nature genetics

Genome variation and evolution of the malaria parasite *Plasmodium falciparum*

Daniel C Jeffares[1], Arnab Pain[2], Andrew Berry[2], Anthony V Cox[1], James Stalker[1], Catherine E Ingle[1], Alan Thomas[3], Michael A Quail[2], Kyle Siebenthall[1,4], Anne-Catrin Uhlemann[5], Sue Kyes[6], Sanjeev Krishna[5], Chris Newbold[6], Emmanouil T Dermitzakis[1] & Matthew Berriman[2]

Infections with the malaria parasite *Plasmodium falciparum* result in more than 1 million deaths each year worldwide[1]. Deciphering the evolutionary history and genetic variation of *P. falciparum* is critical for understanding the evolution of drug resistance, identifying potential vaccine candidates and appreciating the effect of parasite variation on prevalence and severity of malaria in humans. Most studies of natural variation in *P. falciparum* have been either in depth over small genomic regions (up to the size of a small chromosome[2]) or genome wide but only at low resolution[3]. In an effort to complement these studies with genome-wide data, we undertook shotgun sequencing of a Ghanaian clinical isolate (with fivefold coverage), the IT laboratory isolate (with onefold coverage) and the chimpanzee parasite *P. reichenowi* (with twofold coverage). We compared these sequences with the fully sequenced *P. falciparum* 3D7 isolate genome[4]. We describe the most salient features of *P. falciparum* polymorphism and adaptive evolution with relation to gene function, transcript and protein expression and cellular localization. This analysis uncovers the primary evolutionary changes that have occurred since the *P. falciparum*–*P. reichenowi* speciation and changes that are occurring within *P. falciparum*.

In this study, we compare the completely sequenced genome of the *P. falciparum* laboratory-cultured clone 3D7 (ref. 4) with whole-genome sequence data from (i) the chimpanzee malaria parasite *P. reichenowi*, the closest relative of *P. falciparum*; (ii) a non-cultured clinical *P. falciparum* isolate, directly isolated from an individual infected in Ghana (PFCLIN) and (iii) a *P. falciparum* laboratory model, clone IT. Sequence data were produced by whole-genome shotgun sequencing, with fivefold, twofold and onefold coverage for PFCLIN, *P. reichenowi* and IT, respectively. We used SSAHA2 (ref. 5) to align sequence reads to the reference *P. falciparum* genome 3D7. The fractions of the genome that could be reliably analyzed (see Methods) were 74% for PFCLIN, 25% for IT and 42% for *P. reichenowi* (excluding the majority of *var* genes). We identified single-nucleotide differences (SNPs or fixed substitutions) and indels from these alignments based on standard calling strategies and additional filters. We identified 27,169 nonredundant SNPs between 3D7, PFCLIN and IT (nucleotide diversity: 3D7 versus PFCLIN, $\pi = 0.00131$; 3D7 versus IT, $\pi = 0.0011$). We called 216,619 fixed differences between 3D7 and *P. reichenowi* (divergence $d = 0.0203$). Based on coverage by multiple reads, we estimate the false discovery rate for both SNPs and fixed difference calls to be 10% (that is, 90% of the calls are correct (see Methods)). We also identified many insertion-deletion events (indels) (**Table 1**). These data indicate that there is substantial variation in *P. falciparum* genomes and that the divergence between *P. reichenowi* and *P. falciparum* is approximately tenfold greater than within-species polymorphism.

In regions where the prevalence of malaria is high, many individuals harbor multiple *P. falciparum* genotypes that survive at low parasitemia during asymptomatic infection[6]. Clinical episodes of malaria are caused by a new infection, the majority of which are of a single genotype. To estimate the multiplicity of infection in the individual from whom PFCLIN was isolated, we estimated heterozygosity based on overlapping high-quality reads. Approximately 7%–10% of the SNP sites called between 3D7 and PFCLIN have good support for both allelic states in PFCLIN, whereas this is only about 5% for the IT, which is expected to be monomorphic. By subtracting the noise inferred from IT, we estimate that about 2%–5% of the PFCLIN sample (genome-wide) corresponds to other genotypes.

Culture-adapted *P. falciparum* strains may be imperfect models for the genome owing to extensive accumulation or loss of sequences in culture. We investigated whether this had occurred in the culture-adapted 3D7 isolate by examining the indel events we detected in the PFCLIN and *P. reichenowi* comparisons, neither of which has been adapted to *in vitro* culture. Large numbers of small indels were identified in this analysis (**Table 1**). In both comparisons we observed

[1]Informatics Division and [2]Pathogen Sequencing Unit, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, CB10 1SA Hinxton, UK. [3]Biomedical Primate Research Centre, Lange Kleiweg 139, Rijswijk, Postbus 3306, 2280 GH Rijswijk, The Netherlands. [4]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA. [5]St. George's, University of London, Cranmer Terrace, London SW17 ORE, UK. [6]The Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK. Correspondence should be addressed to E.T.D. (md4@sanger.ac.uk) or M.B. (mb4@sanger.ac.uk).

