A molecular barcode to inform the geographical origin and transmission dynamics of *Plasmodium vivax* malaria.

Ernest Diez Benavente
ernest.diezbenavente@lshtm.ac.uk

LONDON
SCHOOL *of*
HYGIENE
&TROPICAL
MEDICINE

The
Applied
Genomics
Centre

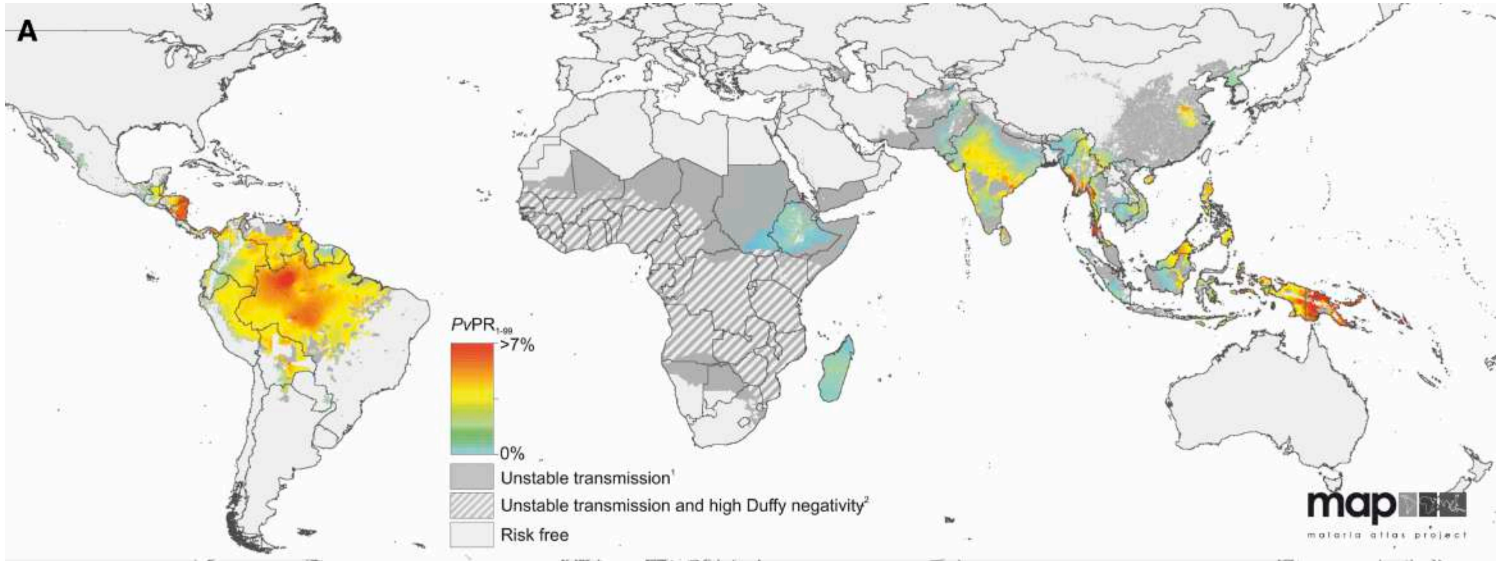
Plasmodium vivax, burden of disease

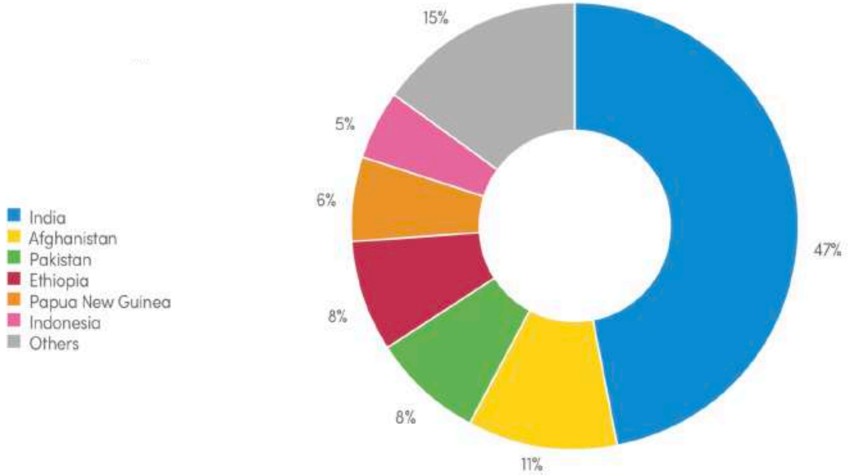
- ***P. vivax*** is one of the six protozoan *Plasmodium* parasites that causes **human malaria** and is transmitted by the bite of **female Anopheles mosquitoes**.

- It has the **widest geographical distribution of the six human malarias** (WHO, 2019).

- Estimated **7.5 million symptomatic cases of *P. vivax* malaria in 2018** (WHO, 2019), **53% of these are in the WHO South-East Asia Region**, with the majority being in India (47%). Most predominant parasite in the WHO Region of the Americas, representing 75% of malaria cases (WHO, 2019).

- The reasons for its distribution include amongst others:
  - ~70 species of *Anopheles* capable of transmitting the disease
  - The wide ranges of temperatures that the disease can survive on
  - Host genetics: Protection associated with Duffy Binding Receptor negativity


A
PvPR1-99
>7%
0%
🟫 Unstable transmission[1]
▨ Unstable transmission and high Duffy negativity[2]
□ Risk free
map
malaria atlas project



Plasmodium vivax, a challenging parasite

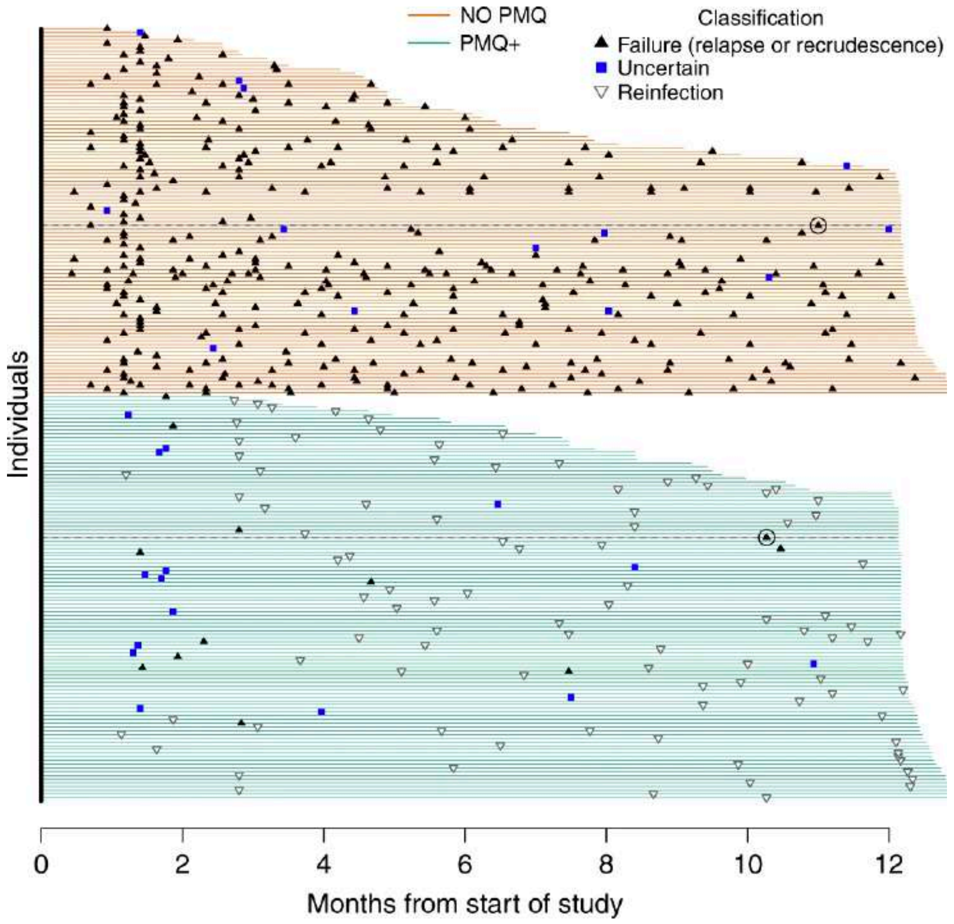
- Despite being **less virulent than *P. falciparum***, it can still **cause life-threatening infections** and due to its hypnozoite formation in the liver, causes **relapsing malaria**.

- **Relapsing *Plasmodium vivax*** can be treated using **primaquine** which kills the liver stages of the parasite, but it is **not recommended for G6PD deficient patients** as it can cause **severe anemia**.

- Some malaria endemic regions have reported an **increase in the proportion of *P. vivax* cases during effective control of *P. falciparum* malaria**, highlighting the resilience of this parasite.

- *Plasmodium vivax* can't be maintained in in-vitro culture.

