Genomic approaches to tackle asymptomatic chronic malaria

Marc-Antoine Guery

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Bioinformatique

École doctorale Sciences Chimiques et Biologiques pour la Santé (CBS2)

Laboratory of Pathogens and Host Immunity (LPHI) - UMR 5294

APPROCHES GÉNOMIQUES POUR ABORDER LA MALARIA CHRONIQUE ASYMPTOMATIQUE

Présentée par Marc-Antoine Guery
Le 5 décembre 2023

Sous la direction d'Antoine Claessens

Devant le jury composé de

Mme Ana Rivero, Directrice de recherche, Université de Montpellier	Présidente
M. Daniel Neafsey, Associate professor, Harvard University	Rapporteur
M. Thomas Otto, Professor, University of Glasgow	Rapporteur
M. Michael Fontaine, Chargé de recherche, Université de Montpellier	Invité
M. Antoine Claessens, Chargé de recherche, Université de Montpellier	Co-Directeur

UNIVERSITÉ DE MONTPELLIER

Remerciements

Tout d'abord je tiens à montrer ma gratitude envers Antoine, mon directeur de thèse qui m'a permis de vivre cette formidable aventure enrichissante. Sa passion, bienveillance, écoute et confiance font partie de ses qualités qui ont été déterminantes dans la grande qualité de mon expérience passée sous sa supervision. Enfin, son importante ambition m'a toujours poussé plus loin dans l'interprétation de cette vaste quantité de données génomiques disponible.

Je tiens à remercier Prince et son expertise dans la culture de parasites, sans qui la moitié de mon travail n'aurait pas été possible. Merci aux autres membres de l'équipe GATAC (Camille, Inayat, Sophia) et ses anciens membres (Lionel, Benoît, Aakanksha), avec une attention particulière pour Balotin et Mathieu pour leurs commentaires avisés sur mon manuscrit.

Merci aux membres actuels et anciens du LPHI et de l'IRD qui m'ont accompagné au long de ma thèse. Je remercie évidement les membres de mon comité de thèse Michaël, Isabelle et David qui m'ont suivi chaque année depuis le début de ma thèse pour leurs remarques et conseils pertinents. Merci également à Ian du Texas Biomedical Research Institute et Will du Sanger Institute pour m'avoir accueilli dans leur institut et permis de gagner une expérience enrichissante dans un nouvel environnement.

À ma famille, maman toujours fière de moi, papa qui veut mon bonheur, Victor mon petit-frère adoré, Charles qui prend régulièrement de mes nouvelles et Rose-Anne qui est toujours ravie de me revoir, pour m'avoir accompagné pendant ma thèse. Merci à mes amis pour leur soutien et les bons moments passés pour se changer les idées.

Finalement, merci à celle que j'aime par dessus tout, Ana, pour avoir illuminé mon quotidien en me faisant découvrir tant de choses aussi bien virtuelles que réelles. Je souhaite que 'Les Hannetons' aient un grand avenir devant eux ! Ana, merci de m'avoir encouragé, motivé et aidé avant et pendant l'écriture de mon manuscrit, notamment pour la traduction en français.

Résumé

En Gambie, la saisonnalité du paludisme se caractérise par environ 5 mois de transmission élevée suivis de 7 mois de transmission faible à nulle, pendant lesquels les cas cliniques sont rares. L'impact des changements constants entre les saisons de transmission forte et faible sur la diversité génétique de *Plasmodium falciparum* est peu étudié. Puisque les cas importés ne peuvent pas à aux seuls expliquer toutes les nouvelles infections à chaque saison de forte transmission, les hôtes humains serviraient donc de réservoir au parasite durant la saison de faible transmission. La longue durée de ces infections chroniques pourrait être provoquée par l'élargissement du répertoire antigénique des parasites grâce à la génération de gènes *var* chimériques pendant l'infection. Nous avons cherché à caractériser la diversité génétique de *Plasmodium falciparum* dans l'est de la Gambie au niveau populationnel et de la cellule individuelle à partir d'une étude longitudinale menée dans quatre villages voisins entre 2014 et 2017.

Des échantillons sanguins de 1505 participants ont été prélevés sur 16 périodes pendant la saison des pluies et la saison sèche. Sur les 436 échantillons positifs pour *Plasmodium falciparum* provenant d'infections asymptomatiques, 89 *single nucleotide polymorphisms* ont été génotypés avec succès, et 334 échantillons ont été séquencés en génome entier. Nous avons utilisé l'*identity by descent* (IBD), une méthode de comparaison permettant d'estimer la parenté entre les isolats génotypés ou séquencés. Les parasites prélevés dans le même foyer étaient significativement plus apparentés génétiquement, en particulier lorsque les prélèvements étaient espacés de moins de trois mois. De plus, les parasites isolés pendant les saisons de faible transmission étaient plus apparentés à la saison de forte transmission précédente qu'à la suivante. Nous avons estimé que la majeure partie de la diversité génétique des parasites se renouvelait après environ un an, avec l'exception notable d'un individu de 9 ans infecté par le même parasite pendant au moins un an et demi.

Pendant la saison sèche 2016/2017, nous avons étudié des infections chroniques asymptomatiques chez 11 individus. Les génomes de *Plasmodium falciparum* provenant de ces infections ont été séquencés en cellule unique *ex vivo*, ou en *bulk* avec des lectures longues après un mois d'adaptation en culture. Les assemblages en lectures longues des clones de parasites ont donné des contigs de grande taille annotés avec la quasi-totalité de gènes attendus chez le parasite, y compris la famille hypervariable des gènes *var*. Le séquençage en cellule unique a été performant, comme l'indique les clusters IBD des génomes des cellules uniques ségréguant efficacement les génotypes de parasites. De plus, nous avons assemblé *de novo* des répertoires de gènes *var* presque complets dans chacune des cellules uniques extraites d'une même infection.

Au final, nous montrons que les saisons de faible transmission agissent comme des réservoirs de parasites issus d'infections asymptomatiques, attendant la saison de transmission élevée pour se propager et se recombiner, renouvelant ainsi la diversité génétique. Nous fournissons aussi une grande quantité de données génomiques sur des infections, y compris de nouvelles séquences de gènes *var* de haute qualité qui peuvent être ajoutées aux bases de données existantes. Nous soutenons que la détection active des cas est essentielle pour comprendre les infections asymptomatiques, le réservoir caché du paludisme.

Abstract

In The Gambia, malaria seasonality is characterized by approximately 5 months of high transmission followed by around 7 months of low to no transmission, during which clinical cases are rare. The impact of constantly shifting from high to low transmission seasons on the genetic diversity of *Plasmodium falciparum* is understudied. Given that imported cases alone are unlikely to account for all new infections during each high transmission season, it appears that human hosts serve as a reservoir for the parasite throughout the low transmission season. The long duration of these chronic infections could be caused by the parasites enhancing their antigenic repertoire through the generation of chimeric *var* genes during the infection. Here, we aimed to characterise the *Plasmodium falciparum* genetic diversity from the single-cell to the population level in eastern Gambia from a longitudinal study conducted in four nearby villages from 2014 to 2017.

Blood samples from 1505 participants over 16 time points were collected during both wet and dry seasons. Out of the 436 *Plasmodium falciparum* positive samples from asymptomatic infections, 89 single nucleotide polymorphisms were successfully genotyped, and 334 samples underwent whole genome sequencing. We used identity by descent (IBD), a pairwise comparison method to estimate the relationships between genotyped or sequenced isolates. Parasite samples collected within the same household were significantly more genetically related especially when distant by three months at most. Also, parasites isolated during the low transmission seasons were more related to the previous high transmission season than to the following one. We could estimate that most of the parasite genetic diversity was renewed after approximately one year. An interesting exception is a 9-year old individual who had been infected with the same parasite for at least one year and a half.

Additionally, during the dry season of 2016/2017, we closely monitored chronic asymptomatic infections in 11 individuals. Genomes of *Plasmodium falciparum* derived from these infections were either single-cell sequenced *ex vivo* or long-read bulk sequenced after one month of culture adaptation. Long-read assemblies of parasite clones yielded chromosomal-long contigs annotated with almost the entire set of genes expected in *Plasmodium falciparum*, including the hypervariable family of *var* genes. The single-cell sequencing performed well, as indicated by the IBD clusters of single-cell genomes effectively segregating parasite genotypes. Finally, we were able to successfully *de novo* assemble the almost complete *var* repertoire of each single-cell from the same infection.

Altogether, our findings demonstrate that low transmission seasons act as reservoirs of parasites from asymptomatic infections, lying in wait for the high transmission season to spread and recombine, thus renewing the genetic diversity at every high transmission season. Moreover, we provide a wealth of quantity of genomic data on individual infections, including new high-quality *var* gene sequences that can enrich existing databases. We argue that active case detection is key to understanding asymptomatic chronic infections, the hidden reservoir of malaria.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

3'-UTR three prime untranslated region

ACD active case detection

ACT artemisinin-based combination therapies

AMA1 apical membrane antigen 1

ATS acidic terminal sequence

BPS base pair substitution

bp base pairs

CCS circular consensus sequencing

CHMI controlled human malaria infections

CIDR cysteine-rich interdomain region

COI complexity of infection

CSA chondroitin sulfate A

CSP circumsporozoite protein

DBL Duffy binding-like

EM electron microscope

EPCR endothelial protein C receptor

HiFi high fidelity

HRP2 histidine-rich protein 2

HRP3 histidine-rich protein 3

IBD identity by descent

ICAM-1 intercellular adhesion molecule 1

iHS integrated haplotype score

INDEL insertion/deletion

iRBC infected red blood cell

ITN insecticide-treated mosquito net

KAHRP knob-associated histidine-rich protein

kb thousand of base pairs

LILRB1 leucocyte immunoglobulin-like receptor B1

Mb million of base pairs

mRNA messenger RNA

MSP2 merozoite surface protein-2

NTS N-terminal segment

PacBio Pacific Biosciences

PAM pregnancy-associated malaria

PCD passive case detection

PCR polymerase chain reaction

pfAAT1 *Plasmodium falciparum* amino acid transporter 1

pfCRT *Plasmodium falciparum* chloroquine resistance transporter

pfDHFR *Plasmodium falciparum* dihydrofolate reductase

pfDHPS *Plasmodium falciparum* dihydropteroate synthase

PfEMP1 *Plasmodium falciparum* erythrocyte membrane protein 1

PfGDV1 *Plasmodium falciparum* gametocyte development protein 1

pfK13 *Plasmodium falciparum* Kelch 13

pfMDR1 *Plasmodium falciparum* multidrug resistance protein 1

PfMSH2-2 *Plasmodium falciparum* MutS homologue 2-2

PfPDE beta *Plasmodium falciparum* 3,5-cyclic nucleotide phosphodiesterase beta

PfPKG *Plasmodium falciparum* cGMP-dependant protein kinase

pLDH *Plasmodium* lactate dehydrogenase

qPCR quantitative real-time PCR

qRT-PCR quantitative reverse-transcription-PCR

RBC red blood cell

RDT rapid diagnostic test

RIFIN repetitive interspersed families of polypeptides

rRNA ribosomal RNA

SMRT single molecule real-time

SNP single nucleotide polymorphism

SP Sulfadoxine/Pyrimethamine

STEVOR subtelomeric variable open reading frame

SURFIN surface-associated interspersed protein

sWGA selective whole genome amplification

TARE telomere associated repetitive element

TRF Tandem Repeat Finder

ups upstream

VSA variant surface antigen

WGS whole genome sequencing

Résumé Substantiel

Introduction

Le paludisme est une maladie qui a marqué les civilisations humaines les plus anciennes, décrivant déjà des symptômes de fièvre et de splénomégalie [1]. En 1880, Charles Louis Alphonse Laveran observa le parasite à l'origine du paludisme dans des globules rouges, résolvant ainsi le cycle asexuel du parasite qui sera connu plus tard sous le nom de *Plasmodium* [1, 2, 3]. Plus tard, Ronald Ross et Giovanni Battista Grassi montreront que ce sont les moustiques qui permettent la transmission du parasite d'un hôte à un autre [1, 4, 5]. Le cycle de vie du parasite fut complètement décrit en 1949 grâce à Henry Shortt et Cyril Garnham qui montrèrent que le parasite séjourne dans le foie pendant plusieurs semaines après la piqûre de moustique [1].

Il y a au total cinq espèces de *Plasmodium* qui infectent spécifiquement les humains : *Plasmodium falciparum*, *Plasmodium malariae*, *Plasmodium ovale curtisi*, *Plasmodium ovale wallikeri* et *Plasmodium vivax*. De plus, une sixième espèce, *Plasmodium knowlesi*, dont l'hôte naturel est le macaque, est responsable de la vaste majorité des infections zoonotiques chez l'humain. *Plasmodium falciparum*, le seul parasite de l'humain du sous-genre *Laverania* (les autres appartenant au sous genre *Plasmodium*), serait apparu en Afrique à partir d'un transfert depuis une espèce de parasite n'infectant pas l'humain, *Plasmodium praefalciparum*, il y a entre 40 000 et 365 000 ans [6, 7].

Les deux parasites du paludisme les plus virulents sont *Plasmodium falciparum*, principalement présent en Afrique et responsable de la majorité des morts dans le monde, et *Plasmodium vivax* qui est localisé dans la plupart des régions endémiques du paludisme, sauf en Afrique [8]. Le cycle érythrocytaire du parasite *Plasmodium falciparum* dure 48 heures entre l'invasion initiale des globules rouges par les mérozoïtes et la libération des nouveaux mérozoïtes ou alternativement gamétocytes dans le sang. Le cycle démarre avec le globule rouge infecté au stade anneau jusqu'à 20 heures post-invasion, puis il poursuit avec le stade trophozoïte jusqu'à 36 heures post-invasion pour finir avec le stade schizonte où les globules rouges infectés libèrent entre 20 et 30 mérozoïtes [9].

Les tests les plus utilisés permettant de diagnostiquer un cas de paludisme comprennent la microscopie où les érythrocytes infectés sont visualisés à l'aide d'un pigment, les RDT (*rapid diagnostic tests*) qui sont basés sur la détection d'antigènes spécifiques du parasite ou encore les tests PCR (*polymerase chain reaction*) qui amplifient des gènes du parasite. Les tests PCR sont les plus sensibles, permettant la détection et la quantification de parasites dans des infections avec une très faible parasitémie, mais ils sont plus coûteux que les RDT ou la microscopie [10, 11].

De nombreuses actions sont menées par les pays dans lesquels le paludisme est endémique pour limiter l'impact de cette maladie affectant 270 millions de personnes et en tuant 619 000 dans le monde en 2021 [12]. Le principal traitement utilisé contre le parasite correspond aujourd'hui aux ACT (*artemisinin-based combination therapies*) qui contient plusieurs anti-paludiques aux mécanismes d'action différents. Les moustiques, étant les vecteurs du paludisme, sont également la cible des campagnes d'élimination à travers l'utilisation de moustiquaires traitées à l'insecticide ou encore la vaporisation systématique d'insecticide en intérieur. L'incidence et la mortalité due au paludisme ont baissé depuis les 20 dernières années grâce à ces stratégies, cependant, la résistance des moustiques aux insecticides et des parasites aux anti-paludiques a entraîné le ralentissement de l'élimination du paludisme [13, 14, 15, 16, 17]. De plus, la pandémie de COVID-19, plus particulièrement à son apogée en 2020 et 2021, a provoqué une perturbation majeure des services dédiés au contrôle du paludisme, ce qui a ramené à des niveaux d'incidence et de mortalité identiques à ceux de 2015 [12, 18].

Les infections causées par *Plasmodium falciparum* peuvent être sévères avec un risque de mortalité élevée, non compliquées lorsque les symptômes sont moins graves ou encore asymptomatiques. Ces dernières présentent généralement une faible parasitémie détectable par PCR seulement et affectent en particulier les individus âgés de plus de 5 ans [19, 20, 21, 22, 23, 24]. Ces infections cachées où les individus ne présentent aucun symptôme peuvent durer pendant plus d'un an [25, 22].

De nombreux anti-paludéens ont été utilisés à travers l'histoire dans une constante course avec la capacité du parasite à générer des phénotypes résistants et les répandre dans le monde. Le premier composé à être massivement utilisé était la quinine, extraite dès 1820 d'un arbre originaire d'Amérique du Sud dont l'écorce était connue pour guérir des fièvres que l'on attribue aujourd'hui au paludisme [26, 27]. Par la suite, la chloroquine remplaça la quinine mais peu de temps après, la première résistance à cette molécule émergea en Asie du Sud-est et en Amérique du Sud dans les années 50 ou 60 pour être finalement présente partout dans le monde en 1980 [28, 13]. Une thérapie combinant la sulfadoxine et pyrimethamine remplaça la chloroquine mais encore une fois, des parasites résistants émergèrent dans les années 70 en Asie du Sud-est, la même région où la résistance à la chloroquine émergea [29, 13]. De manière similaire à la quinine, l'artémisinine a été extraite en 1970 d'une plante utilisée en médecine chinoise pour traiter des fièvres [30, 27]. Rapidement, l'artémisinine a été utilisée en combinaison avec un composé partenaire ayant un mode d'action différent dans les thérapies ACT [31]. Cette molécule n'a pas fait exception à la résistance de *Plasmodium falciparum* qui émergea en Asie du Sud-est dans les années

2000 mais qui n'est pas répandue dans le monde pour le moment [32, 13, 33, 34]. Les ACT étant la principale thérapie utilisée pour traiter le paludisme non sévère, la présence globale de la résistance à l'artémisinine aurait des conséquences graves sur les cas de paludisme dans le monde.

Plasmodium falciparum peut être cultivé *in vitro* à partir de souches adaptées au laboratoire comme 3D7, la plus étudiée. La souche 3D7 est dérivée de NF54, un parasite qui était présent dans un patient des Pays-Bas mais qui est probablement originaire d'Afrique comme le suggère la comparaison avec des isolats cliniques du monde entier [35, 36, 37]. D'autres souches de laboratoire, comme Dd2 ou HB3, ont pu être obtenues de différentes parties du monde et sont couramment utilisées puisqu'elles sont plus facilement cultivables que les souches de terrain. Les 23 millions de bases du génome haploïde de *Plasmodium falciparum* ont été totalement séquencées et assemblées à partir de la souche 3D7 en 2002, qui est devenue le génome de référence [38]. Le parasite possède 14 chromosomes nucléaires constituant l'un des génomes les plus riches en taux de AT parmi toutes les espèces connues, avec un pourcentage de AT d'environ 81 %, atteignant 84 % dans les régions non codantes. Il y a également une mitochondrie et un apicoplaste par parasite avec respectivement 20 copies du génome mitochondrial et 15 copies du génome apicoplastique [39, 40]. Les 16 différents chromosomes du parasite contiennent près de 5600 gènes, dont 5300 codants des protéines. Autour de 10 % du génome correspond à des régions hyper variables et des répétitions subtélomériques qui sont très différentes entre les souches, ce qui rend leur comparaison très difficile [41]. Ces régions de faible complexité ne peuvent pas être assemblées facilement avec le séquençage de deuxième génération car les très longues répétitions ne peuvent pas être couvertes par des courtes lectures de séquençage de 150 bases seulement. En 2018, grâce au séquençage de troisième génération (ou séquençage en lectures longues), les assemblages des régions de faible complexité des chromosomes de 3D7 ont été fortement améliorées [42].

La famille de gènes *var* est la plus polymorphique du génome de *Plasmodium falciparum*. Ces gènes *var* sont transcrits à partir du stade anneau des globules rouges infectés et les protéines qu'ils encodent, *Plasmodium falciparum erythrocyte membrane protein 1* (PfEMP1), sont exprimées à la surface des globules rouges infectés aux stades trophozoïte et schizonte [43]. L'une des fonctions de PfEMP1 est de se lier à la surface des cellules endothéliales, un processus connu sous le nom de séquestration [44, 45]. Les quelques 60 gènes *var* (leur nombre varie entre les génomes de parasite) sont organisés de manière similaire avec deux exons séparés par un intron et sont localisés dans les régions subtélomeriques ou internes des chromosomes au sein des régions hyper variables [41, 46]. De manière similaire, PfEMP1, d'une taille de 1000 à 4000 acides aminés dans 3D7, ont tous la même configuration avec

plusieurs domaines extracellulaires comprenant un segment N-terminal (NTS) suivi d'une succession de domaines variables DBL (*Duffy binding-like*) et CIDR (*cysteine-rich interdomain region*) encodés par l'exon 1 et d'un domaine intracellulaire plus conservé appelé ATS (*acidic terminal segment*) encodé par l'exon 2 [47]. De manière intéressante, au cours du même cycle érythrocytaire, seul un gène *var* est transcrit, ce qui résulte en un seul type de PfEMP1 à la surface du globule rouge infecté.

En se basant sur la similarité de séquence de la région génomique située en amont des gènes (ups), trois principaux groupes de gènes *var* (upsA, upsB, upsC) et un groupe mineur ne comprenant qu'un gène (upsE) ont été définis à partir de 3D7 [38]. Les gènes *var* des groupes A et B (sauf un) sont localisés dans les régions subtélomériques tandis que les gènes *var* du groupe C sont localisés dans les régions internes des chromosomes. Le côté 5 prime de tous les gènes *var* excepté *var2csa* (PF3D7_1200600) contient un domaine DBLα [49]. Dans ce domaine, le 'DBLα tag' est composé d'approximativement 130 acides aminés et contient des séquences variables entourées par des motifs conservés LARSFADIG et DYVPQ[YF]LRW dont les séquences ADN correspondantes peuvent être utilisées pour une amplification PCR [50].

À cause de leur localisation particulière dans des régions répétées et de leur grande variabilité entre les génomes parasitaires, les gènes *var* d'isolats de terrain ne peuvent pas être accessibles en utilisant un classique mapping de lectures courtes sur un unique répertoire de référence. Les séquences de gènes *var* sont plus facilement retrouvées en utilisant des pipelines d'assemblage *de novo*, ce qui nécessite néanmoins des étapes en amont et souvent mènent à des répertoires incomplets [51, 46]. La façon la plus fiable d'obtenir des répertoires de gènes *var* complets est d'utiliser des technologies de séquençage de troisième génération tel que Pacific Biosciences (PacBio) SMRT (*single molecule real-time*) qui produit des lectures d'une longueur supérieure à 12 000 bases, couvrant les gènes *var* en leur totalité [52, 42].

La diversité des gènes *var* peut être générée par des recombinaisons chromosomiques entre différents gènes *var* (plus précisément entre différents exons 1) dans un processus appelé recombinaison ectopique, qui est fréquent pendant la méiose dans le moustique [53, 50]. Ces recombinaisons ectopiques peuvent avoir lieu lors des mitoses du cycle érythrocytaire, générant ainsi des gènes *var* chimères à un taux de 2×10^{-3} par parasite à chaque cycle érythrocytaire [54].

Il y a d'autres familles de gènes qui encodent des protéines qui co-localisent avec PfEMP1 à la surface des globules rouges infectés : RIFIN (*repetitive interspersed families of polypeptides*) et

STEVOR (*subtelomeric variable open reading frame*) encodés respectivement par les gènes *rif* et *stevor* [55]. PfEMP1, RIFIN et STEVOR sont les principaux antigènes de surface de *Plasmodium falciparum*, tous liant divers récepteurs de l'hôte. Les 150 à 200 gènes *rif* et les 28 gènes *stevor* sont organisés en cluster proche des gènes *var*, ce qui suggère qu'ils subissent probablement aussi de fréquentes recombinaisons au sein de ces régions hyper variables [56, 57].

La complexité d'infection représente le nombre de lignées de parasites qui sont distinctes dans un même hôte. Il existe deux voies menant à plusieurs génotypes de parasites infectant un même hôte simultanément. Cela peut passer par des piqûres successives de moustiques qui chacune inocule un parasite distinct génétiquement (appelé une super infection), une seule piqûre de moustique portant différents sporozoïtes générés par la recombinaison de deux gamétocytes (appelé une co-transmission), ou bien les deux voies à la fois [58]. Dans les régions connaissant une forte prévalence de paludisme, la co-tranmission devient très commune alors que la super infection devient de plus en plus rare [59]. Quand la prévalence du paludisme augmente dans une région, la proportion d'infections polyclonales augmente aussi à cause d'une plus grande chance que le moustique porte un ou plusieurs parasites à la fois [60, 61]. Le changement de valeur moyenne de complexité d'infection ou de proportion d'infections multiples au cours du temps est l'un des meilleurs estimateurs (basé sur la diversité génétique) du changement de la prévalence locale du paludisme [62, 63]. Cependant, les valeurs absolues de complexité d'infection ne peuvent pas être utilisées pour estimer directement la prévalence du paludisme ou pour comparer des populations de parasites de deux régions distinctes [64, 65].

L'identification et la comparaison de souches de *Plasmodium falciparum* entre les infections et au sein d'une même infection est nécessaire pour comprendre son histoire évolutive ou la dynamique de transmission en cours dans une région. Il y a plusieurs catégories de marqueurs génétiques capables d'identifier les souches de *Plasmodium falciparum* avec différents niveaux de précision. Puisque beaucoup d'allèles MSP2 (*merozoite surface protein-2* ; PF3D7_0206800) sont différents en terme de taille, le génotypage des souches peut être accompli en utilisant les tailles des produits de PCR sans avoir besoin d'obtenir la séquence entière [66, 67]. De plus, des hauts niveaux de complexité d'infection peuvent être estimés par amplification PCR de MSP2 en comptant le nombre de produits de PCR avec une taille différente au sein du même isolat [68, 24]. Comme décrit plus haut, la région hautement polymorphique des DBLα tag peut être utilisée comme un code-barre génétique qui peut distinguer différentes populations aux niveaux local et mondial [69, 70]. Les parasites peuvent être identifiés par le génotypage de SNP (*single nucleotide polymorphisms*) bi-alléliques répartis dans le génome qui sont choisis pour leur capacité à décrire une grande diversité génétique. Aussi peu que 24 SNP est déjà

suffisant pour obtenir une estimation de la diversité génétique qui corrèle bien avec l'intensité de transmission ou la distance géographique [71, 60, 72, 73]. Les marqueurs micro-satellites, qui correspondent à des courtes répétitions en tandem situées dans de nombreuses régions du génome, sont aussi utilisés communément pour estimer la diversité des parasites ou même la complexité d'infection [74, 75].

Bien que le génotypage de locus est très efficace, il ne peut pas capturer totalement la diversité des parasites, plus particulièrement dans des régions avec un niveau de transmission intermédiaire ou élevé où les très hauts niveaux de complexité d'infection peuvent biaiser la similarité calculée entre les infections [76]. Dans ces cas, le séquençage en génome entier peut être utilisé pour accéder à plus de locus génétiques dans chaque parasite, impliquant des dizaines de milliers de SNP [77, 78]. Si deux parasites partagent un même ancêtre commun, ils posséderont des fragments génétiques avec des SNP identiques, dont la taille dépendra du nombre de générations les précédant selon le taux de recombinaison de *Plasmodium falciparum* de 13.5 kb/cM [41]. Ces paramètres sont utilisés par hmmIBD, un modèle de Markov caché qui recherche des fragments en identité par descendance (*identity by descent* abrégé IBD) partagés entre deux génomes avec un minimum de 200 marqueurs bi-alléliques basés sur leur plus probable historique de recombinaison [79, 80]. En conséquence, chaque paire de génome aura une valeur associée de parenté qui indiquera le pourcentage global d'identité le long des fragments chromosomiques qui sont soit en IBD ou non [60]. Mesurer la complexité d'infection avec un séquençage en *bulk* peut être possible avec les mesures R_{H} et F_{WS}, qui utilisent les fréquences alléliques des locus génomiques polymorphiques [81, 82]. Pour obtenir la séquence précise des souches individuelles de parasites dans une infection complexe, le séquençage en cellule unique est un outil nécessaire pour assurer que des parasites co-transmis avec une forte parenté sont correctement ségrégés [59].

Une fois que les mérozoïtes de *Plasmodium falciparum* entrent dans le sang de leur hôte humain, ils envahissent graduellement les globules rouges. L'hôte possède différentes stratégies pour se débarrasser du parasite, en commençant par les cellules immunitaires mais aussi via la rate. En effet, les globules rouges infectés perdent leur déformabilité d'autant plus après le stade anneau, ce qui les rend sujets à être détruits par la rate qui supprime les globules rouges qui sont trop rigides [83]. Pour éviter d'être immédiatement éliminés du sang, les parasites ont développé la capacité de lier les cellules hôtes via des protéines exportées pour qu'ils puissent compléter leur cycle sans être gênés par la rate ou les cellules immunitaires [84]. Cette adhérence a été associée principalement avec PfEMP1 mais aussi avec d'autres antigènes de surfaces comme les RIFIN ou STEVOR. Quelques RIFIN sont

capables de lier LILRB1 (*leucocyte immunoglobulin-like receptor B1*), un récepteur de surface de plusieurs cellules immunitaires, et réduire la production d'anticorps [85].

L'expression mutuellement exclusive des gènes *var* permet au parasite de développer une variation antigénique au cours du temps pour échapper au système immunitaire adaptatif. Les mécanismes qui gouvernent l'expression d'un seul gène *var* et la répression de tous les autres impliquent les différents états de condensation de la chromatine à la périphérie du noyau et la modification d'histones en amont des gènes *var* [86, 87]. Le taux de changement de gène *var* exprimé d'approximativement 2 % diffère entre les souches de *Plasmodium falciparum* et entre les gènes *var* en fonction de leur localisation chromosomique, avec les internes plus susceptibles d'être sélectionnés par rapport aux subtélomériques [88, 89, 90].

L'incidence du paludisme dû à *falciparum* est très hétérogène à travers le monde, avec le contient africain comprenant le plus de cas [12]. Cela reste vrai au sein des pays comme la Gambie, le Sénégal ou le Laos, où l'incidence du paludisme suit généralement un gradient d'intensité de transmission [91, 92, 93]. À une échelle régionale, au Ghana ou en Tanzanie par exemple, la prévalence du paludisme peut être hétérogène avec des zones de forte transmission séparées par quelques dizaines de kilomètres de régions de faible transmission [94, 95]. Plusieurs parasites de régions faiblement endémiques ont montré des hauts niveaux de similarité par rapport à ceux de régions plus endémiques, comme entre la Gambie de l'Ouest et la Gambie de l'Est, et cela peut expliquer comment le paludisme est maintenu dans ces régions [73, 96, 97]. En général, le niveau de similarité entre les parasites du paludisme tend à décroître avec la distance spatiale et temporelle qui sépare leur échantillonnage, et cette relation semble aller jusqu'au niveau du foyer [73, 98, 99, 100]. Cette connectivité de la population de parasites peut être synthétisée par une propagation clonale avec un niveau de similarité descendant graduellement [101].

La transmission de *Plasmodium falciparum* est modulée par les conditions climatiques, principalement parce que son vecteur, le moustique anophèle, est adapté à seulement certaines températures et taux d'humidité relative, et requiert la présence de poches d'eau pour pondre leurs œufs. La transmission du paludisme est corrélée positivement avec le taux d'humidité relative de plus de 50 à 60 % et avec l'intensité des pluies [102, 103, 104, 105, 106]. Dans certaines régions endémiques du paludisme, les conditions climatiques permettant la transmission du paludismes sont réunies tout au long de l'année. Cependant, d'autres régions sont touchées par un paludisme saisonnier caractérisé par une forte transmission durant la saison des pluies et au commencement de la saison sèche, mais une transmission très faible à nulle le reste de l'année. Dans la Gambie, où le paludisme est plus prévalent dans l'Est du pays, la saison humide commence généralement en juin et finit en septembre, ce qui se

traduit par une transmission du paludisme élevée entre août et décembre [107, 108].

À ce jour, on ignore comment les parasites parviennent à maintenir leur présence dans les régions endémiques caractérisées par de longues périodes de faible transmission. En effet, pendant la saison de faible transmission, qui s'étend sur la majorité de l'année dans certains pays comme la Gambie (environ 7 mois), les cas symptomatiques de paludisme sont pratiquement inexistants. À l'inverse, le pourcentage de porteurs asymptomatiques reste souvent constant tout au long de l'année, quelle que soit la saison de transmission [109, 110, 111, 22]. Les parasites échantillonnés régulièrement entre les différentes saisons de forte transmission sont génétiquement proches et montrent des niveaux de clonalité similaires, ce qui implique qu'il y a une continuité génétique des parasites pouvant traverser les longues saisons de faible transmission [112, 113, 100, 114]. Puisque les infections asymptomatiques peuvent présenter de hauts niveaux de densité de gamétocytes qui ont le potentiel d'infecter les moustiques, il est probable qu'elles représentent le principal réservoir de parasites responsables des cas de paludisme symptomatique pendant la saison de forte transmission, tandis que les parasites importés d'autres régions endémiques y contribuent dans une moindre mesure [22, 115]. De plus, les infections asymptomatiques sont souvent polyclonales avec de nombreux parasites distincts circulant dans le sang, ce qui aide à maintenir une forte diversité génétique de la population de parasites [24, 109, 116].

Toutes ces preuves confirment que les infections asymptomatiques, bien qu'elles ne soient pas nocives pour leur hôte, contribuent fortement aux conséquences du paludisme, ce qui nécessite de leur accorder une attention particulière dans les programmes d'élimination du paludisme. De plus, il pourrait être possible de prédire la durée de l'infection asymptomatique, ainsi que son potentiel de transmission, à partir de la densité parasitaire à la fin de la saison de forte transmission, puisqu'elle est positivement corrélée à la durée de l'infection pendant la saison sèche [24]. Deux hypothèses principales ont été avancées pour expliquer comment les infections chroniques de *Plasmodium falciparum* persistent dans l'hôte durant toute la saison sèche. Premièrement, Andrade *et al.* suggèrent que les parasites des saisons de faible transmission présentent une adhésion vasculaire moins importante que les parasites des saisons de forte transmission, les rendant ainsi plus susceptibles d'être détruits par la rate [117]. Une autre possibilité est que les parasites des saisons de faible transmission ont un taux de multiplication dans le sang si bas qu'ils restent quasiment invisible pour le système immunitaire de l'hôte [118]. Une question qui n'a pas encore été étudiée est l'importance de la génération des gènes *var* chimériques au cours d'une infection dans le maintien des parasites dans le sang pendant plus longtemps, en augmentant la taille du répertoire de gènes *var*, et ainsi améliorant leur capacité d'évasion immunitaire.

Matériels et méthodes

À partir de décembre 2014, nous avons recruté tous les habitants de deux villages (Madina Samako et Njayel), avec deux villages supplémentaires (Sendebu et Karandaba) recrutés à partir de juillet 2016, les villages étant dans la 'Upper River Region' de la Gambie et situés à 5 kilomètres les uns des autres. Plus d'information à propos des participants recrutés peut être trouvée dans une étude précédente [119]. En décembre 2016, une cohorte de 42 porteurs asymptomatiques du paludisme a été échantillonnée pendant 6 mois, comme reporté précédemment [24].

La description complète de l'extraction d'ADN des parasites est fournie dans un travail précédent [119]. Un total de 101 SNP localisés sur les 14 chromosomes ont été génotypés et fusionnés dans un 'code-barre moléculaire'. Les SNP sont tous bi-alléliques et ont été choisis pour leur utilité dans les analyses de relation génétique entre les parasites. Parmi les 442 échantillons génotypés, 334 ont aussi été séquencés en génome entier (Illumina) avec une étape d'amplification sélective [120]. Les paires de lectures d'ADN (150 bases) ont été mappées sur le génome de référence 3D7 version 3 et les variants ont été obtenus à partir d'un script du consortium MalariaGen utilisant GATK HaplotypeCaller [121, 122]. Six marqueurs de résistance à multiples anti-paludéens (amodiaquine, artémisinine, chloroquine, luméfantrine, méfloquine, pyriméthamine, sulfadoxine) ont été obtenus à partir du génotypage et du séquençage [123]. Ces marqueurs correspondent à six changements non synonymes dans six protéines encodées par les gènes : *aat1* (PF3D7_0629500) S528L, *crt* (PF3D7_0709000) K76T, *dhfr* (PF3D7_0417200) S108N, *dhps* (PF3D7_0810800) A437G, *kelch13* (PF3D7_1343700) C580Y et *mdr1* (PF3D7_0523000) N86Y, dont chacun est connu pour réduire la susceptibilité du parasite à multiples anti-paludéens.

Pour estimer de manière précise la similarité des parasites de différentes infections échantillonnées, nous avons estimé la valeur d'IBD entre chaque paire de génomes ou de code-barres en utilisant hmmIBD [79]. La probabilité que deux échantillons soient en IBD représente la fraction partagée de leur génome. Une IBD de 0.9 est considérée comme identique, décrivant ainsi le même génotype de parasite.

La clonalité de chaque isolat a été estimée à partir du séquençage en génome entier par la mesure de F_{WS}, basé sur les fréquences alléliques [82]. Additionnellement, la complexité d'infection à été estimée par la méthode *categorical* du programme THE REAL McCOIL [124].

Nous avons développé un pipeline permettant d'améliorer les code-barres génotypés à l'aide des séquençages en génome entier obtenus sur les mêmes échantillons. Ces

'code-barres consensus' sont par la suite utilisés par le pipeline pour déterminer les relations génétiques spatio-temporelles entre les code-barres.

Les code-barres ont été groupés par leur date de récolte et la localisation de l'échantillonnage. Pour chaque paire de groupes de code-barres, la parenté génétique a été estimée par la proportion de code-barres similaires (IBD > 0.5) par rapport à toutes les combinaisons possibles de code-barres en excluant ceux échantillonnés dans le même individu.

Les fragments génomiques en IBD entre les 19700 paires de génomes non-identiques (IBD < 0.5) ont été combinés pour obtenir, pour chaque position, le nombre de génome en IBD. Les régions génomiques dans le top 5 % des plus couvertes ont été extraites et fusionnées lorsqu'elles étaient espacées de moins de 10000 bases.

Les génomes hautement similaires (IBD > 0.9) et ceux faiblement similaires (IBD entre 0.35 et 0.65) ont été groupés dans des clusters en utilisant le package R igraph (version 1.2.11) [125, 126]. Dans chaque cluster, trois génomes ont été associés en 'triade' quand l'un d'eux, supposé être la progéniture, était relié à deux autres génomes, supposés être les parents (IBD entre parents < 0.2).

Parmi les 42 porteurs d'infection asymptomatique recrutés en décembre 2016, 11 ont été sélectionnés pour un suivi précis de leur infection à l'aide de séquençages en lecture longue et en cellule unique sur différents échantillons sanguins. Prince Nyarko (doctorant dans l'équipe d'Antoine Claessens) était en charge de toutes les étapes nécessaires avant l'obtention des données de séquençages, ses compétences en culture de parasite ont été particulièrement utiles sur ces infections à très faible parasitémie et à clonalité parfois élevée. Pour chacun des 11 individus, au moins un échantillon sanguin a été sélectionné dans le but d'adapter ses parasites à la culture. Quand l'adaptation à la culture a été un succès, les parasites ont été clonés puis séquencés en lecture longue grâce aux lectures PacBio de plus de 10000 bases. En parallèle, deux échantillons les plus espacés en temps d'infection ont été choisis pour un triage de leurs globules rouges infectés au stade schizonte, soit approximativement 40 heures après décongélation de l'échantillon sanguin. Les schizontes ont été séquencés individuellement en utilisant les lectures Illumina de 150 bases.

Les données de séquençages en lecture longue ont été utilisées pour construire un assemblage par clone grâce aux pipelines de l'équipe Tree of Life du Sanger Institute qui utilisent hicanu et hifiasm [127, 128]. Les contigs ont été associés avec les chromosomes de 3D7 en utilisant ragtag (version 2.1.0) [129]. Les gènes ont été prédits avec Augustus (version 3.3.3) [130]. Les gènes *var* ont été extraits à partir de la liste complète des gènes en

recherchant les motifs LARSFADIG et DYVPQYLRW en autorisant deux acides aminés différents. Les domaines des PfEMP1 ont été identifiés avec VarDom (version 1.0) [47].

Pour l'individu DC05, aucun génome de référence n'a pu être généré car les parasites en culture n'ont pas survécu *in vitro*. Les génomes obtenus avec les séquençages en cellules multiples et cellule unique des deux échantillons ont été utilisés pour générer les répertoires de gènes *var* en utilisant un pipeline d'assemblage *de novo* de lecture courte basé sur les méthodes développées par Otto *et al.* (2019) et Andradi-Brown *et al.* (2023) [51, 46]. Les répertoires de PfEMP1 de chaque génome (cellule unique ou cellules multiples) ont été alignés l'un avec l'autre et groupés avec Clustal-Omega (version 1.1.0) [131]. En utilisant igraph, tous les PfEMP1 ont été groupés dans des clusters de haute similarité dans lesquels ils partagent une identité supérieure à 95 % avec au moins un autre PfEMP1 du même cluster [125, 126].

Les lectures séquencées des génomes en cellule unique ou cellules multiples des échantillonnages du mois 6 ont été mappées sur les répertoires de gènes *var* extraits ou assemblés à partir des échantillonnages précédents en utilisant discovarif (https://github.com/marcguery/discovarif, version 0.0.6). Les gènes *var* chimères ont été recherchés avec DELLY (version 1.1.6), qui est capable d'identifier des translocations entre contigs en utilisant la localisation différentielle des paires de lectures après mapping [132].

Résultats

Pour anticiper l'aire de pré-élimination du paludisme, nous avons utilisé des méthodes de génotypage et de séquençage en génome entier pour construire un réseau de parenté génétique de parasites et ainsi d'accéder aux mécanismes de transmission du paludisme en Gambie.

En majorité, les parasites étaient génétiquement différents et ceux similaires étaient partagés entre les différents villages étudiés. Nous avons aussi trouvé que les parasites échantillonnés dans des foyers voisins étaient plus proches génétiquement seulement lorsqu'ils étaient échantillonnés à moins de trois mois d'écart. Quasiment toutes les souches identifiées avaient recombiné l'année suivante, renouvelant complètement la diversité génétique des parasites.

Au cours de cette étude, la plus longue infection que nous avons été capable d'observer a impliqué un individu de 9 ans continuellement infecté, sans aucun symptôme, pendant un an et demi par la même souche de parasite. Presque toutes les régions chromosomiques avec de hauts niveaux d'IBD partagés sont en déséquilibre de liaison avec les marqueurs de résistances aux anti-paludéens connus, ce qui indique que la résistance aux anti-paludéens exerce toujours une forte pression de sélection sur le génome parasitaire en Gambie.

Un aspect clé de notre approche était la détection active de cas d'infection asymptomatique contrairement à la plupart des études focalisées sur des parasites dérivés de cas cliniques. Nos résultats promeuvent la détection passive de cas comme une étape nécessaire pour caractériser correctement la diversité génétique des parasites.

Afin de découvrir les mécanismes de résilience de *Plasmodium falciparum*, 11 individus avec un paludisme asymptomatique durant 3 à 6 mois ont eu leur infection précisément suivie pendant la saison sèche 2016/2017 de la Gambie avec des échantillonnages sanguins mensuels pour chacun d'eux.

Parmi les 11 infections, 7 ont eu un ou plusieurs clones de parasite séquencés en lecture longue et 7 ont eu des cellules uniques ou cellules multiples séquencées en lecture courte à partir de 1 ou 2 échantillons de sang séparés de plusieurs mois. Malgré la faible parasitémie des infections asymptomatiques, les séquençages en lecture courte ou longue ont fait preuve de leur grandes qualités, le premier résultant en une couverture génomique élevée et le dernier en un très faible nombre de contigs presque aussi longs que des chromosomes et contenant la quasi totalité des gènes attendus dans *Plasmodium falciparum*.

Dans deux infections, le même clone parasitaire a été séquencé en cellule unique à partir de deux échantillonnages, le premier en décembre et le dernier en mai, ce qui a rendu possible la

découverte de variants qui ont augmenté en fréquence dans la population de parasites. Cette augmentation des allèles mutants pourrait résulter d'un avantage sélectif d'adaptation à l'hôte humain.

Un total de trois infections avaient à la fois plusieurs cellules uniques séquencées à partir du dernier échantillonnage (mois 6) et un ou plusieurs génomes distincts séquencés en lecture longue à partir d'un échantillonnage précédant. Pour une quatrième infection (DC05) disposant de nombreux génomes séquencés en cellule unique, un répertoire de gènes *var* a été directement assemblé à partir des lectures courtes des 67 cellules uniques afin de compenser le fait qu'aucun génome en lecture longue n'était disponible. L'assemblage *de novo* des gènes *var* à partir des lectures courtes a été un succès à la fois en terme du nombre attendu de gènes et de leur longueur par rapport à 3D7 et des autres génomes assemblés en lecture longue. Pour trois infections (dont DC05), les mappings des cellules uniques des échantillons du mois 6 sur les répertoires des gènes *var* obtenus dans des échantillonnages précédents ont montré une forte similarité entre les deux, suggérant que ces cellules uniques correspondent exactement à la souche de parasite utilisée pour obtenir le répertoire. Les gènes *var* chimères ont été recherchés à partir de ces mappings de haute similarité, mais aucun exemple convainquant n'a pu être reporté pour le moment.

Conclusion et perspectives

L'éradication du paludisme implique la destruction individuelle de chacun des parasites *Plasmodium falciparum*, incluant ceux des infections asymptomatiques à faible parasitémie qui sont généralement indétectables au microscope. Il est crucial que ces 'infections silencieuses', agissant comme un réservoir, soient éradiquées. Aujourd'hui, la grande majorité de la recherche sur le paludisme est concentrée sur une poignée de parasites adaptés aux conditions de laboratoire ou provenant d'infections symptomatiques obtenues en clinique. L'intérêt du projet de la cohorte de Gambie était de précisément décrire les infections chroniques asymptomatiques aux niveaux génomique et transcriptomique, de la cellule unique à la population de parasites, avec une attention particulière sur les variants de surface antigéniques du parasite.

Ce travail présente une large analyse de la diversité génétique aux niveaux populationnel et individuel du parasite *Plasmodium falciparum* présent dans des infections asymptomatiques en Gambie. Une grande variété de méthodes de séquençage a été utilisée pour accéder aux génomes parasitaires, incluant le génotypage de code-barres, le séquençage en lecture courte de cellules multiples et cellules uniques et finalement le séquençage en lecture longue de clones. De nombreuses difficultés ont dû être surmontées à la fois avant et après l'acquisition de données génomiques. La collecte et la culture des échantillons constituaient un défi à cause de la nature cachée des infections asymptomatiques mais aussi à cause de leur très faible parasitémie. Concernant les analyses des séquences génomiques, le génome de *Plasmodium falciparum* a imposé ses contraintes, notamment son extrême richesse en AT et son nombre important de régions hyper variables entourant les familles d'antigènes de surface, dont les gènes *var*.

Choisir le moyen approprié pour comparer les séquences génomiques est essentiel pour arriver à des conclusions biologiquement pertinentes. Grâce à l'IBD, la parenté entre les génomes des parasites peut être obtenue de manière informée à travers l'inclusion des caractéristiques de recombinaison de *Plasmodium falciparum*. Cette méthode a permis de montrer que les parasites étaient plus similaires lorsqu'ils étaient échantillonnés dans les trois mois et dans des individus vivant proches les uns des autres. Cela a été confirmé plus tard grâce aux multiples génomes séquencés en cellule unique. Deux autres facteurs ont affecté la parenté des populations de parasites. Premièrement les anti-paludéens ont pour effet de diminuer la diversité des haplotypes de parasite, puisque les régions les plus partagées entre les parasites sont situées proches des marqueurs de résistance aux anti-paludéens. L'autre facteur important concerne la saisonnalité du paludisme, où le début

des saisons de forte transmission est crucial pour que les parasites puissent recouvrer des longues périodes de faible transmission marquées par peu de recombinaisons.

Pour aller plus loin dans l'analyse des génomes des parasites, quelques individus sélectionnés ont eu leur infection analysée plus en détail. L'intérêt de cette étude a porté sur des infections chroniques durant plusieurs mois sans possibilité de ré-infection pendant les saisons de faible transmission. Certains parasites peuvent en effet rester dans le même hôte durant la totalité de la saison sèche, comme c'est le cas pour un individu de la cohorte infecté par la même souche de parasite pendant au moins un an et demi. Nous avons testé l'hypothèse que des recombinaisons entre les gènes *var* étaient la principale raison pour laquelle les parasites ne sont pas éliminés par le système immunitaire de l'hôte après plusieurs mois. Pour ce faire, des méthodes de séquençages plus précises, en particulier le séquençage de lectures courtes en cellule unique et le séquençage de lectures longues, étaient nécessaires pour rechercher ces événements rares qui ont lieu dans des régions hyper variables du génome.

Tous les génomes séquencés en lecture longue à partir des parasites adaptés à la culture pendant un mois ont été assemblés fructueusement avec de nombreux gènes identifiés dans peu de contigs presque aussi longs que les chromosomes de *Plasmodium falciparum*. Cependant, il est toujours nécessaire de les améliorer manuellement puisqu'il y avait des erreurs d'assemblage détectées lors de la comparaison des contigs avec les chromosomes de 3D7. L'annotation des gènes est actuellement améliorée par Mathieu Quenu (post-doctorant dans l'équipe d'Antoine Claessens) qui a fait le lien entre les gènes prédits et ceux annotés dans le génome de 3D7. De plus, l'un de ses objectifs principaux est d'obtenir l'assemblage de la famille hyper variable des gènes *rif* dans chacun des génomes assemblés. Malheureusement il n'y avait presque aucun génome en cellule unique disponible pour certains échantillons sanguins. Néanmoins, pour les globules rouges infectés au stage schizonte qui ont pu être triés, le séquençage en cellule unique a très bien fonctionné comme l'indique la haute couverture du génome de référence 3D7.

Pour deux infections monoclonales avec des cellules uniques disponibles pour les échantillonnages des mois 1 et 6, plusieurs mutations ont augmenté en fréquence. Pour retracer l'histoire de ces mutations, la prochaine étape se concentrera sur la ségrégations des cellules dans des groupes d'haplotypes. De plus, pour confirmer le potentiel avantage sélectif de ces SNP, leur fréquence allélique dans toutes les cellules devra être vérifiée. Des génomes séquencés en *bulk* ont été obtenus sur de nombreux échantillons sanguins des saisons humide et sèche. Les fréquences alléliques des SNP de ces génomes séquencés en *bulk* pourraient être comparées à celles obtenues à partir des cellules uniques des mêmes

échantillons sanguins. De plus, la totalité des génomes séquencés en *bulk* pourrait être utilisée pour vérifier que les fréquences alléliques des mutations identifiées des génomes séquencés en cellule unique sont aussi différentes entre les saisons sèche et humide, suggérant un rôle dans l'adaptation du parasite aux infections chroniques. Finalement, les mesures F_{WS} et R_{H} calculées sur les échantillons *bulk* pourraient être comparées avec les clusters IBD des cellules uniques.

Pendant ma thèse, j'ai développé deux pipelines qui pourront être utilisés dans le futur pour des travaux similaires. Le premier permet la comparaison et la fusion de données de séquençage en génome entier et de données de génotypage obtenues sur le même échantillon sanguin. Ce pipeline s'étend au calcul de la parenté entre les génomes, leur complexité d'infection et l'identification de génomes possédant des liens parents-progénitures. Le deuxième pipeline assemble des gènes *var* de génomes parasitaires séquencés en cellule unique et les fusionne par similarité, obtenant ainsi un répertoire de gènes *var* par infection. Bien qu'une méthode plus précise pour obtenir des répertoires de gènes *var* était de séquencer en lecture longue, cela a requit que les parasites étaient capables d'être cultivés *in vitro* et était plus coûteux que le séquençage en lecture courte. Pour valider la qualité du pipeline d'assemblage des gènes *var* à partir des cellules uniques, il pourrait être utilisé sur des génomes en cellules uniques qui sont identiques à un génome assemblé afin que les deux répertoires puissent être comparés. Tous ces gènes *var* assemblés *de novo* à partir de lectures courtes ou longues pourraient être ajoutés à la base de données de gènes *var* connus afin que les travaux futurs puissent inclure la diversité des parasites de Gambie.

Introduction

1.1 Malaria burden

1.1.1 Brief history of malaria life cycle discovery

Malaria is an ancient disease that was presumably at the origin of written records describing symptoms of fever and enlarged spleen in ancient Chinese, Egyptian, Indian and Mesopotamian civilizations dating back millennia [1]. These fevers, like many other illnesses at the time, were initially thought to be caused by a sort of poisonous cloud (called miasma) hovering swamps that could affect unfortunate people passing by [2]. It was not until the late 19th century, in a thriving microbiologist world led by Louis Pasteur and Robert Koch, that scientists began thinking a germ might be at the origin of malaria. In 1880, following his intuition of searching near pigmented bodies that were known to be degraded haemoglobin, Charles Louis Alphonse Laveran managed to observe parasites within red blood cells (RBC), discovering the asexual blood cycle of what will be latter known as *Plasmodium* [1, 2, 3]. The process by which *Plasmodium* is transmitted between humans was discovered later by Ronald Ross who was the first to describe mosquito stages of bird malaria parasites in 1897 and by Giovanni Battista Grassi who confirmed that mosquito bites could transmit malaria to humans in 1898 [1, 4, 5]. The complete life cycle was made available in 1949 when Henry Shortt and Cyril Garnham discovered that an exoerythrocytic life cycle of *Plasmodium* occurred in the liver, finally explaining why the parasite was missing from the blood for a few weeks after the mosquito bite (Figure 1.1) [1].

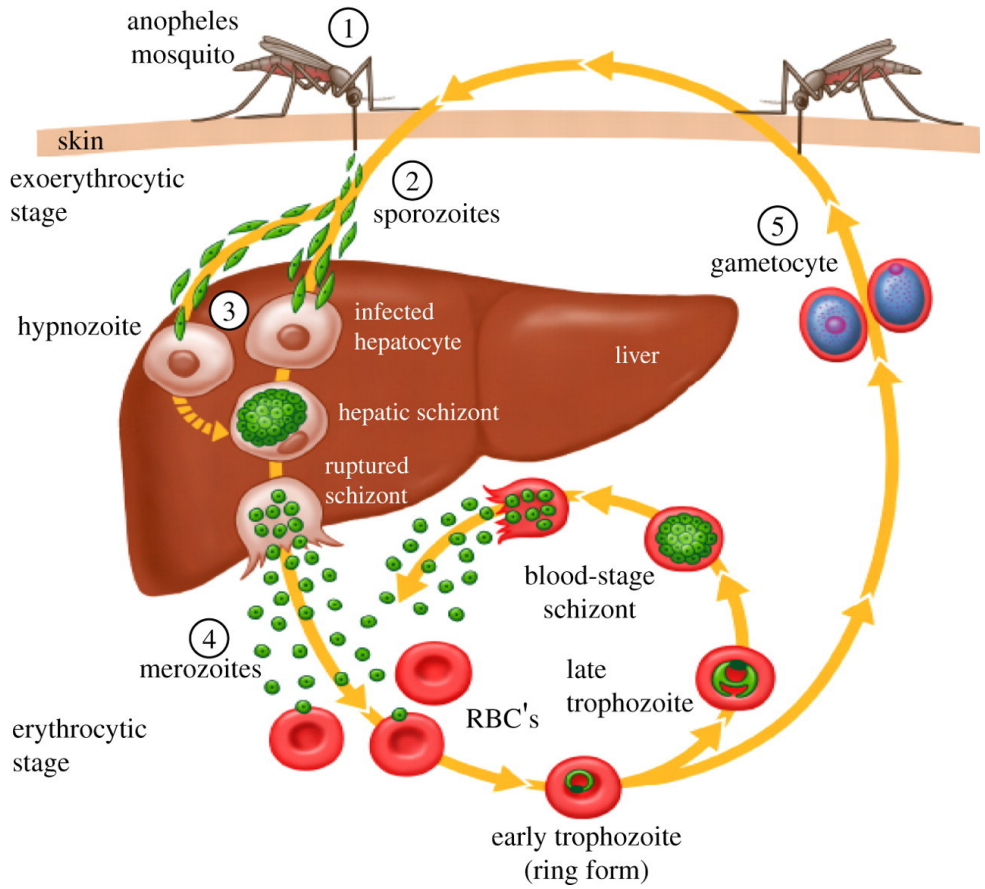


Figure 1.1: **Malaria life cycle**. Laveran discovered that malaria was caused by a parasite infecting RBCs. Ross and Grassi showed that the vector carrying parasites between humans is a mosquito. Finally, Garnham completed the life cycle by describing a exoerythrocytic stage in the liver immediately after the mosquito bite.

1.1.2 Human malaria species

Malaria parasites infect a wide range of species like birds, lizards, rodents and primates. There are in total five *Plasmodium* species specific to humans: *Plasmodium falciparum*, *Plasmodium malariae*, *Plasmodium ovale curtisi*, *Plasmodium ovale wallikeri* and *Plasmodium vivax* (Figure 1.2). A sixth species, *Plasmodium knowlesi*, whose natural hosts are macaques, is responsible for almost all zoonotic malaria infections in humans and is able to survive in a human-mosquito-human transmission setting, although this has only been experimentally demonstrated yet [134].

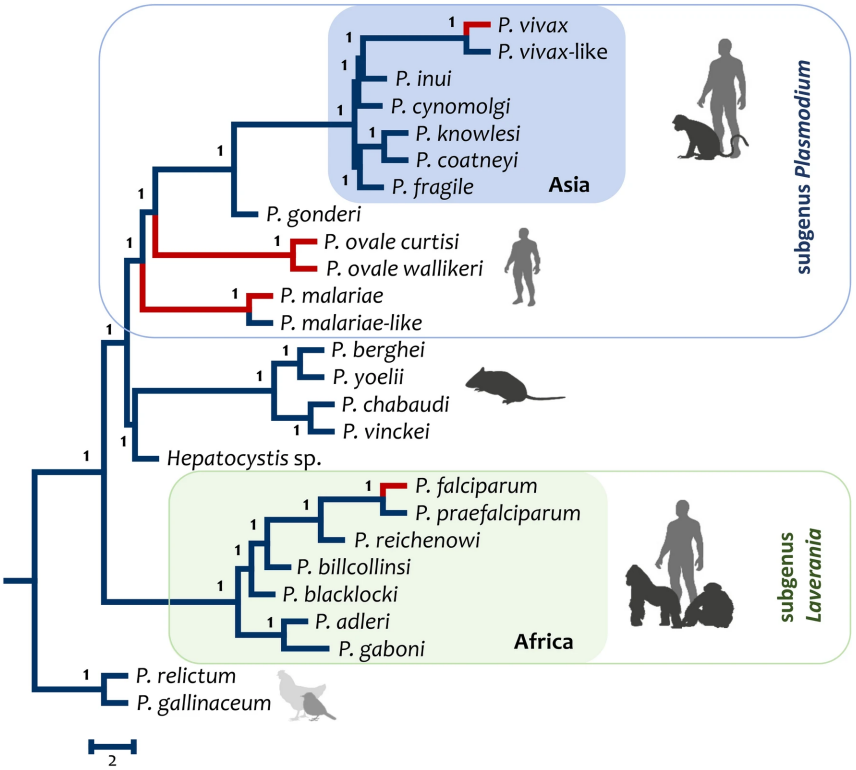


Figure 1.2: **Phylogenetic tree of *Plasmodium* species built by comparing orthologous genes**. The five main species infecting human are highlighted in red.

