Original Research

Defining malaria parasite population subdivisions, transmission dynamics and infection origins using SNP barcodes

Short Title: Using SNP barcodes for malaria genomic epidemiology

G.L. Abby Harrison[1,2], Somya Mehra[1,3], Zahra Razook[1,4], Natacha Tessier[1], Stuart Lee[1], Manuel W. Hetzel[5,6], Livingstone Tavul[7], Moses Laman[7], Leo Makita[8], Roberto Amato[9], Olivo Miotto[10,11], Nicholas Burke[12], Anne Jensen[12], Dominic Kwiatkowski[9]†, Inoni Betuela[7], Peter M. Siba[7,13], Melanie Bahlo[1,2], Ivo Mueller[1,2,14] and Alyssa E. Barry[1,2,3,4]*.

1. Division of Population Health and Immunity, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, AUSTRALIA
2. Department of Medical Biology, University of Melbourne, Carlton, Victoria, AUSTRALIA
3. Burnet Institute, Melbourne, AUSTRALIA
4. IMPACT/School of Medicine, Deakin University, Geelong, AUSTRALIA
5. Swiss Tropical and Public Health Institute, Basel, SWITZERLAND
6. University of Basel, Basel, SWITZERLAND
7. Papua New Guinea Institute of Medical Research, Madang, PAPUA NEW GUINEA
8. National Department of Health, Port Moresby, PAPUA NEW GUINEA
9. Wellcome Sanger Institute, Hinxton, UNITED KINGDOM
10. University of Oxford, Oxford, UNITED KINGDOM
11. Mahidol-Oxford Research Unit, Mahidol University, Bangkok, THAILAND
12. Exxon Mobil PNG Ltd, Port Moresby, PAPUA NEW GUINEA
13. Divine Word University, Madang, PAPUA NEW GUINEA
14. Parasites and Insect Vectors, Institut Pasteur, Paris, FRANCE

*Corresponding author:
Centre for Innovation in Infectious Disease and Immunology Research (CIIDIR), IMPACT/School of Medicine, Deakin University, Health Education Research Building B,
299 Ryrie St, Geelong, Victoria, AUSTRALIA 3216
e: a.barry@deakin.edu.au
p: +61 3 52278777

†deceased 27 April 2023

ABSTRACT

The successful implementation of pathogen genomic surveillance demands rapid, low-cost genotyping solutions for tracking infections. Here we demonstrate the capacity of single nucleotide polymorphism (SNP) barcodes to generate practical information for malaria surveillance and control. The study was conducted in Papua New Guinea (PNG), a country with a wide range of malaria transmission intensities. A panel of 191 candidate SNPs was selected from 5786 SNPs with minor allele frequency greater than 0.1, identified amongst 91 *Plasmodium falciparum* genomes from three provinces of PNG. We then genotyped 772 *P. falciparum* isolates from a 2008 nationwide malaria indicator survey and 31 clinical infections from an outbreak of unknown origin. We assessed the performance of SNP panels with different allele frequency characteristics, and measured population diversity, structure and connectivity using both whole genome data and the SNP barcode. The full SNP barcode captured similar patterns of population structure evident with 5786 'whole genome' SNPs. Geographically informative SNPs (iSNPs, F_{ST}>0.05) show increased population clustering compared to the full barcode whilst randomly selected SNPs (rSNPs) and SNPs with similar allele frequencies (F_{ST}<0.05) amongst different countries (universal, uSNPs) or local PNG populations (balanced, bSNPs) indicated little clustering. Applied to samples from all endemic areas of PNG, this barcode identified variable transmission dynamics, and eight major populations. Genetic diversity was high throughout most areas, however, in the southern region, isolates were either closely related, suggesting highly inbred or near-clonal populations; or, they shared ancestry with other parasite populations, consistent with importation. Applied to outbreak samples, only the full barcode, the iSNPs and bSNPs distinguished between locally acquired and imported infections. The full barcode contains more than 100 SNPs prevalent in other endemic regions, allowing the transfer of this tool to other settings. SNP barcodes must be validated in local settings to ensure they capture the diversity and population structure of the target population. Subsets of geographically informative SNPs will be essential for predicting geographic origins but may bias analyses of population structure and gene flow if used alone.

KEYWORDS: malaria, *Plasmodium falciparum*, genomics, single nucleotide polymorphisms, diversity, population genetics, spatial epidemiology, control,

AUTHOR SUMMARY

Pathogen genomic surveillance is a sensitive approach for mapping pathogen transmission dynamics to support decisions about how to prevent and control infectious diseases. High throughput genotyping tools known as single nucleotide polymorphism (SNP) barcodes are used to measure relationships between individuals and connectivity between populations, however, the barcode design may influence these results. We used whole genome sequences from the malaria parasite *Plasmodium falciparum* to design a barcode that captures the diversity both within and between parasite populations of Papua New Guinea (PNG), where transmission is variable amongst provinces. By investigating the performance of different panels of SNPs, we show that validation for use in the target population is crucial to correctly identifying population genetic structure. Applying the validated SNP barcode on hundreds of samples from all endemic provinces of PNG, high levels of variability in local transmission dynamics, and regions of population subdivision were observed. Some geographic areas show evidence of interrupted transmission, and the substantial genetic differentiation between the northern, eastern and island populations, presents an opportunity to design targeted, subnational control efforts. Application of the barcode to outbreak samples classified cases into imported and locally acquired infections, with substantial local transmission indicating control efforts were not sufficient to prevent the spread of infections. SNP barcodes are useful tools that can be used to supplement existing malaria surveillance tools however careful validation of their effectiveness in different settings is recommended.

INTRODUCTION

Pathogen genomics has become an essential tool for public health to detect, track and contain infectious disease threats [1]. Whole genome sequencing (WGS) is a popular approach for small genomes such as those of viruses and some bacteria. However, for pathogens with larger genomes such as the malaria parasite, *Plasmodium falciparum* [2], targeted genotyping approaches are more affordable and scalable for national malaria control programs. Panels of single nucleotide polymorphisms (SNPs) referred as "SNP barcodes" have been increasingly used for malaria genomic surveillance [3-8]. For malaria genomic surveillance to reach its full potential however, assessment of the capabilities and resolution of these genotyping tools is essential.

Malaria, caused by infection with *Plasmodium falciparum* and transmitted by anopheline mosquitoes, remains a major public health problem for over half the world's population, particularly for residents of low- and middle-income countries in tropical and sub-tropical regions [9]. Reductions in malaria transmission, through intensified control efforts, have stagnated since 2016, resulting in a call for new approaches to continue to progress towards elimination, and to prevent resurgence. Genomic surveillance can play a critical role in malaria control and elimination including quantification of local transmission dynamics, population structure and connectivity, and identification of emerging or imported strains that may undermine control efforts [6, 10-19]. As malaria rates decline, transmission becomes more heterogeneous, resulting in reduced gene flow between areas, and increasing focal inbreeding and relatedness of strains as the population size decreases [15, 20-23]. Maps of transmission zones, population connectivity and identification of "source and sink" populations could define the spatial scale and importation risk of distinct transmission zones and guide the containment of antimalarial drug resistance should it emerge [24, 25]. In pre-elimination areas, genotyping of parasites has been used to determine whether infections were imported or locally acquired, characterise transmission chains and to guide outbreak preparedness and response [26]. Genomic surveillance can therefore be used to monitor malaria control and elimination progress, and to improve the efficiency of control efforts [19, 27, 28].

It is crucial that genotyping tools used in malaria surveillance accurately define parasite population structure. WGS is the most comprehensive approach but is currently too expensive and time consuming to be conducted at the fine spatial scales that would be needed to inform local control efforts. Alternative approaches are urgently needed to estimate genome-wide relatedness and define population structure at a similar resolution to WGS but for a fraction of the cost. Previous approaches have included panels of 10-14 polymorphic microsatellites [10], or barcodes comprising 24 [29], 28 [3], 54 [30] or 101 single nucleotide polymorphisms (SNPs)[6]. However, these genotyping tools lacked the required accuracy or spatial resolution to define local parasite populations, especially where the barcodes contained too few SNPs to accurately capture genome-wide relatedness [16, 31-33]. Indeed, the occurrence of low minor allele frequencies (MAF) for the included SNPs in some populations will reduce the spatial resolution of SNP barcodes. The development of a barcode that captures the diversity and divergence of local parasite populations in a manner similar to WGS may provide the most accurate and relevant insights for local malaria control programs [34]. Furthermore, comparison of barcode surveillance data across different transmission settings within and between geographical regions of endemic countries will be crucial to assess the efficacy of control efforts pre- to post-elimination.

We aimed to develop a SNP barcode to characterise parasite transmission dynamics throughout the Southwest Pacific country of Papua New Guinea (PNG). PNG is a major hotspot of malaria in the Asia Pacific, with heterogeneous *P. falciparum* transmission in different provinces [35]. Control efforts were intensified since 2006 resulting in significant decline of transmission and malaria burden in the period up to 2016 [36-38] with a national resolution aiming at elimination from 2025 onwards [39]. We previously showed that *P. falciparum* populations in the highly endemic northern region (Momase) were structured into subpopulations prior to control intensification [21, 40, 41] which suggested that genomic surveillance could potentially identify transmission zones for targeted elimination. To support the integration of genomic surveillance into the national malaria control program in PNG, a barcode was designed using SNPs identified among *P. falciparum* isolates collected from three endemic areas of PNG. We then barcoded samples collected in a nationwide malaria indicator survey and from an outbreak of unknown origins in a low transmission area near the capital city, Port Moresby. The barcodes differentiated between local and imported infections and provided insights into the transmission dynamics, population structure and connectivity of *P. falciparum* parasite populations throughout PNG.

