nature COMMUNICATIONS

ARTICLE

Received 13 Feb 2014 | Accepted 18 Jul 2014 | Published 9 Sep 2014

DOI: 10.1038/ncomms5754 OPEN

Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts

Thomas D. Otto[1], Julian C. Rayner[1], Ulrike Böhme[1], Arnab Pain[1,2], Natasha Spottiswoode[3,4], Mandy Sanders[1], Michael Quail[1], Benjamin Ollomo[5], François Renaud[6], Alan W. Thomas[7], Franck Prugnolle[5,6], David J. Conway[8], Chris Newbold[1,4,*] & Matthew Berriman[1,*]

Plasmodium falciparum causes most human malaria deaths, having prehistorically evolved from parasites of African Great Apes. Here we explore the genomic basis of *P. falciparum* adaptation to human hosts by fully sequencing the genome of the closely related chimpanzee parasite species *P. reichenowi*, and obtaining partial sequence data from a more distantly related chimpanzee parasite (*P. gaboni*). The close relationship between *P. reichenowi* and *P. falciparum* is emphasized by almost complete conservation of genomic synteny, but against this strikingly conserved background we observe major differences at loci involved in erythrocyte invasion. The organization of most virulence-associated multigene families, including the hypervariable *var* genes, is broadly conserved, but *P. falciparum* has a smaller subset of *rif* and *stevor* genes whose products are expressed on the infected erythrocyte surface. Genome-wide analysis identifies other loci under recent positive selection, but a limited number of changes at the host–parasite interface may have mediated host switching.

[1] Parasite Genomics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK. [2] Biological and Environmental Sciences and Engineering (BESE) Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. [3] National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, Maryland 20852, USA. [4] Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK. [5] Centre International de Recherches Médicales de Franceville, CIRMF, BP 769 Franceville, Gabon. [6] Laboratoire MIVEGEC, UMR 5290 CNRS-IRD-UMI-UMII, IRD, BP 64501, 34394 Montpellier, France. [7] Biomedical Primate Research Centre, Department of Parasitology, 2280 GH Rijswijk, The Netherlands. [8] Department of Pathogen Molecular Biology, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to C.N. (email: chris.newbold@imm.ox.ac.uk) or to M.B. (email: mb4@sanger.ac.uk).

Plasmodium falciparum causes the overwhelming majority of human malaria mortality, and is only distantly related to the other major human *Plasmodium* species. Until recently, the only known close relative of *P. falciparum* was *P. reichenowi*, a malaria parasite of chimpanzees that is morphologically so similar to *P. falciparum* that they were initially thought to be the same species[1]. In recent years, this simple view of the origins of *P. falciparum* has been comprehensively overhauled by a series of molecular studies that have revealed an unexpected diversity of *P. falciparum* related parasites in African apes[2–6]. These studies establish that all human *P. falciparum* parasites fall as a single cluster within a broader network of ape parasites, now collectively referred to as the *Laverania* subgenus, and the closest relative and likely origin of human *P. falciparum* is a clade of parasites found in western gorillas[3]. *Laverania* parasites appear to be largely host specific, with individual clades infecting only chimpanzees or gorillas, even when the apes are sympatric, and there are at least three clades infecting both chimpanzees (*P. reichenowi, P. gaboni* and *P. billcollinsi*) and gorillas (*P. praefalciparum, P. adleri* and *P. blacklocki*). However, to date, comparisons between these new *Laverania* clades have relied primarily on mitochondrial genome sequences, with only fragments of the nuclear genome amplified and sequenced. An understanding of the forces driving the radiation of the *Laverania* subgenus, and the extent of genomic diversity between *Laverania* clades, awaits the generation of whole-genome sequences for all species. While low-coverage capillary sequence data has been generated for *P. reichenowi* previously[7], no complete genome has been generated for any *Laverania* parasite other than *P. falciparum*

In this study, we present the first comparative genomic analysis of *Laverania* parasites by comparing high-quality genome sequences of *P. reichenowi P. falciparum*[8] and partial genome sequence data from *P. gaboni*, another species of the *Laverania* subgenus, obtained from a chimpanzee blood sample following a routine health check. Our analyses show that these genomes are essentially co-linear in the core central regions, allowing us to focus on the small number of significant differences. The most striking of these involve genes associated with red cell invasion implicating them in determining host specificity. In the highly polymorphic multigene families located in the subtelomeric regions, we observe one-to-one orthology in some (*PHIST* and *FIKK*), a commonality in the basic architecture of the *var* family that is involved in antigenic variation and significant copy number variation in the *rif* and *stevor* families[4].

Results

A high-quality reference genome for *P. reichenowi* was produced from Illumina reads using *de novo* assembly and *post hoc* improvement (Supplementary Methods and Supplementary Fig. 1). The 24 Mb sequence extends into the subtelomeres on four chromosomes (Supplementary Fig. 2) and encodes 5,731 genes (Fig. 1 and Table 1). Excluding the subtelomeres, the genome is almost completely co-linear with *P. falciparum* and