Received 6 June; accepted 2 November; published online 10 December 2006; corrected after print 8 February 2007; doi:10.1038/ng1931

Table 1 Comparative genome data

Alignment and SNP data

	PFCLIN	IT	*P. reichenowi*
Reads aligned	121,810	17,753	33,699
Unique coverage (percentage of reference)	17,336,871 (74%)	5,754,163 (25%)	9,818,236 (42%)
SNPs[a]	22,735	6,415	216,619
Indels[a]	27,478	6,461	37,258
Complex differences[a]	3,736	901	19,964

Rates of SNPs and fixed differences[b]

	PFCLIN	IT	*P. reichenowi*
RNA genes	4 (0.75)	2 (0.52)	72 (14.44)
Intergenic regions	7,327 (1.21)	1,818 (0.96)	66,687 (20.28)
Exons	12,730 (1.27)	3,993 (1.16)	132,509 (23.59)
Synonymous sites	3,659 (2.21)	1,243 (2.14)	59,907 (71.11)
Nonsynonymous sites	9,071 (1.08)	2,750 (0.96)	72,602 (15.21)
dN/dS[c]	0.49	0.45	0.21
FFD sites[d]	1031 (1.49)	321 (1.22)	22,160 (55.20)
Introns	2,678 (2.49)	603 (1.76)	17,421 (26.96)
Repeats	1,918 (4.40)	527 (3.82)	6,489 (26.57)
Total	22,735 (1.31)	6,415 (1.11)	216,619 (22.06)

[a]'SNPs' refers to the number of SNPs or fixed differences (derived from SSAHA SNP and multiple SNP POLY calls; see Methods), 'indels' are insertion-deletion events and 'complex differences' are SSAHA-identified differences that include both single-nucleotide differences and indels. Complex differences were not included in our analysis. [b]Number of total differences includes SNP and multiple-SNP POLYs (see Methods). Data represent the total number of differences (rate/kb of read coverage). [c]Genome-wide dN/dS from all complete codons with read coverage. [d]Fourfold degenerate (FFD) sites. Rates of fixed differences and SNPs were higher in G/C FFD sites than A/T FFD sites (PFCLIN rates: A/T 0.91/kb, G/C 2.82/kb; *P. reichenowi* rates: A/T 34/kb, G/C 137/kb).

an excess of 'insertions' in 3D7 (ancestral state not known), indicating that the 3D7 isolate had accumulated DNA in culture at the rate of approximately +1 nt/kb (**Supplementary Methods** online). Indels were underrepresented in coding sequences (0.07 indels/kb, compared with 3.75 indels/kb in other regions), without any bias toward gain or loss of DNA relative to the reference 3D7 and frequently in multiples of three, thus maintaining ORFs. In this study, we excluded much of the subtelomeric regions that have been the focus of previous observations of sequence loss and gain[7,8], and we did not detect any indels longer than ~100 nt owing to the short length of read alignments. Therefore, it is possible that accumulation of sequences in interstitial regions occurs simultaneously with loss and rearrangement events toward the chromosome ends[7,8].

To estimate the selective effects acting on individual genes, we calculated the ratio of the nonsynonymous substitution rate to the synonymous substitution rate (dN/dS) for the three pairwise comparisons for each gene with at least 100 codons of read coverage (n = 4,686, 2,437 and 3,674 genes for PFCLIN, IT and *P. reichenowi*, respectively). This produced 762, 207 and 3,024 genes for which dN/dS estimates could be calculated for PFCLIN, IT and *P. reichenowi*, respectively. All dN/dS distributions were skewed toward 0 (median = 0.27, 0.30 and 0.17, for PFCLIN, IT and *P. reichenowi*), as expected, as most genes are under purifying selection (**Supplementary Fig. 1** online). However, some genes showed elevated dN/dS values (**Supplementary Table 1** online), which in some cases were >1. Many genes that are under positive selection may have dN/dS values <1 if only a fraction of the amino acids are under positive selection[9]. We also performed McDonald-Kreitman tests[10] using coding sequences with read coverage for all three *P. falciparum* isolates (3D7, PFCLIN and IT) and *P. reichenowi* (see Methods). McDonald-Kreitman tests indicated categories of genes that showed either an excess of divergence (probably explained by directional selection) or an excess of polymorphism (probably diversifying selection) (**Supplementary Table 1**) in the isolates we examined.

