.49	18.00	0.67	0.62
TS25	0.25	10.00	0.84	0.82
TS31	0.54	15.00	0.64	0.65
Mean	0.46	12.44	0.70	0.66

different populations and the vaccine strain using the genetic data analysis tool (GDA) (http://lewis.eeb.uconn.edu/lewishome/gda.html).

The standardized index of association (I_A^S) between each group of samples was estimated using the LIAN 3.7 program, as well as, the degree of linkage disequilibrium (LD) within and between populations (Haubold and Hudson, 2000). After each population was studied separately, the samples were pooled and processed as a single dataset.

STRUCTURE 2.3.4 (http://pritchardlab.stanford.edu/structure.html) was used to investigate population structure employing Bayesian clustering analysis with sample sites as a basis and the admixture scenario with linked allele polymorphism (Pritchard et al., 2000; Evanno et al., 2005). Initial runs of one million steps were used to investigate the datasets (burn-in of 20%). For every value of *K* scale from one (considering all are *T. annulata*) to five (assuming all the five populations are genetically distinct), triplicates were performed. To identify which *K* produced the greatest representation of the data, STRUCTUREHARVESTER 0.6.1 (Earl and von Holdt, 2012) was employed. CLUMPP 1.1 (Jakobsson and Rosenberg, 2007) and DISTRUCT 1.1.2 (Rosenberg, 2004) were used to parse and format the data in order to assess the STRUCTURE output. CLUMPP 1.1 aligns cluster assignments across duplicate analyses, while DISTRUCT 1.1.2 assists with visual representation.

Multiplicity of Infection

MOI was considered as "existence of numerous genotypes per isolate" when more than one allele was detected at a locus and the smaller peaks were exceed 33% of the height of the predominant allele expressed (Weir et al., 2011; Salih et al., 2018). The mean number of alleles across all nine loci was determined for every sample, and this number was used to indicate the multiplicity of

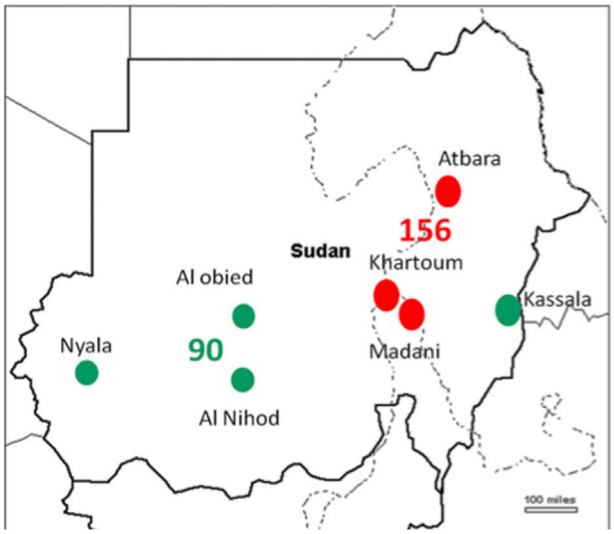


FIGURE 1 | The regions of Sudan where samples were obtained are depicted on the map, the red dots indicates endemic region, while the green ones indicates new extended regions.

infection in that sample. The multiplicity of infection for each population was calculated by taking the overall mean for each sample's index value.

RESULTS

Verification of Positive Samples for *T. annulata* DNA

T. annulata DNA was tested in 530 cattle blood samples. The SSU rRNA PCR assay verified 246 (46.4%) samples positive for *T. annulata* DNA which were subjected to genotyping in addition to the vaccine strain (**Table 1**). Distribution of the positive samples were as follow, endemic regions $n = 156$ (North $n = 36$, Central $n = 120$) and new extension regions $n = 90$ (East $n = 47$, West $n = 43$) (**Figure 1**).

Satellite Marker Diversity

In all of the samples, each marker was highly polymorphic. The polymorphic information content (PIC) of marker TS8 had the highest (0.87), whereas TS9 had the lowest (0.36) (**Table 2**). This finding argued in favor that these markers could be effective in determining linkage disequilibrium analysis in *T. annulata* populations. The existence of more than one allele at one or more loci confirmed the presence of several genotypes in the samples. For each marker, the number of alleles identified varied from four in TS9 to 22 for TS8 with the mean of 12.44 per marker (**Table 2**). The dominant allele frequencies varied from 0.25 (TS25) to 0.65 (TS9), with an average of 0.46 (**Table 2**).

Population Diversity and Structure

Principal components analysis (PCA) revealed that there is no clustering according to geographical origin (**Figure 2**). Two sub-structures with a mix of all four populations in both clusters and the vaccine stain being aligned with left-lower cluster were demonstrated, indicating that the parasite populations are rather distinct, with considerable genetic mixing and gene flow between parasites in the four distinct geographical populations investigated.

The allelic profile data set was examined to see if the *T. annulata* populations observed in Sudan were in linkage equilibrium or disequilibrium. When all the four sub-populations were analysed together (as a single population), the (I_A^S) was positive and greater than zero and the pairwise variance (V_D) was more than the 95% critical value (L) suggesting

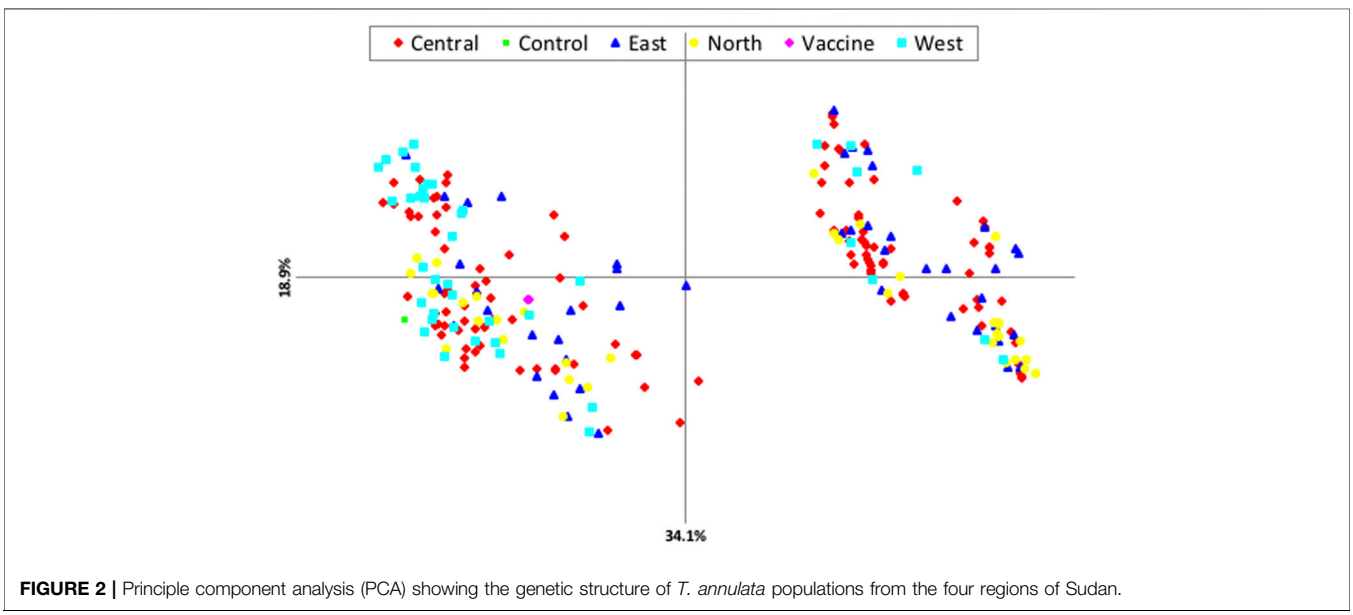


FIGURE 2 | Principle component analysis (PCA) showing the genetic structure of *T. annulata* populations from the four regions of Sudan.

TABLE 3 | Linkage equilibrium analyses in Sudanese population of *Theileria annulata*.

Population	I_A^S	V_D	L_{para}	L_{MC}	Linkage
North	0.0311	2.0792	1.9275	1.9992	LE
Central	0.0055	1.9965	2.1102	2.0798	LD
East	−0.0049	1.7881	2.3669	2.5958	LD
West	0.0037	1.9048	2.7354	2.7937	LD
All population	0.0174	2.1006	1.9591	1.9760	LE

I_A^S, standard index of association; V_D, mismatch variance (linkage analysis); LD, linkage disequilibrium; LE = linkage equilibrium; L_{MC} and L_{para}, upper 95% confidence limits of Monte Carlo simulation and parametric tests respectively (linkage analysis).

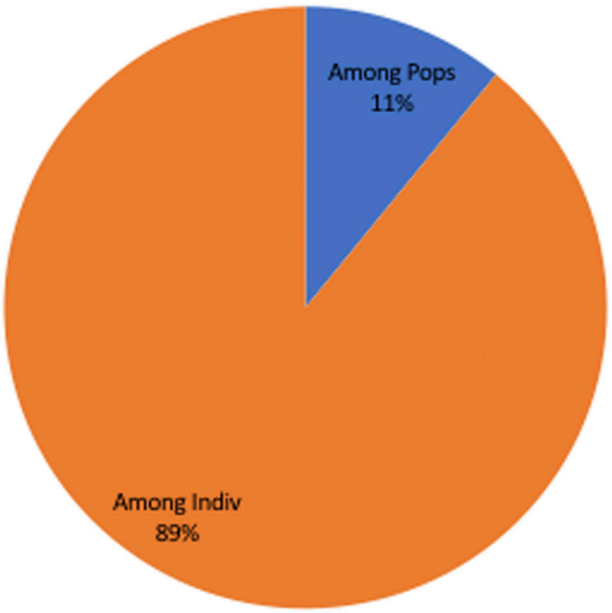


FIGURE 3 | Analysis of Molecular Variation (AMOVA) showing only 11% of the genetic variation was explained by populations differences, despite the fact that population differences account for the majority of genetic variety (89%) observed within populations.

TABLE 4 | The Nei genetic destance between the four populations studied and the vaccine strain.

	Central	East	North	Vaccine	West
Central	1.00	—	—	—	—
East	0.82	1.00	—	—	—
North	0.85	0.83	1.00	—	—
Vaccine	0.51	0.53	0.62	1.00	—
West	0.81	0.64	0.70	0.39	1.00

that the merged populations are in linkage equilibrium (LE) (**Table 3**). The analysis was performed for each population individually to assess for geographic sub-structuring, and three of the populations central, east and west, were shown to be in linkage disequilibrium (LD) (**Table 3**). Only 11% of the genetic variation was explained by variations between populations, which account for a considerable portion of the genetic diversity (89%) detected within populations (**Figure 3**).

Estimating Nei's genetic distance (*D*) between each of the four regionally sampled populations as well as between them and the vaccine strain, was used to evaluate genetic differentiation between the four populations (**Table 4**). The genetic differentiation between central and east populations ($D = 0.82$) was greater than that observed between the east and west populations ($D = 0.64$). The population with the lowest genetic distance from the vaccine genotype was west ($D = 0.39$), while the most genetically similar was north ($D = 0.62$) (**Table 4**).

Based on the Evanno et al. delta K technique, the STRUCTURE results imply that $K = 3$ is the optimal number of genetic groups to define the genotypes of Sudanese *T. annulata* populations as well as in *T. annulata* vaccine strain (**Figure 4**). The three clusters are designated as gene pool 1, 2 and 3 respectively. Gene pool 1 (purple colour) prevailed in central and east, while gene pool 2 (blue colour) were most prevalent in north and vaccine, and pool 3 (yellow colour) predominated in west (**Figure 4**). In the vaccine strain, gene pool 2 appears to be more common than gene pool 1.

Multiplicity of Infection

Multiple genotypes were observed in *T. annulata* populations from the four geographic regions, with multiple alleles being found at one or more loci. The mean number of alleles for the nine loci was determined for each sample, to obtain an index value that denoted multiplicity of infection. **Table 5** summarizes the multiplicity of infection for each population and across all four populations analyzed. North and central populations had high mean of values of 1.75 and 1.59, respectively, while east and west had values of 1.33 and 1.03, respectively.

DISCUSSION

Using microsatellite markers, this study investigated the diversity and population structure of *T. annulata* in Sudan. The study's samples ($n = 246$) were obtained from four different geographical regions. North and central regions known to be endemic of *T. annulata* since the eighties, and the remaining two (west and east) witnessed the spreading of the disease in the nineties. In addition, the *T. annulata* vaccine from Sudan was also included in the study. In order to gain insight into the epidemiology of a parasite, ascertain sources of infection and modes of transmission, it is critical to assess population, genetic diversity and structure (Weir et al., 2007, 2011; Salih et al., 2018).

The genetic diversity and population structure of *T. annulata* found in Sudan were studied using a panel of nine microsatellite markers. The highest mean genetic diversity was observed in north, a finding which could be due to significant tick infestations in this region, where the disease has been established for long time (El Hussein, et al., 2012; Gharbi et al., 2020). The lower degree of *T. annulata* diversity detected in the parasite population from the west corresponded to the recent reported of tropical theileriosis (Mohammed-Ahmed et al., 2020). In other countries where

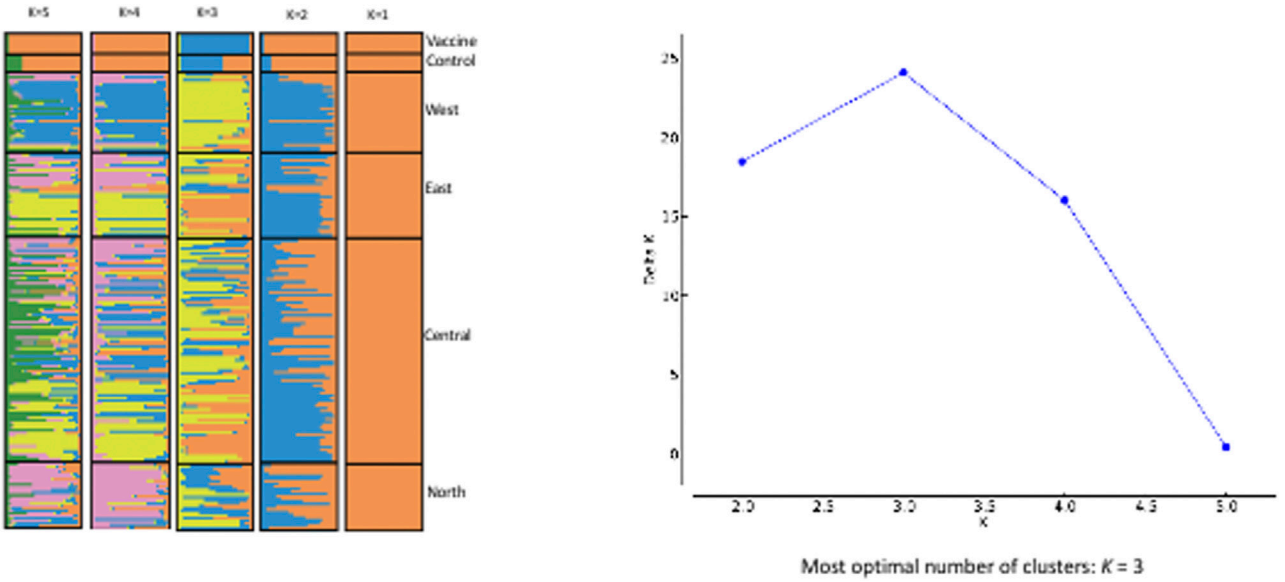


FIGURE 4 | EVANNO Method Delta and STRUCTURE. The graph shows optimal number of clusters from the STRUCTURE analysis; STRUCTURE analysis from K = 2 to K = 5 with samples from the four regions of Sudan.

TABLE 5 | Multiplicity of infection in Sudanese *Theileria annulata* population.

Population	*n*	Multiplicity of infection			
		Mean	SD	Min	Max
North	36	1.75	0.90	0.56	3.22
Central	120	1.59	0.83	0.69	2.91
East	47	1.33	0.88	0.45	2.89
West	43	1.03	0.75	0.17	2.43
All	246	1.43	0.84	0.47	2.86

n, *number of samples and SD, standard deviation.*

tropical theileriosis is endemic, a comparable scale of genetic variation has been observed among *T. annulata* populations (Weir et al., 2011; Al-Hamidhi et al., 2015; Gomes et al., 2016; Yin et al., 2018; Roy et al., 2021).

The results revealed relatively slight geographical sub-structuring among the four populations of *T. annulata* in Sudan with no evidence of grouping based on geographical origin. The fact that resources (feeds and water) are collectively utilized under the nomadic cattle systems prominent in Sudan is essential to enhance genetic uniformity. This result is supported by PCA analysis as well as STRUCRURE results. AMOVA revealed a high percentage of crossing between various *T. annulata* samples as well as recombination within the parasite population. Individual samples, rather than groups derived from a specific geographic region, accounted for the majority of genetic variation. In the future, other aspects such as parasite challenge and quantifying the extent of tick infestation should be examined. PCA and AMOVA results figured out no evident link between population genetic structure and the geographical origin of the isolates investigated. However, PCA analysis revealed a close genetic link between the north and *T. annulata* vaccine genotypes, with the *T. annulata* vaccine and majority of north genotypes clustered together.

When the PCA and STRUCTURE data are combined, it can be expected that there are three potential populations of *T. annulata* in Sudan. It's possible that gene pool 1 is introgressing into gene pool 2 or vice versa, with the two gene pools will eventually merging into one. This conclusion could be a result of cattle migration being unfettered across the country, due to the lack of trade barriers and policies restricting livestock movement (Oura et al., 2005; Roy et al., 2019). The mobility of parasite-infected/tick-infested cattle from one region to another assists in population homogenization.

The extent of linkage equilibrium between alleles at pairs of loci was evaluated, to see if the *T. annulata* populations in the four regions of Sudan constituted a single panmictic population with a high degree of genetic exchange. When the samples from the four regions were analyzed as a single population, an I_A^S value of 0.0174 was obtained as well as a V_D value (2.1006) that was greater than L (1.9591), demonstrating LE. The presence of LE in the combined populations could be due to an epidemic population structure (Smith et al., 1993), or it could be due to occasional genetic exchange, resulting in a clonal population structure (Wier et al., 2011). Other factors that could contribute to the reported LE include inbreeding, recombination rate and the size of the regional parasite functional population (Charlesworth, 2009). More samples from Sudan are needed to clarify which characteristics are most essential, especially because a limited number of genetically identical parasites in the vertebrate host could result in substantial linkage disequilibrium (Anderson et al., 2000).

The highest level of multiplicity of infection (MOI) was identified in the north, with a highest value of 3.22, followed by east, with lowest and maximum values of 0.45 and 2.89, respectively showing a

significant degree of variability in the dataset. In the midgut of the tick vector, multiple infections stimulate cross-mating and recombination among distinct parasite genotypes, as well as the formation of unique recombinant genotypes (Weir et al., 2001; Al-Hamidhi et al., 2015; Salih et al., 2018). The higher number of *T. annulata* genotypes in north could enable a high rate of cross-mating and recombination, resulting in increased genetic diversity in the bovine host (Conway et al., 1999). It could be also explained by the high tick load reported in north compared to the other regions (Salih et al., 2004).

In conclusion, the application of polymorphic microsatellite loci has offered preliminary insight into the population genetic diversity and structure of *T. annulata* population in Sudan. Extensive genetic intermixing between the four *T. annulata* populations studied was indicated, as well as minimal evidence of genetic differentiation and a high level of genetic diversity within each population. The findings show that the vaccine (Atbara strain) could be used in all areas where tropical theileriosis present.

T. annulata populations found in north African countries where tropical theileriosis is currently an economically important disease, should be examined and compared to see how genetically similar they are. Such data can assist veterinary control policy makers in determined if preventative measures, such as immunization, should be deployed at the national, regional or continental level.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The animal study was reviewed and approved by ILRI IACUC (The ILRI's Institutional Animal Care and Use Committee). Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

DS study design, conceptualization, data curation, data analysis, visualization, writing original draft, AA helped in sample collection, data curation MN Methodology, KT helped in sample collection, review manuscript, MM helped in sample collection, JM data analysis, review manuscript NM-K data analysis, AE conceptualization, supervision, manuscript review and editing, RB conceptualization, supervision, RS conceptualization, supervision, funding acquisition. All authors have read and approved the final version of the manuscript.

FUNDING

We gratefully acknowledge the financial support provided to the Biosciences eastern and central Africa Hub at the International Livestock Research Institute (BecA-ILRI Hub, Nairobi) by the Australian Agency for International Development (AusAID) through a partnership between Australia's Commonwealth Scientific and Industrial Research Organization (CSIRO) and the BecA-ILRI Hub; and by the Syngenta Foundation for Sustainable Agriculture (SFSA); the Bill and Melinda Gates Foundation (BMGF); and the Swedish Ministry of Foreign Affairs through the Swedish International Development Agency (Sida), which made this work possible. DAS was a recipient of an Africa Biosciences Challenge Fund (ABCF) Fellowship. The work was also supported in part by a grant from the DFG (Germany-African Cooperation; "Molecular epidemiology network for promotion and support of delivery of live vaccines against *T. parva* and *T. annulata* infection in Eastern and Northern Africa" (SE862/2-1).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.742808/full#supplementary-material

Supplementary Table 1 | Primers information used.

REFERENCES

Al-Hamidhi, S., H. Tageldin, M., Weir, W., Al-Fahdi, A., Johnson, E. H., Bobade, P., et al. (2015). Genetic Diversity and Population Structure of Theileria Annulata in Oman. *PLoS One* 10, e0139581. doi:10.1371/journal.pone.0139581

Allsopp, B. A., Baylis, H. A., Allsoppi, M. T. E. P., Cavalier-smith, T., Bishop, R. P., Carrington, D. M., et al. (1993). Discrimination between Six Species of *Theileria* Using Oligonucleotide Probes Which Detect Small Subunit Ribosomal RNA Sequences. *Parasitology* 107, 157–165. doi:10.1017/s0031182000067263

Anderson, T. J. C., Haubold, B., Williams, J. T., Estrada-Franco§, J. G., Richardson, L., Mollinedo, R., et al. (2000). Microsatellite Markers Reveal a Spectrum of Population Structures in the Malaria Parasite *Plasmodium Falciparum*. *Mol. Biol. Evol.* 17, 1467–1482. doi:10.1093/oxfordjournals.molbev.a026247

Beniwal, R. K., Nichani, A. K., Sharma, R. D., Rakha, N. K., Suri, D., and Sarup, S. (1997). Responses in Animals Vaccinated with theTheileria Annulata (Hisar) Cell Culture Vaccine. *Trop. Anim. Health Prod.* 29 (4Suppl. l), 109S–113S. doi:10.1007/BF02632947

Charlesworth, B. (2009). Effective Population Size and Patterns of Molecular Evolution and Variation. *Nat. Rev. Genet.* 10, 195–205. doi:10.1038/nrg2526

Conway, D. J., Roper, C., Oduola, A. M. J., Arnot, D. E., Kremsner, P. G., Grobusch, M. P., et al. (1999). High Recombination Rate in Natural Populations of *Plasmodium Falciparum*. *Proc. Natl. Acad. Sci.* 96, 4506–4511. doi:10.1073/pnas.96.8.4506

Darghouth, M. A. (2008). Review on the Experience with Live Attenuated Vaccines against Tropical Theileriosis in Tunisia: Considerations for the Present and Implications for the Future. *Vaccine* 26 (Suppl. 6), G4–G10. doi:10.1016/j.vaccine.2008.09.065

Dolan, T. T. (1989). Theileriasis : a Comprehensive Review. *Rev. Sci. Tech. OIE* 8 (1), 11–78. doi:10.20506/rst.8.1.398

Earl, D. A., and vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a Website and Program for Visualizing STRUCTURE Output and Implementing the Evanno Method. *Conservation Genet. Resour.* 4, 359–361. doi:10.1007/s12686-011-9548-7

El Hussein, A. M., Hassan, S. M., and Salih, D. A. (2012). Current Situation of Tropical Theileriosis in the Sudan. *Parasitol. Res.* 111, 503–508. doi:10.1007/s00436-012-2951-5

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the Number of Clusters of Individuals Using the Software Structure: a Simulation Study. *Mol. Ecol.* 14, 2611–2620. doi:10.1111/j.1365-294X.2005.02553.x

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin Suite Ver 3.5: A New Series of Programs to Perform Population Genetics Analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi:10.1111/j.1755-0998.2010.02847.x

Gharbi, M., Darghouth, M. A., Elati, K., AL-Hosary, A. A. T., Ayadi, O., Salih, D. A., et al. (2020). Current Status of Tropical Theileriosis in Northern Africa: A Review of Recent Epidemiological Investigations and Implications for Control. *Transbound Emerg. Dis.* 67 (Suppl. 1), 8–25. doi:10.1111/tbed.13312

Ghosh, S., and Azhahianambi, P. (2007). Laboratory Rearing of *Theileria Annulata*-free *Hyalomma Anatolicum Anatolicum* Ticks. *Exp. Appl. Acarol.* 43, 137–146. doi:10.1007/s10493-007-9100-3

Gomes, J., Salgueiro, P., Inácio, J., Amaro, A., Pinto, J., Tait, A., et al. (2016). Population Diversity of Theileria Annulata in Portugal. *Infect. Genet. Evol.* 42, 14–19. doi:10.1016/j.meegid.2016.04.023

Hashemi-Fesharki, R. (1988). Control of *Theileria Annulata* in Iran. *Parasitol. Today* 4 (2), 36–40. doi:10.1016/0169-4758(88)90062-2

Haubold, B., and Hudson, R. R. (2000). LIAN 3.0: Detecting Linkage Disequilibrium in Multilocus Data. *Bioinformatics* 16, 847–849. doi:10.1093/bioinformatics/16.9.847

Idury, R. M., and Cardon, L. R. (1997). A Simple Method for Automated Allele Binning in Microsatellite Markers. *Genome Res.* 7, 1104–1109. doi:10.1101/gr.7.11.1104

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a Cluster Matching and Permutation Program for Dealing with Label Switching and Multimodality in Analysis of Population Structure. *Bioinformatics* 23, 1801–1806. doi:10.1093/bioinformatics/btm233

Mohammed-Ahmed, G. M., Hassan, S. M., El Hussein, A. M., and Salih, D. A. (2018). Molecular, Serological and Parasitological Survey of *Theileria Annulata* in North Kordofan State, Sudan. *Vet. Parasitol. Reg. Stud. Rep.* 13, 24–29. doi:10.1016/j.vprsr.2018.03.006

Nei, M. (1978). Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. *Genetics* 89, 583–590. doi:10.1093/genetics/89.3.583

Ouhelli, H., Kachani, M., Flach, E., Williamson, S., El Hasnaoui, M., and Spooner, R. (1997). Investigations on Vaccination against Theileriosis in Morocco. *Trop. Anim. Health Prod.* 29, 103S. doi:10.1007/bf02632945

Oura, C. A. L., Asiimwe, B. B., Weir, W., Lubega, G. W., and Tait, A. (2005). Population Genetic Analysis and Sub-structuring of Theileria Parva in Uganda. *Mol. Biochem. Parasitol.* 140, 229–239. doi:10.1016/j.molbiopara.2004.12.015

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research-Aan Update. *Bioinformatics* 28, 2537–2539. doi:10.1093/bioinformatics/bts460

Peakall, R., and Smouse, P. E. (2006). Genalex 6: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research. *Mol. Ecol. Notes* 6, 288–295. doi:10.1111/j.1471-8286.2005.01155.x

Pipano, E., Samish, M., Kriegel, Y., and Yeruham, I. (1981). Immunization of Friesian Cattle against *Theileria Annulata* by the Infection-Treatment Method. *Br. Vet. J.* 137 (4), 416–420. doi:10.1016/s0007-1935(17)31640-8

Pipano, E., and Shkap, V. (2006). Vaccination against Tropical Theileriosis. *Ann. N. Y. Acad. Sci.* 916, 484–500. doi:10.1111/j.1749-6632.2000.tb05328.x

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155, 945–949.