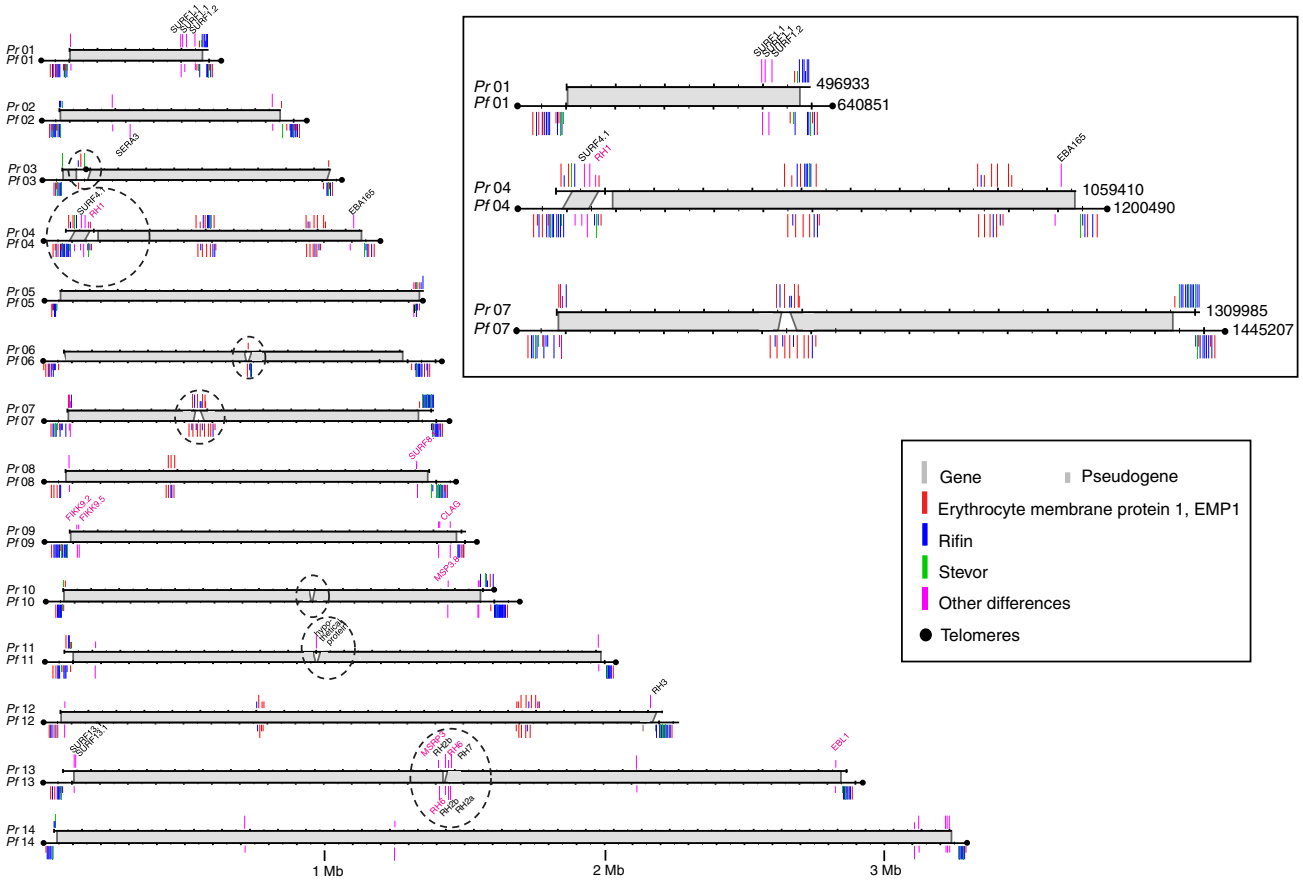

Figure 1 | Alignment of the *P. reichenowi* CDC and *P. falciparum* 3D7 genomes. The 14 chromosomes of the nuclear genome are aligned and regions with shared synteny are shown shaded in grey, divergent loci are circled. Variable extracellular multiprotein families, Stevors (green), Rifins (blue) and erythrocyte membrane protein 1 (red), are shown and pseudogenes for each of these multigene families are shown (shorter lines). Differentially distributed genes or pseudogenes (Supplementary Data 1) are shown in pink with annotations highlighting the *Pr*CDC difference and pseudogenes highlighted with pink text. The values on the right side of the chromosomes indicate the total length of the assembled sequence scaffold in base pairs.

Table 1 | Comparison of genome statistics between *P. reichenowi* CDC and *P. falciparum* 3D7.

	*P. reichenowi*CDC	*P. falciparum*3D7
Genome size (Mb)	24.0	23.2
GC content (%)	19.27	19.34
Sequence scaffolds	261	14
Unassigned contigs	237	0
Gaps	1,574	0
Genes	5,731*	5,418*
Mean gene length (bp)	2,223	2,279
Gene density (bp per gene)	4,232	4,342
Percentage coding	52.9	52.9
Pseudogenes	134	144
tRNA	45	45
ncRNA	89	100

bp, base pairs; tRNA, transfer RNA; Mb: megabases; ncRNA, non-coding RNA.
Statistics for the nuclear genomes are shown. Both species have 5.97 kb mitochondrial genomes containing only three genes and ~30 kb apicoplast genomes (29,226 bp, *Pr*CDC; 29,686 bp, *Pf*3D7) containing 30 genes.
*Number of genes includes pseudogenes and partial genes but excludes ncRNA genes

share almost identical gene content emphasizing their recent shared ancestry (Fig. 1). The vast majority of these ~5,000 'core' genes are common to both parasites, with only three genes present in *P. reichenowi* that are missing in *P. falciparum* and only one in *P. falciparum* that is absent from *P. reichenowi* (Supplementary Fig. 3). In addition, 19 intact genes in the chimpanzee parasite are pseudogenes in *P. falciparum* and there are 16 genes for which the reverse is the case (Supplementary Data 1 and Supplementary Fig. 4). Twenty-seven orthologous loci contain pseudogenes in both species, and these vary in their degree of coding sequence degeneracy.

Significant differences were anticipated for genes that function at the host–parasite interface. In *P. falciparum,* many of these genes are subtelomeric, most notably in a number of multigene families that encode proteins that are exported into the host erythrocyte. *Plasmodium* subtelomeres are highly repetitive and difficult to assemble, but we were able to assemble the majority of *P. reichenowi* coding sequences in these regions, and thereby produced the first complete genome sequence for a *Laverania* parasite other than *P. falciparum*. In several multigene families of exported proteins, including the *PHIST* and *FIKK* families[9], clear orthologues could be identified for almost all genes between *P. falciparum* and *P. reichenowi* (Fig. 2a). The few exceptions (Fig. 2a) occurred where *P. reichenowi* orthologues were pseudogenes (Supplementary Fig. 5).

The extremely polymorphic *var* multigene family encodes the PfEMP1 proteins that are central to the pathogenesis and persistence of malaria infection in humans where their expression on the surface of infected erythrocytes mediates adherence to a variety of receptors on different host cell types[10–12]. Surprisingly, between *P. falciparum* and *P. reichenowi, var* genes are conserved both in number (Table 2) and broadly, in terms of their organization and size (Fig. 2b). PfEMP1 proteins are highly polymorphic and are composed of varying numbers and classes of Duffy-binding-like (DBL) domains and cysteine-rich interdomain regions (CIDR), that mediate adhesion, with the amino terminal DBL domain (DBLα) being the most conserved[13]. The relative orders of these domains vary markedly between gene copies in *P. falciparum,* but the most frequent arrangement in *P. falciparum* (DBLa-CIDRa-DBLd-CIDRb) was also the second-most abundant in *P. reichenowi* (Supplementary Data 2). While their overall organization is conserved, individual PfEMP1 sequences were highly divergent between *P. reichenowi* and *P. falciparum*, just as they are between *P. falciparum* strains.

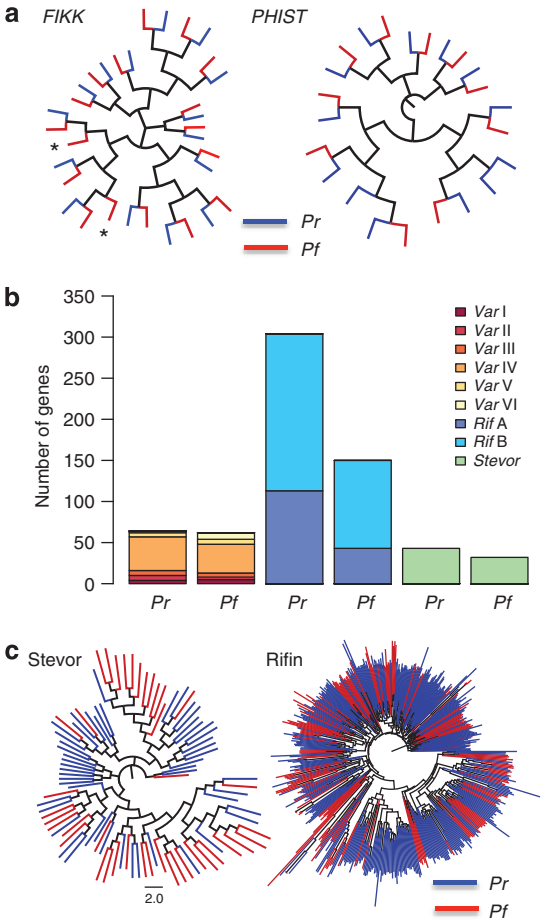

Figure 2 | Number and organization of multigene families encoding proteins exported into the infected erythrocyte. (**a**) Dendrograms of the exported kinase (FIKK) and PHIST domain containing proteins showing one-to-one orthologous pairing between *P. reichenowi* (blue) and *P. falciparum* (red). Two *P. falciparum FIKK* genes have no protein coding orthologue (starred) because the homologous genes have become pseudogenes in *P. reichenowi* (Supplementary Fig. 5). (**b**) Distribution of the number of intact members of the *var, rif* and *stevor* families within each defined subclass in both *P. falciparum* (*Pf*) and *P. reichenowi* (*Pr*). (**c**) Dendrograms of the *rif* and *stevor* protein families in *P. falciparum* and *P. reichenowi*, showing *P. reichenowi*-specific expansions in both cases.