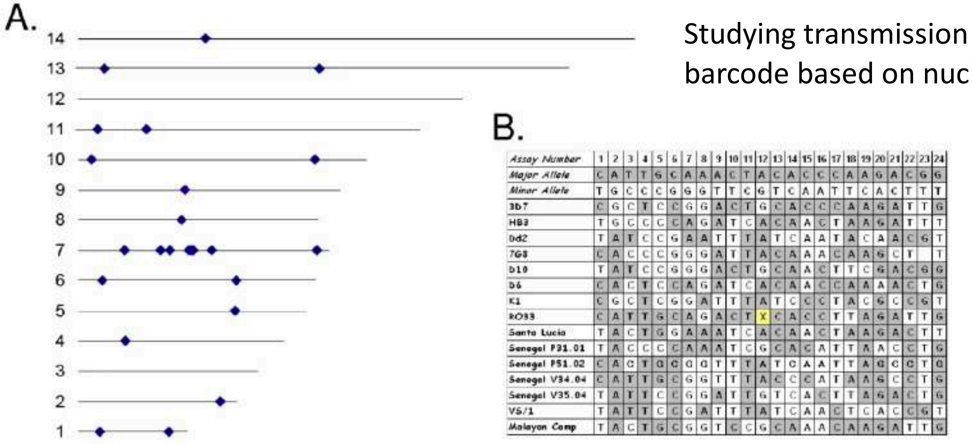
- Reports of **chloroquine drug resistance (first line drug)** in parts of SEA.



Taylor et al. 2019

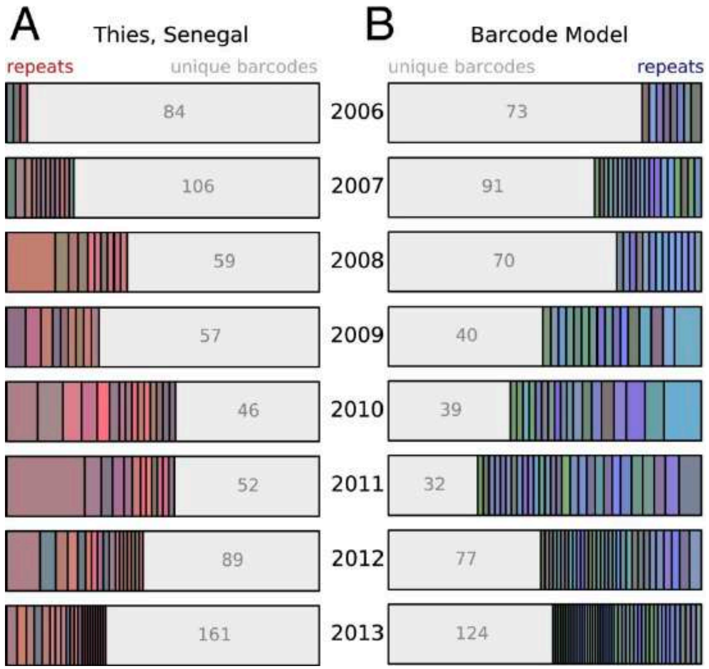
SNP barcoding *Plasmodium* parasites to help malaria control

Studying transmission using SNP barcode based on nuclear genome



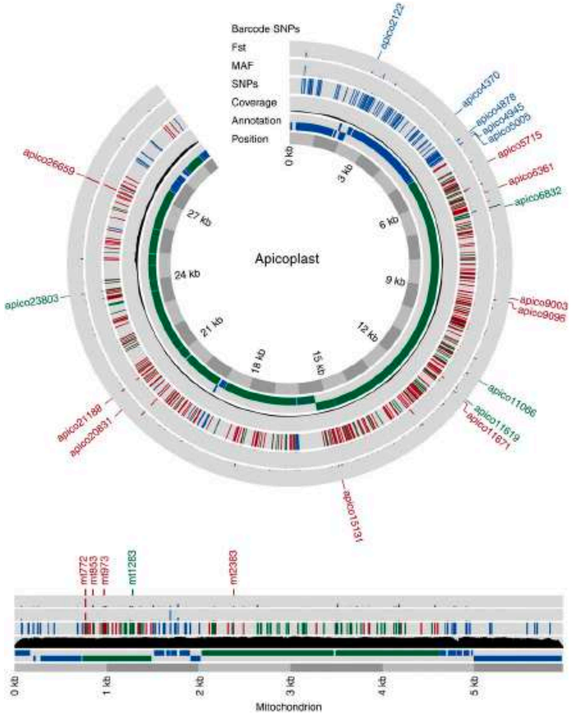
Assay Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Major Allele	C	A	T	T	G	C	A	A	A	C	T	A	C	A	C	C	C	A	A	G	A	C	G	G
Minor Allele	T	G	C	C	C	G	G	G	T	T	C	G	T	C	A	A	T	T	C	A	C	T	T	T
3D7	C	G	C	T	C	C	G	G	A	C	T	G	C	A	C	C	C	A	A	G	A	T	T	G
HB3	T	G	C	C	C	C	A	G	A	T	C	A	C	A	A	C	T	A	A	G	A	T	T	T
Dd2	T	A	T	C	C	G	A	A	T	T	T	A	T	C	A	A	T	A	C	A	A	C	G	T
7G8	C	A	C	C	C	G	G	G	A	T	T	A	C	A	A	A	C	A	A	G	C	T		T
D10	T	A	T	C	C	G	G	G	A	C	T	G	C	A	A	C	T	T	C	G	A	C	G	G
D6	C	A	C	T	C	C	A	G	A	T	C	A	C	A	A	C	C	A	A	A	A	C	T	G
K1	C	G	C	T	C	G	G	A	T	T	T	A	T	C	C	C	T	A	C	G	C	C	G	T
RO33	C	A	T	T	G	C	A	G	A	C	T	X	C	A	C	C	T	T	A	G	A	T	T	G
Santa Lucia	T	A	C	T	G	G	A	A	A	T	C	A	C	A	A	C	T	A	A	G	A	C	T	T
Senegal P31.01	T	A	C	C	C	C	A	A	A	T	C	G	C	A	C	A	T	T	A	A	C	C	T	G
Senegal P51.02	C	A	C	T	G	C	G	G	T	T	T	A	T	C	A	A	T	T	A	G	C	C	T	G
Senegal V34.04	C	A	T	T	G	C	G	G	T	T	T	A	C	C	C	A	T	A	A	G	C	C	T	G
Senegal V35.04	T	A	T	T	C	C	G	G	A	T	T	G	T	C	A	C	T	T	A	G	A	C	T	G
V5/1	T	A	T	T	C	C	G	A	T	T	T	A	T	C	A	A	C	T	C	A	C	C	G	T
Malayon Comp	T	A	C	T	G	C	G	G	T	C	C	G	C	A	A	A	C	A	A	G	A	T	T	G

