Malaria Journal

RESEARCH Open Access

Whole genome sequencing and microsatellite analysis of the *Plasmodium falciparum* E5 NF54 strain show that the *var*, *rifin* and *stevor* gene families follow Mendelian inheritance

Ellen Bruske[1], Thomas D. Otto[2,3*] and Matthias Frank[1*]

Abstract

Background: *Plasmodium falciparum* exhibits a high degree of inter-isolate genetic diversity in its variant surface antigen (VSA) families: *P. falciparum* erythrocyte membrane protein 1, repetitive interspersed family (RIFIN) and subtelomeric variable open reading frame (STEVOR). The role of recombination for the generation of this diversity is a subject of ongoing research. Here the genome of E5, a sibling of the 3D7 genome strain is presented. Short and long read whole genome sequencing (WGS) techniques (Illumina, Pacific Bioscience) and a set of 84 microsatellites (MS) were employed to characterize the 3D7 and non-3D7 parts of the E5 genome. This is the first time that VSA genes in sibling parasites were analysed with long read sequencing technology.

Results: Of the 5733 E5 genes only 278 genes, mostly *var* and *rifin/stevor* genes, had no orthologues in the 3D7 genome. WGS and MS analysis revealed that chromosomal crossovers occurred at a rate of 0–3 per chromosome. *var, stevor* and *rifin* genes were inherited within the respective non-3D7 or 3D7 chromosomal context. 54 of the 84 MS PCR fragments correctly identified the respective MS as 3D7- or non-3D7 and this correlated with *var* and *rifin/stevor* gene inheritance in the adjacent chromosomal regions. E5 had 61 *var* and 189 *rifin/stevor* genes. One large non-chromosomal recombination event resulted in a new *var* gene on chromosome 14. The remainder of the E5 3D7-type subtelomeric and central regions were identical to 3D7.

Conclusions: The data show that the *rifin/stevor* and *var* gene families represent the most diverse compartments of the *P. falciparum* genome but that the majority of *var* genes are inherited without alterations within their respective parental chromosomal context. Furthermore, MS genotyping with 54 MS can successfully distinguish between two sibling progeny of a natural *P. falciparum* cross and thus can be used to investigate identity by descent in field isolates.

Keywords: *var* genes, Recombination, E5, 3D7, NF54, Variant surface antigens, Antigenic variation, Epigenetic, PfEMP1, RIFIN, STEVOR, Microsatellites, Whole genome sequencing, Cross over recombination, Non cross over recombination

*Correspondence: ThomasDan.Otto@glasgow.ac.uk; matthias.frank@praxis-frank-tuebingen.de
[1] Institute of Tropical Medicine, University of Tuebingen, Wilhelmstr. 27, 72074 Tuebingen, Germany[3] Present Address: Centre of Immunobiology, Institute of Infection, Immunity & Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK
Full list of author information is available at the end of the article

Background

The malaria parasite *Plasmodium falciparum* is the most prevalent malaria species found on the African continent [1] and is responsible for 90% of deaths from malaria [2]. The NF54 isolate derives from an infection obtained near Schiphol Airport in the Netherlands [3]. Two sibling parasites, 3D7 and E5, were independently isolated by limiting dilution from the original NF54 culture [3, 4]. The 3D7 clone has been used in the malaria genome sequencing project [5], revealing that the *P. falciparum* genome consists of a 23 Mb nuclear genome with 14 chromosomes and around 5500 genes [6]. E5 was incidentally indentified during transfection experiments of the original NF54 culture [4] and has previously been characterized by PCR cloning and gene specific PCR [7]. Whole genome sequencing of *P. falciparum* is complicated by the special properties of the *P. falciparum* genome: it is very AT-rich and contains many repetitive regions and homopolymer runs, especially in its intergenic regions, complicating the assembly of genome data [8–10]. It has recently been shown that the genome can be divided into a core genome (95%) hypervariable regions (5%) [10, 11]. Unambiguous alignments of sequence data of different strains are only possible in the core regions and not in the hypervariable regions that harbour the majority of the variant surface antigen (VSA) gene families [12–14].

To date, five multicopy gene families that encode VSAs have been described in *P. falciparum*: *stevor* (subtelomeric variable open reading frame) [15], *rif* (repetitive interspersed family) [16], *pfmc-2tm* (*P. falciparum* Maurer's clefts two transmembrane) [17], *surfin* (surface associated interspersed genes) [18] and *var* [19]. The best investigated VSA is *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) [6, 20, 21]. PfEMP1 is encoded by the multicopy *var* gene family that consists of about 60 *var* (variability) genes per *P. falciparum* genome [19]. Antigenic variation is primarily mediated by mutually exclusive expression of 1 of the 60 *var* genes per infected red blood cell. The subtelomeric position of most *var* genes [14] predisposes them to recombination contributing to the diversity of PfEMP1 [22]. PfEMP1 is transported to the surface of infected red blood cells and acts as a receptor for the surface receptors on endothelial host cells. This cytoadhesion prevents clearance of the red blood cells by the spleen. Different forms of PfEMP1 possess different binding specificities and individual PfEMP1 variants have been associated with distinct malaria syndromes such as malaria in pregnancy or cerebral malaria [23–28].

In endemic regions, antibodies to PfEMP1 develop early in life have and have been shown to correlate with the development of protective immunity [29]. To escape the human immune response, *P. falciparum* can switch the PfEMP1-variant expressed on the surface of infected red blood cells. Recent investigations also support a role for the non-PfEMP1 VSA proteins in cytoadhesion, antigenic variation and as targets of the human immune response [30–32]. The non-PfEMP1 VSA families are located in close proximity to the *var* genes within the hypervariable regions of the *P. falciparum* chromosomes. The chromosomal position of the VSA gene families thus complicates their genetic analysis. Because of this position the VSA gene families were excluded from a recent extensive analysis of progenies of experimental *P. falciparum* crosses [10].

The aim of this work was to characterize VSA-gene family inheritance in a NF54 clone with WGS technology. To provide a framework to investigate identity by descent (IBD) in field isolates a set of 84 microsatellites was evaluated for its ability to distinguish between the 3D7 and non-3D7 parts of the E5 genome. Microsatellites are variable numbers of tandem repeats in DNA [33]. They have the advantage that they are locus-specific and highly polymorphic. Because most microsatellites are located in non-coding regions they are not subject to purifying selection. The original work by Walliker, Wellems and Su has generated a large repository of MS primers that were originally used to determine the genetic basis of chloroquine resistance [34] and erythrocyte invasion in progeny of experimental genetic crosses [35] as well as multiple other fundamental aspects of *P. falciparum* biology (summarized in Figan et al. [36]). MS flanking drug resistance loci have also been employed to determine the size of genetic sweeps in population based studies [37, 38]. A 12-locus primer set developed by Anderson et al. [39] has been used by many investigators to assess the genetic diversity of field isolates [40, 41]. Recently, Figan et al. [36] identified 12 MS markers that can reliably differentiate progeny from experimental crosses. However, the small number of MS precludes an analysis of chromosomal inheritance. Therefore, here we evaluate a set of 84 microsatellite alleles distributed over the 14 *P. falciparum* chromosomes to type chromosomal regions as 3D7- or non-3D7.

Genome changes in progeny of a *P. falciparum* cross are a consequence of crossover or non-crossover recombination [10, 42]. Crossover recombination represent, a reciprocal exchange between homologous chromosomes during meiosis, whereas non-cross over recombination results in the duplication of a sequence from a donor sites that replaces a sequence at an acceptor site (also referred to as a gene conversion).

The analysis of E5 offered the opportunity to investigate crossover and non-crossover recombination in a natural sibling of the 3D7 genome clone. Zero to three cross- overs per chromosome were identified. VSA gene

families were inherited in their respective parental chromosomal background. The chromosomal distribution of VSA genes in E5 was virtually identical to 3D7. The *var* and *rifin/stevor* gene families represented the most genetically distinct parts of the E5 genome. However, only one definite non-crossover recombination event among non-3D7 and 3D7 *var* genes was detected.

Methods

Parasites, cell culture and generation of DNA

The NF54-C2 clone (isogenic with 3D7) [43, 44] and the NF 54 E5 [4] clone were used for in vivo microsatellite typing. Parasites were cultivated in RPMI 1640 medium completed with 10% Albumax concentrate (Gibco), 2 mM Glutamine, 0.05 mg/ml Gentamicin and 25 mM Hepes buffer (Sigma) at 2.5% haematocrit of 0+ erythrocytes from a local blood bank. Culture flasks were kept at 37 °C under standard parasite cell culture conditions (5% O_2, 5% CO_2, 90% N_2). At a parasitaemia of ca. 4%, the erythrocytes were spun down and parasite DNA was extracted from the pellet using the QiAmp DNA Blood Midi Kit according to the manufacturer's manual.

var gene PCR and Sanger sequencing

PCR using *var*-specific primers from Salanti et al. [26] with modifications [44] was carried out using the following conditions: 94 °C 3 min, 94 °C 30 s, 48 °C 45 s for 30× cycles, then 70 °C 30 s, 70 °C 3 min. PCR products were run on a 1% agarose gel, bands were cut out and purified with NucleoSpin® Extract Kit (Macherey–Nagel) according to the manufacturer's protocol. Preparation of the DNA for sequencing was as follows: the reaction mix containing 1× sequencing buffer, 10% Big Dye Terminator v1.1 (Applied Biosystems), sequencing buffer (Applied Biosystems), 125 µM primer, 60% dH_2O, 1–5 ng DNA were run with the following conditions: 94 °C 10 s, 50 °C 5 s, 25 cycles, 60 °C 4 min. All samples were purified using a 6.7% Sephadex (w/v) column. Sequences were aligned and a consensus sequence was generated with the help of BioEdit sequence alignment editor (http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

Microsatellite primer selection

For each of the 14 chromosomes microsatellites (MS) were selected with the help of the NCBI map viewer database (http://www.ncbi.nlm.nih.gov/mapview/maps.cgi). The initial selection of MS was based on the unpublished analysis of the 3D7xHB3 cross (kindly provided by Akhil Vaidya, Department of Microbiology and Immunology, Drexel University College of Medicine, Philadelphia, Pennsylvania, USA). A total of 84 MS were evaluated. In order to allow cross reference to MS alleles employed in other MS genotyping investigations 4 microsatellites evaluated by Anderson et al. [39] were also included. Subsequently, 4 MS per chromosome were selected for the genotyping of E5 and NF54C2. In general, 3 MS were located in subtelomeric regions and 1 in a central chromosomal region. For chromosome 7, 5 MS were chosen. The MS 9B12, located 1.4 Kb downstream the chloroquine (CQ) resistance gene *pfcrt* [38] and MS B5M77, which is located 18.1 Kb upstream of the *pfcrt* locus were also included. 9B12 is highly conserved in chloroquine resistant strains. B5M77 is positioned at the 3′ end of the "chloroquine resistance genetic sweep" area and thus exhibits low allele diversity in resistant strains but high allele diversity even in chloroquine sensitive strains [38, 45].