However, even here crucial aspects were conserved. The most highly conserved amino acid sequence motif in the *P. falciparum* DBLα domain, LARSFADIGDI, was completely conserved in *P. reichenowi,* but based on the limited data available only a variant of this sequence is recognizable in *P. gaboni*, and the surrounding sequences appear to be more polymorphic (Supplementary Fig. 6). Although the extent of *var* gene diversity makes direct pairwise comparisons impossible for most of their sequences, *P. falciparum* DBLα domains can be categorized on the basis of cysteine content, length and semi-conserved motifs into six different classes[13]. All six DBLα classes were evident in *P. reichenowi* in approximately the same ratios as in *P. falciparum* (Fig. 2b and Table 2).

In summary, despite the fact that radical polymorphism in *var* genes is essential for their role in immune evasion, their number, organization and some elements of sequence appear to be conserved between chimpanzee and human parasites. By contrast, we found that the size of the *rif* and *stevor* multigene families, which also encode proteins expressed on the surface of infected red blood cells and other stages directly exposed to the host[14,15],

Table 2 | Comparison of gene families between *P. reichenowi* CDC and *P. falciparum* 3D7.

Gene family	*P. reichenowi* CDC	*P. falciparum* 3D7
Erythrocyte membrane protein 1	103	103
Complete genes	64	63
DBLα group 1	4	5
DBLα group 2	6	3
DBLα group 3	6	5
DBLα group 4	41	35
DBLα group 5	5	6
DBLα group 6	2	8
Pseudogenes (fragments)	22 (37)	34 (6)
Rifin	462	184
*Complete genes**	308	157
Type A†	113	46
Type B†	189	110
Pseudogenes (fragments)	49 (105)	27 (0)
Stevor	66	42
Complete genes	43	32
Pseudogenes (Fragments)	10 (13)	10 (0)
Maurer's cleft two transmembrane protein	5	13
Complete genes	5	12
Pseudogenes	0	1
Plasmodium-exported protein, unknown function	201	210
Complete genes	162	175
Pseudogenes (fragments)	28 (11)	35 (0)
FIKK kinase	21	21
Complete genes	18	19
Pseudogenes	3	2
Surfin	12	10
Complete genes	7	7
Pseudogenes (fragments)	2 (3)	3 (0)
Serine repeat antigen	8	9

DBL, Duffy-binding-like domain; *Surfin*, surface-associated interspersed protein.
Complete genes refers to gene predictions with full-length putative protein coding sequences; *Pseudogenes* contain at least one in-frame stop codon or a frameshift; and 'Fragments' refers to partial coding sequences truncated by a contig boundary.
*The following complete genes are neither Type A nor Type B: PRCDC_1038200 and PRCDC_0004700 in *P. reichenowi*, and PF3D7_0401600 in *P. falciparum*, respectively. A further four *P. reichenowi* Rifins were excluded from classification owing to sequence gaps.
†Rifins were classified into types A and B as previously described in ref. 28.

is markedly different between the two species. While the function of these genes is not known, homologues are found in all *Plasmodium* species, implying a universal and ancient role in the relationship between *Plasmodium* parasites and their vertebrate hosts. There are 568 *rif* genes in *P. reichenowi* and only 185 in *P. falciparum*, with the number of pseudogenes differing by a similar ratio (49 and 27, respectively; Table 2 and Fig. 2b). The number of *stevor* genes is also higher in *P. reichenowi* (66) than in *P. falciparum* (42). Successful colonization of humans is therefore clearly possible with a much reduced repertoire of these two important multigene families.

The relative size of the *rif* and *stevor* gene families were the most striking differences between *P. reichenowi* and *P. falciparum* at the genome structure level, but having almost complete sequence data for the *P. reichenowi* genome also enabled us to take a systematic approach to identifying genes that were significantly different in sequence between these two species (Supplementary Note 1), and therefore might be involved in adaptation to human hosts. Unfortunately, a limited representation within the sequencing libraries precluded performing similar systematic analyses in parallel using the *P. gaboni* genome data (Supplementary Note 1 and Supplementary Data 3).

The Hudson–Kreitman–Aguade ratio (HKAr)[16] summarizes the ratio of interspecific nucleotide divergence (K) to intraspecific polymorphism (π). We used an assembly-based approach and five different *P. falciparum* strains to enable π and K to be derived for most of the coding sequences in the *P. falciparum* genome ($n = 4{,}933$; Supplementary Data 4). The distribution of K values was highly skewed across genes (Supplementary Fig. 7a), with a mean of 0.024, whereas for intergenic regions, K values were much higher but less skewed, with a mean of 0.033 (Supplementary Data 5; $P = 2.2e-16$). To identify genes with unusually high levels of polymorphism in *P. falciparum* and to minimize random sampling variance in the HKAr test, we focussed on loci for which the denominator K had ≥ 5 fixed nucleotide differences. Coding regions generally contained a lower ratio of polymorphisms to fixed differences than intergenic regions, reflected in lower π/K ratios (coding, mean = 0.025, $n = 4445$; intergenic, mean = 0.044; Mann–Whitney test, 10^{-16}). Particularly, high HKAr values of over 0.15—likely to reflect diversifying selection within *P. falciparum*—were seen for 72 coding sequences (Fig. 3 and Supplementary Data 6). Of these, 45 had functional annotations, and 10 of these are considered likely to be associated with erythrocyte invasion[17], including three members of the merozoite surface protein 3 (*msp3*) gene family on chromosome 10, a highly variable locus amongst *P. falciparum* strains (Supplementary Fig. 8).

As a separate test for non-neutral evolution, we compared polymorphism versus divergence using ratios of synonymous and non-synonymous differences in coding sequences with the McDonald–Kreitman (MK) test[18]. The individual genes with significantly high ratios (Fig. 3; and Supplementary Data 4 and 7) may either be under balancing selection (maintaining intraspecific polymorphisms) or under weak negative selection (allowing some nearly neutral polymorphism to be retained but preventing fixed differences). Three genes showed extreme values in both the HKAr and MK test, and encode apical membrane antigen-1—a known polymorphic protein crucial to merozoite invasion—as well as two previously uncharacterized proteins (Fig. 3b). An indication of proteins that may have been under positive selection can also be obtained by examining *Ka*/*Ks* ratios of fixed differences between species. Of the 100 proteins with the highest *Ka*/*Ks* ratios, 77 have no known function but significantly ($P < 0.0001$, χ^2-test for enrichment compared with the genome as a whole) 31 of the 100 have predicted motifs for export into the infected erythrocyte. Six of these proteins are involved in erythrocyte invasion, again emphasizing that the strongest selection pressures is acting on the parasite–erythrocyte interface (Supplementary Data 4 and 8). Introgression between species potentially confounds interpreting evolutionary signatures of selection. However, the lack of evidence that the high HKAr loci are spatially clustered in the genome (Supplementary Fig. 9) and the smooth distribution of K values (Supplementary Fig. 7b) suggest that major genetic introgression is not a factor.

