Population Genetic Analysis of Plasmodium Falciparum Circumsporozoite Protein in Two Distinct Ecological Regions in Ghana

Elikplim A Amegashie
Jomo Kenyatta University of Agriculture and Technology
Lucas Amenga-Etego
University of Ghana, Department of Biochemistry, Cell and Molecular Biology, West African Centre for Cell Biology of Infectious Pathogens
Courage Adobor
University of Ghana Noguchi Memorial Institute for Medical Research
Peter Ogoti
Jomo Kenyatta University of Agriculture and Technology
Kevin Mbogo
Jomo Kenyatta University of Agriculture and Technology
Alfred Amambua-Ngwa
Parasite Molecular Biology, Disease Control and Elimination, Medical Research Council Unit, The Gambia at LSHTM, Atlantic Road, Fajara, Banjul, The Gambia
Anita Ghansah (aghansah@noguchi.ug.edu.gh)
University of Ghana Noguchi Memorial Institute for Medical Research https://orcid.org/0000-0003-4639-1249

Research

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

Posted Date: August 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-52164/v1

Version of Record: A version of this preprint was published on November 27th, 2020. See the published version at https://doi.org/10.1186/s12936-020-03510-3.

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/ASO1 vaccine. 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

Pfcsp sequences were retrieved from Illumina generated paired-end short-read sequences of 101 and 131 malaria samples from children aged 6-59 months with clinical malaria presenting at health facilities in Cape Coast (on the coastal belt) and Navrongo (Guinea savannah region) respectively in Ghana. The sequences were mapped to 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 analyzed for allele frequency spectrum, genetic diversity both within host and between the populations and signatures of selection. Population genetics tools were used to determine the extent and dynamics of *Pfcsp* diversity in *P. falciparum* from the two geographically distinct locations in Ghana.

Results

Pfcsp was extensively diverse at the two sites with higher transmission site, Navrongo, recording both higher within host and population level diversity. The vaccine strain C-terminal epitope of *Pfcsp* was found in only 5.9% and 45.7% of the Navrongo and Cape Coast sequences respectively. Amino acid variations ranging between 1 and 6 were observed in the TH2R and TH3R epitope regions of the PfCSP. Tajima's D was negatively skewed especially for the population from Cape Coast given expected historical population expansion. On the contrary, positive Tajima's D was observed for the Navrongo *P. falciparum* population, consistent with balancing selection acting on the immuno-dominant TH2R and TH3R vaccine epitopes.

Conclusion

The low frequencies of *Pfcsp* vaccine haplotype in the populations analyzed calls for additional molecular and immuno-epidemiological studies with temporal and wider geographic sampling in endemic populations targeted for RTS,S application. These results have implications on the efficacy of the vaccine in Ghana and will inform the choice of alleles to include in future multivalent or chimeric vaccines.

Background

Stagnation in the decline of malaria over the last 5 years is indicative that the set global malaria elimination targets may not be achieved without the addition of a broadly effective vaccine to compliment 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 (Burkina Faso, Gabon, Ghana, Kenya, Malawi, Mozambique, and Tanzania) in children (aged 5 to 17 months) and infants (aged 6-12 weeks), who were followed up for a median of 48 and 38 months respectively, revealed that the vaccine confers moderate protective efficacy against clinical disease and severe malaria which waned over time [2]. The efficacy of the vaccine was observed to be lower in the infants than in the children (at first vaccination) [2]. The vaccine conferred only 36.3% protection against clinical malaria and approximately 29% against severe malaria in children aged 5-17 months [2]. Despite this modest efficacy, the vaccine was approved by the European Medicines Agency in 2015 stating that the benefits of protective immunity outweigh the risk and that the impact of this moderate efficacy would potentially be high, given the huge disease burden. Subsequently, Ghana, Kenya and Malawi were selected and currently conducting the pilot Phase IV implementation trials via The Malaria Vaccine Implementation Programme (MVIP) led by WHO.

The RTS,S vaccine is a malaria subunit vaccine that is formulated from a fragment of the circumsporozoite protein (CSP) of *P. falciparum* 3D7 laboratory strain and fused with the Hepatitis B surface antigen and the ASO1 adjuvant [3]. 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 $CD4^+$ and $CD8^+$ T-cell epitopes denoted as TH2R and TH3R respectively [4]. Several studies have reported high levels of polymorphisms at the T-cell epitopes within the C-terminal region of PfCSP in natural parasite populations [5–8]. Although there are variations in the immuno-dominant 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 in the children [6]. Initial ancillary studies on Phase II clinical trials conducted in three sites including The Gambia, Kenya and Mozambique revealed that immune protection of the RTS,S vaccine was not strain specific [9–12]. However, these studies were based on less sensitive and older sequencing technologies such as Sanger sequencing and methodologies such as oligonucleotide probing for genotyping of *Pfcsp* (the gene that encodes *P. falciparum* CSP) and only in a few hundred isolates.

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 mismatch strains (33.3%) [13]**.** Also, recent studies of the population structure of *Pfcsp* suggest that geographically variable levels of diversity and geographic

restriction of specific subgroups may have an impact on the efficacy of *Pfcsp*-based malaria vaccines in specific geographic regions [5].

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[13–15]. Furthermore, evolutionary factors such as selection operating on parasites differ locally owing to varying transmission patterns, ecology and degrees of acquired immunity in humans [16]. Therefore, further characterization of the genetic diversity of immune epitopes of vaccine antigens is important, especially in regions like Ghana where the vaccine is undergoing the Phase IV implementation trial. This should provide broader assessment of the extent to which the local natural diversity could impact its efficacy and wider implementation.

Malaria transmission in Ghana is generally perennial but with marked seasonal effects, that varies with local ecology and overall transmission intensity [17]. For control purposes, malaria transmission across Ghana has been classified eco-epidemiologically into three main zones, namely; forest ecology with perennial but high transmission during the rainy season (May-August and October-November), northern/Guinea savannah with seasonal and intense malaria transmission during the rainy season (June-October), but with periods of very low transmission during dry season, and coastal savannah with low to moderate perennial transmission and a marked seasonal effect during the rainy season [18]. The implementation trial of RTS,S vaccine is being conducted in three regions namely, Brong-Ahafo and Volta regions in the forest ecology and Central region on 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 *Pfcsp* in parasite populations from two geographically distinct sites in Ghana, we evaluated within-host diversity (complexity of infection) and the extent of population specific haplotype diversity of the C-terminal region of *Pfcsp* encompassing the TH2R and TH3R epitopes. Our results provide information on diversity most relevant to vaccine escape and cross-protection. We further explored PfCSP amino acid diversity and conservation. In addition, we assessed evidence of selection in the *Pfcsp* that could be driving and sustaining its diversity observed.

Methods

Study area

The study was conducted in two sites, the Cape Coast Metropolitan area with Cape Coast as the main township and Kassena-Nankana districts (KNDs) with Navrongo as the main township (Fig 9). Cape Coast is located in the Southern coastal savannah region with low to moderate perennial malaria transmission but with marked seasonal effect during the rainy season (May-August and October-November). Estimated annual entomological inoculation rate (EIR) is fewer than 50 infective bites per person per year [18,44]. 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. Estimated annual EIR for the KNDs is up to 157 infective bites per person per year [23,45].

In Cape Coast, *P. falciparum* parasites were isolated from 101 children (aged 6-59 months) with malaria who live within the municipality and attended 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 also with uncomplicated *P. falciparum* malaria living in the KNDs and presenting with fever (i.e. axillary temperature ≥ 37.5 or history of fever during the previous 24 hours) at health facilities in the KNDs in the years 2010 (January to October), 2011 (January to February) and 2013 (August to October) during both dry and wet seasons. Participants with fever were screened with malaria Rapid Diagnostic Test (RDT) and blood smears taken for RDT positive individuals and *P. falciparum* asexual parasites was determined by microscopy.

Fig9. Location of the Navrongo within the Kassena-Nankana Districts and the Cape Coast within the Cape Coast Metropolis. The distance between Cape Coast and Navrongo is approximately 784.8 km (generated by QGIS software).

Genomic DNA extraction and sequencing

Genomic DNA was extracted using the QiaAmp DNA prep kit (Qiagen, Valencia, CA) following manufacturer's protocol and confirmation of *P. falciparum* positive samples was done by amplification using nested PCR with specific primers [46]. 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 (200bp insert) were aligned to the reference *P. falciparum* 3D7 reference genome after which variant calling was done following the customized GATK pipeline. Each sample was genotyped for 797,000 polymorphic bi-allelic coding SNPs across the genome ensuring a minimum of 5x 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 were mined from the MalariaGEN *Plasmodium falciparum* Community (Pf3k) Project release 5.1 Database [47] in variant call format (VCF) and Cape Coast data obtained as VCF files. Genetic variants on chromosome 3 were retrieved for both Navrongo and Cape Coast. For the VCF file of Cape Coast, all genotypes at each SNP position were mono-allelic (monoclonal); we modeled biallelic genotypes using a custom-made python script. This was based on the approach by the MalariaGEN Pf3k Project, where loci with mixed allele calls were modeled using the read and allelic

depth [48]. Briefly, for read depth of two alleles at a SNP position in a sample, at positions with read depth <5, the genotype was undetermined. At all other positions, with read depth ≥ 5, the sample was determined to be heterozygous if the allelic depth of both alleles were greater or equal to 2. All other samples were homozygous for the allele with majority read count. These were either genotyped as the homozygote reference allele or homozygote alternative allele.

Data for both populations were filtered to obtain only biallelic SNPs using Bcftools v1.9 [49]. Data was quality checked as follows : Only SNPs that passed all VCF filters were retained. Isolates with >10% missing SNPs were excluded followed by removal of SNPs with >5% of missing data using PLINK v1.9 [50]. Further, 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) less than 1% were removed. Missing SNPs were imputed and phased using Beagle v5.1 [51]. After quality control, the Cape Coast dataset remained with 2504 SNPs out of 26156 within Chromosome 3, 35 within *Pfcsp* and 92 out of 101 samples. On the other hand, the Navrongo dataset retained 1954 out of 43199 SNPs within Chromosome 3, 55 within *Pfcsp* and 128 out of 131 samples. The *Pfcsp* was then extracted from chromosome 3 (position: 221323 – 222516) and retained SNPs at the CSP loci used were 13 and 22 for Cape Coast and Navrongo respectively.

Population genetics analysis

Minor allele frequency distribution

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

Within host parasite diversity estimation

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

Fws metric estimates the heterozygosity of parasites (H_W) within the individual relative to the heterozygosity within the parasite population (H_S) using the read count of alleles. The Fws metric calculation for each sample was done using the equation:

$Fws = 1 - H_W/H_S$

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 [52,53]. Fws ranges from 0 to 1; a low Fws value indicates low inbreeding rates within the parasite population thus high within host diversity relative to the population. An Fws threshold ≥0.95 implies samples with clonal (single strain) infections while samples with Fws <0.95 are considered highly to have mixed strain infections signifying within host diversity. Fws was calculated using an R package, moimix [54]. Samples with clonal infections were used for selection analysis. The Pearson Chi square 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 done using R software with P values of <0.05 considered statistically significance.