Microsatellite PCR

Microsatellite PCR was performed using the following conditions: 5 µl DNA were used in a 50 µl reaction containing 1× PCR buffer, 1.5 mM $MgCl_2$, 0.08 µM dNTPs and 0.25 µM primer. For fragment analysis, the forward primer was labelled at its 5′ end (Eurofins Genomics). Four different dyes were used per chromosome (Table 1). The program was as follows: 94 °C 5 min, 94 °C 20 s, 45 °C 10 s, 40 °C 10 s, 60 °C 30 s, 40×, 65 °C 2 min. PCR products were checked on a 1.5–2% agarose gel.

MS sequence analysis

To verify the MS position, all MS were amplified from NF54 C2 DNA and the reaction products were sequenced. The obtained sequences were manually aligned with the 3D7 MS sequence in the database. Only MS PCR primers pairs that amplified the 3D7 reference sequence were used for the fragment analysis.

Fragment analysis

PCR products were all diluted 1:200 with water. A master mix was prepared with 5 µl H_2O and 5 µl Formamide (Hi-Di Life Technologies) and −0.1 µl Standard (LIZ 500 Life Technologies). Per singleplex reaction 10 µl master mix and 1 µl diluted PCR product were used. For multiplex fragment analysis 10 µl master mix and 1 µl diluted PCR product of each PCR reaction were used. The mix was heated at 95 °C for 3 min and immediately chilled on ice for a few minutes before fragment analysis. Analysis was done in 96-well plates (Biozym Scientific GmbH) with the Applied Biosystems ABI Prism 3130xl Genetic Analyzer and data were evaluated using GeneMapper v 4.1.

Chimera breakpoint PCR

PCR across the breakpoint of the chimeric *var* gene in E5 containing an E5-like half downstream and a part (approx. 3 kb) of the 3D7 *var* gene Pf3D7_083350

Table 1 Microsatellite coordinates on the respective chromosomes, primer sequences and the dyes used are depicted

Chromosome	Microsatellite	Position on chromosome (3D7/NCBI-Map viewer)[centiMorgan]	Fwd rimer (5´-3´)	Rev primer (5´-3´)	Fluorophore
1	C1M38	124812 .. 124973	GCCATATCATCGGTAATAAT	CTGGTTGAATGATCTAAGAA	ATTO 565
	C1M39	182870 .. 183017	GTAAATCGTTAACATATTCAC	CATGTATGATCTATGTCCAAA	ATTO 550
	B7M97	216,927 .. 217,120	TATCTTCAAACGATTTGGAA	ATGGAAGTCTTCTTCATCATG	YY
	C1M13	551434 .. 551570	GGGATATAATATTATGTTATTG	AATCACCTACACGATACAAC	FAM
2	KPG	105,531 .. 105,694	TCTAATAACGTAAGTTCA	TGGAGTAAATATTGTTCA	ATTO 565
	C2M20	147593 .. 147729	CAGGGTTCATGTTATATTGA	AGGAGAACCTCACAGTAAT	ATTO 550
	C2M11	457780 .. 457922	CATTCAAGTGTATTATCATTA	TGCATTTGGAGTGAGCTT	YY
	BM41	772378 .. 772535	CATGTTTATTATGATTGGGAA	TAATGATCCATGTACCTTTCC	FAM
3	C3M29	148906 ..149074	GAGAGCAAAAATGGCAGAAG	TCATTAATCCTCTTAACTACA	ATTO 550
	C3M27	222118 .. 222274	AGTATCATATTTGGTTAGATC	TTTGGTTAACAAATTTCCTAC	PET
	C3M33	502303 .. 502453	CTTATAAAAGAATTACCTGG	TTGTTACATTTTAAATGGTAC	YY
	C3M45	937398 .. 937557	CGAAAAGATAACTTACACATT	AATCATATCATATATGCAAGC	FAM
4	C4M62	108839 .. 109123	GAATTCACTTAAATGTATTTATTTG	GACACAAGTTATTTTGTGAAT	ATTO 565
	C3M35	796589 .. 796807	GGAAATATATATCATACTTGG	TTTTTGGTGTCGGTTATTTTT	YY
	B5M109	1011330 .. 1011466	AAAAAAATAAATAATAATAATAAC	TGTGGGGAAATATTTGTCG	ATTO 550
	B5M51	1057107 .. 1057335	AACACAACATATGAATTCTCC	TCTTTCCATCTTTATCGTTC	FAM
5	B5M58	140933 .. 141164	AAATGTTATATCATTTGGGGA	AGTGGATCATATATTTAATGC	NED
	B5M96	232350 .. 232518	ATATCAGGAGTATGGTTTTG	AAAAAAGGCTAGGTAAAATTC	ATTO 565
	C5M12	683546 .. 683730	TCAAAGTATAAATATAACCAC	CTAATAAGGTTGATGTTACTTCC	YY
	B5M94	1212784 .. 1212947	GGGTTCTTAATATTTTTTACC	CATATCAAAATTCATCATTTCT	FAM
6	BM70	258517 .. 258701	GGAAAATATCCCAGAAAAGG	GGAACAAAAAAAAAAAGGAAA	FAM
	Ta109	800986 .. 801159	GGTTAAATCAGGACAACAT	CCTATACCAAACATGCTAAA	YY
	TA1	899,844 .. 900,029	CCGTCATAAGTGCAGAGC	TTTTATCTTCATCCCCACA	ATTO 550
	Ta24	1101542 .. 1101734	CATAGATACATCAAACATAA	TAAATAAAAATTTATTCCTG	ATTO 565
7	B5M77	290289 .. 290433	TAAAGTCTTTCAATACATATG	GAAATAATTTCATATACACAC	ATTO 565
	9B12	313,056 .. 313,217	ATATATTCCAGTATGTTCGC	AATGATACAATGGGATTTAC	
	C13M30	599702 .. 599908	CCTTTTGGGGTTTAAATGTA	TGGAGATGGAACATGGAAAA	YY
	BM51	1030760 .. 1030899	TAATAATATTAATGGTGGTGA	TAATTGTATGACTTCGAGAAT	NED
	ebp	1271522 .. 1271662	TTCACAAGCCAAATATCA	ATTCATAACTCCTTCAGA	FAM
8	hrp2	98739 .. 98901	CGTAAGCATTTTAATTGCA	TAAATGCGGAATTATTCTA	ATTO 550
	BM5	123670 .. 123810	GAAAGTAGATTGTAGTATTTA	TACACATGAATGATTTAATCA	FAM
	BM16	575427 .. 575586	GCTCCTCAATAAATTGTTAT	TCTGTTGCCTCTCAGACAAT	YY
	BM62	1201795 .. 1201968	TCTGATGGTATAACCAGATAC	GATATAGCGAATTCTTAAGAG	ATTO 565
9	B7M57	125063 .. 125286	GTAAGAATAGATTTAACAAATG	AAGAGAAAGAGAGAAAAAATG	FAM
	C9M103	743822 .. 743972	ATTAGATATTATAATGAACGG	ATGTGTACGTGTCGAATACT	ATTO 550
	BM54	817042 .. 817208	GATGAATATTATGAGGAAAAC	TCATCATAATTAACAATATGG	YY
	C9M43	1430410 .. 1430541	GACACACATATAGTAATAGA	GATATACATATATGGACATAT	ATTO 565
10	C4M3	57991 .. 58168	GTTGTTTCGGCAATTTACCT	TTAATGCACTTTTATTATCAC	ATTO 565
	B7M101	194437 .. 194637	TATATGGAAGTTTTCTTCAGG	CTATGTTTATGTTAATTATTGC	ATTO 550
	B7M46	763306 .. 763479	AGCCATTCGTAAACTGCCT	CACTATCAATATAAGTATCC	YY
	ta40	1322577 .. 1322793	AAGGGATTGCTGCAAGGT	CATCAATAAAATCACTACTA	FAM
11	Ta119	627614 .. 627856	TCCTCGATTATATTATTGCA	TAATACATTCCCATTAGATG	ATTO 565
	C12M110	1515255 .. 1515393	GATGATAAATATGCACCCATC	TTCATGTATATGCATATATAC	YY
	ta117	1790958 .. 1791137	ATCTCTACCTCAACCACCA	TGTGTTACCACCATTGTTA	ATTO 550
	Resa2	1990946 ..1991063	CTATTTGTTTATAGTTATGTA	TTTAATGTAGTGTCATGAA	FAM
12	C12M30	0	CTTCAATAAGGGAAAATCCA	TAATAGTAAGTAAAGTCACA	ATTO 550
	TA121	168703 .. 168863	ACTTGTCAAGTGCTCATCA	TTTGTAATTTTCACTAGGAT	FAM
	Ta34	1144887 .. 1145007	AACATAGCCAAATCGCAC	CCATTTGATGTGTCATCAC	YY
	Ta48	2032882 .. 2033158	TTTTGATATCTCTCAATCAT	CTTCACGACAGAGGTGTC	ATTO 565
13	B8M6	703669 .. 703794	ATGATGCAGAAAAGAATAAAT	GTGATGGTTTCTCAAATTTGG	FAM
	C1M70	1457131 .. 1457306	ATATCGAAAGGTGAATAGAAA	ATAATTAATATGGTCATATGG	YY
	ta60	2584963 .. 2585166	CTCAAAGAGAAATAATTCA	AAAAAGGAGGATAAATACAT	ATTO 565
	C14M35	2664073 .. 2664255	ATCCCTACATGAATAAAATG	TCCCTTATGTATACTTCCAC	ATTO 550
14	C14M59	125828 .. 126006	AGTACAAAGAAGTATATCCAT	CTTATAATAGATAAATGTGTC	ATTO 565
	RHO1	419854 .. 420048	TGTAAAAATAGACATTTCA	AAAACGAAAATACAACCAA	FAM
	Ta88	1610557 .. 1610781	CTGGTAACGATGGAAAAGC	TACGCTTATTGTTATTACTCA	YY
	Pf9607	2315888 .. 2315994	TTTAAAGACCGATCATCA	AGTAGCACAACAATAACA	ATTO 550

