


RESEARCH

Open Access



# Population genetic analysis of the *Plasmodium falciparum* circumsporozoite protein in two distinct ecological regions in Ghana

Elikplim A. Amegashie<sup>1</sup>, Lucas Amenga-Etego<sup>2†</sup>, Courage Adobor<sup>3</sup>, Peter Ogoti<sup>1</sup>, Kevin Mbogo<sup>1</sup>, Alfred Amambua-Ngwa<sup>4†</sup> and Anita Ghansah<sup>3\*†</sup> 

## Abstract

**Background:** Extensive genetic diversity in the *Plasmodium falciparum* circumsporozoite protein (PfCSP) is a major contributing factor to the moderate efficacy of the RTS,S/AS01 vaccine. The transmission intensity and rates of recombination within and between populations influence the extent of its genetic diversity. Understanding the extent and dynamics of PfCSP genetic diversity in different transmission settings will help to interpret the results of current RTS,S efficacy and Phase IV implementation trials conducted within and between populations in malaria-endemic areas such as Ghana.

**Methods:** *Pfcs*p sequences were retrieved from the Illumina-generated paired-end short-read sequences of 101 and 131 malaria samples from children aged 6–59 months presenting with clinical malaria at health facilities in Cape Coast (in the coastal belt) and Navrongo (Guinea savannah region), respectively, in Ghana. The sequences were mapped onto the 3D7 reference strain genome to yield high-quality genome-wide coding sequence data. Following data filtering and quality checks to remove missing data, 220 sequences were retained and analysed for the allele frequency spectrum, genetic diversity both within the host and between populations and signatures of selection. Population genetics tools were used to determine the extent and dynamics of *Pfcs*p diversity in *P. falciparum* from the two geographically distinct locations in Ghana.

**Results:** *Pfcs*p showed extensive diversity at the two sites, with the higher transmission site, Navrongo, exhibiting higher within-host and population-level diversity. The vaccine strain C-terminal epitope of *Pfcs*p was found in only 5.9% and 45.7% of the Navrongo and Cape Coast sequences, respectively. Between 1 and 6 amino acid variations were observed in the TH2R and TH3R epitope regions of PfCSP. Tajima's D was negatively skewed, especially for the population from Cape Coast, given the expected historical population expansion. In contrast, a positive Tajima's D was observed for the Navrongo *P. falciparum* population, consistent with balancing selection acting on the immunodominant TH2R and TH3R vaccine epitopes.

\*Correspondence: aghansah@noguchi.ug.edu.gh

<sup>†</sup>Lucas Amenga-Etego, Alfred Amambua-Ngwa and Anita Ghansah equally contributed to this work

<sup>3</sup> Parasitology Department, Noguchi Memorial Institute of Medical Research, University of Ghana, , Legon, Ghana

Full list of author information is available at the end of the article



**Conclusion:** The low frequencies of the *Pf*csp vaccine haplotype in the analysed populations indicate a need for additional molecular and immuno-epidemiological studies with broader temporal and geographic sampling in endemic populations targeted for RTS,S application. These results have implications for the efficacy of the vaccine in Ghana and will inform the choice of alleles to be included in future multivalent or chimeric vaccines.

**Keywords:** *Plasmodium falciparum* circumsporozoite protein, Genetic diversity, Selection, Within-host diversity

## Background

Stagnation in the decline of malaria over the last 5 years indicates that global malaria elimination targets may not be achieved without the addition of a broadly effective vaccine to complement the panel of available malaria control tools [1]. However, it has taken over 15 years to finally license a moderately efficacious malaria vaccine for implementation due to extreme levels of antigenic diversity of most vaccine candidates, which reduces their efficacy across a broad range of evolving natural parasite populations.

Efficacy data from a Phase III clinical trial conducted across 11 sites in 7 African countries in children (aged 5 to 17 months) and infants (aged 6–12 weeks) revealed that the vaccine conferred moderate protective efficacy against clinical disease and severe malaria which waned over time [2]. The vaccine conferred only 36.3% protection against clinical malaria and 32.2% against severe malaria in children aged 5–17 months who received 3 primary doses of RTS,S with a booster in the 20th month [2]. The European Medicines Agency gave a favourable scientific opinion in 2015, indicating how the benefits of protective immunity outweigh the risk and the potentially high impact this moderate efficacy could have, given the huge disease burden [3]. Subsequently, Ghana, Kenya and Malawi were selected for pilot Phase IV implementation trials that are currently underway, carried out by the Malaria Vaccine Implementation Programme (MVIP) led by the World Health Organization (WHO) [4].

The RTS,S vaccine is a malaria subunit vaccine that is formulated from a fragment of the circumsporozoite protein (CSP) of *Plasmodium falciparum* 3D7 laboratory strain fused with the Hepatitis B surface antigen and the AS01 adjuvant [5]. For cell-mediated immunity, RTS,S includes a fragment of the central NANP-NVDP repeat polymorphic B-cell epitope region and a highly polymorphic C-terminal non-repeat epitope region of PfCSP, which covers CD4<sup>+</sup> and CD8<sup>+</sup> T-cell epitopes denoted as TH2R and TH3R, respectively [6]. Several studies have reported high levels of polymorphism in the T-cell epitopes within the C-terminal region of PfCSP in natural parasite populations [7–10]. Although there are variations in the immunodominant central repeat region (CRR), it was hoped that antibodies targeting a single dominant epitope based on the

tetrapeptide repeat NANP would provide strain-surpassing immunity. This hope was strengthened by the findings of a molecular epidemiology study in African children that showed no evidence of naturally acquired strain-specific immunity to different variants of CSP obtained using the 454 sequencing platform [8]. In addition, initial ancillary studies on Phase II clinical trials conducted at three sites, including The Gambia, Kenya and Mozambique, revealed that the immune protection conferred by the RTS,S/AS02 vaccine was not strain-specific even after vaccination [11–14]. However, these studies were based on only a few hundred isolates and were not statistically powered to detect moderate effects, such as the strain-specific immune response of the vaccine.

Subsequently, an ancillary next-generation deep sequencing analysis of Phase III trial samples in 2015 showed that the vaccine indeed conferred partial protection against clinical malaria for strain-specific vaccine alleles (50.3%) and poor protection against mismatched strains (33.3%) [15]. Additionally, recent studies of the population structure of *Pf*csp suggest that geographically variable levels of diversity and geographic restriction of specific subgroups may have an impact on the efficacy of *Pf*csp-based malaria vaccines in specific geographic regions [7, 16]. In particular, extreme global genetic diversity of *Pf*csp strains has been reported, with the 3D7 vaccine strain being found only in approximately 5.0% and 0.2% in some African and Asian countries, respectively [16].

The need to explore the extent of genetic diversity and the natural dynamics of malaria vaccine antigens in endemic areas where vaccines will be deployed is a point of focus due to the polymorphic nature of *P. falciparum* antigens [15–17]. Furthermore, evolutionary factors such as selection operating on parasites differ locally owing to varying transmission patterns, ecology and degrees of acquired immunity in humans [18]. Therefore, further characterization of the genetic diversity of immune epitopes of vaccine antigens is important, especially in regions such as Ghana, where the vaccine is undergoing the Phase IV implementation trial. This should provide a broader assessment of the extent to which the local natural diversity could impact efficacy and wider implementation.

Malaria transmission in Ghana is generally perennial but with marked seasonal effects that vary with the local ecology and overall transmission intensity [19]. For control purposes, malaria transmission across Ghana has been eco-epidemiologically classified into three main zones: a forest ecology zone, with perennial but high transmission during the rainy season (May–August and October–November); a northern/Guinea savannah zone, showing seasonal and intense malaria transmission during the rainy season (June–October) but also periods of very low transmission during the dry season; and a coastal savannah zone, with low to moderate perennial transmission and a marked seasonal effect during the rainy season [20].

The implementation trial of RTS,S vaccine is being conducted in three regions, namely, the Brong-Ahafo and Volta regions in the forest ecology zone and the Central region in the coastal belt, with varying transmission levels. Understanding the extent and drivers of diversity in these regions could also have a profound impact on improving the design of future circumsporozoite protein-based vaccines. Using paired-end short-read sequences of the *Pfcspr* in parasite populations from two geographically distinct sites in Ghana, within-host diversity (complexity of infection) and the extent of population-specific haplotype diversity of the c-terminal region of *Pfcspr* encompassing the TH2R and TH3R epitopes were investigated. Information on diversity most relevant to vaccine

