Falciparum malaria from coastal Tanzania and Zanzibar remains highly connected despite effective control efforts on the archipelago

Andrew Morgan
University of North Carolina at Chapel Hill
Nicholas Brazeau
University of North Carolina at Chapel Hill
Billy Ngasala
Muhimbili University of Health and Allied Sciences
Lwidiko Mhamilawa
Muhimbili University of Health and Allied Sciences
Madeline Denton
University of North Carolina at Chapel Hill
Mwinyi Msellem
Mnazi Mmoja Hospital
Ulrika Morris
Karolinska Institutet
Dayne Filer
University of North Carolina at Chapel Hill
Ozkan Aydemir
Brown University
Jeffrey Bailey
Brown University
Jonathan Parr
University of North Carolina at Chapel Hill
Andreas Mårtensson
Uppsala Universitet
Anders Bjorkman
Karolinska Institutet
Jonathan Juliano (jjuliano@med.unc.edu)
University of North Carolina at Chapel Hill https://orcid.org/0000-0002-0591-0850

Research

Keywords: Plasmodium, malaria, population genetics

Posted Date: January 24th, 2020

DOI: https://doi.org/10.21203/rs.2.18557/v3

Version of Record: A version of this preprint was published on January 28th, 2020. See the published version at https://doi.org/10.1186/s12936-020-3137-8.

Abstract

Background Tanzania's Zanzibar archipelago has made significant gains in malaria control over the last decade and is a target for malaria elimination. Despite consistent implementation of effective tools since 2002, elimination has not been achieved. Importation of parasites from outside of the archipelago is thought to be an important cause of malaria's persistence, but this paradigm has not been studied using modern genetic tools. Methods Whole-genome sequencing (WGS) was used to investigate the impact of importation, employing population genetic analyses of Plasmodium falciparum isolates from both the archipelago and mainland Tanzania. Ancestry, levels of genetic diversity and differentiation, patterns of relatedness, and patterns of selection between these two populations were assessed by leveraging recent advances in deconvolution of genomes from polyclonal malaria infections. Results Significant decreases in the effective population sizes were inferred in both populations that coincide with a period of decreasing malaria transmission in Tanzania. Identity by descent analysis showed that parasites in the two populations shared long segments of their genomes, on the order of 5 cM, suggesting shared ancestry within the last 10 generations. Even with limited sampling, two of isolates between the mainland and Zanzibar were identified that are related at the expected level of half-siblings, consistent with recent importation. Conclusions These findings suggest that importation plays an important role for malaria incidence on Zanzibar and demonstrate the value of genomic approaches for identifying corridors of parasite movement to the island.

Background

Despite nearly two decades of progress in control, malaria remains a major public health challenge with an estimated 219 million cases and 435,000 deaths in 2017 globally [1]. The mainland of Tanzania has heterogeneous transmission of mainly *Plasmodium falciparum*, but overall levels of malaria remain high, accounting for approximately 3% of global malaria cases [1]. However, through a combination of robust vector control and access to efficacious anti-malarial treatment, the archipelago of Zanzibar has been deemed a pre-elimination setting, having only low and mainly seasonal transmission [2]. Despite significant efforts, however, elimination has been difficult to achieve in Zanzibar. The reasons for Zanzibar's failure to achieve elimination are complex and likely driven by several key factors: 1) as transmission decreases, the distribution of cases changes and residual transmission is more focal and mainly outdoors [3]; 2) a significant number of malaria infections are asymptomatic and thus untreated and remain a source for local transmission [4–7]; and 3) the archipelago has a high level of connectivity with the mainland, thus imported malaria through human travel may play an increasing relative role in transmission.

Genomic epidemiology can supplement traditional epidemiological measures in studies of malaria transmission and biology, thereby helping to direct malaria elimination strategies [8]. Whole-genome sequencing (WGS) can be particularly useful for understanding the history of parasite populations and movement of closely related parasites over geographical distances [9,10]. Identity by descent (IBD), the sharing of discrete genomic segments inherited from a common genealogical ancestor, has been found

to be a particularly good metric for studying the interconnectivity of parasite populations [11–13]. A major obstacle to studying IBD in microorganisms, and in particular malaria, is the presence of multiple clones in a single infection. In order to address this obstacle, recent algorithms have been developed to deconvolve multiple infections into their respective strains from Illumina sequence data [14,15]. These advances now make it tractable to conduct population genetic analysis of malaria in regions of higher transmission, where infections are often polyclonal.

Decreases in malaria prevalence are hypothesized to be associated with increasing inbreeding in the parasite population, decreased overall parasite genetic diversity and a reduced complexity of infection (COI), defined as a decreased number of infecting clones [8]. This has been shown in pre-elimination settings in Asia as well as in lower transmission regions of Africa [16–18]. It has not been determined if a similar reduction in diversity has occurred in Zanzibar with the significant reduction of malaria in the archipelago. WGS data was used to: 1) characterize the ancestry of parasites in the two regions, 2) determine the levels of genetic diversity and differentiation between archipelago and mainland, 3) determine patterns of relatedness and inbreeding and 4) search for signatures of adaptation and natural selection. Inferred genetic relationships were then examined for evidence of importation of parasites from the higher transmission regions of mainland Tanzania to the lower transmission regions of the Zanzibar archipelago. These findings improve understanding of how importation may affect malaria elimination efforts in Zanzibar.

Methods

Clinical samples

WGS was attempted on 106 *P. falciparum* isolates collected from subjects with uncomplicated malaria or asymptomatic infection from 2015 to 2017. Forty-three of these were leukodepleted blood collected as part of an *in vivo* efficacy study of artemether-lumefantrine (AL) in paediatric uncomplicated malaria patients collected from 2015-2017 in Yombo, Bagamoyo District. A remaining 63 isolates were from dried blood spots (DBS) collected in Zanzibar in 2017. These came from cross-sectional surveys of asymptomatic individuals (n = 34) and an *in vivo* efficacy study of artesunate-amodiaquine (ASAQ) with single low dose primaquine (SLDP) in paediatric uncomplicated malaria patients (n = 29). These isolates essentially represent a convenience sample. Isolates were not selected for sequencing on the basis of specific clinical or epidemiologic characteristics; however, sequencing was more likely to be successful on isolates from subjects with high parasitaemia. Study participants from Zanzibar were asked to report any overnight travel away from home in the past 4 months. Responses were coded as yes (overnight travel to mainland Tanzania or Kenya) or no (no overnight travel off islands of Zanzibar). Clinical characteristics of the attempted and sequenced samples from each cohort from Zanzibar is provided in Supplemental Table 1.

Generation and sequencing of libraries

Leukodepleted blood samples and DBS were extracted using QIAmp 96 DNA blood kits per the manufacturer protocol (Qiagen, Hilden, Germany). DNA from leukodepleted blood was acoustically sheared using a Covaris E220 instrument, prepared for sequencing without enrichment using Kappa Hyper library preps, and individually barcoded per manufacturer's protocol (Kappa Biosystems, Columbus, OH). DNA extracted from DBS was enriched for *P. falciparum* DNA before library prep using two separate selective whole genome amplification (sWGA) reactions. The sWGA approach was adapted from previously published methods and employed two distinct sets of primers designed for *P. falciparum*, including the Probe_10 primer set described previously by Oyola *et al.* and another set of custom primers (JP9) designed using 'swga' [19–21]. Phosphorothioate bonds were included between the two most 3' nucleotides for all primers in both sets to prevent primer degradation. Design and evaluation of these custom primers and the sWGA approach are described in the Supplemental Materials and Supplementary Table 2. The two sWGA reactions were carried out under the same conditions. The products of the two sWGA reactions were pooled in equal volumes and acoustically sheared using a Covaris E220 instrument before library preparation using Kappa Hyper library preps. The indexed libraries were pooled and sequenced on a HiSeq4000 using 2x150 chemistry at the University of North Carolina High Throughput Sequencing Facility. Sequencing reads were deposited into the NCBI SRA (Accession numbers: pending).

Public sequencing data

Illumina short read WGS data for *P. falciparum* isolates was downloaded from public databases. This included 68 isolates from other regions of Tanzania, collected between 2010 and 2013, as well as 179 isolates from other regions, including Southeast Asia, South Asia, East and West Africa (Supplemental Table 3).

Read alignment and quality control

Raw paired-end reads were trimmed for adapter sequences with $cutadapt$ v1.18 and aligned to the *P. falciparum* 3D7 reference genome (assembly version 3, PlasmoDB version 38: https://plasmodb.org/common/downloads/release-38/Pfalciparum3D7/fasta/data/PlasmoDB-38_Pfalciparum3D7_Genome.fasta) with $bwamem$ v0.7.17-r1188. Duplicates were marked with $sambl * er$ v0.1.24. A position was defined as "callable" if it was covered by >= 5 high-quality reads (MQ >= 25, BQ >= 25), and computed the proportion of callable sites in each isolate was calculated with the Genome Analysis Toolkit (GATK) $Callab \leq Loci$ tool v3.8-0. Only isolates with >= 70% of the genome callable were used for further analysis.

Variant discovery and filtering

Short sequence variants (including SNVs, indels and complex multi-nucleotide variants) were ascertained in parallel in each isolate using GATK $HaplotypeCal \leq r$ v.4.0.3.0, then genotyped jointly across the entire cohort with GATK $Ge \neg ypeGVCFs$ according to GATK best practices. Variant discovery was limited to the core (non-hypervariable) nuclear genome as defined by Miles *et al*. [22]. Putative SNVs only were filtered using the GATK Variant Quality Score Recalibration (VQSR) method. For training sets, the

following datasets were used: QC-passing sites from the *P. falciparum* Genetic Crosses Project release 1.0 (ftp://ngs.sanger.ac.uk/production/malaria/pf-crosses/1.0/; [22]) (true positives, prior score Q30); QC-passing sites from the Pf3K release v5.1 (ftp://ngs.sanger.ac.uk/production/pf3k/release_5/5.1/) (true positives + false positives, prior score Q15). Site annotations QD, MQ, MQRankSum, ReadPosRankSum, FS, SOR were used and the model was trained with 4 Gaussian components. A VQSLOD threshold -0.0350 achieved 90% sensitivity for re-discovering known sites in the training sets. All biallelic SNVs with VQSLOD at or above this threshold were retained.