Daniels et al. 2009

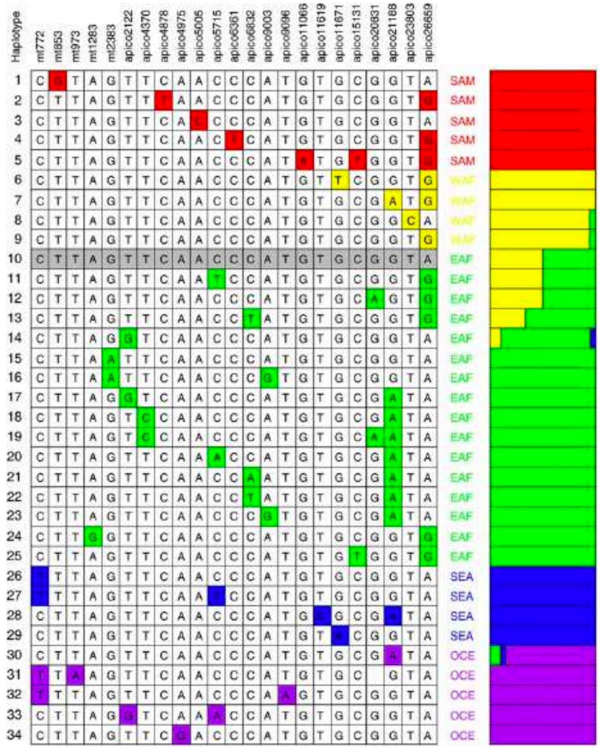


Daniels et al. 2015

Geographic inference using SNP barcode based on organellar genomes



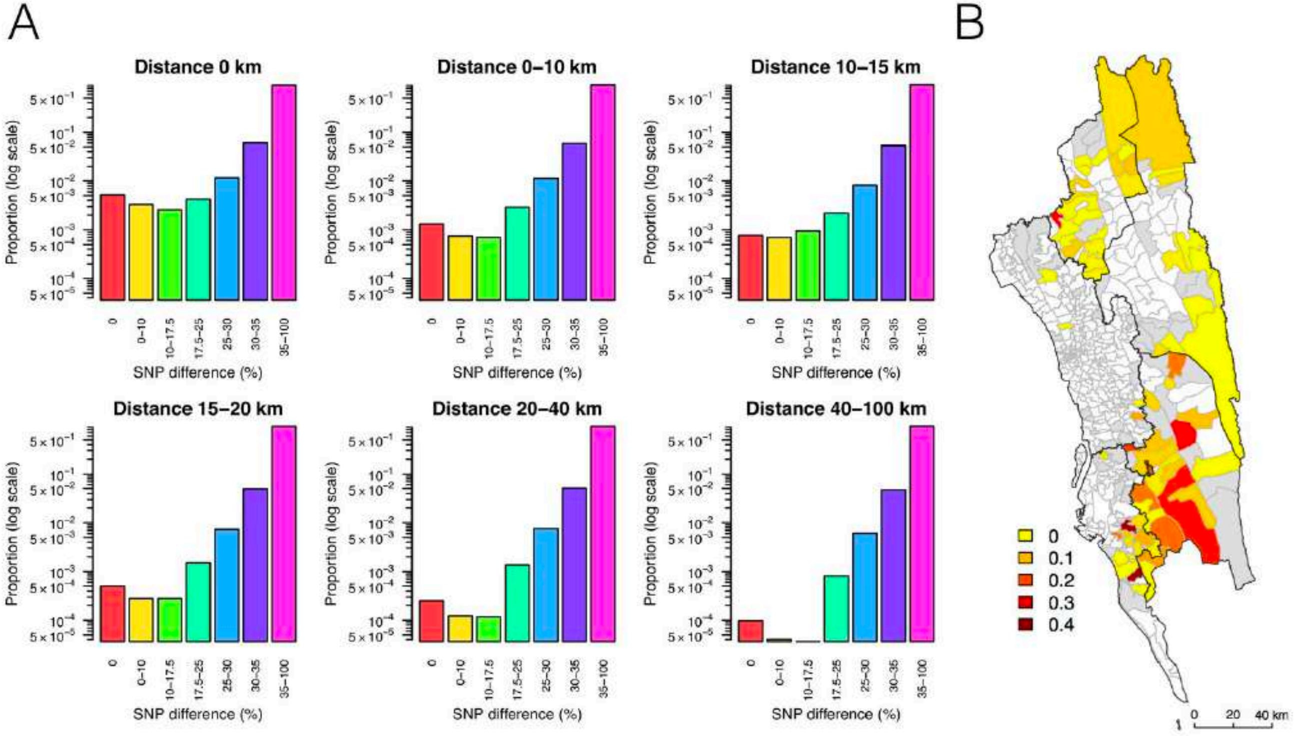
Preston et al. 2014



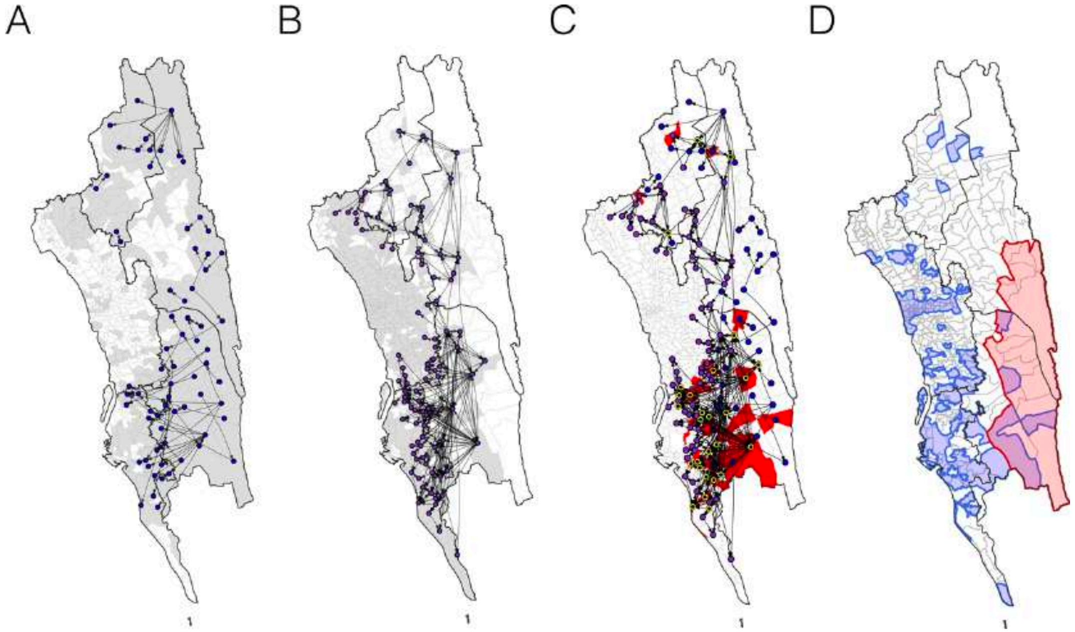
Haplotype	mt772	mt853	mt973	mt1283	mt2383	apico2122	apico4370	apico4878	apico4975	apico5005	apico5715	apico6361	apico6832	apico9003	apico9096	apico11066	apico11619	apico11671	apico15131	apico20831	apico21188	apico23803	apico26659	
1	C	G	T	A	G	T	T	C	A	A	C	C	C	A	T	G	T	G	C	G	G	T	A	SAM
2	C	T	T	A	G	T	T	T	A	A	C	C	C	A	T	G	T	G	C	G	G	T	G	SAM
3	C	T	T	A	G	T	T	C	A	C	C	C	C	A	T	G	T	G	C	G	G	T	A	SAM
4	C	T	T	A	G	T	T	C	A	A	C	T	C	A	T	G	T	G	C	G	G	T	G	SAM
5	C	T	T	A	G	T	T	C	A	A	C	C	C	A	T	A	T	G	T	G	G	T	G	SAM
6	C	T	T	A	G	T	T	C	A	A	C	C	C	A	T	G	T	T	C	G	G	T	G	WAF
7	C	T	T	A	G	T	T	C	A	A	C	C	C	A	T	G	T	G	C	G	A	T	G	WAF
8	C	T	T	A	G	T	T	C	A	A	C	C	C	A	T	G	T	G	C	G	G	C	A	WAF
9	C	T	T	A	G	T	T	C	A	A	C	C	C	A	T	G	T	G	C	G	G	T	G	WAF
10	C	T	T	A	G	T	T	C	A	A	C	C	C	A	T	G	T	G	C	G	G	T	A	EAF
11	C	T	T	A	G	T	T	C	A	A	T	C	C	A	T	G	T	G	C	G	G	T	G	EAF
12	C	T	T	A	G	T	T	C	A	A	C	C	C	A	T	G	T	G	C	A	G	T	G	EAF
13	C	T	T	A	G	T	T	C	A	A	C	C	T	A	T	G	T	G	C	G	G	T	G	EAF
14	C	T	T	A	G	G	T	C	A	A	C	C	C	A	T	G	T	G	C	G	G	T	A	EAF
15	C	T	T	A	A	T	T	C	A	A	C	C	C	A	T	G	T	G	C	G	G	T	A	EAF
16	C	T	T	A	A	T	T	C	A	A	C	C	C	G	T	G	T	G	C	G	G	T	A	EAF
17	C	T	T	A	G	G	T	C	A	A	C	C	C	A	T	G	T	G	C	G	A	T	A	EAF
18	C	T	T	A	G	T	C	C	A	A	C	C	C	A	T	G	T	G	C	G	A	T	A	EAF
19	C	T	T	A	G	T	C	C	A	A	C	C	C	A	T	G	T	G	C	A	A	T	A	EAF
20	C	T	T	A	G	T	T	C	A	A	A	C	C	A	T	G	T	G	C	G	A	T	A	EAF
21	C	T	T	A	G	T	T	C	A	A	C	C	A	A	T	G	T	G	C	G	A	T	A	EAF
22	C	T	T	A	G	T	T	C	A	A	C	C	T	A	T	G	T	G	C	G	A	T	A	EAF
23	C	T	T	A	G	T	T	C	A	A	C	C	C	G	T	G	T	G	C	G	A	T	A	EAF
24	C	T	T	G	G	T	T	C	A	A	C	C	C	A	T	G	T	G	C	G	G	T	G	EAF
25	C	T	T	A	G	T	T	C	A	A	C	C	C	A	T	G	T	G	T	G	G	T	G	EAF
26	T	T	T	A	G	T	T	C	A	A	C	C	C	A	T	G	T	G	C	G	G	T	A	SEA
27	T	T	T	A	G	T	T	C	A	A	T	C	C	A	T	G	T	G	C	G	G	T	A	SEA
28	C	T	T	A	G	T	T	C	A	A	C	C	C	A	T	G	C	G	C	G	A	T	A	SEA
29	C	T	T	A	G	T	T	C	A	A	C	C	C	A	T	G	T	A	C	G	G	T	A	SEA
30	C	T	T	A	G	T	T	C	A	A	C	C	C	A	T	G	T	G	C	G	A	T	A	OCE
31	T	T	A	A	G	T	T	C	A	A	C	C	C	A	T	G	T	G	C		G	T	A	OCE
32	T	T	T	A	G	T	T	C	A	A	C	C	C	A	A	G	T	G	C	G	G	T	A	OCE
33	C	T	T	A	G	G	T	C	A	A	A	C	C	A	T	G	T	G	C	G	G	T	A	OCE
34	C	T	T	A	G	T	T	C	G	A	C	C	C	A	T	G	T	G	C	G	G	T	A	OCE

SNP barcoding *Plasmodium* parasites to help malaria control

Using phone tracking data and 101-SNP barcode based on nuclear genome to map imported malaria in Bangladesh