Studies of other organisms have shown the evolutionary rates (of protein-coding genes, dN/dS) are often correlated with mRNA and protein expression levels[11]. We used *P. falciparum* expression studies to look for similar correlations. Both total protein mass spectrometry counts (corrected for protein length (see Methods))[12] and total transcript levels[13] were highly significantly correlated with *P. reichenowi* and PFCLIN dN/dS estimates (**Fig. 1a,b** and **Supplementary Fig. 2** online). In all cases, rapidly evolving genes were expressed at low levels, and abundantly expressed genes were conserved during

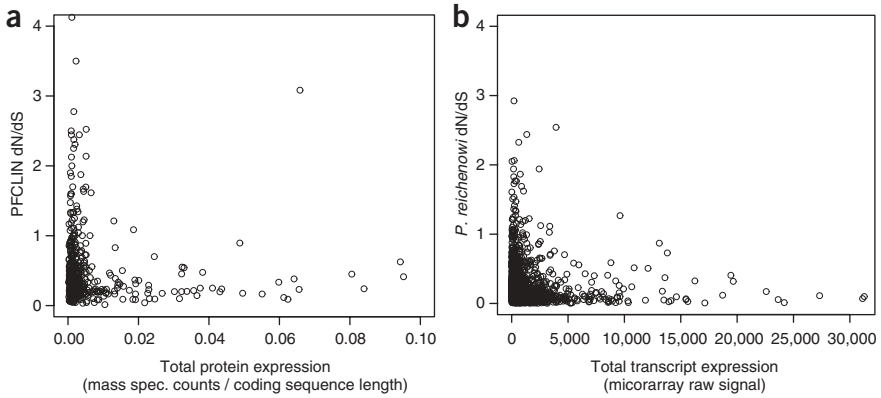

Figure 1 Evolutionary rates correlate with total gene expression levels. (**a**) Scatter plot of PFCLIN dN/dS and total protein expression. Because longer proteins are expected to be detected more frequently in mass spectrometry analysis[12], protein amount (here, measured by mass spectrometry) is corrected for protein length by dividing by predicted coding sequence length. The PFCLIN dN/dS correlates with corrected total protein levels from all developmental stages[12] (Spearman rank correlation, $r = -0.16$, $P = 5.6 \times 10^{-4}$). Some data points with $x > 0.1$ are not shown on this plot. PFCLIN dN/dS estimates were also correlated with total RNA expression levels (**Supplementary Fig. 2**). (**b**) Scatter plot of *P. reichenowi* dN/dS and total transcript expression. The *P. reichenowi* dN/dS correlates with total transcript from all developmental stages[13] (Spearman rank correlation, $r = -0.22$, $P < 2 \times 10^{-16}$). One data point with total transcript >30,000 is not shown on this plot. *P. reichenowi* dN/dS estimates were also correlated with total protein levels (**Supplementary Fig. 2**).