Rosenberg, N. A. (2004). DISTRUCT: A Program for the Graphical Display of Population Structure. *Mol. Ecol. Notes* 4, 137–138.

Roy, S., Bhandari, V., Barman, M., Kumar, P., Bhanot, V., Arora, J. S., et al. (2021). Population Genetic Analysis of the *Theileria Annulata* Parasites Identified Limited Diversity and Multiplicity of Infection in the Vaccine from India. *Front. Microbiol.* 11, 579929. doi:10.3389/fmicb.2020.579929

Roy, S., Bhandari, V., Dandasena, D., Murthy, S., and Sharma, P. (2019). Genetic Profiling Reveals High Allelic Diversity, Heterozygosity and Antigenic Diversity in the Clinical Isolates of the *Theileria Annulata* from India. *Front. Physiol.* 10, 673. doi:10.3389/fphys.2019.00673

Sager, H., Bertoni, G., and Jungi, T. W. (1998). Differences between B Cell and Macrophage Transformation by the Bovine Parasite, *Theileria Annulata*: a Clonal Approach. *J. Immunol.* 1161 (1), 335–341.

Salih, D. A., Hassan, S. M., El Hussein, A. M., and Jongejan, F. (2004). Preliminary Survey of Ticks (Acari: Ixodidae) on Cattle in Northern Sudan. *Onderstepoort J. Vet. Res.* 71, 319–326. doi:10.4102/ojvr.v71i4.252

Salih, D. A., Mwacharo, J. M., Pelle, R., Njahira, M. N., Odongo, D. O., Mbole-Kariuki, M. N., et al. (2018). Genetic Diversity and Population Structure of *Theileria Parva* in South Sudan. *Ticks Tick-borne Dis.* 9, 806–813. doi:10.1016/j.ttbdis.2018.03.002

Sivakumar, T., Hayashida, K., Sugimoto, C., and Yokoyama, N. (2014). Evolution and Genetic Diversity of Theileria. *Infect. Genet. Evol.* 27, 250–263.

Smith, J. M., Smith, N. H., O'Rourke, M., and Spratt, B. G. (1993). How Clonal Are Bacteria? *Proc. Natl. Acad. Sci.* 90, 4384–4388. doi:10.1073/pnas.90.10.4384

Stepanova, N. I., Zablotskii, V. T., Çakmak, A., Inci, A., Yukari, B. A., Vatanserver, Z., et al. (1989). Bovine Theileriosis in the USSR. *Rev. Sci. Tech. OIE* 8 (1), 89–92. doi:10.20506/rst.8.1.396

Taha, K. M., Salih, D. A., Ali, A. M., Omer, R. A., and El Hussein, A. M. (2013). Naturally Occurring Infections of Cattle with Theileria Lestoquardi and Sheep with Theileria Annulata in the Sudan. *Vet. Parasitol.* 191, 143–145. doi:10.1016/j.vetpar.2012.08.003

Viseras, J., García-Fernández, P., and Adroher, F. J. (1997). Isolation and Establishment in *In Vitro* Culture of a Theileria Annulata - Infected Cell Line from Spain. *Parasitol. Res.* 83 (4), 394–396. doi:10.1007/s004360050270

Weir, W., Ben-Miled, L., Karagenç, T., Katzer, F., Darghouth, M., Shiels, B., et al. (2007). Genetic Exchange and Sub-structuring in Theileria Annulata Populations. *Mol. Biochem. Parasitol.* 154 (2), 170–180. doi:10.1016/j.molbiopara.2007.04.015

Weir, W., Karagenç, T., Gharbi, M., Simuunza, M., Aypak, S., Aysul, N., et al. (2011). Population Diversity and Multiplicity of Infection in Theileria Annulata. *Int. J. Parasitol.* 41, 193–203. doi:10.1016/j.ijpara.2010.08.004

Yin, F., Liu, Z., Liu, J., Liu, A., Salih, D. A., Li, Y., et al. (2018). Population Genetic Analysis of Theileria Annulata from Six Geographical Regions in China, Determined on the Basis of Micro- and Mini-Satellite Markers. *Front. Genet.* 9 (9), 50. doi:10.3389/fgene.2018.00050

Zhang, Z. H. (1997). A General Review on the Prevention and Treatment of Theileria Annulata in China. *Vet. Parasitol.* 70 (1-3), 77–81. doi:10.1016/s0304-4017(96)01127-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 December 2021
doi: 10.3389/fgene.2021.742153

Absence of PEXEL-Dependent Protein Export in *Plasmodium* Liver Stages Cannot Be Restored by Gain of the HSP101 Protein Translocon ATPase

Oriana Kreutzfeld[1,2,3], *Josephine Grützke*[1,2,4], *Alyssa Ingmundson*[1,2], *Katja Müller*[1,2] *and Kai Matuschewski*[1,2]*

[1]*Molecular Parasitology, Institute of Biology/Faculty for Life Sciences, Humboldt Universität zu Berlin, Berlin, Germany,* [2]*Parasitology Unit, Max Planck Institute for Infection Biology, Berlin, Germany,* [3]*Department of Medicine, University of California, San Francisco, San Francisco, CA, United States,* [4]*Department of Biological Safety, Federal Institute for Risk Assessment, Berlin, Germany*

Host cell remodeling is critical for successful *Plasmodium* replication inside erythrocytes and achieved by targeted export of parasite-encoded proteins. In contrast, during liver infection the malarial parasite appears to avoid protein export, perhaps to limit exposure of parasite antigens by infected liver cells. HSP101, the force-generating ATPase of the protein translocon of exported proteins (PTEX) is the only component that is switched off during early liver infection. Here, we generated transgenic *Plasmodium berghei* parasite lines that restore liver stage expression of *HSP101. HSP101* expression in infected hepatocytes was achieved by swapping the endogenous promoter with the *ptex150* promoter and by inserting an additional copy under the control of the elongation one alpha (ef*1α*) promoter. Both promoters drive constitutive and, hence, also pre-erythrocytic expression. Transgenic parasites were able to complete the life cycle, but failed to export PEXEL-proteins in early liver stages. Our results suggest that PTEX-dependent early liver stage export cannot be restored by addition of HSP101, indicative of alternative export complexes or other functions of the PTEX core complex during liver infection.

Keywords: malaria, plasmodium, protein export, pre-erythrocytic stage, PEXEL motif, PTEX translocon, ATPase, heat shock protein

OPEN ACCESS

Edited by:
Jun-Hu Chen, National Institute of Parasitic Diseases, China

Reviewed by:
Tania F. De Koning-Ward, Deakin University, Australia
Sreejoyee Ghosh, University of Texas MD Anderson Cancer Center, United States

***Correspondence:**
Kai Matuschewski
Kai.Matuschewski@hu-berlin.de

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *15 July 2021*
Accepted: *18 October 2021*
Published: *08 December 2021*

Citation:
Kreutzfeld O, Grützke J, Ingmundson A, Müller K and Matuschewski K (2021) Absence of PEXEL-Dependent Protein Export in Plasmodium *Liver Stages Cannot Be Restored by Gain of the HSP101 Protein Translocon ATPase. Front. Genet. 12:742153. doi: 10.3389/fgene.2021.742153*

INTRODUCTION

Host cell remodeling is critical for successful *Plasmodium* replication inside erythrocytes and achieved by targeted export of parasite-encoded proteins. Therefore, the parasite exports a large set of proteins to remodel the host erythrocyte, which is particularly important for plasma membrane fluidity and adhesive properties (Jonsdottir et al., 2021). The repertoire of exported proteins, termed exportome, is estimated to be approximately 10% of the *Plasmodium* proteome, and a quarter of the exported proteins are likely essential for parasite survival during blood stage development (Maier et al., 2008; Spielmann and Gilberger, 2015). Exported proteins have three major destinations in host erythrocytes, i) parasite-generated membranous structures, *e.g.* Maurer's clefts (Mundwiler-Pachlatko and Beck, 2013) and caveolae-vesicle complexes (Akinyi et al., 2012), ii) the erythrocyte membrane cytoskeleton (Warncke and Beck, 2019), and iii) the erythrocyte plasma membrane (Wahlgren et al., 2017).

Export of proteins across two apposed membranes, the parasite plasma membrane and the parasitophorous vacuole membrane (PVM), requires a tightly coordinated translocation process. Similar to the classic secretory pathway, *Plasmodium* exported proteins enter the endoplasmic reticulum (ER) and are eventually secreted into the PV, although the underlying mechanism is still inadequately understood (Matthews et al., 2019a). The signature export label for this destination is a small host cell targeting (HT) sequence or *Plasmodium* export element (PEXEL), comprising the amino acid residues RxLxE/Q/D located downstream of the ER signal sequence (Hiller et al., 2004; Marti et al., 2004; Mesén-Ramírez et al., 2016). A different class of exported proteins lack this motif, so-called PEXEL/HT negative exported proteins (PNEPs), but share the downstream translocation pathway (Grüring et al., 2012; Heiber et al., 2013). Both PEXEL and PNEP proteins are believed to be loaded into secretory vesicles to be transported and released into the PV. Once in the PV, proteins need to translocate the PVM to reach the erythrocyte cytoplasm, a process mediated by the *Plasmodium* translocon of exported proteins (PTEX) (de Koning-Ward et al., 2009).

Studies in human and rodent *Plasmodium* species revealed that the PTEX complex consists of three core proteins, EXP2, HSP101, PTEX150, which are flanked by two auxiliary proteins, PTEX88 and TRX2 (de Koning-Ward et al., 2009; Matthews et al., 2013; Matz et al., 2013), as well as the GPI-anchored surface protein P113 (Elsworth et al., 2016). Additional identified auxiliary proteins include the Export Protein Interacting complex, which comprises of the parasitophorous vacuole protein 1 (PV1), PV2 and EXP3 (Batinovic et al., 2017). EXP2 plays a central role as a membrane-spanning component, but simultaneously acts as a nutrient channel in a PTEX-independent manner (Garten et al., 2018). HSP101 is considered to be the force-generating motor protein and a member of the ClpB-type AAA^+-ATPase family (Beck et al., 2014; Matthews et al., 2019b). HSP101 contains a cargo-binding pore loop at the amino-terminal end and is most likely responsible for protein unfolding and active translocation through the complex. Although there is little evidence for a distinct functional role of PTEX150, numerous studies consistently confirmed that PTEX150 appears to play a structural role linking EXP2 and HSP101 (Bullen et al., 2012; Elsworth et al., 2014; Elsworth et al., 2016; Garten et al., 2018). Experimental genetics studies of the three core components in human and rodent *Plasmodium* species corroborated the importance of these proteins for parasite blood stage survival, and loss-of-function mutants failed to propagate during blood infection (Matthews et al., 2013; Matz et al., 2013; Beck et al., 2014; Elsworth et al., 2014). The role of PTEX88 in protein translocation is less well understood. Knockout of *PTEX88* in a murine malaria model reduced experimental cerebral malaria and cytoadherence (Matthews et al., 2013; Matz et al., 2013; Matz et al., 2015; Chisholm et al., 2016). It, thus, might play a role in translocation of specific proteins important for parasite virulence.

Although host cell remodeling and protein export are remarkably expanded in asexual blood stages, they occur both in gametocytes and during liver stage development (Ingmundson et al., 2014). This observation is further corroborated by the presence of syntenic genes in closely related *Hepatocystis* parasites, which only replicate in the liver and lack blood schizogony (Ejotre et al., 2021), yet harbor orthologous genes for *EXP2* (HEP_00110700), *HSP101* (HEP_00056400), *PTEX150* (HEP_ 00473300) and *PTEX88* (HEP_004505500) (Aunin et al., 2020). One hypothesis is that during liver infection the malarial parasite appears to curtail protein export. As metabolically active cells and in contrast to red blood cells, hepatocytes permit the developing parasite to exploit the host cell, likely limiting the need for intensive remodeling and, thereby, directing energy towards population expansion (Prudencio et al., 2006; Silvie et al., 2008a).

Moreover, hepatocytes harbor an active MHCI pathway with the potential to present peptide fragments of intracellular parasitic antigens to the host's immune cells (Chakravarty et al., 2007). To avoid recognition by the host, the parasite might keep protein translocation during early to mid liver stage development to a minimum. To date, only peptides of sporozoite origin could be identified as targets of $CD8^+$ T cells (Bongfen et al., 2007; Hafalla et al., 2013; Müller et al., 2017). Recently, a liver stage specific peptide, Kb-17, was identified, but its contribution to protective immune responses remains to be determined (Pichugin et al., 2018).

To date, three proteins were reported to translocate during liver stage growth, namely circumsporozoite protein (CSP) (Singh et al., 2007), liver-specific protein 2 (LISP2) (Orito et al., 2013), and sporozoite- and liver stage-expressed tryptophan rich protein (SLTRiP) (Jaijyan et al., 2015). The extent of CSP export in infected hepatocytes remains highly controversial (Singh et al., 2007; Cockburn et al., 2011), partly because CSP might be already released into the host's cytoplasm during sporozoite cell traversal and upon hepatocyte entry. Immunofluorescence assays with LISP2 antibodies demonstrated onset of export 24 h after infection, yet parasites harboring a C-terminal mCherry tagged LISP2 failed to export the tagged protein to the hepatocyte cytoplasm (Orito, et al., 2013). SLTRiP export in liver stages was detected as early as 12 h after infection, with 80% of infected host cells showing export of SLTRiP (Jaijyan et al., 2015). A central question is whether a potential liver stage exportome is translocated by similar or distinct mechanisms as compared to translocation during blood infection.

The notion of very restricted protein export during the first schizogony in infected hepatocytes is supported by additional experimental genetics evidence. Fusion of the PEXEL motif of *Plasmodium berghei* CSP to the model antigen ovalbumin and expression under the control of the *UIS4* promoter failed to induce export in infections with transgenic parasites, although $CD8^+$ T cell proliferation was increased (Montagna et al., 2014). Moreover, during liver stage maturation HSP101 is not expressed until 48 h after infection, whereas all other PTEX components are present (Matz et al., 2015), offering a plausible explanation for restricted protein export in these stages. Investigation of $KAHRP_L$-GFP PEXEL dependent export in liver stages showed absence of such export despite the presence of EXP2 and PTEX150. Further experiments with HA-tagged HSP101

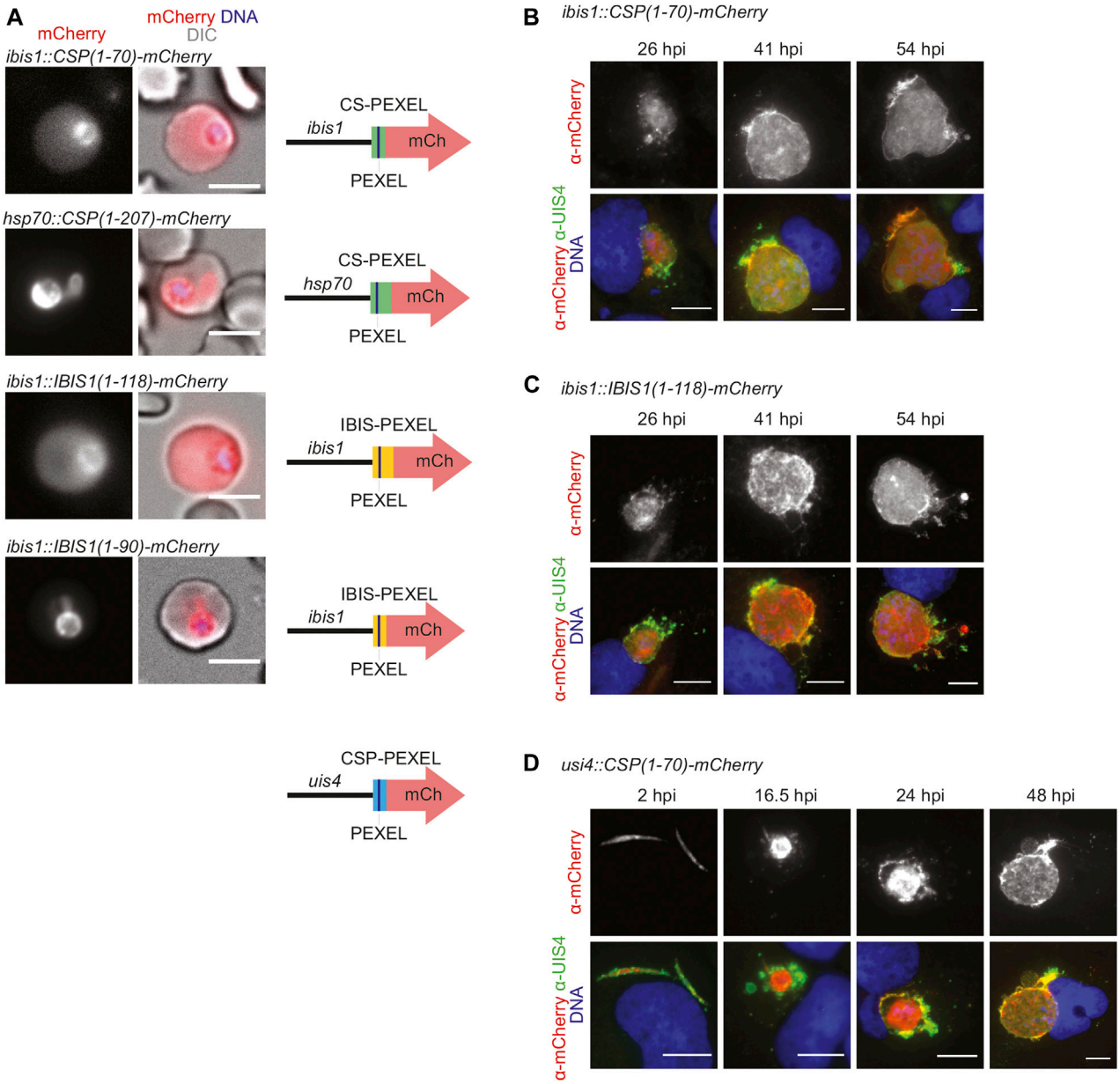


FIGURE 1 | PEXEL-dependent export of mCherry reporter protein occurs during blood but not liver infection. **(A)** Live fluorescence imaging of parasite-infected erythrocytes. Shown are the mCherry signal **(left)**, a merge with DIC images **(center)**, and graphic images of the reporter constructs **(right).** Depicted are promoters, PEXEL regions (green, yellow, blue), the PEXEL motif (line), and the mCherry open reading frame (red arrow). Top row, erythrocyte infected with a transgenic *CSP*$_{1\text{-}70}$-*mCherry* parasite under the control of the *IBIS1* promoter. Upper center row, erythrocyte infected with a transgenic *CSP*$_{1\text{-}70}$-*mCherry* parasite under the control of the *hsp70* promoter. Lower center row, erythrocyte infected with a transgenic *IBIS*$_{1\text{-}118}$-*mCherry* parasite under the control of the *IBIS1* promoter. Bottom row, erythrocyte infected with a transgenic *IBIS*$_{1\text{-}90}$-*mCherry* parasite under the control of the *IBIS1* promoter. Bar, 5 µm. **(B–D)** Immunofluorescent staining of Huh7 cells infected with the parasite lines expressing either **(B)** *CSP*$_{1\text{-}70}$-*mCherry*, **(C)** *IBIS1*$_{11\text{-}118}$-*mCherry*, or **(D)** *CSP-PEXEL-mCherry* under the control of the promoter indicated and at the indicated time points after sporozoite infections. Shown are α-mCherry signals **(top row)** and merge images **(bottom row)** of α-mCherry (red), α-UIS4 (green), and Hoechst 33,342 (blue) stain. Bar, 10 µm.

parasites confirmed the absence of HSP101 in liver stages, an indicator for lack of PEXEL dependent export in liver stages (Kalanon et al., 2016). A valid hypothesis, experimentally addressed in the present study, is that the *Plasmodium* PTEX ATPase HSP101 might be a limiting factor for liver stage protein export. Alternatively, the protein unfolding and translocation functions of HSP101 might be substituted by a distinct, yet unidentified factor.

An important implication is whether increasing antigen export in early liver stages would result in increased immune recognition of *Plasmodium*-infected cells, by presentation of parasite-derived epitopes *via* MHCI. We, therefore, sought to

define liver stage export in *HSP101* over-expressing parasites and to study these parasite lines for potential enhanced $CD8^+$ T cell responses and protection in cell-based immunization.

To this end, we designed *P. berghei* lines that express *HSP101* (PBANKA_0931200) during liver stage maturation. This gain-of-function approach is expected to allow PTEX reconstitution, since the other major PTEX components, PTEX150 and EXP2, and the auxiliary factor PTEX88 are expressed during liver stage development (Matz et al., 2015). Reconstitution of the PTEX export machinery might facilitate protein export in early liver stages and, thus, might promote antigen presentation and recognition of parasites by the host immune system.

RESULTS

Defined PEXEL Regions Lead to Export of mCherry Reporter Constructs in the Blood Stage, but not in the Liver Stage

To compare PEXEL-dependent export in *P. berghei* blood and liver stages we generated several reporter lines, which encode mCherry fusion proteins that contain predicted *CSP-PEXEL* and *IBIS-PEXEL* motifs at their amino-terminal end (**Figure 1A**; **Supplementary Figure S1**). Export of the fluorescent reporter proteins into red blood cells was analyzed by live immunofluorescence imaging. The first set of transgenic parasite lines expressed the PEXEL-containing reporter proteins under the control of the *IBIS1* promoter, which is active in several life cycle stages, including in liver and blood stages. Reporter proteins that included additional amino acids after the RxLxQ/E/D motif, $CSP_{1\text{-}70}$-mCherry and $IBIS_{1\text{-}118}$-mCherry, were exported to the infected erythrocytes (**Figure 1A**). When a truncated version of an IBIS1 fusion protein, $IBIS1_{1\text{-}90}$-mCherry, was expressed under the control of the *IBIS1* promoter, the red fluorescent signal was confined to the parasite and displayed an enrichment at the PV. A similar confinement to the parasite and a stronger signal at the PV was detected for transgenic parasites that expressed $CSP_{1\text{-}70}$-mCherry under the control of the strong and ubiquitous *HSP70* promoter (**Figure 1A**), indicative that excessive protein may congest the PTEX translocon.

We wanted to determine the fate of the two fusion proteins, $CSP_{1\text{-}70}$-mCherry and $IBIS_{1\text{-}118}$-mCherry, that were exported during blood infection, in liver stages. When hepatoma cells were infected with the lines described above that express these proteins from the *IBIS1* promoter, mCherry remained confined within the liver-stage PVM (**Figures 1B,C**). Because *IBIS1* is not expressed until the mid-liver stage (Ingmundson et al., 2012), we generated an additional transgenic line in which $CSP_{1\text{-}70}$-mCherry is under the control of the *UIS4* promoter to examine the localization of the reporter early in the pre-erythrocytic stage (Silvie et al., 2014). Fluorescent imaging of hepatoma cells infected with these transgenic sporozoites revealed expression of mCherry as early as 2 h after infection and throughout liver stage maturation up until 48 h later, but the signal remained within the membranes of the parasitophorous vacuole and associated tubovesicular network (**Figure 1D**).

Together, PEXEL-dependent export of reporter constructs was readily detected in blood stages, but liver stage export was not observed.

Generation of *Plasmodium berghei* Lines That Constitutively Express *HSP101*

To test whether PEXEL-dependent protein export in *P. berghei* liver stages is limited by the absence of the PTEX component HSP101, we next generated transgenic parasite lines that express *HSP101* using two complementary strategies, i) by insertion of an additional *HSP101* copy under the control of a constitutively active promoter, and ii) by swapping the endogenous promoter with a constitutive promoter. Both strategies are predicted to allow full reconstitution of a functional PTEX in liver stages, since PTEX150, EXP2, and PTEX88 are expressed during liver stage maturation (Matz et al., 2015).

Plasmids for *P. berghei* transfection were assembled in the pBAT and the pBART-SIL6 vector (Kooij et al., 2012). These targeting vectors harbor the drug-selectable *hDHFR-yFcu* cassette for positive/negative selection and a GFP expression cassette. The latter is used to acquire isogenic lines by fluorescence-activated cell sorting (FACS) and for live imaging of the parasite cytoplasm during the entire parasite life cycle. The pBART-SIL6 vector also contains two homologous regions to integrate into a silent region of chromosome 6 (SIL6) (**Supplementary Figure S2**).