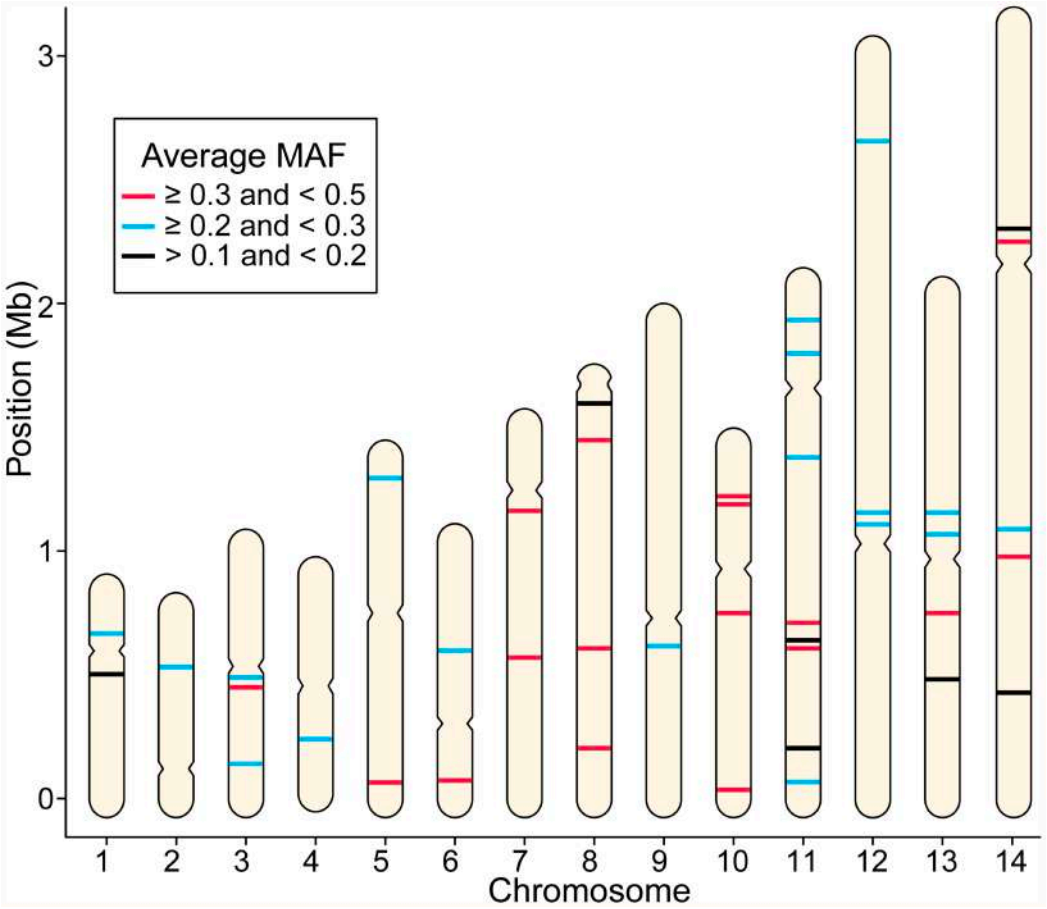
Chang et al. 2019

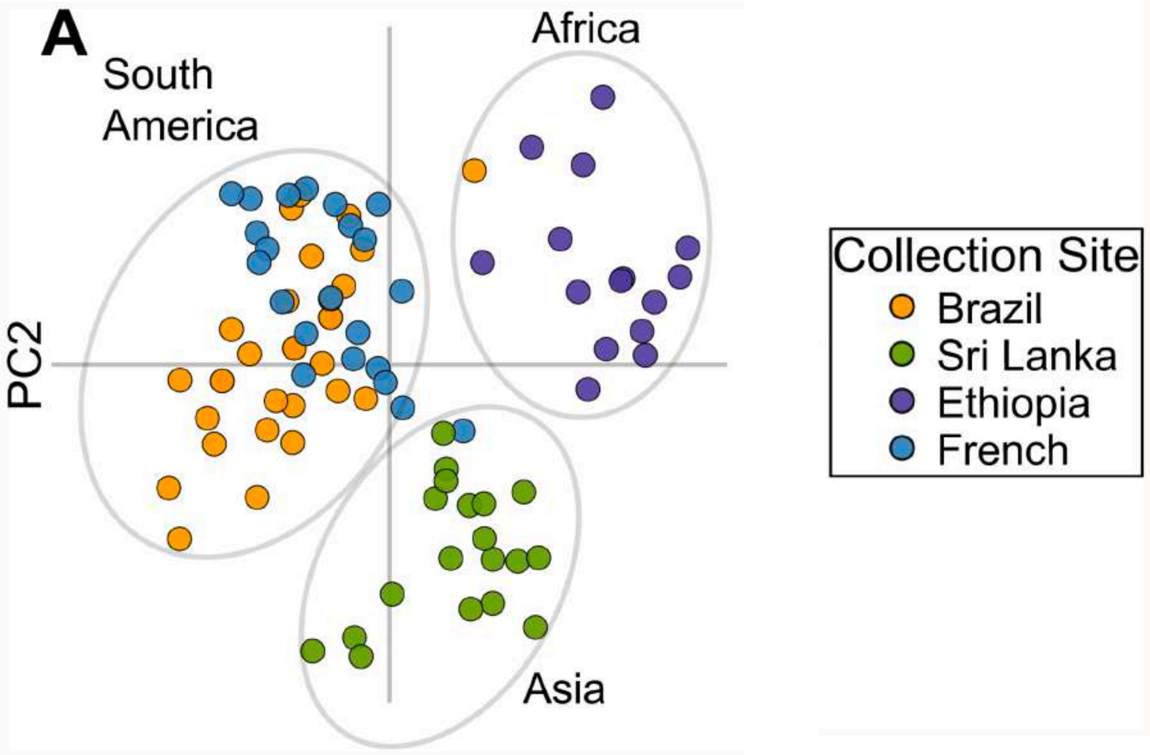


Chang et al. 2019

SNP barcoding *Plasmodium* parasites to help malaria control

42-SNP barcode for *P. vivax* population genetics based on data from 13 isolates



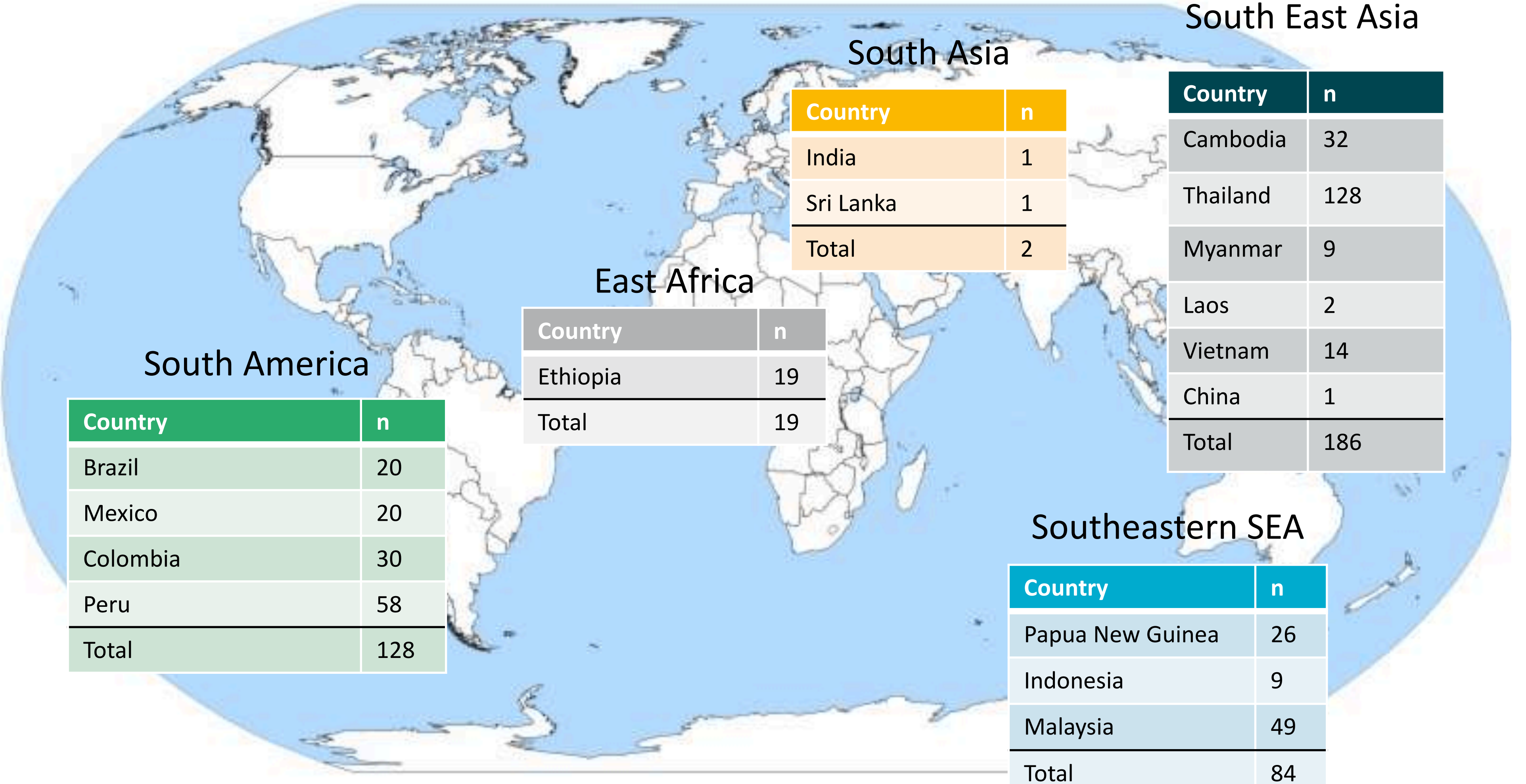


Baniecki et al. 2015

Can we create a **SNP barcode** that can **both** be used to identify **geographic origin** and help characterize **transmission dynamics** of ***Plasmodium vivax*** *infections*?