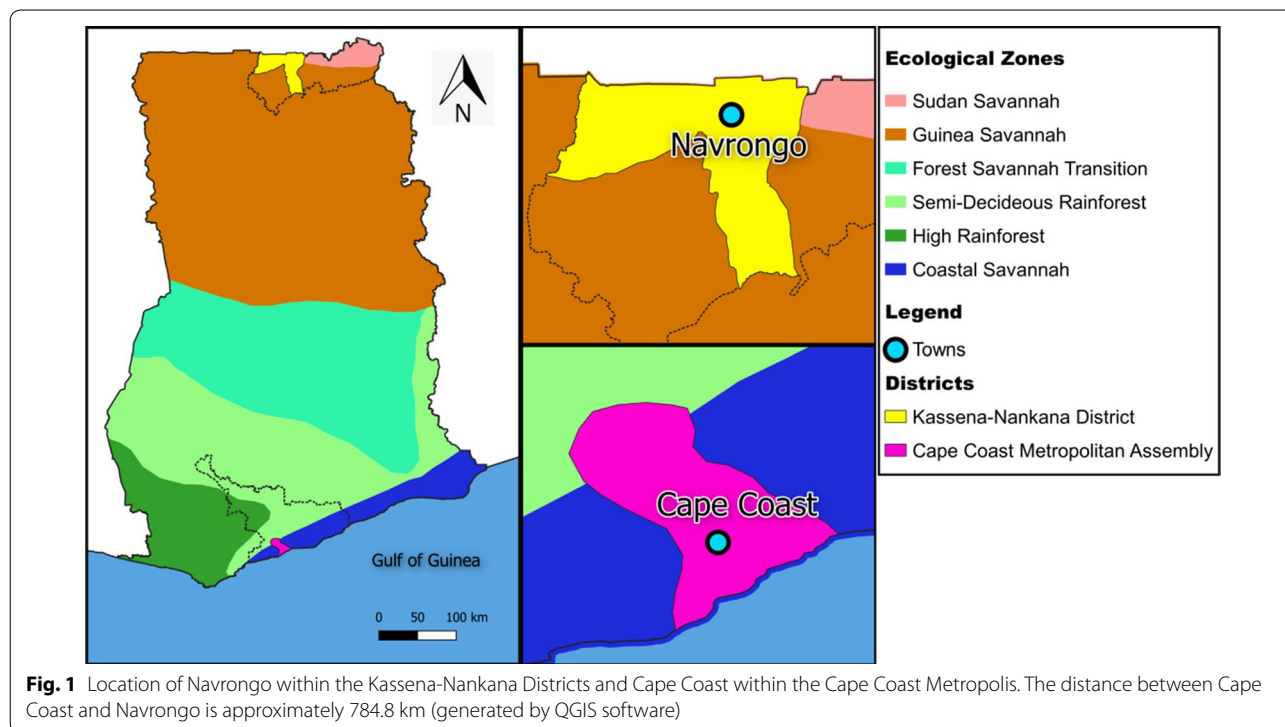
escape and cross-protection are provided. *PfCSP* amino acid diversity and conservation were further explored. In addition, evidence of selection on *Pfcspr* that could be driving and sustaining the observed diversity was assessed.

**Methods**

**Study area**

The study was conducted at two sites, the Cape Coast Metropolitan area, where Cape Coast is the main township, and the Kassena-Nankana districts (KNDs), where Navrongo is the main township (Fig. 1). Cape Coast is located in the Southern coastal savannah region, showing low to moderate perennial malaria transmission with a marked seasonal effect during the rainy season (May–August and October–November). The estimated annual entomological inoculation rate (EIR) is fewer than 50 infective bites per person per year [20]. The KNDs are located in the Upper East Region of Ghana with a Guinea savannah vegetation. Malaria is perennial in the KNDs, with high seasonal malaria transmission during the rainy season (June to October) and minimal transmission during the rest of the year, which are relatively dry months. The estimated annual EIR for the KNDs is up to 157 infective bites per person per year [21].

In Cape Coast, *P. falciparum* parasites were isolated from 101 children (aged 6–59 months) living within the municipality and presenting with clinical malaria at the



Cape Coast District hospital. Samples were collected during the major rainy season (May–August) in 2013. In Navrongo, *P. falciparum* parasites were isolated from 131 children aged between 12–59 months who lived in the KNDs and presented with *P. falciparum* clinical malaria at health facilities in the KNDs in the years 2010 (January to October), 2011 (January to February) and 2013 (August to October) during both the dry and wet seasons. At both study sites, children presenting with fever, i.e., an axillary temperature  $\geq 37.5^\circ$  or a history of fever during the previous 24 h, were screened with malaria rapid diagnostic test (RDT). Blood smears were prepared for RDT-positive individuals and *P. falciparum* asexual parasites were identified by microscopy. Venous blood (2–5 mL) from *P. falciparum*-infected patients who gave consent was collected and archived.

#### Genomic DNA extraction and sequencing

Genomic DNA was extracted using the QiaAmp DNA prep kit (Qiagen, Valencia, CA) following the manufacturer's protocol, and the confirmation of *P. falciparum*-positive samples was performed by amplification using nested PCR with specific primers [22] (see Additional file 1). The Genomic DNA was submitted to the Wellcome Trust Sanger Institute Hinxton, UK, for whole-genome sequencing using the Illumina HiSeq platform as part of the MalariaGEN community project. Illumina sequencing libraries (200 bp insert) were aligned to the reference *P. falciparum* 3D7 genome, after which variant calling was conducted via the customized GATK pipeline. Each sample was genotyped for 797,000 polymorphic biallelic coding SNPs across the genome ensuring a minimum of  $5 \times$  paired-end coverage across each variant per sample. The dominant allele was retained in the genotype file at loci with mixed reads (reference/non-reference). The genotypes were assigned denoting the reference and non-reference nucleotides as 0 and 1, respectively. Polymorphic sites with low call rates and those in hypervariable, telomeric and repetitive sequence regions were excluded.

#### Sequence acquisition and pre-processing

Genome sequences from Navrongo and Cape Coast were mined from the MalariaGEN *Plasmodium falciparum* Community (Pf3k) Project release 5.1 Database [23] in variant call format (VCF). Genetic variants on chromosome 3 were retrieved from both Navrongo and Cape Coast. In the VCF file for Cape Coast, all genotypes at each SNP position were monoallelic (monoclonal); biallelic genotypes were modelled using a custom Python script. This process was based on the approach of the MalariaGEN Pf3k Project, where loci with mixed allele calls were modelled using the read and allelic depth

[24]. Briefly, to account for PCR errors, the genotypes of SNPs with a read depth  $< 5$  were not determined. At SNP positions with a read depth  $\geq 5$ , the samples were genotyped as heterozygous if the allelic depth of both alleles was  $\geq 2$ . The remaining alleles were genotyped as either the homozygote reference allele or homozygote alternative/derived allele.

Data for both populations were filtered to obtain only biallelic SNPs using Bcftools v1.9 and quality checked as follows: only SNPs that passed all VCF filters were retained. Isolates with  $> 10\%$  missing SNPs were excluded, followed by the removal of SNPs with  $> 5\%$  missing data using PLINK v1.9 [25]. Furthermore, heterozygosity was calculated, and 8 isolates with outlier heterozygosity within the Cape Coast population were excluded. No outlier heterozygosity was observed in Navrongo. SNPs with a minor allele frequency (MAF)  $< 1\%$  were removed. The remaining missing SNPs were imputed and phased using Beagle v5.1 [26]. After quality control, in the Cape Coast dataset, 2504 SNPs out of 26,156 on chromosome 3 and 92 out of 101 samples remained. However, the Navrongo dataset retained 1954 out of 43,199 SNPs on chromosome 3 and 128 out of 131 samples. *Pfcspl* was then extracted from chromosome 3 (position: 221,323–222,516) and 13 and 22 SNPs at the selected CSP loci were retained for Cape Coast and Navrongo, respectively.

#### Population genetics analysis

##### Minor allele frequency distribution

Prior to the removal of rare alleles (MAF  $\leq 0.01$ ), the minor allele frequency distribution for all putative SNPs ( $n = 90$ ) within *Pfcspl* for both Cape Coast ( $n = 35$  SNPs) and Navrongo ( $n = 55$  SNPs) *P. falciparum* isolates was determined using Plink1.9. The MAF is the frequency with which the second most common allele occurs at a given SNP position in a population.

##### Within-host parasite diversity estimation and statistical analysis

The genetic diversity within the individuals was assessed by estimating Wright's inbreeding co-efficient ( $F_{ws}$ ). For this analysis, the within-host diversity of *Pfcspl*, which refers to the number of different *Pfcspl* strains contained within an individual infection, was estimated. The retained variants (13 and 22 SNPs) from the 92 Cape Coast isolates and 128 Navrongo isolates were used for this analysis.

The  $F_{ws}$  metric estimates the heterozygosity of parasites ( $H_w$ ) within an individual relative to the heterozygosity within a parasite population ( $H_s$ ) using the read count of alleles.  $F_{ws}$  metric calculation for each sample was performed using the following equation:

$$F_{ws} = 1 - H_W/H_S$$

where  $H_W$  refers to the allele frequency of each unique allele found at specific loci of the parasite sequences within the individual, and  $H_S$  refers to the corresponding allele frequencies of those unique alleles within the population [27, 28].  $F_{ws}$  ranges from 0 to 1; a low  $F_{ws}$  value indicates low inbreeding rates within the parasite population and thus high within-host diversity relative to the population. An  $F_{ws}$  threshold  $\geq 0.95$  indicates samples with clonal (single strain) infections, while samples with an  $F_{ws} < 0.95$  are considered highly likely to come from mixed strain infections, indicating within-host diversity.  $F_{ws}$  was calculated using an R package, *moimix* [29]. Samples with clonal infections were used for selection analysis. The Pearson chi-squared test was used to measure the statistical significance of any differences observed in the within-host diversity estimates between the population pair. The test was performed using R software with  $P$  values  $< 0.05$  considered statistically significant.

#### Genetic diversity within parasite populations

The haplotype diversity (the number of two random strains within the population having different haplotypes) of *Pf**csp* in each population was determined by exploring the variants in the C-terminal region of the gene (909–1140 bp). One hundred and eighty four *Pf**csp* FASTA DNA sequences were re-constructed with the retained variants (13 SNPs) from the 92 Cape Coast isolates and 256 DNA sequences (22 SNPs) from the 128 Navrongo isolates using an in-house Python script.

