A molecular barcode and online tool to identify and map imported infection with *Plasmodium vivax*

Hidayat Trimarsanto[1,2], Roberto Amato[3,4], Richard D Pearson[3,4], Edwin Sutanto[1], Rintis Noviyanti[1], Leily Trianty[1], Jutta Marfurt[5], Zuleima Pava[5], Diego F Echeverry[6,7], Tatiana M Lopera-Mesa[8], Lidia Madeline Montenegro[8], Alberto Tobón-Castaño [8], Matthew J Grigg[5,9], Bridget Barber[5,9], Timothy William[9,10], Nicholas M Anstey[5], Sisay Getachew[11,12], Beyene Petros[11], Abraham Aseffa[12], Ashenafi Assefa[13], Awab Ghulam Rahim[14,15], Nguyen Hoang Chau[16], Tran Tinh Hien[16], Mohammad Shafiul Alam[17], Wasif A Khan[17], Benedikt Ley[5], Kamala Thriemer[5], Sonam Wangchuck[18], Yaghoob Hamedi[19], Ishag Adam[20], Yaobao Liu[21,22], Qi Gao[22], Kanlaya Sriprawat[23], Marcelo U Ferreira[24], Alyssa Barry[25,26, 27,28], Ivo Mueller[28,29], Eleanor Drury[4], Sonia Goncalves[4], Victoria Simpson[3,4], Olivo Miotto[3,4,13], Alistair Miles[3,4], Nicholas J White[13,30], Francois Nosten[13,23,30], Dominic P Kwiatkowski[3,4], Ric N Price*[5,13,30], Sarah Auburn *[5,13,30]

1. Eijkman Institute for Molecular Biology, Jakarta 10430, Indonesia
2. Agency for Assessment and Application of Technology, Jl. MH Thamrin 8, Jakarta 10340, Indonesia
3. Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Old Road Campus, Oxford, OX3 7LF, UK
4. Wellcome Sanger Institute, Hinxton, Cambridge CB10 1SA, UK
5. Global and Tropical Health Division, Menzies School of Health Research and Charles Darwin University, Darwin, Northern Territory 0811, Australia
6. Centro Internacional de Entrenamiento e Investigaciones Medicas, CIDEIM, Cali, Colombia
7. Universidad Icesi, Cali, Colombia
8. Grupo Malaria, Facultad de Medicina, Universidad de Antioquia, Medellin, Colombia

9. Infectious Diseases Society Sabah-Menzies School of Health Research Clinical Research Unit, Kota Kinabalu, Sabah, Malaysia
10. Clinical Research Centre, Queen Elizabeth Hospital, Sabah, Malaysia
11. College of Natural Sciences, Addis Ababa University, P.O. Box 52, Addis Ababa, Ethiopia
12. Armauer Hansen Research Institute, P.O. Box 1005, Jimma Road, Addis Ababa, Ethiopia
13. Ethiopian Public Health Institute, Addis Ababa, Ethiopia
14. Mahidol-Oxford Tropical Medicine Research Unit, Mahidol University, Bangkok 10400, Thailand
15. Nangarhar Medical Faculty, Nangarhar University, Ministry of Higher Education, Afghanistan
16. Centre for Tropical Medicine, Oxford University Clinical Research Unit, 764 Vo Van Kiet, W.1, Dist.5, Ho Chi Minh City, Vietnam
17. Infectious Diseases Division, International Centre for Diarrheal Diseases Research, Bangladesh Mohakhali, Dhaka, 1212, Bangladesh
18. Royal Center for Disease Control, Department of Public Health, Ministry of Health, Thimphu, Bhutan
19. Infectious and Tropical Diseases Research Center, Hormozgan University of Medical Sciences, Bandar Abbas, Hormozgan Province, Iran
20. Faculty of Medicine, University of Khartoum, P.O. Box 102, Khartoum, Sudan
21. Medical College of Soochow University, Suzhou, Jiangsu, People's Republic of China
22. Key Laboratory of National Health and Family Planning Commission on Parasitic Disease Control and Prevention, Jiangsu Provincial Key Laboratory on Parasite and Vector Control Technology, Jiangsu Institute of Parasitic Diseases, Wuxi, Jiangsu, People's Republic of China
23. Shoklo Malaria Research Unit, Mae Sot, Tak 63110, Thailand
24. Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil
25. School of Medicine, Deakin University, Geelong, VIC, Australia

26. Burnet Institute, Melbourne, VIC, Australia
27. Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia
28. Department of Medical Biology, The University of Melbourne, Parkville, VIC, Australia
29. Department of Parasites and Insect Vectors, Institut Pasteur, Paris, France
30. Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, OX3 7LJ, UK

*Corresponding author: Dr Sarah Auburn, Menzies School of Health Research, PO Box 41096, Casuarina, Darwin, NT 0811, Australia; Tel: (+61) 8 8946 8503

Abstract

Imported cases present a considerable challenge to the elimination of malaria. Traditionally, patient travel history has been used to identify imported cases, but the long-latency liver stages confound this approach in *Plasmodium vivax*. Molecular tools to identify and map imported cases offer a more robust approach, that can be combined with drug resistance and other surveillance markers in high-throughput, population-based genotyping frameworks. Using a machine learning approach incorporating hierarchical FST (HFST) and decision tree (DT) analysis applied to 831 *P. vivax* genomes from 20 countries, we identified a 28-Single Nucleotide Polymorphism (SNP) barcode with high capacity to predict the country of origin. The Matthews correlation coefficient (MCC), which provides a measure of the quality of the classifications, ranging from -1 (total disagreement) to 1 (perfect prediction), exceeded 0.9 in 15 countries in cross-validation evaluations. When combined with an existing 37-SNP *P. vivax* barcode, the 65-SNP panel exhibits MCC scores exceeding 0.9 in 17 countries with up to 30% missing data. As a secondary objective, several genes were identified with moderate MCC scores

(median MCC range from 0.54-0.68), amenable as markers for rapid testing using low-throughput genotyping approaches. A likelihood-based classifier framework was established, that supports analysis of missing data and polyclonal infections. To facilitate investigator-lead analyses, the likelihood framework is provided as a web-based, open-access platform (vivaxGEN-geo) to support the analysis and interpretation of data produced either at the 28-SNP core or full 65-SNP barcode. These tools can be used by malaria control programs to identify the main reservoirs of infection so that resources can be focused to where they are needed most.

Keywords

Plasmodium vivax, malaria, imported malaria, geographic origin, surveillance, genotyping, genomics, molecular barcode

Background

The last three World Malaria Reports have revealed a disturbing rise in malaria cases, and, outside Subsaharan Africa, an increasing proportion of malaria due to *Plasmodium vivax*, undermining the painstaking efforts to reduce transmission over the past decade [1]. These trends highlight the urgent need for new surveillance tools, with greater attention to the non-falciparum species. In today's global climate, human populations are highly mobile, with imported cases of malaria confounding local control efforts and enhancing the risks of drug resistance spread and outbreaks. There is thus a critical need to develop tools that can help to determine where patients acquired their infection.

The challenge of imported infections is particularly pertinent for *P. vivax*, in view of the parasite's ability to form dormant liver stages (hypnozoites) that can reactivate weeks to months after the initial infection, as well as highly persistent, low density blood-stage infections [2,3].The re-emergence of *P. vivax* in multiple regions where it was once almost eliminated serves as an important reminder of the need to maintain diligent surveillance of this species [4]. In low endemic settings where the incidence of local infections is declining, the relative proportion of imported cases generally rises, emphasizing the importance for surveillance tools that can identify imported *P. vivax* cases. Traditionally, imported cases have been identified and mapped using information on patient travel history, but the persistent blood stage infections and long-latency liver stages constrain the accuracy of this approach in *P. vivax* infections. Molecular tools to identify and map imported *P. vivax* cases offer an attractive complement to traditional epidemiological tools.