Isolates may contain multiple strains that are haploid resulting in mixed infections with arbitrary effective ploidy. To account for this complexity of infection (COI), prior literature was followed [23] and the following quantities were calculated at each variant site: for each isolate, the within-sample allele frequency (WSAF), the proportion of mapped reads carrying the non-reference allele; the population-level allele frequency (PLAF), the mean of within-sample allele frequencies; and the population-level minor allele frequency (PLMAF), the minimum of PLAF or 1-PLAF. These calculations were performed with $vcfdowsaf$ (https://github.com/IDEELResearch/vcfdo).

Analyses of mutational spectrum

Ancestral *versus* derived alleles at sites polymorphic in *P. falciparum* were assigned by comparison to the outgroup species *Plasmodium reichenowi*. Briefly, an approximation to the genome of the *P. reichenowi* - *P. falciparum* common ancestor (hereafter, "ancestral genome") was created by aligning the *P. falciparum* 3D7 assembly to the *P. reichenowi* CDC strain assembly (version 3, PlasmoDB version 38: https://plasmodb.org/common/downloads/release-38/PreichenowiCDC/fasta/data/PlasmoDB-38_PreichenowiCDC_Genome.fasta) with $vcmer$ v3.1 using parameters "-g 500 -c 500 -l 10" as in [24]. Only segments with one-to-one alignments were retained; ancestral state at sites outside these segments was deemed ambiguous. The one-to-one segments were projected back into the 3D7 coordinate system. Under the assumption of no recurrent mutation, any site polymorphic in *P. falciparum* is not expected to also be mutated on the branch of the phylogeny leading to *P. reichenowi*. Thus, the allele observed in *P. reichenowi* is the ancestral state conditional on the site being polymorphic. Transitions-transversion (Ti:Tv) ratios and mutational spectra were tallied with $bcf \rightarrow olsstats$ v1.19.

Analyses of ancestry and population structure

VQSR-passing sites were filtered more stringently for PCA to reduce artifacts due to rare alleles and missing data. Genotype calls with GQ < 20 or DP < 5 were masked; sites with < 10% missing data and PLMAF >5% after sample-level filters were retained for PCA, which was performed with $aktpca$ v3905c48 [25]. For calculation of f_3 statistics, genotype calls with GQ < 10 or DP < 5 were masked; sites with <10% missing data and PLMAF >1% after sample-level filters were retained. Then f_3 statistics were calculated from WSAFs rather than nominal diploid genotype calls, using $vcfdof3stat$.

Estimation of sequence diversity

Estimates of sequence diversity and differentiation were obtained from the site-frequency spectrum (SFS), which in turn was estimated directly from genotype likelihoods with *ANGSD* 0.921-11-g20b0655 [26] using parameters "-doCounts 1 -doSaf 1 -GL 2 -minDepthInd 3 -maxDepthInd 2000 -minMapQ 20 -baq 1 -c 50." Unfolded SFS were obtained with the *ANGSD* tool *realSFS* using the previously-described ancestral sequence from *P. reichenowi*. All isolates were treated as nominally diploid for purposes of estimating the SFS because systematic bias against mixed isolates was noted when using *ANGSD* in haploid mode. Four-fold degenerate and zero-fold degenerate sites were defined for protein-coding genes in the usual fashion using transcript models from PlasmoDB v38. SFS for all sites, 4-fold and 0-fold degenerate sites were estimated separately in mainland Tanzania and Zanzibar isolates in non-overlapping 100 kb bins across the core genome. Values of sequence diversity (theta_pi) and Tajima's *D* were estimated for these bin-wise SFS using *spy* $\sum$ *marize* (https://github.com/IDEELResearch/sfspy), and confidence intervals obtained by nonparametric bootstrap. F_{st} was calculated from the joint SFS between mainland Tanzania and Zanzibar. The distribution of local F_{st} values was calculated in 5 kb bins for purposes of visualization only.

Strain deconvolution and inheritance-by-descent analyses

Complexity of infection (COI) and strain deconvolution (phasing) were performed jointly using *dEplo* v0.6-beta [14]. These analyses were limited to 125 isolates from mainland Tanzania and Zanzibar (57 new in this paper and 68 previously published). On the basis of the analyses shown in Figs 1 and 2, these isolates appeared to constitute a reasonably homogeneous population, so the set of 125 was used for determination of PLAFs to be used as priors for the phasing algorithm. Phasing was performed using population allele frequencies as priors in the absence of an external reference panel known to be well-matched for ancestry. The analysis was further limited to very high-confidence sites: VQSLOD > 8, 75% of isolates having GQ ≥ 10 and DP ≥ 5, ≥ 10 bp from the nearest indel (in the raw callset), ≥ 10 total reads supporting the non-reference allele, and PLMAF ≥ 1%. The *dEplo* algorithm was run in "-noPanel" mode with isolate-specific dispersion parameters ("-c") set to the median coverage in the core genome, and default parameters otherwise. Within-isolate IBD segments were extracted from the *dEplo* HMM decodings by identifying runs of sites with probability ≥ 0.90 assigned to hidden states where at least two of the deconvoluted haplotypes were IBD. The total proportion of strain genomes shared IBD (within-isolate F_{IBD}) for isolates with COI > 1 was obtained directly from *dEplo* log files, and agreed closely with the sum of within-isolate IBD segment lengths.

Between-isolate IBD segments were identified by applying *ref* ∈ *edIBD* v12Jul18 [27] to the phased haplotypes produced by *dEplo*. For a genetic map, a constant recombination rate of 6.44 x 10^{-5} cM/bp (equal to the total genetic length of the *P. falciparum* map divided by the physical size of the autosomes in the 3D7 assembly) was assumed. Segments >2 cM were retained for analysis. The proportion of the genome shared IBD between phased haplotypes (between-isolate F_{IBD}) was estimated by maximum likelihood described in [28] using *vcfdoibd*.

Demographic inference

Curves of recent historical effective population size were estimated from between-isolate IBD segments with *IBDNe* v07May18-6a4 [29] using length threshold > 3 cM, 20 bootstrap replicates and default parameters otherwise. Local age-adjusted parasite prevalence point estimates ($Pf\mathrm{PR}_{2\text{-}10}$) and credible intervals were obtained from the Malaria Atlas Project [30] via the R package *malariaAtlas* [31].

More remote population-size histories were estimated with *smc++* v1.15.2 [32]. Phased haplotypes from *dEplo* were randomly combined into diploids and parameters estimated separately for mainland Tanzania and Zanzibar populations using 5-fold cross-validation via command *smc++ cv*, with mutation rate set to 10^{-9} bp^{-1} gen^{-1}. Marginal histories from each population were then used to estimate split times using *smc++ split*.

Analyses of natural selection

The distribution of fitness effects (DFE) was estimated within mainland Tanzania and Zanzibar populations with *polyDFE* v2.0 using 4-fold degenerate sites as putatively-neutral and 0-fold degenerate sites as putatively-selected [33]. "Model C" in *polyDFE* parlance – a mixture of a gamma distribution on selection coefficients of deleterious mutations and an exponential distribution for beneficial mutations – was chosen because it does not require *a priori* definition of discrete bins for selection coefficients, and the gamma distribution can accommodate a broad range of shapes for the DFE of deleterious mutations (expected to represent the bulk of polymorphic sites.) Confidence intervals for model parameters were obtained by non-parametric bootstrap via 20 rounds of resampling over the 100 kb blocks of the input SFS. Because *polyDFE* fits nuisance parameters for each bin in the SFS, computation time increased and numerical stability decreased for SFS with larger sample sizes. Input SFS were, therefore, smoothed and rescaled to pre-specified sample size of 10 chromosomes each using an empirical-Bayes-like method (https://github.com/CartwrightLab/SoFoS/) re-implemented in *spysm∞th*. Smoothing of input SFS had very modest qualitative effect on the resulting DFE.

The cross-population extended haplotype homozygosity (XP-EHH) statistic was used to identify candidate loci for local adaptation in mainland Tanzania or Zanzibar. Because the statistic requires phased haplotypes and is potentially sensitive to phase-switch errors, only isolates with COI = 1 were used (n = 18 mainland Tanzania, n = 12 Zanzibar.) XP-EHH was calculated from haploid genotypes at a subset of 103,982 biallelic SNVs polymorphic among monoclonal isolates with the *xpehhb* ∈ utility of *hapb* ∈ v1.3.0-12-gdb383ad [34]. Raw values were standardized to have zero mean and unit variance; the resulting z-scores are known to have an approximately normal distribution [35] so nominal p-values were assigned from the standard normal distribution. The Benjamini-Hochberg method was used to adjust nominal p-values for multiple testing.

Pipelines used for WGS read alignment, variant calling, variant filtering, haplotype deconvolution and SFS estimation are available on Github: https://github.com/IDEELResearch/NGS_Align_QC_Pipelines.

