Identification of complex *Plasmodium falciparum* genetic backgrounds circulating in Africa: a multi-country genomic epidemiology analysis

Olivo Miotto PhD[1,2,*], Prof Alfred Amambua-Ngwa PhD[3,4], Lucas N Amenga-Etego PhD[5], Prof Muzamil M Abdel Hamid PhD[6], Prof Ishag Adam PhD[7], Enoch Aninagyei PhD[8], Tobias Apinjoh PhD[9], Prof Gordon A Awandare PhD[5], Prof Philip Bejon PhD[10], Gwladys I Bertin PhD[11], Prof Marielle Bouyou-Akotet MD[12], Antoine Claessens PhD[13,3], Prof David J Conway PhD[4], Prof Umberto D'Alessandro PhD[3], Prof Mahamadou Diakite DPhil[14], Prof Abdoulaye Djimdé PhD[14], Prof Arjen M Dondorp MD[1,2], Patrick Duffy MD[15], Rick M Fairhurst PhD[15], Caterina I Fanello PhD[1,2], Anita Ghansah PhD[16], Deus S Ishengoma PhD[17], Mara Lawniczak PhD[18], Oumou Maïga-Ascofaré PhD[19], Sarah Auburn PhD[20], Prof Anna Rosanas-Urgell PhD[21], Varanya Wasakul PhD[1], Nina FD White PhD[18], Alexandria Harrott MS[18], Jacob Almagro-Garcia PhD[18], Richard D Pearson PhD[18], Sonia Goncalves PhD[18], Cristina Ariani PhD[18], Prof Zbynek Bozdech PhD[22], William Hamilton MD[18], Victoria Simpson PhD[18], Prof Dominic P Kwiatkowski FRS[23]

[1]Mahidol-Oxford Tropical Medicine Research Unit, Mahidol University, Bangkok, Thailand
[2]Centre for Tropical Medicine and Global Health, University of Oxford, Oxford, UK
[3]Medical Research Council Unit The Gambia at LSHTM, Banjul, The Gambia
[4]London School of Hygiene and Tropical Medicine, London, UK
[5]West African Centre for Cell Biology of Infectious Pathogens (WACCBIP), University of Ghana, Accra, Ghana
[6]Institute of Endemic Diseases, University of Khartoum, Khartoum, Sudan
[7]Department of Obstetrics and Gynaecology, Unaizah College of Medicine and Medical Sciences, Qassim University, Unaizah, Saudi Arabia
[8]Department of Biomedical Sciences of School of Basic and Biomedical Sciences, University of Health and Allied Science, Ho, Ghana
[9]Department of Biochemistry and Molecular Biology, University of Buea, Buea, Cameroon
[10]KEMRI Wellcome Trust Research Programme, Kilifi, Kenya
[11]Institute of Research for Development (IRD), Paris, France
[12]Faculty of Medicine, University of Health Sciences, Libreville, Gabon
[13]LPHI, MIVEGEC, INSERM, CNRS, IRD, University of Montpellier, Montpellier, France
[14]Malaria Research and Training Centre, University of Science, Techniques and Technologies of Bamako, Bamako, Mali
[15]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville MD, USA
[16]Noguchi Memorial Institute for Medical Research (NMIMR), Accra, Ghana
[17]National Institute for Medical Research (NIMR), Dar Es Salaam, Tanzania
[18]Wellcome Sanger Institute, Hinxton, UK
[19]Bernhard Nocht Institute for Topical Medicine (BNITM), Hamburg, Germany
[20]Menzies School of Health Research, Charles Darwin University, Darwin, Australia
[21]Institute of Tropical Medicine Antwerp, Antwerp, Belgium
[22]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore
[23]MRC Centre for Genomics and Global Health, Big Data Institute, Oxford University, Oxford, UK (posthumous)

[*] Corresponding Author: olivo@tropmedres.ac

ABSTRACT

Background

The population structure of the malaria parasite *Plasmodium falciparum* can reveal underlying adaptive evolutionary processes. In particular, selective pressures to maintain complex genetic backgrounds may encourage inbreeding, producing distinct parasite clusters identifiable by population structure analyses.

Methods

We analysed population structure in 3,783 *P. falciparum* genomes from 21 countries across Africa, provided by the MalariaGEN Pf7 dataset. We used Principal Coordinate Analysis to cluster parasites; *identity by descent* (IBD) methods to identify genomic regions shared by cluster members; and linkage analyses to determine their co-inheritance patterns. Structural variants were reconstructed by *de novo* assembly, and verified by long-read sequencing.

Findings

We identified a strongly differentiated cluster of parasites, named AF1, comprising ~1.2% of samples analysed, distributed over 13 countries across Africa, at locations over 7,000 km apart. Members of this cluster share a complex genetic background, consisting of up to 23 loci harbouring a large number of highly differentiated variants, rarely observed outside the cluster. IBD analyses revealed common ancestry at these loci, irrespective of sampling location; outside the shared loci, however, AF1 members appear to outbreed with sympatric parasites. The AF1 differentiated variants comprise structural variations, including a gene conversion involving the DBLMSP and DBLMSP2 genes, and numerous single nucleotide polymorphisms. Several of the genes harbouring these mutations are functionally related, often involved in interactions with red blood cells including invasion, egress and erythrocyte antigen export.

Interpretation

We propose that AF1 parasites have adapted to some unidentified evolutionary niche, probably associated to host erythrocyte interactions, by acquiring a complex compendium of interacting variants that are otherwise rarely seen in Africa, which appears to remain mostly intact despite recombination events. The term *cryptotype* was coined to describe a common background interspersed with genomic regions of local origin.