Amplicon-based sequencing has become a favored approach for targeted genotyping of malaria parasites. Using highly parallel sequencing platforms, such as the latest generation of Illumina sequencers, amplicon-based sequencing can be applied at moderate to high-throughput, with high accuracy and sensitivity. These platforms are flexible, allowing iterative enhancement of the Single Nucleotide Polymorphism (SNP) barcodes, which can provide an affordable genotyping approach, amenable to population-based molecular surveillance.

Previous studies have used mitochondrial and apicoplast markers to resolve imported from local *P. vivax* isolates, but the resolution of these organelles is constrained[5-7]. In 2015, a panel of 42 SNPs was identified to facilitate parasite finger-printing and geographic assignment [8]. The proposed 42-SNP Broad barcode was derived from genomic data available from 13 isolates from 7 countries. In the last 4 years, the repository of genomic data on *P. vivax* has expanded greatly, allowing further refinement of a parsimonious and widely applicable genotyping barcode.

The primary objective of our study was to develop molecular tools for identifying and characterizing imported *P. vivax* cases amenable to population-based surveillance frameworks, so that these data can be used to inform strategic decisions on where and how to deploy malaria control interventions. We tailored our molecular tools primarily to surveillance frameworks using Illumina or other high-throughput genotyping platforms. As a secondary objective, we sought to identify single gene regions permissible to lower throughput approaches for use in settings or situations where high-throughput or centralized approaches are not feasible. In addition, we provide informatics tools to support users in analyzing genotyping data produced at the barcode that can accommodate missing data and polyclonal infections.

Materials and Methods

Overview of the marker selection approach

A flow diagram outlining the steps involved in the marker selection process is provided in Figure 1a. In accordance with the multiplexing features of the Illumina platform, we sought to identify approximately 50 new SNP-based markers to append to the recently published Broad barcode[8], to provide a composite panel with ≤100 markers for country-level geographic assignment of *P. vivax* infections. The decision to append markers to the Broad barcode rather than selecting a de novo panel of SNPs was pragmatic, aimed at promoting consensus and continuity with existing molecular tools available to the vivax community. A likelihood-based classifier approach was chosen for the respective evaluation of marker sets and end-user data analysis tasks. This approach was chosen since it allows manual addition of specific SNP sets, such as the Broad barcode. Two selector algorithms, hierarchical FST (HFST) and decision-tree (DT), were implemented in the likelihood-based classifier framework to select SNPs with high country-level prediction values. The primary SNP selection method was the HFST selector, which leverages on the prior knowledge of the population structure to inform on a relatively parsimonious SNP

set with moderately high prediction. The DT selector, the secondary method, was used to select additional SNPs to append to the Broad barcode and the HFST panel for further enhancement of geographical prediction. A 10-fold cross-validation strategy was used to assess the performance of the selectors with the likelihood-based classification framework.

To achieve the secondary objective of the study, identifying single gene regions with moderate-to-high country-level resolution, simulations were run across individual genes using the HFST-0.75 (HFST with FST threshold of 0.75) selector model with the likelihood classifier. The top 20 genes with the highest pooled median Matthew Correlation Coefficient (MCC) scores for each selector model were reported (Figure 1b).

Overview of the web-based data analysis and sharing platform

To establish accessible informatics tools to support users to analyze and interpret their data, a platform was created incorporating data classification tools for determining the most likely country of origin of a sample using genetic data at a given barcode. Existing source code, developed for a microsatellite-based *P. vivax* data sharing and analysis platform [9], was modified to create a new web-based platform (vivaxGEN-geo), to collate SNP data generated at the geographic barcode. A likelihood-based classifier approach was chosen for geographic assignment within the vivaxGEN-geo platform owing to the ability to i) incorporate manual SNP sets, ii) evaluate barcodes with missing data, and iii) evaluate heterozygous genotype calls.

Likelihood-based classifier framework

The custom classifier was developed to handle bi-allelic heterozygote calls for mix-infection cases by treating the samples as diploid samples, as well as missing data by treating as heterozygote calls. The

classifier was derived from Bernoulli Naive Bayes with modification to the likelihood equation and elimination of prior probability, since the distribution of our dataset did not reflect the distribution in nature, but rather the implication from sample and extracted DNA quality, as well as the characteristics of the original study such as duration and type of the study. Hence the classifier only depends on the likelihood of the SNP data. The likelihood equation was modified to handle the bi-allelic data as follows:

$$\boldsymbol{L(X \mid Ck)} = \prod_i^n \boldsymbol{p_{ki}^{x_i} \cdot (1 - p_{ki})^{(2-x_i)}}$$

where $\boldsymbol{X}$ is the SNP data set of a sample, Ck is a group (or a country), x_i is the number of alternate alleles at position i and p_{ki} is the frequency of the alternate allele at position i of country k counted as diploid samples.

SNP Selection

To select optimal SNPs for country classification, a combination of the HFST and DT selector methods were employed. The DT selector utilized the Python-based scikit-learn package for the decision tree implementation, which employed an optimized version of the CART (Classification And Regression Tree) algorithm and Gini coefficient. To avoid overfitting, a minimum of 3 samples was required for a leaf node. The Hierarchical FST (HFST) approach worked by traversing across a bifurcating guide tree and selecting SNP with the highest FST between the two populations represented by the two nodes of the branch with the assumption that the SNP with the highest FST might differentiate those two populations. If the highest FST of a certain branch was lower than a given threshold during guide tree traversal, the DT method was employed to obtain additional SNPs to separate the given branch. To avoid overfitting, a maximum tree depth of 2 was set for this particular DT step. The HFST analysis in this study was undertaken using a guide tree constructed using Nei's population distance matrix implemented with a neighbor-joining algorithm.

187 The classification performance was measured with MCC for each country [10]. In addition, the pooled
188 median, mean, minimum and first-quartile MCC were collected as additional measurements.

189 Three models, HFST-0.90 (HFST and DT with FST threshold of 0.90), HFST-0.95 (HFST and DT with FST
190 threshold of 0.95), and pure DT were trained with the full dataset. For each of the three models, 500
191 repeats were run to allow for different random seeds of the DT analysis, and the top 25 SNP sets with
192 the highest aggregate minimum MCC score as evaluated by the likelihood classifier were obtained from
193 each model. A stratified 100 repeat, 10-fold cross-validation was run on each of the 25 SNP sets from
194 each model, and the best SNP set from each of model, as indicated by highest aggregate minimum MCC
195 score within a repeat, was selected. To compare the Broad SNP panel to the three new SNP panels
196 identified by the HFST-0.90, HFST-0.95 and pure DT selectors, a 500 repeat, stratified 10-fold cross-
197 validation was undertaken on each SNP panel.

198

199 **Missing data evaluation**

200 To assess the durability of prediction performance of the SNP sets with missing data, a simulation was
201 run by removing genotype data randomly. The Likelihood classifier was trained against the selected SNP
202 sets using all samples. For each country, 25 samples were sampled randomly with replacement and
203 missing genotype calls were added to the SNP sets in 10%, 20% and 30% proportions. The random
204 samples were then subjected to the trained classifier. This process was run in 100 repeats and MCC-
205 score of the prediction for each country was reported.

206

207 **Datasets**

208 The analysis included genomic data on *P. vivax* isolates collected from 26 countries. Published data were
209 included from 19 countries derived from the European Nucleotide Archive[11-15], and new data from 10
210 countries (Supplementary Table 1, Supplementary Figure 1). New genomic data were derived from

patients recruited to partner studies in Afghanistan, Bangladesh, Bhutan, Colombia, Ethiopia, Indonesia, Iran, Malaysia, Sudan, and Vietnam. With the exception of Colombia, the patient sampling frameworks have been described previously [11,12,14,16-20]. The samples from Colombia were collected within the framework of cross-sectional epidemiological surveys conducted between 2013 and 2017. Whole genome sequencing, read alignment and variant calling were undertaken within the framework of a *P. vivax* community study in the Malaria Genomic Epidemiology Network (MalariaGEN) [21]. Data was derived from an open dataset of *Plasmodium vivax* genome variation comprising 2,671,112 discovered variants across 1,366 isolates (MalariaGEN manuscript in preparation). The data were initially filtered to derive a set of 670,962 high-quality bi-allelic SNPs with VQSLOD score >0, and minimum read depth and minimum minor allele count of 2. Individual genotype calls were defined as heterozygotes based on an arbitrary threshold of a minor allele ratio > 0.1 and a minimum of 2 reads for each allele; all other genotype calls were defined as homozygous for the major allele. A pair of isolates with distance less than 0.0005 (0.05%) were considered non-independent. Amongst non-independent sample pairs, the isolate with the higher percentage of genotype failures was removed from the dataset; this removal process was iterated until all non-independent isolates had been removed from the dataset. The samples and SNPs were then subjected to further filtering to eliminate missing data using information derived from a simulation which calculated the total number of SNPs with no missing data and the number of consecutive informative SNPs as defined by SNPs with minimum minor allele count (MAC) >2. The remaining dataset was defined as Dataset 1.