Genome-wide tests thus identified particular selection pressure on erythrocyte invasion. Moreover, of only 16 genes that are missing or pseudogenes in the *P. reichenowi* genome, three are members of invasion-associated multigene families, *SERA3*, *MSP3.8* (Supplementary Fig. 8) and *MSRP3* (on chromosomes 2, 10 and 13, respectively; Supplementary Data 1)[19]. Both genome-wide tests for selection, as well as gene loss and gain, therefore identify erythrocyte invasion loci as being significantly different between the *P. falciparum* and *P. reichenowi* genomes. Erythrocyte recognition is dependent on two multigene families,

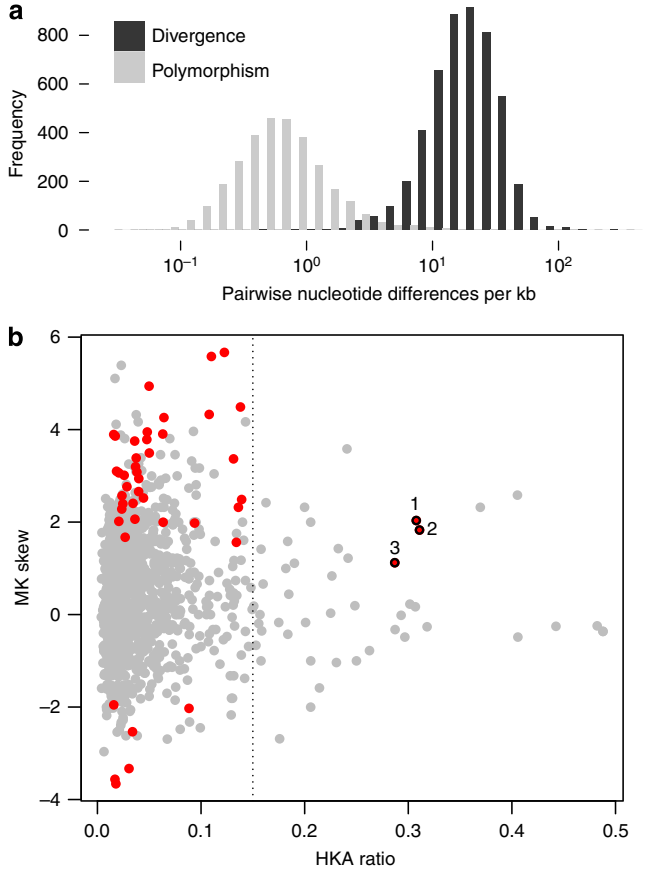

Figure 3 | Genome-wide scan for genes under selection pressure. Where *Pf*3D7–*Pr*CDC orthologues could be identified, polymorphism was calculated from the alignment of five laboratory clones and divergence was calculated from the pairwise alignment of *Pf*3D7–*Pr*CDC, respectively (**a**). From the data in **a**, McDonald–Kreitman (MK) and Hudson–Kreitman–Aguade ratio (HKAr) values were calculated for every orthologous pair of genes across the two genomes (**b**). The MK skew is the log (base 2) of the odds ratio of the 2×2 contingency table of numbers of non-synonymous versus synonymous polymorphisms among five *P. falciparum* lines divided by numbers of non-synonymous versus synonymous fixed differences between *P. reichenowi* and *P. falciparum*. Genes with significant MK skew determined by Fisher's exact test are highlighted in red. The HKAr of nucleotide polymorphism (within *P. falciparum*) divided by divergence (between *P. falciparum* and *P. reichenowi*) is independently distributed from the MK ratio. Individual values for each gene are listed in Supplementary Data 4–7. A threshold of HKAr values >0.15 (dotted line) was used to identify highlight genes under particularly high diversifying selection. Genes with high HKAr, plus significant and high MK skew: (1) apical membrane antigen 1 (PF3D7_1133400); (2) putative conserved membrane protein of unknown function (PF3D7_0104100); and (3) putative conserved membrane protein of unknown function (PF3D7_0710200).

the reticulocyte-binding-like (RBL) and erythrocyte-binding-like (EBL) proteins. Members of these families have been found in every *Plasmodium* species sequenced to date and here the differences between *P. falciparum* and *P. reichenowi* were particularly striking. Within the five-member EBL family, *eba-165*, is a pseudogene in *P. falciparum* (Supplementary Fig. 10a) but not in *P. reichenowi*[20], while EBL1 has a substantial deletion in *P. reichenowi* that may render it non-functional (Supplementary Fig. 10b). Similarly, of six RBLs in the *P. falciparum* genome only three (*Rh2b, Rh4* and *Rh5*) have clear orthologues in *P. reichenowi*. This analysis confirmed previous observations that *Rh1* (ref. 21) is a highly divergent pseudogene in *P. reichenowi*[22], whereas *Rh3* is a pseudogene in *P. falciparum*[23] but has a complete open reading frame in *P. reichenowi*[22].

The *Rh2* locus on chromosome 13 (Fig. 4a) is the most substantially different between the species. *P. falciparum*, it contains two almost identical genes, *PfRh2a* and *PfRh2b*, and a highly degenerate pseudogene, *PfRh6* (refs 24–26), whereas in *P. reichenowi*, *Rh2b* is inverted and there is an inversion and complete divergence of *Rh2a* with a gene so distinct that we refer to it as *PrRh7*. The overall arrangement of the *P. reichenowi* genomic region is also significantly different, with *PrRh2b* and *PrRh7* lying head to tail and separated by *Rh6*, a pseudogene that in *P. falciparum* lies downstream of *Rh2b*. Despite the nature of our *P. gaboni* sample precluding a full-genome assembly, we could identify an orthologue of *PrRh2b* and *PfRh2a* (Fig. 4a), as well as a close orthologue of *PrRh7*, supporting the inference that the inversion/duplication event at this locus has happened only in the *P. falciparum* branch (Fig. 4b). Surprisingly, we found strong evidence that the orthologue of the *Pr-* and *PfRh6* pseudogenes is in fact an intact gene in *P. gaboni*, suggesting further that changes at this locus had also taken place between the chimpanzee *Plasmodium* species (Fig. 4). The radical changes in gene identity and organization at this individual invasion gene locus reinforce the whole-genome analysis, and suggest that extensive changes in the erythrocyte invasion repertoire may have accompanied the expansion of *Laverania* species and their transitions between hosts.

Discussion

By generating the first complete genome sequence of any non-human *Laverania* parasite, as well as partial sequence from a second, we have been able to take the first systematic view of differences between these important pathogens, from which the most significant human malaria parasite evolved. Despite the fact that these are distinct species with no evidence of recent introgression, our data show that the genomes of human and chimpanzee *Laverania* species are remarkably similar. Excluding the subtelomeres, which contain genes associated with antigenic variation and are also highly polymorphic between different *P. falciparum* genotypes, the genomes are effectively co-linear. Against this strongly conserved 'background', differentially distributed genes or diverged genes presented a stark contrast, and we observed two different classes of divergence—genes within the core genome that show strong evidence of selection and differences in multigene families in the subtelomeric regions.

