nature microbiology

Article

https://doi.org/10.1038/s41564-023-01360-8

Babesia duncani multi-omics identifies virulence factors and drug targets

Received: 20 May 2022

Accepted: 14 March 2023

Published online: 13 April 2023

Check for updates

Pallavi Singh[1,10], Stefano Lonardi[2,10] ✉, Qihua Liang[2], Pratap Vydyam[1], Eleonora Khabirova[3], Tiffany Fang[1], Shalev Gihaz[1], Jose Thekkiniath[1], Muhammad Munshi[1], Steven Abel[4], Loic Ciampossin[4], Gayani Batugedara[4], Mohit Gupta[4], Xueqing Maggie Lu[4], Todd Lenz[4], Sakshar Chakravarty[2], Emmanuel Cornillot[5], Yangyang Hu[2], Wenxiu Ma[3], Luis Miguel Gonzalez[6], Sergio Sánchez[7], Karel Estrada[8], Alejandro Sánchez-Flores[8], Estrella Montero[6], Omar S. Harb[9], Karine G. Le Roch[4] ✉ & Choukri Ben Mamoun[1] ✉

Babesiosis is a malaria-like disease in humans and animals that is caused by *Babesia* species, which are tick-transmitted apicomplexan pathogens. *Babesia duncani* causes severe to lethal infection in humans, but despite the risk that this parasite poses as an emerging pathogen, little is known about its biology, metabolic requirements or pathogenesis. Unlike other apicomplexan parasites that infect red blood cells, *B. duncani* can be continuously cultured in vitro in human erythrocytes and can infect mice resulting in fulminant babesiosis and death. We report comprehensive, detailed molecular, genomic, transcriptomic and epigenetic analyses to gain insights into the biology of *B. duncani*. We completed the assembly, 3D structure and annotation of its nuclear genome, and analysed its transcriptomic and epigenetics profiles during its asexual life cycle stages in human erythrocytes. We used RNA-seq data to produce an atlas of parasite metabolism during its intraerythrocytic life cycle. Characterization of the *B. duncani* genome, epigenome and transcriptome identified classes of candidate virulence factors, antigens for diagnosis of active infection and several attractive drug targets. Furthermore, metabolic reconstitutions from genome annotation and in vitro efficacy studies identified antifolates, pyrimethamine and WR-99210 as potent inhibitors of *B. duncani* to establish a pipeline of small molecules that could be developed as effective therapies for the treatment of human babesiosis.

Recent reports suggest an increase in the incidence of tick-borne bacterial, parasitic and viral infections worldwide[1]. This is largely due to changes in environmental factors that led to the expansion of the geographical distribution of the tick vectors, several of which carry multiple human pathogens[1]. Among these pathogens, *Babesia* parasites are considered a serious threat to humans and animals[2,3]. All cases of human babesiosis reported in the United States have been linked to either *Babesia microti* (majority of cases), *B. duncani* or a *B. divergens*-like species called MO1[3]. Available data suggest that *B. duncani* is transmitted by the hard tick *Dermacentor albipictus*[4]. The first clinical isolate of *B. duncani* (WA1) was collected in 1991 from a patient who lacked all of the typical risk factors for fulminating babesiosis (that

A full list of affiliations appears at the end of the paper. ✉e-mail: stelo@cs.ucr.edu; karinel@ucr.edu; choukri.benmamoun@yale.edu

is, old age, splenectomy and weak immune system)[5]. Previous phylogenetic analyses based on the 18S rRNA gene and the mitochondrial *cox-1* genes revealed that *B. duncani* belongs to Clade II of piroplasmids and defines a new lineage that is closely related to *Babesia negevi*, a causative agent of canine babesiosis. This clade is distinct from those encompassing *B. microti* (Clade I), *B. bovis* (Clade VI), *Theileria* spp. (Clade IV and V) and *Cytauxzoon* spp. (Clade III)[6]. Recent studies using a continuous in vitro culture of the parasite in human erythrocytes showed that *B. duncani* has low susceptibility to atovaquone, azithromycin, clindamycin and quinine[7]. Consistent with the severe clinical outcome of *B. duncani* infection in humans[5,8–11], studies in immunocompetent hamsters, and both immunocompetent and immunocompromised mice confirmed the high virulent status of *B. duncani*[12].

Despite the highly virulent properties of *B. duncani*, little is known about its biology, evolution and mechanism of virulence. Here we report presumably the first completed nuclear genome sequence, assembly, three-dimensional (3D) structure and transcriptional annotation of this parasite. We also characterized its epigenomic landscape and localized the active and inactive histone marks. Our analysis further revealed that the parasite has evolved new classes of multigene families (MGFs). Drug target mining and in vitro efficacy studies identified the antifolates, pyrimethamine and WR-99210, as excellent inhibitors of parasite development within human erythrocytes. The high potency of these compounds provides promise for the future development of more effective therapies.

Results

B. duncani genome sequencing and assembly

Babesia duncani asexual development takes place within human erythrocytes or those of the reservoir host (the mule deer *Odocoileus hemionus*), whereas its sexual cycle occurs in the tick vector (*Dermacentor albipictus*) (Extended Data Fig. 1)[3,4]. The parasite can be transmitted from ticks to humans either through a tick-bite or through transfusion of *Babesia duncani*-infected blood (Extended Data Fig. 1a). During its intraerythrocytic life cycle, *B. duncani* undergoes various morphological changes that manifest in four developmental stages: paired pyriform, young rings, mature and filamentous rings and tetrads (Extended Data Fig. 1b). To gain further insight into the genomic content and structure of *B. duncani*, we purified its total genomic DNA from a culture of the parasite in human erythrocytes[7]. A high-quality genome assembly was produced by processing PacBio HiFi WGS long reads (~150x coverage) and Illumina WGS paired-end reads (~130x coverage) (Supplementary Fig. 1). Reads enriched for nuclear DNA reads were assembled with HiCANU[13]. The assembly was polished using Polypolish[14] and scaffolded/corrected using a Bionano optical map (Supplementary Fig. 2 and Table 1). The final assembly had five chromosome-level scaffolds of 3.13 Mb (Chromosome (Chr) I), 1.58 Mb (Chr II), 1.42 Mb (Chr III), 1.07 Mb (Chr IV) and 0.35 Mb (Chr V) with a total of only 35 Kb in two gaps (see Extended Data Table 1 for other statistics). The assembly pipeline is illustrated in Supplementary Fig. 1 and is described in more details in Methods. Analysis using BUSCO v5 (ref. 15) was used to determine that the assembly has 95.1% of the gene models in the apicomplexa_odb10 database (85.2% single-copy and 9.9% duplicated). Long-range chromatin contact frequency information using Hi-C allowed creation of contact maps for *B. duncani* WA1 (shown on Figs. 1a and 2b). The contact maps show that the assembly does not contain large mis-joints or mis-assemblies and is consistent with that of other apicomplexan parasites[16]. The centromeres interact strongly with each other (as illustrated in Fig. 1b(bottom)). Chr I was found to have a metacentric centromere; Chr II and III have a telocentric profile, and Chr IV and V have acrocentric centromeres (Supplementary Fig. 3). A similar organization was previously observed in three of the four chromosomes of *Babesia bovis*[17].

The chromosomal organization of the *B. duncani* nuclear genome was further validated using pulsed field gel electrophoresis (PFGE) analysis. Five bands with approximate sizes of ~3.1 Mb (Chr I), ~1.81 Mb (Chr II), ~1.37 Mb (Chr III), ~1.05 Mb (Chr IV) and <1 Mb (Chr V) were identified (Fig. 1c, lane 1). Using a specific telomeric probe from *Plasmodium berghei*, each of the 5 chromosomes were labelled using Southern-blot analysis (Fig. 1c, lane 2). Altogether, these results demonstrate that *B. duncani* WA1 nuclear genome consists of five chromosomes. A 3D model was constructed from Hi-C contact maps showing genome-wide chromatin organization (Fig. 1d). The five chromosomes showed strong interactions among the centromeres (Fig. 1b,d), an organization similar to that of recently reported 3D structures of apicomplexan parasites including *Babesia microti*[16,18]. Full interactions were observed between the telomere ends, whereas only partial interactions were detected between the centromeres and telomeres (Fig. 1d).

B. duncani genome annotation

To accurately predict gene loci, we generated Illumina RNA-seq and PacBio IsoSeq data using RNA isolated from parasites propagated in vitro in human erythrocytes. The gene annotation pipeline used FunAnnotate and InterProScan to determine gene loci and functional annotations. The final set of gene models included 4,222 gene loci (including 52 transfer (t)RNA genes) with an average length of 1,656 bp (average protein length was 499 amino acids) (Extended Data Table 2). Of the total genes, 63% were multi-exon and 76% had a functional annotation. The *B. duncani* nuclear genome expresses two 18S rRNA encoding genes, one on Chr I and the second on Chr III, three 5S rRNAs genes in tandem on Chr II, and 46 tRNAs distributed on all five chromosomes with another 6 found in unplaced contigs. Transcript analysis using BUSCO v5 (ref. 15) showed that the transcripts' annotations match 88.6% of the transcripts in the apicomplexa_odb10 database (80.5% single-copy and 8.1% duplicated), whereas 5.8% are fragmented and 5.6% are missing. BUSCO analysis further showed that the *B. duncani* protein annotations match 87.7% of the proteins in the apicomplexa_odb10 database (79.6% single-copy and 8.1% duplicated), whereas 5.6% are fragmented and 6.7% are missing. Only 1.2% of the genome is considered repetitive by RepeatMasker.

Genome-wide comparison of *B. duncani* gene models with other Apicomplexa

The *B. duncani* nuclear genome has a GC content of 37.32%, which is similar to that of other piroplasmids[19] and almost double that of *Plasmodium falciparum* (19.34%)[20]. Main genome and gene statistics of *B. duncani* and other Apicomplexa are shown in Extended Data Table 2. The genome size, number of chromosomes and number of protein coding genes in *B. duncani* are higher than those in *B. microti, B. bovis, Theileria. parva, Cryptosporidium. parvum* and *Theileria. annulate*, which could be also attributed to a more complete genome assembly for *B. duncani*. An approximate 60% of the *B. duncani* genome encodes proteins, with ~63% of the genes containing introns with an average number of exons per gene of 3.1.

Out of the total 4,222 *B. duncani* genes, 3,484 mapped evenly across the five chromosomes while the rest were located on unplaced contigs. A summarized comparative genomics analysis between *B. duncani* and other Apicomplexa species is presented in Fig. 2a. The bars depict the number of orthologous proteins shared between the compared species, as well as unique proteins for each organism and their percentage of annotation (upper dot plot). Interestingly, we found 842 core proteins that are shared between *B. duncani* and other Apicomplexa, where 100% of the proteins have a functional annotation. Interestingly, the *B. duncani* genome possesses 1,242 unique proteins only ~30% of which have a functional annotation, suggesting that there could be several predicted proteins with unknown functions. In contrast, *B. bovis* has 565 unique proteins, ~50% of which are annotated, and *P. falciparum* has 3,674 proteins, ~55% of which are annotated. These differences could be due to different genome size (as seen in Fig. 2b),

https://doi.org/10.1038/s41564-023-01360-8

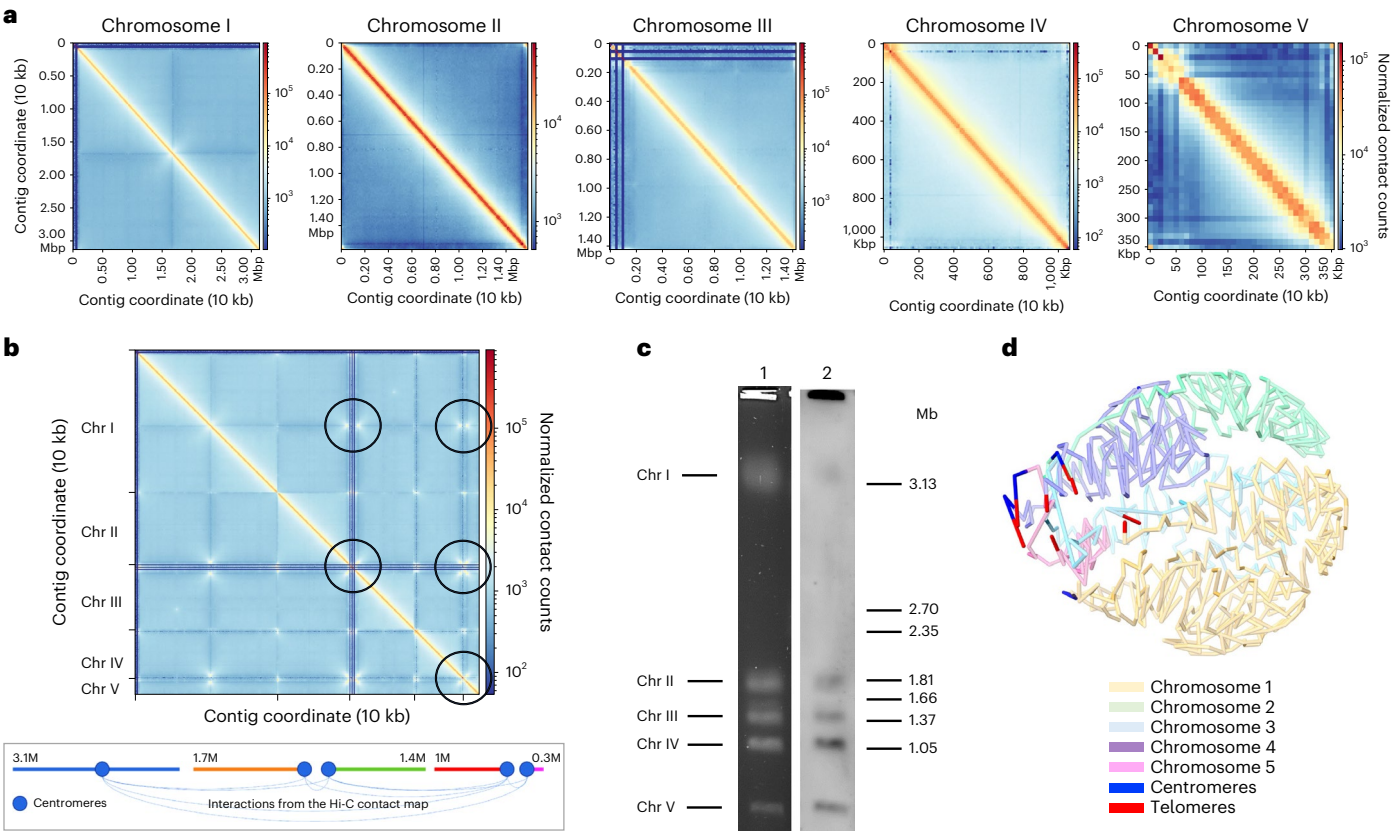


Fig. 1 | *B. duncani* genome organization. a, Hi-C interchromosomal contact maps for the five *B. duncani* chromosomes (10 Kb bins). **b**, Full Hi-C contact map for *B. duncani* (10 Kb bins); black circles highlight strong intra-chromosomal interaction, which are summarized in the bottom panel; X-shaped interactions typically indicate the presence of centromeres. **c**, PFGE (lane 1) and subsequent Southern-blot analyses using a *Plasmodium berghei* telomeric probe (lane 2) showing the number and approximate size of *B. duncani* Chr: ~3.1 Mb, (Chr I), ~1.81, Mb (Chr II), ~1.37 Mb (Chr III), ~1.05 Mb (Chr IV) and <1 Mb (Chr V). *Hansenula wingei* DNA chromosomes were used as DNA markers. The manufacturer's estimates of the sizes of chromosomes are indicated in megabase pairs (Mb) on the right. The experiment was performed in biological duplicates. **d**, 3D genome structure of *B. duncani* derived from the contact map interactions.

quality of assembly and annotation, gene paralogy or species-specific related proteins.