Genetic diversity within parasite populations

We examined the haplotype diversity (the number of two random strains within the population having different haplotypes) of the *Pfcsp* in each population by exploring the variants in the C-terminal region of the gene (909 – 1140bp). Firstly, we re-constructed 440 *Pfcsp* fasta DNA sequences with the retained variants (13 and 22 SNPs) from the 92 Cape Coast isolates and 128 Navrongo isolates using an in-house python script. We then determined the *Pfcsp* C-terminal haplotypes using DnaSP software (version 6.10.01) [55].

The following metrics were then used to assess diversity of *Pfcsp* C-terminal within each parasite population using the DnaSP software (version 6.10.01): number of sequences (n), number of haplotypes (h), segregating sites (S), average number of pairwise nucleotide differences (K), nucleotide diversity (π) and haplotype diversity (Hd) [56,57].

To assess the genealogical relationships between the *Pfcsp* C-terminal haplotypes found in Navrongo and Cape Coast, we constructed a network based on the method described by Templeton, Crandall, and Sing (TCS) [58,59] using PopArt [60]. The haplotypes were denoted as 3D7, Hap 2 up to Hap 66 in the network.

We further explored the amino acid haplotypes within each population by translating all the 440 *Pfcsp* DNA sequences into amino acid sequences and comparing them to the 3D7 reference strain (0304600.1, PlasmoDB [19]) using an in-house python script

Consistent with literature, the frequency of TH2R 311-327 amino acid (PSDKHIKEYLNKIQNSL) and TH3R 352-363 amino acid (NKPKDELDYAND) haplotypes in each parasite population were determined also using a customized python script and plotted. We also explored the patterns in amino acid sequence at the C-terminal region between the two sites by generating sequence logos online with Weblogo [61]. The weblogos generated shows which amino acid positions are conserved or polymorphic within the C-terminal regions of both parasite populations. The amino acids are denoted in the weblogo as a stack of letters and measured in bits with the height of each letter proportional to the observed frequency of the

corresponding amino acid at each position and the overall height of each stack is proportional to the sequence conservation [61].

Population differentiation and structure analysis

Wright Fixation index (Fst) and principal component analysis (PCA) were used for population differentiation and structure analyses. To reduce bias in Fst and PCA analysis, we pruned out SNPs (from the 2504 Cape Coast chromosome 3 retained SNPs and the 1954 Navrongo retained SNPs) with pairwise linkage disequilibrium (LD) value, r^2 >0.5 within a window of 100bp in the entire chromosome 3 dataset 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 *Pfcsp* SNPs.

The *Pfcsp* SNPs were then used to estimate Fst and population structure. Weir and Cockerham's Fst per SNP between Cape Coast and Navrongo parasite isolates, was calculated using Vcftools v0.1.5 [62] and population structure by PCA was done using smartpca (Cambrigde, MA, USA) in EIGENSOFT package v6.1.3 [63]. 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 Cape Coast and 123 samples in Navrongo population after outlier samples were removed.

Signatures of selection

To test for SNP neutrality, the Tajima's D statistical test [64], was done in sliding windows of size 100bp and step size of 10 with *Pfcsp* 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 selection.

To detect loci likely to be under recent positive selection in the Cape Coast and Navrongo monoclonal isolates, we calculated the standardized integrated haplotype score (|iHS|) for each SNP with a MAF >0.05 in chromosome 3 (358 out of the 2504 and 608 out of the 1954 remaining SNPs from Cape Coast and Navrongo respectively) [65]. For this analysis, the Fws metric was used to estimate these monoclonal parasites in chromosome 3 using the R package, moimix [54]. |iHS| measures the amount of extended haplotype homozygosity (EHH) at a given SNP along the ancestral allele relative to the derived allele [65]. The reference and alternate alleles were characterized as ancestral and derived alleles respectively. This was done in R using the rehh package v2.0.4 [66]. Genomic regions under positive selection were identified as those with multiple SNPs having |iHS| values >3 and formed the focal SNPs for extended haplotype homozygosity (EHH) analysis. EHH for both the reference and alternate alleles were calculated and bifurcation plots generated to visualize the decay of EHH at increasing distances from the focal SNP loci [67] using rehh package v2.0.4 in R.

Results

Minor allele frequency distribution of *Pfcsp*

A total of 90 SNPs within the *Pfcsp* were analyzed for minor allele frequency (MAF). The *P. falciparum* population from Navrongo was more variable at *Pfcsp* (55 SNPs) compared to Cape Coast (35 SNPs), (**Fig 1).** The allele frequency distribution of all putative SNPs within the *Pfcsp* loci ranged between 0.001-0.45 in Navrongo and 0.001-0.40 in Cape Coast **(Fig 1).** As expected for natural *P. falciparum* populations in Africa (high transmission settings), the allele frequency spectrum was dominated by very low frequency alleles (MAF ≤ 0.05) in both populations. Rare alleles (MAF ≤ 0.01) were observed at 62.9% (22/35) and 61.8% (34/55) in Cape Coast and Navrongo respectively. Low-frequency variants 20% (7/35) and 10.9% (6/55) (MAF range = [0.01-0.05]) were observed in Cape Coast and Navrongo respectively. However, the remaining alleles were observed in moderate to high MAF in both populations implying some underlining evolutionary events.

Fig 1. A histogram showing the minor allele frequency distribution of a total of 90 SNPs set within the *Pfcsp* loci in samples from both Cape Coast (n= 35 SNPs) and Navrongo (n =55 SNPs). Vertical axis represents the number of SNPs in each category of allele frequency and the horizontal axis shows the binned SNPs set with bin1 to bin8 representing the following MAFs ranges; **[0.0-0.01], [0.01-0.05], [0.05-0.10], [0.10-0.15], [0.15-0.20], [0.20-0.25], [0.35-0.40], [0.40-0.45])** respectively. There were no alleles found within the [0.25-0.30] and [0.30-0.35] bins in both populations.

Within host genetic diversity of *Pfcsp*

In order to assess the within host diversity of *Pfcsp* in the population, the inbreeding co-efficient (Fws) was investigated. Isolates with Fws values ≥0.95 were considered single strain (or monoclonal) infections whilst Fws <0.95 was denoted as diverse multi-gene infections. In Cape Coast, 71.7% of *Pfcsp* isolates (66/92) were single strain infections with high inbreeding potential while 28.3% (26/92) were highly diverse multi-strain infections with high potential for outcrossing **(Fig 2).** For *P. falciparum* infections from Navrongo, 50.8% (65/128) were monoclonal *Pfcsp* isolates, and 49.2% (63/128) harbored multiple *Pfcsp* strains (Fig 2). The Navrongo *Pfcsp* isolates had significantly higher within host diversity compared to Cape Coast (χ^2 = 15.382, p = 0.00009).

Fig 2. **Within host diversity.** Boxplot showing the distribution of Fws estimates in samples from Cape Coast and Navrongo populations. Cape Coast highlighted in red and Navrongo highlighted in blue.

Genetic diversity of *Pfcsp* C-terminal haplotypes

To assess the extent of genetic diversity and similarity within and between the two populations, we investigated the diversity in the C-terminal region of *Pfcsp* (231bp) from 440 DNA sequences from Cape Coast (n = 184) and Navrongo (n = 256) (**Table 1**) and summarized this in a Templeton, Crandall, and Sing (TCS) network **(Fig 3).**

In total, we observed 66 haplotypes from the 440 *Pfcsp* sequences obtained from both populations **(Fig 3).** Of these, 15 and 53 haplotypes were found in the Cape Coast and Navrongo populations respectively. The RTS,S vaccine haplotype (Pf3D7-type) and 1 non-vaccine haplotype (denoted as "Hap 10") were found in both populations (Fig 3). 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 1**). Only a single sample had the Hap 10 (0.4%) but this represented 6.0% of haplotypes in Cape Coast (11/184 isolates), (**Additional file 1**). While the Pf3D7-type haplotype was the most prevalent *Pfcsp* 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 1**).

From genetic diversity indices analyzed, *Pfcsp* C-terminal from Navrongo isolates were generally more diverse than in Cape Coast **(Table 1).** In summary, we observed more nucleotide polymorphisms (K= 3.761) and segregating sites (S =16) in the Navrongo than Cape Coast (K = 1.148, S = 8). Consequently, *Pfcsp* nucleotide diversity (was also higher in the Navrongo isolates (0.0007) than in isolates from Cape Coast (= 0.005 0.0004). Haplotype diversity was also higher in Navrongo (Hd = 0.925 0.009) in comparison with Cape Coast (0.718 0.026) parasite isolates.

Fig 3: Templeton, Crandall, and Sing (TCS) network providing a summary of the diversity of *Pfcsp* haplotypes in the C-terminal region obtained from 440 DNA sequences in both Cape Coast (n = 184) and Navrongo (n = 256). Circles represent each *Pfcsp* C-terminal haplotype, and circles are scaled according to the frequency with which the haplotype was observed. Each haplotype is denoted as "Hap" with the vaccine strain 3D7 (3D7 0304600.1, PlasmoDB [19]) denoted as "3D7". Haplotypes obtained from the Navrongo sequences are color coded red and haplotypes obtained from Cape Coast coded green.

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

Population	n	Calculated indices[1]				
		h	S	K	$\pi \pm$ S.D	Hd$\pm$ S.D
Cape Coast	184	15	8	1.15	0.005 ± 0.0004	0.718 ± 0.026
Navrongo	256	53	16	3.76	0.016 ± 0.0007	0.925 ± 0.009

[1]Note; n= number of sequences, h = number of unique haplotypes, S= number of segregating sites, K = average number of pairwise nucleotide differences, π = nucleotide diversity, Hd = haplotype diversity

[1]Note; n= number of sequences, h = number of unique haplotypes, S= number of segregating sites, K = average number of pairwise nucleotide differences, π = nucleotide diversity, Hd = haplotype diversity

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 all the isolates in both TH2R and TH3R epitope regions with implications for cross-protection. Of the 92 (184 amino acid haplotypes) and 128 (256 amino acid haplotypes) isolates from Cape Coast and Navrongo, there were 8 and 27 non-vaccine TH2R haplotypes respectively (see **Additional file 2**). There were also 2 and 10 non-vaccine TH3R haplotypes in Cape Coast and Navrongo respectively, with 1 non-vaccine haplotype (NKPKDELNYAND) shared between the two populations (**Additional file 2**). The frequency of the Pf3D7-type TH2R vaccine haplotype (PSDKHIKEYLNKIQNSL) was 56.5% and 7.4% in Cape Coast and Navrongo respectively (**Fig 4A**). While there was 79.3% and 18.4% for the Pf3D7-type TH3R vaccine haplotype (NKPKDELDYAND) (**Fig 4B**) 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 between 1 - 6 in both epitope regions.