The following metrics were then used to assess the diversity of the *Pf**csp* C-terminus within each parasite population using *DnaSP* software (version 6.10.01) [30]: number of sequences ( $n$ ), number of haplotypes ( $h$ ), segregating sites ( $S$ ), average number of pairwise nucleotide differences ( $K$ ), nucleotide diversity ( $\pi$ ) and haplotype diversity ( $H_d$ ) [31, 32].

To assess the genealogical relationships between *Pf**csp* C-terminal haplotypes found in Navrongo and Cape Coast, a network based on the method described by Templeton, Crandall, and Sing (TCS) [33, 34] was constructed using *PopArt* [35]. The haplotypes were denoted as 3D7, Hap 2 up to Hap 66 in the network.

In addition, amino acid haplotypes within each population were explored by translating all 440 *Pf**csp* DNA sequences (Cape Coast (184) and Navrongo (256)) into amino acid sequences and comparing them to the 3D7 reference strain (0304600.1, PlasmoDB [36]) using in-house Python scripts. The frequency of TH2R 311–327 amino acid (PSDKHIKEYLNKIQNSL) and TH3R 352–363 amino acid (NKPKDELDDYAND) haplotypes in each

parasite population were determined also using a customized Python script and plotted.

#### Population differentiation and structure analysis

The Wright Fixation index ( $F_{st}$ ) and principal component analysis (PCA) were used for population differentiation and structure analyses. To reduce bias in  $F_{st}$  analysis and PCA, SNPs (from the 2504 Cape Coast chromosome 3 retained SNPs and the 1954 Navrongo retained SNPs) with pairwise linkage disequilibrium (LD) values  $r^2 > 0.5$  within a window of 100 bp in the entire chromosome 3 dataset were pruned out using a step size of 10. The remaining SNPs set at chromosome 3 shared between the populations after pruning was 516, of which 10 were *Pf**csp* SNPs.

The *Pf**csp* SNPs were then used to estimate  $F_{st}$  and population structure. The Weir and Cockerham  $F_{st}$  per SNP between Cape Coast and Navrongo parasite isolates was calculated using *Vcftools* v0.1.5 [37] and population structure by PCA was performed using *smartpca* (Cambridge, MA, USA) in *EIGENSOFT* package v6.1.3 [38]. Principal components were computed with the number of outlier removal iterations set at 10 while maintaining other parameters. In all, 10 PCs were computed with 5 and 9 outlier samples removed from the 92 and 128 isolates from Cape Coast and Navrongo, respectively. Thus, there remained 83 samples in the Cape Coast population and 123 samples in the Navrongo population after outlier samples were removed.

#### Signatures of selection

To test for SNP neutrality, the Tajima's D statistical test [39] was performed in sliding windows with a size of 100 bp and a step size of 10 with *Pf**csp* monoclonal samples from each population using *Vcftools* v0.1.5. Tajima's D test compares the average pairwise differences ( $\pi$ ) and the total number of segregating sites ( $S$ ). Negative values indicate directional or purifying selection, while positive values indicate balancing the selection.

To detect loci likely to be under recent positive selection in the Cape Coast and Navrongo monoclonal chromosome 3 isolates, the standardized integrated haplotype score ( $|iHS|$ ) was calculated for each SNP with an  $MAF > 0.05$  on chromosome 3 (358 out of the 2504 and 608 out of the 1954 remaining SNPs from Cape Coast and Navrongo, respectively) [40]. Again, for the purpose of this analysis, the  $F_{ws}$  metric was used to estimate these monoclonal chromosome 3 isolates in the retained variants within the chromosome 3 region (2504 in Cape Coast and 1954 SNPs in Navrongo).

$|iHS|$  measures the amount of extended haplotype homozygosity (EHH) at a given SNP in the ancestral allele relative to the derived allele [40]. The reference

and alternate alleles were characterized as ancestral and derived alleles, respectively. This was performed in R using the *rehh* package v2.0.4 [41]. Genomic regions under positive selection were identified as those with multiple SNPs showing  $|iHS|$  values  $> 3$  and provided the focal SNPs for extended haplotype homozygosity (EHH) analysis. EHH for both the reference and alternate alleles was calculated, and bifurcation plots were generated to visualize the decay of EHH at increasing distances from the focal SNP loci [42] using *rehh* package v2.0.4 in R.

**Results**

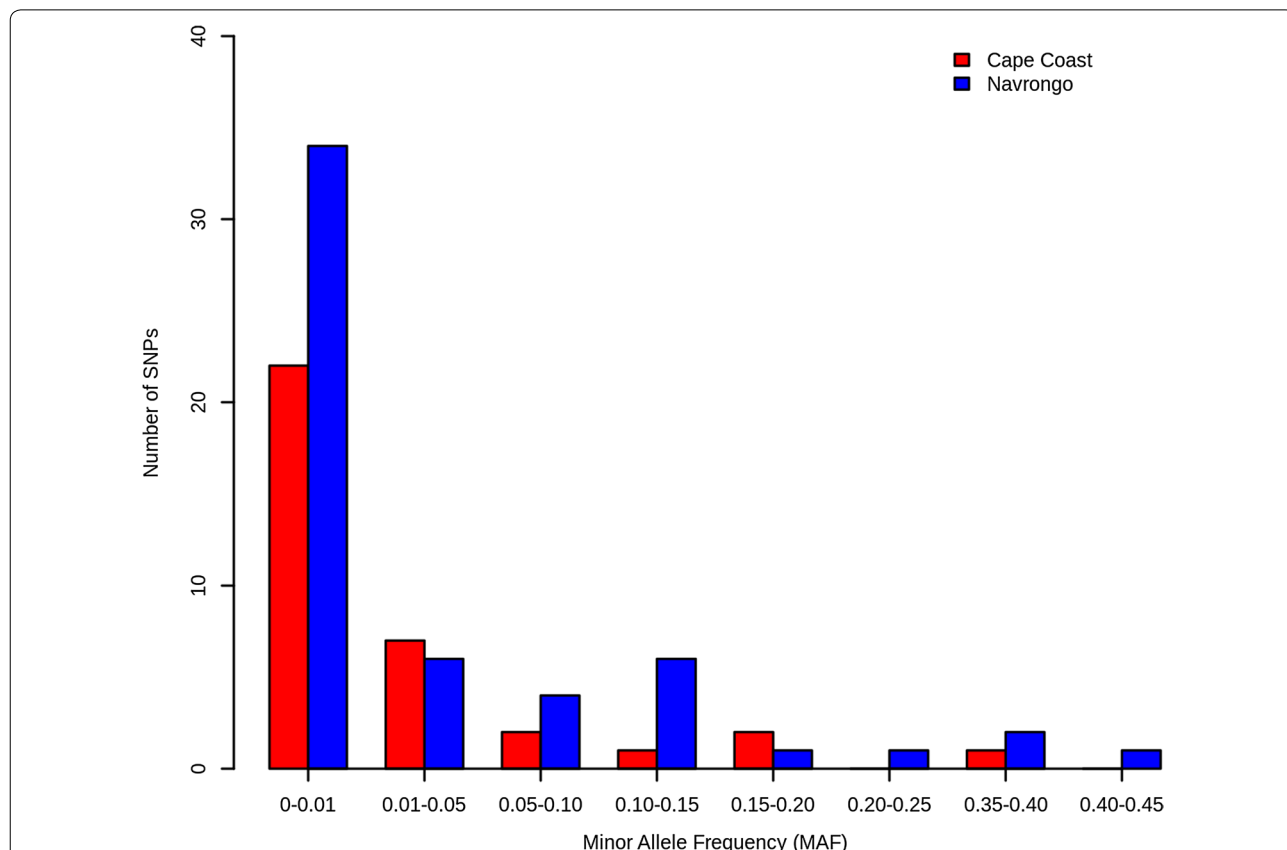
**Minor allele frequency distribution of *Pf**csp***

A total of 90 SNPs within *Pf**csp* were analysed for the minor allele frequency (MAF). The *P. falciparum* population from Navrongo showed more variability in *Pf**csp* (55 SNPs) than the Cape Coast population (35 SNPs) (Fig. 2). The allele frequency distribution of all putative SNPs within the *Pf**csp* loci ranged from 0.001–0.45 in Navrongo and 0.001–0.40 in Cape Coast (Fig. 2). As expected for natural *P. falciparum* populations in Africa (high