To investigate the evolutionary distance and relationships among piroplasms, the phylogeny for 19 species (18 different Apicomplexa with *Tetrahymena hominis* as the outgroup) was determined using a maximum-likelihood approach and 190 single-copy orthologous proteins (see Methods). As shown in Fig. 2c, *B. duncani* is a defining member of a separate clade in Piroplasmida. This analysis provided further support to previous clade classification of piroplasms[21] using other genetic markers, which showed that *B. duncani* belongs to a clade (Clade II) distinct from that of *B. microti* (Clade I)[22] and other human *Babesia* parasites (Fig. 2c). *B. duncani* also shares the same ancestor with *Babesia* species of Clade VI, which includes *B. bovis, B. bigemina* and *B. divergens* (Fig. 2c). Synteny-based analyses (Extended Data Fig. 2) indicated that *B. duncani* is least sytenic with *B. microti* compared with *T. parva, B. bovis* and *B. bigemina* (Extended Data Fig. 2).

The metabolic atlas of *B. duncani* was reconstituted from the annotated genome. All enzymes of the glycolytic pathway and the tricarboxylic acid cycle were found in the annotated proteome (Extended Data Tables 3 and 4). An estimated 11.4% of the coding sequences harbour an N-terminal signal peptide, thus highlighting the importance of protein secretion in *B. duncani* intraerythrocytic development. Of all the predicted proteins of *B. duncani*, 2.5% of coding sequences carry a mitochondrial targeting sequence, 19% have at least one predicted transmembrane domain and 0.1% have 18 transmembrane domains. The most conserved genes between *B. duncani* and other *Babesia* species are those involved in DNA replication, transcription and protein translation. Our analysis identified 19 members of the Apicomplexan Apetala 2 (ApiAP2) family and 17 GPI-anchored proteins (Extended Data Tables 5 and 6, and Supplementary Fig. 4).

New multigene families of *B. duncani*

The extremely high virulence of *B. duncani* and the cytokine storm it triggers in immunocompetent mice and hamsters suggest that the parasite produces virulence factors during its blood stage development that are delivered into the host and trigger a strong host response[12]. Interestingly, analysis of the *B. duncani* genome identified 747 genes that belong to MGFs, each with at least three members (Fig. 3). These families can be divided into two classes: BdUMGFs (unique to *B. duncani*) and BdOMGFs (with orthologues in other Apicomplexa) (Fig. 3a,b). The BdUMGF class includes 397 genes grouped into 73 gene families, whereas the BdOMGFs comprise 420 genes grouped into 105 gene families, each belonging to an orthologous group among Apicomplexa (Fig. 3b). The components of each of the gene families and their properties are available in the supplementary datasets. The chromosomal location of the genes in the three largest BdUMGF and BdOMGF families, as well as the activator protein (AP) family, is illustrated in Fig. 3c. Our analysis showed that genes in the BdUMGF1 strongly co-localize and are on the telomeres of Chr I and IV, which also appear to be in close proximity in the 3D genome model (Fig. 3c and also Fig. 4a and Supplementary Fig. 4). All genes in the BdUMGF3 family (shown in blue) are clustered together on the same strand in the

https://doi.org/10.1038/s41564-023-01360-8

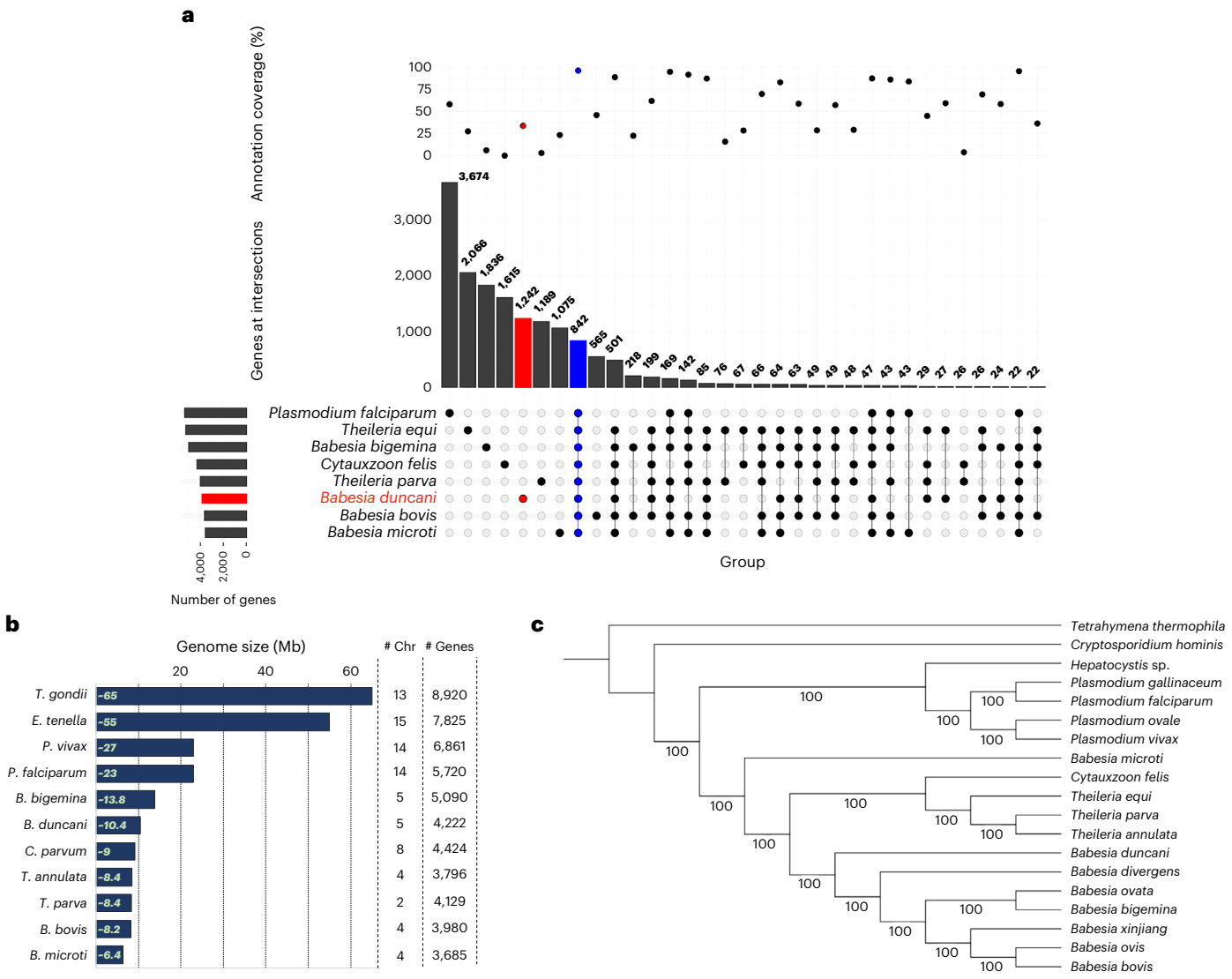


Fig. 2 | Evolutionary analysis of *B. duncani* genome. a, Upset plot for orthology results between *B. duncani* and other organisms. Top: the percentage of annotated proteins for shared or unique ones from a given organism is presented. Middle: the total number of shared proteins or unique ones from a given organism is depicted. Bottom: the intersection or uniqueness of a given species, with horizontal bars at the left side representing the total number of genes for a given species. **b**, Schematic showing genome sizes in megabases of various Apicomplexan species. Gene numbers include both coding and non-coding genes. **c**, Maximum-likelihood gene orthology-based phylogenetic tree analysis using single-copy orthologues from *B. duncani* and other species. Bootstrap values are shown at each node.

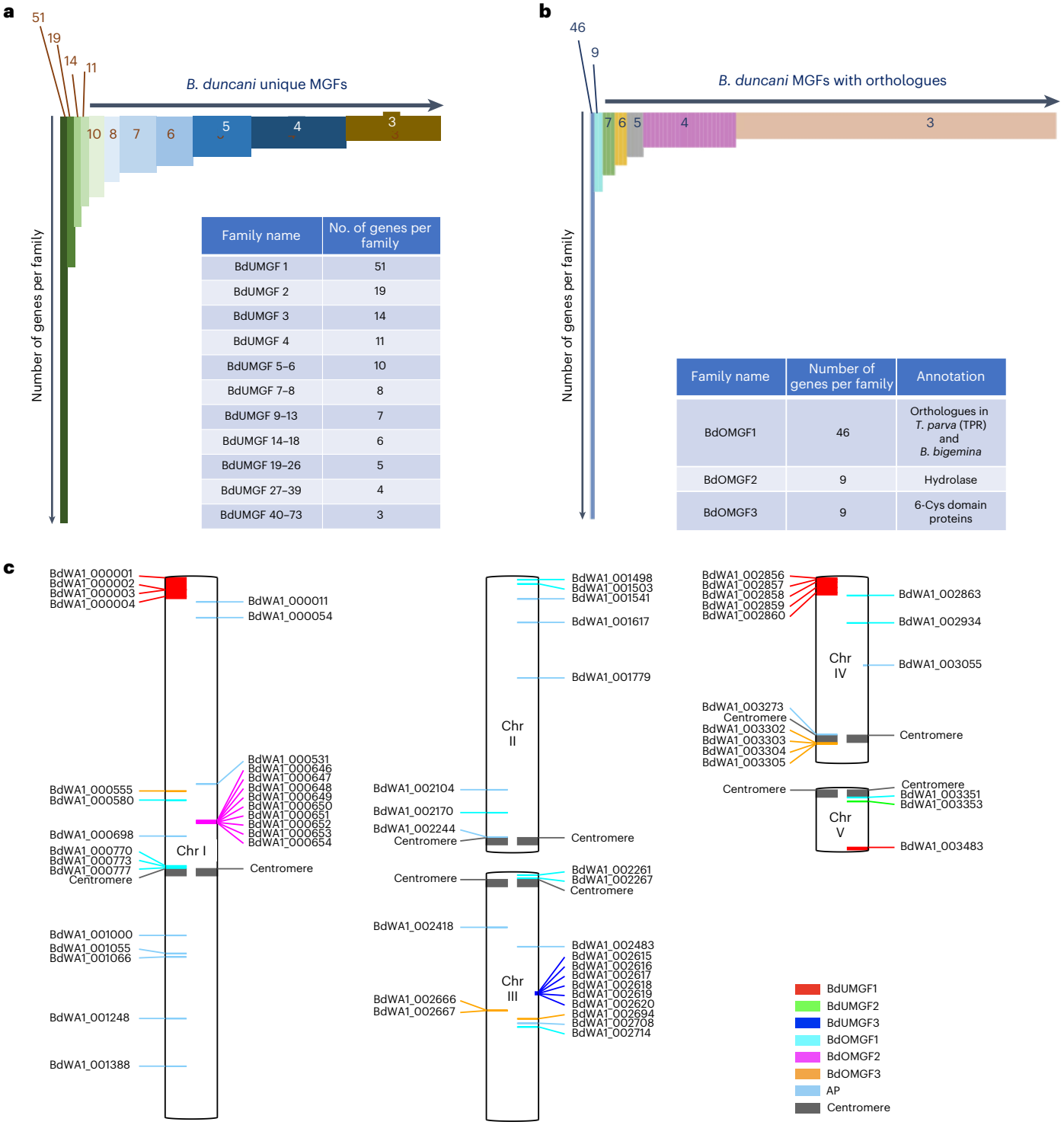
middle of Chr III. All genes in the BdOMGF2 family (shown in magenta) are clustered together on the same strand in the middle of Chr I. The genes in the AP family are uniformly distributed along the chromosomes.

The distribution of BdUMGF1 gene family members near the telomeres as determined by Hi-C analysis (Fig. 4a) suggests a possible heterochromatin clustering associated with transcriptional silencing surrounding these genes as described for the *var* or *SICAvar* gene families of *P. falciparum* and *P. knowlesi*, respectively[23,24]. The 3D genome model of the *B. duncani* genome based on Hi-C analysis also determined that members of the BdUMGF3 multigene family are present outside the telomeric and centromeric regions (Fig. 4a). Interestingly, despite this localization, RNA-seq showed that these genes are repressed (Supplementary Data file 1). DNA fluorescence in situ hybridization (FISH) using probes specific to BdUMGF1 to mark Chr I and IV, and a probe specific to BdUMGF3 to mark Chr III demonstrated that BdUMGF1 and BdUMGF3 are distant from each other in the nucleus since two distinct non-overlapping signals were observed for each family (Fig. 4b and Supplementary Fig. 5). In contrast, the members of BdUMGF1 on Chr I and IV are in close proximity to each other as two partially overlapping signals were observed using DNA FISH (Fig. 4b and Supplementary Fig. 5). These data suggest that the transcriptional repression of the BdUMGF3 family is not due to their proximity to the members of BdUMGF1 family (Fig. 4a,b).

Regulation of gene expression in *B. duncani*

Using transcript abundance values based on the mean depth of coverage (see Methods), we found that the total range of transcriptional activity captured using the continuous in vitro growth conditions varied by more than four orders of magnitude (Fig. 5a–d and supplementary datasets). Overall, RNA-seq data captured close to 90% of the predicted annotated genes in the assembled *B. duncani* genome (Fig. 5a,b), indicating that most genes are expressed during the intraerythrocytic stages of the parasite life cycle and are potentially needed for parasite

https://doi.org/10.1038/s41564-023-01360-8



Family name	No. of genes per family
BdUMGF 1	51
BdUMGF 2	19
BdUMGF 3	14
BdUMGF 4	11
BdUMGF 5–6	10
BdUMGF 7–8	8
BdUMGF 9–13	7
BdUMGF 14–18	6
BdUMGF 19–26	5
BdUMGF 27–39	4
BdUMGF 40–73	3

Family name	Number of genes per family	Annotation
BdOMGF1	46	Orthologues in *T. parva* (TPR) and *B. bigemina*
BdOMGF2	9	Hydrolase
BdOMGF3	9	6-Cys domain proteins

Fig. 3 | Multigene families and chromosomal localization in the *B. duncani* genome. a, Gene families unique to *B. duncani*. **b**, Gene families with orthologues in other Apicomplexa. TPR, *Theileria parva* repeat gene family. **c**, Localization of the genes in the gene families BdUMGF1, BdUMGF2, BdUMGF3, BdOMGF1, BdOMGF2, BdOMGF3 and AP on the five *B. duncani* chromosomes (genes localized on unplaced contigs are ignored); genes on the right side of a chromosome are on the positive strand, genes on the left side are on the negative strand.