Sequence logos were generated to assess patterns and conservation of these amino acid polymorphisms within the CS epitopes (**Fig 4C and 4D**). The regions outside the TH2R and TH3R epitopes remained highly conserved with only two polymorphic sites (298 and 301) in both populations. The sequence logos indicate that although the two geographically distinct Ghanaian parasite populations had varying allele frequencies at specific polymorphic loci, the amino acids were conserved between populations. In Cape Coast, four coding sites within the TH2R epitope regions namely 314, 321, 324, 327 and two within the TH3R epitope regions were polymorphic, namely; 352 and 359. In Navrongo, five sites within the TH2R epitopes; 314, 317, 321, 324, 327 and also five sites in TH3R epitopes; 352, 354, 356, 357 and 359 were polymorphic. Between the two parasite populations, there were common polymorphic amino acid loci, namely; 314, 321, 324 and 327 in the TH2R epitope and 352 and 359 in the TH3R epitope region. Overall, there were similar non-Pf3D7 alleles at particular amino acids positions in the TH2R and TH3R epitopes in both parasite populations, namely; K314Q, L327I, and D359N (**Fig 4C, 4D and 5**).

Fig 4: Plot **(A and B)** showing the percentage of isolates sharing specific amino acid haplotypes within the TH2R (311-327aa) and TH3R (352-363aa) epitope regions in both Cape Coast and Navrongo population respectively. Colored 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 blue colored column from the bottom. The proportion of samples having the non-vaccine haplotypes is shown in the rest of the colored column bars**.** Images **(C and D)** are weblogos showing amino acid sequence conservation and polymorphisms of the C-terminal region of the circumsporozoite protein from Cape Coast (**panel C**) and Navrongo (**panelD)** respectively**.** The TH2R region and TH3R region are underlined in black and red in both Panels. The sequence conservation is measured in bits. Lower values represent

polymorphism at that amino acid position while higher values represent conservation. The TH2R and TH3R sites appear to be very polymorphic in both populations meanwhile the rest of the amino acids in both populations appear to be more conserved.

Fig 5. A plot showing the distribution of each amino acid in the TH2R **(A)** and TH3R **(B)** epitopes. The reference 3D7 amino acid sequences are shown on the x axis. The amino acid column represents the frequency of amino acids seen in isolates from Cape Coast (left half of the column) and Navrongo (right half of the column).

Population differentiation and structure of *Pfcsp*

Overall Weir and Cockerham's Fst between Cape Coast and Navrongo *Pfcsp* populations was <0.05 (**Fig 6A**) which signifies minimal population differentiation due to genetic structure and suggesting 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**).

Fig 6. Population differentiation and structure. (A). Weir and Cockerham's Fst calculated at SNP loci between Cape Coast (n=92) and Navrongo (n=128) population samples for 10 SNPs. The red line shows the borderline of 0.05 which signifies moderate population differentiation **(B).** Plot of first (PC1) and second principal component (PC2) of samples in both Cape Coast (n= 83) and Navrongo (n= 123) population after outlier samples were removed.

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 *Pfcsp* (221,422-221,583) for a subpopulation of monoclonal isolates from Navrongo (**Fig7A**) suggesting balancing selection. However, a Tajima's D <0 was seen in the Cape Coast population at these loci suggest likely directional selection or clonal expansion in the population. Alleles at SNP locus 221554 which is within the segment coding for 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 locus 221554 in the Navrongo population, but no long-range haplotypes extended beyond 221554 (**Fig 8A & 8B**).

Fig 7: (A) Tajima's D plot in sliding windows of 100 and a step size of 10 of 66 monoclonal samples (Cape Coast) and 65 monoclonal samples (Navrongo). Region highlighted in grey represents SNPs located in the C-terminal region (221,422-221,583) of *Pfcsp.* Plots **(B)** and **(C)** shows evidence of signatures of positive directional selection in chromosome 3 using the standardized integrated haplotype score (|iHS|) plotted as -log10 (P-value) for monoclonal isolates (66 and 65 samples) obtained in both Cape Coast **(B)** and Navrongo **(C)** respectively. IHS were calculated for SNPs with no missing data and a

minor allele frequency of >0.05. SNPs on the chromosome 3 are identified as red in Cape Coast **(B)** and blue in Navrongo **(C)**. Horizontal lines indicate threshold for high scoring SNPs with standardized |iHS| >3. Vertical lines indicate the position of the *Pfcsp* in both Cape Coast and Navrongo respectively. Evidence of positive directional selection was observed at *Pfcsp* loci in Navrongo however there was no evidence of selection seen in the *Pfcsp* loci in Cape Coast.

Fig 8:Extended Haplotype Homozygosity (EHH) and Bifurcation diagram. (A) Plots of EHH showing extended haplotypes from a focal SNP locus (221554). **(B)** Bifurcation diagrams showing the breakdown of these extended haplotypes from increasing distances in Navrongo parasite population. Evidence of positive directional selection was observed at this focal SNP locus which is found in the TH2R epitope loci of *Pfcsp*.

Discussion

The RTS,S/ASO1 malaria vaccine is based only on the *Pfcsp* sequence of the *P. falciparum* 3D7 clone [20] and strain specific immunity has been confirmed for the licensed vaccine [13]. To provide new insights into how well RTS,S/AS01 may perform if implemented 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 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, we observed a higher within host malaria parasite diversity in Navrongo with 49.2% of infections having *Fws* <0.95 than in Cape Coast where only 28.3% of infections had *Fws* <0.95. This high genetic diversity is known for high transmission areas where infected individuals usually harbor more polyclonal infections compared to those living in low transmission areas, where infections are often monoclonal [21,22]. Malaria transmission in Navrongo is much higher (EIR=157) than in Cape Coast (EIR= 50) [18,23]. These findings are consistent with high outcrossing potential in the parasite population in the KNDs compared to that in the coastal town of Cape Coast. This marked difference in within-host diversity is noteworthy for future region-specific vaccine intervention strategies.

We observed high genetic diversity in the C-terminal TH2R and TH3R amino acid epitopes in the two sites. Notably, the vaccine specific Pf3D7-type haplotype in the TH2R and TH3R epitope represent about 56.5% and 79.3% respectively of the observed haplotypes in Cape Coast and only about 7.4% and 18.4% of 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 [6–8,24–26]. Such polymorphisms at the T-cell epitopes have been suggested to be due to an immune evasion mechanism, in response to host T-cell immune responses [8] or selection in the mosquito host during the malaria transmission cycle [24]. Another hypothesis drawn from a previous study suggest that polymorphism at

the T cell epitopes could also be driven by an evolutionary response to intermolecular interactions at the surface of CSP [27].

The high degree of location specific *Pfcsp* 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. Monitoring differential vaccine efficacy by *Pfcsp* haplotypes during the RTS,S/AS01 implementation programs will be valuable for such a high transmission area, where post-vaccination expansion of non-vaccine haplotypes in the population is likely to be observed and this could lead to reduced vaccine efficacy and vaccine breakthrough infections.

Adaptive immunity from the RTS,S/AS01 vaccine is partly mediated by T cells epitopes (TH2R and TH3R) localized in the C-terminal region of PfCSP [28,29]. The conformation of epitopes may therefore be altered by non-synonymous substitutions consequently compromising the host immune response [30,31]. Our analysis revealed high levels of non-synonymous substitutions in the immunogenic TH2R and TH3R epitopes. We observed amino acid differences within the TH2R and TH3R epitope regions ranging from 1 to 6 at each epitope in both parasite populations which is similar to amino acid haplotype differences observed in the C-terminal region in the Zambian and DRC population, ranging from 2 to 10 [32]. However, there were more amino acid substitutions in Navrongo parasite populations than in Cape Coast, which is consistent with the lower frequency of vaccine haplotype observed in the network analysis for the Navrongo parasite population and this will have implications for the vaccine efficacy in comparable high malaria burden populations in Ghana.

An ancillary study conducted following Phase III RTS,S/AS01 vaccine clinical trials showed that immunity of the RTS,S vaccine waned significantly for parasites not matching Pf3D7-type haplotype in the *Pfcsp* C-terminal region, and this correlates with increasing amino acid differences [13]. Therefore, in populations where the Pf3D7-type vaccine haplotype is less prevalent, rapid decay in immunity may result in selection of vaccine escape mutants and undermine the overall effectiveness of RTS,S vaccine.

Our data shows that abundance of rare alleles in both Cape Coast and Navrongo contribute to the parasite population. Despite this high level of genetic diversity resulting from non-synonymous nucleotide and amino acid substitutions observed, there remained a shared gene pool between the two sites that resulted in a largely homogeneous parasite population. Over the sampled range of 784.4 km kilometers between the two sites, there was gene flow between the local populations of *P. falciparum* based on *Pfcsp* sequence analysis, with pairwise index of differentiation (Fst) being less than 0.05 [33]. 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 of *P. falciparum* parasites [33, 34]. Despite the ecological and epidemiological diversity between the 2 sites**,** human movement between the two sites is significant and could be accounting for *Pfcsp* gene flow within Country.

We observed negative Tajima's D in the Cape Coast isolates signifying 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 results corroborates finding from Thiès, Senegal where increased deployment of malaria control interventions resulted in an increase in the frequency of clonal strains and decrease in the probability of multiple infections [36]. Evidence of recent positive and balancing selection was observed in the Navrongo parasite isolates. Majority of alleles present at the C-terminal region in the Navrongo parasite population had a positive Tajima's D score and were highly polymorphic, which is likely due to balancing selection in response to host immune pressure on this immunogenic epitope [24,37,38]. Evidence of balancing selection on *Pfcsp* had been reported previously for a population from Malawi [24]. Balancing selection is common for immune targets and has been reported in other vaccine antigen candidates such as in the domain I epitope of Pf38 gene (found on the merozoite surface) in Papau New Guinea and the Gambia [39] and also in the extracellular domains of AMA1, a target of allele-specific immune responses [40]. However, seasonal genetic drift among loci attributable to sampling across multiple transmission seasons in Navrongo population may contribute to the balancing selection observed. We also observed evidence of recent positive directional selection (iHS >3) at the T-cell epitope loci in the Navrongo parasite population, this could be due to the addition of new and useful alleles to the already existing repertoire of alleles which are being maintained by balancing selection in the population [16]. On the other hand, the signature of positive selection observed in Navrongo could likely be attributed to the dominance of one allele against the others at the T-cell epitope region in the Navrongo parasite population. Considering the differences in 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.