The isolates in Dataset 1 were initially assigned to national groups based on the country in which the patient presented at the clinic with the infection. The national-level groupings were evaluated further using country-level assignments derived from the genome-wide data classification with the likelihood classifier. Infections presenting with country classifications differing from the country of presentation were considered suspected imported infections and removed to produce Dataset 2.

Of the 42 Broad barcode SNPs, 37 SNPs were present in the 670K dataset (bi-allelic high-quality SNPs) and exhibited successful amplicon-based sequencing assays (personal communication, Wellcome Sanger Institute Core Sequencing Facility); these 37 SNPs were not present in dataset 1 or 2. A new dataset (Dataset 3) was prepared for evaluation of the Broad barcode comprising of samples with complete data across the 37 Broad barcode SNPs and partial data across SNPs selected from the HFST and DT algorithms.

Software and Web Service Availability

All custom, in-house scripts used for data filtering, simulation, analyses and visualization are available from http://github.com/trmznt/vivaxgen-geo. The VivaxGEN-geo web service provides a user-friendly online tool for country classification with all SNP sets, and is accessible at https://geo.vivaxgen.org/. The likelihood classifier provided on the online platform has been trained with 809 samples (dataset 4), consisting of all samples with complete data at all SNP sets. The classifier tool reports the three highest likelihoods for country of origin and their associated probabilities.

Ethics

All samples were collected with written informed consent from patients or their legal guardians. Ethical approvals for the published samples are detailed in the original papers [11-15]. Approvals for the newly represented studies are outlined in Supplementary Document 1.

Results

Geographic clustering patterns using the genome-wide dataset

The primary dataset (Dataset 1) was derived using the missing data simulation, to minimise genotype failures (Supplementary Figure 3), it comprised 854 high-quality samples and 294,628 high-quality

informative SNPs, with no missing data. The median percentage of heterozygous calls in each country ranged from 0.02% to 0.08%. Details on the geographic locations of the samples in dataset 1 are presented in Supplementary Table 1. Neighbor-joining analysis on dataset 1 revealed distinct geographic clustering of most countries (Supplementary Figure 4). Exceptions included the isolates from Afghanistan, Iran, India and Sri Lanka, which appeared to form a single cluster, warranting further analysis of this geographic region with larger sample sets. Although several isolates in border regions including Vietnam relative to Cambodia, and Thailand relative to Myanmar, overlapped between countries, the majority of isolates in these countries could be differentiated by national boundaries. Visual inspection of the neighbour-joining tree revealed potential imported cases. Using country-level assignments derived from the genome-wide data classification with the likelihood classifier and manual confirmation of the neighbor-joining tree, 21 isolates presented country classifications differing from the country of presentation (Supplementary Table 1). After exclusion of the imported cases, and countries represented by a single sample, a total of 831 isolates and 20 countries remained, constituting Dataset 2 (Supplementary Table 1).

Performance of the Broad barcode, HFST and DT SNP selection

When the HFST selector was applied with an FST threshold of 0.90 (HFST-0.90), a set of 28 SNPs (listed in Supplementary Table 2) were identified. This dataset exhibited median MCC scores exceeding 0.9 in all countries with the exception of Vietnam (0.75) and Cambodia (0.80). On increasing the FST threshold to 0.95 (HFST-0.95), the HFST model identified 51 SNPs (listed in Supplementary Table 3), which displayed MCC scores exceeding 0.95 in all countries except for Vietnam (0.85) and Cambodia (0.87). Using the DT selector alone, 50 SNPs (listed in Supplementary Table 4) displayed comparable performance to the 51-SNP panel, with a slightly lower aggregate minimum MCC score.

The results of cross-validation of the classification performance of the five SNP panels (37-SNP Broad barcode, 28-SNP, 28-SNP plus Broad barcode (65-SNP), 50-SNP and 51-SNP panels) are illustrated in

Figure 2, and the MCC and F scores reflecting the consensus results of the cross-validation are summarized in Table 1. The performance, ranked from lowest to highest, was: 37-SNP Broad barcode (median MCC = 0.82), 28-SNP (MCC = 0.99), 50-SNP (MCC = 1.00), 65-SNP (MCC = 1.00), and 51-SNP (MCC = 1.00).

Missing data simulations

In the missing data simulations, genotyping failures had the greatest impact on the classification of samples from Cambodia and Vietnam (Figure 3). With 10% missing data, the median MCCs of the 28-SNP panel exceeded 0.9 in all countries, with exception of Vietnam (MCC = 0.80) and Cambodia (MCC = 0.77). With this level of missing data, the addition of the 37 Broad SNPs (65-SNP panel) increased the median MCC to 0.83 in Vietnam and 0.82 in Cambodia. When missing data increased to 30%, the 65-SNP panel achieved median MCCs above 0.9 in most countries, with exception of Vietnam (MCC = 0.79) and Cambodia (MCC = 0.75). The 50- and 51-SNP panels both achieved MCC scores exceeding 0.95 for all countries except Cambodia (0.80-0.82) and Vietnam (0.83-0.85) with 10-30% missing data.

Evaluation of single gene regions to predict country classification

The suitability of single genes to predict country classifications were assessed by simulations of individual genes using HFST-0.75 selector model with the likelihood classifier framework. The top 20 genes with the highest pooled median MCC scores for the HFST-0.75 are presented in Supplementary Table 5. The highest prediction capacity, with median MCC score of 0.68, was PVP01_0302600, a gene coding a 11.5 Kb conserved protein with unknown function. The gene list also included three members of the *cysteine repeat modular protein family* (CRMP): CRMP1 (median MCC = 0.63), CRMP3 (MCC = 0.57) and CRMP4 (MCC = 0.56).

Discussion

The primary objective of the study was to develop molecular tools amenable to population-based surveillance frameworks to identify and map imported *P. vivax* infections. Using machine-learning methods, 3 new SNP panels were identified with high country classification performance, able to distinguish imported *P. vivax* infections across a range of endemic scenarios. The most parsimonious panel, the 28-SNP barcode, exhibited high country classification, and can be appended to the 37 bi-allelic, assayable Broad barcode SNPs for marginal improvement in predictive capacity in samples with moderate levels of missing data. The combined 65-SNP barcode generated robust country classification in most endemic areas, even when the proportion of missing data rose to 30%. However, the validity of the 65-SNP barcode was lower in Cambodia and Vietnam, a likely reflection of the porous border between these two countries. Although the 50- and 51-SNP panels achieved better resolution in these areas, characterization of parasite transmission across borders with high levels of gene flow may be addressed better by the addition of markers suited to an analysis of identity-by-descent [22]. The application and wider validation of the 65-SNP barcode is underway, with amplicon-based sequencing assays already established for the 37 Broad barcode SNPs, and under development for the 28 new markers.

The analysis and interpretation of "real-world" genotyping data raises significant challenges from low-quality samples such as those collected on dried blood spots. In anticipation of these needs we established a likelihood-based framework with the capacity to deal with polyclonal infections as well as missing data. This framework has been integrated into the vivaxGEN-geo online platform, so that users can analyze and interpret their data without needing complex bioinformatics skills and avoiding the subjective visual inspection of neighbour-joining trees or principal component plots. Whilst the informatics tools implemented in vivaxGEN-geo are tailored to *P. vivax*, we anticipate that a similar approach can be adapted to other species. To facilitate wider application the source code will be made publicly available.