survival within host erythrocytes. We further examined the possible relationship between gene expression and genome organization. The genes of *B. duncani* were binned into 20 groups on the basis of their distance from the centroid of telomeres. The average gene expression using RNA-seq data was calculated (Fig. 5c) and the normalized average gene expression values across the whole chromosomes were colour-coded (Fig. 5b). Similar to what was observed in the *P. falciparum* genome that possesses gene families involved in antigenic variation, we detected a significant relationship between gene expression and 3D location relative to the telomeres, with noticeable repression near the telomeric ends of Chr I, IV and V, which harbour gene clusters belonging to the BdUMGF1 (Fig. 3c), and suggests that the *B. duncani* genome may

https://doi.org/10.1038/s41564-023-01360-8

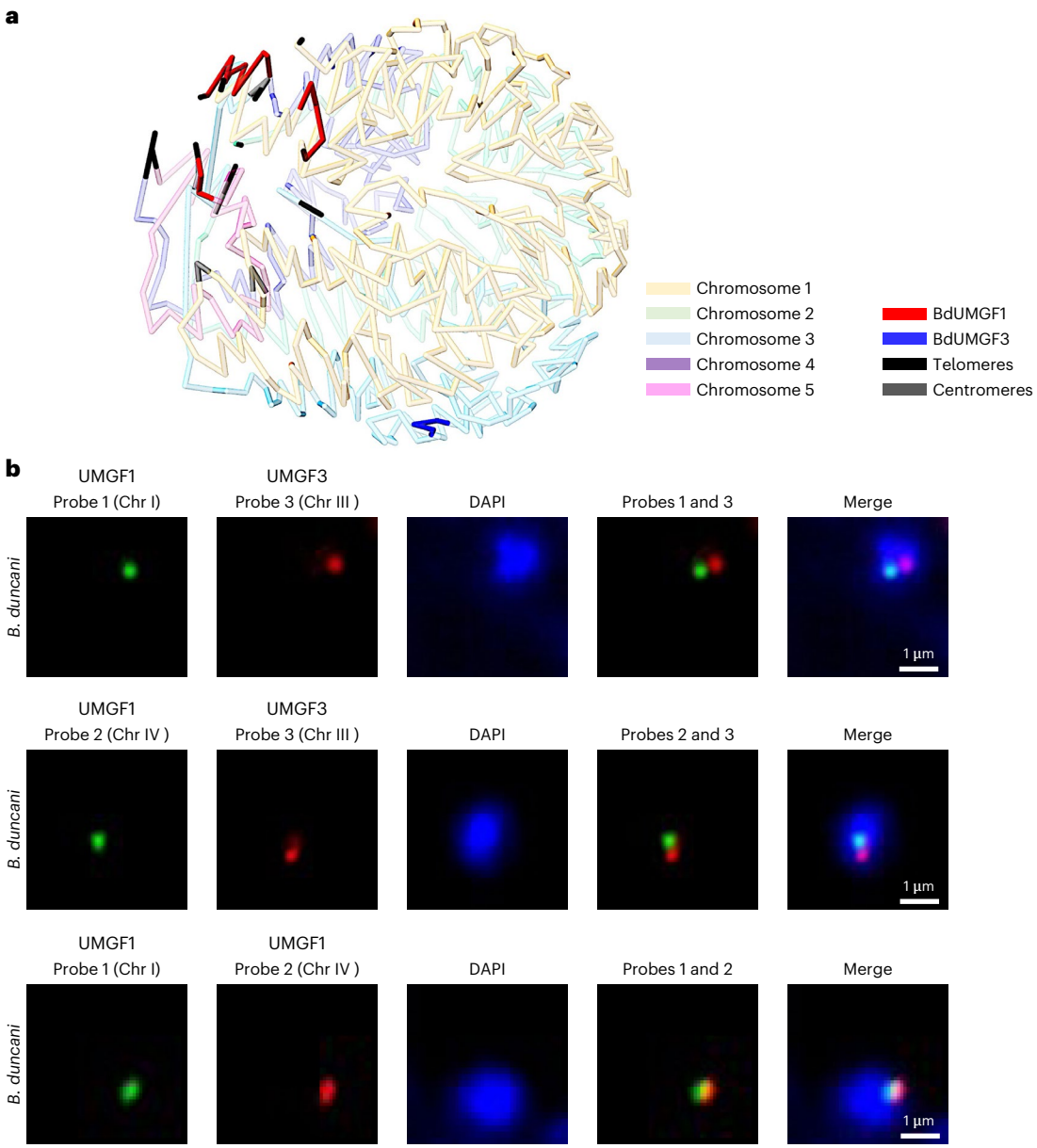


Fig. 4 | *B. duncani* unique multigene family (BdUMGF) members are spatially distinct from each other in the genome. a, *B. duncani* WA1 3D genome model showing the location of BdUMGF1 (red) and BdUMGF3 (dark blue) on different chromosomes. **b**, The detection of BdUMGF1 (located on Chr I and Chr IV) and BdUMGF3 (located on Chr III) using a DNA probe labelled with either fluorescein (Probes 1 and 2, green) or biotin/avidin-rhodamine (Probe 3, red). The detection of BdUMGF1 on Chr I with DNA probe labelled with fluorescein (Probe 1) and on Chr IV using a DNA probe labelled with biotin/avidin-rhodamine (Probe 2, red). Nuclear DNA is stained with DAPI (blue). Representative images from two independent experiments performed in duplicates.

contain a heterochromatin cluster near the telomere ends allowing for mono-allelic expression of the BdUMGF1.

Our analysis revealed that among the most highly expressed genes are genes involved in protein translation, protein degradation, cell cycle, ion transport, carbohydrate metabolism and histone core proteins (Fig. 5d).

Epigenomic landscape of *B. duncani*

The expression profile of the *B. duncani* multigene families is reminiscent of that of the *P. falciparum var* gene family, which consists of ~60 genes that encode a major virulence antigen, PfEMP1, with a single variant expressed in each infected red blood cell[25]. PfEMP1-mediated antigenic variation has been suggested to play a role in *P. falciparum* immune evasion[26]. The control of mutually exclusive *var* gene expression in the parasite relies on epigenetic and chromatin structural changes that are critical for pathogenesis and immune evasion[27,28]. To assess whether members of the BdUMGF family undergo similar epigenetic regulation and thus could undergo antigenic variation, we conducted chromatin immunoprecipitation assays (ChIP) to identify both the specific histone marks and the affected genes. ChIP was conducted using antibodies against tri-methylated histone 3 lysine 9 (H3K9me3) and acetylated histone 3 lysine 9 (H3K9ac) as markers for heterochromatin and euchromatin marks, respectively. The immuno-precipitated DNA and input used as a positive control were purified, amplified and subjected to next-generation sequencing on the Illumina Novaseq sequencing platform. Reads were mapped to the *B. duncani* genome and normalized per million of mapped reads for each sample (Fig. 5e). Pearson correlation coefficients between each ChIP-seq

https://doi.org/10.1038/s41564-023-01360-8

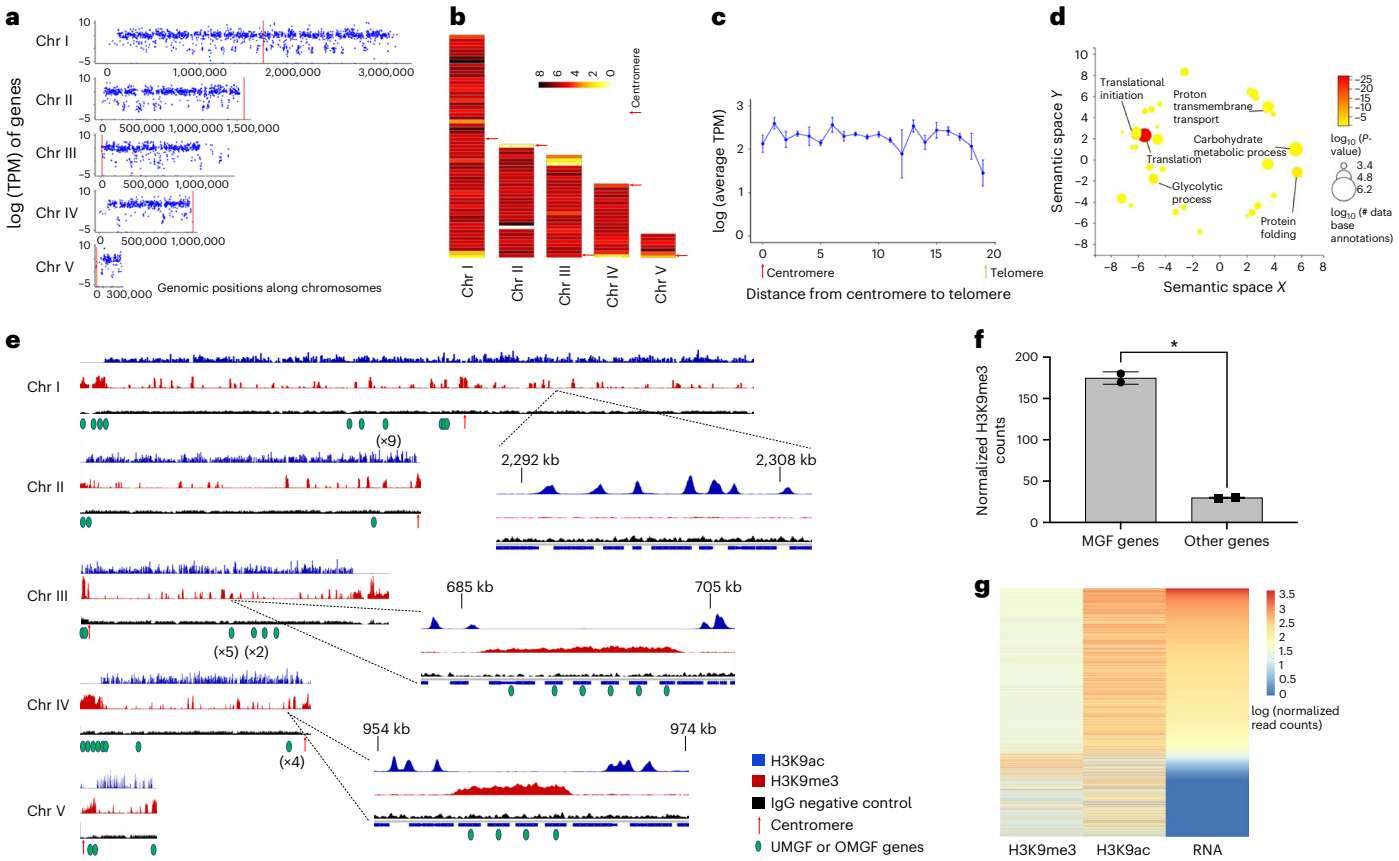


Fig. 5 | Transcriptomic and epigenomic profiles of *B. duncani*. **a**, Logarithms of the TPM counts were used as expression values0 for each gene across the 5 chromosomes using the R package ggplot2. **b**, RNA-seq data as normalized heat maps across the 5 chromosomes. Chromosomes were divided into 50 kb bins and the average of the log TPM of genes within each bin was calculated. **c**, Relation between gene expression and distance from the centromeres to the telomeres. Chromosomes 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 for each chromosome. $n = 2$ biologically independent samples. **d**, Gene ontology (GO) enrichment of the most highly expressed genes. GO was calculated using the R package TopGO with the weight01 algorithm and the biological process tree. REVIGO was used to visualize the GO results of the most highly expressed genes (500 TPM or higher). Fisher's exact test, one-sided, with the weight01 algorithm. **e**, Heterochromatin and euchromatin distribution across all five *B. duncani* chromosomes. Tracks correspond to H3K9me3-ChIP, H3K9ac-ChIP and IgG control and are normalized to millions of mapped reads. **f**, Normalized H3K9me3 counts in multigene families, and other genes encoded in the *B. duncani* genome (unpaired *t*-test with Welch's correction, $P = 0.022$). $n = 2$ biologically independent samples. **g**, Comparison between epigenetic marks and gene expression. Heat maps were generated using normalized $\log_2$ H3K9me3 and H3K9ac read counts as well as the RNA-seq TPM levels for each gene. Read counts for H3K9me3 and H3K9ac were normalized to millions of mapped reads and gene length, while TPM was determined by Stringtie. Genes were sorted from high to low TPM highlighting the correlation and anti-correlation between transcript abundance and the H3K9ac and H3K9me3 marks, respectively.

sample confirmed the reproducibility of our data (Supplementary Table 2) and demonstrated that euchromatin and heterochromatin marks are mutually exclusive (Fig. 5e–g). To identify genomic regions that are associated with hetero- and euchromatin, we visualized the ChIP-seq data using the Integrative Genomic Viewer (IGV) (Fig. 5e). Similar to findings in *Plasmodium*, the heterochromatin is localized to telomeric and subtelomeric regions of all chromosomes, except for the right end of Chr I which is depleted of multigene families. Heterochromatin marks were also observed on multigene families found in internal chromosome clusters as well as in centromeric regions, a profile similar to that observed in *T. gondii*[29]. Statistical analyses were then used to determine whether genes from multigene families were significantly enriched with H3K9me3 marks compared with other genes encoded in the *B. duncani* genome. Our data demonstrate that in *B. duncani*, genes that belong to multigene families (UMGFs or OMGFs) are significantly enriched in H3K9me3 marks (Fig. 5f). Additional histone H3K9me3 marks were also observed in genes throughout the genome (Fig. 5g); however, most of these genes were not expressed during the intraerythrocytic life cycle. Our analysis further identified additional genes that are annotated as hypothetical proteins as they lack any homologues in other organisms and potentially belong to novel multigene families. Some of them are localized on the right end of chromosome III and in the Hi-C experiment were shown to interact with telomeric and subtelomeric regions of the four other chromosomes, allowing possible mono-allelic expression between the BdUMGF and BdOMGF genes. The euchromatic mark, H3K9ac, on the other hand, is detected in all other chromosomal regions and found to be enriched in the promoters of active genes (Fig. 5g). Our analysis further showed that H3K9ac intensity correlates with transcript abundance (Fig. 5g). Altogether, the epigenetic analysis provides new insights into *Babesia* gene expression and regulation. This study which revealed a unique epigenetic silencing associated with multigene families in *B. duncani*, suggests that antigenic variation might occur in this parasite.

