Classification: Biological Sciences - Microbiology

Comparative 3D Genome Organization in Apicomplexan Parasites

Short title: 3D Genome Organization in Apicomplexan Parasites

Evelien M. Bunnik[1], Aarthi Venkat[2,*], Jianlin Shao[2,#,*], Kathryn E. McGovern[3,$], Gayani Batugedara[4], Danielle Worth[3], Jacques Prudhomme[4], Stacey A. Lapp[5,6,7], Chiara Andolina[8,%], Leila S. Ross[9], Lauren Lawres[10], Declan Brady[11], Photini Sinnis[12], Francois Nosten[8], David A. Fidock[9], Emma H. Wilson[3], Rita Tewari[11], Mary R. Galinski[5,6,7], Choukri Ben Mamoun[10], Ferhat Ay[2,13,&], and Karine G. Le Roch[4,&]

[1] Department of Microbiology, Immunology & Molecular Genetics, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229, USA.

[2] Division of Vaccine Discovery, La Jolla Institute for Allergy & Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA.

[3] School of Medicine, Division of Biomedical Sciences, University of California Riverside, 900 University Ave, Riverside, CA 92521, USA.

[4] Department of Molecular, Cell and Systems Biology, University of California Riverside, 900 University Ave, Riverside, CA 92521, USA.

[5] International Center for Malaria Research, Education and Development, Emory Vaccine Center, Yerkes National Primate Research Center, Emory University, 201 Dowman Drive, Atlanta, GA 30329, USA.

[6] Department of Medicine, Division of Infectious Diseases, Emory University, 1648 Pierce Dr NE, Atlanta, GA 30329, USA.

[7] Malaria Host-Pathogen Interaction Center, 954 Gatewood Road, Atlanta, GA 30329, USA.

[8] Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine Research building, University of Oxford, Old Road campus, Roosevelt Drive, Headington, Oxford, OX3 7FZ, UK and Shoklo Malaria Research Unit, Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Mae Sot, Tak 63110, Thailand.

[9] Department of Microbiology and Immunology, Columbia University Medical Center, New York, NY 10032, USA.

[10] Department of Internal Medicine, Section of Infectious Diseases, Yale School of Medicine, 330 Cedar St, Boardman 110, New Haven, CT 06520, USA.

[11] School of Life Sciences, Queens Medical Centre, University of Nottingham, Nottingham NG7 2UH, UK.

[12] Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, E5132, Baltimore, MD 21205, USA.

[13] School of Medicine, University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92521, USA.

[#] Present address: Zhejiang Provincial Center for Cardio-Cerebro-Vascular Disease Control and Prevention, Zhejiang Hospital, Zhejiang Province, China

[$] Present address: BIO5 Institute, University of Arizona, Tucson, AZ, USA

[%] Present address: Department of Medical Microbiology, Radboud University Medical Centre, Nijmegen, The Netherlands.

*A.V. and J.S. contributed equally to this work.

[&]F.A. and K.G.L.R. contributed equally to this work.

Corresponding authors:

Ferhat Ay

Division of Vaccine Discovery, La Jolla Institute for Allergy & Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA.

Karine Le Roch

Department of Molecular, Cell and Systems Biology, University of California Riverside, 900 University Ave, Riverside, CA 92521, USA.

Keywords: Malaria, genome organization, virulence, Hi-C, epigenomics

ABSTRACT (250 words)

The positioning of chromosomes in the nucleus of a eukaryotic cell is highly organized and has a complex and dynamic relationship with gene expression. In the human malaria parasite *Plasmodium falciparum*, the clustering of a family of virulence genes correlates with their coordinated silencing and has a strong influence on the overall organization of the genome. To identify conserved and species-specific principles of genome organization, we performed Hi-C experiments and generated 3D genome models for five *Plasmodium* species and two related apicomplexan parasites. *Plasmodium* species mainly showed clustering of centromeres, telomeres and virulence genes. In *P. falciparum*, the heterochromatic virulence gene cluster had a strong repressive effect on the surrounding nuclear space, while this was less pronounced in *P. vivax* and *P. berghei*, and absent in *P. yoelii*. In *P. knowlesi*, telomeres and virulence genes were more dispersed throughout the nucleus, but its 3D genome showed a strong correlation with gene expression. The *Babesia microti* genome showed a classical Rabl organization with colocalization of subtelomeric virulence genes, while the *Toxoplasma gondii* genome was dominated by clustering of the centromeres and lacked virulence gene clustering. Collectively, our results demonstrate that spatial genome organization in most *Plasmodium* species is constrained by the colocalization of virulence genes. *P. falciparum* and *P. knowlesi*, the only two *Plasmodium* species with gene families involved in antigenic variation, are unique in the effect of these genes on chromosome folding, indicating a potential link between genome organization and gene expression in more virulent pathogens.

SIGNIFICANCE STATEMENT (120 words)

From yeast to human cells, genome organization in eukaryotes has a tight relationship with gene expression. We investigated the three-dimensional organization of chromosomes in malaria parasites to identify possible connections between genome architecture and pathogenicity. Genome organization was dominated by the clustering of *Plasmodium*-specific gene families in 3D space. In particular, the two most pathogenic human malaria parasites shared unique features in the organization of gene families involved in antigenic variation and immune escape. Related human parasites *Babesia microti* and *Toxoplasma gondii* that are less virulent lacked the correlation between gene expression and genome organization observed in human *Plasmodium* species. Our results suggest that genome organization in malaria parasites has been shaped by parasite-specific gene families and correlates with virulence.

INTRODUCTION

Apicomplexans are obligate intracellular parasites that can be highly pathogenic and are responsible for a wide range of diseases in humans and animals. The phylum consists of at least 6,000 species, with potentially many undiscovered members (1). Among apicomplexan parasites that infect humans, *Plasmodium* spp., the causative agents of malaria, have the highest health and economic impact. The most prevalent and deadly human malaria parasite is *P. falciparum*, responsible for an estimated 445,000 deaths per year (2). Other *Plasmodium* species that infect humans include *P. vivax* and *P. knowlesi*. *P. vivax* is widespread predominantly outside Africa, and *P. knowlesi* is a zoonosis in Southeast Asia. The natural hosts of *P. knowlesi* are long-tailed and pig-tailed macaques. However, transmission from monkeys to humans through a mosquito vector has been widely reported in Malaysia and can cause severe, potentially lethal disease (3). Other human-relevant apicomplexans include *Babesia microti* (4), the causative agent of human babesiosis, a malaria-like illness endemic in the US but with worldwide distribution, and *Toxoplasma gondii*, the causative agent of toxoplasmosis, an opportunistic infection commonly encountered among individuals with weakened immune systems (5).

During millions of years of co-evolution with their hosts, apicomplexan parasites have developed species-specific large multigene families that are involved in host-parasite interactions (6). These gene families are important for parasite survival, pathogenesis, virulence, and immune evasion. All *Plasmodium* species contain virulence genes that belong to the *Plasmodium* interspersed repeat (*pir*) superfamily. In addition, *P. falciparum* and *P. knowlesi*, have evolved unique gene families that orchestrate these parasites to undergo antigenic variation, called *var* and *SICAvar*, respectively (7, 8). During the process of antigenic variation, the parasite can escape from host immune responses by changing which members of these large families of parasite antigens are expressed. The ability of these parasites to switch their antigenic profile correlates with their high virulence and persistence in the face of adaptive immune responses.