Conserved genetic material between available malaria species genomes like the mitochondria or orthologous genes are commonly used to infer the evolutionary history of *Plasmodium* [135]. There were seemingly two successive major speciation events that led to malaria in primates. A split from the bird malaria parasites branch that led to the *Laverania* subgenus occurred some 40 to 50 million years ago and was followed by another speciation event leading to the subgenus *Plasmodium* specific to primates [136, 137].

Plasmodium falciparum, the only human malaria parasite from the subgenus *Laverania*, may have originated in Africa from a transfer of its closest non-human malaria parasite found to date, *Plasmodium praefalciparum*, between 40 000 and 365 000 years ago [6, 7].

Despite their indistinguishable morphologies and their specificity to the same host, the two *Plasmodium ovale* species exhibit a high genetic divergence and may have split 5 times earlier than the split between *Plasmodium falciparum* and *Plasmodium reichenowi* which occurred between 140 000 and a few million years ago [137, 138, 139].

Plasmodium vivax geographical origin in human has not reached a strong consensus to date mainly because of its ability to remain dormant in the liver for an extensive time under

a hypnozoite form, which makes divergence values harder to estimate [7]. Some support a South-East Asian origin because of the highest genetic diversity of *Plasmodium vivax* observed in this region which gradually decreases with the geographical distance [140]. Other favour an African origin because multiple African ape malaria parasites (such as *Plasmodium vivax-like*) are genetically close to *Plasmodium vivax* and the indirect evidence that a human mutation conferring resistance to *Plasmodium vivax*, Duffy-negative, is widely spread in Central and Western Africa, suggesting a past history of an important selective pressure [141].

The two most virulent human malaria parasites are *Plasmodium falciparum*, mainly active in Africa and responsible for the vast majority of deaths worldwide, and *Plasmodium vivax* which is located in almost all malaria-endemic regions except Africa (Figure 1.3) [8].

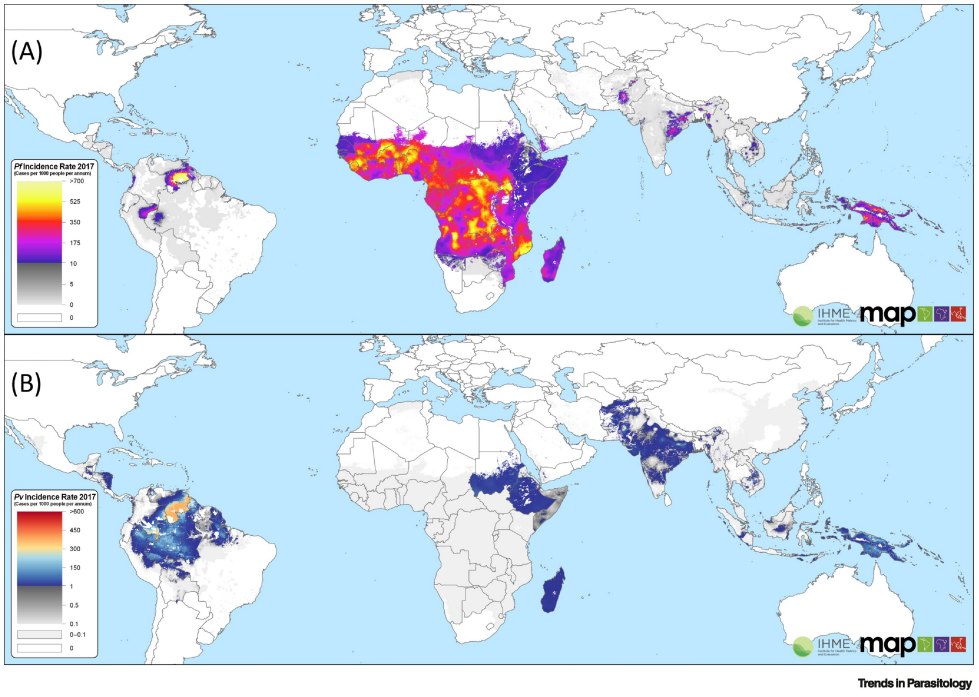


Figure 1.3: ***Plasmodium falciparum* (top) and *Plasmodium vivax* (bottom) malaria incidence in 2017**. *Plasmodium falciparum* is highly prevalent in sub-Saharan Africa while *Plasmodium vivax* is more globally homogeneously distributed. [8].

Although all human malaria species share the same host and vectors, their erythrocytic cycles are not all identical in terms of duration or infected red blood cell (iRBC) morphology. More specifically, *Plasmodium falciparum* erythrocytic life cycle lasts for 48 hours from the initial merozoite invasion to the release of new merozoites or alternatively gametocytes in the blood. Its cycle starts with rings until around 20 hours post invasion, then it follows with pigmented-trophozoites until 36 hours post invasion to finally end with the schizont form

which releases between 20 to 30 merozoites during a process called egress (Figure 1.4) [9]. The RNA and DNA content of the iRBC increase with the different stages, allowing for developmental stage determination [142]. The two biggest bottlenecks drastically reducing the population of parasites are during the exoerythrocytic phase in the liver with sometimes just one sporozoite making it and during fertilisation in the mosquito leading to very few oocysts [143, 144].

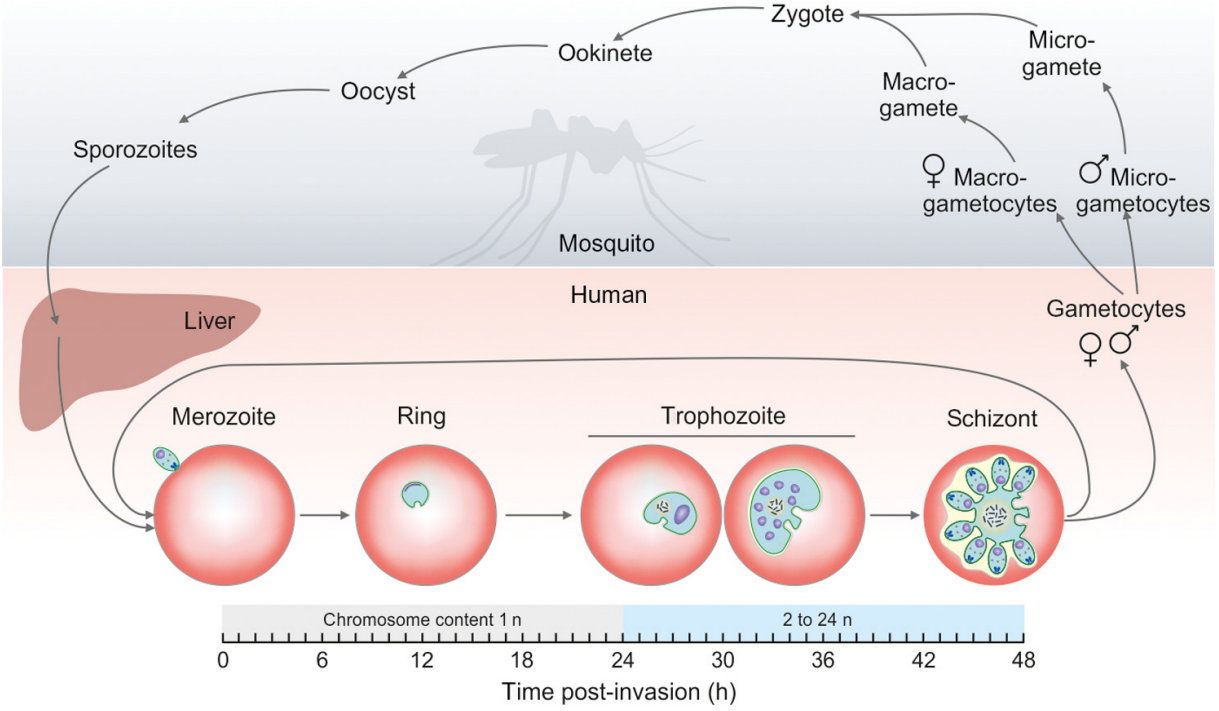


Figure 1.4: ***Plasmodium falciparum* erythrocytic cycle timing**. *Plasmodium falciparum* infects the human host (pink area) and is transmitted by an anopheles vector (grey area). After sporozoites enter the liver, they release merozoites in the blood which starts the erythrocytic cycle.

1.1.3 Malaria detection tools

Microscopy: The most direct way to diagnose malaria is by screening iRBCs in stained blood films through a light microscope. This provides a quantitative measure of parasitaemia and an experienced microscopist can distinguish malaria species that are morphologically different [146].

Rapid Diagnostic Tests (RDTs): RDTs detect *Plasmodium* specific antigens in blood pricks such as histidine-rich protein 2 (HRP2; PF3D7_0831800), specific to *Plasmodium falciparum* or *Plasmodium* lactate dehydrogenase (pLDH; PF3D7_1324900), with a conserve region similar

in all human malaria species and variable regions able to differentiate *falciparum* from non-*falciparum* malaria [147, 148]. These tests offer a cheaper and faster alternative to microscopy, however they have a similar (using HRP2 RDTs) to lower (using pLDH RDTs) sensitivity and cannot be used for quantification (Figure 1.5) [149, 11]. Moreover, the increasing number of strains with a deletion of *hrp2* poses a major threat as this antigen is by far the most used for *Plasmodium falciparum* malaria detection [150].

Polymerase Chain Reaction (PCR) tests: The two main Polymerase Chain Reaction (PCR) tests are used to amplify the 18S ribosomal RNA (rRNA) gene shared by all malaria species, or the *var* gene acidic terminal sequence (*var*ATS), specific to *Plasmodium falciparum* [151, 152]. Thanks to quantitative real-time PCR (qPCR), both PCR tests can be used to obtain an accurate estimation of parasitaemia and the 18S rRNA qPCR can even be used to distinguish the different malaria species with a single pair of primers [153, 154, 155]. Although much more sensitive than both RDTs and microscopy tools, performing PCR tests requires more human and financial resources which is not always affordable in malaria endemic regions (Figure 1.5) [10, 11].

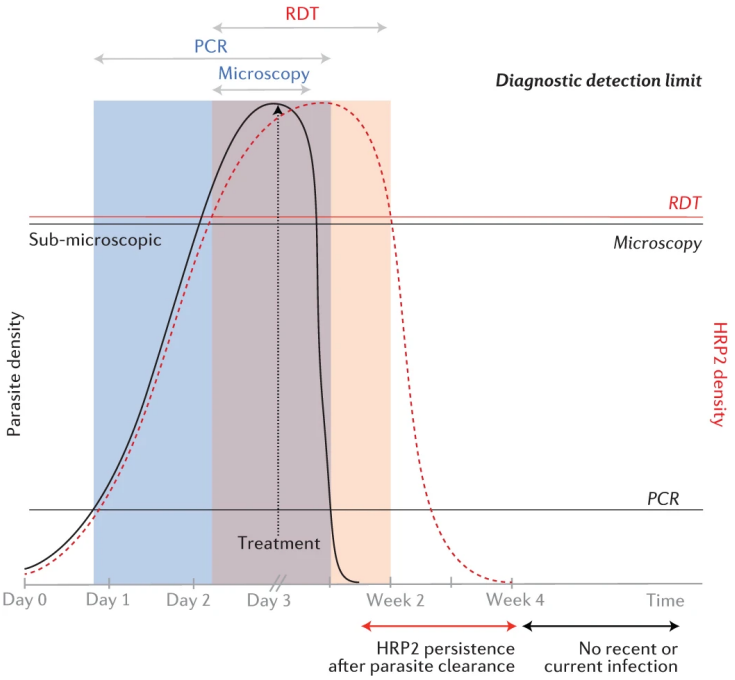
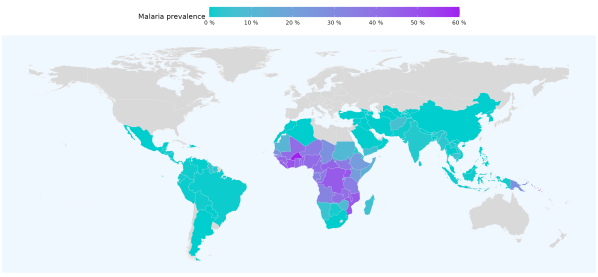


Figure 1.5: **Schematic representation of optimal usage of malaria detection tools regarding parasitaemia in an infected individual**. Of all three commonly used methods, PCR is by far the most sensitive. Because they are based on antigens, RDTs can still detect an infection that had been cleared a few weeks earlier as opposed to PCR or microscopy.

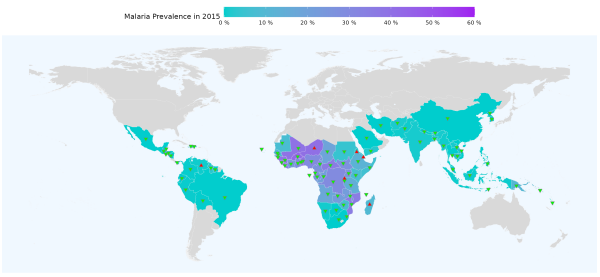
1.1.4 Control and elimination efforts

Multiple actions are led by malaria endemic countries to limit the malaria burden estimated to have affected 270 millions humans and taken 619 000 lives worldwide in 2021 [12]. Malaria surveillance is undertaken through passive case detection (PCD), where symptomatic people directly seek care and through active case detection (ACD), where health workers test individuals in a generally low transmission area with (reactive case detection) or without (proactive case detection) malaria infection previously reported.

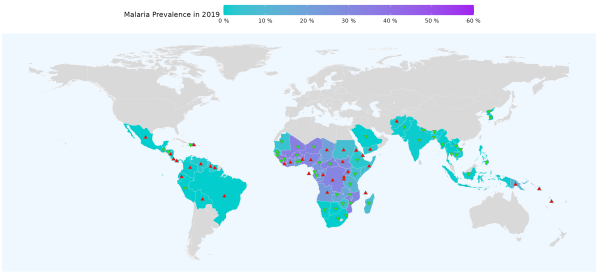
Malaria-endemic countries use mainly artemisinin-based combination therapies (ACT), which contains two distinct anti-malarial drugs to treat infected people. Vector control comprises measures such as the widespread distribution of insecticide-treated mosquito nets (ITNs) or indoor residual spraying of insecticides. Both global malaria case incidence and mortality rate have been declining in the last two decades thanks to these strategies, however some limitations emerged and slowed down the progress towards malaria elimination (Figure 1.6) [12, 156]. First, ACD require massive human, logistic and financial resources, more so in very low endemic regions as more sensitive and consequently expensive tests (PCR-based) are needed to be truly efficient [157]. Over the years *Plasmodium falciparum* has developed a resistance to most of the antimalarials which triggered the switch from monotherapies to ACTs with two or even three distinct drugs combined [13, 14]. Mosquitoes also are showing increasing level of resistance to different insecticides, including the highly efficient pyrethroid-based ITNs which represents almost 80 % of all insecticide usage in Africa [15, 16, 17]. Malaria elimination is also threatened by less direct factors, the most notable being the COVID-19 pandemic, especially at its peaks in 2020 and 2021, which caused a major disruption of malaria control services, leading to both case incidence and mortality rate falling back to their 2015 levels (Figure 1.6) [12, 18]. Other more regional political and economical instabilities, like the one currently hitting Venezuela, can interfere with malaria programs not only from within the country but also from neighbouring countries [158, 159].



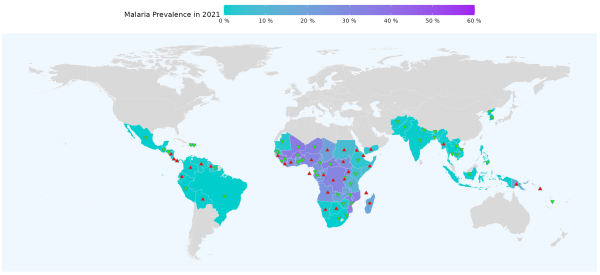
(a) Malaria prevalence in 2000.



(b) Prevalence in 2015 with changes from 2000.



(c) Prevalence in 2019 with changes from 2015.



(d) Prevalence in 2021 with changes from 2015.

Figure 1.6: **Malaria point prevalence**. For each selected year among 2000 (a), 2015 (b), 2019 (c) and 2021 (d), the number of malaria cases available in the WHO 2022 malaria report were divided by the country population at that time to obtain the point prevalence [12]. For years after 2000, increase (green triangle) and decrease (red triangle) in prevalences from a chosen previous year are shown. The most significant progress towards malaria elimination was made between 2000 and 2015. However, malaria prevalence remained stable or even sometimes slightly increased after 2015, although some countries achieved malaria elimination during that time. Made with Natural Earth.

1.1.5 Malaria disease and treatments

Plasmodium falciparum infections are generally classified into three categories based on the severity of symptoms.

Severe: Of all human malaria species, *falciparum* is the most likely to induce serious medical conditions. An infection that induces high risks of death is classified as severe [160, 161, 162]. The deadliest form of the disease, cerebral malaria, is characterised by an impaired consciousness with neurological symptoms such as coma, seizures or headache. Even with treatment, cerebral malaria has a mortality rate ranging from 10 to 20 % and patients are at high risk of developing sustained neurological impairments [163, 164].

Uncomplicated: More frequently, *Plasmodium falciparum* malaria can cause a cyclic fever with intermediate to high level of parasitaemia but without any severe symptoms.

Asymptomatic: Although positive for *Plasmodium falciparum*, an individual can be free of any symptom for an extensive period of time with often low parasite densities that are not detectable with microscopy or RDT but only PCR. This type of infection is described as a

subpatent chronic infection [19, 20]. In some endemic regions, asymptomatic infections are highly prevalent in individuals older than 5 years who sometimes present a parasite density nearing that of symptomatic infections [21, 22, 23, 24]. Asymptomatic infections tend to last longer in children between 5 and 15 years old and more generally in male individuals, with durations of infection of more than on year reported multiple times [165, 25, 22]. However, much longer durations of infections have been reported in several documented cases of blood donation where the donor transmitted *Plasmodium falciparum* malaria to their recipients through iRBC, although neither of them had been in a malaria endemic region for up to 13 years [166].

Multiple drugs have been used throughout the history of antimalarial therapies in a constant race with the parasite and its ability to generate resistant phenotypes and spread them globally.

Quinine: The first antimalarial drug to be widely used was quinine, chemically extracted in 1820 from the cinchona, a tree originating from South-America whose bark was already known to cure fever-like symptoms that we can nowadays attribute to malaria [26, 27].

Chloroquine: The search for synthetic quinine analogues led to the discovery of chloroquine in 1934, which gradually replaced quinine as the main antimalarial used worldwide [26]. However, it did not take long before the first resistance to chloroquine emerged in *Plasmodium falciparum* both in South East Asia and South America in the 1950s or 1960s. In the 1980s, chloroquine resistance was widespread globally [28, 13]. Chloroquine resistance is known to be induced by multiple mutations in positions 72 to 76 of the *Plasmodium falciparum* chloroquine resistance transporter (*pfCRT*; PF3D7_0709000) that generate resistant haplotypes CVIET or SVMNT instead of the sensitive form CVMNK [167, 28]. Two other mutations are also commonly found in chloroquine resistant parasites: N86Y of the *Plasmodium falciparum* multidrug resistance protein 1 (pfMDR1; PF3D7_0523000) and S258L of the *Plasmodium falciparum* amino acid transporter (pfAAT1; PF3D7_0629500) [168, 169].

Sulfadoxine/Pyrimethamine: The global spread of chloroquine resistance made obvious that other drugs were needed to reduce the malaria burden. Sulfadoxine/Pyrimethamine (SP), which contains two drugs acting synergically, is the first combination therapy used to treat malaria as a replacement for chloroquine. Unfortunately, shortly after its commercial use, *Plasmodium falciparum* resistance arose in the 1970s in South East Asia, the same region where chloroquine resistance emerged [29, 13]. The main non-synonymous mutations associated with resistance were here found in positions 51, 59, 108 and 164 of *Plasmodium falciparum* dihydrofolate reductase (pfDHFR; PF3D7_0417200) and positions 436, 437, 540, 581 and 613

of *Plasmodium falciparum* dihydropteroate synthase (pfDHPS; PF3D7_0810800) [28]. A high number of distinct haplotypes are present in different proportions between malaria-endemic regions, with in majority parasites containing 2 or 3 substitutions in pfDHFR and 1, 2 or 3 substitutions in pfDHPS [170, 171].

Artemisinin: Similarly to quinine, artemisinin was extracted in 1971 from a plant that had been used to treat fevers in Chinese medicine [30, 27]. Due to its short half life, artemisinin was rapidly used in combination with a partner drug that had a distinct mode of action in ACT (artemisinin-based combination therapies) [31]. Drug resistance to artemisinin emerged in South East Asia in the 2000s, with parasites displaying a range of non-synonymous mutations in the gene *Plasmodium falciparum* Kelch 13 (pfK13; PF3D7_1343700), the most prevalent being C580Y [32, 13, 33, 34]. Additionally, *Plasmodium falciparum* harbouring several gene copies of pfMDR1 were shown to be less sensitive to artemisinin and several of its partner drugs [172, 173]. Artemisinin resistance was also recently spotted in Uganda with a small proportion of patients carrying parasite strains requiring a longer treatment before complete clearance [174]. As ACTs are the main therapeutic strategy used today against uncomplicated malaria infections, the worldwide spread of artemisinin resistance would severely hinder global efforts against malaria elimination.

Most of the resistant strains for the different antimalarials emerged in regions with a relatively low prevalence of *Plasmodium falciparum* infections like in South East Asia or South America (Figure 1.7) [28]. This could be explained by a lower immunity and a better access to healthcare of the human populations in these regions, therefore more likely to seek treatment and thus leading to the selection of resistant parasites. The intra-host competition between strains should favour sensitive over resistant strains as most resistant mutations come with a fitness cost. However, in these regions of low prevalence, the low complexity of infection makes this competition almost non existent and this allows resistant strains to acquire even more polymorphisms that might lower the cost of the resistance [175, 176]. But even highly prevalent resistant strains can be out competed by sensitive ones in the absence of the drug pressure, as it was observed in several countries that discontinued the use of chloroquine and have now a majority of parasites susceptible to this antimalarial [177, 178, 179].

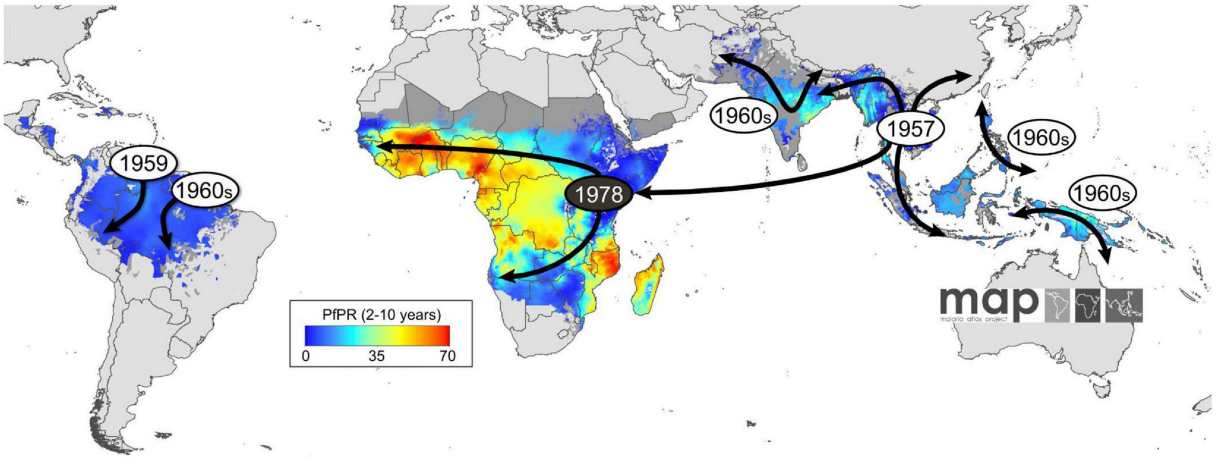


Figure 1.7: **Spread of chloroquine resistance**. Chloroquine resistance most likely emerged independently first in South East Asia and then in South America to finally spread from Asia to Africa. The resistance to SP probably had the same route as chloroquine resistance [29, 13].

1.2 The diversity of *Plasmodium falciparum*

1.2.1 *Plasmodium falciparum* genome

Plasmodium falciparum can be cultivated *in vitro*, the most studied laboratory-adapted strain is 3D7. It is derived from NF54, a parasite that infected a patient from the Netherlands, but likely originated from Africa as is suggested by the comparison with global clinical isolates [35, 36, 37]. Other laboratory-adapted strains were obtained from different parts of the world. Dd2 was cloned from a strain cultivated in the presence of the antimalarial mefloquine (W2-mef) that was itself derived from W2, collected from a patient in former Indochina [180, 181]. HB3 is another commonly used strain which was collected in Honduras. As these 3 strains originate from distant parts of the world, they have several specificities. W2 and Dd2 are resistant to many antimalarials such as quinine, chloroquine, sulfadoxine and pyrimethamine while 3D7 and HB3 are only resistant to sulfadoxine and pyrimethamine respectively [182, 183, 184, 185]. Additionally, Dd2 lacks the *hrp2* gene and HB3 lacks the *hrp3* (histidine-rich protein 3; PF3D7_1372200) gene (HRP3 shares similar epitopes with HRP2), reducing both their sensitivity to RDT tests [186].

Plasmodium falciparum 23 Mb (million of base pairs) haploid genome was fully sequenced and assembled from the 3D7 strain in 2002 which then became the reference genome [38]. The parasite has 14 nuclear chromosomes constituting one of the most AT-rich genome among all known species with an overall AT content of 81 %, reaching 84 % in non-coding regions. There are additionally one mitochondrion and one apicoplast per parasite with respectively around 20 copies of 8 kb (thousand of base pairs) long mitochondrial genome and 15 copies of 35 kb long apicoplast genome [39, 40]. The 16 different chromosomes of the parasite contain almost 5600 genes, among which 5300 code a protein. The base pair substitution (BPS) rate of the nuclear genome was estimated to be between 5×10^{-3} and 1×10^{-2} non-deleterious mutations per erythrocytic life cycle and was similar for all strains tested (3D7, Dd2, HB3, KH01 and KH02) [187, 54, 188]. G:C to A:T mutations are about 5 times more likely to occur than A:T to G:C mutations, which explains the 81 % AT content of *Plasmodium falciparum* genome [188]. Around 10 % of the genome corresponds to hypervariable regions (regions including genes with a high degree of polymorphism) and subtelomeric repeats (non-coding repeated regions near the end of chromosomes) that are highly variable between strains and complicates the comparison of these regions [41]. Repeated regions cover roughly 23 % of *Plasmodium falciparum* genome and are differentially present in exonic (10 %), intronic (48 %) and intergenic regions (38 %) (Figure 1.8).

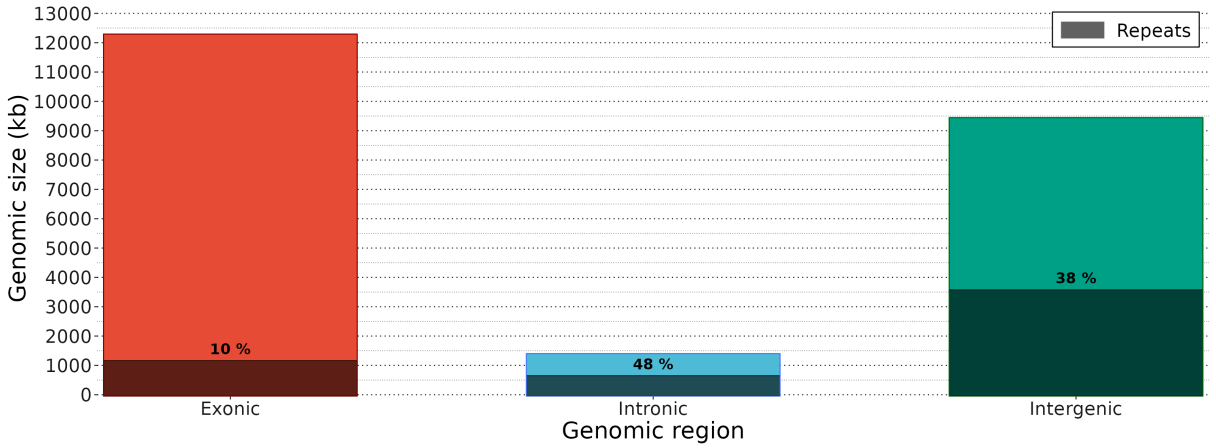


Figure 1.8: **Repeated sequences in *Plasmodium falciparum* 3D7 genome**. Roughly half of the 23 Mb of the genome of *Plasmodium falciparum* contains exons. This region is the one that has the smallest proportion of repeated regions (10 %) behind intergenic regions (38 %) and intronic regions (48 %). Repeats were identified with Tandem Repeats Finder, as described in *C.1.2 Repeat motif identification* [189].

These low complexity regions cannot be easily assembled using next generation sequencing technologies because the very long repeats cannot be bridged by short-reads of typically 150 bases. In 2018, thanks to third generation sequencing (or long-read sequencing), the assemblies of the low complexity regions of 3D7 chromosomes were dramatically improved and supplementary assemblies of other strains such as Dd2 and HB3 were produced [42].

1.2.2 Hypervariable genes

To successfully invade a human host, *Plasmodium falciparum* exports certain proteins to the surface of either its own plasma membrane or the membrane of the RBCs it infects. These genes have evolved to carry highly polymorphic sequences, allowing them to adapt to diverse immune systems and hinder the rapid elimination of the parasite by the host.

Plasmodium falciparum possesses several genes exposed at its plasma membrane surface such as CSP (circumsporozoite protein; PF3D7_0304600), which is exported at the surface of sporozoites, AMA1 (apical membrane antigen 1; PF3D7_1133400) or MSP2 (merozoite surface protein-2; PF3D7_0206800), which are both exported at the surface of merozoites. A total of 66 non-synonymous mutations were found in the worldwide available sequences of CSP, with most of the polymorphisms observed in Th2R and Th3R C-terminal regions [190]. AMA1 contains 62 polymorphic amino acids, especially those located in the c1L domain that are the most diverse among parasites worldwide [191]. The MSP2 gene is also very diverse with numerous size specific alleles belonging either to 3D7 or FC27 families [151, 192].

A very particular gene family called *var* is also very diverse within and between individuals. *var* genes are transcribed from the early ring-stage of the erythrocytic cycle and the proteins that they encode, *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1), are expressed in knob structures on the surface of iRBCs from the trophozoite and schizont stages [43]. One function of PfEMP1 is to bind to host receptors at the surface of endothelial cells, a cytoadherence process known as sequestration [44, 45]. The roughly 60 *var* genes (their number varies between parasite genomes) are all organised similarly with two exons separated by one intron and are located in subtelomeric or internal parts of chromosomes within hypervariable regions (Figure 1.9) [41, 46]. PfEMP1s, with a size of 1000 to 4000 amino acids in 3D7, all have the same configuration with several extracellular domains comprising a N-terminal segment (NTS) and successive highly variable DBL (Duffy binding-like) and CIDR (cysteine-rich interdomain region) domains encoded by the exon 1, and a more conserved intracellular domain called ATS (acidic terminal segment) encoded by the exon 2 [47]. All *var* genes start with a specific DBL domains called DBLα, with the exception of the only strain-transcending *var* gene, *var2csa* (PF3D7_1200600). Interestingly, only one *var* gene per erythrocytic cycle is transcribed, resulting in only one type of PfEMP1 at the surface of the iRBC.

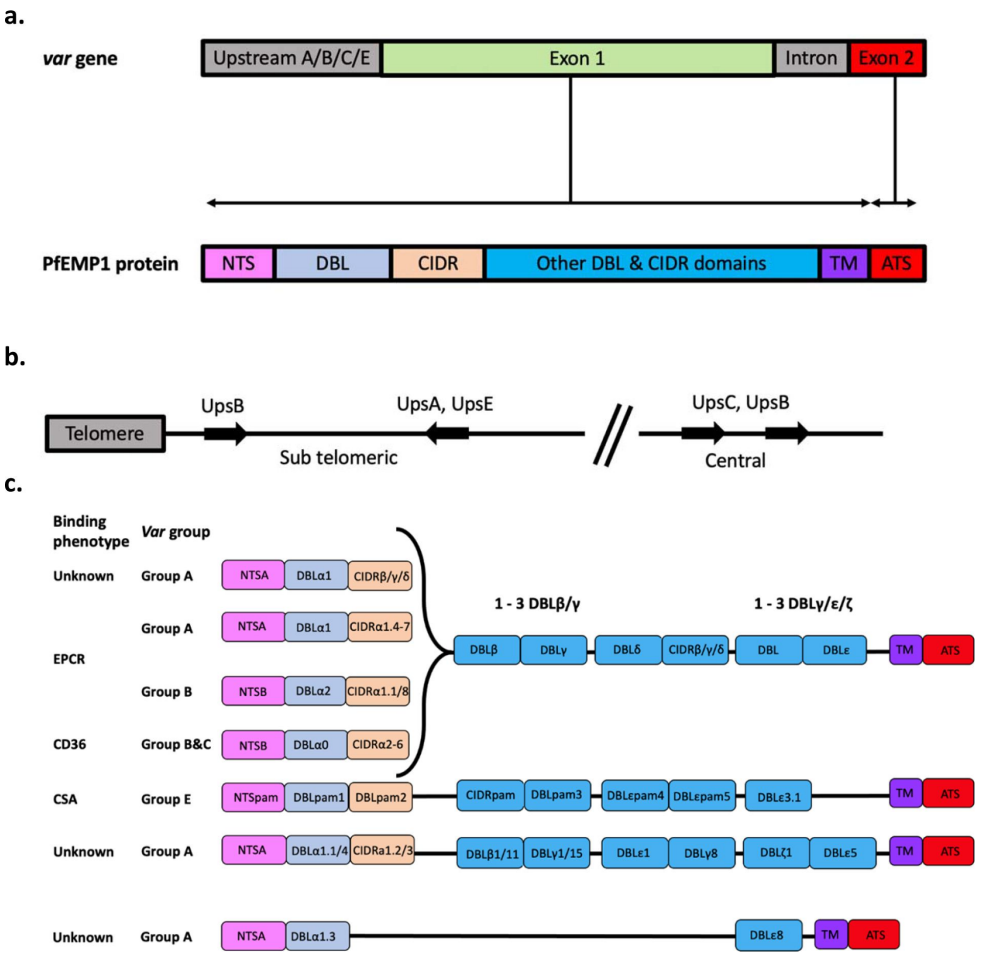


Figure 1.9: ***var* gene classification**. (a): Each *var* gene contains one intron and two exons which encode a protein called PfEMP1. (b): *var* genes are generally organised in cluster in subtelomeric or internal regions of most chromosomes. (c): PfEMP1 have a wide variety of possible domain configurations and all but the one belonging to group E contain a DBLα domain.

Based on the sequence similarity of the upstream (ups) genomic region, three major groups of *var* genes (upsA, upsB, upsC) and one minor group with one *var* gene (upsE) were defined using 3D7 genome [38]. Interestingly, upsA and upsC *var* genes are exclusively located in respectively subtelomeric or internal regions of chromosomes while upsB *var* genes are located in both regions (Figure 1.10). As this classification was too limited to characterize the full diversity of *var* genes, deeper classifications were made available. The similarity between DBL-CIDR, ATS or downstream regions showed that several upsB *var* genes contain domains that are typically observed in upsA or upsC *var* genes [48]. Even though *var* genes are highly heterogeneous, there are two *var* genes that are particularly conserved between *Plasmodium falciparum* strains. Belonging to upsA group, *var1csa* (PF3D7_0533100), which is truncated

with a missing exon 2 in 3D7, is transcribed mostly at the trophozoite stage and not the ring stage like all other *var* genes [193, 194]. *var2csa*, the sole member of the upsE group, is the only *var* gene that is not starting with a DBLα domain and is highly expressed in parasites from pregnancy-associated malaria (PAM) cases [195]. About 30 % of all *Plasmodium falciparum* genomes available, including HB3, present multiple copies of *var2csa*, while the remaining genomes, including 3D7, only have one copy [196, 197].

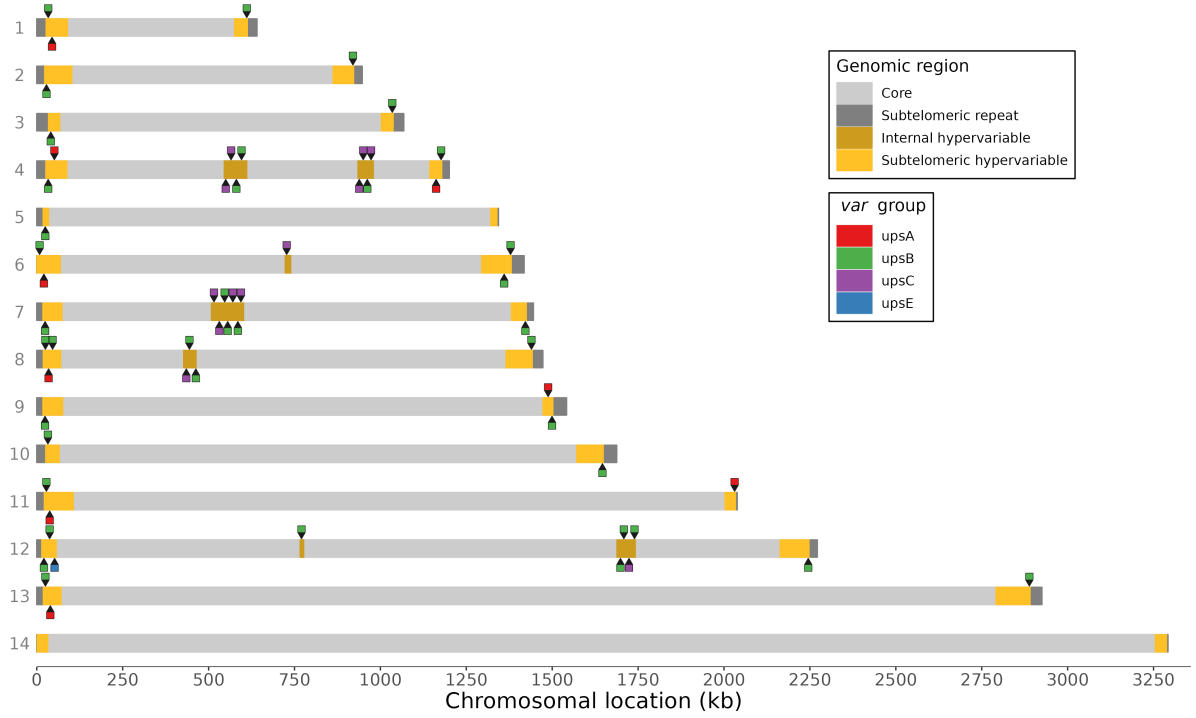
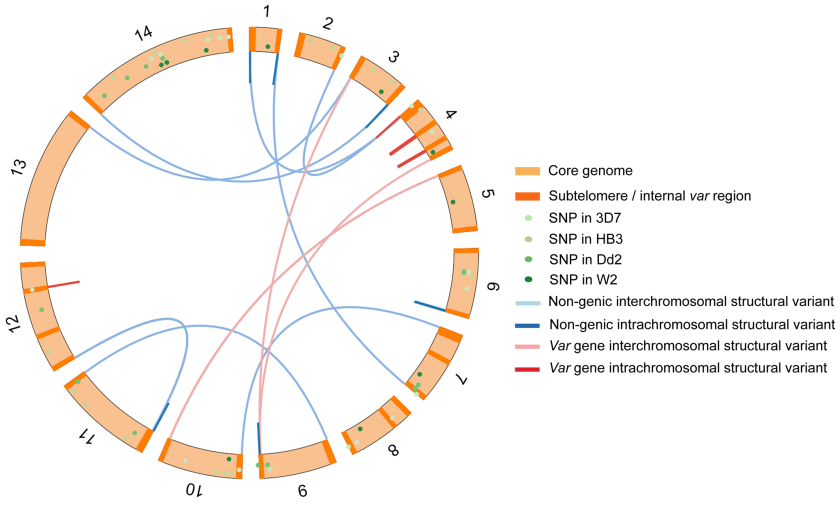


Figure 1.10: ***var* gene chromosomal location in 3D7**. Each *var* gene belongs to upsA, upsB, upsC or upsE group. upsA and upsC *var* genes are exclusively located in respectively subtelomeric and internal regions, while upsB *var* genes are present in both regions. *var* genes can be organised in clusters in subtelomeric and more particularly in internal regions of most chromosomes. Categories of genomic regions were extracted from Miles *et al* (2016) and *var* genes were grouped using the classification of Lavstsen *et al.* (2003) [41, 48].

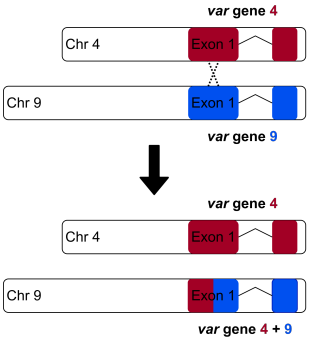
Because of their particular location in repeated regions and their high variability between parasite genomes, *var* gene sequences from clinical isolates cannot be accessible using a typical mapping of short-read sequences to a unique reference repertoire. Instead, *var* gene sequences are better retrieved using *de novo* assembly pipelines, which, despite the gradual improvement of their efficacy, still require many pre-processing steps and often lead to incomplete repertoires [51, 46]. The most reliable way to obtain complete high quality *var* gene repertoires is to use third generation sequencing technologies, such as Pacific Biosciences (PacBio) single molecule real-time (SMRT) sequencing which can produce reads

exceeding 12 kb, covering *var* genes from end to end [52, 42]. Circular consensus sequencing (CCS) overcomes the high error rate associated with long-read sequencing by building a consensus read from multiple sub-reads of the same DNA template, resulting in high fidelity (HiFi) reads [198]. These consensus reads exhibit error rates comparable to those of short-read sequencing, enabling the precise detection of even rare structural variants [199, 200, 201].

The diversity of *var* genes can be generated via chromosomal recombination between different *var* genes (more precisely between different *var* exons 1) in a process called ectopic recombination, which is frequent during the meiosis in the mosquito [53, 50]. These ectopic recombinations can partly explain why multiple *var* genes from a ups group display domains typically observed in another ups group. Chimeric *var* genes can also be generated from ectopic recombinations during mitosis with a rate of about 2×10^{-3} per parasite for every asexual erythrocytic cycle (Figure 1.11) [54]. There are multiple potential breaking points within *var* genes corresponding to regions with a higher homology that are generally located between two identical domain classes encoded by the exon 1 [54].



(a) Genetic variants detected in laboratory-adapted strains.



(b) *var* exon 1 unbalanced ectopic recombination.

Figure 1.11: **Mitotic *var* gene recombinations from the erythrocytic life cycle**. (a): The laboratory-adapted strains 3D7, HB3, Dd2 and W2 were cultivated for several months and regularly cloned before being sequenced in the aim of finding genetic variants. The overall detected structural variants of 3D7 were the result of intra-chromosomal and inter-chromosomal recombinations either between *var* genes (exon 1) or between intergenic regions. For the remaining strains, *var* gene structural variants could not be mapped as high quality chromosomal assemblies were not available at the time. (b): An unbalanced ectopic recombination was detected between the first exons of two *var* genes located in chromosomes 4 and 9.

The 5 prime end of almost all *var* genes contains a DBLα domain [49]. Within that domain, the 'DBLα tag' is about 130 amino acids long and contains a variable sequence surrounded by the conserved motifs LARSFADIG (N terminal side) and DYVPQ[YF]LRW (C terminal side) whose conserved corresponding DNA sequences can be used for PCR amplification (Figure 1.12) [50]. As *var* genes sometimes show a higher degree of similarity between them, several classifications based on mutually exclusive semi-conserved motifs or on homology blocks emerged [202, 47]. Classifications obtained with just these 130 amino acids have been shown to correlate well with the consistent presence of features from whole *var* genes such as domains or ups regions [203]. Knowing the extent of the information brought by DBLα tag classification alone is essential as the PCR amplification of this short sequence is much less cost-intensive than the sequencing of full length *var* genes.

Figure 1.12: **Frequency of DBLα tag amino acids**. The DBLα tag from the DBLα domain of *var* genes is defined as a highly variable sequence surrounded by conserved motifs LARSFADIG and DYVPQ[YF]LRW. The amino acid frequency was calculated from a random subset of 1000 sequences from the varDB database [51].

There are other similar families of genes that encode proteins that co-localize with PfEMP1 around knobs of the iRBCs: RIFIN (repetitive interspersed families of polypeptides) and STEVOR (subtelomeric variable open reading frame) encoded respectively by *rif* and *stevor* genes (Figure 1.13) [55]. Together PfEMP1, RIFIN and STEVOR are the main variant surface antigens (VSAs) of *Plasmodium falciparum*, all binding to various host receptors. The 150 to 200 *rif* genes and the 28 *stevor* genes are organised in clusters close to *var* genes, suggesting that they might also undergo frequent recombination events lying in these hypervariable regions [56, 57].

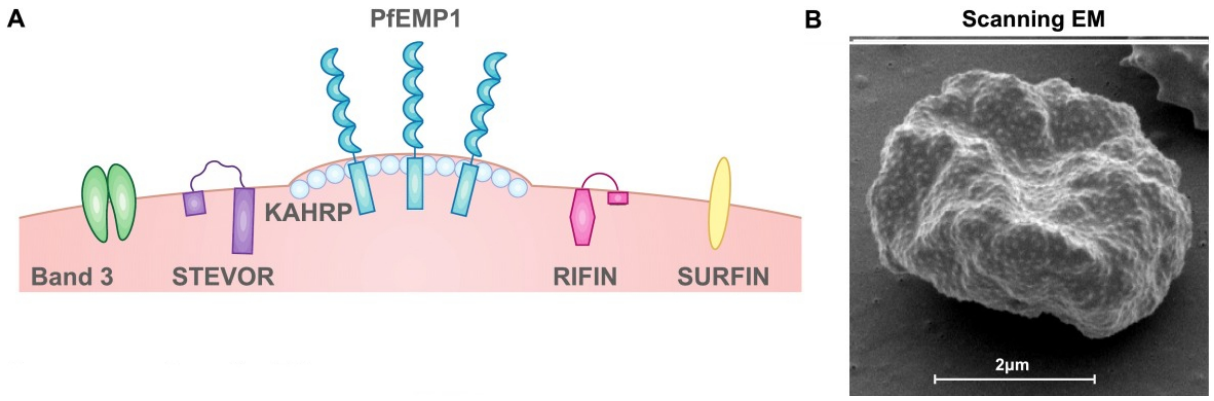


Figure 1.13: **Variant surface antigens location**. (A): PfEMP1, RIFIN and STEVOR are three VSAs located in or next to knobs on the surface of iRBCs. The other proteins shown are the human Band 3 and the *Plasmodium falciparum* KAHRP (knob-associated histidine-rich protein; PF3D7_0202000) and SURFIN (surface-associated interspersed protein). (B): Knobs at the surface of an iRBC visible under an electron microscope (EM).

1.2.3 Complexity of infections

The complexity of infection (COI) represents the count of distinct clones within an individual infection. There are two ways leading to multiple parasite genotypes infecting a human host simultaneously. This can be achieved by successive bites of mosquitoes that each inoculates a genetically distinct parasite (called a superinfection), a single bite of a mosquito carrying different sporozoites generated from the recombination of two gametocytes (called a co-transmission), or both ways concurrently (Figure 1.14) [58]. In regions with a high malaria prevalence, co-transmission becomes very common while superinfection is surprisingly much more rare [59]. When malaria prevalence increases in a region, the proportion of polyclonal (or poly-genotype) infections will also increase as a result of a higher chance of mosquito carrying one or more parasites [60, 61]. In fact, the change in values of mean COI or proportion of mixed infections over time offer one of the best genetic diversity-based estimators of the change in local malaria prevalence with a change lagging just a few weeks behind its effect on malaria prevalence [62, 63]. However, absolute values of COI cannot be used to estimate malaria prevalence or even to compare population of parasites from two distinct locations [64, 65]. This is mainly because populations of parasites geographically isolated do not share the same level of genetic diversity, those highly inbred cannot allow for many distinct parasite strains even in high transmission settings.

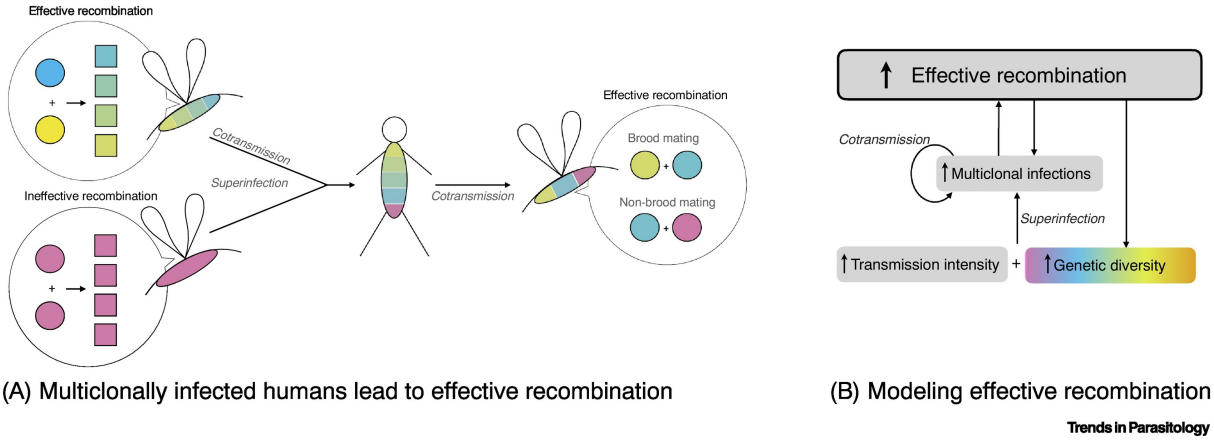


Figure 1.14: **Complexity of infection**. (A): A superinfection occurs when a mosquito transmits identical sporozoites originating from identical gametes to an already infected human host who then becomes multiclonally infected. When a mosquito bites a multiclonally infected host, it has a high chance to carry and then co-transmit different but related sporozoites originating from distinct gametes. (B): In high transmission settings, both co-tranmssion and superinfection increase the genetic diversity, provided that the initial population is sufficiently diverse.

1.2.4 Identifying parasite strains

The identification and comparison of *Plasmodium falciparum* genotypes between infections and within infections is necessary to understand its evolutionary history or the on-going transmission dynamics in a region. There are multiple categories of genetic markers able to identify specific *Plasmodium falciparum* strains with different level of precision.

Highly polymorphic genes: The two genes *csp* and *ama1* can be amplified by PCR using primers of the conserved regions and sequenced to identify individual parasite strains [205, 206]. Another commonly used marker is the gene *msp2*. As *msp2* alleles are size specific, genotyping can be achieved simply using PCR product sizes without the need to sequence them [66, 67]. High levels of COI can be estimated from PCR amplification of *msp2* by counting the number of distinct sizes of the PCR products within the same isolate [68, 24]. As detailed above, the highly polymorphic region within the DBLα tag can act as a genetic 'barcode'. An amplicon-sequencing approach of DBLα tags from field isolates can distinguish geographic population structure at the local and worldwide level [69, 70]. Similarly, the comparison of full-length *var* gene repertoires between isolates can be used to infer the level of similarity between parasites [51].

Locus Genotyping: Parasites can be identified by genotyping bi-allelic single nucleotide polymorphisms (SNPs) scattered in the genome. Levels of similarity between parasites are then obtained by calculating their pairwise distance which is an estimate of their level of

genetic similarity. As little as 24 SNPs is already enough to obtain an estimate of population genetic diversity that correlates well with transmission intensity or geographical distance [71, 60, 72, 73]. A better resolution of genetic diversity can be achieved by using more SNPs, for example 96, 101 or even 250 [61, 123, 98]. The COI can be estimated by providing these bi-allelic SNP calls to THE REAL McCOIL, an algorithm that estimates the population allele frequencies [124]. Microsatellite markers, which correspond to short tandem repeats located in many places throughout the genome, are also commonly used to estimate the parasite genetic diversity or even the complexity of infections [74, 75].

Whole genome sequencing (WGS): Although locus genotyping is highly effective and relatively cost-efficient, it may not fully capture the diversity of parasites, especially in regions with intermediate or high transmission intensities, where the very high complexity of infection (COI) can bias the calculated relatedness between infections [76]. In such cases, whole genome sequencing (WGS) can be used to access much more genetic loci in each parasite, typically involving tens of thousands of SNPs [77, 78]. These SNPs are generally located in the core genome as these conserved regions are similar and thus comparable between all parasite genomes [41]. With a high number of SNPs, a straightforward pairwise comparison might not accurately reflect the actual genetic relatedness between two parasites because they are not independent. Each meiosis fragments and shuffles two initially distinct genomes, with a fixed recombination rate of 13.5 kb/cM [41]. If two parasites share a recent common ancestor, they will consequently possess genetic fragments with identical SNPs, the size of which depends on the number of generations that precede them, given the recombination rate of *Plasmodium falciparum*. These parameters are employed by hmmIBD, a hidden Markov model that seeks genetic fragments in identity by descent (IBD) shared between two genomes (with a minimum of 200 bi-allelic markers) based on their most likely recombination history (Figure 1.15) [79, 80]. A pairwise SNP comparison can also be biased by the level of linkage disequilibrium in a population of parasites or the presence of markers under non-neutral selection, such as those involved in drug resistance. Consequently, each pair of genomes will have an associated value of relatedness, indicating the overall percentage of identity along with the coordinates of chromosomal fragments that are either in IBD or not [60]. Measuring COI with bulk WGS data can be achieved using R_{H} or F_{WS}, which both take advantage of the within-sample allele frequencies of polymorphic genomic loci [81, 82]. The F_{WS} metric is calculated for each individual as $1 - H_W/H_S$ (with H_W the within-sample heterozygosity and H_S the local population heterozygosity) in a similar way to the inbreeding coefficient F_{IS}. A higher F_{WS} value indicates lower heterozygosity (and therefore lower polyclonality) in the sample. For example, an F_{WS} value above 0.95 typically

indicates that a sample is monoclonal [207]. R_H is very similar to F_{WS} but has also the ability to distinguish co-transmissions from superinfections. To obtain the precise sequence of individual parasite strains in a complex infection, single-cell sequencing is a necessary tool to ensure that highly related co-transmitted parasites can be segregated [59]. WGS can also be used to highlight genomic regions with a lower heterozygosity than for the rest of a genome, which is generally observed around genetic loci under recent positive selection: the integrated haplotype score (iHS) was developed for that purpose [208].

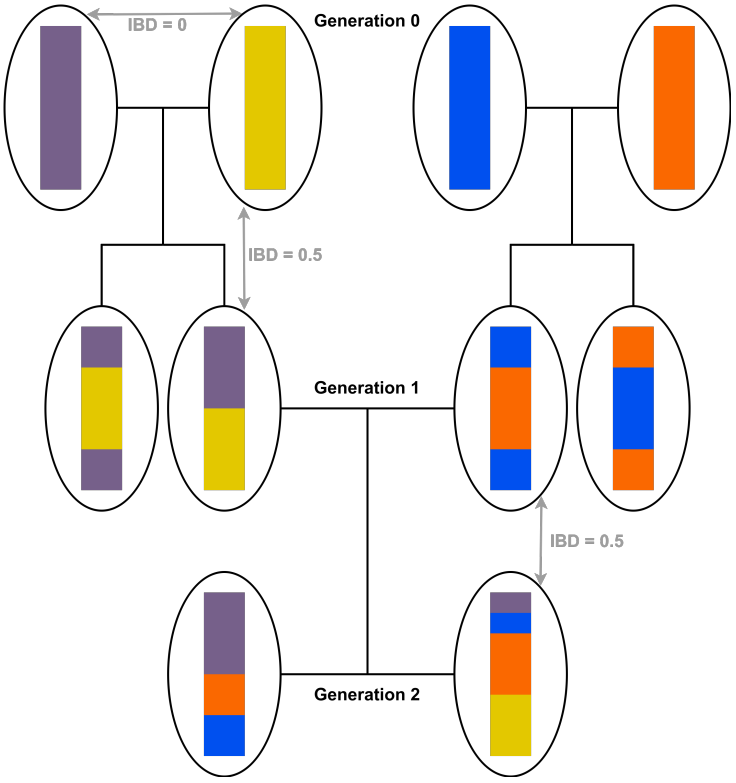


Figure 1.15: **Identity by descent**. IBD refers to the proportion of genetic material shared between two individuals that has been inherited from a common ancestor. In this example, two pairs of parasites from generation 0 recombine their haploid genomes during meiosis to produce four new parasites in generation 1. The IBD between a parent and its direct offspring is expected to be close to 0.5, assuming that the parents are not closely related (IBD = 0).

1.3 The resilience of *Plasmodium falciparum*

We describe here how thousands of years of evolution have shaped the host-pathogen interaction and identify key elements that are currently missing from our understanding of this interaction.

1.3.1 Interacting with the human host

Once *Plasmodium falciparum* merozoites enter the blood system of their human host, they gradually invade RBCs. The host has different strategies to get rid of the parasite, starting by the immune cells but also via the spleen. Indeed, iRBCs lose their deformability after the ring stage, which makes them subject to removal by the spleen that clears RBCs that are too rigid [83]. To avoid being immediately eliminated from the blood system, parasites have developed the ability to bind to host cells via exported proteins so that they can complete their cycle unworried by the spleen or the immune cells [84]. This cytoadherence has been associated mainly with PfEMP1 but also with other VSAs such as RIFIN and STEVOR. Some RIFINs are able to bind LILRB1 (leucocyte immunoglobulin-like receptor B1), a receptor at the surface of several immune cells, and induce a reduced antibody production [85]. STEVOR binding to host receptor Glycophorin C can enhance rosetting, a process in which *Plasmodium falciparum* iRBCs surround themselves with multiple uninfected RBCs, which could possibly protect from phagocytosis and splenic clearance [209, 210].

The binding of PfEMP1 with endothelial cell receptors ICAM-1 (intercellular adhesion molecule 1) and EPCR (endothelial protein C receptor) is responsible for the parasite sequestration in the microvasculature which helps parasites avoid splenic clearance and is associated with acute malaria symptoms potentially leading to cerebral malaria [211]. The binding of PfEMP1 to CD36 receptors can lead to various effects depending on the host cell type carrying the receptor: binding to macrophages can lead to parasite phagocytosis while binding to uninfected RBCs causes the formation of a rosette [211, 210]. All upsA and some upsB *var* genes encode PfEMP1 that preferentially binds EPCR and are generally observed in cases of severe malaria, including cerebral malaria [212, 213, 214, 215]. Group B and C *var* genes encode PfEMP1 types that tend to bind CD36 and induce rather asymptomatic malaria [216, 215]. Although both *var1csa* and *var2csa* are able to bind the placental chondroitin sulfate A (CSA) *in vitro*, only *var2csa* is more expressed in PAM cases [217, 218, 195].