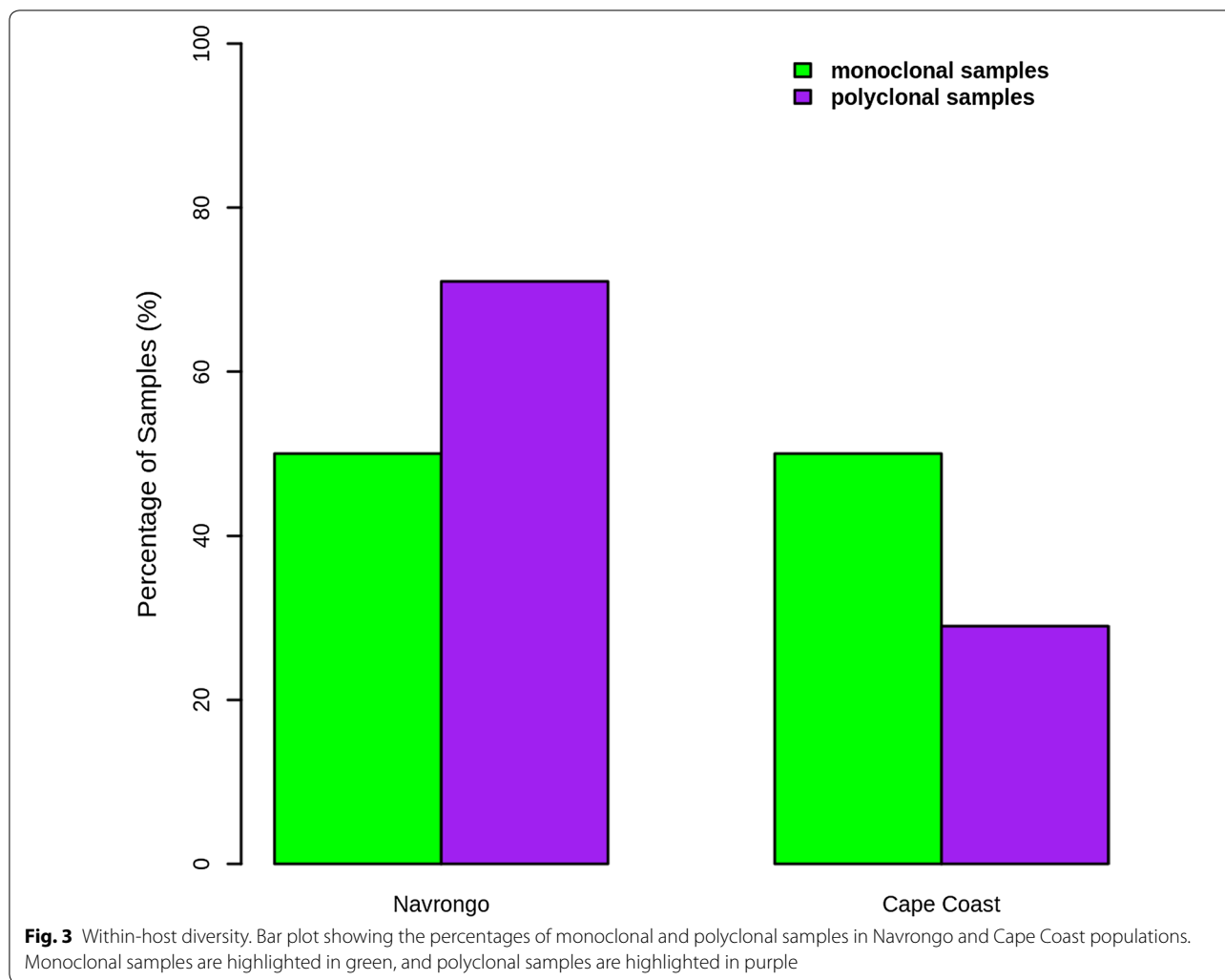
transmission settings), the allele frequency spectrum was dominated by very-low-frequency alleles ( $MAF \leq 0.05$ ) in both populations. Rare alleles ( $MAF \leq 0.01$ ) were observed at frequencies of 62.9% (22/35) and 61.8% (34/55) in Cape Coast and Navrongo, respectively. A total of 20% (7/35) and 10.9% (6/55) low-frequency variants [ $MAF$  range = (0.01–0.05)] were observed in Cape Coast and Navrongo, respectively. However, the remaining alleles showed a moderate to high MAF in both populations, implying some underlying evolutionary events.

**Within-host genetic diversity of *Pf**csp***

To assess the within-host diversity of *Pf**csp* in the population, the inbreeding coefficient (*F**WS*) was investigated. Isolates with *F**ws* values  $\geq 0.95$  were considered single strain (or monoclonal) infections, while *F**ws*  $< 0.95$  indicated diverse multigenic infections. In Cape Coast, 71.7% of *Pf**csp* isolates (66/92) came from single-strain infections with high inbreeding potential, while 28.3% (26/92) came from highly diverse multistrain infections with high potential for outcrossing (Fig. 3). For *P. falciparum*



**Fig. 2** A histogram showing the minor allele frequency distribution of a total of 90 SNPs located within *Pf**csp* loci in samples from both Cape Coast (n = 35 SNPs) and Navrongo (n = 55 SNPs). The vertical axis represents the number of SNPs in each category of allele frequency, and the horizontal axis shows the SNPs set in the respective MAFs range. There were no alleles found within the [0.25–0.30] and [0.30–0.35] MAF ranges in both populations



infections from Navrongo, 50.8% (65/128) were monoclonal *Pf*csp isolates, and 49.2% (63/128) harboured multiple *Pf*csp strains (Fig. 3). The Navrongo *Pf*csp isolates exhibited significantly higher within-host diversity than those from Cape Coast ( $\chi^2 = 15.382, p = 0.00009$ ).

**Genetic diversity of *Pf*csp C-terminal haplotypes**

To assess the extent of genetic diversity and similarity within and between the two populations, the diversity in

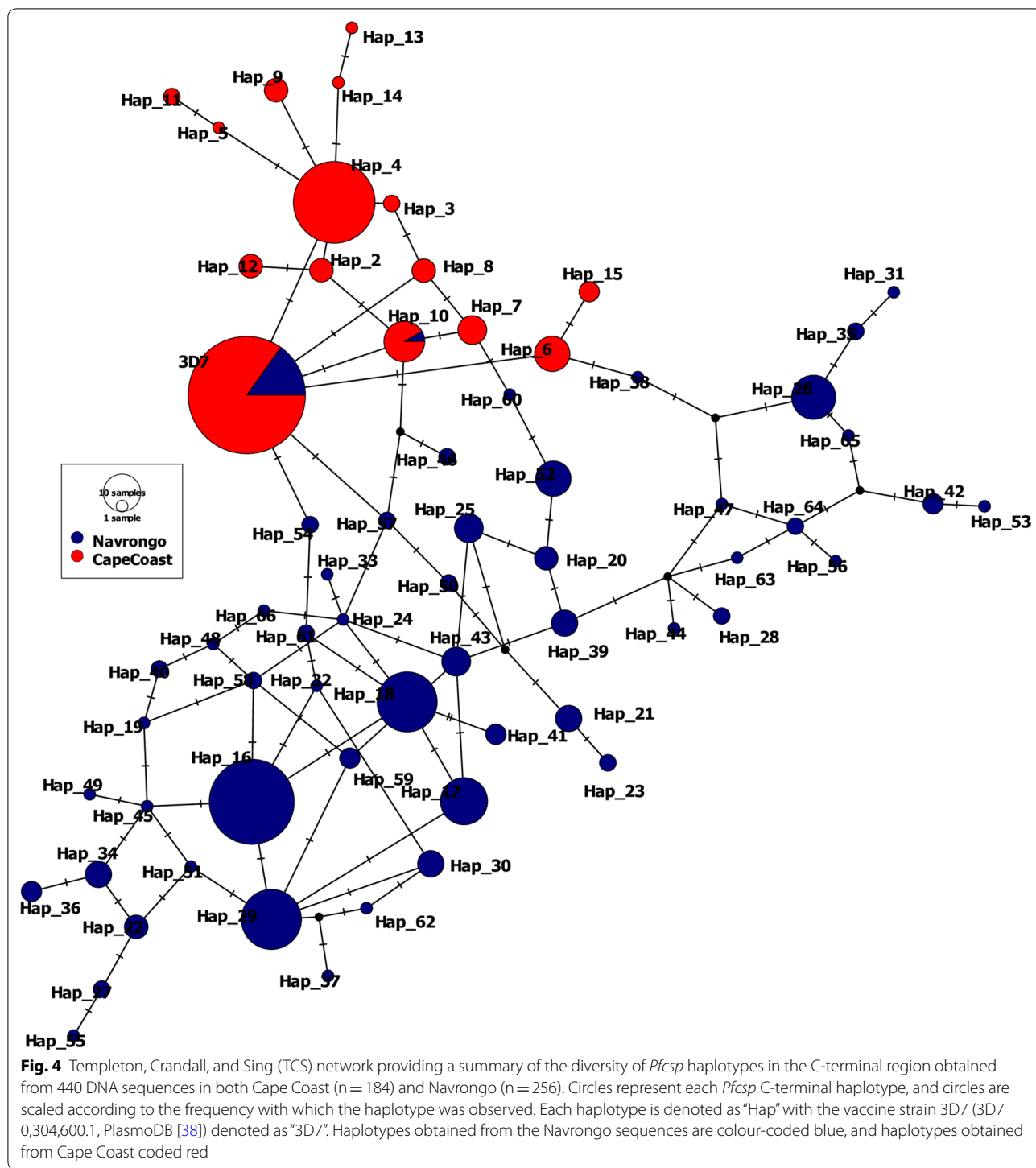
the C-terminal region of *Pf*csp (231 bp) was investigated from a total of 440 DNA sequences from Cape Coast (n = 184) and Navrongo (n = 256) (Table 1) and summarized in a Templeton, Crandall, and Sing (TCS) network (Fig. 4).