The variants in the 28-SNP panel are located in genes representing a range of functions, some of which may be unstable over time. Although our dataset represents one of the most geographically diverse panels of *P. vivax* isolates currently available, with representation of all of the major vivax-endemic regions, the predictive capacity of the derived tools are likely to be constrained by the geographic representation of the reference panel. In particular, representation from central and south America and the Indian subcontinent were limited in our data set. Despite this limitation the dataset used comprises good representation of isolates from areas of public health relevance, including the epicenter of chloroquine-resistant *P. vivax* in Papua Indonesia [23,24]. The likelihood-based framework is able to re-evaluate the predictive potential of current marker sets as new genomic data become available, so that the selected SNP panels can be refined further in an iterative process. Furthermore, as the reference panel expands with increasing data generated at the barcode SNPs, the accuracy of the likelihood-based classifications will improve.

In addition to the independent selection of SNPs, a number of informative genes were identified, each of which had moderately high geographic resolution power. Genotyping of these genes or gene regions are amenable to standard capillary sequencing, offering an alternative approach, albeit with slightly lower resolution, to define a parasite's geographic origin in settings where high-throughput genotyping facilities are unavailable. The genes with the greatest geographic resolution, included members of the *cysteine repeat modulator protein* (CRMP) family (CRMP1, CRMP3 and CRMP4) implicated in essential roles in parasite transmission from the mosquito to the human host [25]. It is plausible that the CRMP genes have maintained high geographic differentiation to ensure parasite adaptation to the local vector species. Although adaptations of these genes are likely to temporally stable, the resolution of these loci may be constrained by the distributions of host Anopheles vector species.

In 2017, up to 100% of all confirmed malaria cases in 17 malaria-endemic countries in the Asia-Pacific region, the Middle East and the Americas, where *P. vivax* infections predominate, were reported as being infections [1]. Malaria control programs in these countries can utilize the information derived from

the molecular tools provided from our analysis to assess the efficacy of ongoing interventions in reducing local transmission, and to determine the major routes of infection importation. The tools have potential to reduce ambiguity for certificating malaria elimination by the World Health Organization, where one of the key requirements is the demonstration that all malaria cases detected in-country over at least three consecutive years were imported. For this purpose, countries approaching elimination will need to maintain archival samples for future molecular comparisons against putatively imported cases.

The molecular *P. vivax* geographic classification tools presented are designed to empower users in malaria-endemic countries to analyze and interpret locally produced genotyping data with comparison to globally available datasets. Amplicon-based sequencing of the full 65-SNP barcode is being developed and will be combined with other surveillance markers at central laboratories in endemic partner countries of the Asia Pacific Malaria Elimination Network (www.apmen.org). The data generated from these centers will inform researchers, National Malaria Control Programs and other key stakeholders on the incidence, epidemiology and key reservoirs of imported malaria, and, in doing so, help to target resources to where they are needed most.

References

1. WHO. World Malaria Report 2018. World Health Organization; Geneva 2018. (2018).
2. White, N.J. & Imwong, M. Relapse. *Adv Parasitol* **80**, 113-50 (2012).
3. Tripura, R. *et al.* Persistent Plasmodium falciparum and Plasmodium vivax infections in a western Cambodian population: implications for prevention, treatment and elimination strategies. *Malar J* **15**, 181 (2016).
4. Sattabongkot, J., Tsuboi, T., Zollner, G.E., Sirichaisinthop, J. & Cui, L. Plasmodium vivax transmission: chances for control? *Trends Parasitol* **20**, 192-8 (2004).
5. Iwagami, M. *et al.* Geographical origin of Plasmodium vivax in the Republic of Korea: haplotype network analysis based on the parasite's mitochondrial genome. *Malar J* **9**, 184 (2010).
6. Rodrigues, P.T. *et al.* Using mitochondrial genome sequences to track the origin of imported Plasmodium vivax infections diagnosed in the United States. *Am J Trop Med Hyg* **90**, 1102-8 (2014).
7. Diez Benavente, E. *et al.* Genomic variation in Plasmodium vivax malaria reveals regions under selective pressure. *PLoS One* **12**, e0177134 (2017).

8. Baniecki, M.L. *et al.* Development of a single nucleotide polymorphism barcode to genotype Plasmodium vivax infections. *PLoS Negl Trop Dis* **9**, e0003539 (2015).
9. Trimarsanto, H. *et al.* VivaxGEN: An open access platform for comparative analysis of short tandem repeat genotyping data in Plasmodium vivax Populations. *PLoS Negl Trop Dis* **11**, e0005465 (2017).
10. Jurman, G., Riccadonna, S. & Furlanello, C. A comparison of MCC and CEN error measures in multi-class prediction. *PLoS One* **7**, e41882 (2012).
11. Auburn, S. *et al.* Genomic analysis of a pre-elimination Malaysian Plasmodium vivax population reveals selective pressures and changing transmission dynamics. *Nat Commun* **9**, 2585 (2018).
12. Auburn, S. *et al.* Genomic analysis of Plasmodium vivax in southern Ethiopia reveals selective pressures in multiple parasite mechanisms. *J Infect Dis* (2019).
13. Hupalo, D.N. *et al.* Population genomics studies identify signatures of global dispersal and drug resistance in Plasmodium vivax. *Nat Genet* **48**, 953-8 (2016).
14. Pearson, R.D. *et al.* Genomic analysis of local variation and recent evolution in Plasmodium vivax. *Nat Genet* **48**, 959-64 (2016).
15. Parobek, C.M. *et al.* Selective sweep suggests transcriptional regulation may underlie Plasmodium vivax resilience to malaria control measures in Cambodia. *Proc Natl Acad Sci U S A* **113**, E8096-E8105 (2016).
16. Wangchuk, S. *et al.* Where chloroquine still works: the genetic make-up and susceptibility of Plasmodium vivax to chloroquine plus primaquine in Bhutan. *Malar J* **15**, 277 (2016).
17. Ley, B. *et al.* G6PD Deficiency and Antimalarial Efficacy for Uncomplicated Malaria in Bangladesh: A Prospective Observational Study. *PLoS One* **11**, e0154015 (2016).
18. Hamedi, Y. *et al.* Molecular Epidemiology of P. vivax in Iran: High Diversity and Complex Sub-Structure Using Neutral Markers, but No Evidence of Y976F Mutation at pvmdr1. *PLoS One* **11**, e0166124 (2016).
19. Getachew, S. *et al.* Variation in Complexity of Infection and Transmission Stability between Neighbouring Populations of Plasmodium vivax in Southern Ethiopia. *PLoS One* **10**, e0140780 (2015).
20. Taylor, W.R.J. *et al.* Short-course primaquine for the radical cure of Plasmodium vivax malaria: a multicentre, randomised, placebo-controlled non-inferiority trial. *Lancet* (2019).
21. Malaria Genomic Epidemiology, N. A global network for investigating the genomic epidemiology of malaria. *Nature* **456**, 732-7 (2008).
22. Taylor, A.R., Jacob, P.E., Neafsey, D.E. & Buckee, C.O. Estimating Relatedness Between Malaria Parasites. *Genetics* **212**, 1337-1351 (2019).
23. Ratcliff, A. *et al.* Therapeutic response of multidrug-resistant Plasmodium falciparum and P. vivax to chloroquine and sulfadoxine-pyrimethamine in southern Papua, Indonesia. *Trans R Soc Trop Med Hyg* **101**, 351-9 (2007).
24. Price, R.N. *et al.* Global extent of chloroquine-resistant Plasmodium vivax: a systematic review and meta-analysis. *Lancet Infect Dis* **14**, 982-91 (2014).
25. Douradinha, B. *et al.* Plasmodium Cysteine Repeat Modular Proteins 3 and 4 are essential for malaria parasite transmission from the mosquito to the host. *Malar J* **10**, 71 (2011).
26. Battle, K.E. *et al.* Mapping the global endemicity and clinical burden of Plasmodium vivax, 2000-17: a spatial and temporal modelling study. *Lancet* **394**, 332-343 (2019).