A SNP barcode to inform geographic origin and transmission dynamics in *Plasmodium vivax*.

- Starting with **837 isolates WGS** from the different endemic areas of *P. vivax*.
- Identified 1,522,046 variants using a bioinformatics pipeline using ***trimmomatic*, *bwa*** **and** ***samtools***.
- After quality filtering using different thresholds for **coverage, missing data, mixed calls and low quality base calls** the final high quality data consisted of **433 isolates and 720,340 SNPs**.



South America

Country	n
Brazil	20
Mexico	20
Colombia	30
Peru	58
Total	128

East Africa

Country	n
Ethiopia	19
Total	19

South Asia

Country	n
India	1
Sri Lanka	1
Total	2

South East Asia

Country	n
Cambodia	32
Thailand	128
Myanmar	9
Laos	2
Vietnam	14
China	1
Total	186

Southeastern SEA

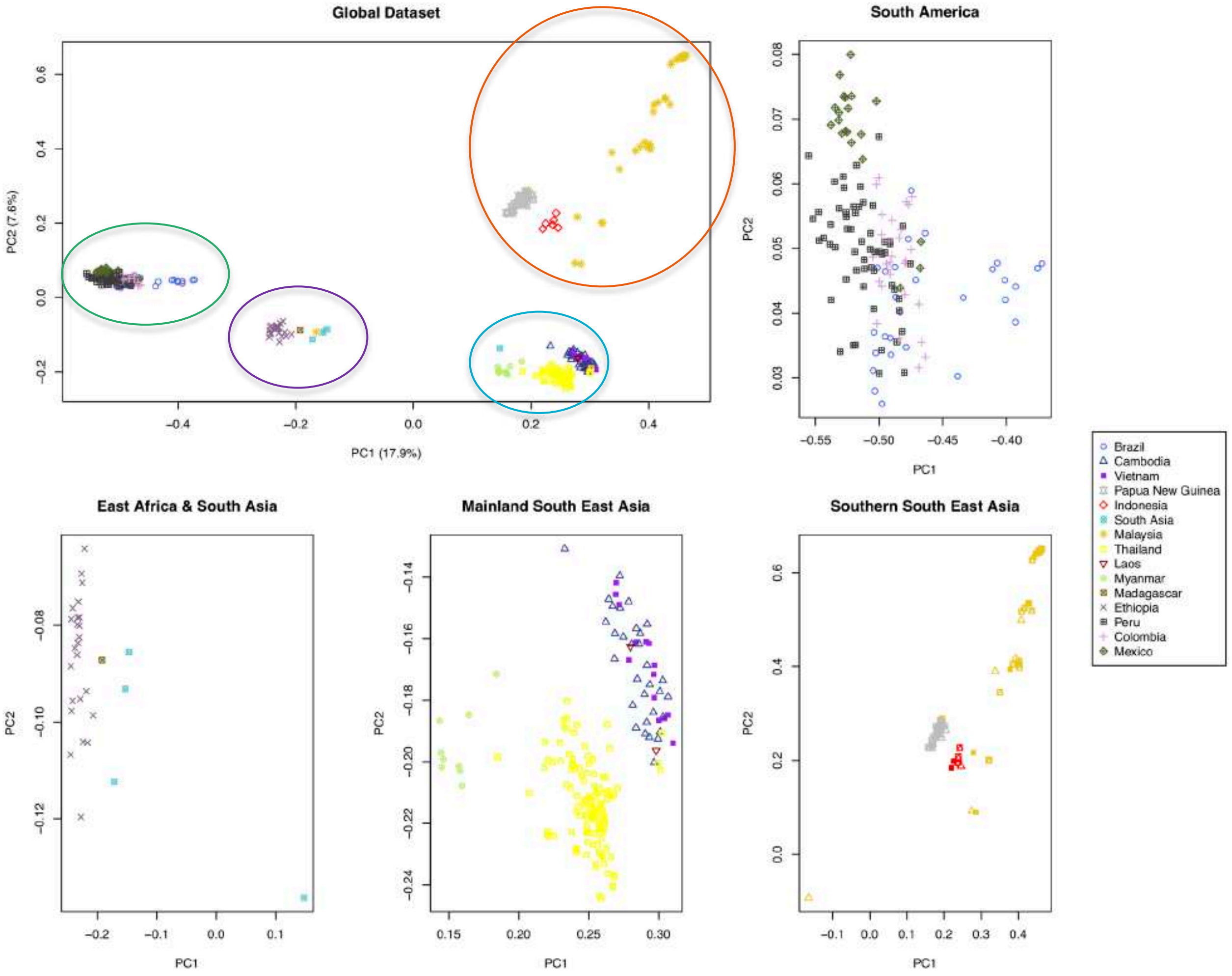
Country	n
Papua New Guinea	26
Indonesia	9
Malaysia	49
Total	84

An overview of population structure using the 720K SNPs to create a PCA

- PCA generated with 720K SNPs **shows good clustering of isolates by region** and to **relatively good separation of isolates by country of origin**.

- The **whole genome pairwise genetic distance between isolates** obtained using the Manhattan method (number of **SNPs differences** in pairwise comparisons) was used as the **gold standard for assessing the performance in the measurement of relatedness of the different subsets of SNPs**.

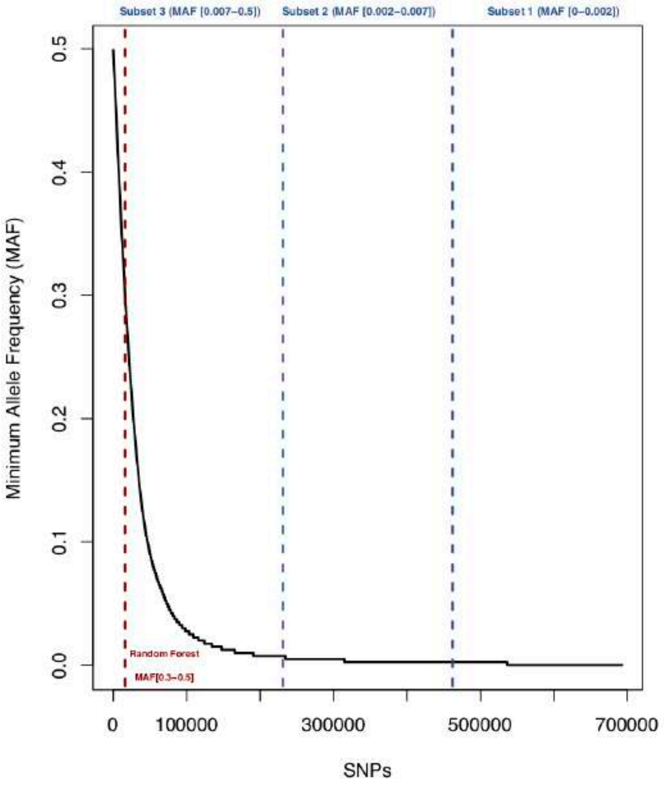
- **NOTE: The genetic distance used to generate the PCA plots has been divided by the sum of the MAF of all the SNPs considered in order to make the distances comparable across subset of SNPs.**