In total, 66 haplotypes were observed among the 440 *Pf*csp sequences obtained from both populations (Fig. 4). Among these haplotypes, 15 and 53 were found in the Cape Coast and Navrongo populations,

**Table 1** Diversity indices of the *Pf*csp C-terminal region of samples included in the network analysis

Population	n	Calculated indices				
		h	S	K	$\pi \pm S.D$	Hd $\pm S.D$
Cape Coast	184	15	8	1.15	0.005 $\pm$ 0.0004	0.718 $\pm$ 0.026
Navrongo	256	53	16	3.76	0.016 $\pm$ 0.0007	0.925 $\pm$ 0.009

n number of sequences, h number of unique haplotypes, S number of segregating sites, K average number of pairwise nucleotide differences,  $\pi$  nucleotide diversity, Hd haplotype diversity



respectively. The RTS,S vaccine haplotype (Pf3D7-type) and 1 nonvaccine haplotype (denoted as "Hap 10") were found in both populations (Fig. 4). The Pf3D7-type haplotype represented only 5.9% (n = 15/256) of haplotypes in Navrongo but 45.7% (n = 84/184) in Cape Coast (see Additional file 2). Only a single sample exhibited

Hap 10 from Navrongo (0.4%), but this haplotype represented 6.0% of the total haplotypes in Cape Coast (11/184 isolates) (Additional file 2). While the Pf3D7-type haplotype was the most prevalent *Pfmsp* C-terminal haplotype (45.7%) in isolates from Cape Coast, the most frequent haplotype in the Navrongo isolates was



“Hap 16”, representing 20.3% (52/256) of the haplotypes detected (Additional file 2).

According to the analysed genetic diversity indices, the *Pf*csp C-termini of the Navrongo isolates were generally more diverse than those from Cape Coast (Table 1). In summary, more nucleotide polymorphisms ( $K=3.761$ ) and segregating sites ( $S=16$ ) were observed in Navrongo than in Cape Coast ( $K=1.148$ ,  $S=8$ ). Consequently, *Pf*csp nucleotide diversity ( $\pi$ ) was higher in the Navrongo isolates ( $\pi = 0.016 \pm 0.0007$ ) than in the isolates from Cape Coast ( $\pi = 0.005 \pm 0.0004$ ). Haplotype diversity was also higher in Navrongo ( $Hd = 0.925 \pm 0.009$ ) in comparison with Cape Coast ( $0.718 \pm 0.026$ ) parasite isolates.

**TH2R and TH3R amino acid haplotype diversity**

The TH2R and TH3R sites were more polymorphic in both populations than the remaining amino acid sequence in the C-terminal region of PfCSP. In general, non-synonymous mutations predominated in all the isolates in both TH2R and TH3R epitope regions, with implications for cross-protection. Among the 92 (184 amino acid haplotypes) and 128 (256 amino acid haplotypes) isolates from Cape Coast and Navrongo, there were 8 and 27 nonvaccine TH2R haplotypes, respectively (see Additional file 3). There were also 2 and 10 nonvaccine TH3R haplotypes in Cape Coast and Navrongo, respectively, with 1 nonvaccine haplotype (NKPKDELNYAND) being shared between the two populations (Additional file 3). The frequencies of the Pf3D7-type TH2R vaccine haplotype (PSDKHIKEYLNKIQNLSL) were 56.5% and 7.4% in Cape Coast and Navrongo, respectively

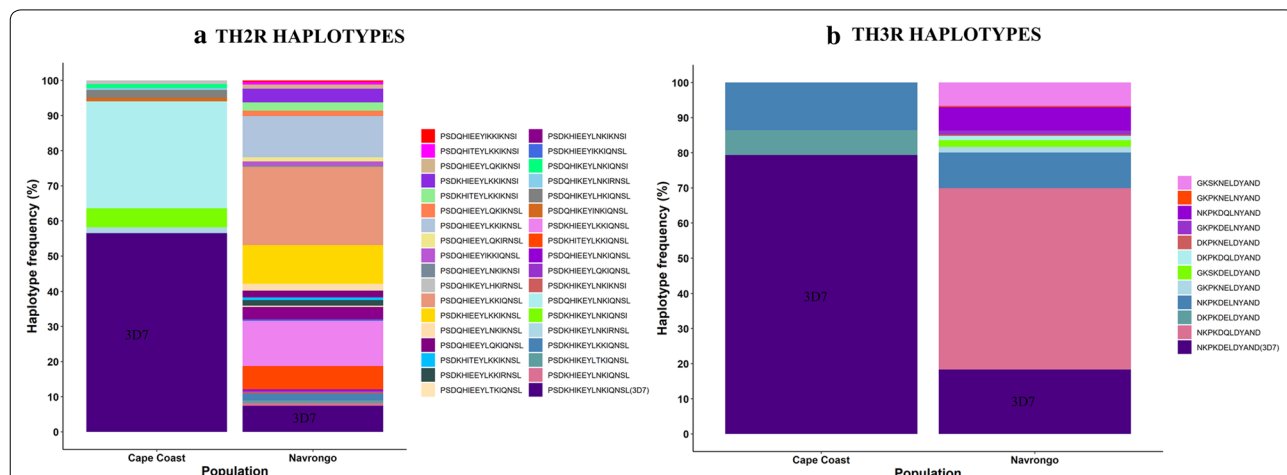
(Fig. 5a), while the frequencies were 79.3% and 18.4% for the Pf3D7-type TH3R vaccine haplotype (NKPKDELNYAND) (Fig. 5b) in the Cape Coast and Navrongo isolates, respectively. The amino acid differences observed between Pf3D7 reference (3D7 0304600.1, PlasmoDB) and the Ghanaian isolates ranged from 1 to 6 in both epitope regions (see Additional file 3).

**Population differentiation and structure of *Pf*csp**

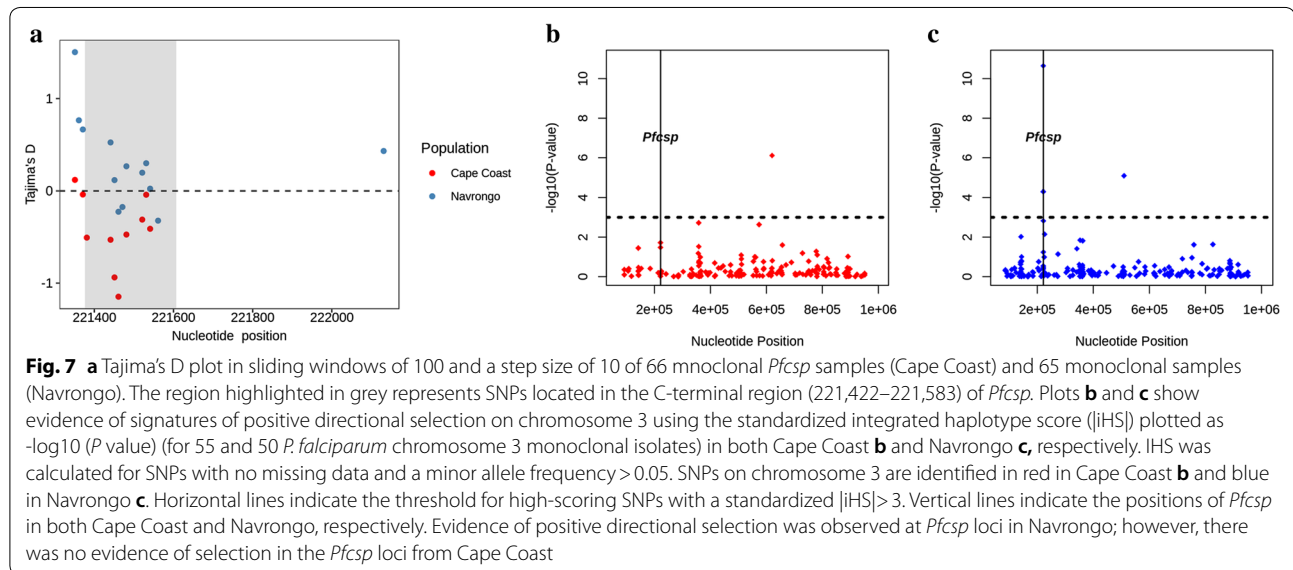
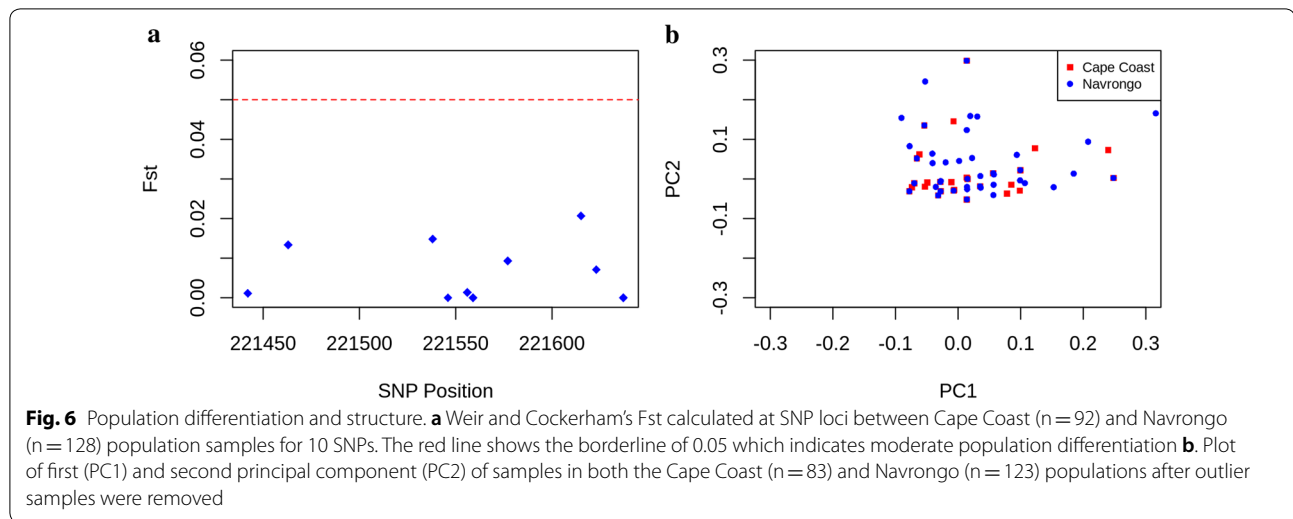
The overall Weir and Cockerham’s  $F_{st}$  between the Cape Coast and Navrongo *Pf*csp populations was  $< 0.05$  (Fig. 6a), which indicates minimal population differentiation due to genetic structure and suggests gene flow between the populations, despite the geographic distance between the sites. This also confirms the lack of genetic structure observed between Cape Coast and Navrongo parasite isolates through principal component analysis (Fig. 6b).