In the case of multigene families, many families such as the *PHIST* and *FIKK* genes show clear orthology between *P. reichenowi* and *P. falciparum*. Rodent *Plasmodium* genomes contain only a single *FIKK* gene, so it appears that both these families expanded before the *P. reichenowi/P. falciparum* divergence, and selection pressure to preserve the expanded number of genes has been maintained after speciation. Surprisingly, similarities between *P. falciparum* and *P. reichenowi* multigene families extended to the highly polymorphic *var* gene family that is central to parasite survival in the vertebrate host through mediating the process of antigenic variation and immune evasion. Although sequence similarity is low between the two species (as it is between different *P. falciparum* isolates), the relative ratio of different DBLα sequence types in *P. reichenowi* is very close to that in the human parasite, suggesting that maintaining the relative ratios of DBLα classes is functionally important in infections of both apes and humans. A previous analysis[27] of low coverage genome survey sequencing data from *P. reichenowi*[7] showed that many of the basic building blocks are conserved in DBLα, but this complete genome sequence shows

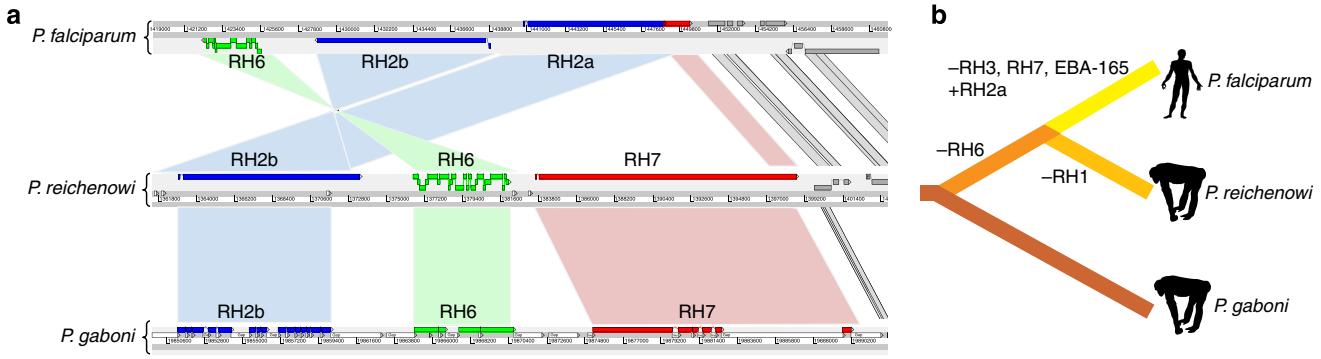

Figure 4 | Evolution of erythrocyte invasion ligands across *Laverania* parasite species. (**a**) Comparison of the RH locus on chromosome 13 between *P. falciparum* 3D7, *P. reichenowi* CDC and *P. gaboni*. The forward and reverse DNA strands are shown as thick horizontal lines (grey), with sequence gaps (white boxes), where appropriate. Annotated genes (red/green/blue) are shown above or below the DNA lines, in forward or reverse reading frames, respectively, and connected by coloured quadrilaterals to indicate conservation between homologous genes. At this one locus, there have been two incidents of gene loss (*RH7* between *Pf*3D7 and *Pr*CDC; *Rh6* between Pg and Pr) and one incident of gene gain (*Rh2a* in *Pf*). Multiple frameshifts in *Pr* and *Pf* are shown but none could be detected in *Pg*. (**b**) Stylized phylogram showing the relationship between *P. falciparum, P. reichenowi* and *P. gaboni*, indicating the incidents of gene loss and gain within the *EBL* and *Rh* gene families across the species radiation.

for the first time conservation in the domain type ratios and their organization. Our limited data from *P. gaboni* provide the first indication that there might be more sequence variation within the *var* gene family in this organism, although traces of the same conserved short amino acid block clearly remain.

The most notable difference in the gene complement in the subtelomeres is seen between the *rif* and *stevor* multigene families where the numbers are much lower in the human parasite. The *rif* genes have previously been divided into A and B subtypes based on the presence or absence of a 25-codon insert in the second exon[28]. Despite the difference in numbers of *rif* genes, both species contain A and B subtypes with ratios of A and B of 0.6 and 0.4 in *P. reichenowi* and *P. falciparum*, respectively, supporting a gene family split that predates speciation (Fig. 2b). Orthologues of the *rif* and *stevor* families are present in all *Plasmodium* spp. sequenced to date and are known collectively as the *Plasmodium* interspersed repeat (*pir*) family. It is noteworthy that *P. falciparum* has one of the lowest complements of the *pir* family of any species yet sequenced. We believe that this is a genuine difference between species rather than a chance finding due to copy number variation in *P. falciparum*. Although only a single clone of *P. falciparum* is fully sequenced and assembled, draft assemblies are available for a number of other strains (IT, HB3 and Dd2). Despite the fact that the subtelomeres are incompletely assembled in these sequences, searching the assembled contigs for motifs specific to each of these *rif* and *stevor* multigene families results in very similar numbers for all *P. falciparum* strains. A recent contraction of the *rif* and *stevor* families in the human parasite (Fig. 2c) is therefore likely and, unlike other exported multigene families, these two families appear to be under quite different pressures in *P. falciparum* and *P. reichenowi*. Evolution has thus had different effects on subtelomeric multigene families, but critically the founder members and basic organization occurred before the emergence of *P. falciparum*.

With respect to genes in the core region of the genome, we found it striking that among the 100 most divergent genes between species, the majority (77) encode proteins of unknown function. Only 38% of genes in the *P. falciparum* genome have this definition, highlighting that the least explored content of the genome includes genes that may have important roles relating to host differences. This work therefore immediately identifies a number of new genes for which functional work is urgently required. Of the genes with ascribed functions, six are known to be involved in erythrocyte invasion, and tests for genes under selection (MK and HKAr) also identified several genes associated with this essential step in the *Plasmodium* life cycle. These widespread changes in invasion loci is most notable at the *Rh2* locus on chromosome 13 where significant rearrangement and diversification has occurred, not only just between *P. falciparum* and *P. reichenowi* but also between the two chimpanzee parasite species, *P. reichenowi* and *P. gaboni*. Given the otherwise relatively limited changes both in the core genome and in the functional subdivisions within the exported protein families, we hypothesize that changes in the sequence and arrangement of genes in the EBL and RBL family may be directly associated with *Laverania* host adaptation. Support for this hypothesis with respect to the *Rh* genes comes from the recent data that shows that *Rh5*, a member of the RBL family that is essential for erythrocyte invasion in *P. falciparum* and has previously been implicated in host specificity[29,30], is strikingly primate specific in its receptor-binding preferences[31]. Further studies will be needed to directly establish the role that these new RBL family members also have a role in primate-specific erythrocyte recognition, and thus may contribute to *Laverania* host restriction.

The recent discovery of multiple species of *P. falciparum* related parasites in African apes has raised numerous questions about the potential of these *Laverania* parasites to act as zoonotic pathogens of humans. While molecular surveys and epidemiological work will contribute to the answers to these questions[32], a molecular understanding of the factors that restrict *Laverania* parasites to specific hosts is also needed. This first complete genome analysis of a *Laverania* parasite very clearly identifies changes at the host–parasite interface during the blood stages of their life cycle, both during erythrocyte invasion and on the surface of the infected erythrocyte, as well as a short list of genes of unknown function, as potential key contributors to host specificity. The relative contribution of each of these candidates to speciation and host switching awaits further functional study.

Methods

***P. reichenowi* genome sequencing.** The only extant isolate of *P. reichenowi* (*Pr*CDC) was obtained from a chimpanzee imported to the United States of America from the Belgian Congo (now the Democratic Republic of Congo) in the 1950s (ref. 33). Before commencement of the present study, parasites were collected (in 2001), with approval of the Institutional Ethics Committee and according to Dutch law, from a chimpanzee called Dennis that had been infected with *P. reichenowi* from a chimpanzee called Oscar at the Center for Disease Control, Atlanta. Twelve days after infection, parasitemia had reached 0.5%, and

blood was collected and filtered to reduce white blood cell numbers (Plasmodipur, the Netherlands). The chimpanzee was subsequently cured of infection by treatment with chloroquine. DNA was isolated from infected erythrocytes using the PureGene gDNA isolation kit (Gentra Systems, Minneapolis, USA) and subsequently stored at 4°C.

Genomic DNA was sheared into 300–500 base pair (bp) fragments by focused ultrasonication (Covaris Inc., Woburn, USA). Amplification-free Illumina libraries were prepared[34] and 76-bp-end reads were produced on an Illumina Genome Analyzer IIx following the manufacturer's standard cluster generation and sequencing protocols[35]. A 3–4-kb mate-pair library was prepared using the Illumina mate-pair library prep kit (v1) and sequenced as above.