Results

WGS and variant discovery

Genomic data for *P. falciparum* was generated using leukodepleted blood collected from 43 subjects from Yombo, Tanzania ("mainland") and from DBS collected from 63 subjects from the Zanzibar archipelago ("Zanzibar"; Fig. 1A) using selective whole-genome amplification (sWGA) followed by Illumina sequencing. Thirty-six isolates (84%) from the mainland and 21 isolates (33%) from Zanzibar yielded sufficient data for analysis. These 57 genomes were combined with an additional 68 published genomes from other sites in Tanzania in the MalariaGEN *P. falciparum* Community Project (PfCP) and 179 genomes from other sites in Africa and Asia, representing a broad geographic sampling of Africa and Asia [36]. Single-nucleotide variants (SNVs) were ascertained jointly in the global cohort. After stringent quality control on 1.3 million putative variant sites, a total of 387,646 biallelic SNVs in the "core genome" – the 20.7 Mb of the 3D7 reference assembly lying outside hypervariable regions and accessible by short-read sequencing [22] – were retained for further analysis. The frequency spectrum was dominated by rare alleles: 151,664 alleles (39.1%) were singletons and 310,951 (80.2%) were present in <1% of isolates in the dataset. Ancestral and derived states at 361,049 sites (93.1%) were assigned by comparison to the *P. reichenowi* (CDC strain) genome, treating the *reichenowi* allele as ancestral. Similar biases were observed in the mutational spectrum as have been estimated directly from mutation-accumulation experiments [37]: transitions are more common transversions (Ti:Tv = 1.12; previous estimate 1.13), with a large excess of G:C > A:T changes even after normalizing for sequence composition (Supplementary Fig. 1). Consistency in the mutational spectrum between independent studies, using different methods for sample preparation and different bioinformatics pipelines, supports the accuracy of genotype calls.

Ancestry of mainland Tanzania and Zanzibar isolates

In order to place new isolates in the context of global genetic variation in *P. falciparum*, principal components analysis (PCA) was performed with existing isolates from around the globe (Fig. 1B). A subset of 7,122 stringently-filtered sites with PLMAF > 5% (see Methods) were retained for PCA to minimize distortion of axes of genetic variation by rare alleles or missing data. Consistent with existing literature, isolates separated into three broad clusters corresponding to southeast Asia, east Africa and west Africa. Mainland Tanzania and Zanzibar isolates fell in the east Africa cluster. This observation was formalized using f_3 statistics [38,39], which measure shared genetic variation in a pair of focal populations *A* and *B* relative to an outgroup population *O*. By calculating f_3 across different combinations of comparator populations and holding the outgroup fixed, one can build up an idea of the ancestry of the populations of interest: pairs with relatively larger positive values of f_3 are more genetically similar than pairs with relatively smaller f_3. The new isolates from Yombo and Zanzibar and published Tanzanian isolates shared mutually greater genetic affinity for each other than for other populations in the panel (Fig. 1C-E); isolates from neighbouring countries Malawi and Kenya were next-closest. Together these analyses support an east African origin for parasites in mainland Tanzania and in Zanzibar.

Genetic diversity and differentiation

In order to better understand the population demography and effects of natural selection in the parasite populations, indices of genetic diversity within populations, and the degree to which that diversity is shared across populations, were examined. The genome was partitioned into four sequence classes – all sites in the core genome; 4-fold degenerate ("synonymous") sites; 0-fold degenerate ("nonsynonymous") sites; and coding sites in genes associated with resistance to antimalarial drugs – and several estimators of sequence diversity were calculated in each class (see Methods). Levels of sequence diversity at synonymous (putatively neutral) sites were very similar within mainland Tanzania and Zanzibar isolates (theta_pi = 9.0 x 10^{-4} [95% CI 8.6 x 10^{-4} – 9.4 x 10^{-4}] *vs* 8.4 [95% CI 8.0 x 10^{-4} – 8.7 x 10^{-4} per site) and 1.3-fold lower than among previously-published Tanzanian isolates (Fig. 2A). As expected, diversity was lower at non-synonymous sites, which are more likely to be under purifying selection. Tajima's *D* took negative values in all three populations and across all sites classes (Fig. 2B); demographic explanations for this pattern are investigated later in the manuscript. Minimal evidence was found for differentiation between parasites in mainland Tanzania and Zanzibar. Genome-wide F_{st} was just 0.0289 (95% bootstrap CI 0.0280 – 0.0297); the distribution of F_{st} in 5 kb windows is shown in Fig. 2C. For comparison, genome-wide F_{st} between southeast Asian and African isolates is on the order of 0.20 [23]. Thus there exists minimal evidence for genetic differentiation between parasites in mainland Tanzania and Zanzibar.

Patterns of relatedness and inbreeding

Long segments of the genome shared identical by descent (IBD) – that is, inherited intact from the same recent common ancestor – provide a powerful and fine-grained view of relationships in the recent past. Recent methodological innovations [14] allow estimation of complexity of infection (COI) – the number of distinct parasite strains in a single infection – and simultaneous deconvolution the component haplotypes. The F_{ws} statistic, an index of within-host diversity that is conceptually similar to traditional inbreeding coefficients, was also calculated for comparison [23]. Approximately half of isolates had COI = 1 ("clonal") and half had COI > 1 ("polyclonal" or "mixed") in both populations, and the distribution of COI was similar between the mainland and Zanzibar (chi squared = 0.27 on 2 df, p = 0.87; Supplemental Table 4). Ordinal trends in F_{ws} were qualitatively consistent with COI but show marked variation for COI > 1 (Fig. 3A). Phased haplotypes were used to identify segments shared IBD between isolates and, in the case of mixed infections, within isolates. This revealed substantial relatedness between infecting lineages within mixed isolates (Fig. 3B): the median fraction of the genome shared IBD (F_{IBD}) within isolates was 0.22 among mainland and 0.24 among Zanzibar isolates, with no significant difference between populations (Wilcoxon rank-sum test, p = 0.19). The expected sharing is 0.50 for full siblings and 0.25 for half-siblings with unrelated parents [40]. F_{IBD} was then estimated between all pairs of phased haplotypes. F_{IBD} between pairs of *isolates* was then defined as the maximum over the values for all combinations of haplotypes inferred from the isolates (Fig. 3C). As expected, most pairs were effectively unrelated (median F_{IBD} <= 0.001, on the boundary of the parameter space), but a substantial fraction were related at the level of half-siblings or closer (F_{IBD} > 0.25, 4.0% of all pairs), including 1.3% of mainland-Zanzibar pairs.

Long segments of the genome are shared IBD both within and between isolates. Mean within-isolate segment length was 5.7 cM (95% CI 4.1 – 7.3 cM, *n* = 117) on the mainland and 3.7 cM (95% CI 2.8 – 4.6 cM, *n* = 80) on Zanzibar in a linear mixed model with individual-level random effects; the full distributions are shown in Fig. 3D. Segments shared between isolates within the mainland population (6.2 cM, 95% CI 5.9 – 6.6 cM, *n* = 3279) were longer than segments shared within Zanzibar (4.5 cM, 95% 4.1 – 4.8 cM, *n* = 592) or between mainland and Zanzibar populations (4.1 cM, 95% CI 3.9 – 4.3 cM, *n* = 6506). After accounting for differences in segment length by population, difference in lengths of IBD segments detected between versus within individuals are not significant (mean difference -0.038 cM, 95% CI -0.10 – 0.023 cM). In a random-mating population the length of a segment shared IBD between a pair of individuals with last common ancestor *G* generations in the past is exponentially-distributed with mean 100/(2*G) cM. The shared haplotypes that observed, with length on the order of 5 cM, are thus consistent with shared ancestry in the past 10 generations – although as many as half of such segments probably date back at least 20 generations [41]. In the presence of inbreeding, IBD sharing persists even longer in time.

Close relationships between isolates from the archipelago and the mainland suggest recent genetic exchange. A threshold of $F_{IBD} > 0.25$ (half-siblings) was chosen because it implies that two isolates shared at least one common parent in the last outcrossing generation and, therefore, are related as recently as the last 1-2 transmission cycles, depending on background population dynamics. In principle this could result from importation of either insect vectors or human hosts. To investigate the latter possibility, a travel-history questionnaire completed by subjects from Zanzibar was used. Nine subjects reported travel to the mainland in the month before study enrollment; their destinations are shown in Fig. 4A. Ten pairs with $F_{IBD} > 0.25$ (marked by orange triangles in histogram in Fig. 4B) were identified; all involved a single Zanzibar isolate from a patient who travelled to the coastal town of Mtwara (orange arc in Fig. 4A). It is very likely that this individual represents an imported case. Overall, isolates from travellers had slightly higher mean pairwise relatedness to isolates from the mainland (mean F_{IBD} = 0.0020, 95% CI 0.0018 – 0.0021) than did isolates from non-travellers (mean F_{IBD} = 0.0015, 95% CI 0.0014 – 0.0016; Wilcoxon rank-sum test $p = 1.8 \times 10^{-12}$ for difference). But these relationships – spanning 10 or more outcrossing generations – are far too remote to be attributed to the period covered by the travel questionnaire. The pattern likely represents instead the presence of subtle population structure within Zanzibar.

Demographic history of parasite populations

The distribution of IBD segment lengths carries information about the trajectory of effective population size in the recent past, up to a few hundred generations before the time of sampling. The site frequency spectrum and patterns of fine-scale linkage disequilibrium carry information about the more remote past. Complementary methods were used to infer recent and remote population demography from phased haplotypes. First, a non-parametric method was applied [29] to infer recent effective population size (N_e) from IBD segment lengths separately in mainland Tanzania and Zanzibar populations (Fig. 5A). The

method infers a gradual decline of several orders of magnitude in N_e over the past 100 generations to a nadir at N_e ~= 5,000 around 15-20 outcrossing generations before the time of sampling. Although the confidence intervals are wide, similar trajectories are inferred in all three populations (Zanzibar, new mainland Tanzania isolates and published Tanzanian isolates).