The biological functions of *pir* genes are largely unknown, but it has been suggested that they have many different roles in virulence, signaling, trafficking, protein folding, adhesion, and establishment of chronic infections (9-11). The availability of genome sequences for selected apicomplexan parasites has revealed the genomic landscape of these virulence gene families. Copy numbers for the virulence gene families typically range between 150 and 300 genes per organism, although there are some exceptions (for example, 980 *yir* genes in *P. yoelii* 17X) (6,

12-14). These genes are located close to the telomere ends of most (sometimes all) of the 14 chromosomes of the various *Plasmodium* genomes. Similarly, the subtelomeric regions of *B. microti* chromosomes contain several small gene families encoding exported proteins. These proteins are targets of the antibody response in *B. microti*-infected humans and may be involved in antigenic variation (15-17). *T. gondii* also has multiple parasite-specific gene families involved in pathogenesis and immune evasion. Some of these are subtelomeric, but most are dispersed among the genome, either as individual genes, or in smaller and larger arrays (18). *T. gondii* does not use classic antigenic variation, although some of these *Toxoplasma*-specific gene families may be involved in escape from immune responses (19).

First discovered in *P. knowlesi* (20-23), the P. falciparum *var* and *P. knowlesi SICAvar* gene families mediate antigenic variation and immune escape and may be one of the factors that make *P. knowlesi* and *P. falciparum* so lethal in humans. The *var* and *SICAvar* gene families encode *P. falciparum* Erythrocyte Membrane Protein 1 (PfEMP1) and Schizont Infected Cell Agglutination variant antigen, respectively, which are expressed on the surface of the infected red blood cell. As a result, these proteins are exposed to the host immune system and elicit strong antibody responses. Each parasite expresses a single PfEMP1 or a limited repertoire of SICAvar antigens (24, 25). Parasites can rapidly and efficiently escape from the host immune response by switching the gene variant that is expressed, resulting in successive cycles of antibody production and parasite escape (22, 26). In addition, PfEMP1 mediates adherence of infected red blood cells to the vasculature, which prevents clearance of the parasite by the spleen. Cytoadherence in vital organs causes tissue damage and is a major cause of pathology in *P. falciparum* malaria (27).

Similar to the *pir* genes, most *var* genes are located in the subtelomeric regions of all 14 *P. falciparum* chromosomes, although several chromosomes also harbor internal *var* genes. The *SICAvar* genes are distributed more evenly along the chromosomes, with only a few located in subtelomeric regions (7). One of the mechanisms involved in the complex network of clonal *var* gene expression is localization of these genes in perinuclear clusters of heterochromatin (28, 29). In previous studies (30, 31), we observed that the requirement for *var* genes to come together in 3D space has a strong influence on the overall organization of the *P. falciparum* genome. Chromosomes with internal *var* gene clusters form loops to accommodate the perinuclear localization of all *var* genes. In addition, we observed a strong association between three-dimensional genome organization and gene expression.

Based on these observations, we hypothesized that gene families involved in antigenic variation need to be tightly regulated and therefore have a strong influence on genome organization. Organisms that do not undergo antigenic variation may thus have less stringent requirements with respect to the structure of their genome in the nucleus. To test this hypothesis, we studied the genome architecture of five different *Plasmodium* species parasites, two that are known to undergo antigenic variation (*P. falciparum* and *P. knowlesi*), and three that are not (*P. vivax*, *P. berghei*, and *P. yoelii*). In addition, we studied two related apicomplexan parasites (*B. microti* and *T. gondii*) to identify characteristics of genome organization that are specific to the *Plasmodium* genus. When considering genome-wide clustering of all virulence genes, we observed that all organisms studied here, with the exception of *T. gondii* and *P. knowlesi*, showed significant colocalization for these genes. In *P. knowlesi*, even though contact counts between *SICAvar* genes show a moderate enrichment compared to random, these were found to be scattered throughout the nucleus. However, *SICAvar* loci showed cross-shaped patterns in intra-chromosomal contact with enriched contact counts to other *SICAvars* on the same chromosome similar to interaction patterns created by *P. falciparum var* genes, suggesting that colocalization of these genes is a conserved feature and may play role in mutually exclusive expression of *SICAvars*. *T. gondii* virulence genes are not located subtelomerically and did not cluster in 3D space, pointing towards differences in regulation of gene expression in this apicomplexan parasite.

RESULTS

Profiling genome organization for apicomplexan parasites

We performed Hi-C experiments for five different apicomplexan parasites using the *in situ* Hi-C methodology (32) (see Methods): *Plasmodium knowlesi*, *P. yoelii*, *P. berghei*, *Toxoplasma gondii*, and *Babesia microti* (**Table 1** and **Fig. 1**). Hi-C data for *P. falciparum* and *P. vivax* were available from previous studies (30, 31). For all *Plasmodium* species except *P. vivax*, we performed the Hi-C experiment on human blood stage parasites. The only *P. vivax* sample available to this study was from mosquito salivary gland sporozoites, and we therefore also included *P. falciparum* sporozoites to allow comparisons between different developmental stages in different parasite species. Even though the global features of genome organization are comparable in *P. falciparum* trophozoites and gametocytes, we also included the *P. falciparum* gametocyte stage to allow a direct comparison with *P. berghei* gametocytes. *B. microti* tick isolate (LabS1) and clinical isolate (Bm1438) samples were obtained from non-synchronized blood stage infections in mice and from

cultures that were predominantly at the last stage of intraerythrocytic development (tetrads). Finally, for *T. gondii* we included the rapidly replicating tachyzoite stage and the more dormant bradyzoite stage, both cultured in vitro.

For several of these samples, we performed Hi-C experiments for two or more biological replicates or highly related strains (**Table 1** and **Materials and Methods**), which showed a high degree of similarity as quantified by a Hi-C-specific reproducibility measure (33) (***SI Appendix,* Fig. S1**). For *P. falciparum,* we observed higher reproducibility among replicates of the same stage as compared to samples from different stages. Therefore, we combined the two gametocyte samples (early and late) as well as the two sporozoite replicates for subsequent analysis to obtain higher resolution. For *P. vivax*, the two sporozoite replicates showed very high reproducibility and were therefore also combined. For *P. berghei*, all three conditions (WT and Smc2/Smc4 mutants) were highly similar and, hence, were combined for further analysis. For *B. microti*, synchronous and asynchronous samples had higher reproducibility within compared to across these two groups regardless of whether the samples were from the tick or field isolates. Hence, we combined all synchronous samples together and similarly the asynchronous samples. After combining specimens with high correlations, we obtained a total of eleven samples from the various stages, strains, and conditions of the seven different apicomplexan parasites (**Table 1**).

For each sample, the Hi-C reads were processed (i.e., mapping, filtering, pairing and removing duplicates) using HiCPro package (34), resulting in a total of ~300 million unique valid read pairs out of more than 1.5 billion pairs sequenced (***SI Appendix,* Table S1**). The observed intrachromosomal and interchromosomal contacts were aggregated into contact count matrices at 10-kb resolution and were then normalized using the ICE method to correct for experimental and technical biases (35) (normalized contact count heatmaps are accessible at http://apicomplexan3d.lji.org/). Next, we inferred a consensus 3D genome structure for each of the eleven samples using the negative binomial model from the PASTIS 3D modeling toolbox (36) (https://github.com/hiclib/pastis) (**Fig. 2** and ***SI Appendix,* Fig. S2**). When comparing this modeling approach to the multidimensional scaling method for PASTIS, all major hallmarks of genome organization are conserved (***SI Appendix,* Fig. S3**). However, the negative binomial distribution model better captures the overdispersion of contact counts and was therefore used for inferring all models in this study. For assessing the stability of the consensus structures, we used 100 random initializations of the 3D coordinates and clustered the resulting 100 structures for each sample (***SI Appendix,* Fig. S4**). For most of our samples, we did not find any subset of

3D models that distinctly cluster with each other suggesting the resulting models are robust to differences in initialization. For *P. vivax* which showed some sign of clustering, we sampled representative models from each potentially distinct cluster. Our comparison of 3D structures from the four most prominent clusters showed no striking differences in genome organization (***SI Appendix,* Fig. S5**). Since we observed conservation of all of the main structural features among the different initializations in either case, we used a single representative model for each of our Hi-C samples. Raw and ICE-normalized contact count heatmaps, Fit-Hi-C p-value heatmaps (37), PASTIS Euclidean distance matrices and PASTIS 3D models for each chromosome are shown in **Dataset 1**.