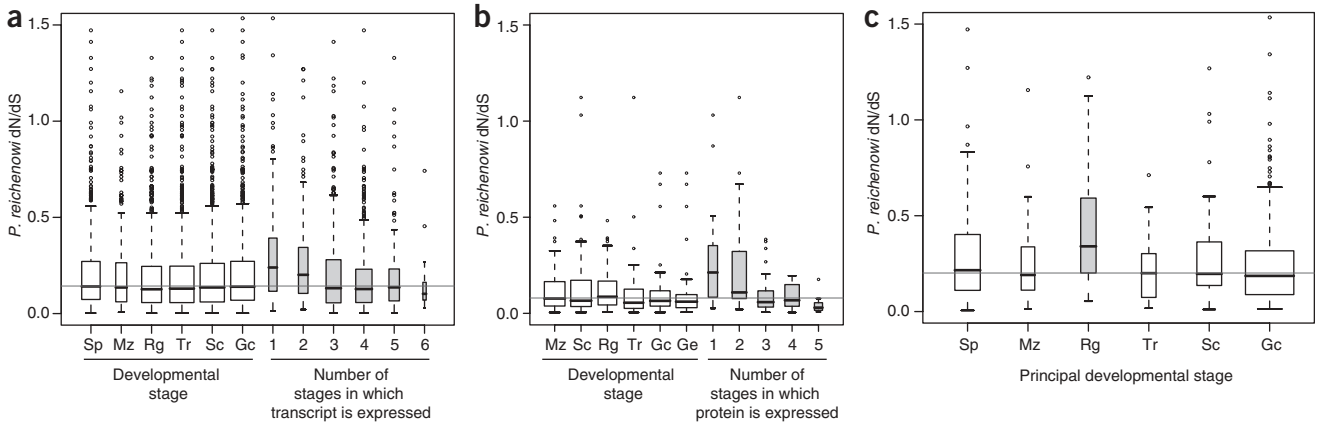

Figure 2 Evolutionary rates in the context of gene expression during development. *P. reichenowi* dN/dS distributions of genes grouped by developmental expression. Gray horizontal lines indicate the median dN/dS of all plotted genes. In boxes, bold line represents median, and upper and lower boundaries represent 75th and 25th percentiles, respectively. Box widths are proportional to group size. Developmental stages include sporozoite (Sp), merozoite (Mz), schizont (Sc), ring stage (Rg), trophozoite (Tr), gametocyte (Gc) and gamete stage (Ge). (**a**,**b**) Developmental expression profiles using transcript expression data[13] (**a**) or protein expression data[13] (**b**). We grouped genes by developmental stage (left) and by the number of developmental stages (one to six) in which they are expressed (right). We defined a gene as 'present' in a given stage if ≥10% of the gene's expression occurred in that stage; see **Supplementary Methods** for details. dN/dS distributions of the groups of genes grouped by developmental stage did not differ significantly from one another (Kruskal-Wallis test: transcript, $P = 0.11$; protein, $P = 0.35$) (open boxes). However, the dN/dS distributions of genes grouped by the number of stages in which they were present did differ significantly (Kruskal-Wallis test: transcript, $P = 5.05 \times 10^{-15}$; protein, $P = 2.6 \times 10^{-4}$) (filled gray boxes). (**c**) Principal developmental stage (transcript data). We grouped genes that were primarily present in one developmental stage (meaning that ≥50% of the gene's expression was in one stage; data from ref. 13). Ring-stage transcripts have the highest *P. reichenowi* dN/dS distribution (Mann-Whitney test, $P = 2.2 \times 10^{-3}$). We observed consistent relative differences in evolutionary rates between stage-specific genes using protein data from ref. 13 and using data from another microarray study[30] (**Supplementary Fig. 3** online).

evolution. The duration of gene expression (measured by the number of developmental stages in which a gene is expressed (data from ref. 13)) was also negatively correlated with the 3D7-*P. reichenowi* dN/dS values (Spearman rank correlations, protein $r = -0.24$, $P < 2.2 \times 10^{-16}$, transcript $r = -0.22$, $P < 2.2 \times 10^{-16}$) (**Fig. 2a**,**b**). These observations were very robust even after correction for total expression effects (**Supplementary Methods**) or with the use of different expression and genetic variation data sets (**Supplementary Fig. 2**). Although the incomplete coverage and relatively high SNP false detection rate preclude detailed analysis of individual genes, these results indicate that global analysis of our dN/dS distributions provides biologically meaningful information. We also observed heterogeneity between dN/dS estimates of genes that were primarily expressed primarily in a single developmental stage of the *Plasmodium* life cycle (**Fig. 2c**). Therefore, highly specialized genes in the *Plasmodium* genome are more likely to undergo accelerated selective processes either because of relaxed constraint or because of directional selection.

The events that occur between the invasion of the host erythrocyte and its eventual rupture and release of daughter merozoites determine the clinical manifestations of malaria. We used a detailed study of the intraerythrocytic developmental cycle (IDC)[14] to examine this stage in more detail. In that study, it was found that ~80% of the nuclear-encoded IDC-expressed genes could be clustered into 13 groups that were both coexpressed and were functionally related. When we compared the dN/dS values split in these clusters, we found that two of these clusters, 'merozoite invasion' and 'early ring transcripts', showed significantly elevated rates of nonsynonymous change (high dN/dS) in the 3D7–*P. reichenowi* comparison (**Fig. 3a**). McDonald-Kreitman tests indicated that both these clusters contained a significant excess of nonsynonymous polymorphisms in *P. falciparum* relative to fixed differences (merozoite neutrality index (NI) = 2.5, Fisher's exact test $P = 0.009$; ring NI = 10.4, $P = 1.59 \times 10^{-9}$), indicating that these proteins were also highly variable within *P. falciparum* isolates.

To test whether proteins that were directly interacting with host cells were under adaptive evolution, we examined the evolutionary rates of genes grouped by their intracellular localization using the cellular component aspect of the Gene Ontology classification[15] and definitions of predicted exported proteins[16]. Proteins localized to the nucleus, cytoplasm and the mitochondrion were generally conserved, whereas apicoplast-localized proteins, predicted membrane-spanning proteins and predicted exported proteins had evolved significantly more rapidly (**Fig. 3b**). Exported and transmembrane proteins also show elevated evolutionary rates in rodent parasites[17]. These observations, and previous studies indicating that the strongest evidence for positive selection between human and chimpanzee is related to immunity and defense[18], are consistent with the model of an 'evolutionary arms race' between the mammalian immune system and the exposed proteins of *Plasmodium* parasites. As expected given this 'arms race' model, genes annotated as 'antigens' have significantly higher dN/dS distributions from the *P. reichenowi* comparison (Mann-Whitney test, $P = 2.67 \times 10^{-3}$, antigen median = 0.29, median for all others = 0.17).