Funding

Bill & Melinda Gates Foundation

RESEARCH IN CONTEXT

Evidence before this study

This study builds on previous work by the authors to elucidate regional population structure- notably sub-populations driven by artemisinin resistance in the Greater Mekong Subregion (GMS), and resistance to drugs in Africa and Oceania. Here, we sought to identify new population structure patterns in Africa, applying methods based on identity-by-descent (IBD) algorithms. These methods are based on IBD techniques developed in population structure analyses in New Guinea and southern Laos. We searched PubMed, without language restrictions up to 30 January 2024, for literature pertaining *Plasmodium* IBD-based population structure analysis. A search for relevant literature (terms: falciparum, ("population structure" OR subpopulations), "identity by descent") yielded 9 peer-reviewed publications, including 4 studies that analysed data from the MalariaGEN whole-genome sequence dataset. Although most studies were on a national scale, we reviewed one global study, and regional studies from the GMS, South America and Africa; the latter describing results complementary to those presented here.

Added value of this study

We analysed population structure by clustering *P. falciparum* genomes by similarity and by extent of IBD. Due to high transmission and frequent recombination, African parasites are mostly expected to exhibit low levels of similarity, except in highly inbred, geographically confined. Contrary to this, we found a group of parasites (named AF1), present at low frequency across the continent, whose members share several portions of the genome. The genomic regions forming this complex genetic background appear to be co-inherited and in strong linkage disequilibrium. They are also strongly differentiated, comprising many loci (>20) that carry alleles rarely seen in other African parasites, including large structural variants. In spite of this constellation of co-inherited loci, AF1 parasites appear to recombine with local non-AF1 individuals, such that some degree of geographical differentiation is seen within the group. The most commonly shared loci within AF1 contain genes known to interact with host erythrocytes, participating in invasion and egress, or exporting antigens to the red blood cell surface.

Implications of all the available evidence

This study has identified a novel phenomenon in malaria genetic epidemiology, which we dubbed *cryptotype* since identifying AF1 required specific analyses of ancestry. While previous studies have

identified subpopulations of highly similar parasites, these were typically localized geographically and driven by recent selection. The geographical extent of the AF1 population, from Madagascar to Mauritania, indicates it is neither localized nor recent. Its discovery suggests we need to rethink our understanding of *Plasmodium falciparum* epidemiology and evolution. How is such a complex constellation of mutated loci maintained, in spite of the extremely low likelihood of passing it on intact to the progeny after recombination? One possible explanation is that AF1 occupies a niche where the ensemble of mutations provides an adaptation, conferring a survival advantage. The functions of the genes involved suggest that this involves host-parasite interactions, but further studies will be required to elucidate the underlying biology. Meanwhile, the present work provides experimental parasitologists with a catalogue of candidate interacting variants that can form the basis for new investigations.

INTRODUCTION

The protozoan *Plasmodium falciparum*, a leading cause of malaria, is responsible for hundreds of thousands of deaths yearly in Sub-Saharan Africa.[1] This parasite has shown great propensity for genetic changes in response to human interventions, often undermining malaria control and elimination efforts.[2] The recent availability of high-throughput genome sequencing has made it possible to study such changes in near-real time, providing important insights into the dynamics of evolution at the population level.[3-5] In particular, studies of *P. falciparum* population structure– the differences in the distribution of genetic variation between populations– have revealed insights into *P. falciparum* demography by identifying patterns associated with deviations from random mating.

Where malaria transmission is high, large parasite populations and frequent infection rates provide frequent mating opportunities for genetically distinct parasites, maintaining high levels of genetic variation through outbreeding. Hence, genetic distances within these populations tend to be evenly distributed, without significant population structure, as seen in parts of Africa.[6] In areas of low malaria transmission, on the other hand, mosquitoes often acquire parasites from a single infected individual, which results in mating between clones with identical genomes, or *selfing*. High levels of selfing result in inbred populations, which exhibit lower genetic distances between individuals, and can be detected in population genomics analyses. High levels of inbreeding may also be observed when selfing is advantageous for parasite survival: on average, a single variants will only propagate to half of all offspring when mating with a wild-type parasite, but to all offspring when selfing. Population structure driven by drug-resistant mutations was observed in Southeast Asia, where inbred artemisinin-resistant populations were associated to mutations in the *kelch13* gene.[7,8] The benefits of high selfing rates are even greater when transmitting complex genetic backgrounds, e.g. when a drug-resistant mutation is detrimental to parasite development unless accompanied by multiple compensatory mutations.[9] In the case of artemisinin resistance in the GMS, at least five loci were found to be co-inherited with key *kelch13* mutations.[8] Due to recombination, a greater number of co-inherited loci has a lower probability of passing a complete set of variants to offspring when outbreeding. However, if the full set of variants strongly increases survival likelihood, then lineages from selfing parasites may undergo selection, resulting in reduced genetic variation.

Analyses of population structure in sub-Saharan Africa have shown high levels of genetic variations in high-transmission regions, with gradual genetic differentiation between East and West Africa.[10] Population structure can be observed at the margins of endemicity, in lower transmission regions such as the Gambia and the Horn of Africa.[10,11] To date, however, no published analyses have

reported population structure driven by the selection of complex co-inherited multi-locus genetic backgrounds.