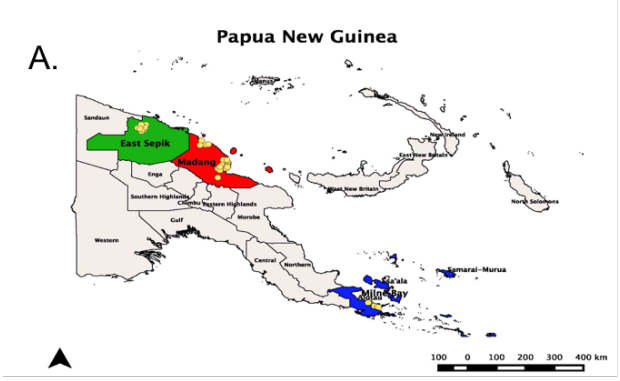
RESULTS

Barcode design and validation

SNP candidates were identified among 91 high quality genomes remaining after processing WGS of 125 *P. falciparum* isolates collected in Milne Bay, East Sepik and Madang Provinces (**Figure 1A,B, Text S1,** Supporting Information: **Figure S1**). This data was generated in collaboration with the MalariaGEN *P. falciparum* Community Project and is publicly available from the European Nucleotide Archive (Supporting Information: **Table S1)**[42, 43] The final SNP barcode included 95 SNPs with mean F_{ST} values of 0.05 or more considered to be geographically informative (iSNPs) and 97 SNPs with F_{ST} values of less than 0.05 with similar (balanced) allele frequencies among the PNG populations (bSNPs), which included 56 from an expanded version of a 'Universal' barcode that also met the above criteria (uSNPs) [29, Baro et al. personal communication, 30] (**Figure S2,** Supporting Information: **Table S2**).

Barcode performance was first assessed by comparing barcodes amongst the 91 high quality genomes (**Figure 1B,** Supporting Information: **Text S1**). We extracted the genotypes for each SNP locus and compared multidimensional scaling (MDS) clustering patterns of 'all SNPs' which included those with minor allele frequency (MAF) of greater than 0.1 in PNG (n=5786 SNPs), the complete uSNP barcode [29, (n= 96)], the complete local barcode (n=191), iSNPs (n=95) and bSNPs (n=97) (**Figure 1**). Using all SNPs, isolates from Milne Bay formed a distinct cluster, reflecting the relative geographic isolation to the other two locations, while East Sepik and Madang overlapped indicating significant gene flow between these neighbouring provinces (**Figure 1**). Two groups of outliers in the East Sepik population suggest importation from another geographic area with a separate analysis showing those isolates clustering with those from the neighbouring Indonesian province of West Papua (data not shown, R. Amato, personal communication). The uSNPs and bSNPs did not capture the observed population structure. While the full local barcode distinguished between the three populations, as expected, the iSNPs showed greater resolution of isolates by geographic origin. This improved resolution suggests a trade-off between the number of SNPs in barcodes and their allele frequency characteristics (**Figure 1C**).

A.



B.

Province	Catchment	Samples
East Sepik	Maprik	36
Madang	Madang area	31
Milne Bay	Alotau	24
PAPUA NEW GUINEA		91

C.

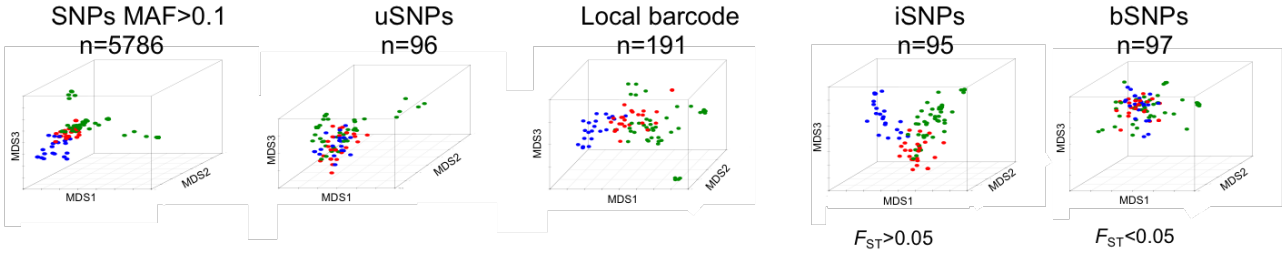


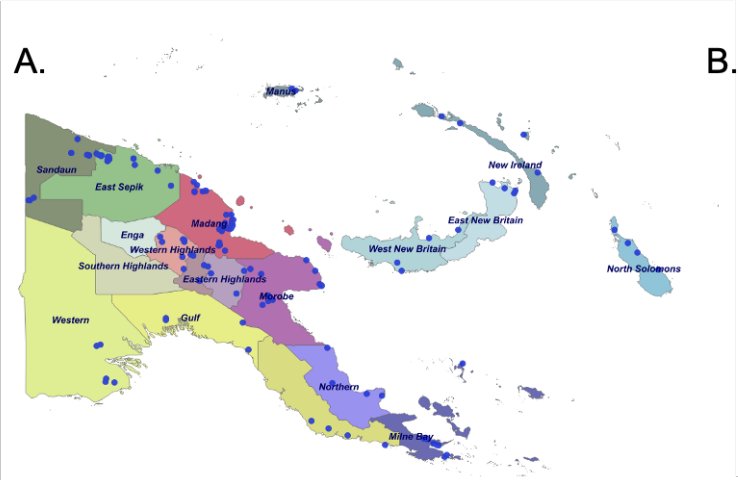
Figure 1. Clustering patterns of P. falciparum isolates from three PNG provinces based on different barcodes. *A) Map of PNG indicating the village of residence of sample donors (yellow dots) and locations (green = Maprik area, East Sepik, red = Madang area, Madang Province, blue = Alotau area, Milne Bay Province). B) Number of high quality P. falciparum genomes used for the analysis and their geographic distribution. C) Multidimensional Scaling (MDS) plots showing clustering patterns of different barcodes (n = number of SNPs). Coloured dots correspond to the province in which samples were collected.*

Barcode performance for Papua New Guinean field isolates

P. falciparum isolates from 68 villages across 16 of the 20 provinces of PNG (**Figure 3A,B**) were genotyped using the 191 SNP local barcode (Supporting Information: **Table S2**). After removing low quality samples and SNP loci, the dataset comprised 636 samples successfully genotyped at 155 SNPs. This "typable" panel contained comparable proportions of iSNPs (n=73) and bSNPs (n=81), of which 46 were uSNPs [29] (Supporting Information: **Table S2**). From the 155 SNPs, a random (rSNP) panel was also selected by combining 35 iSNPs and 35 bSNPs. The data was pooled with corresponding barcodes extracted from the WGS data providing a total of 727 high quality SNP haplotypes in total. The dataset was stratified according to province (n=16) or catchment area by combining proximal villages (n=34 'geocodes') (Supporting Information: **Figure S3, Table S3**), and explored at the village level if subpopulations were identified. Most samples within the final dataset had less than 10% of SNPs missing (Supporting Information: **Figure S4**).

We then investigated the population structure of *P. falciparum* in PNG observed using different barcode panels. MDS showed overlap of clusters by geographic location, although geographical separation was visible for the full barcode and iSNPs (**Figure 2C**). DAPC density plots further indicated that the genetic differentiation of parasite populations was greatest with both the full barcode and with iSNPs (**Figure 2D**). The rSNPs captured some population structure but did not identify Milne Bay as a distinct

population, which was at odds with the WGS data (**Figure 1**). The uSNPs and bSNPs failed to identify any significant population structure (**Figure 2C,D**).



B.

Region	Province	Catchments	Samples
MOMASE (n=472)	West Sepik	3	44
	East Sepik	4	110
	Madang	4	140
	Morobe	4	178
ISLANDS (n=197)	Manus	1	16
	West New Britain	2	97
	East New Britain	1	22
	New Ireland	2	59
	Bougainville	1	3
SOUTHERN (n=50)	Oro	1	14
	Milne Bay	3	29
	Central	1	2
	Gulf	2	3
	Western	2	2
HIGHLANDS (n=8)	Western Highlands	2	7
	Eastern Highlands	1	1
	Totals	34	727

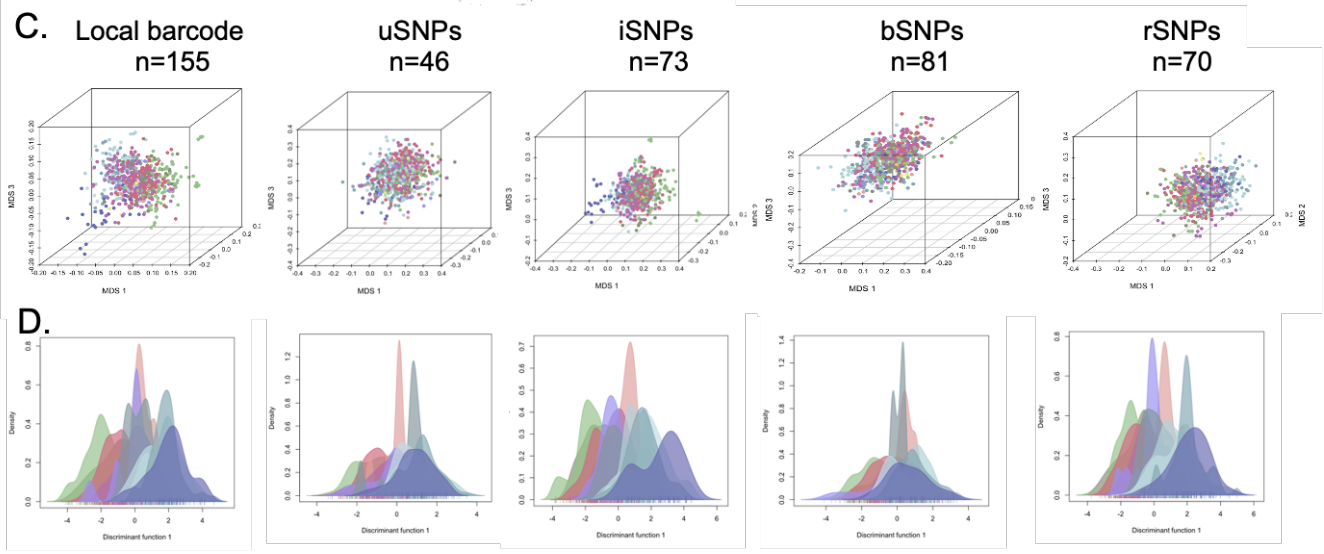


Figure 2. SNP barcode performance on Papua New Guinean P. falciparum isolates

A) Map of Papua New Guinea with sampling locations. Dots indicate villages, colours indicate different provinces. B) Numbers of catchment areas and successfully genotyped samples for each of the endemic provinces. C) Multidimensional Scaling (MDS) analysis of the samples using different panels of SNPs. n= number of SNPs. D) Discriminant Analysis of Principle Components (DAPC) density plots showing the dispersion of pairwise diversity values for populations with at least 7 samples. Colours correspond to the map.