The samples analyzed here were non-randomly selected from the population and this may have some limitations and bias the inferences that can be drawn from *Pfcsp* 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, geographic and temporal distributions are continually updated and expanded. Another limitation that may affect the data interpretation is the small sample sizes analyzed. Finally, the Navrongo sequences obtained from Pf3k represent different time periods, in comparison with sequences from Cape Coast which were sampled from same periods, which may prevent samples from these two regions from being optimally comparable. We however, observed no population structure within and between the two populations, an indication that timespan did not affect the results obtained here. 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. A larger sample size, with wider geographic and temporal analysis will further reveal the full extent of the diversity of *Pfcsp* locally and across Africa. This will help inform strategies for a wider implementation of the RTS,S vaccine.

Conclusion

The extent of polymorphisms of CSP observed in our study sites would likely implicate an allele-specific immune response during the pilot Phase IV implementation trials being conducted in Ghana. Consequentially, 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 [41]. This might slowly result in a directional selective advantage of unmatched CSP haplotypes because the vaccine does not target them. This lays emphasis on the need for a polyvalent malaria vaccine [42,43].

With the ongoing Phase IV RTS,S vaccine implementation trials in Ghana, which includes populations from Cape Coast and Navrongo, the findings from this study provides prior information on the extent of diversity in *Pfcsp* and the evolutionary forces driving these variations within Ghanaian natural parasite populations. This will inform the vaccine implementation outcomes and contribute to future vaccine designs. 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.

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
PDNA:	Plasmodium Network Diversity Africa
VCF:	Variant Call Format
HBsAg:	Hepatitis B Surface Antigen
SNP:	Single Nucleotide Polymorphism
NANP:	Asparagine-Alanine-Asparagine-Proline
NVDP:	Asparagine-Valine-Aspartate-Proline
GATK:	Genome Analysis Tool Kit
PCA:	Principal Component Analysis
MAF:	Minor Allele Frequency
HIS:	Integrated Haplotype Score
EHH:	Extended Haplotype Homozygosity

Declarations

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 (NHRCIRB203). Written informed consent was obtained from individuals, parents or guardians of all children before enrollment.

Consent for publication

Not Applicable

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)

Competing interests

Authors declare no conflict of interest

Funding

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

Authors' contributions

EAA analyzed 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, supported data analysis and the writing of the manuscript. PO and KM supervised the data analysis and contributed to the drafting of the manuscript.

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 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 and AAN LA, worked on this project as part of their DELGEME aspiring leadership fellowship and as co-applicants 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' information

Not Applicable

References

1. World Health Organization. World Malaria Report. 2019.
2. 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(9988):31–45.
3. Stoute JA, Kester KE, Krzych U, Wellde BT, Hall T, White K, Glenn G, Ockenhouse CF, Garcon N, Schwenk R, Lanar DE, Sun P, Momin P, Wirtz RA, Golenda C, Slaoui M, Wortmann G, Holland C, Dowler M, Cohen J, Ballou WR. Long-Term Efficacy and Immune Responses following Immunization with the RTS,S Malaria Vaccine. J Infect Dis. 1998; (4):1139-1144
4. Casares S, Brumeanu TD, Richie TL. The RTS,S malaria vaccine. Vaccine. 2010;28(31):4880–94.
5. 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(12).
6. Gandhi K, Thera MA, Coulibaly D, Traoré K, Guindo AB, Ouattara A, Takala-Harrison S, Berry AA, Doumbo OK, Plowe C V. Variation in the Circumsporozoite Protein of Plasmodium falciparum: Vaccine Development Implications. PLoS One. 2014 Jul 3;9(7):e101783.
7. 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(1):1–9.
8. Zeeshan M, Alam MT, Vinayak S, Bora H, Tyagi RK, Alam MS, Choudhary V, Mittra P, Lumb V, Bharti PK, Udhayakumar V, Singh N, Jain V, Singh PP, Sharma YD. Genetic Variation in the Plasmodium falciparum Circumsporozoite Protein in India and Its Relevance to RTS,S Malaria Vaccine. PLoS One. 2012;7(8):e43430.
9. Alloueche A, Milligan P, Conway DJ, Pinder M, Bojang K, Doherty T, Tornieporth N, Cohen J, Greenwood BM. Protective efficacy of the RTS,S/AS02 Plasmodium falciparum malaria vaccine is not strain specific. Am J Trop Med Hyg. 2003;68(1):97–101.
10. Bojang KA, Milligan PJM, Pinder M, Vigneron L, Alloueche A, Kester KE, Ballou WR, Conway DJ, Reece WHH, Gothard P, Yamuah L, Delchambre M, Voss G, Greenwood BM, Hill A, McAdam KPWJ, Tornieporth N, Cohen JD, Doherty T. 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(9297):1927–34.
11. Enosse S, Dobaño C, Quelhas D, Aponte JJ, Lievens M, Leach A, Sacarlal J, Greenwood B, Milman J, Dubovsky F, Cohen J, Thompson R, Ballou WR, Alonso PL, Conway DJ, Sutherland CJ. RTS,S/AS02A Malaria Vaccine Does Not Induce Parasite CSP T Cell Epitope Selection and Reduces Multiplicity of Infection. PLoS Clin Trials. 2006;1(1):e5.

12. Waitumbi JN, Anyona SB, Hunja CW, Kifude CM, Polhemus ME, Walsh DS, Ockenhouse CF, Heppner DG, Leach A, Lievens M, Ballou WR, Cohen JD, Sutherland CJ. 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(11):1–8.
13. Neafsey DE, Juraska M, Bedford T, Benkeser D, Valim C, Griggs A, Lievens M, Abdulla S, Adjei S, Agbenyega T, Agnandji ST, Aide P, Anderson S, Ansong D, Aponte JJ, Asante KP, Bejon P, Birkett AJ, Bruls M, Connolly KM, D'Alessandro U, Dobaño C, Gesase S, Greenwood B, Grimsby J, Tinto H, Hamel MJ, Hoffman I, Kamthunzi P, Kariuki S, Kremsner PG, Leach A, Lell B, Lennon NJ, Lusingu J, Marsh K, Martinson F, Molel JT, Moss EL, Njuguna P, Ockenhouse CF, Ogutu BR, Otieno W, Otieno L, Otieno K, Owusu-Agyei S, Park DJ, Pellé K, Robbins D, Russ C, Ryan EM, Sacarlal J, Sogoloff B, Sorgho H, Tanner M, Theander T, Valea I, Volkman SK, Yu Q, Lapierre D, Birren BW, Gilbert PB, Wirth DF. Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine. N Engl J Med. 2015;373(21):2025–37.
14. Takala SL, Plowe C V. efficacy : Preventing and overcoming " vaccine resistant malaria ." Parasite Immuno. 2010;31(9):560–73.
15. Pringle JC, Carpi G, Almagro-Garcia J, Zhu SJ, Kobayashi T, Mulenga M, Bobanga T, Chaponda M, Moss WJ, Norris DE. RTS,S/AS01 malaria vaccine mismatch observed among Plasmodium falciparum isolates from southern and central Africa and globally. Sci Rep. 2018;8(1):1–8.
16. Duffy CW, Assefa SA, Abugri J, Amoako N, Owusu-Agyei S, Anyorigiya T, MacInnis B, Kwiatkowski DP, Conway DJ, Awandare GA. Comparison of genomic signatures of selection on Plasmodium falciparum between different regions of a country with high malaria endemicity. BMC Genomics. 2015;16(1):527.
17. Koram KA, Owusu-Agyei S, Fryauff DJ, Anto F, Atuguba F, Hodgson A, Hoffman SL, Nkrumah FK. Seasonal profiles of malaria infection, anaemia, and bednet use among age groups and communities in northern Ghana. Trop Med Int Heal. 2003;8(9):793–802.
18. Abuaku B, Duah-Quashie NO, Quaye L, Matrevi SA, Quashie N, Gyasi A, Owusu-Antwi F, Malm K, Koram K. Therapeutic efficacy of artesunate-amodiaquine and artemether-lumefantrine combinations for uncomplicated malaria in 10 sentinel sites across Ghana: 2015-2017. Malar J. 2019;18(1):1–12.
19. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ, Treatman C, Wang H. PlasmoDB: A functional genomic database for malaria parasites. Nucleic Acids Res. 2009;37(SUPPL. 1):539–43.
20. Pringle JC, Carpi G, Almagro-Garcia J, Zhu SJ, Kobayashi T, Mulenga M, Bobanga T, Chaponda M, Moss WJ, Norris DE. RTS,S/AS01 malaria vaccine mismatch observed among Plasmodium falciparum isolates from southern and central Africa and globally. Sci Rep. 2018;8(1):1–8.
21. Adjah J, Fiadzoe B, Ayanful-Torgby R, Amoah LE. Seasonal variations in Plasmodium falciparum genetic diversity and multiplicity of infection in asymptomatic children living in southern Ghana.

BMC Infect Dis. 2018;18(1):1–10.

22. Peyerl-Hoffmann G, Jelinek T, Kilian A, Kabagambe G, Metzger WG, Von Sonnenburg F. Genetic diversity of Plasmodium falciparum and its relationship to parasite density in an area with different malaria endemicities in West Uganda. Trop Med Int Heal. 2001;6(8):607–13.
23. Oduro AR, Wak G, Azongo D, Debpuur C, Wontuo P, Kondayire F, Welaga P, Bawah A, Nazzar A, Williams J, Hodgson A, Binka F. Profile of the Navrongo health and demographic surveillance system. Int J Epidemiol. 2012;41(4):968–76.
24. Bailey JA, Mvalo T, Aragam N, Weiser M, Congdon S, Kamwendo D, Martinson F, Hoffman I, Meshnick SR, Juliano JJ. 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(4):580–7.
25. Escalante AA, Grebert HM, Isea R, Goldman IF, Basco L, Magris M, Biswas S, Kariuki S, Lal AA. 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(1–2):83–90.
26. Kumkhaek C, Phra-ek K, Singhasivanon P, Looareesuwan S, Hirunpetcharat C, Brockman A, Grüner AC, Lebrun N, Rénia L, Nosten F, Snounou G, Khusmith S. 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(2):281–7.
27. Aragam NR, Thayer KM, Nge N, Hoffman I, Martinson F, Kamwendo D, Lin FC, Sutherland C, Bailey JA, Juliano JJ. Diversity of T Cell Epitopes in Plasmodium falciparum Circumsporozoite Protein Likely Due to Protein-Protein Interactions. PLoS One. 2013;8(5):1–13.
28. Good MF, Pombo D, Quakyi IA, Riley EM, Houghten RA, Menon A, Alling DW, Berzofsky JA, Miller LH. Human T-cell recognition of the circumsporozoite protein of Plasmodium falciparum: immunodominant T-cell domains map to the polymorphic regions of the molecule. Proc Natl Acad Sci U S A. 1988 Feb 1;85(4):1199–203.
29. Riley EM, Allen SJ, Bennet S, Thomas PJ, Donnell AO, Lindsay SW, Good MF, Greenwood BM. Recognition of dominant T cell-stimulating circumsporozoite protein of Plasmodium malaria morbidity in Gambian children epitopes falciparum from the and relationship to malaria morbidity in Gambian children. Transactions of the Royal Society of Tropical Medicine and Hygiene.1990;84(5):648-57.
30. Kumar KA, Sano G, Boscardin S, Nussenzweig RS, Nussenzweig MC, Zavala F, Nussenzweig V. The circumsporozoite protein is an immunodominant protective antigen in irradiated sporozoites. Nature. 2006;444(7121):937-40.
31. Chenet SM, Tapia LL, Escalante AA, Durand S, Lucas C, Bacon DJ. Genetic diversity and population structure of genes encoding vaccine candidate antigens of Plasmodium vivax. Malar J. 2012 Mar 14;11(1):68.
32. Pringle JC, Carpi G, Almagro-Garcia J, Zhu SJ, Kobayashi T, Mulenga M, Bobanga T, Chaponda M, Moss WJ, Norris DE. RTS,S/AS01 malaria vaccine mismatch observed among Plasmodium

falciparum isolates from southern and central Africa and globally. Sci Rep. 2018;8(1):1–8.