Second, more remote population size histories were inferred jointly for mainland Tanzania and Zanzibar and used to estimate the split time between these populations using a sequentially Markovian coalescent method [32]. This family of models has good resolution for relatively remote events, but less precision in the recent past than models based on IBD segments. The result (Fig. 5B) supports a common ancestral population with N_e ~= 10^5 individuals that underwent a sharp bottleneck followed by rapid growth around 50,000 generations before the present. The time at which the mainland and Zanzibar populations diverged could not be estimated precisely and may have been as recent as 50 or as ancient as 50,000 generations before the present. Trends in N_e were compared to local trends in parasite prevalence from the Malaria Atlas Project [30] (Fig. 5C). Assuming an interval of approximately 12 months per outcrossing generation [42], the contraction in N_e may correspond in time to the decrease in prevalence brought about by infection-control measures over the past two decades.

Natural selection and adaptation

Finally, several approaches were taken to characterize the effects of natural selection on sequence variation in mainland and Zanzibar populations. The fate of a new mutation – whether it spreads and ultimately becomes fixed, or is lost – is determined by its selection coefficient (s), scaled by the effective population size (N_e). The distribution of fitness effects (DFE) describes the distribution of s and can be estimated from the frequency spectrum at putatively-neutral (synonymous) and putatively-selected (non-synonymous) sites (Fig. 6A). Building on previous work in other organisms, the DFE was modelled in each population as a mixture of a gamma distribution (for deleterious mutations, $N_e s < 0$) and an exponential distribution (for beneficial mutations, $N_e s > 0$) [33]. The inference was performed using both the raw SFS and a smoothed representation of the SFS that is more numerically stable and found that results to be similar with both methods. Fitted parameter values are provided in Supplementary Table 5, but the discretized representation of the DFE is more amenable to qualitative comparisons (Fig. 6B).

Differences in the DFE between mainland Tanzania and Zanzibar populations are not statistically significant. The great majority of new mutations (mainland: 74%; Zanzibar: 76%) are expected to be very weakly deleterious ($-0.01 < 4N_e s < 0$), and only a small minority are expected to be beneficial ($4N_e s > 0$) (mainland: 4.5% [95% CI 2.7 – 29%]; Zanzibar: 2.4% [95% CI 0.56 – 50%]). The DFE also allows us to estimate that 8.8% (mainland) and 5.2% (Zanzibar) of substitutions since the common ancestor with *P. reichenowi* have been fixed by positive selection; this quantity is known in some contexts as the “rate of adaptive evolution.”

Although the DFE tells us the proportion of polymorphic sites under positive selection, it does not pinpoint which sites those are. To identify signals of recent, population-specific positive selection, the XP-EHH

statistic between mainland and Zanzibarian isolates were used [35]. Outliers in the XP-EHH scan, defined as standardized XP-EHH scores above the 99.9th percentile, represent candidates for local adaptation (Supplementary Fig. 2). One-hundred four biallelic SNPs in 20 distinct genes passed this threshold (Supplementary Table 6). None of these have been associated with resistance to anti-malarial drugs – an important form of local adaptation in this species – but one (PF3D7_0412300) has been identified in a previous selection scan [43]. Prevalences of 54 known drug-resistance alleles are shown in Supplementary Table 7 and are similar to previous reports in East Africa [44–46]. None of these loci had F_{st} > 0.05 between mainland Tanzania and Zanzibar.

Discussion

Zanzibar has been the target of intensive malaria control interventions for nearly two decades following the early implementation of ACT therapies in 2003 [2]. Despite sustained vector control practices and broad access to rapid testing and effective treatment, malaria has not been eliminated from the archipelago [2]. Here WGS of *P. falciparum* isolates from Zanzibar and nearby sites on the mainland was used to investigate ancestry, population structure and transmission in local parasite populations. These data place Tanzanian parasites in a group of east African populations with broadly similar ancestry and level of sequence diversity. There was minimal genome-wide signal of differentiation between mainland and Zanzibar isolates.

The most parsimonious explanation for these findings is a source-sink scenario, similar to a previous report in Namibia [47], in which importation of malaria from a region of high but heterogeneous transmission (the mainland) is inhibiting malaria elimination in a pre-elimination area (Zanzibar). Using WGS it is shown that the parasite population on the islands remains genetically almost indistinguishable from regions on the mainland of Tanzania. Numerous long haplotypes could be identified that are shared between the populations, on the order of 5 cM, suggesting that genetic exchange between the populations has occurred within the last 10-20 sexual generations. In addition, a Zanzibar isolate is identified that is related at the half-sibling level to a group of mutually-related mainland isolates. This likely represents an imported case and provides direct evidence for recent, and likely ongoing, genetic exchange between the archipelago and the mainland. These observations suggest that parasite movement from the mainland to the archipelago is appreciable and may be a significant hurdle to reaching elimination.

Human migration is critical in the spread of malaria [48], thus the most likely source for importation of parasites into Zanzibar is through human travel to high-risk malaria regions. Multiple studies have been conducted on travel patterns of Zanzibarian residents as it relates to importation of malaria [49–51], one of which estimated that there are 1.6 incoming infections per 1000 inhabitants per year. This is also in accordance with the estimate of about 1.5 imported new infections out of a total of 8 per 1000 inhabitants in a recent epidemiological study [2]. None of these studies have leveraged parasite population genetics to understand importation patterns. Though this study is small, the findings are proof of principle for using genetics to identify specific importation events. These data provides a platform for

future genetic surveillance efforts by, for example, design of targeted assays for sequence variants that discriminate mainland from Zanzibari parasites. Such surveillance, including of asymptomatic individuals, would clarify the role of importation *versus* endemic transmission and potentially identify specific travel corridors to target for interventions. Larger sample sizes would also likely begin to reveal subtle population structure that is not obvious when examining a few dozen isolates.

Malarial infections in Africa are highly polyclonal. This within-host diversity poses technical challenges but also provides information on transmission dynamics. Approximately half of isolates from both the mainland and Zanzibar represent mixed infections (COI > 1), similar to estimates in Malawian parasites with similar ancestry [15]. It is clear that a widely-used heuristic index (F_{ws}) is qualitatively consistent with COI estimated by haplotype deconvolution [52], but has limited discriminatory power in the presence of related lineages in the same host. Furthermore, median within-host relatedness (F_{IBD}) is ~0.25, the expected level for half-siblings, in both mainland and Zanzibar populations. This strongly suggests frequent co-transmission of related parasites in both populations [40]. Estimates of F_{IBD} are within the range of estimates from other African populations and add to growing evidence that mixed infections may be predominantly due to co-transmission rather than superinfection even in high-transmission settings [53,54]. An important caveat of this work is its dependence on statistical haplotype deconvolution. Direct comparison of statistical deconvolution to direct sequencing of single clones has shown that methods like *dEplo* have limited accuracy for phasing the minority haplotype(s) in a mixed infection. Phasing errors tend to limit power to detect IBD between infections, and may cause underestimation of between-host relatedness.

Intensive malaria surveillance over the past several decades provides an opportunity to compare observed epidemiological trends to parasite demographic histories estimated from contemporary genetic data. Estimates of historical effective population size (N_e) support an ancestral population of approximately 10^5 individuals that grew rapidly around 10^4 generations ago, then underwent sharp contraction within the past 100 generations to a nadir around 10-20 generations before the present. Stable estimates of the split time between the mainland and Zanzibar populations could not be obtained, either with a coalescent-based method (Fig. 5B) or with method based on the diffusion approximation to the Wright-Fisher process [55]. This is not surprising given that the shape of joint site frequency spectrum (Supplementary Fig. 3), summarized in low F_{st} genome-wide, is consistent with near-panmixia. The timing and strength of the recent bottleneck appears similar in mainland Tanzania and Zanzibar isolates and coincides with a decline in the prevalence of parasitemia. However, it should be remembered that the relationship between genetic and census population size – for which prevalence is a proxy – is complex, and other explanations may exist for the observed trends.

Finally, this paper makes the first estimates of the distribution of fitness effects (DFE) in *P. falciparum*. Although the impact of selection on genetic diversity in this species has long been of interest in the field, previous work has tended to focus on positive selection associated with resistance to disease-control interventions. The DFE is a more fundamental construct that has wide-ranging consequences for the

evolutionary trajectory of a population and the genetic architecture of phenotypic variation [56]. Purifying selection is pervasive, but most new alleles (~75%) are expected to have sufficiently small selection coefficients that their fate will be governed by drift. The proportion of new mutations expected to be beneficial – the “target size” for adaption– is small, on the order 1-2%. Together these observations imply that even in the presence of ongoing human interventions, patterns of genetic variation in the Tanzanian parasite population are largely the result of drift and purifying selection rather than positive selection. It should be noted that these conclusions are based on the core genome and may not hold for hypervariable loci thought to be under strong selection such as erythrocyte surface antigens. Furthermore, the complex lifecycle of *Plasmodium* species also departs in important ways from the assumptions of classical population-genetic models [57]. The qualitative impact of these departures conclusions is hard to determine.

Conclusion

The elimination of malaria from Zanzibar has been a goal for many years. This paper pesents genomic evidence of continued recent importation of *P. falciparum* from mainland Tanzania to the archipelago. Reducing this importation is likely to be an important component of reaching elimination. Investigation of approaches to limit importation, such as screening of travellers or mass drug treatment, is needed. However, the high degree of connectivity between the mainland and the Zanzibar archipelago will make this challenging. It is encouraging that parasite populations in the region appear to be contracting (Fig. 5). These declines are likely due to decreasing transmission but nonetheless need to be interpreted with caution, as they may also be due to other factors that impact effective population size estimates, including violation of model assumptions. The data suggests that larger studies of the relationship between Zanzibarian and mainland parasites will enable further more precise estimates of corridors of importation based on parasite genetics. Genomic epidemiology has the potential to supplement traditional epidemiologic studies in Zanzibar and to aid efforts to achieve malaria elimination on the archipelago.