Detection of genome assembly problems and their correction with Hi-C data

Hi-C data has been used to detect translocation events or to improve genome assemblies based on Illumina and/or PacBio reads in many organisms, including *Arabidopsis thaliana, Aedes aegypti*, and human (38-40). Therefore, we first scanned our samples using a metric developed to detect genome assembly errors in Hi-C data (38) to avoid biases in our downstream analysis of 3D genome organization. Small misassemblies were observed in *P. yoelii* chr9 and in *B. microti* chr1 for the tetrad sample (***SI Appendix,* Fig. S6**), but these are unlikely to influence our results. For *P. knowlesi*, we recently published an assembly of the *P. knowlesi* genome based our Hi-C data in combination with sequencing data generated using Illumina and Pacbio platforms (41). All analyses presented here were performed using the updated version of the *P. knowlesi* genome assembly. All other *Plasmodium* and *Babesia* samples were error-free. However, several issues were detected for *T. gondii* that were handled as described below.

The current genome assembly of *T. gondii* consists of 14 chromosomes, the same number as for *Plasmodium* species and close apicomplexan relatives, such as *Neospora caninum*. For 13 chromosomes, the location of the centromere has previously been identified by chromatin immunoprecipitation of centromeric and pericentromeric proteins (42, 43). For chromosome VIIb, the centromere has thus far remained elusive. In the interchromosomal heatmap using the original genome assembly with 14 chromosomes, it can be observed that chrVIIb and chrVIII have a higher number of interactions than any other combination of chromosomes, and that the number of contacts is highest between the right telomere of chrVIIb and the left telomere of chrVIII (**Fig. 3A**). These observations suggest that chrVIIb and chrVIII could be physically linked. After we computationally stitched chrVIIb and chrVIII together to create one large chromosome, the interchromosomal contact counts were at the expected levels, while no apparent discrepancies

were observed in either the interchromosomal heatmap (**Fig. 3B**) or the intrachromosomal heatmap of the combined chrVIIb and chrVIII (**Fig. 3C**). This stitched chromosome showed a single centromere interaction with every other chromosome. In addition, we did not detect any signs of misassembly in the stitched chromosome (**Fig. 3D**). The small signal observed in the tachyzoite sample is not at the junction of the stitched chrVIIb and chrVIII. Based on these observations, we propose that chrVIIb and chrVIII are in fact a single chromosome. This would explain the unusual interaction pattern in our original interchromosomal interaction plot, as well as the apparent absence of a centromere in chrVIIb. We have used this stitched chromosome in all of our analyses and refer to this chromosome as chrVIII.

The misassembly metric detected several other issues in the *T. gondii* genome (***SI Appendix,* Fig. S6**). Most prominently, chrXII showed an unusual pattern that is most likely caused by an inversion of a segment spanning from bin 272 to bin 499 (***SI Appendix,* Fig. S7A**). The fraction of the parasite population harboring this inversion is approximately 10% in the tachyzoite sample and is close to 100% in the bradyzoite population. The tachyzoite and bradyzoite samples were generated using an unmodified and a transgenic ME49 strain, respectively. This result indicates that this inversion most likely arose spontaneously in the ME49 strain and can be selected during bottleneck events such as the generation of a transgenic strain. For the purpose of this study, the contact count patterns caused by this inversion were not considered a misassembly of the *T. gondii* genome.

Lastly, chrIX showed an unexpected signal around bin 500 (***SI Appendix,* Fig. S7B**). Upon inspection of the contact counts in this region, we noted that bins 498 and 505 showed a much higher number of interactions than expected based on the surrounding bins, suggesting amplification of genomic sequences within these two bins. This effect was observed in both the tachyzoite and bradyzoite samples and may point towards an error in the *T. gondii* genome assembly. Since Hi-C does not provide sufficient resolution to resolve such potentially tandem array of repeats, we did not attempt to correct the reference genome for this region. Finally, smaller potential misassemblies were observed in chrIV and chrV.

Colocalization of functional elements and gene families in *Plasmodium*

In a previous study, we showed that during the blood stages in the human host, the *P. falciparum* genome has a more complex organization than the *Saccharomyces cerevisiae* genome, which is organized in a classical Rabl conformation. In *P. falciparum*, similar to yeast, we observed

clustering of telomeres and centromeres on opposite sides of the nucleus but in addition we also observed domain-like structures (DLS), similar to multicellular organisms, surrounding genes involved in pathogenicity (30, 31). Clustering of telomeres and centromeres has previously been observed in budding and fission yeast (44, 45), plants, *Drosophila*, and recently also in mammalian cells using advanced single-cell Hi-C analysis and genome modeling (46). It can also clearly be observed in 3D models for the *P. falciparum* trophozoite and gametocyte stages (**Fig. 2B**).

Here, we analyzed the organization of these genomic hallmarks in all analyzed apicomplexan parasites by testing for an enrichment in interactions between loci of interest, as well as for colocalization in our 3D models (**Table 2**). The centromeres interacted with each other in *P. knowlesi*, *P. berghei*, and *P. yoelii*, although the interchromosomal heatmaps showed that these interactions were relatively localized in *P. knowlesi* and *P. yoelii* (***SI Appendix,* Fig. S8**). In agreement with this finding, the centromeres strongly clustered in the 3D models of these organisms. As highlighted in a previous study (31), and unlike the blood stages of any of the *Plasmodium* species where all pairs of centromeres significantly interact with each other, the centromeres in *P. vivax* salivary gland sporozoites showed weaker contacts and enrichment for only a subset of centromere pairs. Salivary gland sporozoites from *P. falciparum* showed no clustering of centromeres, suggesting that the loss of centromere interactions could be a general feature of the sporozoite stage. For *P. vivax*, the subset of significantly interacting centromere pairs was sufficient to drive significance in the contact count-based Witten-Noble colocalization test (47), but not for strong colocalization in 3D models when evaluated visually or using 3D distances (Table 2).

The telomeres of *P. yoelii* and *P. vivax* showed strong enrichment in contact counts, while the telomeres in *P. berghei* and *P. knowlesi* did not. However, the telomeres in *P. berghei* did come together in 3D space, although to a lesser extent than those in *P. yoelli* and *P. vivax*. All *Plasmodium* species that were analyzed in this study harbor virulence genes at the subtelomeric regions of nearly every chromosome. In line with clustering of the telomeres, we also observed colocalization of virulence genes in all of these organisms, with the exception of the *P. knowlesi SICAvar* genes (**Table 1**). The strong clustering of these genes was recapitulated in the 3D models of *P. falciparum*, *P. vivax*, *P. berghei* and *P. yoelii*. In conclusion, while we observed varying degrees of clustering of telomeres and centromeres, all *Plasmodium* genomes, except for *P. knowlesi*, showed colocalization of *pir* genes.