Genome mining to identify antibabesial inhibitors

Mining of the annotated *B. duncani* proteome identified several potential drug targets (Fig. 6a and Supplementary Table 3). One of these is dihydrofolate reductase-thymidylate synthase (DHFR-TS)

https://doi.org/10.1038/s41564-023-01360-8

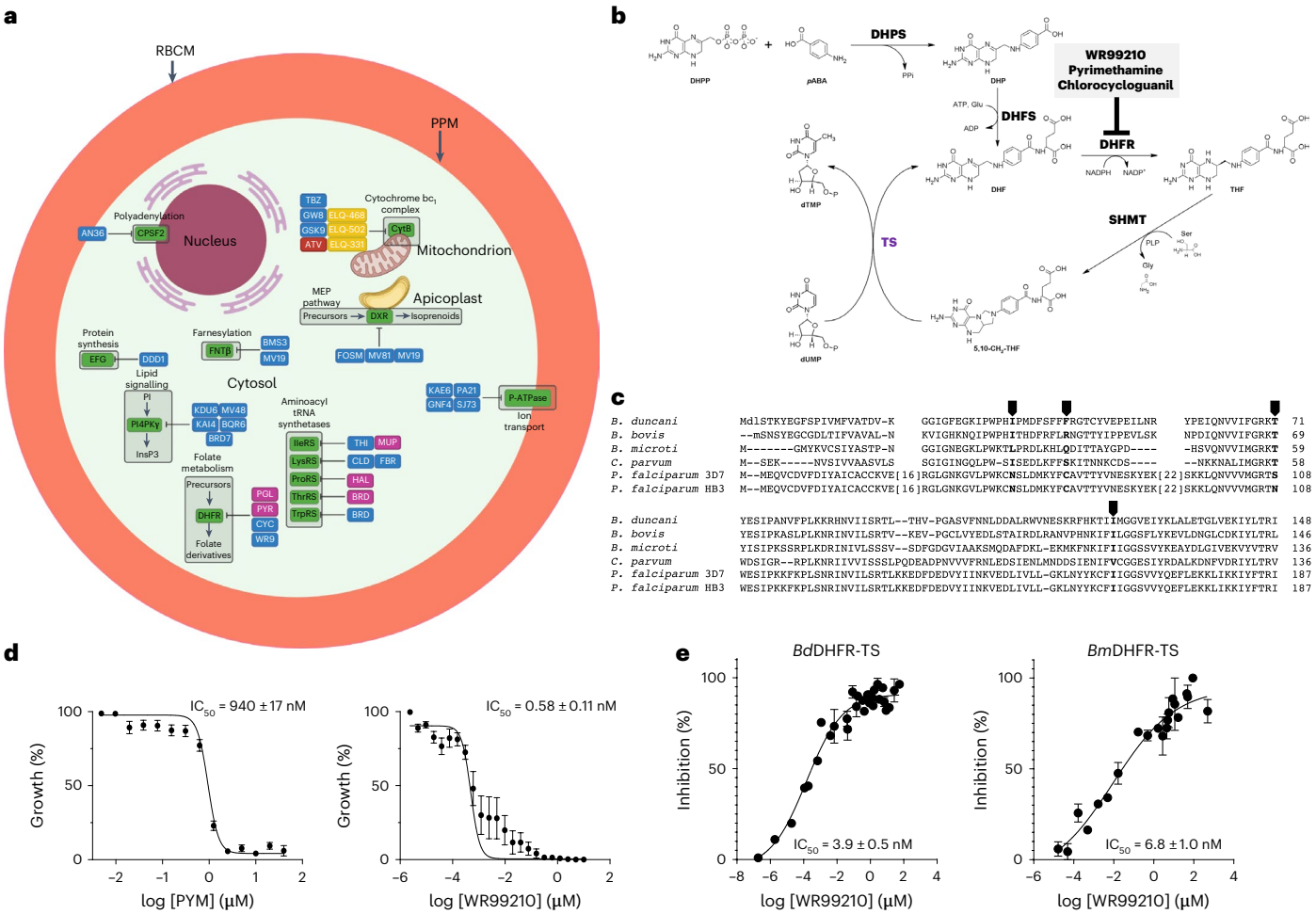


Fig. 6 | Potential targets for therapeutics development in *B. duncani*. **a**, Drugs are categorized as follows: Red, effective against *Babesia* parasites and used clinically; Yellow, effective against *Babesia* parasites but have not been evaluated clinically; Pink, clinically approved drugs for other diseases but have not been tested against *Babesia* parasites; and Blue, drugs under clinical evaluation for the treatment of other diseases but have not yet been tested against *Babesia* parasite inhibitors. Drug and protein abbreviations are as follows: Lys-TRNAS, Lysyl tRNA Synthase; Pro-TRNAS, Prolyl tRNA Synthase; Thre-TRNAS, Theronyl/alanyl tRNA Synthase; Tryp-TRNAS, Tryptophanyl tRNA Synthase; DHFR, Dihydrofolate reductase; P-ATPase, P-type ATPase; CytB, Cytochrome bc1 complex subunit 7 superfamily; Translation EFG/EF2, Elongation factor EFG (EFG); PI4PKγ, Phosphatidylinositol 4-kinase gamma; FNTB, Farnesyltransferase subunit beta; CPSF2, Polyadenylation specificity factor subunit 2; DXR, MEP Synthase; FBR, Febrifugine; HAL, Halofuginone; CLD, Cladosporin; MUP, Mupirocin; BRD, Borrelidin; DDD1, DDD107498; KDU6, KDU691; MV48, MMV048; KAI4, KAI4KAI407; BQR6, BQR695; BRD, BRD73842; ATV, Atovaquone; TBZ, Tetracyclic Benzothiazepine; GW8, GW844520; GSK9, GSK932121; ELQ, Endochin-like quinolones; KAE6, KAE609; GNF4, GNF-Pf-4492; PA21, PA21A092; SJ73, SJ733; AN36, AN3661; FOSM, Fosmidomycin; MV81, MMV008138; BMS3, BMS-388891; MV19, MMV019066; PYR, Pyrimethamine; PGL, Proguanil; MV27, MMV027634. **b**, Overview of folate metabolism and inhibitors. Synthesis of DHF precursor by DHP(F)S and reduction of DHF to tetrahydrofolate (THF) catalysed by the bifunctional enzyme DHFR by using NADPH as electron donor. The thymidylate synthase (TS) catalyses the reductive methylation of deoxyuridine monophosphate (dUMP) in deoxythymidine monophosphate (dTMP) using the tetrahydrofolate (5,10-Methylene THF) as cofactor via a hydroxymethyl transfer, mediated by serine hydroxy methyltransferase (SHMT). Common DHFR inhibitors are marked in grey box. **c**, Sequence comparison of DHFR-TS sequences between *B. duncani, B. microti, B. bovis* and *P. falciparum* pyrimethamine-sensitive (3D7) and resistant (HB3) parasites with highlighted resistance-associated residues in their primary sequences (black arrows). **d**, Dose-response curves with IC_{50} values of antifolates pyrimethamine and WR-99210 and their inhibitory effect on *B. duncani* intraerythrocytic parasite growth. Data presented as mean ± s.d. from three independent experiments performed in triplicates. **e**, Enzymatic activity inhibition of BdDHFR-TS and BmDHFR-TS by WR-99210. Data presented as mean ± s.d. from three independent experiments performed in triplicates.

(Fig. 6b). Residue T71 in *B. duncani* DHFR-TS (BdDHFR-TS) sequence is equivalent to residue 108 in DHFR-TS from *P. falciparum* (PfDHFR-TS) (Fig. 6c)[30–34]. In *P. falciparum*, the serine residue in this position is associated with susceptibility to pyrimethamine, whereas an asparagine residue is associated with resistance to this drug[35]. Consistent with our genome analysis, drug sensitivity and in vitro safety assays revealed that *B. duncani* is sensitive to pyrimethamine and WR-99210, with 50% inhibitory concentration (IC_{50}) values of 940 and 0.58 nM, respectively, with both compounds showing excellent in vitro therapeutic indices when tested against four human cell lines (Fig. 6d and Extended Data Table 7). The specificity of inhibition by these drugs was further examined biochemically using recombinant DHFR-TS enzymes from *B. duncani* and *B. microti* (BmDHFR-TS) (Fig. 6e). WR-99210 was found to inhibit BdDHFR-TS and BmDHFR-TS enzymes, with IC_{50} of 3.9 and 6.8 nM, respectively. Our biochemical assay results showed that both enzymes have similar kinetic properties (K_m and V_{max}), but different turnover numbers (k_{cat}) (Extended Data Table 8).

The availability of a complete genome sequence and annotation enables the understanding of the mechanism of action of antiparasitic drugs. As proof of concept, we used a single-step drug selection

approach[36], we selected *B. duncani* parasites that are refractory to atovaquone in vitro (Supplementary Fig. 6), and their drug susceptibility was compared to that of the parental isolate WA1 (IC_{50} of 0.072 µM). Whole genome sequencing of clone $BdATV^R$, which was found to be 15x more resistant to atovaquone compared with WA1 (Supplementary Fig. 6a), identified a single nucleotide polymorphism associated with the resistance profile in the form of a non-synonymous mutation L117F in the parasite *Bdcyt-B* gene, which encodes a component of the mitochondrial bc1 complex and is the main target of atovaquone and endochin-like quinolones (Supplementary Fig. 6b)[37,38]. A similar mutation was previously found to confer resistance to atovaquone in *P. falciparum*[39]. Altogether, these data highlight the advantages of genomic analyses in the search for potential new therapies for the treatment of human babesiosis and in the identification of the mechanism of action of antiparasitic drugs.

Discussion

In this study, we report the first genome sequence assembly, 3D structure, annotation, transcriptomic profile and metabolic reconstitution of the human intraerythrocytic pathogen *B. duncani*. This parasite species is unique within Piroplasmida and belongs to Clade II[21]. The high GC content of this parasite, its ability to propagate continuously in human red blood cells in vitro and the availability of a mouse model of virulence make it an ideal system to study the basic requirements for intraerythrocytic parasitism and to identify conserved and essential metabolic processes that could be targeted for the development of pan-antiparasitic drugs. Inhibitors of such processes could then be evaluated in mice for safety and efficacy to select the best candidates to advance towards clinical evaluation. Several new targets and potential drugs that could inhibit their function have been identified in this study and will stimulate future efforts to evaluate the activity of new inhibitors using the *B. duncani* in culture-in mouse model system (Fig. 6a). As a proof of feasibility, we have validated the DHFR-TS enzyme of *B. duncani* and *B. microti* as excellent targets for the development of antibabesial drugs (Fig. 6e and Extended Data Table 8). Our study further examined the utility of the *B. duncani* genome sequence in validating the mode of action and determining possible mechanisms of resistance of known and newly developed antibabesial drugs.

Our annotation of the *B. duncani* genome provides invaluable insights into the biology, pathogenesis and drug susceptibility of this important human pathogen. Notable among these is the presence of classes of multigene families that could be important in host–pathogen interactions and virulence of the parasite in mice and humans. The BdOMGF families encompass genes with orthologues in other hematozoan parasites and thus might have evolved to support survival of the parasite in mammalian red blood cells and escape from host immune attacks as was shown for the *var* and variant erythrocyte surface antigen (*vesa*) gene families of *P. falciparum* and *B. bovis*[26,40].

Evolutionary analysis of the *B. duncani* genome (Fig. 2) revealed its unique genetic relationships with other haemoparasites. This analysis, which confirmed the membership of *B. duncani* in clade II of Piroplasmida[21], is also consistent with a recent phylogenomic analysis of the parasite[41]. The evolution of Piroplasm parasites could be shaped by several factors such as geography, tropism, as well as other genomic features such as genome size and chromosome number. However, our analysis using orthologous proteins with functional annotation and shared among all analysed species suggests that *B. duncani* has a distant relationship with the rest of *Babesia* species, as well as unique metabolic, virulence and pathogenic features that can be studied through the genomic information presented in this work. A classification of selected organisms and OrthoMCL groups based on the presence or absence of these orthologues is consistent with the proposed evolution of *Babesia* species and the phylogenetic position of *B. duncani*. Binary distance between OrthoMCL groups places *B. duncani* at the root of the clade of *Babesia* (Supplementary Fig. 7).

Another unique property of *B. duncani* is the presence of a large array of tandemly duplicated functions. We found 139 genes with orthology groups that contain unique genes in other piroplasmida but have two or up to four genes in *B. duncani*. These genes vary in size and sequence, suggesting that duplication events occurred early during the evolution of the parasite and continue to occur in the genome. No specific enrichment of specific functions was observed. Two DNA polymerase subunits and the RPB11 subunit of polymerase II were duplicated, as well as transcription factor IIA (TFIIA) basal factor and other RNA binding and/or modifying proteins. Duplicated functions were also associated with cellular trafficking, ubiquitinylation and stress response. No enzymes from the central metabolic pathways or subunits of the ribosome were part of this specific group of duplicated genes. We also identified 73 unique gene families in *B. duncani* (BdUMGFs) that have no homologues in other Apicomplexa based on BLAST and OrthMCL analyses (Fig. 3a). RNA-seq data indicate that some of the BdUMGFs, such as BdUMGF3, are not expressed during the asexual life cycle of the parasite and are potentially important for parasite development during other stages of the life cycle. Interestingly, transcriptional repression was observed in the case of the BdUMGF3 gene family, which are distant from the telomeres and centromeres. DNA FISH and Hi-C analyses confirmed the distribution of the BdUMGF3 gene family on Chr III at a position distinct from that of BdUMGF1 on Chr I and IV (Fig. 4).

We further performed ChIP-seq analysis in *B. duncani* to investigate genomic histone marks, which are often associated with regulated cell biological events such as gene expression, cell differentiation and antigenic variation (Fig. 5e–g). The dynamic distribution of acetyl or methyl marks on the histone tails are essential to the maintainence of the heterochromatin and euchromatin structure and are known to regulate gene expression needed for parasite survival. In the *P. falciparum* asexual blood stages, a transcriptionally silent heterochromatin is defined by the presence of H3K9me3 and is restricted to the telomeric and subtelomeric regions, as well as some internal chromosome clusters associated with multigene protein families involved in antigenic variation or genes involved in sexual differentiation[42]. Conversely, euchromatin marks are characterized by a different set of histone marks including H3K9ac localized in the promoter regions of active genes. The binding of transcription factors with their DNA binding domains is heavily dependent on the state of the chromatin structure surrounding the promoter regions. Inhibition of histone writers and erasers such as histone acetyltransferases, histone methyltransferase, histone deacetylases and histone demethylases has been shown to lead to inhibition of *Plasmodium* growth in vitro and in vivo[43–45]. Our ChIP-seq analysis in *B. duncani* revealed trends similar to those observed in *T. gondii* and *P. falciparum*[28,46]. In particular, the epigenetic profile of multigene families in *B. duncani* is reminiscent of that of *var* genes in *P. falciparum*, suggesting that a similar mechanism of antigenic variation might occur in this parasite and could account for its high virulence.

In summary, our analysis of the *B. duncani* nuclear genome and transcriptome has shed light into the biology, evolution and drug susceptibility of this important human pathogen. It is anticipated that this knowledge will help advance new strategies to develop reliable, sensitive and specific diagnostic tools as well as therapeutic strategies for better management of human babesiosis. The close relationship between *B. duncani* and other haemoparasites including malaria parasites will help future efforts to better understand the evolution of virulence in Apicomplexa.

Methods

Parasite strain

The *B. duncani* WA1 used in this study was obtained from BEI Resources (NR-12311) and propagated either in vitro or in mice and hamsters as previously described[12].

In vitro parasite culture of *B. duncani*

The in vitro *B. duncani* parasites were cultured in human red blood cells (hRBCs) as reported earlier[12]. Briefly, parasites were cultured in a complete HL-1 medium (base medium of HL-1 (Lonza, 344017) supplemented with 20% heat-inactivated FBS, 2% 50X HT Media Supplement Hybrid-MaxTM (Sigma, H0137), 1% 200 mM L-glutamine (Gibco, 25030-081), 1% 100X penicillin/streptomycin (Gibco, 15240-062) and 1% 10 mg ml^{-1} gentamicin (Gibco, 15710-072)) in 5% haematocrit A^+ RBCs. The cultures were maintained at 37 °C under a 2% O_2/5% CO_2/93% N_2 atmosphere in a humidified chamber. The culture medium was changed daily and parasitemia was monitored by light microscope examination of Giemsa-stained thin-blood smears.

Chemicals

Unless otherwise stated, chemicals were purchased from commercial suppliers and used as received. WR-99210 was ≥95% pure as determined by reversed-phase high-performance liquid chromatography (HPLC) and was purchased from Jacobus.

DNA prep for PacBio sequencing

Genomic DNA was isolated from 100 ml in vitro culture of *B. duncani* (15% parasitemia and 5% haematocrit) using DNeasy Blood and Tissue kit (Qiagen, 69506), and quality control and concentration determination were performed using nanodrop and qubit. DNA integrity was evaluated using Blue Pippin pulse gel and the DNA was then used for library preparation using Pacific Biosciences SMRTbell Express template prep kit 2.0 (100-938-900) according to manufacturer instructions. The Pacific Biosciences Smart Link software was used to determine loading concentration and proper stoichiometric measurements. The gDNA library was then annealed to the Pacific Biosciences V5 primer for 1 h at 20 °C. The annealed library was then bound to polymerase using Pacific Biosciences Polymerase 2.2 for 1–4 h at 30 °C and loaded onto the Sequel II Instrument as an adaptive sequencing run. At least one smart cell was sequenced for each gDNA library with a move time of 30 h and a pre-extension of 2 h. After the DNA library sequencing was completed, the loading metrics were evaluated by mean read length, polymerase read length, data yield and P1 values (a productive zero-mode waveguide (ZMW) with a high quality (HQ) sequencing region detected within the read) to ensure the sample ran as expected and data had met Yale's gold standards (polymerase read length between 50–60 kb, data yield (HiFi) of around 2–4 million reads of a total of 10–20 Gb and P1 between 60–70%).