We generated two targeting plasmids, which express an additional copy *P. berghei HSP101* (*Pb*ANKA_0931200) under the control of the constitutive promoter of elongation factor 1*α* (*ef1*α, PBANKA_113330) either as a fusion protein with a triple Myc tag only (*ef1α:HSP101-myc*) or an additional mCherry tag (*ef1α:HSP101-mCherry*) (**Supplementary Figure S3A**). Upon transfection and positive selection transgenic parasites were readily obtained and labeled *ef1α:HSP101-myc* or *-mCherry,* respectively. As a complementary approach, we replaced the endogenous promoter with the promoter of PTEX150 (PBANKA_1008500), which is expressed during liver stage development (Matz et al., 2015) and is expected to permit the appropriate temporal expression of *HSP101* (**Supplementary Figure S3C**). The corresponding parasite line, which was selected after transfection was labeled *HSP101:5′ptex150* to denote the promoter swap and retainment of *HSP101* at the original locus. For all three parasite lines the desired integration was confirmed by diagnostic PCR, demonstrating 5′ and 3′ integration and absence of WT parasites (**Supplementary Figures S3B,D**). Successful generation of the two *ef1α:HSP101* parasite lines and *HSP101:5′ptex150* parasites already indicated that additional expression from a second copy or exchange of the endogenous *HSP101* does not interfere with parasite blood stage development. *HSP101:5′ptex150* parasites harbor the endogenous *HSP101* copy under the *ptex150* promoter. As correct PTEX assembly is vital for parasite survival (Matthews et al., 2013; Matz et al., 2013; Beck et al., 2014; Elsworth et al., 2014), the ability to generate and grow *HSP101:5′ptex150* parasites indicated that correct PTEX assembly was achieved with *HSP101* expressed

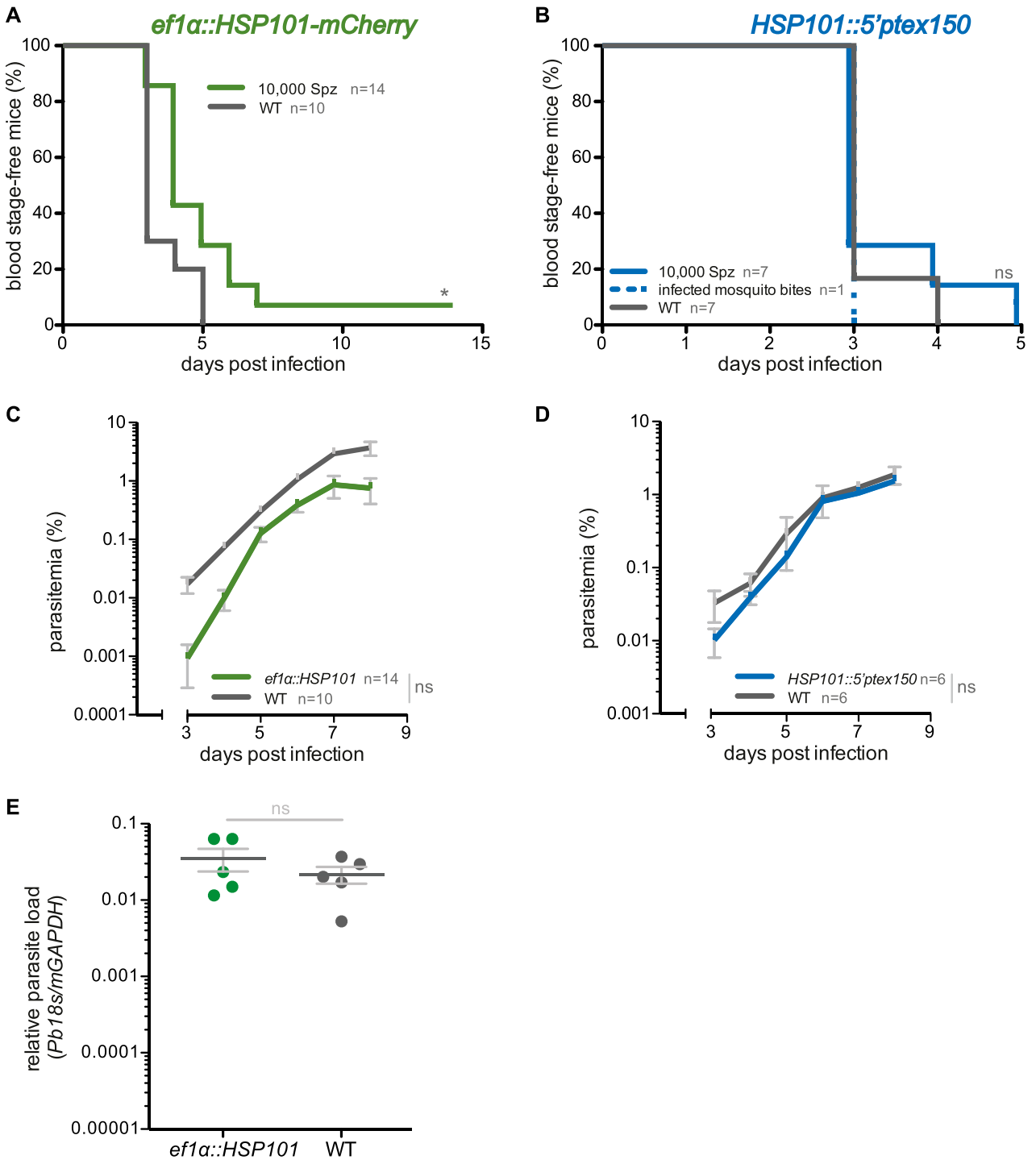


FIGURE 2 | Blood infections induced by *ef1α:HSP101-mCherry* and *hsp101:5'ptex150* sporozoites. **(A,B)** Shown are Kaplan-Meier analyses of blood infection over time after intravenous injection of 10,000 sporozoites isolated from **(A)** *ef1α:HSP101-mcherry* or **(B)** *hsp101:5'ptex150*-infected *Anopheles stephensi* mosquitoes. Exposure to 15 *hsp101:5'ptex150*-infected mosquitoes is included in **(B)**. **(C,D)** Blood stage growth in **(C)** *ef1α:HSP101-mCherry* or **(D)** *hsp101:5'ptex150*-infected mice. **(E)** qPCR analysis of relative parasite liver load 48 h after sporozoite infection. Shown are mean values (±SD). Blood infection was monitored daily by microscopic examination of Giemsa-stained blood films. *, $p < 0.05$; n. s., non-significant (Log rank test for Kaplan-Meier analysis and slope analysis; Mann-Whitney-U test for relative parasite liver load).

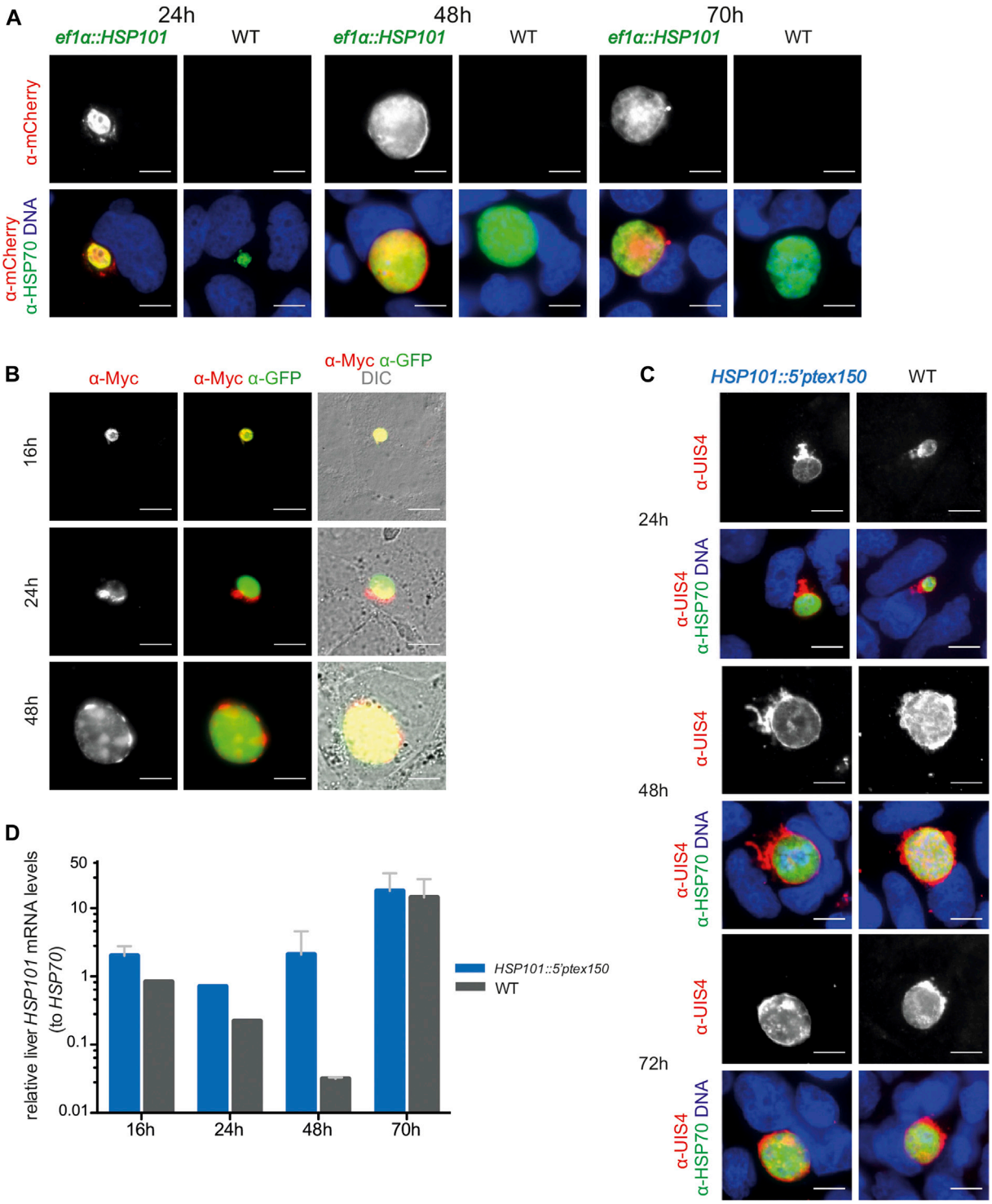


FIGURE 3 | Ectopic expression of HSP101 in liver stages results in PV localization. **(A)** Localization of HSP101-mCherry in *ef1α:HSP101-mCherry*-infected hepatoma cells at 24, 48, and 72 h after infection. Shown are immunofluorescent images with α-mCherry antibody **(top row)** and merge images **(bottom row)** displaying the signals of α-mCherry (red), α-HSP70 (green), and Hoechst 33,342 (blue). Bars, 10 µm. **(B)** Localization of HSP101-myc in *ef1α:HSP101-myc*-infected hepatoma cells at 16, 24, 48 h after infection. Shown are immunofluorescent images with an α-myc antibody **(left column)**, α-myc (red) and α-GFP (green) antibodies **(center column)**, and merge images **(right column)** displaying the signals of α-myc (red), α-GFP (green), and DIC. Bars, 10 µm. **(C)** Liver stage development of *hsp101:5'ptex150* parasites at 24, 48, and 72 h after infection. Shown are immunofluorescent images with an α-UIS4 antibody **(top)**, and merge images **(bottom)** displaying the signals of α-UIS4 (red), α-HSP70 (green), and Hoechst 33,342 (blue). Bars, 10 µm. **(D)** Steady state *HSP101* mRNA levels in *hsp101:5'ptex150* liver stages. Shown are relative mRNA levels determined qRT-PCR from total RNA of *hsp101:5'ptex150* or WT-infected hepatoma cells at indicated time points after infection. ΔΔCT values of *HSP101* mRNA values were normalized to *HSP70*. Depicted are mean values (±SD) (n = 2, except for 24h, n = 1).

under the *ptex150* promoter. Hence, we could characterize life cycle progression and analyze protein export during parasite maturation in the liver.

We next examined life cycle progression of the transgenic parasite lines. *ef1α:HSP101-mCherry* blood stages developed normally and displayed an mCherry signal consistent with PV localization (**Supplementary Figures S4A,B**), indicative of proper localization of the HSP101-mCherry fusion protein. Female *Anopheles stephensi* mosquitoes were allowed to feed on mice infected with *ef1α:HSP101-mCherry* parasites. Successful development in the mosquito was determined on day 10 for oocysts, day 14 for midgut sporozoites and day 17 for salivary gland sporozoite formation and results were compared to *P. berghei* WT parasites (*bergreen*). *ef1α:HSP101-mCherry*-infected mosquitoes showed normal oocyst and sporozoite development, and HSP101-mCherry was moderately expressed in ef*1α:HSP101-mCherry* oocysts and sporozoites (**Supplementary Figure S4C**). Midgut infectivity and sporozoite numbers in freshly dissected salivary glands of all three transgenic lines were similar to WT (**Supplementary Table S1**).

Overall, transgenic parasites underwent normal life cycle progression enabling analysis of liver stage development and protein export.

Ectopically Expressed HSP101 Localizes to the Parasitophorous Vacuole in Liver Stages

We first examined whether *ef1α:HSP101-mCherry* and *HSP101:5′ptex150* sporozoites are able to undergo pre-erythrocytic development and establish a patent blood infection. Groups of C57BL/6 mice were infected intravenously with 10,000 of WT, *ef1α:HSP101-mCherry* or *HSP101:5′ptex150* sporozoites and monitored for the time to occurrence of the first blood stage parasite, termed pre-patency, and parasite growth dynamics (**Figure 2**). We noticed a slight delay of approximately 1 day in pre-patency in mice infected with *ef1α:HSP101-mCherry* parasites in comparison to WT (**Figure 2A**). The delayed onset of blood infection was clearly visible in the analysis of blood stage propagation (**Figure 2C**).

We independently confirmed this finding by intravenous injection of the second transgenic line, *ef1α:HSP101-myc*, which also displayed a 1 day delay in patency (**Supplementary Figure S5**). But exposure to 15 *ef1α:HSP101-myc*-infected mosquitoes resulted in a normal prepatent period of 3 days (**Supplementary Figure S5**). This finding prompted us to quantify the parasite RNA levels in livers infected with mature liver stages, *i.e.* 48 h after sporozoite inoculation (**Figure 2E**). In this analysis, no significant differences were detectable between *ef1α:HSP101-mCherry* and WT-infected mice, indicating that the observed delay is not due to slower liver stage maturation. Remarkably, and in contrast to *ef1α:HSP101-mCherry* infections, there was no difference in *HSP101:5′ptex150* sporozoites parasite growth as compared to WT (**Figures 2B,D**).

TABLE 1 | Export ability of reporter lines in transgenic HSP101-expressing liver stages.

Transgenic parasites	Reporter line	Export (Yes/No)
ef1α:HSP101-mCherry	*uis4:CS-PEXEL-OVA*	*N*
ef1α:HSP101-myc	*uis4:CS-PEXEL-mCherry*	*N*
	uis4:IBIS1-PEXEL-mCherry	*N*
	ibis1:IBIS1-PEXEL-mCherry	*N*
	uis4:CS-PEXEL-OVA	*N*
HSP101:5′ptex150	*uis4:CS-PEXEL-mCherry*	*N*
	uis4:IBIS1-PEXEL-mCherry	*N*
	ibis1:IBIS1-PEXEL-mCherry	*N*
	uis4:CS-PEXEL-OVA	*N*

We infected hepatoma cells with freshly dissected salivary gland sporozoites and monitored liver stage size, morphology, HSP101 expression, and numbers of infected cells (**Figure 3**). The sizes of liver stages were similar for all transgenic lines and WT parasites. In *ef1α:HSP101-mCherry* parasites successful expression of HSP101 could be detected, using the red mCherry and myc tags as indicator. HSP101-mCherry localized to the liver stage PVM (**Figure 3A**). This localization was independently confirmed in hepatoma cells infected with *ef1α:HSP101*-myc sporozoites (**Figure 3B**). For *HSP101:5′ptex150* parasites growth was monitored by immunofluorescence analysis (**Figure 3C**). All three transgenic lines showed similar growth rates compared to WT (**Figure 3**).

In contrast to *ef1α:HSP101-mCherry* parasites, HSP101 expression in *HSP101:5′ptex150* could not be verified by immunofluorescence, since the endogenous open reading frame was retained. As a proxy for expression, we determined steady-state *HSP101* mRNA levels (**Figure 3D**). As intended, *HSP101* mRNA levels were significantly elevated at 24 and 48 h in *HSP101:5′ptex150* liver stages.

In conclusion, gain of *HSP101* expression resulted in successful liver stage development and HSP101 localization to the PVM in infected liver cells. While onset of blood infection was delayed in *ef1α:HSP101-mCherry* parasites, *HSP101:5′ptex150* infections developed parasite infections indistinguishable from WT.

Absence of PEXEL-Dependent Export in *ef1α:HSP101-mCherry* and *HSP101:5′ptex150* Liver Stages

To determine whether *ef1α:HSP101-mCherry* or *HSP101:5′ptex150* parasites display export of PEXEL-containing proteins during early liver stage development, we performed genetic crosses of the three transgenic parasite lines with four parasite lines that express reporter proteins (**Table 1**). Three of the lines express reporter proteins that were successfully exported during blood infection, namely *uis4:CSP$_{1\text{-}70}$-mCherry, uis4:IBIS$_{1\text{-}118}$-mCherry,* and *ibis1:IBIS$_{1\text{-}118}$-mCherry* (**Figure 1**), and one line, *uis4:CSP$_{1\text{-}70}$-OVA,* expresses the model antigen ovalbumin fused to the predicted CSP PEXEL motif (Montagna et al., 2014), which

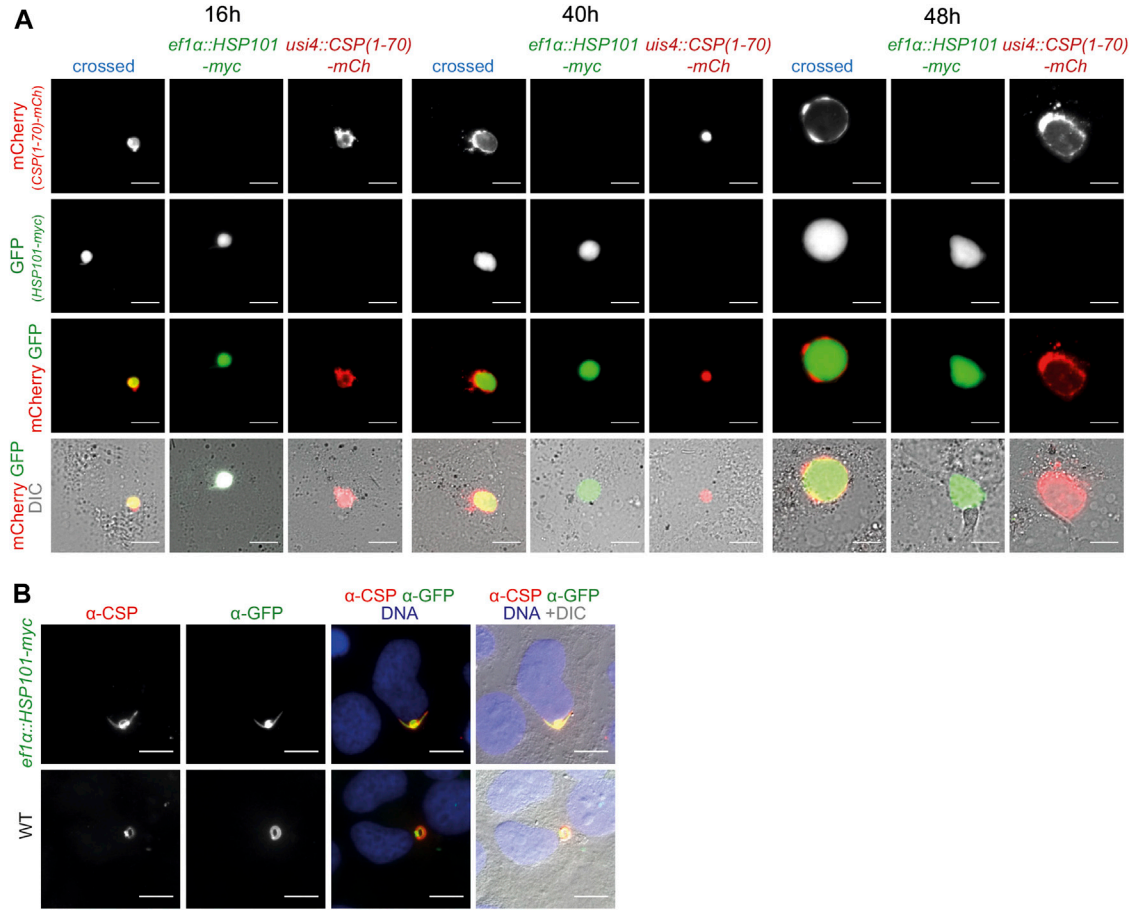


FIGURE 4 | Ecctopic expression of HSP101 in liver stages results in PV localization. **(A)** Live imaging of hepatoma cells infected with sporozoites from a genetic cross of *ef1α:HSP101-myc* and *UIS4:CSP-PEXEL-mCherry* (crossed, blue), and the two individual strains, *ef1α:HSP101-myc* (green), *UIS4:CSP-PEXEL-mCherry* (red). Shown are the PEXEL model protein fused to mCherry **(top row,** red), cytoplasmic GFP of the *ef1α:HSP101-myc* parasite **(upper center row,** green), the merge of both signals **(lower center row)**, and a merge of both fluorescent signals with DIC images **(bottom row)**. Note that the mCherry-tagged model antigen remains confined to the PV in HSP101-expressing parasites. Bars, 10 µm. **(B)** Immunofluorescence analysis of CSP in *ef1α:HSP101-myc* **(top row)** and WT **(bottom row)** parasites, 4 h after infection. Shown are the signals of α-CSP **(left)** and α-GFP **(centre left)**, the merge of both signals together with Hoechst 33,342 (blue; **centre right)**, and a merge of all fluorescent signals with DIC images **(right)**. Bars, 10 µm.

is also predicted to be exported in the blood stage. Genetic crosses were obtained by feeding *Anopheles stephensi* mosquitoes on mice, which were double infected with either *ef1α:HSP101-myc* or *HSP101:5′ptex150* parasites and the export reporter line (**Supplementary Figure S6**). Sporozoites were then isolated for hepatoma cell infection.

After sexual recombination in the brief phase of the diploid/tetraploid zygote of the otherwise haploid parasite life cycle a quarter of the mixed sporozoite population carries both desired genetic modification, since they are coded on different chromosomes. They can be directly visualized by fluorescent imaging, with GFP only in *HSP101* expressing parasites and mCherry or ovalbumin in the reporter constructs. We evaluated protein export in liver stages by both, live imaging and after antibody staining. As exemplified for one cross, *ef1α:HSP101-myc* and *UIS4:CSP$_{1-70}$-mCherry,* the tagged protein stayed confined to the PV at all time points during liver stage maturation, similar to the reporter line in a WT background (**Figure 4A**). We systematically determined protein export in all nine crosses and noted similar observations of no indication for export into the host hepatocyte. We independently tested the distribution of CSP (Singh, et al., 2007; Cockburn, et al., 2011) by antibody staining. Even in the presence of HSP101 we could not capture CSP signal outside the area of the *ef1α:HSP101-myc* parasite (**Figure 4B**).

Thus, protein export could not be induced in *ef1α:HSP101-myc* or *HSP101:5′ptex150* parasites, despite a range of PEXEL sequences and assessment of different time points during liver stage maturation. Accordingly, HSP101 is not a limiting factor for PEXEL-dependent protein export during this life cycle stage.

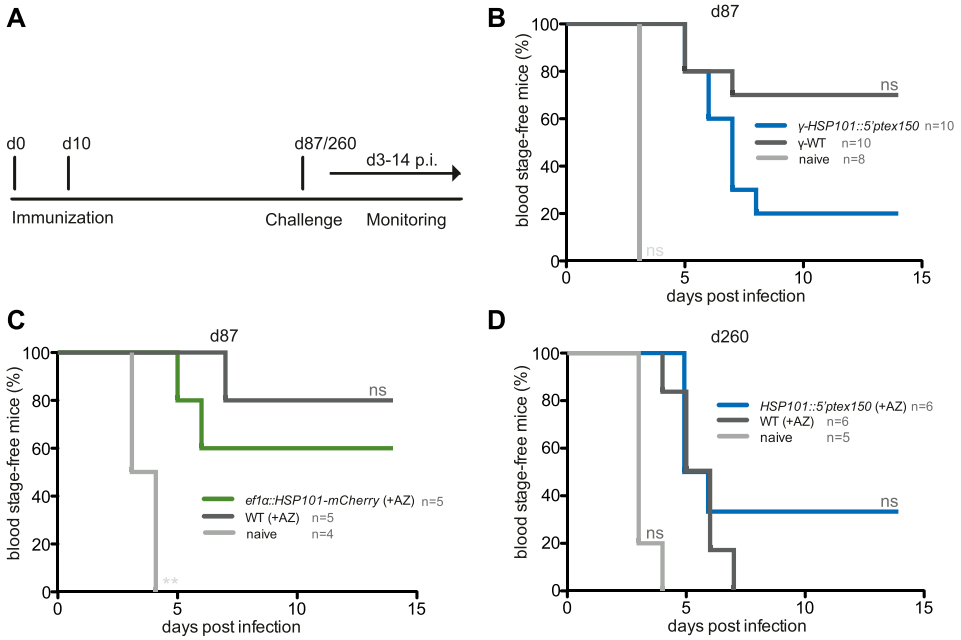


FIGURE 5 | Inferior protection in mice vaccinated with attenuated *ef1α:HSP101-mCherry* or *HSP101:5′ptex150* sporozoites. **(A)** Overview of immunization and challenge protocol. C57bl/6 mice were immunized twice at a 10-days interval with 10,000 intravenously injected (i.v.) sporozoites. Challenge infections with 10,000 i. v. WT sporozoites was done at indicated time points. Monitoring of blood infection was done daily by microscopic examination of Giemsa-stained blood films. **(B–D)** Kaplan-Meier analysis of blood infection over time after intravenous sporozoite injection. **(B)** Immunizations with irradiated *HSP101:5′ptex150* (blue) and WT (black) sporozoites. **(C)** Immunizations with azithromycin-arrested *ef1α:HSP101-mCherry* (green) and WT (black) sporozoites. **(B,C)** Challenge infections were done on day 87 (77 days after the booster immunization). **(D)** Immunizations with azithromycin-arrested *HSP101:5′ptex150* (blue) and WT (black) sporozoites. Challenge infections were done on day 260 (250 days after the booster immunization). Naive mice (grey) served as infection controls in all experiments. n. s, non-significant (Log rank (Mantel-Cox) test).

Attenuated *ef1α:HSP101-mCherry* and *HSP101:5′ptex150* Sporozoites Elicit Weak Vaccine-Induced Protection

We finally wanted to explore whether gain of HSP101 function modulated vaccine-induced protection. One hypothesis is that a larger repertoire of yet unidentified parasite antigens might increase immunogenicity during liver stage development, and processing of exported parasite proteins might be an important contributing factor. To test this notion, protective efficacy of immunization with live attenuated *ef1α:HSP101-mCherry* or *HSP101:5′ptex150* parasites was compared to WT immunizations (**Figure 5**). Groups of C57BL/6 mice received a prime/boost vaccination scheme of 10,000 sporozoites, followed by azithromycin treatment to elicit late liver stage arrest (Friesen et al., 2010). Challenge was done 3 months later by intravenous injection of 10,000 WT sporozoites (**Figure 5A**). This protocol has consistently shown incomplete sterile protection permitting evaluation of superior or inferior vaccination schemes (Friesen and Matuschewski, 2011). *ef1α:HSP101-mCherry* and WT sporozoite -immunized mice displayed substantial sterile protection, with three and four out of five mice, respectively, blood infection-free (**Figure 5B**). In contrast, only two out of 10 *HSP101:5′ptex150* -immunized mice were protected in comparison to seven out of 10 in WT-immunized mice (**Figure 5C**). Protection in *HSP101:5′ptex150* -immunized mice under azithromycin cover vanished completely when challenged at a late time point, 260 days after immunization (**Figure 5D**).