We also analyzed the evolutionary rates of genes with respect to their Gene Ontology biological process annotations. We mapped 2,098 genes with detailed Gene Ontology annotation to a reduced version of the full ontology (GO-slim) with 19 broad categories (**Supplementary Table 1**). Of the 19 categories, 7 did not have *P. reichenowi*–3D7 dN/dS values significantly different from other Gene Ontology–defined genes; 10 had significantly lower dN/dS values, mainly involved in metabolism and cell cycle (data not shown), and 2 GO-slim categories ('cell communication' and 'entry into host cell') had significantly higher dN/dS estimates (**Fig. 3c**). The 'cell communication' category

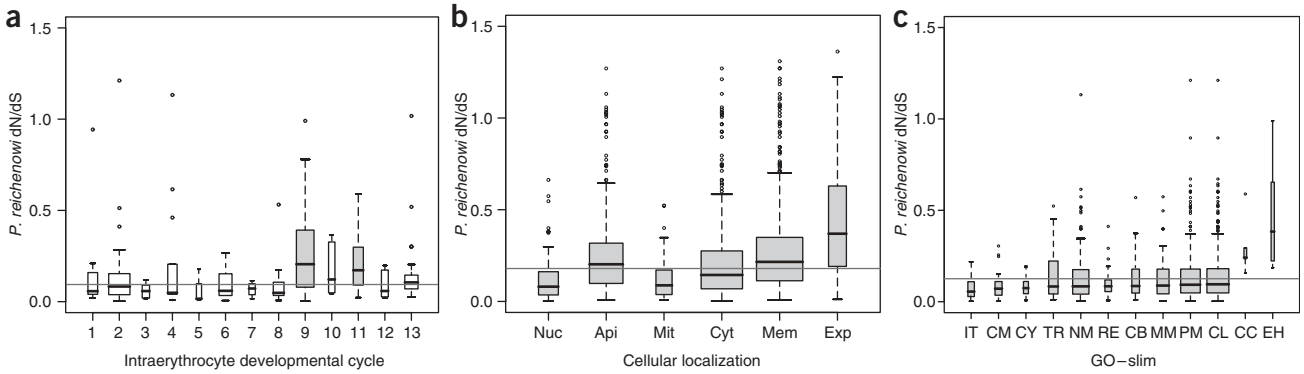

Figure 3 Evolutionary rate in the context of gene function. *P. reichenowi* dN/dS distributions of genes grouped by protein function. Plots are as in **Figure 2**. (**a**) Intraerythrocyte developmental cycle (IDC). Temporal and functional IDC groups[15] are transcription (1), cytoplasmic translation (2), glycolysis (3), ribonucleotide synthesis (4), deoxyribonucleotide synthesis (5), DNA replication (6), tricarboxylic acid cycle (7), proteosome (8), merozoite invasion (9), actin-myosin motors (10), early ring transcripts (11), mitochondria (12) and organelle translation (13). Merozoite invasion (9) and early ring cluster (11) showed significantly elevated dN/dS distributions (Mann-Whitney test, $P = 1.7 \times 10^{-5}$ and $P = 0.024$). (**b**) Cellular localization. We grouped genes according to Gene Ontology localization and predicted export motif[17]. All groups had significantly different dN/dS distributions. Categories (and Mann-Whitney test *P* values) are nucleus (Nuc, $P < 2.2 \times 10^{-16}$), apicoplast (Api, $P = 0.013$), mitochondria (Mit, $P = 1.238 \times 10^{-9}$), cytoplasm (Cyt, $P = 7.283 \times 10^{-8}$), membrane-spanning (Mem, $P < 2.2 \times 10^{-16}$) and exported (Exp, $P = 1.776 \times 10^{-15}$). (**c**) GO-slim categories with significantly different dN/dS distributions (compared with all genes not in the given category) are shown. Categories showing significantly lower dN/dS distributions were intracellular transport (IT), carbohydrate metabolism (CM); cell cycle (CY); transport (TR); nucleobase, nucleoside, nucleotide and nucleic acid metabolism (NM); regulation of cellular physiological process (RE); cell organization and biogenesis (CB); macromolecule metabolism (MM); protein metabolism (PM) and cellular metabolism (CL). Cell communication (CC) and entry into host cell (EH) had significantly higher dN/dS distributions (Mann-Whitney test, $P = 2.0 \times 10^{-3}$ and $P = 0.043$).

also showed a significant excess of nonsynonymous polymorphisms with the McDonald-Kreitman test (NI = 2.8, Fisher's exact test, $P = 7.1 \times 10^{-3}$).