upstream was done using the primers PFE5_F1 (5′-CGC CATAGTATCACCAATGC-3′) and PF3D7083350_R2 (5′-CCCGACGTGGTACACCTG-3′) with the following conditions: 3 min 94 °C, 10 s 94 °C, 30 s 56 °C, 30 s 72 °C, 3 min 72 °C, 40 cycles, using 5 µl DNA template (concentration up to 5 ng), 2.5 µM primer, 0.2 mM $MgCl_2$, 0.2 mM dNTPs. PCR products were checked on an 1% agarose gel.

Whole genome sequencing

Illumina

Genomic DNA of E5 was sheared into 250–350 bp fragments by focused ultrasonication [Covaris Adaptive Focused Acoustics technology (AFA Inc., Woburn, USA)]. An amplification-free Illumina library [46] was prepared and sequenced on a Illumina GAII (150 bp) platform according to the manufacturer's standard sequencing protocol. Reads were mapped with smalt (ftp://ftp.sanger.ac.uk/pub/resources/software/smalt/,parameter–x–a 1000). Variants were called with gatk against the *P. falciparum* version 3 assembly from geneDB [47, 48].

Pacific bioscience reads

From the same DNA, SMRTbell template library using the Pacific Biosciences issued protocol (20 kb Template Preparation Using BluePippin Size-Selection System) were generated. Five SMRT cells were sequenced on the PacBio RS II platform using P5 polymerase and chemistry version 3. Raw sequence data were deposited in the European Nucleotide Archive under accession number ERS500965.

Sequence processing

Sequence data from the SMRT cells were assembled with HGAP. As expected genome size 23.5 Mb was used. Next, the contigs were further improved with IPA (https://github.com/ThomasDOtto/IPA). The script performs following steps: delete small contigs, identify overlapping contigs with low Illumina coverage, order contigs against the *P. falciparum* 3D7 reference using ABACAS2 [49], corrects errors with Illumina reads using iCORN2 [50], circularizes the two plastid genomes with circulator [51] and renames the chromosomes and contigs. The draft genome was annotated with companion [52], using *P. falciparum* 3D7 version 3 from October 2015 as reference.

Bioinformatic sequence analysis

Using Artemis [53] and bamview [54], a free genome browser and annotation tool, and the Artemis Comparison Tool (ACT) [54, 55], a pairwise comparison tool of DNA sequences from the Wellcome Trust Sanger Institute (Hinxton, UK) to visualize similarities and differences between genomes, the genome of 3D7 (serving as reference) was compared with E5. By laying one sequence over the other, coverage and single nucleotide polymorphism (SNP) maps of the E5 reads over the 3D7 genome can be loaded into the program. The SNP map served as a tool to detect differences between 3D7 and E5. Areas with low SNP frequency and even Illumina read coverage were defined as 3D7 chromosomal areas. To find shared proteins between PfE5 and Pf3D7, they were compared (ignoring alternative splicing) with a BLASTp (E-value cutoff 1e−6) and then clustered with orthomcl version 1.4, default parameter [56].

Results

Analysis of chromosomal inheritance in E5

To characterize chromosomal regions of E5 as 3D7 or non-3D7 a set 84 MS primer pairs was evaluated for multiplex PCR with fluorescent probes. First, all MS were amplified under the same set of PCR conditions from 3D7 DNA. Most of the reactions resulted in a clear product after gel electrophoresis. Sanger sequencing and manual alignment of the sequences to the 3D7 reference genome showed that 57 of the 84 MS PCR fragments could be unambiguously aligned to the target MS sequence, whereas 27 MS resulted in sequences that either did not amplify or could not be aligned with the target sequence (Additional file 1).

57 MS were, therefore, employed for genotyping (Table 1). All MS were amplified from E5 and 3D7 DNA and MS length was determined by capillary fragment analysis. Each chromosome was characterized at 4–5 MS loci. An MS allele was designated as 3D7-type, if the fragment length difference was ≤ 3 bp between E5 and 3D7. E5 had 37 3D7-type and 20 non-3D7 MS alleles (Table 2). On chromosomes 1, 3, 6 and 9 all MS alleles were 3D7-type. The remaining 10 E5 chromosomes were composed of 3D7 and non-3D7 MS alleles. The pattern of MS allele distribution suggested that large chromosomal haplotypes were inherited together, however the distant spacing of the MS markers precluded fine scale mapping. The latter was only achieved for the area of chromosome 7 that harbours the chloroquine resistance gene *pfcrt*. The microsatellites 9B12, located 1.4 Kb downstream of *pfcrt* [34, 45] and B5M77, located 18.1 Kb upstream of this locus, had 3D7-type alleles in E5. Thus, MS in the "genetic sweep" area were the same in E5 and 3D7, suggesting that in both clones this chromosomal fragment was inherited as a 3D7-type haplotype.

To further evaluate chromosomal inheritance we next characterized E5 and 3D7 with a 3D7 specific *var* gene primer set. This revealed overall good correlation between the presence of 3D7 *var* genes and 3D7 MS alleles in the respective chromosomal areas (Table 2).

Table 2 Genotyping based on microsatellite fragment length and 3D7 specific *var* gene PCR

Chromosome	MS	*Var* gene	Expected length (NCBI)	Length 3D7	Length E5
1		PF3D7_0100100			
		PF3D7_0100300			
	C1M38		160	165	165
	C1M39		150	151	151
	B7M97		194	190	190
	C1M13		137	136	136
		PF3D7_0115700			
2		PF3D7_0200100			
	KPG		161	167	163
	C2M20		127	139	123
	C2M11		143	139	139
	BM41		158	159	159
		PF3D7_0223500			
3		PF3D7_0300100			
	C3M29		169	ich	174
	C3M27		157	160	160
	C3M33		151	150	150
	C3M45		210	160	160
		PF3D7_0324900			
4		PF3D7_0400100			
		PF3D7_0400400			
	C4M62		285	291	221
		PF3D7_0412900			
		PF3D7_0413100			
		PF3D7_0412400			
		PF3D7_0412700			
	C3M35		219	216	174
		PF3D7_0420900			
		PF3D7_0420700			
		PF3D7_0421100			
		PF3D7_0421300			
	B5M109		137	140	158
	B5M51		229	227	214
		PF3D7_0425800			
		PF3D7_0426000			
5		PF3D7_0500100			
	B5M58		232	236	215
	B5M96		169	173	173
	C5M12		185	183	183
	B5M94		164	163	131
6		PF3D7_0600200			
		PF3D7_0600400			
	BM70		185	185	185
		PF3D7_0617400			
	Ta109		174	173	174
	Ta1		168	188	188
	Ta24		193	194	194
		PF3D7_0632500			
		PF3D7_0632800			
7		PF3D7_0700100			
	B5M77		145	148	148
	9B12		165	160	160
		PF3D7_0711700			
		PF3D7_0712300			
		PF3D7_0712400			
		PF3D7_0712600			
		PF3D7_0712800			
		PF3D7_0712900			
	C13M30		207	205	205
	BM51		140	144	140
	ebp		141	138	138
		PF3D7_0733000			
8		PF3D7_0800100			
		PF3D7_0800200			
		PF3D7_0800300			
	hrp2		163	163	164
	BM5		141	140	140
		PF3D7_0808600			
		PF3D7_0808700			
		PF3D7_0809100			
	BM16		160	159	159
	BM62		174	176	164
		PF3D7_0833500			
9		PF3D7_0900100			
	B7M57		224	225	225
	C9M103		151	154	154
	BM54		167	167	167
	C9M43		132	134	134
		PF3D7_0937600			
		PF3D7_0937800			
10		PF3D7_1000100			
	C4M3		178	179	174
	B7M101		201	204	185
	B7M46		174	170	151
	Ta40		217	215	212
		PF3D7_1041300			
11		PF3D7_1100100			
		PF3D7_1100200			
	Ta119		243	247	247
	C12M110		139	160	134
	Ta117		180	181	180
	Resa2		118	115	115
		PF3D7_1150400			
12	C12M30		158	114	114
		PF3D7_1200100			
		PF3D7_1200400			
		PF3D7_1200600/ var2csa			
	Ta121		161	158	164
		PF3D7_1219300			
	Ta34		121	116	117
		PF3D7_1240300			
		PF3D7_1240400			
		PF3D7_1240600			
		PF3D7_1240900			
	Ta48		277	276	270
		PF3D7_1255200			
13		PF3D7_1300100			
		PF3D7_1300300			
	B8M6		126	122	125
	C1M70		176	174	174
	ta60		204	206	199
	C14M35		183	181	151
		PF3D7_1373500			
14	C14M59		179	181	158
	RHO1		195	195	195
	Ta88		225	223	223
	Pf9607		107	111	111