Variable local transmission dynamics across PNG

To determine the transmission dynamics of *P. falciparum* populations in PNG, using the complete panel of 'typable' SNPs (n=155), we then measured population nucleotide diversity (π) and pairwise relatedness using expected identity by descent (eIBD) [44]. Nucleotide diversity was high in all populations across the country, though slightly lower in the island populations than on the mainland suggesting lower transmission or a more isolated population (**Figure 3**). Overall, PNG parasites showed

low proportions of pairwise relatedness as would be expected for a subsample of the parasite population in a high transmission, endemic setting where high rates of recombination of genetically distinct individuals reduces the chance of related genotypes being found. Higher proportions of related pairs with eIBD greater than 0.25 were found in West Sepik (GeoCodes 1, 2), East Sepik (GeoCode. 4), Manus (GeoCode 19) and other island populations (GeoCodes 20-26) (**Figure 3**). Larger proportions of closely related parasite pairs with eIBD values greater than 0.55 were found in East Sepik (GeoCode 4) and Manus (GeoCode 19) suggesting low transmission with less frequent outcrossing and clonal transmission (**Figure 3**). The Central Province (samples collected east of the capital city, Port Moresby), and Gulf Province had small sample sizes (n=2) due to the low prevalence of *P. falciparum* in these areas [35] and therefore were not shown on the plot. Central Province parasites were not related (eIBD<0.50) suggesting these infections may have been imported, whilst in Gulf Province the genotypes were identical (eIBD not calculated due to low sample size), consistent with their originating from the same source and local transmission (not shown in Figure 3 due to small sample size).

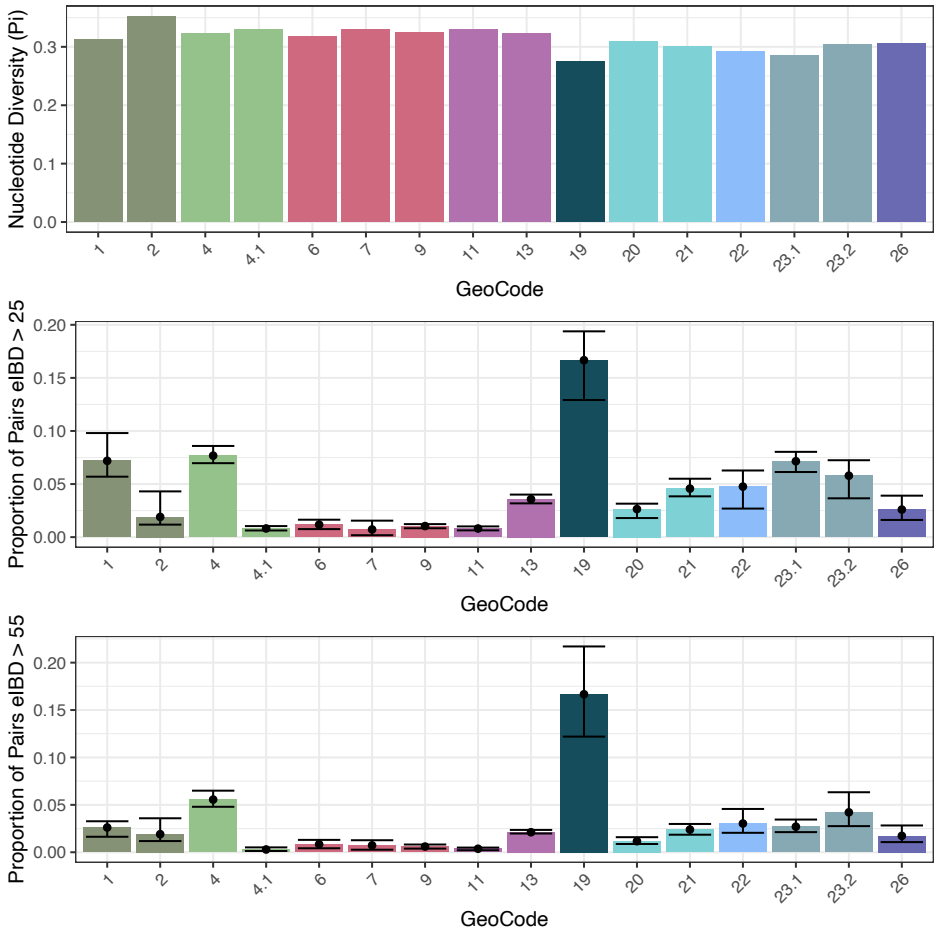


Figure 3. Population diversity and pairwise relatedness of Papua New Guinean P. falciparum isolates
Nucleotide diversity (π) (top) and relatedness (proportion of pairs with IBD > 25% (middle) and 55% (bottom). Error bars indicate 95% confidence intervals after bootstrapping (500 iterations). Only geocodes with 15 or more samples were included in this analysis. Names of catchment areas (numbered 'geocodes') are in Table S4. Colours indicate province as shown in the map in Figure 2A.

We then drew 'haplotype networks' using the pairwise eIBD values to connect related genotypes, with networks drawn using differing eIBD thresholds. These networks identified several clusters of highly related (eIBD>0.55) and closely related (eIBD>0.75) parasites (**Figure 4**). All but one genotype cluster was comprised of parasites from the same geographic area, with only one cluster containing isolates from different geographic areas, suggesting they may be imported. At eIBD thresholds above 0.75, clusters are detected within villages or between neighbouring villages only. This indicates focal transmission due to intensifying control pressure and pockets of residual transmission [21, 37]. As eIBD thresholds for were relaxed from 0.55 to 0.45 to 0.35, parasite population-connectivity remained within the villages, or between neighbouring villages only, indicating that lower eIBD thresholds are necessary to resolve population connectivity at sub-provincial spatial scales. At an eIBD threshold of 0.30, there was evidence of connectivity between neighbouring provinces within Momase and the Outlying Islands, and limited evidence of relatedness between more distant provinces. More extensive evidence of population connectivity at sub-provincial spatial scales was conspicuous at an eIBD threshold of 0.25, however this is the lower limit of resolution for the local barcode and prone to higher false-positive rates (Mehra et al. unpublished data).

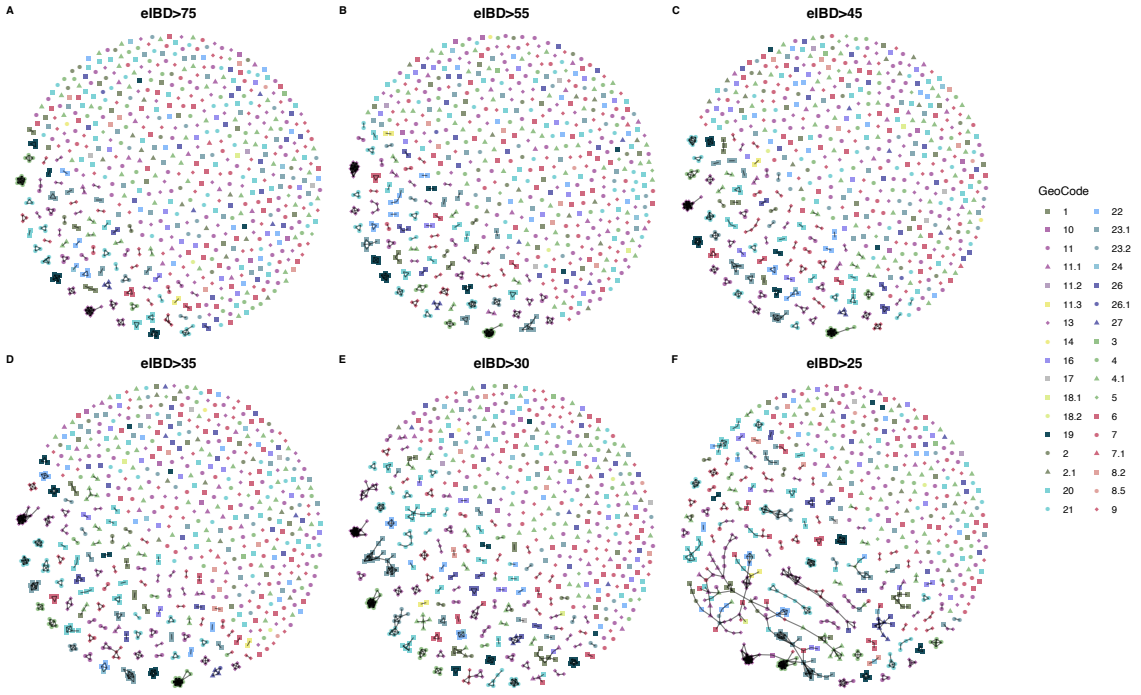


Figure 4. Relatedness networks of Papua New Guinean P. falciparum isolates. *Expected Identity by Descent parameter (eIBD). Networks of within population connectivity for different eIBD cut-offs. Nodes indicate parasite genotypes, coloured by geographical origins and lines indicate pairs with significant levels of IBD with a threshold of A) 0.75, B) 0.55, C) 0.45, D) 0.35, E) 0.30 and F) 0.25.* Names of geographic catchment areas numbers in the key are shown in Table S4. Colours indicate province as shown in the map in Figure 2A.