Collectively, our evolutionary analysis indicates that the most significant functional changes between *P. falciparum* and *P. reichenowi* have been to membrane and exported proteins and proteins involved in merozoite invasion and subsequent ring-stage parasite growth. Although our data do not allow us to distinguish which lineage has undergone positive selection, adaptive changes in these functions in one or both of these species are consistent with differences in the biology of the parasites. *P. reichenowi* produces 10–12 merozoites, in contrast to the 20–30 merozoites produced by *P. falciparum*[19], and some data suggests that *P. reichenowi* is less pathogenic to chimpanzees than *P. falciparum* is to humans[19]. Both *P. falciparum* and *P. reichenowi* have a strong preference for their host species[19,20], which may be due to differences in the sialic acid modifications to erythrocyte surface proteins between human and chimpanzee, which are major determinants of the invasion of *Plasmodium* parasites[20].

Our analysis describes genome-wide patterns of variation but can also indicate specific genes of interest. The list of the genes with the highest number of nonsynonymous polymorphisms in the PFCLIN isolate comprises mostly uncharacterized genes (**Supplementary Table 1**). However, it also contains examples of well-characterized surface antigens, such as two VAR-like genes (*PFL0030c* and *PFE1640w*), *PfEMP2* (*PFE0040c*) and *liver stage antigen 3* (*PFB0915w*). Thus, the other members of the list may include additional uncharacterized antigens. Of the uncharacterized genes, *PF10_0355* is particularly notable; it has the second-highest total SNP number, indicating that is may be under immune selection, and, like VAR genes, it contains a DBL (duffy binding–like) domain. The chromosomal location of *PF10_0355* contains a cluster of genes that have previously received much attention: the vaccine candidate *liver stage antigen 1* (*PF10_0356*) is downstream of *PF10_0355*, and several genes encoding merozoite surface proteins are upstream of *PF10_0355*. We identified an additional 15 genes that had a significant excess of nonsynonymous polymorphisms (McDonald-Kreitman test: $P < 0.05$, **Supplementary Table 1**). These 15 genes encode two known polymorphic antigens: SURFIN13.2 (encoded by *PF13_0074*), which has been identified on the cell membrane fractions of merozoites and infected erythrocytes[21], and a CSP and TRAP-related protein (CTRP, encoded by *PFC0640w*), a transmission-blocking vaccine candidate[22]. The samples of genes mentioned above highlight the usefulness of genome-wide surveys of nucleotide variation for the identification of medically important targets for vaccine and drug development.

The present analysis provides an unbiased view of natural variation in the *Plasmodium falciparum* genome and its divergence from *P. reichenowi*. This study and its companion papers[23,24] will further our understanding of the biology of this complex organism. The increasing efficiency of gene manipulation strategies for *P. falciparum* means that hypotheses about gene function derived from evolutionary data can rapidly be tested in the future. More extensive exploration of *P. falciparum* polymorphism will be required for fine-scale analysis of the evolution in the genome. This information, combined with accurate data on geographical origin and parasite phenotype, should allow us to understand both the complex population structure of this important pathogen and to carry out genotype-phenotype association studies. The latter will be vital for elucidating mechanisms underlying processes such as virulence and drug resistance within human populations and are likely to lead to new medical intervention strategies.

METHODS

Detailed methods are available in **Supplementary Methods**.

Parasite collection and DNA extraction. The clinical parasite sample (PFCLIN) was obtained by erythrocytapheresis from a nonimmune woman who had recently returned from Ghana (a detailed case report is available in **Supplementary Methods**). Molecular studies were approved by the Wandsworth Local Research Ethics Committee (UK). DNA was extracted from about 400 ml of enriched infected red blood cells (**Supplementary Methods**).

Laboratory isolate IT DNA was prepared from trophozoite-stage culture of IT subclone P1B5 (ref. 25) using standard overnight proteinase K digestion followed by phenol and chloroform extractions. The *P. reichenowi* study was approved by the Institutional Ethics Committee according to Dutch law.

Sequencing. Randomly sheared DNA was sequenced from small-insert (1.4–5.0 kb) pUC19 clone libraries by whole-genome shotgun sequencing as previously described[26]. From the PFCLIN, IT and *P. reichenowi* libraries, 201,068, 36,138 and 78,442 reads were produced, respectively.

Read alignment and identification of differences. Paired shotgun clone reads were aligned to the *P. falciparum* 3D7 genome[4] (3D7 version 2.0) with SSAHA2 (ref. 5). GFF-format SSAHA outputs are available upon request. Only alignments that mapped to a single location on the reference genome and were opposed by their read pair were used in further analysis (single-location paired reads). SNPs and polynucleotide differences (POLYs), including insertion-deletion events, differences of two or nucleotides within 5 nt of each other, and more complex polynucleotide differences (containing both insertion-deletion events and polymorphisms and substitutions), were identified from these alignments using neighborhood quality standard[27].

Data quality filters. We excluded all read alignments and difference calls (SNPs and POLYs) that mapped to the non-unique regions of the genome[4]; for each project (PFCLIN, IT, *P. reichenowi*) we excluded any 10-kb block of the genome where the 'uniqueness' (U = (single-location paired reads mapped to block)/(all reads mapped to the block)) was ≥ 0.5.