Acknowledgements

We would like to thank the patients who contributed their samples to the study, local health facilities, and the health workers and field teams who assisted with the sample collections. We also thank the staff of the Wellcome Sanger Institute Sample Logistics, Sequencing, and Informatics facilities for their contributions.

Financial Support

Financial support for the study was provided by the Wellcome Trust (Senior Fellowship in Clinical Science awarded to R.N.P., 200909), the Australian Department of Foreign Affairs and Trade (TDCRRI 72904), the Australian National Health and Medical Research Council (NHMRC) ('Improving Health Outcomes in the Tropical North: A multidisciplinary collaboration 'HOT North' Career Development Fellowship awarded to S.A.), and the Bill and Melinda Gates Foundation (OPP1164105). D.F.E received financial support from Colciencias -Colombia, call 656-2014 "Es Tiempo de Volver" award FP44842-503-2014. The patient sampling and metadata collection was funded by the Asia-Pacific Malaria Elimination Network (108-07), the Malaysian Ministry of Health (BP00500420), and the NHMRC (1037304 and 1045156; Fellowships to N.M.A. [1042072 and 1135820], B.E.B. [1088738] and M.J.G. [1074795]). M.J.G was also supported by a 'Hot North' Earth Career Fellowship (1131932). M.U.F is supported by a senior researcher scholarship from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil. The whole genome sequencing component of the study was supported by grants from the Medical Research Council and UK Department for International Development (M006212) and the Wellcome Trust (206194, 204911) awarded to D.P.K. This work was supported by the Australian Centre for Research Excellence on Malaria Elimination (ACREME), funded by the NHMRC (APP 1134989).

Figures

Figure 1. Overview of the marker selection approaches

Flow diagrams illustrating the datasets, selector models and classification approaches used to identify and evaluate independent SNP panels (A) and single gene regions (B). Decision Tree (DT), HFST-0.90 (HFST and DT with FST threshold of 0.90), HFST-0.95 (HFST and DT with FST threshold of 0.95) and HFST-0.75 (HFST and DT with FST threshold of 0.75) represent the SNP selector models. The DT, HFST-0.9 and HFST-0.95 SNP selector models were run in 500 repeats for SNP selection (A), and the HFST-0.75 model was run in 5 repeats for gene selection (B). For SNP selection (A), the top-25 SNP sets were selected from each model for a further 100 repeats of stratified cross-validation from which one SNP set was selected from each of the DT, HFST-0.9 and HFST-0.95 SNP selector models.

Figure 2. Comparison between the 37-SNP Broad barcode, new marker panels and combined SNP sets

The Broad-37 SNP set reflects 37 of the 42 Broad SNPs represented amongst the 294K high-quality SNPs. The SNP-28 SNP set reflects 28 high-performance SNPs derived from the HFST selector with FST threshold of 0.9. The SNP-28+Broad SNP set reflects the combined Broad-37 and SNP-28 SNP sets for a total of 65 SNPs. The SNP-50 set reflects the 50 SNPs selected by the Decision Tree selector. The SNP-51 set reflects 51 high-performance SNPs from the HFST selector with threshold FST of 0.95. The boxplots present the MCC scores from 500 repeats with stratified 10-fold cross validation for each SNP set. Country labels are provided on the y-axis; MEDIAN, MEAN, Q1 (1st percentile) and MIN reflect the respective summary statistics for the pooled MCC scores across all countries.

Figure 3. Simulation of missing data in the 28-SNP, 65-SNP, 50-SNP and 51-SNP panels

Result of 200 repeats, 25 samples per country simulation of missing data (genotype fails) of 10%, 20% and 30% against the 37-SNP Broad barcode, 28-SNP set, 65-SNP set (28-SNP + Broad panel), 51-SNP and 50-SNP set. The 65-SNP set demonstrated marginally better performance relative to the 28-SNP set with missing data. However, both the 50-SNP and 51-SNP panels outperformed the 65-SNP panel with missing data.

Supplementary Figure 1. *P. vivax* prevalence map pinpointing the countries included in the study

P. vivax prevalence map from the Malaria Atlas Project (*Plasmodium vivax* parasite rate in all ages globally (1-99) from (2000-2017)[26], with counties included in dataset 2 demarked by stars.

Supplementary Figure 2. Overview of the datasets

Supplementary Figure 3. Output from the data quality simulation

The upper panel shows the number of complete SNPs (green), complete informative SNPs with minor allele count (MAC) = 1 (orange) and complete informative SNPs with MAC = 2 (red) against the number of samples. The lower panel shows the number of differences in SNPs between consecutive number of samples, with informative SNPs with MAC = 1 (blue) and informative SNPs with MAC = 2 (orange). The maximum of both MAC=1 and MAC=2 were 958 samples.

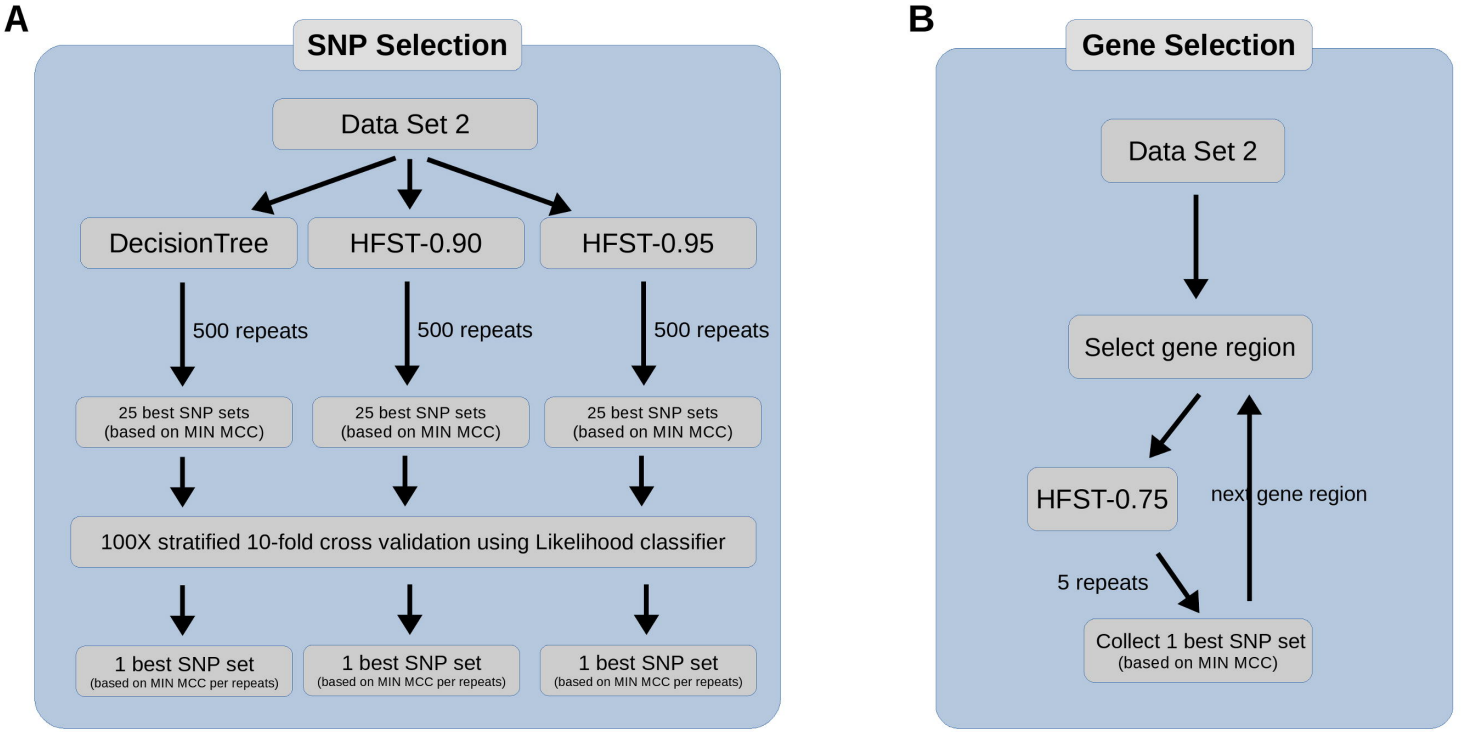
Supplementary Figure 4. Neighbour-joining tree of the global dataset