Antigenic variation genes are associated with formation of domain-like structures in *Plasmodium* species

In *P. knowlesi, SICAvar* genes are located in subtelomeric regions, but are also found scattered throughout the genome, either individually or in small groups of up to four genes. Due to the highly repetitive nature of their sequences, 31 of the *SICAvar* genes have low mappability. We were unable to detect strong co-localization for the 136 remaining *SICAvar* genes. However, our co-localization test measures whether all loci in the genome cluster and does not pick up localized clusters that consist of only a limited subset of genes. Although Hi-C signals were weak for loci that consist of only one or a few genes, the contact count heatmaps showed additional interactions between large internal and subtelomeric *SICAvar* gene clusters, most clearly observed in chr4 (**Fig. 4** and ***SI Appendix,* Fig. S9**). This interaction is reminiscent of *var* gene interaction patterns observed in *P. falciparum* that give rise to domain-like structures (30). We calculated a domain score to quantify the insulation of each locus from its neighborhood (see Materials and Methods) and observed a total of 6 domain-like structures (DLS) in *P. knowlesi*. Using the same metric, we identified 3 DLS in *P. falciparum* and 2 DLS in *P. vivax*, but none in any of the other organisms (**Fig. 4** and **Dataset 1**). It should be noted that these DLS are distinct from topologically associating domains observed in metazoan genomes, both in size and in mechanism of formation. Homologs of CCCTC-binding factor (CTCF) have not been identified in *Plasmodium* spp., and it is therefore unlikely that chromatin loop exclusion or any similar mechanism is involved in the formation and maintenance of these DLS. These results suggest that subsets of *SICAvar* genes throughout the genome may interact with each other, similar to the *var* genes in *P. falciparum*. Such interactions may be crucial to coordinate mutually exclusive gene expression necessary for antigenic variation.

Genome organization in the apicomplexan parasites *B. microti* and *T. gondii*

Babesia microti has a relatively small genome with four chromosomes that showed strong interactions among the telomeres and the centromeres, resulting in a classic Rabl conformation (**Fig. 2B** and ***SI Appendix,* Fig. S2**). Consequently, subtelomeric multi-gene families were strongly clustered, but additional virulence genes were localized away from the telomeric cluster. The 3D models of the asynchronous and the synchronous sample were very similar (***SI Appendix,* Fig. S2**). However, the contact count heatmaps showed additional interaction patterns within chromosomes for the synchronized samples obtained at the tetrad stage, but not for the asynchronous samples (***SI Appendix,* Fig. S10**). For chr3, these patterns may partially be caused

by a virulence gene of the BMN1 family located at position 272,813 - 273,799. These results suggest that these interactions may not be maintained during the complete cell cycle and underscore the importance of using tightly synchronized cell populations to be able to observe transient interactions.

In agreement with a previous microscopy study (42), *T. gondii* centromeres showed strong interactions and colocalized in the 3D models of both the tachyzoite and the bradyzoite stage. The telomeres also interacted, but to a much lower extent as compared to the centromeres, and this interaction was not readily apparent from the 3D models (***SI Appendix,* Fig. S2**). At the resolution used in our models, no significant differences were observed between tachyzoites and bradyzoites (***SI Appendix,* Fig. S2**). Most virulence genes in *T. gondii* are not located in subtelomeric regions, but are found on every chromosome arranged as single genes as well as smaller and larger arrays of up to 19 genes. As a consequence, virulence genes were scattered throughout the 3D model of the genome and we were unable to detect any significant colocalization pattern for these genes (**Table 2**).

Another difference between genome organization in *Plasmodium* species and *T. gondii* was observed for the strength of chromosome territories. The 3D models for *T. gondii* exhibited more territorialized chromosomes whereas *Plasmodium* chromosomes were more stretched out through the nucleus (see ***SI Appendix,* Fig. S11**). In order to quantify these differences directly from Hi-C data, we interrogated the relationship between distance and contact probability, which has been shown to depend on the arrangement of chromosomes within the nucleus (48, 49). However, this scaling relationship has been mainly used to compare different samples of the same organism or organisms with similar genome sizes. Here we adapted this analysis for comparing all our samples ranging from ~4.5Mb (*B. microti*) to ~70Mb (*T. gondii*) in size to each other as well as to human (32) and budding yeast genomes (50) as reference points. In order to do that, at 10kb resolution, we focused only on intra-chromosomal interactions within 1Mb distance and we computed the log2 fold change of the average contact for each genomic distance with respect to the overall average contact count of all possible pairs within 1Mb (***SI Appendix,* Fig. S12**). Our results show that, similar to human genome organization, the contacts within 50kb to 500kb are highly enriched for both *T. gondii* samples compared to all other apicomplexan parasites and budding yeast. Interestingly, this genomic distance range corresponds to topological domains (TADs) for human genome, which are known to be enriched for higher within-domain interactions. However, we have not seen clear TAD patterns in *T. gondii* genome,

suggesting that this trend may be related to the larger chromosomes and genome size of *T. gondii* as compared to other apicomplexan parasites. Further analysis may be necessary to understand the relationship between genome size and chromosome territory formation in the presence and absence of TAD-like structures.

Relation between gene expression and genome organization

P. falciparum gene expression has been shown to be associated with the position of genes in the nucleus (30, 31). To determine the relation between gene expression and genome organization in other apicomplexan parasites, we binned the genes of each organism into 20 groups based on their distance from the centroid of all telomeres. For each bin, we calculated the average gene expression using stage-specific transcriptomics data sets and plotted these values against the average distances from the telomere centroid (**Fig. 5A**). We also colored the 20 bins in the 3D models based on normalized average gene expression values (**Fig. 5B**).

As expected based on previous work, *P. falciparum* trophozoites showed the lowest gene expression in the bin closest to the centroid of the telomere, and a gradient of increasing gene expression in the next three bins. The remaining 16 bins showed relatively comparable levels of gene expression. The results for *P. falciparum* gametocytes were almost identical. For *P. vivax* sporozoites and *P. berghei* gametocytes, a similar but much weaker pattern was observed, with reduced gene expression only in the bin closest to the telomere centroid. *P. yoelii* and *B. microti* did not show any relation between gene expression and 3D location relative to the telomeres. Surprisingly, the *P. knowlesi* trophozoite stage displayed a gene expression gradient across the entire genome, with the lowest gene expression close to the centroid of the telomere. In the absence of strong telomere and centromere clustering, the 3D model of *P. knowlesi* looked somewhat unorganized. However, the gene expression gradient observed here suggests that genome organization of this parasite is in fact strongly correlated with expression with highly expressed genes preferentially localizing on one side of the 3D genome and the telomeres localizing on the other side.

Finally, for the *T. gondii* samples, we observed a decrease in gene expression in the bin furthest away from the telomere centroid, in stark contrast to all other organisms analyzed here. This decrease was reproducible between the two stages. As can be observed in the 3D models, these genes are located at the nuclear periphery. In both tachyzoites and bradyzoites, the genes in this bin were strongly enriched for merozoite-specific gene expression (51) ($p<0.00001$, two-tailed

Fisher's exact test), including 19 of 33 members of *T. gondii* family A proteins. Collectively, these results suggest that the organization of *Plasmodium* genomes is to a large extent driven by virulence genes. On the other hand, *T. gondii* has adopted different ways to organize gene expression in relation to its nuclear architecture, although the nuclear periphery may also function as a site of gene silencing.