We conducted an analysis of African genomes from the MalariaGEN Pf7 dataset[6] to search for patterns of population structure associated with complex genetic backgrounds. By applying methods based on *identity by descent* (IBD), we characterized a group of parasites, labelled AF1, which share a complex multi-locus genetic background, whose components appear to be co-inherited. AF1 parasites are found at low frequency across Africa, from Mauritania to Madagascar. We defined the term *cryptotype* to describe their genetic background, reflecting the fact that it is "hidden" by large portions of the genome that bear similarities to other local parasites. We investigated functional relationships between the cryptotype component loci, and the forces that may be contributing to the maintenance of this complex and geographically widespread genetic background.

METHODS

The process of selection of samples and variants is detailed in the Supplementary Text; the following is a brief summary. We began with the MalariaGEN Pf7 dataset,[6] which comprises 20,864 samples. We selected "essentially clonal" samples (F_{WS} ≥0.95) from Africa, discarding those that had high genotyping missingness, resulting in a set of 3,783 samples, organized by macroregions: West, Central and East Africa- labelled WAF, CAF and EAF respectively (Table 1). Samples were genotyped at 743,584 high-quality biallelic SNPs that had a minor allele frequency (MAF) ≥0.1% in at least one macroregion. Samples were genotyped at each SNP with the allele supported by the most reads. Allele frequencies were estimated at each SNP by calculating the proportion of samples carrying each allele, disregarding samples with missing genotypes. Pairwise F_{ST} was estimated at each SP as previously described.[8] The AF1 mean F_{ST} mean was calculated as the arithmetic mean of F_{ST} between AF1 and each of the macroregions WAF, CAF and EAF. F_{ST} estimation was also performed at 68,360 additional SNPs that had high levels of missingness in sWGA-amplified samples (see Supplementary Text).

Genotype analyses were performed using bespoke software programs written in Java and R (v4.4.0, https://www.r-project.org/). PCoA analyses were conducted using `cmdscale` in the R `stats` package with a NxN pairwise genetic distance matrix (N=3,783); genetic distances were estimated by the proportion of the 743,584 SNPs where two samples carry different alleles, after discarding SNPs where one or both of the two samples have a missing genotype. AF1 proportions and 95% confidence intervals were calculated by R `DescTools` package (v0.99.5419) using the Agresti-Coull

method. The linkage disequilibrium measure r^2 was computed for all pairs of SNPs with mean F_{ST}≥0.2 (Supplementary Text). Circular genome linkage disequilibrium plots were generated using circos (v0.69).[12]

Identity by descent (IBD) analysis was performed using the program hmmIBD[13] with default parameters. We filtered out extremely low-frequency variants, retaining coding SNPs with MAF ≥0.1 in at least one macroregion, and at least one sample with a non-reference genotype. High-IBD regions were defined by identifying uninterrupted sequences of SNPs where ≥50% of all AF1 pairs were in IBD; neighbouring high-IBD regions separated by gaps ≤50kbp were subsequently merged.

De novo assemblies of genomic sequencing reads were performed using Cortex v1.0.5.21[14] with kmer_size=61. The generated contigs were aligned against reference sequences provided by the Pf3k project (https://www.malariagen.net/projects/pf3k) using BioEdit (v7.2.5, https://thalljiscience.github.io/). Sequencing reads coverage visualizations were produced using the LookSeq web application[15] and JBrowse2 (v2.10.3).[16] MSP1 gene references were obtained from GenBank, accession numbers X03371.1 (K1), AB276005.1 (RO33) and X05624.2 (MAD30). Functional information about genes was obtained from PlasmoDB (https://plasmodb.org/plasmo/app/) and literature searches.

Role of the funding source

The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

RESULTS

Population structure analysis of African *P. falciparum* parasites

From all African genomes in the MalariaGEN Pf7 dataset,[6] we selected 3,783 samples from the quality-filtered Pf7 *analysis set*, which were essentially clonal (F_{WS} ≥ 0.95), and had low genotype missingness (Table 1). We estimated allele frequencies in three macroregions: West, Central and East Africa (labelled WAF, CAF and EAF respectively) for all high-quality biallelic single nucleotide polymorphisms (SNPs) in Pf7, and discarded SNPs with minor allele frequency (MAF) <0.1% in all three macroregions, yielding a set of 743,584 SNPs to be used in our analyses.

We performed principal coordinate analysis (PCoA), a method that maps samples onto a series of dimensions (*principal components*) to explain variance in a genetic distance matrix, clustering together highly similar genomes. The first component (PC1) was driven by the differentiation

between WAF and EAF parasites (Figure 1), as reported previously.[10] Unexpectedly, the second component (PC2) was driven by a diverging cluster, which we named AF1, composed of parasites from multiple countries across Africa, rather than from sites in close geographic proximity. The broad geographical distribution of AF1, which includes regions of high transmission, rules out population structure driven by low endemicity, suggesting that AF1 members share a high degree of similarity in a substantial portion of the genome, which differentiates them from other individuals from the same countries.

We labelled samples with PC2 ≥ 0.025 as AF1 members (Figure 1), while parasites with PC2 ≤ 0.01 were labelled according to their macroregion (WAF, CAF or EAF); samples with intermediate PC2 values (n=14) were disregarded. AF1 members comprised 47 samples (1.2% of the sample set), sampled from 13 countries across all macroregions, up to 7,500 km apart (Figure 2). Within most countries, AF1 accounts for 1-6% of samples, with significantly higher proportions in Guinea and Malawi only (Table 1). AF1 frequencies were also consistent by year, except for a higher proportion in 2011 (Supplementary Table 1), which is difficult to interpret since it coincided with the collections of samples in Guinea and Malawi. To a first level of approximation, AF1 appears to be evenly distributed at low frequency across the continent.