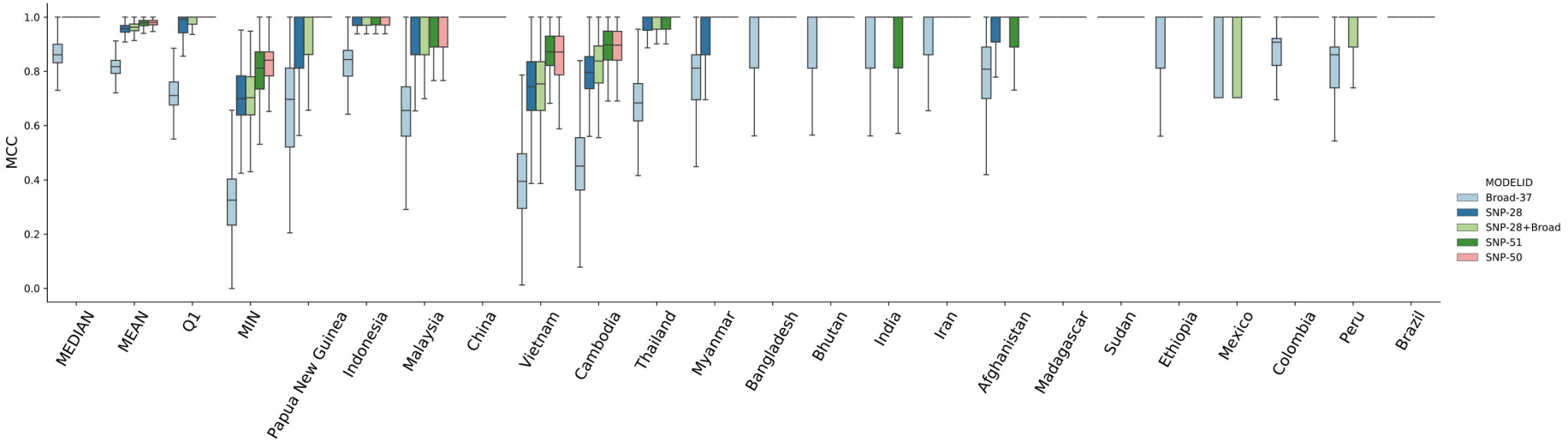
The tree was constructed using genotyping data from 854 samples at 294K SNPs.

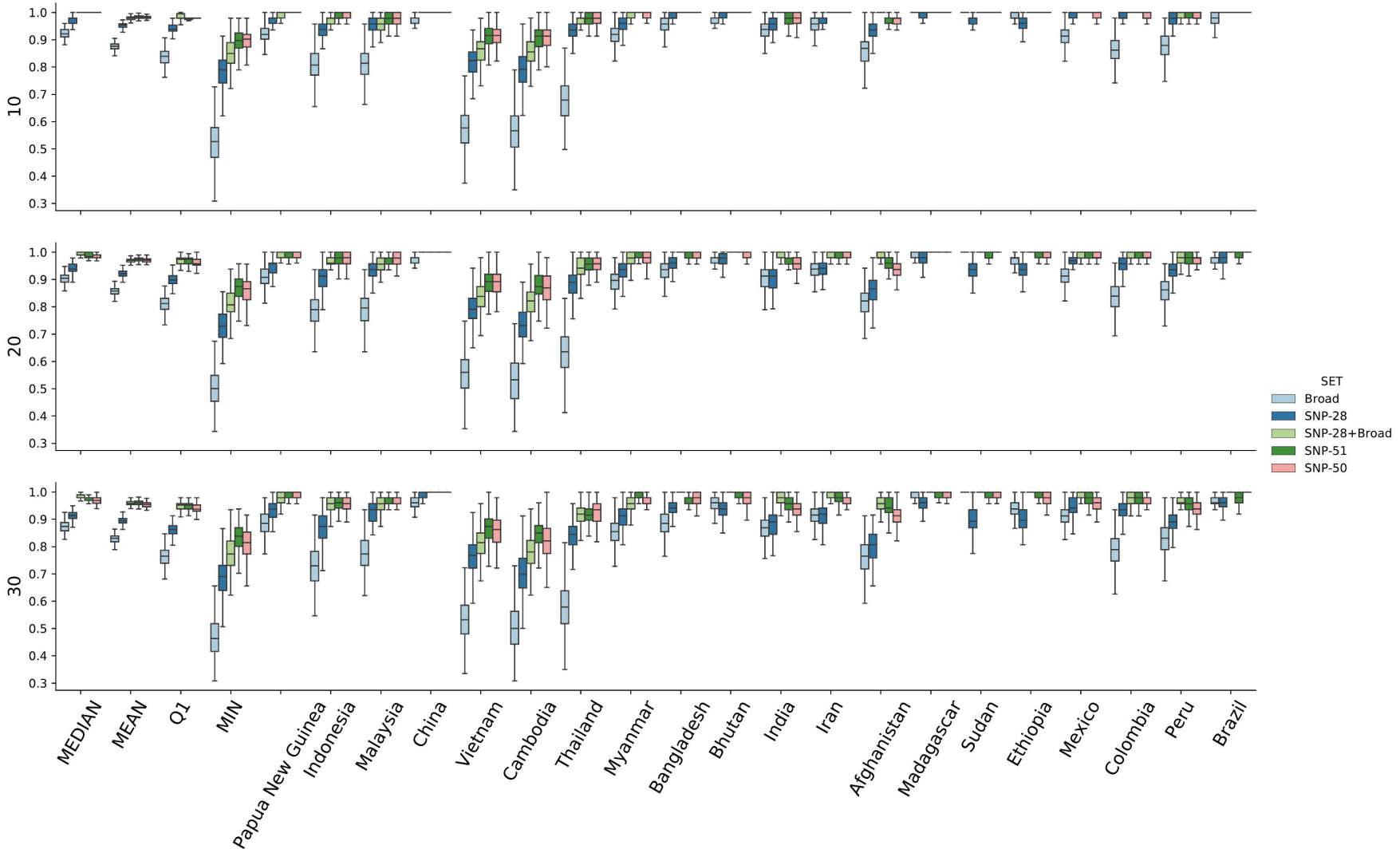
Tables

Table 1. Summary of MCC and F-scores from the consensus results of 500 repeats of the stratified 10-fold cross-validation of the SNP panels