DISCUSSION

From yeast to human cells, genome organization in eukaryotes has a tight relationship to gene expression. In particular, evidence is accumulating that the compartmentalized architecture of a cell nucleus is critical to many biological functions. Evolutionarily conserved mechanisms in the compartmentalization and function of yeast and mammalian genomes have been identified, but genome organization in these organisms also shows differences due to the fact that the nuclei of single cell organisms are typically 1000 times smaller than those of mammals. While nuclear architecture in single cell eukaryotes has been extensively studied in yeast, little is known about the genome organization of other unicellular organisms. We therefore investigated the genome architecture of various apicomplexan parasites. Our goal was to identify common features of genome organization and possible connections between genome architecture and pathogenicity.

In this study, we demonstrated that the Hi-C methodology can be employed to correct genome assemblies and study its 3D organization at the same time. While Hi-C experiments detect many long-range interactions, genomic segments that are physically linked and in close proximity along the DNA strand are preferentially ligated to one another. This results in a highly reproducible relationship between genomic distance and contact probability. Using this property, we detected that chromosomes VIIb and VIII of the *T. gondii* genome are likely to be physically linked. Our proposed genome assembly provides a coherent explanation of the apparent absence of a centromere in our original chrVIIb contact count matrix as well as the absence of centromeric protein TgCenH3 and pericentromeric protein TgChromo1 in chrVIIb in previous ChIP-on-chip analyses (42, 43). Our corrected *T. gondii* assembly has one centromere for each of the 13 chromosomes and the interaction between all of these centromeres is clearly visible in interchromosomal contact maps as well as 3D models.

Once genome assemblies were corrected, we focused our analysis on the identification of commonalities and differences in genome organization between members of the *Plasmodium* family, as well as related apicomplexan parasites. We previously showed (30) that although the genomes of *P. falciparum* and *Saccharomyces cerevisae* are similar in size, the *P. falciparum* genome has a more complex three-dimensional structure. In particular, we noted that spatial complexity was added by the requirement for virulence genes of the *var* family to colocalize.

The results from this study demonstrate that the organization of other *Plasmodium* genomes is also largely driven by their virulence genes. However, in contrast to *P. falciparum* and *P. knowlesi*, the rodent malaria species harbor virulence genes only in subtelomeric regions. The organization of their genomes is therefore relatively simple and similar to fission and budding yeast, with clustering of pericentromeric and subtelomeric heterochromatin islands driving the overall structure. On the other hand, the internal localization of *var* and *SICAvar* genes necessitates the formation of chromosome loops to bring distal loci in close spatial proximity. Similarly, *P. vivax* chr5 harbors a locus that interacts with subtelomeric regions, generating similar domain-like patterns as compared to internal *var* gene islands (31). This locus contains genes of the *Pv-fam-e* family, encoding exported proteins involved in erythrocyte remodeling. The precise function of this gene family and the reasons for its association with subtelomeric heterochromatin remain to be discovered. The clustered organization of virulence genes allows for increased rates of recombination to generate additional diversity and coordination of gene expression. We can only speculate why certain genes in the human malaria parasites are located away from subtelomeric regions, but the requirements for domain formation and complex chromosome structure highlight that additional layers of control are necessary to orchestrate mutually exclusive gene expression.

The *pir* genes of *P. vivax* and the rodent malaria parasites are not subject to the epigenetically driven mutually exclusive expression that is observed for the *var* and *SICAvar* gene families, pointing towards additional differences in regulation between these gene families. While clonal expression of *var* and *SICAvar* genes is thought to enable escape of the parasite from host antibody responses, the expression of certain *pir* genes has been associated with the establishment of acute and chronic infections, independent of adaptive immunity (9). This difference in virulence gene expression patterns between primate and rodent malaria parasites is reflected by the absence of several nuclear proteins in the genome of rodent malaria parasites. Of these, both the C-terminal extension of Rpb1 (52) and the histone methyltransferase PfSET2 (53, 54) have been shown to contribute to *var* gene regulation. These observations suggest that

the various parasite lineages have evolved different strategies to survive in their respective hosts. One contributing factor could be the life span of the host. Parasites infecting long-lived primates may require escape from adaptive immune responses to establish chronic infections and ensure transmission to a susceptible host, while parasites infecting short-lived rodent hosts may benefit from more flexible expression of their virulence genes.

The unique features of *Plasmodium* genome organization as described above are highlighted by the analysis of *B. microti* and *T. gondii*. Both parasites show distinct differences in the structure of their genomes. The *B. microti* genome showed a classical Rabl organization, with colocalization of a subset of virulence genes, those located in subtelomeric regions, in 3D space. From this observation, we conclude that genome organization in *B. microti* is not as strongly involved in regulation of virulence gene expression as in *Plasmodium* parasites. In fact, we observed no association between genome structure and gene expression in general in this parasite. In *T. gondii*, virulence genes were scattered throughout the genome and did not colocalize in the nucleus. The only association between genome organization and gene expression that we could detect was the possible repression of stage-specific genes by the formation of perinuclear heterochromatin. Our observation that tachyzoites and bradyzoites showed similar genome organization is in line with the relatively small differences in gene expression between these two life cycle stages, in particular when compared to gene expression profiles of merozoites and oocysts (55). However, it should be mentioned that the bradyzoites used here were obtained from in vitro culture and may not fully represent bradyzoites in a tissue cyst. Attempts to prepare Hi-C libraries from bradyzoites isolated from mouse brain tissue cysts were not successful.

The distinct organization of virulence genes suggests that *T. gondii* has adopted different mechanisms for gene regulation as compared to *Plasmodium* species, which may be related to its much broader host range and cell tropism. This is reflected in the larger number of ApiAP2 transcription factors encoded by the *T. gondii* genome (67 versus ~27 for *Plasmodium* species), which are used for both transcriptional activation and repression (56). During the evolution from the ancestor of Apicomplexan organisms, the *Plasmodium* lineage has lost many genes that were retained in *T. gondii* (57). *Plasmodium* species have thus evolved into highly specialized organisms that control their virulence through a highly restricted mechanism, whereas *T. gondii* has retained more properties of their free-living ancestor and is more flexible when it comes to gene regulation.

A limitation of this study is that Hi-C experiments were performed using millions of cells as input. The contact count matrices and 3D models therefore represent the average of a population of possible genome organizations. In *P. falciparum*, various microscopy approaches have shown clustering of telomeres and virulence genes in 2 - 5 foci spread around the nuclear periphery (28, 29, 58-60). The nature of our data does not allow us to draw a conclusion about whether virulence genes are organized into a single large cluster as observed in our 3D model, or into multiple clusters that vary in *var* gene content among the parasite population.

Another limitation is the dependency of our 3D modeling on strong contact count signals to establish hallmarks of genome organization. In the 3D models, strongly interacting centromeres and telomeres (as seen for example in the trophozoite stage of *P. falciparum* and in *B. microti*) were organized as clusters located at the nuclear periphery. In organisms with weaker centromere interactions (*P. yoelii* and possibly *P. knowlesi*), the 3D models showed clustering of the centromeres in the center of the nucleus. Similarly, weaker telomere interactions (as seen in *P. berghei* and *T. gondii*) also formed clusters in the nuclear center. To further explore this behavior, we deleted the centromeres (all bins containing centromeric sequence) or the telomeres (40 kb region) from each genome, and generated 3D models of these modified genomes. Removal of one hallmark (centromeres or telomeres) resulted in distortion of the 3D structure and loss of clustering of the other hallmark (***SI Appendix,* Fig. S11**). For *T. gondii*, FISH experiments have shown the colocalization of telomeres at the nuclear periphery (43). We therefore believe that the central location of centromere and telomere clusters may be an artifact of our modeling approach and consider it possible that these clusters are in fact located at the nuclear periphery. For *P. knowlesi*, the telomeres were mostly not mappable and the absence of a signal from the telomeres may explain the unusual genome structure that we obtained for this parasite. Even though the centromeres showed localized interactions in the contact count heatmaps, the *P. knowlesi* 3D model displayed only weak centromere clustering. In the absence of data from telomeres due to mappability issues, the centromeres may be more dispersed in our 3D model than in reality. The mappable telomeres showed ~3-fold higher contact counts as compared to background, and it is therefore not unthinkable that the telomeres in fact cluster as well. Additional approaches, such as immunofluorescence microscopy or FISH may be necessary to confirm and further investigate our Hi-C based observations.