SNP locations with multiple read coverage were used to determine the minimum base Phred score for which 90% of SNP calls were validated (minimum validated score, MVS). All subsequent analysis used only SNPs and 'multiple-SNP' POLYs with a Phred score $\geq$ MVS or a cumulative Phred score from all reads covering the location >60 (1×10^{-6} probability of error; note that Phred quality scores are logarithmically linked to error probabilities; see URL below). All indel POLYs were retained. Estimates of the total rate of genetic change, dN/dS and McDonald-Kreitman tests were calculated from a predicted variant sequence generated using SNP and multiple-SNP POLYs with SSAHA2 scores $\geq$ MVS.

We manually scrutinized 50 SNPs called from PFCLIN over 12 genes (44,124 nt in length) by examining Gap4-aligned electrophoretograms (Gap4 version 4.10b.3) from all sequence reads from PFCLIN covering these genes, and further reads were generated by sequencing PCR products obtained from the initial PFCLIN DNA extraction. We located and verified 48 SNPs, including one heterozygous SNP. One was identified but not found to be a SNP in this alignment, and one SNP was not aligned by Gap4.

Detection of heterozygous sites. Sites with $\geq$1 MVS SNP call and at least N ($N \geq 2$) reads of high-quality read length (**Supplementary Methods**) were used to estimate the proportion of heterozygous sites. A site qualified as heterozygous if the cumulative Phred score from all reads covering that site was $\geq H$ (where $H = 60$, 70 or 80) for each of two alleles. For various values of N and H, estimates of the percentages of SNP locations that were heterozygous were 4%– 6% for IT and 7%–10% for PFCLIN.

Synonymous-nonsynonymous rate ratios (dN/dS). Pairwise alignments were generated using the predicted variant sequences for each comparison (IT, PFCLIN, *P. reichenowi*) to the reference for each protein-coding gene with $\geq$100 codons of unique read coverage, excluding regions containing indels. dN/dS was calculated with the yn00 program of the PAML package[28] using the Yang and Nielsen algorithm. Counts of the number of synonymous and nonsynonymous changes in codons used the same method, except that all codons that were completely included in unique read coverage were included.

McDonald-Kreitman tests. We carried out McDonald-Kreitman tests according to ref. 10 using alignments as above. The neutrality index (NI) was calculated from the equation[29] NI = (Pn/Ps)/(Fn/Fs), where 'Pn' represents the number of nonsynonymous polymorphisms, 'Ps' the number of synonymous polymorphisms, 'Fn' nonsynonymous fixed differences and 'Fs' synonymous fixed differences. The NI values for groups of genes were calculated by concatenating all coding sequences for the group of genes. Significance was evaluated from concatenated gene sequences using Fisher's exact test.

Statistics. All statistics and plots were produced in R (v1.11) (http://www.R-project.org). Tests for differences in evolutionary rate between groups of genes or correlations used only those genes included in the classification or detected in the expression study. For grouped data, if the Kruskal-Wallis rank sum null hypothesis (that all groups were the same dN/dS distribution) was rejected ($P < 0.05$), then Mann-Whitney tests were performed on each group to examine whether the dN/dS distribution of the genes in that group differed from the dN/dS distribution of all other genes that were included in the expression data or classification.

Accession codes. dbSNP: PFCLIN, 65863114–65917062; IT, 65849337–65863113.

URLs. dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP. Sequence trace files are available at http://www.sanger.ac.uk/Projects/P_falciparum/. The polymorphism data will be available for browsing and querying in PlasmoDB (http://www.plasmodb.org). The *P. falciparum* 3D7 genome sequence can be found at ftp://ftp.sanger.ac.uk/pub/pathogens/malaria2/3D7/. SSAHA2 software can be found at http://www.sanger.ac.uk/Software/analysis/SSAHA2/. Phred can be found at http://www.phrap.com/phred.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the Pathogen Sequencing teams for producing the sequence data used in this study, P. Horrocks and B. Pinches for the supply of DNA from the IT isolate and M. Marti for the list of PEXEL motif–containing genes. This study was funded by the Wellcome Trust through its support of the Pathogen Sequencing Unit and E.T.D.'s group at the Wellcome Trust Sanger Institute. Sequencing of the *P. falciparum* IT isolate was funded by a European Union 6th Framework Program grant to the BioMalPar Consortium (contract number LSHP-LT-2004-503578).