Table 2 (continued)

The MS individual fragment lengths are shown for 3D7 and E5 as well as for the in silico 3D7 microsatellite lengths in the NCBI database. MS alleles with > 3 bp size difference between 3D7 and E5 are typed as non-3D7 (grey). 3D7 alleles are white, non-3D7 alleles are grey. 3D7 *var* gene amplification on E5 DNA was verified by targeted Sanger sequencing of PCR fragments. Note that the table differs from Table 1 in Frank et al. [7] at the 5′ end of Chromosome 7 (reannotation of PF3D7_0700100 previously Mal8P1.220) and at the 3′ end of Chromosome 13: PF3D7_1373500 (previously MAL13P1.356)

However, in a few telomeric areas MS and *var* gene typing did not correlate. This was most noticeable on chromosome 7 (*ebp*), 8 (*hrpII*) and 12 (C12M30) where the MS genotyping suggested a 3D7 chromosomal haplotype but the adjacent 3D7 *var* genes were not amplified from E5.

Comparison of the PCR fragment lengths of the 57 MS after amplification on 3D7 DNA with the "in silico" length of the respective MS in the 3D7 genome (version 3) revealed, that 21 MS exhibited a size difference of > 3 bp thus raising the question of the validity of some of the MS typing results.

To validate the MS and *var* gene haplotyping results E5 was characterized by whole genome Illumina sequencing resulting in a E5 genome with > 100× coverage. Mapping of E5 onto the 3D7 reference sequence revealed areas with many SNPs and low coverage, which were defined as non-3D7 (E5-type) and areas with only few SNPs and even median coverage which were regarded as 3D7-type (Fig. 1a). Fragment analysis as well as WGS both showed the same pattern of cross overs in chromosomes 1–14 in clone E5 (compared to 3D7 reference) (Fig. 1b) and confirmed that chromosomes 1, 3, 6 and 9 of E5 were identical with 3D7 and thus appeared to have been inherited without cross overs. In the remaining 10 chromosomes between one to three cross overs per chromosome were detected by Illumina WGS. WGS haplotyping of the telomeric areas of chromosome 7, 8 and 12 typed these areas as non-3D7, confirming the *var* PCR genotype. In contrast MS genotyping of the same areas (*ebp*, *hrp2* and C12M30) showed identical alleles between E5 and 3D7 suggesting a 3D7-haplotype. Together the data suggested that these 3 MS were not sufficiently diverse to allow haplotyping of sibling parasites and they were, therefore, excluded from further analysis. In summary, 54 MS allele typing results were confirmed by WGS as being 3D7 or non-3D7. Overall, the MS and WGS data showed that genetic exchange in the progeny of a natural genetic cross occurred at a rate of 0–3 crossover per chromosome. Allthough *var* gene fragment haplotyping correlated well with microsatellite and WGS haplotyping the Illumina read length precluded a exact anaylsis of VSA gene family inheritance in E5.

Anaylsis of VSA gene family inheritance in E5

Short read assemblies do not permit the accurate assembly of non-reference subtelomeric regions, therefore VSA family inheritance in E5 was assessed by long read Pacific Bioscience sequencing technology. This resulted in an assembly of 58 contigs. Using IPA, the number of contigs was reduced to 29, including the apicoplast and mitochondrial genomes, resulting in a total assembly length of 23.3 Mb.

Annotation of the assembly generated 5733 genes (Table 3). Overall the E5 genome showed a highly conserved structure compared to the 3D7 reference genome. Of the 5733 E5 genes all except 278 had orthologues in the 3D7 genome. These 278 genes were, therefore, designated as singletons (Fig. 2). The *rifin/stevor* and *var* families had 58 singletons and 11 singletons respectively and together represented the largest group of genes with known function among the singletons (Additional file 2).

The Pacific Bioscience genome assembly confirmed the Illumina haplotyping and furthermore enabled a detailed analysis of the chromosomal areas harbouring VSA gene families. VSA family genes consisted of: 62 full length *var* genes (61 LARSFADIG motifs), 32 *var* pseudo-genes, 189 *rifin*/*stevor* genes, 20 *rifin*/*stevor* pseudo-genes, 8 *surfins*, 6 *Pfmc-2TM* genes and 2 *Pfmc-2TM* pseudo-genes. Comparison of the Pacific Bioscience E5 genome with the 3D7 genome showed that the VSA antigen families have virtually the same size (Table 4). Both clones share the same distribution of VSA genes into subtelomeric and central regions as 3D7 (genome sequence vs.3) (Fig. 3). One miss-assembly in the first *var* gene cluster of chromosome 4 was detected (Additional file 3).

3D7 VSA genes were surrounded by 3D7-type chromosomal areas. Similarly, the non-3D7 (E5 specific) VSA sequences mapped to non-3D7 chromosomal areas. Interestingly for the E5 part that is identical to 3D7 the majority was co-linear with the respective areas in 3D7. Only one large scale recombination event was detected in the 3D7-type subtelomeric regions of E5 (see below).

Based on the Pacific Bioscience assembly the correlation of telomeric MS genotype and *var* gene genotyping was reevaluated (Fig. 3). Of the 54 MS, 27 were located in vicinity of telomeric regions (distance range 50–1,000,000 kb from the telomeres). 11 MS carried 3D7 alleles and 16 non-3D7 alleles. In 26 of the corresponding 27 telomeric areas the *var* gene alleles correlated with the MS alleles. The only exception was the 5′ end of

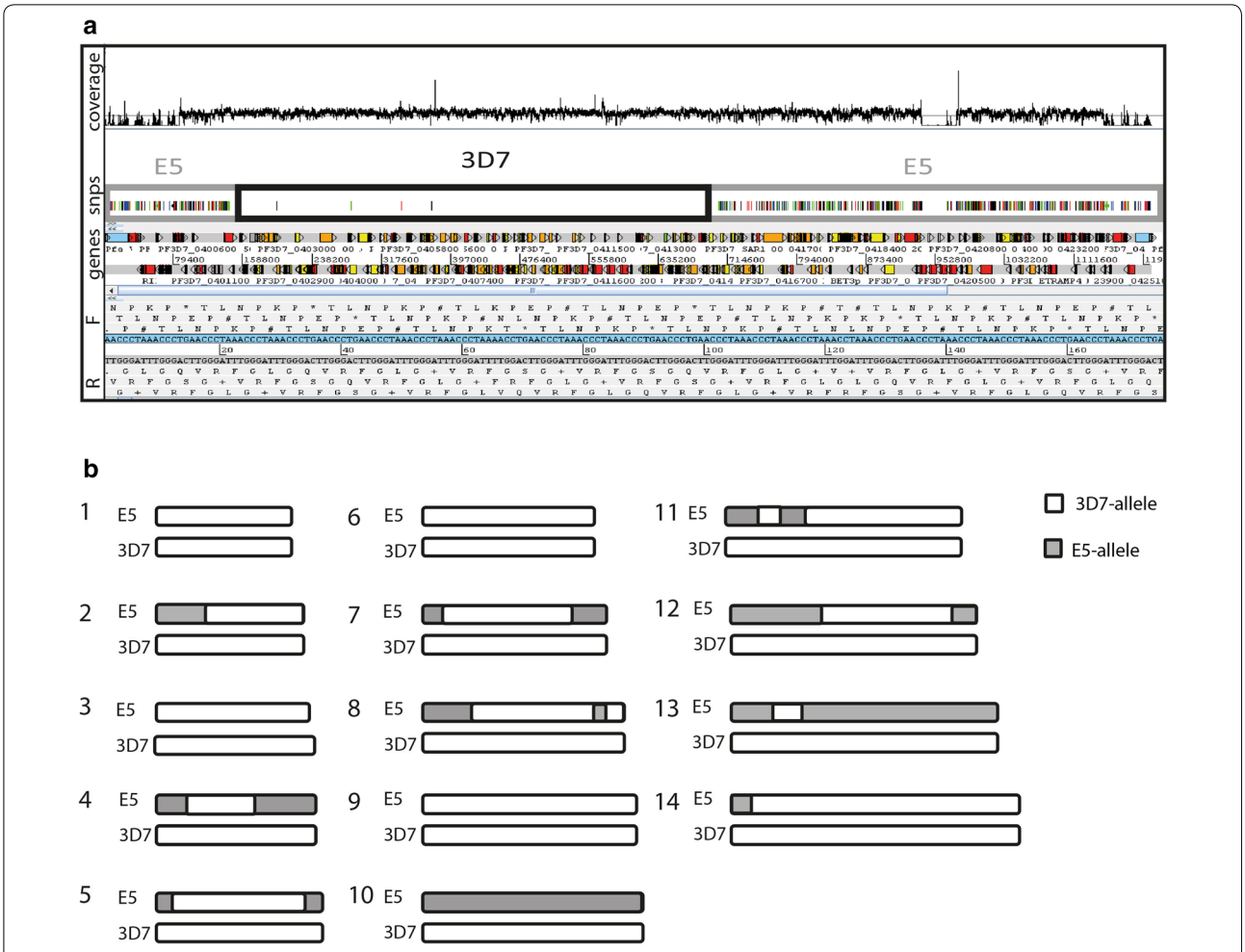

Fig. 1 **a** 3D7 Artemis view of chromosome 4 showing snp plots and coverage in comparison to E5. Areas with many SNPs and low coverage were defined as non-3D7 (E5-type), those with only few SNPs and a high coverage were regarded as 3D7-type. **b** Chromosome map deriving from illumina whole genome sequencing data showing putative crossovers in the individual chromosomes of 3D7 compared to E5. 3D7-alleles are depicted in white, parts distinct from 3D7 ("E5-alleles") in grey

Table 3 *Plasmodium falciparum* NF54 E5 genome characteristics

Number of annotated regions/sequences	29
Number of genes	5733
Gene density (genes/megabase)	240.97
Number of coding genes	5607
Number of pseudogenes	126
tRNA	105
Overall GC%	19.28
Coding GC%	23.9

chromosome 11 that carried non-3D7 *var* genes but the next MS TA119 located at approximately 600 kb had a 3D7 allele. WGS showed that a chromosomal crossover had occurred 5′ to TA119. WGS and MS data thus clearly showed that *var* gene inheritance followed a Mendelian pattern.