Collectively, immunizations with live *ef1α:HSP101-mCherry* or *HSP101:5′ptex150* sporozoites followed by attenuation with azithromycin does not lead to superior protection, but instead results in reduced vaccine efficacy.

DISCUSSION

In this study, we systematically analyzed PEXEL-dependent protein export during liver stage maturation and determined whether HSP101 is the limiting factor for the apparent paucity of secreted proteins. We generated and investigated transgenic *Plasmodium berghei* parasite lines that express fluorescent reporter proteins or HSP101 in liver stages, and, by genetic crosses, both simultaneously. We could show that reporter proteins, which are exported to the erythrocyte cytoplasm, fail to cross the membranes of the parasitophorous vacuole and tubovesicular network during liver stage development. For instance, *P. berghei* IBIS1 contains a conventional PEXEL motif that mediates export in *Plasmodium falciparum* blood stages (Petersen et al., 2015), but this is apparently not efficiently recognized for protein export in liver stages. Accordingly, any protein export

occurring in the liver stages is not analogous to protein export in the blood infection stage.

We also analyzed an unconventional PEXEL sequence, of the sporozoite surface antigen CSP. However, when mCherry was linked to this motif the corresponding fusion protein remained confined to the PV of developing liver stages, although it was efficiently targeted to host erythrocytes during blood infection. Our data corroborate the notion that CSP is constantly shed during sporozoite transmigration, but not actively exported after conversion to liver stage development (Cockburn, et al., 2011), and refute the suggestion that CSP is exported to the host nucleus to reprogram the infected hepatocyte (Singh et al., 2007). As of now, LISP2 and SLTriP remain the only liver stage proteins reported to translocate during mid-liver stage development. Carboxy-terminal mCherry tagging of LISP2 leads to accumulation of mCherry in the PV, and it was suggested that carboxy-terminal processing of LISP2 is required for export (Orito, et al., 2013). Signatures of exported but PEXEL-negative proteins remain poorly understood (Spielmann and Gilberger 2010), and in perspective, discovery of liver stage-specific export signals might help to define the liver stage exportome.

In other pathogens, protein translocons fail to export larger, stably folded proteins, therefore limiting translocation. For instance, bacterial type III secretion systems fail to translocate stably folded fluorescent proteins (Akeda and Galan 2005). In this study, reporter constructs harbored a PEXEL linked to mCherry, which is a relatively packed and large (29 kDa) protein, potentially limiting export of the reporter constructs in the liver stage in spite of efficient export in the blood. However, the model antigen OVA can be translocated by bacterial type III secretion systems (Wiedig et al., 2005), and is also not detected outside the liver stage PV when fused to the region of CSP that drives protein translocation across the blood stage PVM. Tagging with fluorescent proteins does not appear to limit protein translocation in the blood, and export of the mCherry reporters we used in our study was consistently detected in asexual blood stages. Translocation of fluorescent proteins in the blood stage requires functional HSP101 for protein unfolding (Matthews et al., 2019b), so the lack of mCherry export in the liver in the absence of HSP101 is perhaps expected. However, we were able to demonstrate that expression of untagged HSP101 in the liver stage from the *PTEX150* promoter still did not allow liver-stage parasites to export the mCherry reporter constructs. Clearly, more work is warranted to decipher protein export by liver stage *Plasmodium*. To this end, smaller tags that permit immunodetection might further minimize constraints by protein folding.

Plasmodium is an obligate intracellular pathogen and has to avoid immune recognition inside the only nucleated host cell, the hepatocyte, during the first population expansion phase in the liver. Parasite replication inside a PV is a powerful shield to minimize antigen exposure at the surface of the infected cell. The corresponding trade-offs are nutrient depletion and waste accumulation. During blood infection extensive erythrocyte remodeling poses no danger, since these terminally differentiated host cells are devoid of MHC class I -dependent antigen presentation. Previous studies have proposed a concept of regional PV export protein storage in asexual blood stages (Bullen, et al., 2012). These export areas are proposed to localize to specific bulges in the PV, which are defined by PTEX assembly, which in turn might be initiated by EXP2 localization determined by host cell factors (McMillan et al., 2013; Meibalan et al., 2015; de Koning-Ward et al., 2016). Proteins resistant to unfolding get trapped in loop-like PVM extensions in blood stages, and unblocking the unfolding capacity only partially recovers protein export, as cargo remains spatially segregated from the export machinery (Charnaud et al., 2018). Although our results indicate localization of cargo proteins in the PV, previous studies propose that proteins need to remain in defined export zones for efficient translocation, which might be determined by yet unknown proteins. PV-localized accessory proteins are involved in the export of proteins in blood stages (Elsworth et al., 2016; Batinovic et al., 2017). Whether this complex system relies on further accessory proteins, thereby raising the question whether these proteins are present in liver stages, remains to be determined.

Proper PTEX assembly plays a major role in effective protein export. Despite introducing HSP101 expression in liver stages to facilitate PTEX assembly, the mCherry signal of reporter constructs were confined to the parasite PV and not detectable in the hepatocyte cytoplasm. Thus, we conclude that transgenic parasites were not able to recreate the PEXEL-dependent export machinery in liver stages. Although expression of all core components was induced in the transgenic parasites, we were unable to confirm proper assembly of the complex in liver stages. Carboxy-terminal truncation of PTEX150, the linker between the pore spanning EXP2 and the AAA^+ ATPase HSP101, led to decrease in quantity of other PTEX components (Elsworth et al., 2016), suggesting complex destabilization. In this study, HSP101 was carboxy-terminally tagged, possibly interfering with correct protein folding. Indeed, recent structural analysis of PTEX complex assembly revealed stable binding of the EXP2 protomer to the carboxy-terminal ends of the HSP101 protomer, thereby placing the HSP101 cargo tunnel directly on the top of the EXP2-PTEX150 pore (Ho, et al., 2018). This suggests that interfering with the HSP101 carboxy-terminus, and as a consequence, with proper protein folding may prevent PTEX assembly; incomplete or unstable complex assembly has been proposed to inhibit protein translocon activity (de Koning-Ward, et al., 2016). Tagging HSP101 with mCherry has been shown to disrupt its function (Matthews et al., 2019b). Specifically, C-terminal mCherry tagging has been shown to interfere with unfolding of soluble translocated proteins and to interfere with parasite growth in the blood. Our findings of reduced growth of the *ef1α:HSP101-mCherry* are consistent with these findings. A smaller tag was well tolerated in this study (Matthews et al., 2019b). We note that we also expressed not only a myc-tagged HSP101, but also a native HSP101 protein in our promoter swap approach. Since these parasites displayed similar paucity of protein export, we consider dysfunctional PTEX assembly due to their carboxy-terminal tags less likely.

In vitro infection of hepatoma cells with *ef1α:HSP101-mCherry* parasites revealed HSP101 expression. However, the mCherry tag signal in the PV was weak, questioning if reconstruction of the machinery is feasible. As *HSP101:*

5′ptxe150 parasites lack a tagged version of HSP101, mRNA levels confirmed increased liver *HSP101* transcription compared to WT parasites, but whether HSP101 localizes to the PV remains to be determined. While we cannot rule out the possibility of post-translational repression, we believe this is unlikely to be the case because the regulatory regions used are well characterized, and visualization of *HSP101-mCherry* in early liver stages was possible in *ef1α:HSP101-mCherry* parasites. In contrast, endogenous *HSP101* is controlled at the transcriptional level in this stage; however, low mRNA levels were detectable in WT parasites, whereas HSP101 protein is absent in early liver stages (Matz et al., 2015), raising the questions whether endogenous *HSP101* is post-transcriptionally repressed.

It remains to be determined if further, yet unidentified, parasite and host proteins can contribute to protein export in liver stages. The core components EXP2, PTEX150 and HSP101 are essential for PTEX assembly and blood stage parasite survival (de Koning-Ward et al., 2009; Matthews et al., 2013; Matz et al., 2013). Previous work established that decreased HSP101 levels blocked export and led to vesicular accumulation of proteins in blood stages (Beck et al., 2014; Elsworth et al., 2014). Ho et al. indicate that the stoichiometry of 6:7:7 for HSP101-PTEX150-EXP2 is a prerequisite for functional PTEX assembly (Ho et al., 2018). Presently, it remains unknown whether the *ef1α* promoter or the promoter swap with the *ptex150* promoter yields sufficient protein. This can be experimentally tested by placing *HSP101* under the control of a strong pre-erythrocytic promoter.

In vivo infections demonstrated half a day delay in pre-patency for *ef1α:HSP101-mCherry* and *HSP101:5′ptex150*. Similarly, subsequent blood stage growth was delayed and parasitemia remained lower, however slope analysis demonstrated no significant differences in growth rate. Thus, it can be postulated that intra-hepatic growth is slightly affected, whereas blood stage development remains unaffected by *HSP101* overexpression. This notion is strengthened by parasite loads of infected mouse livers, which were slightly elevated at 48 h after infection in *ef1α:HSP101-mCherry* compared to WT -infected cohorts. As budding of merozoite-filled merosomes initiates as early as 48 h after infection in WT parasites (Sturm et al., 2006), slightly elevated parasite loads in *ef1α:HSP101-mCherry* infected mice could indicate a delay in maturation and budding of merosomes leading to the observed delay in pre-patency. Whether this is actually caused by *HSP101* overexpression itself or subsequent up- or downregulation or obstruction of other liver stage proteins remains to be determined.

To address PEXEL-dependent export in liver stages it is of interest to determine PTEX assembly with the additional HSP101 copy in engineered parasites. Lack of *P. berghei*-specific PTEX component antibodies prevented confirmation of PTEX assembly with immunoprecipitation. However, despite the lack of PTEX assembly verification, we observed localization of HSP101 to the PV of the parasite in liver stages and could confirm elevated *HSP101* mRNA levels in early liver stages. Additionally, PTEX assembly is vital for parasite survival in blood stages, therefore, a functional PTEX complex and correct HSP101 positioning in the complex is present in the promoter swap parasites *HSP101:5′ptex*. Localization of the tagged HSP101 copy in *ef1a:HSP101-mCherry* parasites was observed in the cytoplasm and a more prominent stain in the PV of the parasite. We note that the bulky mCherry tag could interfere with the correct localization of HSP101. This might affect PTEX assembly in parasites harboring an additional copy of HSP101.

Taking advantage of pathogen stage conversion for immunization strategies is one way to attack the parasite at its developmental bottleneck (Kreutzfeld et al., 2017). Studies have shown that superior immunity can be achieved by a developmental arrest shortly before liver-to-blood stage conversion, as a larger and broader array of antigens is presented compared to early arresting GAPs (Butler et al., 2011; Haussig et al., 2011). Therefore, we performed immunizations with azithromycin attenuated late arresting liver stages and challenged vaccinated animals to explore whether restoring HSP101 in liver stages can elicit superior immune responses. We, however, detected inferior protection, and whether this can be attributed to alteration of protein export during liver stage development needs to be explored further.

In conclusion, a functional PTEX complex, as seen in infected erythrocytes, cannot be reconstructed solely by ectopic HSP101. Our work also highlights that liver stage-host cell interactions remain relatively ill defined. The tight control of protein export during liver development likely contributes to poor recognition and cytolysis by the host immune system. Ultimately, identifying candidate liver stage-specific protein export signals, defining the contribution of PTEX in liver stage protein export, and characterizing the liver stage exportome by proteomic approaches are key to a better mechanistic understanding of immunity and vaccine development against *Plasmodium* pre-erythrocytic stages.

METHODS

Ethics Statement

All animal work was conducted in accordance with the German "Tierschutzgesetz in der Fassung vom 18. Mai 2006 (BGBl. I S. 1207)", which implements the directive 86/609/EEC from the European Union and the European Convention for the protection of vertebrate animals used for experimental and other scientific purposes. The ethics committee of MPI-IB and the Berlin state authorities (LAGeSo Reg# G0469/09 and G0294/15) approved the protocol.

Parasites and Experimental Animals

P. berghei ANKA cl507 parasites that constitutively express GFP under the control of the *EF1α* promoter (Franke-Fayard et al., 2004) and *P. berghei* ANKA bergreen parasites that constitutively express GFP under the control of the *HSP70* promoter (Matz et al., 2013) were used in our experiments. 6–8 weeks old female NMRI or C57BL/6 mice used in this study were either purchased from Charles River Laboratories, Janvier, or bred in-house. NMRI mice were used for transfection experiments, blood stage

infections and transmission to *Anopheles stephensi* mosquitos. C57BL/6 mice were used for sporozoite infections and subsequent immunological studies.

Generation of Transfection Vectors

To express an additional copy of *PbHSP101* (PbANKA_0931200) under a constitutive promoter the expression cassette was inserted into a silent locus on chromosome 6, using the pBART-SIL6 vector (Kooij et al., 2012). This plasmid harbors a drug selectable *hDHFR-yFcu* cassette for positive/negative selection, a GFP under the *PbHSP70* promoter for parasite life cycle analysis, an mCherry/triple c-Myc (3xMyc) tag sequence for protein tagging under the control of the 3′ region of *PbPPPK-DHPS* for mRNA stability, and two homologous sequences for stable integration into the silent locus on chromosome 6. The *HSP101* open reading frame was amplified using primers HSP101for and HSP101rev (**Supplementary Table S2**). The fragment was inserted *via Xba*I and *Age*I 5′ to the mCherry/3xMyc tag. For the plasmids *EF1α:HSP101-mCherry* and *EF1α:HSP101-myc* the 5′ region of *EF1α* (PbANKA_113330) was amplified using primers EF1for and EF1rev (**Supplementary Table S2**). The fragment was inserted *via BssH*II and *Xba*I 5′ next to the *HSP101* open reading frame. mCherry was excised in the plasmids *EF1α:HSP101-mCherry*, and replaced by a linker sequence, generated with Linkfor und Linkrev (**Supplementary Table S2**), fusing the *HSP101* open reading frame and the 3xMyc tag (**Supplementary Table S2**). For swap of the endogenous promoter with the 5′ region of *PTEX150* a pBART-based plasmid containing the *PTEX150* promoter linking *HSP101* ORF was generated. In brief, the amino-terminal portion of the *HSP101* ORF (primers HSP101_amino_F, HSP101_amino_R in **Supplementary Table S2**) was integrated under the control of *PTEX150* promoter (primers PTEX150-5′-PrimerF, PTEX150-5′-PrimerR in **Supplementary Table S2**) integrated with *Sac*II and *Hpa*I. The 5′UTR of HSP101 amplified with HSP101-5′-F and HSP101-5′-R (**Supplementary Table S2**). The 5′UTR together with the *HSP101* amino-terminus served as homologous recombination sites (**Supplementary Figure S3**).

The mCherry-based export reporters were constructed in the B3D+mCherry vector (Silvie et al., 2008b). The sequence encoding the first 70 amino acids of *CSP* was amplified with CSPfor and CSP_{70}rev (**Supplementary Table S2**) and cloned into the *Xba*I and *Spe*I sites of B3D+mCherry, and the 5′ genomic regions directly upstream of *UIS4*, *HSP70* and *IBIS1* were amplified with UIS4-5′for and UIS4-5′rev, HSP70-5′for and HSP70-5′rev, and IBIS1-5′for and IBIS1-5′rev (**Supplementary Table S2**), respectively, and inserted into the *Sac*II and *Not*I sites. Plasmids were linearized with *Bsm*I, *Hpa*I or PacI, respectively, prior to transfection. $IBIS1_{1\text{-}90}$ and $IBIS1_{1\text{-}118}$ were amplified together with the *IBIS1* 5′ region using primers IBIS1-5′for and $IBIS1_{90}$rev or $IBIS1_{118}$rev (**Supplementary Table S2**), respectively and inserted into the *Sac*II and *Spe*I sites of B3D + mCherry (**Supplementary Figure S2**). The IBIS1 plasmids were linearized with *Pac*I prior to transfection.

Generation of Transgenic Parasites

In order to generate stable transgenic parasite lines, NMRI mice were infected with 100 µl thawed cryopreserved *P.b.* ANKA blood stage parasites peritoneal and parasites were grown until a parasitemia of ~3–5%. Parasite culture and schizont isolation were done as described (Janse et al., 2006). Schizonts were transfected with 5–10 µg *Apa*LI linearized plasmids using the AMAXA Nucleofector device (program U33). Electroporated parasites were mixed with additional 50 µl transfection medium and intravenously injected into NMRI mice. The linearized *ef1α:HSP101-mCherry* and *ef1α:HSP101-myc* plasmids integrated in the silent locus in chromosome 6 (**Supplementary Figure S3**), and positive recombination events were selected with pyrimethamine (70 mg/L) in the drinking water, followed by FACS of GFP-positive iRBCs to generate isogenic transgenic parasite lines. The linearized *hsp101:5′ptex150* plasmid integrated between the endogenous *HSP101* 5′UTR and ORF resulting in a new, constitutively active promoter for *HSP101* expression (**Supplementary Figure S3C**). Recombinant *hsp101:5′ptex150* were selected by pyrimethamine, and an isogenic line was established by flowcytometric (FACS) cloning as described (Kenthirapalan et al., 2012). To verify successful integration, transgenic parasites DNA was extracted from blood stage cultures, and genotyping was performed with oligonucleotides specific for 5′ and 3′ integration (**Supplementary Table S2**) as well as for the WT locus (**Supplementary Table S2**, **Supplementary Figure S3B+D**).

Immunofluorescence Staining and Live Imaging of Infected Red Blood Cells

Parasite blood stages were visualized in immunofluorescence assays (IFA). Infected RBCs were retrieved by tail vein punctures of infected mice, mixed with 1x PBS and placed on a poly-L-lysine covered 12 mm ø cover slips in a 24-well plate to settle. Cells were fixed with 4% PFA/0.0075% fresh glutaraldehyde (GA) at RT. After three washing steps cells were permeabilized with 0.2% Trition X-100 (in PBS) for 20 min at RT. Cells were washed and blocked with 3% BSA, 0.2% Triton X-100 (in PBS) for 30 min at RT on a shaker. Primary antibody was added in blocking solution for 1 hour at RT, followed by three washing steps and incubation with fluorophore labeled secondary antibody and Hoechst 33,342 in blocking solution for 1 h at RT. Additional washing steps removed remaining secondary antibody before cells were mounted with Fluoromount G on a glass slide. Live imaging of iRBC was performed to access expression of mCherry and GFP as well as adequate growth of the parasites. iRBCs were placed on a glass slide and covered with a cover slip. Cells were analyzed with an Zeiss Axio Observer immunofluorescence microscope, and images were taken at 630x magnification.

Salivary Gland Sporozoite Isolation

WT GFP-expressing *P. berghei* ANKA and transgenic strains were maintained by continuous cycling in NMRI mice and female *Anopheles stephensi* mosquitoes (Vanderberg, 1975). Mosquitoes were kept at 28°C (non-infected) or 20°C (infected) at 80% humidity. 10–14 days after mosquito infection, midguts were isolated and oocyst development analyzed by fluorescent

microscopy. Salivary gland sporozoites were isolated on day 17 post infection from mosquitos in DMEM medium containing 10% fetal calf serum (FCS).

Immunofluorescence Staining of Salivary Gland Sporozoites

Salivary gland sporozoites were visualized by immunofluorescence assay (IFA). Briefly, 10,000 salivary gland sporozoites suspended in 3% BSA-RPMI were added into each ring of a Medco glass slide pre-coated with 3% BSA-RPMI and incubated for 15 min at 37°C. Slides were fixed with 4% paraformaldehyde (PFA), permeabilized with 0.2% TritonX-100 and blocked with 3% BSA. Sporozoites were stained with chicken anti-GFP antibody, followed by anti-chicken Alexa Fluor 488-coupled antibody (Invitrogen, OR, United States), rat anti mCherry antibody followed by anti-rat Alexa Fluor 546-coupled antibody (Invitrogen, OR, United States) and the nuclear stain Hoechst 33,342 (1:5,000; Invitrogen, OR, United States). Slides were mounted with Fluoromount-G (SouthernBiotech) prior to analysis by fluorescence microscopy using a Zeiss Axio Observer immunofluorescence microscope, and images were taken at 630x magnification.

Parasite Infectivity *in vivo*

For analysis of pre-patency and blood stage growth after sporozoite infection, 10,000 WT or transgenic sporozoites were injected intravenously into to C57BL/6 mice (n = 5 each). Tail punctures for blood smears were performed daily from day 3 after infection onwards until day 14. Smears were stained with a 10% Giemsa solution and analyzed by microscopic examination for presence and percentage of blood stage parasites. For natural mosquito infections 15 infected mosquitoes were kept in a separate container allowing a targeted blood meal on one C57BL/6 mouse. Mice were anesthetized with ketamine/xylazine intraperitoneal (i.p.), and mosquitoes were allowed to feed for 15 min. Mosquitoes were analyzed for successful feeding by examining blood uptake in the mosquito midgut.

Quantification of Parasite Burden in Mouse Livers

Groups of 5 C57BL/6 mice were infected with 10,000 WT or *EF1α:HSP101* sporozoites each, and the livers were harvested 42 h later. Livers were homogenized in TriZol (Thermo Fischer), and total RNA was extracted. cDNA was generated by reverse transcription with the RETROScript Kit (Ambion). Quantitative real-time PCR (qPCR) analysis was done with primers specific for *Pb*18S rRNA and mouse *GAPDH* for normalization (**Supplementary Table S2**). Relative liver parasite levels were measured applying the ΔΔCt method.

Crossing of Recombinant Parasite Lines

For crossing two parasite lines in the mosquito, female NMRI mice were infected intravenously (i.v.) with the parasite lines. On day 4, parasitemia was determined and five million parasites of each line were intravenously injected into one female NMRI mouse for subsequent mosquito infection. After confirmation of exflagellation of male gametes mosquitoes were allowed to take a blood meal on the infected mouse.

Hepatoma Cell Infection *in vitro* and Immunofluorescence Staining

Cell lines were kept according to standard cell culture conditions for human hepatoma cell lines (Huh7). In brief, Huh7 cells were cultured in DMEM complete (DMEM, 10% FCS, 1% Penicillin-Streptomycin) at 37°C and 5% CO2. Nunc Lab-Tek II 8-well Chamber Slides (ThermoFischer, NY, United States) were seeded with 30,000 Huh7 cells per well 24 h before infection. 10,000 WT or transgenic salivary gland sporozoites suspended in 100 µl DMEM-complete were added, centrifuged and allowed to settle for 2 h at RT. Cells were washed 2 h later to remove non-invaded sporozoites, and then medium was changed daily. Growth was terminated at 2, 16, 24, 40, 48, 54 and 72 h after infection by fixation with 4% PFA for 10 min at RT. Infected cells were blocked with PBS/10% FCS and permeabilized with 0.2% Triton (Roth, Karlsruhe, Germany) for 1 h at 37 °C. Primary antibodies (α-GFP, α-USI4, α-mCherry, α-Myc, α-*Pb*HSP70, and α-CSP) were added for 1 hour at 37 °C. After washing, cells were incubated with the respective fluorescently labeled secondary antibody (α-chicken Alexa Fluor 488, α-mouse Alexa Fluor 546, α-mouse Alexa Fluor 488, α-rat Alexa Fluor 546, α-rat Alexa Fluor 488, α-rabbit Alexa Fluor 546, α-rabbit Alexa Fluor 488; Invitrogen, OR, United States) and Hoechst 33,342 (Invitrogen, OR, United States) for 1 hour at 37 °C. Subsequently, cells were washed and mounted with Fluoromount G (SouthernBiotech) and imaged by fluorescence microscopy (Zeiss Axio Imager Z2) with 630x magnification.

Live Imaging of Infected Hepatoma Cells

For protein export analysis of infected Huh7 cells with crossed transgenic parasite lines, live imaging of parasites was performed. Infected cells were cultured in Ibidi slides, washed once with PBS and remained in PBS solution for the time of imaging. Live imaging was performed with a Zeiss Axio Observer immunofluorescence microscope with 1,000 magnification. Duration of analysis was kept at a minimum to ensure cell survival.

Quantification of mRNA Levels in Liver Stages

To determine expression of liver stage specific genes, relative mRNA levels were determined in WT and transgenic parasites. 300,000 Huh7 cells/well were plated in a 24-well plate. One day later, salivary gland sporozoites were isolated from infected mosquitoes and Huh7 cells were infected with 100,000 sporozoites/well. For parasite sedimentation plates were centrifuged for 10 min at 3,000 rpm, no brake, and incubated at 37 °C and 5% CO_2. Two hours later, cells were washed once to remove extracellular sporozoites. After 16, 24, 48 or 72 h cells were trypsinized with 200 µl for 3 min at 37 °C, and carefully

resuspended in DMEM and washed with additional DMEM. After 5 min centrifugation at 1200 rpm with brake, the pellet was resuspended in 500 µl TRIzol, and RNA was extracted for subsequent cDNA synthesis and qPCR analysis as described previously. qPCR analysis was performed with *HSP101* and *HSP70* primers (**Supplementary Table S2**).

Immunization With Attenuated WT and Transgenic Sporozoites

To examine vaccine efficacy mice were immunized mice in a prime boost regimen (day 0 and day 7) with 10,000 sporozoites delivered i. v. Salivary gland sporozoites were attenuated either by a γ-irradiation dose of 12×10^4 cGy, or with azithromycin (AZ) cover (4.8 mg in 200 µl of sodium chloride solution per mouse) (Friesen et al., 2010). For challenge experiments, immunized and control mice were i.v. injected with 10,000 WT *P. berghei* ANKA sporozoites 87 or 260 days post immunization. All mice were checked for parasitemia by microscopical examination of Giemsa-stained blood smears starting at day 3 after challenge inoculation until day 14. Animals that remained parasitemia-free were continuously checked for parasitemia up till 45 days after the challenge.

Statistics

Statistical significance was analyzed by non-parametrical Whitney-U test for non-normally distributed data (infected hepatoma cells) using GraphPad Prism V5.0c. Log rank (Mantel-Cox) test was applied for Kaplan-Meier analyses, and slope of parasite growth was analyzed using R (RStudio version 0.99.879).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The animal study was reviewed and approved by The ethics committee of MPI-IB, Berlin.

AUTHOR CONTRIBUTIONS

OK, AI, KMü, and KMa designed the experiments; OK, KMü and JG performed experiments and analyzed data; OK and KMa wrote the paper; all authors commented and revised the manuscript.