AUTHOR CONTRIBUTIONS

D.J. processed SSAHA data, produced diversity and evolutionary measures, analyzed the data and wrote the manuscript. E.T.D. and M.B. directed the project and assisted with analysis of the data and writing of the manuscript. A.P. and A.B. assisted with analysis and processing of the data and biological interpretation of the data. A.T. collected the *P. reichenowi* sample and extracted DNA. K.S. assisted with data processing and analysis. A.C. provided SSAHA mapping. J.S. assisted with data processing. C.I. resequenced genes and manually verified SNPs. A.-C.U. assisted with parasite DNA extraction. S. Krishna assisted in biological interpretation of the data and parasitology. C.N. shaped some of the initial ideas for the project, assisted in biological interpretation of the data and assisted with parasite DNA extraction. S. Kyes grew the IT parasite and purified and extracted DNA from parasites.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Korenromp, E., Miller, J., Nahlen, B., Wardlaw, T. & Young, M. *World Malaria Report 2005* (Roll Back Malaria Partnership, Geneva, 2005).
2. Mu, J. *et al.* Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*. *Nature* **418**, 323–326 (2002).
3. Anderson, T.J. Mapping drug resistance genes in *Plasmodium falciparum* by genome-wide association. *Curr. Drug Targets Infect. Disord.* **4**, 65–78 (2004).
4. Gardner, M.J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511 (2002).
5. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
6. Anderson, T.J. *et al.* Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol. Biol. Evol.* **17**, 1467–1482 (2000).
7. Volkman, S.K. *et al.* Excess polymorphisms in genes for membrane proteins in *Plasmodium falciparum*. *Science* **298**, 216–218 (2002).
8. Carret, C.K. *et al.* Microarray-based comparative genomic analyses of the human malaria parasite *Plasmodium falciparum* using Affymetrix arrays. *Mol. Biochem. Parasitol.* **144**, 177–186 (2005).

9. Yang, Z. & Bielawski, J.P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496–503 (2000).
10. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
11. Rocha, E.P. The quest for the universals of protein evolution. *Trends Genet.* **22**, 412–416 (2006).
12. Florens, L. *et al.* A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520–526 (2002).
13. Le Roch, K.G. *et al.* Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Res.* **14**, 2308–2318 (2004).
14. Bozdech, Z. *et al.* The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* **1**, E5 (2003).
15. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
16. Marti, M., Good, R.T., Rug, M., Knuepfer, E. & Cowman, A.F. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306**, 1930–1933 (2004).
17. Hall, N. *et al.* A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science* **307**, 82–86 (2005).
18. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170 (2005).
19. Garnham, P.C.C. & Duggan, A.J. *Malaria Parasites and Other Haemosporidia* (Blackwell Scientific, Oxford, 1996).
20. Martin, M.J., Rayner, J.C., Gagneux, P., Barnwell, J.W. & Varki, A. Evolution of human-chimpanzee differences in malaria susceptibility: relationship to human genetic loss of N-glycolylneuraminic acid. *Proc. Natl. Acad. Sci. USA* **102**, 12819–12824 (2005).
21. Winter, G. *et al.* SURFIN is a polymorphic antigen expressed on *Plasmodium falciparum* merozoites and infected erythrocytes. *J. Exp. Med.* **201**, 1853–1863 (2005).
22. Li, F. *et al.* Plasmodium ookinete-secreted proteins secreted through a common micronemal pathway are targets of blocking malaria transmission. *J. Biol. Chem.* **279**, 26635–26644 (2004).
23. Mu, J. *et al.* Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nat. Genet.* advance online publication 10 December 2006 (doi:10.1038/ng1924).
24. Volkman, S.K. *et al.* A genome-wide map of diversity in *Plasmodium falciparum*. *Nat. Genet.* advance online publication 10 December 2006 (doi:10.1038/ng1930).
25. Horrocks, P., Kyes, S., Pinches, R., Christodoulou, Z. & Newbold, C. Transcription of subtelomerically located var gene variant in *Plasmodium falciparum* appears to require the truncation of an adjacent var gene. *Mol. Biochem. Parasitol.* **134**, 193–199 (2004).
26. Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400**, 532–538 (1999).
27. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
28. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
29. Rand, D.M. & Kann, L.M. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol. Biol. Evol.* **13**, 735–748 (1996).
30. Young, J.A. *et al.* The *Plasmodium falciparum* sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol. Biochem. Parasitol.* **143**, 67–79 (2005).

CORRIGENDUM: Genome variation and evolution of the malaria parasite *Plasmodium falciparum*

Daniel C Jeffares, Arnab Pain, Andrew Berry, Anthony V Cox, James Stalker, Catherine E Ingle, Alan Thomas, Michael A Quail, Kyle Siebenthall, Anne-Catrin Uhlemann, Sue Kyes, Sanjeev Krishna, Chris Newbold, Emmanouil T Dermitzakis & Matthew Berriman

***Nat. Genet.* 39, 120–125 (2007); published online 10 December 2006; corrected after print 8 February 2007**

In the original version of this paper, the authors failed to acknowledge that sequencing of the *P. falciparum* IT laboratory isolate was funded by a European Union 6th Framework Program grant to the BioMalPar Consortium (contract number LSHP-LT-2004-503578). This error has been corrected in the PDF version of the article.