To estimate the contribution of non-crossover recombination to VSA diversity the *var* gene family of E5 was evaluated. E5 specific *var* genes that shared sequences between 50 and 500 bp with 3D7 were identified and then manually verified using the ACT program. This identified one previously described *var* gene in E5 (PfE5_120005800) that shares 105 bp with Pf3D7_0937800 that is present in E5 and 3D7 on chromosome 9.

A new chimera (preliminary nomenclature: PfE5_232200) was located on chromosome 14. It shares one half of exon 1 of the *var* gene Pf3D7_0833500 (MAL7P1.212, approx. 3 kb) and the remainder of the subtelomeric area with 3D7. The rest of the *var* gene is E5-specific (Additional file 4). The same but complete

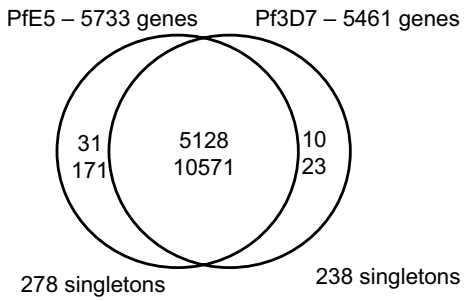

Fig. 2 Venn diagrams displaying shared and species-specific orthologue clusters and their proteins in the target genome *P. falciparum* E5 and the *P falciparum* 3D7 reference. Singletons, i.e. genes without orthologues and paralogues in either species, are placed outside the Venn diagram to the left and right. The numbers within the Venn diagram that belongs to both genomes represent the number of orthologue groups (upper number) and the number of genes in orthologue groups (lower number). The numbers within the E5 and 3D7 specific Venn diagram circles represent paralogue groups and the number of genes within paralogue groups. The assembly can be found at ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/falciparum/E5/Version1

Table 4 3D7 and E5 VSA-gene families

VSA-gene family	Genes in 3D7	Genes in E5
var (≥ *4* kb)	61	62
rifin + *stevors*	190	189
pfmc-2tm	12	6
Surfin	7	8

3D7 data was retrieved from GeneDB [47]

var gene Pf3D7_0833500 is also found on chromosome 8 of 3D7 and E5. PCR analysis across the breakpoint in E5, coming from the E5-specific part of the *var* gene and going into the 3D7-type *var* gene, yielded a product and thus verified the result in vivo (Fig. 4). Together the data suggest that the telomeric end of chromosomes 14 up to the middle of the *var* gene is duplicated. Both chimeric genes thus appear to have resulted from a partial duplication of a 3D7 *var* gene.

Discussion

3D7 and E5 were both cloned from the original NF54 isolate [3, 4] and thus represent progeny of a natural genetic cross. Although the parents of this cross are not known, a previous analysis of 32 progeny of the 7G8XGB4 experimental cross [57] has shown that the two parental genomes are inherited on average at a ratio of 1:1 per progeny. Given that approximately 50% of the E5 genome is identical to 3D7 this suggests that 3D7 is isogenic with one parent of this cross. Thus, analysis of E5 allowed an assessment of chromosomal crossovers as well as non-crossover recombination in a progeny clone of a natural genetic cross.

In this work, the E5 genome was characterized with MS genotyping as well as short and long read WGS techniques. All genotyping approaches suggested a chromosomal recombination rate of 0–3 crossovers per chromosome, consistent with previously reported crossover rates in progeny of experimental genetic crosses [10, 57]. Similarly, all methods indicated that inheritance of VSA gene families occurred within the context of the respective parental haplotypes. A comprehensive analysis of VSA inheritance was however only possible with long read Pacific Biosciene WGS, because the readlength of > 8000 base pairs enabled an accurate assembly of the highly variable telomeric and central chromosomal parts that harbour the VSA gene families. This analysis showed that the VSA gene families have almost the same number of genes in E5 and 3D7.

Annotation of the E5 genome revealed a total of 5733 genes. This number is slightly higher than the 5500 genes in the 3D7 reference genome and is explained by the fact that companion annotation tool overpredicts open reading frames [52]. Genome wide comparison by orthomcl-analysis revealed that the E5 and 3D7 genomes consisted of > 95% genes that had orthologues in both genomes. Only approximately 4% of the E5 and 3D7 genes were singletons and the *rifin/stevor* and *var* genes represented the largest group of genes with known functions among the singletons. Despite this, the total number of identified singleton *var* genes was lower in the orthomcl-analysis than the number of unique E5 *var* genes identified by direct sequence alignment. The underestimation of *var* gene diversity by the orthomcl-analysis is likely due to highly conserved exon II sequences. Overall the data are clearly consistent with the previously reported high genetic diversity of VSA gene families compared to the highly conserved *P. falciparum* core genome.

The *var* gene family has long been shown to be prone to recombination during meiosis [7, 42, 58–60] and mitosis [9, 61, 62]. Furthermore, several investigations have recently quantified mitotic *var* gene recombination rates [9, 62] in different strains. Analysis of the 3D7 and E5 genomes revealed that E5 had a total of 62 *var* genes (compared to 61 *var* genes in the 3D7 reference genome). The "additional" new *var* gene was generated by recombination between a 3D7 *var* gene on chromosome 8 and an E5 specific *var* gene on chromosome 14. 3D7 has no full *var* gene on chromosome 14, but recently Otto et al. showed that 8 of 10 field isolates carry a *var* gene in this subtelomere of chromosome 14 [11]. This shows that non-chromosomal recombination can expand the *var* gene repertoire of individual strains but that the sites of

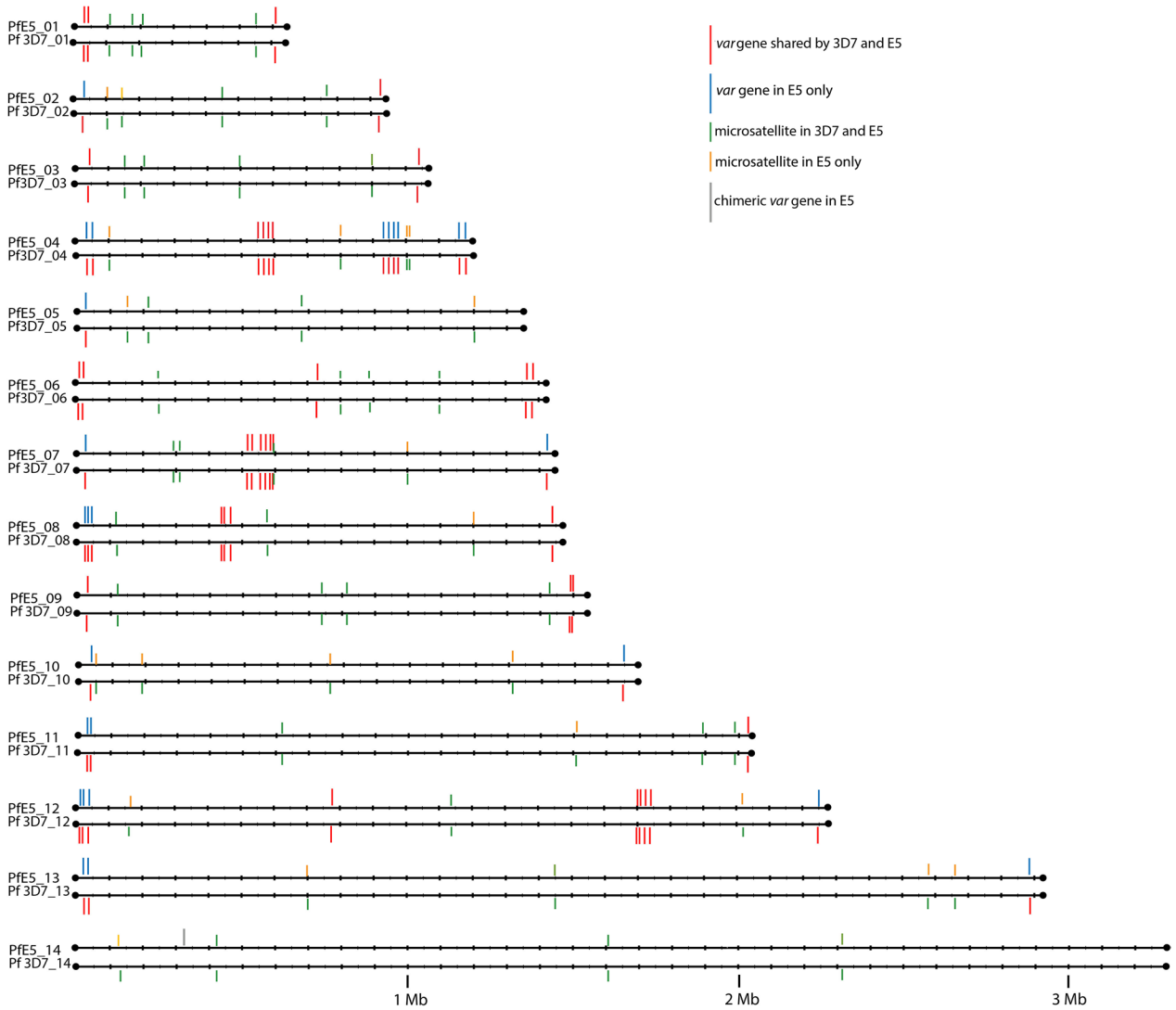

Fig. 3 E5 and 3D7 have the same *var* gene distribution into telomeric repeats and central clusters. *var* genes that are identical between E5 and 3D7 are coloured in red. *var* genes only found in E5 are blue. MS that are identical between E5 and 3D7 are coloured in green. MS that are only found in E5 are coloured in orange. The E5 chimeric *var* gene on chromosome 14 is depicted in grey. Note that for clarity reasons only the *var* genes are depicted. The exact chromosomal location of the *rifin/stevor* gene positions can be found at ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/falciparum/E5/Version1