The mutually exclusive expression of *var* genes enables the parasite to develop an antigenic variation over time which allows them to remain longer in the host by keeping at hand unseen antigens from the adaptive immune system (Figure 1.16). The mechanisms that

govern the expression of a single *var* gene and the silencing of all others involve the differential condensation states of chromatin at the nuclear periphery and histone modifications of the upstream region of *var* genes [86, 87]. The switching rate of about 2 % differs between *Plasmodium falciparum* strains and between *var* genes depending on their chromosomal location, with those internal more likely to be selected that those subtelomeric [88, 89, 90]. The choice of the next expressed *var* gene could be random or coordinated. Indeed, populations of parasites showed two expression patterns *in vitro*: either a single dominant *var* gene is expressed or many simultaneously [219, 220]. In this model, some genes such as *var2csa* would act as sink nodes thanks to the high competitiveness of their promoter, which could then favour *var* gene switching by making the bridge between two successive *var* genes [220]. However, even under the most optimistic models, the antigenic *var* switching process alone cannot explain how parasites can remain in the same host for several years without being cleared by the immune system [166]. One possibility is that the parasite is able to progressively reshuffle its antigenic repertoire size by ectopic recombinations of its *var* gene family and more generally its VSAs (Figure 1.16). Although this basic model is an oversimplification of highly complex host-parasite interaction, it provides a testable hypothesis on *Plasmodium falciparum* isolates derived from chronic infections.

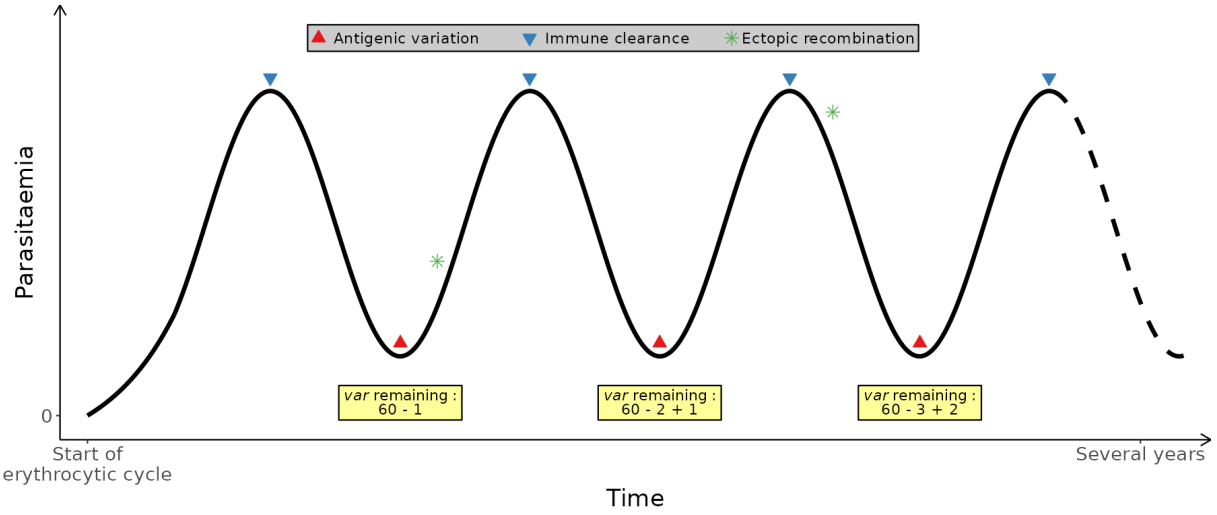


Figure 1.16: **Hypothetical example of antigenic variation in chronic infections**. After the first rise in parasitaemia, parasites are being cleared by the immune system that recognises the major PfEMP1 expressed at the surface of iRBCs. Before being completely cleared by the immune system, parasites undergo *var* gene switching which enables them to express a different non recognised PfEMP1 at the surface of iRBCs and the parasitaemia rises again. The new antigen is recognised and infected cells cleared until the cycle repeats for an extended period of time leading to a chronic infection. During the process, some parasites might undergo ectopic recombination of *var* genes which increase the size of the repertoire of available *var* genes unseen by the immune system and consequently the duration of infection.

Controlled human malaria infections (CHMI) studies are able to inspect the behaviour of the parasite *in vivo* by monitoring very frequently infection-related metrics such as symptoms, parasitaemia, host antibody levels or parasite antigenic levels in volunteers of different immune backgrounds in whom *Plasmodium falciparum* was inoculated [221]. One CHMI study showed a different distribution of *var* group expression related to host immunity, with more upsA *var* genes expressed in naïve individuals compared to semi-immune individuals able to control the parasitaemia, adding a new evidence of the role of upsA *var* genes in inducing a severe malaria [222].

1.3.2 Spatio-temporal connectivity

The incidence of *falciparum* malaria is highly heterogeneous throughout the world, with the African continent reporting most of the cases [12]. At a finer regional scale in Ghana or Tanzania for example, malaria prevalence can be heterogeneous with high transmission areas localised just a few kilometres away from low transmission areas [94, 95]. Several parasites from low endemic regions have shown high level of similarity with those from more endemic regions, like in East versus West Gambia, and this might explain how malaria is sustained in these regions [73, 96, 97]. In general, the level of similarity between malaria parasites tends to decrease with both the spatial and temporal distances that separate their sampling and this relationship seems to extend up to the household level [73, 98, 99, 100]. The parasite population connectivity can be summarised as a clonal propagation with a gradually decreasing level of similarity [101].

This parasite connectivity makes the idea of targeting only malaria hotspots appealing at first; this is the basis of the strategy of surveillance and containment that led to smallpox eradication [223]. Indeed, targeted approaches could rely on distribution of antimalarial drugs to members and neighbours of a household with an individual presenting to a health facility with malaria [224]. However, this is without considering the extent of asymptomatic malaria, whose cases are harder to diagnose and can present a different spatial distribution than that of symptomatic cases [225]. Furthermore, the location of hotspots in a region might vary between the different transmission seasons, which makes them even harder to predict [226]. Finally, targeting hotspots only might not always be the best strategy to achieve malaria elimination, as there is a lack of evidence on its efficacy, especially in high transmission settings [227, 228].

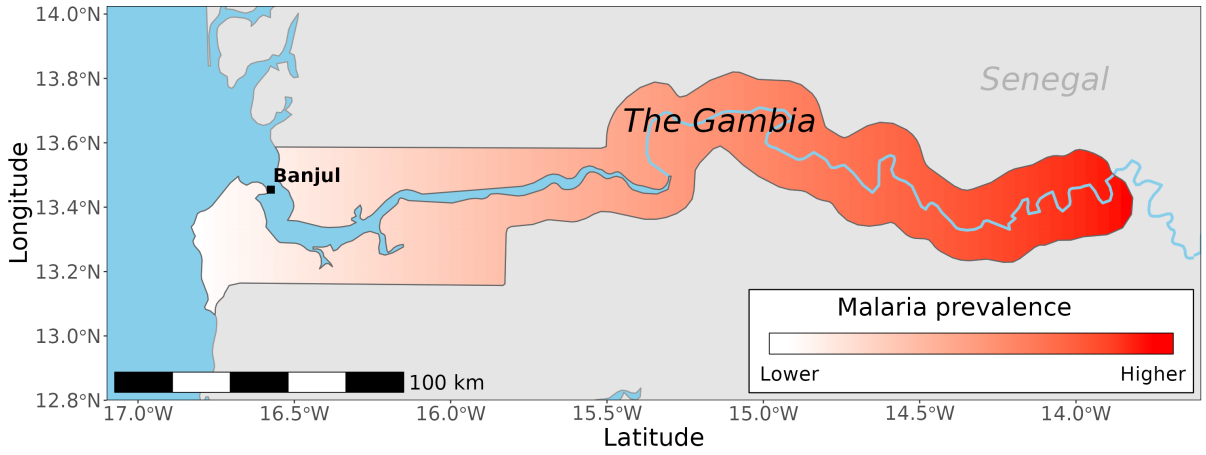
1.3.3 Climate conditions and seasonal malaria

The transmission of *Plasmodium falciparum* is modulated by climate conditions mostly because its vector, the anopheles mosquito, is adapted to only certain temperature and

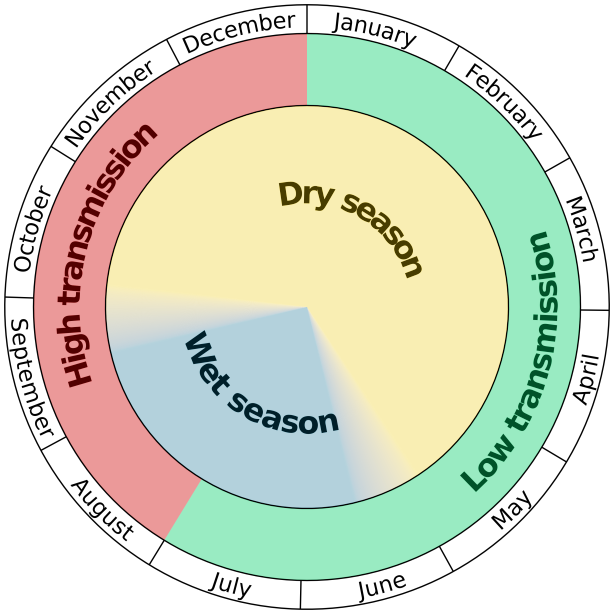
relative humidity levels and requires the presence of pockets of water to lay their eggs. The optimal temperature for malaria transmission differs between studies, but overall, they tend to agree that when below 15°C or above 35°C, parasites cannot invade properly mosquitoes which on the top of that do not survive such conditions [229, 230, 231]. Malaria transmission is positively correlated with relative humidity levels of above 50 to 60 % and with the intensity of rainfall [102, 103, 104, 105, 106]. The best correlation with the increase of malaria incidence is generally observed when these climate parameters lag a few months behind malaria incidence [232].

Temporary changing climate conditions are able to shift transmission hotspots from one region to another, making it difficult to plan the most efficient malaria elimination campaign [226]. At the global level, climate change is expected to actively reshape regional malaria transmission dynamics, which could lead to regions with a contraction or expansion of the transmission season, a shift later or earlier in the year, or even the emergence of malaria in currently unaffected regions [233, 234, 235].

In some malaria-endemic regions, climate conditions allowing malaria transmission are met throughout the entire year. However, other regions are hit by a seasonal malaria characterised by a high transmission during the wet season and at the beginning of the dry season but a very low to non-existent transmission for the rest of the year. In The Gambia, where malaria is more prevalent in the eastern part of the country, the wet season generally starts in June and ends in September, which translates into a high malaria transmission between August and December (Figure 1.17) [107, 108].



(a) The Gambia.



(b) Seasonality of malaria.

Figure 1.17: **Seasonality of malaria in The Gambia**. (a): The Gambia is a West African country where malaria greatly affects its eastern part every year during the high transmission season. (b): The high transmission season is lagging 2 months behind the wet season which lasts from June to September. For the majority of the year (around 7 months), there is little to no malaria transmission, yet every year, the country is hit by malaria again. Made with Natural Earth.

How parasites are able to remain prevalent in malaria-endemic regions with long-lasting low transmission seasons is not fully understood to this date. Indeed, during the low transmission season that last the majority of the year in some countries, around seven months for The Gambia, there are almost no symptomatic malaria infections reported. Contrary to symptomatic cases which are absent during low transmission seasons, the percentage of asymptomatic carriers is often constant throughout the year regardless of the

transmission season [109, 110, 111, 22]. Parasites sampled regularly across the different high transmission seasons are genetically close and display similar levels of clonalities, which implies that there is a genetic continuity of parasites even across long low transmission seasons [112, 113, 100, 114]. As asymptomatic infections can present high levels of gametocyte density which have the potential to infect mosquitoes, it is likely that they represent the main reservoir of parasites responsible for cases of symptomatic malaria during the high transmission seasons while imported parasites from other endemic regions contribute to a lesser extent [22, 115]. Furthermore, asymptomatic infections are often polyclonal with many distinct parasites circulating in the blood, which helps maintaining a high genetic diversity of the parasite population [24, 109, 116].

All these evidences support the fact that asymptomatic infections, while not directly harmful to their host, largely contribute to malaria burden and thus require specific attention in malaria elimination programs. However, predicting the duration of an asymptomatic infection, and consequently its transmission potential, may be possible from the parasite density at the end of the high transmission season, as it was positively correlated with the duration of *Plasmodium falciparum* carriage across the dry season [24]. Two main hypotheses have been proposed to explain how *Plasmodium falciparum* chronic infections persist in the host for the entire dry season. First, Andrade *et al.* suggests that parasites from the low transmission season exhibit a reduced vascular adhesion phenotype compared to those from high transmission seasons, making them more susceptible to clearance by the spleen [117]. Another possibility is that parasites from the low transmission season have such a low multiplication rate in the blood that they remain virtually invisible from the host immune system [118]. An aspect yet to be explored is the significance of chimeric *var* gene generation during the course of an infection in extending parasites presence in the blood. This increase in the *var* repertoire size might enhance their immune evasion capabilities, contributing to a longer duration of infection.

Aims of the Thesis

The goal of this thesis is to understand the genetic mechanisms enabling the parasite *Plasmodium falciparum* to establish an asymptomatic chronic infection.

Project 1

The primary objective of this project is to elucidate the variations in transmission dynamics within a wet-dry season climate by analysing genetic relatedness. Hundreds of blood samples were collected for inhabitants of 4 nearby villages in the Upper River Region of The Gambia between 2014 and 2017. We will employ genotyping and whole genome sequencing techniques to assess the relatedness among parasites sampled across different spatial and temporal distances. Additionally, we will determine the minimum duration of infections caused by the same parasite genotype and utilize shared genetic fragments' locations to infer recombinatorial histories at various scales.

Project 2

We aim to characterise the impact of long-lasting chronic infections on *Plasmodium falciparum* genome, including its most polymorphic gene family, *var* genes. We will be using a combination of long-read and single-cell short-read sequencing data from parasite isolates obtained from asymptomatic individuals with chronic infections, spanning time intervals of up to 6 months. The long-read genome assemblies to be generated will encompass the entire genome of each individual infection, including *var* genes. Additionally, our single-cell sequencing approach will allow us to identify exceptionally rare variants that might be overlooked in bulk sequencing. Initially, we will compare uncultured single-cell genomes from two time points separated by several months within the same infection, aiming to identify mutations conferring a selective advantage. Subsequently, we will compare the uncultured single-cell genomes from later time points with the complete genome assembly obtained from the same infection at an earlier stage. This comparison will help us identify putative structural variants involving *var* genes.

Collaborative projects

During my thesis, I could collaborate with various teams from within the same laboratory I was hosted in, or from other laboratories with common interests in *Plasmodium falciparum* malaria. A thorough description of each collaborative project is provided in the appendices (*Appendix C Collaborative Projects*).

Mutation rates: During this project in collaboration with William L. Hamilton[1], I performed an in-depth analysis of 254 *de novo* micro-insertions/deletions (INDEL) called from 354 whole genome sequencing of parasites cultivated from 5 distinct laboratory-adapted strains. I was involved in the co-supervision of a master's student, Aakanksha Singh[2], who used and improved my pipeline of variant calling to identify *de novo* micro-INDELs. My responsibility included pinpointing the locations and insertion/deletion biases associated with these *de novo* micro-INDELs (*Appendix C.1 Mutation rates*).

Identification of a drug resistant marker: A team led by Rachel Cerdan[2] cultivated thirteen 3D7 strains in the presence of a newly synthetic antimalarial compound and one other 3D7 strain in the absence of any antimalarial compound, serving as a negative control. Over time, the thirteen 3D7 lines acquired different phenotypes of resistance with varying degrees of sensitivity to the synthetic drug. My role was to discover mutations potentially associated with the observed drug resistance phenotype. This involved pinpointing variants present in resistant strains while absent from the sensitive one (*Appendix C.2 Identification of a drug resistant marker*).

Duration of asymptomatic infections: In a published collaborative work of Collins *et al.*, multiple blood samples of 42 individuals were genotyped with MSP2 fragment sizes during the dry season 2016/2017 of The Gambia in four nearby villages [24]. My role was to estimate the number of *Plasmodium falciparum* genotypes present in each blood isolate and to follow, for each individual, the resurgence of parasite genotypes over time by grouping those with the same MSP2 fragment length (*Appendix C.3 Duration of asymptomatic infections*).

Risk of clinical malaria: I reviewed the work of Fogang *et al.*, who were interested in following malaria infection status and parasitaemia from individuals that were tested in four nearby villages of The Gambia between 2014 and 2016 (*Appendix C.4 Risk of clinical malaria*) [119].

DBLα diversity: I published with Antoine Claessens[2] a perspective article focused on the work of Tonkin-Hill *et al.* who used the DBLα tag, a variable region of the DBLα domain surrounded by conserved motifs, to describe the genetic diversity of *Plasmodium falciparum* at the country level (*Appendix C.5 DBLα diversity*) [69, 236].

[1]Malaria Programme, Wellcome Sanger Institute, Hinxton, United Kingdom
[2]LPHI, MIVEGEC, Université de Montpellier, CNRS, INSERM, Montpellier, France

Project 1: Spatio-temporal Relatedness of Parasites

2.1 Introduction

The Gambia has achieved a more substantial reduction of *Plasmodium falciparum* malaria prevalence than most other countries in Africa [237]. The climate is characterised by a wet season from late June to September, followed by a 8 month-long dry season, typically without rainfall. Malaria transmission and cases peak from September to November. The prevalence of the infection is highest in the eastern part of the country, and the persisting *Plasmodium falciparum* reservoir during the dry season is thought to be due to asymptomatic chronic infections [91, 107]. Previous studies in the area between 2012 and 2016 reported that roughly half of asymptomatic infected individuals at the end of the wet season were still infected at the end of the dry season, showing that these infections can establish a chronic infection for an extended period of time [238, 24]. Importantly, asymptomatic infections may carry gametocytes that have been shown experimentally to transmit to mosquitoes [22, 118]. In Mali, where malaria is also seasonal, parasites isolated from long-lasting chronic infections in the dry season are on average older in terms of hours post invasion, indicating a reduced cytoadherence phenotype [117].

However, little is known about how the dry season affects the parasite population at the genetic level. A crude measure of the parasite prevalence can be inferred from the average complexity of infection (COI, the number of distinct parasite genotypes within an infection) in the population, as COI and transmission are correlated [239, 240, 241, 63, 101]. We previously showed that the COI, estimated by MSP2 genotyping of 42 individuals followed monthly during a single dry season in The Gambia, was also positively associated with both parasitaemia and duration of infection [24].

More accurately, the analysis of genetic relatedness between parasites can shed light on fine-scale spatio-temporal transmission dynamics. Genetic relatedness is obtained from comparing pairs of parasite genotypes which can be accessed by whole genome sequencing (WGS) or the cost-effective barcode genotyping which can identify a unique parasite strain with as few as 24 loci [71]. From pairwise distances between such barcodes, it has been shown that relatedness tends to decrease with time and distance of the sampled infections both at the national level in The Gambia and at the local level in neighbouring Senegal [73, 101]. Although the parasite population genetic diversity has been well characterized from infections isolated within a single wet season, the impact of the low transmission intensity dry seasons on the long-term genetic diversity of the parasite populations is still poorly understood.

Identity by descent (IBD) can be used as a metric of the genetic relatedness between parasites that not only considers the pairwise distances but also incorporates the organism recombination rate. While studies based on F_{ST} are well adapted to characterize the haplotype diversity of populations of parasites across countries or for a long period of time, IBD has the advantage of highlighting evidences of a recent recombination between two parasites with varying levels of relatedness [74, 242, 243]. Indeed, an IBD analysis can provide the overall relatedness of two genomes and, with a sufficient number of loci typically obtained from WGS data, the more precise location of related genomic fragments within chromosomes [98, 60]. To date, most genomic epidemiology studies have only sampled clinical cases, yet such symptomatic cases only represent a minority of all *Plasmodium falciparum* infections [244, 245]. Whether the *Plasmodium falciparum* allelic diversity observed from clinical cases is representative of the total parasite population is unknown.

Here, we genotyped *Plasmodium falciparum* isolates from a longitudinal study in four nearby villages in the Upper River Region of The Gambia from 2014 to 2017 already reported elsewhere [24, 119]. Blood samples from 1505 participants over 16 time points were collected during both wet and dry seasons. In total, 442 *Plasmodium falciparum* positive samples from asymptomatic infections were successfully genotyped by sequencing 89 good quality SNPs, and 334 were whole genome sequenced. We used the IBD to estimate the parasite genetic diversity throughout the study period.

2.2 Material and methods

2.2.1 Study design and participants

Starting in December 2014, we recruited all residents from two villages (Madina Samako and Njayel, identified respectively with the letters 'K' and 'J'), with two additional villages (Sendebu and Karandaba, identified respectively with the letters 'P' and 'N') recruited from July 2016, all four villages being in the Upper River Region in The Gambia within 5 km of each other. More information about the recruited participants can be found in a previous study [119]. In December 2016, a cohort of 42 asymptomatic *Plasmodium falciparum* carriers was recruited and sampled monthly for 6 months, as it was previously reported [24].

2.2.2 Sampling and molecular detection of parasites

The full description of DNA extraction of parasites is provided in a previous work [119]. Briefly, fingerpricks blood samples were collected and tested for *Plasmodium falciparum*. From July 2016 onwards, individuals testing positive for *Plasmodium falciparum* were invited to provide an additional 5 to 8 mL venous blood sample before being treated with CoArtem. The venous blood samples were leucodepleted with cellulose-based columns and frozen immediately. DNA was extracted with QIAgen Miniprep kit following manufacturer procedure.

2.2.3 Genotyping and sequencing

All *Plasmodium falciparum* DNA positive samples, from fingerprick and from venous blood, were genotyped using the mass-spectrometry based platform from Agena, as part the MalariaSpot consortium [123]. A total of 101 SNPs located on the 14 chromosomes were genotyped and merged into a 'molecular barcode'. SNPs, all biallelic, were chosen for their usefulness in analyses of relationship between parasites. Out of 442 genotyped samples, 334 were also whole genome sequenced (Illumina) with a selective whole genome amplification (sWGA) step [120]. Paired-end DNA sequencing reads (150 bp) were aligned to 3D7 reference genome version 3. Variants were called by a script from the MalariaGen consortium using GATK HaplotypeCaller [121, 122]. Six markers of resistance to multiple antimalarials (amodiaquine, artemisinin, chloroquine, lumefantrine, mefloquine, pyrimethamine, sulfadoxine) were obtained by both genotyping and sequencing [123]. These markers correspond to 6 non-synonymous changes in 6 proteins encoded by the genes: *aat1* (PF3D7_0629500) S528L, *crt* (PF3D7_0709000) K76T, *dhfr* (PF3D7_0417200) S108N, *dhps* (PF3D7_0810800) A437G, *kelch13* (PF3D7_1343700) C580Y and *mdr1* (PF3D7_0523000) N86Y,

each of which is known to reduce the susceptibility to one or multiple antimalarial drugs. Two markers, *aat1* S528L and *kelch13* C580Y were only available in sequencing data and absent from genotyping data.

2.2.4 Parasite relatedness

To accurately assess the parasite genetic similarity between different sampled infections, we estimated pairwise mean posterior probabilities of identity by descent (IBD) between genomes or barcodes using hmmIBD, a hidden Markov model-based software relying on meiotic recombination events given a recombination rate of *Plasmodium falciparum* of 13.5 kb/cM [79, 41]. The probability of two samples to be in IBD represents the expected shared fraction of their genomes. An IBD of more than 0.9 is considered identical, hence describing the same parasite genotype.

2.2.5 Complexity of infections

The clonality of each isolate was estimated from whole genome sequenced samples by the F_{WS} metric based on allelic frequencies [82]. Additionally, the complexity of infection was estimated by the categorical method of THE REAL McCOIL ('maxCOI = 30, threshold_ind = 0, threshold_site = 0, totalrun = 10000, burnin = 1000, M0 = 15, e1 = 0.01, e2 = 0.01') from consensus barcodes [124]. To evaluate the correlation between F_{WS} and THE REAL McCOIL, a genomic barcode generated from WGS filtered SNPs was used as an input for THE REAL McCOIL. A locus was considered mixed if the within-sample minor allele frequency was above 0.05, while loci with missing data from more than 20 % of samples were filtered out. To minimise the likelihood of linkage disequilibrium between nearby loci within these sets, only loci that were more than 5 kb apart were kept for both mass-spectrometry SNP genotyping (44 remaining loci) and WGS filtered SNP barcodes (819 remaining loci) [246]. If the clonality estimated by THE REAL McCOIL was too uncertain (with a 95 % confidence interval bigger than 2), the sample was excluded (1 consensus barcode excluded).

2.2.6 Multi-locus genotype barcode data analysis pipeline

To analyse whole genome data formatted in a VCF format, we developed the following pipeline to (Figure 2.1b):

1. Filter out genomic positions for which 'QUAL' is inferior to 10000, with more than 2 alleles identified in the population or that are located outside of the core genome [41].

2. Remove genomes comprising less than 4000 SNPs covered by at least

5 reads.

3. Estimate the proportion of polyclonal samples using the F_{WS} metric and THE REAL McCOIL.

4. Merge the SNPs into genomic barcodes and format them to a binary matrix as required by hmmIBD (mixed and unknown positions set to 0). SNP calls were considered mixed if the within-sample minor allele frequency (MAF) was greater than 0.2. The MAF of 0.2 was chosen according to the good agreement between molecular barcodes and genomic barcodes (high number of mixed locus matches and low number of mixed locus mismatches) (Figure A.1).

5. Use the paired IBD values obtained from hmmIBD and build a network of genome relatedness with pairs of genomes having more than 100 informative loci. A locus is deemed informative when it is available in both genomes and that at least one of them is the minor allele.

6. Identify lineages by grouping genomes related with an IBD between 0.35 and 0.65. The 'parents' are defined as two genetically unrelated genomes (IBD < 0.2) in high IBD with 'offspring' genomes. The latter shared at least 35 % of their genome with each parent.

The second part of the pipeline imputes missing SNPs in the molecular barcode from the WGS data to build a 'consensus barcode'. Then, it estimates the duration of infection by the same genotype (Figure 2.1):

1. From the initial 101 genotyped SNPs, remove positions not present in filtered WGS SNPs data.

2. Replace unknown positions of molecular barcode with filtered WGS SNPs, which generates what we refer to as 'consensus barcodes'. The comparison between barcodes and WGS SNPs showed a discrepancy for 21 of the 101 SNPs for samples obtained after May 2016 (Figure A.2). As a result, these 21 barcode SNPs were considered unknown for all the samples obtained after May 2016.

3. Remove consensus barcodes with fewer than 21 SNPs.

4. Estimate the proportion of polyclonal samples using the THE REAL McCOIL.

5. Format consensus barcodes into a binary matrix (mixed and unknown positions set to 0) and run hmmIBD.

6. Use the paired IBD values obtained from hmmIBD and build a network of barcode relatedness with pairs of consensus barcodes having more than 10 informative positions. A locus is deemed informative when it is available in both barcodes and that at least one of them is the minor allele.

2.2.7 Genetic relatedness between groups of multi-locus genotype barcodes

All 'consensus barcodes', hereafter referred to as 'barcodes', available after December 2016 were excluded from the genetic relatedness analysis as they were obtained exclusively from the 42 individuals recruited in the cohort from the December 2016 screening [24]. Barcodes were grouped by their collection date (11 time points from December 2014 to December 2016), sampling location (same household, different households from the same village and different households from different villages) or collection date split by sampling location. For each pair of groups of barcodes, the genetic relatedness was estimated by the proportion of related barcodes (IBD > 0.5) over all possible combinations of barcodes excluding those sampled from the same individual. Pairs of sampling location with less than 5 comparisons (941/2933 removed pairs) and pairs of collection dates split by sampling location with less than 5 comparisons (25/193 removed pairs) were filtered out. All pairs of sampling locations contained at least 10 comparisons.

2.2.8 Most conserved genomic regions

The genomic fragments in IBD between the 19700 pairs of non-identical genomes (IBD < 0.5) were combined to obtain, for each position, the number of genomes in IBD. The top 5 % most covered regions (with at least 440/19700 pairs of genomes in IBD or 2.2 % of all pairs) were extracted and merged if they were less than 10 kb apart.

2.2.9 Natural cross offspring identification

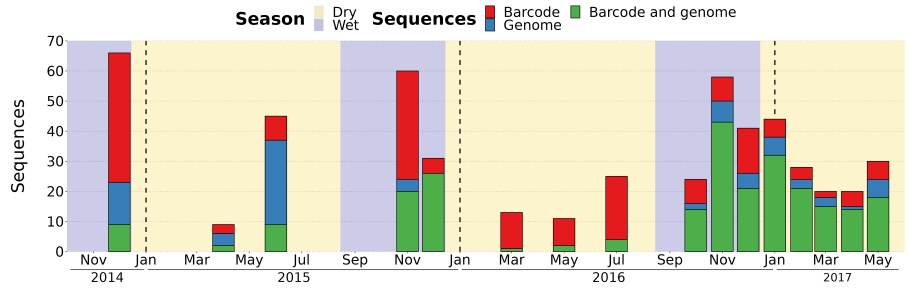
Highly related genomes (IBD > 0.9) and slightly related genomes (IBD between 0.35 and 0.65) were grouped into clusters (corresponding to components in graph theory) using igraph (version 1.2.11) R package [125, 126]. Each cluster may contain identical parasites (IBD > 0.9) and genetically related parasites separated by a recent recombination event from a natural cross (IBD between 0.35 and 0.65). Within each cluster, three genomes were associated in

a 'triad' when one of them, assumed to be an offspring, was related to two other genomes assumed to be parents (IBD between parents < 0.2). The two parental genomes picked were those that, if combined, maximized the total size of chromosomal fragments in IBD. Putative offsprings that had less than 85 % of the accessible parts of their genome in IBD with their combined parents were excluded.

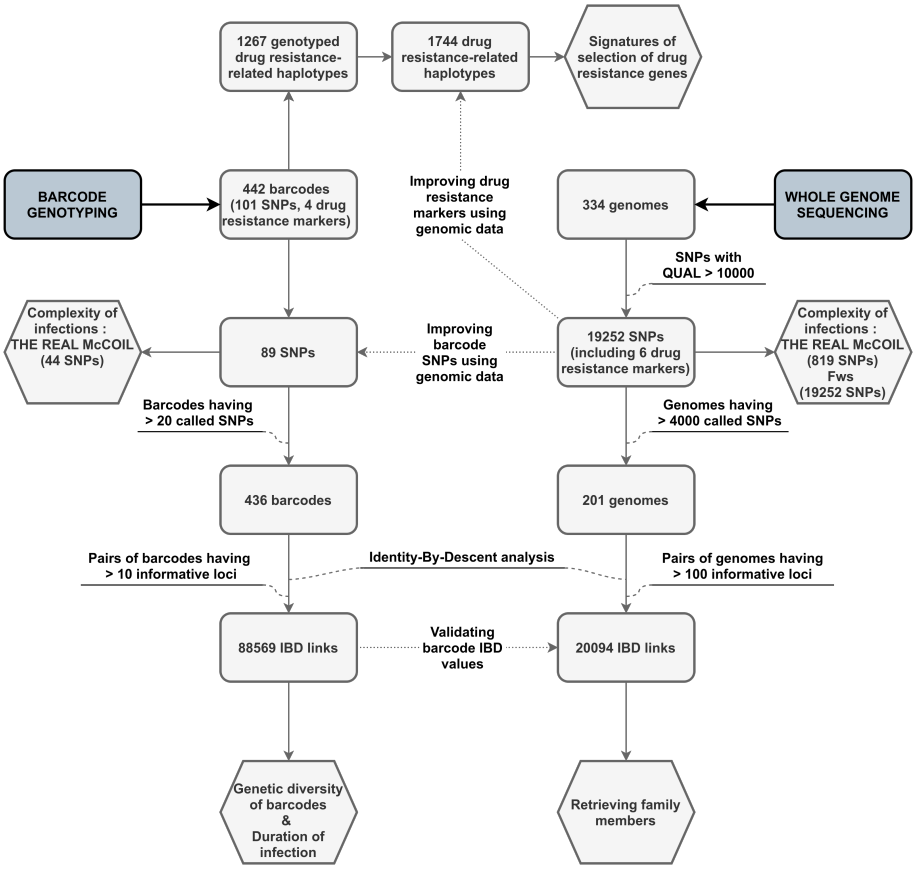
2.3 Results

2.3.1 Combined barcode and whole genome analysis pipeline

To characterise the parasite genetic diversity and identify the impact of antimalarial drugs, all *Plasmodium falciparum* isolates were genotyped and whole genome sequenced (Figure 2.1a). After concatenation of the two datasets, we obtained a high-quality genomic barcode comprising 89 SNPs for 436 isolates (Figure 2.1b). Out of these, whole parasite genomes were also successfully sequenced for 201 isolates, with 19252 filtered SNPs called. The barcode data analysis pipeline is freely available at https://github.com/marcguery/malaria-barcodes-analysis.



(a) Study design.



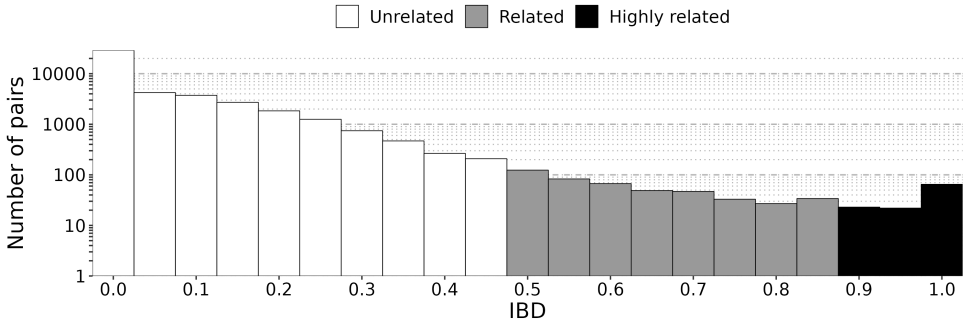
(b) Analysis pipeline.

Figure 2.1: **Study design and analysis pipeline**. (a): Regular blood samplings successfully genotyped and/or whole genome sequenced before being processed in the analysis pipeline (https://github.com/marcguery/malaria-barcodes-analysis). (b): Barcodes of 89 high-quality SNPs spanning 14 chromosomes were generated for 436 samples for spatio-temporal analysis of *Plasmodium falciparum* genetic diversity. In parallel, 201 high quality (> 4000 called SNPs) *Plasmodium falciparum* genomes were sequenced to validate barcodes quality, measure IBD values and identify natural crosses and potential clonal lineage expansions. Additionally, 6 drug resistance markers genotyped in 431 samples (4/6 drug resistance markers available, 1267 haplotypes in total) and called in 203 *Plasmodium falciparum* genomes (6/6 drug resistance markers available, 1022 haplotypes in total), leading to a merged total of 1744 drug resistance-related haplotypes, were used to determine the prevalence of resistant haplotypes over time and to estimate the effect of drug resistance on genomic fragment relatedness. Finally, complexity of infections (number of distinct *Plasmodium falciparum* clones from the same blood sample) were estimated using F_{WS} metric on all genomic SNPs and THE REAL McCOIL on unlinked SNPs from both genotyping and sequencing data.

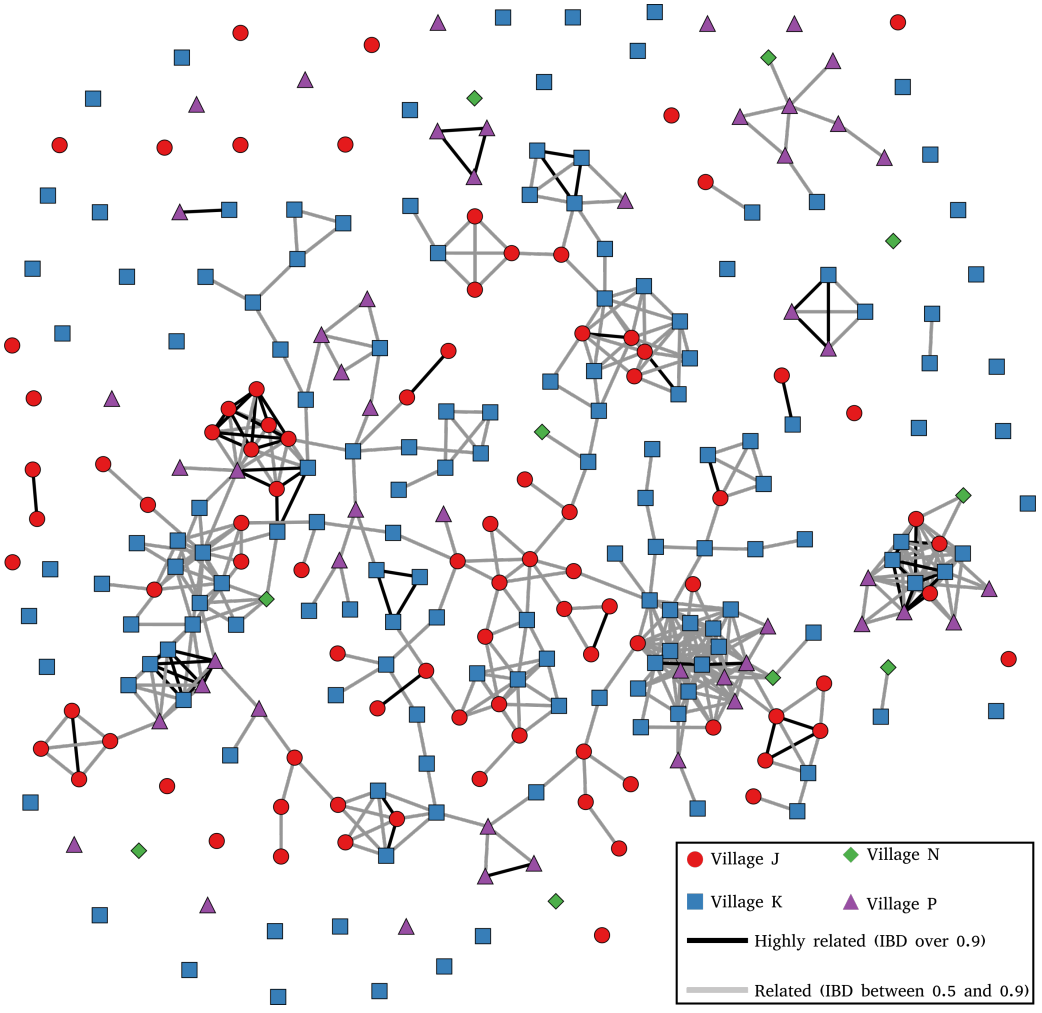
2.3.2 High genetic complexity in asymptomatic infections

One indicator of parasite strain diversity is the level of complexity (or multiplicity) of infection (COI), denoting the number of unique genotypes/genomes within an isolate. We estimated COI using F_{WS} values from genomes and THE REAL McCOIL from both barcodes and genomes (Figure A.3). The population minor allele frequency estimated by THE REAL McCOIL on both barcodes and genomes showed a good correlation with the true population minor allele frequency calculated from allelic depths (Figure A.4). Overall, the proportions of isolates with polygenotype infections were estimated as 39 % using F_{WS} (79/201), 23 % using THE REAL McCOIL on genomes (47/201) and 9 % using THE REAL McCOIL on barcodes (41/435) and were evenly distributed across all time points, meaning that the average level of genetic complexity of infections remained stable in the area over the 2.5 years of the study (Figure A.5).

To determine the relatedness between isolates, identity by descent was calculated pairwise between all barcodes and all genomes. The accuracy of the IBD calculated from barcode data was assessed using IBD calculated from WGS data as the gold standard (Figure A.6). IBDs from barcode and WGS data showed a strong linear correlation when both were above 0.5 (R^2 = 0.76); this cut-off was chosen to discern between related and unrelated pairs of samples. Out of 88569 pairs of infection samples, only 1301 were related with an IBD above 0.5 (1.5 %), demonstrating a very large recombinatorial genetic diversity throughout the parasite population (Figure 2.2a). Out of the 294 barcodes sampled at least once in each individual (barcodes within an individual are considered identical if they have an IBD above 0.9), 68 (23 %) were identical (IBD > 0.9) between different individuals while 229 (78 %) were related (IBD > 0.5) between different individuals (Figure 2.2b). This result is consistent with a longitudinal clinical study from Senegal in which the proportions of repeated barcodes ranged from 10 to 55 % from 2006 to 2013 [60]. There is no apparent evidence of genetically isolated strains at the village level as all villages are interconnected.



(a) Distribution of IBD values between barcodes.



(b) Barcode IBD relatedness network.

Figure 2.2: **High parasite recombinatorial diversity in 4 villages in The Gambia inferred from inter-individual genetic relatedness**. (a): Log-scale distribution of IBD values between barcodes. (b): Relatedness network of parasites genotyped between December 2014 and December 2016, with barcodes represented as nodes and IBD values represented as edges. Sets of nearly identical barcodes (IBD > 0.9) within a volunteer were merged into a single one by keeping the earliest genotyped barcode, resulting in one barcode per continuous infection represented here as a node. Barcodes are grouped into clusters using the compound spring embedder layout algorithm from Cytoscape (version 3.9.0) [247].

2.3.3 Pattern of relatedness between infections is shaped by seasonality

The spatio-temporal relationship between pairs of barcodes (only between distinct individuals to exclude chronic infections) was investigated. To do so, the proportion of related barcodes (IBD > 0.5) was calculated for all pairs of time points split by their sampling locations (same household, different households from the same village, or different households from different villages) and grouped by time between sample collections (Figure 2.3). The average proportion of similar barcodes is ten times lower between barcodes sampled less than two months apart and those sampled more than 12 months apart (average proportions of 0.048 and 0.0043, Welch t-test value = 4.8979, p value < 0.0001). At the spatial level, barcodes sampled less than 2 months apart and from the same household were twice more related than those from different households of the same village (average proportions of 0.096 and 0.042, Welch t-test value = 2.3471, p value = 0.033) and seven times more related than those sampled from different villages (average proportions of 0.096 and 0.014, Welch t-test value = 4.0519, p value = 0.0025). When barcodes were sampled more than two months apart, the correlation of genetic relatedness with sampling location disappeared. Altogether, the overall large parasite recombinatorial genetic diversity combined with the increased proportion of related isolates within the same household indicate a scenario in which the same infectious mosquito has infected two or more individuals who live together, or a direct transmission chain between two household members.

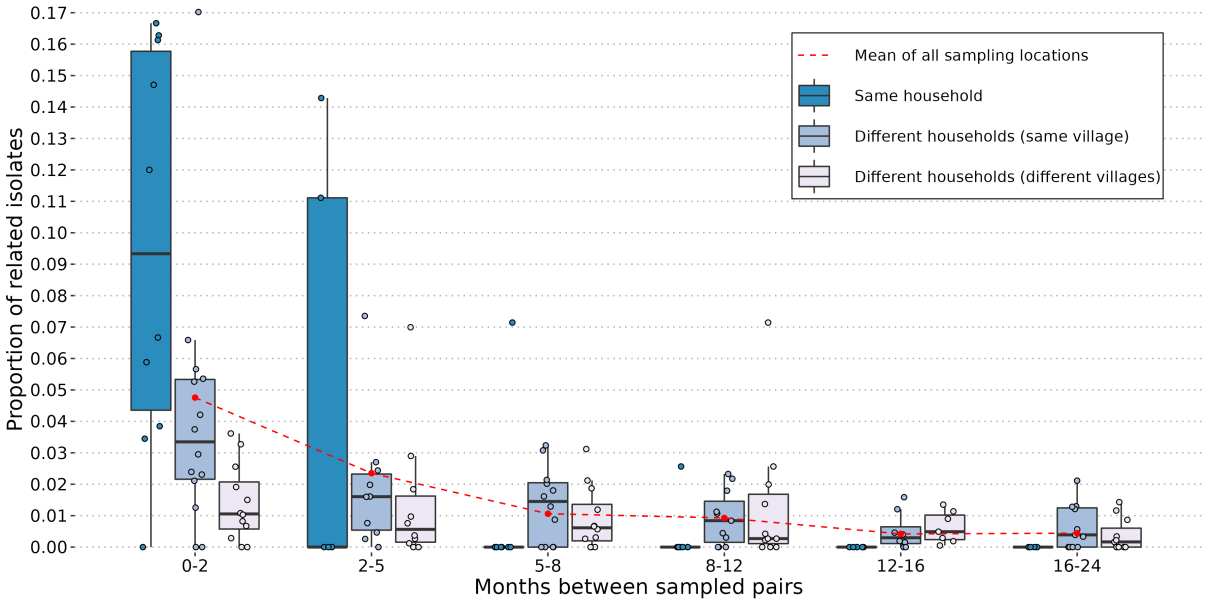
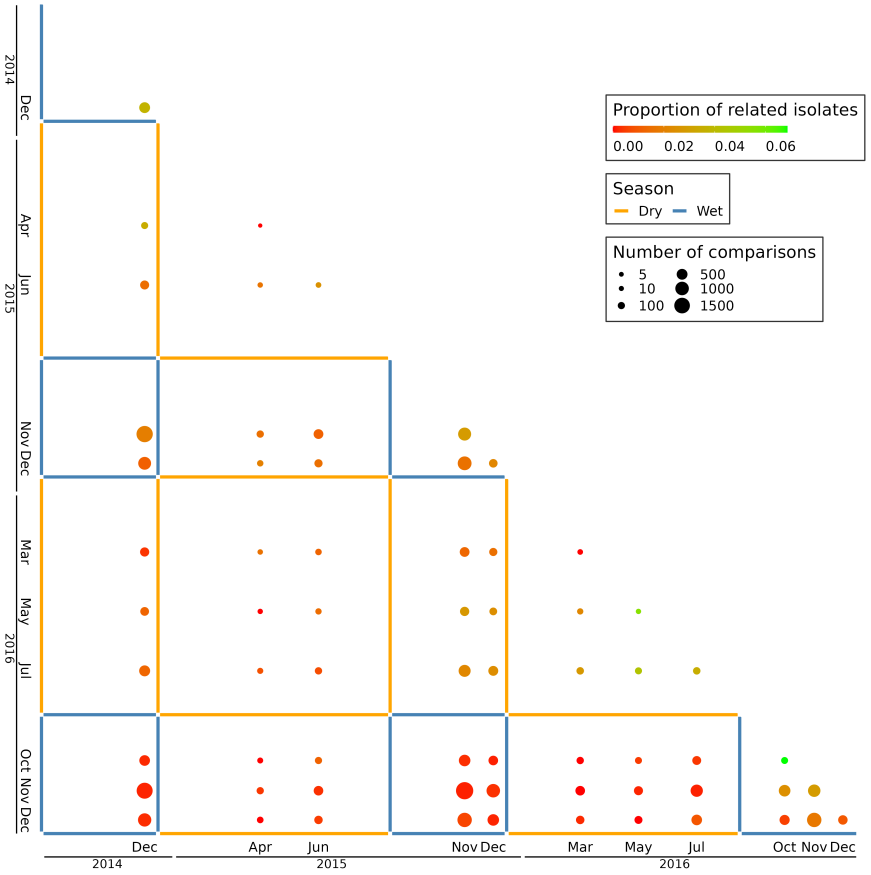


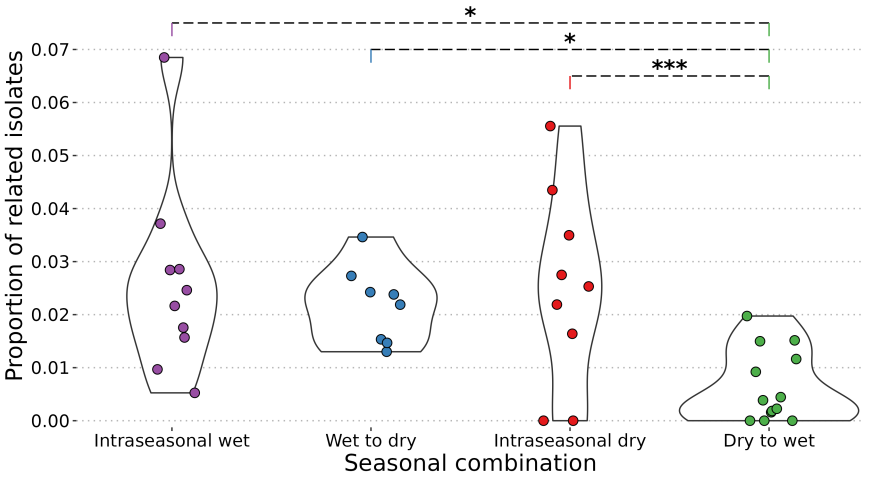
Figure 2.3: **Combined effects of spatial and temporal distances on parasite relatedness**. The proportion of related isolates between each pair of households is grouped according to their relative spatial distance and binned into time intervals of various lengths such that the number of observations in each bin is similar. The mean proportion of related barcodes for each time interval is shown as a dotted red line.

The impact of seasonality on the parasite recombinatorial genetic diversity was assessed by comparing the proportion of related barcodes (IBD > 0.5) between groups of collection dates within or between seasons. Barcodes sampled during the wet seasons from 2014, 2015 and 2016, as well as the dry seasons of 2015 and 2016, were grouped into intraseasonal pairs if they were collected during the same season and interseasonal pairs otherwise. This resulted in 5 groups for intraseasonal pairs and ten groups for interseasonal pairs, with the most distant groups (wet 2014 and wet 2016) being 4 seasons apart. As shown in Figure 2.3, pairs of collection dates relatively close in time exhibited greater similarity than more distant collection dates (Figure 2.4a). This suggests a continuous recombination process among all parasites, rather than the transmission of one or more specific strains. To determine the specific time of the year the average genetic relatedness declines, the proportion of related barcodes was compared between pairs of sample collections from the exact same season or exactly one season apart (Figure 2.4). Parasites belonging to the 'wet intraseasonal', 'dry intraseasonal' and 'wet to dry' groups displayed similar genetic relatedness with average proportions of related barcodes of 0.026, 0.025 and 0.022 respectively. However, the 'dry to wet' group exhibited a 3-fold lower genetic relatedness, with an average proportion of related barcodes of 0.007, indicating that the majority of parasite differentiation occurs during the transition from the dry season to the subsequent wet season. This corresponds to

the increase in transmission rate at the onset of the wet season, with parasite genetic diversity being reshuffled after sexual reproduction in the mosquito. This increase in genetic diversity is not observed in the 'wet to dry' season transition, demonstrating a lack of transmission in the dry season.



(a) All pairs of dates.



(b) Closest pairs of dates.

Figure 2.4: **Effect of seasonality on parasite recombinatorial genetic diversity**. (a): Proportion of similar barcodes (IBD > 0.5) between all pairs of sample collection dates from December 2014 to December 2016. (b): Proportion of similar barcodes (IBD > 0.5) between the closest pairs of sample collection dates that are within the same season ('intraseasonal wet' and 'intraseasonal dry') and one season apart when the wet season precedes the dry season and conversely (respectively 'wet to dry' and 'dry to wet'). Genetic similarities were compared between the 'dry to wet' group and all other groups with Welch t-tests (*: p value < 0.05, ***: p value < 0.0005).

2.3.4 *Plasmodium falciparum* chronic infections with persisting genotypes

To differentiate 'true' chronic infections with the same parasite genotype from reinfections, we measured the minimal duration of infection using IBD values between barcodes obtained from the same individual sampled at different time points, which were all ignored for the spatio-temporal analysis of genetic relatedness (*Section 2.3.3 Pattern of relatedness between infections is shaped by seasonality*). The beginning and end of an infection by the same parasite were attributed to the two most further away time points separating two related barcodes (IBD > 0.5). Forty individuals were infected with the same dominant *Plasmodium falciparum* genotype for at least two months and up to one year and a half (Figure 2.5). Within this group of chronically infected individuals, 26 were male and 14 females (Figure A.7). However the duration of infection did not significantly differ between gender nor age groups (Figure A.7).

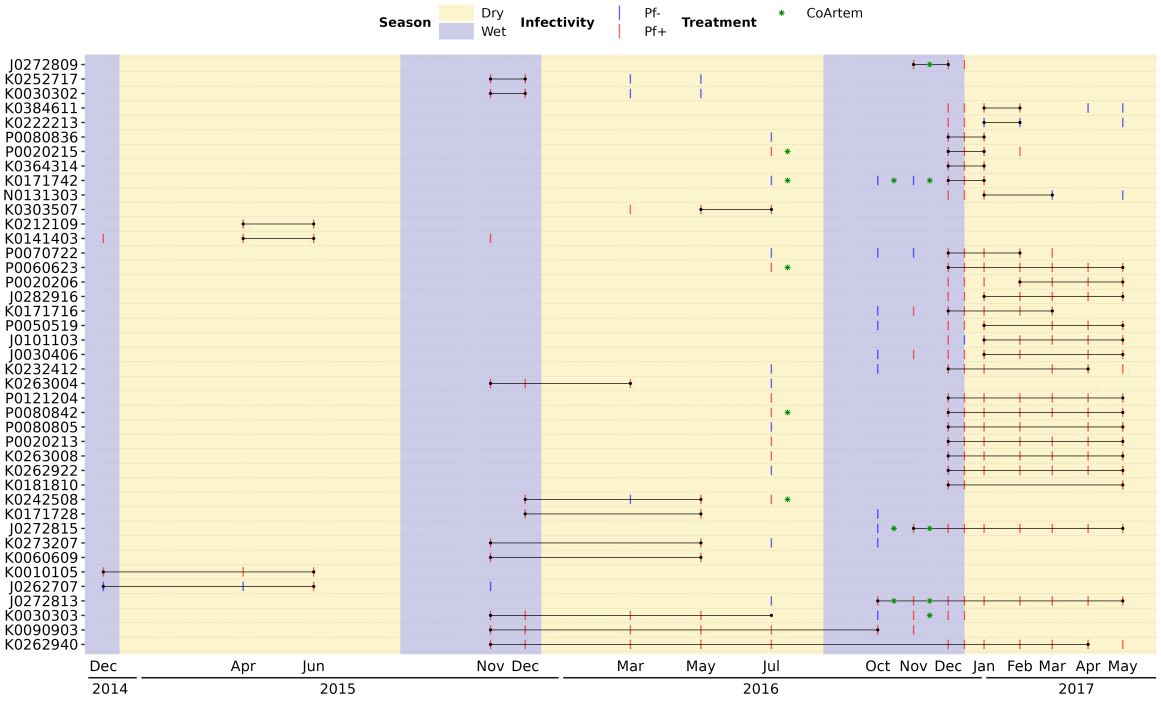


Figure 2.5: **Continuous *Plasmodium falciparum* infections with the same dominant genotype**. Forty individuals were infected with related barcodes (IBD > 0.5) at two or more time points. Duration of continuous infections are ranked from the shortest (top: J0272809) to the longest (bottom: K0262940), and linked with a black line. Unlinked *Plasmodium falciparum* positive tick marks indicate a different barcode (IBD < 0.5) or barcode not available. Some individuals received occasionally an antimalarial treatment.

2.3.5 Independence of seasonality and drug resistance prevalence

Six drug resistance-related haplotypes were obtained by molecular genotyping and called from whole genome sequencing in genes *aat1*, *crt*, *dhfr*, *dhps*, *kelch13* and *mdr1* respectively for 431 barcodes (1267 halpotypes) and 203 genomes (1022 haplotypes) for a total of 2289 haplotypes. The comparison of the 545 haplotypes that were obtained with both molecular genotyping and WGS on the same samples showed that the two methods had a very good agreement with 476 (86 %) haplotypes matching (Figure A.8). For the remaining 77 haplotypes that did not match, only the call of WGS was considered for the rest of the analysis. Overall, the percentage of isolates with resistant alleles was stable over time for 6 haplotypes with 91 % for *aat1* S258L, 91 % for *dhfr* S108N (pyrimethamine resistance), 47 % for *dhps* A437G (sulfadoxine resistance), 0 % for *kelch13* C580Y (artemisinin resistance) and 12 % for *mdr1* N86Y (multi-drug resistance) (Figure 2.6). The remaining haplotype, *crt* K76T (chloroquine resistance), had a constant prevalence with 51 % of resistant parasites between December 2014 and July 2016 which then increased and remained stable at 73 % afterwards (Figure 2.6). These prevalences are on par with the 2008 estimates in the country [248]. Although it was previously reported that the prevalence of mutations associated with chloroquine resistance declined during the dry season in The Gambia [249], here no drug resistance allele frequency consistently decreased nor increased across the six seasons from 2014 to 2017 (Figure 2.6). In our dataset, there is no evidence for seasonality-related drug resistance fitness cost that could impact the population haplotype diversity.

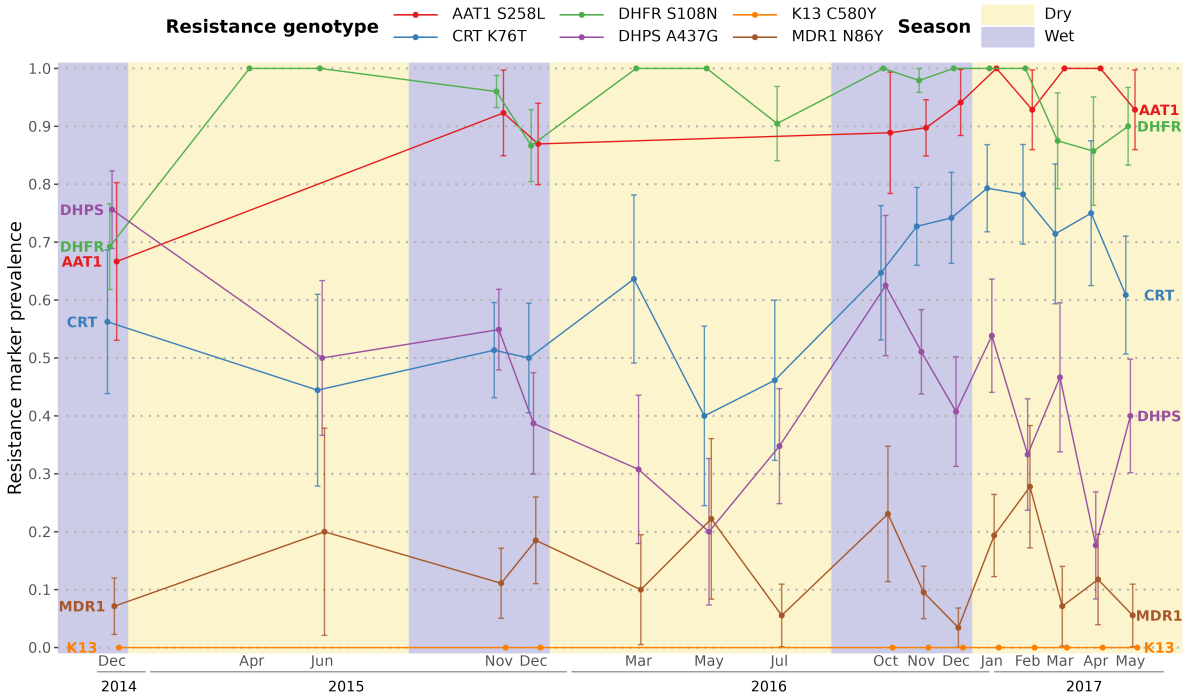


Figure 2.6: **Prevalence with standard errors of 6 drug resistance-related haplotypes for each time point with at least 5 observations over the study time period**. The 6 variants induce non-synonymous changes in *aat1* (S528L), *crt* (K76T), *dhfr* (S108N), *dhps* (A437G), *kelch13* (C580Y) and *mdr1* (N86Y) which are known to reduce the susceptibility to multiple antimalarials. The proportions of all variants except *crt* K76T remain stable over time with different degrees of prevalence. Two variants are almost fixed in the population of parasites: 91 % of parasites are *aat1* S528L mutants, 91 % *dhfr* S108N mutants and one is absent: no parasite is a *kelch13* C580Y mutant. The prevalence of variants *dhps* A437G and *mdr1* N86Y oscillates around the same value over time (respectively 0.47 and 0.12) while the prevalence of variant *crt* K76T went from an average of 0.51 to 0.73 after July 2016.

2.3.6 Signature of selection around drug-resistance markers

To investigate genome-wide signatures of selection, IBD values between all pairs of the 201 genomes were calculated along with the location of the shared chromosomal fragments. The locations of the most conserved fragments in the population were identified by overlapping identical by descent genomic fragments between all pairs of unrelated genomes (IBD < 0.5) (Figure 2.7). Seven fragments of 2 to 116 kb make the top 5 % of the most shared regions. Two of them are located around known drug resistance markers *crt* and *aat1*, on chromosomes 7 and 6 respectively, and were identical in 6.0 % and 3.8 % of all pairs of unrelated genomes (Figure A.9).

The most conserved genomic fragments were also identified in subset of genomes with ('Resistant') or without ('WT') a mutated allele reducing the susceptibility to antimalarials for 6 different genes: *aat1*, *crt*, *dhfr*, *dhps* and *mdr1* (Figure A.10). As expected, genomes

carrying the haplotype *crt* K76T share two large conserved regions around *crt* and *aat1* that are absent in genomes with a wild-type *crt* haplotype, showing that these two regions are strongly linked in *Plasmodium falciparum* strains resistant to chloroquine. Interestingly, genomes carrying the *mdr1* N86Y allele have more conserved regions around *mdr1*, *aat1*, *crt* and *dhps*, which may indicate that *mdr1* N86Y haplotype is linked to regions able to confer a resistance to diverse antimalarials.

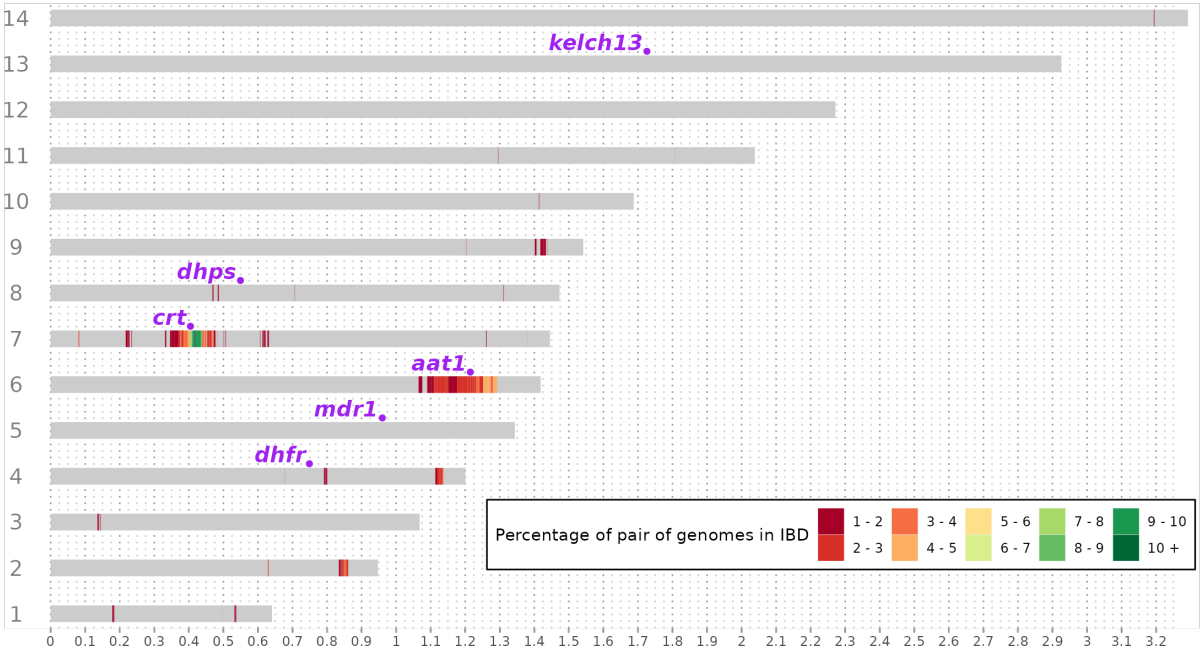


Figure 2.7: **Identical chromosomal regions most frequently detected**. Genomic regions identical by descent between unrelated genomes (IBD < 0.5) using all 19700 pairs of genomes. The 14 *Plasmodium falciparum* chromosomes were annotated with the 6 genes known to reduce susceptibility to antimalarials *aat1*, *crt*, *dhfr*, *dhps*, *kelch13* and *mdr1*. Two large regions showing the highest percentage of IBD are centered around the *aat1* locus on chromosome 6 and the *crt* locus on chromosome 7. These two regions are nearly identical to two of the regions identified by Nwakanma *et al.* (2014) with high standardized integrated haplotype scores [248].

2.3.7 Natural cross offspring and lineages expansion

To identify progeny and parental genomes of putative natural crosses, we sought pairs of unrelated genomes that, if involved in a natural cross, would lead to an IBD pattern similar to that of one other genome present in the dataset. In total, 6 putative parent-offspring clusters were observed, each containing the same two unrelated genomes (IBD = 0.17) that could be parents and at least one putative progeny genome in IBD from 0.35 to 0.65 with each parent (Figure A.11). The 6 putative offspring genomes were identified in different human hosts, with two of them (K0030301 and K0030302) being young children siblings living in the same house. Among the 6 distinct putative offspring with identical parental genomes K0141403_1504 and

K0374502_1412, J0060701_1511 is the one with the smallest proportion of genomic fragments with a non-parental origin and displays large segregated blocks of genomes from the parents, separated by meiotic breakpoints (Figure 2.8 & Figure A.12). Each of the 6 putative progeny genome is unique (IBDs between the putative offspring are inferior or equal to 0.68), with multiple meiotic breakpoint coordinates not found in the other 5 progeny, demonstrating that none of these 6 genomes result from a clonal lineage expansion. However we cannot rule out that one or more natural crosses occurred between K0141403_1504 and K0374502_1412 before these 6 progeny genomes were identified.

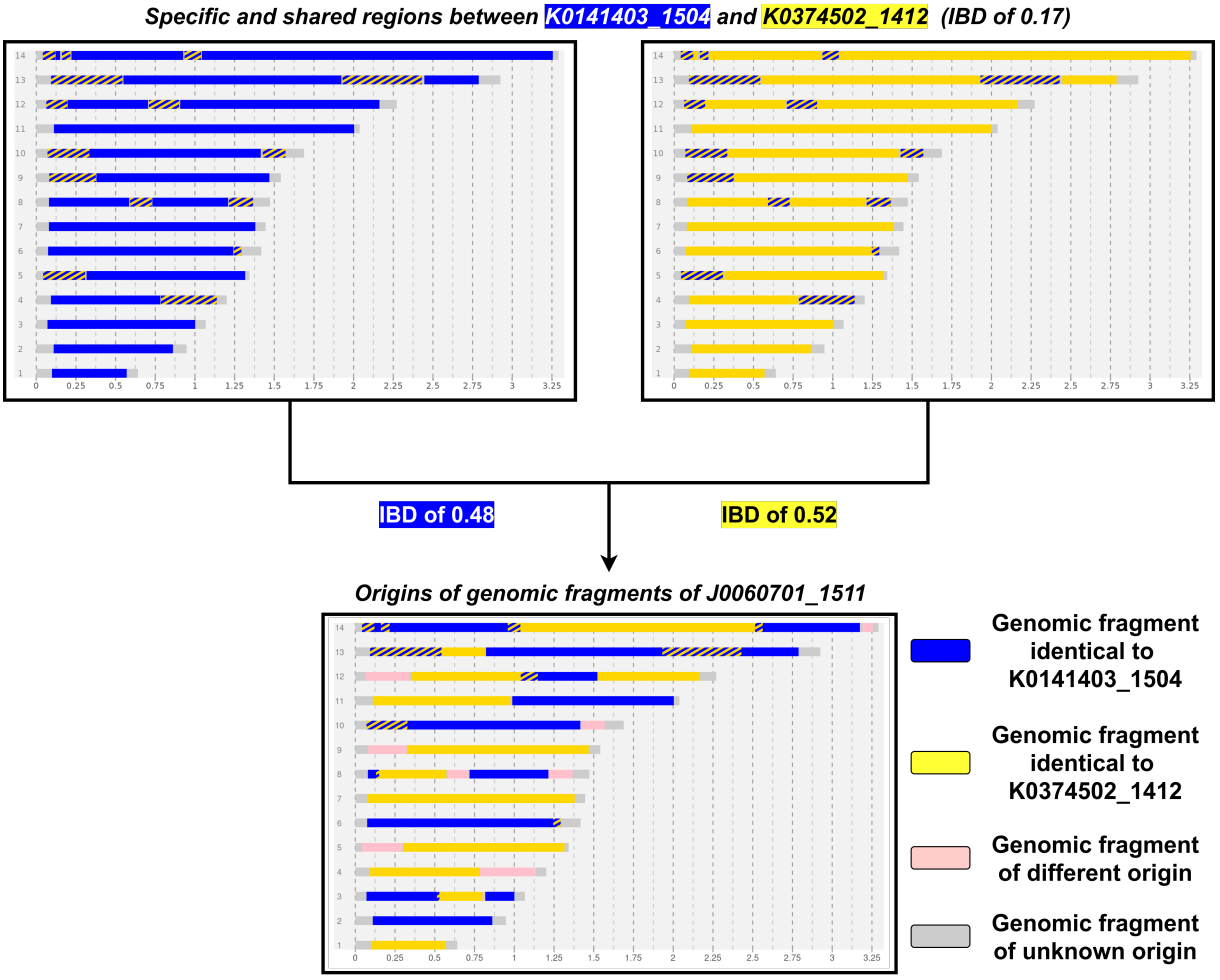


Figure 2.8: **Example of a likely natural cross between two unrelated parental genomes**. The 14 nuclear chromosomal origins of J0060701_1511 are inferred from IBD with unrelated genomes (IBD = 0.17) K0141403_1504 and K0374502_1412. Each chromosomal fragment can be in IBD with K0141403_1504 (blue), K0374502_1412 (yellow), both of them (hatched) or neither of them (pink). Some regions (grey) located generally near or in telomeric regions had too few high quality SNPs, thus preventing an estimation of the IBD. Surprisingly, almost all the fragments that are unrelated to neither of the parental genomes occur near identical regions between the two parental genomes. This most likely means that the actual strains that gave birth to J0060701_1511 are not K0141403_1504 and K0374502_1412 but are highly related to them.

2.4 Discussion

To anticipate the pre-elimination era of malaria, we developed barcode genotyping and WGS methods to build a genetic relatedness network of parasites and access the transmission mechanisms of malaria in The Gambia. The vast majority of parasites were genetically unrelated and those related were shared across the different villages studied. Almost every strain identified has recombined the following year, completely renewing the parasite genetic diversity. This sharply contrasts with neighbouring Senegal, in the area of Thiès, where parasite haplotype diversity is low, with multiple identical *Plasmodium falciparum* genomes recurring year after years, indicative of self-mating transmission and low outcrossing levels [60, 250, 251].

We estimated the complexity of infections (COI) using both allelic frequencies and locus heterozygosity and observed a constant level of polyclonal infections [82, 124]. Variations of the proportion of polyclonal infections has been positively correlated with changes in transmission intensity [252, 241, 63, 101]. In this studied area of The Gambia, COI was stable over time, correlating with recent data on prevalence [238]. Most infections were sub-microscopic, which is typical, especially when prevalence falls below 20 % [253, 254]. These low parasitaemia infections that may persist for months are typically asymptomatic unless the host immune system is compromised [166]. In this study, the longest infection that we were able to observe was in a 9-year-old individual continuously infected, asymptomatically, for one and a half year by the exact same parasite strain. The age of this individual falls in the range of school-age children who contribute the most the malaria reservoir through their higher carriage of asymptomatic infections and their ability to effectively infect mosquitoes [255, 22, 238, 68].

In The Gambia, drug resistant allele frequencies increased dramatically from the late 1980's until the early 2000's, then plateaued until 2008 [248]. Almost all chromosomal regions with higher levels of shared IBD are in linkage disequilibrium with known drug resistance markers, indicating that drug resistance is currently the strongest genomic selective pressure on the parasite genome in The Gambia. We could confirm that the two regions surrounding *aat1* and *crt* were strongly linked as they were identical between genomes carrying *aat1* S258L and *crt* K76T haplotypes [169]. Furthermore, 3 distinct conserved regions surrounding *aat1* were identified between genomes carrying the haplotypes *aat1* wild-type, *aat1* S258L and *mdr1* N86Y suggesting that three distinct haplotypes surrounding *aat1* might coexist in the population. This is likely explained by constant use of Sulfadoxine/Pyrimethamine prophylaxis since the late 90s and Artemether/Lumefantrine as treatment. While in South-East Asia the large-scale

usage of ACTs has led to development of drug-resistance and drastic reduction of parasite population haplotype diversity, this is not (yet) the case in The Gambia.

Amambua-Ngwa *et al.* showed previously that parasite genetic similarity is inversely correlated with each spatial and temporal distances in The Gambia during the 2013 transmission season [73]. We also find that parasites sampled in close households are more genetically similar but only when parasites were sampled less than three months apart. Similarly, Lee *et al.* observed parasite strains with varying levels of spatio-temporal propagation in Thiès, Senegal which suggests that some parasites are more actively transmitted than others [101]. These results add evidences that anti-malarial strategies could prioritise all members of a household with an infected individual.

A follow-up study carried out in rural villages of eastern Gambia between 2012 and 2016 reported that roughly half of asymptomatic infected individuals at the end of the wet season were still infected at the end of the dry season [238]. We showed that parasites of a dry season share significantly more genetic similarity with those from the previous wet season than those from the next wet season. The genetic similarity between parasites sampled 8 to 12 months apart is five times lower than the one between parasites sampled less than 2 months apart. Also, parasites sampled more than one year apart had very low level of similarity, implying that the whole population had been replaced in just one year. A study conducted in Colombia (with a 10-fold lower malaria prevalence than The Gambia) showed that the same level of decrease of genetic similarity was reached in about 7 years [252, 98].

A key aspect of our approach was the active case detection of asymptomatic infections, as opposed to previous studies sequencing parasites solely derived from clinical cases. It is essential to consider asymptomatic infections when dealing with *Plasmodium falciparum* malaria as we showed that they likely represent the main reservoir of parasites surviving the dry season in The Gambia. Furthermore, we described a 9-year old individual infected with the same parasite genotype for one year and a half, showing the important immune evasion capabilities of *Plasmodium falciparum*. Our results advocate for active case detection as a necessary step to comprehensively characterize the true parasite genetic diversity. Thanks to whole genome sequencing, the recombinatorial history of parasites can be retrieved from a few generations back, allowing to understand the precise origin of each genomic fragment. Using IBD status of each fragment in the whole population, we showed that drug resistance has a major impact on the parasite genome.

2.5 Acknowledgements

This publication uses data from the MalariaGEN SpotMalaria project as described in 'Jacob CG *et al.*; Genetic surveillance in the Greater Mekong Subregion and South Asia to support malaria control and elimination; eLife 2021;10:e62997 DOI: 10.7554/eLife.62997' [123]. The project is coordinated by the MalariaGEN Resource Centre with funding from Wellcome (206194, 090770). We would like to thank Julia Mwesigwa, Michael Fontaine, Franck Prugnolle, Virginie Rougeron and all fieldworkers involved in the study. The authors would like to thank the staff of Wellcome Sanger Institute Sample Management, Genotyping, Sequencing and Informatics teams for their contribution. The authors acknowledge the ISO 9001 certified IRD i-Trop HPC (member of the South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: https://bioinfo.ird.fr/ - http://www.southgreen.fr.

Project 2: Variants Generated During the Course of an Infection

3.1 Introduction

As detailed in the previous chapter (*Chapter 2 Project 1: Spatio-temporal Relatedness of Parasites*), malaria in The Gambia is highly seasonal, with a 7-month long dry season with very few *Plasmodium falciparum* cases. The lack of transmission during the dry season implies that *Plasmodium falciparum* parasites are able to remain hidden in an asymptomatic human host, therefore surviving for many months without being cleared by the immune system [111, 114]. The extent of these *Plasmodium falciparum* malaria asymptomatic chronic infections has just recently been seriously considered thanks to more sensitive malaria diagnosis tools able to detect very low parasite densities in the blood [152, 256].

The main *Plasmodium falciparum* antigens triggering an immune reaction towards iRBCs are PfEMP1s, encoded by the hypervariable family of *var* genes, containing approximately 60 genes per parasite genome [257, 211]. Interestingly, iRBCs display at their surface only one type of PfEMP1 throughout the intra-erythrocytic life cycle, effectively concealing most of the other PfEMP1s from the immune system. At the end of each cycle, parasites switch of *var* gene expressed at a rate of about 2 %, generating a pool of parasites with antigens unknown to the immune system [88]. One model suggests that the parasite-host relationship has evolved to favour short-lived immune responses that allow the parasite to persist and the host to survive [258]. However, to the best of our knowledge, a detailed analysis of *var* gene expression and antibody recognition from the same chronic infection over a long period of time has never been performed. Mathematical models predicting durations of infection given the antigenic variation of the *var* gene family cannot explain the relatively high frequency of infections lasting longer than a few months [259]. An interesting hypothesis that could explain the efficient immune evasion of the parasites is that chimeric *var* genes can be generated through ectopic recombinations between their exons 1 (encoding the extracellular part of the PfEMP1). These new antigens can be added to the existing pool of PfEMP1s, effectively extending the duration of infection.

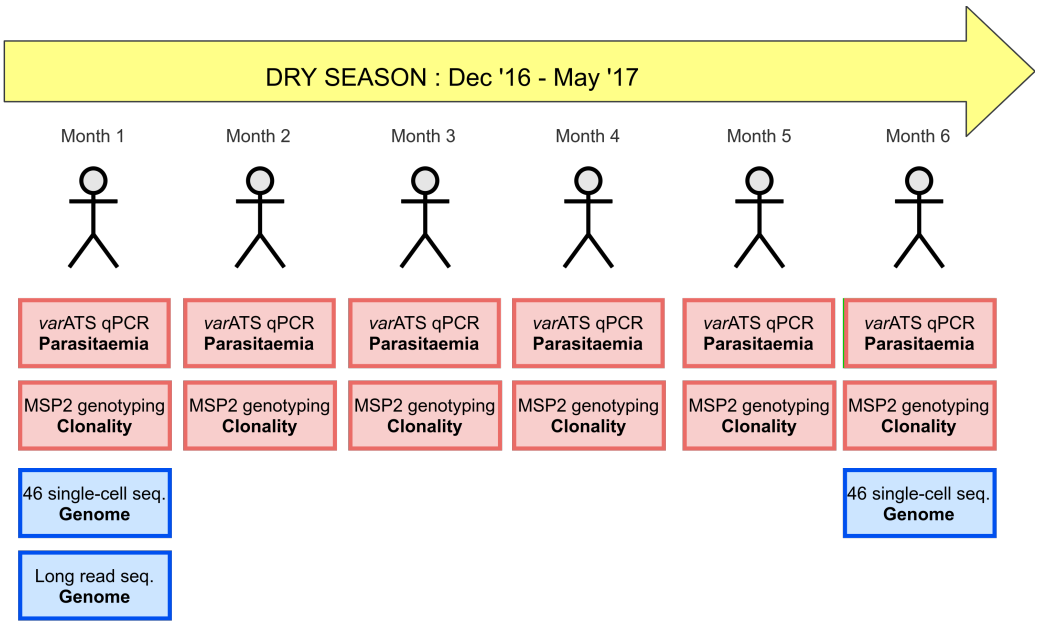
Although the generation of chimeric *var* genes without meiosis was only observed from *in vitro* culturing of several strains, it is likely that they are also generated during the course of an *in vivo* infection [54, 260]. Multiple challenges arise when looking for occurrences of chimeric *var* genes generated *in vivo*. First, an infection has to be monitored for a sufficient time to let a parasite have a chance to undergo a successful ectopic *var* gene recombination between two samplings without any risk for the volunteer. Second, the most thorough approach to assemble full-length *var* genes is via long-read sequencing as these genes are all located in internal and subtelomeric hypervariable regions of the chromosomes, inaccessible to short-

reads of typically 150 bp [38, 41]. The technique requires a relatively large amount of parasite DNA, which is not directly available from an asymptomatic infection [20]. Culturing isolates *in vitro* to bulk up DNA parasite is not trivial and can only be limited to a few weeks before the parasite genome acquire mutations conferring a growth advantage [261, 262]. Finally, as these *in vivo de novo* chimeric *var* genes are assumed to be rare and do not necessarily confer a growth advantage to the parasites carrying the chimera, it is necessary to sequence each single-cell genome individually instead of sequencing in bulk.

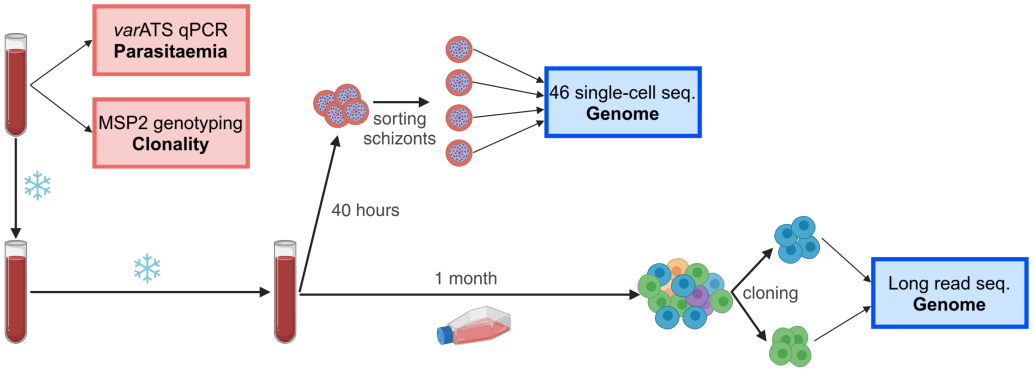
To characterise *de novo* mutations occurring in long-lasting chronic infections, we took advantage of the cohort recruited in The Gambia, as described previously (*Chapter 2 Project 1: Spatio-temporal Relatedness of Parasites*). Briefly, in December 2016, inhabitants of 4 nearby villages in Eastern Gambia were tested for *Plasmodium falciparum* by *var*ATS qPCR and only those positive were recruited. A cohort of 42 asymptomatic *Plasmodium falciparum* carriers was followed monthly with *var*ATS qPCR tests and venous blood samplings until May 2017. The follow-up was ended for individuals who became symptomatic and received treatment immediately, or who naturally cleared their infection (2 successive months of being *Plasmodium falciparum* negative) or withdrew from the cohort. Thanks to this survey, the clonality of infections could be estimated and followed monthly by using MSP2 genotyping on each blood sample. One important conclusion was that long infections tend to carry more clones. The full details of this cohort study are available in Collins *et al.*, 2022 [24].

Out of the 42 individuals, 11 were chosen to have their infection precisely monitored with both long-read and single-cell sequencing of the parasites at different time points. Prince Nyarko[1], a PhD student of Antoine Claessens[1], was in charge of all laboratory work (including parasite culture) prior to obtaining sequencing data (Figure 3.1). In each of the 11 individuals, at least one blood sample was selected to attempt to culture-adapt its parasites. When the culture adaptation was successful, parasites were cloned and subsequently sequenced using the highly accurate PacBio CCS HiFi long-reads, which have an average read length of around 10 kb. In parallel, two blood isolates sampled the furthest away in infection time (5 to 6 months) were chosen to undergo *ex vivo* cell-sorting, after 40 hours of culture to reach the schizont stage. The sorted cells were individually sequenced using Illumina short-reads of 150 bp.

[1]LPHI, MIVEGEC, Université de Montpellier, CNRS, INSERM, Montpellier, France



(a) Chronology of blood sample collection during the dry season.



(b) Parasitaemia, clonality and sequencing workflow of the parasites.

Figure 3.1: **Sample collection and analysis workflow of blood samples**. (a): During the dry season in 2016/2017, 11 individuals were followed monthly with parasitaemia and clonality estimations (red boxes). For the majority of individuals, single-cell sequencing ('seq.') was attempted on the most distant time points of the infection while long-read sequencing was attempted on the very first time point available of each infection. For individuals for whom the sequencing was not successful, the time points that underwent single-cell or long-read sequencing (blue boxes) could be different from those presented here. (b): In more detail, parasitaemia and clonality were initially gathered just after the blood samplings. Following that, blood vials were kept frozen until thawed in 2022 or 2023 to undergo either single-cell short-read or clonal long-read sequencing ('seq.'). The single-cells correspond to schizonts sorted about 40 hours after thawing the samples, while the clones were obtained from culture-adapted parasites grown for about one month. Created with BioRender.com.

3.2 Material and methods

3.2.1 Parasitaemia quantification

The parasitaemia of each blood sample was quantified using qRT-PCR of mRNA (messenger RNA) transcripts specific to gametocytes or ring-stage iRBCs [263, 24].

3.2.2 Culturing

The culturing of the parasites was performed by Prince Nyarko[1] and Catherine Jett[2]. To compare genomic data of the short-term-hosted with the long-term-hosted parasites, 2 different blood samples (one from an early time point, the other from a late time point) were picked for each of 11 selected individuals. After thawing, blood samples were cultured for about 40 hours to reach late-pigmented trophozoite or schizont stages, which both contain the highest amount of DNA material. These samples were then purified by MACS column and sorted by flow cytometry. For each isolate, 48 cells were sorted in individual wells, plus a well with 48 pooled cells. In parallel, 10 blood isolates from 7 individuals were cultured *in vitro* for one month. Cultures of parasites were cloned by limiting dilution in multiple independent subcultures and expanded until sufficient cells were obtained. Two of the 10 cultures yielded two distinct clones by MSP2 genotyping, making the total number of clonal parasite lines to 12.

3.2.3 Sequencing

The 22 pairs of sorted schizonts of the 11 individuals underwent multiple displacement amplification of their DNA content and were sent for single-cell sequencing (48 cells) with Illumina 150 bases paired-end reads at the Texas Biomedical Institute (San Antonio, USA). An additional sequencing of 48 cells pooled together, refer to as 'pooled-cells' hereafter, was also performed on each sample of sorted schizonts with Illumina 150 bases paired-end reads. In parallel, 12 clonal parasite lines were sent for whole genome sequencing with Pacific Biosciences long-read sequencing using the CCS HiFi reads at the Sanger Institute (Hinxton, UK).

3.2.4 Whole genome long-read assembly and annotation

The long-read sequencing data from culture-adapted lines was used to build one assembly per clone thanks to both Tree of Life team at the Sanger Institute and custom pipelines using hicanu to trim the CCS reads and hifiasm (version 0.16.1-r375, '--hg-size 24m -l0 -f0') to

[1]LPHI, MIVEGEC, Université de Montpellier, CNRS, INSERM, Montpellier, France
[2]Host Pathogen Interactions Program, Texas Biomedical Research Institute, San Antonio, TX, USA

assemble them [127, 128]. Contigs were assembled into longer scaffolds with the 3D7 chromosomes as a reference using ragtag (version 2.1.0) [129]. Genes were predicted with Augustus (version 3.3.3) from the *Plasmodium falciparum* trained model available, with the help of Mathieu Quenu[1] for some of the assemblies [130]. *var* genes were extracted from the full list of identified proteins by screening them for LARSFADIG and DYVPQYLRW motifs with 2 mismatches allowed (these conserved motifs are found in all *Plasmodium falciparum* PfEMP1s, with the exception of the PfEMP1 encoded by *var2csa*). The different PfEMP1 domains were identified by the web version of VarDom (version 1.0) [47].

3.2.5 Mapping and variant calling

Ian Cheeseman[2] mapped single-cell and pooled-cells sequencing reads on the 3D7 reference genome using bwa and called variants using a custom GATK variant discovery pipeline focused on 3D7 core genome [264, 121, 41]. To check for non-*falciparum* malaria, reads were mapped to the reference genomes of *Plasmodium malariae*, *Plasmodium ovale curtisi* and *Plasmodium vivax* using FastQ Screen (version 0.15.2). Similarly to project 1 (*Section 2.2.4 Parasite relatedness*), we estimated the genetic similarity by computing pairwise mean posterior probabilities of identity by descent (IBD) between single-cell genomes using hmmIBD [79]. The probability of two samples to be in IBD represents the expected shared fraction of their genomes. To identify variants that increased in frequency between the single-cells obtained from different time points of the same infection, SNPs covered by at least 10 reads with a within-cell MAF below 0.1 were extracted using the publicly available program varif (https://github.com/marcguery/varif, version 0.4.0.dev2).

3.2.6 *de novo* assembly of *var* genes

For one individual (DC05), no reference genome could be generated as the cultured parasites failed to grow *in vitro*. The pooled-cells and single-cell genomes available from both time points sequenced were therefore used to generate repertoires of *var* genes using a custom short-read *de novo* assembly pipeline based on methods developed by Otto *et al.* (2019) and Andradi-Brown *et al.* (2023) [51, 46].

The assembly pipeline performs the following steps on each individual genome sequenced:

1. Extract reads (and their mate) aligned in regions annotated as *var* gene in the 3D7 genome.

[1]LPHI, MIVEGEC, Université de Montpellier, CNRS, INSERM, Montpellier, France
[2]Host Pathogen Interactions Program, Texas Biomedical Research Institute, San Antonio, TX, USA

2. Extract unmapped reads (and their mate).
3. Merge reads from steps 1. and 2. and keep only unique pairs.
4. Obtain a *var* assembly of each genome with rnaSPADES (-k 71) [265].
5. Remove similar contigs (identity > 95 % and both sequences overlap > 90 %) with cd-hit [266].
6. Merge contigs with SSPACE using the mapping of sequencing reads on their own *var* assembly [267].
7. Keep contigs encoding a protein containing both motifs LARSFADIG and DYVPQYLRW (with 2 mismatches allowed) with at least 400 uninterrupted amino acids following LARSFADIG (included).
8. Extract *var* genes from contigs using the delimitation of the detected PfEMP1 domains by the web version of VarDom (version 1.0) [47].
9. Merge similar *var* genes (identity > 95 % and shortest *var* gene covering 90 % of the longest).
10. Translate all *var* genes and extract the open reading frames containing LARSFADIG motif which corresponds to the part of the protein encoded by exons 1 of *var* genes

To identify identical PfEMP1 sequences found in multiple time points, all PfEMP1 repertoires of each individual (single-cell or pooled-cells) genome were aligned with one another and clustered using Clustal-Omega (version 1.1.0) [131]. Using igraph (version 1.2.11) R package, all PfEMP1s were grouped into high similarity clusters in which they share an identity above 95 % with any other PfEMP1 from the same cluster, corresponding to a component in graph theory [125, 126].

3.2.7 Chimeric *var* gene screening

As described above, *var* repertoires were extracted from long-read assemblies or directly assembled *de novo* from short-read sequencing of single-cell and pooled-cells genomes. Reads sequenced from single-cell and pooled-cells parasite genomes obtained from the late time points of infections (month 6) were processed through discovarif (https://github.com/marcguery/discovarif, version 0.0.6), a custom pipeline that I developed to map reads and call for variants, including large structural variants such as those generating chimeric *var* genes. During the first step of this pipeline, reads were trimmed

with Trimmomatic (version 0.39) and mapped with bwa (0.7.17) on their corresponding *var* repertoire which was made out of one contig per *var* gene [268, 264]. Duplicate reads were then removed from the mappings with Picard (version 2.25.5) before chimeric *var* genes were searched using DELLY (version 1.1.6), which can identify inter-contig translocations events using the differential locations of read pairs after the mapping [269, 132]. Hits obtained from DELLY were retained if the variant was annotated as 'PASS', the read depth was above 10 and the proportion of reads supporting the translocation event was above 0.5.

3.3 Results

3.3.1 Asymptomatic infections followed in 11 individuals

Between December 2016 and May 2017, 11 *Plasmodium falciparum* PCR-positive individuals (DC01, DC03, DC04, DC05, DC07, DC08, DC09, DC10, DC11, DC12, DC13) aged 9 to 27 years old were followed monthly and selected for this analysis (Table 3.1). Three pairs of individuals (DC01 and DC04, DC07 and DC08, DC09 and DC10) came from the same household identified as J027, P008 or P002. Parasite densities (rings and gametocytes) were determined monthly by qRT-PCR. At their enrolment in December 2016 or January 2017, individuals had a median concentration of ring-stage iRBCs at 9×10^5 parasites/mL. The majority of the individuals had more than 90 % of their initial parasitaemia made out of ring-stage iRBCs while one of them (DC12) had 39 % of its initial parasitaemia made out of gametocytes.

Table 3.1: ***Plasmodium falciparum* asymptomatic individuals selected for this analysis**. Out of the 42 individual recruited in the *Plasmodium falciparum* asymptomatic cohort, 11 were selected for this analysis. Their parasitaemia (ring and gametocyte levels) shown here were those from their first blood sampling in December 2016 or January 2017.

Participant ID	Sex	Age	Household ID	Gametocytes/mL	Rings/mL	Rings (%)
DC01	Male	9	J027	71	9.3×10^6	100
DC03	Female	13	P006	1.1×10^3	3.7×10^6	100
DC04	Female	15	J027	76×10^1	1.7×10^6	100
DC05	Male	13	K026	2.5×10^4	1.8×10^6	99
DC07	Female	23	P008	84×10^1	4.9×10^5	100
DC08	Male	10	P008	22×10^1	1.2×10^7	100
DC09	Male	12	P002	35×10^1	9.0×10^5	100
DC10	Male	12	P002	6.6×10^3	3.9×10^5	98
DC11	Female	27	P005	70×10^1	9.7×10^3	93
DC12	Female	19	K023	1.2×10^4	1.9×10^4	61
DC13	Male	11	P012	29×10^1	1.1×10^5	100

Out of the 11 individuals, 10 had infections lasting for at least 6 months and 1 (DC10) had an infection lasting 3 months. Multiple monthly blood samplings were obtained from each individual and labelled m1 to m6. With the exception of DC01 and DC10, each individual had 2 blood samples separated by at least five months that were selected for single-cell processing, hereafter referred to as 'early' and 'late' time points. In total, 949 genomes were short-read sequenced from 933 single-cells and 16 'pooled-cells', the latter corresponding to 48 cells sorted together in the same well, as a positive control (Table 3.2). Five individuals (DC01, DC03, DC04, DC07 and DC10) had one time point long-read sequenced early in the follow-up study (month 1 or 2), and three (DC08, DC09 and DC10) had a long-read sequencing performed

from blood samplings at month 3 or 5. For individuals DC08 and DC09, two distinct clones were identified from the same time point (both at month 5) and were consequently individually long-read sequenced. Individuals DC08 and DC10 had respectively 2 and 3 distinct time points processed by long-read sequencing.

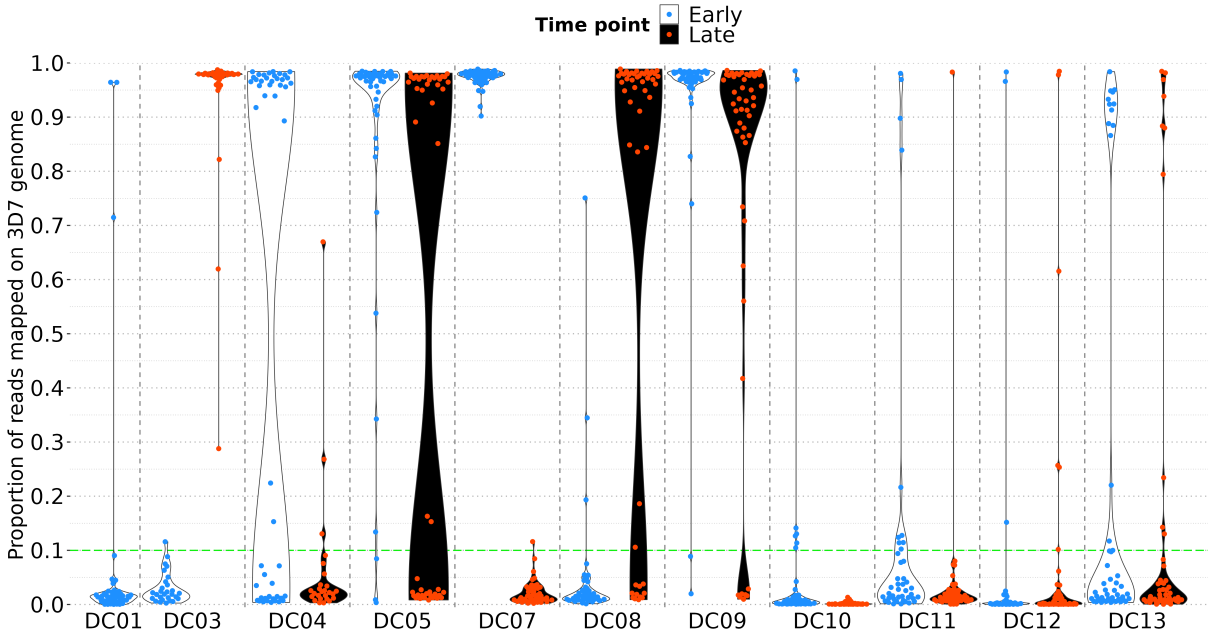
Table 3.2: **Individual blood samples single-cell or long-read sequenced**. In total, 25 blood samples from the 11 selected individuals underwent single-cell or long-read sequencing. Part of the blood samples were directly thawed and sorted to enrich for schizont-stages iRBCs before being sent for single-cell sequencing. The remaining parasites were cultivated for one month and cloned by limiting dilution before being individually sent for long-read sequencing. For two of the blood samples (DC08_m5 and DC09_m5), two distinct clones were identified and separately long-read sequenced.

Participant ID	Month (*time point*)	Sample ID	Sequencing
DC01	December (*early*)	DC01_m1	Long-Read + Single-Cell
DC03	December (*early*)	DC03_m1	Long-Read + Single-Cell
	May (*late*)	DC03_m6	Single-Cell
DC04	January (*early*)	DC04_m2	Long-Read + Single-Cell
	May (*late*)	DC04_m6	Single-Cell
DC05	December (*early*)	DC05_m1	Single-Cell
	May (*late*)	DC05_m6	Single-Cell
DC07	December (*early*)	DC07_m1	Long-Read + Single-Cell
	May (*late*)	DC07_m6	Single-Cell
DC08	December (*early*)	DC08_m2	Single-Cell
	February	DC08_m3	Long-Read
	April	DC08_m5[1]	Long-Read
	May (*late*)	DC08_m6	Single-Cell
DC09	January (*early*)	DC09_m2	Single-Cell
	April	DC09_m5[2]	Long-Read
	May (*late*)	DC09_m6	Single-Cell
DC10	December (*early*)	DC10_m1	Long-Read
	January	DC10_m2	Long-Read + Single-Cell
	February (*late*)	DC10_m3	Long-Read + Single-Cell
DC11	December (*early*)	DC11_m1	Single-Cell
	May (*late*)	DC11_m6	Single-Cell
DC12	December (*early*)	DC12_m1	Single-Cell
	May (*late*)	DC12_m6	Single-Cell
DC13	December (*early*)	DC13_m1	Single-Cell
	May (*late*)	DC13_m6	Single-Cell

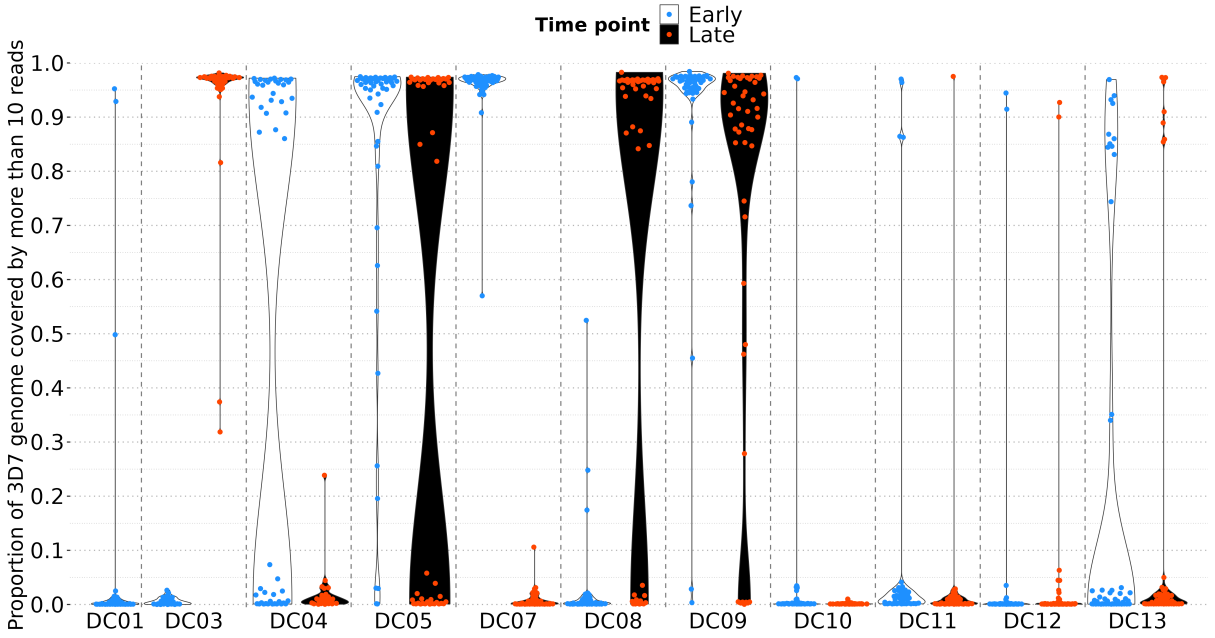
3.3.2 Successful single-cell sequencing of several asymptomatic infections

The 949 (933 single-cell and 16 pooled-cells) genomes sequenced from all individual infections at early or late time points were mapped to 3D7 reference genome, and their

quality was assessed (Table 3.2 & Figure 3.2). Out of the 949 genomes, 384 had more than 10 % of their reads mapped on the 3D7 reference genome (Figure 3.2a). Most of the genomes (883/949) had either more than 90 % (291) or less than 5 % (592) of the 3D7 reference genome covered by at least 10 reads (Figure 3.2b). Out of the 565 genomes with a mapping percentage below 10 %, 562 had a genomic coverage below 5 %, indicating that these two metrics are highly correlated and can be equally used to filter out low quality genomes (Figure B.1). In most cases, the vast majority of the genomes sequenced from the same time point had all quasi identical levels of genomic coverage or mapping percentage between one another, resulting in either almost all genomes (DC03_m6, DC05_m1, DC07_m1 and DC09_m1) or almost no genomes (DC01_m1, DC03_m1, DC04_m6, DC07_m6, DC08_m1, DC10, DC11 and DC12) successfully sequenced for a given time point (Figure 3.2).



(a) 3D7 genome mapping.



(b) 3D7 genome coverage.

Figure 3.2: **Mapping and coverage proportions of 3D7 from single-cell and pooled-cells reads**. The 949 (933 single-cell and 16 pooled-cells) genomes from early (blue points) and late (red points) time points of the infections were sequenced. (a): The resulting reads were mapped to 3D7 reference genome in varying proportions, 384 genomes have more than 10 % (green dotted line) of their reads mapped to 3D7 reference genome. (b): The mappings resulted in different levels of coverage at 10 reads depth.

Overall, 89962 bi-allelic SNPs were called from the 384 *ex vivo* parasite genomes (371 single-cells and 13 pooled-cells) with at least 10 % of their reads mapped to 3D7. Out of these 384 genomes, 341 (330 single-cells and 11 pooled-cells) had more than 30 % of the 54505 high quality SNPs (QUAL > 10000) covered by 10 reads or more (Figure 3.3). The vast majority of these genomes (326/341) had more than 80 % of the 3D7 reference genome covered by at least 10 reads (Figure B.1b).

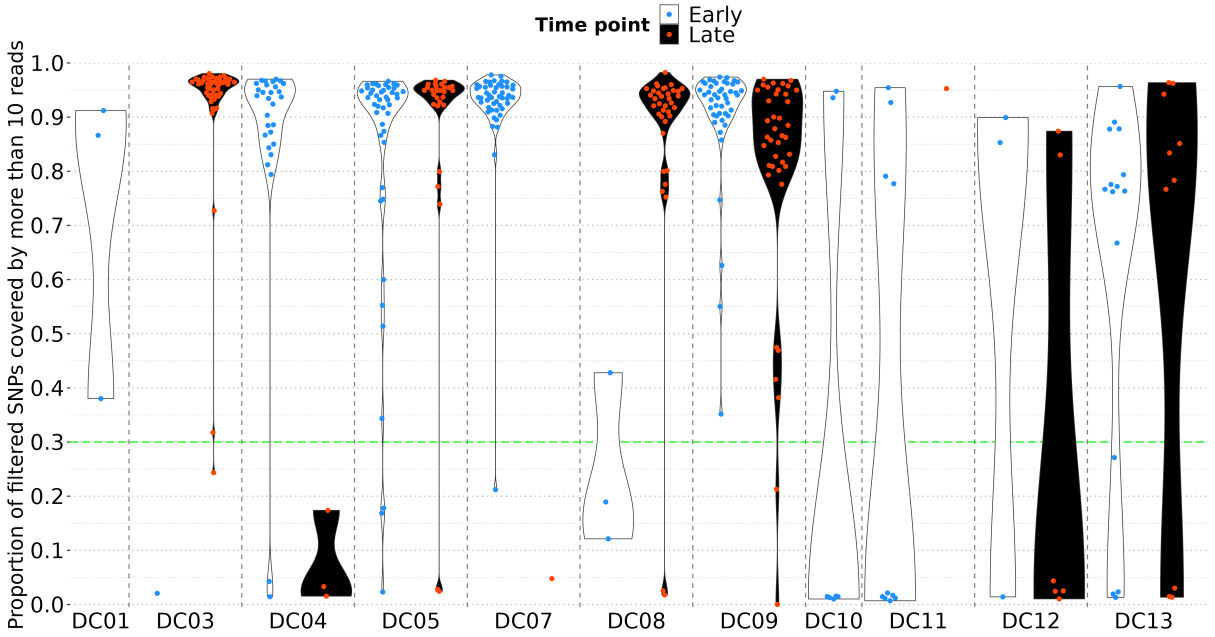


Figure 3.3: **Coverage of 54505 SNPs in single-cell and pooled-cells genomes**. Overall, 54505 bi-allelic SNPs were called from the 371 single-cell and 13 pooled-cells genomes collected from early (blue points) and late (red points) time points of the infection that had at least 10 % of their reads mapping to 3D7. The majority of the genomes (341/384) had more than 30 % of the total set of 54505 SNPs covered by 10 read or more (green dotted line).

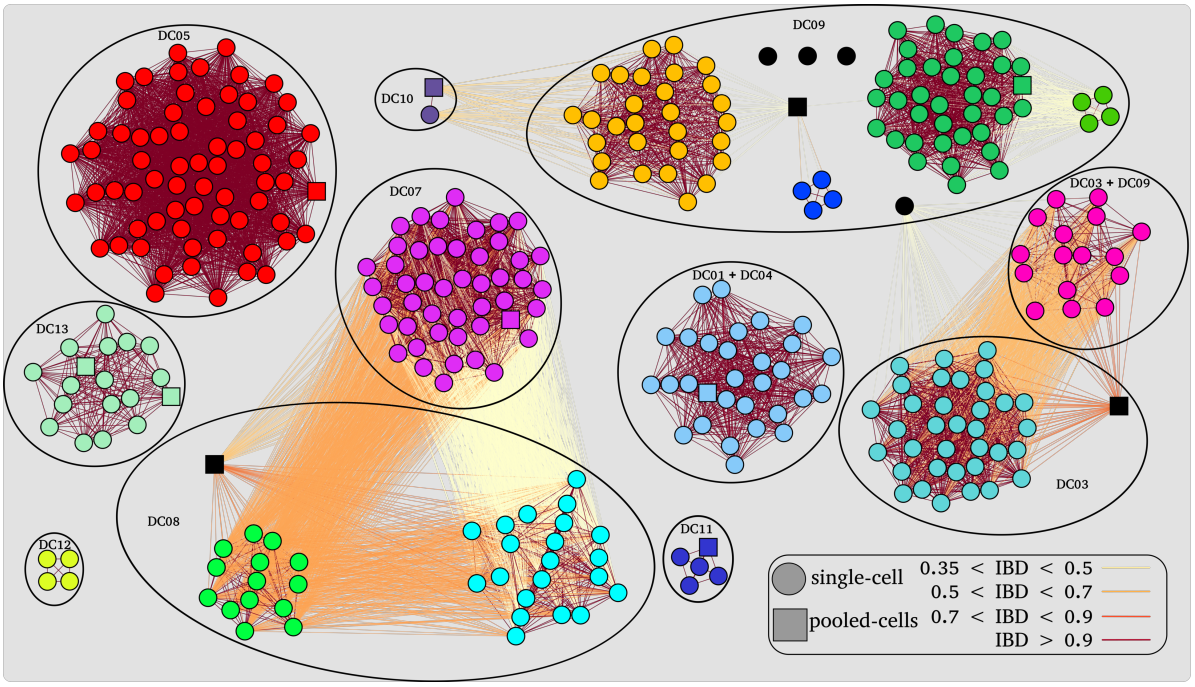
3.3.3 Precise segregation of parasite genotypes

As expected from single-cell genomes of a haploid organism, 99 % of their SNPs covered by more than 10 reads displayed a within-sample MAF under 0.1. The relatedness between each pair of *ex vivo* parasite genomes with at least 1000 informative positions (a locus is deemed informative when it is available in both genomes and that at least one of them is the minor allele) was estimated by hmmIBD using the 54505 high quality SNPs, excluding for each sample the few SNPs with a within-sample MAF above 0.1. The different genomes were grouped into high relatedness clusters by considering the amount of IBD links over 0.9 that they share with one another (Figure 3.4). As expected, most of the IBD clusters contain cells obtained from the same individual. All pooled-cells genomes are preferentially related to the cells obtained from

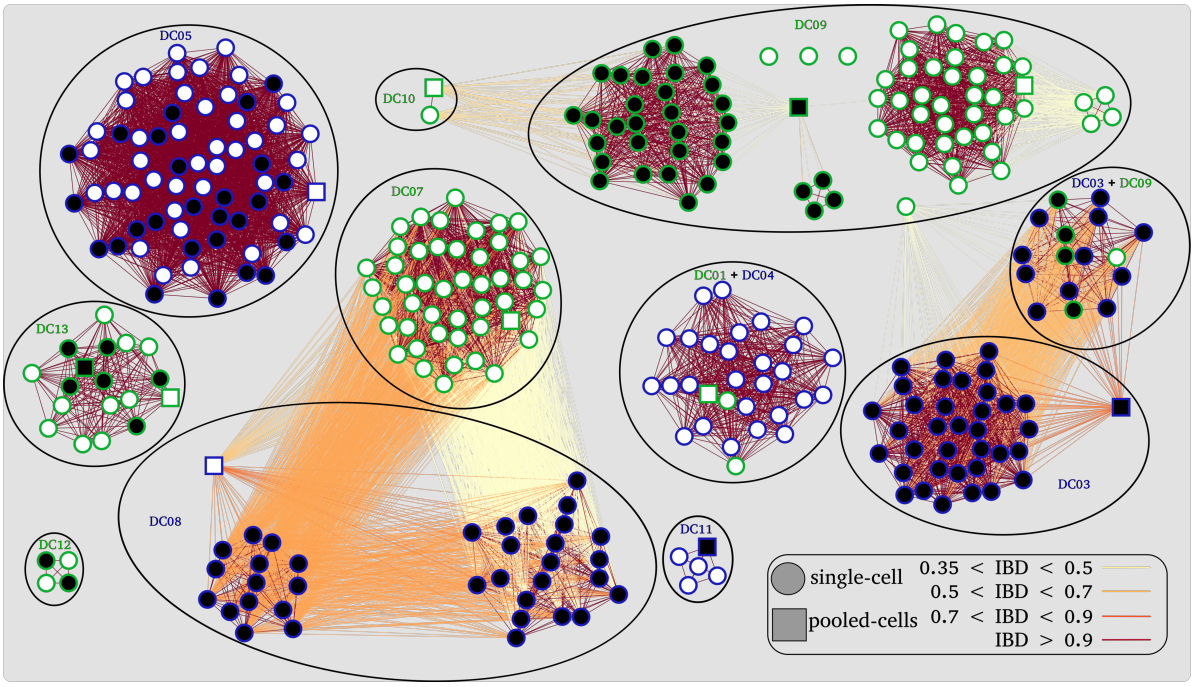
the same individual infection. Three individuals (DC05, DC12, DC13) had multiple cells from both time points all showing high IBD between each other, implying that these individuals had monoclonal infections that survived for several months. Among all three monoclonal infections, individual DC05 is the one with the highest number of cells from both the early and late time points with 41 and 26, respectively.

Interestingly, two clusters contain identical genomes (IBD > 0.9) from different individuals. Specifically, one cluster combines DC01 (2 early time point cells and one early time point pooled-cells genome) with DC04 (27 early time point cells), two individuals living in the same household (J027). Secondly, a cluster combines DC03 (11 early time point cells) with DC09 (5 early and late time point cells) with an IBD in the range of 0.5 to 0.7 between all single-cell genomes.

Individuals DC03, DC08 and DC09 each contains an array of genomes that share varying levels of relatedness with one another. DC08 contains 35 single-cell genomes from the late time point only that cluster in two subpopulations (14 and 21 cells) with an IBD around 0.6. Interestingly, DC07 cells (47 cells from the early time point only) all display low to intermediate (IBD between 0.35 and 0.7) level of similarity with each of the DC08 subpopulations. Of note, DC08 and DC07 live in the same household. DC03 contains 47 single-cell genomes from the late time point, they cluster in two subpopulations that share an IBD of about 0.6. As expected, the pooled-cells genome is in IBD with both subpopulations. DC09 (45 and 37 cells from early and late time points) is the individual with the highest level of polyclonality with cells grouped in 5 different clusters that have a low (IBD < 0.5) relatedness with one another and 4 individual cells that were not clustered with any other genome. Three out of the four isolated cells had the highest level of relatedness with the pooled-cells genomes DC09_m2 and DC09_m6 (IBD values of 0.27, 0.26 and 0.31). Altogether, over the two time points in DC09, 9 distinct genomes have been identified.



(a) Genomes coloured by high relatedness (IBD > 0.9) cluster.

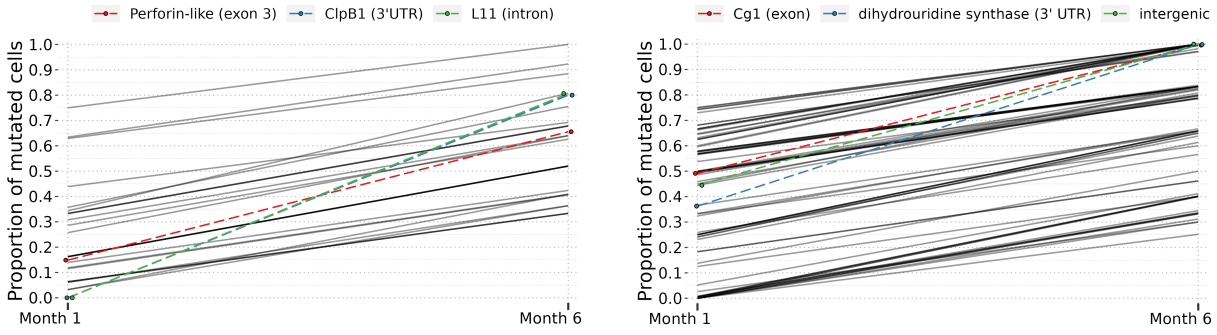


(b) Genomes coloured by time point of sampling and individual.

Figure 3.4: **Relatedness network of single-cell and pooled-cells parasite genomes**. The network of IBD values above 0.35 between *ex vivo* parasite single-cell (circle) and pooled-cells (square) genomes is represented by nodes linked by edges coloured by the level of relatedness between genomes estimated by IBD. Highly similar genomes were grouped so that every genome in the same cluster shares an IBD above 0.9 with at least one other member from the same cluster. The resulting clusters were arranged with the compound spring embedder layout algorithm from Cytoscape (version 3.9.0) and manually annotated with the different individuals (DC) composing them [247]. (a): Genomes are coloured by the high relatedness cluster they belong to, with in black the genomes that did not belong to any cluster. (b): Genomes are coloured by the date of sampling (early in white and late in black) and by individuals.

3.3.4 Selective advantages of variants during *in vivo* infections

Individuals DC05 and DC13 had a continuous monoclonal infection with numerous cells available for two time points corresponding to months 1 and 6 of the infections. As such, the prevalence of each SNP was compared between single-cell genomes of the two time points for each individual to identify markers of selection responsible for adaptation to *in vivo* long-term infection. The proportion of mutated cells was calculated for each month separately using each SNP covered by more than 10 reads and present in more than 80 % of cells of both time points. For DC05 and DC13 respectively, there were 30 and 78 SNPs that had a change of more than 0.25 in proportion of cells between months 1 and 6, with each having 3 SNPs that had a change of more than 0.5 (Figure 3.5, Table B.1 & Table B.2). For DC05, the proportion of two SNPs increased between months 1 and 6 from 0 % of the cells to 81 % (the same cells for both variants) in the late time point (Figure 3.5a). They are located in the 3'-UTR (three prime untranslated region) of the gene encoding ClpB1 (PF3D7_0816600) and in an intron of the gene encoding the ribosomal protein L11 (PF3D7_1110600). The third SNP of DC05, inducing a non-synonymous mutation K165N in the perforin-like protein 1 (PF3D7_0408700), change from a proportion of 15 % to 66 % of the cells between months 1 and 6 (Figure 3.5a). For individual DC13, one SNP located in the exon of *cg1* (PF3D7_0709100) gene and responsible for a non-synonymous mutation (P676S), change from a proportion of 36 % to 100 % of cells between month 1 and month 6 of the infection (Figure 3.5b). Two other SNPs, in the 3'-UTR of the gene encoding the dihydrouridine synthase (PF3D7_0918800) and in an intergenic region, change from a proportion of respectively 44 % and 49 % to 100 % of the cells between month 1 and 6 (Figure 3.5b).



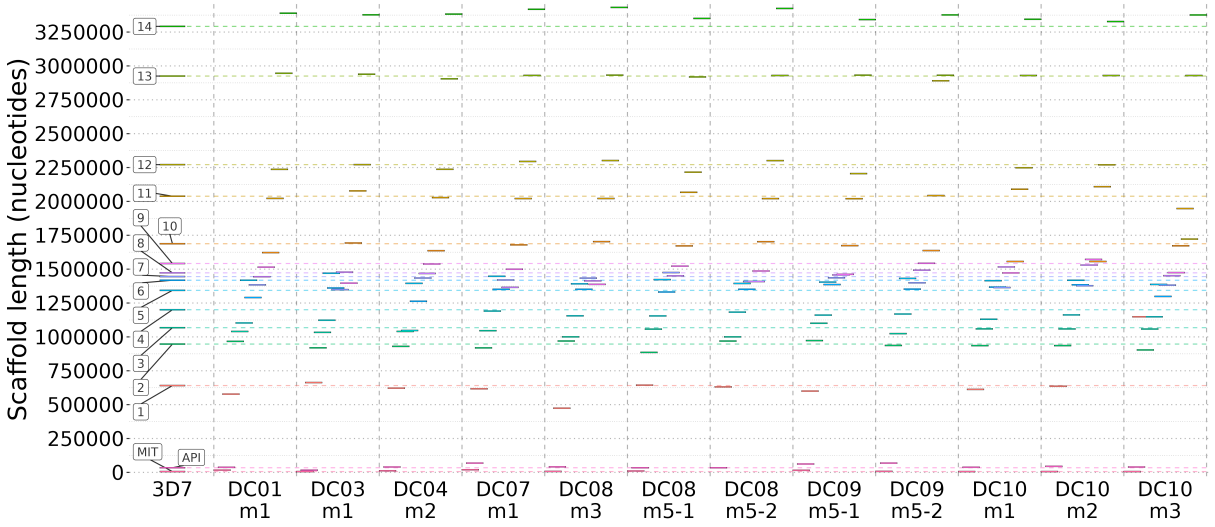
(a) DC05 SNP proportion change.

(b) DC13 SNP proportion change.

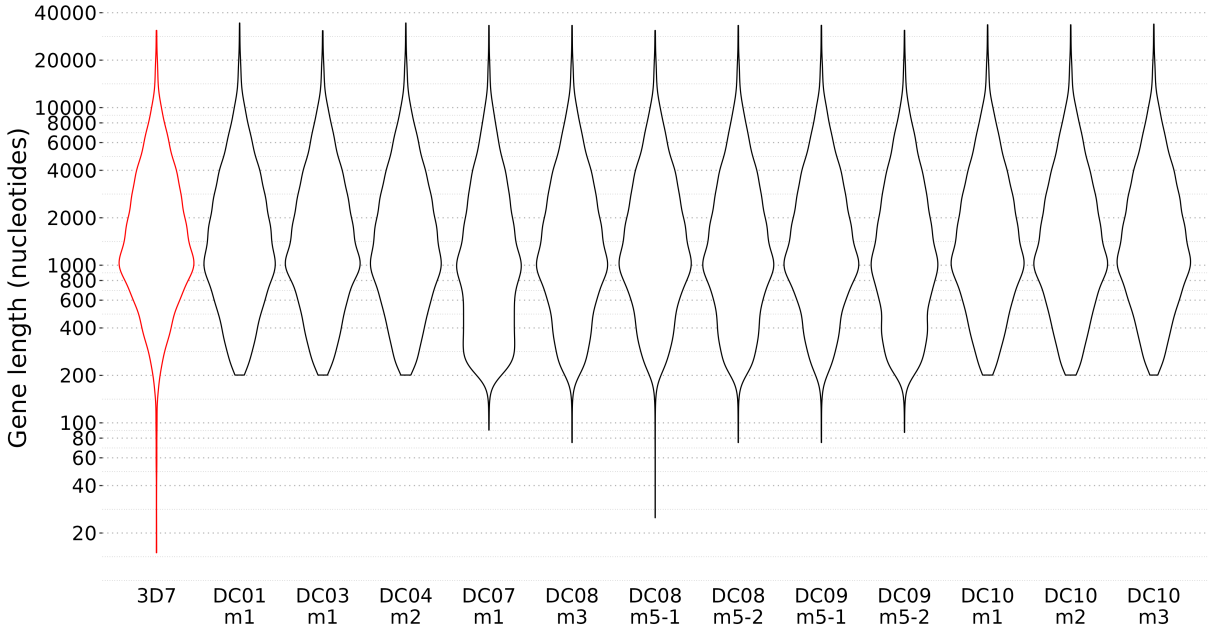
Figure 3.5: **Mutation frequency change during the infections**. The proportion of core genome SNPs was calculated in groups of single-cells corresponding to the blood sampling obtained in months 1 and 6 of individuals DC05 (a) and DC13 (b). The SNPs that have an increase in their proportion between month 1 and month 6 of more than 0.25 are shown in black and those above 0.5 in colours with their annotation.

3.3.5 *de novo* assemblies of parasite genomes from asymptomatic infections

Overall, 7 (DC01, DC03, DC04, DC07, DC08, DC09, DC10) of the 11 asymptomatic individuals had one or more blood samplings cultured for one month, cloned and long-read sequenced to obtain a better resolution of their genomes (Table 3.2). The resulting 12 long-read *de novo* assemblies obtained were made out of 17 to 185 contigs for a total assembly size of 23 to 26 Mb. After the contigs were assembled into scaffolds using 3D7 chromosomes as a reference, all assemblies contained 15 to 16 merged contigs (or scaffolds) for a total size of 23 Mb, similar to the 16 (including apicoplast and mitochondria) 3D7 chromosomes totalling also 23 Mb (Figure 3.6a). The 5600 to 7000 genes identified showed a very similar length distribution to that of the approximately 5500 genes of 3D7 (Figure 3.6b).



(a) Scaffold lengths.

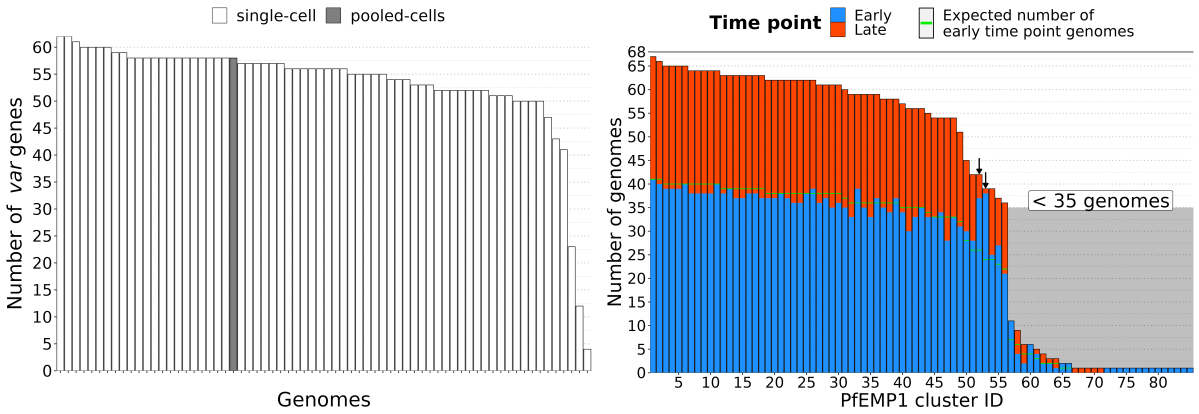


(b) Length of all genes.

Figure 3.6: **Quality control of long-read assemblies**. (a): Scaffolds were built from assembly contigs by merging them using 3D7 chromosomes as a reference. The scaffold sizes of assemblies are compared to the chromosomal sizes of 3D7 with colours corresponding to each of the 16 3D7 chromosomes (including apicoplast and mitochondria) identified during the scaffolding. Compared to 3D7 genome, chromosomes 1 and 12 had unexpected high or low nucleotide sizes for two assemblies, with DC09_m5-2 having no chromosome 1 but a 600 Mb longer chromosome 12 while DC10_m2 had a 500 Mb longer chromosome 1 and 500 Mb shorter chromosome 12. (b): The distribution of the nucleotide size of all genes is also compared to that of 3D7 (red).

3.3.6 Extraction of *var* gene repertoires from individual infections

For one individual (DC05), no genome could be long-read sequenced at any time point, although the infection appears to be monoclonal with a high number of cells available for both early and late time points (Table 3.2 & Figure 3.4). Consequently, an effort was made to obtain the full sequences of all *var* genes from this infection, inspired by previously published methods by Otto *et al.* (2019) and Andradi-Brown *et al.* (2023) [51, 46]. Among the 68 genomes sequenced from both time points, 65 (comprising 64 single-cells and one pooled-cells genome) had more than 40 unique *var* genes, with an average protein length superior to 1500 amino acids (Figure 3.7a & Figure B.2). The 3628 PfEMP1 sequences obtained from all the genomes were combined and categorized into clusters of high similarity (identity > 95 %). Only the longest PfEMP1 was retained for each cluster, leaving a unique PfEMP1 repertoire for both early and late time points combined, referred here as DC05_m1+m6. In total, 85 reference PfEMP1s (labelled 'DC05_m1+m6_1' to 'DC05_m1+m6_85') could be identified among all genomes available in DC05 infection, with 66 % (56/85) present in more than 35 genomes and 22 % (19/85) present in only one genome (Figure 3.7b). While 41 % (12/29) of the rare PfEMP1s (present in less than 35 genomes) were relatively short (< 1000 amino acids), the remaining ones had sizes comparable to the most frequent PfEMP1s, often reaching 1500 amino acids. Interestingly, out of the 56 most frequent PfEMP1s, 2 (DC05_m1+m6_52 and DC05_m1+m6_53) were almost exclusive to the early time point genomes while all the remaining ones were equally shared between both time points.



(a) Number of *var* genes per genome.

(b) Number of genomes containing a PfEMP1 per cluster.

Figure 3.7: **Quality control of short-read assemblies of *var* genes in DC05 genomes**. (a): *var* genes were short-read assembled and grouped by similarity (identity > 95 % and shortest *var* gene covering 90 % of the longest) for each parasite genome (67 single-cells and one pooled-cells) obtained from both time points of the DC05 monoclonal infection. Genomes were sorted from the ones with the highest to the ones with the lowest number of *var* genes. (b): PfEMP1s extracted from every genome were all pooled together and grouped into clusters of high similarity. Clusters of PfEMP1s were sorted from the ones present in the highest to the ones present in the lowest number of genomes. The 85 PfEMP1 clusters are labelled for the rest of this chapter by using their rank in this figure, ranging from 1 to 85 ('DC05_m1+m6_1' to 'DC05_m1+m6_85'). The two arrows highlight PfEMP1 clusters DC05_m1+m6_52 and DC05_m1+m6_53, which are almost exclusive to early time point genomes.

In the 12 long-read assemblies of the 7 individual infections, proteins containing a LARSFADIG motif were extracted and their originating gene was considered as belonging to the *var* family. Overall, the 811 PfEMP1-like proteins (51 to 76 per isolate) obtained from all long-read assemblies had a very similar distribution of size to that of 3D7 61 PfEMP1s (Figure 3.8). Note that the 85 reference PfEMP1s of DC05_m1+m6 (short-read *de novo* assembly) contain only the part of the protein encoded by the exon 1, thus are lacking the roughly 475 amino acids encoded by exon 2.

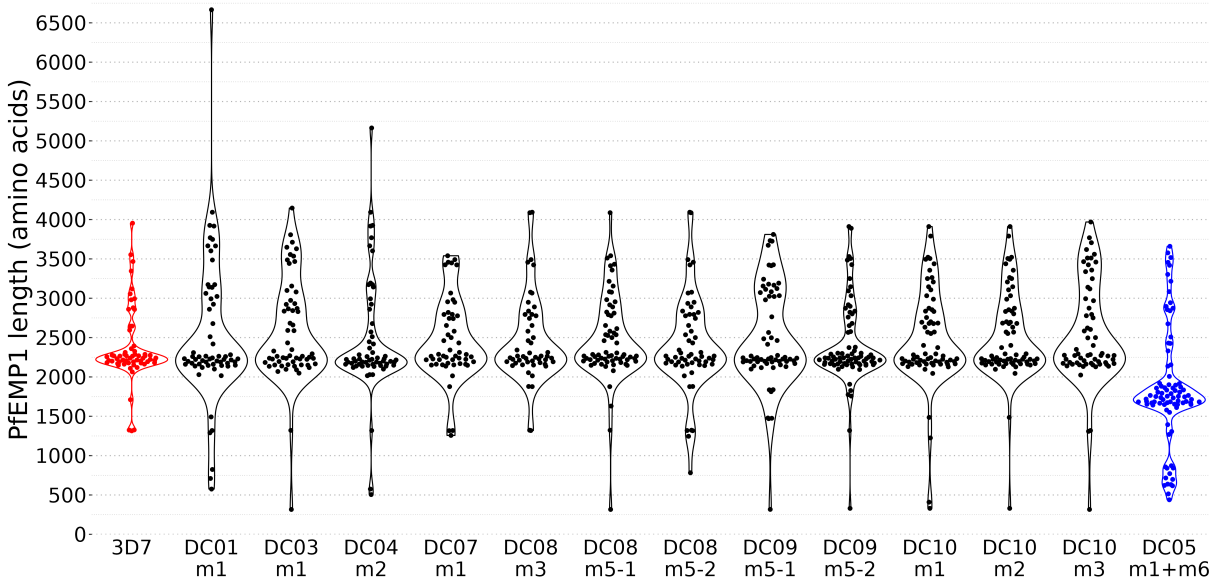
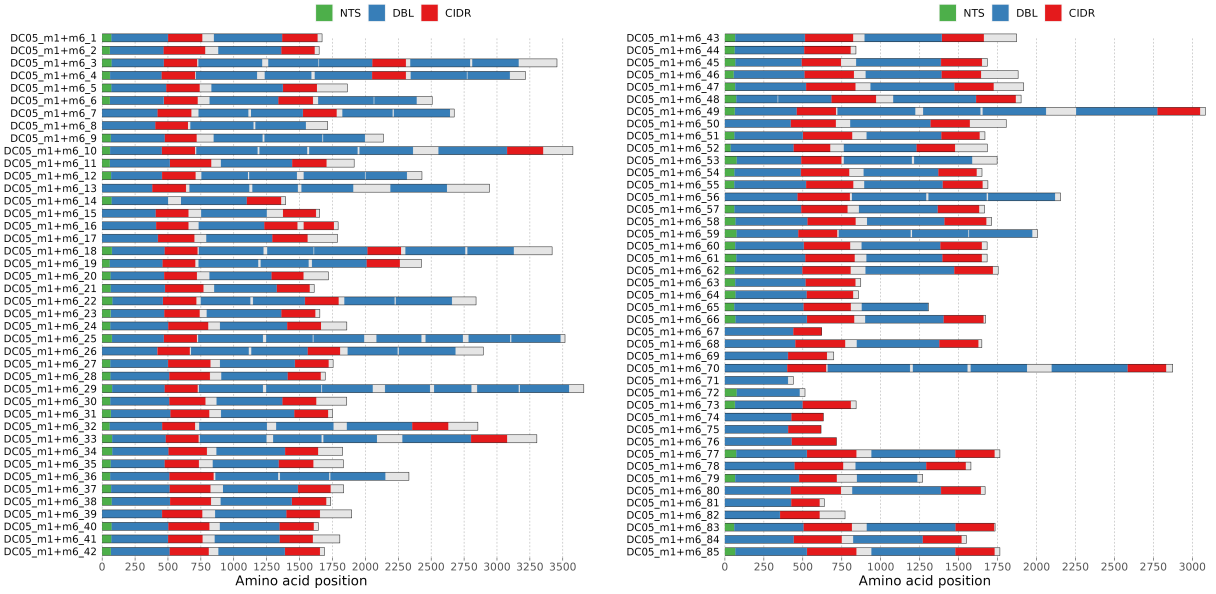


Figure 3.8: **Quality control of *var* gene assemblies**. The distribution of the amino-acid size of PfEMP1s extracted from the 12 long-read assemblies (black) and the unique short-read assembly (blue) is compared to that of 3D7 (red). Of note, only the open reading frames containing the LARSFADIG motif were extracted from PfEMP1s of the short-read assembly, thus excluding the part of the protein encoded by the exon 2 (about 475 amino acids in 3D7).

In PfEMP1s assembled from the long-read assemblies, the domains identified showed that the proteins were generally completely assembled from their two terminal domain segments NTS to ATS, with only 5% (42/811) of the assemblies missing their ATS terminal domain. As expected, each PfEMP1 is a succession of diverse DBL and CIDR domains. In the short-read assembly of DC05_m1+m6, the assembled *var* gene contigs were found to encode almost complete PfEMP1s which contains all domains except ATS and which all starts with a DBLα sub-domain as expected (Figure 3.9 & Figure B.3).



(a) PfEMP1s 1 to 42.

(b) PfEMP1s 43 to 85.

Figure 3.9: **Annotated reference PfEMP1 domains of DC05_m1+m6**. The 85 reference PfEMP1s from individual DC05 were built by *de novo* assembly of the short-reads obtained from the sequencing of one pooled-cells and 67 single-cell samples. For each PfEMP1, the coordinates of NTS, DBL and CIDR domains (the ATS domains are all missing) are extracted from the annotation provided by VarDom. For a better visualization, the first 42 (a) and the last 43 (b) PfEMP1s are shown separately (same order as Figure 3.7b).

3.3.7 Associating late time point genomes with a *var* repertoire

To assess the potential generation of chimeric *var* genes during the course of infection, reads from late time point genomes DC05_m6, DC08_m6 and DC09_m6 were mapped against *var* genes short-read (DC05_m1+m6) or long-read (DC08_m3, DC08_m5-1, DC08_m5-2, DC09_m5-1 and DC09_m5-2) assembled from an earlier time point. DC03_m6 was excluded as its mapping results on the whole DC03_m1 genome assembly showed that the late time point cells were clearly different from the long-read assembly with many SNPs present in the core genome and a poor mapping quality in hypervariable regions.

The different genomes belonging to the same IBD cluster (IBD > 0.9) had always similar proportions of coverage of *var* genes, confirming that these genomes are indeed almost identical to one another (Figure 3.10). The proportion of coverage was generally different between genomes belonging to distinct IBD clusters mapped on the same *var* repertoire, except for the mapping of DC09_m6 genomes on DC09_m5-1 *var* genes, where genomes from three different IBD clusters had a similar proportion of coverage below 0.7. Interestingly, the only DC09_m6 genome that had a higher proportion of coverage on DC09_m5-1 *var* genes was the one sequenced from the pooled-cells, implying that some of

the pooled cells of DC09_m6 had a higher similarity with DC09_m5-1. The highest proportion of coverage was observed for DC08_m6 genomes mapped on DC08_m5-1 *var* genes, DC09_m6 genomes on DC09_m5-2 *var* genes and DC05_m6 genomes on DC05_m1+m6 *var* genes, each resulting in cells from a unique IBD cluster with a coverage proportion of more than 0.9.

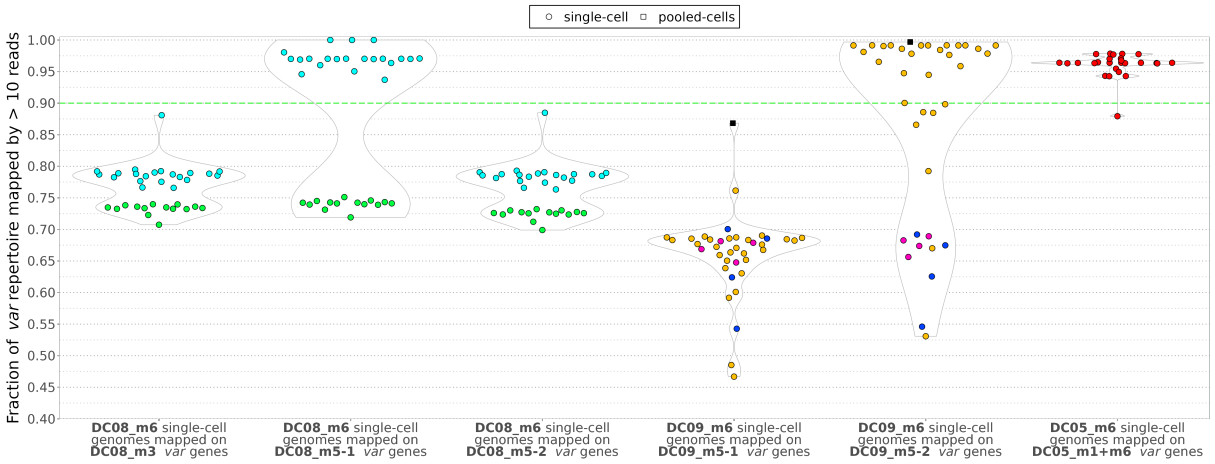


Figure 3.10: **Overall coverage of *var* repertoires from single-cell and pooled-cells reads**. Each single-cell (circle) or pooled-cells (square) genome had their reads mapped on *var* repertoires assembled from long-reads (DC08_m3, DC08_m5-1, DC08_m5-2, DC09_m5-1 and DC09_m5-2 clones) or short-reads (DC05_m1+m6). Genomes were coloured according to the IBD cluster they belong to, with colours identical to Figure 3.4a. In each genome, the fraction of *var* nucleotides that are covered by more than 10 reads serves as a proxy for the number of *var* genes present in each genome. Genomes that have a coverage proportion of the assembled *var* repertoire above 0.9 were considered to contain the same *var* genes as those from that *var* repertoire. For example, isolate DC08_m6 was shown to contain two IBD clusters with an IBD 0.6 with each other, in light-blue and green (Figure 3.4a). All light-blue single-cell genomes mapped almost perfectly against DC08_m5-1 *var* gene repertoire, but not to the other two *var* gene repertoires (DC08_m3 and DC08_m5-2). Therefore *var* genes from light-blue genomes belong to the DC08_m5-1 group, and will be used for further analysis on chimeric *var* gene (*Section 3.3.8 Searching for chimeric var genes generated in the course of infection*).

3.3.8 Searching for chimeric *var* genes generated in the course of infection

The mapping of late time points cells DC05_m6, DC08_m6 and DC09_m6 on a matching *var* repertoire obtained from respectively DC05_m1+m6, DC08_m5-1 and DC09_m5-2 were screened for chimeric *var* genes. The first approach was to use DELLY to identify 'translocation reads' (read pairs mapping each to a distinct *var* gene) potentially indicating a recombination event. For DC08_m6 and DC09_m6, no potential translocation events were detected between *var* exons 1. For DC05_m1+m6, one potential translocation event was found between two *var* exons 1. To check the presence of this translocation earlier in the infection of DC05, early time point single-cells DC05_m1 were mapped on DC05_m1+m6. The screening of chimeric *var* genes revealed the presence of the exact same translocation event in early time point single-

cells, meaning that the translocation did not happen during the course of infection between months 1 and 6.

A second approach to search for chimeric *var* genes involved the screening of the similarity network obtained from the pairwise alignments of all 3628 PfEMP1 sequences assembled from all 67 single-cells and the unique pooled-cells reads in DC05_m1+m6. As chimeric *var* genes are the result of a recombination between two 'parental' *var* genes, we here assume that a chimeric PfEMP1 will share identities with its two 'parents'. Also, the chimeric *var* gene is probably rare in the population, likely found in only one single-cell genome. Pairwise similarities between all PfEMP1s belonging to the 56 most frequent PfEMP1 clusters and the 5 PfEMP1s each found in only one genome from the late time point (DC05_m1+m6_67 to DC05_m1+m6_71) were retained (Figure 3.7b). Out of the 56 most frequent PfEMP1 clusters, 31 contained at least one PfEMP1 that was similar (identity > 70 %) to one of the 5 PfEMP1s of the late time point or to a PfEMP1 from another cluster (Figure 3.11). Out of the 5 PfEMP1s from the late time point, 2 were similar to PfEMP1s from only one cluster and were thus discarded. The remaining three were similar either to 2 or to 3 distinct PfEMP1s and were then potentially generated by a recombination from distinct *var* genes. The first one (DC05_m1+m6_67) was short with only 612 amino acids and was discarded. Another one (DC05_m1+m6_68) was 1649 amino acids long and was similar to 5/88 PfEMP1s belonging to cluster DC05_m1+m6_43 and 29/40 PfEMP1s belonging to cluster DC05_m1+m6_56. The last one (DC05_m1+m6_70) was 2874 amino acids long and was similar to 2/77 PfEMP1s from cluster DC05_m1+m6_18, 62/77 PfEMP1s from cluster DC05_m1+m6_25 and 3/64 PfEMP1s from cluster DC05_m1+m6_29.

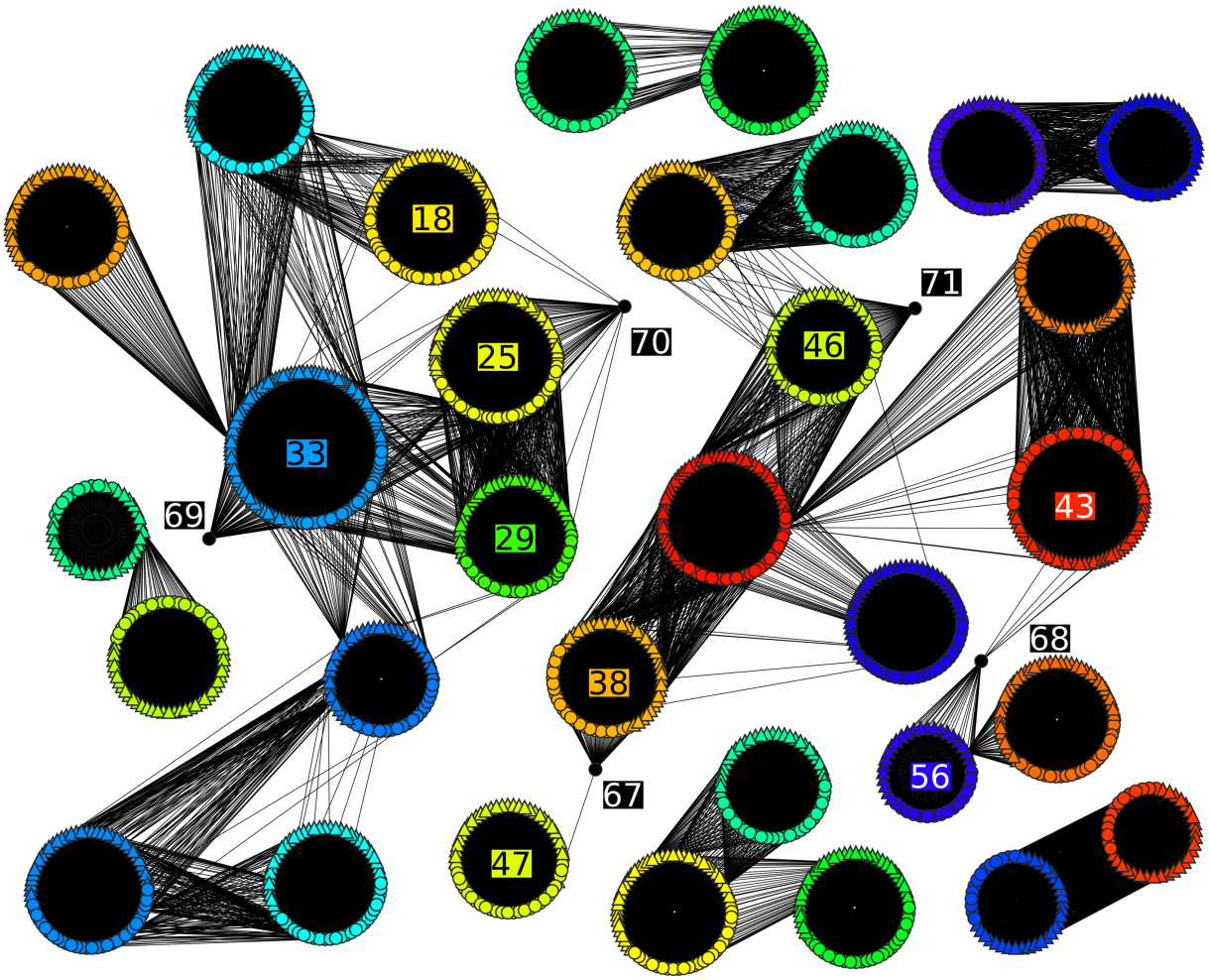


Figure 3.11: **Similarity network of *de novo* assembled *var* genes of DC05_m1+m6**. Overall, there were 56 clusters of nearly identical (identity > 95 %) *de novo* assembled PfEMP1s from two different time points of one monoclonal infection (DC05) that were retrieved in at least 35 of 68 genome sequenced. Only the 31 clusters containing at least one protein that is similar (identity > 70 %) to another one from a distinct cluster are shown. Each PfEMP1 assembled from a genome of the early (triangle) or late (circle) time point is represented by a node coloured by the cluster it belongs to, with in black, the 5 PfEMP1s each present in only one single-cell genome of the late time point. Pairwise similarities between PfEMP1s, represented by edges, are shown only if they are above 70 %. Clusters of PfEMP1s are annotated if they were related with one of the 5 PfEMP1s from the late time point genomes. The network was visualised with Cytoscape (version 3.10.1) [247].

3.4 Discussion

As of today, it remains unclear how *Plasmodium falciparum* parasites persist within the same human host for several months without being eliminated by the immune system. To unravel the mechanisms behind the resilience of these parasites, 11 individuals with *Plasmodium falciparum* asymptomatic malaria lasting 3 to 6 months were closely monitored during the 2016/2017 dry season in The Gambia. Out of these 11 cases, seven individuals had one or more parasite clones sequenced using long-read technology (DC01, DC03, DC04, DC07, DC08, DC09, and DC10), and seven individuals had numerous single-cells sequenced from one or two time points separated by several months (DC03, DC04, DC05, DC07, DC08, DC09, and DC13). Despite the very low parasitaemia observed in asymptomatic infections, both short-read and long-read sequencing demonstrated exceptional quality.

Most of the single-cell mappings yielded either a very good coverage of 3D7 genome (> 80 %) or a very bad one (< 10 %), with very few intermediate values (Figure 3.2b). During single-cell sorting, an 'event' (a detected object interpreted as a single iRBC) could be a false positive that actually contains an uninfected RBC or even another biological material totally deprived of *Plasmodium falciparum* parasite. In this scenario, any available DNA (typically from bacteria) would be amplified and sequenced but would fail to map 3D7 reference genome, which is what is observed here. Single-cell sorting of several blood isolates completely failed with almost no iRBC successfully obtained (Figure 3.2b). This could be caused by a low level of parasitaemia which would increase the risks of sorting false positive events compared to those with a higher parasitaemia. Indeed, the four individuals (DC10, DC11, DC12 and DC13) with the lowest levels of initial parasitaemia ($< 4 \times 10^5$ rings/mL) had all very few single-cells sorted that were of good quality (Table 3.1). Consistent with these observations, mapping quality has previously been shown to depend on parasitaemia levels, particularly below a certain level [270, 271]. Another hypothesis is that, during the roughly 40 hours culturing step after thawing an isolate, some parasites might not develop to schizont stages. The high quality of successfully sequenced single-cells obtained from low to very low parasitaemia levels confirms the applicability of this new sequencing approach, even in cases of asymptomatic chronic low-density infections, without the need to culture parasites to increase their DNA content.

Two monoclonal infections (DC05 and DC13) with multiple cells (respectively 67 and 16 cells) available for both time points were used to determine *de novo* mutations occurring during the infection and likely subsequently provide a selective advantage to parasites. The higher number of SNPs that changed in frequency between the two time points observed in

DC13 (78) compared to DC05 (30) is probably due to the lower number of cells sequenced in DC13 that necessarily increases the variance of allele frequencies in the population (false positives). Furthermore, many SNPs changing in frequency were in linkage disequilibrium (within a few hundreds of nucleotides of each other) and thus not independently generated *de novo* (Table B.1 & Table B.2). In the individual DC05, 3 SNPs had an important change in frequency (> 0.5) in the population of cells during the course of infection. One variant modifies the 3'-UTR region of ClpB1, a protein located in the apicoplast that might be involved in protein conformation when the parasite is subjected to an environmental stress, that could correspond to RBC invasion [272, 273]. A second SNP occurs in the intron of the ribosomal protein L11, a protein located in the mitochondria involved in protein synthesis [274]. The two SNPs were completely absent from the population at the earliest time point, indicating that they might have happened *de novo* during the course of infection. Although the two mutations occurred outside coding region, it is plausible that they modify the level of gene transcription or even the mRNA splicing by preventing the excision of an intron as it may happen in some *Plasmodium falciparum* genes [275]. The third SNP changing in frequency in DC05 infection results in an amino-acid change in the perforin-like protein 1, which enables the permeabilization of iRBCs during the egress, favouring the efficient release of merozoites from the schizont [276]. These three SNPs could be signs of adaptations of the best fitted parasites to their host, especially the one located in the perforin-like protein 1 which is directly involved in invasion. Regarding DC13, three SNPs have a high change in frequency (> 0.5) between cells of the two time points, with one located in an intergenic region and the other one in the 3'-UTR region of the gene encoding the dihydrouridine synthase. The third SNP is present in one exon of *cg1*, that is in linkage disequilibrium with the *pfcrt* gene responsible for chloroquine resistance [277, 167]. Another SNP inducing a non-synonymous mutation in the same gene was previously found to be a strong marker of differentiation between parasite populations from Guinea and The Gambia, suggesting that *cg1* generally exhibits a high polymorphism [278]. However, a severe limitation of our study is the limited number of single-cell genomes available from DC13, 10 and 6 cells from the early and late time points, respectively. It is premature to speculate whether these six variants might confer a direct growth advantage during an infection, especially considering that they have not been identified in other infections in our dataset. Nonetheless, our results demonstrate that, for the first time, we can identify genetic diversity between single-cell genomes even within monoclonal infections. Moreover, the proportion of such 'mutant genomes' significantly fluctuated over time during the chronic infection. The driving force behind this selection could be a genetic variant or, alternatively, an epigenetic factor, such as

the expression of a specific PfEMP1 protein. Future analyses will have to focus on assessing all SNP allele frequencies at the population level using available *Plasmodium falciparum* genomes from West Africa. This analysis will identify alleles that significantly change in frequency during low and high transmission seasons, similar to what has been observed with the *crt* alleles in The Gambia [249]. The hypothesis to be tested is whether the genes identified in DC05 and DC13 also undergo changes in allele frequencies with seasonality at the population level.

The generation of long-read assemblies of parasites from asymptomatic infections was highly dependent on the ability of parasites to survive *in vitro* conditions during their culture adaptation. Furthermore, the high clonality of some infections render the correct assembly of reads more complex. Although multiple long-read-based tools are designed to accurately retrieve full genomes from a strain mixture, they are generally developed for bacterial communities, which contain single circular chromosomes of about 1 Mb, making them much easier to assemble than the 14 nuclear chromosomes of *Plasmodium falciparum* totalling 23 Mb [279, 280]. An alternative would be to use haplotype resolution used in long-read sequencing of non-haploid organisms, but again these methods are mostly limited to diploid organisms and often fail in low complexity regions. Had the analysis been focused on the core genome only, segregating the different strains into different clonal cultures would not have been necessary. However we were interested in the hypervariable family of *var* genes for which paralogs can not be identified between strains. The quality of the long-read assemblies obtained from these cultured-adapted isolates derived from asymptomatic infections was very high, equivalent in terms of number of contigs and genes identified with PacBio long-read assemblies obtained from lab strains (Figure 3.6) [52, 42]. In particular, the hypervariable family of *var* genes was successfully assembled for all genomes as indicated by their number per genome, their size and the domain organisation of the PfEMP1 that they encode (Figure 3.8). These more than 800 new *var* gene sequences will be added to the database of genes involved in antigenic variations, varDB [281]. Future analysis includes determining the orientation, location and upstream sequences of these *var* genes. Although assembly contigs were for the most part correctly identified to 3D7 chromosomes, an assembly error was observed in DC09_m5-2 and DC10_m2: an incorrect merging between chromosomes 1 and 12 (Figure 3.6a). Long repetitive sequences present at the edges of all chromosomes, called telomere associated repetitive elements (TARE), might induce these false merging between contigs originating from two chromosomes [282]. The assembly contigs and the annotation of all the genes are currently being improved by Mathieu Quenu[1],

including the assembly of the hypervariable family of *rif*.

Thanks to the availability of single-cell genomes, the relatedness between polyclonal individuals can be precisely described using their different IBD cell clusters (DC07 with DC08, or DC03 with DC09), whereas this would have been difficult to achieve with bulk sequencing (Figure 3.4). While tools like DEploid are able to retrieve haplotypes of the different strains in a polyclonal sample bulk sequenced, they are generally unable to discern strains that are too related or to detect those that are too rare [283, 59].

As one monoclonal infection (DC05) had many single-cell successfully sequenced from both time points, *de novo* assembly of *var* genes from short-reads was attempted. By building short-read *de novo* assemblies of *var* genes in each individual cell and merging them later by similarity, we obtained a unique set of 56 long (about 1500 amino acids) and frequent (> 35 cells) PfEMP1s (Figure 3.7 & Figure B.2). The customized *de novo* assembly pipeline was inspired by published works of Otto *et al.* (2019) and Andradi-Brown *et al.* (2023), and notably contains a step where reads are assembled with rnaSPADES, a tool developed specifically for transcriptome assembly rather than genome assembly [51, 46, 265]. This choice was made after attempts with classical genome assemblers such as MaSuRCA failed to produce a high quality *var* repertoire [284]. Also, because the *var* gene family consists in tens of genes spread in a single *Plasmodium falciparum* genome that share multiple homology blocks, they can be assimilated as much fewer genes (like the three main ups groups) possessing multiple transcript isoforms. In the end, the resulting *var* repertoire contained *var* genes with similar sizes as 3D7 and other assemblies when excluding the exon 2 (Figure 3.8). Furthermore the location of annotated PfEMP1 domains and sub-domains were consistent with what is expected for PfEMP1, confirming that the assembly pipeline did not generate *var* sequences that are biologically meaningless (Figure 3.9 & Figure B.3). The multiple protein alignment performed by clustal-omega includes gap opening penalties but no gap extension penalties, as such, two *de novo* assembled *var* genes could be considered highly related (identity > 95 %) even if they have different protein sequences that contain long identical fragments, provided that the alignment is not too much fragmented. However this behaviour is necessary to be able to merge identical proteins that were partially assembled with those more complete.

The mapping of the short-read sequences from the late time point cells on *var* repertoires obtained from long-read or short-read assemblies of earlier samplings was carried out with the aim of identifying single-cell genomes containing a near identical *var* repertoire to the

[1]LPHI, MIVEGEC, Université de Montpellier, CNRS, INSERM, Montpellier, France

one from a long-read assembly. Most of the cells within the same IBD clusters exhibited nearly identical coverage percentage of each *var* repertoire, confirming the effectiveness of IBD in segregating cells based on their similarity (Figure 3.10). In *var* repertoires obtained from DC08_m5-1, DC09_m5-2 and DC05_m1+m6, single-cells from the same IBD cluster showed coverage proportions above 90 %, suggesting that the clone that was long-read sequenced was from the same parasite line as the cells clustered. This also shows that the assemblies of *var* genes was successful both with short and long-reads.

Our search for chimeric *var* genes using the mapping of the cells which could be associated with a *var* repertoire at an earlier time point yielded no highly confident hit. This could be explained by multiple reasons. First, long-read assembled *var* repertoires (DC08_m5-1 and DC09_m5-2) matched single-cell clusters distant by just one month in infection time. As chimeric *var* genes are relatively rare and are not necessarily conferring a selective advantage because of the mutual exclusive expression feature of the *var* family, one month might not be enough to be able to find a recombination event. Based on the in vitro recombination rate, we expect in the order of half of all single-cell genome to have generated at least one chimeric *var* gene [54]. Also, we were hampered by the lack of long-read sequencing genomes for some isolates, more of them are currently being sequenced. Due to time constraint, some analyses could not be repeated. It would be better to perform the mapping on the entire long-read assembly (this was done for DC03 only) to confirm that the cells and long-read sequenced clone are truly identical, by showing that there are almost no SNPs detected throughout the whole genome for instance. The mapping of single-cell reads of DC05_m6 cells was done on all 85 unique *var* genes *de novo* assembled which could already contain some chimeric *var* genes already. Single-cell reads should have been mapped on just the 56 most frequent *var* genes instead of all 85 so that reads have a chance to indicate a translocation between two *var* genes instead of mapping directly to the potential chimera.

Another approach focused on looking for chimeric *var* genes using the similarity network of PfEMP1s obtained from the short-read *de novo* assembly of *var* genes from DC05 early and late time point genomes. Out of the 85 reference PfEMP1s, 5 were present in a single parasite genome of the late time point: this rarity could be a sign of a chimeric *var* gene generated during the infection and thus absent from the early time point. Three of these rare late time point *var* genes were similar to 2 or three different clusters of frequent (present in > 35 genomes) *var* genes which would have been expected if these were indeed chimeric *var* genes (Figure 3.11). Further screening is obviously needed to confirm or infirm that chimeric *var* genes were identified. For example, the differential location of identical genomic fragments between the potential chimera and the two parents may indicate where the

breaking points occurred between the two parental *var* genes. Out of the 56 most frequent *var* genes, 2 (DC05_m1+m6_52 and DC05_m1+m6_53) were almost exclusive to the early time point genomes while all the remaining ones were equally shared between both time points (Figure 3.7b). An interesting hypothesis could be that they recombined with one another in one parasite genome just after the blood sampling of the early time point and yielded two new *var* genes afterwards. However, if this was the case, there should be two frequent *var* genes that are exclusive the late time point single-cells, which was not observed here.

Even though no chimeric *var* gene generated during the course of an infection could be definitely validated, this important and diverse quantity of genomic data is still at the dawn of its full exploration. As for now, only a tiny fraction of the high quality long-read assemblies and the numerous single-cell genomes has been exploited and many subsequent interesting analyses will most certainly follow this work.

3.5 Acknowledgements

We would like to thank Prince Nyarko, Catherine Jett, Ian Cheeseman, Will Hamilton, Mathieu Quenu and all fieldworkers involved in the study. The authors would like to thank the staff of Wellcome Sanger Institute Sample Management and Tree of Life teams for their contribution. The authors acknowledge the ISO 9001 certified IRD i-Trop HPC (member of the South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: https://bioinfo.ird.fr/ - http://www.southgreen.fr.

Conclusion and Outlooks

The eradication of malaria implies the destruction of very single *Plasmodium falciparum* parasites, even from low parasite density infections that are asymptomatic and typically below microscopy detection level. These 'silent infections', that act as a reservoir, will be challenging to eliminate. Yet the vast majority of malaria research so far is based on a handful of laboratory-adapted strain or clinical isolates from symptomatic infections. The motivation behind the Gambian cohort project was to decipher asymptomatic chronic infections at the genomic and transcriptomic level, and from single-cell to population level, with a particular focus on the parasite variant surface antigens.

This thesis focused on *Plasmodium falciparum* genetics, both at the population and the individual level by sampling asymptomatic infections from four nearby villages of The Gambia. A wide variety of sequencing methods were employed to access the parasite genomes, including SNP barcode genotyping, bulk WGS, single-cell short-read sequencing and finally clonal long-read sequencing. The low parasite densities, implying a low amount of DNA, were the main challenge to overcome to gather genomic data. Secondly, the polyclonality of many field isolates adds a layer of complexity for data analysis. Thirdly, the *Plasmodium falciparum* genome itself is inherently difficult to study because of the extremely high AT-richness and high variability within VSAs, particularly *var* genes.

Choosing the right methods to compare genomic sequences is essential to draw biological meaningful conclusions. Using IBD, relatedness between parasite genomes can be obtained in an informed manner by including the recombinatorial characteristics of *Plasmodium falciparum*. Here, this method was able to show that parasites were more similar when sampled within 3 months and in individuals who live close to each other. This trend also appeared in the single-cell genome dataset, with individuals living in the same household being infected with genetically-related *Plasmodium falciparum* isolates. Two other factors affect the overall relatedness of population of parasites. First, antimalarial drugs have the effect of decreasing the haplotype diversity of parasites, as the most shared regions between parasites occurred near drug resistance markers. Secondly, the cycle of seasons, hence the intensity of transmission, also affects the number of recombinations that generate the parasite genetic diversity.

To dig deeper into parasite genomic features, 42 individuals were selected to have their infections more thoroughly analysed. The main focus was made on chronic infections lasting several months, with much lower chances of reinfection during the low transmission season. To the best of our knowledge, we gathered the highest number of WGS derived from asymptomatic infections, the first ones sequenced at the single-cell level and the largest collection of long-read sequence genomes. This unique dataset allowed us an unprecedented

characterisation of the parasite genome and its changes over time, by analysing interactions between genotypes within a host, and *de novo* mutations within a single clonal lineage. First, we are now able to identify *de novo* SNPs that occur during the course of an infection and analyse their frequency over time. Second, we can now test the hypothesis that mitotic recombinations between the highly immunogenic *var* genes may occur during the course of an infection.

For two monoclonal infections with single-cells available at both early and late time points of their infection, several mutations increased in frequency. To trace back the history of these mutations, the next step will be to segregate cells into haplotype groups and identify the change in proportion between each of these haplotype groups. To address the potential selective advantage of these SNPs, their allele frequency in each haplotype group will have to be checked. Bulk genomes were obtained previously on most blood isolates from both wet and dry seasons (*Section 2.2.3 Genotyping and sequencing*), including some of the blood isolates that were single-cell sequenced. The allele frequencies of SNPs from the bulk genomes could be compared to the ones obtained from the single-cells of the same blood isolates. Additionally, the whole set of bulk genomes could be used to check if the allele frequencies of the identified mutations from single-cell genomes are also different between low and high transmission seasons, suggesting a role in parasite adaptation to chronic infections. Finally, the F_{WS} and R_H metrics calculated on bulk samples could be compared with the IBD clusters from the single-cells.

Due to time constraint, we have not yet investigated whether the duration of infection of a specific *Plasmodium falciparum* genotype can be inferred from the number of *de novo* mutations in single-cell genomes. As of now, there is no known method for estimating the age of a malaria infection, starting when sporozoites were injected by the mosquito vector. Using the known constant mutation rate measured *in vitro* (*Appendix C.1 Mutation rates*), one can infer the duration of the infection based on the accumulation of *de novo* mutations since the last round of meiosis in the mosquito. We expect that the older the infection, the more *de novo* mutations (SNPs or INDELs) would have accumulated in each single-cell genome. We will easily validate our data with the known time elapsed between time points (5 to 6 months between early and late time points). This original idea could have applications in epidemiological studies, for example to differentiate a treatment failure from a new infection after a mass drug administration campaign.

During my thesis, I developed two pipelines that may be used in the future for other related works. The first one enables the comparison and merging of genotyping and whole genome sequencing data acquired on the same blood isolate. It extends into the calculation of

the relatedness between genomes, their complexity of infection and the retrieval of parental-offspring related genomes. The second pipeline successfully assembles *var* genes from individual parasite genomes single-cell sequenced and merges them by similarity, obtaining a *var* gene repertoire for an infection. Although an even more accurate method to obtain *var* gene repertoires was to perform long-read sequencing, this required that parasites were able to grow *in vitro* and was costly compared to short-read sequencing. To further validate the high quality of the single-cell *var* gene assembly pipeline, it could be also run on single-cell genomes that are identical to a long-read assembled genome so that the two repertoires can be compared. All these *de novo* assembled *var* genes from either long or short-reads will be added to the database of known *var* genes, varDB. This will be particularly useful for improving the precision of Varia, a software that predicts a full length *var* gene based on DBLα expression tag.

Additional Material of Project 1

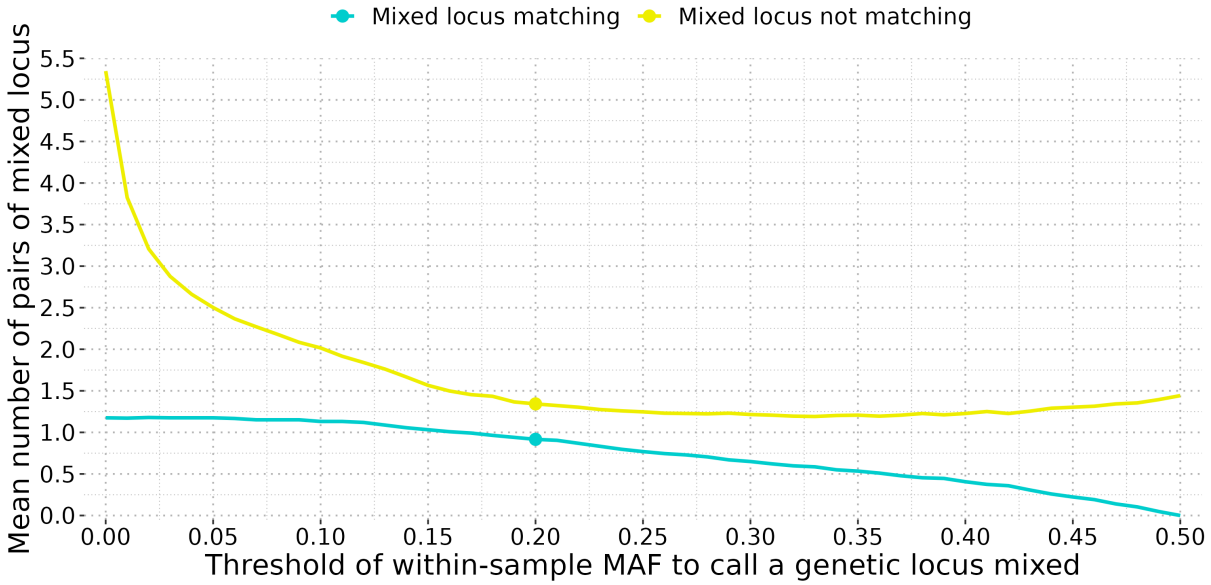


Figure A.1: **Determination of the minor allele frequency to call a locus mixed** (*Section 2.2.6 Multi-locus genotype barcode data analysis pipeline*). Comparisons of mixed loci between each pair of molecular barcode obtained by genotyping and their corresponding genomic barcode built from allelic frequencies obtained from WGS. WGS loci were considered mixed if their within-sample Minor Allele Frequencies (MAF) were above a threshold ranging from 0 to 0.5. The threshold of within-sample MAF of 0.2 was finally retained because it yields to a high number of matches and a low number of mismatches of mixed loci, which makes the genomic barcode as close as possible to the molecular one.

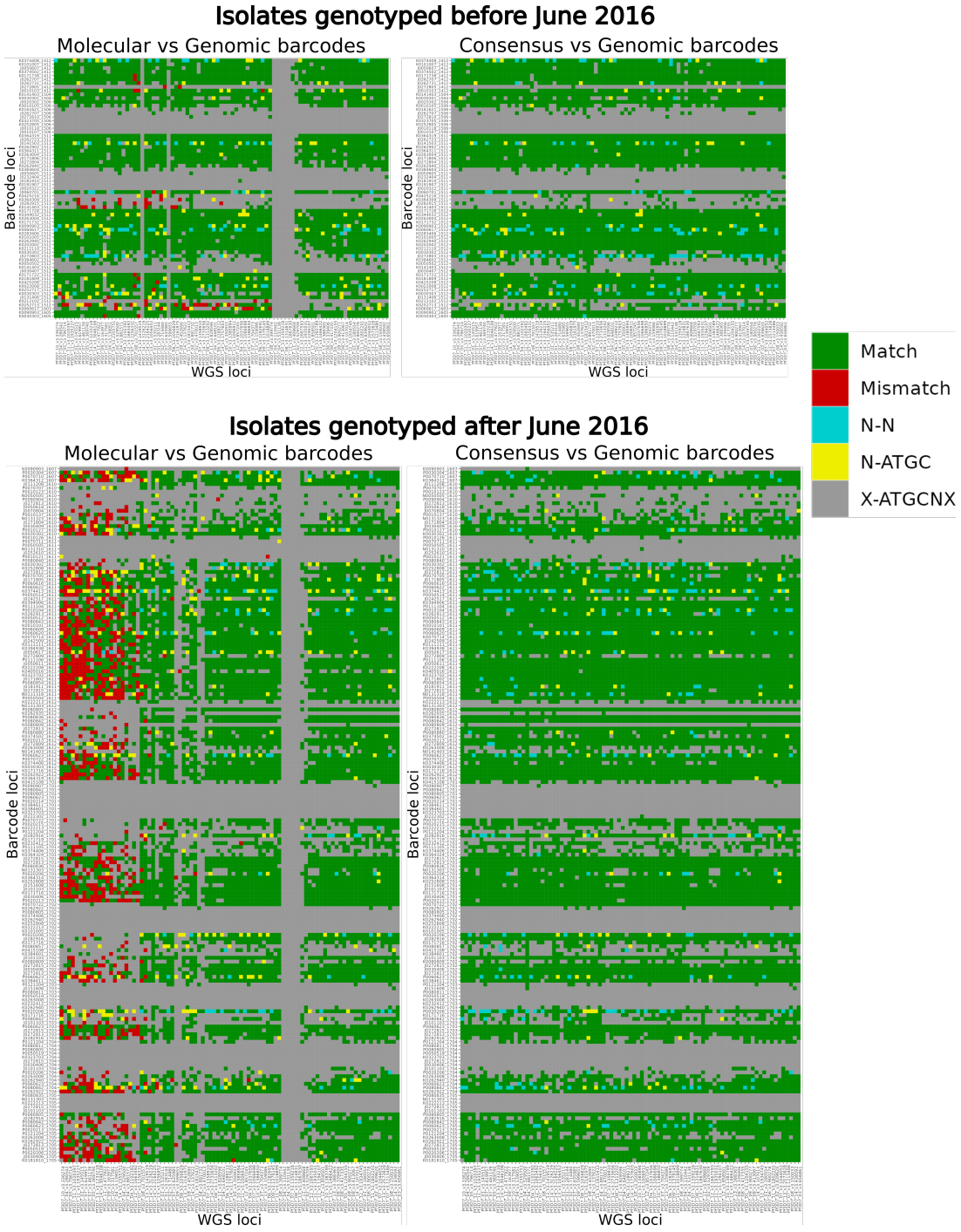
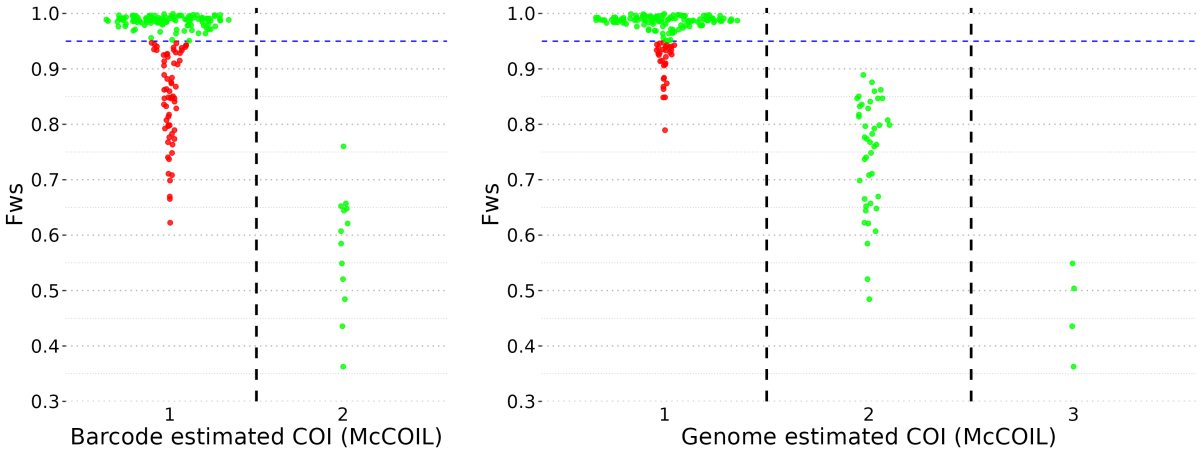


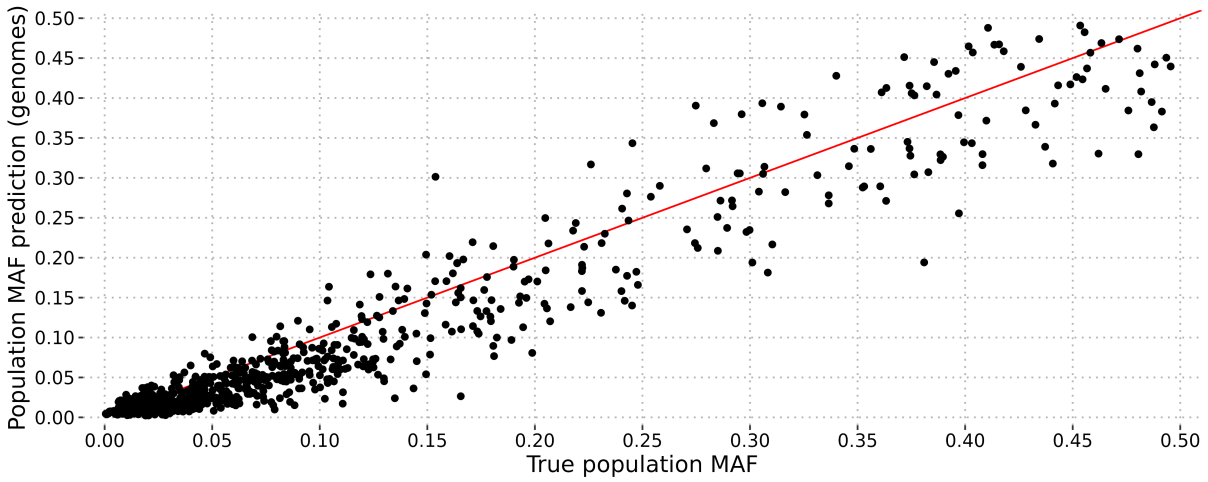
Figure A.2: **Molecular and genomic barcodes discrepancies** (*Section 2.2.6 Multi-locus genotype barcode data analysis pipeline*). Pairwise comparison of molecular barcodes loci (built from genotyped SNPs) and consensus barcodes loci (combining molecular and genomic barcodes) with genomic barcodes loci (built from WGS loci) that were matching (green), not matching (red), matching a mixed call (blue), not matching a mixed call (yellow) or incomparable when at least one call is unknown (grey). WGS loci were considered mixed if the within-sample MAF was above 0.2 (Figure A.1). Overall, loci match between the two methods when the genotyping was done before June 2016 while 21 genotyped loci show a consistent mismatch with WGS loci for all samples genotyped after June 2016. As such, all those genotyped loci are likely incorrect calls that would result from a protocol change. To increase the number of available loci from barcodes, all unknown and incorrectly genotyped loci were replaced by WGS loci, resulting in what we refer to as 'consensus barcodes'.



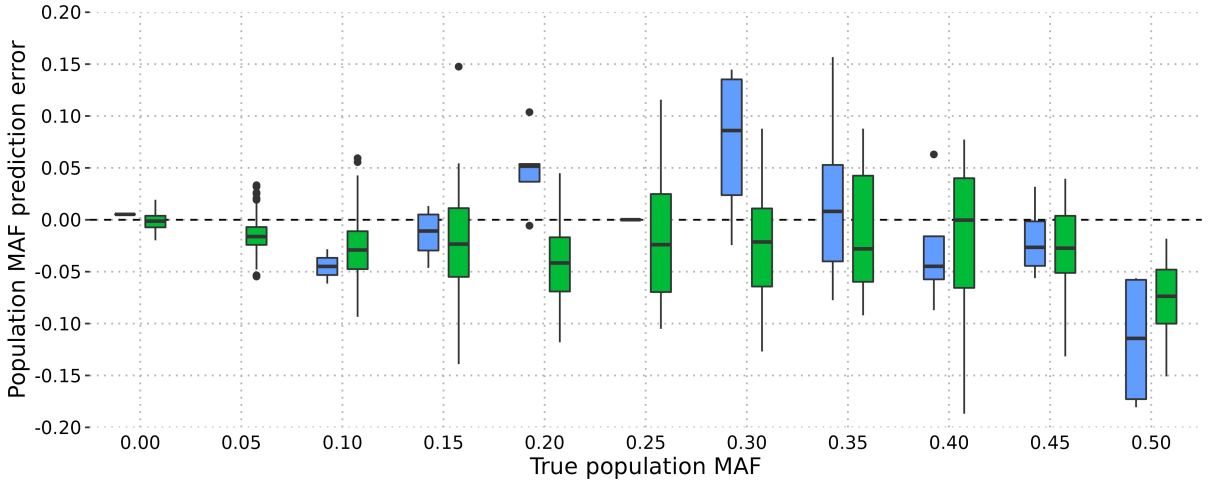
(a) COI estimated on barcodes (44 loci).

(b) COI estimated on genomes (819 loci).

Figure A.3: **Distribution of F_{WS} values grouped by the value of clonality** (*Section 2.3.2 High genetic complexity in asymptomatic infections*). The clonality (COI) was obtained from THE REAL McCOIL on barcodes (44 loci) (a) and genomes (819 loci) (b). F_{WS}, limited to genomic data, is a within-isolate fixation index ranging between 0 and 1 and is inversely proportional to the level of clonality. A value over 0.95 suggests that the isolate has only one genotype, while lower values indicate that the isolate contains mixed genotypes. All barcodes and genomes that had a predicted clonality of two by THE REAL McCOIL had a F_{WS} value below 0.95. Genomes predicted with a clonality of three had even lower F_{WS} values than those with a clonality of two. Overall, 16 % and 21 % of the monoclonal isolates called by THE REAL McCOIL were polyclonal according to F_{WS} (F_{WS} < 0.95) for barcode (65/394) and genomic (32/154) data. These percentages dropped to 10 and 5 % respectively for barcode (41/394) and genome data (8/154) when F_{WS} values were below 0.9. These evidences confirm that although the higher number of loci available with genomic data better correlates with F_{WS} values, the only 44 loci available with barcodes are already enough to get an accurate prediction of clonality.



(a) Prediction accuracy of population MAF.



(b) Prediction error of population MAF.

Figure A.4: **Population minor allele frequency prediction** (*Section 2.3.2 High genetic complexity in asymptomatic infections*). The population Minor Allele Frequency (MAF) was estimated by the categorical method of THE REAL McCOIL using consensus barcodes (44 loci) or genomes (819 loci). The true population MAF was calculated from WGS data. (a): Population MAF prediction accuracy from THE REAL McCOIL on genomes. (b): Population MAF prediction error from THE REAL McCOIL on consensus barcodes (blue) and genomes (green).

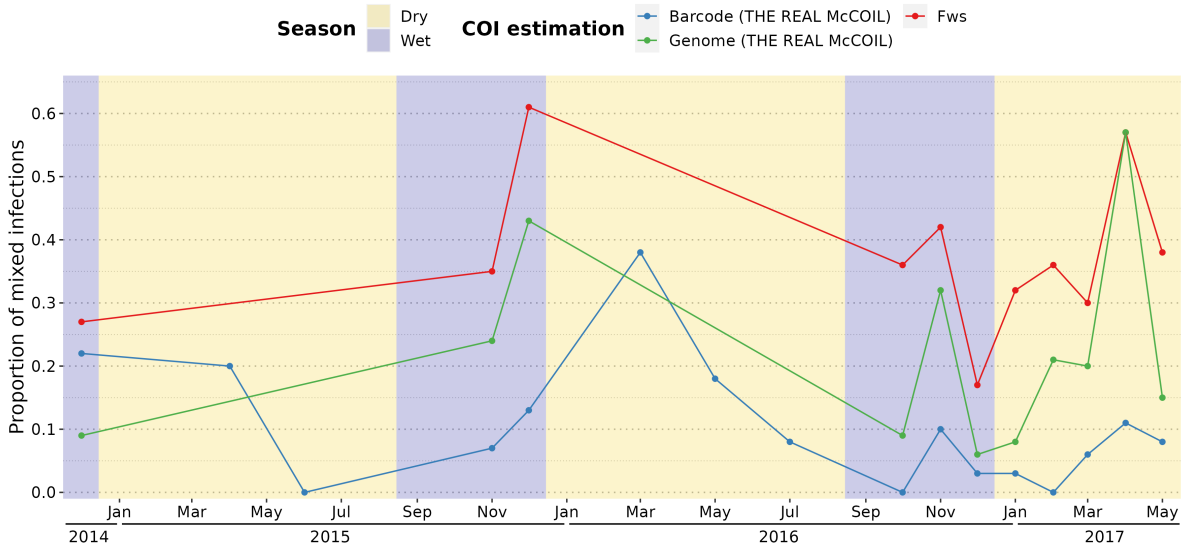


Figure A.5: **Proportion of polyclonal isolates over time** (*Section 2.3.2 High genetic complexity in asymptomatic infections*). Proportion of polyclonal isolates over time estimated by F_{WS} (F_{WS} < 0.95) and THE REAL McCOIL on either molecular (44 loci) or genomic (819 loci) barcode. The percentage of polyclonal isolates is stable over time for all methods (around 39 % for F_{WS}, 9 % with THE REAL McCOIL on a 'molecular' barcode, 23 % with THE REAL McCOIL on a 'genomic' barcode). From the December 2016 screening, a cohort of 42 *Plasmodium falciparum* positive asymptomatic individuals was selected, as previously described [24].

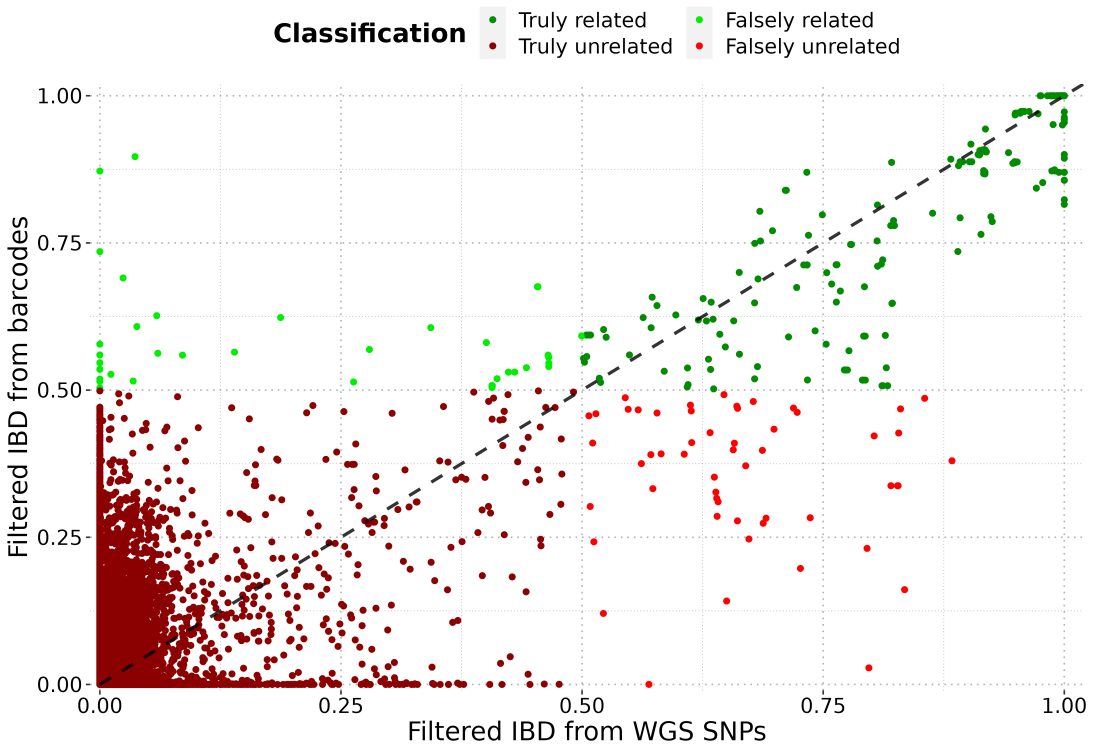


Figure A.6: **Agreement of pairwise IBD values calculated from barcodes or from genomes** (*Section 2.3.2 High genetic complexity in asymptomatic infections*). For each pair of samples, IBD from genomic data is considered the gold standard. All pairs are coloured by their relatedness agreement with a cutoff of 0.5 of IBD. Among the 347 pairs of samples with barcode IBD above 0.5, 303 were indeed related (WGS IBD above 0.5, dark green) while 44 were actually not related (WGS IBD below 0.5, bright green).

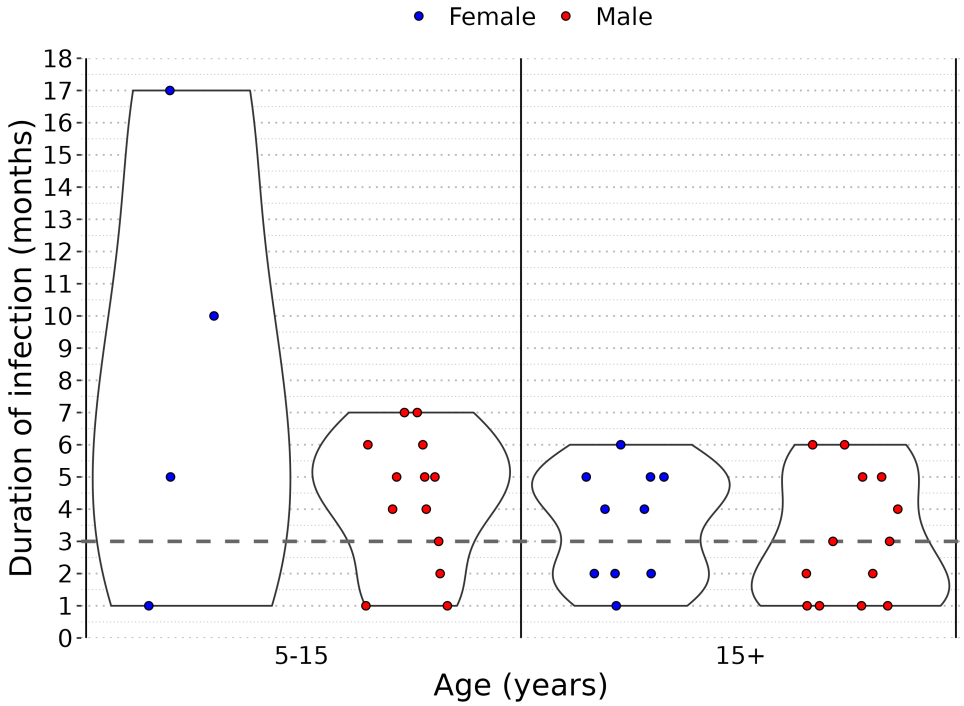


Figure A.7: **Durations of infection by the same parasite strain** (*Section 2.3.4 Plasmodium falciparum chronic infections with persisting genotypes*). The number of short (less than 3 months) and long infections (more than 3 months) were compared between genders (female versus male) and age groups (5-15 years old versus 15 years old or older). No significant difference of the duration of infection was found between genders, which had very similar percentages of short infection with 36 % and 42 % of infections for female and male respectively (χ^2 value = 0.16484, p value > 0.6). Long infections were more common in the group 5-15 years old (76 % of all infections) compared to the 15+ age group (48 % of all infections) but this was not statistically significant (χ^2 value = 3.3419, p value < 0.07) due to relatively small sample size.

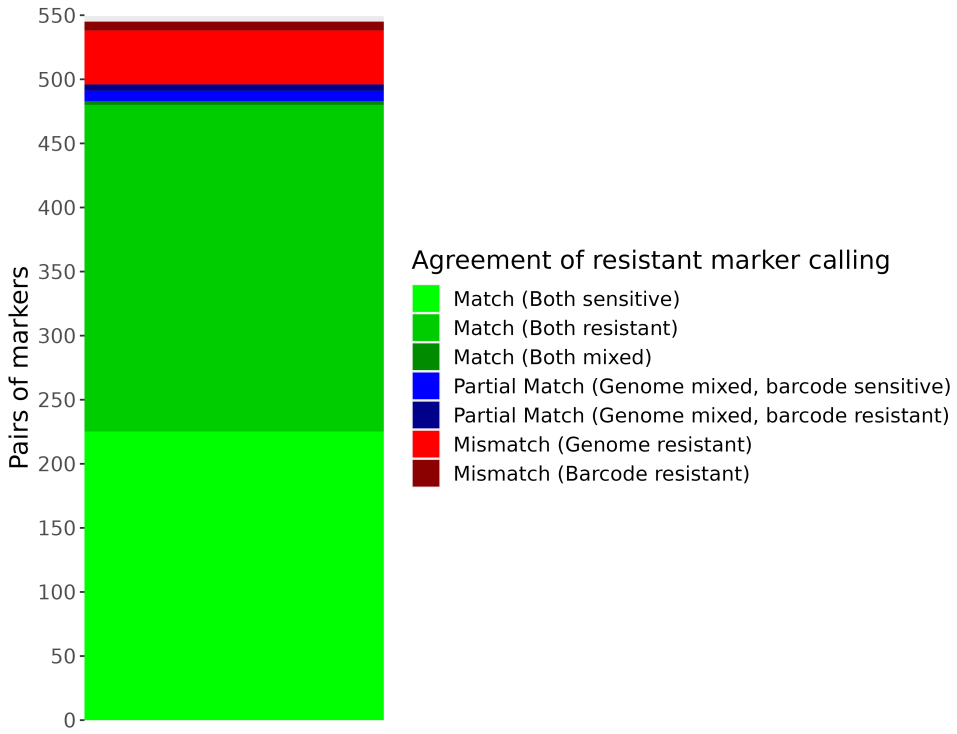


Figure A.8: **Calling agreement of drug resistance-related haplotypes obtained from genotyping and whole genome sequencing** (*Section 2.3.5 Independence of seasonality and drug resistance prevalence*). The vast majority (483; 89 %) of the 545 pairs of haplotypes obtained with the two methods are identical, which is an evidence that both methods are accurate. While there are multiple partial matches (one haplotype is mixed according to a method and either sensitive or resistant by the other) with mixed haplotypes called from genomes, no partial match with mixed haplotypes called from barcodes were found, showing that WGS is more sensitive than molecular genotyping.

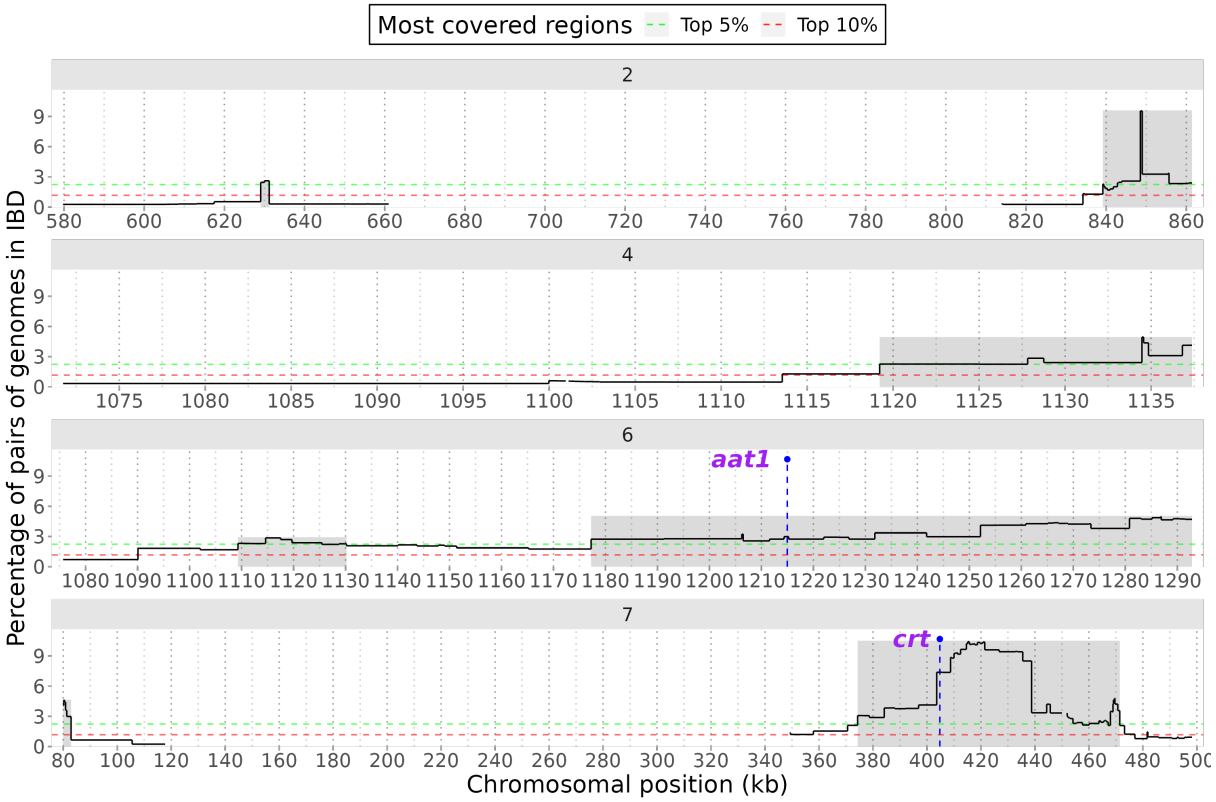
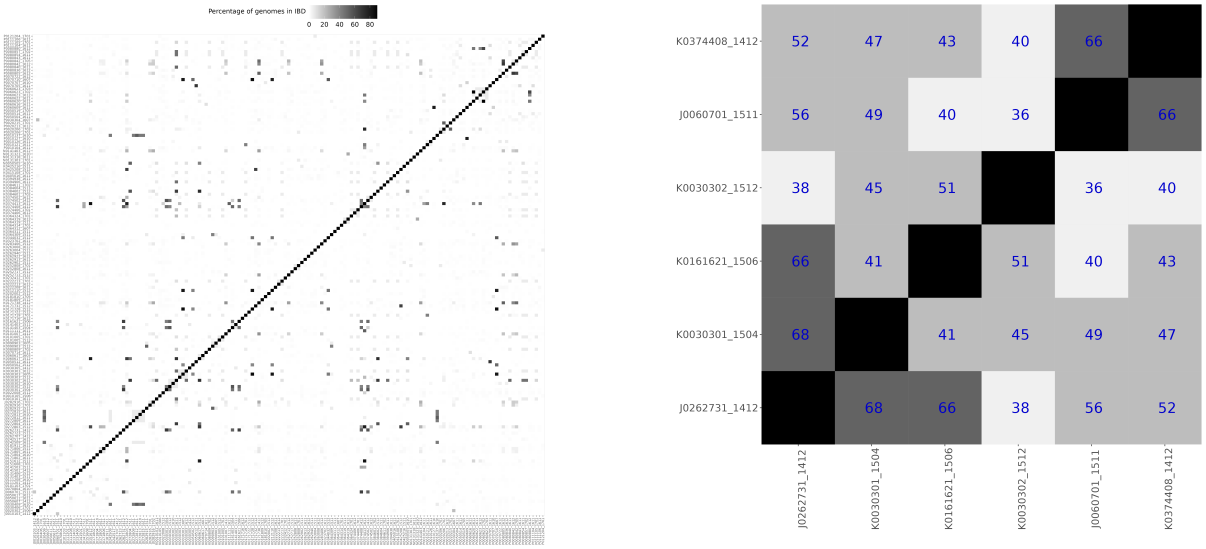


Figure A.9: **Most shared genomic fragments between non-identical genomes** (*Section 2.3.6 Signature of selection around drug-resistance markers*). Seven regions (2:629060-631190, 2:839178-861423, 4:1119215-1137378, 6:1109310-1130185, 6:1177239-1292831, 7:79935-82889, 7;374311-471442) were identified by extracting the top 5 % most shared fragments (with at least 2.2 % - 440/19700 - of pairs, green dotted line) and merging them when they were less than 10 kb apart. The percentages of pairs of genomes in IBD in these regions are shown along with those of their 50 kb flanking regions. For comparison, the top 10 % coverage throughout the whole genome is also shown (1.2 % - 231/19700 - of pairs, red dotted line). Two out of the seven regions are surrounding genes known to reduce the susceptibility to multiple antimalarials: *aat1* and *crt*.

Figure A.10: **Identical chromosomal regions most frequently detected** (*Section 2.3.6 Signature of selection around drug-resistance markers*). Genomic regions identical by descent between unrelated genomes (IBD < 0.5) using only pairs of genomes (at least 100 pairs) both with (Res) or without (WT) the same drug resistance-related haplotype among *aat1* S528L (WT: 100 pairs; Resistant: 12437 pairs), *crt* K76T (WT: 1759 pairs; Resistant: 7091 pairs), *dhps* A437G (WT: 3015 pairs; Resistant: 4661 pairs) and *mdr1* N86Y (WT: 12291 pairs; Resistant: 143 pairs). The haplotypes *dhfr* S108N, virtually fixed in the population, and *kelch13* C580Y, absent from the population, are not shown. The 14 *Plasmodium falciparum* chromosomes were annotated with the 6 genes known to reduce susceptibility to antimalarials (in blue when either their WT or resistant form was used to subset genomes, in purple otherwise).



(a) All genomes.

(b) Six genomes of identical putative parents.

Figure A.11: **Percentage of genomic regions in IBD between pairs of genomes** (*Section 2.3.7 Natural cross offspring and lineages expansion*). (a): All genomes, after merging identical pairs (IBD > 0.9) from the same individual, were clustered according to their IBD values converted to a distance matrix. Most of the clusters are small in size with just 2 or 3 genomes clustered. (b): Focus on 6 genomes likely originating from a genetic cross between the two same parental genomes. The presumed offspring identified in our dataset are generally not or mildly related (IBD $\leq$ 0.68), suggesting that they are originating from distinct events of genomic cross from the same initial two parasite strains.

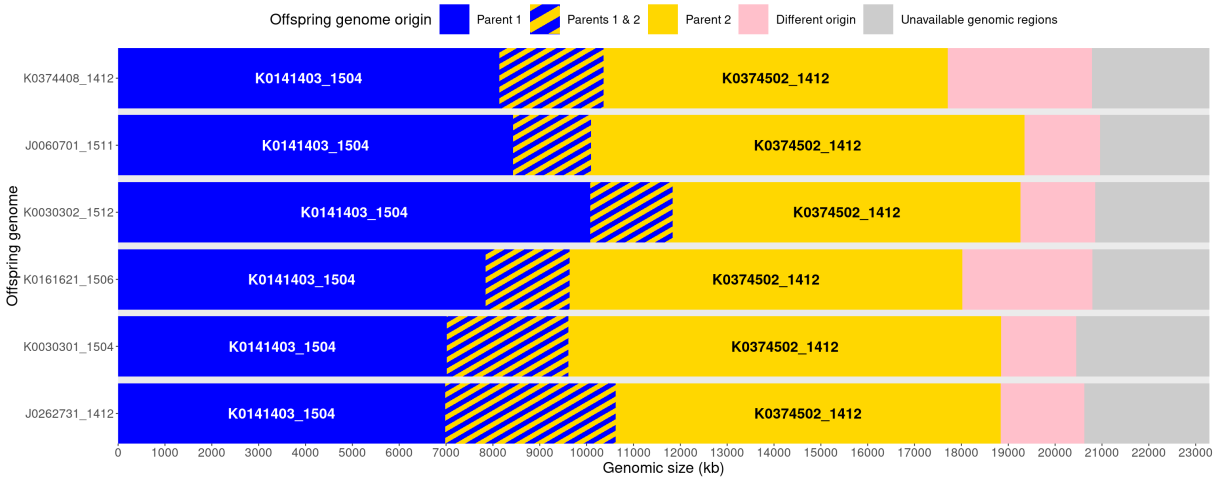
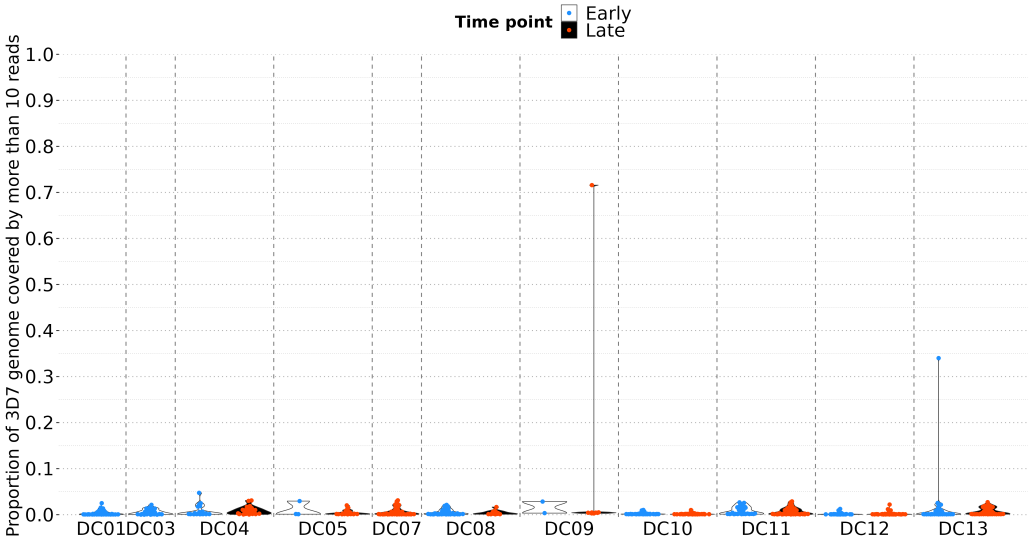
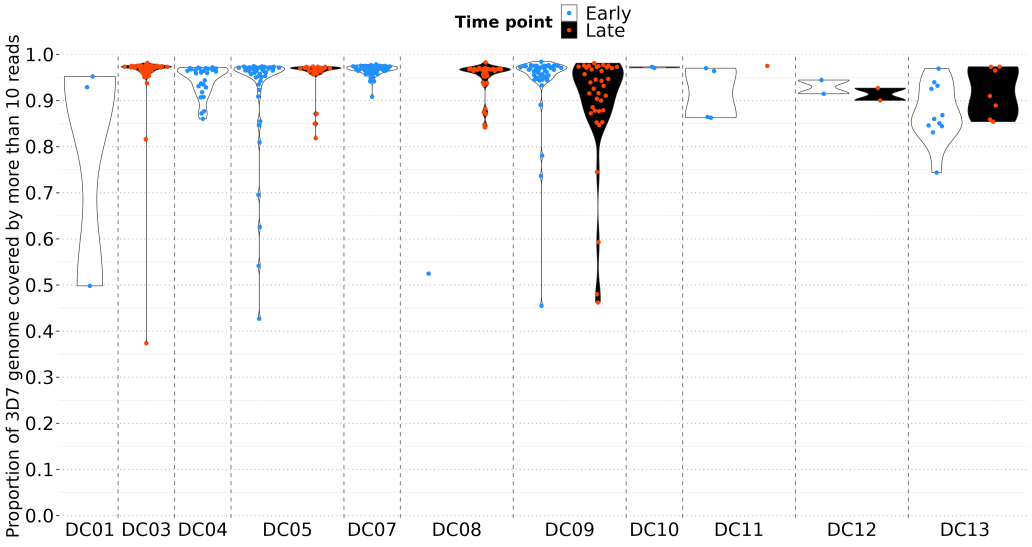


Figure A.12: **Fragmented origin of 6 genomes relative to their attributed parents** (*Section 2.3.7 Natural cross offspring and lineages expansion*). Genomic fragments can be in IBD with their parents (blue and yellow), neither of them (red) or can have an unknown origin (grey) when too few SNPs can be compared in an area. A larger area of different or unknown origin indicates that one or both of the attributed parental genomes are not exactly the ones that were involved in the meiotic cross but are probably closely related to them. The 6 offspring are sharing more than 85 % of the accessible parts of their genomes with their parents which have little overlap with each other.

Additional Material of Project 2



(a) Low mapping quality.



(b) High mapping quality.

Figure B.1: **Coverage of single-cell and pooled-cells genomes by mapping quality** (*Section 3.3.2 Successful single-cell sequencing of several asymptomatic infections*). (a): Out of the 949 (933 single-cell and 16 pooled-cells) genomes from early (blue) and late (red) time points of the infections sequenced, 565 were poorly mapped with less than 10 % of all reads mapping 3D7 reference genome. Almost all poorly mapped genomes had a very low level of coverage (under 5 %) of 3D7 genome with 10 reads depth, the remaining two resulting in around 70 % and 35 % of the 3D7 genome covered. (b): 341 out of the 949 genomes contained at least 16351 SNPs (30 % of 54505 SNPs) covered by at least 10 reads were consequently considered as high-quality genomes. Almost all genomes with a high mapping quality (326/341) had at least 80 % of the 3D7 genome covered by 10 reads at least, with the remaining ones (15/341) showing coverage proportion lying between 35 % and 80 %.

Table B.1: **SNPs differentially present in DC05 cells from months 1 and 6** (*Section 3.3.4 Selective advantages of variants during in vivo infections*). The proportion of genomes with a particular alternate nucleotide (*Alt*) from the reference (*Ref*) in 3D7 genome was compared between cells sequenced from month 1 (*AFm1*) and 6 (*AFm6*) of the infection in individual DC05. SNPs that have a change of more than 0.25 (0.5 in bold) in frequency with 80 % of cells in each time point covered by at least 10 reads are registered here. Some of the SNPs are located within genes, in which case they are associated with a gene (*Annotation*), and optionally within coding regions, in which case the change in amino acid from the 3D7 reference (*AAchange*) is also available.

Chromosome	Position	Ref	Alt	AAchange	Annotation	AFm1	AFm6
Pf3D7_01_v3	530617	C	A	D1134E	DBL containing protein unknown function (PF3D7_0113800)	0.44	0.69
Pf3D7_01_v3	530667	G	A	G1151D	DBL containing protein unknown function (PF3D7_0113800)	0.37	0.08
Pf3D7_01_v3	573135	T	C	–	–	0.71	0.37
Pf3D7_04_v3	340646	G	T	L1137I	NYN domain-containing protein putative (PF3D7_0406500)	0.33	0.68
Pf3D7_04_v3	340656	G	T	D1133E	NYN domain-containing protein putative (PF3D7_0406500)	0.33	0.68
Pf3D7_04_v3	340662	A	T	S1131R	NYN domain-containing protein putative (PF3D7_0406500)	0.33	0.68
Pf3D7_04_v3	340667	A	T	F1130I	NYN domain-containing protein putative (PF3D7_0406500)	0.31	0.64
Pf3D7_04_v3	340698	G	T	D1119E	NYN domain-containing protein putative (PF3D7_0406500)	0.26	0.64
Pf3D7_04_v3	**420049**	**A**	**C**	**N165K**	**perforin-like protein 1 (PF3D7_0408700)**	**0.85**	**0.34**
Pf3D7_05_v3	134411	C	T	–	serine/arginine-rich splicing factor 12 (PF3D7_0503300)	0.88	0.59
Pf3D7_06_v3	1214350	G	T	I552I	amino acid transporter AAT1 (PF3D7_0629500)	0.86	0.58
Pf3D7_07_v3	251447	T	A	–	–	0.94	0.67
Pf3D7_07_v3	251449	T	A	–	–	0.94	0.67
Pf3D7_07_v3	251451	T	A	–	–	0.94	0.67
Pf3D7_07_v3	251453	T	A	–	–	0.94	0.64
Pf3D7_07_v3	251455	T	A	–	–	0.97	0.64
Pf3D7_08_v3	223482	T	A	–	conserved Plasmodium protein unknown function (PF3D7_0803600)	0.89	0.59
Pf3D7_08_v3	**761855**	**A**	**C**	**–**	**chaperone protein ClpB1 (PF3D7_0816600)**	**0**	**0.8**
Pf3D7_08_v3	886767	C	A	–	ubiquitin-like protein putative (PF3D7_0819600)	0.84	0.48
Pf3D7_08_v3	886772	G	T	–	ubiquitin-like protein putative (PF3D7_0819600)	0.84	0.48
Pf3D7_08_v3	886774	A	T	–	ubiquitin-like protein putative (PF3D7_0819600)	0.84	0.48
Pf3D7_08_v3	886780	G	T	–	ubiquitin-like protein putative (PF3D7_0819600)	0.84	0.48
Pf3D7_08_v3	886788	T	A	–	ubiquitin-like protein putative (PF3D7_0819600)	0.84	0.48
Pf3D7_08_v3	886794	G	A	–	ubiquitin-like protein putative (PF3D7_0819600)	0.84	0.48
Pf3D7_10_v3	802927	G	C	D192E	conserved Plasmodium protein unknown function (PF3D7_1019700)	0.75	1
Pf3D7_10_v3	1135146	G	T	D337E	U3 small nucleolar ribonucleoprotein protein MPP10 putative (PF3D7_1027100)	0.36	0.76
Pf3D7_10_v3	1146459	T	C	–	D−-directed R−polymerase II subunit RPB7 putative (PF3D7_1027400)	0.63	0.88
Pf3D7_11_v3	**426094**	**T**	**C**	**–**	**ribosomal protein L11 mitochondrial putative (PF3D7_1110600)**	**0**	**0.81**
Pf3D7_13_v3	1976258	C	G	–	–	0.66	0.19
Pf3D7_13_v3	2143334	A	T	–	Ran-binding protein putative (PF3D7_1353400)	0.97	0.59

Table B.2: **SNPs differentially present in DC13 cells from months 1 and 6** (*Section 3.3.4 Selective advantages of variants during in vivo infections*). The proportion of genomes with a particular alternate nucleotide (*Alt*) from the reference (*Ref*) in 3D7 genome was compared between cells sequenced from month 1 (*AFm1*) and 6 (*AFm6*) of the infection in individual DC13. SNPs that have a change of more than 0.25 (0.5 in bold) in frequency with 80 % of cells in each time point covered by at least 10 reads are registered here. Some of the SNPs are located within genes, in which case they are associated with a gene (*Annotation*), and optionally within coding regions, in which case the change in amino acid from the 3D7 reference (*AAchange*) is also available.

Chromosome	Position	AAchange	AAref	AAalt	Annotation	AFm1	AFm6
Pf3D7_01_v3	227872	–	–	–	–	0.12	0.4
Pf3D7_01_v3	573644	–	–	–	–	0.62	1
Pf3D7_01_v3	573645	–	–	–	–	0.62	1
Pf3D7_01_v3	573649	–	–	–	–	0.62	1
Pf3D7_02_v3	161537	–	–	–	conserved protein unknown function (PF3D7_0203400)	1	0.66
Pf3D7_02_v3	**293647**	**–**	**–**	**–**	**–**	**0.49**	**1**
Pf3D7_03_v3	609665	S261S	S (261/612)	S (261/612)	conserved Plasmodium membrane protein unknown function (PF3D7_0314900)	0.73	0.99
Pf3D7_03_v3	718965	S806S	S (806/1620)	S (806/1620)	kinesin-5 (PF3D7_0317500)	0.18	0.46
Pf3D7_03_v3	718966	S806N	S (806/1620)	N (806/1620)	kinesin-5 (PF3D7_0317500)	0.18	0.46
Pf3D7_03_v3	751701	S1079S	S (1079/2458)	S (1079/2458)	D–-directed R–polymerase II subunit RPB1 (PF3D7_0318200)	0.23	0.61
Pf3D7_03_v3	848972	G2845S	G (2845/3086)	S (2845/3086)	oocyst capsule protein Cap380 (PF3D7_0320400)	0	0.25
Pf3D7_04_v3	227564	R1390R	R (1390/2272)	R (1390/2272)	AP2 domain transcription factor AP2-SP2 putative (PF3D7_0404100)	0	0.33
Pf3D7_04_v3	235503	D308D	D (308/632)	D (308/632)	conserved Plasmodium protein unknown function (PF3D7_0404300)	0.26	0.57
Pf3D7_04_v3	237364	–	–	–	conserved Plasmodium protein unknown function (PF3D7_0404300)	0.54	0.8
Pf3D7_04_v3	908786	I577I	I (577/582)	I (577/582)	serine/threonine protein kinase RIO2 (PF3D7_0420100)	1	0.67
Pf3D7_04_v3	1101237	D476G	D (476/2381)	G (476/2381)	surface-associated interspersed protein 4.2 (SURFIN 4.2) (PF3D7_0424400)	0.43	0.17
Pf3D7_04_v3	1101248	R480G	R (480/2381)	G (480/2381)	surface-associated interspersed protein 4.2 (SURFIN 4.2) (PF3D7_0424400)	0.43	0.17
Pf3D7_04_v3	1101251	N481D	N (481/2381)	D (481/2381)	surface-associated interspersed protein 4.2 (SURFIN 4.2) (PF3D7_0424400)	0.43	0.17
Pf3D7_04_v3	1101253	N481N	N (481/2381)	N (481/2381)	surface-associated interspersed protein 4.2 (SURFIN 4.2) (PF3D7_0424400)	0.43	0.17
Pf3D7_04_v3	1101256	C482C	C (482/2381)	C (482/2381)	surface-associated interspersed protein 4.2 (SURFIN 4.2) (PF3D7_0424400)	0.42	0.17
Pf3D7_04_v3	1101264	G485A	G (485/2381)	A (485/2381)	surface-associated interspersed protein 4.2 (SURFIN 4.2) (PF3D7_0424400)	0.42	0.17
Pf3D7_04_v3	1101271	Y487_	Y (487/2381)	_ (487/2381)	surface-associated interspersed protein 4.2 (SURFIN 4.2) (PF3D7_0424400)	0.42	0.17
Pf3D7_04_v3	1101276	N489T	N (489/2381)	T (489/2381)	surface-associated interspersed protein 4.2 (SURFIN 4.2) (PF3D7_0424400)	0.42	0.17
Pf3D7_05_v3	374678	C2553G	C (2553/3135)	G (2553/3135)	protein AAP6 (PF3D7_0508900)	0.74	1
Pf3D7_05_v3	374679	C2553Y	C (2553/3135)	Y (2553/3135)	protein AAP6 (PF3D7_0508900)	0.74	1
Pf3D7_05_v3	374725	D2568E	D (2568/3135)	E (2568/3135)	protein AAP6 (PF3D7_0508900)	0.67	1
Pf3D7_05_v3	374731	E2570D	E (2570/3135)	D (2570/3135)	protein AAP6 (PF3D7_0508900)	0.67	1
Pf3D7_05_v3	511633	D8344D	D (8344/10062)	D (8344/10062)	R–pseudouridylate synthase putative (PF3D7_0511500)	0.97	0.7
Pf3D7_05_v3	629713	N245D	N (245/2134)	D (245/2134)	phosphatidylinositol 3-kinase (PF3D7_0515300)	0.5	0.79
Pf3D7_05_v3	629715	N245N	N (245/2134)	N (245/2134)	phosphatidylinositol 3-kinase (PF3D7_0515300)	0.5	0.79
Pf3D7_05_v3	629717	K246T	K (246/2134)	T (246/2134)	phosphatidylinositol 3-kinase (PF3D7_0515300)	0.5	0.79
Pf3D7_05_v3	629720	Y247C	Y (247/2134)	C (247/2134)	phosphatidylinositol 3-kinase (PF3D7_0515300)	0.5	0.79
Pf3D7_05_v3	678684	–	–	–	tR–pseudouridine synthase putative (PF3D7_0516300)	0	0.4
Pf3D7_07_v3	**410240**	**P676S**	**P (676/1249)**	**S (676/1249)**	**Cg1 protein (PF3D7_0709100)**	**0.36**	**1**
Pf3D7_07_v3	410348	S712P	S (712/1249)	P (712/1249)	Cg1 protein (PF3D7_0709100)	0.64	1
Pf3D7_07_v3	410446	T744T	T (744/1249)	T (744/1249)	Cg1 protein (PF3D7_0709100)	0.6	1
Pf3D7_07_v3	447043	–	–	–	–	0.44	0.8
Pf3D7_07_v3	447138	–	–	–	–	0.5	0.8
Pf3D7_07_v3	609353	N19N	N (19/408)	N (19/408)	erythrocyte membrane protein 1 (PfEMP1) pseudogene (PF3D7_0713300)	0.99	0.67
Pf3D7_07_v3	977226	Q487Q	Q (487/881)	Q (487/881)	conserved protein unknown function (PF3D7_0723300)	0.75	1
Pf3D7_07_v3	977227	R488R	R (488/881)	R (488/881)	conserved protein unknown function (PF3D7_0723300)	0.75	1
Pf3D7_08_v3	130673	H609D	H (609/4663)	D (609/4663)	lysine-specific histone demethylase putative (PF3D7_0801900)	0.5	0.17
Pf3D7_08_v3	841265	–	–	–	zinc finger protein putative (PF3D7_0818500)	0.37	0
Pf3D7_08_v3	1164359	–	–	–	–	0.65	0.98
Pf3D7_08_v3	1334715	S434S	S (434/962)	S (434/962)	DnaJ protein putative (PF3D7_0831200)	0.49	0.8
Pf3D7_09_v3	79311	–	–	–	–	0	0.3
Pf3D7_09_v3	79321	–	–	–	–	0	0.31
Pf3D7_09_v3	79459	–	–	–	–	0.05	0.41
Pf3D7_09_v3	228861	D1324G	D (1324/2569)	G (1324/2569)	copper-transporting ATPase (PF3D7_0904900)	0.58	0.83
Pf3D7_09_v3	228976	V1286I	V (1286/2569)	I (1286/2569)	copper-transporting ATPase (PF3D7_0904900)	0.49	0.82
Pf3D7_09_v3	763641	–	–	–	–	0.55	0.17
Pf3D7_09_v3	**772579**	**–**	**–**	**–**	**dihydrouridine synthase putative (PF3D7_0918800)**	**0.44**	**1**
Pf3D7_10_v3	749679	–	–	–	conserved protein unknown function (PF3D7_1018800)	1	0.6
Pf3D7_10_v3	800556	H983N	H (983/1075)	N (983/1075)	conserved Plasmodium protein unknown function (PF3D7_1019700)	0.68	0.97
Pf3D7_10_v3	800557	E982D	E (982/1075)	D (982/1075)	conserved Plasmodium protein unknown function (PF3D7_1019700)	0.68	0.97
Pf3D7_10_v3	830645	I940T	I (940/2135)	T (940/2135)	conserved Plasmodium membrane protein unknown function (PF3D7_1020600)	0.6	1
Pf3D7_10_v3	880669	C4548Y	C (4548/6935)	Y (4548/6935)	VPS13 domain-containing protein putative (PF3D7_1021700)	1	0.66
Pf3D7_10_v3	880673	N4547D	N (4547/6935)	D (4547/6935)	VPS13 domain-containing protein putative (PF3D7_1021700)	1	0.6
Pf3D7_10_v3	880674	D4546D	D (4546/6935)	D (4546/6935)	VPS13 domain-containing protein putative (PF3D7_1021700)	1	0.6
Pf3D7_10_v3	880675	D4546G	D (4546/6935)	G (4546/6935)	VPS13 domain-containing protein putative (PF3D7_1021700)	0.99	0.6
Pf3D7_10_v3	1261458	K383K	K (383/920)	K (383/920)	SAE2 domain-containing protein putative (PF3D7_1031300)	0.99	0.66
Pf3D7_10_v3	1464383	E503D	E (503/2227)	D (503/2227)	D–polymerase zeta catalytic subunit putative (PF3D7_1037000)	0.54	0.19
Pf3D7_10_v3	1464407	E511D	E (511/2227)	D (511/2227)	D–polymerase zeta catalytic subunit putative (PF3D7_1037000)	0.43	0.16
Pf3D7_11_v3	1865528	N610N	N (610/2941)	N (610/2941)	sporozoite and liver stage asparagine-rich protein (PF3D7_1147000)	0.33	0.6
Pf3D7_11_v3	1977902	V457V	V (457/1091)	V (457/1091)	ring-infected erythrocyte surface antigen (PF3D7_1149200)	0.25	0.64
Pf3D7_11_v3	1977910	Q460L	Q (460/1091)	L (460/1091)	ring-infected erythrocyte surface antigen (PF3D7_1149200)	0.25	0.66
Pf3D7_11_v3	1977913	N461S	N (461/1091)	S (461/1091)	ring-infected erythrocyte surface antigen (PF3D7_1149200)	0.25	0.66
Pf3D7_11_v3	1977925	S465N	S (465/1091)	N (465/1091)	ring-infected erythrocyte surface antigen (PF3D7_1149200)	0.25	0.66
Pf3D7_11_v3	1977928	G466V	G (466/1091)	V (466/1091)	ring-infected erythrocyte surface antigen (PF3D7_1149200)	0.25	0.66
Pf3D7_11_v3	1977933	Q468E	Q (468/1091)	E (468/1091)	ring-infected erythrocyte surface antigen (PF3D7_1149200)	0.33	0.66
Pf3D7_11_v3	1977940	S470N	S (470/1091)	N (470/1091)	ring-infected erythrocyte surface antigen (PF3D7_1149200)	0.24	0.65
Pf3D7_11_v3	1977943	D471V	D (471/1091)	V (471/1091)	ring-infected erythrocyte surface antigen (PF3D7_1149200)	0.24	0.65
Pf3D7_12_v3	1031167	–	–	–	rR–-processing protein FCF1 putative (PF3D7_1225300)	0.86	0.5
Pf3D7_12_v3	1931578	–	–	–	myosin A-tail interacting protein (PF3D7_1246400)	0.37	0.66
Pf3D7_13_v3	794542	K149R	K (149/869)	R (149/869)	tetratricopeptide repeat protein putative (PF3D7_1319200)	0.49	0.82
Pf3D7_13_v3	2042409	–	–	–	–	0.68	0.33
Pf3D7_14_v3	2423128	–	–	–	ATP-dependent R–helicase DBP5 (PF3D7_1459000)	0.66	1
Pf3D7_14_v3	2423131	–	–	–	ATP-dependent R–helicase DBP5 (PF3D7_1459000)	0.67	1

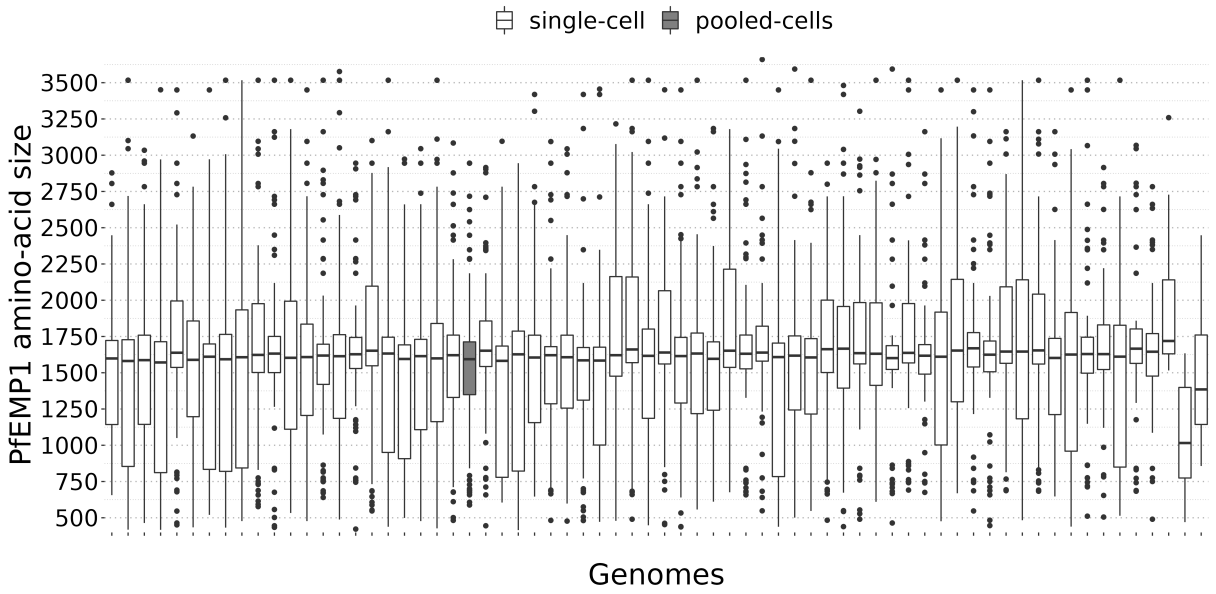
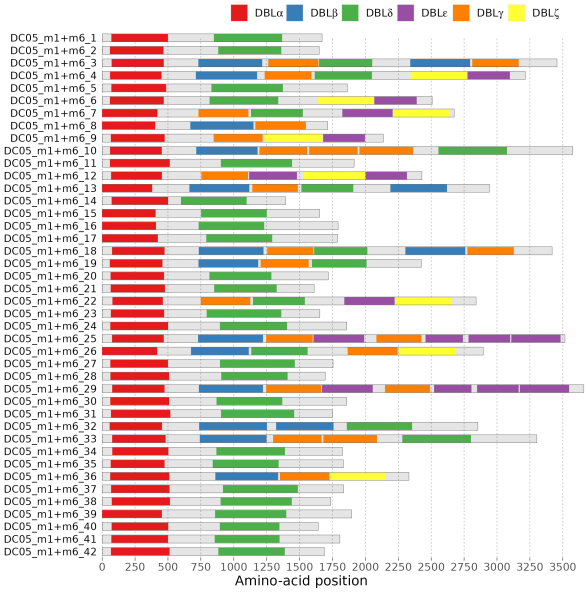
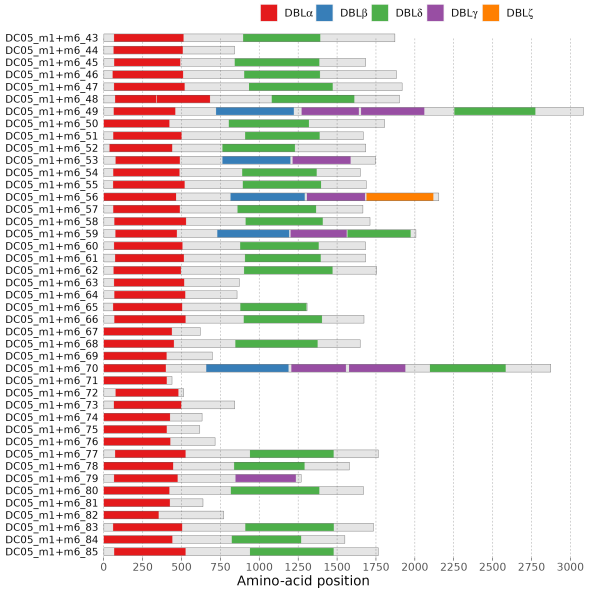


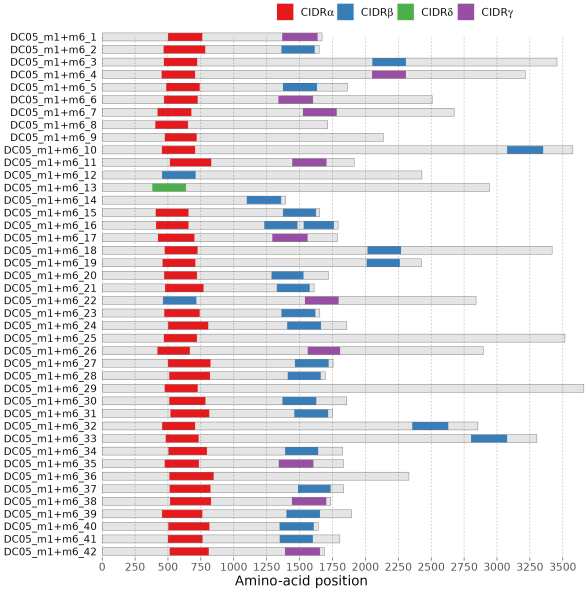
Figure B.2: **PfEMP1 amino acid sizes within each DC05 genome** (*Section 3.3.6 Extraction of var gene repertoires from individual infections*). For each of the parasite genome sequenced from both time points of the monoclonal infection of DC05 (67 single-cell and one pooled-cells genome), unique *var* genes (identity > 95 % and shortest *var* gene covering 90 % of the longest) were translated into PfEMP1s and their amino-acid sizes were calculated. Genomes were sorted from the ones with the highest to the ones with the lowest number of PfEMP1s (same order as Figure 3.7a).



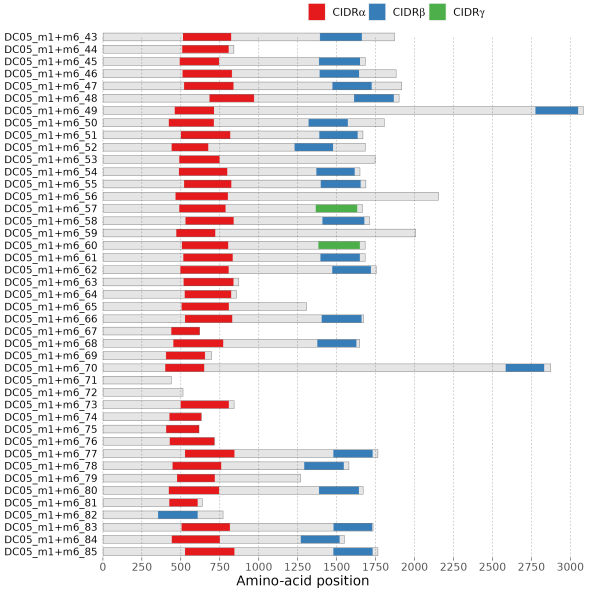
(a) DBL sub-domains of PfEMP1s 1 to 42.



(b) DBL sub-domains of PfEMP1s 43 to 85.



(c) CIDR sub-domains of PfEMP1s 1 to 42.



(d) CIDR sub-domains of PfEMP1s 43 to 85.

Figure B.3: **Annotated reference PfEMP1 sub-domains of DC05_m1+m6** (*Section 3.3.6 Extraction of var gene repertoires from individual infections*). The full set of reference PfEMP1s from individual DC05 was built by *de novo* assembly of the short-reads obtained form the sequencing of one pooled-cells and 67 single-cell samples. For each reference PfEMP1, the coordinates of DBL (a & b) and CIDR (c & d) sub-domains are extracted from the annotation provided by VarDom. For a better visualization, the first 42 (a & c) and the last 43 (b & d) PfEMP1s are shown separately.

Collaborative Projects

C.1 Mutation rates

C.1.1 Introduction

Elimination of malaria is still to this day challenged by the ability of its most deadly parasite, *Plasmodium falciparum*, to adapt to various external pressures applied by antimalarial drugs or by the host immune system. Historically, genetic markers of drug resistance have first been acquired in parasites from South-East Asia [28]. Secondly, the parasite is able to evade the immune system through regular switching of its PfEMP1, encoded by 60 *var* genes undergoing mutually exclusive expression. Mutation accumulation lines previously indicated a high rate of mitotically generated novel *var* gene sequences [188]. However the micro-insertion/deletion (INDEL) and structural variant mutation rates were limited to the 3D7 strain. Here, using four novel 'reference genomes', we re-analysed 354 whole genomes originated from four distinct *Plasmodium falciparum* strains from various geographical origin, cultivated for over four years of combined *in vitro* culturing [42].

This study was conducted by William L. Hamilton[1,2], Aakanksha Singh[3], Marc-Antoine Guery[3], Balotin Fogang[1,2], Dominic Kwiatkowski[1,5], Julian C. Rayner[1] and Antoine Claessens[3,4].

Collaborators work The different *Plasmodium falciparum* strains were cultivated and successively cloned by limiting dilution to obtain one clone tree per strain, totalling 4 years of combined culturing. An additional HB3-derived strain possessing a disruption in the MutS homologue 2-2 gene (*pfmsh2-2*; PfHB3_070010700), involved in DNA mismatch repair, followed also the same experimental procedures. The mutation rate of the disrupted HB3-derived line was compared with those of the wild-type strains and showed no difference, indicating that this gene might not be the only one involved in DNA repair. A total of 96 structural variants were identified with 40 % of them involving genes, 90 % of which encode variant surface antigens (VSAs). The first chimeric *rif* generated *in vitro* was also observed. The rates of structural variants were similar between the different strains except for HB3 and disrupted HB3-derived strains which showed respectively none and a single structural variant. Further characterization of that line should shed light on how *Plasmodium falciparum* generates novel antigens.

[1]Malaria Programme, Wellcome Sanger Institute, Hinxton, United Kingdom
[2]University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Hills Road, Cambridge, United Kingdom
[3]LPHI, MIVEGEC, Université de Montpellier, CNRS, INSERM, Montpellier, France
[4]Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom
[5]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

My involvement I co-supervised Aakanksha, a master's student, who used and improved one of my pipelines (https://github.com/marcguery/discovarif) to call and filter *de novo* mutations from the mapping with the new reference assemblies available to obtain 254 *de novo* micro-INDELs [42]. Following this, I was interested in determining the insertion to deletion biases of the 254 *de novo* micro-INDELs and their preferential location in genomic regions.

C.1.2 Repeat motif identification

Tandem repeat finder (TRF) was used to find all tandem repeats in each of the strains 3D7, Dd2, HB3, KH1, KH2, disrupted HB3-derived (MH2D) and W2 [189]. The minimum alignment score was set to 24 (allowing for repeats at least 12 nucleotides long) and only repeat motifs repeated at least 5 times were conserved. The alignment score is the result of the alignment between the sequence obtained from a repetition of a given repeat motif and an actual genomic sequence. Because the weight of a match is 2 and the minimum score is set to 24, only repeats whose total size is 12 or more nucleotides are kept. For each strain, the initial repeat motifs detected were grouped with one another if any combination between their actual, reverse, complementary or reverse complementary sequences were identical, a subset or resulting from a shift of the repeat motif (Figure C.1). All strains show similar number of repeats covering 24 % of their cumulated genome size (27140177/114843219 nucleotides) for a total of 700012 repeated regions in all the strains. Similarly, almost identical numbers of unique repeat motifs were found for all the strains totalling at 11906 unique repeats in all the strains combined.

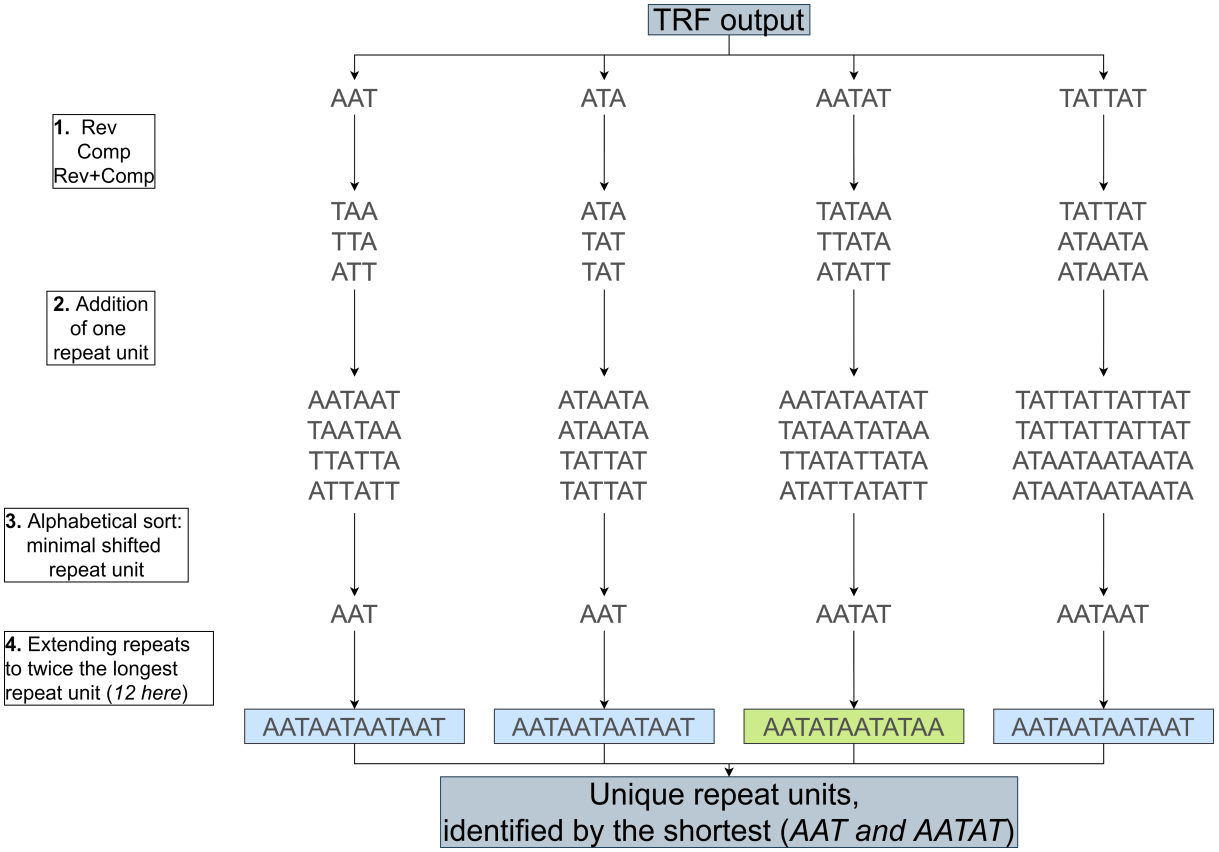


Figure C.1: **Pipeline of repeat motif merging**. First, the reverse, complement and reverse-complement sequences of each motif are associated with the original sequence so that the strand or direction of the repeat can be ignored. Then, each sequence is extended by one repeat unit. The minimal alphabetically-sorted shifted repeat unit is then extracted and extended so that all repeats are the same length. Completely identical sequences are merged and identified by the shortest repeat unit for the rest of the analysis.

C.1.3 Results

Repeat motifs have a median length of 2 and 95 % of the motifs are less than 12 nucleotides long. The median total size of repeated sequences is 35 and 95 % of the repeated sequences are less than 136 nucleotides long. The 11906 repeat motif groups were ranked by their abundance, first using the proportion of repeats attributed to each motif, then using the coverage of unique nucleotides involved in each motif among all unique nucleotides in repeated regions. Both estimations of abundance resulted in the same top 5 motifs AT, A, AAT, AAAT and AAAAT for all the strains; they are responsible for 78 % (543245/700012 repeats) of repeats and cover 71 % (19364783/27140177 nucleotides) of all the repeated regions. The proportion and coverage of the top 5 repeat motifs were all similar except for AT and A repeats which had virtually the same number of repeats but AT repeats were generally longer than A repeats resulting in a higher part of the repeated regions covered by AT repeats.

The genomic locations of 254 INDELs detected in the strains 3D7, Dd2, HB3, KH1, KH2, disrupted HB3-derived (MH2D) and W2 were screened for the presence of repeats previously identified with TRF. The repeat retained from multiple matches with the same INDEL is the one with the highest ratio of their score over an hypothetical optimal alignment score obtained if the corresponding motif is identically repeated over the total length of the repeat. More than 98 % of all INDELs are found in a repeated region (250/254). Considering all the strains combined, the top 5 repeat motifs found in INDELs were AT, A, AAT, AAAT and AATAT with only AT being the most abundant repeat motif in each strain individually (Figure C.2).

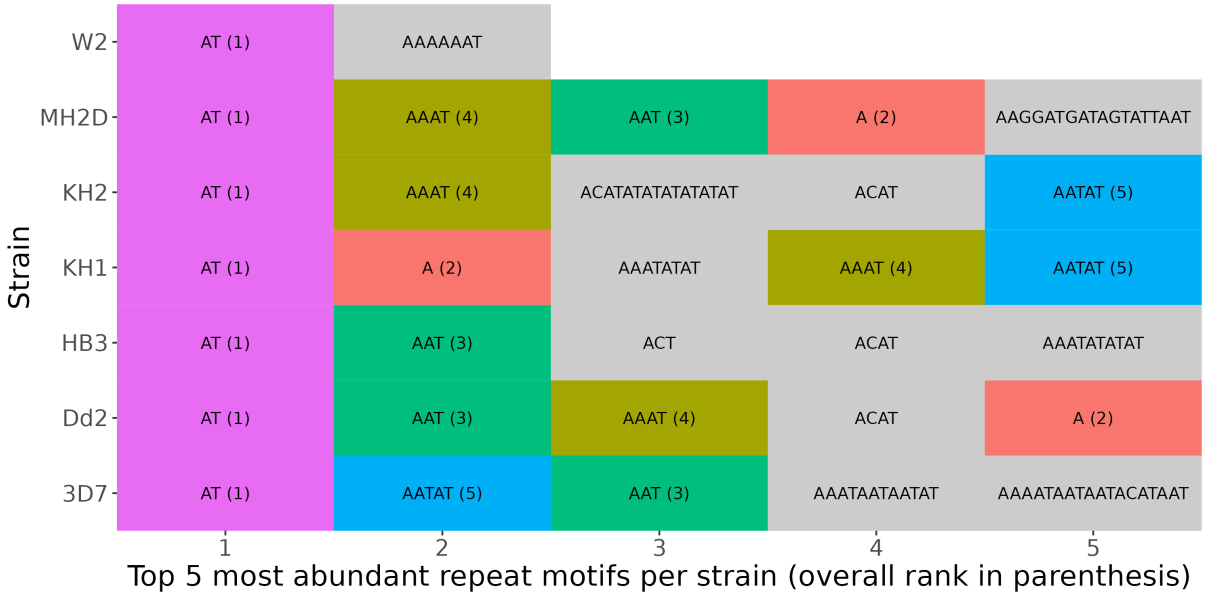
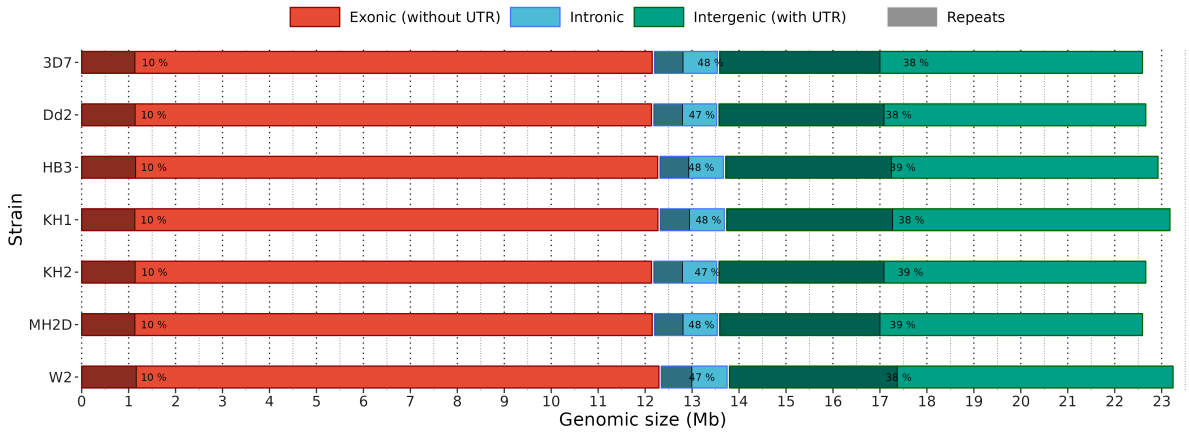
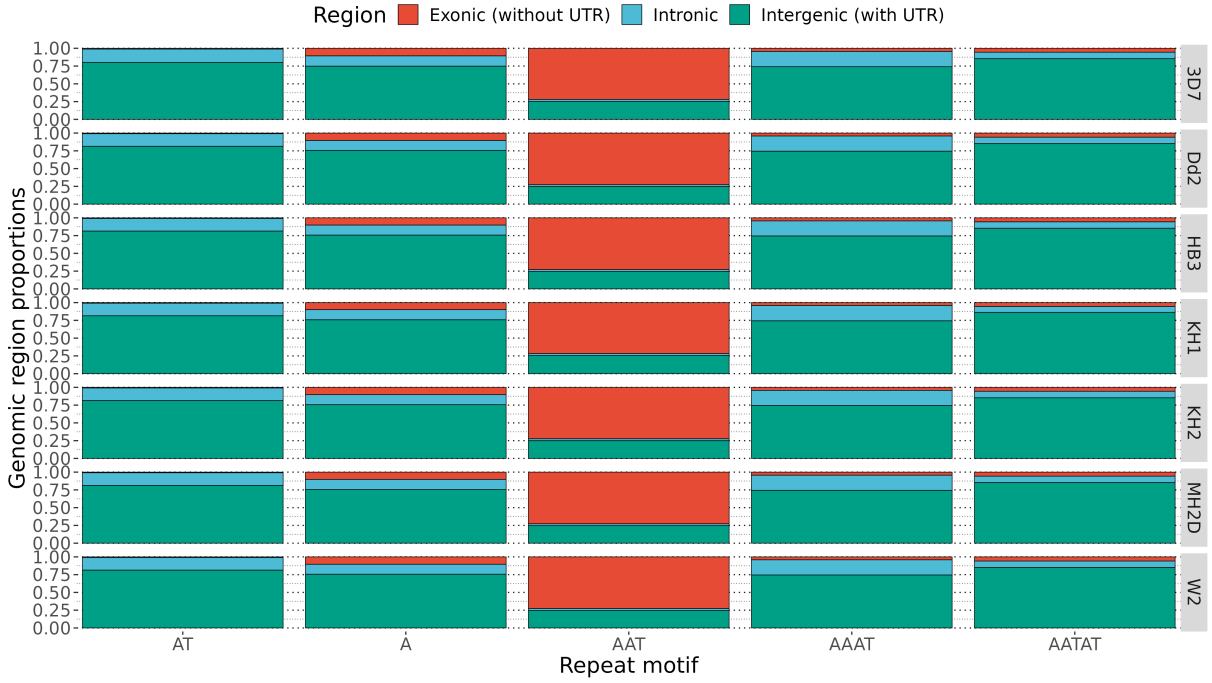


Figure C.2: **Top 5 most abundant repeat motifs**. The top 5 most abundant repeat motifs found in all INDELs (AT, A, AAT, AAAT and AATAT) are shared in all strains used, with the AT motif being the most abundant in every strain.

The overall and top 5 repeat motifs distributions were calculated across exonic, intronic and intergenic genomic regions (Figure C.3). Repeats associated with the top 5 most abundant repeat motifs except AAT are largely present in intergenic (79 % on average) and intronic regions (17 % on average) and almost absent in exonic regions (4 % on average) for all strains. Repeats associated with the repeat motif AAT are almost all located in exonic (72 %) regions with the remaining in intergenic (25 %) and intronic (3 %) regions. As all the strains used display an almost identical architecture of repeats, the INDELs of all the strains were pooled together for the subsequent analyses.



(a) Genomic distribution of all repeats (exonic, intronic and intergenic regions).



(b) Genomic distribution of repeats from the top 5 most abundant motifs found in INDELs.

Figure C.3: **Distribution of repeats in laboratory-adapted strains**. (a): All strains have a near identical distribution of repeats from all repeat motifs in the three different genomic regions. (b): Similarly, the individual distribution of each of the top 5 most abundant repeat motifs found in INDELs (AT, A, AAT, AAAT and AATAT) is almost identical in all strains. The repeat motif AAT has a particular genomic distribution compared to other repeat motifs as it is mostly present in exonic regions and not intergenic regions.

The proportions of the top 5 repeat motifs over all repeat motifs found in all INDELs were compared to their expected value according to their coverage in all the strains combined (Figure C.4). Despite its high coverage of repeated regions, the most abundant repeat motif AT was significantly more prevalent than what was expected at random (χ^2 test value of

89.6; p value = 2.8×10^{-21}), suggesting a preference for INDELs to be conserved in AT repeated regions. In contrary, INDELs were less present in A repeats than what was expected at random (χ^2 test value of 35.6; p value = 2.5×10^{-9}). All other repeat motifs were either not significantly differentially present in INDELs or not enough data was available to perform the test.

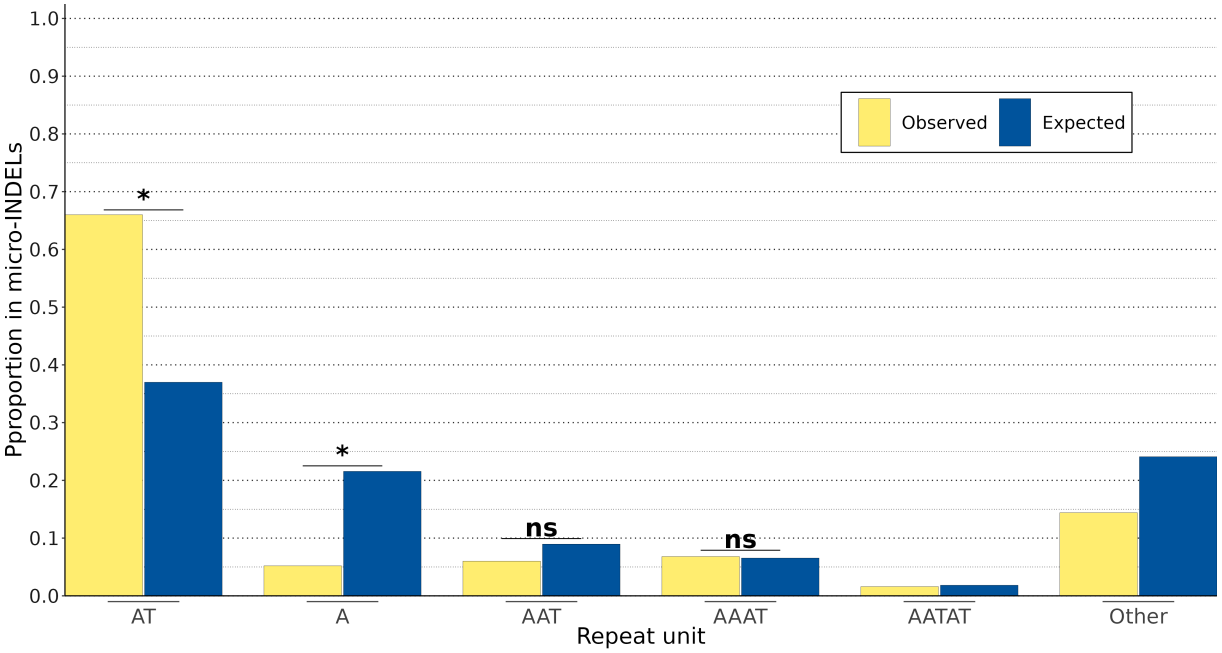


Figure C.4: **Expected vs observed repeat unit coverage in INDELs**. The observed proportion of INDELs with a particular repeat motif was compared to its expected proportion according to its coverage in genomic repeats using χ^2 (* = p value < 0.01, 'ns' = non significant). Given the genomic coverage of each repeat motif, INDELs of AT repeats are more prevalent than what was expected while INDELs of A homopolymers are less prevalent than what was expected.

The distribution of insertion and deletion sizes was compared using all repeat units. Short insertions are far more frequent than longer insertions while the deletion sizes distribution is more flattened and allows for longer deletions (Figure C.5). This difference is more visible when including only AT INDELs and less visible or almost reversed when considering only A INDELs. Even though there are much less deletions than insertions (92/250 INDELs are deletions), INDELs tend to decrease the genomic size (855 bases removed and 483 bases inserted in total). Three frequent repeat units (AT, AAT and AAAT) have a negative balance (respectively 58 %, 78 % and 69 % of total INDEL size corresponding to deletions) while A homopolymer repeats have a positive one (68 % of total INDEL size corresponding to insertions).

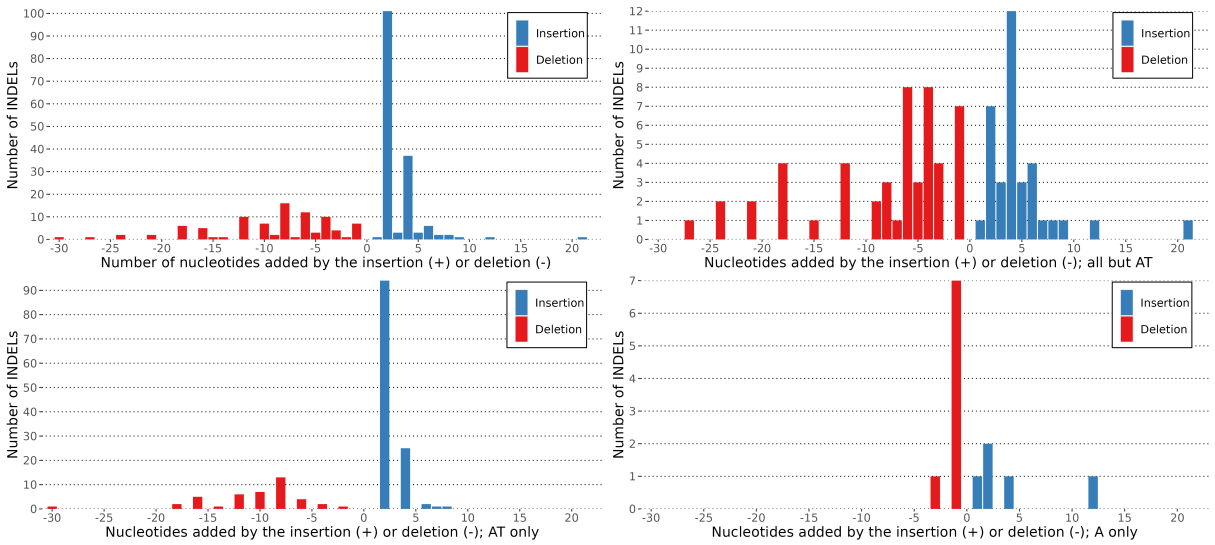


Figure C.5: **Distribution of insertion vs deletion sizes**. The number of bases inserted (in red) or deleted (in blue) is shown for INDELs containing any repeat motif (top left panel), any motif except AT (top right panel), only AT motifs (bottom left panel) or only A motifs (bottom right panel). INDELs with AT and A repeat motifs have an opposite distribution with mostly short deletions for the former and mostly short insertions for the latter.

C.1.4 Conclusion

All the results of this project are already obtained and most of the sections of the draft manuscript are being completed. I will be included as a co-second author for this work.

C.2 Identification of a drug resistant marker

C.2.1 Introduction

Collaborators work A team led by Rachel Cerdan[1] cultivated thirteen 3D7 strains in the presence of a newly synthetic antimalarial compound (a purine analogue) and one other 3D7 strain in the absence of any antimalarial compound, serving as a negative control (Figure C.6). Over time, the thirteen 3D7 lines acquired different phenotypes of resistance with varying degrees of sensitivity to the synthetic drug.

My involvement My role was to find mutations that could be linked to the drug resistance phenotype observed by finding variants present in resistant strains while absent from the sensitive one.

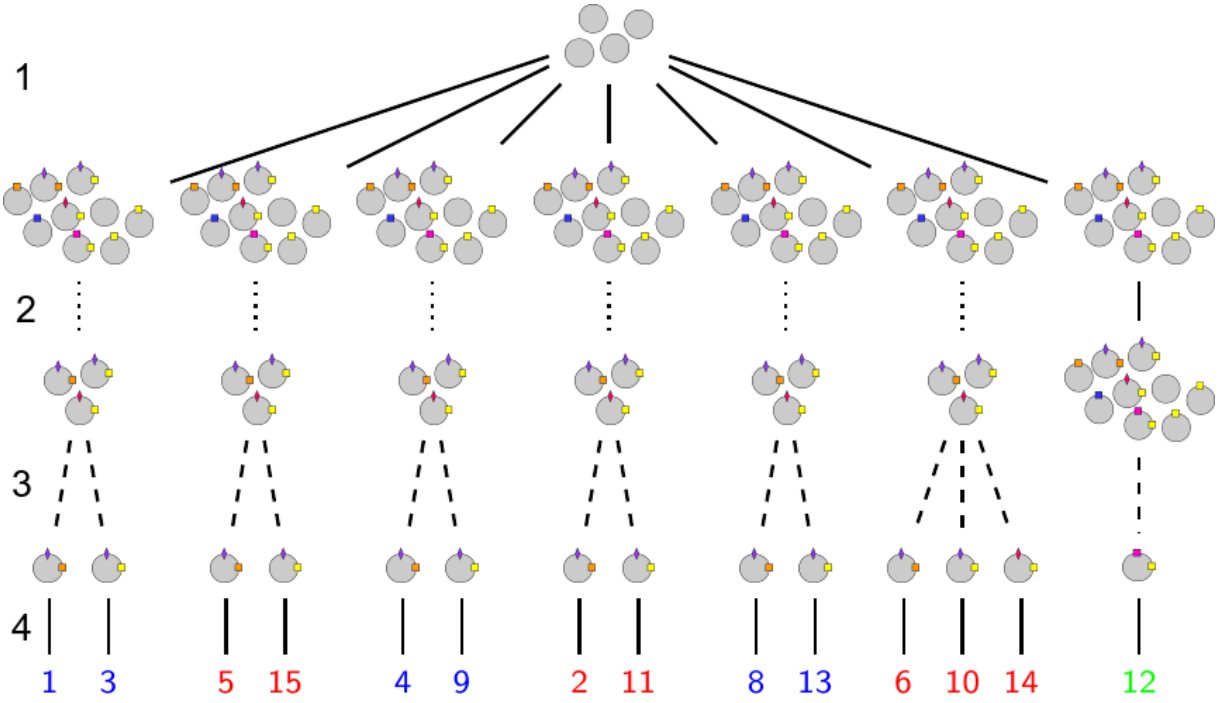


Figure C.6: **Clone trees of drug selected parasites**. Clone trees were obtained after rounds of (1) growth, (2) drug selection (dotted lines) or not (solid line), (3) cloning and (4) sequencing after sufficient DNA is available.

C.2.2 Material and methods

The sequencing was performed in two independent runs of 150 bp paired-end Illumina miniseq by Montpellier GenomiX. The sequencing reads obtained from both runs were merge together for each sample to increase the average depth. Sequence adapters were

[1]LPHI, MIVEGEC, Université de Montpellier, CNRS, INSERM, Montpellier, France

trimmed with Trimmomatic (version 0.39) and filtered reads were mapped on the 3D7 reference genome (version 46) with bwa 0.17.17 [268, 264]. Single-nucleotide polymorphisms (SNPs) and micro-insertions/deletions (INDELs) were discovered with GATK 4.2.0.0 HaplotypeCaller, CombineGVCFs and GenotypeGVCFs and were later filtered with varif (https://github.com/marcguery/varif, version 0.2.2) that was parametrized to keep only variants whose allele frequency was above 0.8 in at least one drug-resistant sample and below 0.02 in the drug-sensitive sample [121]. Filtered variants were used to cluster samples by similarity of mutations from the cluster (version 2.1.3) R package [285, 125]. Structural variants (macro-INDELs, duplications, inversions and translocations) were discovered and filtered with DELLY (version 0.8.7, '-p -f somatic -a 0.5 -r 0.5 -v 10 -c 0.01') [132].

C.2.3 Results

The 8182 variants (1591 SNPs and 6591 micro-INDELs) detected in all 14 samples combined were reduced to 1216 variants (258 SNPs and 958 micro-INDELs) present in drug resistant samples and absent in the drug sensitive one. Among those 1216 variants, 246 (48 SNPs and 198 micro-INDELs) were located within a gene with 41 leading to a non-synonymous mutation (20 SNPs and 21 micro-INDELs). The 14 proteins displaying non-synonymous mutations induced by SNPs are ordered by their number of mutations among all samples (Figure C.7). The cGMP-dependant protein kinase (PfPKG; PF3D7_1436600) showed 4 independent amino acid substitutions (R420I, H524Y, H524N and D597Y) in 12 out of 13 drug-resistant samples. Interestingly, the only sample with an absence of mutation in PfPKG has a non-synonymous mutation (Y539D) in a protein also involved in a cyclic nucleotide related pathway, the 3,5-cyclic nucleotide phosphodiesterase beta (PfPDE beta; PF3D7_1321500). The second most prevalent mutated protein, gametocyte development protein 1 (PfGDV1; PF3D7_0935400), has a stop codon within its reading frame (K561_) for 8 out of 13 drug-resistant samples. However, as PfGDV1 is involved in gametocyte production, its mutation is unlikely to be directly related to the drug resistance. Another protein of unknown annotation (PF3D7_1462400) has two independent stop codons inserted within its reading frame (Y344_ and Y246_) as a result of the two SNPs in 4 out of 13 drug-resistant samples.

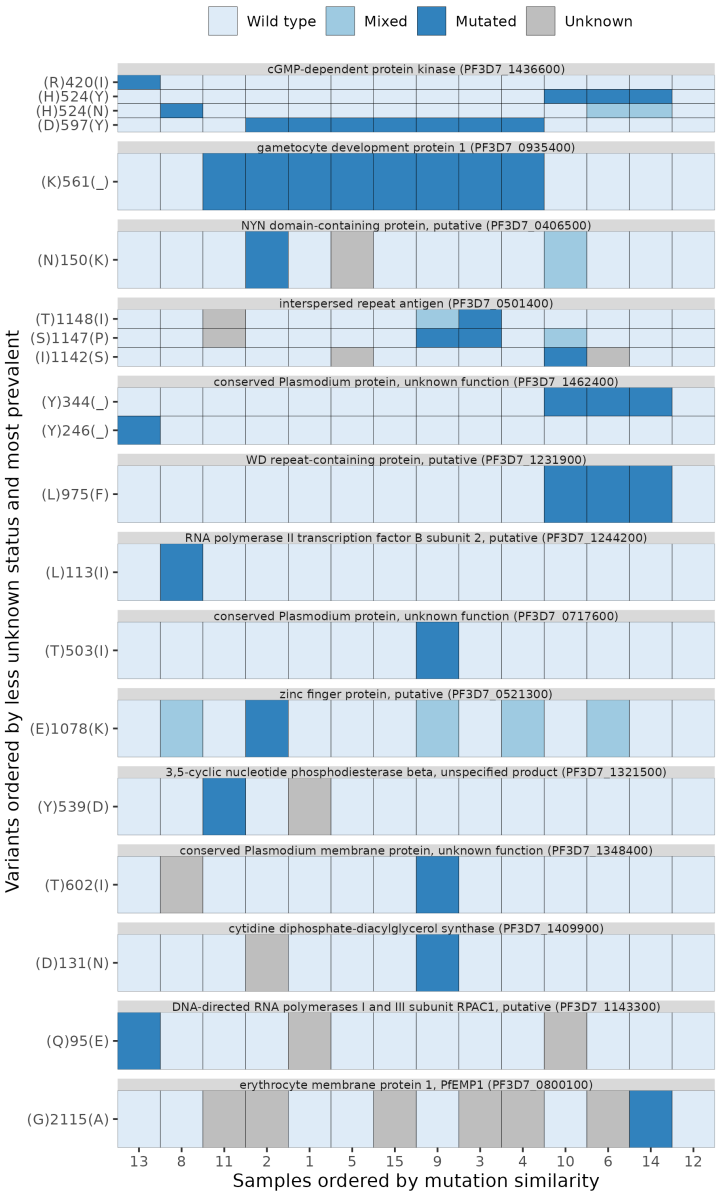


Figure C.7: **Non synonymous mutations induced by SNPs**. SNPs were kept only if they were found in at least one drug resistant sample and absent from the drug sensitive sample. Samples were ordered by their similarity over all 8182 initial variants detected, regardless of their potential relationship to the drug resistant phenotype.

Other proteins with either SNPs or micro-INDELs are unlikely to be directly responsible for the drug resistance because their annotation is unrelated to cyclic nucleotide related pathways, they are present in only a few drug resistant samples or they affect a low complexity region of a protein. Regarding structural variants, a duplication of the end part of the chromosome 7 was found in a majority of drug resistant samples.

C.2.4 Conclusion

The 5 non-synonymous mutations found in PfPKG and PfPDE beta, two proteins involved in cyclic nucleotide related pathways, were promising candidates that might be involved in the observed reduction of sensitivity to the anti-malarial compound, being a purine analogue. This work was followed by the project of a PhD student in molecular biology, Marie Ali[1], who was interested in finding the targets of the antimalarial synthetic compound. Thanks to her work that is now completed, the 5 potential drug resistance-related SNPs identified in this project were all confirmed to confer a resistance to the synthetic compound when introduced by CRISPR/Cas9 in a sensitive 3D7 strain. Another PhD student, Rea Dura[1], is currently looking at the three-dimensional conformation of one of the identified mutated protein while binding to the antimalarial compound.

[1]LPHI, MIVEGEC, Université de Montpellier, CNRS, INSERM, Montpellier, France

C.3 Duration of asymptomatic infections

C.3.1 Introduction

I participated in a work that was published by Katharine A. Collins[1], Sukai Ceesay[2], Sainabou Drammeh[2], Fatou K. Jaiteh[2], Marc-Antoine Guery[3], Kjerstin Lanke[1], Lynn Grignard[4], Will Stone[4], David J. Conway[4], Umberto D'Alessandro[2], Teun Bousema[1,4] and Antoine Claessens[2,3] in 'K. A. Collins *et al.*, A Cohort Study on the Duration of Plasmodium falciparum Infections During the Dry Season in The Gambia, The Journal of Infectious Diseases, vol. 226, no. 1, pp. 128–137, Jul. 2022, doi: 10.1093/infdis/jiac116' [24].

Collaborators work In December 2016, 42 individuals of 4 nearby villages of the eastern part of The Gambia were PCR positive and presented no symptoms. The enrolled individuals were followed monthly with *var*ATS qPCR and quantitative reverse-transcription (qRT)-PCR tests until the end of the low transmission season in May 2017 if they did not show any symptom (if symptomatic, they received an antimalarial treatment) and did not clear the infection (PCR negative 2 months in a row). Parasitaemia was quantified by *var*ATS qPCR and the different blood stages were quantified by (qRT)-PCR. Parasite genotyping was obtained using the differential fragment sizes of MSP2 inferred from the migration of PCR bands.

My involvement My role was to estimate the number of *Plasmodium falciparum* genotypes present in each blood isolate and to follow, for each individual, the resurgence of parasite genotypes over time by grouping those with the same MSP2 fragment length.

C.3.2 Material and methods

The detailed material and methods for this study are available in Collins *et al.*, 2021 [24]. In short, only fragment lengths between 193 and 506 associated with a relative fluorescent unit above 200 were considered. Two MSP2 fragments from the same individual infection sampled at different time points are assumed to correspond to the same genotype if they differ by less than 5 bases in size.

[1]Radboud university medical center, Radboud Institute for Health Sciences, Department of Medical Microbiology, Nijmegen, Netherlands
[2]Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, Banjul, The Gambia
[3]LPHI, MIVEGEC, Université de Montpellier, CNRS, INSERM, Montpellier, France
[4]Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom

C.3.3 Results

Thanks to the comparison of sizes between the different MSP2 fragments, I could produce a supplementary figure (Figure S3) published in the article of Collins *et al.* (2022) that is available at https://doi.org/10.1093/infdis/jiac116 [24]. For each individual infection, parasite strains were compared and considered identical between time points when they had a similar size. The short-lived infections (less than 3 months) have a lower level of clonality, with only two different genotypes at most, compared to persistent infections often displaying more than 2 genotypes, one infection even reaching 15 genotypes. One persistent infection had however a single parasite genotype that remained throughout the dry season in December, January, February, March and April (no information is available for May). In some volunteers, novel genotypes appear over time, indicating a too low percentage of circulating parasites of this genotype earlier or an unlikely new infection during the dry season. In the majority of persistent infections (12/22), the same MSP2 genotype was detected twice at 5-month interval at least, indicating that *Plasmodium falciparum* asymptomatic infections can last several months without being cleared by the host immune system.

C.3.4 Conclusion

This visualisation of the different genotypes over time in each individual infection serves readers of the associated article that wish to seek more information about the study. It was also used to determine the clonality of infections and the level of overlap between strains identified at different time points in the same individual to aid in the selection of the optimal combination of time points when planning a single-cell sequencing of individual infections regularly blood sampled (*Chapter 3 Project 2: Variants Generated During the Course of an Infection*).

C.4 Risk of clinical malaria

I was involved in reviewing the work of Balotin Fogang[1], Lionel Lellouche[1], Sukai Ceesay[2], Sainabou Drammeh[2], Fatou K. Jaiteh[2], Marc-Antoine Guery[1], Jordi Landier[3], Cynthia Haanappel[4], Janeri Froberg[4], David Conway[5], Umberto D'Alessandro[2], Teun Bousema[4] and Antoine Claessens[1,2,4], who were interested in following malaria infection status and parasitaemia from individuals that were tested in four nearby villages of The Gambia between 2014 and 2016. One of the findings was that individuals were at higher risk of developing a clinical malaria after asymptomatically carrying parasites during a low transmission season. The individuals followed in this study were also the ones that had their parasites genotyped, bulk sequenced, single-cell sequenced and long-read sequenced in both my thesis projects (*Chapter 2 Project 1: Spatio-temporal Relatedness of Parasites* & *Chapter 3 Project 2: Variants Generated During the Course of an Infection*). This work has been submitted to a journal for publication and is currently available as a pre-print in 'B. Fogang *et al.*, Asymptomatic carriage of Plasmodium falciparum in children living a hyperendemic area occurs independently of IgG responses but is associated with induction of IL-10. medRxiv, p. 2022.05.04.22274662, May 07, 2022. doi: 10.1101/2022.05.04.22274662' [119].

[1]LPHI, MIVEGEC, Université de Montpellier, CNRS, INSERM, Montpellier, France
[2]Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, Banjul, The Gambia
[3]Aix Marseille Univ, IRD, INSERM, SESSTIM, ISSPAM, 27 boulevard Jean Moulin, 13005, Marseille, France
[4]Radboud university medical center, Radboud Institute for Health Sciences, Department of Medical Microbiology, Nijmegen, The Netherlands
[5]Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, UK

C.5 DBLα diversity

I published a review of the work of Tonkin-Hill *et al.* with my thesis supervisor Antoine Claessens[1] on the use of *var* genes DBLα domains to build homogeneous clusters of samples correlating with their originated country [69, 236]. We highlighted the fact that the use of DBLα sequencing is an affordable and efficient way to accurately separate samples into group corresponding to their country of origin (Figure C.8). The pairwise comparison method that Tonkin-Hill *et al.* developed was sensitive enough to show that South American samples were closer to African than Asian or Oceanian samples, confirming previous findings [286].

[1] LPHI, MIVEGEC, Université de Montpellier, CNRS, INSERM, Montpellier, France

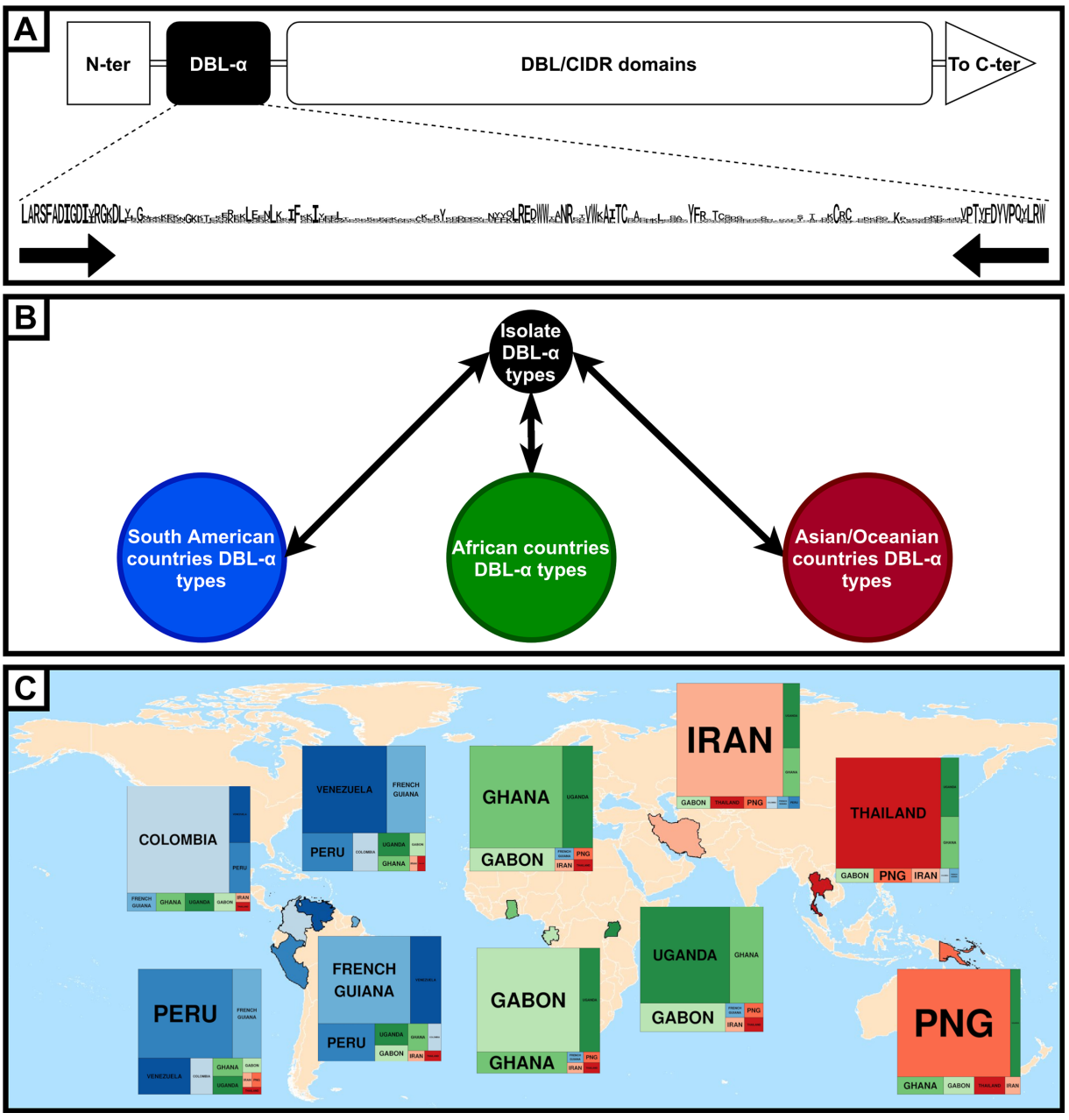


Figure C.8: **DBLα diversity across continents**. A: The variable region of the DBLα domain surrounded by conserved motifs LARSFADIG and DYVPQ[YF]LRW (DBLα tag) was used to identify each parasite genotype by a single DBLα type. B: The different DBLα types present in each isolate were compared to those observed in each of the 10 countries across three continents (South American countries in blue, African countries in green and Asian/Oceanian countries in red) to estimate the geographical origin of the isolate. C: Isolates were grouped by sampling location with their overall estimated geographical origin weighted by rectangle areas. Geographical origins of isolates inferred with DBLα types are consistent with the sampling origin of isolates, meaning that DBLα tags are able to describe the genetic diversity of *Plasmodium falciparum* at the country level. Data from Tonkin-Hill *et al.* [69]. Made with Natural Earth.

Bibliography

[1] Cox FE. History of the discovery of the malaria parasites and their vectors. Parasites & Vectors. 2010 Feb;3(1):5. Available from: https://doi.org/10.1186/1756-3305-3-5.

[2] Hempelmann E, Krafts K. Bad air, amulets and mosquitoes: 2,000 years of changing perspectives on malaria. Malaria Journal. 2013 Jul;12:232. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3723432/.

[3] Laveran AAdt. Nature parasitaire des accidents de l'impaludisme : description d'un nouveau parasite trouvé dans le sang des malades atteints de fièvre palustre / par A. Laveran,...; 1881. Available from: https://gallica.bnf.fr/ark:/12148/bpt6k9761344p.

[4] Boualam MA, Pradines B, Drancourt M, Barbieri R. Malaria in Europe: A Historical Perspective. Frontiers in Medicine. 2021;8. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2021.691095.

[5] Grassi B. Studi di uno zoologo sulla malaria. 2nd ed. Roma: R. Accademia dei lincei; 1901. Pages: 1-354. Available from: https://www.biodiversitylibrary.org/bibliography/37999.

[6] Prugnolle F, Durand P, Neel C, Ollomo B, Ayala FJ, Arnathau C, et al. African great apes are natural hosts of multiple related malaria species, including Plasmodium falciparum. Proceedings of the National Academy of Sciences. 2010 Jan;107(4):1458-63. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/10.1073/pnas.0914440107.

[7] Rougeron V, Boundenga L, Arnathau C, Durand P, Renaud F, Prugnolle F. A population genetic perspective on the origin, spread and adaptation of the human malaria agents Plasmodium falciparum and Plasmodium vivax. FEMS Microbiology Reviews. 2022 Jan;46(1):fuab047. Available from: https://doi.org/10.1093/femsre/fuab047.

[8] Price RN, Commons RJ, Battle KE, Thriemer K, Mendis K. Plasmodium vivax in the Era of the Shrinking P. falciparum Map. Trends in Parasitology. 2020 Jun;36(6):560-70. Publisher: Elsevier. Available from: https://www.cell.com/trends/parasitology/abstract/S1471-4922(20)30074-X.

[9] Garg S, Agarwal S, Dabral S, Kumar N, Sehrawat S, Singh S. Visualization and quantification of Plasmodium falciparum intraerythrocytic merozoites. Systems and Synthetic Biology. 2015 Dec;9(Suppl 1):23-6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4688406/.

[10] Nsanzabana C. Strengthening Surveillance Systems for Malaria Elimination by Integrating Molecular and Genomic Data. Tropical Medicine and Infectious Disease. 2019 Dec;4(4):139. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6958499/.

[11] Wu L, van den Hoogen LL, Slater H, Walker PGT, Ghani AC, Drakeley CJ, et al. Comparison of diagnostics for the detection of asymptomatic Plasmodium falciparum infections to inform control and elimination strategies. Nature. 2015 Dec;528(7580):S86-93. Number: 7580 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature16039.

[12] World Health Organization. World malaria report 2022. World Health Organization; 2022. Available from: https://apps.who.int/iris/handle/10665/365169.

[13] Blasco B, Leroy D, Fidock DA. Antimalarial drug resistance: linking Plasmodium falciparum parasite biology to the clinic. Nature Medicine. 2017 Aug;23(8):917-28. Number: 8 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nm.4381.

[14] Pluijm RWvd, Amaratunga C, Dhorda M, Dondorp AM. Triple Artemisinin-Based Combination Therapies for Malaria – A New Paradigm? Trends in Parasitology. 2021 Jan;37(1):15-24. Publisher: Elsevier. Available from: https://www.cell.com/trends/parasitology/abstract/S1471-4922(20)30254-3.

[15] Wangdi K, Furuya-Kanamori L, Clark J, Barendregt JJ, Gatton ML, Banwell C, et al. Comparative effectiveness of malaria prevention measures: a systematic review and network meta-analysis. Parasites & Vectors. 2018 Mar;11(1):210. Available from: https://doi.org/10.1186/s13071-018-2783-y.

[16] Riveron JM, Tchouakui M, Mugenzi L, D Menze B, Chiang MC, Wondji CS, et al. Insecticide Resistance in Malaria Vectors: An Update at a Global Scale. In: Towards Malaria Elimination - A Leap Forward. IntechOpen; 2018. Available from: https://www.intechopen.com/chapters/62169.

[17] van den Berg H, da Silva Bezerra HS, Al-Eryani S, Chanda E, Nagpal BN, Knox TB, et al. Recent trends in global insecticide use for disease vector control and potential implications for resistance management. Scientific Reports. 2021 Dec;11(1):23867. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41598-021-03367-9.

[18] Weiss DJ, Bertozzi-Villa A, Rumisha SF, Amratia P, Arambepola R, Battle KE, et al. Indirect effects of the COVID-19 pandemic on malaria intervention coverage, morbidity, and mortality in Africa: a geospatial modelling analysis. The Lancet Infectious Diseases. 2021 Jan;21(1):59-69. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7505634/.

[19] Björkman A, Morris U. Why Asymptomatic Plasmodium falciparum Infections Are Common in Low-Transmission Settings. Trends in Parasitology. 2020 Nov;36(11):898-905. Publisher: Elsevier. Available from: https://www.cell.com/trends/parasitology/abstract/S1471-4922(20)30191-4.

[20] Malpartida-Cardenas K, Moser N, Ansah F, Pennisi I, Ahu Prah D, Amoah LE, et al. Sensitive Detection of Asymptomatic and Symptomatic Malaria with Seven Novel Parasite-Specific LAMP Assays and Translation for Use at Point-of-Care. Microbiology Spectrum. 2023 May;11(3):e05222-2. Publisher: American Society for Microbiology. Available from: https://journals.asm.org/doi/10.1128/spectrum.05222-22.

[21] Fogang B, Schoenhals M, Maloba F, Abite MF, Essangui E, Donkeu C, et al.. Asymptomatic carriage of Plasmodium falciparum in children living a hyperendemic area occurs independently of IgG responses but is associated with induction of IL-10. medRxiv; 2022. Pages: 2022.05.04.22274662. Available from: https://www.medrxiv.org/content/10.1101/2022.05.04.22274662v1.

[22] Andolina C, Rek JC, Briggs J, Okoth J, Musiime A, Ramjith J, et al. Sources of persistent malaria transmission in a setting with effective malaria control in eastern Uganda: a longitudinal, observational cohort study. The Lancet Infectious Diseases. 2021 Nov;21(11):1568-78.

[23] Agaba BB, Rugera SP, Mpirirwe R, Atekat M, Okubal S, Masereka K, et al. Asymptomatic malaria infection, associated factors and accuracy of diagnostic tests in a historically high transmission setting in Northern Uganda. Malaria Journal. 2022 Dec;21(1):392. Available from: https://doi.org/10.1186/s12936-022-04421-1.

[24] Collins KA, Ceesay S, Drammeh S, Jaiteh FK, Guery MA, Lanke K, et al. A Cohort Study on the Duration of Plasmodium falciparum Infections During the Dry Season in The Gambia. The Journal of Infectious Diseases. 2022 Jul;226(1):128-37. Available from: https://doi.org/10.1093/infdis/jiac116.

[25] Briggs J, Teyssier N, Nankabirwa JI, Rek J, Jagannathan P, Arinaitwe E, et al. Sex-based differences in clearance of chronic Plasmodium falciparum infection. eLife. 2020;9. Publisher: eLife Sciences Publications, Ltd. Available from: https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC7591246/.

[26] Achan J, Talisuna AO, Erhart A, Yeka A, Tibenderana JK, Baliraine FN, et al. Quinine, an old anti-malarial drug in a modern world: role in the treatment of malaria. Malaria Journal. 2011 May;10:144. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3121651/.

[27] Renslo AR. Antimalarial Drug Discovery: From Quinine to the Dream of Eradication. ACS Medicinal Chemistry Letters. 2013 Dec;4(12):1126-8. Publisher: American Chemical Society. Available from: https://doi.org/10.1021/ml4004414.

[28] Roux AT, Maharaj L, Oyegoke O, Akoniyon OP, Adeleke MA, Maharaj R, et al. Chloroquine and Sulfadoxine–Pyrimethamine Resistance in Sub-Saharan Africa—A Review. Frontiers in Genetics. 2021;12. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2021.668574.

[29] Roper C, Pearce R, Nair S, Sharp B, Nosten F, Anderson T. Intercontinental Spread of Pyrimethamine-Resistant Malaria. Science. 2004 Aug;305(5687):1124-4. Publisher: American Association for the Advancement of Science. Available from: https://www.science.org/doi/10.1126/science.1098876.

[30] Tse EG, Korsik M, Todd MH. The past, present and future of anti-malarial medicines. Malaria Journal. 2019 Mar;18(1):93. Available from: https://doi.org/10.1186/s12936-019-2724-z.

[31] Li G, Guo X, Arnold K, Jian H, Fu L. RANDOMISED COMPARATIVE STUDY OF MEFLOQUINE, QINGHAOSU, AND PYRIMETHAMINE-SULFADOXINE IN PATIENTS WITH FALCIPARUM MALARIA. The Lancet. 1984 Dec;324(8416):1360-1. Available from: https://www.sciencedirect.com/science/article/pii/S0140673684920579.

[32] Amato R, Pearson RD, Almagro-Garcia J, Amaratunga C, Lim P, Suon S, et al. Origins of the current outbreak of multidrug-resistant malaria in southeast Asia: a retrospective genetic study. The Lancet Infectious Diseases. 2018 Mar;18(3):337-45. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5835763/.

[33] Hamilton WL, Amato R, Pluijm RWvd, Jacob CG, Quang HH, Thuy-Nhien NT, et al. Evolution and expansion of multidrug-resistant malaria in southeast Asia: a genomic epidemiology study. The Lancet Infectious Diseases. 2019 Sep;19(9):943-51. Publisher: Elsevier. Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(19)30392-5/fulltext.

[34] Imwong M, Dhorda M, Tun KM, Thu AM, Phyo AP, Proux S, et al. Molecular epidemiology of resistance to antimalarial drugs in the Greater Mekong subregion: an observational study. The Lancet Infectious Diseases. 2020 Dec;20(12):1470-80. Publisher: Elsevier. Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30228-0/fulltext.

[35] Walliker D, Quakyi IA, Wellems TE, McCutchan TF, Szarfman A, London WT, et al. Genetic Analysis of the Human Malaria Parasite Plasmodium falciparum. Science. 1987 Jun;236(4809):1661-6. Publisher: American Association for the Advancement of Science. Available from: https://www.science.org/doi/10.1126/science.3299700.

[36] Su Xz. Tracing the geographic origins of Plasmodium falciparum malaria parasites. Pathogens and Global Health. 2014 Sep;108(6):261-2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4216747/.

[37] Moser KA, Drábek EF, Dwivedi A, Stucke EM, Crabtree J, Dara A, et al. Strains used in whole organism Plasmodium falciparum vaccine trials differ in genome structure, sequence, and immunogenic potential. Genome Medicine. 2020 Jan;12(1):6. Available from: https://doi.org/10.1186/s13073-019-0708-9.

[38] Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature. 2002 Oct;419(6906):498-511.

[39] Preiser PR, Wilson RJ, Moore PW, McCready S, Hajibagheri MA, Blight KJ, et al. Recombination associated with replication of malarial mitochondrial DNA. The EMBO journal. 1996 Feb;15(3):684-93.

[40] Matsuzaki M, Kikuchi T, Kita K, Kojima S, Kuroiwa T. Large amounts of apicoplast nucleoid DNA and its segregation in Toxoplasma gondii. Protoplasma. 2001;218(3-4):180-91.

[41] Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, et al. Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. Genome Research. 2016 Sep;26(9):1288-99. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.203711.115.

[42] Otto TD, Böhme U, Sanders M, Reid A, Bruske EI, Duffy CW, et al. Long read assemblies of geographically dispersed Plasmodium falciparum isolates reveal highly structured subtelomeres. Wellcome Open Research. 2018 May;3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5964635/.

[43] Howard RJ, Barnwell JW, Rock EP, Neequaye J, Ofori-Adjei D, Lee Maloy W, et al. Two approximately 300 kilodalton Plasmodium falciparum proteins at the surface membrane of infected erythrocytes. Molecular and Biochemical Parasitology. 1988 Jan;27(2):207-23. Available from: https://www.sciencedirect.com/science/article/pii/0166685188900400.

[44] Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, et al. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes. Cell. 1995 Jul;82(1):89-100.

[45] Crabb BS, Cooke BM, Reeder JC, Waller RF, Caruana SR, Davern KM, et al. Targeted gene disruption shows that knobs enable malaria-infected red cells to cytoadhere under physiological shear stress. Cell. 1997 Apr;89(2):287-96.

[46] Andradi-Brown C, Wichers-Misterek JS, Thien Hv, Höppner YD, Scholz JAM, Hansson HS, et al.. A novel computational pipeline for var gene expression augments the discovery of changes in the Plasmodium falciparum transcriptome during transition from in vivo to short-term in vitro culture. bioRxiv; 2023. Pages: 2023.03.21.533599

Section: New Results. Available from: https://www.biorxiv.org/content/10.1101/2023.03.21.533599v1.

[47] Rask TS, Hansen DA, Theander TG, Pedersen AG, Lavstsen T. Plasmodium falciparum Erythrocyte Membrane Protein 1 Diversity in Seven Genomes – Divide and Conquer. PLOS Computational Biology. 2010 Sep;6(9):e1000933. Publisher: Public Library of Science. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000933.

[48] Lavstsen T, Salanti A, Jensen AT, Arnot DE, Theander TG. Sub-grouping of Plasmodium falciparum 3D7 var genes based on sequence analysis of coding and non-coding regions. Malaria Journal. 2003 Sep;2(1):27. Available from: https://doi.org/10.1186/1475-2875-2-27.

[49] Smith JD, Subramanian G, Gamain B, Baruch DI, Miller LH. Classification of adhesive domains in the Plasmodium falciparum Erythrocyte Membrane Protein 1 family. Molecular and Biochemical Parasitology. 2000 Oct;110(2):293-310. Available from: http://www.sciencedirect.com/science/article/pii/S0166685100002796.

[50] Taylor HM, Kyes SA, Newbold CI. Var gene diversity in Plasmodium falciparum is generated by frequent recombination events. Molecular and Biochemical Parasitology. 2000 Oct;110(2):391-7.

[51] Otto TD, Assefa SA, Böhme U, Sanders MJ, Kwiatkowski D, Berriman M, et al. Evolutionary analysis of the most polymorphic gene family in falciparum malaria. Wellcome Open Research. 2019 Dec;4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7001760/.

[52] Vembar SS, Seetin M, Lambert C, Nattestad M, Schatz MC, Baybayan P, et al. Complete telomere-to-telomere de novo assembly of the Plasmodium falciparum genome through long-read (>11 kb), single molecule, real-time sequencing. DNA research: an international journal for rapid publication of reports on genes and genomes. 2016 Aug;23(4):339-51.

[53] Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, et al. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of P. falciparum. Nature. 2000 Oct;407(6807):1018-22. Number: 6807 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/35039531.

[54] Claessens A, Hamilton WL, Kekre M, Otto TD, Faizullabhoy A, Rayner JC, et al. Generation of Antigenic Diversity in Plasmodium falciparum by Structured Rearrangement of Var Genes During Mitosis. PLoS Genetics. 2014 Dec;10(12):e1004812. Available from: http://dx.plos.org/10.1371/journal.pgen.1004812.

[55] Chan JA, Fowkes FJI, Beeson JG. Surface antigens of Plasmodium falciparum-infected erythrocytes as immune targets and malaria vaccine candidates. Cellular and Molecular Life Sciences. 2014 Oct;71(19):3633-57. Available from: https://doi.org/10.1007/s00018-014-1614-3.

[56] Joannin N, Abhiman S, Sonnhammer EL, Wahlgren M. Sub-grouping and sub-functionalization of the RIFIN multi-copy protein family. BMC Genomics. 2008 Jan;9:19. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2257938/.

[57] Frank M, Kirkman L, Costantini D, Sanyal S, Lavazec C, Templeton TJ, et al. Frequent recombination events generate diversity within the multi-copy variant antigen gene families of Plasmodium falciparum. International Journal for Parasitology. 2008 Aug;38(10):1099-109.

[58] Neafsey DE, Taylor AR, MacInnis BL. Advances and opportunities in malaria population genomics. Nature Reviews Genetics. 2021 Aug;22(8):502-17. Number: 8 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41576-021-00349-5.

[59] Nkhoma SC, Trevino SG, Gorena KM, Nair S, Khoswe S, Jett C, et al. Co-transmission of Related Malaria Parasite Lineages Shapes Within-Host Parasite Diversity. Cell Host & Microbe. 2020 Jan;27(1):93-103.e4. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1931312819306304.

[60] Daniels RF, Schaffner SF, Wenger EA, Proctor JL, Chang HH, Wong W, et al. Modeling malaria genomics reveals transmission decline and rebound in Senegal. Proceedings of the National Academy of Sciences. 2015 Jun;112(22):7067-72. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1505691112.

[61] Nkhoma SC, Nair S, Al-Saai S, Ashley E, McGready R, Phyo AP, et al. Population genetic correlates of declining transmission in a human pathogen. Molecular Ecology. 2013 Jan;22(2):273-85. Available from: http://doi.wiley.com/10.1111/mec.12099.

[62] Watson OJ, Okell LC, Hellewell J, Slater HC, Unwin HJT, Omedo I, et al. Evaluating the Performance of Malaria Genetics for Inferring Changes in Transmission Intensity Using Transmission Modeling. Molecular Biology and Evolution. 2021 Jan;38(1):274-89.

[63] Hendry JA, Kwiatkowski D, McVean G. Elucidating relationships between P.falciparum prevalence and measures of genetic diversity with a combined genetic-epidemiological model of malaria. PLoS computational biology. 2021 Aug;17(8):e1009287.

[64] Miller RH, Hathaway NJ, Kharabora O, Mwandagalirwa K, Tshefu A, Meshnick SR, et al. A deep sequencing approach to estimate Plasmodium falciparum complexity

of infection (COI) and explore apical membrane antigen 1 diversity. Malaria Journal. 2017 Dec;16(1):490. Available from: https://doi.org/10.1186/s12936-017-2137-9.

[65] Lopez L, Koepfli C. Systematic review of Plasmodium falciparum and Plasmodium vivax polyclonal infections: Impact of prevalence, study population characteristics, and laboratory procedures. PLOS ONE. 2021 Jun;16(6):e0249382. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0249382.

[66] Somé AF, Bazié T, Zongo I, Yerbanga RS, Nikiéma F, Neya C, et al. Plasmodium falciparum msp1 and msp2 genetic diversity and allele frequencies in parasites isolated from symptomatic malaria patients in Bobo-Dioulasso, Burkina Faso. Parasites & Vectors. 2018 May;11(1):323. Available from: https://doi.org/10.1186/s13071-018-2895-4.

[67] Oboh MA, Ndiaye T, Diongue K, Ndiaye YD, Sy M, Deme AB, et al. Allelic diversity of MSP1 and MSP2 repeat loci correlate with levels of malaria endemicity in Senegal and Nigerian populations. Malaria Journal. 2021 Jan;20(1):38. Available from: https://doi.org/10.1186/s12936-020-03563-4.

[68] Mueller I, Schoepflin S, Smith TA, Benton KL, Bretscher MT, Lin E, et al. Force of infection is key to understanding the epidemiology of Plasmodium falciparum malaria in Papua New Guinean children. Proceedings of the National Academy of Sciences of the United States of America. 2012 Jun;109(25):10030-5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3382533/.

[69] Tonkin-Hill G, Ruybal-Pesántez S, Tiedje KE, Rougeron V, Duffy MF, Zakeri S, et al. Evolutionary analyses of the major variant surface antigen-encoding genes reveal population structure of Plasmodium falciparum within and between continents. PLOS Genetics. 2021 Feb;17(2):e1009269. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009269.

[70] Tan MH, Shim H, Chan Yb, Day KP. Unravelling var complexity: Relationship between DBLα types and var genes in Plasmodium falciparum. Frontiers in Parasitology. 2023;1. Available from: https://www.frontiersin.org/articles/10.3389/fpara.2022.1006341.

[71] Daniels R, Volkman SK, Milner DA, Mahesh N, Neafsey DE, Park DJ, et al. A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking. Malaria Journal. 2008;7(1):223. Available from: http://malariajournal.biomedcentral.com/articles/10.1186/1475-2875-7-223.

[72] Bankole BE, Kayode AT, Nosamiefan IO, Eromon P, Baniecki ML, Daniels RF, et al. Characterization of Plasmodium falciparum structure in Nigeria with malaria SNPs barcode. Malaria Journal. 2018 Dec;17. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6296064/.

[73] Amambua-Ngwa A, Jeffries D, Mwesigwa J, Seedy-Jawara A, Okebe J, Achan J, et al. Long-distance transmission patterns modelled from SNP barcodes of Plasmodium falciparum infections in The Gambia. Scientific Reports. 2019 Dec;9(1):13515. Available from: http://www.nature.com/articles/s41598-019-49991-4.

[74] Mobegi VA, Loua KM, Ahouidi AD, Satoguina J, Nwakanma DC, Amambua-Ngwa A, et al. Population genetic structure of Plasmodium falciparum across a region of diverse endemicity in West Africa. Malaria Journal. 2012 Jul;11(1):223. Available from: https://doi.org/10.1186/1475-2875-11-223.

[75] Touray AO, Mobegi VA, Wamunyokoli F, Herren JK. Diversity and Multiplicity of P. falciparum infections among asymptomatic school children in Mbita, Western Kenya. Scientific Reports. 2020 Apr;10(1):5924. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41598-020-62819-w.

[76] Argyropoulos DC, Tan MH, Adobor C, Mensah B, Labbé F, Tiedje KE, et al. Performance of SNP barcodes to determine genetic diversity and population structure of Plasmodium falciparum in Africa. Frontiers in Genetics. 2023;14. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2023.1071896.

[77] Volkman SK, Sabeti PC, DeCaprio D, Neafsey DE, Schaffner SF, Milner DA, et al. A genome-wide map of diversity in Plasmodium falciparum. Nature Genetics. 2007 Jan;39(1):113-9. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/ng1930.

[78] Amambua-Ngwa A, Amenga-Etego L, Kamau E, Amato R, Ghansah A, Golassa L, et al. Major subpopulations of *Plasmodium falciparum* in sub-Saharan Africa. Science. 2019 Aug;365(6455):813-6. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aav5427.

[79] Schaffner SF, Taylor AR, Wong W, Wirth DF, Neafsey DE. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. Malaria Journal. 2018 Dec;17(1):196. Available from: https://malariajournal.biomedcentral.com/articles/10.1186/s12936-018-2349-7.

[80] Taylor AR, Jacob PE, Neafsey DE, Buckee CO. Estimating Relatedness Between Malaria Parasites. Genetics. 2019 Aug;212(4):1337-51. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6707449/.

[81] Wong W, Volkman S, Daniels R, Schaffner S, Sy M, Ndiaye YD, et al. R H: a genetic metric for measuring intrahost Plasmodium falciparum relatedness and distinguishing cotransmission from superinfection. PNAS nexus. 2022 Sep;1(4):pgac187.

[82] Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, et al. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature. 2012 Jul;487(7407):375-9. Available from: http://www.nature.com/articles/nature11174.

[83] Cranston HA, Boylan CW, Carroll GL, Sutera SP, Williamson JR, Gluzman IY, et al. Plasmodium falciparum Maturation Abolishes Physiologic Red Cell Deformability. Science. 1984 Jan;223(4634):400-3. Publisher: American Association for the Advancement of Science. Available from: https://www.science.org/doi/10.1126/science.6362007.

[84] Lee WC, Russell B, Rénia L. Sticking for a Cause: The Falciparum Malaria Parasites Cytoadherence Paradigm. Frontiers in Immunology. 2019;10. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2019.01444.

[85] Saito F, Hirayasu K, Satoh T, Wang CW, Lusingu J, Arimori T, et al. Immune evasion of Plasmodium falciparum by RIFIN via inhibitory receptors. Nature. 2017 Dec;552(7683):101-5. Number: 7683 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature24994.

[86] Ralph SA, Scheidig-Benatar C, Scherf A. Antigenic variation in Plasmodium falciparum is associated with movement of var loci between subnuclear locations. Proceedings of the National Academy of Sciences. 2005 Apr;102(15):5414-9. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/10.1073/pnas.0408883102.

[87] Lopez-Rubio JJ, Gontijo AM, Nunes MC, Issar N, Hernandez Rivas R, Scherf A. 5' flanking region of var genes nucleate histone modification patterns linked to phenotypic inheritance of virulence traits in malaria parasites. Molecular Microbiology. 2007 Dec;66(6):1296-305.

[88] Roberts DJ, Craig, Berendt AR, Pinches R, Nash G, Marsh K, et al. Rapid switching to multiple antigenic and adhesive phenotypes in malaria. Nature. 1992 Jun;357(6380):689-92. Number: 6380 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/357689a0.

[89] Ye R, Zhang D, Chen B, Zhu Y, Zhang Y, Wang S, et al. Transcription of the var genes from a freshly-obtained field isolate of Plasmodium falciparum shows more variable switching patterns than long laboratory-adapted isolates. Malaria Journal. 2015 Feb;14(1):66. Available from: https://doi.org/10.1186/s12936-015-0565-y.

[90] Frank M, Dzikowski R, Amulic B, Deitsch K. Variable switching rates of malaria virulence genes are associated with chromosomal position. Molecular Microbiology. 2007;64(6):1486-98. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2958.2007.05736.x. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2958.2007.05736.x.

[91] Mwesigwa J, Okebe J, Affara M, Di Tanna GL, Nwakanma D, Janha O, et al. On-going malaria transmission in The Gambia despite high coverage of control interventions: a nationwide cross-sectional survey. Malaria Journal. 2015 Aug;14(1):314. Available from: https://doi.org/10.1186/s12936-015-0829-6.

[92] Fall P, Diouf I, Deme A, Sene D. Assessment of Climate-Driven Variations in Malaria Transmission in Senegal Using the VECTRI Model. Atmosphere. 2022 Mar;13(3):418. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/2073-4433/13/3/418.

[93] Jorgensen P, Nambanya S, Gopinath D, Hongvanthong B, Luangphengsouk K, Bell D, et al. High heterogeneity in Plasmodium falciparum risk illustrates the need for detailed mapping to guide resource allocation: a new malaria risk map of the Lao People's Democratic Republic. Malaria Journal. 2010 Feb;9(1):59. Available from: https://doi.org/10.1186/1475-2875-9-59.

[94] Amratia P, Psychas P, Abuaku B, Ahorlu C, Millar J, Oppong S, et al. Characterizing local-scale heterogeneity of malaria risk: a case study in Bunkpurugu-Yunyoo district in northern Ghana. Malaria Journal. 2019 Mar;18:81. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6420752/.

[95] Bousema T, Drakeley C, Gesase S, Hashim R, Magesa S, Mosha F, et al. Identification of Hot Spots of Malaria Transmission for Targeted Malaria Control. The Journal of Infectious Diseases. 2010 Jun;201(11):1764-74. Available from: https://doi.org/10.1086/652456.

[96] Daniels RF, Schaffner SF, Dieye Y, Dieng G, Hainsworth M, Fall FB, et al. Genetic evidence for imported malaria and local transmission in Richard Toll, Senegal. Malaria Journal. 2020 Dec;19(1):276. Available from: https://malariajournal.biomedcentral.com/articles/10.1186/s12936-020-03346-x.

[97] Bei AK, Niang M, Deme AB, Daniels RF, Sarr FD, Sokhna C, et al. Dramatic Changes in Malaria Population Genetic Complexity in Dielmo and Ndiop, Senegal, Revealed Using Genomic Surveillance. The Journal of Infectious Diseases. 2018 Jan;217(4):622-7. Available from: https://academic.oup.com/jid/article/217/4/622/4793403.

[98] Taylor AR, Echeverry DF, Anderson TJC, Neafsey DE, Buckee CO. Identity-by-descent with uncertainty characterises connectivity of Plasmodium falciparum populations on

the Colombian-Pacific coast. PLOS Genetics. 2020 Nov;16(11):e1009101. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009101.

[99] Briggs J, Kuchta A, Murphy M, Tessema S, Arinaitwe E, Rek J, et al. Within-household clustering of genetically related Plasmodium falciparum infections in a moderate transmission area of Uganda. Malaria Journal. 2021 Feb;20(1):68. Available from: https://doi.org/10.1186/s12936-021-03603-7.

[100] Fola AA, Moser KA, Aydemir O, Hennelly C, Kobayashi T, Shields T, et al. Temporal and spatial analysis of Plasmodium falciparum genomics reveals patterns of parasite connectivity in a low-transmission district in Southern Province, Zambia. Malaria Journal. 2023 Jul;22(1):208. Available from: https://doi.org/10.1186/s12936-023-04637-9.

[101] Lee A, Ndiaye YD, Badiane A, Deme A, Daniels RF, Schaffner SF, et al. Modeling the levels, trends, and connectivity of malaria transmission using genomic data from a health facility in Thiès, Senegal. medRxiv; 2021. ISSN: 2126-3639. Available from: https://www.medrxiv.org/content/10.1101/2021.09.17.21263639v2.

[102] Duque C, Lubinda M, Matoba J, Sing'anga C, Stevenson J, Shields T, et al. Impact of aerial humidity on seasonal malaria: an ecological study in Zambia. Malaria Journal. 2022 Nov;21(1):325. Available from: https://doi.org/10.1186/s12936-022-04345-w.

[103] Santos-Vega M, Martinez PP, Vaishnav KG, Kohli V, Desai V, Bouma MJ, et al. The neglected role of relative humidity in the interannual variability of urban malaria in Indian cities. Nature Communications. 2022 Jan;13(1):533. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-022-28145-7.

[104] Yé Y, Louis VR, Simboro S, Sauerborn R. Effect of meteorological factors on clinical malaria risk among children: an assessment using village-based meteorological stations and community-based parasitological survey. BMC Public Health. 2007 Jun;7(1):101. Available from: https://doi.org/10.1186/1471-2458-7-101.

[105] Krefis AC, Schwarz NG, Krüger A, Fobil J, Nkrumah B, Acquah S, et al. Modeling the Relationship between Precipitation and Malaria Incidence in Children from a Holoendemic Area in Ghana. The American Journal of Tropical Medicine and Hygiene. 2011 Feb;84(2):285-91. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3029183/.

[106] Herdiana H, Fuad A, Asih PB, Zubaedah S, Arisanti RR, Syafruddin D, et al. Progress towards malaria elimination in Sabang Municipality, Aceh, Indonesia. Malaria

Journal. 2013 Jan;12(1):42. Available from: https://doi.org/10.1186/1475-2875-12-42.

[107] Mwesigwa J, Achan J, Tanna GLD, Affara M, Jawara M, Worwui A, et al. Residual malaria transmission dynamics varies across The Gambia despite high coverage of control interventions. PLOS ONE. 2017 Nov;12(11):e0187059. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0187059.

[108] Ceesay SJ, Casals-Pascual C, Erskine J, Anya SE, Duah NO, Fulford AJ, et al. Changes in malaria indices between 1999 and 2007 in The Gambia: a retrospective analysis. The Lancet. 2008 Nov;372(9649):1545-54. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673608616542.

[109] Baliraine FN, Afrane YA, Amenya DA, Bonizzoni M, Vardo-Zalik AM, Menge DM, et al. A cohort study of Plasmodium falciparum infection dynamics in Western Kenya Highlands. BMC Infectious Diseases. 2010 Dec;10(1):283. Available from: https://bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-10-283.

[110] Amoah LE, Abukari Z, Dawson-Amoah ME, Dieng CC, Lo E, Afrane YA. Population structure and diversity of Plasmodium falciparum in children with asymptomatic malaria living in different ecological zones of Ghana. BMC Infectious Diseases. 2021 May;21:439. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8120845/.

[111] Coulibaly D, Travassos MA, Tolo Y, Laurens MB, Kone AK, Traore K, et al. Spatio-Temporal Dynamics of Asymptomatic Malaria: Bridging the Gap Between Annual Malaria Resurgences in a Sahelian Environment. The American Journal of Tropical Medicine and Hygiene. 2017 Dec;97(6):1761-9.

[112] Portugal S, Tran TM, Ongoiba A, Bathily A, Li S, Doumbo S, et al. Treatment of Chronic Asymptomatic Plasmodium falciparum Infection Does Not Increase the Risk of Clinical Malaria Upon Reinfection. Clinical Infectious Diseases. 2017 Mar;64(5):645-53. Available from: https://academic.oup.com/cid/article/64/5/645/2739519.

[113] Sondo P, Derra K, Rouamba T, Nakanabo Diallo S, Taconet P, Kazienga A, et al. Determinants of Plasmodium falciparum multiplicity of infection and genetic diversity in Burkina Faso. Parasites & Vectors. 2020 Aug;13(1):427. Available from: https://doi.org/10.1186/s13071-020-04302-z.

[114] Gwarinda HB, Tessema SK, Raman J, Greenhouse B, Birkholtz LM. Parasite genetic diversity reflects continued residual malaria transmission in Vhembe District, a hotspot in the Limpopo Province of South Africa. Malaria Journal. 2021 Feb;20(1):96. Available from: https://doi.org/10.1186/s12936-021-03635-z.

[115] Oduma CO, Ogolla S, Atieli H, Ondigo BN, Lee MC, Githeko AK, et al. Increased investment in gametocytes in asymptomatic Plasmodium falciparum infections in the wet season. BMC Infectious Diseases. 2021 Jan;21(1):44. Available from: https://doi.org/10.1186/s12879-020-05761-6.

[116] Biabi MFAB, Fogang B, Essangui E, Maloba F, Donkeu C, Keumoe R, et al. High Prevalence of Polyclonal Plasmodium falciparum Infections and Association with Poor IgG Antibody Responses in a Hyper-Endemic Area in Cameroon. Tropical Medicine and Infectious Disease. 2023 Jul;8(8):390. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10459087/.

[117] Andrade CM, Fleckenstein H, Thomson-Luque R, Doumbo S, Lima NF, Anderson C, et al. Increased circulation time of Plasmodium falciparum underlies persistent asymptomatic infection in the dry season. Nature Medicine. 2020 Oct:1-12. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41591-020-1084-0.

[118] Barry A, Bradley J, Stone W, Guelbeogo MW, Lanke K, Ouedraogo A, et al. Higher gametocyte production and mosquito infectivity in chronic compared to incident Plasmodium falciparum infections. Nature Communications. 2021 Apr;12(1):2443. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-021-22573-7.

[119] Fogang B, Lellouche L, Ceesay S, Drammeh S, Jaiteh FK, Guery MA, et al.. Asymptomatic Plasmodium falciparum Carriage at the End of the Dry Season is Associated with Subsequent Infection and Clinical Malaria in Eastern Gambia. medRxiv; 2023. Pages: 2023.09.29.23296347. Available from: https://www.medrxiv.org/content/10.1101/2023.09.29.23296347v1.

[120] Oyola SO, Ariani CV, Hamilton WL, Kekre M, Amenga-Etego LN, Ghansah A, et al. Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. Malaria Journal. 2016 Dec;15(1):597. Available from: https://doi.org/10.1186/s12936-016-1641-7.

[121] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010 Sep;20(9):1297-303. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. Available from: https://genome.cshlp.org/content/20/9/1297.

[122] MalariaGEN, Ahouidi A, Ali M, Almagro-Garcia J, Amambua-Ngwa A, Amaratunga C, et al. An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples. Wellcome Open Research. 2021;6:42.

[123] Jacob CG, Thuy-Nhien N, Mayxay M, Maude RJ, Quang HH, Hongvanthong B, et al. Genetic surveillance in the Greater Mekong subregion and South Asia to support malaria control and elimination. eLife. 2021 Aug;10:e62997. Publisher: eLife Sciences Publications, Ltd. Available from: https://doi.org/10.7554/eLife.62997.

[124] Chang HH, Worby CJ, Yeka A, Nankabirwa J, Kamya MR, Staedke SG, et al. THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. PLOS Computational Biology. 2017 Jan;13(1):e1005348. Available from: https://dx.plos.org/10.1371/journal.pcbi.1005348.

[125] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: https://www.R-project.org/.

[126] Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Systems:1695. Available from: https://igraph.org.

[127] Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Research. 2020 Sep;30(9):1291-305. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab. Available from: https://genome.cshlp.org/content/30/9/1291.

[128] Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods. 2021 Feb;18(2):170-5. Number: 2 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41592-020-01056-5.

[129] Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biology. 2022 Dec;23(1):258. Available from: https://doi.org/10.1186/s13059-022-02823-7.

[130] Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Research. 2006 Jul;34(suppl_2):W435-9. Available from: https://doi.org/10.1093/nar/gkl200.

[131] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011 Oct;7:539.

[132] Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012 Sep;28(18):i333-9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3436805/.

[133] Hill AVS. Vaccines against malaria. Philosophical Transactions of the Royal Society B: Biological Sciences. 2011 Oct;366(1579):2806-14. Publisher: Royal Society. Available from: https://royalsocietypublishing.org/doi/10.1098/rstb.2011.0091.

[134] Ruiz Cuenca P, Key S, Lindblade KA, Vythilingam I, Drakeley C, Fornace K. Is there evidence of sustained human-mosquito-human transmission of the zoonotic malaria Plasmodium knowlesi? A systematic literature review. Malaria Journal. 2022 Mar;21(1):89. Available from: https://doi.org/10.1186/s12936-022-04110-z.

[135] Escalante AA, Cepeda AS, Pacheco MA. Why Plasmodium vivax and Plasmodium falciparum are so different? A tale of two clades and their species diversities. Malaria Journal. 2022 May;21(1):139. Available from: https://doi.org/10.1186/s12936-022-04130-9.

[136] Hayakawa T, Culleton R, Otani H, Horii T, Tanabe K. Big bang in the evolution of extant malaria parasites. Molecular Biology and Evolution. 2008 Oct;25(10):2233-9.

[137] Pacheco MA, Battistuzzi FU, Junge RE, Cornejo OE, Williams CV, Landau I, et al. Timing the origin of human malarias: the lemur puzzle. BMC Evolutionary Biology. 2011 Oct;11:299. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3228831/.

[138] Hawadak J, Dongang Nana RR, Singh V. Epidemiological, Physiological and Diagnostic Comparison of Plasmodium ovale curtisi and Plasmodium ovale wallikeri. Diagnostics (Basel, Switzerland). 2021 Oct;11(10):1900.

[139] Rutledge GG, Böhme U, Sanders M, Reid AJ, Cotton JA, Maiga-Ascofare O, et al. Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. Nature. 2017 Feb;542(7639):101-4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5326575/.

[140] Daron J, Boissière A, Boundenga L, Ngoubangoye B, Houze S, Arnathau C, et al. Population genomic evidence of Plasmodium vivax Southeast Asian origin. Science Advances. 2021 Apr;7(18):eabc3713.

[141] Loy DE, Plenderleith LJ, Sundararaman SA, Liu W, Gruszczyk J, Chen YJ, et al. Evolutionary history of human Plasmodium vivax revealed by genome-wide analyses of related ape parasites. Proceedings of the National Academy of Sciences of the United

States of America. 2018 Sep;115(36):E8450-9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6130405/.

[142] Dia A, Jett C, McDew-White M, Li X, Anderson TJC, Cheeseman IH. Efficient transcriptome profiling across the malaria parasite erythrocytic cycle by flow sorting. bioRxiv; 2020. Pages: 2020.11.10.377168 Section: New Results. Available from: https://www.biorxiv.org/content/10.1101/2020.11.10.377168v1.

[143] Pickford AK, Michel-Todó L, Dupuy F, Mayor A, Alonso PL, Lavazec C, et al. Expression Patterns of Plasmodium falciparum Clonally Variant Genes at the Onset of a Blood Infection in Malaria-Naive Humans. mBio. 2021 Aug;12(4):e0163621.

[144] Smith RC, Vega-Rodríguez J, Jacobs-Lorena M. The Plasmodium bottleneck: malaria parasite losses in the mosquito vector. Memorias Do Instituto Oswaldo Cruz. 2014 Aug;109(5):644-61.

[145] Molnár P, Marton L, Izrael R, Pálinkás HL, Vértessy BG. Uracil moieties in Plasmodium falciparum genomic DNA. FEBS Open Bio. 2018;8(11):1763-72. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2211-5463.12458. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/2211-5463.12458.

[146] Warhurst DC, Williams JE. ACP Broadsheet no 148. July 1996. Laboratory diagnosis of malaria. Journal of Clinical Pathology. 1996 Jul;49(7):533-8. Publisher: BMJ Publishing Group Section: Research Article. Available from: https://jcp.bmj.com/content/49/7/533.

[147] Moody A. Rapid Diagnostic Tests for Malaria Parasites. Clinical Microbiology Reviews. 2002 Jan;15(1):66-78. Publisher: American Society for Microbiology. Available from: https://journals.asm.org/doi/10.1128/CMR.15.1.66-78.2002.

[148] Rock EP, Marsh K, Saul AJ, Wellems TE, Taylor DW, Maloy WL, et al. Comparative analysis of the Plasmodium falciparum histidine-rich proteins HRP-I, HRP-II and HRP-III in malaria parasites of diverse origin. Parasitology. 1987 Oct;95(2):209-27. Publisher: Cambridge University Press. Available from: https://www.cambridge.org/core/journals/parasitology/article/abs/comparative-analysis-of-the-plasmodium-falciparum-histidinerich-504249B92065D81E73050082D1C0F3C2.

[149] Hopkins H, Kambale W, Kamya MR, Staedke SG, Dorsey G, Rosenthal PJ. Comparison of HRP2- and pLDH-based rapid diagnostic tests for malaria with longitudinal follow-up in Kampala, Uganda. The American Journal of Tropical Medicine and Hygiene. 2007 Jun;76(6):1092-7.

[150] Gatton ML, Dunn J, Chaudhry A, Ciketic S, Cunningham J, Cheng Q. Implications of Parasites Lacking Plasmodium falciparum Histidine-Rich Protein 2 on Malaria Morbidity and Control When Rapid Diagnostic Tests Are Used for Diagnosis. The Journal of Infectious Diseases. 2017 Apr;215(7):1156-66.

[151] Snounou G, Viriyakosol S, Zhu XP, Jarra W, Pinheiro L, do Rosario VE, et al. High sensitivity of detection of human malaria parasites by the use of nested polymerase chain reaction. Molecular and Biochemical Parasitology. 1993 Oct;61(2):315-20.

[152] Hofmann N, Mwingira F, Shekalaghe S, Robinson LJ, Mueller I, Felger I. Ultra-Sensitive Detection of Plasmodium falciparum by Amplification of Multi-Copy Subtelomeric Targets. PLOS Medicine. 2015 Mar;12(3):e1001788. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001788.

[153] Kamau E, Alemayehu S, Feghali KC, Saunders D, Ockenhouse CF. Multiplex qPCR for Detection and Absolute Quantification of Malaria. PLOS ONE. 2013 Aug;8(8):e71539. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0071539.

[154] Ballard E, Wang CYT, Hien TT, Tong NT, Marquart L, Pava Z, et al. A validation study of microscopy versus quantitative PCR for measuring Plasmodium falciparum parasitemia. Tropical Medicine and Health. 2019 Aug;47(1):49. Available from: https://doi.org/10.1186/s41182-019-0176-3.

[155] Mangold KA, Manson RU, Koay ESC, Stephens L, Regner M, Thomson RB, et al. Real-Time PCR for Detection and Identification of Plasmodium spp. Journal of Clinical Microbiology. 2005 May;43(5):2435-40. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1153761/.

[156] Weiss DJ, Lucas TCD, Nguyen M, Nandi AK, Bisanzio D, Battle KE, et al. Mapping the global prevalence, incidence, and mortality of Plasmodium falciparum, 2000-17: a spatial and temporal modelling study. Lancet (London, England). 2019 Jul;394(10195):322-31.

[157] Perera R, Caldera A, Wickremasinghe AR. Reactive Case Detection (RACD) and foci investigation strategies in malaria control and elimination: a review. Malaria Journal. 2020 Nov;19(1):401. Available from: https://doi.org/10.1186/s12936-020-03478-0.

[158] Conn JE, Grillet ME, Correa M, Sallum MAM, Conn JE, Grillet ME, et al. Malaria Transmission in South America—Present Status and Prospects for Elimination. In: Towards Malaria Elimination - A Leap Forward. IntechOpen; 2018. Available from: https://www.intechopen.com/chapters/62219.

[159] Grillet ME, Moreno JE, Hernández-Villena JV, Vincenti-González MF, Noya O, Tami A, et al. Malaria in Southern Venezuela: The hottest hotspot in Latin America. PLoS

Neglected Tropical Diseases. 2021 Jan;15(1):e0008211. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7861532/.

[160] Chu CS, Stolbrink M, Stolady D, Saito M, Beau C, Choun K, et al. Severe Falciparum and Vivax Malaria on the Thailand-Myanmar Border: A Review of 1503 Cases. Clinical Infectious Diseases. 2023 May:ciad262. Available from: https://doi.org/10.1093/cid/ciad262.

[161] Wångdahl A, Wyss K, Saduddin D, Bottai M, Ydring E, Vikerfors T, et al. Severity of Plasmodium falciparum and Non-falciparum Malaria in Travelers and Migrants: A Nationwide Observational Study Over 2 Decades in Sweden. The Journal of Infectious Diseases. 2019 Oct;220(8):1335-45. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6743839/.

[162] Organization WH. Management of Severe Malaria: A Practical Handbook. 3rd ed. Geneva: World Health Organization; 2013.

[163] Idro R, Marsh K, John CC, Newton CR. Cerebral Malaria; Mechanisms Of Brain Injury And Strategies For Improved Neuro-Cognitive Outcome. Pediatric research. 2010 Oct;68(4):267-74. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3056312/.

[164] Trampuz A, Jereb M, Muzlovic I, Prabhu RM. Clinical review: Severe malaria. Critical Care. 2003;7(4):315-23. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC270697/.

[165] Felger I, Maire M, Bretscher MT, Falk N, Tiaden A, Sama W, et al. The Dynamics of Natural Plasmodium falciparum Infections. PLOS ONE. 2012 Sep;7(9):e45542. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0045542.

[166] Ashley EA, White NJ. The duration of Plasmodium falciparum infections. Malaria Journal. 2014 Dec;13(1):500. Available from: https://malariajournal.biomedcentral.com/articles/10.1186/1475-2875-13-500.

[167] Fidock DA, Nomura T, Talley AK, Cooper RA, Dzekunov SM, Ferdig MT, et al. Mutations in the P. falciparum Digestive Vacuole Transmembrane Protein PfCRT and Evidence for Their Role in Chloroquine Resistance. Molecular cell. 2000 Oct;6(4):861-71. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2944663/.

[168] Shrivastava SK, Gupta RK, Mahanta J, Dubey ML. Correlation of molecular markers, Pfmdr1-N86Y and Pfcrt-K76T, with in vitro chloroquine resistant Plasmodium falciparum, isolated in the malaria endemic states of Assam and Arunachal Pradesh, Northeast India. PloS One. 2014;9(8):e103848.

[169] Amambua-Ngwa A, Button-Simons KA, Li X, Kumar S, Brenneman KV, Ferrari M, et al. Chloroquine resistance evolution in Plasmodium falciparum is mediated by the putative amino acid transporter AAT1. Nature Microbiology. 2023 Jul;8(7):1213-26.

[170] Chaturvedi R, Chhibber-Goel J, Verma I, Gopinathan S, Parvez S, Sharma A. Geographical spread and structural basis of sulfadoxine-pyrimethamine drug-resistant malaria parasites. International Journal for Parasitology. 2021 Jun;51(7):505-25.

[171] McCollum AM, Poe AC, Hamel M, Huber C, Zhou Z, Shi YP, et al. Antifolate resistance in Plasmodium falciparum: multiple origins and identification of novel dhfr alleles. The Journal of Infectious Diseases. 2006 Jul;194(2):189-97.

[172] Koenderink JB, Kavishe RA, Rijpma SR, Russel FGM. The ABCs of multidrug resistance in malaria. Trends in Parasitology. 2010 Sep;26(9):440-6.

[173] Gil JP, Krishna S. pfmdr1 (Plasmodium falciparum multidrug drug resistance gene 1): a pivotal factor in malaria resistance to artemisinin combination therapies. Expert Review of Anti-Infective Therapy. 2017 Jun;15(6):527-43.

[174] Balikagala B, Fukuda N, Ikeda M, Katuro OT, Tachibana SI, Yamauchi M, et al. Evidence of Artemisinin-Resistant Malaria in Africa. New England Journal of Medicine. 2021 Sep;385(13):1163-71. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa2101746. Available from: https://doi.org/10.1056/NEJMoa2101746.

[175] Walliker D, Hunt P, Babiker H. Fitness of drug-resistant malaria parasites. Acta Tropica. 2005 Jun;94(3):251-9. Available from: https://www.sciencedirect.com/science/article/pii/S0001706X05000860.

[176] Rosenthal PJ. The interplay between drug resistance and fitness in malaria parasites. Molecular Microbiology. 2013;89(6):1025-38. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mmi.12349. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/mmi.12349.

[177] Frosch AEP, Laufer MK, Mathanga DP, Takala-Harrison S, Skarbinski J, Claassen CW, et al. Return of Widespread Chloroquine-Sensitive Plasmodium falciparum to Malawi. The Journal of Infectious Diseases. 2014 Oct;210(7):1110-4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6281358/.

[178] Ndam NT, Basco LK, Ngane VF, Ayouba A, Ngolle EM, Deloron P, et al. Reemergence of chloroquine-sensitive pfcrt K76 Plasmodium falciparum genotype in southeastern Cameroon. Malaria Journal. 2017 Mar;16(1):130. Available from: https://doi.org/10.1186/s12936-017-1783-2.

[179] Asare KK, Africa J, Mbata J, Opoku YK. The emergence of chloroquine-sensitive Plasmodium falciparum is influenced by selected communities in some parts of the

Central Region of Ghana. Malaria Journal. 2021 Nov;20(1):447. Available from: https://doi.org/10.1186/s12936-021-03985-8.

[180] Oduola AMJ, Milhous WK, Weatherly NF, Bowdre JH, Desjardins RE. Plasmodium falciparum: Induction of resistance to mefloquine in cloned strains by continuous drug exposure in vitro. Experimental Parasitology. 1988 Dec;67(2):354-60. Available from: https://www.sciencedirect.com/science/article/pii/0014489488900823.

[181] van Schalkwyk DA, Burrow R, Henriques G, Gadalla NB, Beshir KB, Hasford C, et al. Culture-adapted Plasmodium falciparum isolates from UK travellers: in vitro drug sensitivity, clonality and drug resistance markers. Malaria Journal. 2013 Sep;12(1):320. Available from: https://doi.org/10.1186/1475-2875-12-320.

[182] Rathod PK, McErlean T, Lee PC. Variations in frequencies of drug resistance in Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America. 1997 Aug;94(17):9389-93. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC23200/.

[183] Calçada C, Silva M, Baptista V, Thathy V, Silva-Pedrosa R, Granja D, et al. Expansion of a Specific Plasmodium falciparum PfMDR1 Haplotype in Southeast Asia with Increased Substrate Transport. mBio. 2020 Dec;11(6):10.1128/mbio.02093-20. Publisher: American Society for Microbiology. Available from: https://journals.asm.org/doi/10.1128/mbio.02093-20.

[184] Sidhu ABS, Verdier-Pinard D, Fidock DA. Chloroquine Resistance in Plasmodium falciparum Malaria Parasites Conferred by pfcrt Mutations. Science (New York, NY). 2002 Oct;298(5591):210-3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2954758/.

[185] Wang P, Read M, Sims PF, Hyde JE. Sulfadoxine resistance in the human malaria parasite Plasmodium falciparum is determined by mutations in dihydropteroate synthetase and an additional factor associated with folate utilization. Molecular Microbiology. 1997 Mar;23(5):979-86.

[186] Kong A, Wilson SA, Ah Y, Nace D, Rogier E, Aidoo M. HRP2 and HRP3 cross-reactivity and implications for HRP2-based RDT use in regions with Plasmodium falciparum hrp2 gene deletions. Malaria Journal. 2021 Apr;20(1):207. Available from: https://doi.org/10.1186/s12936-021-03739-6.

[187] Bopp SER, Manary MJ, Bright AT, Johnston GL, Dharia NV, Luna FL, et al. Mitotic Evolution of Plasmodium falciparum Shows a Stable Core Genome but Recombination in Antigen Families. PLoS Genetics. 2013 Feb;9(2):e1003293. Available from: http://dx.plos.org/10.1371/journal.pgen.1003293.

[188] Hamilton WL, Claessens A, Otto TD, Kekre M, Fairhurst RM, Rayner JC, et al. Extreme mutation bias and high AT content in Plasmodium falciparum. Nucleic Acids Research. 2017 Feb;45(4):1889-901. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5389722/.

[189] Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research. 1999 Jan;27(2):573-80. Available from: https://doi.org/10.1093/nar/27.2.573.

[190] Huang HY, Liang XY, Lin LY, Chen JT, Ehapo CS, Eyi UM, et al. Genetic polymorphism of Plasmodium falciparum circumsporozoite protein on Bioko Island, Equatorial Guinea and global comparative analysis. Malaria Journal. 2020 Jul;19(1):245. Available from: https://doi.org/10.1186/s12936-020-03315-4.

[191] Takala SL, Coulibaly D, Thera MA, Batchelor AH, Cummings MP, Escalante AA, et al. Extreme Polymorphism in a Vaccine Antigen and Risk of Clinical Malaria: Implications for Vaccine Development. Science translational medicine. 2009 Oct;1(2):2ra5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2822345/.

[192] Eisen D, Billman-Jacobe H, Marshall VF, Fryauff D, Coppel RL. Temporal Variation of the Merozoite Surface Protein-2 Gene of Plasmodium falciparum. Infection and Immunity. 1998 Jan;66(1):239-46. Publisher: American Society for Microbiology. Available from: https://journals.asm.org/doi/10.1128/IAI.66.1.239-246.1998.

[193] Kraemer SM, Kyes SA, Aggarwal G, Springer AL, Nelson SO, Christodoulou Z, et al. Patterns of gene recombination shape var gene repertoires in Plasmodium falciparum: comparisons of geographically diverse isolates. BMC Genomics. 2007 Feb;8:45. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1805758/.

[194] Kyes SA, Christodoulou Z, Raza A, Horrocks P, Pinches R, Rowe JA, et al. A well-conserved Plasmodium falciparum var gene shows an unusual stage-specific transcript pattern. Molecular microbiology. 2003 Jun;48(5):1339-48. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2869446/.

[195] Ndam NGT, Salanti A, Bertin G, Dahlbäck M, Fievet N, Turner L, et al. High Level of var2csa Transcription by Plasmodium falciparum Isolated from the Placenta. The Journal of Infectious Diseases. 2005;192(2):331-5. Publisher: Oxford University Press. Available from: https://www.jstor.org/stable/30086214.

[196] Sander AF, Salanti A, Lavstsen T, Nielsen MA, Magistrado P, Lusingu J, et al. Multiple var2csa-Type PfEMP1 Genes Located at Different Chromosomal Loci Occur in Many Plasmodium falciparum Isolates. PLoS ONE. 2009 Aug;4(8):e6667. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723927/.

[197] Benavente ED, Oresegun DR, de Sessions PF, Walker EM, Roper C, Dombrowski JG, et al. Global genetic diversity of var2csa in Plasmodium falciparum with implications for malaria in pregnancy and vaccine development. Scientific Reports. 2018 Oct;8:15429. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6193930/.

[198] Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Research. 2010 Aug;38(15):e159. Available from: https://doi.org/10.1093/nar/gkq543.

[199] Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology. 2019 Oct;37(10):1155-62. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: DNA sequencing;Genome informatics;Genomics;Machine learning;Next-generation sequencing Subject_term_id: dna-sequencing;genome-informatics;genomics;machine-learning;next-generation-sequencing. Available from: https://www.nature.com/articles/s41587-019-0217-9.

[200] Hon T, Mars K, Young G, Tsai YC, Karalius JW, Landolin JM, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. Scientific Data. 2020 Nov;7(1):399. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41597-020-00743-4.

[201] Kucuk E, van der Sanden BPGH, O'Gorman L, Kwint M, Derks R, Wenger AM, et al. Comprehensive de novo mutation discovery with HiFi long-read sequencing. Genome Medicine. 2023 May;15(1):34. Available from: https://doi.org/10.1186/s13073-023-01183-6.

[202] Bull PC, Kyes S, Buckee CO, Montgomery J, Kortok MM, Newbold CI, et al. An approach to classifying sequence tags sampled from Plasmodium falciparum var genes. Molecular and Biochemical Parasitology. 2007 Jul;154(1):98-102.

[203] Githinji G, Bull PC. A re-assessment of gene-tag classification approaches for describing var gene expression patterns during human Plasmodium falciparum malaria parasite infections. Wellcome Open Research. 2017;2:86.

[204] Camponovo F, Buckee CO, Taylor AR. Measurably recombining malaria parasites. Trends in Parasitology. 2023 Jan;39(1):17-25. Publisher: Elsevier. Available from: https://www.cell.com/trends/parasitology/abstract/S1471-4922(22)00262-8.

[205] Enosse S, Dobaño C, Quelhas D, Aponte JJ, Lievens M, Leach A, et al. RTS,S/AS02A malaria vaccine does not induce parasite CSP T cell epitope selection and reduces multiplicity of infection. PLoS clinical trials. 2006 May;1(1):e5.

[206] Escalante AA, Grebert HM, Chaiyaroj SC, Magris M, Biswas S, Nahlen BL, et al. Polymorphism in the gene encoding the apical membrane antigen-1 (AMA-1) of Plasmodium falciparum. X. Asembo Bay Cohort Project. Molecular and Biochemical Parasitology. 2001 Apr;113(2):279-87. Available from: https://www.sciencedirect.com/science/article/pii/S0166685101002298.

[207] Auburn S, Campino S, Miotto O, Djimde AA, Zongo I, Manske M, et al. Characterization of Within-Host Plasmodium falciparum Diversity Using Next-Generation Sequence Data. PLoS ONE. 2012 Feb;7(2):e32891. Available from: https://dx.plos.org/10.1371/journal.pone.0032891.

[208] Voight BF, Kudaravalli S, Wen X, Pritchard JK. A Map of Recent Positive Selection in the Human Genome. PLOS Biology. 2006 Mar;4(3):e72. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040072.

[209] Niang M, Bei AK, Madnani KG, Pelly S, Dankwa S, Kanjee U, et al. STEVOR Is a Plasmodium falciparum Erythrocyte Binding Protein that Mediates Merozoite Invasion and Rosetting. Cell Host & Microbe. 2014 Jul;16(1):81-93. Publisher: Elsevier. Available from: https://www.cell.com/cell-host-microbe/abstract/S1931-3128(14)00218-2.

[210] Lee WC, Russell B, Rénia L. Evolving perspectives on rosetting in malaria. Trends in Parasitology. 2022 Oct;38(10):882-9. Publisher: Elsevier. Available from: https://www.cell.com/trends/parasitology/abstract/S1471-4922(22)00178-7.

[211] Gowda DC, Wu X. Parasite Recognition and Signaling Mechanisms in Innate Immune Responses to Malaria. Frontiers in Immunology. 2018;9. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2018.03006.

[212] Lavstsen T, Turner L, Saguti F, Magistrado P, Rask TS, Jespersen JS, et al. Plasmodium falciparum erythrocyte membrane protein 1 domain cassettes 8 and 13 are associated with severe malaria in children. Proceedings of the National Academy of Sciences of the United States of America. 2012 Jun;109(26):E1791-800.

[213] Avril M, Tripathi AK, Brazier AJ, Andisi C, Janes JH, Soma VL, et al. A restricted subset of var genes mediates adherence of Plasmodium falciparum-infected erythrocytes to brain endothelial cells. Proceedings of the National Academy of Sciences of the United States of America. 2012 Jun;109(26):E1782-90.

[214] Claessens A, Adams Y, Ghumra A, Lindergard G, Buchan CC, Andisi C, et al. A subset of group A-like var genes encodes the malaria parasite ligands for binding to human brain endothelial cells. Proceedings of the National Academy of Sciences of the United States of America. 2012 Jun;109(26):E1772-81.

[215] Cabrera A, Neculai D, Kain KC. CD36 and malaria: friends or foes? A decade of data provides some answers. Trends in Parasitology. 2014 Sep;30(9):436-44.

[216] Kaestli M, Cockburn IA, Cortés A, Baea K, Rowe JA, Beck HP. Virulence of Malaria Is Associated with Differential Expression of Plasmodium falciparum var Gene Subgroups in a Case-Control Study. The Journal of infectious diseases. 2006 Jun;193(11):1567-74. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2877257/.

[217] Buffet PA, Gamain B, Scheidig C, Baruch D, Smith JD, Hernandez-Rivas R, et al. Plasmodium falciparum domain mediating adhesion to chondroitin sulfate A: A receptor for human placental infection. Proceedings of the National Academy of Sciences of the United States of America. 1999 Oct;96(22):12743-8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC23079/.

[218] Srivastava A, Gangnard S, Dechavanne S, Amirat F, Bentley AL, Bentley GA, et al. Var2CSA Minimal CSA Binding Region Is Located within the N-Terminal Region. PLOS ONE. 2011 May;6(5):e20270. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0020270.

[219] Recker M, Buckee CO, Serazin A, Kyes S, Pinches R, Christodoulou Z, et al. Antigenic Variation in Plasmodium falciparum Malaria Involves a Highly Structured Switching Pattern. PLoS Pathogens. 2011 Mar;7(3):e1001306. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3048365/.

[220] Zhang X, Florini F, Visone JE, Lionardi I, Gross MR, Patel V, et al. A coordinated transcriptional switching network mediates antigenic variation of human malaria parasites. eLife. 2022 Dec;11:e83840. Publisher: eLife Sciences Publications, Ltd. Available from: https://doi.org/10.7554/eLife.83840.

[221] Stanisic DI, McCarthy JS, Good MF. Controlled Human Malaria Infection: Applications, Advances, and Challenges. Infection and Immunity. 2018 Jan;86(1):e00479-17.

[222] Bachmann A, Bruske E, Krumkamp R, Turner L, Wichers JS, Petter M, et al. Controlled human malaria infection with Plasmodium falciparum demonstrates impact of naturally acquired immunity on virulence gene expression. PLOS Pathogens. 2019 Jul;15(7):e1007906. Publisher: Public Library of Science. Available from: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1007906.

[223] Small PA. Smallpox: The Death of a Disease and ?House on Fire: The Fight to Eradicate Smallpox. Emerging Infectious Diseases. 2011 Nov;17(11):2085-6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3310594/.

[224] Bousema T, Griffin JT, Sauerwein RW, Smith DL, Churcher TS, Takken W, et al. Hitting Hotspots: Spatial Targeting of Malaria for Control and Elimination. PLoS Medicine. 2012 Jan;9(1):e1001165. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3269430/.

[225] Shannon KL, Shields T, Ahmed S, Rahman H, Prue CS, Khyang J, et al. Temporal and Spatial Differences between Symptomatic and Asymptomatic Malaria Infections in the Chittagong Hill Districts, Bangladesh. The American Journal of Tropical Medicine and Hygiene. 2022 Dec;107(6):1210-7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9768271/.

[226] Dieng S, Ba EH, Cissé B, Sallah K, Guindo A, Ouedraogo B, et al. Spatio-temporal variation of malaria hotspots in Central Senegal, 2008–2012. BMC Infectious Diseases. 2020 Jun;20(1):424. Available from: https://doi.org/10.1186/s12879-020-05145-w.

[227] Stresman GH, Mwesigwa J, Achan J, Giorgi E, Worwui A, Jawara M, et al. Do hotspots fuel malaria transmission: a village-scale spatio-temporal analysis of a 2-year cohort study in The Gambia. BMC Medicine. 2018 Dec;16(1):160. Available from: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-018-1141-4.

[228] Stresman G, Bousema T, Cook J. Malaria Hotspots: Is There Epidemiological Evidence for Fine-Scale Spatial Targeting of Interventions? Trends in Parasitology. 2019 Oct;35(10):822-34.

[229] Mordecai EA, Paaijmans KP, Johnson LR, Balzer C, Ben-Horin T, de Moor E, et al. Optimal temperature for malaria transmission is dramatically lower than previously predicted. Ecology Letters. 2013;16(1):22-30. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.12015. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.12015.

[230] Beck-Johnson LM, Nelson WA, Paaijmans KP, Read AF, Thomas MB, Bjørnstad ON. The Effect of Temperature on Anopheles Mosquito Population Dynamics and the Potential for Malaria Transmission. PLOS ONE. 2013 Nov;8(11):e79276. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0079276.

[231] Shapiro LLM, Whitehead SA, Thomas MB. Quantifying the effects of temperature on mosquito and parasite traits that determine the transmission potential of human malaria. PLoS biology. 2017 Oct;15(10):e2003489.

[232] Reiner RC, Geary M, Atkinson PM, Smith DL, Gething PW. Seasonality of Plasmodium falciparum transmission: a systematic review. Malaria Journal. 2015 Sep;14:343. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4570512/.

[233] Ryan SJ, Lippi CA, Zermoglio F. Shifting transmission risk for malaria in Africa with climate change: a framework for planning and intervention. Malaria Journal. 2020 May;19(1):170. Available from: https://doi.org/10.1186/s12936-020-03224-6.

[234] Wang Z, Liu Y, Li Y, Wang G, Lourenço J, Kraemer M, et al. The relationship between rising temperatures and malaria incidence in Hainan, China, from 1984 to 2010: a longitudinal cohort study. The Lancet Planetary Health. 2022 Apr;6(4):e350-8. Publisher: Elsevier. Available from: https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(22)00039-0/fulltext.

[235] Diouf I, Adeola AM, Abiodun GJ, Lennard C, Shirinde JM, Yaka P, et al. Impact of future climate change on malaria in West Africa. Theoretical and Applied Climatology. 2022 Feb;147(3):853-65. Available from: https://doi.org/10.1007/s00704-021-03807-6.

[236] Guery MA, Claessens A. Order within chaos: Harnessing Plasmodium falciparum var gene extreme polymorphism for malaria epidemiology. PLOS Genetics. 2021 Feb;17(2):e1009344. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009344.

[237] WHO. World malaria report 2021; 2021. Available from: https://www.who.int/publications-detail-redirect/9789240040496.

[238] Ahmad A, Mohammed NI, Joof F, Affara M, Jawara M, Abubakar I, et al. Asymptomatic Plasmodium falciparum carriage and clinical disease: a 5-year community-based longitudinal study in The Gambia. Malaria Journal. 2023 Mar;22(1):82. Available from: https://doi.org/10.1186/s12936-023-04519-0.

[239] Nkhoma SC, Nair S, Cheeseman IH, Rohr-Allegrini C, Singlam S, Nosten F, et al. Close kinship within multiple-genotype malaria parasite infections. Proceedings of the Royal Society B: Biological Sciences. 2012 Jul;279(1738):2589-98. Available from: https://royalsocietypublishing.org/doi/10.1098/rspb.2012.0113.

[240] Daniels R, Chang HH, Séne PD, Park DC, Neafsey DE, Schaffner SF, et al. Genetic Surveillance Detects Both Clonal and Epidemic Transmission of Malaria following Enhanced Intervention in Senegal. PLoS ONE. 2013 Apr;8(4):e60780. Available from: https://dx.plos.org/10.1371/journal.pone.0060780.

[241] Pacheco MA, Forero-Peña DA, Schneider KA, Chavero M, Gamardo A, Figuera L, et al. Malaria in Venezuela: changes in the complexity of infection reflects the increment in transmission intensity. Malaria Journal. 2020 May;19(1):176.

[242] Taylor AR, Schaffner SF, Cerqueira GC, Nkhoma SC, Anderson TJC, Sriprawat K, et al. Quantifying connectivity between local Plasmodium falciparum malaria parasite

populations using identity by descent. PLOS Genetics. 2017 Oct;13(10):e1007065. Available from: http://dx.plos.org/10.1371/journal.pgen.1007065.

[243] Noviyanti R, Miotto O, Barry A, Marfurt J, Siegel S, Thuy-Nhien N, et al. Implementing parasite genotyping into national surveillance frameworks: feedback from control programmes and researchers in the Asia–Pacific region. Malaria Journal. 2020 Jul;19(1):271. Available from: https://doi.org/10.1186/s12936-020-03330-5.

[244] Lindblade KA, Steinhardt L, Samuels A, Kachur SP, Slutsker L. The silent threat: asymptomatic parasitemia and malaria transmission. Expert Review of Anti-infective Therapy. 2013 Jun;11(6):623-39. Available from: http://www.tandfonline.com/doi/full/10.1586/eri.13.45.

[245] Stone W, Gonçalves BP, Bousema T, Drakeley C. Assessing the infectious reservoir of falciparum malaria: past and future. Trends in Parasitology. 2015 Jul;31(7):287-96.

[246] Amambua-Ngwa A, Jeffries D, Amato R, Worwui A, Karim M, Ceesay S, et al. Consistent signatures of selection from genomic analysis of pairs of temporal and spatial Plasmodium falciparum populations from The Gambia. Scientific Reports. 2018 Dec;8(1):9687. Available from: http://www.nature.com/articles/s41598-018-28017-5.

[247] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003 Nov;13(11):2498-504. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC403769/.

[248] Nwakanma DC, Duffy CW, Amambua-Ngwa A, Oriero EC, Bojang KA, Pinder M, et al. Changes in Malaria Parasite Drug Resistance in an Endemic Population Over a 25-Year Period With Resulting Genomic Evidence of Selection. The Journal of Infectious Diseases. 2014 Apr;209(7):1126-35. Available from: https://academic.oup.com/jid/article-lookup/doi/10.1093/infdis/jit618.

[249] Ord R, Alexander N, Dunyo S, Hallett R, Jawara M, Targett G, et al. Seasonal carriage of pfcrt and pfmdr1 alleles in Gambian Plasmodium falciparum imply reduced fitness of chloroquine-resistant parasites. The Journal of Infectious Diseases. 2007 Dec;196(11):1613-9.

[250] Sy M, Deme AB, Warren JL, Early A, Schaffner S, Daniels RF, et al. Plasmodium falciparum genomic surveillance reveals spatial and temporal trends, association of genetic and physical distance, and household clustering. Scientific Reports. 2022 Jan;12(1):938. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41598-021-04572-2.

[251] Redmond SN, MacInnis BM, Bopp S, Bei AK, Ndiaye D, Hartl DL, et al. De Novo Mutations Resolve Disease Transmission Pathways in Clonal Malaria. Molecular Biology and Evolution. 2018 Jul;35(7):1678-89.

[252] Echeverry DF, Nair S, Osorio L, Menon S, Murillo C, Anderson TJ. Long term persistence of clonal malaria parasite Plasmodium falciparum lineages in the Colombian Pacific region. BMC Genetics. 2013;14(1):2. Available from: http://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-14-2.

[253] Okell LC, Bousema T, Griffin JT, Ouédraogo AL, Ghani AC, Drakeley CJ. Factors determining the occurrence of submicroscopic malaria infections and their relevance for control. Nature Communications. 2012;3:1237.

[254] Slater HC, Ross A, Felger I, Hofmann NE, Robinson L, Cook J, et al. The temporal dynamics and infectiousness of subpatent Plasmodium falciparum infections in relation to parasite density. Nature Communications. 2019 Mar;10(1):1433. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-019-09441-1.

[255] Bylicka-Szczepanowska E, Korzeniewski K. Asymptomatic Malaria Infections in the Time of COVID-19 Pandemic: Experience from the Central African Republic. International Journal of Environmental Research and Public Health. 2022 Mar;19(6):3544. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8951439/.

[256] Zhao Y, Zhao Y, Lv Y, Liu F, Wang Q, Li P, et al. Comparison of methods for detecting asymptomatic malaria infections in the China–Myanmar border area. Malaria Journal. 2017 Apr;16(1):159. Available from: https://doi.org/10.1186/s12936-017-1813-0.

[257] Chan JA, Howell KB, Reiling L, Ataide R, Mackintosh CL, Fowkes FJI, et al. Targets of antibodies against Plasmodium falciparum–infected erythrocytes in malaria immunity. The Journal of Clinical Investigation. 2012 Sep;122(9):3227-38. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3428085/.

[258] Recker M, Nee S, Bull PC, Kinyanjui S, Marsh K, Newbold C, et al. Transient cross-reactive immune responses can orchestrate antigenic variation in malaria. Nature. 2004 Jun;429(6991):555-8. Number: 6991 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/nature02486.

[259] Childs LM, Buckee CO. Dissecting the determinants of malaria chronicity: why within-host models struggle to reproduce infection dynamics. Journal of The Royal Society Interface. 2015 Mar;12(104):20141379. Publisher: Royal Society. Available from: https://royalsocietypublishing.org/doi/10.1098/rsif.2014.1379.

[260] Nyarko PB, Claessens A. Understanding Host-Pathogen-Vector Interactions with Chronic Asymptomatic Malaria Infections. Trends in Parasitology. 2021 Mar;37(3):195-204.

[261] Claessens A, Affara M, Assefa SA, Kwiatkowski DP, Conway DJ. Culture adaptation of malaria parasites selects for convergent loss-of-function mutants. Scientific Reports. 2017 Jan;7(1):41303. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/srep41303.

[262] Claessens A, Stewart LB, Drury E, Ahouidi AD, Amambua-Ngwa A, Diakite M, et al. Genomic variation during culture adaptation of genetically complex Plasmodium falciparum clinical isolates. Microbial Genomics. 2023;9(5):001009. Publisher: Microbiology Society,. Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.001009.

[263] Meerstein-Kessel L, Andolina C, Carrio E, Mahamar A, Sawa P, Diawara H, et al. A multiplex assay for the sensitive detection and quantification of male and female Plasmodium falciparum gametocytes. Malaria Journal. 2018 Nov;17(1):441. Available from: https://doi.org/10.1186/s12936-018-2584-y.

[264] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul;25(14):1754-60. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp324.

[265] Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience. 2019 Sep;8(9):giz100. Available from: https://doi.org/10.1093/gigascience/giz100.

[266] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec;28(23):3150-2. Available from: https://doi.org/10.1093/bioinformatics/bts565.

[267] Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011 Feb;27(4):578-9. Available from: https://doi.org/10.1093/bioinformatics/btq683.

[268] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug;30(15):2114-20. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/.

[269] Institute B. Picard toolkit. Broad Institute; 2019. Publication Title: Broad Institute, GitHub repository. Available from: https://broadinstitute.github.io/picard/.

[270] Auburn S, Campino S, Clark TG, Djimde AA, Zongo I, Pinches R, et al. An effective method to purify Plasmodium falciparum DNA directly from clinical blood samples for whole genome high-throughput sequencing. PloS One. 2011;6(7):e22213.

[271] Shah Z, Adams M, Moser KA, Shrestha B, Stucke EM, Laufer MK, et al. Optimization of parasite DNA enrichment approaches to generate whole genome sequencing data for Plasmodium falciparum from low parasitaemia samples. Malaria Journal. 2020 Mar;19(1):135. Available from: https://doi.org/10.1186/s12936-020-03195-8.

[272] Ngansop F, Li H, Zolkiewska A, Zolkiewski M. Biochemical characterization of the apicoplast-targeted AAA+ ATPase ClpB from Plasmodium falciparum. Biochemical and Biophysical Research Communications. 2013 Sep;439(2):191-5. Available from: https://www.sciencedirect.com/science/article/pii/S0006291X13014113.

[273] AhYoung AP, Koehl A, Cascio D, Egea PF. Structural mapping of the ClpB ATPases of Plasmodium falciparum: Targeting protein folding and secretion for antimalarial drug design. Protein Science : A Publication of the Protein Society. 2015 Sep;24(9):1508-20. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4570544/.

[274] Gupta A, Shah P, Haider A, Gupta K, Siddiqi MI, Ralph SA, et al. Reduced ribosomes of the apicoplast and mitochondrion of Plasmodium spp. and predicted interactions with antibiotics. Open Biology. 2014 May;4(5):140045.

[275] Lu F, Jiang H, Ding J, Mu J, Valenzuela JG, Ribeiro JM, et al. cDNA sequences reveal considerable gene prediction inaccuracy in the Plasmodium falciparum genome. BMC Genomics. 2007 Jul;8:255. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1978503/.

[276] Garg S, Agarwal S, Kumar S, Shams Yazdani S, Chitnis CE, Singh S. Calcium-dependent permeabilization of erythrocytes by a perforin-like protein during egress of malaria parasites. Nature Communications. 2013 Apr;4(1):1736. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/ncomms2725.

[277] Fidock DA, Nomura T, Cooper RA, Su X, Talley AK, Wellems TE. Allelic modifications of the cg2 and cg1 genes do not alter the chloroquine response of drug-resistant Plasmodium falciparum. Molecular and Biochemical Parasitology. 2000 Sep;110(1):1-10.

[278] Mobegi VA, Duffy CW, Amambua-Ngwa A, Loua KM, Laman E, Nwakanma DC, et al. Genome-Wide Analysis of Selection on the Malaria Parasite Plasmodium falciparum in West African Populations of Differing Infection Endemicity. Molecular Biology and Evolution. 2014 Jun;31(6):1490-9. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msu106.

[279] Feng X, Cheng H, Portik D, Li H. Metagenome assembly of high-fidelity long reads with hifiasm-meta. Nature methods. 2022 Jun;19(6):671-4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9343089/.

[280] Kim CY, Ma J, Lee I. HiFi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. Nature Communications. 2022 Oct;13(1):6367. Number: 1 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41467-022-34149-0.

[281] Hayes CN, Diez D, Joannin N, Honda W, Kanehisa M, Wahlgren M, et al. varDB: a pathogen-specific sequence database of protein families involved in antigenic variation. Bioinformatics (Oxford, England). 2008 Nov;24(21):2564-5.

[282] Hernandez-Rivas R, Pérez-Toledo K, Herrera Solorio AM, Delgadillo DM, Vargas M. Telomeric Heterochromatin in Plasmodium falciparum. Journal of Biomedicine and Biotechnology. 2010;2010:290501. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2821646/.

[283] Zhu SJ, Almagro-Garcia J, McVean G. Deconvolution of multiple infections in Plasmodium falciparum from high throughput sequencing data. Bioinformatics. 2018 Jan;34(1):9-15. Available from: https://academic.oup.com/bioinformatics/article/34/1/9/4091117.

[284] Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013 Nov;29(21):2669-77. Available from: https://doi.org/10.1093/bioinformatics/btt476.

[285] Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions; 2022. Available from: https://CRAN.R-project.org/package=cluster.

[286] Yalcindag E, Elguero E, Arnathau C, Durand P, Akiana J, Anderson TJ, et al. Multiple independent introductions of Plasmodium falciparum in South America. Proceedings of the National Academy of Sciences. 2012 Jan;109(2):511-6. Publisher: National Academy of Sciences Section: Biological Sciences. Available from: https://www.pnas.org/content/109/2/511.