33. Duffy CW, Ba H, Assefa S, Ahouidi AD, Deh YB, Tandia A, Kirsebom FCM, Kwiatkowski DP, Conway DJ. Population genetic structure and adaptation of malaria parasites on the edge of endemic distribution. Mol Ecol. 2017;26(11):2880–94.
34. Amambua-Ngwa A, Amenga-Etego L, Kamau E, Amato R, Ghansah A, Golassa L, Randrianarivelojosia M, Ishengoma D, Apinjoh T, Maïga-Ascofaré O, Andagalu B, Yavo W, Bouyou-Akotet M, Kolapo O, Mane K, Worwui A, Jeffries D, Simpson V, D'Alessandro U, Kwiatkowski D, Djimde AA. Major subpopulations of Plasmodium falciparum in sub-Saharan Africa . Science. 2019;365(6455):813–6.
35. Henden L, Lee S, Mueller I, Barry A, Bahlo M. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. Vol. 14, PLoS Genetics. 2018. 1–31 p.
36. Daniels R, Chang HH, Séne PD, Park DC, Neafsey DE, Schaffner SF, Hamilton EJ, Lukens AK, Van Tyne D, Mboup S, Sabeti PC, Ndiaye D, Wirth DF, Hartl DL, Volkman SK. Genetic Surveillance Detects Both Clonal and Epidemic Transmission of Malaria following Enhanced Intervention in Senegal. PLoS One. 2013;8(4):e60780.
37. Tetteh KKA, Stewart LB, Ochola LI, Amambua-Ngwa A, Thomas AW, Marsh K, Weedall GD, Conway DJ. Prospective identification of malaria parasite genes under balancing selection. PLoS One. 2009;4(5) :e5568.
38. Weedall GD, Conway DJ. Detecting signatures of balancing selection to identify targets of anti-parasite immunity. Trends in Parasitology. 2010;26(7):363-9.
39. 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(1):126.
40. Polley SD, Conway DJ. Strong diversifying selection on domains of the Plasmodium falciparum apical membrane antigen 1 gene. Genetics. 2001;158(4):1505–12.
41. Thera MA, Doumbo OK, Coulibaly D, Diallo DA, Kone AK, Guindo AB, Traore K. Safety and Immunogenicity of an AMA-1 Malaria Vaccine in Malian Adults : Results of a Phase 1 Randomized Controlled Trial. PloS one. 2008;3(1):e1465.
42. Barry AE, Arnott A. Strategies for designing and monitoring malaria vaccines targeting diverse antigens. Front Immunol. 2014;5:1–16.
43. Dutta S, Seung YL, Batchelor AH, Lanar DE. Structural basis of antigenic escape of a malaria vaccine candidate. Proc Natl Acad Sci U S A. 2007;104(30):12488–93.
44. Mensah-Brown HE, Abugri J, Asante KP, Dwomoh D, Dosoo D, Atuguba F, Conway DJ, Awandare GA. Assessing the impact of differences in malaria transmission intensity on clinical and haematological indices in children with malaria. Malar J. 2017;16(1):1–11.
45. Kasasa S, Asoala V, Gosoniu L, Anto F, Adjuik M, Tindana C, Smith T, Owusu-Agyei S, Vounatsou P. Spatio-temporal malaria transmission patterns in Navrongo demographic surveillance site, northern Ghana. Malar J. 2013;12(1):1–10.

46. Snounou G, Viriyakosol S, Xin Ping Zhu, Jarra W, Pinheiro L, do Rosario VE, Thaithong S, Brown KN. High sensitivity of detection of human malaria parasites by the use of nested polymerase chain reaction. Mol Biochem Parasitol. 1993;61(2):315–20.
47. MalariaGEN Database. http://www.malariagen.net/projects/parasite/pf. Accessed 2 January 2019.
48. Amato R, Miotto O, Woodrow CJ, Almagro-Garcia J, Sinha I, Campino S, Mead D, Drury E, Kekre M, Sanders M, Amambua-Ngwa A, Amaratunga C, Amenga-Etego L, Andrianaranjaka V, Apinjoh T, Ashley E, Auburn S, Awandare GA, Baraka V, Barry A, Boni MF, Borrmann S, Bousema T, Branch O, Bull PC, Chotivanich K, Conway DJ, Craig A, Day NP, Djimdé A, Dolecek C, Dondorp AM, Drakeley C, Duffy P, Echeverry DF, Egwang TG, Fairhurst RM, Faiz A, Fanello CI, Hien TT, Hodgson A, Imwong M, Ishengoma D, Lim P, Lon C, Marfurt J, Marsh K, Mayxay M, Michon P, Mobegi V, Mokuolu OA, Montgomery J, Mueller I, Kyaw MP, Newton PN, Nosten F, Noviyanti R, Nzila A, Ocholla H, Oduro A, Onyamboko M, Ouedraogo JB, Phyo APP, Plowe C, Price RN, Pukrittayakamee S, Randrianarivelojosia M, Ringwald P, Ruiz L, Saunders D, Shayo A, Siba P, Takala-Harrison S, Thanh TNN, Thathy V, Verra F, Wendler J, White NJ, Ye H, Cornelius VJ, Giacomantonio R, Muddyman D, Henrichs C, Malangone C, Jyothi D, Pearson RD, Rayner JC, McVean G, Rockett KA, Miles A, Vauterin P, Jeffery B, Manske M, Stalker J, Macinnis B, Kwiatkowski DP. Genomic epidemiology of artemisinin resistant malaria. Elife. 2016;5:1–29.
49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. 2009;
50. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
51. Browning BL. Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. The American Journal of Human Genetics. 2018;103(3):338-48.
52. Auburn S, Campino S, Miotto O, Djimde AA, Zongo I, Manske M, Maslen G, Mangano V, Alcock D, MacInnis B, Rockett KA, Clark TG, Doumbo OK, Ouédraogo JB, Kwiatkowski DP. Characterization of within-host plasmodium falciparum diversity using next-generation sequence data. PLoS One. 2012;7(2):3–9.
53. Manske M, Miotto O, Campino S, Auburn S, Zongo I, Ouedraogo J, Michon P, Mueller I, Su X, Amaratunga C, Fairhurst R, Socheat D, Imwong M, White NJ, Sanders M, Anastasi E, Rubio VR, Jyothi D, Amenga-etego L. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature. 2012;487(7407):375-9.
54. Lee Stuart. Bahlo M. moimix: an R package for assessing clonality in high-throughput sequencing data. 2016.
55. Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sanchez-Gracia A. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol. 2017;34(12):3299–302.

56. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 1979;76(10):5269–73.
57. Nei M. Molecular Evolutionary Genetics. New York: Columbia University Press; 1987.
58. Clement M, Snell Q, Walker P, Posada D, Crandall K. HICOMB2002-03.pdf. 1997;
59. 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(2):619–33.
60. Leigh JW, Bryant D. POPART: Full-feature software for haplotype network construction. Methods Ecol Evol. 2015;6(9):1110–6.
61. Crooks G, Hon G, Chandonia J, Brenner S. NCBI GenBank FTP Site\nWebLogo: a sequence logo generator. Genome Res. 2004;14:1188–90.
62. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
63. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):2074–93.
64. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;
65. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4(3):0446–58.
66. Gautier M, Vitalis R. Rehh An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics. 2012;28(8):1176–7.
67. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko J V., Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419(6909):832–7.