Eight genetically distinct parasite sub-populations in Papua New Guinea

Multidimensional scaling (MDS) and Bayesian clustering (STRUCTURE) analyses indicated a spatial "gradient" of parasite population structure along the north coast (West Sepik (Sandaun) > East Sepik > Madang > Morobe) with admixture between neighbouring provinces (West/East Sepik, and

Madang/Morobe, **Figure 5A**) and low pairwise F_{ST} values (0.01-0.05, Supporting Information: **Figure S5**) consistent with a single contiguous parasite population with individual catchment areas isolated by distance. The outlying Islands (New Britain, New Ireland, Manus Island) and Milne Bay isolates however showed limited overlap with mainland isolates and were genetically distinct populations as confirmed by the elevated levels of genetic differentiation (mean F_{ST} = 0.03-0.18, Supporting Information: **Figure S5**). Overall, the STRUCTURE analysis identified eight genetically distinct clusters amongst the genotypes that were asymmetrically distributed throughout the country. The data suggest three major regional blocks, namely the mainland (East/West Sepik, Madang/Morobe and Milne Bay) as well as distinct populations for Manus and the other outlying Islands, with multiple subpopulations within Oro Province aligning with the villages suggesting lower and focal transmission (**Figure 5B**). Morobe isolates clustered by village and isolates from one village shared ancestry with Island isolates. These genetic clusters may indicate distinct malaria transmission 'zones' and defines the spatial scale that may be needed for targeted control efforts. The results also indicate a potential parasite migration route between the Islands and Mainland via Morobe, which would also be important to consider in targeted control efforts. There were some island genotypes with shared ancestry with the north coast (green and red) indicating recent importation from provinces on the north coast. Highlands and South Coast isolates shared ancestry within either Island, Sepik or Madang/Morobe populations, consistent with multiple importation events.

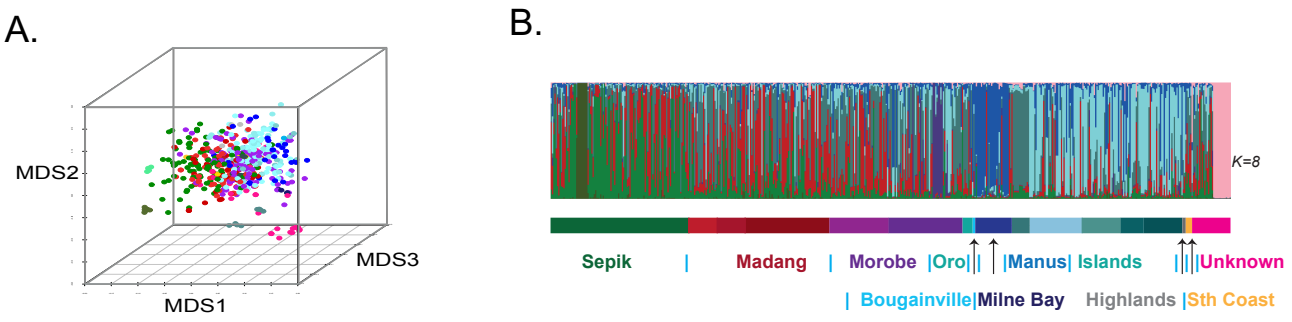


Figure 5. Population structure of P. falciparum in Papua New Guinea. *A) MDS analysis of the outbreak samples with all PNG samples. B) Ancestry coefficients (min = 0 max =1) of outbreak samples with all PNG samples using Bayesian cluster analysis. Colours match geographical origins as indicated on the map in Figure 2A. Unknown genotypes (pink) correspond to clinical samples of unknown geographical origin from migrating workers as described in the next section.*

Identifying the origins of infections in malaria outbreaks

Given the presence of genetically distinct populations nationally, and higher levels of relatedness between isolates from the same village, we hypothesised that SNP barcodes may distinguish between imported cases and local transmission in lower transmission settings, and potentially determine the geographical origins of infections. To assess whether SNP barcodes with different characteristics can provide this information, we genotyped 32 *P. falciparum* clinical isolates from cases collected during a malaria outbreak that occurred amongst migrant workers over a 12-month period. The camp was situated in a hypo-endemic region just outside the capital city Port Moresby and workers periodically travelled to their home provinces. Therefore, we wanted to determine whether infections were imported from another endemic area or acquired locally at the camp, and if so, whether they originated from an index case at the camp.

In the MDS analysis (**Figure 5A**), comprising of all samples collected in PNG, a subset of the outbreak samples were closely related outliers, whilst others clustered with samples from other endemic areas include in this study. This pattern of differentiation was consistent with the Bayesian cluster analysis, which showed a distinct genetic cluster (pink) of a subset of the outbreak samples. Admixture clustering of the other subsets indicated ancestry with north coast and island populations (**Figure 5B**).

We also assessed the ability of different SNP panels to classify infections as local or imported using phylogenetic analyses (**Figure 6**). The local barcode, iSNPs and bSNPs showed similar clustering patterns, dividing related (eIBD>0.45, Figure S6) infections (pink) into a distinct clade from the more diverse imported infections. However, uSNPs were only able to identify three related infections, and the rSNPs indicated two independent sources of the outbreak. The phylogenies identified a monophyletic clade with 15 related haplotypes suggesting a clonal expansion from genetically related parasites (Figure 6, pink nodes) with eIBD values of more than 0.45 (**Figure S6**).

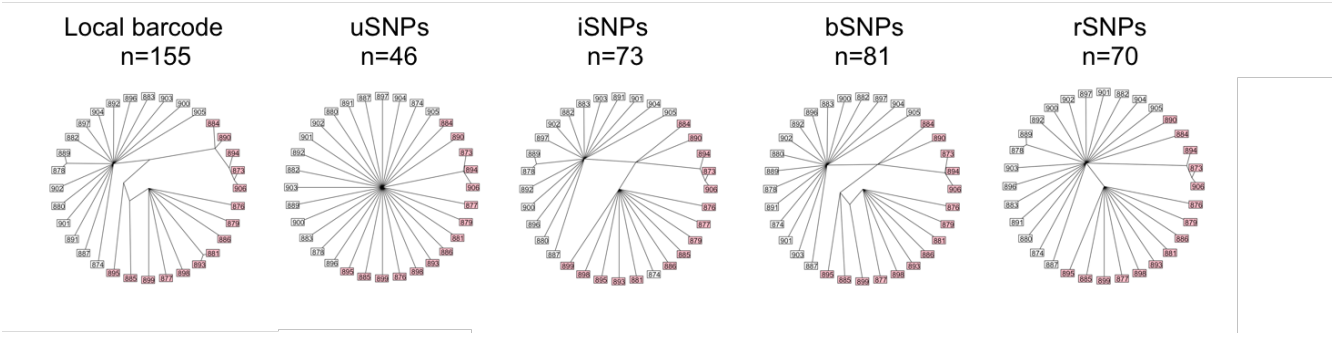


Figure 6. Phylogenetic analysis of P. falciparum haplotypes from a malaria outbreak using different SNP barcodes

Neighbour joining trees for different SNP barcodes. Neighbour joining trees for different SNP barcodes were created using the bionj function in R statistical software. Bootstrapping was performed by collapsing weak nodes with bootstrap support less than 70% over 1000 iterations. Samples indicated in pink were identified to be part of the same cluster by IBD analysis, using a threshold of eIBD>0.45 to connect samples (only one distinct cluster was detected) (Supporting Information: ***Figure S6****).*

We screened for related isolates in the nationwide dataset with more than 0.25 eIBD sharing. Detected eIBD connections for eight isolates however all showed low IBD values of between 0.25-0.29 (Supporting Information: **Table S5**). Three of the 15 haplotypes within the monophyletic clade above appeared to be related to isolates from the south coast of West New Britain and the Huon Peninsula of Morobe, suggesting that the outbreak was seeded by imported cases from these provinces. The remaining 17 haplotypes within the outbreak network were diverse and exhibited limited eIBD sharing with other infections within the outbreak.

Utility of the barcode for other malaria endemic regions

We also sought to assess the applicability of our SNP barcode to define parasite populations in geographic regions with different malaria transmission dynamics to PNG. We extracted the 155 validated SNPs from a global *P. falciparum* WGS dataset. This comprised all genomes available in Pf3k

version 5.0 from 14 countries in addition to the genomes from the three PNG populations described in this study (Total n=2598 isolates, [45]). Based on 742,365 high-quality SNPs identified in this dataset, parasites clustered by region – Africa, Southeast Asia, Bangladesh, and PNG (Figure 7). Most SNPs in the barcode were polymorphic in parasite populations that existed in these regions, with at least 90 SNPs showing MAF > 0.05 in African and Southeast Asian countries (Supporting Information: **Figure S7, Table S6**).

The MDS output showed similar population structures irrespective of the SNP barcode type used, albeit with type-specific clinal differentiation (**Figure 7**). Barcodes based on bSNPs and rSNPs differentiated PNG parasite populations from Asia Pacific and Africa, whilst uSNPs and iSNPs differentiated African and Southeast Asian isolates only, with PNG and Bangladesh showing clinal differentiation to Southeast Asian populations (**Figure 7**). This pattern of differentiation suggested that the iSNPs used in this study lacked sufficient discriminatory power to accurately differentiate parasite populations in PNG from other Asia-Pacific regions.