In conclusion, this study highlights the association between spatial organization of virulence gene families and gene expression in *Plasmodium* species. In *P. falciparum* and *P. knowlesi*, gene

families involved in antigenic variation provide a potential link between genome organization and pathogenicity. In contrast, related human parasites *Babesia microti* and *Toxoplasma gondii* lack the correlation between gene expression and genome organization observed in human *Plasmodium* species. Our results emphasize the importance of 3D genome organization in eukaryotes and suggest that genome organization in malaria parasites has been shaped by parasite-specific gene families that affect virulence and clinical phenotypes. Identifying the molecular components regulating these parasite-specific genes at the chromatin structure level will assist the identification of new targets for novel therapeutic strategies.

MATERIALS AND METHODS

We provide a brief description of methods below. Full details are available in SI Appendix, Supplementary Materials and Methods.

In situ Hi-C procedure

Parasites were crosslinked in 1.25% formaldehyde in warm PBS for 25 min on a rocking platform in a total volume between 1 and 10 ml, depending on the number of parasites harvested. Glycine was added to a final concentration as 150 mM, followed by 15 minutes of incubation at 37°C and 15 minutes of incubation at 4°C, both steps on a rocking platform. The parasites were centrifuged at 660 x g for 20 min at 4°C, resuspended in 5 volumes of ice-cold PBS, and incubated for 10 min at 4°C on a rocking platform. Parasites were centrifuged at 660 x g for 15 min at 4°C, washed once in ice-cold PBS, and stored as a pellet at -80°C. To map the inter- and intrachromosomal contact counts, crosslinked parasites were subjected to the *in situ* Hi-C procedure (32), using MboI for restriction digests.

Hi-C data processing

For each sample, the Hi-C reads were processed (i.e., mapping, filtering, pairing and removing duplicates) using HiCPro package (34). The observed intrachromosomal and interchromosomal contacts were aggregated into contact count matrices at 10-kb resolution and were then normalized using the ICE method to correct for experimental and technical biases (35).

Reproducibility score for samples from same organisms

To compute the reproducibility scores for Hi-C data from the same organisms, we used 3DChromatin_ReplicateQC and method GenomeDISCO (33) with default parameters. GenomeDISCO employs graph diffusion and random walks for transformation, and compares the smoothed contact maps between pairs to estimate global similarity. The reproducibility scores in **Supplementary Figure 1** are genome-wide concordance scores, averaged across all chromosomes.

3D modeling and visualization

We inferred a consensus 3D genome structure for each organism using the negative binomial model from the PASTIS 3D modeling toolbox (36) (https://github.com/hiclib/pastis). For assessing the stability of the consensus structures, we used 100 random initializations of the 3D coordinates to generate a set of eleven 3D models for Hi-C data from various stages, strains and conditions of the seven different apicomplexan parasites. In order to check the robustness of inferred 3D models to initialization differences, we calculated a disparity score for each possible pair of 3D models (100 choose 2) by first performing Procrustes transformation to find the best alignment and then computing the sum of the squares of the pointwise differences between the pair of structures. To determine stability of disparity, we clustered these disparity scores in cluster maps and observed very little variation among models with different initializations as well as conservation of all of the main structural features among these initializations (Supplementary Fig. 3). This allowed us to use only a single representative model for each of our Hi-C samples, which we visualized using Jmol: an open-source Java viewer for chemical structures in 3D (jmol.sourceforge.net/). PDB files and a manual to recreate the structures as displayed in Figure 2 is available at http://apicomplexan3d.lji.org/.

Colocalization tests on 3D distances

For each set of genes or functional annotations such as centromeres or telomeres, we characterized each locus/gene/annotation by including every 10kb bin that it overlaps with. For assessing the p-value of colocalization of loci within a given set, we computed the median pairwise distance for all pairs of loci within the set. Then, we randomly generated the same number of bins on the same chromosomes while preserving the genomic distance relationships for loci on each chromosome. The latter part is done by selecting an initial random locus and pairing it with an anchor locus in the real set and then choosing all other random loci that are at the same distance offset to the random anchor locus as compared to the distance of their counterparts and the real anchor locus. This approach could be considered a generalization of Witten-Noble colocalization

test (47) to 3D models instead of contact counts and to handle intra-chromosomal pairs as well as inter. We performed the randomization a 100,000 times and computed the median pairwise distance between all pairs of loci for each random selection. This median is compared to the corresponding median from the real input set to compute the p-value of observing a random set of loci that are at least as proximal to each other as the real set. The smaller the p-value the more significantly colocalized the real set of loci.

Colocalization tests on contact maps

The colocalization test for contact counts was performed using the Witten-Noble test (47). Statistical confidences of pairwise contacts were computed using Fit-Hi-C (37) and thresholded at 1% FDR to identify pairs of loci with significant interactions. The representation of loci by bins, the number of randomizations and the process to get a randomized set of loci was identical to the 3D distance test. The only difference, in this case, is that the number of significant pairs were used for the calculation of p-values, which corresponds to the chance of observing at least as many such contacts among pairs of random set of loci compared to the real set.

Quantification of insulation of each locus from neighborhood

We computed a coverage-based domain score to quantify the extent of insulation of each region from their nearest neighbors, which deviates from expected or average for regions of virulence genes in *P. falciparum* and several other parasites. For each bin x, this domain score was calculated as the average normalized contact count from bin x to upstream and downstream bins that were between 100kb to 200kb distance to x and were mappable (not filtered out by ICE normalization) in order to avoid near diagonal interactions. For bins that were unmappable (filtered by ICE) the domain score was linearly interpolated from the scores of neighboring mappable bins. The running median of domain score was computed for each bin +/-2 bins around it (5 bins total) and this smoothed score was further scaled to stay in between 0 and 1 using the overall range of scores computed for each Hi-C dataset (organism + stage). The smoothed and scaled domain score was utilized for visualization and for calling of domain-like structures using scipy.signal package find_peaks module with default parameters (score is multiplied by -1 to find dips instead of peaks). The dips corresponding to centromeric regions (+/- 10 bins) and each telomeric end of a chromosome (10% of total chromosome length) were filtered to identify internal dip regions. We then further filtered these internal dips using a prominence – i.e., the amplitude of the dip relative to the scores of surrounding regions – of 0.25, a maximum size of 10 adjacent bins.

Distance scaling plots

The relationship between genomic distance between a pair of loci and the expected number of contacts for this pair was computed using Fit-Hi-C's equal occupancy binning of contact counts (prior to the spline fit) with up to 100 distance bins within 30kb to 1Mb interval. In order to account for differences across genome size, chromosome length and chromosome number for each parasite, we normalized the average number of contact counts per pair at each distance bin by the overall average for the whole distance range up to 1Mb. We then log2 transformed these values for y-axis. X-axis is simply the log10 of the genomic distance.