List Of Abbreviations

AL: artemether-lumefantrine

ASAQ: artesunate-amodiaquine

cM: centimorgan

COI: complexity of infection

DBS: dried blood spots

DFE: distribution of fitness effects

F_{IBD}: median fraction of the genome shared IBD

IBD: identity by descent

N_e: effective population size

PCA: principal components analysis

PfCP: Pf Community Project

PLAF: population-level allele frequency

PLMAF: population-level minor allele frequency

SFS: site-frequency spectrum

SLDP: single low dose primaquine

SNV: single nucleotide variant

sWGA: selective whole genome amplification

VQSR: Variant Quality Score Recalibration

WGS: whole genome sequencing

WSAF: within-sample allele frequency

Declarations

ETHICAL APPROVALS AND CONSENT TO PARTICIPATE

This analysis was approved by the IRBs at the University of North Carolina at Chapel Hill, Muhimbili University of Health and Allied Sciences (MUHAS), Zanzibar Medical Research Ethical Committee and the Regional Ethics Review Board, Stockholm, Sweden.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIAL

Sequencing reads were deposited into the NCBI SRA (Accession numbers: pending). Code is available through GitHub (https://github.com/IDEELResearch). This publication uses data from the MalariaGEN *P. falciparum* Community Project (www.malariagen.net/projects/p-falciparum-community-project) as described in [36]. Genome sequencing was performed by the Wellcome Trust Sanger Institute and the

Community Projects is coordinated by the MalariaGEN Resource Centre with funding from the Wellcome Trust (098051, 090770). This publication uses data generated by the Pf3k project (www.malariagen.net/pf3k) which became open access in September 2016.

COMPETING INTERESTS

The authors have no competing interests to declare.

FUNDING

This research was funded by the National Institutes of Health, grants R01AI121558, F30AI143172 (NFB), F30MH103925 (APM), and K24AI134990. Funding was also contributed from the Swedish Research Council and Erling-Persson Family Foundation.

AUTHOR CONTRIBUTIONS

APM, NFB and JBP designed experiments, conducted analysis and wrote the manuscript. BN, EL, MM, and UM collected samples and participated in manuscript preparation. MD conducted laboratory work and participated in manuscript preparation. DLF helped develop software and participated in manuscript preparation. JAB, AM, AB and JJJ helped conceive the study, contributed to the experimental design and wrote the manuscript.

ACKNOWLEDGEMENTS

We would like to thank the communities and participants who took part in these studies. We would also like to thank Molly Deutsch-Feldman for helping to optimize the sWGA protocol.

References

1. WHO. World Malaria Report 2018. Geneva, World Health Organization, 2019.
2. Björkman A, Shakely D, Ali AS, Morris U, Mkali H, Abbas AK, et al. From high to low malaria transmission in Zanzibar-challenges and opportunities to achieve elimination. BMC Med. 2019;17:14.
3. Björkman A, Cook J, Sturrock H, Msellem M, Ali A, Xu W, et al. Spatial distribution of falciparum malaria infections in Zanzibar: implications for focal drug administration strategies targeting asymptomatic parasite carriers. Clin Infect Dis. 2017;64:1236–43.
4. Bousema JT, Gouagna LC, Drakeley CJ, Meutstege AM, Okech BA, Akim INJ, et al. *Plasmodium falciparum* gametocyte carriage in asymptomatic children in western Kenya. Malar J. 2004;3:18.
5. Mawili-Mboumba DP, Nikiéma R, Bouyou-Akotet MK, Bahamontes-Rosa N, Traoré A, Kombila M. Sub-microscopic gametocyte carriage in febrile children living in different areas of Gabon. Malar J. 2013;12:375.

6. 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. Nat Commun. 2012;3:1237.
7. The malERA Consultative Group on Diagnoses and Diagnostics. A research agenda for malaria eradication: diagnoses and diagnostics. PLoS Med. 2011;8:e1000396.
8. Neafsey DE, Volkman SK. Malaria genomics in the era of eradication. Cold Spring Harb Perspect Med. 2017;7:a025544.
9. Wesolowski A, Taylor AR, Chang H-H, Verity R, Tessema S, Bailey JA, et al. Mapping malaria by combining parasite genomic and epidemiologic data. BMC Med. 2018;16:190.
10. Chang H-H, Park DJ, Galinsky KJ, Schaffner SF, Ndiaye D, Ndir O, et al. Genomic sequencing of *Plasmodium falciparum* malaria parasites from Senegal reveals the demographic history of the population. Mol Biol Evol. 2012;29:3427–39.
11. 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 Genet. 2017;13:e1007065.
12. Henden L, Lee S, Mueller I, Barry A, Bahlo M. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. PLoS Genet. 2018;14:e1007279.
13. Schaffner SF, Taylor AR, Wong W, Wirth DF, Neafsey DE. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. Malar J. 2018;17:196.
14. Zhu SJ, Almagro-Garcia J, McVean G. Deconvolution of multiple infections in *Plasmodium falciparum* from high throughput sequencing data. Bioinformatics. 2018;34:9–15.
15. Zhu SJ, Hendry JA, Almagro-Garcia J, Pearson RD, Amato R, Miles A, et al. The origins and relatedness structure of mixed infections vary with local prevalence of *P. falciparum* malaria [Internet]. bioRxiv. 2019 [cited 2019 Jun 12]. p. 387266. Available from: https://www.biorxiv.org/content/10.1101/387266v4
16. Auburn S, Benavente ED, Miotto O, Pearson RD, Amato R, Grigg MJ, et al. Genomic analysis of a pre-elimination Malaysian *Plasmodium vivax* population reveals selective pressures and changing transmission dynamics. Nat Commun. 2018;9:2585.
17. Daniels RF, Schaffner SF, Wenger EA, Proctor JL, Chang H-H, Wong W, et al. Modeling malaria genomics reveals transmission decline and rebound in Senegal. Proc Natl Acad Sci USA. 2015;112:7067–72.
18. 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. J Infect Dis. 2018;217:622–7.
19. 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. Malar J. 2016;15:597.

20. Sundararaman SA, Plenderleith LJ, Liu W, Loy DE, Learn GH, Li Y, et al. Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. Nat Commun. 2016;7:11078.
21. Clarke EL, Sundararaman SA, Seifert SN, Bushman FD, Hahn BH, Brisson D. swga: a primer design toolkit for selective whole genome amplification. Bioinformatics. 2017;33:2071–7.
22. 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 Res. 2016;26:1288–99.
23. 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;487:375–9.
24. Otto TD, Rayner JC, Böhme U, Pain A, Spottiswoode N, Sanders M, et al. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. Nat Commun. 2014;5:4754.
25. Arthur R, Schulz-Trieglaff O, Cox AJ, O'Connell J. AKT: ancestry and kinship toolkit. Bioinformatics. 2017;33:142–4.
26. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of next generation sequencing data. BMC Bioinformatics. 2014;15:356.
27. Browning BL, Browning SR. Improving the malaria genomics in the era of eradication Genetics. 2013;194:459–71.
28. Verity R, Aydemir O, Brazeau NF, Watson OJ, Hathaway NJ, Mwandagalirwa MK, et al. The impact of antimalarial resistance on the genetic structure of *Plasmodium falciparum* in the DRC [Internet]. bioRxiv. 2019 [cited 2019 Jun 18]. p. 656561. Available from: https://www.biorxiv.org/content/10.1101/656561v1.abstract
29. Browning SR, Browning BL. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. Am J Hum Genet. 2015;97:404–18.
30. Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, et al. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. Nature. 2015;526:207–11.
31. Pfeffer DA, Lucas TCD, May D, Harris J, Rozier J, Twohig KA, et al. malariaAtlas: an R interface to global malariometric data hosted by the Malaria Atlas Project. Malar J. 2018;17:352.
32. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 2017;49:303–9.
33. Tataru P, Mollion M, Glémin S, Bataillon T. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. Genetics. 2017;207:1103–19.
34. Maclean CA, Chue Hong NP, Prendergast JGD. Hapbin: an efficient program for performing haplotype-based scans for positive selection in large genomic datasets. Mol Biol Evol. 2015;32:3027–9.
35. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449:913–8.

36. MalariaGEN *Plasmodium falciparum* Community Project. Genomic epidemiology of artemisinin resistant malaria. Elife. 2016;5:e08714.
37. 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 Res. 2017;45:1889–901.
38. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192:1065–93.
39. Peter BM. Admixture, population structure, and f-statistics. Genetics. 2016;202:1485–501.
40. Wong W, Wenger EA, Hartl DL, Wirth DF. Modeling the genetic relatedness of *Plasmodium falciparum* parasites following meiotic recombination and cotransmission. PLoS Comput Biol. 2018;14:e1005923.
41. Speed D, Balding DJ. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet. 2015;16:33–44.
42. Huber JH, Johnston GL, Greenhouse B, Smith DL, Perkins TA. Quantitative, model-based estimates of variability in the generation and serial intervals of *Plasmodium falciparum* malaria. Malar J. 2016;15:490.
43. Samad H, Coll F, Preston MD, Ocholla H, Fairhurst RM, Clark TG. Imputation-based population genetics analysis of *Plasmodium falciparum* malaria parasites. PLoS Genet. 2015;11:e1005131.
44. Kavishe RA, Kaaya RD, Nag S, Krogsgaard C, Notland JG, Kavishe AA, et al. Molecular monitoring of *Plasmodium falciparum* super-resistance to sulfadoxine-pyrimethamine in Tanzania. Malar J. 2016;15:335.
45. Ngondi JM, Ishengoma DS, Doctor SM, Thwai KL, Keeler C, Mkude S, et al. Surveillance for sulfadoxine-pyrimethamine resistant malaria parasites in the Lake and Southern Zones, Tanzania, using pooling and next-generation sequencing. Malar J. 2017;16:236.
46. Baraka V, Ishengoma DS, Fransis F, Minja DTR, Madebe RA, Ngatunga D, et al. High-level *Plasmodium falciparum* sulfadoxine-pyrimethamine resistance with the concomitant occurrence of septuple haplotype in Tanzania. Malar J. 2015;14:439.
47. Tessema S, Wesolowski A, Chen A, Murphy M, Wilheim J, Mupiri A-R, et al. Using parasite genetic and human mobility data to infer local and cross-border malaria connectivity in Southern Africa. Elife. 2019;8:e43510.
48. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, et al. Quantifying the impact of human mobility on malaria. Science. 2012;338:267–70.
49. Tatem AJ, Qiu Y, Smith DL, Sabot O, Ali AS, Moonen B. The use of mobile phone data for the estimation of the travel patterns and imported *Plasmodium falciparum* rates among Zanzibar residents. Malar J. 2009;8:287.
50. Tatem A, Qiu Y, Smith D, Sabot O, Ali A, Moonen B. Travel patterns and imported *Plasmodium falciparum* rates among Zanzibar residents. In: Hospitality and Health: issues and developments. 2011:58–72. Available from: http://dx.doi.org/10.1201/b12232-9