Figures

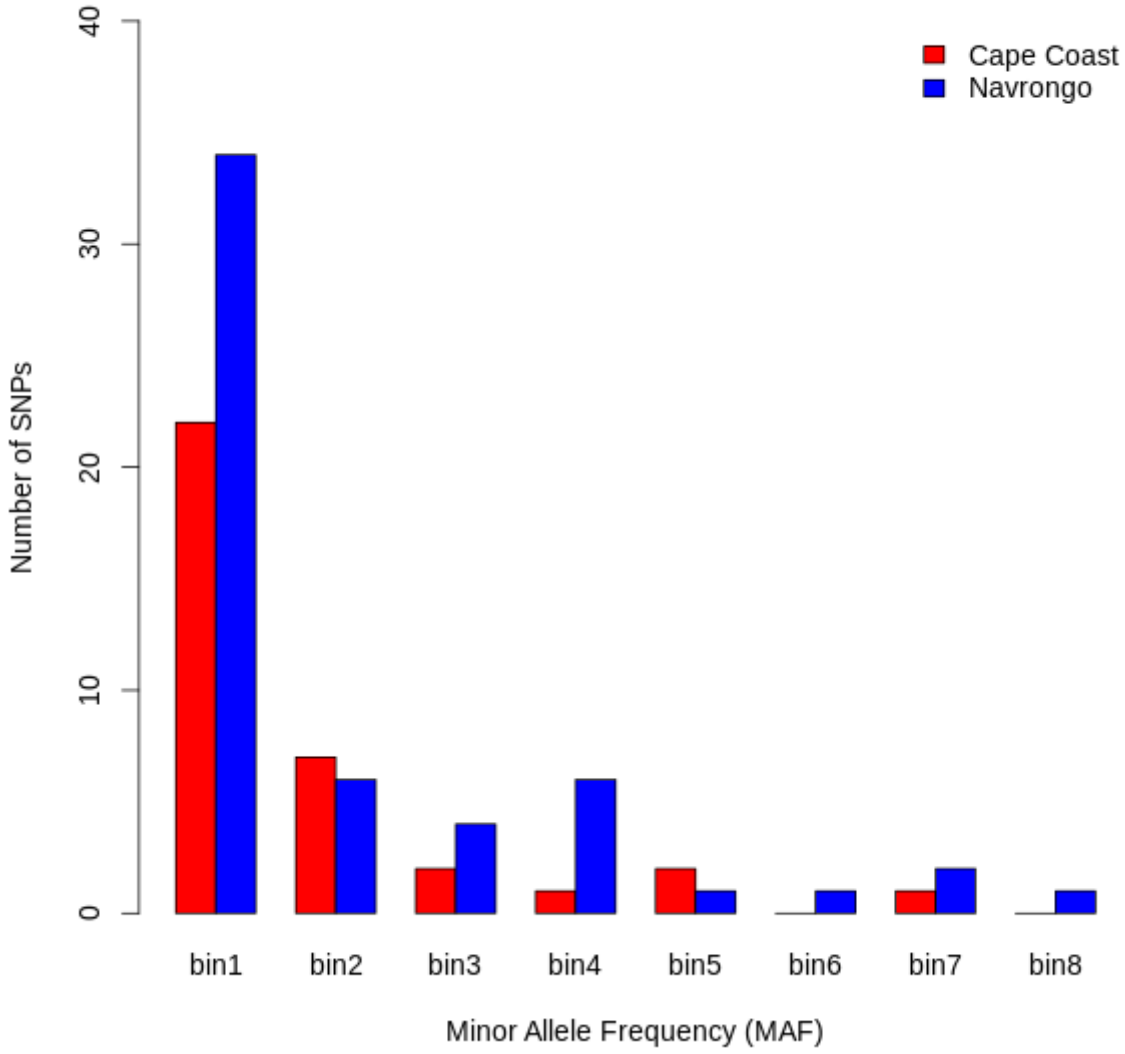

Figure 1

A histogram showing the minor allele frequency distribution of a total of 90 SNPs set within the Pfcsp loci in samples from both Cape Coast (n= 35 SNPs) and Navrongo (n =55 SNPs). Vertical axis represents the number of SNPs in each category of allele frequency and the horizontal axis shows the binned SNPs set with bin1 to bin8 representing the following MAFs ranges; [0.0-0.01], [0.01-0.05], [0.05-0.10], [0.10-0.15], [0.15-0.20], [0.20-0.25], [0.35-0.40], [0.40-0.45]) respectively. There were no alleles found within the [0.25-0.30] and [0.30-0.35] bins in both populations.

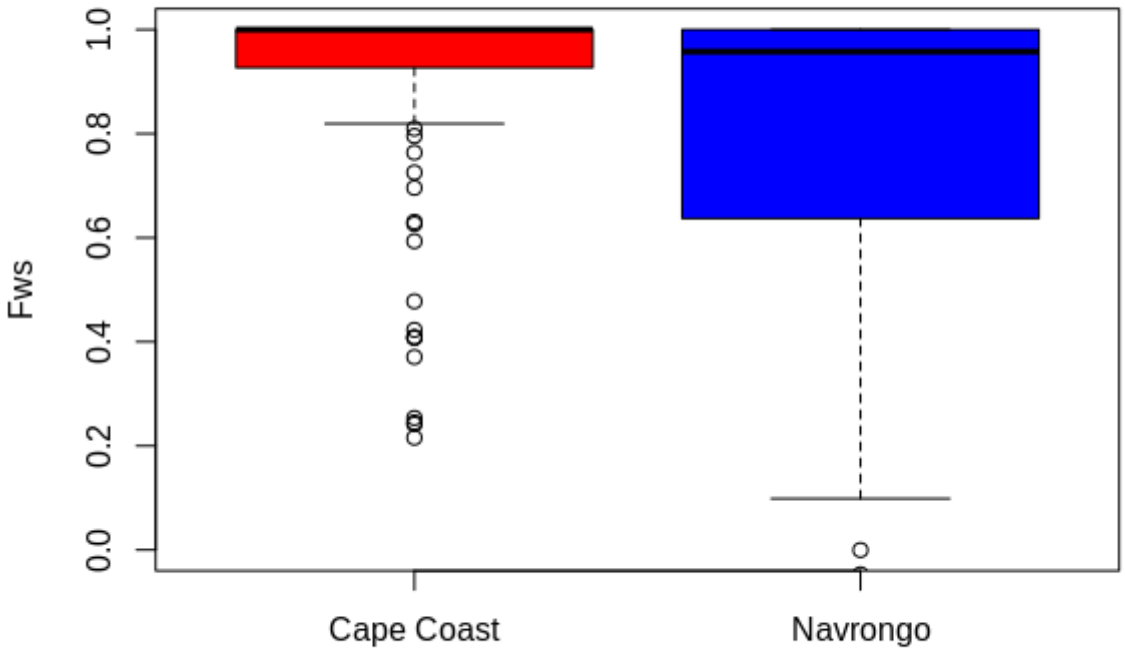

Figure 2

Within host diversity. Boxplot showing the distribution of Fws estimates in samples from Cape Coast and Navrongo populations. Cape Coast highlighted in red and Navrongo highlighted in blue.

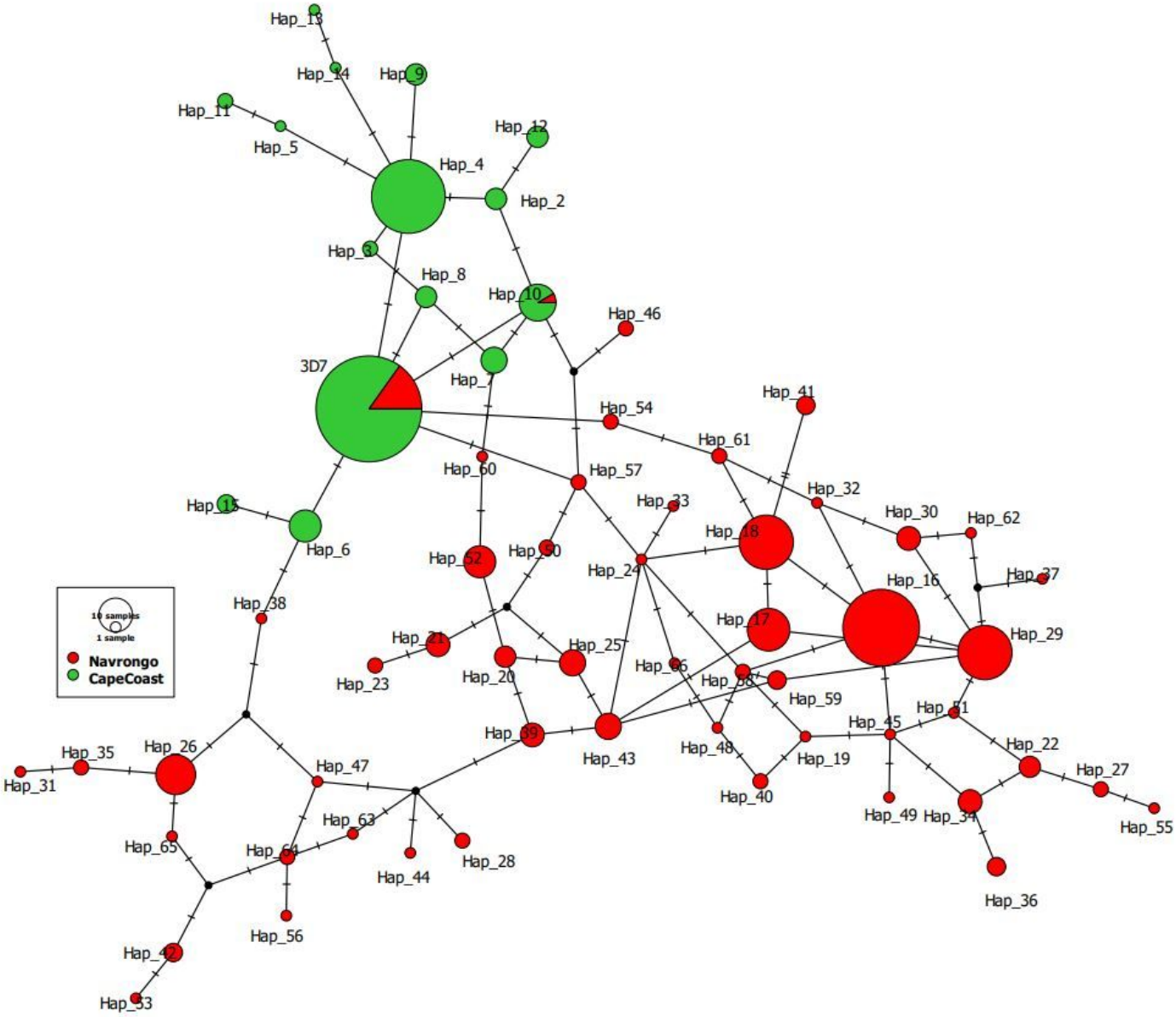

Figure 3

Templeton, Crandall, and Sing (TCS) network providing a summary of the diversity of Pfcsp haplotypes in the C-terminal region obtained from 440 DNA sequences in both Cape Coast (n = 184) and Navrongo (n = 256). Circles represent each Pfcsp C-terminal haplotype, and circles are scaled according to the frequency with which the haplotype was observed. Each haplotype is denoted as “Hap” with the vaccine strain 3D7 (3D7 0304600.1, PlasmoDB [19]) denoted as “3D7”. Haplotypes obtained from the Navrongo sequences are color coded red and haplotypes obtained from Cape Coast coded green.