Correlation between gene expression and 3D models

Gene expression was represented as a function of distance to telomeres. To generate these plots, all genes are first sorted by increasing distance to the centroid of telomeres (x-axis). Then the distance was binned into 20 equal width quantiles and the log average expression value together with the range of values in the bin (y-axis) was plotted for genes in each quantile. For the coloring of 3D models, all 10kb bins are first sorted by increasing distance to the centroid of telomeres (x-axis) and these distances are also binned into 20 equal width quantiles. For each quantile, the representative expression value that is used as the color gradient in 3D visualization was computed as the overall average of the average value of gene expression for all 10kb bins in the quantile.

Data availability

Sequence reads have been deposited in the NCBI Sequence Read Archive with accession number SRP151138. Hi-C data for *P. falciparum* trophozoites from a previous study are available from the NCBI Gene Expression Omnibus (GEO) under accession number GSE50199. Hi-C data for *P. falciparum* gametocytes and sporozoites, as well as *P. vivax* sporozoites from a previous study (31) are available from the NCBI Sequence Read Archive (SRA) under accession number SRP091967.

ACKNOWLEDGEMENTS

We thank Nelle Varoquaux (University of California,Berkeley) for help with generating 3D genome models; Kate Cook (Univer-sity of Washington, Seattle) for help with implementation of the misassembly metrics; Clay Clark, John Weger, and Glenn Hicks (Institute for Integrative Genome

Biology, University of California, Riverside) for their assistance in Illumina sequencing; the Insectary and Parasitology Core Facilities at the Johns Hopkins Malaria Research Institute, in particular Abai Tripathi, Godfree Mlambo, and Chris Kizito for their outstanding work; and The Bloomberg Family Foundation for supporting these facilities. Shoklo Malaria Research Unit is part of the Mahidol-Oxford University Research Unit, supported by the Wellcome Trust. The following reagent was obtained through the MR4 as part of the BEI Resources Repository, National Institute of Allergy and Infectious Diseases (NIAID), NIH: NF54 (MRA-1000), deposited by Megan Dowler, Walter Reed Army Institute of Research. This work was supported by NIH Grants R21 AI142506, R01 AI085077, R01AI06775, and R01 AI136511 (to K.G.L.R.), R35 GM128938 (to F.A.), and R01 AI056840 (to P.S.); the NIAID/NIH Department of Health and Human Services (Contract HHSN272201200031C), which established the Malaria Host-Pathogen Interaction Center; the Office of Research Infrastructure Programs/Office of the Director Grant P51OD011132 (to M.R.G.); University of California, Riverside Grant NIFA-Hatch-225935 (to K.G.L.R.); Bill & Melinda Gates Foundation Grant OPP1040938 (to D.A.F.); Medical Research Council Grant MR/K011782/1 (to R.T.); and the University of Texas Health Science Center at San Antonio (E.M.B.).

REFERENCES

1. Adl SM*, et al.* (2007) Diversity, nomenclature, and taxonomy of protists. *Syst Biol* 56(4):684-689.
2. WHO (2017) The World Malaria Report. http://www.who.int/malaria/publications/world-malaria-report-2017/en/.
3. Cox-Singh J*, et al.* (2008) Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening. *Clin Infect Dis* 46(2):165-171.
4. Beugnet F & Moreau Y (2015) Babesiosis. *Rev Sci Tech* 34(2):627-639.
5. Flegr J, Prandota J, Sovickova M, & Israili ZH (2014) Toxoplasmosis--a global threat. Correlation of latent toxoplasmosis with specific disease burden in a set of 88 countries. *PLoS One* 9(3):e90203.
6. Reid AJ (2015) Large, rapidly evolving gene families are at the forefront of host-parasite interactions in Apicomplexa. *Parasitology* 142 Suppl 1:S57-70.
7. Galinski MR*, et al.* (2018) Plasmodium knowlesi: a superb in vivo nonhuman primate model of antigenic variation in malaria. *Parasitology* 145(1):85-100.
8. Deitsch KW & Dzikowski R (2017) Variant Gene Expression and Antigenic Variation by Malaria Parasites. *Annu Rev Microbiol* 71:625-641.
9. Brugat T*, et al.* (2017) Antibody-independent mechanisms regulate the establishment of chronic Plasmodium infection. *Nat Microbiol* 2:16276.
10. Yam XY*, et al.* (2016) Characterization of the Plasmodium Interspersed Repeats (PIR) proteins of Plasmodium chabaudi indicates functional diversity. *Sci Rep* 6:23449.
11. Cunningham D, Lawton J, Jarra W, Preiser P, & Langhorne J (2010) The pir multigene family of Plasmodium: antigenic variation and beyond. *Mol Biochem Parasitol* 170(2):65-73.
12. Gardner MJ*, et al.* (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* 419(6906):498-511.

13. Otto TD*, et al.* (2014) A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biol* 12:86.
14. Carlton JM*, et al.* (2008) Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature* 455(7214):757-763.
15. Lodes MJ*, et al.* (2000) Serological expression cloning of novel immunoreactive antigens of Babesia microti. *Infect Immun* 68(5):2783-2790.
16. Cornillot E*, et al.* (2013) Whole genome mapping and re-organization of the nuclear and mitochondrial genomes of Babesia microti isolates. *PLoS One* 8(9):e72657.
17. Silva JC*, et al.* (2016) Genome-wide diversity and gene expression profiling of Babesia microti isolates identify polymorphic genes that mediate host-pathogen interactions. *Sci Rep* 6:35284.
18. Lorenzi H*, et al.* (2016) Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic Toxoplasma gondii genomes. *Nat Commun* 7:10147.
19. Lekutis C, Ferguson DJ, Grigg ME, Camps M, & Boothroyd JC (2001) Surface antigens of Toxoplasma gondii: variations on a theme. *Int J Parasitol* 31(12):1285-1292.
20. Howard RJ, Barnwell JW, & Kao V (1983) Antigenic variation of Plasmodium knowlesi malaria: identification of the variant antigen on infected erythrocytes. *Proc Natl Acad Sci U S A* 80(13):4129-4133.
21. Eaton MD (1938) The Agglutination of Plasmodium Knowlesi by Immune Serum. *J Exp Med* 67(6):857-870.
22. Brown KN & Brown IN (1965) Immunity to malaria: antigenic variation in chronic infections of Plasmodium knowlesi. *Nature* 208(5017):1286-1288.
23. al-Khedery B, Barnwell JW, & Galinski MR (1999) Antigenic variation in malaria: a 3' genomic alteration associated with the expression of a P. knowlesi variant antigen. *Mol Cell* 3(2):131-141.
24. Lapp SA*, et al.* (2013) Spleen-dependent regulation of antigenic variation in malaria parasites: Plasmodium knowlesi SICAvar expression profiles in splenic and asplenic hosts. *PLoS One* 8(10):e78014.
25. Chen Q*, et al.* (1998) Developmental selection of var gene expression in Plasmodium falciparum. *Nature* 394(6691):392-395.
26. Brown KN & Hills LA (1974) Antigenic variation and immunity to Plasmodium knowlesi: antibodies which induce antigenic variation and antibodies which destroy parasites. *Trans R Soc Trop Med Hyg* 68(2):139-142.
27. Autino B, Corbett Y, Castelli F, & Taramelli D (2012) Pathogenesis of malaria in tissues and blood. *Mediterr J Hematol Infect Dis* 4(1):e2012061.
28. Freitas-Junior LH*, et al.* (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. *Nature* 407(6807):1018-1022.
29. Lopez-Rubio JJ, Mancio-Silva L, & Scherf A (2009) Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell host & microbe* 5(2):179-190.
30. Ay F*, et al.* (2014) Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res* 24(6):974-988.
31. Bunnik EM*, et al.* (2018) Changes in genome organization of parasite-specific gene families during the Plasmodium transmission stages. *Nat Commun* 9(1):1910.
32. Rao SS*, et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159(7):1665-1680.
33. Ursu O*, et al.* (2018) GenomeDISCO: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics*.
34. Servant N*, et al.* (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* 16:259.