Genetic features of AF1

The clustering of AF1 parasites suggests they share alleles that are uncommon in other African populations. To identify differentiated sites, we estimated allele frequencies in AF1, WAF, CAF and EAF at all coding SNPs, to calculate the mean F_{ST} between AF1 and each of the other populations. For this task, we included 68,360 additional SNPs that had low coverage in sWGA-amplified samples (Supplementary Text). This analysis revealed 198 coding non-synonymous SNPs with mean F_{ST} ≥0.5, 71 of which had mean F_{ST} ≥0.75 (Supplementary Table 2). The differentiated SNPs are not evenly distributed across the genome, but clustered in several regions on multiple chromosomes (Supplementary Figure 1). We found high-F_{ST} variant clusters in chromosomes 1, 2, 4, 9, 10, 11, 13 and 14, while other chromosomes showed lower differentiation levels. The clustering of high-F_{ST} SNPs suggests that these *AF1 characteristic loci* contain highly differentiated long haplotypes. Although most SNP clusters occupy regions <100kbp, one locus on chromosome 10 stretches over ~250kbp, possibly indicating a haplotype under selection, or a large structural variant.

Given the marked differentiation at the AF1 characteristic loci, we predicted a strong correlation between alleles found in these regions. This was confirmed by computing r^2, a commonly used linkage disequilibrium measure,[17] for all distal pairs of SNPs with mean F_{ST}≥0.2. Several loci

contained highly correlated distal SNPs (r^2≥0.2); mapping these associations across the genome shows a complex linkage disequilibrium network (Figure 3). Seven differentiated loci each contained at least one SNP very strongly associated (r^2 ≥0.4) with SNPs at all other loci (Supplementary Table 3). This provides clear evidence that AF1 parasites possess a multi-component genetic background, carried as a complete set by most members. However, the exact composition of this background requires further analysis, since high r^2 values only occur where AF1 alleles are very rare outside AF1, which is not a requisite for a component locus.

Ancestry analysis

To address this question of whether AF1 shared alleles originate from different sources in different countries, or they have been co-inherited from common ancestry, we conducted an analysis of identity by descent (IBD) for all sample pairs. This analysis identifies genomic regions where parasites pairs are identical to an extent not explainable unless the two parasites have a common ancestry. AF1 parasites exhibited pairwise IBD at a much higher fraction of their genomes (median [IQR]: 22.4% [18-28%]) than non-AF1 parasites in WAF, CAF and EAF (medians [IQR]: 0.05% [0.00-0.79%], 0.8% [0.00-1.3%] and 1.2% [0.23-1.8%] respectively, Supplementary Figure 2A), suggesting that AF1 is differentiated by haplotypes with shared ancestries. This was confirmed by PCoA, using a distance measure derived from IBD genome fractions (Supplementary Figure 3). Although pairwise IBD levels are well above those in other African populations, AF1 is not a clonally expanding population: west African AF1 genomes shared significantly higher IBD fractions with WAF genomes than with EAF ones (medians [IQR]: 0.67% [0-1.3%] vs 0.18% [0-0.69%]), and vice versa (medians [IQR]: 0.59% [0-1.1%] vs 0.0% [0-0.73%], Supplementary Figure 2B), indicating that recombination occurs between AF1 parasites and non-AF1 local populations.

Hypothesizing that IBD is restricted to specific genomic regions, we mapped the frequency of IBD segments, identifying 23 “high-IBD” regions where >50% of all AF1 pairs were in IBD (Supplementary Figure 4). High-IBD regions were present in all chromosomes except chromosome 12, often near subtelomeric regions. Each high-IBD region contained one or more SNPs with mean F_{ST}>0.5 and AF1 characteristic allele frequency >0.5 (Supplementary Table 4). The high-F_{ST} SNPs, ranked by allele frequency, are effective markers for identifying AF1 members: 42/47 samples carry the AF1 characteristic alleles at all top 7 ranked SNPs, and no more than one non-AF1 allele at the top 13 SNPs (Figure 3B). Conversely, only one non-AF1 samples carried AF1 alleles at more than half of the six top ranked SNPs, suggesting that AF1 members can be discriminated by simple genetic tests.

Taken together, results from analyses of IBD, differentiation and correlation show that highly differentiated loci are mostly located in high-IBD regions and strongly linked across chromosomes (Figure 3A). We can deduce that AF1 parasites carry a constellation of variants that differentiate them from other African parasites; these appear to be inherited together, even though AF1 genomes recombine with sympatric strains. It appears that not all the loci involved are equally important: most AF1 members carry a "core" set of ~13 characteristic haplotypes, while other loci seem to be less critical components. All evidence suggests that the variant constellation is co-inherited, rather than having different ancestries in different countries.