FUNDING

This work was funded by the Deutsche Forschungsgemeinschaft through the research training group–PhD program 2046 "Parasite infections: From experimental models to natural systems" (project B1) and in part by the Max Planck Society.

ACKNOWLEDGMENTS

We thank Laura Golusda for her help with the mouse infections and Joachim Matz for expertise advice and help during the cloning process of the *HSP101:5′ptex150 l*ine. We acknowledge support by the German Research Foundation (DFG) and the Open Access Publication Fund of Humboldt-Universität zu Berlin.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.742153/full#supplementary-material

REFERENCES

Akeda, Y., and Galán, J. E. (2005). Chaperone Release and Unfolding of Substrates in Type III Secretion. *Nature* 437, 911–915. doi:10.1038/nature03992

Akinyi, S., Hanssen, E., Meyer, E. V. S., Jiang, J., Korir, C. C., Singh, B., et al. (2012). A 95 kDa Protein of *Plasmodium Vivax* and *P. cynomolgi* Visualized by Three-Dimensional Tomography in the Caveola-Vesicle Complexes (Schüffner's Dots) of Infected Erythrocytes Is a Member of the PHIST Family. *Mol. Microbiol.* 84, 816–831. doi:10.1111/j.1365-2958.2012.08060.x

Aunin, E., Böhme, U., Sanderson, T., Simons, N. D., Goldberg, T. L., Ting, N., et al. (2020). Genomic and Transcriptomic Evidence for Descent from *Plasmodium* and Loss of Blood Schizogony in *Hepatocystis* Parasites from Naturally Infected Red colobus Monkeys. *Plos Pathog.* 16, e1008717. doi:10.1371/journal.ppat.1008717

Batinovic, S., McHugh, E., Chisholm, S. A., Matthews, K., Liu, B., Dumont, L., et al. (2017). An Exported Protein-Interacting Complex Involved in the Trafficking of Virulence Determinants in *Plasmodium*-Infected Erythrocytes. *Nat. Commun.* 8, 16044. doi:10.1038/ncomms16044

Beck, J. R., Muralidharan, V., Oksman, A., and Goldberg, D. E. (2014). PTEX Component HSP101 Mediates export of Diverse Malaria Effectors into Host Erythrocytes. *Nature* 511, 592–595. doi:10.1038/nature13574

Bongfen, S. E., Torgler, R., Romero, J. F., Renia, L., and Corradin, G. (2007). *Plasmodium berghei*-Infected Primary Hepatocytes Process and Present the Circumsporozoite Protein to Specific CD8+T Cells *In Vitro*. *J. Immunol.* 178, 7054–7063. doi:10.4049/jimmunol.178.11.7054

Bullen, H. E., Charnaud, S. C., Kalanon, M., Riglar, D. T., Dekiwadia, C., Kangwanrangsan, N., et al. (2012). Biosynthesis, Localization, and Macromolecular Arrangement of the *Plasmodium falciparum* Translocon of Exported Proteins (PTEX). *J. Biol. Chem.* 287, 7871–7884. doi:10.1074/jbc.M111.328591

Butler, N. S., Schmidt, N. W., Vaughan, A. M., Aly, A. S., Kappe, S. H. I., and Harty, J. T. (2011). Superior Antimalarial Immunity after Vaccination with Late Liver Stage-Arresting Genetically Attenuated Parasites. *Cell Host & Microbe* 9, 451–462. doi:10.1016/j.chom.2011.05.008

Chakravarty, S., Cockburn, I. A., Kuk, S., Overstreet, M. G., Sacci, J. B., and Zavala, F. (2007). CD8+ T Lymphocytes Protective against Malaria Liver Stages Are Primed in Skin-Draining Lymph Nodes. *Nat. Med.* 13, 1035–1041. doi:10.1038/nm1628

Charnaud, S. C., Jonsdottir, T. K., Sanders, P. R., Bullen, H. E., Dickerman, B. K., Kouskousis, B., et al. (2018). Spatial Organization of Protein export in Malaria Parasite Blood Stages. *Traffic* 19, 605–623. doi:10.1111/tra.12577

Chisholm, S. A., McHugh, E., Lundie, R., Dixon, M. W. A., Ghosh, S., O'Keefe, M., et al. (2016). Contrasting Inducible Knockdown of the Auxiliary PTEX

Component PTEX88 in *P. falciparum* and *P. berghei* Unmasks a Role in Parasite Virulence. *PLoS ONE* 11, e0149296. doi:10.1371/journal.pone.0149296

Cockburn, I. A., Tse, S.-W., Radtke, A. J., Srinivasan, P., Chen, Y.-C., Sinnis, P., et al. (2011). Dendritic Cells and Hepatocytes Use Distinct Pathways to Process Protective Antigen from *Plasmodium In Vivo*. *Plos Pathog.* 7, e1001318. doi:10.1371/journal.ppat.1001318

de Koning-Ward, T. F., Dixon, M. W. A., Tilley, L., and Gilson, P. R. (2016). *Plasmodium* Species: Master Renovators of Their Host Cells. *Nat. Rev. Microbiol.* 14, 494–507. doi:10.1038/nrmicro.2016.79

de Koning-Ward, T. F., Gilson, P. R., Boddey, J. A., Rug, M., Smith, B. J., Papenfuss, A. T., et al. (2009). A Newly Discovered Protein export Machine in Malaria Parasites. *Nature* 459, 945–949. doi:10.1038/nature08104

Ejotre, I., Reeder, D. M., Matuschewski, K., and Schaer, J. (2021). *Hepatocystis*. *Trends Parasitol.* 37, 456–457. doi:10.1016/j.pt.2020.07.015

Elsworth, B., Matthews, K., Nie, C. Q., Kalanon, M., Charnaud, S. C., Sanders, P. R., et al. (2014). PTEX Is an Essential Nexus for Protein export in Malaria Parasites. *Nature* 511, 587–591. doi:10.1038/nature13555

Elsworth, B., Sanders, P. R., Nebl, T., Batinovic, S., Kalanon, M., Nie, C. Q., et al. (2016). Proteomic Analysis Reveals Novel Proteins Associated with the *Plasmodium protein* Exporter PTEX and a Loss of Complex Stability upon Truncation of the Core PTEX Component, PTEX150. *Cell Microbiol.* 18, 1551–1569. doi:10.1111/cmi.12596

Franke-Fayard, B., Trueman, H., Ramesar, J., Mendoza, J., van der Keur, M., van der Linden, R., et al. (2004). A *Plasmodium berghei* Reference Line that Constitutively Expresses GFP at a High Level throughout the Complete Life Cycle. *Mol. Biochem. Parasitol.* 137, 23–33. doi:10.1016/j.molbiopara.2004.04.007

Friesen, J., and Matuschewski, K. (2011). Comparative Efficacy of Pre-erythrocytic Whole Organism Vaccine Strategies against the Malaria Parasite. *Vaccine* 29, 7002–7008. doi:10.1016/j.vaccine.2011.07.034

Friesen, J., Silvie, O., Putrianti, E. D., Hafalla, J. C. R., Matuschewski, K., and Borrmann, S. (2010). Natural Immunization against Malaria: Causal Prophylaxis with Antibiotics. *Sci. Transl. Med.* 2, ra49. doi:10.1126/scitranslmed.3001058

Garten, M., Nasamu, A. S., Niles, J. C., Zimmerberg, J., Goldberg, D. E., and Beck, J. R. (2018). EXP2 Is a Nutrient-Permeable Channel in the Vacuolar Membrane of *Plasmodium* and Is Essential for Protein export via PTEX. *Nat. Microbiol.* 3, 1090–1098. doi:10.1038/s41564-018-0222-7

Grüring, C., Heiber, A., Kruse, F., Flemming, S., Franci, G., Colombo, S. F., et al. (2012). Uncovering Common Principles in Protein export of Malaria Parasites. *Cell Host & Microbe* 12, 717–729. doi:10.1016/j.chom.2012.09.010

Hafalla, J. C. R., Bauza, K., Friesen, J., Gonzalez-Aseguinolaza, G., Hill, A. V. S., and Matuschewski, K. (2013). Identification of Targets of CD8+ T Cell Responses to Malaria Liver Stages by Genome-wide Epitope Profiling. *Plos Pathog.* 9, e1003303. doi:10.1371/journal.ppat.1003303

Haussig, J. M., Matuschewski, K., and Kooij, T. W. A. (2011). Inactivation of a *Plasmodium* Apicoplast Protein Attenuates Formation of Liver Merozoites. *Mol. Microbiol.* 81, 1511–1525. doi:10.1111/j.1365-2958.2011.07787.x

Heiber, A., Kruse, F., Pick, C., Grüring, C., Flemming, S., Oberli, A., et al. (2013). Identification of New PNEPs Indicates a Substantial Non-PEXEL Exportome and Underpins Common Features in *Plasmodium falciparum* Protein Export. *Plos Pathog.* 9, e1003546. doi:10.1371/journal.ppat.1003546

Hiller, N. L., Bhattacharjee, S., van Ooij, C., Liolios, K., Harrison, T., Lopez-Estraño, C., et al. (2004). A Host-Targeting Signal in Virulence Proteins Reveals a Secretome in Malarial Infection. *Science* 306, 1934–1937. doi:10.1126/science.1102737

Ho, C.-M., Beck, J. R., Lai, M., Cui, Y., Goldberg, D. E., Egea, P. F., et al. (2018). Malaria Parasite Translocon Structure and Mechanism of Effector export. *Nature* 561, 70–75. doi:10.1038/s41586-018-0469-4

Ingmundson, A., Alano, P., Matuschewski, K., and Silvestrini, F. (2014). Feeling at home from Arrival to Departure: Protein export and Host Cell Remodelling during *Plasmodium liver* Stage and Gametocyte Maturation. *Cell. Microbiol.* 16, 324–333. doi:10.1111/cmi.12251

Ingmundson, A., Nahar, C., Brinkmann, V., Lehmann, M. J., and Matuschewski, K. (2012). The Exported *Plasmodium berghei* Protein IBIS1 Delineates Membranous Structures in Infected Red Blood Cells. *Mol. Microbiol.* 83, 1229–1243. doi:10.1111/j.1365-2958.2012.08004.x

Jaijyan, D. K., Singh, H., and Singh, A. P. (2015). A Sporozoite- and Liver Stage-Expressed Tryptophan-Rich Protein Plays an Auxiliary Role in *Plasmodium* Liver Stage Development and Is a Potential Vaccine Candidate. *J. Biol. Chem.* 290, 19496–19511. doi:10.1074/jbc.M114.588129

Janse, C. J., Franke-Fayard, B., Mair, G. R., Ramesar, J., ThielEngelmann, C. S., Engelmann, S., et al. (2006). High Efficiency Transfection of *Plasmodium berghei* Facilitates Novel Selection Procedures. *Mol. Biochem. Parasitol.* 145, 60–70. doi:10.1016/j.molbiopara.2005.09.007

Jonsdottir, T. K., Gabriela, M., Crabb, B. S., F. de Koning-Ward, T., and Gilson, P. R. (2021). Defining the Essential Exportome of the Malaria Parasite. *Trends Parasitol.* 37, 664–675. doi:10.1016/j.pt.2021.04.009

Kalanon, M., Bargieri, D., Sturm, A., Matthews, K., Ghosh, S., Goodman, C. D., et al. (2016). The *Plasmodium* Translocon of Exported Proteins Component EXP2 Is Critical for Establishing a Patent Malaria Infection in Mice. *Cell Microbiol.* 18, 399–412. doi:10.1111/cmi.12520

Kenthirapalan, S., Waters, A. P., Matuschewski, K., and Kooij, T. W. A. (2012). Flow Cytometry-Assisted Rapid Isolation of Recombinant *Plasmodium berghei* Parasites Exemplified by Functional Analysis of Aquaglyceroporin. *Int. J. Parasitol.* 42, 1185–1192. doi:10.1016/j.ijpara.2012.10.006

Kooij, T. W. A., Rauch, M. M., and Matuschewski, K. (2012). Expansion of Experimental Genetics Approaches for *Plasmodium berghei* with Versatile Transfection Vectors. *Mol. Biochem. Parasitol.* 185, 19–26. doi:10.1016/j.molbiopara.2012.06.001

Kreutzfeld, O., Müller, K., and Matuschewski, K. (2017). Engineering of Genetically Arrested Parasites (GAPs) for a Precision Malaria Vaccine. *Front. Cel. Infect. Microbiol.* 7, 198. doi:10.3389/fcimb.2017.00198

Maier, A. G., Rug, M., O'Neill, M. T., Brown, M., Chakravorty, S., Szestak, T., et al. (2008). Exported Proteins Required for Virulence and Rigidity of *Plasmodium falciparum*-infected Human Erythrocytes. *Cell* 134, 48–61. doi:10.1016/j.cell.2008.04.051

Marti, M., Good, R. T., Rug, M., Knuepfer, E., and Cowman, A. F. (2004). Targeting Malaria Virulence and Remodeling Proteins to the Host Erythrocyte. *Science* 306, 1930–1933. doi:10.1126/science.1102452

Matthews, K., Kalanon, M., Chisholm, S. A., Sturm, A., Goodman, C. D., Dixon, M. W. A., et al. (2013). The *Plasmodium translocon* of Exported Proteins (PTEX) Component Thioredoxin-2 Is Important for Maintaining normal Blood-Stage Growth. *Mol. Microbiol.* 89, 1167–1186. doi:10.1111/mmi.12334

Matthews, K. M., Kalanon, M., and de Koning-Ward, T. F. (2019b). Uncoupling the Threading and Unfoldase Actions of *Plasmodium* HSP101 Reveals Differences in Export between Soluble and Insoluble Proteins. *mBio* 10, e01106–19. doi:10.1128/mBio.01106-19

Matthews, K. M., Pitman, E. L., and Koning-Ward, T. F. (2019a). Illuminating How Malaria Parasites export Proteins into Host Erythrocytes. *Cell Microbiol.* 21, e13009. doi:10.1111/cmi.13009

Matz, J. M., Goosmann, C., Brinkmann, V., Grützke, J., Ingmundson, A., Matuschewski, K., et al. (2015). The *Plasmodium berghei* Translocon of Exported Proteins Reveals Spatiotemporal Dynamics of Tubular Extensions. *Sci. Rep.* 5, 12532. doi:10.1038/srep12532

Matz, J. M., Matuschewski, K., and Kooij, T. W. A. (2013). Two Putative Protein export Regulators Promote *Plasmodium* Blood Stage Development *In Vivo*. *Mol. Biochem. Parasitol.* 191, 44–52. doi:10.1016/j.molbiopara.2013.09.003

McMillan, P. J., Millet, C., Batinovic, S., Maiorca, M., Hanssen, E., Kenny, S., et al. (2013). Spatial and Temporal Mapping of the PfEMP1 export Pathway in *Plasmodium falciparum*. *Cel. Microbiol.* 15, 1401–1418. doi:10.1111/cmi.12125

Meibalan, E., Comunale, M. A., Lopez, A. M., Bergman, L. W., Mehta, A., Vaidya, A. B., et al. (2015). Host Erythrocyte Environment Influences the Localization of Exported Protein 2, an Essential Component of the *Plasmodium* Translocon. *Eukaryot. Cell* 14, 371–384. doi:10.1128/EC.00228-14

Mesén-Ramírez, P., Reinsch, F., Blancke Soares, A., Bergmann, B., Ullrich, A.-K., Tenzer, S., et al. (2016). Stable Translocation Intermediates Jam Global Protein export in *Plasmodium falciparum* Parasites and Link the PTEX Component EXP2 with Translocation Activity. *Plos Pathog.* 12, e1005618. doi:10.1371/journal.ppat.1005618

Montagna, G. N., Beigier-Bompadre, M., Becker, M., Kroczek, R. A., Kaufmann, S. H. E., and Matuschewski, K. (2014). Antigen export during Liver Infection of the Malaria Parasite Augments Protective Immunity. *mBio* 5, e01321–14. doi:10.1128/mBio.01321-14

Müller, K., Gibbins, M. P., Matuschewski, K., and Hafalla, J. C. R. (2017). Evidence of Cross-Stage CD8+ T Cell Epitopes in Malaria Pre-erythrocytic and Blood Stage Infections. *Parasite Immunol.* 39, e12434. doi:10.1111/pim.12434

Mundwiler-Pachlatko, E., and Beck, H.-P. (2013). Maurer's Clefts, the enigma of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci.* 110, 19987–19994. doi:10.1073/pnas.1309247110

Orito, Y., Ishino, T., Iwanaga, S., Kaneko, I., Kato, T., Menard, R., et al. (2013). Liver-specific Protein 2: a *Plasmodium* Protein Exported to the Hepatocyte Cytoplasm and Required for Merozoite Formation. *Mol. Microbiol.* 87, 66–79. doi:10.1111/mmi.12083

Petersen, W., Matuschewski, K., and Ingmundson, A. (2015). Trafficking of the Signature Protein of Intra-erythrocytic *Plasmodium berghei*-Induced Structures, IBIS1, to *P. falciparum* Maurer's Clefts. *Mol. Biochem. Parasitol.* 200, 25–29. doi:10.1016/j.molbiopara.2015.04.005

Pichugin, A., Zarling, S., Perazzo, L., Duffy, P. E., Ploegh, H. L., and Krzych, U. (2018). Identification of a Novel CD8 T Cell Epitope Derived from *Plasmodium berghei* Protective Liver-Stage Antigen. *Front. Immunol.* 9, 91. doi:10.3389/fimmu.2018.00091

Prudêncio, M., Rodriguez, A., and Mota, M. M. (2006). The Silent Path to Thousands of Merozoites: the *Plasmodium* Liver Stage. *Nat. Rev. Microbiol.* 4, 849–856. doi:10.1038/nrmicro1529

Silvie, O., Briquet, S., Müller, K., Manzoni, G., and Matuschewski, K. (2014). Post-transcriptional Silencing of *UIS4* in *Plasmodium berghei* sporozoites Is Important for Host Switch. *Mol. Microbiol.* 91, 1200–1213. doi:10.1111/mmi.12528

Silvie, O., Mota, M. M., Matuschewski, K., and Prudêncio, M. (2008a). Interactions of the Malaria Parasite and its Mammalian Host. *Curr. Opin. Microbiol.* 11, 352–359. doi:10.1016/j.mib.2008.06.005

Silvie, O., Goetz, K., and Matuschewski, K. (2008b). A Sporozoite Asparagine-Rich Protein Controls Initiation of *Plasmodium* Liver Stage Development. *PLoS Pathog.* 4, e1000086. doi:10.1371/journal.ppat.1000086

Singh, A. P., Buscaglia, C. A., Wang, Q., Levay, A., Nussenzweig, D. R., Walker, J. R., et al. (2007). *Plasmodium* Circumsporozoite Protein Promotes the Development of the Liver Stages of the Parasite. *Cell* 131, 492–504. doi:10.1016/j.cell.2007.09.013

Spielmann, T., and Gilberger, T.-W. (2010). Protein export in Malaria Parasites: Do Multiple export Motifs Add up to Multiple export Pathways? *Trends Parasitol.* 26, 6–10. doi:10.1016/j.pt.2009.10.001

Sturm, A., Amino, R., van de Sand, C., Regen, T., Retzlaff, S., Rennenberg, A., et al. (2006). Manipulation of Host Hepatocytes by the Malaria Parasite for Delivery into Liver Sinusoids. *Science* 313, 1287–1290. doi:10.1126/science.1129720

Vanderberg, J. P. (1975). Development of Infectivity by the *Plasmodium berghei* Sporozoite. *J. Parasitol.* 61, 43–50. doi:10.2307/3279102

Wahlgren, M., Goel, S., and Akhouri, R. R. (2017). Variant Surface Antigens of *Plasmodium falciparum* and Their Roles in Severe Malaria. *Nat. Rev. Microbiol.* 15, 479–491. doi:10.1038/nrmicro.2017.47

Warncke, J. D., and Beck, H.-P. (2019). Host Cytoskeleton Remodeling throughout the Blood Stages of *Plasmodium falciparum*. *Microbiol. Mol. Biol. Rev.* 83, e00013–19. doi:10.1128/MMBR.00013-19

Wiedig, C. A., Kramer, U., Garbom, S., Wolf-Watz, H., and Autenrieth, I. B. (2005). Induction of CD8+ T Cell Responses by *Yersinia* Vaccine Carrier Strains. *Vaccine* 23, 4984–4998. doi:10.1016/j.vaccine.2005.05.027

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 January 2022
doi: 10.3389/fgene.2021.785934

Comprehensive Genomic Discovery of Non-Coding Transcriptional Enhancers in the African Malaria Vector *Anopheles coluzzii*

Inge Holm[1], Luisa Nardini[1], Adrien Pain[1,2], Emmanuel Bischoff[1], Cameron E. Anderson[3], Soumanaba Zongo[4], Wamdaogo M. Guelbeogo[4], N'Fale Sagnon[4], Daryl M. Gohl[5,6], Ronald J. Nowling[7], Kenneth D. Vernick[1] and Michelle M. Riehle[3]**

[1]Institut Pasteur, Université de Paris, CNRS UMR 2000, Unit of Insect Vector Genetics and Genomics, Department of Parasites and Insect Vectors, Paris, France, [2]Institut Pasteur, Université de Paris, Hub de Bioinformatique et Biostatistique, Paris, France, [3]Department of Microbiology and Immunology, Medical College of Wisconsin, Milwaukee, WI, United States, [4]Centre National de Recherche et de Formation sur le Paludisme (CNRFP), Ministry of Health, Ouagadougou, Burkina Faso, [5]University of Minnesota Genomics Center, Minneapolis, MN, United States, [6]Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, MN, United States, [7]Department of Electrical Engineering and Computer Science, Milwaukee School of Engineering (MSOE), Milwaukee, WI, United States

OPEN ACCESS

Edited by:
Moses Okpeku,
University of KwaZulu-Natal, South Africa

Reviewed by:
Alejandra Medina-Rivera,
National Autonomous University of Mexico, Mexico
Andreas Pfenning,
Carnegie Mellon University, United States

***Correspondence:**
Michelle M. Riehle
mriehle@mcw.edu
Kenneth D. Vernick
kenneth.vernick@pasteur.fr

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: 29 September 2021
Accepted: 10 December 2021
Published: 10 January 2022

Citation:
Holm I, Nardini L, Pain A, Bischoff E, Anderson CE, Zongo S, Guelbeogo WM, Sagnon N, Gohl DM, Nowling RJ, Vernick KD and Riehle MM (2022) Comprehensive Genomic Discovery of Non-Coding Transcriptional Enhancers in the African Malaria Vector Anopheles coluzzii. Front. Genet. 12:785934. doi: 10.3389/fgene.2021.785934

Almost all regulation of gene expression in eukaryotic genomes is mediated by the action of distant non-coding transcriptional enhancers upon proximal gene promoters. Enhancer locations cannot be accurately predicted bioinformatically because of the absence of a defined sequence code, and thus functional assays are required for their direct detection. Here we used a massively parallel reporter assay, Self-Transcribing Active Regulatory Region sequencing (STARR-seq), to generate the first comprehensive genome-wide map of enhancers in *Anopheles coluzzii*, a major African malaria vector in the Gambiae species complex. The screen was carried out by transfecting reporter libraries created from the genomic DNA of 60 wild *A. coluzzii* from Burkina Faso into *A. coluzzii* 4a3A cells, in order to functionally query enhancer activity of the natural population within the homologous cellular context. We report a catalog of 3,288 active genomic enhancers that were significant across three biological replicates, 74% of them located in intergenic and intronic regions. The STARR-seq enhancer screen is chromatin-free and thus detects inherent activity of a comprehensive catalog of enhancers that may be restricted *in vivo* to specific cell types or developmental stages. Testing of a validation panel of enhancer candidates using manual luciferase assays confirmed enhancer function in 26 of 28 (93%) of the candidates over a wide dynamic range of activity from two to at least 16-fold activity above baseline. The enhancers occupy only 0.7% of the genome, and display distinct composition features. The enhancer compartment is significantly enriched for 15 transcription factor binding site signatures, and displays divergence for specific dinucleotide repeats, as compared to matched non-enhancer genomic controls. The genome-wide catalog of *A. coluzzii* enhancers is publicly available in a simple searchable graphic format. This enhancer catalogue will be valuable in linking genetic and phenotypic variation, in identifying regulatory elements that could be employed in vector manipulation, and in better

targeting of chromosome editing to minimize extraneous regulation influences on the introduced sequences.

Importance: Understanding the role of the non-coding regulatory genome in complex disease phenotypes is essential, but even in well-characterized model organisms, identification of regulatory regions within the vast non-coding genome remains a challenge. We used a large-scale assay to generate a genome wide map of transcriptional enhancers. Such a catalogue for the important malaria vector, *Anopheles coluzzii*, will be an important research tool as the role of non-coding regulatory variation in differential susceptibility to malaria infection is explored and as a public resource for research on this important insect vector of disease.

Keywords: mosquito, STARR-seq, anopheles, non-coding, regulatory element, enhancer

INTRODUCTION

Transcriptional enhancers are non-coding *cis*-regulatory elements that are responsible for most of the regulated gene expression in eukaryotic genomes. Promoters, located just upstream of transcription start sites, mediate only basal levels of unregulated gene expression. Enhancers bind transcription and other protein factors and subsequently interact with physically distant promoters. This enhancer-promoter interaction drives the majority of regulated gene expression of enhancer target genes (Robson et al., 2019; Schoenfelder and Fraser, 2019; Andersson and Sandelin, 2020).

Genetic variation in enhancer sequences can differentially influence gene expression and appears to underlie complex medically and agriculturally important phenotypes (Bird et al., 2006; Hrdlickova et al., 2014; Gloss and Dinger, 2018; Zheng et al., 2019). Indeed, genetic variation in coding regions of the genome have been shown to explain little of the variation in phenotype. Results from genetic mapping methods such as genome wide association surveys (GWAS) indicate that at least 90% of significant GWAS candidate variants in humans are found within the non-coding genome, and most of these significant variants appear to be located within enhancers (Neph et al., 2012; Schaub et al., 2012; Farh et al., 2015).

A challenge in identifying transcriptional enhancers is the absence of an amino acid-like code that could facilitate their accurate recognition. An additional obstacle to developing accurate computational algorithms for enhancer prediction is the paucity of experimental data from different systems. The low sequence complexity and repetitive nature of non-coding regions of the genome also make computational prediction more challenging (Kleftogiannis et al., 2016; Lim et al., 2018; Hong et al., 2021).