DNA preparation for Bionano optical mapping

A total of 60 ml in vitro culture of *B. duncani* at 15% parasitemia and 5% haematocrit was collected and centrifuged at 500 × *g* to obtain a packed pellet. For analysis, pellet was used to isolate ultra-high molecular weight gDNA for use in genomic optical mapping by Histogenetics using the Bionano Prep Blood and Cell Culture DNA Isolation kit (Bionano Genomics, 80004). Following this, DNA was quantified using Qubit dsDNA BR Assay kit. A total of 0.75 g of high molecular weight DNA was then labelled using the Bionano Prep direct label and stain method (Bionano Genomics, 80005) and loaded onto a flow cell to run on the Saphyr optical mapping system (Bionano Genomics). Approximately 1,177 Gb of data were generated per run. Raw optical mapping of molecules in the form of BNX files were run through a preliminary bioinformatics pipeline that filtered out molecules less than 150 kb in size and less than 9 motifs per molecule to generate a de novo assembly of the genome maps.

Genome sequencing and assembly

An in vitro culture of *B. duncani* (100 ml) in hRBCs was propagated to a parasitemia of 10% at 5% haematocrit. DNA was isolated from infected hRBCs and sequenced at the Yale Center for Genome Analysis using Illumina HiSeq 2500 and PacBio HiFi (CCS) sequencing. A total of 177,189 PacBio HiFi reads were obtained. The average HiFi read length was 8,814 bp, and the longest read was 29,067 bp. The HiFi reads totalled 1.56 B bases, which translated to an expected ~156x coverage of the *B. duncani* genome (assuming a ~13 Mb genome). HiFi reads were mapped using Minimap2[47] to the *B. duncani* mitochondrion and apicoplast. Only 0.41% of the reads mapped to these organelles, which were then discarded to enrich the data for nuclear DNA. HiFi reads were also mapped to the human genome to determine possible host contaminations, but only 1.48% of them were flagged. Cleaned reads were assembled using HiCANU v2.2[13] using default parameters. HiFi reads were also tested on HiFiASM v0.16[48] and Wengan[49], but they generated slightly inferior assemblies (based on assembly statistics, comparison with the optical map and transcript isoform annotations). The Bionano Hybrid Scaffolding pipeline v1.7 was used to compare the draft assembly with the Bionano optical map to detect possible mis-joins and create scaffolds. The scaffolded assembly was polished using two rounds of PolyPolish[14]. At every step of the assembly (draft, scaffolds, polished), a series of quality control steps were carried out to ensure that no imperfections were introduced. The quality control steps included (1) mapping of all the Illumina WGS and HiFi reads to the assembly, (2) carrying out a BUSCO[15] completeness analysis (at the genome level) and (3) determining the number of detected gene loci from the mapping of RNA-seq IsoSeq reads.

DNA preparation for Hi-C

An in vitro culture of *B. duncani* WA1 (100 ml) in A^+ hRBCs was collected at parasitemia of 15% and 5% haematocrit. The culture was centrifuged at 500 × *g*, and the pellet was crosslinked in 1.25% formaldehyde for 25 min at 37 °C. Crosslinking was quenched in a final concentration of 150 mM glycine for 15 min at 37 °C, followed by a 15 min incubation at 4 °C. Parasite pellets were then resuspended in lysis buffer (10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 2 mM 4-(2-aminoethyl) benzenesulfonyl fluoride HCl, 0.25% Igepal CA-360 (v/v) and EDTA-free protease inhibitor cocktail (Roche)), and incubated on ice for 30 min. Nuclei were isolated after homogenization by 15 gauge needle passages. In situ Hi-C protocol was conducted as previously described[50]. Briefly, nuclei were permeabilized using 0.5% SDS; the DNA was digested with 100 units of MboI (NEB), and the ends of restriction fragments were filled using biotinylated nucleotides and ligated using T4 DNA ligase (NEB). After reversal of crosslinks, the ligated DNA was purified and sheared to a length of ~300–500 bp using the Covaris ultrasonicator S220 (settings: 10% duty factor, 200 cycles per burst and a peak incident power of 140). Ligated fragments were pulled down using streptavidin beads (Invitrogen) and prepared for Illumina sequencing by subsequent end-repair addition of A-overhangs and adapter ligation. Libraries were amplified for a total of 12 PCR cycles (45 s at 98 °C, 12 cycles of 15 s at 98 °C, 30 s at 55 °C, 30 s at 62 °C and a final extension of 5 min at 62 °C) and sequenced with the NOVASeq platform (Illumina), generating 100 bp paired-end sequence reads at the University of California San Diego core facility.

Hi-C data processing

The Illumina sequencing of the four *B. duncani* clones (WA1, A6, B7 and B1) yielded over one billion Hi-C paired-end reads in total. Hi-C reads were processed using the command-line version of the HiCExplorer pipeline. HiCExplorer is a comprehensive and versatile pipeline that processes Hi-C reads and generates normalized chromatin conformation contact maps. The pipeline started by mapping all Hi-C single reads to the *B. duncani* assembly using 'BWA mem -A1 -B4 -E50 -L0'[51]. The percentage of reads mapped ranged between 83% and 96%. After sorting and merging the reads from the four data sets, the command 'hicBuildMatrix' was used with default parameters and a bin size of 10 Kb. A diagnostic plot created using 'hicCorrectMatrix diagnostic_plot' indicated that a threshold of −4.5 would be appropriate to remove GC and open chromatin biases. The correction step, carried out by the command 'hicCorrectMatrix correct', also normalized the

number of restriction sites per bin. The contact maps were plotted using 'hicPlotMatrix' as part of the same pipeline.

3D modelling

The 3D model of the *B. duncani* genome was generated and visualized using PASTIS[52] and ChimeraX[53]. Briefly, interaction data were first manually fitted to a three-column interaction count matrix where the first two columns indicate the two separate bins within the genome and the third is the number of interactions between those two bins. The 3D coordinate matrices were then generated using PASTIS and converted to .PDB format, with each line being one of the coordinate outputs by PASTIS. The data were visualized using ChimeraX[53]. To investigate the spatial conformation of the *B. duncani* genome, a 3D model of the three chromosomes was first built using PASTIS[52] and then improved using the Bezier curve smoothing method. PASTIS infers a consensus 3D structure from the genome-wide Hi-C contact frequency matrix using the following probability model. The programme models the observed chromatin contact frequencies as independent Poisson random variables and infers the model parameters as well as the 3D coordinates of the genome structure via the maximum-likelihood estimation approach. Specifically, PASTIS assumes the Poisson parameter λ_{ij} for chromatin interaction frequency c_{ij} between loci i and j as a decreasing function of d_{ij},

$$\lambda_{ij} = \beta\, d_{ij}\,(X)^{\alpha} \tag{1}$$

with parameters $\beta > 0$ and $\alpha < 0$. Here, d_{ij} is the Euclidean distance between loci i and j in structure X[52]. Therefore, the likelihood of observing c_{ij}s can be formulated as:

$$L\,(X,\alpha,\beta) = \prod_{-}(i,j)\mathrm{Poisson}\,(\lambda_{ij}) = \prod_{-}(i,j)\frac{\left(\beta d_{ij}^{\alpha}\right)^{c_{ij}}}{c_{ij}!}e^{\left(-\beta d_{ij}^{\alpha}\right)} \tag{2}$$

By maximizing the log-likelihood function, PASTIS optimizes the 3D structure X as well as the distance decay parameters α and β. Since the 3D model predicted by PASTIS contained a set of discrete points, cubic Bezier curve smoothing was applied for the purpose of visualization. Cubic Bezier curves guarantee that the interpolated curve passes through all points in the original structure and the first-order geometric continuity at all points is maintained.

PacBio IsoSeq processing

PacBio IsoSeq reads were mapped to the *B. duncani* genome using Minimap2[47] with options 'splice:hq -uf –secondary=no -C5'. The resulting SAM files were sorted using 'sort -k 3,3 -k 4,4n', then fed into the PacBio cDNA_Cupcake ToFU pipeline (https://github.com/Magdoll/cDNA_Cupcake) using the Python code 'collapse_isoforms_by_sam.py' to collapse redundant isoforms (more details about this pipeline can be found at https://github.com/Magdoll/cDNA_Cupcake/wiki). The resulting isoform sequences were used in the gene finding pipeline below.

Gene prediction and annotation

The genome was first soft masked with RepeatMasker (http://www.repeatmasker.org), then processed using the FunAnnotate (v1.8.9) gene annotation pipeline (https://github.com/nextgenusfs/funannotate). As transcript evidence, we provided the IsoSeq-based isoforms computed via the cDNA_cupcake pipeline (see PacBio IsoSeq processing). As protein sources, we provided the annotated protein sets of *B. bigemina*, *B. bovis*, *B. microti*, *P. falciparum*, *T. gondii*, *T. orientalis*, *T. parva* as well as all the UniProt/SwissProt protein models. FunAnnotate was run with default parameters and weights (augustus:2 hiq:4 transcripts:4 proteins:4). Functional annotations were obtained via InterProScan (v5.55-88) with default parameters[54]. The output of InterProScan was parsed by custom scripts for the downstream analyses.

Gene orthology-based classification analysis

A maximum-likelihood gene orthology-based classification analysis was generated using 190 single-copy orthologous proteins from 19 species. Orthology was generated with Proteinortho (v6.0.33)[55] Proteinortho: Detection of (Co-)Orthologs in large-scale analysis[55] with the following parameters: -cpus=64 -ram=462144 -p=blastp+-singles. Proteins were concatenated for each species and a multiple alignment was generated using MAFFTv7.505 with the –maxiterate 1,000 parameter[56]. The resulting multiple alignment was used as input to generate a tree with Raxml (v8.2.12) with the following parameters: -p 070378 -T 48 -m PROTCATIJTTF -# 1000 -f a -x 070378. Finally, the tree was visualized and generated using iTOL (v6[57]). An upset plot was created using the UpSetR (v1.4.0) R package and a csv matrix created from the ProteinOrtho orthologous protein analysis[55]. Basic instruction to create the plot can be found at https://cran.r-project.org/web/packages/UpSetR/vignettes/basic.usage.html

Synteny and gene localization plots

Synteny plots were obtained using mummer2circos (https://github.com/metagenlab/mummer2circos) with the promer algorithm, and Circos[58]. Gene localization plots were produced using custom Python scripts and the Biopython Bio.graphics library.

Orthology detection and database searches

B. duncani genes were assigned to OrthoMCL (https://OrthoMCL.org) groups using the orthology assignment tool available through the VEuPathDB (https://VEuPathDB.org) Galaxy workspace[59,60]. Translated *B. duncani* proteins (4,170) in FASTA format were assigned to groups on the basis of the OG6r9 blast database using default settings. Output files generated by the OrthoMCL pipeline included a mapping file between *B. duncani* gene IDs and OrthoMCL v.6 group IDs (this file was used to query the OrthoMCL database to determine degrees of evolutionary conservation) and a file of leftover proteins that did not map to any OrthoMCL groups but did cluster with at least one other *B. duncani* protein (these were considered species-specific gene duplications or families). To identify genes that do not have any orthologues or paralogues, the original input FASTA file was parsed for *B. duncani* IDs that are not present in any of the OrthoMCL mapping output files. VEuPathDB resources including PlasmoDB.org, ToxoDB.org, CryptoDB.org and PiroplasmaDB.org were used to retrieve genome size, gene content and chromosome numbers.

A matrix containing the number of genes per OrthoMCL groups was made from the annotation of *B. duncani* and species present in PiroplasmaDB. Annotations were compared by selecting OrthoMCL groups presenting annotation in at least 4 species. *B. microti* drastically reduced the number of OrthoMCL IDs and was removed from the analysis. OrthoMCL IDs were compared using Euclidian distance and hierarchical classification was performed using the Ward method. Species annotations were compared using presence/absence of the OrthoMCL ID. Jaccard distance and Ward methods were used to make the tree. All analyses were performed in R. The ComplexHeatmap function was used to generate the heat map.

Probe generation for DNA FISH

The following PCR primers were designed for the amplification of ~1 kb region from the genomic loci of interest: Probe 1_For (Chr I; UMGF1): ACTCAGAATGTTAATGGAACGTC, Probe 1_Rev (Chr I; UMGF1): AATATCACTTTGAATCTTGCTAAC; Probe 2_For (Chr IV, UMGF1): TCAATCGTTCAATTCATCATCG, Probe 2_Rev (Chr IV, UMGF1): TCAATAACTTGCTGCAAATCAC; Probe 3_For (Chr III, UMGF3): GAAAC TAGCGTAATCCTGTG and Probe 3_Rev (Chr III, UMGF3): AATAATCA GAAATGGTGGGTTCG. Fluorescein-labelled Probes 1 and 2 were generated by PCR using *B. duncani* WA1 genomic DNA as template and HighFidelity Fluorescein PCR labelling kit (APP-101-FAMX-L, Jena Biosciences). Biotin-labelled Probe 3 was generated by PCR using

B. duncani WA1 genomic DNA as template and Biotin-16-dUTP kit (11093070910, Roche).

DNA FISH

B. duncani WA1 parasites were cultured in vitro in 20 ml volume and maintained until 20% parasitemia was reached. The culture was treated with 0.015% saponin for 20 min in cold PBS, followed by centrifugation at 4,200 × *g* for 10 min. Following this, the parasite pellet was washed 7 times with 1X PBS. After the final wash, the cells were resuspended in 4% paraformaldehyde in PBS at room temperature and incubated on ice for 15 min. The parasite pellet was washed twice in 10 ml ice-cold PBS and resuspended in 0.25 ml ice-cold PBS. A monolayer of parasites (roughly 1×10^7 parasites) was deposited on a 9 × 9 mm frame-seal slide chamber (AB-0576, Thermo Fisher) on a standard microscopy slide and allowed to dry for 1–2 h at r.t. Following this, the slides were washed with 50 µl PBS for 5 min at r.t. The parasites were then permeabilized in 50 µl 0.1% Triton X-100 (A16046.AP, Thermo Fisher) in PBS for 5 min at r.t. and then washed with 50 µl PBS for 5 min at r.t. The slides were equilibrated with 50 µl of hybridization solution (50% formamide (F5903, Teknova), 10% dextran sulfate (D6001, Sigma), 2x sodium saline phosphate EDTA (15591-043, Invitrogen), 250 µg ml^{-1} single-stranded DNA from salmon testes (D7656, Sigma)) for 30 min at 37 °C. In parallel, the PCR amplified DNA FISH probes (fluorescein-labelled probe 1, fluorescein-labelled probe 3 and biotin-labelled probe 2) were diluted to 1:20 in the hybridization solution, denatured at 95 °C for 5 min and immediately placed on ice for 5 min. The probes were then applied onto the slides; the slides were sealed with plastic frames (AB-0576, Thermo Fisher) and incubated at 80 °C for 30 min and then at 37 °C overnight. Next morning, the coverslip was removed, and the hybridization solution was discarded. The frame was also stripped out and the slides were washed in 15 ml 50% formamide/2xSSC (F5903, Teknova and 15557044, Invitrogen) buffer for 30 min at 37 °C. The slides were then washed sequentially in 15 ml each of 1xSSC, 2xSSC and 4xSSC at 50 °C for 15 min, and then equilibrated in M solution (100 mM maleic acid (M0375, Sigma-Aldrich), 150 mM NaCl (AB01915-01000, AmericanBio) and 1% bovine serum albumin (A9647, Sigma-Aldrich)) for 5 min at r.t. in a humid chamber protected from light. M solution was replaced with a new M solution containing avidin-rhodamine (1:5,000; A-6378, Thermo Fisher) and slides were incubated for 30 min at r.t. in a humid chamber protected from light. Following this step, the slides were washed 3 times in 30 ml TNT solution (100 mM Tris-HCl, pH 7.4 (AB14044-01000, AmericanBio), 150 mM NaCl (AB13198-01000, AmericanBio) and 0.5% Tween 20 (9005-64-5, Sigma-Aldrich)) at 10 min each at r.t. The slides were then mounted with coverslips using Vectashield mounting solution containing DAPI (H-1200-10, Vector Laboratories), and coverslips were sealed using nail polish. The slides were observed using a Nikon ECLIPSE TE2000-E microscope. A ×100 oil immersion objective was used for image acquisition. Excitation at 465–495 nm was used to detect fluorescein. Excitation at 510–560 nm was used to detect rhodamine, and excitation at 340–380 nm was used to detect DAPI positive cells. The images were acquired using Meta Vue with image size of 1,392 × 1,040 pixels and subsequently analysed using ImageJ.