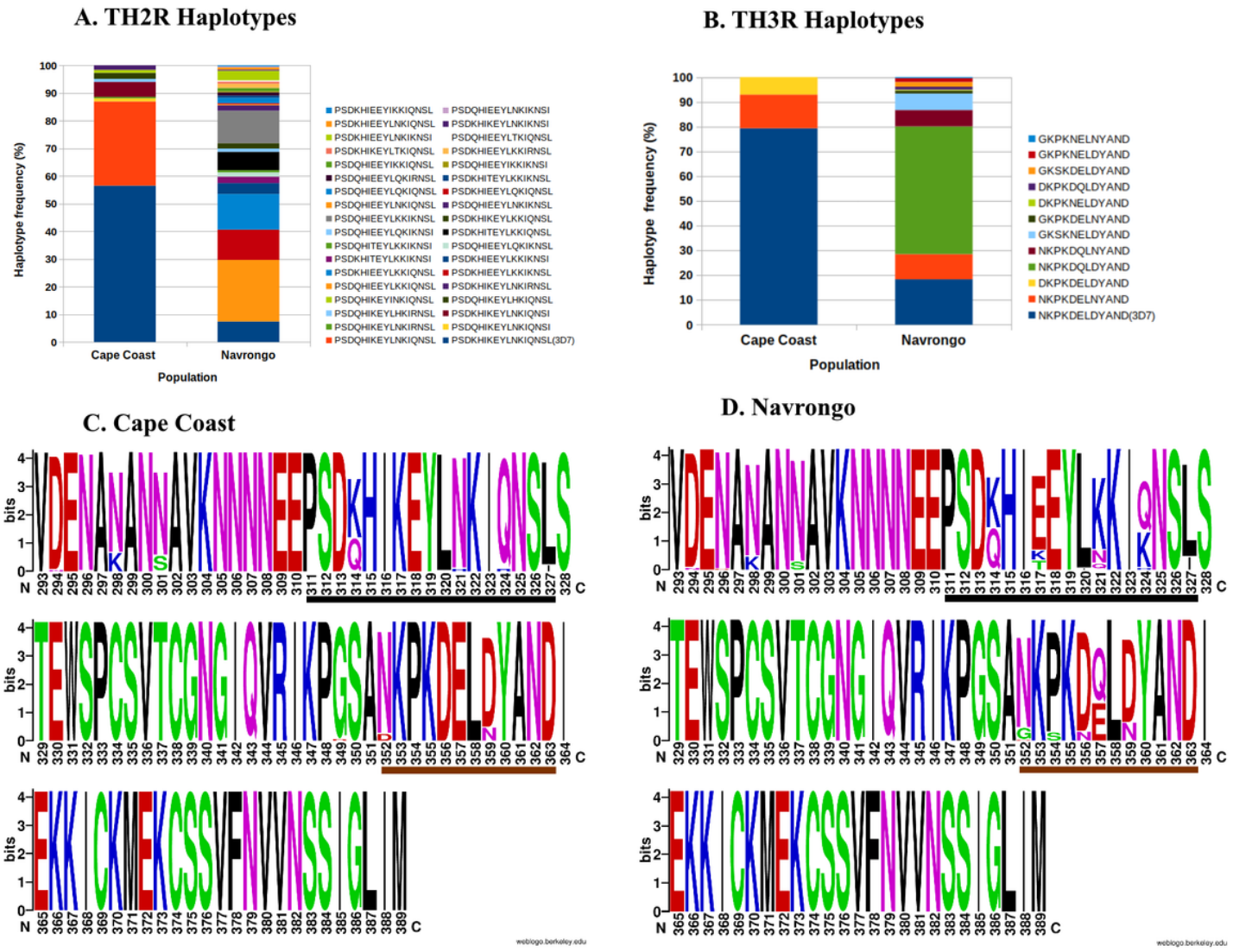

Figure 4

Plot (A and B) showing the percentage of isolates sharing specific amino acid haplotypes within the TH2R (311-327aa) and TH3R (352-363aa) epitope regions in both Cape Coast and Navrongo population respectively. Colored 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 blue colored column from the bottom. The proportion of samples having the non-vaccine haplotypes is shown in the rest of the colored column bars. Images (C and D) are weblogos showing amino acid sequence conservation and polymorphisms of the C-terminal region of the circumsporozoite protein from Cape Coast (panel C) and Navrongo (panel D) respectively. The TH2R region and TH3R region are underlined in black and red in both Panels. The sequence conservation is measured in bits. Lower values represent polymorphism at that amino acid position while higher values represent conservation. The TH2R and TH3R sites appear to be very polymorphic in both populations meanwhile the rest of the amino acids in both populations appear to be more conserved.

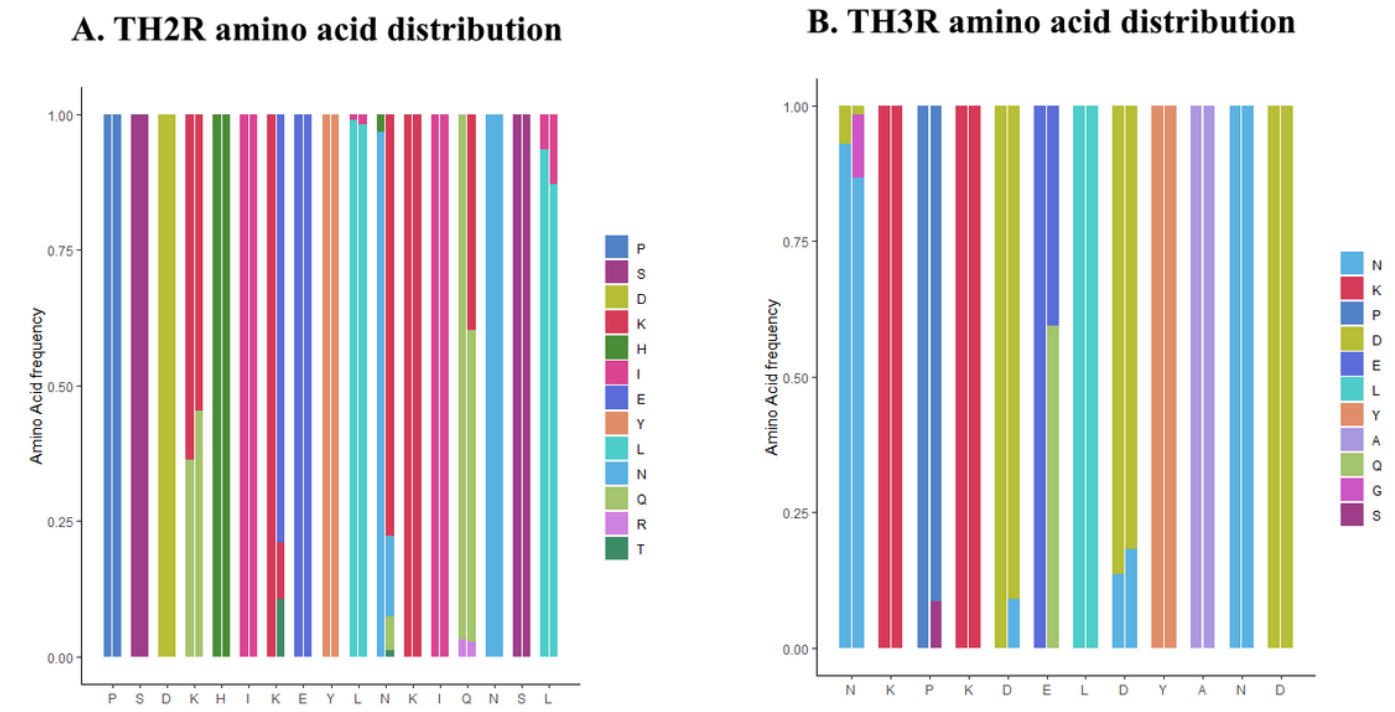

Figure 5

A plot showing the distribution of each amino acid in the TH2R (A) and TH3R (B) epitopes. The reference 3D7 amino acid sequences are shown on the x axis. The amino acid column represents the frequency of amino acids seen in isolates from Cape Coast (left half of the column) and Navrongo (right half of the column).

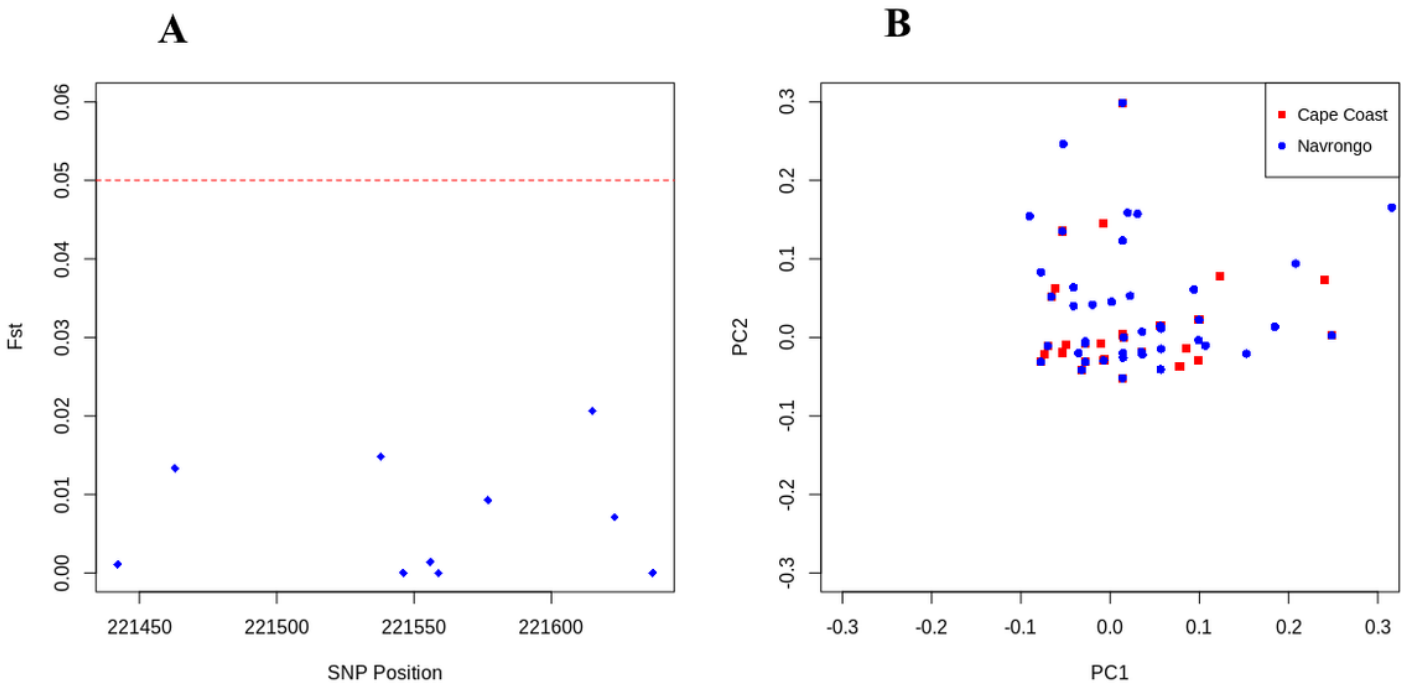

Figure 6

Population differentiation and structure. (A). Weir and Cockerham's Fst calculated at SNP loci between Cape Coast (n=92) and Navrongo (n=128) population samples for 10 SNPs. The red line shows the borderline of 0.05 which signifies moderate population differentiation (B). Plot of first (PC1) and second principal component (PC2) of samples in both Cape Coast (n= 83) and Navrongo (n= 123) population after outlier samples were removed.