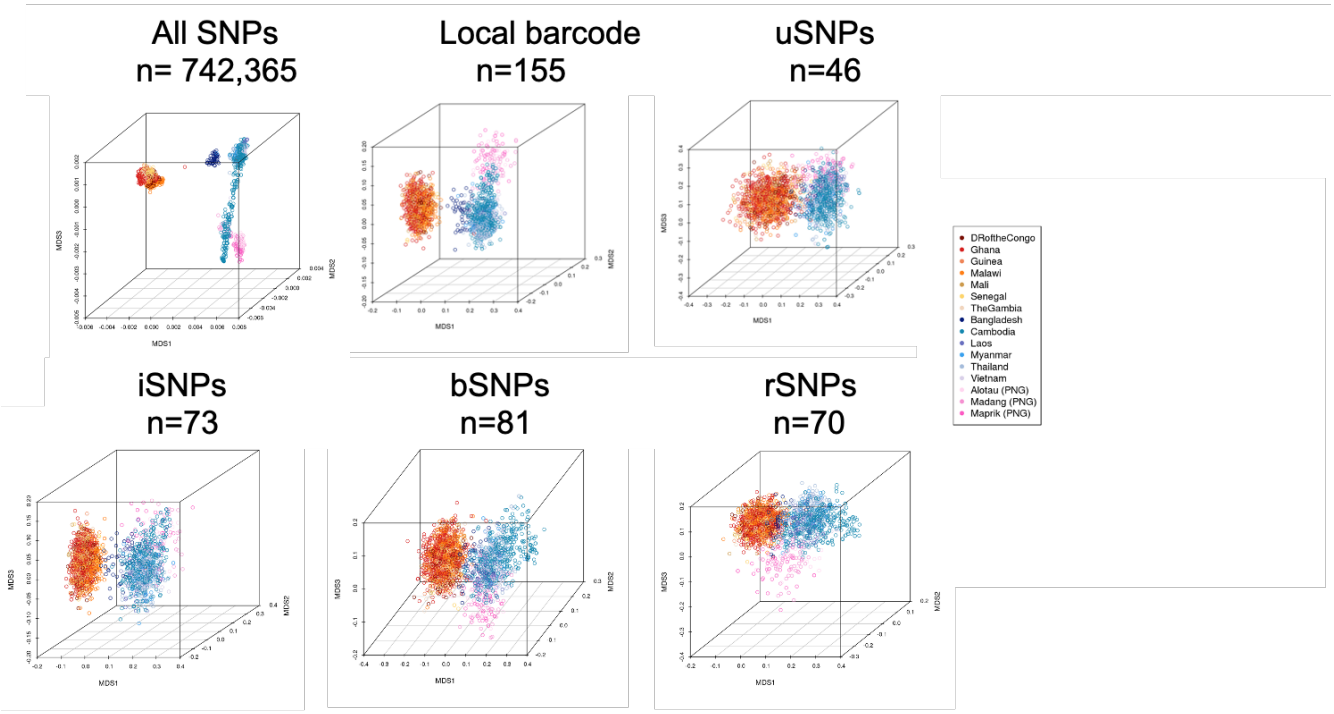


Figure 7. Clustering patterns of worldwide P. falciparum isolates based on different SNP barcodes. *SNP genotypes from Pf3K v5.0 genomes were combined with data from the three PNG populations (n=2789 isolates). Isolates with genotype missingness <20% across the relevant panel of SNPs were selected, and MDS analysis conducted for all high-quality SNPs in the Pf3k dataset (n=742365 SNPs, n=2598 isolates), and validated SNPs from the local barcode (n=155 SNPs, n=2630 isolates), uSNPs (n=46 SNPs, n=2628 isolates), iSNPs (n=73 SNPs, n=2632 isolates), bSNPs (n=81 SNPs, n=2629 isolates)* and *rSNPs (n=70 SNPs, n=2628 isolates).*

DISCUSSION

With the advent of more affordable and high throughput tools for genotyping such as SNP barcodes, parasite genomic surveillance is becoming critical to inform malaria control and elimination efforts. SNP

barcoding can be done at a fraction of the cost of WGS and has immense utility for detecting and tracking the parasite transmission dynamics in different endemic areas. We conducted this study to develop and validate a SNP barcode tailored to capture the diversity of the local PNG parasite population, and to assess the capabilities of SNP barcodes with different allele frequency characteristics. In this study, we developed and validated a SNP barcode comprised of 155 genome-wide SNPs identified in local parasite populations and sampled across provinces with malaria. The barcode was used to define the local parasite population structure and was found to have utility to track outbreaks among clinical cases in PNG. Genetic resolution in the global parasite population suggested the SNP marker panel could be used for different geographical regions, albeit validation is recommended and refinements in the SNP panel may be required to ensure optimal output. Our SNP barcode together with the data generated in this study will inform control activities conducted by the PNG national malaria control programme.

The results suggest that a universal SNP barcode (uSNPs), designed using sequence diversity in non-target populations, may lack sufficient discriminatory power to define local parasite populations. Previous studies suggested that barcodes of 100-200 SNPs can provide accurate measures of genome wide IBD for measuring population connectivity [33]. Here, we aimed to develop a barcode that would have broader practical and geographical utility and typical of SNP allele frequency characteristics within three 'reference' high transmission PNG parasite populations. Geographically informative SNPs (iSNPs) within the developed barcode were included to maximise classification of sample origins at sub-national scale in PNG (F_{ST}>0.05, iSNPs), whilst others had more balanced allele frequencies between the reference populations (F_{ST} <0.05, bSNPs). The barcode was deliberately designed so that the proportions of SNPs with these characteristics were representative of those found for WGS data, on the basis that they could provide an accurate estimate of both the underlying transmission dynamics and structure of parasite populations (all barcode SNPs) and for determining geographic origins of individual malaria cases (iSNPs). SNPs were selected from neutrally evolving regions of the genome to avoid the inclusion of SNPs under selective pressure, either drug or immune selection, that could skew insights into the underlying transmission dynamics. Comparison of the full SNP barcode to the rSNPs, uSNPs, bSNPs and iSNPs showed contrasting discrimination of population structure. Differences appear dependent on the numbers of SNPs, the allele frequencies within (MAF) and between (F_{ST}) populations, where SNPs with F_{ST} values greater than 0.05 are assumed to provide sufficient genetic differentiation among local parasite populations to identify genetically distinct parasite populations.

The complete local barcode was suitable for characterising nationwide transmission dynamics by measuring genetic diversity and relatedness patterns between genotypes and defining the parasite population structure in PNG using MDS, Bayesian clustering and F_{ST}, with utility to track local outbreaks and imported cases using phylogenetic analysis and relatedness against the national dataset. As expected, the iSNPs barcode, with unbalanced allele frequencies between populations, identified increased population structure with fewer SNPs, however the rSNPs which were selected randomly and uSNPs and bSNPs selected based on balanced allele frequencies, were not able to define the population structure observed using the gold standard WGS SNPs. Therefore, refined barcodes containing SNPs with allele frequency characteristics that are representative of the parasite genome, as used in this study, and validation of SNP barcodes against WGS is a more cautious approach than random selections of markers. The cost involved in SNP-barcoding infections could be reduced by reducing the number of SNPs in a panel. However, this could also compromise the genetic resolution in

parasite populations, with limited resolution in high compared to low transmission settings [16, 34]. That is, in regions with higher transmission such as those in PNG and sub-Sahara Africa with little-to-no LD in the parasite population, accurate IBD estimates will require larger SNP panels [34, 46] and should be assessed against WGS for ascertainment bias [8]. In low to middle income countries where malaria control programmes may be under-resourced to conduct parasite genomic surveillance, smaller panels of SNP barcodes could be tailored to the local parasite populations and the research question. Given the limited access to high-throughput genotyping platforms in endemic countries, optimised SNP barcodes with a smaller number of SNPs, could be operationally feasible for parasite genomic surveillance activities. Where SNPs are randomly selected, their frequencies in the local parasite population need to be assessed and regularly monitored to validate their performance in response to local control interventions.

The complete local SNP barcode identified parasite population structures within and between provinces of PNG, a utility that can be integrated into the PNG national malaria control program to define major parasite population subdivisions throughout the country and local transmission dynamics. The SNP barcode identified major population subdivisions within PNG, on the mainland-north coast (a region covering multiple provinces known as Momase), Eastern Tip (Milne Bay) and outlying Islands as well as the local transmission dynamics in low transmission settings (Manus Island, Northern (Oro) and Central Provinces). These population structures, which were also observed between East Sepik and Madang previously, corroborate previous findings for *P. falciparum* and *P. vivax* using microsatellite markers [14, 21, 40] and may be explained by the relatively limited human movement between these regions at the time the surveys were done. Amongst the Islands, despite the dense sampling, we observed limited population structure, except for Manus, which is a relatively isolated island and has a highly structured parasite population with high levels of relatedness amongst genotypes. The Milne Bay parasite population was also genetically distinct from other populations indicating that parasite migration from other areas into and from this region is limited and may reflect a past population bottleneck. There was also evidence of imported infections in the East Sepik with a related cluster of parasites that were outliers to the entire PNG parasite population. This suggests international importation, with subsequent clonal transmission in the local community.

Recent studies have indicated significant mixing among parasite populations in Wewak, East Sepik and from the West Papuan (Indonesian) side of the island of New Guinea [47], and microsatellite data indicated focal transmission with clusters of related parasites within certain villages [21]. East Sepik implemented multiple LLIN distributions prior to this study, suggesting focal clustering may indicate interrupted transmission due to extended control efforts relative to other north coast provinces. Morobe parasites were structured according to village origins with some evidence of importation from the Islands, indicating a potential transmission zone boundary and migration point between Islands and the Mainland, that would need to be considered in subnational control efforts. Sample sizes from the Southern region (Western, Gulf and Central Provinces) and Highlands were very small, limiting the conclusions made about these parasite populations. However, the relatedness estimates of these samples suggested importation (not related) and/or clonal transmission (related), thus, demonstrating a significant interruption to transmission in these areas. While sample numbers were too low to confirm, the low nucleotide diversity and high relatedness between genotypes from the Gulf Province suggests clonal transmission within the local area. Conversely, in Western, Central and multiple highlands

provinces with low parasite relatedness and where the parasite population showed similarity to parasites from highly endemic areas, case importation may be the major contributor to local malaria episodes. Within-country movement of people makes low transmission areas like the Highlands and Southern regions potential sinks for infections. As only low levels of relatedness were detected between geographic areas, levels of connectivity were not inferred using eIBD methods as others have done [15, 31]. Instead, we base the inference of connectivity between populations on F_{ST} values and levels of admixture or shared ancestry in the STRUCTURE analysis.