**Evidence of selection within populations**

Tajima’s D values were greater than zero in the TH2R and TH3R epitope regions of the C-terminal loci of *Pf*csp (221,422–221,583) for the population of monoclonal *Pf*csp isolates from Navrongo (Fig. 7a), suggesting balancing selection. However, a Tajima’s  $D < 0$  was seen in the Cape Coast population at these loci, suggesting likely directional selection or clonal expansion in the population. Alleles at SNP locus 221,554, which is within the segment encoding the TH2R epitope, had an  $|iHS| > 3$  in the Navrongo population, suggesting recent positive selection (Fig. 7c). The extended haplotype homozygosity revealed some extended haplotypes from the focal SNP



**Fig. 5** Plot a and b showing the percentage of isolates sharing specific amino acid haplotypes within the TH2R (311–327 aa) and TH3R (352–363 aa) epitope regions in both Cape Coast and Navrongo population, respectively. Coloured columns in the bar graph represent the haplotypes. The proportion of samples in each population having the 3D7 haplotype (vaccine haplotype) is represented in the first purple-coloured column from the bottom. The proportions of samples with non-vaccine haplotypes are shown in the rest of the coloured column bars



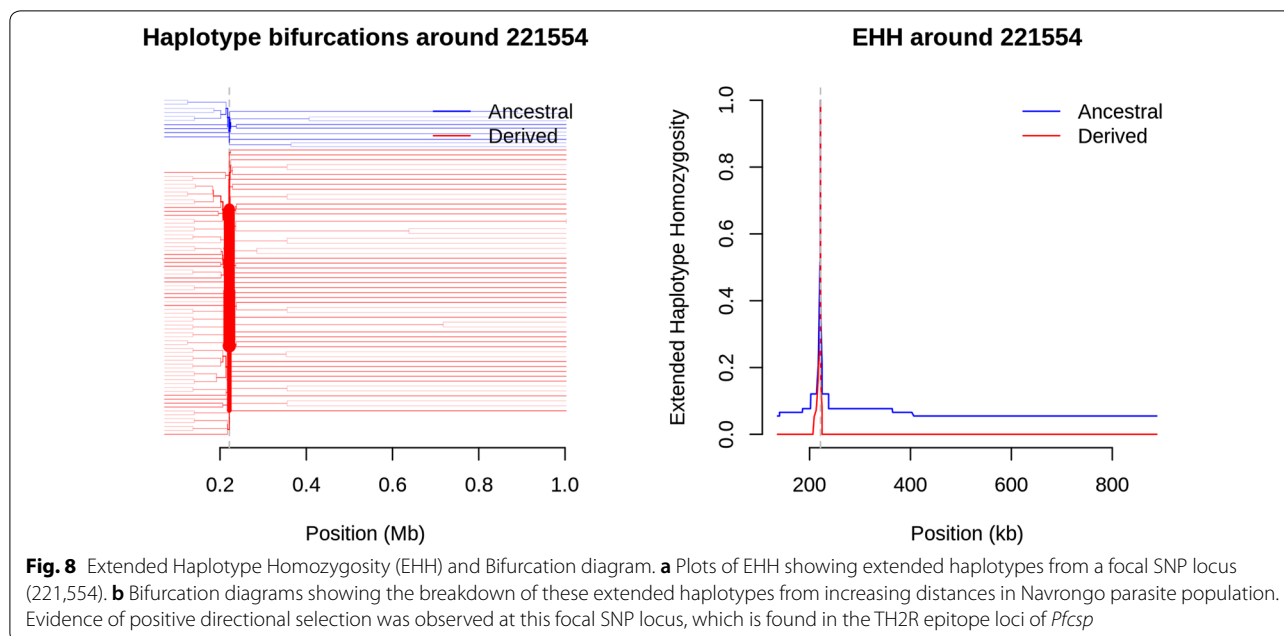
locus 221,554 in the Navrongo population, but no long-range haplotypes extended beyond 221,554 (Fig. 8a, b).

## Discussion

The RTS,S/AS01 malaria vaccine is based only on the *Pfcsp* sequence of the *P. falciparum* 3D7 clone [16], and strain-specific immunity has been confirmed for the licensed vaccine [15]. To provide new insights into how well RTS,S/AS01 may perform if administered on a large scale in different malaria-endemic regions, it is important to assess the intra-host diversity and extent of diversity in circulating parasites from different transmission settings.

Using *Pfcsp* sequence data generated from the whole-genome sequencing of 92 and 128 clinical parasite

isolates from Cape Coast in the coastal savanna region and the Kassena-Nankana districts (KNDS) in the Guinea savannah zone of the Upper East Region of Ghana, higher within-host malaria parasite diversity was observed in Navrongo, where 49.2% of infections exhibited an  $F_{ws} < 0.95$ , than in Cape Coast, where only 28.3% of infections exhibited an  $F_{ws} < 0.95$ . This high genetic diversity is known to occur in high-transmission areas, where infected individuals usually harbour more polyclonal infections compared to those living in low-transmission areas, where infections are often monoclonal [43]. Malaria transmission is much higher in Navrongo (EIR = 157) than in Cape Coast (EIR = 50) [20, 21]. These findings are consistent with high outcrossing potential



in the parasite population in the KNDs compared to the parasite population in the coastal town of Cape Coast. This marked difference in within-host diversity is noteworthy for future region-specific vaccine intervention strategies.

High genetic diversity in the C-terminal TH2R and TH3R amino acid epitopes was observed at the two sites. Notably, the vaccine-specific Pf3D7-type haplotype in the TH2R and TH3R epitopes represented approximately 56.5 and 79.3%, respectively, of the observed haplotypes in Cape Coast and only approximately 7.4 and 18.4% of those in the Navrongo isolates. The observed variance in location-specific diversity in these epitopes, which correlates with malaria transmission intensity, is consistent with findings from previous studies [8–10, 44–46]. Such polymorphisms in T-cell epitopes have been suggested to be due to an immune evasion mechanism in response to host T-cell immune responses [10] or selection in the mosquito host during the malaria transmission cycle [44]. From 1 to 6 amino acid differences were observed within the TH2R and TH3R epitope regions at each epitope in both parasite populations, with implications for the duration of vaccine efficacy [13]. This situation is similar to the amino acid haplotype differences observed in the C-terminal region in the Zambian and DRC populations, ranging from 2 to 10 [16]. In addition, there were more amino acid substitutions in the Navrongo parasite population than in the Cape Coast parasite population, which is consistent with the lower frequency of the vaccine haplotype observed in the network analysis for the Navrongo parasite population, and this will have implications for

vaccine efficacy in comparison with high-malaria-burden populations in Ghana. Another hypothesis drawn from a previous study suggests that polymorphism in the T cell epitopes could also be driven by an evolutionary response to intermolecular interactions at the surface of CSP [47].

The high degree of location-specific *Pfcs*p diversity observed in Ghana might result in differences in vaccine efficacy, potentially reducing RTS,S/AS01 vaccine effectiveness, particularly in Navrongo, where the vaccine haplotype was less prevalent. The monitoring of differential vaccine efficacy according to *Pfcs*p haplotypes during RTS,S/AS01 implementation programmes will be valuable for such high-transmission areas, where post-vaccination expansion of nonvaccine haplotypes in the population is likely to be observed, and this could lead to reduced vaccine efficacy and vaccine breakthrough infections.

The abundance of rare alleles shown in both Cape Coast and Navrongo contributes to the parasite population. Despite this high level of genetic diversity resulting from nonsynonymous nucleotide and amino acid substitutions observed, a shared gene pool remained between the two sites that resulted in a largely homogeneous parasite population. Over the sampled range of 784.4 km between the two sites, there was gene flow between the local populations of *P. falciparum* according to *Pfcs*p sequence analysis, with a pairwise index of differentiation ( $F_{st}$ ) below 0.05. The principal component analysis further confirmed the lack of population structure or genetic isolation. Previous studies have indicated that human population mixing is likely to cause gene flow

among *P. falciparum* parasites [48, 49]. Despite the ecological and epidemiological diversity between the 2 sites, human movement between the two sites is significant and could account for *Pfcsf* gene flow within the country, with implications for the spread of any emerging vaccine-resistant parasite. High levels of genetic recombination in the high-transmission area may explain the observed differences in haplotype diversity in Navrongo in comparison with Cape Coast [50] despite the observed gene flow between the two sites.