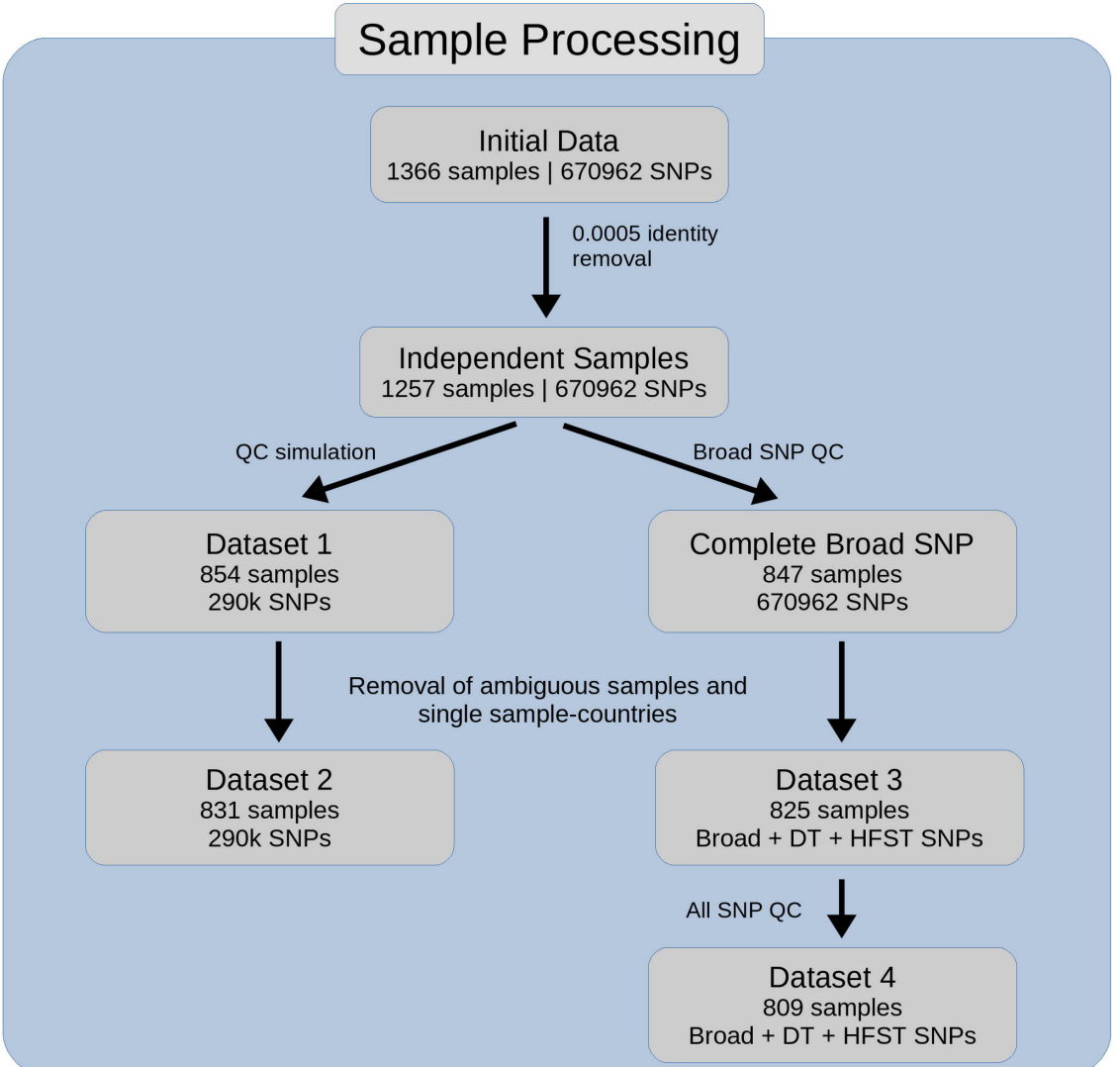
Population	37-SNP (Broad)		28-SNP		65-SNP		50-SNP		51-SNP	
	MCC	F	MCC	F	MCC	F	MCC	F	MCC	F
Afghanistan	0.852	0.857	0.987	0.988	1	1	0.974	0.975	0.987	0.988
Bangladesh	0.796	0.778	0.881	0.875	1	1	1	1	1	1
Bhutan	0.665	0.615	1	1	1	1	0.894	0.889	1	1
Brazil	0.815	0.8	1	1	1	1	1	1	1	1
Cambodia	0.456	0.518	0.78	0.809	0.813	0.837	0.893	0.907	0.898	0.911
China	0.912	0.909	1	1	1	1	1	1	1	1
Colombia	0.895	0.903	1	1	1	1	1	1	1	1
Ethiopia	0.929	0.931	1	1	1	1	1	1	1	1
India	0.714	0.706	1	1	1	1	0.935	0.933	0.925	0.923
Indonesia	0.819	0.857	0.971	0.978	0.984	0.988	0.987	0.99	0.99	0.993
Iran	0.865	0.857	0.925	0.923	1	1	1	1	1	1
Madagascar	0.894	0.889	1	1	1	1	1	1	1	1
Malaysia	0.614	0.627	0.923	0.927	0.949	0.95	0.962	0.963	0.962	0.963
Mexico	0.92	0.919	1	1	0.92	0.919	1	1	1	1
Myanmar	0.529	0.48	0.835	0.824	0.881	0.875	0.935	0.933	1	1
Papua New Guinea	0.616	0.6	0.888	0.889	0.934	0.933	0.976	0.977	0.975	0.976
Peru	0.839	0.846	1	1	0.961	0.962	1	1	1	1
Sudan	1	1	1	1	1	1	1	1	1	1
Thailand	0.681	0.725	0.971	0.975	0.985	0.988	0.99	0.992	0.985	0.987
Vietnam	0.389	0.446	0.714	0.74	0.733	0.757	0.861	0.872	0.866	0.876
Mean	**0.76**	**0.763**	**0.944**	**0.946**	**0.958**	**0.96**	**0.97**	**0.972**	**0.979**	**0.981**
Median	**0.817**	**0.823**	**0.994**	**0.994**	**1**	**1**	**0.995**	**0.996**	**1**	**1**
Q1	**0.653**	**0.624**	**0.914**	**0.915**	**0.945**	**0.946**	**0.955**	**0.956**	**0.983**	**0.984**
Min	**0.389**	**0.446**	**0.714**	**0.74**	**0.733**	**0.757**	**0.861**	**0.872**	**0.866**	**0.876**

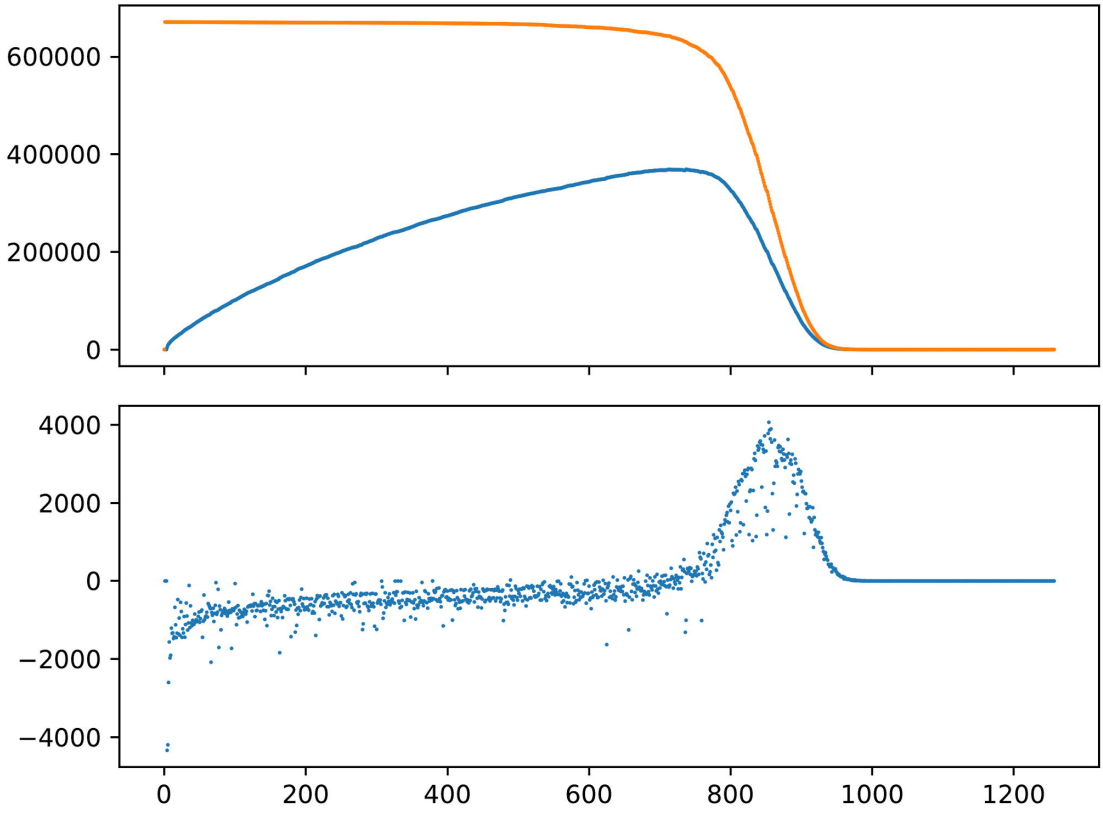

A
SNP Selection
Data Set 2
DecisionTree
HFST-0.90
HFST-0.95
500 repeats
500 repeats
500 repeats
25 best SNP sets (based on MIN MCC)
25 best SNP sets (based on MIN MCC)
25 best SNP sets (based on MIN MCC)
100X stratified 10-fold cross validation using Likelihood classifier
1 best SNP set (based on MIN MCC per repeats)
1 best SNP set (based on MIN MCC per repeats)
1 best SNP set (based on MIN MCC per repeats)
B
Gene Selection
Data Set 2
Select gene region
HFST-0.75
next gene region
5 repeats
Collect 1 best SNP set (based on MIN MCC)