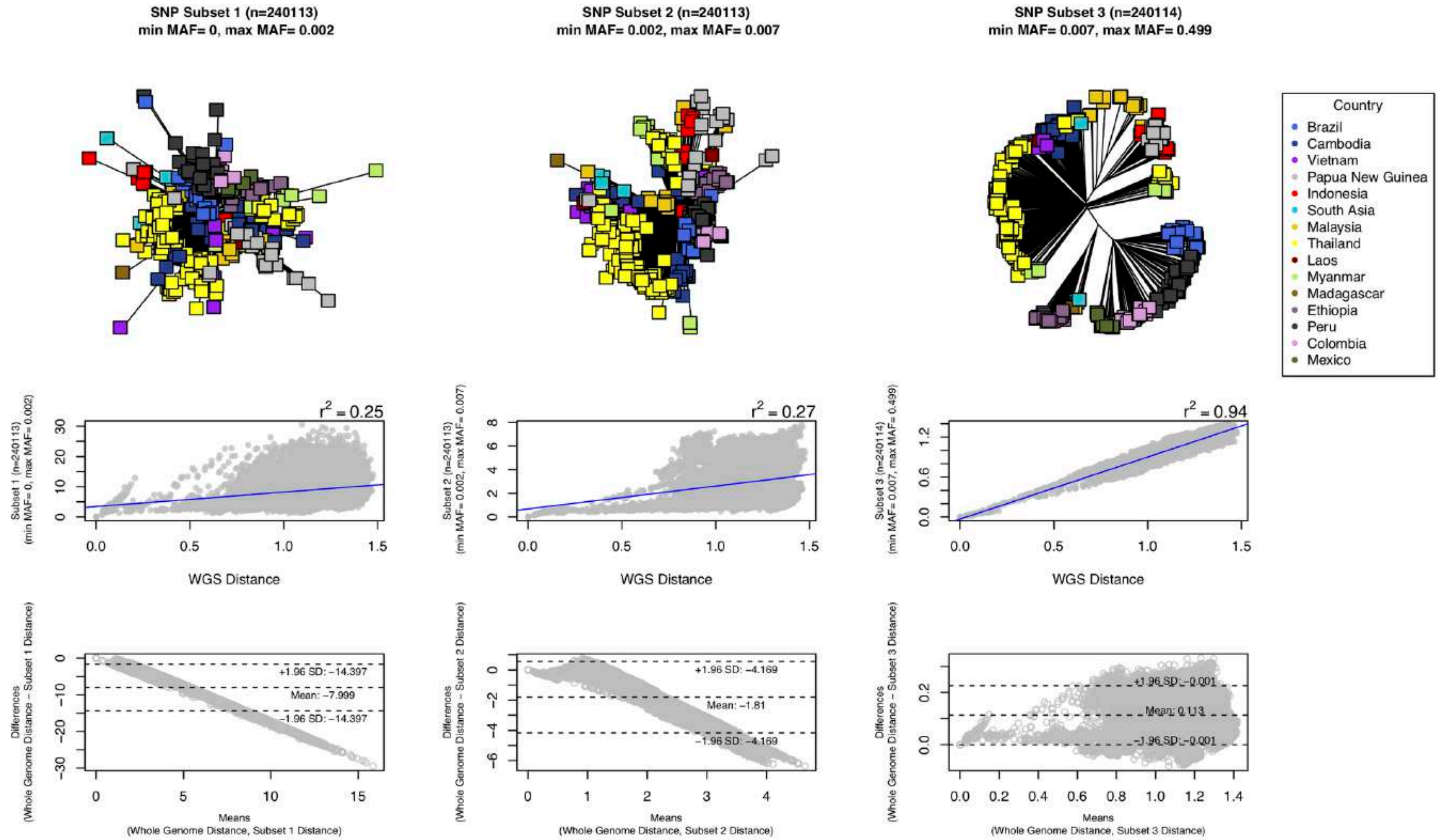


Diez Benavente et al. PLOS Genetics 2020

Using SNPs with high minimum allele frequency (MAF) as drivers of population structure.

SNPs partitioned in **three subsets of equal number of SNPs sorted by MAF**.





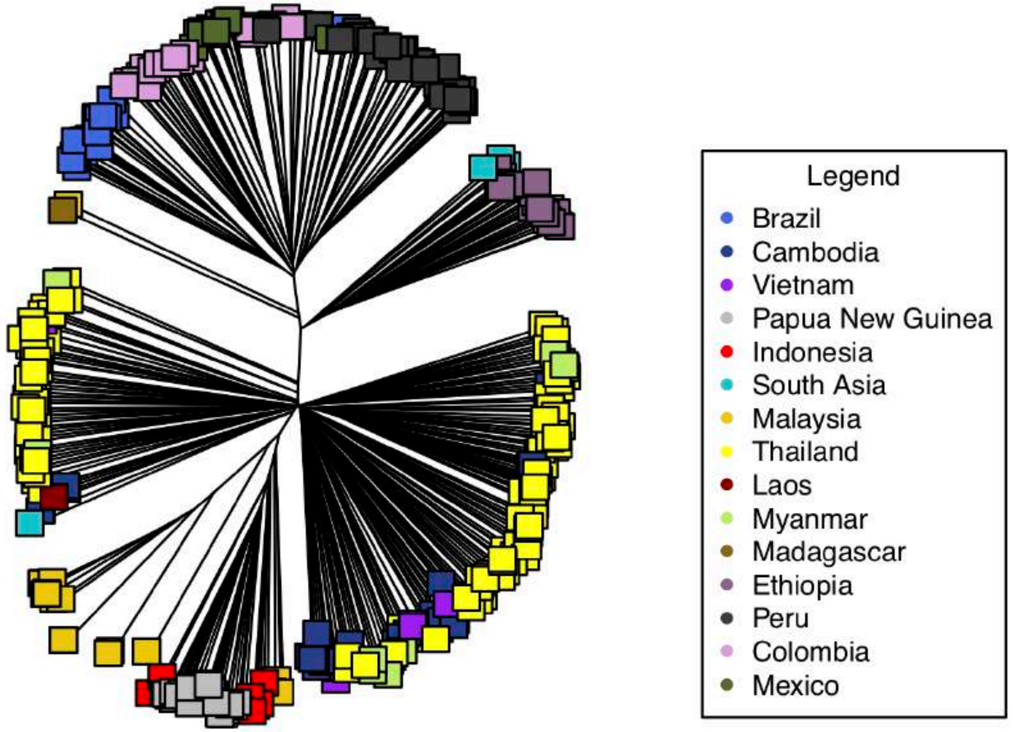
Diez Benavente et al. PLOS Genetics 2020

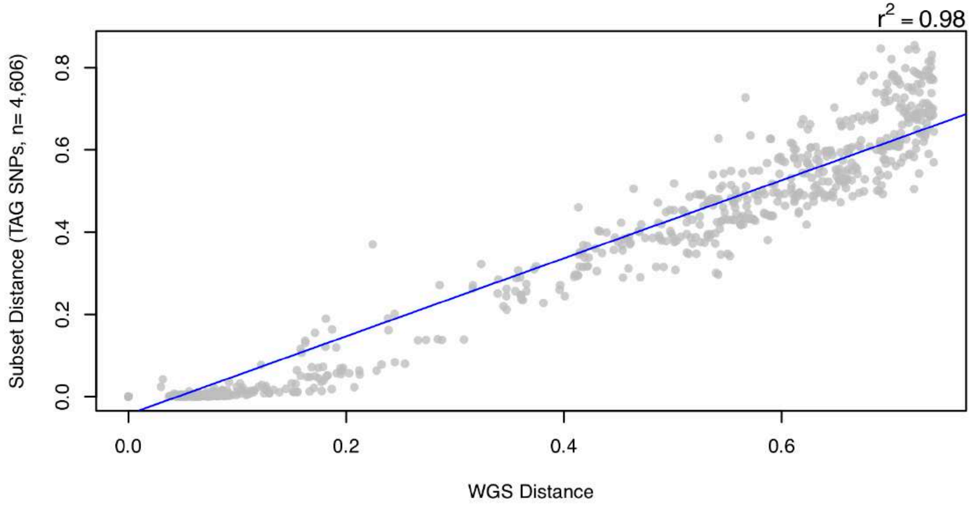
Using tagging software *tagster* to identify independent markers that can capture other SNPs variability

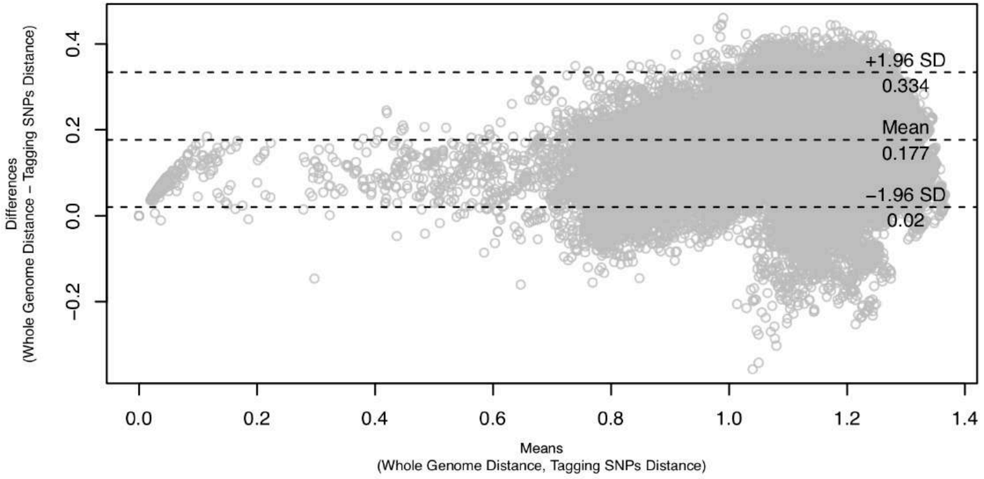
- After MAF selection **16,110 SNPs** were used for further selection.
- We are interested in obtaining independent markers in order to avoid oversampling SNPs that might be high MAF due to sampling differences in our population of isolates.
- *Tagster*, a software originally designed for SNP selection for genotyping chip design was used.
- Tagster identifies SNPs that ”tag” other SNPs in vicinity, so we used this with relatively high threshold for distance between SNPs (500 Kb) and la LD cutoff of 0.7.
- After selection we identified **1,173 independent SNPs** that were able to capture the variability of 40% of the SNPs in our dataset.

Diez Benavente et al. PLOS Genetics 2020

Are we still on track in terms of relatedness (genetic distance) comparability using this subset?







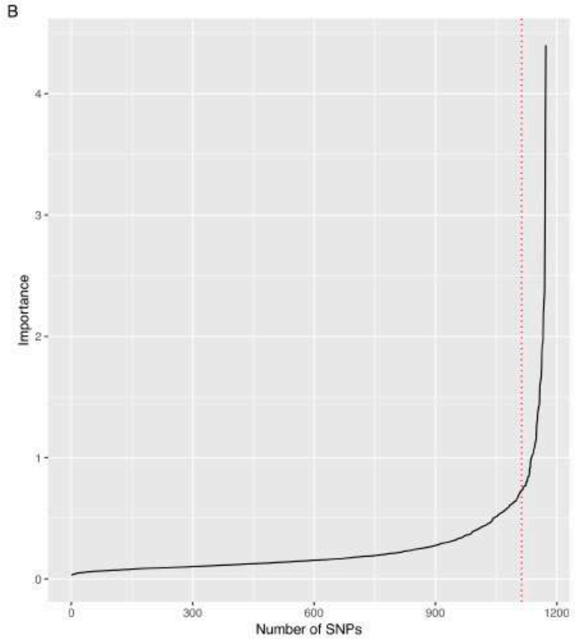
Diez Benavente et al. PLOS Genetics 2020

Using Random Forests to predict geographic origin and identify most relevant SNPs for geographic origin.