The SNP barcode provided utility to detect and track the origins of local outbreaks in low transmission settings. Barcoding of the clinical isolates obtained from migrant workers in Port Moresby revealed a local outbreak that spread from two origins: a single case that was likely imported and multiple imported cases. Our data suggested that the outbreak was seeded by two independent transmission events with an admixture of SNP barcodes carried clusters of parasites that were likely imported from the north coast and island populations. Thus, while it is important to maintain surveillance locally in each endemic province, there is also a need to monitor for imported cases that may potentially seed transmission in provinces moving towards elimination.

Limitations of this study included that SNP candidates were identified from only three parasite populations including one relatively geographically isolated population (Milne Bay). It is possible that the selection of SNPs based on mean F_{ST} among these three populations may have identified SNPs that are private to one of these populations, however none of the SNPs chosen were private to any population (**Table S2**). Another limitation is the success of the genotyping approach, resulting in the loss of almost 25% of the candidate SNPs. Some markers were removed due to batch effects, and others because they had a less than 70% genotyping success rate amongst the samples. This could be improved by validating additional SNP candidates, however mathematical modelling has indicated that 100 SNPs will be adequate for IBD estimation [33]. In addition, the samples used in this study were archival asymptomatic samples that are low density infections and may have reduced DNA quality with multiple freeze thaw cycles, contributing to genotype failures. The analysis of nationwide population structure was limited by sampling bias with large numbers of samples from the high transmission Momase and Islands regions and very few samples from the relatively low transmission Highlands and Southern regions of the country. Sampling in these areas may be boosted through inclusion of cases from malaria clinics and hospitals in the national malaria indicator surveys conducted in PNG every three years, to confirm the origins and dynamics of transmission.

In summary, our data demonstrate the utility of SNP barcodes for parasite genomic surveillance, to inform malaria control and elimination in PNG and globally. The performance of the different SNP barcodes in PNG and among other malaria endemic countries within and outside the Asia-Pacific underscores the need to optimise SNP panels for local and/or regional parasite populations, factoring in the locus-specific MAF and F_{ST} estimates to ensure subnational parasite population structure can be detected and monitored accurately. Integration of the information provided by SNP barcodes into disease surveillance maps to map malaria transmission patterns over space and time will be useful to define regions for targeted elimination, to measure the success of control interventions, detect and track the spread of drug resistant parasites and enhance outbreak preparedness across different transmission settings. Considering the recent recognition that new tools will be required to progress beyond the

current stagnating global progress against malaria, this research provides critical new evidence supporting the use of parasite genomic surveillance. Until WGS becomes widely accessible and affordable for malaria control, SNP barcodes present tremendous opportunities to guide malaria elimination in PNG, and eradication globally.

MATERIALS AND METHODS

Ethics and Informed Consent. All samples were collected as part of ongoing studies in PNG investigating the impact of intensified malaria control efforts since 2004. All samples were collected with written informed consent from individuals or if children, consent was obtained from their parents, and guardians. Ethical approval for the study was obtained from the Papua New Guinea Institute of Medical Research Institutional Review Board (IRB 11/21, 12/29), the PNG Medical Research Advisory Council (MRAC 12/03, 13/08), the Walter and Eliza Hall Institute Human Research Ethics Committee (HREC 12/06, 13/14) and Deakin Human Research Ethics Committee (2020-282, 2020-283).

Study design. The objective of this study was to assess population structure at genome-wide resolution using WGS, and a SNP barcode tailored to local PNG parasites. For WGS, we collected whole blood samples from clinical malaria cases, thus the sample size was limited by the clinical cases and samples available from the original study. For SNP typing we utilised *P. falciparum* qPCR-positive blood samples collected from households in 68 villages as part of a national malaria indicator survey, and symptomatic, laboratory-confirmed clinical cases from four sentinel sites, and during a malaria outbreak amongst migrating workers residing in a camp in a low endemic area. For the samples from the national survey, we aimed to produce dense spatial sampling in high endemicity areas with a maximum of 30 samples per village, and in low endemicity areas we genotyped as many isolates as were available. Data collection was stopped when this number of samples was reached for each village. Data was included only if fewer than 30% of allele calls were missing from a genotype (sample) or marker (SNP). Using Principal Components Analysis (PCA), 'outlier' SNPs and samples were identified, but only outlier SNPs were excluded, since 'outlier' samples were likely to be imported infections, and thus included in the investigation. The units of study were groups of samples (parasite populations) stratified at large to small spatial scales including (i) region (n=4, Momase, Islands, Highlands, Southern), (ii) province (n=16) and (iii) catchment (GeoCode, n=34) comprising samples from proximal villages.

Samples and study area. PNG is located on the eastern half of the island of New Guinea in the WHO Western Pacific region just to the north of Australia and shares a border to the West with Indonesia (West Papua). It is a major hotspot for malaria in the Asia-Pacific region with intense year-round malaria transmission ranging from hyper- to holo-endemic in the lowland and coastal areas of the Momase and Islands and hypo-endemic to epidemic in the Highlands and the less populated or peri-urban Southern regions [35].

For whole genome sequencing (WGS), *P. falciparum* isolates were collected from individuals diagnosed with clinical malaria during (i) a severe disease case control study conducted in 2005 in the Madang area, Madang Province [48], and (ii) an antimalarial efficacy trial conducted in 2012-13 in Maprik, East Sepik Province (7 villages) and Alotau, Milne Bay Province (9 Villages)[49]. The Madang samples (n=55, from 3 villages) were sequenced in previous studies[50]. For Maprik and Alotau, new isolates were

collected for this study, and have been described in a previous publication[49]. In summary, a total of 1 mL of fresh venous blood was processed using a CF11 filtration procedure to obtain purified erythrocytes for DNA extraction according to published protocols [51]. More than 202 *P. falciparum* positive clinical samples from Alotau and Maprik were submitted for sequencing, of those, 157 met the quality control threshold, and 122 met the quality thresholds for inclusion in the study (Supporting Information: **Text S1**, [42]). This included removal of one individual of the pairs the samples identified as clonal or closely related to another through multidimensional scaling (MDS) and known location of collection and familial relationship (siblings from the same household). SNP candidates were initially selected using 56 of the highest quality genomes (Maprik = 16, Madang = 22, Alotau = 18) and further validated using a total of 91 isolates as the data became available (Maprik = 36, Madang = 31, Alotau = 24, Supporting Information: **Table S1**).

P. falciparum isolates for the barcoding were selected from a nationwide malaria indicator survey conducted by the PNG Institute of Medical Research between October 2008 and August 2009 [35]. The survey included the collection of 6646 finger prick dried blood spots from asymptomatic individuals of all ages through a household survey in 49 villages, in addition to 2290 samples from symptomatic individuals residing in 19 villages across six sentinel sites ([35, 36], Supporting Information: **Table S1, Table S3**). In total the sample set included 8936 samples from individuals living in 68 villages in 16 of the 20 provinces of PNG. For molecular diagnosis of *P. falciparum*, DNA was extracted from dried blood spots using Qiagen or Favorgen extraction kits and a semi-quantitative post-polymerase chain reaction, ligase detection reaction/fluorescent microsphere assay conducted as described [52]. *P. falciparum* isolates identified were previously genotyped to determine multiplicity of infection (MOI) [53] and selected for SNP genotyping if only one clone was present (MOI=1). If sample number was low for a particular location, low complexity multiple clone infections were included (MOI=2). In the previous study, 3784 *P. falciparum* positive samples were identified by qPCR and of those, 2400 were successfully genotyped using the highly polymorphic marker *Pfmsp2* [53]. We selected 772 *P. falciparum* isolates for genotyping (MOI1= 585, 75.7%; MOI2=187, 24.3%) aiming to maximise sampling density across different provinces. Data cleaning included removing samples with low SNP genotyping success rates (n=167 samples with more than 30% missing SNP loci) and removing SNP loci with low quality output and batch effects (n=34) from all samples. This resulted in 636 samples successfully genotyped at 155 SNPs. For the population-level analyses we combined villages at two different spatial scales: province (n=16) and geographic catchment area (n=34) defined based on the spatial proximity of villages and local knowledge of transport networks and topographical features that may influence parasite population movement (Supporting Information: **Table S3, Figure S3**).

To investigate infections of unknown origins, we barcoded 32 *P. falciparum* isolates from individuals diagnosed with clinical malaria based at a resource company camp approximately 20 km from the capital city, Port Moresby. Although malaria is at very low prevalence [35], this region is receptive to malaria transmission, and therefore the company has strict malaria control procedures in place. However, workers leave the camp for social activities and to return to their home province, where they may acquire malaria infections. Travel history was not available for the participants.

Positive control samples for the SNP genotyping included pure *P. falciparum* laboratory strains 3D7 and D10. Mixtures combining different ratios of these clones were made as standards to calibrate the

Fluidigm raw data analysis and to enable the detection of major alleles in the case of multiple infections.

SNP candidate selection and validation.

Whole genome sequencing. Leucocyte-depleted clinical *P. falciparum* isolates were sent to the Sanger Institute (Hinxton, UK) for Illumina-based whole genome sequencing supported by the Malaria Genomic Epidemiology Network (MalariaGEN) *P. falciparum* Community Project [50]. This data has now been made publicly available [43]. Illumina short read data was processed to extract high quality variant calls by the MalariaGEN team [50]. Samples with poor coverage or diagnosed as containing multiple infections as determined by *Pfmsp2* genotyping [54] and/or F_{ws} values of less than 0.80 [55] were removed. The missingness threshold was determined using Poisson regression model using Mallows or Huber type robust estimators with the missingness counts as the response variable, and a cutoff of greater than 70% missing genotypes).