51. Le Menach A, Tatem AJ, Cohen JM, Hay SI, Randell H, Patil AP, et al. Travel risk, malaria importation and malaria transmission in Zanzibar. Sci Rep. 2011;1:93.
52. O'Brien JD, Amenga-Etego L, Li R. Approaches to estimating inbreeding coefficients in clinical isolates of *Plasmodium falciparum* from genomic sequence data. Malar J. 2016;15:473.
53. Nkhoma Standwell C, Nair S, Cheeseman IH, Rohr-Allegrini C, Singlam S, Nosten F, et al. Close kinship within multiple-genotype malaria parasite infections. Proc Biol Sci. 2012;279:2589–98.
54. Wong W, Griggs AD, Daniels RF, Schaffner SF, Ndiaye D, Bei AK, et al. Genetic relatedness analysis reveals the cotransmission of genetically related *Plasmodium falciparum* parasites in Thiès, Senegal. Genome Med. 2017;9:5.
55. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5:e1000695.
56. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet. 2007;8:610–8.
57. Chang H-H, Moss EL, Park DJ, Ndiaye D, Mboup S, Volkman SK, et al. Malaria life cycle intensifies both natural selection and random genetic drift. Proc Natl Acad Sci USA. 2013;110:20129–34.

Figures

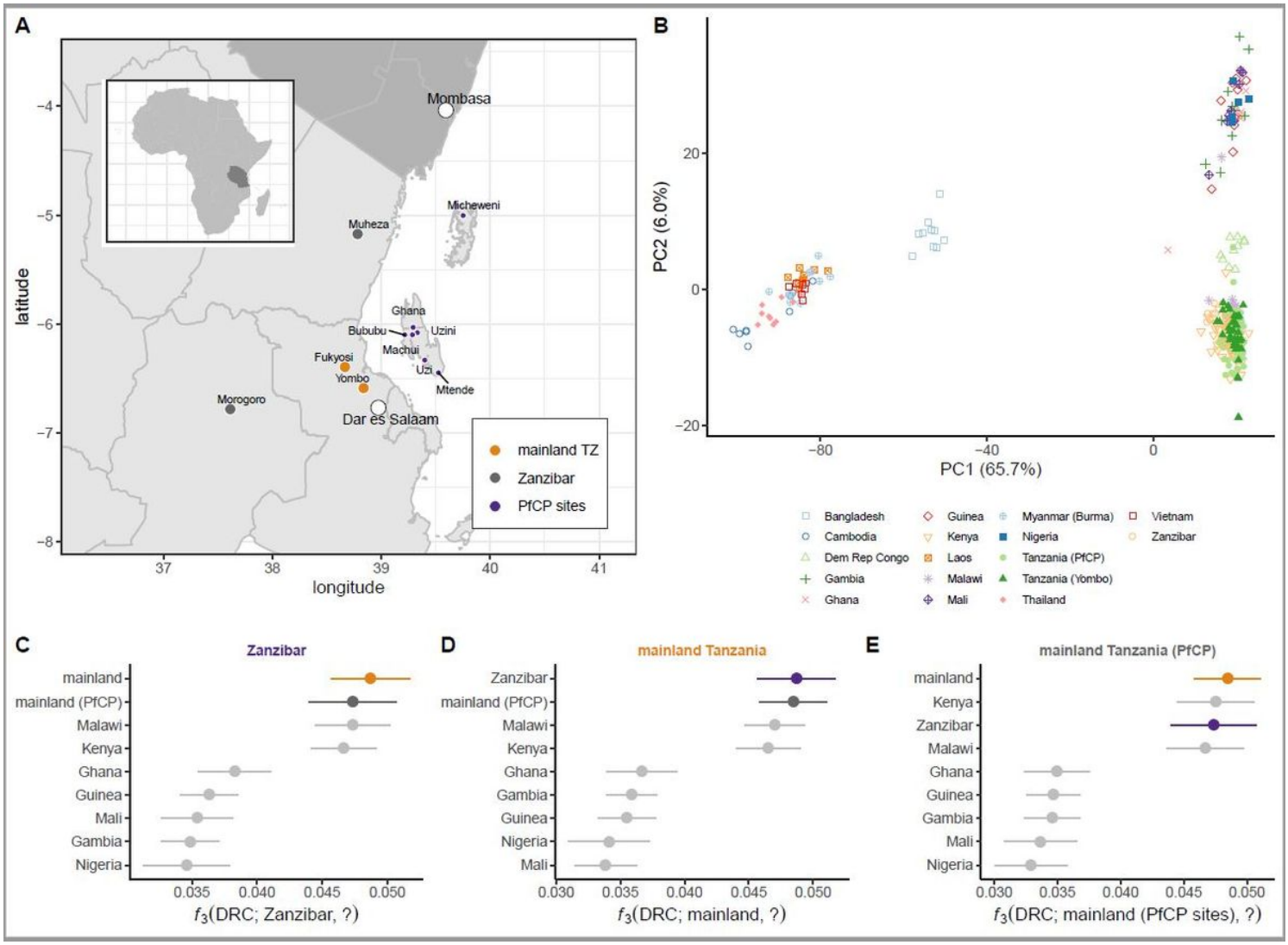

Figure 1

Ancestry of P. falciparum in Zanzibar and mainland Tanzania. (A) Location for samples used in this study, coloured by population: orange, mainland Tanzania; purple, Zanzibar; dark grey, published mainland Tanzania isolates from the MalariaGEN P. falciparum Community Project. Other major regional cities show with open circles. (B) Major axes of genetic differentiation between global P. falciparum populations demonstrated by principal components analysis (PCA) on genotypes at 7,122 SNVs with PLMAF > 5%. Each point represents a single isolate (n = 304) projected onto the top two principal components (71% cumulative variance explained); colour-shape combinations indicate country of origin. (C-E) Population relationships assessed by f3 statistics with focal population indicated at the top of each panel, comparator populations on the vertical axis, and Congolese population as an outgroup. Error bars show 3 times the standard error computed by block-jacknife.