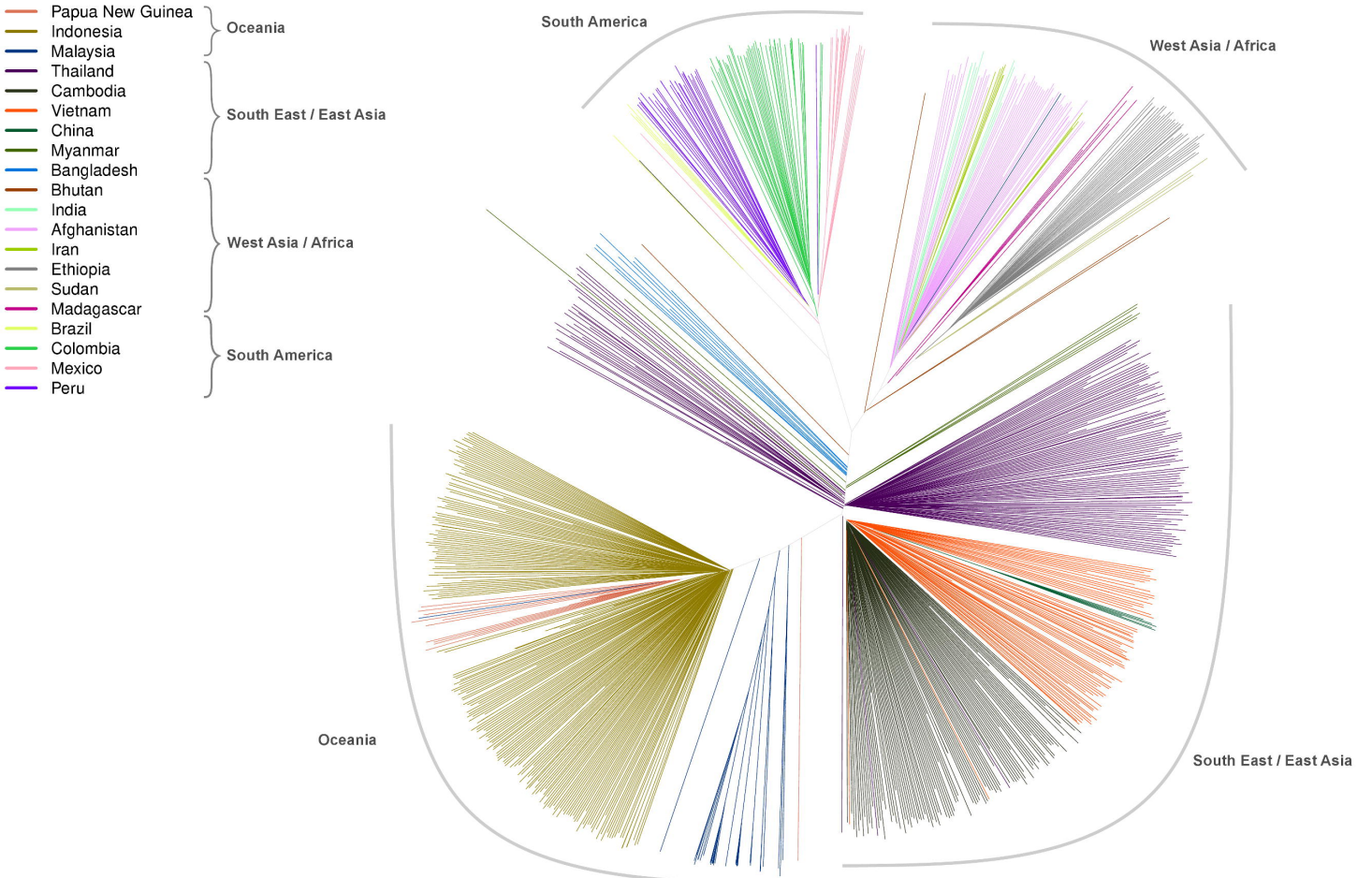

MCC
1.0
0.8
0.6
0.4
0.2
0.0
MEDIAN
MEAN
Q1
MIN
Papua New Guinea
Indonesia
Malaysia
China
Vietnam
Cambodia
Thailand
Myanmar
Bangladesh
Bhutan
India
Iran
Afghanistan
Madagascar
Sudan
Ethiopia
Mexico
Colombia
Peru
Brazil
MODELID
Broad-37
SNP-28
SNP-28+Broad
SNP-51
SNP-50


10
20
30
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
SET
Broad
SNP-28
SNP-28+Broad
SNP-51
SNP-50
MEDIAN
MEAN
Q1
MIN
Papua New Guinea
Indonesia
Malaysia
China
Vietnam
Cambodia
Thailand
Myanmar
Bangladesh
Bhutan
India
Iran
Afghanistan
Madagascar
Sudan
Ethiopia
Mexico
Colombia
Peru
Brazil


United States
Mexico
Cuba
Guatemala
Costa Rica
Venezuela
Colombia
Suriname
Ecuador
Peru
Brazil
Bolivia
Paraguay
Chile
Uruguay
Argentina
Falkland Islands (Islas Malvinas)
France
Spain
Italy
Croatia
Romania
Bulgaria
Greece
Turkey
Morocco
Tunisia
Algeria
Libya
Egypt
Western Sahara
Mauritania
Mali
Niger
Chad
Sudan
Eritrea
Yemen
Senegal
Guinea
Nigeria
Ghana
Cameroon
Gabon
Democratic Republic of the Congo
Kenya
Tanzania
Angola
Zambia
Zimbabwe
Namibia
South Africa
Madagascar
Mauritius
Kazakhstan
Uzbekistan
Azerbaijan
Turkmenistan
Iraq
Iran
Kuwait
Saudi Arabia
Oman
Afghanistan
Pakistan
India
Nepal
Bangladesh
Sri Lanka
Maldives
Mongolia
China
North Korea
South Korea
Japan
Taiwan
Laos
Vietnam
Philippines
Malaysia
Brunei
Indonesia
Papua New Guinea
Australia
New Zealand


Sample Processing
Initial Data
1366 samples | 670962 SNPs
0.0005 identity removal
Independent Samples
1257 samples | 670962 SNPs
QC simulation
Broad SNP QC
Dataset 1
854 samples
290k SNPs
Complete Broad SNP
847 samples
670962 SNPs
Removal of ambiguous samples and single sample-countries
Dataset 2
831 samples
290k SNPs
Dataset 3
825 samples
Broad + DT + HFST SNPs
All SNP QC
Dataset 4
809 samples
Broad + DT + HFST SNPs


600000
400000
200000
0
4000
2000
0
−2000
−4000
0
200
400
600
800
1000
1200


Papua New Guinea
Indonesia
Malaysia
Oceania
Thailand
Cambodia
Vietnam
China
Myanmar
Bangladesh
South East / East Asia
Bhutan
India
Afghanistan
Iran
Ethiopia
Sudan
Madagascar
West Asia / Africa
Brazil
Colombia
Mexico
Peru
South America
South America
West Asia / Africa
Oceania
South East / East Asia