SNP candidates. Candidate SNPs with minor allele frequencies (MAF) greater than 0.1, calculated directly in R using the adegenet version 1.3-8 software [56, 57] were selected to represent both the distribution of geographically informative (iSNPs, F_{ST}>0.05) and balanced allele frequencies (bSNPs, F_{ST}<0.05), and to include SNPs from an extended version of a universal *P. falciparum* barcode (uSNPs) developed based on African and Asian parasite populations [29] that were also found to be polymorphic in the PNG population. Full details of the SNP candidate selection and final panel are outlined in the Supporting Materials and Methods (Supporting Information: **Text S1, Figure S1, S2A, Table S2**).

SNP barcoding.

SNP genotyping assays were developed for the final panel of 191 SNPs using the Fluidigm SNPType® system using the Fluidigm BioMark platform and 96.96 Dynamic Array Integrated Fluidic Circuit (IFC) genotyping (Supporting Information: **Figure S2B**), following the manufacturer's instructions with minor modifications. Two IFC chips per batch of 84 samples were needed to genotype all SNPs in the barcode. Samples on each chip included a total of 84 field isolates, plus 12 controls including pure 3D7 and D10 DNA and mixtures in ratios of 90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80 and 10:90, and a no DNA (water) template as the negative control. The assay was first optimised using the controls, and then used to genotype field samples.

SNP barcoding assays. Due to the low parasitemia of many samples and low volume required, 30 μL of extracted DNA was first reduced to 3 μL using a desiccator and 1.25 μL was added to each primary multiplex reaction containing all 191 primer pairs. Primary reactions were carried out in a total volume of 5 μL in 96 well plates on a T100, BIO-RAD thermal cycler using the following conditions: 2.5 μL (1x) Qiagen 2X Multiplex PCR Master Mix (Cat No:206143, USA) and 0.5uL (0.2 μM) each of the Specific Target Amplification (STA) & Locus Specific (LSP) primers. The cycling protocol included an initial denaturation for 15 min at 95 ºC, followed by 16 cycles of 15 sec denaturation at 95 ºC and annealing & extension together for 4 min at 60 ºC. The secondary reaction is a qPCR end point fluorescence assay using Biotium 2x Fast Probe master mix (Cat No. 31005, Biotium, US), Rox (Cat No.12223-012, Invitrogen, US) and SNPtype genotyping reagent kit 96.96 (Cat No.100-4134, Fluidigm, US). Priming and loading of IFC into BioMark Platform was performed according to manufacturer's instructions. IFC chips (Cat No. BMK-M-96.96GT, Fluidigm, USA) contain 96 separate reservoirs on each side of the microfluidic device, one for samples, into which 5uL of the sample mix containing 1:25

diluted primary PCR product and the other side 4 μL of assay mix containing Allele Specific Probes(ASP) and Locus Specific (LSP) primers were aliquoted following the manufacturer's instructions (Supporting Information: **Table S3**). Within the IFC, 9216 individual qPCR reactions are assembled automatically via microfluidics using the IFC controller. The cycling protocol used on the Biomark TM instrument was "SNPtype 96.96 v1(Fluidigm, US). Each IFC chip was set up to genotype 96 SNPs in 96 samples, which included different ration of mock multiple infection controls comprised of single and mixed proportions of the reference strains 3D7 and D10 (90:10, 70:30, 60:40, 50:50, 30:70, 40:60, 10:90).

SNP calling. Allele specific PCR products were generated and imaged with scatter plots using Fluidigm SNP genotyping analysis Software (Fluidigm, United States). Briefly, a scatterplot is produced for all samples on the chip per SNP to enable automated allele calling by the software. In the case of multiple infections, the dominant allele was identified by the spectra/location on the scatter plot relative to dominant alleles in the control mixed infections (Supporting Information: **Figure S2B**).

Data analysis.

SNP barcode data was analysed using multiple approaches. Briefly, nucleotide diversity (π) [58], was calculated using the function *nuc.div* in Pegas (V0.12) [59] with pairwise deletion of missing loci. Multilocus F_{ST} was calculated for each pair of populations using the function *varcomp.glob* in Hierfstat (V0.04-22) [60].

Bayesian cluster analysis was done using STRUCTURE version 2.3 [61] using the following parameters: K=1 to 20 with 20 runs with different seeds, an MCMC burn-in period of 5000 iterations and total MCMC iterations up to 50,000. Optimal K was determined by using the inflection point in final estimated log-likelihood and the 'Evanno et al' method [62]. MCMC diagnostics were also checked by looking at change in admixture parameter over MCMC runs. Samples were labelled by population, but population origin was not set as a prior in the analysis to avoid overestimation of population structure.

Clustering and phylogenetic analyses for SNP data were based on genotypic distances. The distance between a pair of isolates was defined to be the proportion of loci with differing genotypes with pairwise deletion of missing loci, as implemented in the R package ape (V5.0) [63]. Classical multidimensional scaling analysis (MDS), as described by [64], was performed on the pairwise distance matrix using cmdscale function in the base statistics package (V3.4.0) in R. Discriminant Analysis of Principal Components (DAPC) was also performed on the pairwise distance matrix using Adegenet (V2.1.1) [57], with isolates stratified into provincial populations. The number of principal components retained for DAPC was informed by the 'a-score' (defined to be the difference between observed and random discrimination) to optimise discrimination power and avoid excessive overfitting. Phylogenetic analysis of isolates collected during an outbreak in migrant workers was done by constructing Neighbour Joining trees using the BIONJ algorithm implemented in ape (V 5.0) [63]. Bootstrapping was performed to collapse nodes with bootstrapping support below 70% over 1000 replicates.

Pairwise relatedness (IBD sharing) between isolates was inferred using *IsoRelate* [44], setting the genetic distance to be 1cM=13.5kbp [65] and assuming a genotyping error rate of 0.001. Briefly, the posterior probability of IBD sharing at each locus was averaged for each pair of isolates to obtain the

expected fraction IBD (eIBD), [31]; posterior probability estimates output by *IsoRelate* were corrected by a factor of 2. Adjacency matrices for various eIBD thresholds (namely, 0.25, 0.30, 0.35, 0.45, 0.55 and 0.75) were used to construct IBD networks using the R package *igraph* (V1.2.4.1) [66]. As a measure of diversity, we calculated the proportions of pairs with eIBD sharing above 0.25 and 0.55 for each subpopulation. Bootstrapping (500 iterations) was performed over eIBD values to generate 95% confidence intervals for the proportion of pairs IBD.

Data Availability

All raw whole genome sequence data is available at the European Nucleotide Archive using the accession numbers indicated in Table S1. The final SNP barcode dataset for 727 isolates and for 32 outbreak samples will be made available in a public repository upon publication, and in the interim will be available by contacting the corresponding author.

Acknowledgements

We are grateful to the communities that participated in the study and the staff of the Papua New Guinea Institute of Medical Research that contributed to the field studies and sample collections. Dr. D. Hill and Dr. M. Hapsari Hazarin contributed technical assistance during the study and Drs. S. Volkman and N. Baro shared co-ordinates of uSNPs. Dr. M. Manske assisted with genomic sequencing logistics. Drs. K. McCann and C. Narh assisted with editing the manuscript. Sample collections, DNA extractions and molecular diagnosis were funded by the Global Fund to end AIDS Tuberculosis and Malaria (GFATM). Genetic and genomic data collection for this was funded through a National Health and Medical Research Council (NHMRC) of Australia Project Grant Number GNT1027108. IM and MB are supported by NHMRC Research Fellowships (GNT1155075, GNT1102971). The authors acknowledge the Victorian State Government Operational Infrastructure Support and Australian Government NHMRC Independent Research Institute Infrastructure Support Scheme (IRIISS).

References

1. Grad, Y.H. and M. Lipsitch, *Epidemiologic data and pathogen genome sequences: a powerful synergy for public health.* Genome Biol, 2014. **15**(11): p. 538.

2. Gardner, M.J., et al., *Genome sequence of the human malaria parasite Plasmodium falciparum.* Nature, 2002. **419**(6906): p. 498-511.

3. Kattenberg, J.H., et al., *Malaria Molecular Surveillance in the Peruvian Amazon with a Novel Highly Multiplexed Plasmodium falciparum AmpliSeq Assay.* Microbiol Spectr, 2023. **11**(2): p. e0096022.

4. Trimarsanto, H., et al., *A molecular barcode and web-based data analysis tool to identify imported Plasmodium vivax malaria.* Commun Biol, 2022. **5**(1): p. 1411.

5. Fola, A.A., et al., *SNP barcodes provide higher resolution than microsatellite markers to measure Plasmodium vivax population genetics.* Malar J, 2020. **19**(1): p. 375.

6. Chang, H.H., et al., *Mapping imported malaria in Bangladesh using parasite genetic and human mobility data.* Elife, 2019. **8**.

7. Coll, F., et al., *A robust SNP barcode for typing Mycobacterium tuberculosis complex strains.* Nat Commun, 2014. **5**: p. 4812.

8. Argyropoulos, D.C., et al., *Performance of SNP barcodes to determine genetic diversity and population structure of Plasmodium falciparum in Africa.* Front Genet, 2023. **14**: p. 1071896.

9. World Health Organization, *World Malaria Report 2022*. 2022: Geneva, Switzerland.

10. Anderson, T.J., et al., *Microsatellite markers reveal a spectrum of population structures in the malaria parasite Plasmodium falciparum.* Mol Biol Evol, 2000. **17**(10): p. 1467-82.

11. Miotto, O., et al., *Genetic architecture of artemisinin-resistant Plasmodium falciparum.* Nat Genet, 2015. **47**(3): p. 226-34.