Transcriptional enhancers can be identified by either direct or indirect experimental approaches (Farley et al., 2021). Direct approaches measure enhancer activity of nucleotide sequence through assays such as manual luciferase reporter assays. A manual luciferase reporter assay queries the ability of a candidate sequence to regulate reporter gene transcription, which is quantified by luminescence produced by the luciferase protein product. Direct detection measures the inherent transcriptional regulatory activity of a given nucleotide sequence isolated from its dynamic chromatin context (Arnold et al., 2013; Arnold et al., 2014). In contrast, indirect approaches identify enhancers as correlates of genomic regions differing in their chromatin accessibility or histone modification markers. The indirect strategies such as ChIP-seq and ATAC-seq (Gomez-Diaz et al., 2017; Ruiz et al., 2019; Farley et al., 2021; Ruiz et al., 2021) infer enhancer presence based on chromatin properties such as chromatin accessibility and histone modifications, but do not directly or functionally test or confirm enhancer activity. Methods used to survey histone signatures and chromatin accessibility may also detect transcriptional silencers, insulators and other features in addition to enhancers, and the functional category of detected elements is not necessarily distinguishable without additional work, for example direct functional assays.

In order to understand the phenotypic significance of enhancers and their nucleotide variation, it is essential to be able to filter the genome for functional enhancer elements. Enhancers have typically been studied using manual plasmid reporter assays to query a cloned candidate enhancer fragment. However, such manual reporter assays are not suitable for genome-wide screening because each assay assesses the activity of a single enhancer candidate and they are not multiplexable or scalable to the genome level (Muerdter et al., 2015).

Here, Self-Transcribing Active Regulatory Region sequencing (STARR-seq) was used to generate a genome wide catalog of transcriptional enhancers in the malaria vector *Anopheles coluzzii*. STARR-seq is a massively parallel reporter assay that provides the scale necessary for genome-wide screening and activity measurement of transcriptional enhancers. The approach is essentially a highly multiplexed version of the standard plasmid-based manual luciferase reporter assay that measures the transcriptional regulatory activity of a cloned nucleotide sequence. Briefly, in the STARR-seq screen, randomly sheared candidate genomic fragments are cloned into the 3' UTR of an irrelevant reporter gene in a plasmid with a basal promoter, resulting in the candidate sequence being transcribed as part of the reporter gene transcription unit. Candidate fragments that possess enhancer activity stimulate increased transcription from the basal promoter, which

generates a transcript of the reporter gene as well as the inserted candidate enhancer sequence. Rather than measuring luciferase activity, here RNA-seq is used to detect, in a massively parallel way, all of the reporter transcripts generated from the cloned genomic library transfected into cells. Trimming the flanking reporter sequence from the transcripts yields the RNA-seq reads originating from the collection of genomic fragments that were cloned in the plasmid library, which are then mapped to the reference genome assembly. The RNA-seq reads, originally the randomly sheared genomic DNA library, tile across the genome, and the normalized counts of each mapped window as compared to the control reveals enrichment that defines genomic peaks indicating functional enhancers, as well as a quantitative measure of their level of enhancer activity by normalized sequence read counts.

This massively parallel screening approach evaluates the enhancer activity of nucleotide sequences removed from their native chromatin structure, and thus generates a comprehensive catalogue of enhancers that may be differentially active across different cell types and developmental times. STARR-seq has been implemented in *Drosophila melanogaster* (Arnold et al., 2014), plants (Benoit, 2020) and human (Arnold et al., 2013), but has not been used in mosquitoes. Complementary approaches examining chromatin accessibility including FAIRE-seq (Perez-Zamorano et al., 2017) and ATAC-seq (Ruiz et al., 2021) have been used in *Anopheles* mosquitoes, as well as experiments for computational prediction of enhancers (Asma and Halfon, 2019; Schember and Halfon, 2021).

A comprehensive understanding of mosquito regulatory biology and the regulation of gene expression will require a careful combination and cross validation of data generated using direct, indirect and *in silico* approaches. The 3,288 candidate enhancers identified in *A. coluzzii* will be valuable in linking genetic and phenotypic variation where efforts focused on the coding genome have not yielded answers, in finding new regulatory elements, and in informing the choice of target sites for chromosome editing while minimizing disruptive secondary effects upon gene regulation in the chromosomal domain. A greater understanding of mosquito regulatory networks will add important tools to the malaria vector control arsenal.

MATERIALS AND METHODS

Source of Mosquito Genomic Deoxyribonucleic Acid and Generation of Input Genomic Library

Genomic libraries were prepared using a modified version of the methodology previously presented (Arnold et al., 2013). Mosquito DNA was derived from field material collected from two localities in Burkina Faso from 2007–2009 (Riehle et al., 2011). Samples were collected as larvae and reared to adulthood before being typed for species by the SINE200 × 6.1 assay (Santolamazza et al., 2008). DNA from 60 *A. coluzzii* were pooled (at equal volume) and sheared using a S220 Focused-ultrasonicator (Covaris) to produce fragments ~800 bp-1kb.

After shearing, an Illumina TruSeq Nano DNA Library Prep Kit was used according to the manufacturer's instructions to end-repair, A-tail, and prepare two independent sequencing libraries differing only in the sequence index tag using Illumina TruSeq indices 5 (ACAGTG) and 19 (GTGAAA). Enrichment of adapter-ligated molecules was carried out by PCR in order to add cloning adaptors as follows: an initial denaturation at 95°C for 5 min 10 cycles of 98°C for 15 , 60°C for 30 s, and 72°C for 30 s followed by a final extension of 72°C for 5 min. Each reaction was comprised of 5 ng DNA template, 10 uM 2.5 ul STARR_Seq_PCR1_For primer (10 uM), 2.5 ul STARR_Seq_PCR1_Rev primer (10 uM), 25 ul 2x KAPA ReadyMix (KAPA Biosystems) and nuclease-free H_2O to a final reaction volume of 50ul. Primer sequences for this and all experimental procedures are provided in **Supplementary Table S1**. Amplification products were purified with AMPure XP beads (Beckman Coulter) and the libraries were quantified using a Quant-IT PicoGreen dsDNA assay (Thermo Fisher Scientific) and the resulting fragment size distribution was assessed using a Bioanalyzer (Agilent Technologies).

Cloning and Transfection of Genomic Libraries

Libraries were cloned into the screening vector, pSTARR-seq_fly, kindly supplied by Alexander Stark (available from AddGene as pSTARR-seq_fly, vector #71499). Prior to cloning, the vector was linearized by digestion with SalI and AgeI for 5 h at 37°C, gel purified using the QIAquick Gel Extraction Kit (Qiagen), and further purified and concentrated using the QIAquick PCR Purification Kit and MinElute PCR Purification Kit respectively (Qiagen). Final elution was carried out in nuclease-free H_2O.

In-Fusion Cloning was performed in a 10ul reaction containing 30 ng mosquito DNA (15 ng from each of the uniquely indexed library and 30 ng of linearized vector. In total, 54 In-Fusion reactions were prepared in order to ensure comprehensive cloning of the *A. coluzzii* genome. Reactions were ethanol precipitated in batches of 4-5 reactions, resuspended in nuclease-free H_2O and pooled to prepare a final stock that was used for transformations into MegaX DH10B T1R Electrocomp Cells (Invitrogen). Cells were electroporated (2.0 kV, 200 Ω, 25 uF) in cooled Gene Pulser cuvettes with 0.1 cm electrode gap size using the Gene Pulser Xcell Electroporation System (Bio-Rad), and recovered in 1 ml recovery medium. After 1 h recovery at 37°C/220 rpm, 5 transformations were used to inoculate 500 ml LB with ampicillin (1ug/ml) and incubated at 37°C shaking at 220 rpm until the OD_{600} reached 0.8–1.0. plasmid libraries were extracted using the Plasmid Plus Mega Kit (Qiagen), pooled, and quantified. A total of 60 transformations were prepared in order to sufficiently capture the library.

Pre- and Post-Cloning Sequencing

Library sequencing was carried out before and after cloning in order to ensure no loss of genome coverage due to a bottleneck in the cloning step. For the pre-cloning DNA, the following PCR reaction was set up to prepare the samples for sequencing: 5 ng

template DNA, 2.5 ul Multiplexing PCR Primer 1.0 (10 uM, AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT), p7 primer (10uM, CAAGCAGAAGACGGCATACGA), 25 ul 2x KAPA HiFi HotStart ReadyMix (KAPA Biosystems) and nuclease-free H_2O to a final reaction volume of 50 ul. This reaction was amplified using the following PCR conditions: 95°C for5 min., 10 cycles of 98°C for15 s, 58°C for30 s, 72°C for 30 s. and a final extension of 72°C for 5 min.

For the post-cloning DNA, the following PCR reaction was set up to prepare the samples for sequencing: 32.5 ng of template DNA (this DNA amount reflects the same number of insert molecules as the reaction described above, but accommodates the pSTARR-seq_fly vector), 2.5 ul Multiplexing PCR primer 1.0 (10 uM), p7 primer (10 uM), 25 ul 2x KAPA HiFi HotStart ReadyMix (KAPA Biosystems) and nuclease-free H_20 to a final reaction volume of 50 ul. This reaction was amplified using the following PCR conditions: 95°C/5 min, 10 cycles of [98°C/15 s, 58°C/30 s, 72°C/30s] with a final extension at 72°C/5 min.

The PCR reactions were cleaned up with 1x AMPureXP beads, resuspended in 30 ul Elution Buffer (EB) (10 mM Tris-HCl, pH 8.5). The final pooled sample was quantified using a Quant-IT PicoGreen dsDNA assay (Thermo Fisher Scientific) and the resulting fragment size distribution was assessed using a Bioanalyzer (Agilent Technologies). The libraries were denatured and diluted according to Illumina's guidelines and sequenced on the Illumina HiSeq 2,500 in high output mode using 2 × 125 bp reads.

Cell Culture and Transfection With Cloned Genomic Library

The STARR-seq enhancer screen was carried out in three biological replicates, where each replicate is defined as an independent instance of the entire pipeline including transfection of *Anopheles* 4a3A cells (Muller et al., 1999), cDNA and plasmid isolation, and Illumina sequencing, each of these steps as described below. 4a3A cells were maintained on Insect X-Press media (Lonza) supplemented with 10% heat inactivated Fetal Bovine Serum, at 27°C. No antibiotics were used. Cells were species-typed by a molecular diagnostic assay (Santolamazza et al., 2008) and determined to be derived from *A. coluzzii*, which was not yet described as distinct from *A. gambiae* at the time the 4a3A cell line was established (**Supplementary Figure S1**).

Across the three replicates, a total of 4.8×10^8 4a3A cells were prepared for transfection using Lipofectamine 3,000 Reagent (Invitrogen). Of these, ~4×10^8 cells (40 transfections) were used for RNA extraction, and ~8×10^7 cells (8 transfections) were used for plasmid extraction for the input control. Cells were seeded (1×10^7 cells/25 cm^2 flask) on day 1, transfected with 10ug library after 24 h (day 2), and RNA or plasmid was extracted after an additional 24 h post transfection (day 3). The transfection solution was prepared as a master mix, where for one transfection, a 'DNA mix' (140ul Opti-MEM [Thermo Fisher Scientific] + 10 ug plasmid library +20 ul P3000) and 'Lipofectamine 3,000 Mix' (140 ul Opti-MEM + 10 ul Lipofectamine) were prepared separately. Once the mixes were combined, the master mix was incubated at room temperature for 15 min, and the full volume added to a flask of cells.

RNA and Plasmid Isolation From Transfected Cells

Cells were scraped from the base of each flask. Cells from 10 flasks were pooled and pelleted by centrifugation (2000 g/2 min), washed in 1x phosphate buffered saline (PBS) and re-pelleted (2000 g/2 min). The supernatant was removed once more, and any excess PBS was aspirated from the pellet. Total RNA was extracted from the cell pellet using the RNeasy Midi Kit (Qiagen) with an on-column DNA digestion using the RNase-Free dnase Set (Qiagen). Isolation of mRNA was carried out using the Dynabeads mRNA Purification Kit following the standard protocol. The mRNA was eluted off the beads at 70°C for 2 min, concentration measured on a NanoDrop, and stored at −80°C following the addition of ribonuclease inhibitor, RNasin [40 U/ul] (Promega). Using cells collected in the same manner, plasmid DNA was isolated using the Plasmid Plus Midi or Mini Kits (Qiagen). Plasmid extractions were quantified and stored at −20°C.

Reverse Transcription

Reverse transcription was performed using SuperScript IV First-Strand cDNA Synthesis System (Invitrogen) on 5 ug mRNA per replicate. These reactions were carried out using 400 ng mRNA/reaction according to supplier instructions with the following modifications: a construct specific primer was used (RT_Rev, **Supplementary Table S1**, as described in (Arnold et al., 2013) and reactions (transcription reaction mix plus annealed RNA-primer mix) were incubated at 50°C for 10 min after which RNA was removed by the simultaneous addition of 1ul rnase H (2 U/ul) and 1 ul rnase A (10 mg/ml) (Thermo Scientific) per reaction, incubated at 37°C for 30 min. Finally, the cDNA was cleaned using QIAQuick PCR Purification Kit (Qiagen) and eluted in Elution Buffer (EB).

cDNA and Input Control Plasmid Amplification and Sequencing

Primers used for specific amplification of cDNA, Report_Fwd and Report_Rev, are as in (Arnold et al., 2013) and also in **Supplementary Table S1**. cDNA from each sample was amplified for sequencing in four separate PCR reactions containing 35 ng template each, according to the method of (Arnold et al., 2013). The following reaction composition was used for the cDNA amplifications included 35 ng template DNA, 25ul KAPA HiFi HotStart ReadyMix (KAPA Biosystems), 0.25 ul Report_Fwd primer (100uM), 0.25 ul Report_Rev primer (100uM). These reactions were amplified using the following PCR conditions: 98°C/45 s, 15 cycles of [98°C/15 s, 65°C/30 s, 72°C/70 s]. Next, the PCR reactions were cleaned up with 1x AMPureXP beads and resuspended in 20 ul EB (10 mM Tris-HCl, pH 8.5), and a second PCR was performed to generate the final sequencing libraries. For each of the four reactions per sample, all

20ul of template from the cleaned-up initial cDNA PCR was used to set up the following PCR reactions: 20 ul DNA template, 25 ul KAPA HiFi HotStart ReadyMix (KAPA Biosystems), 2.5 ul D50X forward indexing primer (10 uM), 2.5 ul p7 primer (10 uM). The reactions were amplified using the following PCR conditions: 95°C/5 min, 10 cycles of [98°C/15 s (sec.), 58°C/30 s, 72°C/30 s] with a final extension at 72°C/5 min. The four reactions per sample were pooled, cleaned up with 1x AMPureXP beads, resuspended in 30 ul EB (10 mM Tris-HCl, pH 8.5). The final pooled sample was quantified using a Quant-IT PicoGreen dsDNA assay (Thermo Fisher Scientific) and the resulting fragment size distribution was assessed using a Bioanalyzer (Agilent Technologies). The libraries were denatured and diluted according to Illumina's guidelines and sequenced on the Illumina HiSeq 2,500 in high output mode using 2 × 125 bp reads.

Plasmid control samples were amplified and sequenced in the same manner as the cDNA samples described above with the exception that the primers used for specific amplification of the plasmid template were Plasmid_Fwd and Plasmid_Rev, (Arnold et al., 2013), and listed in **Supplementary Table S1**.

Sequence Analysis, Enhancer Peak Calling and Enrichment

The quality of sequencing reads was tested using FastQC version 0.11.5 (Wingett and Andrews, 2018). Following quality control, reads were mapped against the *Anopheles gambiae* AgamP4 reference genome assembly using BWA MEM (Li and Durbin, 2009) (version 0.7.7) with default parameters. Samtools version 1.6 (Li et al., 2009; Danecek et al., 2021) was used to select only the properly mapped reads (option -f 2), to filter supplementary alignments (option -F 2048) and to convert the sam files to bam files.

Prior to peak calling, reads from sequencing libraries run across multiple sequencing runs were merged and duplicate reads were removed from merged libraries, resulting in one sequencing library per biological replicate, which was then subjected to peak detection. Peak calling to detect significant enrichment in cDNA reads as compared to plasmid DNA reads was performed with R version 3.4.1 (R Core Team, 2016) using the package BasicSTARRseq version 1.4.0 (Buerger, 2015) with the function getPeaks and the following parameters: minQuantile = 0.99, peakWidth = 500, maxPval = 0.001, deduplicate = F and model = 2. For this step, reads mapped to the Y, UNKN and mitochondrial chromosomes were not analyzed. Only peaks displaying sequence read enrichment (cDNA reads/plasmid DNA reads) ≥ 3 were retained (**Supplementary Figure S2**). The common peaks across the three biological replicates were called with the function findOverlaps from the GenomicRanges package (Lawrence et al., 2013) with a minimal overlap set to 250.

Genome Location of Enhancer Peaks

For each of the 3,288 enhancers, UTR, exon and mRNA scores corresponding to the fraction of the enhancer overlapping each genomic feature were calculated using bedtools intersect (Quinlan and Hall, 2010) (v 2.26.0). For enhancers overlapping multiple features of the same type (e.g., two different exons), only the larger score was retained. Then, the bedtools closest tool was used to determine the gene in closest proximity to the left and right of the candidate enhancer. A distance equal to zero indicates an overlap between the gene and the enhancer. Enhancers with no overlap with an annotated gene were classified as "intergenic" while those overlapping a UTR region or a CDS feature (intron or exon) were assigned to the corresponding category (enhancers overlapping both features, CDS and UTR, were assigned to the feature with the higher score). For enhancers overlapping the mRNA feature but no CDS or UTR features, bedtools closest was used to determine the intron in which they are located.

Candidate Validation by Manual Luciferase Reporter Assays

PCR amplicons spanning candidate enhancer regions were amplified from the DNA pool of 60 *A. coluzzii* mosquitoes used for construction of the library. Resulting PCR fragments were cloned into the pCR8/GW/TOPO vector (Invitrogen) The PCR fragments were Gateway cloned into pGL-Gateway-DSCP, kindly supplied by Alexander Stark (available at AddGene, vector # 71,506) using Gateway LR clonase II (Invitrogen) as per the manual. Gateway clones were transformed into OneShot OmniMax 2T1 Phage-Resistant Cells, grown up overnight, plasmid purified (PureLink Quick Plasmid Miniprep Kit, Invitrogen) and sequenced with the primers LucNrev 5′ CCT TATGCAGTTGCTCTCC and RVprimer3 5' CTAGCAAAA TAGGCTGTCCC to verify the insert sequence. Primers used for amplification of candidate enhancers and negative control fragments are available in **Supplementary Table S1**.

Transfections were prepared in 96 well plates. *A. coluzzii* 4a3A hemocyte-like insect cells were seeded at 1×10^5cells/well, in a total of 65ul with growth media. Cells were gently agitated on a MixMate (Eppendorf) for 30 s at 350 rpm to ensure even distribution and incubated for 24 h at 27°C. Transfections were carried out using Lipofectamine 3,000 (Invitrogen) and 2 vectors for dual luciferase assays: a renilla control vector pRL-ubi-63E (AddGene, #74280), and the pGL-Gateway-DSCP (described above) carrying a single amplified haplotype of the candidate enhancer upstream of a firefly luciferase gene. Renilla and firefly plasmids were transfected at a ratio of 1:5 (renilla:firefly) which equated to 18 ng renilla and 90 ng luciferase vector in 10ul volume per well. Plates were agitated for 30 s at 350 rpm on a MixMate (Eppendorf) to ensure mixing and incubated for 24 h at 27°C. The Dual-Glo luciferase Assay System (Promega) was used for luciferase assays, according to supplier instructions. Measurements were recorded on the GloMax Discover (Promega) at 25°C, with two 20 min incubations, one after the addition of Dual-Glo luciferase reagent and another after the addition of Stop & Glo reagent. All test plates contained a previously reported negative control fragment which was a size matched fragment within intron 1 of AGAP007058, and a highly active positive control enhancer peak nearby in AGAP008980 (Nardini et al., 2019). All samples were run in 6-fold replication within a single plate and across at least two

independent plates for at least two biological replicates. In order to eliminate any effects of evaporation, only the 60 internal wells of the plate were used for transfection and measurement of luciferase activity. The external 36 wells were filled with cell media to a volume matching the internal wells. Firefly luciferase measurements (relative light units, RLUs) were corrected against the renilla measurements for the same well. These measurements were then normalized to the firefly/renilla mean for the negative control on the same plate to combine results across replicates. Luciferase activity was compared using a non-parametric ANOVA (Kruskal-Wallis) with post hoc pairwise comparisons.

To test the enhancer activity in multiple cellular backgrounds, a subset of enhancers was screened for activity in Ag55 cells, an *A. coluzzii* first-instar larval cell line (Pudney et al., 1979). Like the 4a3A cell line, the Ag55 line was originally described as *A. gambiae*, but molecular diagnostics confirm that, like 4a3A, it is also derived from *A. coluzzii* (**Supplementary Figure S1**). Cells were grown as described previously (Hire et al., 2015) and luciferase assays were performed as described above. Relative light unit output was compared between 4a3A and Ag55 cells using 2-way ANOVA followed by pairwise comparison.

Genomic Sequence Characteristics of Candidate Enhancers

All 3,288 candidate enhancers were analyzed computationally for underlying sequence composition. For each candidate enhancer peak a matching control sequence was chosen by randomly selecting a window of the same size within 0.5 Mb on the same chromosome. When the randomly chosen window overlapped with either an exon or a candidate enhancer peak or when >1% of the sequence were unknown nucleotides (N's), a new random sample was generated. This process was repeated 10 times for each peak, resulting in 10 control data sets, each containing 3,288 genome wide control fragments.

GC Content. GC content was calculated for each sequence in the control and candidate enhancer sets. Distributions of the GC content were compared across the control sets using a Kruskal-Wallis. If the 10 sets of control sequences were not statistically distinguishable at $p < 0.001$, a second Kruskal-Wallis test was run on the 10 control sets plus the treatment set (11 groups total) using the same significance level. If a statistically significant difference was observed, pair-wise post-hoc tests between each control set and the treatment set were performed with non-parametric Mann-Whitney U tests at $p < 0.001$. Finally, the GC content was considered significantly different if all 10 post-hoc tests were statistically significant and all controls were changed in the same direction (all increased or all decreased) relative to the candidate enhancers.

Perfect Repeat Analysis. Candidate enhancers and control fragments were analyzed for the presence of perfect mononucleotide (e.g., AAAAAA), dinucleotide (e.g., ATATAT), and trinucleotide (e.g., ATCATC) repeats. Repeats were required to be ≥ 6 nucleotides in length; 6 repeat units in the case of a mononucleotide repeat, 3 for a dinucleotide repeat and 2 for a trinucleotide repeat. Mononucleotide repeats were excluded from analyses of dinucleotide and trinucleotide repeats and dinucleotide repeats excluded from the analysis of trinucleotide repeats. The analysis was performed in two stages. First the fraction of the 3,288 sequences containing each type of repeat was examined. For each sequence repeat (e.g., TTTTTT or ACACAC), ten sets of control fragments were also examined. Kruskal-Wallis tests were used to compare perfect repeat presence across the 10 sets of control fragments. If the 10 groups of control fragments were not statistically distinguishable at $p < 0.001$, an additional Kruskal Wallis test was run using the 10 control sets and the candidate enhancer set (11 groups in total) with the same significant level. It the Kruskal Wallis test detected a significant difference among the 11 groups, pair-wise post-hoc tests between each control set and the candidate enhancer set were performed with non-parametric Mann-Whitney U tests as $p < 0.001$. Finally, a given repetitive element was considered significantly different if all 10 post-hoc tests were statistically significant and all control averages were changes in the same direction (all increased or all decreased) relative to the candidate enhancer average.

Secondly, for each specific repeat, control and treatment sequences containing the specific repeat were further analyzed examining repeat counts (the number of repeat occurrences per fragment), repeat length distribution, and the total fraction of the sequence involved in the repeat. The same statistical testing and filtering criteria describe above for repeat presence were applied in the examination of these attributes.

Enrichment of the four dinucleotide repeats CA, GA, GC, and TA and their reverse complements were calculated for candidate enhancers and the non-coding control sequences. GC and TA repeats are their own reverse complements and were counted twice. Enrichment of the dinucleotide repeat was calculated as $\log_2$ (fraction of enhancer sequences containing the dinucleotide perfect repeat/fraction of control sequences containing the same dinucleotide perfect repeat). As with work done in fruit flies and humans, these repeats needed to be at least 6bp in length. The analysis was repeated 10 times (once against each control) and results used to calculate means and standard deviations.

Transcription Factor Binding Site Motifs. Transcription factor binding sites (TFBS) were identified using 143 JASPAR 2020 core insect TFBS motifs from *D. melanogaster* (Fornes et al., 2020). Motifs were identified using the default paraments in AME (McLeay and Bailey, 2010) from MEME Suite (Bailey et al., 2015). The treatment sequences were tested against each set of control sequences for a total of 10 independent runs. The ten independent comparisons were assessed and only those motifs enriched in at least 8 of the 10 AME comparisons of candidate enhancers with controls were considered enriched.

In addition to comparing to known TFBS motifs from *D. melanogaster*, we also identified motifs enriched in candidate enhancers as compared to the 10 controls. Motif discovery was performed with STREME (--minw 6, otherwise default parameters) (Bailey, 2021)), also from MEME Suite. Discovered motifs were assessed for enrichment based on the same criteria described above (enriched relative to at least 8 of 10 control sets) using AME. To determine which of these discovered motifs were also present in JASPAR, the motifs were aligned with the 143 JASPAR motifs using TOMTOM (Gupta et al., 2007)), also from MEME Suite, using the default parameters.

Lastly, to determine if repetitive content differed between *de novo* discovered motifs as compared to those that align to JASPAR motifs, the nucleotide repeat content of the motif matches was compared. Repeats in the enhancers were defined and identified as described above and match coordinates were written to a BED file. The 36 discovered motifs were divided into the 22 that matched JASPAR motifs and the remaining 14 *de novo* motifs. The enhancer sequences were searched for motif matches using FIMO (Grant et al., 2011), from MEME Suite, and match coordinates were converted to BED files. Overlapping match regions in each BED file were merged using BEDtools (bedtools merge, default parameters). For each sequence, the number of nucleotides involved in a motif match and the number of nucleotides involved in both a motif match and a repeat match were counted and used to calculate the fraction of motif matched nucleotides that were repetitive per sequence.