these changes appear to be conserved across different isolates. The presence of an intact "3D7 donor" sequence suggests that the chimeric *var* gene is the result of a gene conversion event as it has been reported previously for the *var* gene family [42, 61]. Recently Calhoun et al. [63] showed that experimentaly induced double stranded breaks are repaired by the "telomerase healing" pathway. Indeed their work showed a similar non-crossover recombination event resulting in the replacement of a chromosome 13 telomere by a chromosome 9 telomere, thereby creating a new chimeric *var* gene on chromosome 13. The data presented here thus support a role for telomere healing in the generation of VSA gene family genetic diversity. A previously described chimeric *var* gene sequence [7] that carries a 105 bp 3D7 fragment within the DBL of the E5 *var* gene was reidentified in the current analysis and the corresponding "3D7 donor" *var* gene was localized to chromosome 9. This chimeric sequence is located within a hypervariable DBL block that has been shown to exhibit high sequence variability in field isolates [64]. Larger population based studies with long read WGS are necessary to determine if this type short chimeric sequence represent

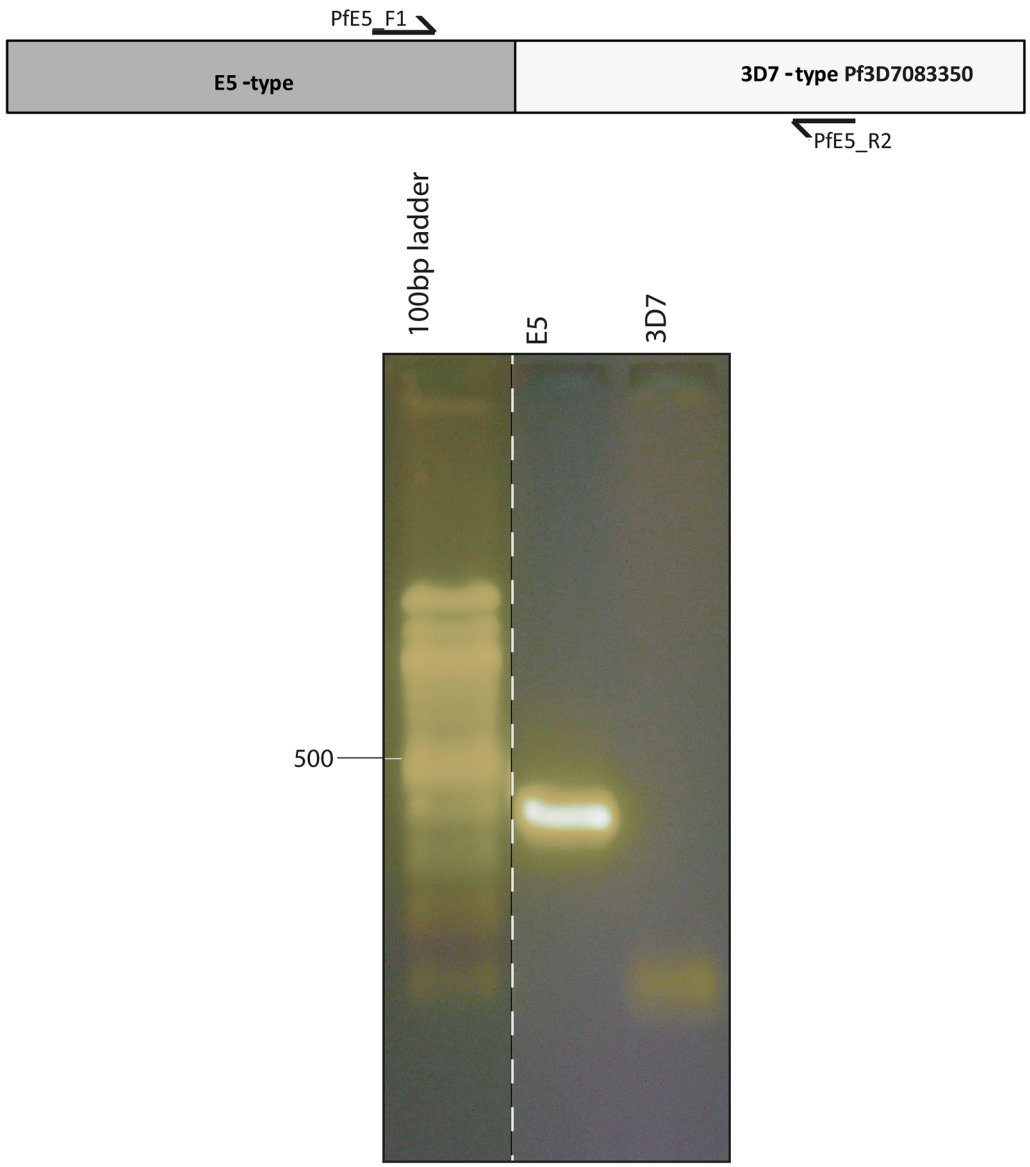

Fig. 4 New chimeric *var* gene on chromosome 14 of strain E5. The upper part of the figure displays the localisation of primers F1 and R2 to amplify the crossing point with PCR. The lower part of the figure displays the reaction product after amplification on E5 DNA. No reaction product was obtained after amplification on 3D7 DNA. The order of the gel was rearranged for reasons of clarity and is marked by a dotted line

true non-chromosomal recombination or simply random sharing of sequences among the global *var* gene population.

The VSA gene families of *P. falciparum* are located in subtelomeric regions and internal clusters. The boundaries between the VSA containing areas and the stable core genome have recently been newly defined by Otto et al. [11], through the analysis of 10 newly cultured field isolates from different geographic regions, by long read Pacific Bioscience sequencing technology. The beginning of the subtelomeric region was defined as the point were newly assembled genomes stop aligning with the 3D7 reference genome, however recombination within the subtelomeric regions was not able to be assessed because the

analysed strains were not genetically related. In contrast in this work the analysis of the 3D7-type subtelomeric and central areas of the E5 genome with short and long read WGS enabled an assessment of recombination in the VSA harbouring parts of the E5 genome. Analysis of the 3D7-like subtelomeres and internal clusters by short read WGS exhibited moderate SNP frequency and low coverage and thus suggested relatively frequent sequence alterations compared to the 3D7 refrence sequence. This likely reflects the difficulty of short read sequencing technology in the characterization of DNA sequences with high AT content and an abundance of repetitive DNA elements. In contrast long read WGS data of the subtelomeres and central clusters only identified one large scale recombination event showing that most of the 3D7-type subtelomeric sequences were indeed co-linear with the original 3D7 sequences. Together these data indicate that the majority of subtelomeres of *P. falciparum* are highly conserved across progeny from genetic crosses and that long read sequencing technology is more appropriate for the characterization of the genome areas harbouring VSA gene families.

3D7 and E5 both originate from the same NF54 culture and, therefore, have been in tissue culture for approximately the same time. The highly conserved nature of the E5 genome parts harbouring the 3D7-VSA gene families suggests that mitotic non-chromosomal recombination alone is insufficient to explain the global genetic diversity of the *var* gene family [65]. This suggests that the selective pressure of the host immune system is essential for the expansion of parasite populations with new chimeric *var* genes and thus for the generation of the seemingly endless diversity of the global *var* gene repertoire. Furthermore, the high degree of genetic diversity in the *rifin/stevor* gene families indicates that these non-PfEMP1 VSAs may be under similar diversifying selection as the *var* gene family [29–32].

Larger studies of progeny from natural genetic crosses with long read sequencing technology are necessary to examine the possible role of acquired immunity in the generation the *var* gene and *rifin/stevor* genetic diversity at the population level.

While there has been a long standing interest in the analysis of VSA families from different laboratory strains, recently field isolate VSA gene families have moved into the focus. In this context it has become clear that progeny of natural genetic crosses that show IBD are far more prevalent than previously thought [66].

In order to establish a method that can reliably differentiate between different progeny of a natural genetic cross, a set of 84 MS primers from the NIH database was evaluated for its ability to identify the 3D7 and non-3D7 parts of the E5 genome. 27 MS primers resulted in erroneous genotyping with the PCR conditions applied in this work. This is likely due to the fact that one standardized set of PCR conditions was applied for all primers and no attempts to optimize individual reaction conditions were made. However, even with these standard PCR conditions, 54 of 57 MS genotyping results were confirmed by WGS. 3 MS loci (*ebp, hrp2* and C12M30) showed the same alleles in E5 and 3D7, despite being located in the non-3D7 part of E5. Two of these MS were located within the open reading frames of *ebp* and *hrp2* indicating that these genes are not sufficiently diverse to distinguish between sibling parasites.