RNA-seq processing for gene-expression analysis

RNA-seq FASTQ files were assessed for quality using FastQC (v0.11.8). Adapter sequences as well as the first 11 bp of each read were trimmed using Trimmomatic (v0.39). Tails of reads were trimmed using Sickle with a Phred base quality threshold of 25, and reads shorter than 18 bp were removed. Reads were then aligned to the *B. duncani* genome assembly using HISAT2 (v2.2.1). Only properly paired reads with a mapping quality score of 40 or higher were retained, with filtering done using Samtools (v1.11). StringTie (v2.2.1) was run with the -e parameter to estimate the abundance of each gene in TPM (transcripts per million).

ChIP-seq data analysis

Quality of the reads was analysed by FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Sequence adaptors were removed using Trimmomatic v0.39 (http://www.usadellab.org/cms/?page=trimmomatic). Bases with Phred quality scores below 20 were trimmed using Sickle v1.330 (https://github.com/najoshi/sickle). Reads were mapped against the *B. duncani* assembly using Bowtie2 (v2.4.4) using default parameters, and PCR duplicates were removed by PicardTools MarkDuplicates v2.18.0 (Broad Institute). Only properly paired reads were retained using Samtools v1.11 (http://samtools.sourceforge.net/). To obtain the read coverage per nucleotide and generate genome browser tracks, we used BedTools (v2.27.1) and custom scripts to normalize the counts by millions of mapped reads. Chromosome tracks were viewed using IGV (Broad Institute). To compare H3K9me3 levels between MGF genes and other genes, we used bedtools multicov to calculate H3K9me3 (and IgG control) read counts within each gene body. These counts were then normalized to millions of mapped reads for each library and gene length in kb. Background signal from the IgG control was subtracted from H3K9me3 counts, with negative values set to 0. Similarly, H3K9ac read counts were also determined by bedtools multicov, but 300 bp on the 5′ side were included in addition to the gene body since histone H3K9 acetylation often occurs in promoter regions. Pearson correlations between ChIP samples were determined by taking genome-wide per-nucleotide read counts for each sample, normalizing by millions of mapped reads using R. To compare histone modifications with gene expression profiles, we generated a heat map using H3K9me3 and H3K9ac read counts, as well as the RNA-seq TPM levels for each gene. Read counts for H3K9me3 and H3K9ac were normalized to millions of mapped reads and gene length, while TPM was determined by Stringtie (see RNA-seq methods). Log scaling of all counts was used for the heat map, and genes were sorted from high to low TPM.

A separate Excel spreadsheet of ChIP-seq data is provided in Supplementary Information (Supplementary Data file 2).

Evaluation of drug efficacy in vitro

Drug efficacy against *B. duncani* in vitro was determined using the SYBR Green assay as previously described[7]. Briefly, a culture of *B. duncani*-infected human erythrocytes at 0.1% parasitemia and 5% haematocrit was maintained for 60 h in vitro in a 96-well plate in the absence or presence of varying concentrations of the drug of interest at concentrations ranging between 0.6 nM and 10 µM. An equal volume of lysis buffer (0.008% saponin, 0.08% Triton X-100, 20 mM Tris-HCl (pH 7.5) and 5 mM EDTA) containing SYBR Green-I (0.01%) was added to each well and the plate incubated at 37 °C for 1 h in the dark. Fluorescence counts in each well were measured at 480 nm (excitation) and 540 nm (emission) using a BioTek Synergy Mx microplate reader and background fluorescence from control wells containing uninfected RBCs was subtracted. A dose-response curve representing the fluorescence counts against the logarithm of the drug concentration was fitted by nonlinear regression and used to determine the IC_{50} of each drug. Analysis was conducted using GraphPad Prism 9.2.1. For each drug, two independent experiments, each with biological triplicates, were conducted. Data are shown as mean ± s.d.

Assessment of drug cytotoxicity on human cell lines

HeLa, HepG2, HEK and HCT116 cell lines were obtained from the American Type Culture Collection and maintained in Dulbecco's modified Eagle's medium (Invitrogen 11995-065) containing 25 mM glucose, 1 mM sodium pyruvate and supplemented with 5 mM HEPES, 10% FBS and penicillin-streptomycin (50 U ml^{-1} penicillin, 50 µg ml^{-1} streptomycin). An estimated 20,000 cells per well were seeded in a 96-well tissue culture plate and allowed to adhere. After 24 h, cells were treated with either vehicle alone (0.1% dimethylsulfoxide) or vehicle containing different concentrations of the target drugs using a 2-fold serial dilution

starting at 10 mM. Dimethylsulfoxide at 10% was used as a positive control. Cells were incubated at 37 °C for 72 h, after which each well was incubated with 0.5 mg ml^{-1} of MTT reagent (Cayman Chemical, 10009591) for 4 h in the dark at 37 °C. Following addition of 100 µl dimethylsulfoxide to each well, absorbance was measured at optical density $(OD)_{590\,nm}$ using Spectra max plate reader. From the obtained OD values, percent cell viability was calculated by normalizing to the mean of 10% dimethylsulfoxide wells (set as 100% toxicity) and the mean of the vehicle control wells (set as 0% toxicity). Dose-response curves were plotted using GraphPad Prism v9.1.2

Steady-state kinetics of DHFR-TS activity

DHFR kinetic experiments were performed by incubating purified enzyme (0.1 µM) with a saturating concentration of DHFR ligand (NADPH or dihydrofolate (DHF) at 300 µM) and measuring the reaction rate at varying concentrations of the complementary ligand (DHF or NADPH, respectively). The experiments were performed in a 96-well plate in a reaction buffer consisting of 25 mM Tris pH 8, 20 mM NaCl, 50 mM L-arginine, 0.5% glycerol, 1 mM dithiothreitol and 1 mM EDTA. Enzyme activity was determined by measuring the decrease in absorbance at 340 nm as NADPH is converted to NADP+. The plates were incubated at 37 °C, and OD_{340} measurements were taken using a BioTek SynergyMx microplate reader for 1 min. Rate constants for steady-state kinetic experiments were estimated by fitting the data to a Michaelis–Menten hyperbolic curve ($(\upsilon = V_{max}[S]/(K_m + [S]))$, where υ is the reaction rate, $[S]$ the concentration of the substrate, and K_m is the Michaelis constant) using the curve-fitting programme in GraphPad Prism.

The IC_{50} of DHFR enzymatic activity

The inhibition of selected compounds on DHFR activity was measured by incubating purified enzyme (0.1 µM) with rising concentrations of the drugs. The reaction buffer contained 300 µM DHF and 300 µM NADPH, and the OD_{340} reduction rate (OD_{340} min^{-1}) was documented for further calculation. The % inhibition was calculated and normalized to dimethylsulfoxide (100% activity) and no enzyme (0% activity) wells accordingly: $1 - (((OD_{340}\ min^{-1})_{sample} - (OD_{340}\ min^{-1})_{neg.\ control})/((OD_{340}\ min^{-1})_{pos.\ control} - (OD_{340}\ min^{-1})_{neg.\ control}))$. The IC_{50} was determined from a sigmoidal dose-response curve using GraphPad Prism v9.1.2.

Additional methods used in this study are detailed in Supplementary Information.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All the datasets generated for the current study are available in the NCBI/SRA repository, as Bioproject PRJNA821606, as follows: PacBio HiFi reads (SRA accession number SRR18778747), *B. duncani* genome assembly (NCBI Genome submission: SUB11253661), RNA-seq (SRA accession number SRR18907291), RNA Isoseq (SRA accession number SRR18902718), Hi-C reads (SRA accession number SRR19325692). Source data are provided with this paper.

References

1. Paules, C. I., Marston, H. D., Bloom, M. E. & Fauci, A. S. Tickborne diseases - confronting a growing threat. *N. Engl. J. Med.* **379**, 701–703 (2018).
2. Allred, D. R. & Ben Mamoun, C. Babesiosis. *eLS* https://doi.org/10.1002/9780470015902.a0001945.pub2 (2018).
3. Renard, I. & Ben Mamoun, C. Treatment of human babesiosis: then and now. *Pathogens* https://doi.org/10.3390/pathogens10091120 (2021).
4. Swei, A. et al. Evidence for transmission of the zoonotic apicomplexan parasite *Babesia duncani* by the tick *Dermacentor albipictus*. *Int. J. Parasitol.* https://doi.org/10.1016/j.ijpara.2018.07.002 (2018).
5. Quick, R. E. et al. Babesiosis in Washington State: a new species of *Babesia*? *Ann. Intern. Med.* **119**, 284–290 (1993).
6. Baneth, G. et al. A new piroplasmid species infecting dogs: morphological and molecular characterization and pathogeny of *Babesia negevi* n. sp. *Parasit. Vectors* **13**, 130 (2020).
7. Abraham, A. et al. Establishment of a continuous in vitro culture of *Babesia duncani* in human erythrocytes reveals unusually high tolerance to recommended therapies. *J. Biol. Chem.* **293**, 19974–19981 (2018).
8. Herwaldt, B. L. et al. Transfusion-transmitted babesiosis in Washington State: first reported case caused by a WA1-type parasite. *J. Infect. Dis.* **175**, 1259–1262 (1997).
9. Herwaldt, B. et al. A fatal case of babesiosis in Missouri: identification of another piroplasm that infects humans. *Ann. Intern. Med.* **124**, 643–650 (1996).
10. Bloch, E. M. et al. The third described case of transfusion-transmitted *Babesia duncani*. *Transfusion* **52**, 1517–1522 (2012).
11. Kjemtrup, A. M. et al. Investigation of transfusion transmission of a WA1-type babesial parasite to a premature infant in California. *Transfusion* **42**, 1482–1487 (2002).
12. Pal, A. C. et al. *Babesia duncani* as a model organism to study the development, virulence and drug susceptibility of intraerythrocytic parasites in vitro and in vivo. *J. Infect. Dis.* https://doi.org/10.1093/infdis/jiac181 (2022).
13. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
14. Wick, R. R. & Holt, K. E. Polypolish: short-read polishing of long-read bacterial genome assemblies. *PLoS Comput. Biol.* **18**, e1009802 (2022).
15. Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
16. Bunnik, E. M. et al. Comparative 3D genome organization in apicomplexan parasites. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1810815116 (2019).
17. Brayton, K. A. et al. Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. *PLoS Pathog.* **3**, 1401–1413 (2007).
18. Bunnik, E. M. et al. Changes in genome organization of parasite-specific gene families during the *Plasmodium* transmission stages. *Nat. Commun.* **9**, 1910 (2018).
19. Cornillot, E. et al. Sequencing of the smallest Apicomplexan genome from the human pathogen *Babesia microti*. *Nucleic Acids Res.* **40**, 9102–9114 (2012).
20. Gardner, M. J. et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
21. Schnittger, L. et al. The Piroplasmida *Babesia*, *Cytauxzoon*, and *Theileria* in farm and companion animals: species compilation, molecular phylogeny, and evolutionary insights. *Parasitol. Res.* **121**, 1207–1245 (2022).
22. Goethert, H. K. What *Babesia microti* is now. *Pathogens* https://doi.org/10.3390/pathogens10091168 (2021).
23. Rubio, J. P., Thompson, J. K. & Cowman, A. F. The var genes of *Plasmodium falciparum* are located in the subtelomeric region of most chromosomes. *EMBO J.* **15**, 4069–4077 (1996).
24. Lapp, S. A., Korir, C. C. & Galinski, M. R. Redefining the expressed prototype SICAvar gene involved in *Plasmodium knowlesi* antigenic variation. *Malar. J.* **8**, 181 (2009).

25. Kraemer, S. M. & Smith, J. D. A family affair: var genes, PfEMP1 binding, and malaria disease. *Curr. Opin. Microbiol.* **9**, 374–380 (2006).
26. Ralph, S. A., Scheidig-Benatar, C. & Scherf, A. Antigenic variation in *Plasmodium falciparum* is associated with movement of var loci between subnuclear locations. *Proc. Natl Acad. Sci. USA* **102**, 5414–5419 (2005).
27. Gomez-Diaz, E. et al. Epigenetic regulation of *Plasmodium falciparum* clonally variant gene expression during development in *Anopheles gambiae*. *Sci. Rep.* **7**, 40655 (2017).
28. Witmer, K., Fraschka, S. A., Vlachou, D., Bartfai, R. & Christophides, G. K. An epigenetic map of malaria parasite development from host to vector. *Sci. Rep.* **10**, 6354 (2020).
29. Brooks, C. F. et al. *Toxoplasma gondii* sequesters centromeres to a specific nuclear region throughout the cell cycle. *Proc. Natl Acad. Sci. USA* **108**, 3767–3772 (2011).
30. Cowman, A. F., Morry, M. J., Biggs, B. A., Cross, G. A. & Foote, S. J. Amino acid changes linked to pyrimethamine resistance in the dihydrofolate reductase-thymidylate synthase gene of *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **85**, 9109–9113 (1988).
31. Peterson, D. S., Walliker, D. & Wellems, T. E. Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in falciparum malaria. *Proc. Natl Acad. Sci. USA* **85**, 9114–9118 (1988).
32. Muller, I. B. & Hyde, J. E. Folate metabolism in human malaria parasites–75 years on. *Mol. Biochem. Parasitol.* **188**, 63–77 (2013).
33. Japrung, D., Leartsakulpanich, U., Chusacultanachai, S. & Yuthavong, Y. Conflicting requirements of *Plasmodium falciparum* dihydrofolate reductase mutations conferring resistance to pyrimethamine-WR99210 combination. *Antimicrob. Agents Chemother.* **51**, 4356–4360 (2007).
34. Yuvaniyama, J. et al. Insights into antifolate resistance from malarial DHFR-TS structures. *Nat. Struct. Biol.* **10**, 357–365 (2003).
35. Hankins, E. G., Warhurst, D. C. & Sibley, C. H. Novel alleles of the *Plasmodium falciparum* dhfr highly resistant to pyrimethamine and chlorcycloguanil, but not WR99210. *Mol. Biochem. Parasitol.* **117**, 91–102 (2001).
36. Ng, C. L. & Fidock, D. A. *Plasmodium falciparum* in vitro drug resistance selections and gene editing. *Methods Mol. Biol.* **2013**, 123–140 (2019).
37. Chiu, J. E. et al. Effective therapy targeting cytochrome bc1 prevents *Babesia* erythrocytic development and protects from lethal infection. *Antimicrob. Agents Chemother.* **65**, e0066221 (2021).
38. Lawres, L. A. et al. Radical cure of experimental babesiosis in immunodeficient mice using a combination of an endochin-like quinolone and atovaquone. *J. Exp. Med.* **213**, 1307–1318 (2016).
39. Korsinczky, M. et al. 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 (2000).
40. Allred, D. R. Variable and variant protein multigene families in *Babesia bovis* persistence. *Pathogens* https://doi.org/10.3390/pathogens8020076 (2019).
41. Wang, J. et al. Comparative genomic analysis of *Babesia duncani* responsible for human babesiosis. *BMC Biol.* **20**, 153 (2022).
42. Guizetti, J. & Scherf, A. Silence, activate, poise and switch! Mechanisms of antigenic variation in *Plasmodium falciparum*. *Cell Microbiol.* **15**, 718–726 (2013).
43. Potluri, V. et al. Discovery of FNDR-20123, a histone deacetylase inhibitor for the treatment of *Plasmodium falciparum* malaria. *Malar. J.* **19**, 365 (2020).
44. Malmquist, N. A., Moss, T. A., Mecheri, S., Scherf, A. & Fuchter, M. J. Small-molecule histone methyltransferase inhibitors display rapid antimalarial activity against all blood stage forms in *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **109**, 16708–16713 (2012).
45. Miao, J. et al. A unique GCN5 histone acetyltransferase complex controls erythrocyte invasion and virulence in the malaria parasite *Plasmodium falciparum*. *PLoS Pathog.* **17**, e1009351 (2021).
46. Nardelli, S. C. et al. Genome-wide localization of histone variants in *Toxoplasma gondii* implicates variant exchange in stage-specific gene expression. *BMC Genomics* **23**, 128 (2022).
47. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
48. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
49. Di Genova, A., Buena-Atienza, E., Ossowski, S. & Sagot, M. F. Efficient hybrid de novo assembly of human genomes with WENGAN. *Nat. Biotechnol.* **39**, 422–430 (2021).
50. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
51. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).
52. Varoquaux, N., Ay, F., Noble, W. S. & Vert, J. P. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* **30**, i26–i33 (2014).
53. Goddard, T. D. et al. UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).
54. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
55. Lechner, M. et al. Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC Bioinformatics* **12**, 1–9 (2011).
56. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
57. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
58. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
59. Amos, B. et al. VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res.* **50**, D898–D911 (2022).
60. Wolff, J. et al. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177–W184 (2020).