Comparison of Candidate STARR-Seq Enhancer Peaks With ChIP-Seq and ATAC-Seq Peaks

H3K27ac, H3K4me3, H3K9ac, and H3K9me3 ChIP-seq peak boundaries from uninfected *A. gambiae* samples were downloaded from **Supplementary Table S3** of (Ruiz et al., 2019) and saved as four BED files containing the chromosome and peak position for each type of ChIP-seq experiment. ATAC-seq peaks for salivary glands and midguts from malaria-infected *A. gambiae* (Ruiz et al., 2021) were downloaded from NIH GEO repository GSE152924. The ATAC-seq peaks were separated by tissue and saved as BED files containing the chromosome, and peak position information. The ATAC-seq peaks were generated from multiple separate samples and contained overlapping peaks. Peaks from the two ATAC-seq data sets were merged using BEDTools (bedtools merge, default parameters), which reduced the number of peaks from 93,921 (midguts) and 99,084 (salivary glands) to 68,016 and 58,515, respectively. Coordinates of 20,578 computationally predicted enhancers were downloaded from **Supplementary Table S1** of (Schember and Halfon, 2021) and saved as a single BED file. The coordinates were merged using BEDTools (bedtools merge, default parameters), resulting in 9,861 computationally predicted enhancer features. Overlaps between the ChIP-seq/ATAC-seq/computationally predicted peaks and STARR-seq peaks were performed using BEDTools (bedtools intersect -f 0.1 -r -u), which outputs each peak with 1 or more matches.

RESULTS

Implementation of a Transcriptional Enhancer Screen in *Anopheles Coluzzii*

Methods developed for a massively parallel functional enhancer screen developed in *D. melanogaster* were modified and optimized for use in *A. coluzzii*. Efficiency of the screen was evaluated at multiple quality control steps to verify genome representation. First, the sheared and adaptor-ligated genomic DNA library was Illumina sequenced prior to cloning to measure genome coverage (**Supplementary Figure S3A**). Second, the library was Illumina sequenced after cloning into the STARR-seq reporter plasmid, but prior to transfection of 4a3A cells, to control for effects of cloning upon representative genome coverage (**Supplementary Figure S3B**), and sequenced again after transfection of three replicates and growth in 4a3A cells for 24 h, to control for culture effects on representation (**Supplementary Figure S3C**). The library retained high-quality genome coverage after each of these experimental steps. Median genome coverage was equivalent in the post-cloning sequence (median = 68) and the post-transfection sequences across three biological replicates (average median of 71, median range 66–78). Finally, the cDNA generated from the STARR-seq reporter plasmid displayed lower median genome coverage than the entire genome sequences above (**Supplementary Figure S3D**), which is consistent with the small fraction of the genome represented by enhancers. These sequencing controls indicate that the cloned reporter plasmid library of randomly-sheared *A. coluzzii* genomic fragments was representative and sufficiently powered to query the entire genome for detection of candidate enhancers.

A Comprehensive Genome-Wide Map of Candidate Transcriptional Enhancers

Three biological replicates of the enhancer screen yielded a total of 3,288 candidate transcriptional enhancers present in all three replicates, hereafter referred to as the candidate enhancers. These candidate enhancers are located across the genome without enrichment on any particular chromosome arm (**Figure 1A**; chi-square = 1.170, df = 4, p = 0.88). The majority of candidate enhancers occur in either intergenic or intronic regions of the genome, with a small fraction overlapping exons (**Figure 1B**). A majority of the intronic candidate enhancers are found in the first intron (**Figure 1B**), similar to at least *D. melanogaster* (Arnold et al., 2013). The candidate enhancers vary in strength of activity, as measured by the normalized cDNA read depth values (**Figure 1C**), a distribution that does not differ by chromosome arm (**Figure 1D**). The complete catalog of candidate enhancers with position and read enrichment information is available in tabular format (**Supplementary Table S2**) as well as a graphic format (example in **Supplementary Figure S4**) generated using the Integrative Genomics Viewer tool (Robinson et al., 2011).

Experimental Validation of Candidate Enhancers

Manual luciferase reporter assays were used to test a validation panel comprised of 28 candidate enhancers and 9 size-matched control fragments that did not display enhancer activity in the STARR-seq screen, and thus were predicted to not be enhancers. The candidates in the validation panel were selected solely on the criterion of location, with one candidate approximately each 10 MB across the genome (candidate coverage: chromosome X, 3; chromosome arm 2R, 7; chromosome arm 2L, 6;

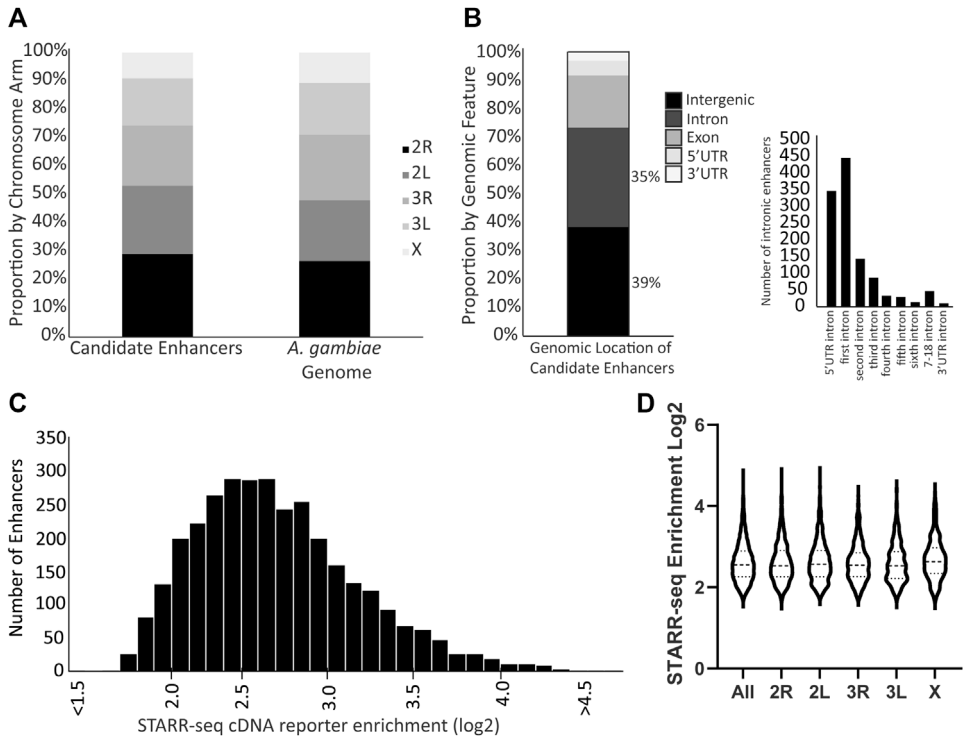


FIGURE 1 | Genome distribution and activity of *A. coluzzii* candidate enhancers. **(A)** Candidate enhancers display a distribution across chromosome arms that equivalent to the total genome composition of chromosome arms. The *A. gambiae* reference genome assembly is used throughout because there is no chromosome-level assembly for *A. coluzzii*. **(B)** 74% of candidate enhancers are located in either intergenic or intronic regions. Of the 1,156 enhancers located in introns, nearly 70% are located in either the intron between the 5′UTR and the first coding exon (5′ UTR intron), or between the first and second coding exons (first intron). **(C)** Candidate enhancers vary in their activity levels as measured by the enrichment of reads of the enhancer sequence transcribed from the reporter plasmid (cDNA reporter) as compared to the same sequence in the control plasmid DNA samples. A $\log_2$ enrichment value of 2 represents a 4-fold increase in normalized read depth, $\log_2$ enrichment of 3 is an 8-fold increase and $\log_2$ enrichment of 4 a 16-fold increase. **(D)** Violin plot indicates that there is no significant difference in distribution of candidate enhancer strength across chromosome arms (middle horizontal lines represent median enrichment, ANOVA F = 1.631, p = 0.15).

chromosome arm 3R, 7; chromosome arm 3L, 5). Control fragments were distributed to include one on chromosome X and two on each autosomal chromosome arm.

When tested by manual luciferase reporter assays, over ninety percent (93%, 26 of 28) of the tested candidate enhancers displayed higher average relative luciferase activity than any of the 9 tested negative controls (**Figure 2A**). All of the nine negative control fragments that were not enriched in the enhancer screen also did not display luciferase activity above background. These results indicate that the screen efficiently identified genome-wide enhancers in *A. coluzzii*. Researchers should specifically confirm specific candidate enhancers of interest before use.

The quantitative levels of enhancer activity of the validation panel, measured by manual luciferase reporter assays, were compared to the cDNA sequence enrichment values for the same enhancers as detected in the STARR-seq screen (**Figure 2B**). Manual luciferase activity values for an enhancer (*y*-axis, relative luciferase activity) were significantly correlated with STARR-seq cDNA enrichment values ($r^2 = 0.68$, $p < 0.001$). Enhancers displaying the greatest cDNA enrichment also display higher activity in the manual luciferase reporter assay. Thus, the enrichment measurements in the STARR-seq assay were not only a qualitative method suitable for identification of transcriptional enhancers, but were also accurate quantitative measurements of enhancer activity levels.

Enhancer Activity in an Independent Cell Line

Enhancer activity for six candidate enhancers and one negative control was also tested in a second independently derived *A. coluzzii* cell line, Ag55. At a qualitative level, the results were independent of cell line, because fragments that displayed enhancer activity above background in 4a3A cells also did so in Ag55 cells. At a quantitative level, three enhancers showed statistically indistinguishable activity levels across cell lines, 2 enhancers had significantly higher activity in 4a3A cells and one enhancer displayed higher activity in Ag55 cells (**Figure 3**).

Genome-wide Enhancer Catalog in Genome Browser Format

The list of the 3,288 detected candidate enhancer peaks is available in **Supplementary Table S2**. BED files (Datasheet 5 & Datasheet 6) are also provided that allow the visualization of detected peaks alongside the gene models currently available for the PEST genome

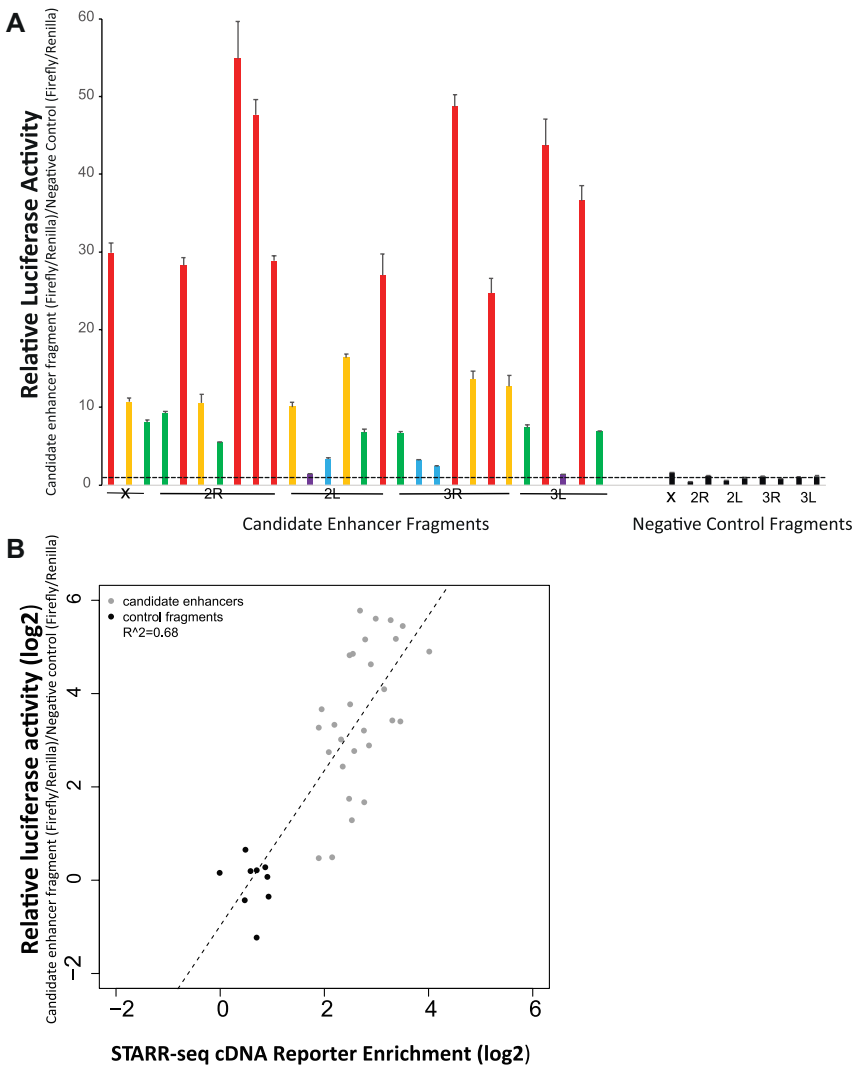


FIGURE 2 | Experimental validation by manual luciferase reporter assays confirms accuracy of the genome-wide enhancer screen. **(A)** Twenty-eight candidate enhancers randomly chosen at ~10 Mb spacing across the genome were tested for enhancer activity with a luciferase assay. Twenty six of the 28 displayed greater average activity than the nine negative controls, which were not predicted to be enhancer peaks in the genome-wide screen. None of the nine negative controls displayed measurable activity above background. Relatively luciferase activity was measured across two biological replicates, with six technical replicates per biological replicate. Relative activity of candidate enhancers is indicated by color: violet, relative luciferase activity <2, blue, activity 2-5, green, activity 5–10, yellow, activity 10–20, and red, activity >20. Negative control values are indicated in black. The horizontal dotted line indicates relative luciferase activity of 1, equal to that of the internal negative control used across all plates as previously (Nardini et al., 2019). **(B)** The correlation for validation panel enhancers between enrichment of cDNA reads in the genome-wide screen (*x*-axis) and relative luciferase activity in manual luciferase reporter assay (*y*-axis) indicates a significant relationship ($p < 0.001$) between the genome-wide screen and manual luciferase reporter assay, indicating the quantitative accuracy of the massively parallel screen for detection of relative activity levels of enhancers.

assembly of *A. gambiae*. To visualize and search the genome wide map of enhancers open the above BED files (**Supplementary Figure S4**) in Integrative Genomics Viewer (Robinson et al., 2011).

Comparative Sequence Analysis of Enhancers and Non-Coding Control Fragments

The 3,288 functionally detected enhancers from the screen together occupy a total of approximately 1.85 Mb, or 0.7% of the *A. coluzzii* 281.38 Mb genome. We evaluated the enhancer compartment in comparison to matched genomic controls in order to identify distinct composition features and sequence patterns of the enhancer compartment of the genome that would be expected to underlie the functional properties of the *A. coluzzii* enhancers. The controls were ten independent genomic control sets extracted from the fraction of the genome that did not display enhancer activity peaks in the genome-wide screen. The control sets were matched to the enhancer compartment for equivalent fragment size and total size as the summed genomic enhancers.

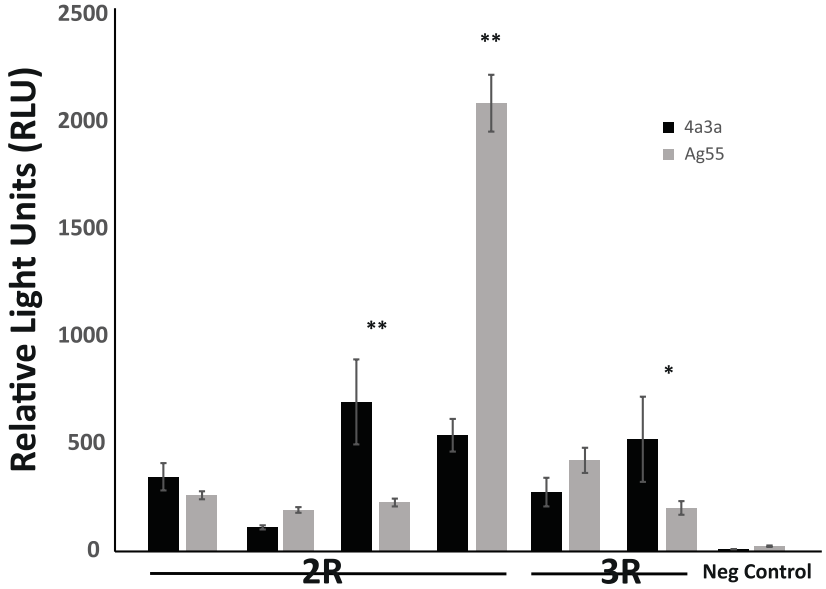


FIGURE 3 | Enhancer activity is qualitatively consistent in independent cell lines, with quantitative differences. Activity of a panel of enhancers was measured in *A. coluzzii* cell lines 4a3A and Ag55 using dual glow luciferase assay with relative light units normalized to renilla. All fragments that displayed significant activity above baseline in 4a3A cells, also did so in Ag55 cells and the negative control fragment showed no activity in either cellular environment. Of the 6 tested enhancers, 3 displayed statistically indistinguishable activity in 4a3A and Ag55 cells, while the other 3 were differentially active (*$p < 0.05$, **$p \leq 0.001$).

GC Content Analysis. The enhancer sequence compartment displayed significantly higher GC content than the non-enhancer controls (**Supplementary Figure S5**, enhancer compartment 49.9 ± 4.5%, control set 43.8 ± 6.8%, Mann-Whitney U p-value = 1.6×10^{-304}). In addition to being more GC-rich relative to control sequences, the enhancer compartment displayed smaller variation in GC across the set of candidate enhancers.

Perfect Repeat Analysis. Perfect mononucleotide, dinucleotide, and trinucleotide repeats of 6 or more nucleotides in length were analyzed. As a whole, there were no significant differences in the types of repeats found in candidate enhancer sequences as compared to genome matched control sequences (**Figures 4A,B**). There were, however, specific repeats that were over- or under -represented amongst enhancer sequences as compared to controls. Specifically, repeats rich in A or T (A, T, AT, TA, AAT, ATA, ATT, TAA, TAT, TTA) were significantly enriched in all control data sets in terms of repeat presence. Conversely, those rich in G and C (AC, CA, CG, CT, GC, GT, TG, ACG, AGC, CAC, CAG, CCG, CGA, CGC, CGG, CGT, CTG, GCA, GCC, GCG, GCT, GGA, GGC, GGT, GTG, TCC, TCG, TGC) were significantly enriched in the enhancer sequences. Enrichment in GC rich repeats is consistent with the difference in overall GC content presented above.

Enrichment of CA, GA, GC, and TA dinucleotide repeats was analyzed in candidate enhancer sequences. *A. coluzzii* enhancers detected in the current screen displayed an enrichment of CA, GA, and GC repeats and depletion of TA (The GA enrichment in candidate enhancer fragments was statistically significant in 8 of 10 post-hoc tests but did not meet the criteria used above for enrichment). GC was enriched at two times the rate of GC or CA (**Figure 4C**). The CA repeat was present in 42.4% of enhancer sequences compared to only 33.9 ± 0.7% of controls ($p = 6.3^{*}10^{-58}$). For the GA repeat, these numbers were 20.0% in candidate enhancer sequences and 15.6 ± 0.4% in controls ($p = 2.9^{*}10^{-26}$) and for GC, 21.8% in candidate enhancers and 12.8 ± 0.5% in controls ($p = 2.4^{*}10^{-108}$). Conversely with the TA repeat, only 4.7% of enhancer sequences contained this sequence while 10.5 ± 0.4% of control sequences did ($p = 3.4^{*}10^{-65}$).

Transcription Factor Binding Sites (TFBS) Definitions of 143 TFBS from the JASPAR CORE 2020 insect release were tested for enrichment in the enhancer compartment as compared to genome matched controls. Each definition was tested in 10 iterations, once against each control. TFBS that were enriched in at least 8 of 10 tests were deemed enriched. In total, 15 of the 143 TFBS tested were enriched in candidate enhancer sequences as compared to control sequences by our criteria (**Figure 4D**).

In addition to searching against the TFBS annotated in JASPAR, motif discovery was also performed. A total of 36 motifs were discovered, all of which were enriched according to our criteria (**Supplementary Table S3**). Twenty-two of the 36 discovered motifs aligned with one or more of 143 JASPAR motifs, including all 15 JASPAR motifs found enriched in the above analysis. Fourteen of the discovered motifs did not match any of the 143 JASPAR motifs and potentially represent uncharacterized TFBS active in the 4a3A cells. The *de novo* motif content was more repetitive (35.6 ± 24.9%; Rank Sums test, $p = 2.2^{*}10^{-37}$) than the JASPAR motif content (27.9 ± 17.8%), suggesting that not all enriched *de novo* motifs may be functional binding sites given the lack of complexity of the highly-

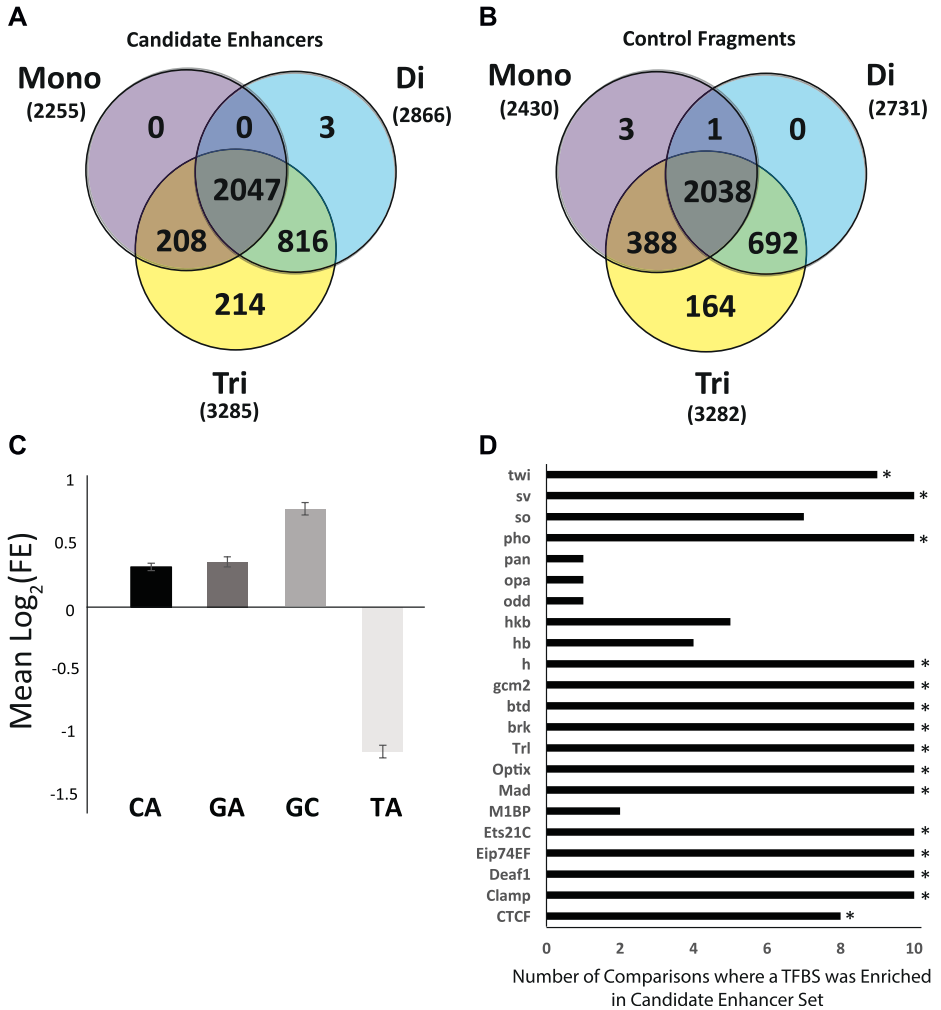


FIGURE 4 | Sequence characteristics of the *Anopheles coluzzii* enhancer compartment as compared to control sequences. The enhancer compartment represents ~0.7% of the total genome sequence. **(A)** Venn diagram of the number of candidate enhancers that contain perfect mono, di and tri nucleotide repeats, defined as sequences ≥6 bp in length (e.g., AAAAAA, ATATAT, ATCATC). **(B)** As in **(A)**, but for 3288 control fragments not detected as enhancers by the genome-wide screen (two control sequences lacked any repetitive sequence meeting the definition described and are not displayed). **(C)** Mean $\log_2$ of the fold-enrichment (FE) in perfect dinucleotide repeats between candidate enhancer sequences and controls. Numbers greater than 0 indicate enrichment in candidate enhancer sequences and numbers less than 0 indicate enrichment in control sequences. **(D)** Sequences of candidate enhancers and controls were searched for predicted transcription factor binding sites (TFBS) using AME from MEME Suite and enrichment of computationally predicted TFBS was computed. Graph indicates the number of control data sets where the TFBS was depleted as compared to candidate enhancers. Fifteen TFBS (indicated by *) were enriched in the candidate enhancer sequences above the defined threshold (enhancer sequences were tested against each of 10 control sequence sets for a total of 10 runs, threshold for motif enrichment was a positive test in at least 8 of 10 runs).

repetitive sequence (**Supplementary Figure S6**), but may serve other functions that otherwise influence enhancer activity.

Comparison of Candidate Enhancers With ChIP-Seq, ATAC-Seq and Computationally Predicted Peaks

The *A. coluzzii* candidate enhancers detected by STARR-seq were compared with peaks from published ChIP-seq and ATAC-seq analyses in *A. gambiae* and with computational enhancer predictions (**Supplementary Table S4**). The ChIP-seq data detected four histone modifications (H3K27ac, H3K9ac, H3K9me3 and H3K4me3) associated with enhancers, promoters, and gene activation and gene silencing, respectively, in a comparison of mosquitoes infected with the malaria parasite *Plasmodium falciparum* or uninfected controls (Ruiz et al., 2019). The largest overlap with STARR-seq candidate enhancers occurred with the 19,321 H3K27ac ChIP-seq peaks, which are indicative of active enhancers and promoters, because 22.8% of the 3,288 *A. coluzzii* STARR-seq peaks overlapped with the H3K27ac ChIP-seq peaks. STARR-seq candidate enhancer peaks were also compared with data collected by ATAC-seq on

midguts and salivary glands from malaria-infected *A. gambiae* samples (Ruiz et al., 2021). Of the 3,288 *A. coluzzii* STARR-seq peaks, 33.6 and 35.3% overlapped with ATAC-seq peaks identified from midguts and salivary gland samples, respectively. Schember and Halfon (Schember and Halfon, 2021) performed *in silico* computational prediction of enhancers with training sets derived from *D. melanogaster* and organized by active tissues. After merging, 9,861 unique peaks were computationally predicted to be enhancers, and 6.8% of the STARR-seq candidate enhancer peaks overlapped these predicted enhancers. Across all of the data sets, most of the overlaps were one-to-one, because a single STARR-seq peak overlapped with a single peak identified by another method. The partial overlap between the different data sets may tend to support the candidate enhancers identified by STARR-seq, but further work will be required to determine the proportion of ChIP-seq and ATAC-seq peaks that represent functional enhancers, as compared to other categories of chromatin features. Indeed, the STARR-seq data could serve as a training set to refine computational models, which are built upon training data.