Comparative genotyping of E5 and 3D7 with 54 MS genotyping was accomplished within a few days and the use of different fluorophores for different MS on each chromosome enabled "head to head" genotyping of individual E5 and 3D7 chromosomes by multiplex PCR-reactions. This is the first time that MS genotyping has been directly compared to WGS. MS length differences of < 3 bp diffrences between E5 and 3D7 correctly identified the 3D7-type parts of the E5 genome. In some of these 3D7-type MS alleles the PCR fragment length differed from the in silico length of the respective MS in the 3D7 genome (version 3). This is most likely due to DNA slippage during PCR DNA replication. However, given the fact the PCR fragment length of these MS were identical after amplification of E5 and 3D7 this phenomenon appears to be higly reproducible and does not lead to erroneous genotyping.

Recently, Figan et al. [36] identified a set of 12 different microsatellite markers that reliably distinguish between progeny of 4 different experimental genetic crosses. The PCR conditions employed by Figan et al. and the PCR conditions in this work were almost identical suggesting that the two primer sets could be combined for rapid genotyping of field isolates.

SNP barcoding has recently emerged as a genome wide typing technique and has been used to investigate *Plasmodium* and the origin of its genotypes [67, 68]. The barcoding genotyping technique, which is based on a 23 single nucleotide polymorphisms (SNPs) and on high-quality raw sequence data [69], detects differences in the organelle genomes of *P. falciparum* and thus is not suitable for characterization of chromosomal inheritance. Similarly, another SNP assay developed some years earlier, is based on 24 SNP loci that are distributed unevenly across the genome, i.e. some chromosomes do not have SNP markers and others only 1 marker, thus tracking chromosomal cross over events is not possible [70].

SNP and WGS analysis are expensive and depend on the availability of high quality sequence data as well as extensive bioinformatic expertise. Therefore SNP and WGS can only be applied to subsets of *P. falciparum* lines

and are usually carried out in specialized centres with extensive resources. In contrast MS genotyping and data analysis can be carried out in smaller centres, potentially enabling investigator driven analysis and identification of *P. falciparum* strains most suitable for subsequent WGS analyses in specialized centres.

The vast majority of the confirmed 54 MS are located in the non-coding parts of the *P. falciparum* genome. Consequently, they are not under purifying selection and may reflect the underlying genetic plasticity of the *P. falciparum* genome more accurately than methods that are based on the detection of SNPs of coding regions.

Future analysis of natural *P. falciparum* cross progeny from semi-immune and non-immune individuals may allow insights into the factors that drive crossover and and non-crossover recombination in *P. falciparum*. In this context MS genotyping may be used to determine IBD in field isolate progeny and to identify parasites clones most suitable for WGS analysis.

Conclusion

The data presented in this work show that the *var* and *rifin/stevor* gene families represent the most diverse parts of the *P. falciparum* genome, but that the majority of the VSA genes are inherited without alteration in a Mendelian fashion. Furthermore, MS genotyping data correlate well with WGS data suggesting that MS genotyping can be employed to define IBD in progeny of natural *P. falciparum* crosses.

Additional files

Additional file 1. MS Primers that generated PCR products that could not be aligned to the 3D7 MS refrence sequence.

Additional file 2. Singelton genes within the E5 genome.

Additional file 3. ACT view showing a miss-assembly between E5 and 3D7 in the first *var* gene cluster of chromosome 4. The blue bars at the top represent the E5 bin contig, matching to an area on E5.

Additional file 4. ACT screenshot of *var* chimera, box. The top sequence (chromosome 8 of PfE5) is identical to Pf3D7 (middle track, chromosome 8), but does not finish with a telomer. The sequence left hand site of the *var* gene in 3D7 up to the chromosome end (telomer repeat marked with T) is shared to chromosome 14 of PfE5 (lowest track). The black blast hits between the identity of 95–100%. *For visualisation reasons, the chromosome 14 of PfE5 was complemented. So the *var* chimera in PfE5 is on the left hand site of chr14 and on the forward strand.

Authors' contributions

MF conceived the project and integrated the different datasets. EB conducted the MS genotyping experiments and analyzed the MS and the short read WGS data. TO assembled the E5 genome with short and long WGS data and performed the de novo annotation of the E5 genome based on long read sequencing data. All authors wrote the manuscript. All authors read and approved the final manuscript.

Author details

[1] Institute of Tropical Medicine, University of Tuebingen, Wilhelmstr. 27, 72074 Tuebingen, Germany. [2] Malaria Programme, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. [3] Present Address: Centre of Immunobiology, Institute of Infection, Immunity & Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK.

Acknowledgements

We are indebted to Chris Newbold without whom this paper would never have been possible. Chris facilitated the sequencing of E5 and brought us (EB, MF and TO) together as scientists. We thank Kathrin Vrankovijc, Johanna Volk, Sandra Dimonte and Andrea Weierich and all trainees for the their assistance in evaluating the microsatellite analysis by targeted Sanger sequencing and establishing the multiplex PCR assay and Matt Berriman and Mandy Sanders for the help with the Illumina and Pacific Bioscience sequencing. We thank Akhil Vaidya for providing us with the MS set that was utilized for the analysis of the 3D7xHB3 cross.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The NF54 E5 clone can be obtained from MF upon request. The E5 genome can be found at: ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/falciparum/E5/Version1.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

MF and EB were funded by the BMBF-Grant 01KA110 of the German ministry for education and research (BMBF). TO was supported by the Wellcome Trust (098051).

Publisher's Note

Received: 21 May 2018 Accepted: 3 October 2018
Published online: 22 October 2018

References

1. WHO | Malaria. WHO. http://www.who.int/mediacentre/factsheets/fs094/en/. Accessed 16 July 2017.
2. WHO | World Malaria Report 2016. WHO. http://www.who.int/malaria/publications/world_malaria_report_2016/en/. Accessed 1 Dec 2017.
3. Ponnudurai T, Leeuwenberg AD, Meuwissen JH. Chloroquine sensitivity of isolates of Plasmodium falciparum adapted to in vitro culture. Trop Geogr Med. 1981;33:50–4.
4. Frank M, Dzikowski R, Costantini D, Amulic B, Berdougo E, Deitsch K. Strict pairing of *var* promoters and introns is required for var gene silencing in the malaria parasite *Plasmodium falciparum*. J Biol Chem. 2006;281:9942–52.
5. Hall N, Pain A, Berriman M, Churcher C, Harris B, Harris D, et al. Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. Nature. 2002;419:527–31.
6. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 2002;419:498–511.
7. Frank M, Kirkman L, Costantini D, Sanyal S, Lavazec C, Templeton TJ, et al. Frequent recombination events generate diversity within the multi-copy variant antigen gene families of *Plasmodium falciparum*. Int J Parasitol. 2008;38:1099–109.
8. Vembar SS, Seetin M, Lambert C, Nattestad M, Schatz MC, Baybayan P, et al. Complete telomere-to-telomere *de novo* assembly of the