Structural variants in chromosome 10 and 9

The top-ranked high-IBD region, on chromosome 10 (Figure 3C), is also the largest; due to its size, we hypothesized that it may harbour a structural variant. Sequencing read coverage showed that AF1 members had few or no reads mapping to genes MSP6 (PF3D7_1035500) and H101 (PF3D7_1035600), suggesting a large deletion (Supplementary Figure 5). The adjacent DBLMSP gene (PF3D7_1035700) was also poorly covered at the 5' end, but the presence of a proximal paralog (DBLMSP2, PF3D7_1036300) raised the possibility of short read mismapping. To clarify, we performed *de novo* assembly of the sequencing reads of an AF1 member from Mali (PM0293-C), mapping the resulting contigs to multiple DBLMSP and DBLMSP2 reference sequences. The AF1 DBLMSP sequence shows marked sequence similarity to PfIT (a South American strain), but a very different organization, being almost identical to the PfIT DBLMSP2 gene at the 5' end (Figure 4A). This suggests a gene conversion event, where the AF1 DBLMSP gene acquired the 5' portion of DBLMSP2; this explains the absence of coverage in that segment when aligning against the DBLMSP reference; this was confirmed by a realignment against the *de novo* assembled AF1 sequences (Supplementary Text, Supplementary Figure 6). We tested AF1 sequencing reads for a sequence containing the recombination breakpoint and flanking regions (Figure 4B, Supplementary Text), which confirmed the gene conversion in 42/47 samples. Both the gene conversion and the deletion of genes MSP6 and H101 were also confirmed by long-read assembly of an AF1 parasite from a different study[18] (Supplementary Text, Supplementary Figures 7 and 8). We observed that other genes in this region contain AF1 high-F_{ST} SNPs, including MSP3 (PF3D7_1035400) and the glutamate-rich protein GLURP (PF3D7_1035300).

The second-ranked high-IBD region, on chromosome 9 (Figure 3C), exhibits a highly differentiated haplotype in the merozoite surface protein MSP1 gene (PF3D7_0930300). Low coverage in some MSP1 regions was observed when aligning AF1 reads against the 3D7 reference (Supplementary Figure 9), suggesting that the AF1 sequence differ substantially from the reference. MSP1 is known

to consist of frequently recombining blocks, and has been classified based on the variants present in four blocks.[19] Alignments against reference strains showed that PM0293-C has a MAD20/K1/K1/K1 MSP1 sequence, uncommon in non-AF1 African populations, but more frequent in South America and Southeast Asia (Supplementary Table 2). The PM0293-C amino acid sequence is near-identical to that of PfHB3, a Mesoamerican strain. We confirmed this result by long-read re-sequencing of an amplicon spanning the entire MSP1 gene of an AF1 sample (Supplementary Figure 10, Supplementary Text), and by inspecting long-read assemblies from an earlier study (Supplementary Text).[18]

Functional analysis of AF1 characteristic loci

The large number of loci and the low frequency of the characteristic alleles in non-AF1 parasites suggest an extremely low probability of inheriting a full complement of AF1 alleles when recombining with non-AF1 parasites. Given that the complete allele constellation circulates at detectable frequency over a large territory, it is likely that it is under selection thanks to some fitness advantage, conferred by functionally related mutations. There are known relationships between the chromosome 10 and 9 loci: on the merozoite surface, MSP1 binds with other surface proteins, including DBLMSP, DBLMSP2 and MSP6.[20] The resulting complex plays a critical role in merozoite egress and invasion of erythrocytes, also involving SERA5 (PF3D7_0207600) and SERA6 (PF3D7_0207500), whose genes carry high-F_{ST} SNPs on chromosome 2.[21,22] High-F_{ST} variants in genes involved in erythrocyte interaction were found at other AF1 characteristic loci, including those encoding other merozoite surface proteins (MSP7 and MSP10), several PHIST gene family members,[23] and a number of proteins involved in exporting to the erythrocyte membrane, such as RESA3 (PF3D7_1149200), PfD80 (PF3D7_0401800), MAHRP1 (PF3D7_1370300),[24] Pf332 (PF3D7_1149000),[25] and the ring-exported proteins REX1 and REX2 (PF3D7_0935900 and PF3D7_0936000 respectively).[26,27] In addition, several genes encoding erythrocyte-exported proteins carry AF1 differentiated alleles, e.g. members of the FIKK and SURFIN families, and the cytoadherence-linked asexual gene CLAG9 (PF3D7_0935800). Thus, several AF1 characteristic variants are associated with common functional categories (Figure 3C). A functional enrichment analysis (Supplementary Text) confirmed that significant proportions of AF1 high-F_{ST} variants are associated with host cell surface, surface binding and processes of erythrocyte invasion and egress, as well as interactions with the immune system and regulatory functions (Supplementary Table 5). The evidence points to a constellation of variants that are functionally linked, and related to host-parasite interactions.

DISCUSSION

The analyses presented here, based on 3,783 high-quality *P. falciparum* genomes, identified a genetic background of remarkable complexity, circulating across the breadth of the African continent and maintaining its integrity without solely relying on inbreeding. To our knowledge, this is the first report of what we describe as a *cryptotype*, a complex inherited genetic background "hidden" within genomes that are otherwise similar to their sympatric parasites. Differently from clonally expanding populations,[5] IBD is not evenly distributed across the AF1 genome, but concentrated in numerous distal regions. The cryptotype's ability to retain identity at its characteristic loci, over the long period of time it must have taken to achieve its geographic spread, is hard to reconcile with the extremely low probability of retaining variant constellations intact through outbreeding. Therefore, it seems likely that the AF1 genomes are maintained through both frequent inbreeding and, far more rarely, acquisition of non-AF1 genes through outbreeding.