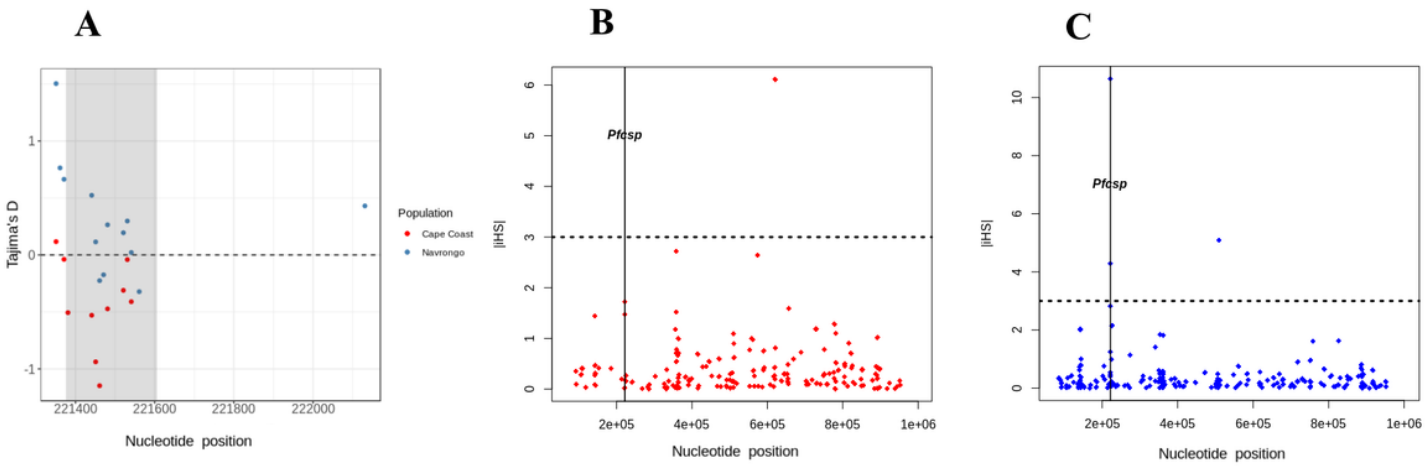

Figure 7

(A) Tajima's D plot in sliding windows of 100 and a step size of 10 of 66 monoclonal samples (Cape Coast) and 65 monoclonal samples (Navrongo). Region highlighted in grey represents SNPs located in the C-terminal region (221,422-221,583) of Pfcsp. Plots (B) and (C) shows evidence of signatures of positive directional selection in chromosome 3 using the standardized integrated haplotype score (|iHS|) plotted as -log10 (P-value) for monoclonal isolates (66 and 65 samples) obtained in both Cape Coast (B) and Navrongo (C) respectively. IHS were calculated for SNPs with no missing data and a minor allele frequency of >0.05. SNPs on the chromosome 3 are identified as red in Cape Coast (B) and blue in Navrongo (C). Horizontal lines indicate threshold for high scoring SNPs with standardized |iHS| >3. Vertical lines indicate the position of the Pfcsp in both Cape Coast and Navrongo respectively. Evidence of positive directional selection was observed at Pfcsp loci in Navrongo however there was no evidence of selection seen in the Pfcsp loci in Cape Coast.

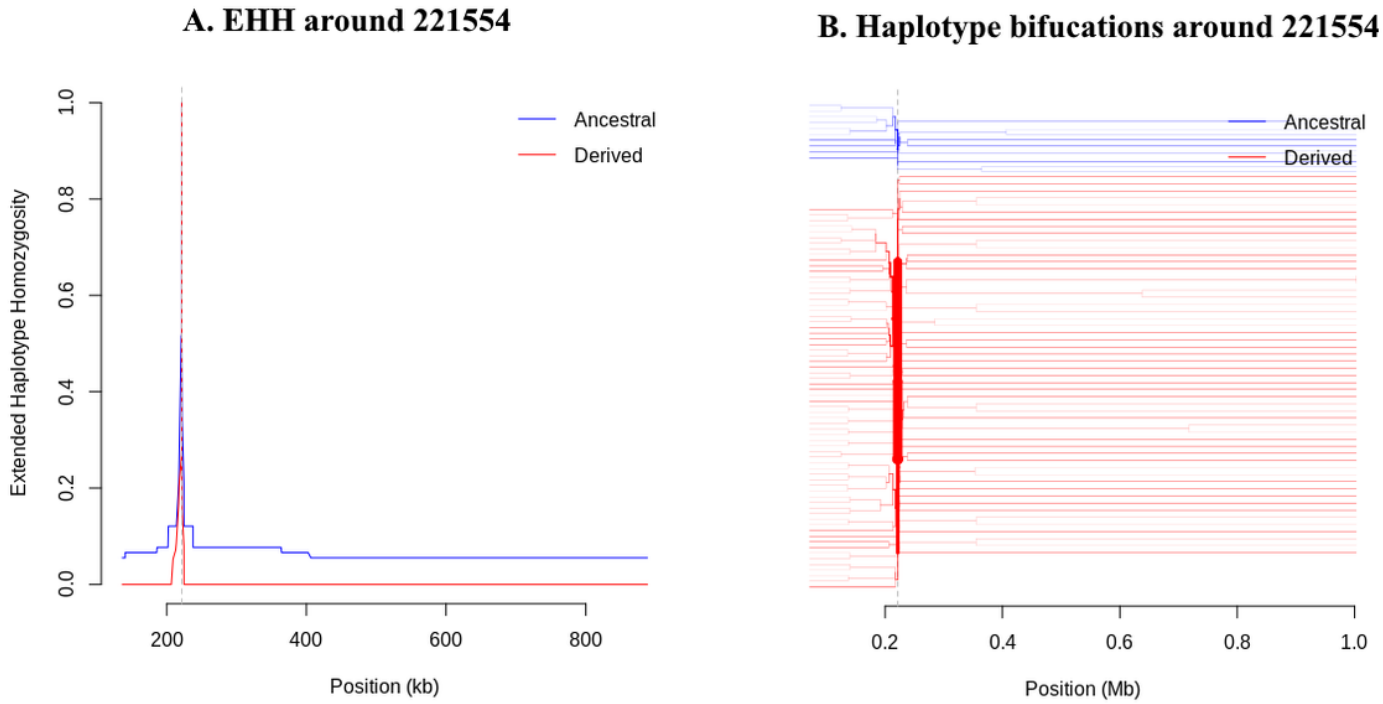

Figure 8

Extended Haplotype Homozygosity (EHH) and Bifurcation diagram. (A) Plots of EHH showing extended haplotypes from a focal SNP locus (221554). (B) Bifurcation diagrams showing the breakdown of these extended haplotypes from increasing distances in Navrongo parasite population. Evidence of positive directional selection was observed at this focal SNP locus which is found in the TH2R epitope loci of Pfcsp.

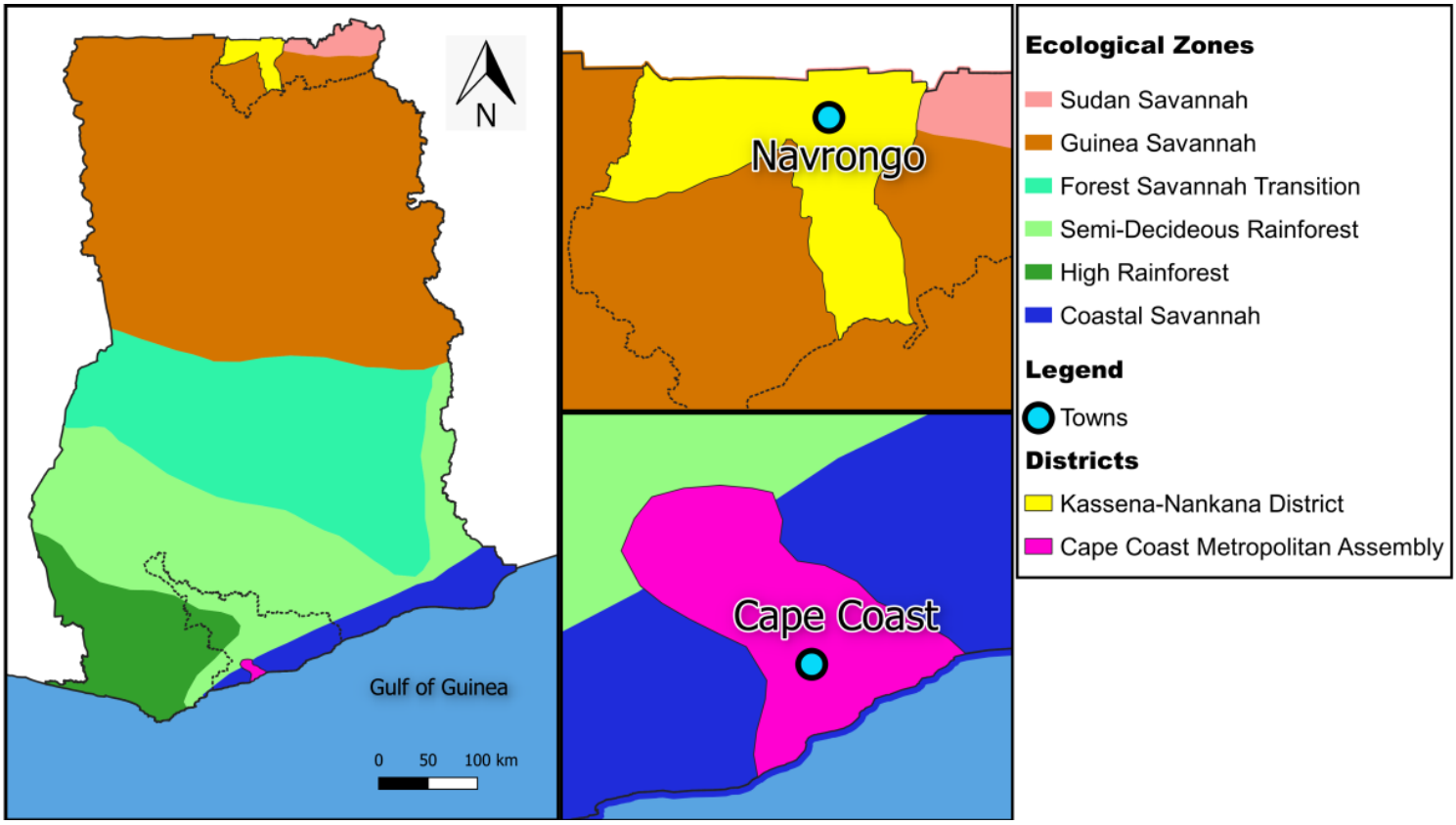

Figure 9

Location of the Navrongo within the Kassena-Nankana Districts and the Cape Coast within the Cape Coast Metropolis. The distance between Cape Coast and Navrongo is approximately 784.8 km (generated by QGIS software).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile2.xlsx
- Additionalfile1.nex