Acknowledgements

We thank R. Gao for her contribution to the initial efforts to sequence the *B. duncani* genome. C.B.M.'s research was supported by grants from the National Institutes of Health (AI097218, GM110506, AI123321 and R43AI136118), the Steven and Alexandra Cohen Foundation (Lyme 62 2020), and the Global Lyme Alliance. S.L.'s research was supported by grants by the US National Science Foundation (IIS 1814359) and the National Institutes of Health (1R01AI169543-01). K.G.L.R.'s research was supported by the National Institutes of Allergy and Infectious Diseases (R01 AI136511, R01 AI142743-01 and R21 AI142506-01), the University of California, Riverside (NIFA-Hatch-225935) and the Health Institute Carlos III (PI20CIII/00037).

Author contributions

C.B.M., S.L. and K.G.L.R. conceived the study. C.B.M., S.L., K.G.L.R., E.M., A.S.-F., P.S., P.V. and S.G. designed the experiments. P.S., P.V., J.T., L.C., G.B., M.G., E.M. and L.M.G. conducted the laboratory

experiments. S.L., P.S., Q.L., P.V., E.K., T.F., S.G., M.M., S.A., M.G., X.M.L., T.L., S.C., E.C., Y.H., W.M., L.M.G., S.S., K.E., A.S.-F., O.S.H., K.G.L.R. and C.B.M. performed analysis of various data sets. S.L., C.B.M., K.G.L.R., P.S., P.V. and S.G. wrote the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41564-023-01360-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-023-01360-8.

Correspondence and requests for materials should be addressed to Stefano Lonardi, Karine G. Le Roch or Choukri Ben Mamoun.

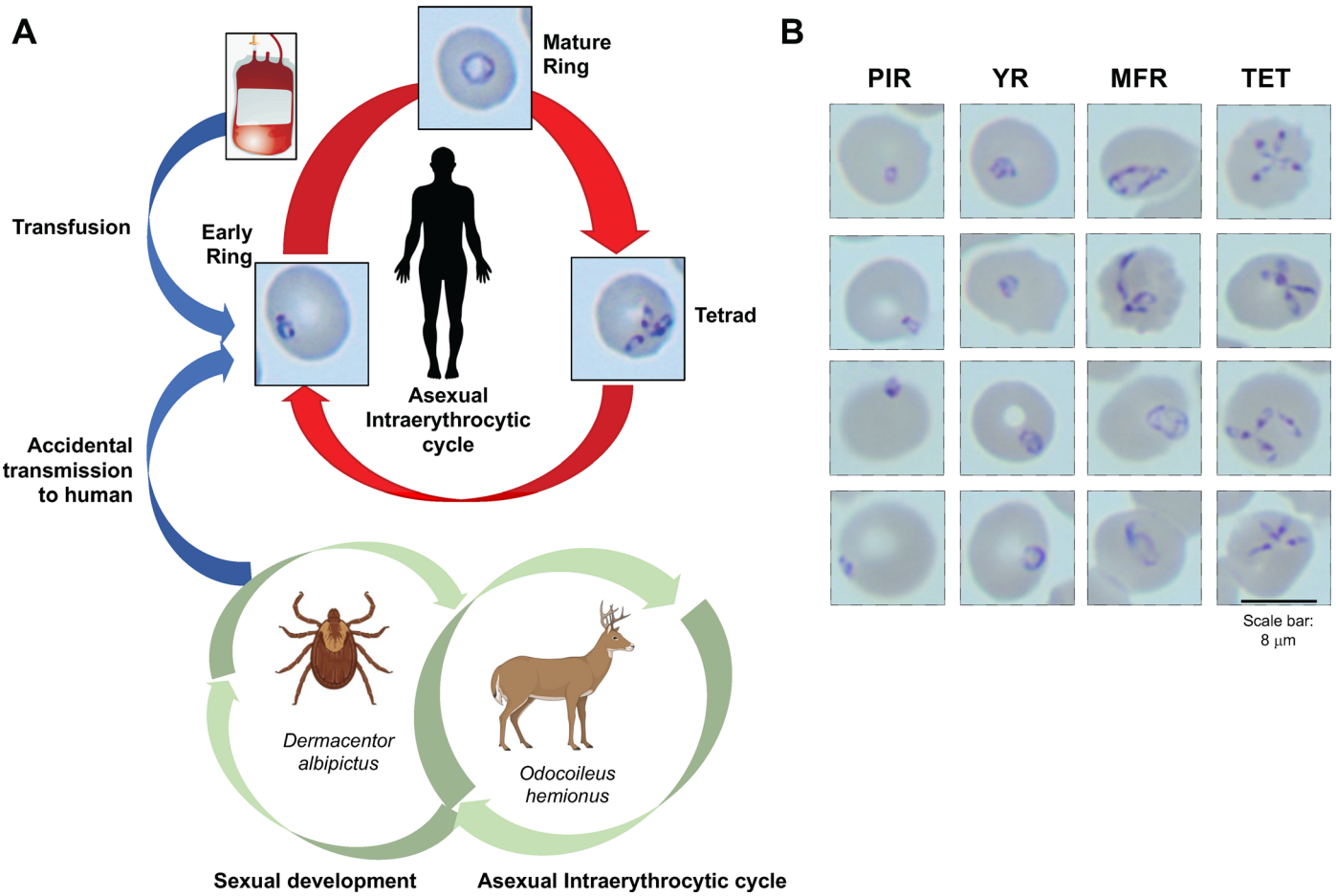
Peer review information *Nature Microbiology* thanks Ellen Knuepfer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

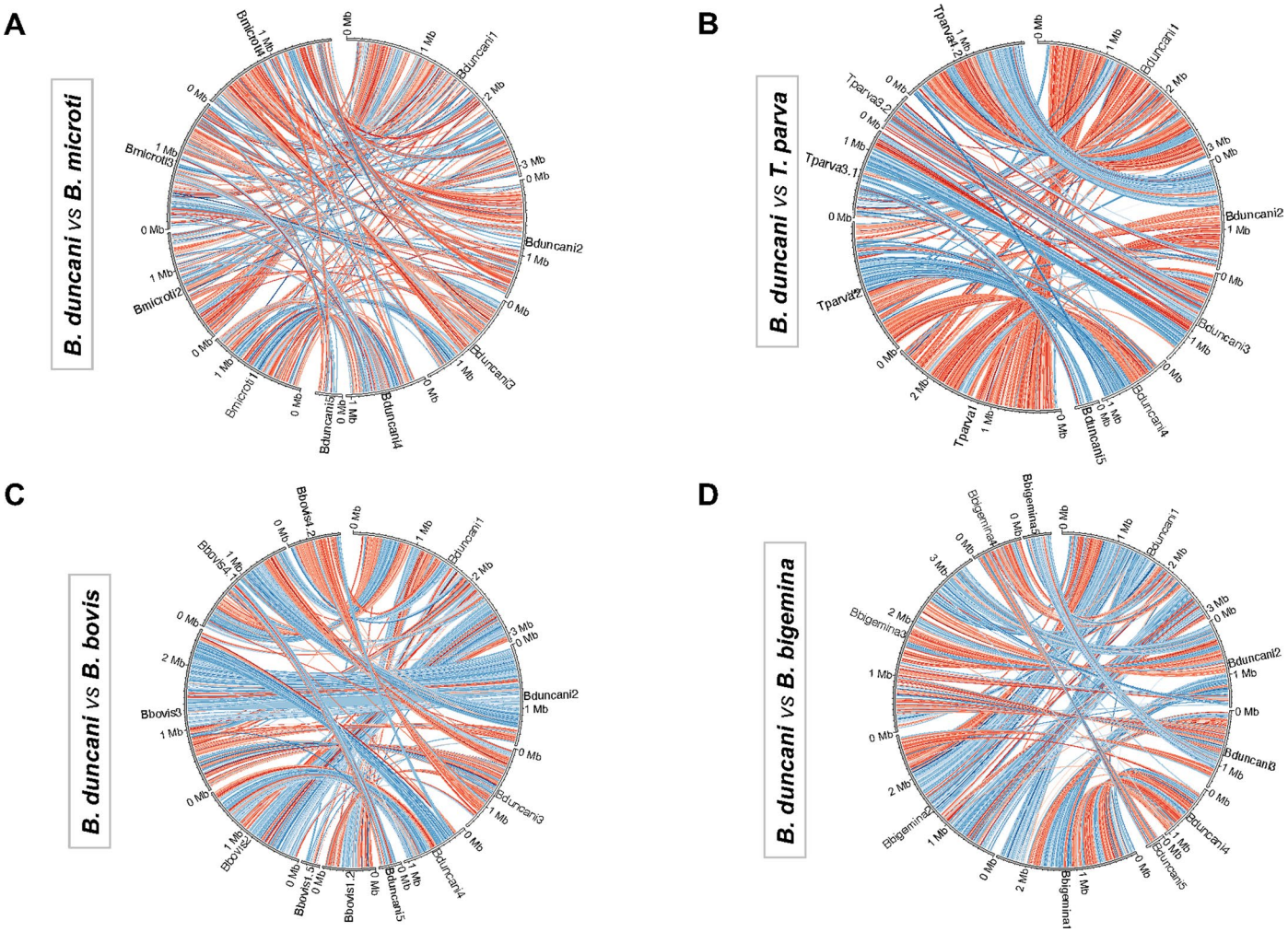
[1]Department of Internal Medicine, Section of Infectious Diseases, Yale School of Medicine, New Haven, CT, USA. [2]Department of Computer Science and Engineering, University of California, Riverside, CA, USA. [3]Department of Statistics, University of California, Riverside, CA, USA. [4]Department of Molecular, Cell and Systems Biology, University of California, Riverside, CA, USA. [5]Institut de Biologie Computationnelle (IBC), and Institut de Recherche en Cancérologie de Montpellier (IRCM - INSERM U1194), Institut régional du Cancer Montpellier (ICM) and Université de Montpellier, Montpellier, France. [6]Parasitology Reference and Research Laboratory, National Centre for Microbiology, Instituto de Salud Carlos III, Majadahonda, Madrid, Spain. [7]Reference and Research Laboratory on Food and Waterborne Bacterial Infections, National Centre for Microbiology, Instituto de Salud Carlos III, Majadahonda, Madrid, Spain. [8]Unidad Universitaria de Secuenciación Masiva y Bioinformática, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, México. [9]Department of Biology, University of Pennsylvania, Philadelphia, PA, USA. [10]These authors contributed equally: Pallavi Singh, Stefano Lonardi. ✉e-mail: stelo@cs.ucr.edu; karinel@ucr.edu; choukri.benmamoun@yale.edu

https://doi.org/10.1038/s41564-023-01360-8



Extended Data Fig. 1 | Life-cycle of *B. duncani*. The *B. duncani* life cycle. **A.** Schematic representation of the life cycle of *B. duncani* in humans, tick vector and mule deer. **B.** Representative images of the various forms of *B. duncani* grown in human erythrocytes *in vitro*. PIR: Post-invasion rings; YR: young rings; MFR: mature and filamentous rings; TET: tetrads. The representative images from three independent experiments performed in duplicates.

https://doi.org/10.1038/s41564-023-01360-8



Extended Data Fig. 2 | Synteny circos plots. Circos synteny plots. The chromosomes of *B. duncani* are illustrated on the right semicircle on all circular plots, and the chromosomes of the other organisms are on the left semicircle (**A**: *B. microti*, **B**: *T. parva*, **C**: *B. bovis*, **D**: *B. bigemina*); blue arcs indicate syntenies, red arcs indicate syntenies involved in a reversals; the intensity of the color is proportional to the level of collinearity; the number after the species' name refers to the chromosome number (when chromosomes are broken into pieces, fragments.

https://doi.org/10.1038/s41564-023-01360-8

Extended Data Table 1 | Genome & transcriptome data

Genomic Data	
gDNA PacBio HiFi reads (#)	176,365
gDNA PacBio HiFi total length (bp)	1,561,750,735
gDNA HiFi read average length (bp)	8,814
HiFi read SRA accession	SRR18778747
Genome assembly accession	BioProject PRJNA821606 BioSample SAMN27116006
Assembly length (bp)	10,408,054
Contigs in assembly (#)	5 chromosomes 160 unplaced contigs
N50 (bp)	1,428,345
Chromosome 1 (bp) Chromosome 2 (bp) Chromosome 3 (bp) Chromosome 4 (bp) Chromosome 5 (bp)	3,133,627 1,587,141 1,428,345 1,068,143 353,949
Transcriptomic data	
Illumina RNA-Seq reads	~119 million reads
Illumina SRA accession	SRR18907291
PacBio IsoSeq reads	~52,000 reads
PacBio SRA accession	SRR18902718

https://doi.org/10.1038/s41564-023-01360-8

Extended Data Table 2 | Genome comparison and gene statistics

	B. duncani	*B. microti*	*B. bovis*	*T. parva*	*T. annulata*	*P. falciparum*	*C. parvum*
Genome							
Total Size (Mbp)	10.41	6.4	8.2	8.3	8.4	23.3	9.1
Number of chromosomes	5	4	4	4	4	14	8
G+C content (%)	37.3	36	41.5	34.1	32.5	19.3	30.8
Genes							
Number of protein-coding genes	4,222	3,567	3,706	3,796	4,082	5,324	3,805
Mean CDS length (bp)	1,500	1,327	1,503	1,407	1,602	2,326	1,844
Mean CDS length including introns (bp)	1,656	1,472	1,609	1,654	1,802	2,590	1,851
Coding regions (%)	60	74	68	68	73	53	76
Coding regions including introns (%)	67	82	73	80	82	59	77
Genes with introns (%)	63	75	60	75	71	54	4
Exons							
Average number per gene	3.1	3.8	2.8	2.7	3.9	2.6	1.1
Mean length (bp)	356	347	547	514	416	904	1,748
Total length (%)	60	74	68	68	73	53	77
Introns							
Average number per gene	2.1	2.8	1.7	2.6	2.9	1.6	0.1
Mean length (bp)	83	51	60	94	70	167	96
Total length (%)	7	8	5	12	9	6	0.02
RNAs							
Number of tRNA genes	57	44	70	71	47	72	45
Number of 5S rRNA genes	3	2	N/A	1	3	3	6
Number of 5.8S/18S/28S rRNA units	2	2	5	8	1	13	9

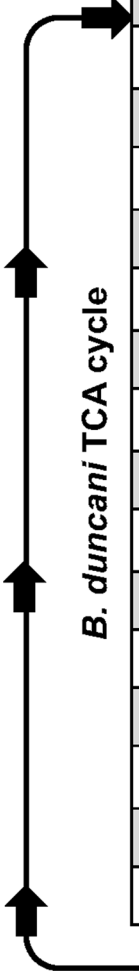
https://doi.org/10.1038/s41564-023-01360-8

Extended Data Table 3 | Glycolytic pathway in *B. duncani*

Glycolysis Steps	Enzyme	Gene ID	Protein Length	RNA expression level (TPM)
Glucose				
↓	Hexokinase	BdWA1_000031-T1	486	73
Glucose-6P				
↓	Phosphoglucose Isomerase	BdWA1_002840-T1	422	511
Fructose-6P				
↓	6-Phosphofructokinase	BdWA1_003208-T1	1313	506
Fructose-1,6P2				
↓	Fructose-1,6-bisphosphate aldolase	BdWA1_000083-T1	356	1884
Glyceraldehyde-3P				
↓	Glyceraldehyde-3P dehydrogenase	BdWA1_001300-T1	336	8497
Glycerate-1,3P2				
↓	Phosphoglycerate Kinase	BdWA1_002139-T1	416	2018
Glycerate-3P				
↓	Phosphoglycerate mutase	BdWA1_003538-T1#	299	50
Glycerate-2P				
↓	Enolase	BdWA1_003622-T1	442	15
Phosphoenolpyruvate				
↓	Phosphoenolpyruvate carboxykinase	BdWA1_000439-T1	556	425
Pyruvate				
↓	Lactate dehydrogenase	BdWA1_000746-T1	320	22
Lactate				

https://doi.org/10.1038/s41564-023-01360-8

Extended Data Table 4 | TCA cycle of *B. duncani*



Krebs Cycle steps	Enzyme	Predicted Gene ID	Protein Length	C/M/S	TM	RNA expression level (TPM)
Citrate						
	Aconitate Hydratase	BdWA1_001117-T1	806	M	-	745
Isocitrate						
	Isocitrate dehydrogenase	BdWA1_001763-T1	454	M	-	773
2-oxaloglutarate						
	2-Oxoglutarate dehydrogenase	BdWA1_000507-T1	940	M	-	293
Succinyl-CoA						
	Succinyl-CoA synthetase	BdWA1_001709-T1# BdWA1_002351-T1	309 423	M M	- -	0.58 204
Succinate						
	Succinate dehydrogenase	BdWA1_004214-T1* (α) BdWA1_000189-T1 (β)	617 280	M -	- -	0 93
Fumarate						
	Fumarase	BdWA1_000547-T1	578	-	-	175
Malate						
	Malate dehydrogenase	BdWA1_002484-T1	489	-	1	165
Oxaloacetate						
	Citrate synthase	BdWA1_003132-T1	553	-	-	485

https://doi.org/10.1038/s41564-023-01360-8

Extended Data Table 5 | Predicted AP2 factors of *B. duncani*

Gene Name	Gene ID #	RNA expression (TPM)	Protein length	Protein MW (kDa)	Domains
BdAP2-1	BdWA1_000011-T1	130	239	27.66	AP2
BdAP2-2	BdWA1_000054-T1	9	619	69.6	AP2
BdAP2-3	BdWA1_000531-T1	178	480	55.1	AP2
BdAP2-4	BdWA1_000698-T1	89	605	67.32	AP2, ACDC
BdAP2-5	BdWA1_001000-T1	66	224	26.76	AP2
BdAP2-6	BdWA1_001055-T1	532	829	92.62	AP2
BdAP2-7	BdWA1_001066-T1*	0	307	35.83	AP2
BdAP2-8	BdWA1_001248-T1	275	494	56.83	AP2
BdAP2-9	BdWA1_001388-T1	289	381	42.57	AP2
BdAP2-10	BdWA1_001541-T1	311	602	66.71	AP2
BdAP2-11	BdWA1_001617-T1	545	359	40.99	AP2, ACDC
BdAP2-12	BdWA1_001779-T1*	0.27	268	30.25	AP2
BdAP2-13	BdWA1_002104-T1	65	249	28.56	AP2, ACDC
BdAP2-14	BdWA1_002244-T1	3.72	152	17.83	AP2
BdAP2-15	BdWA1_002418-T1	307	190	22.60	AP2
BdAP2-16	BdWA1_002483-T1	303	391	45.61	Clat_adaptor_s, AP2
BdAP2-17	BdWA1_002708-T1	38	413	47.66	AP2, AP2, AP2
BdAP2-18	BdWA1_003055-T1	403	598	68.47	RPT1, AP2
BdAP2-19	BdWA1_003273-T1#	5382	441	50.28	AP2

https://doi.org/10.1038/s41564-023-01360-8

Extended Data Table 6 | Predicted GPI-anchored proteins of *B. duncani*

				SignalP 5.0	PredGPI		BLAST	
GPI-AP ID	Protein ID	Length (aa)	RNA expression (TPM)	Score	Prediction	Specificity	Description	Taxonomy distribution
BdGPI01	BdWA1_000266-T1	498	22	0.604043	Highly probable	100%	AAY83299, putative novel antigen from *Babesia* sp. WA1	WA1
BdGPI02	BdWA1_000268-T1	514	40	0.604043	Highly probable	100%	AAY83299, putative novel antigen from *Babesia* sp. WA1	WA1
BdGPI03	BdWA1_000269-T1	506	1.8	0.604043	Highly probable	100%	AAY83299*, putative novel antigen from *Babesia* sp. WA1	WA1
BdGPI04	BdWA1_000270-T1	514	34	0.604043	Highly probable	100%	AAY83299, putative novel antigen from *Babesia* sp. WA1	WA1
BdGPI05	BdWA1_000500-T1	189	24	0.948747	Highly probable	99.9%	No significant similarity found	NA
BdGPI06	BdWA1_000716-T1	236	1.4	0.989215	Probable	99.8%	BmGPI14, B. microti R1	Piroplasmida
BdGPI07	BdWA1_001123-T1	377	278	0.846581	Highly probable	100%	AAY83302*, putative novel antigen from *Babesia* sp. WA1	WA1
BdGPI08	BdWA1_001649-T1	457	123	0.904527	Weakly probable	99.4%	hypothetical protein	Piroplasmida (not *B.microti*)
BdGPI09	BdWA1_001913-T1	729	285	0.992360	Highly probable	100%	carbonic anhydrase	Apicomplexa (+)
BdGPI10	BdWA1_001963-T1	1851	5	0.972058	Probable	99.6%	hypothetical protein	WA1
BdGPI11	BdWA1_002110-T1	330	32	0.998113	Highly probable	100%	No significant similarity found	NA
BdGPI12	BdWA1_002112-T1	378	160	0.980121	Highly probable	99.9%	No significant similarity found	NA
BdGPI13	BdWA1_002223-T1	242	1656	0.962141	Highly probable	99.9%	AAY83296*, putative novel antigen from *Babesia* sp. WA1	WA1
BdGPI14	BdWA1_002224-T1	518	408	0.969359	Highly probable	99.9%	AAY83296, putative novel antigen from *Babesia* sp. WA1	WA1
BdGPI15	BdWA1_002538-T1	631	17	0.985101	Probable	99.8%	No significant similarity found	NA
BdGPI16	BdWA1_003372-T1	122	8441	0.736131	Highly probable	99.9%	No significant similarity found	NA
BdGPI17	BdWA1_003373-T1	127	8200	0.758117	Highly probable	99.9%	No significant similarity found	NA

https://doi.org/10.1038/s41564-023-01360-8

Extended Data Table 7 | Therapeutic indexes of antifolates against different human cell lines cultured and tested in vitro

Drug	*B. duncani* IC_{50}	HeLa LC_{50}	HepG2 LC_{50}	HEK293 LC_{50}	HCT LC_{50}	TI
WR-99210	0.5 ± 0.11 nM	> 0.9 mM	1.1 mM	1 mM	1.05 mM	>15K
Pyrimethamine	940 ± 17 nM	> 0.9 mM	1.69 mM	0.5 mM	>0.03 mM	3-180

https://doi.org/10.1038/s41564-023-01360-8

Extended Data Table 8 | Kinetic parameters of purified recombinant DHFR-TS enzyme

Substrate	DHFR-TS Enzyme	K_m [μM]	V_{max} [μmol · min^{-1} ×10^4]	k_{cat} [s^{-1}]	k_{cat}/K_m [s^{-1} μM^{-1} ×10^2]
DHF	*B. duncani*	37.1 ± 0.9	7.2 ± 0.6	0.3 ± 0.03	0.8
	B. microti	12.2 ± 1.6	6.9 ± 1.1	1.2 ± 0.2	9.2
NADPH	*B. duncani*	32.9 ± 1.0	9.8 ± 2.1	0.4 ± 0.1	1.2
	B. microti	27.3 ± 0.2	10.1 ± 1.6	1.7 ± 0.3	6.0

nature portfolio

Corresponding author(s): DBPR, NMICROBIOL-22051126B

Last updated by author(s): Feb 22, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

DNA Prep for PacBio
* Pacific Biosciences Smart Link v10.2 (https://www.pacb.com/support/software-downloads/)

DNA Prep for Optical Map
* Bionano Solve v3.71 (https://bionanogenomics.com/support/software-downloads/)

Data analysis

Assembly
* Minimap2 v2.17 (https://github.com/lh3/minimap2)
* HiCANU v2.2 (https://github.com/marbl/canu)
* HiFiASM v0.16 (https://hifiasm.readthedocs.io/en/latest/index.html)
* Wengan v0.2 (https://github.com/adigenova/wengan)
* Bionano Hybrid Scaffolding pipeline v1.7 (https://bionanogenomics.com/support/software-downloads/)
* PolyPolish v0.50 (https://github.com/rrwick/Polypolish)
* BUSCO v5.3.2 (https://busco.ezlab.org/)

* Hi-C analysis
* HiCExplorer v3.7.2 (https://hicexplorer.readthedocs.io/en/latest/)
* BWA v0.7.17 (https://github.com/lh3/bwa)

3D modeling
* PASTIS v0.5.0 (https://github.com/hiclib/pastis/)
* ChimeraX v1.3 (https://www.cgl.ucsf.edu/chimerax/)

PacBio IsoSeq
* PacBio cDNA_Cupcake v28.0.0 (https://github.com/Magdoll/cDNA_Cupcake)

Annotation
* RepeatMasker v4.0.6 (http://repeatmasker.genome.washington.edu/)
* FunAnnotate v1.8.9 (https://funannotate.readthedocs.io/en/latest/)
* InterProScan v5.55-88 (https://www.ebi.ac.uk/interpro/download/)

Synteny and gene localization plots
* mummer2circos docker image 9/7/2021 (https://github.com/metagenlab/mummer2circos)
* Circos v0.69-8 (http://circos.ca/)
* Biopython v1.79 (https://biopython.org/)
* Gene localization custom scripts (https://github.com/ucrbioinfo/GeneLocalizationBduncani)

Orthology detection
* OrthoMCL v6.9 (https://orthomcl.org/)
* OrthoMCL mapping pipeline implemented as a Galaxy workflow (https://orthomcl.org/orthomcl/app/galaxy-orientation)
* Database searches were conducted in:
  * VEuPathDB releases 56 and 57 (https://VEuPathDB.org)
  * PiroplasmaDB releases 56 and 57 (https://PiroplasmaDB.org)
  * PlasmoDB releases 56 and 57releases 56 and 57 (https://PlasmoDB.org)
  * ToxoDB releases 56 and 57 (https://ToxoDB.org)
  * CryptoDB releases 56 and 57 (https://CryptoDB.org)
* Venn diagrams of orthology group overlap InteractiVenn (http://www.interactivenn.net)

RNA-Seq processing and analysis
* FastQC v0.11.8 (https://github.com/s-andrews/FastQC)
* Trimmomatic v0.39 (https://github.com/timflutre/trimmomatic)
* Sickle v1.33 (https://github.com/najoshi/sickle)
* HISAT2 v2.2.1 (http://daehwankimlab.github.io/hisat2/)
* Samtools v1.11 (https://github.com/samtools/samtools)
* StringTie v2.2.1 (https://github.com/gpertea/stringtie)
* TopGO v2.40.0 (https://bioconductor.org/packages/release/bioc/html/topGO.html)
* Custom scripts (https://github.com/Sabel14/Babesia_RNAseq_2022)

Drug efficacy study, Assessment of drug cytotoxicity, Steady-state kinetics, half-maximal inhibitory concentration
* Graph Pad Prism v9.2.1 (https://www.graphpad.com/)

Phylogenetic tree reconstruction
* MUSCLE v5.1 (https://github.com/rcedgar/muscle)
* MEGA X v11 (https://www.megasoftware.net/)

Gene orthology based classification analysis
*MAFFTv7.505 (https://mafft.cbrc.jp/alignment/software/)
* UpSetR v1.4.0 (https://cran.r-project.org/web/packages/UpSetR/index.html)
*ProteinOrtho v6 (https://gitlab.com/paulklemm_PHD/proteinortho/)
* Raxml v8 (https://cme.h-its.org/exelixis/web/software/raxml/)
* iTOL v6 (https://itol.embl.de/)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All the datasets generated for the current study are available in the NCBI/SRA repository, as Bioproject PRJNA821606, as follows
* PacBio HiFi reads (SRA accession number SRR18778747)
* B. duncani genome assembly (NCBI Genome submission: SUB11253661)
* RNA-Seq (SRA accession number SRR18907291)
* RNA Iso-Seq (SRA accession number SRR18902718)
* Hi-C reads (SRA accession number SRR19325692)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Different biological samples from the same Babesia duncani WA1 strain were used for each of the following applications: PacBio HiFi sequencing, Nanopore sequencing (but not included in the study), Illumina sequencing, IsoSeq, RNA-Seq, Hi-C, Bionano optical map.
Data exclusions	No data was excluded from the analysis.
Replication	Each of the drug efficacy experiment was repeated three times for each compound, with each experiment including three biological replicates.
Randomization	No animal study, clinical data, or study/ies that require randomization was conducted.
Blinding	No animal study, clinical data, or study/ies that require blinding was conducted.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	anti-H3K9me3 (abcam; catalog #ab8898), anti-H3K14ac (diagenode; catalogue #C15210005)
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HeLa (CCL-2) from ATCC, Hep G2 [HEPG2] from ATCC, HEK293T (CRL-11268) from ATCC, HCT116 (CCL-247) from ATCC.
Authentication	Authenticated cell lines were purchased from ATCC.
Mycoplasma contamination	All cell lines are negative of Mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	Data provided in the excel spreadsheer format.
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*