DISCUSSION

We present a genome-wide map of 3,288 transcriptional enhancers identified by screening in wild samples of the African malaria vector mosquito, *A. coluzzii*. This work generates a resource that will help make the functional non-coding regulatory DNA accessible to study in this mosquito. Manual testing of a validation panel from the genome-wide catalog demonstrated the high accuracy of the screen for enhancer detection, as well as the agreement of quantitative enhancer activity levels between the highly multiplexed genome-wide screen and values obtained for the same enhancers from individual manual tests. The identified enhancers in *A. coluzzii* display transcriptional enhancing activity levels over a wide dynamic range of at least two-fold above baseline and up to at least 16-fold higher activity. These results demonstrate that genomic regions with significant transcriptional regulatory potential were detected and that the quantitative measurement of their activities by the genome-wide screen are likely to correspond to their biological activity levels *in vivo*.

The two methods used in this study for enhancer detection and quantitation, cDNA enrichment in STARR-seq or manual luciferase reporter assays, share the property that, as ectopic plasmid-based assays, they measure the innate sequence-based enhancer activity of the candidate fragment within the cellular environment, but without the influence of chromatin and its chemical modifications. The ectopic plasmids carrying the candidate genome fragments are not chromatinized, while enhancers in the chromosomal environment are modulated by natural chromatin. Thus, the genome-wide enhancers detected in the current study represent a catalog of genomic enhancers expected to be active in diverse cell types or developmental stages of the organism in which the same transcription factors are active, while only a subset of enhancers are expected to be biologically active in a specific cell type and/or developmental stage. Cross validation of a small panel of enhancers in a second *A coluzzii* cell line indicates that the tested enhancers were active in both cell lines. This observation suggests that, at least for this test panel, the cellular machinery of transcription factors and binding partners for transcriptional regulation is comparable among two independently-derived *A. coluzzii* cell lines from the same tissue origin. Three of the six enhancers tested displayed different levels of positive activity among the 2 cell lines. The most likely explanation is that there are abundance differences in the cell lines of transcription factors and partners relevant to each specific enhancer. However, the importance of the activity differences for natural mosquito biology, if any, should be interpreted with caution. The activity differences between cell lines could result from capturing natural allelic variation for transcription factor abundance in the 2 cell lines, which would be interesting, but could also be an *in vitro* artifact caused by genetic drift and/or slightly divergent adaptation of the 2 cell lines to the culture environment. Enhancers with cell-type or developmentally restricted activity result from the different combinations of transcription factors expressed in a given cell type (Kheradpour et al., 2013). Biologically active enhancers are found in open chromatin, which exposes them to binding by transcription factors and other factors, while closed chromatin obscures the enhancers that are not active in the specific cell such that they are not accessible for transcription factor binding.

The subset of enhancers that are biologically active in a given cell can only be detected by indirect means in nuclei isolated from a pure sample of the target cell type or stage. Previous work on mosquito non-coding regulatory elements has employed chromatin-based approaches (Behura et al., 2016; Perez-Zamorano et al., 2017; Mysore et al., 2018; Ruiz et al., 2019; Lezcano et al., 2020; Nowling et al., 2021; Ruiz et al., 2021). However, for phenotypic studies, pure samples of the relevant cell type may not be feasible to obtain. Moreover, during initial exploration of a phenotype, the relevant cell type or timing may not be known, or multiple sites may be responsible. Thus, use of the current comprehensive catalog of *A. coluzzii* enhancers in phenotype studies can be combined with subsequent chromatin-based cell-specific assays once the target cell type is known, in order to clearly define the genome regulatory space underlying the phenotype. Previous work has focused on computational prediction of enhancers in *Anopheles* (Schember and Halfon, 2021) and in highly diverged insects (Kazemian et al., 2014). Integration of the current genome-wide comprehensive catalog with chromatin-based cell and developmental data can be used to train algorithms which may be able to more efficiently predict transcriptional enhancers from sequence context.

STARR-seq is a method that determines and quantifies the ability of a DNA sequence to enhance transcription in a chromatin-independent episomal context, while ChIP-seq and ATAC-seq determine biologically active chromatin features, including but not specifically limited to enhancers. We demonstrated a degree of overlap of candidate enhancers functionally detected using STARR-seq with *Anopheles* chromatin features. Almost all (93%) of the STARR-seq enhancer peaks displayed enhancer activity by manual standard luciferase assays, while just 26% of ENCODE

enhancer predictions made using ChIP-seq displayed regulatory activity (Kwasnieski et al., 2014). Thus, chromatin modifications have the potential to indicate enhancer sequences in the genome, but direct activity assays that measure transcriptional enhancing activity from candidate containing TFBS sequences are more sensitive and precise for identifying enhancers. The combination of functional assays such as STARR-seq coupled with chromatin-based methods will be essential in the dissection of regulated gene expression.

Computational analyses of the current enhancer catalog revealed relative enrichment and depletion of particular dinucleotide repeats within the enhancer space of the *A. coluzzii* genome. Similar patterns of repetitive sequence enrichment and depletion have also been demonstrated in humans and *D. melanogaster*. The functional role or importance of these repeats is not yet clearly demonstrated in any organism, but repeated discovery of the same pattern across invertebrates and vertebrates suggests a level of evolutionary functional conservation (Andolfatto, 2005; Cande et al., 2009; Glassford and Rebeiz, 2013; Schwaiger et al., 2014; Villar et al., 2015). The simplest interpretation is that the nucleotide composition and dinucleotide repeat enrichment or depletion in the enhancer compartment of the genome is a consequence of functional constraints imposed by binding of proteins such as transcription factors to TFBS, which underlies enhancer function as currently understood (Yao et al., 2015; Jacobs et al., 2018; Lambert et al., 2019). Fifteen motifs defining insect TFBS from the *D. melanogaster* JASPAR database and an addition 14 *de novo* motifs were significantly enriched in the *A. coluzzii* enhancer compartment. Further work will be necessary to confirm which of these potential TFBS are functional in enhancers of *A. coluzzii*, because *Anopheles* transcription factors and TFBS have not yet been comprehensively characterized. Thirteen of the fifteen putative transcription factors enriched in STARR-seq candidate enhancers (**Figure 4D**, TF identities inferred from the TFBS) have orthologues in both *A. coluzzii* and *A. gambiae*. Twelve of these 13 are one-to-one orthologues among the two anophelines, and the other is a *D. melanogaster* TF that matches two predicted proteins in both *A. coluzzii* and *A. gambiae* (**Supplementary Table S5**). The evolutionary conservation supports the conclusion that these TFs and their cognate binding sites are likely to be functional in *A. coluzzii*.

Human genetic mapping data suggest that non-coding variation is responsible for the majority of phenotypic variation. Of the significant SNPs associated with phenotypes in genome-wide association, studies (GWAS), only 5–10% are protein-coding variants, and >90% of significant GWAS hits are non-coding SNPs (Neph et al., 2012; Schaub et al., 2012; Farh et al., 2015). Mutations in human enhancers are associated with risk states for arthritis and growth disorders (Capellini et al., 2017), neurodevelopmental disorders (Short et al., 2018) and susceptibility to infection (Telenti and di Iulio, 2020), among others. In yeast, variable non-coding regions contribute to phenotypic diversity (Salinas et al., 2016), and a transcribed enhancer in *C. elegans* regulates the development of egg laying muscles (Goh and Inoue, 2018). In a recent association mapping study of *Anopheles* desiccation resistance, most of the significant SNPs were located in non-coding regions (Ayala et al., 2019), and therefore could be located in regulatory elements. The enhancer catalog presented here will be a useful tool to help decode the phenotypic effects of polymorphic non-coding DNA in *Anopheles*, particularly the important *Anopheles* traits underlying malaria transmission and vector biology such as *Plasmodium* infection susceptibility, insecticide resistance, and others.

However, understanding phenotypic readouts of enhancer function, or the differential phenotypes caused by allelic enhancers, is complicated by the incomplete current understanding of the mechanism of enhancer function. Most importantly, enhancers function by *cis*-activating the promoters of a panel of target genes that can be located tens or hundreds of kilobases distant, and the regulated target genes are often not the nearest gene(s) to the enhancer. Similar to the inability to accurately predict enhancer location computationally due the absence of a defined sequence code, there is also no current capacity to computationally predict the target genes of an enhancer. Target gene detection and modeling is most advanced in the human system, where a combination of empirical chromatin-based assays and computational modeling of interaction networks has indicated that most enhancers regulate up to approximately ten target genes, but that a given gene can be regulated by multiple enhancers with presumably contrasting effects (Mumbach et al., 2017; Baxter et al., 2018; Gasperini et al., 2020; Reilly et al., 2021). Thus, understanding the phenotypic outcomes of enhancer regulatory function is still in the early stages even in the most characterized model organisms.

The current genome-wide catalogue of transcriptional enhancers will provide a useful tool to identify enhancer locations, understand the important drivers of gene expression, and functionally filter the effects of genetic variation in *Anopheles* genotype-phenotype studies. Knowledge of the genomic locations of enhancers can also aid in the design of CRISPR/Cas9-based genome editing experiments, because chromosomal enhancers could alter the behavior of a gene integrated nearby, or an unknown enhancer included in an integrated cassette could produce unexpected expression results. Overall, the current catalog of *cis*-regulatory enhancer elements will contribute to developing a more complete picture of *cis*-regulatory modules within *Anopheles*, and provides a new tool for biological investigation aimed at the design of improved vector control methods.

DATA AVAILABILITY STATEMENT

All sequence files are available from the EBI European Nucleotide Archive database (http://www.ebi.ac.uk/ena/) under ENA study accession number PRJEB34434. The genome wide enhancer catalog generated in this study is available in **Supplementary Table S2**.

AUTHOR CONTRIBUTIONS

Designed research: KV, MR. Performed research: IH, LN, CA, DG, SZ, WG, N'FS, MR. Analyzed data: AP, EB, RN, MR. Wrote the paper: KV, MR.

FUNDING

This work received financial support to KV from the European Commission, Horizon 2020 Infrastructures #731060 Infravec2; European Research Council, Support for frontier research, Advanced Grant #323173 AnoPath; Agence Nationale de la Recherche, #ANR-19-CE35-0004 ArboVec; National Institutes of Health, NIAID #AI145999; and French Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" #ANR-10-LABX-62-IBEID, to RN from National Science Foundation, IIS#194727, and to MR from National Institutes of Health, NIAID #AI121587; National Institutes of Health, NIAID #AI145999. Funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We thank Alexander Stark, Research Institute of Molecular Pathology, Vienna for plasmids and helpful advice. We thank Michael Adang, University of Georgia, United States for the Ag55 cell line.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.785934/full#supplementary-material

REFERENCES

Andersson, R., and Sandelin, A. (2020). Determinants of Enhancer and Promoter Activities of Regulatory Elements. *Nat. Rev. Genet.* 21, 71–87. doi:10.1038/s41576-019-0173-8

Andolfatto, P. (2005). Adaptive Evolution of Non-Coding DNA in Drosophila. *Nature* 437, 1149–1152. doi:10.1038/nature04107

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., and Stark, A. (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-Seq. *Science* 339, 1074–1077. doi:10.1126/science.1232542

Arnold, C. D., Gerlach, D., Spies, D., Matts, J. A., Sytnikova, Y. A., Pagani, M., et al. (2014). Quantitative Genome-Wide Enhancer Activity Maps for Five Drosophila Species Show Functional Enhancer Conservation and Turnover during Cis-Regulatory Evolution. *Nat. Genet.* 46, 685–692. doi:10.1038/ng.3009

Asma, H., and Halfon, M. S. (2019). Computational Enhancer Prediction: Evaluation and Improvements. *BMC Bioinformatics* 20, 174. doi:10.1186/s12859-019-2781-x

Ayala, D., Zhang, S., Chateau, M., Fouet, C., Morlais, I., Costantini, C., et al. (2019). Association Mapping Desiccation Resistance within Chromosomal Inversions in the African Malaria Vector *Anopheles G. Mol. Ecol.* 28, 1333–1342. doi:10.1111/mec.14880

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res.* 43, W39–W49. doi:10.1093/nar/gkv416

Bailey, T. L. (2021). STREME: Accurate and Versatile Sequence Motif Discovery. *Bioinformatics* 37 (18), 2834–2840. doi:10.1093/bioinformatics/btab203

Baxter, J. S., Leavy, O. C., Dryden, N. H., Maguire, S., Johnson, N., Fedele, V., et al. (2018). Capture Hi-C Identifies Putative Target Genes at 33 Breast Cancer Risk Loci. *Nat. Commun.* 9, 1028. doi:10.1038/s41467-018-03411-9

Behura, S. K., Sarro, J., Li, P., Mysore, K., Severson, D. W., Emrich, S. J., et al. (2016). High-Throughput Cis-Regulatory Element Discovery in the Vector Mosquito *Aedes A. BMC genomics* 17, 341. doi:10.1186/s12864-016-2468-x

Benoit, M. (2020). Shooting for the STARRs: A Modified STARR-Seq Assay for Rapid Identification and Evaluation of Plant Regulatory Sequences in Tobacco Leaves. *Plant Cell* 32, 2057–2058. doi:10.1105/tpc.20.00392

Bird, C. P., Stranger, B. E., and Dermitzakis, E. T. (2006). Functional Variation and Evolution of Non-coding DNA. *Curr. Opin. Genet. Develop.* 16, 559–564. doi:10.1016/j.gde.2006.10.003

Buerger, A. (2015). *BasicSTARRseq: Basic Peak Calling on STARR-Seq Data*. R package version 180. Vienna, Austria: R Foundation for Statistical Computing.

Cande, J., Goltsev, Y., and Levine, M. S. (2009). Conservation of Enhancer Location in Divergent Insects. *Proc. Natl. Acad. Sci.* 106, 14414–14419. doi:10.1073/pnas.0905754106

Capellini, T. D., Chen, H., Cao, J., Doxey, A. C., Kiapour, A. M., Schoor, M., et al. (2017). Ancient Selection for Derived Alleles at a GDF5 Enhancer Influencing Human Growth and Osteoarthritis Risk. *Nat. Genet.* 49, 1202–1210. doi:10.1038/ng.3911

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve Years of SAMtools and BCFtools. *Gigascience* 10 (2), giab008. doi:10.1093/gigascience/giab008

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., et al. (2015). Genetic and Epigenetic Fine Mapping of Causal Autoimmune Disease Variants. *Nature* 518, 337–343. doi:10.1038/nature13835

Farley, E. J., Eggleston, H., and Riehle, M. M. (2021). Filtering the Junk: Assigning Function to the Mosquito Non-Coding Genome. *Insects* 12 (2), 186. doi:10.3390/insects12020186

Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., et al. (2020). JASPAR 2020: Update of the Open-Access Database of Transcription Factor Binding Profiles. *Nucleic Acids Res.* 48, D87–D92. doi:10.1093/nar/gkz1001

Gasperini, M., Tome, J. M., and Shendure, J. (2020). Towards a Comprehensive Catalogue of Validated and Target-Linked Human Enhancers. *Nat. Rev. Genet.* 21, 292–310. doi:10.1038/s41576-019-0209-0

Glassford, W. J., and Rebeiz, M. (2013). Assessing Constraints on the Path of Regulatory Sequence Evolution. *Phil. Trans. R. Soc. B* 368, 20130026. doi:10.1098/rstb.2013.0026

Gloss, B. S., and Dinger, M. E. (2018). Realizing the Significance of Noncoding Functionality in Clinical Genomics. *Exp. Mol. Med.* 50, 1–8. doi:10.1038/s12276-018-0087-0

Goh, K. Y., and Inoue, T. (2018). A Large Transcribed Enhancer Region Regulates *C. E* Bed-3 and the Development of Egg Laying Muscles. *Biochim. Biophys. Acta (Bba) - Gene Regul. Mech.* 1861, 519–533. doi:10.1016/j.bbagrm.2018.02.007

Gómez-Díaz, E., Yerbanga, R. S., Lefèvre, T., Cohuet, A., Rowley, M. J., Ouedraogo, J. B., et al. (2017). Epigenetic Regulation of Plasmodium Falciparum Clonally Variant Gene Expression during Development in *Anopheles G. Sci. Rep.* 7, 40655. doi:10.1038/srep40655

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: Scanning for Occurrences of a Given Motif. *Bioinformatics* 27, 1017–1018. doi:10.1093/bioinformatics/btr064

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. (2007). Quantifying Similarity between Motifs. *Genome Biol.* 8, R24. doi:10.1186/gb-2007-8-2-r24

Hire, R. S., Hua, G., Zhang, Q., Mishra, R., and Adang, M. J. (2015). *Anopheles G* Ag55 Cell Line as a Model for Lysinibacillus Sphaericus Bin Toxin Action. *J. Invertebr. Pathol.* 132, 105–110. doi:10.1016/j.jip.2015.09.009

Hong, J., Gao, R., and Yang, Y. (2021). CrepHAN: Cross-Species Prediction of Enhancers by Using Hierarchical Attention Networks. *Bioinformatics* 37 (20), 3436–3443. doi:10.1093/bioinformatics/btab349

Hrdlickova, B., de Almeida, R. C., Borek, Z., and Withoff, S. (2014). Genetic Variation in the Non-Coding Genome: Involvement of Micro-RNAs and Long Non-Coding RNAs in Disease. *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1842, 1910–1922. doi:10.1016/j.bbadis.2014.03.011

Jacobs, J., Atkins, M., Davie, K., Imrichova, H., Romanelli, L., Christiaens, V., et al. (2018). The Transcription Factor Grainy Head Primes Epithelial Enhancers for Spatiotemporal Activation by Displacing Nucleosomes. *Nat. Genet.* 50, 1011–1020. doi:10.1038/s41588-018-0140-x

Kazemian, M., Suryamohan, K., Chen, J.-Y., Zhang, Y., Samee, M. A. H., Halfon, M. S., et al. (2014). Evidence for Deep Regulatory Similarities in Early Developmental Programs across Highly Diverged Insects. *Genome Biol. Evol.* 6, 2301–2320. doi:10.1093/gbe/evu184

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., et al. (2013). Systematic Dissection of Regulatory Motifs in 2000 Predicted Human Enhancers Using a Massively Parallel Reporter Assay. *Genome Res.* 23, 800–811. doi:10.1101/gr.144899.112

Kleftogiannis, D., Kalnis, P., and Bajic, V. B. (2016). Progress and Challenges in Bioinformatics Approaches for Enhancer Identification. *Brief. Bioinformatics* 17, 967–979. doi:10.1093/bib/bbv101

Kwasnieski, J. C., Fiore, C., Chaudhari, H. G., and Cohen, B. A. (2014). High-throughput Functional Testing of ENCODE Segmentation Predictions. *Genome Res.* 24, 1595–1602. doi:10.1101/gr.173518.114

Lambert, S. A., Yang, A. W. H., Sasse, A., Cowley, G., Albu, M., Caddick, M. X., et al. (2019). Similarity Regression Predicts Evolution of Transcription Factor Sequence Specificity. *Nat. Genet.* 51, 981–989. doi:10.1038/s41588-019-0411-1

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for Computing and Annotating Genomic Ranges. *Plos Comput. Biol.* 9, e1003118. doi:10.1371/journal.pcbi.1003118

Lezcano, Ó. M., Sánchez-Polo, M., Ruiz, J. L., and Gómez-Díaz, E. (2020). Chromatin Structure and Function in Mosquitoes. *Front. Genet.* 11, 602949. doi:10.3389/fgene.2020.602949

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Lim, L. W. K., Chung, H. H., Chong, Y. L., and Lee, N. K. (2018). A Survey of Recently Emerged Genome-Wide Computational Enhancer Predictor Tools. *Comput. Biol. Chem.* 74, 132–141. doi:10.1016/j.compbiolchem.2018.03.019

McLeay, R. C., and Bailey, T. L. (2010). Motif Enrichment Analysis: a Unified Framework and an Evaluation on ChIP Data. *BMC Bioinformatics* 11, 165. doi:10.1186/1471-2105-11-165

Muerdter, F., Boryń, Ł. M., and Arnold, C. D. (2015). STARR-seq - Principles and Applications. *Genomics* 106, 145–150. doi:10.1016/j.ygeno.2015.06.001

Müller, H.-M., Dimopoulos, G., Blass, C., and Kafatos, F. C. (1999). A Hemocyte-Like Cell Line Established from the Malaria VectorAnopheles Gambiae Expresses Six Prophenoloxidase Genes. *J. Biol. Chem.* 274, 11727–11735. doi:10.1074/jbc.274.17.11727

Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., et al. (2017). Enhancer Connectome in Primary Human Cells Identifies Target Genes of Disease-Associated DNA Elements. *Nat. Genet.* 49, 1602–1612. doi:10.1038/ng.3963

Mysore, K., Li, P., and Duman-Scheel, M. (2018). Identification of *Aedes A* Cis-Regulatory Elements that Promote Gene Expression in Olfactory Receptor Neurons of Distantly Related Dipteran Insects. *Parasites Vectors* 11, 406. doi:10.1186/s13071-018-2982-6

Nardini, L., Holm, I., Pain, A., Bischoff, E., Gohl, D. M., Zongo, S., et al. (2019). Influence of Genetic Polymorphism on Transcriptional Enhancer Activity in the Malaria Vector Anopheles Coluzzii. *Sci. Rep.* 9, 15275. doi:10.1038/s41598-019-51730-8

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., et al. (2012). An Expansive Human Regulatory Lexicon Encoded in Transcription Factor Footprints. *Nature* 489, 83–90. doi:10.1038/nature11212

Nowling, R. J., Behura, S. K., Halfon, M. S., Emrich, S. J., and Duman-Scheel, M. (2021). PeakMatcher Facilitates Updated *Aedes A* Embryonic Cis-Regulatory Element Map. *Hereditas* 158, 7. doi:10.1186/s41065-021-00172-2

Pérez-Zamorano, B., Rosas-Madrigal, S., Lozano, O. A. M., Castillo Méndez, M., and Valverde-Garduño, V. (2017). Identification of Cis-Regulatory Sequences Reveals Potential Participation of Lola and Deaf1 Transcription Factors in *Anopheles G* Innate Immune Response. *PLoS One* 12, e0186435. doi:10.1371/journal.pone.0186435

Pudney, M., Varma, M. G. R., and Leake, C. J. (1979). Establishment of Cell Lines from Larvae of Culicine (Aedes Species) and Anopheline Mosquitoes. *Tca Man.* 5, 997–1002. doi:10.1007/bf00919719

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics* 26, 841–842. doi:10.1093/bioinformatics/btq033

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reilly, S. K., Gosai, S. J., Gutierrez, A., Mackay-Smith, A., Ulirsch, J. C., Kanai, M., et al. (2021). Direct Characterization of Cis-Regulatory Elements and Functional Dissection of Complex Genetic Associations Using HCR-FlowFISH. *Nat. Genet.* 53, 1166–1176. doi:10.1038/s41588-021-00900-4

Riehle, M. M., Guelbeogo, W. M., Gneme, A., Eiglmeier, K., Holm, I., Bischoff, E., et al. (2011). A Cryptic Subgroup of *Anopheles G* Is Highly Susceptible to Human Malaria Parasites. *Science* 331, 596–598. doi:10.1126/science.1196759

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative Genomics Viewer. *Nat. Biotechnol.* 29, 24–26. doi:10.1038/nbt.1754

Robson, M. I., Ringel, A. R., and Mundlos, S. (2019). Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D. *Mol. Cel* 74, 1110–1122. doi:10.1016/j.molcel.2019.05.032

Ruiz, J. L., Yerbanga, R. S., Lefèvre, T., Ouedraogo, J. B., Corces, V. G., and Gómez-Díaz, E. (2019). Chromatin Changes in *Anopheles G* Induced by Plasmodium Falciparum Infection. *Epigenetics & Chromatin* 12, 5. doi:10.1186/s13072-018-0250-9

Ruiz, J. L., Ranford-Cartwright, L. C., and Gómez-Díaz, E. (2021). The Regulatory Genome of the Malaria Vector *Anopheles G*: Integrating Chromatin Accessibility and Gene Expression. *NAR Genom Bioinform* 3, lqaa113. doi:10.1093/nargab/lqaa113

Salinas, F., de Boer, C. G., Abarca, V., García, V., Cuevas, M., Araos, S., et al. (2016). Natural Variation in Non-Coding Regions Underlying Phenotypic Diversity in Budding Yeast. *Sci. Rep.* 6, 21849. doi:10.1038/srep21849

Santolamazza, F., Mancini, E., Simard, F., Qi, Y., Tu, Z., and della Torre, A. (2008). Insertion Polymorphisms of SINE200 Retrotransposons within Speciation Islands of *Anopheles G* Molecular Forms. *Malar. J.* 7, 163. doi:10.1186/1475-2875-7-163

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking Disease Associations with Regulatory Information in the Human Genome. *Genome Res.* 22, 1748–1759. doi:10.1101/gr.136127.111

Schember, I., and Halfon, M. S. (2021). Identification of New *Anopheles G* Transcriptional Enhancers Using a Cross-species Prediction Approach. *Insect Mol. Biol.* 30, 410–419. doi:10.1111/imb.12705

Schoenfelder, S., and Fraser, P. (2019). Long-Range Enhancer-Promoter Contacts in Gene Expression Control. *Nat. Rev. Genet.* 20, 437–455. doi:10.1038/s41576-019-0128-0

Schwaiger, M., Schönauer, A., Rendeiro, A. F., Pribitzer, C., Schauer, A., Gilles, A. F., et al. (2014). Evolutionary Conservation of the Eumetazoan Gene Regulatory Landscape. *Genome Res.* 24, 639–650. doi:10.1101/gr.162529.113

Short, P. J., McRae, J. F., Gallone, G., Sifrim, A., Won, H., Geschwind, D. H., et al. (2018). De Novo mutations in Regulatory Elements in Neurodevelopmental Disorders. *Nature* 555, 611–616. doi:10.1038/nature25983

Telenti, A., and di Iulio, J. (2020). Regulatory Genome Variants in Human Susceptibility to Infection. *Hum. Genet.* 139, 759–768. doi:10.1007/s00439-019-02091-9

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., et al. (2015). Enhancer Evolution across 20 Mammalian Species. *Cell* 160, 554–566. doi:10.1016/j.cell.2015.01.006

Wingett, S. W., and Andrews, S. (2018). FastQ Screen: A Tool for Multi-Genome Mapping and Quality Control. *F1000Res* 7, 1338. doi:10.12688/f1000research.15931.2

Yao, L., Shen, H., Laird, P. W., Farnham, P. J., and Berman, B. P. (2015). Inferring Regulatory Element Landscapes and Transcription Factor Networks from Cancer Methylomes. *Genome Biol.* 16, 105. doi:10.1186/s13059-015-0668-3

Zheng, X. M., Chen, J., Pang, H. B., Liu, S., Gao, Q., Wang, J. R., et al. (2019). Genome-wide Analyses Reveal the Role of Noncoding Variation in Complex Traits during rice Domestication. *Sci. Adv.* 5, eaax3619. doi:10.1126/sciadv.aax3619

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership