Population Structure and Genetic Variation of *Plasmodium falciparum* in West and Central Africa

TUM, Section of population genetics, Prof. Aurelién Tellier
Research internship final report, WS 2023

Philip L. Wolper

December 12, 2023

Abstract

This report presents insights on the genetic diversity and population structure of the malaria-causing parasite *Plasmodium falciparum* using whole-genome SNP data from various African countries. We confirm previous knowledge of high admixture and limited population substructure between African countries. Our *de novo* clustering analysis generally confirms this. Interestingly, we find chromosome 11, to exhibit higher signatures of isolation-by-distance than chromosome 2. Additionally, we calculate the unfolded site-frequency spectrum (SFS) for both chromosomes and show it to be U-shaped. We provide some discussion on evolutionary mechanisms leading this U-shaped SFS in *P. falciparum* and emphasize how population genetics, particularly multiple-merger coalescents, can be applied to understand the evolutionary dynamics of *P. falciparum* in the future, contributing to better malaria control and drug intervention strategies.

Introduction

Plasmodium falciparum, an apicomplexan human parasite and the most important disease agent of malaria, is a major global health threat, with nearly half the worlds population at risk for the disease. Especially tropical and subtropical regions suffer the highest disease burden, and nearly 95% of all cases and deaths occur in Africa. Fatalities in this region mainly affect young children, as 80% of malaria mortality occurs in children aged 5 years or younger (WHO 2022). Although impoved control measures have contributed to a steady decline of malaria cases and death rates over the past 20 years, increasing resistances to major anti-malaria drugs, such as artemisin and chloroquinine, are constantly threatening to undermine eradication progress of the disease. Continued effective malaria control measures, rely on disentangling the effects of population structure, selection, recombination and random processes on genome evolution and population dynamics.

Understanding the genetic diversiy and population structure of *P. falciparum* is crucial to make valid inferences of selection on parasite traits, such as drug resistance. This allows us to seperate adaptive changes in parasite allele frequencies from non-adaptive changes. Importantly, conclusions about selective processes require comprehensive evolutionary 'null'-models, to test hypothesis. Such models should account for population structure, demographic history, as well as other biological or life-history traits, that could potentially lead to signatures of genetic variation similar to selection. Therefore it is important to assess the structure of a sampled parasite population when making further inferences about adaption.

The population genetic structure of *P. falciparum* has been shown to vary greatly when looking at different global regions, and is closely associated with transmission intensity and disease epidemiology (Anderson et al. 2000). Regions with low transmission, such as South America or South East Asia, typically show low levels of multiple co-infection, leading to predominantly clonal populations, with low genetic diversity and high levels of linkage disequilibrium (LD). This

LD often results from epidemic expansion of a particular clone. On the other hand, high levels of parasite transmission, such as in malaria endemic regions across Africa, leads to high levels of co-infection in the host, high rates of outcrossing and rapid decay in LD (Mu et al. 2005). This variation of population genetic structure makes generalized studies about evolutionary patterns in *P. falciparum* difficult to conduct and interpret reliably.

Most population genetics inferences are based on the Wright-Fisher model and Kingman coalescent framework to connect the population genetic theory to real-world population genetics data (eg. Li and Wiehe 2013). In order to assess the genetic variation in a population, summary statistics can be calculated. One such commonly used statistics is the site frequency spectrum (SFS), a histogram of allele frequencies in the population. Neutral theory predicts most allelic variants expected to be rare, because they do not influence an oranisms fitness and will be lost by genetic drift. For a panmictic population, the expected SFS for can be calculated, using coalescent model simulations following the assumptions of Wright-Fisher populations. Typically, the observed SFS spectrum differs from the expected ones, and the type of this skew can reveal details about selective or demographic processes occuring in the population. For example, an excess in rare variants (such as singletons), can be interpreted as signals of positive or negative selection or population expansion.

In nature many SFS are observed to be U-shaped (uSFS) (Freund et al. 2023), with an excess of rare and common derived alleles. One common explaination for the uSFS is positive selection at several loci with neutral variants hitchhiking to high-frequencies as well (Bustamante et al. 2001). This phenoma is predicted to be reinforced by temporally fluctuating selection (Huerta-Sanchez, Durrett, and Bustamante 2008). Many other alternative explaination for a uSFS have been proposed, such as gene flow from non-sampled populations (Marchi and Excoffier 2020), population subdivision (Lapierre et al. 2016), biased gene conversion (Pouyet et al. 2018), misorientation of ancestral and derived alleles (Baudry and Depaulis 2003) and well as various reproductive strategies (Tellier and Lemaire 2014).

This great number of factor affecting SFS distributions can make it difficult to effectively identify which forces are dominantly acting on the evolution of an organism. While the Kingman coalescent models can deal with many different assumption such as population structure or changes in population size, it is less suited to deal with large variances in offspring numbers (sweepstake reproduction) and merging of more than two lineages. Sweepstake reproduction applies to species, which produce a high number of offspring together with high early-life mortality, leading to a very recent common ancestor. Similarly, strong and recurrent selective sweeps can lead to the coalescence of multiple branches in the genealogies simultaneously. Because of the limitations the Kingman coalescent is faced with regarding the such life-history traits, species violating these assumptions are better modelled with a more class of coalescent models using multiple merger coalescents (MMC) (Tellier and Lemaire 2014). Genealogies originating from MMC models, have been observed across all kingdoms of life (Freund et al. 2023). Inferences methods based on the classical Wright-Fisher assumptions can lead to erroneous interpretations of different evolutionary mechanisms, if applied to species with strong multiple merger phenomena. Thus for organism with life-cycles that produce MMC genealogies, such as many pathogens or marine organisms, multiple-merger coalescent models have been recommended for inferring the expected SFS (Freund et al. 2023).

In this report we analyze population structure and genetic variation of *Plasmodium falciparum* in select African countries. We calculate the unfolded site-frequency spectrum of the population and provide some context to interpret it in.

Data and Methods

In this section, we provide a comprehensive overview of the data source, data preprocessing steps, and the analytical methods employed for this investigation.

Pf7 data availability

The analysis of population genetics data for *Plasmodium falciparum* presented in this report was derived from whole-genome Single Nucleotide Polymorphism (SNP) data. The dataset used in this study was obtained from the MalariaGEN Pf7 data resource, a large release of genome variation data from *Plasmodium falciparum* (MalariaGEN et al. 2023). In whole, Pf7 consists of over 16,000 samples high-quality from over 33 countries and 82 partner studies. By using a method called selective whole genome amplification (sWGA) prior to sequencing, samples could be collected from dried blood spots. While showing the use of sWGA does not indroduce any biases in coverage or population structure, the Pf7 resource is the largest publicly available resource for genomic information of *P.falciparum*. Pf7 provides data on genomic variation as variant calling format (VCF) files per chromosome. The vcf files are freely available from an FTP server using an FTP client (`ftp://ngs.sanger.ac.uk/production/malaria/Resource/34/Pf7_vcf/`).

Data

For this report, we chose to analyze 1846 sample from 4 different african countries: Democratic Republik of Congo (DRC), The Gambia, Kenya and Tanzania. The number of samples per country is 520, 452, 285 and 598, respectively and all of them come from 2010-2017, never more than 4 years for any given country. Because of the size of the data, our analysis was conducted for chromosomes 2 and 11, which we downloaded from the FTP server.

Filtering and processing

Filtering steps and manipulation of the vcf files were done using the software *bcftools* (Danecek et al. 2021). The VCF files for chromosomes 2 and 11 were filtered to include only biallelic SNPs, which where designated with the filter 'PASS' by the Pf7 data resource. Additionally, we retained only the SNPs with a quality score VQSLOD > 5.0. In total this yielded data sets with 16,269 and 44,040 SNPs, for chromosomes 2 and 11, respectively. The proportion of heterozygous genotypes in the VCF files was relatively small: 0.2% and 0.3% for chromosome 2 and 11, respectively. To avoid possible effects of multiple infection, we removed all heterozygous genotypes (including both phased and unphased GT fields), by setting them to missing data (./.).

Population genetics statisics

Basic population genetic measures, such as θ or Tajima' D statistic, were calculated using the R package *PopGenome* (Pfeifer et al. 2014). Statistics were calculated in sliding windows of 10 Kb for both chromosomes. Additionally, we calculated calculated recombination events accross each chromosome, according to the four-gamete test, also implemented in *PopGenome*.

Population structure

Local Population structure of the selected data was investigated using discriminant analysis of principal components (dapc), performed using the R package *adegenet* (Jombart, Devillard, and Balloux 2010). The dapc analysis was conducted according to the guidelines outlined by Thia 2022. To determine the number of clusters suited for discrimination, we performed a K-means clustering algorithm prior to the dapc using *adegenet*'s in-built function *find.clusters()*. The K-means algorithm implemented in *find.cluster()* uses proir tranformation by principal component analysis (PCA), retaining principal component axis which approximately 80-90% of the variance in the data set. The number of clusters to be used for the dapc was selected based on the Bayesian Information Criterion (BIC). The dapc analysis was run with K-1 PCA axes and K-1 discriminant axes, using the K-mean group assignments as priors.

Extending the dapc analysis, we did a population structure analysis using the software *ADMIXTURE* (Alexander, Novembre, and Lange 2009). The VCF files had their chromosome renamed numerically and using *plink2* we converted them to .bed files. We conducted admixture analysis for a wide range of K values, including the value used for the dapc. The results of the admixture analysis were validated by comparing the CV error values of each run of K.

Site-frequency spectrum

In order to calculate an unfolded site-frequency spectrum for the SNP data, we used *Plasmodium reichenowi* as an outgroup. The genome assembly *PrCDC* of *P.reichenowi* was obtained from (`ftp://ftp.sanger.ac.uk/pub/project/pathogens/reichenowi/2018/October_2018`). Since the PrCDC assembly was assembled by using the *Plasmodium falciparum* Pf7 in a refernce guided assembly, we compare the two using genome alignment. We aligned the *P.falciparum* reference genome (Pf7) with the *P.reichowi* reference (PrCDC) using *mummer4* (Marçais et al. 2018), and extracted the segregating sites between both assemblies. Using an R script (`https://taylorreiter.github.io/2019-05-11-Visualizing-NUCmer-Output/`), we plotted the mapping coverage to ensure sufficient overlap between the genomes (Fig. 10 and 11)

To polarize the SNPs found in the Pf7 VCF data set, we compare the segregating sites between out Pf7 refernce and the outgroup with the polymorphic sites found in the vcf file. From the genome alignment we found 927 and 2082 segregating positions, respectively for chromosome 2 and 11, which were also found to be polymorphic in the *P.falciparum* data set. Of these, 524 and 1207 SNPs, respectively, had an alternative (ALT) alleles matching the allele found in the outgroup. These SNPs had their alternative allel redesignated as the anchestral allele, while the reference allele is assumed to be the 'new' alternative (derived) allele. For sites, that were not segregating between the reference and the outgroup, polarisation cannot be performed, and we assume the reference allele to be anchestral. The idea is that the allele found on the outgroup can be considered the anchestral allele if it matches an allele found in the *P. falciparum* population, because the probability of the same allele arising in the outgroup and our population is very small under the infinte-sites model.

After each allele recieved a designation of anchestral or derived, we removed sites, that either had no alternative allele or a fixed alternative allele, since these would skew the site-frequency spectrum. Additionally, we divided the number of allele counts and total counts by two to reflect the haploid nature of *P.falciparum.* This resulted in final sets of 5399 and 13847 SNPs, respectively for chromosome 2 and 11, which could be used to plot the unfolded site-frequency spectrum.

Results

Across both chromosomes analyzed, we found that diversity statisic, such as θ_{W}, θ_{pi}, the number of segregating sites S and Tajima's D to vary quite a lot (Fig. 1 and 2), as is to be expect from different regions of the genome. Interestingly, we found overall very low values of Tajima's

D across all samples (Fig. 1A and 2A). When comparing the nucleotide diversity measure of Watterson, θ_W to an estimate of recombination events Hudson-Kaplan-RM, we find a higher diversity for windows with more recombination events (Fig. 1E and 2E). This is expected under linked selection, where selection between linked loci will drag down polymorphism. Less linked regions with higher recombination frequencies, therefore are estimated to show more polymorphism. At high recombination rates, this effect, θ_W values flatten out, because polymorphism ultimately is constrained by mutation.

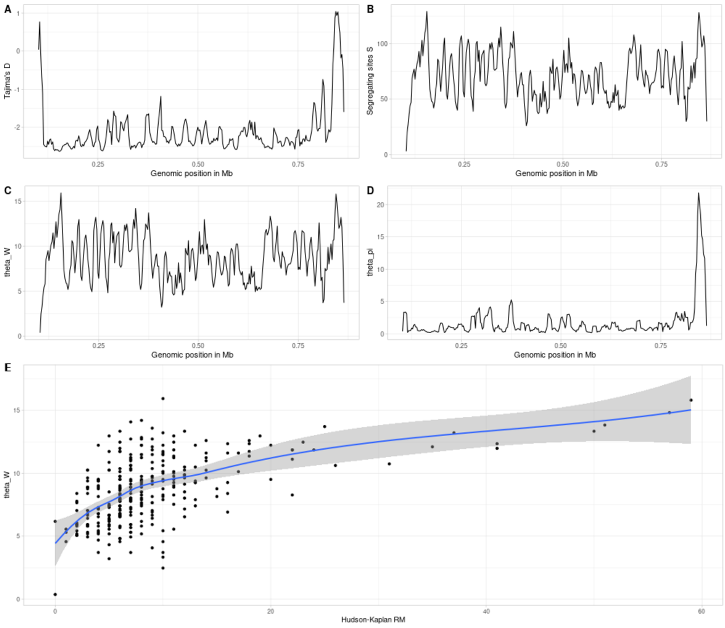

Figure 1: Diversity statistics of chromosome 2, calculated in sliding windows of 10 Kb length. (A) Tajima's D, (B) Segregating sites S, (C) $\theta_{\text{Watterson}}$, (D) θ_{pi} and (E) θ_W plotted against Hudson-Kaplan RM in 10 Kb windows.

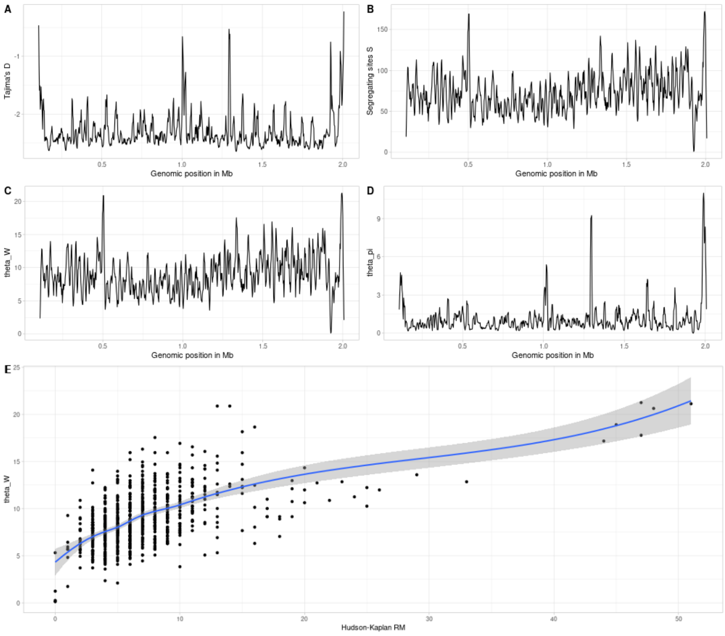

Figure 2: Diversity statistics of chromosome 11, calculated in sliding windows of 10 Kb length. (A) Tajima's D, (B) Segregating sites S, (C) $\theta_{\text{Watterson}}$, (D) θ_{pi} and (E) θ_W plotted against

Hudson-Kaplan RM in 10 Kb windows.

We conducted two parallel approaches to investigate possible population structure in our *P. falciparum* samples, discriminant of principle components (dapc) and bayesian structure analysis using the software ADMIXTURE. We first chose to analyze if population structure follows the sampled locations, using these as our a priori labels. Generally we found high levels of admixture between populations, although we could observe some signals of geographic structuring (Fig. 3 and 4). Specifically, the Gambian samples showed the most differentiation in the admixture analysis for the other locations. While the distruct plots of te structure analysis show similar low, but present population structure for both chromosomes, it is noteworthy, that chromosome 2 seems to have much less population structure than chromosome 11, as can be seen from the dapc (Fig. 3B) and (Fig. 4B).

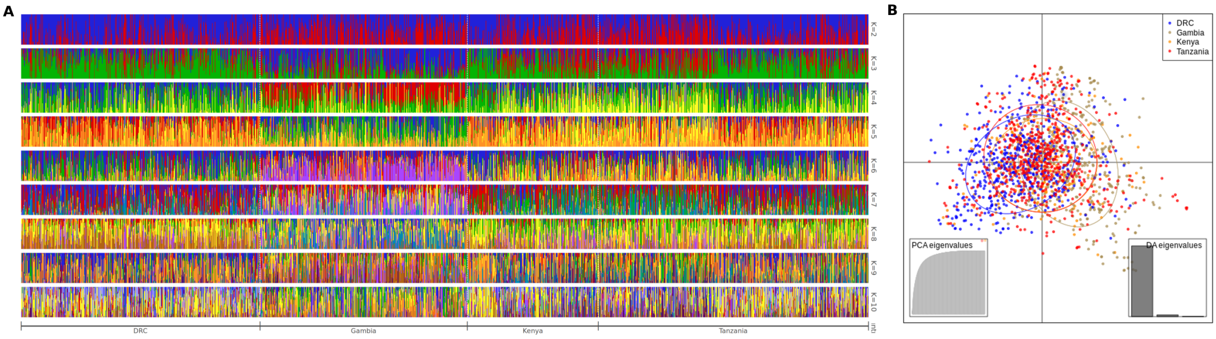

Figure 3: Population structure analysis of *P. falciparum* chromosome 2, showing the clustering by predefined labels based on sampling location.

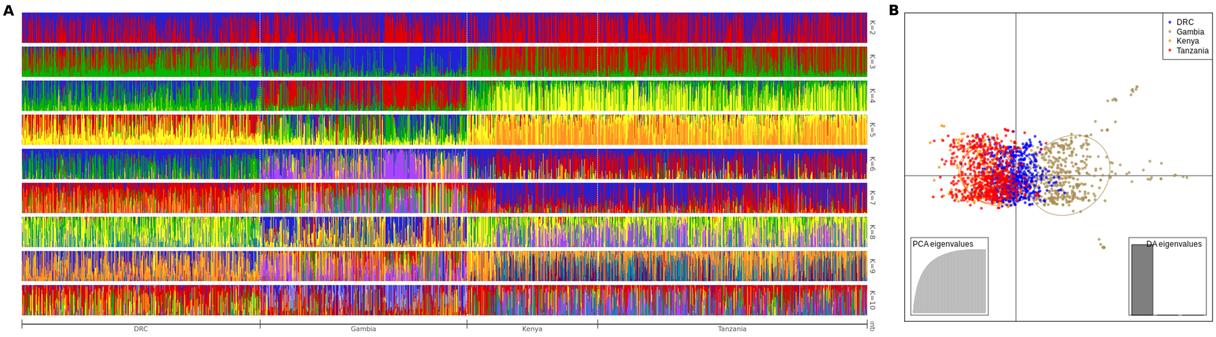

Figure 4: Population structure analysis of *P. falciparum* chromosome 11, showing the clustering by predefined labels based on sampling location.

Next, we were interested in possible other causes of population structure in african *P. falciparum* other than the geographic locations. We inferred the number of groups and membership *de novo* using a K-means clustering algorithm, *find.clusters()*, described in the dapc manual (Jombart, Devillard, and Balloux 2010). The optimal value of K was determined to be 9 and 10, respectively for chromosome 2 and 11, significantly more than the a priori groups based on sampling location. We performed dapc analysis on these newly inferred groups, and clustering of both chromosomes into groups, but some of the clusters seem to overlap quite significantly. Chromosome 2 shows more distinct clustering by inferred group labels, than chromosome 11, but they do not correlate with the sampling location or year (Fig. 5). In contrast, the inferred clusters of chromosome 11 seem to correlate roughly with the sampling location and year. (Fig. 6). When sorting the individuals in the admixture analysis by these inferred group labels, we found the results to agree with the dapc (Fig. 8).

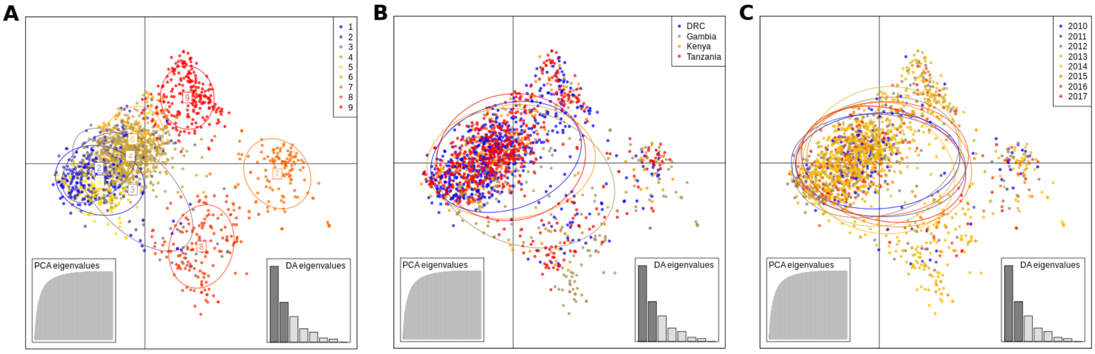

Figure 5: DAPC analysis of nine population clusters inferred *de novo* with a K-means algorithm, chromosome 2. (A) Individuals colored according to assigned clusters. (B) Individuals colored by their sample location. (C) Individuals are colored by the sampled year.

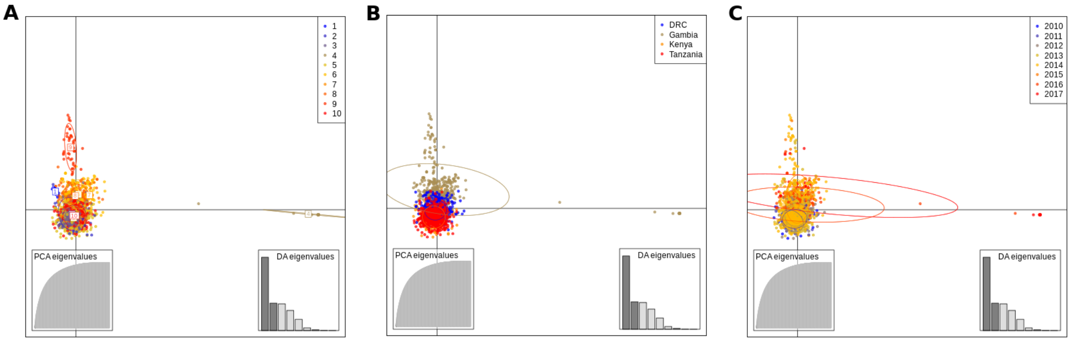

Figure 6: DAPC analysis of nine population clusters inferred *de novo* with a K-means algorithm, chromosome 2. (A) Individuals colored according to assigned clusters. (B) Individuals colored by their sample location. (C) Individuals are colored by the sampled year.

In order to allow for more detailed inferences about the demographic history of the population and evolutionary forces such as selection we calculated the a polarized site frequency spectrum (SFS), for both chromosomes (Fig. 7). The SFS we found, both had a U-shaped character, with an excess of singletons compared the the neutral expectation and increased high-frequency alleles. Because hidden population can influence the shape of the SFS, when samples from different populations are pooled together (Lapierre et al. 2016), we calculated the SFS for samples from Gambia as well, as Gambia was the most distinct country from the others (Fig. 9). We found the U-shape in the Gambian SFS as well, which indicates the discovered U-shape is not due to unidentified substructuring of the sampled parasites.

Discussion

When calculating genome-wide levels of polymorphism, we found levels of both θ_W and pi to vary significantly for different regions of the genome. While this is expected due to varying levels of mutations and recombination, selection of certain genes and region is also a plausible explaination. It remains to be indentified if a significant proportion of putative genes under selection coincide with the observed regions of reduced polymorphism found. Furthermore, we found Tajima's D to vary across the genome, but generally being strongly negative for both

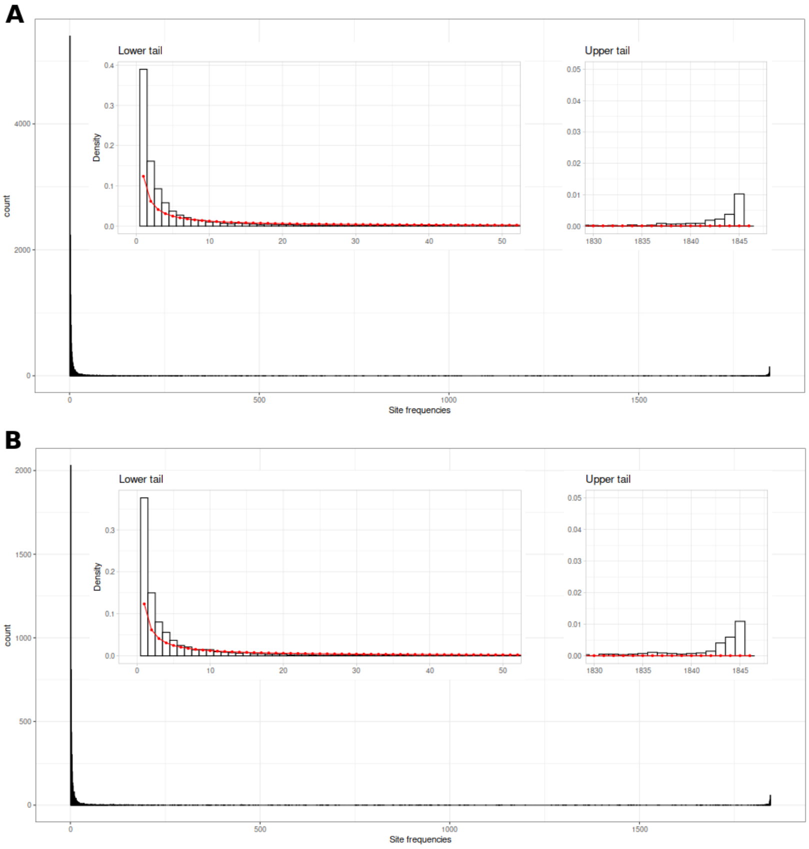

Figure 7: Unfolded site frequency for chromosome 2 (A) and 11 (B).

chromosomes analyzed in each subpopulation separately, as well as for the total population. Negative values of Tajima's D across multiple chromosomes can be sign of population expansion, which would be consistent with findings of other studies (Rich and Ayala 2000). This is likely linked to expansion of human populations and increased vectorial capacity since the last glacial maximum (Donnelly, Licht, and Lehmann 2001).

We investigated population structure in *P. falciparum* samples from the African countries of Gambia, Democratic Republic of Congo, Kenya and Tanzania. Applying DAPC analysis, we found very little population structure, with samples from Gambia being the most differntiated in one of the two chromosomes, supporting a weak form of isolation by distance, as Gambia is the most distant from the other three countries. It is feasible, that this relative lack of population structure is due to strong migration from humans, as human movement has shown to influence parasite movement at smaller geographic scale in Kenya and Uganda (Nderu et al. 2019).

Complementing the DAPC analysis, we performed ADMIXTURE analysis, and found high levels of admixture between all subpopulations of *P. falciparum*. These results can be interpreted in the context of high malaria transmission leading to increased recombination between different strains during co-infection (Anderson et al. 2000). It remains to be investigated, how recombination rates correlate with levels of within-host diversity F_{WS} and genome-wide linkage disequilibrium.

We assigned ancestral states using the outgroup species *P. reichenowi* and calculated the unfolded site-frequency spectrum (SFS) for chromosomes 2 and 11. The number of SNPs we were able to polarize, was limited to divergent sites between the outgroup and ingroups which were segregating in *P.falciparum*. While we deem our method of assigning ancestral states to alleles to be robust under the infinite-sites model, we acknowledge misorientations could happen due incomplete lineage sorting and sites not fully diverged between the two parasite species or recurrent mutation at the same site (i.e. homplasy) (Baudry and Depaulis 2003). Including more samples of the outgroup can increase the certainty of selecting sites, that are divergent between *P. reichenowi* and *P. falciparum*, reducing cross-species polymorphic sites. Additionally, including polarization of alleles using a more distant outgroups, such as *P. vivax*, might yield more sites, that can be assigned anchestral states. We believe that allele misorientations due to sequencing errors of the outgroup are rather unlikely with our method of polarisation, because the allele to be polarized not only should be fixed in the outgroup sequence, but also segregating in the ingroup species. While sequencing errors, that produce the same allele as also found polymorphic in the ingroup are possible, we conclude these to be rather unlikely and sufficient to produce the pervasive u-shaped SFS we observe. We propose investigating measures such as the probability of misorientation used in other studies (Baudry and Depaulis 2003, Fay and Wu 2000), to assess the impact if has on the SFS.

Another factor, which might affect affect the genetic variation in *Plasmodium* parasites is their unusually high AT-content, reported to range between 75-85% (Videvall 2018). This makes *Plasmodium* species one of the eukaryotic organisms, with the lowest known genomic GC-content. While the reason for this is still unknown, research has shown, that the common ancestor between *P. falciparum* and *P. vivax*, was already extremly AT-rich, but recently *P.vivax* has been gaining GC-content (Nikbakht, Xia, and Hickey 2014). Such AT bias can be caused by mutational or gene conversion bias, leading to effects of molecular variation not expected by neutral theory. For example, GC- biased gene conversion in bacteria has been shown to skew the site-frequency spectrum and influence demographic inference made from reconstructed genealogies (Lapierre et al. 2016). While we cannot say if the high AT content is leading to the U-shaped SFS we oberve in *P. falciparum*, further analysis excluding SNPs known for this bias, can help characterize the effect of AT-content. Furthermore, the dynamic and highly variable nature of nucleotide content in closely related species of *Plasmodium* need to be accounted for when doing comparative analysis of molecular polymorphism.

A common interpretation of an excess of high-frequency derived variants in the genome involves selective sweeps, driving the increase of these alleles. The U-shaped upper tail in the SFS results from selection at multiple loci, including hitchhiking of neutral variants linked to beneficial alleles, while the lower tail excess of singletons is due to negative selection keeping deleterious mutations at low frequency (Bustamante et al. 2001). Similar to positive selection, fluctuation selection is also predicted to affect the SFS in a similar way, leading to increased high-frequency derived alleles (Huerta-Sanchez, Durrett, and Bustamante 2008). Comparing the locations of some of these high-frequency SNPs to published candidate genes under selection, can be help identify the contribution of selection to the U-shaped SFS. Additionally, genome-wide recombination rates, can uncover how of genetic hitchhiking affects to levels of diversity in different malaria transmission regions. Because of the parasitic nature of *P. falciparum*, we would expect strong selection to act of genes involved in drug resistance and transmission, as survival of the parasite heavily depends on its ability to infect hosts.

Indeed, several studies have found both directional positive selection and balancing types of selection in *P. falciparum* (Naung et al. 2022, Mobegi et al. 2014). Numerous parasite surface antigens, have showed to harbour signatures of balancing selection and high levels of diversity, as would be expected for proteins at the host-parasite interface (Naung et al. 2022). Knowledge of the selective mechanisms mainting diversity at antigens, will be crucial for effective vacchine development. Usage of antimalarial drugs, such as chloroquine has been shown to drive allele frequency changes over a period of 25 years (Nwakanma et al. 2013). Interestingly, patterns of selection for drug resistance reflect historical usage of antimalaria drugs varying between countries. Similarly, different regional selection pressures can be expected from geographical differences in transmission intensity and malaria endemicity (Mobegi et al. 2014), which is reflected in a high divergence (high F_{ST}), between genomic regions between subpopulations. It is also feasible, that seasonal differences in transmission intensity can cause selection pressures to fluctuate temporally. Particularly the effect of transmission intensity on selection remains to be studied, since transmission intensity has a strong influence on ecological and life-history traits, such as intra-host competition, reproductive strategies and possibly even dormancy. These traits can have strong effects on selection and the observed SFS.

Paraites, such as *P. falciparum* are often display unique life cycles from their free-living counterparts, characterized by recurrent infection and transmission bottlenecks, driving the survival of few infectious individuals. Frequent clonal reproduction phases and a high number of infectious propagules can lead to sweepstake reproduction. Strong intra-host selection and resulting selective bottlenecks from host-parasite coevolution also have the potential to produce multiple merger coalescent signatures (Tellier and Lemaire 2014). The extent of multiple merger genealogies in *Plasmodium* has yet to be explored and surely varies with transmission intensity. Low transmisstion rates would likely increase the rates on multiple mergers happening, since clonal expansion of the pathogen increases the variance in the offspring distribution. This could mean multiple mergers in non-malaria endemic regions are more wide-spread than in Africa, although the connection of multiple merger genealogies to transmission intensity and parasite outcrossing remains a question for further research.

In summary, characterizing genetic variation in *P. falciparum* is highly important for evalutating the efficacy of control measures and drug interventions, as genetically diverse population have a greater reservoir of alleles, which can selection can act on. In addition to quantifying the levels of genetic variation, the structure of the parasite population can inform us about the levels of gene flow and outcrossing, which in turn allow alleles to spread between (sub-)populations. General multiple merger coalescents can allow research on *Plasmodium* population genetics to incorperate special life-history traits and improve inferences and interpretations of parasite evolution.

Bibliography

References

Alexander, David H., John Novembre, and Kenneth Lange (2009). "Fast Model-Based Estimation of Ancestry in Unrelated Individuals". In: *Genome Research* 19.9, pp. 1655–1664. DOI: 10.1101/gr.094052.109. URL: http://dx.doi.org/10.1101/gr.094052.109.

Anderson, Timothy J. C. et al. (Oct. 2000). "Microsatellite Markers Reveal a Spectrum of Population Structures in the Malaria Parasite Plasmodium falciparum". In: *Molecular Biology and Evolution* 17.10, pp. 1467–1482. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a026247. URL: https://doi.org/10.1093/oxfordjournals.molbev.a026247.

Baudry, Emmanuelle and Frantz Depaulis (Nov. 2003). "Effect of Misoriented Sites on Neutrality Tests With Outgroup". In: *Genetics* 165.3, pp. 1619–1622. ISSN: 1943-2631. DOI: 10.1093/genetics/165.3.1619. eprint: https://academic.oup.com/genetics/article-pdf/165/3/1619/42052236/genetics1619.pdf. URL: https://doi.org/10.1093/genetics/165.3.1619.

Bustamante, Carlos D et al. (2001). "Directional Selection and the Site-Frequency Spectrum". In: *Genetics* 159.4, pp. 1779–1788. DOI: 10.1093/genetics/159.4.1779. URL: http://dx.doi.org/10.1093/genetics/159.4.1779.

Danecek, Petr et al. (2021). "Twelve Years of Samtools and Bcftools". In: *GigaScience* 10.2, nil. DOI: 10.1093/gigascience/giab008. URL: http://dx.doi.org/10.1093/gigascience/giab008.

Donnelly, Martin J., Monica C. Licht, and Tovi Lehmann (2001). "Evidence for Recent Population Expansion in the Evolutionary History of the Malaria Vectors Anopheles Arabiensis and Anopheles Gambiae". In: *Molecular Biology and Evolution* 18.7, pp. 1353–1364. DOI: 10.1093/oxfordjournals.molbev.a003919. URL: http://dx.doi.org/10.1093/oxfordjournals.molbev.a003919.

Fay, Justin C and Chung-I Wu (2000). "Hitchhiking Under Positive Darwinian Selection". In: *Genetics* 155.3, pp. 1405–1413. DOI: 10.1093/genetics/155.3.1405. URL: http://dx.doi.org/10.1093/genetics/155.3.1405.

Freund, Fabian et al. (2023). "Interpreting the Pervasive Observation of U-Shaped Site Frequency Spectra". In: *PLOS Genetics* 19.3, e1010677. DOI: 10.1371/journal.pgen.1010677. URL: http://dx.doi.org/10.1371/journal.pgen.1010677.

Huerta-Sanchez, Emilia, Rick Durrett, and Carlos D Bustamante (2008). "Population Genetics of Polymorphism and Divergence Under Fluctuating Selection". In: *Genetics* 178.1, pp. 325–337. DOI: 10.1534/genetics.107.073361. URL: http://dx.doi.org/10.1534/genetics.107.073361.

Jombart, Thibaut, Sébastien Devillard, and François Balloux (2010). "Discriminant Analysis of Principal Components: a New Method for the Analysis of Genetically Structured Populations". In: *BMC Genetics* 11.1, p. 94. DOI: 10.1186/1471-2156-11-94. URL: http://dx.doi.org/10.1186/1471-2156-11-94.

Knaus, Brian J. and Niklaus J. Grünwald (2017). "VCFR: a Package To Manipulate and Visualize Variant Call Format Data in R". In: *Molecular Ecology Resources* 17.1, pp. 44–53. ISSN: 757. URL: https://dx.doi.org/10.1111/1755-0998.12549.

Lapierre, Marguerite et al. (2016). "The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography". In: *Molecular Biology and Evolution* 33.7, pp. 1711–1725. DOI: 10.1093/molbev/msw048. URL: http://dx.doi.org/10.1093/molbev/msw048.

Li, Haipeng and Thomas Wiehe (2013). "Coalescent Tree Imbalance and a Simple Test for Selective Sweeps Based on Microsatellite Variation". In: *PLoS Computational Biology* 9.5, e1003060. DOI: 10.1371/journal.pcbi.1003060. URL: http://dx.doi.org/10.1371/journal.pcbi.1003060.

MalariaGEN et al. (2023). "Pf7: an Open Dataset of Plasmodium Falciparum Genome Variation in 20,000 Worldwide Samples [version 1; Peer Review: 3 Approved]". In: *Wellcome Open Research* 8.22. DOI: 10.12688/wellcomeopenres.18681.1.

Marçais, Guillaume et al. (2018). "Mummer4: a Fast and Versatile Genome Alignment System". In: *PLOS Computational Biology* 14.1, e1005944. DOI: 10.1371/journal.pcbi.1005944. URL: http://dx.doi.org/10.1371/journal.pcbi.1005944.

Marchi, Nina and Laurent Excoffier (2020). "Gene Flow As a Simple Cause for an Excess of High-frequency-derived Alleles". In: *Evolutionary Applications* 13.9, pp. 2254–2263. DOI: 10.1111/eva.12998. URL: http://dx.doi.org/10.1111/eva.12998.

Mobegi, Victor A. et al. (2014). "Genome-Wide Analysis of Selection on the Malaria Parasite Plasmodium Falciparum in West African Populations of Differing Infection Endemicity". In: *Molecular Biology and Evolution* 31.6, pp. 1490–1499. DOI: 10.1093/molbev/msu106. URL: http://dx.doi.org/10.1093/molbev/msu106.

Mu, Jianbing et al. (Sept. 2005). "Recombination Hotspots and Population Structure in Plasmodium Falciparum". In: *PLOS Biology* 3.10, null. DOI: 10.1371/journal.pbio.0030335. URL: https://doi.org/10.1371/journal.pbio.0030335.

Naung, Myo T. et al. (2022). "Global Diversity and Balancing Selection of 23 Leading Plasmodium Falciparum Candidate Vaccine Antigens". In: *PLOS Computational Biology* 18.2, e1009801. DOI: 10.1371/journal.pcbi.1009801. URL: http://dx.doi.org/10.1371/journal.pcbi.1009801.

Nderu, David et al. (2019). "Genetic Diversity and Population Structure of <i>plasmodium Falciparum</i> in Kenyan-Ugandan Border Areas". In: *Tropical Medicine & International Health* 24.5, pp. 647–656. DOI: 10.1111/tmi.13223. URL: http://dx.doi.org/10.1111/tmi.13223.

Nikbakht, Hamid, Xuhua Xia, and Donal A. Hickey (2014). "The Evolution of Genomic Gc Content Undergoes a Rapid Reversal Within the Genus<i>plasmodium</i>". In: *Genome* 57.9, pp. 507–511. DOI: 10.1139/gen-2014-0158. URL: http://dx.doi.org/10.1139/gen-2014-0158.

Nwakanma, Davis C. et al. (2013). "Changes in Malaria Parasite Drug Resistance in an Endemic Population Over a 25-year Period With Resulting Genomic Evidence of Selection". In: *The Journal of Infectious Diseases* 209.7, pp. 1126–1135. DOI: 10.1093/infdis/jit618. URL: http://dx.doi.org/10.1093/infdis/jit618.

Pfeifer, Bastian et al. (2014). "Popgenome: an Efficient Swiss Army Knife for Population Genomic Analyses in R". In: *Molecular Biology and Evolution* 31.7, pp. 1929–1936. DOI: 10.1093/molbev/msu136. URL: http://dx.doi.org/10.1093/molbev/msu136.

Pouyet, Fanny et al. (2018). "Background Selection and Biased Gene Conversion Affect More Than 95 % of the Human Genome and Bias Demographic Inferences". In: *eLife* 7.nil, nil. DOI: 10.7554/elife.36317. URL: http://dx.doi.org/10.7554/eLife.36317.

Rich, Stephen M. and Francisco J. Ayala (2000). "Population Structure and Recent Evolution of <i>plasmodium Falciparum</i>". In: *Proceedings of the National Academy of Sciences* 97.13, pp. 6994–7001. DOI: 10.1073/pnas.97.13.6994. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.97.13.6994. URL: https://www.pnas.org/doi/abs/10.1073/pnas.97.13.6994.

Tellier, Aurélien and Christophe Lemaire (2014). "Coalescence 2.0: a Multiple Branching of Recent Theoretical Developments and Their Applications". In: *Molecular Ecology* 23.11, pp. 2637–2652. DOI: 10.1111/mec.12755. URL: http://dx.doi.org/10.1111/mec.12755.

Thia, Joshua A. (2022). "Guidelines for Standardizing the Application of Discriminant Analysis of Principal Components To Genotype Data". In: *Molecular Ecology Resources* 23.3, pp. 523–

538. DOI: 10.1111/1755-0998.13706. URL: http://dx.doi.org/10.1111/1755-0998.13706.

Videvall, Elin (2018). “Plasmodium Parasites of Birds Have the Most At-Rich Genes of Eukaryotes”. In: *Microbial Genomics* 4.2, nil. DOI: 10.1099/mgen.0.000150. URL: http://dx.doi.org/10.1099/mgen.0.000150.

WHO (2022). *World malaria report 2022*. World Health Organization.

Appendix

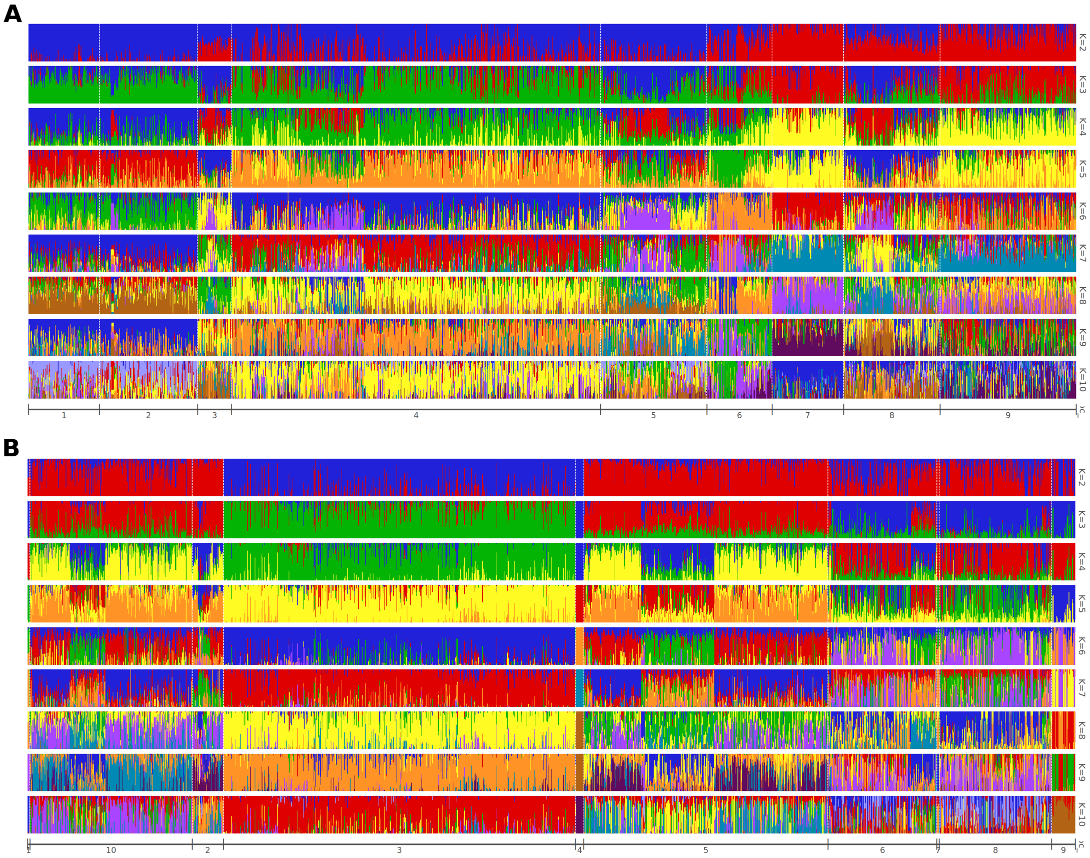

Figure 8: Admixture analyis of *P. falciparum* samples, sorted by *de novo* clusters, identified using K-means. (A) Chromosome 2 and (B) 11.

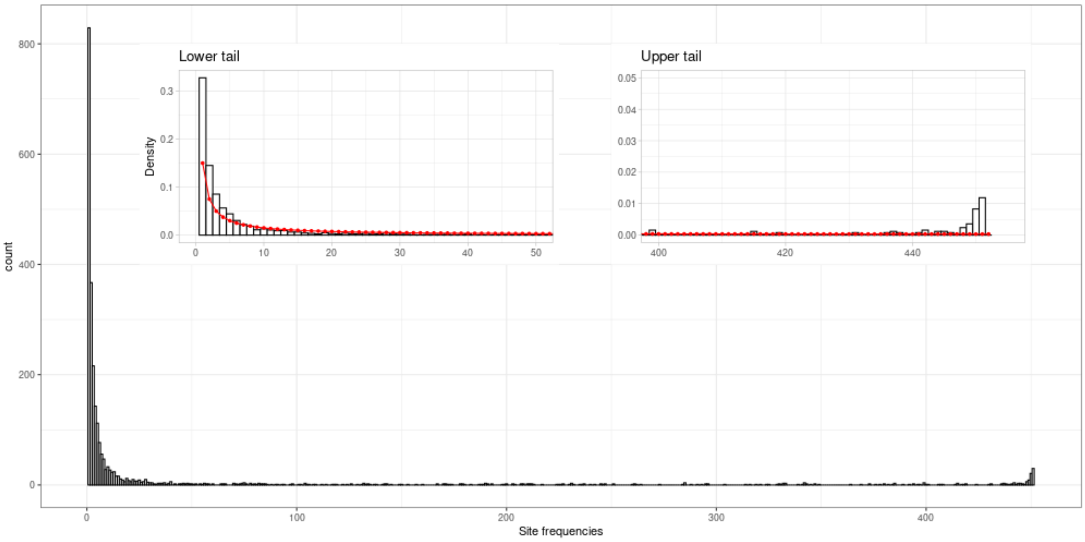

Figure 9: Unfolded site frequency spectrum of Gambian *P. falciparum* samples.

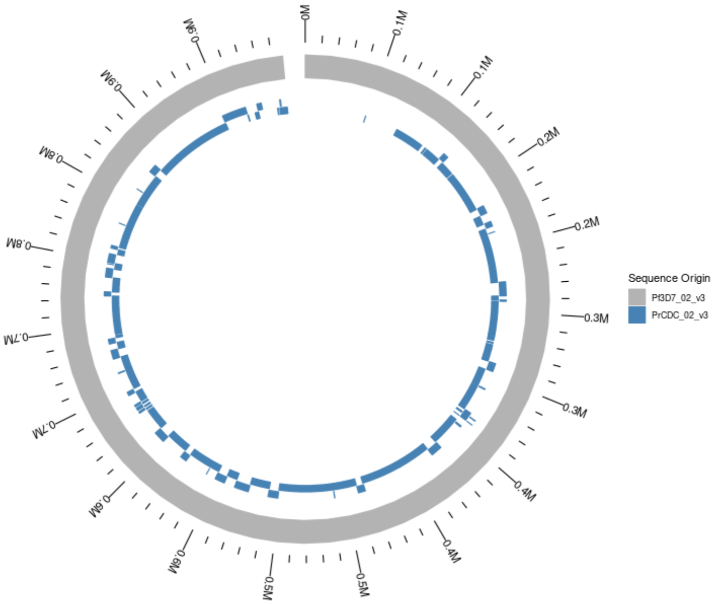

Figure 10: Mapping coverage of *P. reichenowi* PrCDC aligned to the *P. falciparum* Pf3D7 reference genome (*Pf7*), chromosome 2.

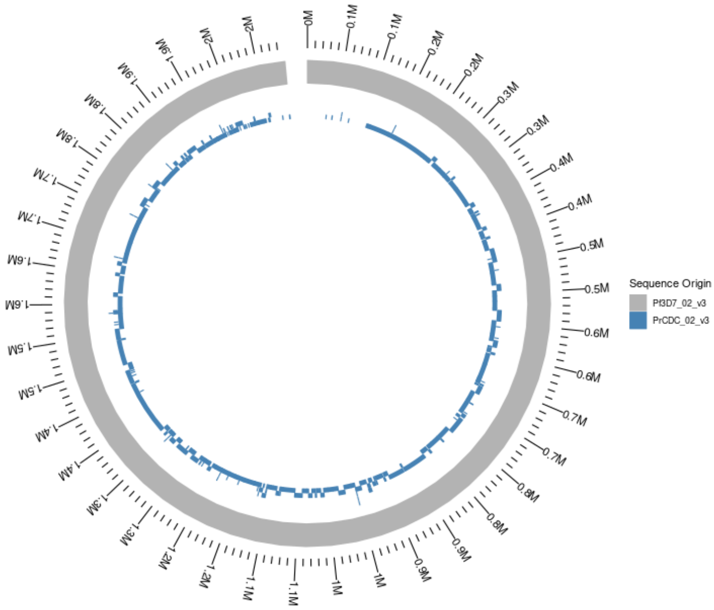

Figure 11: Mapping coverage of *P. reichenowi* PrCDC aligned to the *P. falciparum* Pf3D7 reference genome (*Pf7*), chromosome 11.

Code

The complete code used for this report can be found in the git repository:
`https://github.com/pwolper/Pf7-coevolution`

Filtering

Downloaded VCF files where filtered and reindexed.

```
bcftools view \
    --include 'FILTER="PASS" && N_ALT=1 && TYPE="snp" && VSQLOD>5.0'\
    --samples-file $samples_dir/Pf7_african_samples_ids.txt \
    --output-type z\
    --output-file  Pf3D7_02_v3.SNP.vcf \
    Pf3D7_02_v3.pf7.vcf.gz

bcftools index -t Pf3D7_02_v3.SNP.vcf
```

Heterozygous genotypes were removed using the following code as an example:

```
bcftools filter -S . -e 'GT=="het"' chr2/Pf3D7_02_v3.afr_samples.qSNP.vcf.gz\
    -o chr2/Pf3D7_02_v3.afr_samples.qSNP.GT_filtered.vcf -Oz
```

DAPC

In order to run the dapc on our vcf files we used the R package *vcfR* (Knaus and Grünwald 2017) to read in the file and convert it to a genlight object. Subsequently, we read in a vector of country membership and set haploidy. The dapc is conducted according to guideslines stated in Thia 2022.

```
Pf7.snp <- read.vcfR(vcf_path, verbose = TRUE)

Pf7.snp.gl <- vcfR2genlight(Pf7.snp)

pop(Pf7.snp.gl) <- Pf7.metadata$Country
ploidy(Pf7.snp.gl) <- 1

clust <- find.clusters(Pf7.snp.gl, max.n.clust = 30) #de novo cluster inference
dapc <- dapc(Pf7.snp.gl, n.pca = 3, n.da = 3)
```

ADMIXTURE

Before converting the vcf files were converted to .bed files using *plink2*, we had to rename the chromosome from Pf3D7_x_v3 to x. We used a names.txt file according to the bcftools documentation.

```
plink2 --vcf ../vcf/chr11/Pf3D7_11_v3.afr_samples.qSNP.GT_filtered.renamed.vcf.gz \
    --make-bed \
    --allow-extra-chr \
    -out Pf3D7_11_v3.afr_samples.qSNP.GT_filtered.renamed
```

The admixture analysis was performed with the following options:

```
for K in {2..10}; do
    admixture $bed_file $K --cv | tee log{$K}.out; done
```

Unfolded SFS

The genome sequence of the outgroup *P. reichenowi* was aligned to the *P. falciparum* reference sequence Pf7 using the following code:

```
nucmer --mum -p chr11/Pf7_PrCDC \
    ../Pf7/seqs/Pf3D7_11_v3.fasta ../P.reichenowi/embl.PrCDC/PrCDC_11_v3.fasta

show-coords -c -l -r -T chr11/Pf7_PrCDC.delta > chr11/Pf7_PrCDC_coords.txt

show-snps -T chr11/Pf7_PrCDC.delta > chr11/Pf7_PrCDC_SNPs.txt

awk \
    'NR>3 && $2 !~ /\./ && $3 !~ /\./ {print}' \
    chr11/Pf7_PrCDC_SNPs.txt > chr11/Pf7_PrCDC_11_SNPs_formatted.txt
```

SNP frequencies and counts were extracted with the following code:

```
bcftools query -f '%POS %REF %ALT %AF %AC %AN\n'\
    Pf3D7_02_v3.afr_samples.qSNP.GT_filtered.vcf.gz > Pf7.02.vcf.qSNP.AF_AC_AN.txt
```