35. Imakaev M*, et al.* (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* 9(10):999-1003.
36. Varoquaux N, Ay F, Noble WS, & Vert JP (2014) A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* 30(12):i26-33.
37. Ay F, Bailey TL, & Noble WS (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res* 24(6):999-1011.
38. Dudchenko O*, et al.* (2017) De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* 356(6333):92-95.
39. Jiao WB*, et al.* (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res* 27(5):778-786.
40. Kaplan N & Dekker J (2013) High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat Biotechnol* 31(12):1143-1147.
41. Lapp SA*, et al.* (2017) PacBio assembly of a Plasmodium knowlesi genome sequence with Hi-C correction and manual annotation of the SICAvar gene family. *Parasitology* 145(1):1-14.
42. Brooks CF*, et al.* (2011) Toxoplasma gondii sequesters centromeres to a specific nuclear region throughout the cell cycle. *Proc Natl Acad Sci U S A* 108(9):3767-3772.
43. Gissot M*, et al.* (2012) Toxoplasma gondii chromodomain protein 1 binds to heterochromatin and colocalises with centromeres and telomeres at the nuclear periphery. *PLoS One* 7(3):e32671.
44. Duan Z*, et al.* (2010) A three-dimensional model of the yeast genome. *Nature* 465(7296):363-367.
45. Tanizawa H*, et al.* (2010) Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic acids research* 38(22):8164-8177.
46. Stevens TJ*, et al.* (2017) 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 544(7648):59-64.
47. Witten DM & Noble WS (2012) On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic acids research* 40(9):3849-3855.
48. Lieberman-Aiden E*, et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289-293.
49. Mirny LA (2011) The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res* 19(1):37-51.
50. Eser U*, et al.* (2017) Form and function of topologically associating genomic domains in budding yeast. *Proc Natl Acad Sci U S A* 114(15):E3061-E3070.
51. Hehl AB*, et al.* (2015) Asexual expansion of Toxoplasma gondii merozoites is distinct from tachyzoites and entails expression of non-overlapping gene families to attach, invade, and replicate within feline enterocytes. *BMC Genomics* 16:66.
52. Kishore SP, Perkins SL, Templeton TJ, & Deitsch KW (2009) An unusual recent expansion of the C-terminal domain of RNA polymerase II in primate malaria parasites features a motif otherwise found only in mammalian polymerases. *J Mol Evol* 68(6):706-714.
53. Ukaegbu UE*, et al.* (2014) Recruitment of PfSET2 by RNA polymerase II to variant antigen encoding loci contributes to antigenic variation in P. falciparum. *PLoS Pathog* 10(1):e1003854.
54. Jiang L*, et al.* (2013) PfSETvs methylation of histone H3K36 represses virulence genes in Plasmodium falciparum. *Nature* 499(7457):223-227.
55. Behnke MS, Zhang TP, Dubey JP, & Sibley LD (2014) Toxoplasma gondii merozoite gene expression analysis with comparison to the life cycle discloses a unique expression state during enteric development. *BMC Genomics* 15:350.

56. Radke JB*, et al.* (2018) Transcriptional repression by ApiAP2 factors is central to chronic toxoplasmosis. *PLoS Pathog* 14(5):e1007035.
57. Woo YH*, et al.* (2015) Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *Elife* 4:e06974.
58. Brancucci NM*, et al.* (2014) Heterochromatin protein 1 secures survival and transmission of malaria parasites. *Cell Host Microbe* 16(2):165-176.
59. Flueck C*, et al.* (2010) A major role for the Plasmodium falciparum ApiAP2 protein PfSIP2 in chromosome end biology. *PLoS Pathogens* 6(2):e1000784.
60. Ralph SA, Scheidig-Benatar C, & Scherf A (2005) Antigenic variation in Plasmodium falciparum is associated with movement of var loci between subnuclear locations. *Proc Natl Acad Sci U S A* 102(15):5414-5419.
61. Bensch S*, et al.* (2016) The Genome of Haemoproteus tartakovskyi and Its Relationship to Human Malaria Parasites. *Genome Biol Evol* 8(5):1361-1373.

FIGURE LEGENDS

Figure 1: Overview of samples and protocol. A) Phylogenetic tree showing the genetic relationship between the seven different apicomplexan parasites used in this study (adapted from (61). **B)** Light microscopy images of the various parasites. **C)** Schematic representation of the in situ Hi-C protocol.

Figure 2: Hi-C data and 3D genome modeling. A) Normalized interchromosomal contact count heatmaps at 10-kb resolution. Chromosomes are lined up in numerical order starting with chr1 in the bottom left corner. Individual chromosomes are delineated by dashed lines. Intrachromosomal contacts are not displayed, hence the white squares along the diagonal of each heatmap. Larger versions of the interchromosomal heatmaps as well as all intrachromosomal heatmaps are accessible at http://apicomplexan3d.lji.org/. **B)** Representative 3D models for genomes of all organisms studied here. Chromosomes are shown as transparent white ribbons. Centromeres are indicated with red spheres, telomeres with blue spheres and virulence genes with orange spheres. Representative 3D models for asynchronous *B. microti* blood stages and *T. gondii* tachyzoites are shown in Supplementary Figure 2.

Figure 3: Correction of the *T. gondii* genome assembly by Hi-C data. A) Normalized interchromosomal contact count heatmap plotted using the current version of the *T. gondii* genome showing unusually high levels of contact counts between chrVIIb and chrVIII (inside the red box). **B)** Normalized interchromosomal contact count heatmap plotted using an updated version of the *T. gondii* genome in which chrVIIb and chrVIII have been stitched to form one large chrVIII. All interchromosomal contacts are at expected levels and the newly formed chromosome shows a single centromeric interaction with each other chromosome. **C)** The intrachromosomal contact count heatmap of the newly formed chromosome chrVIII, showing no discrepancies in contact counts along the chromosome or at the junction. **D)** Misassembly metric for the newly formed chrVIII, showing no signs of misassembly. The junction is indicated with a dashed line.

Figure 4: Formation of domain-like structures and chromosome loops by *var* and *SICAvar* genes. Top row: normalized intrachromosomal contact count heatmaps at 10-kb resolution for representative chromosomes, showing a canonical “X” shape for chromosomes of *P. berghei* and *P. yoelii*, and domain-like structures in chromosomes with internal *var* and *SICAvar* genes in *P. falciparum* and *P. knowlesi*, respectively. Second row: domain score tracks. Dips in the tracks

that reach the threshold of a domain-like structure are marked with a black box and centromeres are marked with a black dashed line. Third row: mappability tracks. Bottom row: individual chromosome conformation extracted from the 3D model of the full genome. *P. berghei* and *P. yoelii* chromosomes show a folded structure anchored at the centromere, with both chromosome arms arranged in parallel. *P. falciparum* and *P. knowlesi* chromosomes show additional folding structures to bring virulence genes in close spatial proximity.

Figure 5: Correlation between genome organization and gene expression. A) Relation between gene expression and distance from the centroid of the telomeres. For each organism, genes were divided into 20 bins. For each bin, the average gene expression value was plotted. Error bars denote the range of expression values within each bin. **B)** Average gene expression values of each bin were projected onto the 3D models, using a color scale ranging from dark blue (low gene expression) to white (high gene expression). Centromeres are shown as yellow spheres, while telomeres are depicted as red spheres.