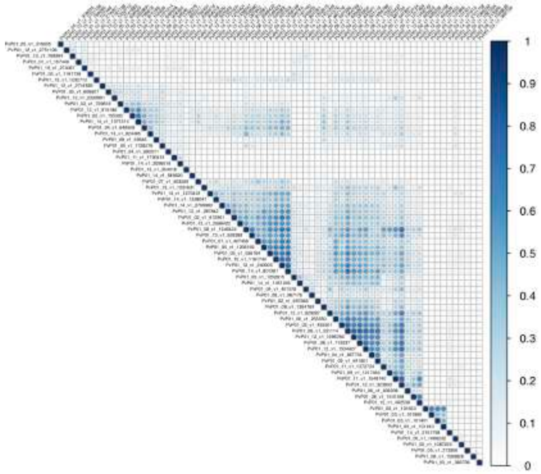
- Random forest classification using country as the outcome to be predicted, we partitioned the dataset randomly into 80% training (n=346) and 20% for validation (n=87), obtaining a out-of-bag error rate of 17.1%.

- We identify the **60 SNPs with the highest importance** for classification and combine this with 11 SNPs with F_{ST} >0.7 (highly differentiating) for the populations with the highest misclassification rates (Thailand/Myanamar).

- We retrain the model using the training set but in this case using only the 71 SNP barcode positions, obtaining an overall accuracy **of 91.4% when predicting in the validation set**.

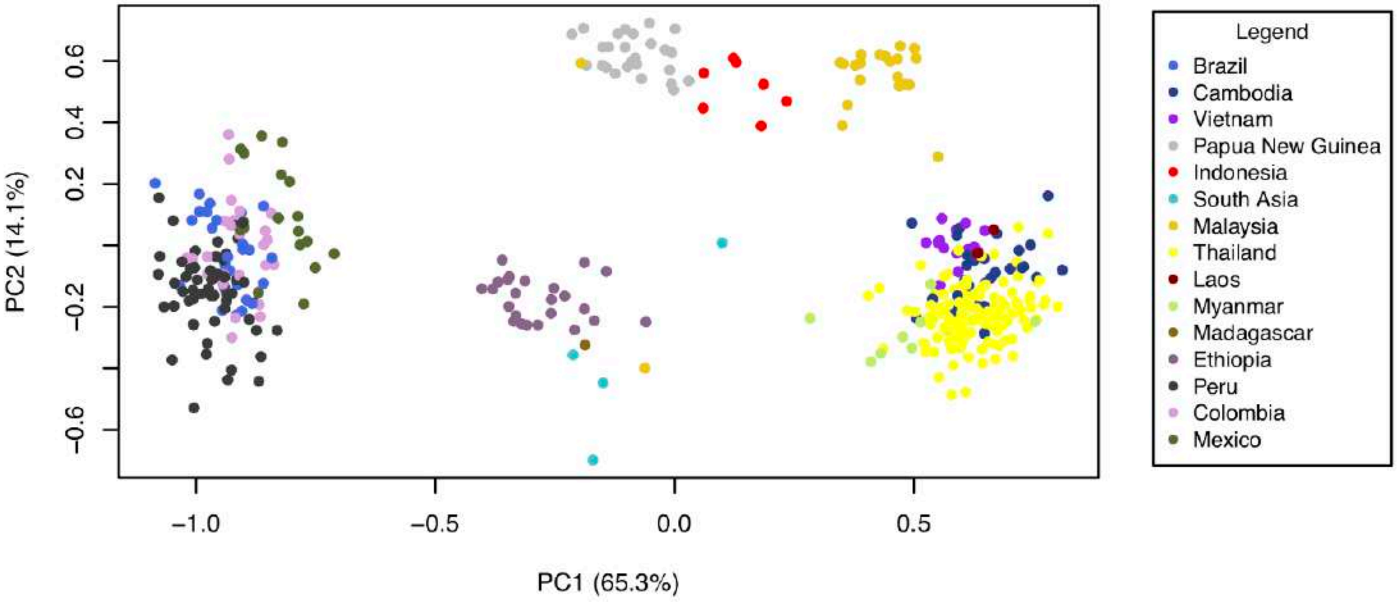


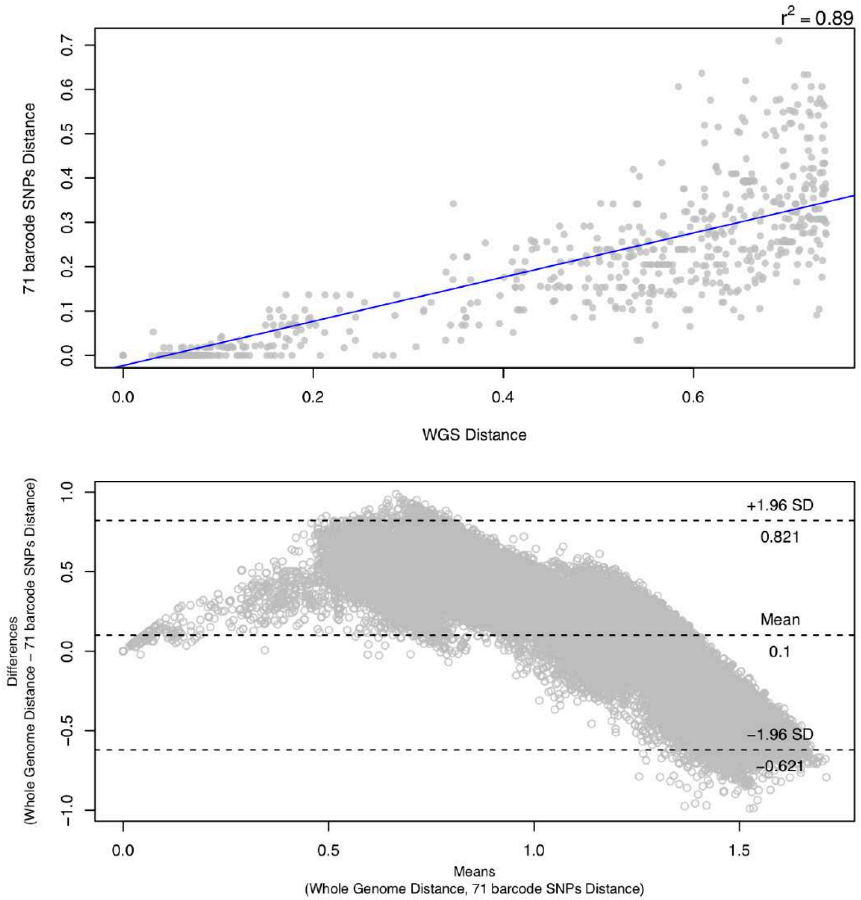
Mean 0.15, IQR = 0.02 - 0.24



Diez Benavente et al. PLOS Genetics 2020

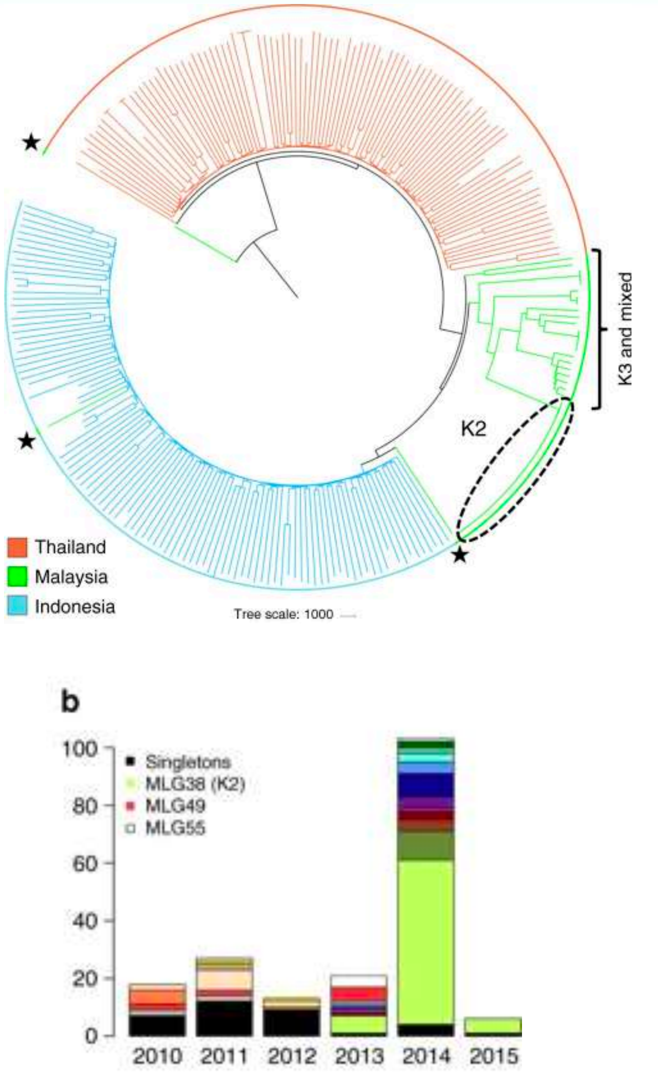
How does the 71 SNP barcode perform?