The negative Tajima's D observed in the Cape Coast isolates indicates a likely population expansion of the 3D7 major haplotype in an area with moderate malaria transmission after over 15 years of enhanced nationwide malaria control interventions (chemotherapy and vector control). This result corroborates findings from Thiès, Senegal, where increased deployment of malaria control interventions resulted in an increase in the frequency of clonal strains and a decrease in the probability of multiple infections [51]. Evidence of recent positive and balancing selection was observed in the Navrongo parasite isolates. The majority of the alleles present in the C-terminal region in the Navrongo parasite population exhibited a positive Tajima's D score and were highly polymorphic, likely due to balancing selection in response to host immune pressure on this immunogenic epitope [44, 52, 53]. Evidence of balancing selection on *Pfcsf* had been reported previously for a population from Malawi [44]. Balancing selection is common for immune targets and has been reported in other vaccine antigen candidates, such as in the domain I epitope of the Pf38 gene (found on the merozoite surface) in Papua, New Guinea, and The Gambia [54] and in the extracellular domains of AMA1, a target of allele-specific immune responses [55]. However, seasonal genetic drift among loci attributable to sampling across multiple transmission seasons in the Navrongo population may contribute to the observed balancing selection. The evidence of recent positive directional selection ( $iHS > 3$ ) observed at the T-cell epitope loci in the Navrongo parasite population could be due to the addition of new and useful alleles to the already existing repertoire of alleles being maintained by balancing selection in the population [18]. However, the signature of positive selection observed in Navrongo could likely be attributed to the dominance of one allele over the others at the T-cell epitope region in the Navrongo parasite population. Considering the differences in the eco-epidemiological background and the EIRs of these two populations, the intensity of transmission at these two ecologically distinct sites could account for differences in selection signals observed [45].

The samples analysed here were nonrandomly selected from the population, and this lack of randomness may

have some limitations and bias the inferences that can be drawn from *Pfcsf* and PfCSP diversity. Notably, the Navrongo and Cape Coast isolates were opportunistic samples whose sequence data were deposited into the Pf3K database at different times, leading to a geographically biased set of sequences, possibly over-representing limited genotypes from a small number of geographic foci and, in turn, under-representing large higher frequency SNPs. Furthermore, the conclusions drawn from sequences obtained from any given sequence repository are subject to change as sample sizes and geographic and temporal distributions are continually updated and expanded. Another limitation that may affect the interpretation of data is the small sample sizes analysed. Finally, the Navrongo sequences obtained from Pf3k come from different periods than the sequences from Cape Coast, which were sampled from the same periods. These limitations may prevent the samples from these two regions from being optimally comparable. There was, however, no population structure found within or between the two populations, indicating that the timespan did not affect the results obtained here. In addition, samples from Cape Coast were collected from a single district hospital, whereas the samples from Navrongo came from three health facilities. This circumstance may potentially have an impact on the results obtained, although the Cape Coast district hospital serves a wider catchment area, comparable to that of the three Navrongo sites. Despite these inherent limitations, the sequence analysis elaborated here is a powerful approach capable of elucidating local patterns in vaccine candidate genetic diversity and would be useful for monitoring the effect and efficacy of interventions. The examination of a larger sample size allowing a geographically and temporally broader analysis will further reveal the extent of the diversity of *Pfcsf* both locally and across Africa. This will help inform strategies for a wider implementation of the RTS,S vaccine.

## Conclusions

The extent of CSP polymorphism observed at the study sites likely indicates an allele-specific immune response during the pilot Phase IV implementation trials being conducted in Ghana. Similar to observations made in a study of an AMA-1 vaccine, vaccine efficacy during this trial in Ghana may be dependent on the degree of homology between the amino acid haplotypes circulating in the natural parasite populations and the 3D7 vaccine haplotype [56]. This situation might gradually result in a directional selective advantage of unmatched CSP haplotypes because the vaccine does not target them, emphasizing the need for a polyvalent malaria vaccine [57, 58].

With the ongoing Phase IV RTS,S vaccine implementation trials in Ghana, which include populations from Cape Coast and Navrongo, the findings from this study provide prior information on the extent of diversity in *Pfcs*p and the evolutionary forces driving these variations within Ghanaian natural parasite populations. These data will inform vaccine implementation outcomes and contribute to future vaccine design. These findings further emphasize the need for incorporating large-scale prevalence and population genetic analysis of vaccine candidate antigens into future malaria vaccine design to predict malaria vaccine outcomes.

## Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s12936-020-03510-3>.

**Additional file 1.** This file contains the primers used in nested PCR amplification of *Plasmodium falciparum* genomic DNA.

**Additional file 2.** This file is a Nexus file containing C-terminal *Plasmodium falciparum* circumsporozoite protein gene (*Pfcs*p) haplotype sequences that were included in the Templeton, Crandall, and Sing (TCS) network.

**Additional file 3.** TH2R and TH3R amino acid haplotype frequencies.

## Abbreviations

WHO: World Health Organization; RDT: Rapid Diagnostic Test; CSP: Circumsporozoite Protein; Pfcsp: Plasmodium falciparum Circumsporozoite Protein; NMCP: National Malaria Control Program; MVIP: Malaria Vaccine Implementation Program; AEIR: Annual Entomological Inoculation Rate; EIR: Entomological Inoculation Rate; AMA1: Apical Membrane Antigen 1; MSP: Merozoite Surface Protein; CRR: Central Repeat Region; PCR: Polymerase Chain Reaction; VCF: Variant Call Format; HBsAg: Hepatitis B Surface Antigen; SNP: Single Nucleotide Polymorphism; GATK: Genome Analysis Tool Kit; PCA: Principal Component Analysis; MAF: Minor Allele Frequency; IHS: Integrated Haplotype Score; EHH: Extended Haplotype Homozygosity.

## Acknowledgements

We are grateful to the Developing Excellence in Leadership and Genetic Training for Malaria Elimination (DELGEME) for the funding and support given to EAA as a Masters Fellow to enable her to complete this study. The authors, LA, AAN and AG, are currently supported through the DELTAS Africa Initiative an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [DELGEME Grant #107740/Z/15/Z] and the UK government. AG, AAN and LA worked on this project as part of their DELGEME aspiring leadership fellowship and as coapplicants on the grant. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust, H3Africa or the UK government.

## Authors' contributions

EAA analysed the data and prepared the first draft of the manuscript. LA, AAN and AG conceived the idea, the analysis plan, supervised the analysis and critically reviewed the manuscript. CA, wrote the in-house Python scripts and supported the data analysis and the writing of the manuscript. PO and KM supervised the data analysis and contributed to the drafting of the manuscript. All authors read and approved the final manuscript.

## Funding

Wellcome Trust [DELGEME Grant #107740/Z/15/Z].

## Availability of data and materials

The datasets generated during the current study are available in the MalariaGEN *Plasmodium falciparum* Community (Pf3k) Project release 5.1 (<http://www.malariagen.net/projects/parasite/pf>).

## Ethics approval and consent to participate

The study protocol was reviewed and approved by the Institutional Review Board of the Noguchi Memorial Institute for Medical Research, University of Ghana (056/12-13) and the Institutional Review Board of the Navrongo Health Research Centre (NHR CIRB203). Written informed consent was obtained from individuals, parents or guardians of all children before enrolment.

## Consent for publication

Not Applicable.

## Competing interests

Authors declare no conflict of interest.

## Author details

<sup>1</sup> Department of Biochemistry, Jomo Kenyatta University of Agriculture and Technology, Juja, Kenya. <sup>2</sup> Department of Biochemistry, Cell and Molecular Biology, West African Centre for Cell Biology of Infectious Pathogens, University of Ghana, Legon, Ghana. <sup>3</sup> Parasitology Department, Noguchi Memorial Institute of Medical Research, University of Ghana, Legon, Ghana. <sup>4</sup> Disease Control and Elimination, Medical Research Council Unit, The Gambia Unit, Bakau, The Gambia.

Received: 11 August 2020 Accepted: 19 November 2020

Published online: 27 November 2020

## References

- World Health Organization. World malaria report. Geneva: World Health Organization; 2019a.
- Clinical Trial Partnership. Efficacy and safety of RTS, S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet*. 2015;386:31–45.
- European Medicines Agency. First malaria vaccine receives positive scientific opinion from EMA. *Pharm J*. 2015;44:30–2.
- World Health Organization. Malaria Vaccine Implementation Programme (MVIP): proposed framework for policy decision on RTS, S/AS01 malaria vaccine. Geneva: World Health Organization; 2019b.
- Stoute JA, Slaoui M, Heppner DG, Momin P, Kester KE, Desmons P, et al. A preliminary evaluation of a recombinant circumsporozoite protein vaccine against *Plasmodium falciparum* malaria. *N Engl J Med*. 1997;336:86–91.
- Casares S, Brumeau TD, Richie TL. The RTS, S malaria vaccine. *Vaccine*. 2010;28:4880–94.
- Barry AE, Schultz L, Buckee CO, Reeder JC. Contrasting population structures of the genes encoding ten leading vaccine-candidate antigens of the human malaria parasite *Plasmodium falciparum*. *PLoS ONE*. 2009;4:e8497.
- Gandhi K, Thera MA, Coulibaly D, Traoré K, Guindo AB, Ouattara A, et al. Variation in the circumsporozoite protein of *Plasmodium falciparum*: vaccine development implications. *PLoS ONE*. 2014;9:e101783.
- Jalloh A, Jalloh M, Matsuoka H. T-cell epitope polymorphisms of the *Plasmodium falciparum* circumsporozoite protein among field isolates from Sierra Leone: age-dependent haplotype distribution? *Malar J*. 2009;8:120.
- Zeeshan M, Alam MT, Vinayak S, Bora H, Tyagi RK, Alam MS, et al. Genetic variation in the *Plasmodium falciparum* circumsporozoite protein in India and its relevance to RTS,S malaria vaccine. *PLoS ONE*. 2012;7:e43430.
- Allouche A, Milligan P, Conway DJ, Pinder M, Bojang K, Doherty T, et al. Protective efficacy of the RTS, S/AS02 *Plasmodium falciparum* malaria vaccine is not strain specific. *Am J Trop Med Hyg*. 2003;68:97–101.
- Bojang KA, Milligan PJM, Pinder M, Vigneron L, Allouche A, Kester KE, et al. Efficacy of RTS, S/AS02 malaria vaccine against *Plasmodium falciparum* infection in semi-immune adult men in The Gambia: a randomised trial. *Lancet*. 2001;358:1927–34.