12. MalariaGEN *P. falciparum* Community Project, *Genomic epidemiology of artemisinin resistant malaria.* Elife, 2016. **5**: p. e08714.

13. Auburn, S., et al., *Genomic analysis of a pre-elimination Malaysian Plasmodium vivax population reveals selective pressures and changing transmission dynamics.* Nat Commun, 2018. **9**(1): p. 2585.

14. Fola, A.A., et al., *Nationwide genetic surveillance of Plasmodium vivax in Papua New Guinea reveals heterogeneous transmission dynamics and routes of migration amongst subdivided populations.* Infect Genet Evol, 2018. **58**: p. 83-95.

15. Tessema, S.K., et al., *Using parasite genetic and human mobility data to infer local and cross-border malaria connectivity in Southern Africa.* Elife, 2019. **8**: p. e43510.

16. Daniels, R.F., et al., *Modeling malaria genomics reveals transmission decline and rebound in Senegal.* Proc Natl Acad Sci U S A, 2015. **112**(22): p. 7067-72.

17. Dalmat, R., et al., *Use cases for genetic epidemiology in malaria elimination.* Malar J, 2019. **18**(1): p. 163.

18. Neafsey, D.E., A.R. Taylor, and B.L. MacInnis, *Advances and opportunities in malaria population genomics.* Nat Rev Genet, 2021. **22**(8): p. 502-517.

19. Auburn, S. and A.E. Barry, *Dissecting malaria biology and epidemiology using population genetics and genomics.* Int J Parasitol, 2017. **47**(2-3): p. 77-85.

20. Li, Y., et al., *Dynamics of Plasmodium vivax populations in border areas of the Greater Mekong sub-region during malaria elimination.* Malar J, 2020. **19**(1): p. 145.

21. Kattenberg, J.H., et al., *Monitoring Plasmodium falciparum and Plasmodium vivax using microsatellite markers indicates limited changes in population structure after substantial transmission decline in Papua New Guinea.* Mol Ecol, 2020. **29**(23): p. 4525-4541.

22. Pringle, J.C., et al., *Genetic Evidence of Focal Plasmodium falciparum Transmission in a Pre-elimination Setting in Southern Province, Zambia.* J Infect Dis, 2019. **219**(8): p. 1254-1263.

23. Waltmann, A., et al., *Increasingly inbred and fragmented populations of Plasmodium vivax associated with the eastward decline in malaria transmission across the Southwest Pacific.* PLoS Negl Trop Dis, 2018. **12**(1): p. e0006146.

24. Shetty, A.C., et al., *Genomic structure and diversity of Plasmodium falciparum in Southeast Asia reveal recent parasite migration patterns.* Nat Commun, 2019. **10**(1): p. 2665.

25. Knudson, A., et al., *Spatio-temporal dynamics of Plasmodium falciparum transmission within a spatial unit on the Colombian Pacific Coast.* Sci Rep, 2020. **10**(1): p. 3756.

26. Obaldia, N., 3rd, et al., *Clonal outbreak of Plasmodium falciparum infection in eastern Panama.* J Infect Dis, 2015. **211**(7): p. 1087-96.

27. Kwiatkowski, D., *Malaria genomics: tracking a diverse and evolving parasite population.* Int Health, 2015. **7**(2): p. 82-4.

28. Wesolowski, A., et al., *Mapping malaria by combining parasite genomic and epidemiologic data.* BMC Med, 2018. **16**(1): p. 190.

29. Daniels, R., et al., *A general SNP-based molecular barcode for Plasmodium falciparum identification and tracking.* Malar J, 2008. **7**: p. 223.

30. Amambua-Ngwa, A., et al., *Long-distance transmission patterns modelled from SNP barcodes of Plasmodium falciparum infections in The Gambia.* Sci Rep, 2019. **9**(1): p. 13515.

31. Taylor, A.R., et al., *Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent.* PLoS Genet, 2017. **13**(10): p. e1007065.

32. Schaffner, S.F., et al., *hmmIBD: software to infer pairwise identity by descent between haploid genotypes.* Malar J, 2018. **17**(1): p. 196.

33. Taylor, A.R., et al., *Estimating Relatedness Between Malaria Parasites.* Genetics, 2019. **212**(4): p. 1337-1351.

34. Lo, E., et al., *Selection and utility of single nucleotide polymorphism markers to reveal fine-scale population structure in human malaria parasite plasmodium falciparum.* Frontiers in Ecology and Evolution, 2018. **6**: p. 145.

35. Hetzel, M.W., et al., *Prevalence of malaria across Papua New Guinea after initial roll-out of insecticide-treated mosquito nets.* Trop Med Int Health, 2015. **20**(12): p. 1745-55.

36. Hetzel, M.W., et al., *Changes in malaria burden and transmission in sentinel sites after the roll-out of long-lasting insecticidal nets in Papua New Guinea.* Parasit Vectors, 2016. **9**(1): p. 340.

37. Kattenberg, J.H., et al., *The epidemiology of Plasmodium falciparum and Plasmodium vivax in East Sepik Province, Papua New Guinea, pre- and post-implementation of national malaria control efforts.* Malar J, 2020. **19**(1): p. 198.

38. Koepfli, C., et al., *Sustained Malaria Control Over an 8-Year Period in Papua New Guinea: The Challenge of Low-Density Asymptomatic Plasmodium Infections.* J Infect Dis, 2017. **216**(11): p. 1434-1443.

39. Health, P.M.o. *Papua New Guinea National Malaria Strategic Plan (2014-2018)*. 2014.

40. Schultz, L., et al., *Multilocus haplotypes reveal variable levels of diversity and population structure of Plasmodium falciparum in Papua New Guinea, a region of intense perennial transmission.* Malar J, 2010. **9**: p. 336.

41. Jennison, C., et al., *Plasmodium vivax populations are more genetically diverse and less structured than sympatric Plasmodium falciparum populations.* PLoS Negl Trop Dis., 2015. **9**.

42. Tessema, S.K., et al., *Antibodies to Intercellular Adhesion Molecule 1-Binding Plasmodium falciparum Erythrocyte Membrane Protein 1-DBLβ Are Biomarkers of Protective Immunity to Malaria in a Cohort of Young Children from Papua New Guinea.* Infect Immun, 2018. **86**(8).

43. MalariaGEN, et al., *Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples.* Wellcome Open Res, 2023. **8**: p. 22.

44. Henden, L., et al., *Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens.* PLoS Genet, 2018. **14**(5): p. e1007279.

45. MalariaGEN, *The Pf3K Project: pilot data release 5*. 2016: www.malariagen.net/data/pf3k-5.

46. Morin, P.A., K.K. Martien, and B.L. Taylor, *Assessing statistical power of SNPs for population structure and conservation studies.* Mol Ecol Resour, 2009. **9**(1): p. 66-73.

47. Miotto, O., et al., *Emergence of artemisinin-resistant Plasmodium falciparum with kelch13 C580Y mutations on the island of New Guinea.* PLoS Pathog, 2020. **16**(12): p. e1009133.

48. Manning, L., et al., *Severe anemia in Papua New Guinean children from a malaria-endemic area: a case-control etiologic study.* PLoS Negl Trop Dis, 2012. **6**(12): p. e1972.

49. Tavul, L., et al., *Efficacy of artemether-lumefantrine and dihydroartemisinin-piperaquine for the treatment of uncomplicated malaria in Papua New Guinea.* Malar J, 2018. **17**(1): p. 350.

50. Manske, M., et al., *Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing.* Nature, 2012. **487**(7407): p. 375-9.

51. Venkatesan, M., et al., *Using CF11 cellulose columns to inexpensively and effectively remove human DNA from Plasmodium falciparum-infected whole blood samples.* Malar J, 2012. **11**: p. 41.

52. Kasehagen, L.J., et al., *Changing patterns of Plasmodium blood-stage infections in the Wosera region of Papua New Guinea monitored by light microscopy and high throughput PCR diagnosis.* Am J Trop Med Hyg, 2006. **75**(4): p. 588-96.

53. Fola, A.A., et al., *Higher Complexity of Infection and Genetic Diversity of.* Am J Trop Med Hyg, 2017. **96**(3): p. 630-641.

54. Falk, N., et al., *Comparison of PCR-RFLP and Genescan-based genotyping for analyzing infection dynamics of Plasmodium falciparum.* Am J Trop Med Hyg, 2006. **74**(6): p. 944-50.

55. Auburn, S., et al., *Characterization of within-host Plasmodium falciparum diversity using next-generation sequence data.* PLoS One, 2012. **7**(2): p. e32891.

56. R Core Team, *R: A language and environment for statistical computing*. 2013, R Foundation for Statistical Computing: https://www.R-project.org/.

57. Jombart, T., *adegenet: a R package for the multivariate analysis of genetic markers.* Bioinformatics, 2008. **24**(11): p. 1403-5.

58. Nei, M., *Molecular Evolutionary Genetics*. 1987: Colombia University Press.

59. Paradis, E., *pegas: an R package for population genetics with an integrated-modular approach.* Bioinformatics, 2010. **26**(3): p. 419-20.

60. Goudet, J., *Hierfstat, a package for R to compute and test hierarchical F-statistics.* Molecular Ecology Notes, 2005. **5**(1): p. 184-186.

61. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-59.

62. Evanno, G., S. Regnaut, and J. Goudet, *Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study.* Mol Ecol, 2005. **14**(8): p. 2611-20.

63. Paradis, E. and K. Schliep, *ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.* Bioinformatics, 2019. **35**(3): p. 526-528.

64. Gower, J.C., *Some distance properties of latent root and vector methods used in multivariate analysis.* Biometrika, 1966. **53**(3--4): p. 325--338.

65. Miles, A., et al., *Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum.* Genome Res, 2016. **26**(9): p. 1288-99.

66. Csardi, G. and T. Nepusz, *The igraph software package for complex network research.* InterJournal, 2006. **Complex Systems**: p. 1695.