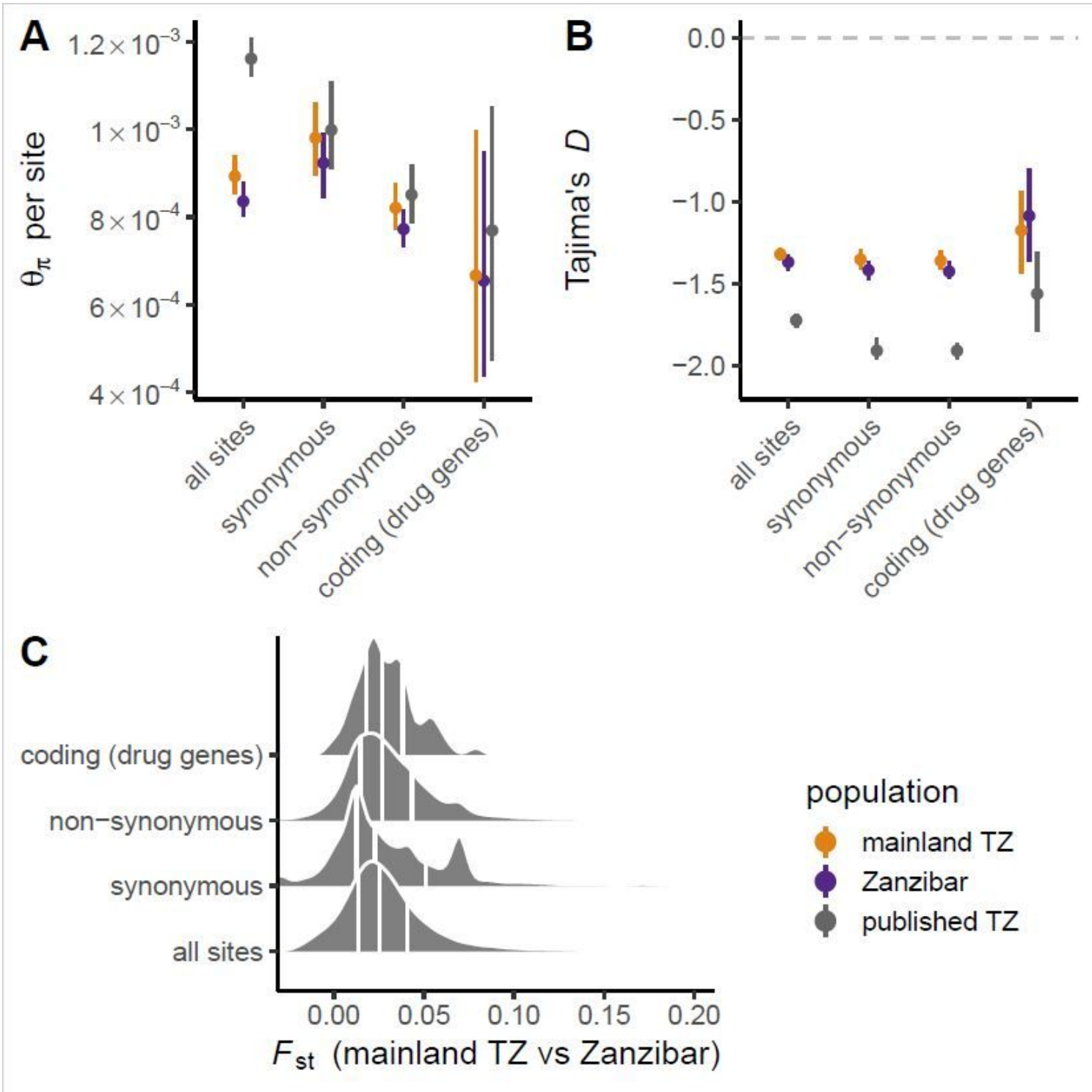

Figure 2

Diversity and differentiation of P. falciparum in mainland Tanzania and Zanzibar. (A) Average pairwise sequence diversity (theta_pi) per base pair in different compartments of the core genome: all sites, 4-fold degenerate ("synonymous") sites, 0-fold degenerate ("non-synonymous") sites, and coding regions of putative drug-resistance genes. Points are coloured by population; error bars give 95% bootstrap CIs. (B) Tajima's D in same classes of sites as in panel A. (C) Distribution of Fst between mainland Tanzania and Zanzibar isolates, calculated in 5 kb windows. Vertical lines mark 25th, 50th and 75th percentiles.

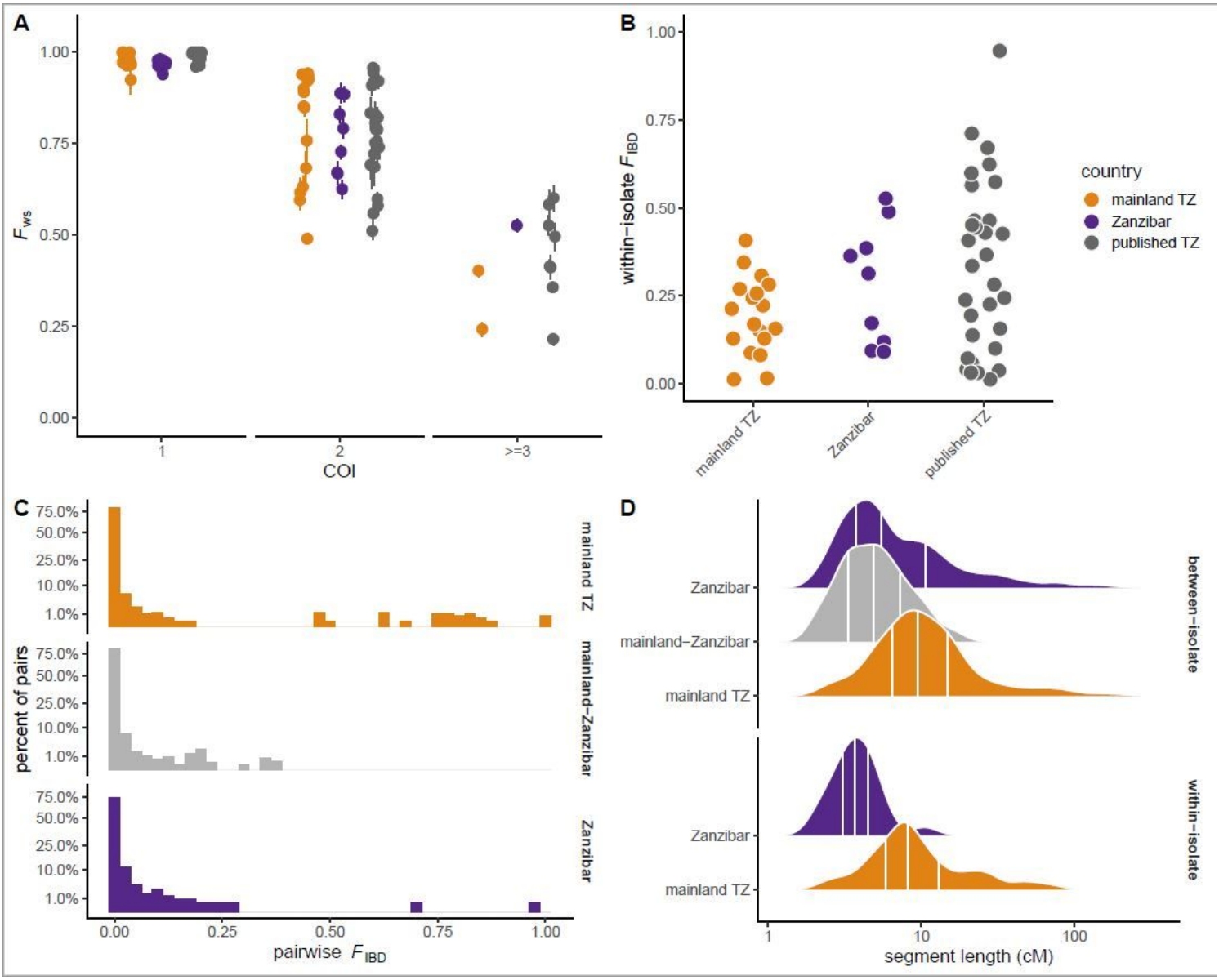

Figure 3

Complexity of infection and patterns of within- and between-host relatedness. (A) The Fws index of within-host diversity, binned by complexity of infection (COI) estimated from genome-wide SNVs. Points coloured by population. (B) Distribution of within-host relatedness, measured as the proportion of the genome shared IBD (FIBD) between strains, for isolates with COI > 1. Note that y-axis is on square-root scale. (C) Distribution of between-host relatedness, calculated from haplotype-level IBD. (D) Distribution of the length of segments shared IBD between (top) or within hosts (bottom). Segment lengths given in centimorgans (cM). Vertical lines mark 25th, 50th and 75th percentiles.

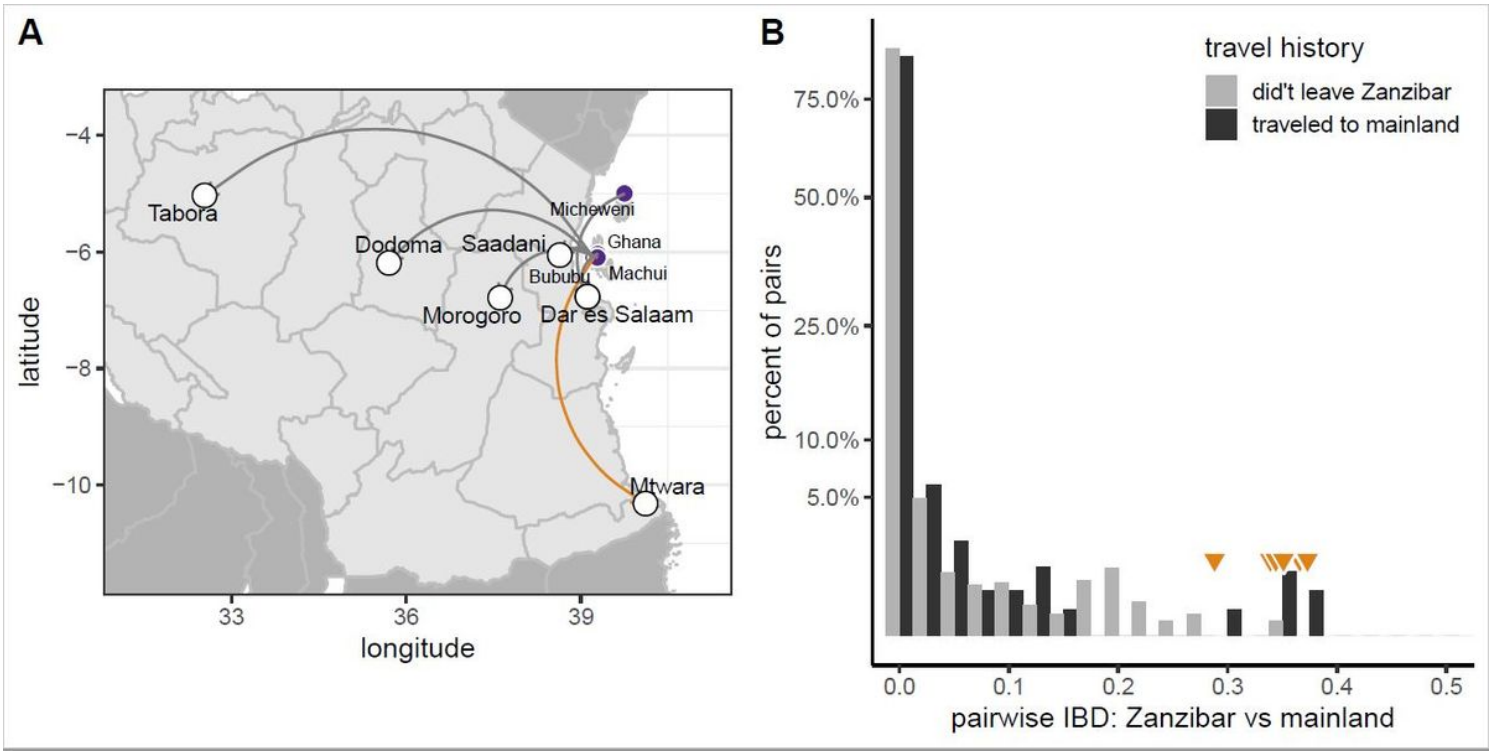

Figure 4

Travel history and parasite relatedness. (A) Reported destinations for 9 residents of Zanzibar who travelled to mainland Tanzania in the month before study enrollment. Orange arc shows destination of suspected imported case. (B) Pairwise IBD sharing between Zanzibar isolates from hosts with recent travel (dark bars) versus non-travellers (light bars). Values > 0.25 highlighted by orange triangles. Note that y-axis is on square-root scale.