The fact that more than 20 identical AF1 variants are found in parasites from Madagascar, Ghana and Congo suggests a fine-tuned functional interplay between these loci, and a phenotypic benefit of carrying the complete constellation. Such functional benefit would help maintain AF1 at significant frequency- for example, by bestowing a selected fitness advantage, or by providing adaptation to a specific niche where AF1 is particularly competitive. Occupying an exclusive niche, e.g. a particular vector species or host population, would provide some level of reproductive isolation, promoting inbreeding and helping maintain the variant constellation. Although at this point we cannot determine the functional advantage conferred by the cryptotype, we note that many AF1 differentiated variants are functionally related. Several of the genes encode proteins that participate in erythrocyte egress and invasion, or export of parasite antigens to the red blood cell surface. Taken together, these lines of evidence suggest that the AF1 variant ensemble underpins phenotypic changes related to host erythrocyte interactions. We hypothesize that AF1 parasites have adapted to a specific erythrocyte-related host niche, e.g. a hemoglobinopathy that reduces invasion[28] or prevents erythrocyte remodelling.[29] Although the broad geographic distribution makes it unlikely that the cryptotype is fine-tuned to a specific human population, it is possible that its evolutionary niche involves a non-human host.

Our analysis opens several questions that will require further investigation. Culturing *in vitro* field isolates can elucidate the biological mechanisms underpinning the cryptotype and the properties conferring its selective advantage, and provide material for high-quality, high coverage long-read sequencing to investigate structural rearrangements. Identifying patients infected with AF1 parasites

may help characterize the cryptotype's evolutionary niche and understand its epidemiology. Given AF1's low prevalence, such studies will be challenging, but may produce important shifts in our understanding of invasion mechanisms and of protective human blood phenotypes. The wide-ranging catalogue of variants identified here can already provide experimental parasitologists with candidates for studying gene interactions and synergies.

A question emerging from this work is whether AF1 is the only cryptotype in Africa, or globally. AF1 parasites separate clearly in PCoA plots largely because their differentiated variants are mostly absent from other African populations, resulting in high levels of differentiation. However, the absence of characteristic alleles from the general population is not a requisite for cryptotypes. Clusters of individuals carrying co-inherited variants may be hard to detect by PCoA when these variants are common outside the clusters; alternative approaches may be needed, e.g. based on sensitive IBD detection algorithms. Furthermore, detecting cryptotypes at a low frequency may require larger genomic datasets. We have shown that analysing genomic data shared by a multitude of studies can lead to important discoveries. We advocate that repositories providing such data in organized and usable forms, such as those managed by MalariaGEN,[6] must continue to be supported by funders and contributing researchers alike, to power advancements in understanding of epidemiological phenomena.

Contributors

AAN, LAE, MMAH, IA, EA, TA, GAA, PB, GIB, MBA, AC, DJC, UD, MD, AD, AMD, PD, RF, CF, AG, DI, ML, OMA, SA, ARU organized or carried out sample collections. SG, AH conducted laboratory analyses. JA, RP, SG, AH, CA produced genomic data. OM, VW, NW, ZB, WH performed data analyses. OM, VS, DPK designed and coordinated the project. OM, ZB drafted the manuscript. OM accessed and verified all the data. All authors provided critical revision of the manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

We declare no competing interests.

Data sharing

Sequencing data used in the present study is publicly available as part of the open-access MalariaGEN Plasmodium falciparum Community Project v7 (Pf7)[6]

Acknowledgments

This study was funded by the Bill & Melinda Gates Foundation (grant numbers OPP11188166 and OPP1204268). This publication uses open-access data from the MalariaGEN *Plasmodium falciparum* Community Project v7 (Pf7) as described in https://doi.org/10.12688/wellcomeopenres.18681.1. The authors wish to thank all the patients and guardians who generously agreed to provide blood samples. We are indebted to all researchers who contributed samples to the Community Project since its inception, including the samples analyzed in the present work.

REFERENCES

1. World Health Organization. World Malaria Report 2023. Geneva: World Health Organization, 2023.
2. Malisa AL, Pearce RJ, Abdulla S, et al. Drug coverage in treatment of malaria and the consequences for resistance evolution--evidence from the use of sulphadoxine/pyrimethamine. *Malar J* 2010; **9**: 190.
3. Hamilton WL, Amato R, van der Pluijm RW, et al. Evolution and expansion of multidrug-resistant malaria in southeast Asia: a genomic epidemiology study. *Lancet Infect Dis* 2019; **19**(9): 943-51.
4. Conrad MD, Asua V, Garg S, et al. Evolution of Partial Resistance to Artemisinins in Malaria Parasites in Uganda. *N Engl J Med* 2023; **389**(8): 722-32.
5. Wasakul V, Disratthakit A, Mayxay M, et al. Malaria outbreak in Laos driven by a selective sweep for Plasmodium falciparum kelch13 R539T mutants: a genetic epidemiology analysis. *Lancet Infect Dis* 2023; **23**(5): 568-77.
6. MalariaGen, Abdel Hamid MM, Abdelraheem MH, et al. Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples. *Wellcome Open Res* 2023; **8**: 22.
7. Miotto O, Almagro-Garcia J, Manske M, et al. Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia. *Nature Genetics* 2013; **45**(6): 648-55.
8. Miotto O, Amato R, Ashley EA, et al. Genetic architecture of artemisinin-resistant Plasmodium falciparum. *Nat Genet* 2015; **47**(3): 226-34.
9. Amambua-Ngwa A, Button-Simons KA, Li X, et al. Chloroquine resistance evolution in Plasmodium falciparum is mediated by the putative amino acid transporter AAT1. *Nat Microbiol* 2023; **8**(7): 1213-26.
10. Amambua-Ngwa A, Amenga-Etego L, Kamau E, et al. Major subpopulations of Plasmodium falciparum in sub-Saharan Africa. *Science* 2019; **365**(6455): 813-6.
11. Amambua-Ngwa A, Jeffries D, Amato R, et al. Consistent signatures of selection from genomic analysis of pairs of temporal and spatial Plasmodium falciparum populations from The Gambia. *Sci Rep* 2018; **8**(1): 9687.
12. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res* 2009; **19**(9): 1639-45.
13. Schaffner SF, Taylor AR, Wong W, Wirth DF, Neafsey DE. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *Malar J* 2018; **17**(1): 196.
14. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics* 2012; **44**(2): 226-32.
15. Manske HM, Kwiatkowski DP. LookSeq: a browser-based viewer for deep sequencing data. *Genome Res* 2009; **19**(11): 2125-32.
16. Diesh C, Stevens GJ, Xie P, et al. JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol* 2023; **24**(1): 74.
17. Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theoret Appl Genet* 1968; **38**: 226-31.
18. Dara A, Drabek EF, Travassos MA, et al. New var reconstruction algorithm exposes high var sequence diversity in a single geographic location in Mali. *Genome Med* 2017; **9**(1): 30.
19. Ferreira MU, Kaneko O, Kimura M, Liu Q, Kawamoto F, Tanabe K. Allelic diversity at the merozoite surface protein-1 (MSP-1) locus in natural Plasmodium falciparum populations: a brief overview. *Memorias do Instituto Oswaldo Cruz* 1998; **93**(5): 631-8.
20. Lin CS, Uboldi AD, Epp C, et al. Multiple Plasmodium falciparum Merozoite Surface Protein 1 Complexes Mediate Merozoite Binding to Human Erythrocytes. *J Biol Chem* 2016; **291**(14): 7703-15.
21. Das S, Hertrich N, Perrin AJ, et al. Processing of Plasmodium falciparum Merozoite Surface Protein MSP1 Activates a Spectrin-Binding Function Enabling Parasite Egress from RBCs. *Cell Host Microbe* 2015; **18**(4): 433-44.