In total, 202 million reads were produced and deposited in the European Nucleotide Archive with the accession number ERP000299.

P. gaboni genome sequencing. During a routine medical investigation, *P. gaboni* was isolated from a young male chimpanzee in Koulamoutou, Ogooué-Lolo Province, Gabon, and conserved in 7 ml EDTA vacutainers. The investigation was approved by the Government of the Republic of Gabon and by the Animal Life Administration of Libreville, Gabon (no. CITES 00956). All animal work was conducted according to relevant national and international guidelines. Red cells and plasma were separated by centrifugation and stored at −20°C until transportation to the Centre International de Recherches Médicales de Franceville, Gabon, where they were stored at −80 °C until processed.

To overcome host contamination, flow cytometry and cell sorting[36] were performed. Two sets of 10,000 parasite nuclei (Gab.1.4. and Gab.1.5.) were obtained. Whole-genome amplification was performed on each set using an Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Uppsala, Sweden), followed by purification through Illustra Sephadex G-50 DNA Grade F columns (GE Healthcare). The two final products were then pooled for subsequent sequencing.

Genomic DNA was sheared into 300–400-bp fragments by focused ultrasonication (Covaris Adaptive Focused Acoustics technology (AFA Inc., Woburn, USA)). A standard Illumina library was then prepared following the manufacturer's protocol[35]. Using an Illumina Genome Analyzer IIx, 76-bp-end reads were initially produced from the library. The depth of coverage was subsequently increased using an Illumina HiSeq 2000 to produce 76-bp-end reads. In total, 447 million raw reads were produced and deposited in the European Nucleotide Archive with the accession number ERP000135.

Limited template coverage of the *P. gaboni* genome precluded a systematic genome-wide analysis of sequence variation, but the limited number of genes for which we could calculate *Ka*/*Ks* ratios are given in Supplementary Data 3.

Assembly of _P. reichenowi_. Before sequence assembly, contaminating host-derived sequences were excluded by mapping to the chimpanzee genome using SSAHA[37] and BWA[38]. Putative sequencing errors in the reads were removed using SGA[39].

A working reference of *P. reichenowi* was produced by iteratively mapping the Illumina reads against the *P. falciparum* 3D7 genome (*Pf*3D7) and changing the sequence of the latter to match apparent fixed differences using iCORN[40]. Against this working reference, *P. reichenowi* reads were remapped using SMALT and a guided assembly was performed using Velvet Columbus[41]. Contigs were joined into scaffolds using SSPACE[42] and paired reads from the PCR-free the amplified 3 kb libraries. Scaffolds ≥500 bp were ordered and oriented against the *Pf*3D7 reference genome using ABACAS[43] and gaps closed by further mapping of Illumina reads using IMAGE[44]. Where contig–contig joins were supported by <3 mapped mate pairs, scaffolds were broken. Iteratively remapping Illumina reads and correcting erroneous bases using iCORN further improved the *P. reichenowi* assembly.

Assembly of _P. gaboni_. The assembly of *P. gaboni* resulted in many fragmented small contigs owing to the low amount of template DNA and high level of host and adapter contamination. Although we tested the same methodology as used with the *Pr*CDC genome (that is, transforming the genome), better results were obtained by assembling the corrected reads *de novo* using Velvet.

The *P. gaboni* assembly is available from: ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/gaboni/Assembly/Version1/.

Assembly of _P. falciparum_ laboratory clones. The genome of *P. falciparum* IT clone was produced by Illumina sequencing from a PCR-free small fragment library together with a 2.5-kb mate-pair library. The data were assembled *de novo*, improved using the PAGIT pipeline[45] and stored in GeneDB.

For the three other laboratory clones (DD2, HB3 and 7G8), read quality was insufficient to produce complete *de novo* assemblies, as many coding regions were not generated. Therefore, a morphing approach was used, whereby reads from the above genomes were iteratively mapped to the genome of 3D7 and SNPs, and small indels incorporated until no further differences could be called[40].

Raw reads of the four isolates are ERS013737 (IT), ERS005005 (DD2), ERS010000 (HB3) and ERS006631 (7G8). The annotated assemblies for the four laboratory clones can be found at the following FTP sites:

ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/falciparum/IT_strain/version_2/June_2012/, ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/falciparum/DD2/Assembly/V1_morphed, ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/falciparum/7G8/Assembly/V1_morphed, ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/falciparum/HB3/Assembly/V1_morphed.

Annotation. From 5,418 genes in *Pf*3D7, 5,111 were transferred to *P. reichenowi* using RATT[46]. In regions with no shared synteny, genes were predicted *ab initio* using Augustus[47], with the transferred gene models as a training set, and manually inspected in Artemis. *P. gaboni* was not annotated owing to its highly fragmented state and was instead used for focused comparisons where contigs were aligned to specific loci of interest. To annotate the laboratory clones, annotation was transferred using RATT and, in the case of *P. falciparum* IT, gene models were systematically corrected manually using Artemis and the Artemis Comparison Tool[48].

The alignments for genes with extreme HKAr or MK values (below) were manually checked, and in some cases resulted in additional edits to gene models to update their structures in accordance with the June 2012 annotation.

Putative exported proteins were predicted using Exportpred v 2.0 (ref. 49).

Analysis of gene family trees. For each gene family (*Surfin, PHIST, FIKK, rif* and *stevor*), we constructed phylogenetic trees. Multiple alignments of each family were produced with Clustal W[50]. For the Rifin and Stevor families, only full-length uninterrupted translations were used. Four *P. reichenowi* Rifins were excluded from analysis owing to internal gaps within their sequences, and in both *P. reichenowi* (PRCDC_1038200 and PRCDC_0004700) and *P. falciparum* (PF3D7_0401600) contain divergent Rifin sequences that align poorly to either type A or type B Rifins. For other families, protein sequences that differed in length by >15% from the median were excluded, along with pseudogenes. From the obtained alignment, trees were calculated with PhyML version 3.0.1 (ref. 51) (LG model, aLRT branch support; optimized 'Across site rate variation' NNI Tree searching operation; BioNJ Starting tree (enabling optimized tree topology)). For colouring, we used a PERL script and FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Analysis of _var_ genes. The *var* genes in *P. reichenowi* were annotated during the automatic functional annotation. In general, the *var* genes of *Pr*CDC are longer than those of *Pf*3D7 but not significantly different to *P. falciparum* when other isolates are taken into account. Not all of them are completely assembled as many are missing the second, highly conserved, exon (Supplementary Data 2). Two *var* genes that are known to be more conserved in *P. falciparum* isolates than the rest of the repertoire—*var1csa* and *var2csa* (the latter known to be associated with malaria in pregnancy[52])—were also present but differed in length. Compared with its *Pf*3D7 orthologue, the *var2csa* gene of *Pr*CDC is ∼5 kb shorter and runs into telomeric repeats despite being located in an interstitial region. *Pr*VAR1CSA is 1,584 bp longer than its *Pf*3D7 orthologous.

The similarity between *Pf*3D7 and *Pr*CDC also includes the arrangement of the internal *var* gene clusters on chromosomes 4, 7, 8 and 12. The internal *var* gene locus on Chr 6 in 3D7 (*var* gene PF3D7_0617400) is, however, not present in *Pr*CDC. On chromosome 3, the *var* gene locus that is a pseudogene in 3D7 (PF3D7_0302300) is intact in *Pr*CDC.