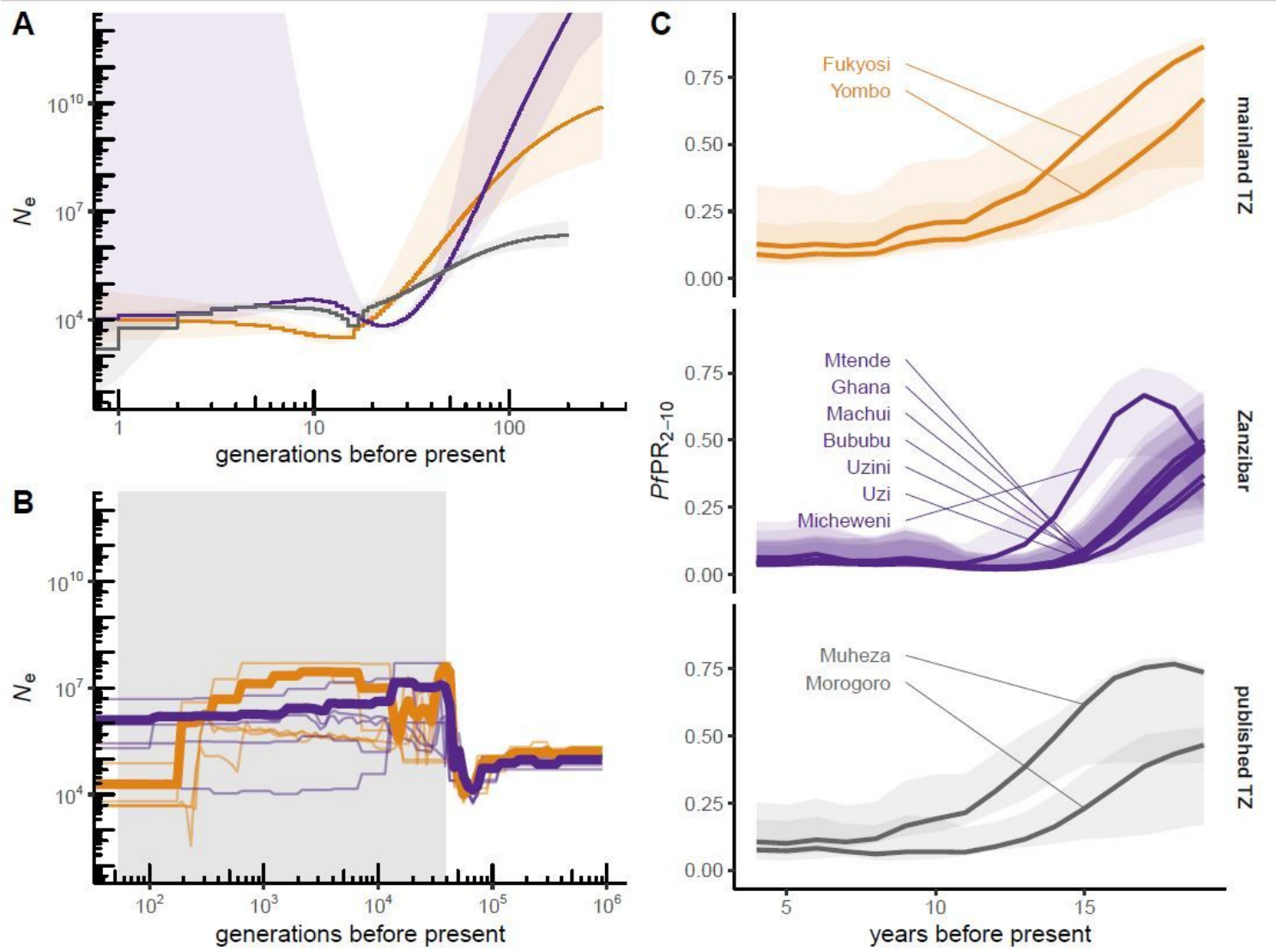

Figure 5

Comparison of historical parasite demography and infection prevalence. (A) Curves of recent historical effective population size (Ne) reconstructed from IBD segments; shaded regions give 95% bootstrap CIs. (B) Effective population size in the more remote past, reconstructed from phased haplotypes. Thin lines, independent model runs; bold lines, model averages (see Methods). Shaded region, range of inferred split times between mainland and Zanzibar populations. Scale of y-axis matches panel A. (C) Estimated prevalence of P. falciparum infection from the Malaria Atlas Project at sampling sites for cohorts (expressed as age-standardized prevalence rate among children aged 2-10 years, PfPR2-10, in cross-sectional surveys); shaded regions give 95% credible intervals. Present = 2019.

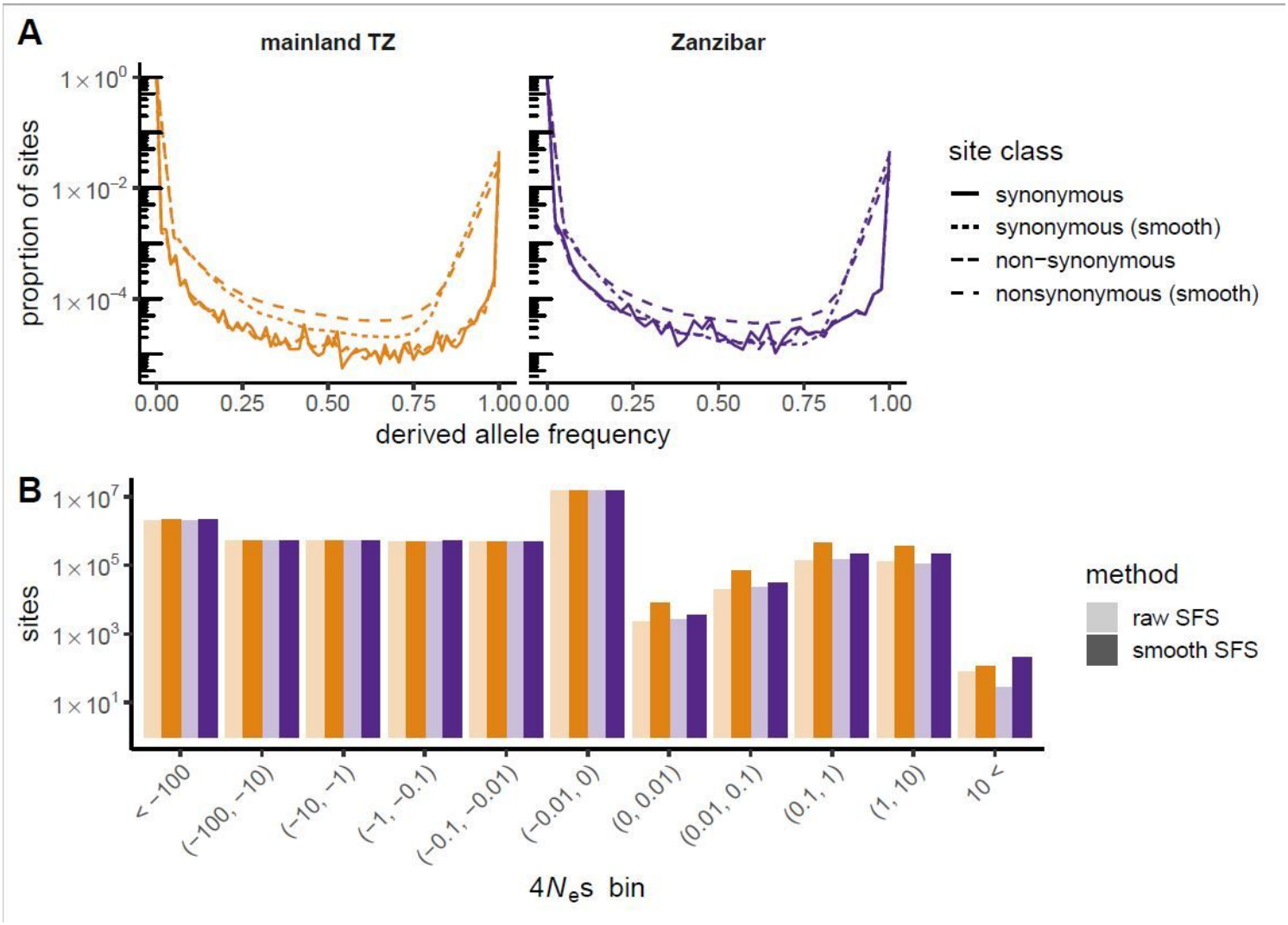

Figure 6

Characterizing the impact of natural selection on sequence variation. (A) Site-frequency spectra for putatively neutral (4-fold degenerate) and putatively-selected (0-fold degenerate) sites. (B) Inferred distribution of population-scaled selection coefficients (4Nes) for each population, shown in discrete bins. Dark bars, estimates from raw SFS; light bars, estimates from smoothed SFS. Note logarithmic scale for vertical axis in both panels.

Supplementary Files

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

- TableS8candidateswgaprimers.xlsx
- Morgan2019MalariaJournalSupplementalMaterial.pdf
- Suppmaterialcomparedocument.pdf
- TableS3publicgenomes.xlsx

- TableS7TZZBdruglociprev.xlsx
- TableS6xpehhgeneGO.xlsx