22. Koussis K, Withers-Martinez C, Yeoh S, et al. A multifunctional serine protease primes the malaria parasite for red blood cell invasion. *EMBO J* 2009; **28**(6): 725-35.
23. Sargeant TJ, Marti M, Caler E, et al. Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biol* 2006; **7**(2): R12.
24. Spycher C, Rug M, Pachlatko E, et al. The Maurer's cleft protein MAHRP1 is essential for trafficking of PfEMP1 to the surface of Plasmodium falciparum-infected erythrocytes. *Molecular Microbiology* 2008; **68**(5): 1300-14.
25. Nilsson S, Angeletti D, Wahlgren M, Chen Q, Moll K. Plasmodium falciparum antigen 332 is a resident peripheral membrane protein of Maurer's clefts. *PLoS One* 2012; **7**(11): e46980.
26. Spielmann T, Hawthorne PL, Dixon MW, et al. A cluster of ring stage-specific genes linked to a locus implicated in cytoadherence in Plasmodium falciparum codes for PEXEL-negative and PEXEL-positive proteins exported into the host cell. *Mol Biol Cell* 2006; **17**(8): 3613-24.
27. Dixon MW, Kenny S, McMillan PJ, et al. Genetic ablation of a Maurer's cleft protein prevents assembly of the Plasmodium falciparum virulence complex. *Mol Microbiol* 2011; **81**(4): 982-93.
28. Taylor SM, Cerami C, Fairhurst RM. Hemoglobinopathies: slicing the Gordian knot of Plasmodium falciparum malaria pathogenesis. *PLoS Pathog* 2013; **9**(5): e1003327.
29. Cyrklaff M, Sanchez CP, Kilian N, et al. Hemoglobins S and C interfere with actin remodeling in Plasmodium falciparum-infected erythrocytes. *Science* 2011; **334**(6060): 1283-6.

TABLES

Table 1 – Summary of sample counts by country.

Each row represents one African country where *P. falciparum* samples analysed in this study were samples. The columns show: the macroregion in which the country is located (West, Central or East Africa); the name of the country and its ISO 3166 code; the total number of analysed samples from that country; the number of AF1 samples identified in the country, their percentage of the samples analysed (with 95% confidence interval), and the p-value of a Fisher's exact test comparing the proportion within the country against the proportion in the rest of the continent ($p<0.01$ shown in bold type).

Macroregion	Country Name	Country Code	Sample Count	AF1 Count	AF1 %	95% C.I.	*p*
West Africa (WAF)	Mauritania	MR	49	1	2.0%	[0.0%, 12%]	0.46
	Mali	ML	534	4	0.75%	[0.20%, 2.0%]	0.40
	Senegal	SN	110			[0.0%, 4.1%]	0.65
	Gambia	GM	462	3	0.65%	[0.13%, 2.0%]	0.27
	Guinea	GN	70	5	7.1%	[2.7%, 16%]	**0.0012**
	Ghana	GH	1,191	19	1.6%	[1.0%, 2.5%]	0.21
	Ivory Coast	CI	43	1	2.3%	[0.0%, 13%]	0.42
	Burkina Faso	BF	11			[0.0%, 30%]	1.00
	Benin	BJ	88			[0.0%, 5.0%]	0.63
	Nigeria	NG	52			[0.0%, 8.2%]	1.00
	Cameroon	CM	127			[0.0%, 3.5%]	0.41
Central Africa (CAF)	Gabon	GA	33	1	3.0%	[0.0%, 17%]	0.34
	DR Congo	CD	186	1	0.54%	[0.0%, 3.3%]	0.73
East Africa (EAF)	Sudan	SD	24			[0.0%, 16%]	1.00
	Ethiopia	ET	19			[0.0%, 20%]	1.00
	Kenya	KE	356	1	0.28%	[0.0%, 1.7%]	0.12
	Uganda	UG	5	1	20%	[2.0%, 64%]	0.061
	Tanzania	TZ	290	4	1.4%	[0.36%, 3.6%]	0.78
	Malawi	MW	100	5	5.0%	[1.9%, 11%]	**0.0044**
	Mozambique	MZ	15			[0.0%, 24%]	1.00
	Madagascar	MG	18	1	5.6%	[0.0%, 28%]	0.20
Total			**3,783**	**47**	**1.2%**	**[0.8%, 1.4%]**	