13. Enosse S, Dobaño C, Quelhas D, Aponte JJ, Lievens M, Leach A, et al. RTS, S/AS02A malaria vaccine does not induce parasite CSP T cell epitope selection and reduces multiplicity of infection. *PLoS Clin Trials*. 2006;1:e5.
14. Waitumbi JN, Anyona SB, Hunja CW, Kifude CM, Polhemus ME, Walsh DS, et al. Impact of RTS, S/AS02A and RTS, S/AS01B on genotypes of *P. falciparum* in adults participating in a malaria vaccine clinical trial. *PLoS ONE*. 2009;4:e7849.
15. Neafsey DE, Juraska M, Bedford T, Benkeser D, Valim C, Griggs A, et al. Genetic diversity and protective efficacy of the RTS, S/AS01 malaria vaccine. *N Engl J Med*. 2015;373:2025–37.
16. Pringle JC, Carpi G, Almagro-García J, Zhu SJ, Kobayashi T, Mulenga M, et al. RTS, S/AS01 malaria vaccine mismatch observed among *Plasmodium falciparum* isolates from southern and central Africa and globally. *Sci Rep*. 2018;8:6622.
17. Takala SL, Plowe CV. Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming “vaccine resistant malaria.” *Parasit Immunol*. 2010;31:560–73.
18. Duffy CW, Assefa SA, Abugri J, Amoako N, Owusu-Agyei S, Anyorigiya T, et al. Comparison of genomic signatures of selection on *Plasmodium falciparum* between different regions of a country with high malaria endemicity. *BMC Genomics*. 2015;16:527.
19. Koram KA, Owusu-Agyei S, Fryauff DJ, Anto F, Atuguba F, Hodgson A, et al. Seasonal profiles of malaria infection, anaemia, and bednet use among age groups and communities in northern Ghana. *Trop Med Int Health*. 2003;8:793–802.
20. Abuaku B, Duah-Quashie NO, Quaye L, Mtrevi SA, Quashie N, Gyasi A, et al. Therapeutic efficacy of artesunate-amodiaquine and artemether-lumefantrine combinations for uncomplicated malaria in 10 sentinel sites across Ghana: 2015–2017. *Malar J*. 2019;18:206.
21. Oduro AR, Wak G, Azongo D, Debpur C, Wontuo P, Kondayire F, et al. Profile of the Navrogon health and demographic surveillance system. *Int J Epidemiol*. 2012;41:968–76.
22. Snounou G, Viriyakosol S, Zhu XP, Jarra W, Pinheiro L, do Rosario VE, et al. High sensitivity of detection of human malaria parasites by the use of nested polymerase chain reaction. *Mol Biochem Parasitol*. 1993;61:315–20.
23. The Pf3K Project (2016): pilot data release 5. <http://www.malariagenet.net/data/pf3k-5>.
24. Amato R, Miotto O, Woodrow CJ, Almagro-García J, Sinha I, Campino S, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:207.
25. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–75.
26. Browning BL, Zhou Y, Browning SR. A One-penny imputed genome from next-generation reference panels. *Am J Hum Genet*. 2018;103:338–48.
27. Auburn S, Campino S, Miotto O, Djimde AA, Zongo I, Manske M, et al. Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. *PLoS ONE*. 2012;7:e32891.
28. Manske M, Miotto O, Campino S, Auburn S, Zongo I, Ouedraogo J, et al. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature*. 2013;487:375–9.
29. Lee S, Harrison A, Tessier N, Tavul L, Miotto O, Siba P, et al. Assessing clonality in malaria parasites from massively parallel sequencing data. *F1000Research*. 2015;4:1043.
30. Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol Biol Evol*. 2017;34:3299–302.
31. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci USA*. 1979;76:5269–73.
32. Nei M. *Molecular evolutionary genetics*. New York: Columbia University Press; 1987.
33. Clement M, Posada D, Crandall KA. TCS: a computer program to estimate gene genealogies. *Mol Ecol*. 2000;9:1657–9.
34. Templeton AR, Crandall KA, Sing CF. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data III Cladogram estimation. *Genetics*. 1992;132:619–33.
35. Leigh JW, Bryant D. POPART: full-feature software for haplotype network construction. *Methods Ecol Evol*. 2015;6:1110–6.
36. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res*. 2009;37:539–43.
37. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156–8.
38. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006;2:2074–93.
39. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989;123:585.
40. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS Biol*. 2006;4:e72.
41. Gautier M, Rehh RV. An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*. 2012;28:1176–7.
42. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. 2002;419:832–7.
43. Abukari Z, Okonu R, Nyarko SB, Lo AC, Dieng CC, Salifu SP, et al. The diversity, multiplicity of infection and population structure of *P. falciparum* parasites circulating in asymptomatic carriers living in high and low malaria transmission settings of Ghana. *Genes*. 2019;10:434.
44. Bailey JA, Mvalo T, Aragam N, Weiser M, Congdon S, Kamwendo D, et al. Use of massively parallel pyrosequencing to evaluate the diversity of and selection on *Plasmodium falciparum* csp T-cell epitopes in Lilongwe. *Malawi J Infect Dis*. 2012;206:580–7.
45. Escalante AA, Grebert HM, Isea R, Goldman IF, Basco L, Magris M, et al. A study of genetic diversity in the gene encoding the circumsporozoite protein (CSP) of *Plasmodium falciparum* from different transmission areas - XVI Asembo Bay Cohort Project. *Mol Biochem Parasitol*. 2002;125:83–90.
46. Kumkhaek C, Phra-ek K, Singhasivanon P, Looareesuwan S, Hirunpetcharat C, Brockman A, et al. A survey of the Th2R and Th3R allelic variants in the circumsporozoite protein gene of *P. falciparum* parasites from Western Thailand. *Southeast Asian J Trop Med Public Health*. 2004;35:281–7.
47. Aragam NR, Thayer KM, Nge N, Hoffman I, Martinson F, Kamwendo D, et al. Diversity of T Cell Epitopes in *Plasmodium falciparum* circumsporozoite protein likely due to protein-protein interactions. *PLoS ONE*. 2013;8:e62427.
48. Amambua-Ngwa A, Tetteh KKA, Manske M, Gomez-Escobar N, Stewart LB, Deerhake ME, et al. Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genet*. 2012;8:e1002992.
49. Duffy CW, Ba H, Assefa S, Ahouidi AD, Deh YB, Tandia A, et al. Population genetic structure and adaptation of malaria parasites on the edge of endemic distribution. *Mol Ecol*. 2017;26:2880–94.
50. Yuan L, Zhao H, Wu L, Li X, Parker D, Xu S, et al. *Plasmodium falciparum* populations from northeastern Myanmar display high levels of genetic diversity at multiple antigenic loci. *Acta Trop*. 2013;125:53–9.
51. Daniels R, Chang HH, Séné PD, Park DC, Neafsey DE, Schaffner SF, et al. Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. *PLoS ONE*. 2013;8:e60780.
52. Tetteh KKA, Stewart LB, Ochola LI, Amambua-Ngwa A, Thomas AW, Marsh K, et al. Prospective identification of malaria parasite genes under balancing selection. *PLoS ONE*. 2009;4:e5568.
53. Weedall GD, Conway DJ. Detecting signatures of balancing selection to identify targets of anti-parasite immunity. *Trends Parasitol*. 2010;26:363–9.
54. Reeder JC, Wapling J, Mueller I, Siba PM, Barry AE. Population genetic analysis of the *Plasmodium falciparum* 6-cys protein Pf38 in Papua New Guinea reveals domain-specific balancing selection. *Malar J*. 2011;10:126.
55. Polley SD, Conway DJ. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. *Genetics*. 2001;158:1505–12.
56. Thera MA, Doumbo OK, Coulibaly D, Diallo DA, Kone AK, Guindo AB, et al. Safety and immunogenicity of an AMA-1 malaria vaccine in Malian adults: results of a phase 1 randomized controlled trial. *PLoS ONE*. 2008;3:e1465.
57. Barry AE, Arnott A. Strategies for designing and monitoring malaria vaccines targeting diverse antigens. *Front Immunol*. 2014;5:359.
58. Dutta S, Seung YL, Batchelor AH, Lanar DE. Structural basis of antigenic escape of a malaria vaccine candidate. *Proc Natl Acad Sci USA*. 2007;104:12488–93.

## Publisher's Note

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