The *var*-encoded protein domains were classified using the VARdom server[53] and the output parsed with a bespoke PERL script. DBLα domains were classified according to ref. 54, using a bespoke PERL script (Supplementary Data 2).

To identify *var* genes in *P. gaboni*, we also searched the *P. gaboni* assembly for matches to the *P. falciparum var* genes using BLASTX (E-value cutoff of 10^{-6}). All contigs that hit a 3D7 *var* gene were extracted and open frames ≥1,500 bp were identified. This resulted in 15 *var* gene fragments, with a mean length of 973 amino acids (largest 3130 amino acids). The largest *var* gene had as first hit a weak similarity to the *var1csa* of *P. falciparum* IT and 3D7.

The putative *var* genes of *P. gaboni* were aligned with those of *Pr*CDC and *Pf*3D7 using Clustal W. Interestingly, the conserved LARSFADIGDI motif from the latter two species, which is present in the first DBL domain and has been universally used for the design of PCR primers, appears to be different and degenerate in *P. gaboni* (A[LIM][KR][YN]SF[A/Y]DIGDI) (Supplementary Fig. 6).

Classifying the *P. gaboni var* genes with the VARdom server returned 8, 3, 1 and 1 hits to the DBLe, DBLz, DBLb and DBLg domains, respectively. As those results are different to the obtained alignment, we assume that the domains are too divergent to be classified with the VARdom server.

Polymorphism versus divergence tests. Amino acid repeats and low complexity regions in genes were identified by blasting each protein against itself. Any amino acid repeats with an identity of at least 95% and longer than 20 amino acids were masked for subsequent analysis. To mask low complexity regions, we used the SEG programme (from NCBI BLAST).

A multiple amino acid alignment was generated with MUSCLE[55] for each orthologous group (defined through the RATT transfer) of *P. reichenowi* and the five *P. falciparum* lab clones: 3D7, HB3, DD2, IT and 7G8. From the amino acid alignment, we generated a nucleotide alignment. We excluded orthologous clusters if one gene was a pseudogene (contains an internal stop codon or frameshift).

Alignment positions with gaps were excluded, as well as masked regions. From the resulting alignments, we calculated HKA[16], MK[18] and *Ka*/*Ks*[56] values (Supplementary Data 4).

HKA values (Supplementary Data 4–6) were calculated by counting the proportion of pairwise differences in the intraspecific samples (laboratory clones) and the interspecific comparison, averaging across all pairwise intraspecific comparisons to get the overall nucleotide diversity π, and taking the average of the comparison of each of the *P. falciparum* sequences versus PrCDC to get the nucleotide divergence *K*. The HKAr is the π/K ratio. Intergenic regions were also analysed between orthologous genes that shared gene order. These regions were identified by the gene on their left hand side (Supplementary Data 5). Smith–Waterman alignments were performed using SSEARCH[57] and indels or gap positions were deleted.

For the MK test (Supplementary Data 4 and 7), the number of fixed and polymorphic positions was generated using previously described software[58] that also provides a *P* value. The MK skew is the $\log_2((N_{poly}/S_{poly})/(N_{fix}/S_{fix}))$, where *N* and *S* are the number of non-synonymous and synonymous sites, respectively, and fix and poly refer to fixed differences and polymorphisms, respectively; Supplementary Data 7).

To calculate the *Ka*/*Ks* ratio (Supplementary Data 4 and 8), we took the cleaned alignments of the MK test, extracting the sequences of *Pf*3D7 and *Pr*CDC. A PERL script used the Bio::Align::DNAStatistics module to calculate the *Ka*/*Ks* values[56].

Erythrocyte invasion genes. Loci containing erythrocyte invasion genes were manually inspected using the Artemis Comparison Tool (Fig. 4)[48]. By inspecting mapped 3 kb mate pairs, the orientation of the Rh2 locus in *Pr*CDC was confirmed; with the locus in its new orientation, 3-kb mate pairs mapped evenly across the locus boundaries, whereas no 3-kb mate could be mapped when the locus was flipped to match the orientation in *Pf*3D7.