FIGURE LEGENDS

Figure 1 – Principal Coordinate Analysis (PCoA) of African samples, revealing population structure.

In this figure, we show plots of the second principal component against the first (PC2 vs PC1). Along PC1 (explaining 1.9% of variance), samples separate geographically, so that macroregions EAF, CAF and WAF can be distinguished as labelled. A cluster of AF1 parasites, from different countries, separates along PC2 (0.9% of variance). Two horizontal dotted lines indicate the thresholds for defining the AF1 population. Samples with PC2 > 0.025 were classified as AF1; those with PC2 < 0.01 as non-AF1; the remaining parasites were disregarded in further analysis, since their AF1 membership status is uncertain.

Figure 2 – Geographic distribution of AF1 parasite samples.

In the map above, countries from which parasites were sampled are shown with a coloured background and a label showing the country name. Countries where AF1 parasites were found are shown with an orange background; for each of these countries, the number of AF1 samples and the total number of analyzed samples are separated by a slash, and the percentage of AF1 samples is shown in brackets. The map uses data from Natural Earth (https://www.naturalearthdata.com/)

Figure 3 – AF1 characteristic loci.

A – Linkage disequilibrium between AF1 characteristic loci. The circular plot maps all 14 nuclear chromosomes (starting from top, clockwise, each chromosome is represented by a coloured segment in the outer ring). The ring with a green background shows a plot of mean F_{ST} between AF1 and the three African macroregions (WAF, CAF and EAF) at non-synonymous coding SNPs with F_{ST} ≥0.01, with orange markers showing F_{ST} ≥0.5 and larger red markers indicating F_{ST} ≥0.75 (see Supplementary Table 2). Inside this ring, red strips delimit high-IBD regions, in which ≥50% of all AF1 sample pairs are in IBD (see Supplementary Figure 4). Inner blue lines show the r^2 measure of linkage disequilibrium (LD) between pairs of high-F_{ST} SNPs (F_{ST} >0.2), estimated using all African parasites. Three types of line represent three LD ranges: 0.2 ≤ r^2 <0.4 (lightest blue, thinnest lines), 0.4 ≤ r^2 <0.5, and r^2 ≥0.5 (deepest blue, thickest line).

B – Presence of characteristic haplotypes in AF1 parasites. This panel shows a matrix of genotypes at the 23 SNPs with the highest F_{ST} in the 23 high-IBD regions identified in the AF1 population. Each row represents an AF1 sample; the sample ID and the country of provenance are shown. Blue cells indicate that the sample carried the reference allele, while orange cells indicate the characteristic AF1 allele (non-reference). White cells indicate a missing genotype.

C – Genes at AF1 characteristic loci. This panel shows maps of gene positions for the 10 highest-

ranked high-IBD regions identified in the AF1 population. The x-axis represents positions on the high-IBD region's chromosome, with a pink line showing the extent of the region. Each gene in the region is shown by a rectangle, labelled with the gene's name and coloured according to its function (where they are known). Non-synonymous coding SNPs with mean F_{ST}≥0.5 and shown by blue diamond markers below the gene boxes; the highest-F_{ST} SNP in each region (see Supplementary Table 3) is denoted by n orange marker.

Figure 4 – DBLMSP gene sequence crossover in AF1 parasites.

A – Schematic of the gene conversion underpinning the AF1 variant of DBLMSP. The diagram shows as colour blocks the sequences of DBLMSP and DBLMSP2 in four Pf genomes: Pf3D7 (reference), PfIT, PfKH02 (both long-read sequenced) and AF1 (*de novo* assembly of sample PM0293-C). Blocks of the same colour indicate highly similar (near-identical) sequences. Coordinates shown (not to scale) correspond to the Pf3D7 positions in DBLMSP2 (above) and DBLMSP (below). The AF1 DBLMSP sequence is near-identical to that of PfIT DBLMSP2 at the 5' end, and of PfIT DBLMSP after position 991. The AF1 DBLMSP2 sequence, on the other hand, is near-identical to the DBLMSP2 sequence of PfKH02. The gray region is a 19-nt sequence identical in DBLMSP and DBLMSP2, likely to be the recombination breakpoint.

B – Detail of the AF1 DBLMSP/DBLMSP2 breakpoint region. This panel shows an alignment of the AF1 DBLMSP sequence (middle) against the DBLMSP2 (above) and DBLMSP (below) sequences of PfIT. The 19-nt region of 100% identity is shown in green; to the left, the AF1 sequence is identical to PfIT DBLMSP2, while to the right it is identical to PfIT DBLMSP. Allele differences with respect to the AF1 sequence are highlighted in red. The underlined 62-nt portion of the AF1 sequence was used as search query to confirm the presence of the conversion breakpoint in the AF1 parasites.