Plasmodium falciparum genome through long-read (>11 kb), single molecule, real-time sequencing. DNA Res Int J Rapid Publ Rep Genes Genomes. 2016;23:339–51.
9. Hamilton WL, Claessens A, Otto TD, Kekre M, Fairhurst RM, Rayner JC, et al. Extreme mutation bias and high AT content in *Plasmodium falciparum*. Nucleic Acids Res. 2017;45:1889–901.
10. Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, et al. Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. Genome Res. 2016;26:1288–99.
11. Otto TD, Böhme U, Sanders M, Reid A, Bruske EI, Duffy CW, et al. Long read assemblies of geographically dispersed *Plasmodium falciparum* isolates reveal highly structured subtelomeres. Wellcome Open Res. 2018;3:52.
12. Rubio JP, Thompson JK, Cowman AF. The var genes of *Plasmodium falciparum* are located in the subtelomeric region of most chromosomes. EMBO J. 1996;15:4069–77.
13. Hernandez-Rivas R, Mattei D, Sterkers Y, Peterson DS, Wellems TE, Scherf A. Expressed var genes are found in *Plasmodium falciparum* subtelomeric regions. Mol Cell Biol. 1997;17:604–11.
14. Fischer K, Horrocks P, Preuss M, Wiesner J, Wünsch S, Camargo AA, et al. Expression of var genes located within polymorphic subtelomeric domains of *Plasmodium falciparum* chromosomes. Mol Cell Biol. 1997;17:3679–86.
15. Cheng Q, Cloonan N, Fischer K, Thompson J, Waine G, Lanzer M, et al. stevor and rif are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. Mol Biochem Parasitol. 1998;97(1–2):161–76.
16. Fernandez V, Hommel M, Chen Q, Hagblom P, Wahlgren M. Small, clonally variant antigens expressed on the surface of the *Plasmodium falciparum*–infected erythrocyte are encoded by the rif gene family and are the target of human immune responses. J Exp Med. 1999;190:1393–404.
17. Sam-Yellowe TY, Florens L, Johnson JR, Wang T, Drazba JA, Le Roch KG, et al. A Plasmodium gene family encoding maurer's cleft membrane proteins: structural properties and expression profiling. Genome Res. 2004;14:1052–9.
18. Winter G, Kawai S, Haeggström M, Kaneko O, von Euler A, Kawazu S, et al. SURFIN is a polymorphic antigen expressed on *Plasmodium falciparum* merozoites and infected erythrocytes. J Exp Med. 2005;201:1853–63.
19. Su X, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, et al. The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. Cell. 1995;82:89–100.
20. Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, Feldman M, et al. Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. Cell. 1995;82:77–87.
21. Smith JD, Rowe JA, Higgins MK, Lavstsen T. Malaria's deadly grip: cytoadhesion of *Plasmodium falciparum* infected erythrocytes. Cell Microbiol. 2013;15:1976–83.
22. Kyes S, Horrocks P, Newbold C. Antigenic variation at the infected red cell surface in malaria. Annu Rev Microbiol. 2001;55:673–707.
23. Bernabeu M, Danziger SA, Avril M, Vaz M, Babar PH, Brazier AJ, et al. Severe adult malaria is associated with specific PfEMP1 adhesion types and high parasite biomass. Proc Natl Acad Sci USA. 2016;113:E3270–9.
24. Nunes-Silva S, Dechavanne S, Moussiliou A, Pstrąg N, Semblat J-P, Gangnard S, et al. Beninese children with cerebral malaria do not develop humoral immunity against the IT4-VAR19-DC8 PfEMP1 variant linked to EPCR and brain endothelial binding. Malar J. 2015;14:493.
25. Avril M, Bernabeu M, Benjamin M, Brazier AJ, Smith JD. Interaction between endothelial protein C receptor and intercellular adhesion molecule 1 to mediate binding of *Plasmodium falciparum*-infected erythrocytes to endothelial cells. mBio. 2016;7:e00615–6.
26. Salanti A, Staalsoe T, Lavstsen T, Jensen ATR, Sowa MPK, Arnot DE, et al. Selective upregulation of a single distinctly structured var gene in chondroitin sulphate A-adhering *Plasmodium falciparum* involved in pregnancy-associated malaria. Mol Microbiol. 2003;49:179–91.
27. Lau CKY, Turner L, Jespersen JS, Lowe ED, Petersen B, Wang CW, et al. Structural conservation despite huge sequence diversity allows EPCR binding by the PfEMP1 family implicated in severe childhood malaria. Cell Host Microbe. 2015;17:118–29.
28. Turner L, Lavstsen T, Berger SS, Wang CW, Petersen JEV, Avril M, et al. Severe malaria is associated with parasite binding to endothelial protein C receptor. Nature. 2013;498:502–5.
29. Chan J-A, Howell KB, Reiling L, Ataide R, Mackintosh CL, Fowkes FJI, et al. Targets of antibodies against *Plasmodium falciparum*–infected erythrocytes in malaria immunity. J Clin Invest. 2012;122:3227–38.
30. Bruske EI, Dimonte S, Enderes C, Tschan S, Flötenmeyer M, Koch I, et al. In Vitro variant surface antigen expression in *Plasmodium falciparum* parasites from a semi-immune individual is not correlated with var gene transcription. PLoS ONE. 2016;11:e0166135.
31. Tan J, Pieper K, Piccoli L, Abdi A, Perez MF, Geiger R, et al. A LAIR1 insertion generates broadly reactive antibodies against malaria variant antigens. Nature. 2016;529:105–9.
32. Niang M, Yan Yam X, Preiser PR. The *Plasmodium falciparum* STEVOR multigene family mediates antigenic variation of the infected erythrocyte. PLoS Pathog. 2009;5:e1000307.
33. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004;5:435–45.
34. Wellems TE, Walker-Jonah A, Panton LJ. Genetic mapping of the chloroquine-resistance locus on *Plasmodium falciparum* chromosome 7. Proc Natl Acad Sci U S A. 1991;88:3382–6.
35. Hayton K, Gaur D, Liu A, Takahashi J, Henschen B, Singh S, et al. Erythrocyte binding protein PfRH5 polymorphisms determine species-specific pathways of *Plasmodium falciparum* invasion. Cell Host Microbe. 2008;4:40–51.
36. Figan CE, Sá JM, Mu J, Melendez-Muniz VA, Liu CH, Wellems TE. A set of microsatellite markers to differentiate *Plasmodium falciparum* progeny of four genetic crosses. Malar J. 2018;17:60.
37. Roper C, Pearce R, Nair S, Sharp B, Nosten F, Anderson T. Intercontinental spread of pyrimethamine-resistant malaria. Science. 2004;305:1124.
38. Wootton JC, Feng X, Ferdig MT, Cooper RA, Mu J, Baruch DI, et al. Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. Nature. 2002;418:320–3.
39. Anderson TJC, Su X-Z, Bockarie M, Lagog M, Day KP. Twelve microsatellite markers for characterization of *Plasmodium falciparum* from finger-prick blood samples. Parasitology. 1999;119:113–25.
40. Nabet C, Doumbo S, Jeddi F, Konaté S, Manciulli T, Fofana B, et al. Genetic diversity of *Plasmodium falciparum* in human malaria cases in Mali. Malar J. 2016;15:353.
41. Hong NV, Delgado-Ratto C, Thanh PV, Van den Eede P, Guetens P, Binh NTH, et al. Population genetics of *Plasmodium vivax* in four rural communities in Central Vietnam. PLoS Negl Trop Dis. 2016;10:e0004434.
42. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, et al. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. Nature. 2000;407:1018–22.
43. Frank M, Dzikowski R, Amulic B, Deitsch K. Variable switching rates of malaria virulence genes are associated with chromosomal position and gene subclass. Mol Microbiol. 2007;64:1486–98.
44. Enderes C, Kombila D, Dal-Bianco M, Dzikowski R, Kremsner P, Frank M. Var Gene promoter activation in clonal *Plasmodium falciparum* isolates follows a hierarchy and suggests a conserved switching program that is independent of genetic background. J Infect Dis. 2011;204:1620–31.
45. Frank M, Lehners N, Mayengue PI, Gabor J, Dal-Bianco M, Kombila DU, et al. A 13-year analysis of *Plasmodium falciparum* populations reveals high conservation of the mutant pfcrt haplotype despite the withdrawal of chloroquine from national treatment guidelines in Gabon. Malar J. 2011;10:304.
46. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods. 2009;6:291–5.
47. Logan-Klumpler FJ, De Silva N, Boehme U, Rogers MB, Velarde G, McQuillan JA, et al. GeneDB—an annotation database for pathogens. Nucleic Acids Res. 2012;40(Database issue):D98–108.
48. DePristo MA, Banks E, Poplin RE, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.

49. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009;25:1968–9.
50. Otto TD, Sanders M, Berriman M, Newbold C. Iterative correction of reference nucleotides (iCORN) using second generation sequencing technology. Bioinformatics. 2010;26:1704–7.
51. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol. 2015;16:294.
52. Steinbiss S, Silva-Franco F, Brunk B, Foth B, Hertz-Fowler C, Berriman M, et al. Companion: a web server for annotation and analysis of parasite genomes. Nucleic Acids Res. 2016;44(Web Server issue):W29–34.
53. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. Artemis: sequence visualization and annotation. Bioinformatics. 2000;16:944–5.
54. Carver T, Harris SR, Otto TD, Berriman M, Parkhill J, McQuillan JA. BamView: visualizing and interpretation of next-generation sequencing read alignments. Brief Bioinf. 2013;14:203–12.
55. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: the artemis comparison tool. Bioinformatics. 2005;21:3422–3.
56. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
57. Jiang H, Li N, Gopalan V, Zilversmit MM, Varma S, Nagarajan V, et al. High recombination rates and hotspots in a *Plasmodium falciparum* genetic cross. Genome Biol. 2011;12:R33.
58. Deitsch KW, del Pinal A, Wellems TE. Intra-cluster recombination and var transcription switches in the antigenic variation of *Plasmodium falciparum*. Mol Biochem Parasitol. 1999;101:107–16.
59. Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, Lavstsen T. *Plasmodium falciparum* erythrocyte membrane protein 1 diversity in seven genomes—divide and conquer. PLoS Comput Biol. 2010;6:e1000933.
60. Sander AF, Lavstsen T, Rask TS, Lisby M, Salanti A, Fordyce SL, et al. DNA secondary structures are associated with recombination in major *Plasmodium falciparum* variable surface antigen gene families. Nucleic Acids Res. 2014;42:2270–81.
61. Bopp SER, Manary MJ, Bright AT, Johnston GL, Dharia NV, Luna FL, et al. Mitotic evolution of *Plasmodium falciparum* shows a stable core genome but recombination in antigen families. PLoS Genet. 2013;9:e1003293.
62. Claessens A, Hamilton WL, Kekre M, Otto TD, Faizullabhoy A, Rayner JC, et al. Generation of antigenic diversity in *Plasmodium falciparum* by structured rearrangement of *var* genes during mitosis. PLoS Genet. 2014;10:e1004812.
63. Calhoun SF, Reed J, Alexander N, Mason CE, Deitsch KW, Kirkman LA. Chromosome end repair and genome stability in *Plasmodium falciparum*. mBio. 2017;8:e00547–617.
64. Bull PC, Buckee CO, Kyes S, Kortok MM, Thathy V, Guyah B, et al. *Plasmodium falciparum* antigenic variation. Mapping mosaic var gene sequences onto a network of shared, highly polymorphic sequence blocks. Mol Microbiol. 2008;68:1519–34.
65. Barry AE, Leliwa-Sytek A, Tavul L, Imrie H, Migot-Nabias F, Brown SM, et al. Population genomics of the immune evasion (*var*) genes of *Plasmodium falciparum*. PLoS Pathog. 2007;3:e34.
66. Taylor AR, Schaffner SF, Cerqueira GC, Nkhoma SC, Anderson TJC, Sriprawat K, et al. Quantifying connectivity between local *Plasmodium falciparum* malaria parasite populations using identity by descent. PLoS Genet. 2017;13:e1007065.
67. Su X. Tracing the geographic origins of *Plasmodium falciparum* malaria parasites. Pathog Glob Health. 2014;108:261–2.
68. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, et al. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. Nature. 2012;487:375–9.
69. Preston MD, Campino S, Assefa SA, Echeverry DF, Ocholla H, Amambua-Ngwa A, et al. A barcode of organellar genome polymorphisms identifies the geographic origin of *Plasmodium falciparum* strains. Nat Commun. 2014;5:4052.
70. Daniels R, Volkman SK, Milner DA, Mahesh N, Neafsey DE, Park DJ, et al. A general SNP-based molecular barcode for *Plasmodium falciparum* identification and tracking. Malar J. 2008;7:223.