References

1. Reichenow, E. Uber das Vorkommen der Malariaparasiten des Menschen bei den Afrikanischen Menschenaffen. *Centralbl. f Bakt. I Abt. Orig.* **85,** 9 (1920).
2. Duval, L. *et al.* African apes as reservoirs of Plasmodium falciparum and the origin and diversification of the Laverania subgenus. *Proc. Natl Acad. Sci. USA* **107,** 10561–10566 (2010).
3. Liu, W. *et al.* Origin of the human malaria parasite Plasmodium falciparum in gorillas. *Nature* **467,** 420–425 (2010).
4. Ollomo, B. *et al.* A new malaria agent in African hominids. *PLoS Pathog.* **5,** e1000446 (2009).
5. Prugnolle, F. *et al.* African great apes are natural hosts of multiple related malaria species, including Plasmodium falciparum. *Proc. Natl Acad. Sci. USA* **107,** 1458–1463 (2010).
6. Rich, S. M. *et al.* The origin of malignant malaria. *Proc. Natl Acad. Sci. USA* **106,** 14902–14907 (2009).
7. Jeffares, D. C. *et al.* Genome variation and evolution of the malaria parasite Plasmodium falciparum. *Nat. Genet.* **39,** 120–125 (2007).
8. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419,** 498–511 (2002).
9. Maier, A. G., Cooke, B. M., Cowman, A. F. & Tilley, L. Malaria parasite proteins that remodel the host erythrocyte. *Nat. Rev. Microbiol.* **7,** 341–354 (2009).
10. Baruch, D. I. *et al.* Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* **82,** 77–87 (1995).
11. Smith, J. D. *et al.* Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82,** 101–110 (1995).
12. Su, X. Z. *et al.* The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. *Cell* **82,** 89–100 (1995).
13. Smith, J. D., Subramanian, G., Gamain, B., Baruch, D. I. & Miller, L. H. Classification of adhesive domains in the Plasmodium falciparum erythrocyte membrane protein 1 family. *Mol. Biochem. Parasitol.* **110,** 293–310 (2000).
14. Cheng, Q. *et al.* stevor and rif are Plasmodium falciparum multicopy gene families which potentially encode variant antigens. *Mol. Biochem. Parasitol.* **97,** 161–176 (1998).
15. Kyes, S. A., Rowe, J. A., Kriek, N. & Newbold, C. I. Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with Plasmodium falciparum. *Proc. Natl Acad. Sci. USA* **96,** 9333–9338 (1999).
16. Innan, H. Modified Hudson-Kreitman-Aguade test and two-dimensional evaluation of neutrality tests. *Genetics* **173,** 1725–1733 (2006).
17. Hu, G. *et al.* Transcriptional profiling of growth perturbations of the human malaria parasite Plasmodium falciparum. *Nat. Biotechnol.* **28,** 91–98 (2010).
18. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351,** 652–654 (1991).
19. Cowman, A. F., Berry, D. & Baum, J. The cellular and molecular basis for malaria parasite invasion of the human red blood cell. *J. Cell Biol.* **198,** 961–971 (2012).
20. Rayner, J. C., Huber, C. S. & Barnwell, J. W. Conservation and divergence in erythrocyte invasion ligands: Plasmodium reichenowi EBL genes. *Mol. Biochem. Parasitol.* **138,** 243–247 (2004).
21. Rayner, J. C., Vargas-Serrato, E., Huber, C. S., Galinski, M. R. & Barnwell, J. W. A Plasmodium falciparum homologue of Plasmodium vivax reticulocyte binding protein (PvRBP1) defines a trypsin-resistant erythrocyte invasion pathway. *J. Exp. Med.* **194,** 1571–1581 (2001).
22. Rayner, J. C., Huber, C. S., Galinski, M. R. & Barnwell, J. W. Rapid evolution of an erythrocyte invasion gene family: the Plasmodium reichenowi Reticulocyte Binding Like (RBL) genes. *Mol. Biochem. Parasitol.* **133,** 287–296 (2004).
23. Taylor, H. M. *et al.* Plasmodium falciparum homologue of the genes for Plasmodium vivax and Plasmodium yoelii adhesive proteins, which is transcribed but not translated. *Infect. Immun.* **69,** 3635–3645 (2001).
24. Dvorin, J. D., Bei, A. K., Coleman, B. I. & Duraisingh, M. T. Functional diversification between two related Plasmodium falciparum merozoite invasion ligands is determined by changes in the cytoplasmic domain. *Mol. Microbiol.* **75,** 990–1006 (2010).
25. Rayner, J. C., Galinski, M. R., Ingravallo, P. & Barnwell, J. W. Two Plasmodium falciparum genes express merozoite proteins that are related to Plasmodium vivax and Plasmodium yoelii adhesive proteins involved in host cell selection and invasion. *Proc. Natl Acad. Sci. USA* **97,** 9648–9653 (2000).
26. Triglia, T. *et al.* Identification of proteins from Plasmodium falciparum that are homologous to reticulocyte binding proteins in Plasmodium vivax. *Infect. Immun.* **69,** 1084–1092 (2001).
27. Zilversmit, M. M. *et al.* Hypervariable antigen genes in malaria have ancient roots. *BMC. Evol. Biol.* **13,** 110 (2013).
28. Joannin, N., Abhiman, S., Sonnhammer, E. L. & Wahlgren, M. Sub-grouping and sub-functionalization of the RIFIN multi-copy protein family. *BMC Genomics* **9,** 19 (2008).
29. Hayton, K. *et al.* Erythrocyte binding protein PfRH5 polymorphisms determine species-specific pathways of Plasmodium falciparum invasion. *Cell Host Microbe* **4,** 40–51 (2008).
30. Crosnier, C. *et al.* Basigin is a receptor essential for erythrocyte invasion by Plasmodium falciparum. *Nature* **480,** 534–537 (2011).
31. Wanaguru, M., Liu, W., Hahn, B. H., Rayner, J. C. & Wright, G. J. RH5-Basigin interaction plays a major role in the host tropism of Plasmodium falciparum. *Proc. Natl Acad. Sci. USA* **110,** 20735–20740 (2013).
32. Sundararaman, S. A. *et al.* Plasmodium falciparum-like parasites infecting wild apes in southern Cameroon do not represent a recurrent source of human malaria. *Proc. Natl Acad. Sci. USA* **110,** 7020–7025 (2013).
33. Collins, W. E., Skinner, J. C., Pappaioanou, M., Broderson, J. R. & Mehaffey, P. The sporogonic cycle of Plasmodium reichenowi. *J. Parasitol.* **72,** 292–298 (1986).
34. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G + C)-biased genomes. *Nat. Methods* **6,** 291–295 (2009).
35. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
36. Boissiere, A. *et al.* Isolation of Plasmodium falciparum by flow-cytometry: implications for single-trophozoite genotyping and parasite DNA purification for whole-genome high-throughput sequencing of archival samples. *Malar. J.* **11,** 163 (2012).
37. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11,** 1725–1729 (2001).
38. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
39. Simpson, J. T. & Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* **22,** 549–556 (2012).
40. Otto, T. D., Sanders, M., Berriman, M. & Newbold, C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26,** 1704–1707 (2010).
41. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
42. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578–579 (2011).
43. Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: Algorithm Based Automatic Contiguation of Assembled Sequences. *Bioinformatics* **25,** 2 (2009).
44. Tsai, I. J., Otto, T. D. & Berriman, M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome. Biol.* **11,** R41 (2010).
45. Swain, M. T. *et al.* A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat. Protoc.* **7,** 1260–1284 (2012).
46. Otto, T. D., Dillon, G. P., Degrave, W. S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.* **39,** e57 (2011).
47. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34,** W435–W439 (2006).

48. Carver, T. *et al.* Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24,** 2672–2676 (2008).
49. Boddey, J. A. *et al.* Role of plasmepsin V in export of diverse protein families from the Plasmodium falciparum exportome. *Traffic* **14,** 532–550 (2013).
50. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
51. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).
52. Buffet, P. A. *et al.* Plasmodium falciparum domain mediating adhesion to chondroitin sulfate A: a receptor for human placental infection. *Proc. Natl Acad. Sci. USA* **96,** 12743–12748 (1999).
53. Rask, T. S., Hansen, D. A., Theander, T. G., Gorm Pedersen, A. & Lavstsen, T. Plasmodium falciparum erythrocyte membrane protein 1 diversity in seven genomes--divide and conquer. *PLoS Comput. Biol.* **6** pii e1000933 (2010).
54. Bull, P. C. *et al.* An approach to classifying sequence tags sampled from Plasmodium falciparum var genes. *Mol. Biochem. Parasitol.* **154,** 98–102 (2007).
55. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
56. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3,** 418–426 (1986).
57. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85,** 2444–2448 (1988).
58. Holloway, A. K., Lawniczak, M. K., Mezey, J. G., Begun, D. J. & Jones, C. D. Adaptive gene expression divergence inferred from population genomics. *PLoS Genet.* **3,** 2007–2013 (2007).

Acknowledgements

This work was supported by the Wellcome Trust (grant number WT 098051) with additional funding to T.D.O. from the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement number 242095; J.C.R., from the National Institutes of Health (R01 AI091595); B.O. from Centre International de Recherches Médicales de Franceville; F.R. and F.P., from CNRS and IRD; F.R., B.O. and F.P. from the Agence Nationale de la Recherche (grant ANR JCJC SVSE 7-2012 ORIGIN); D.J.C. from an ERC Advanced Award (grant number 294428); and C.N. from the Wellcome Trust (grant number WT 082130/Z/07/Z).

Author contributions

T.D.O. carried out the sequence assembly and data analysis; N.S. and A.P. performed preliminary gene annotation and analysis, which were manually updated, checked, and curated by UCB; M.S. coordinated sequencing; M.Q. directed production of sequencing libraries; B.O., F.R. and F.P. sorted the *P. gaboni* nuclei and provided whole-genome-amplified DNA; A.W.T. provided *P. reichenowi* DNA; J.C.R. coordinated the analysis of invasion-associated genes; D.J.C. coordinated the analysis of signatures of selection; C.N., D.J.C., J.C.R., M.B. and T.D.O. drafted the manuscript; A.P., C.N., D.J.C., F.P., J.C.R., M.B., N.S. and T.D.O. critically revised the manuscript for intellectual content; and C.N. and M.B. directed the study.

Additional information

Accession codes: *P. reichenowi* and *P. gaboni* sequence data have been deposited in the European Nucleotide Archive (ENA) under the accession codes ERP000299 and ERP000135, respectively. Raw reads of *P. falciparum* IT, Dd2, HB3 and 7G8 have been deposited in the ENA under the accession codes ERS013737, ERS005005, ERS010000 and ERS006631, respectively. The assembled *contigs* are available under the following accession codes: CBXM010000001 to CBXM010002055 (for *P. reichenowi*) and CBUG010000001 to CBUG010025756 (for *P. gaboni*). The assembled and annotated genome scaffolds of *P. reichenowi* have been deposited in EMBL under the accession codes HG810406 to HG810777.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Otto, T. D. *et al.* Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat. Commun.* 5:4754 doi: 10.1038/ncomms5754 (2014).