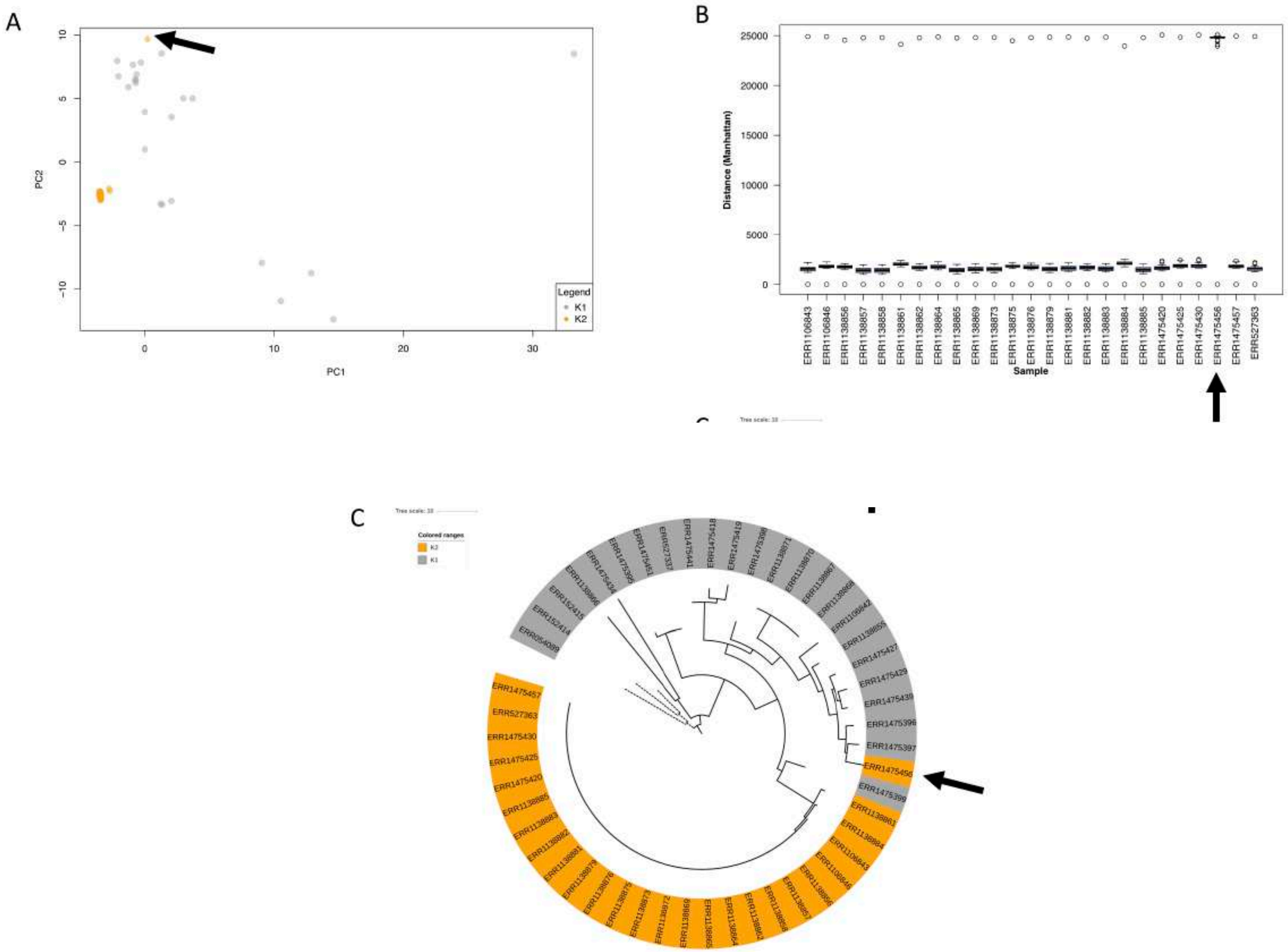


Diez Benavente et al. PLOS Genetics 2020

Use case example for the 71 SNPs barcode, studying clonal transmission in a pre-elimination setting in Sabah, Malaysia.



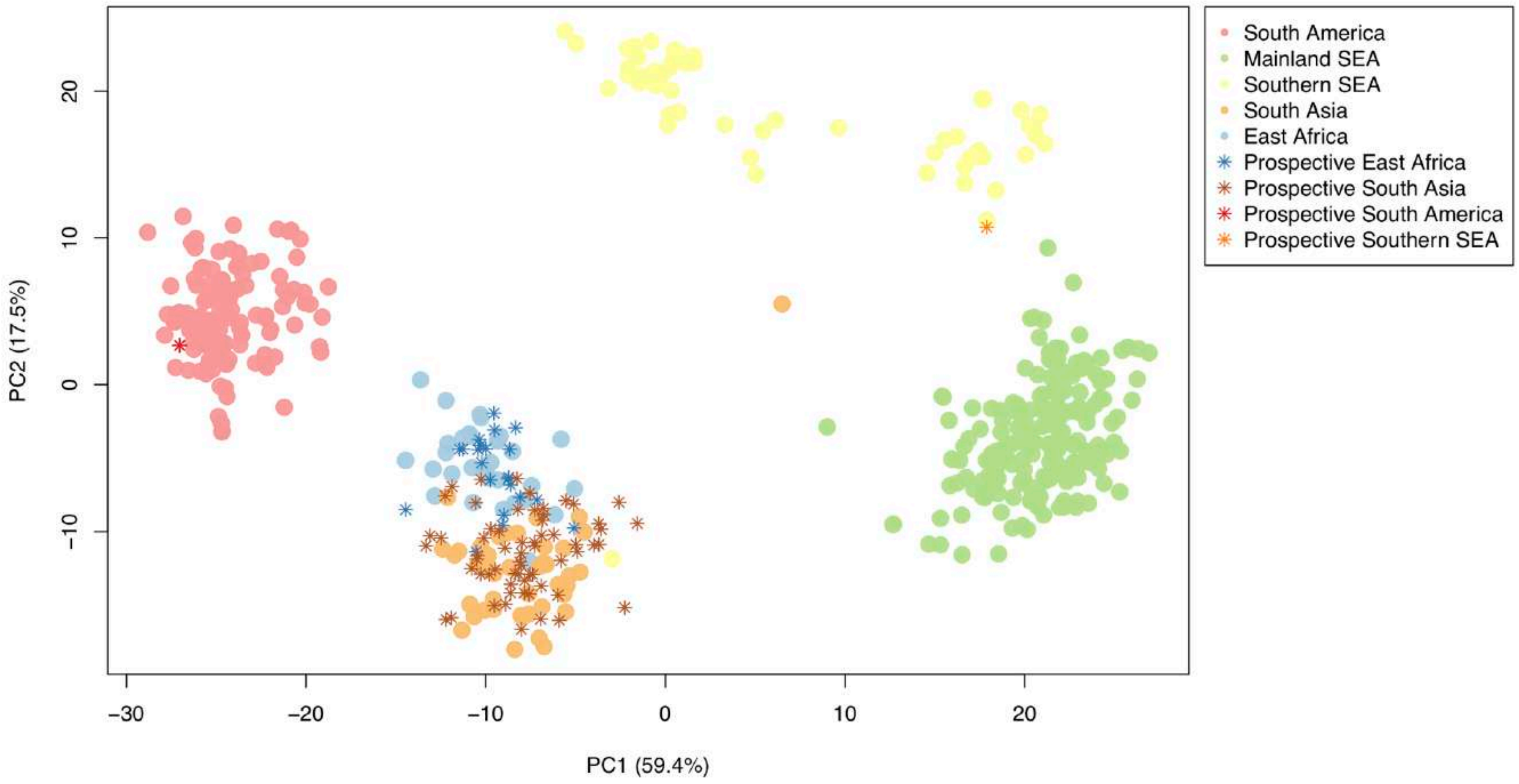
Auburn et al. Nat Comms 2018



Diez Benavente et al. PLOS Genetics 2020

Use case example for the 71 SNPs barcode, origin classification of isolates from returning travelers to the UK (MRL).

Country	n
Afghanistan	31
Bangladesh	2
India	37
Pakistan	35
Eritrea	17
Ethiopia	7
Sudan	10
Uganda	5
The Philippines	1
Guyana	3



Diez Benavente et al. PLOS Genetics 2020

Conclusions

- SNPs barcodes can be designed to be used for geographic origin prediction and to study transmission dynamics of *P. vivax* simultaneously.
- Whole genome sequencing can be used to inform barcode design for *Plasmodium sp*.
- Our in-silico 71-SNP barcode for *P. vivax* showed strong regional classification and promising country level geographical origin prediction.

- Some caveats:
  - We have not assessed the impact of mixed infections, this work assumes only mono infections are considered.
  - Random Forests might not be the best method to perform variable selection, other models could be used.
  - We used genetic distance as our measure of relatedness, but other measures have shown stronger power to understand transmission dynamics (i.e. identity by descent) although it is estimated that at least 200 SNPs would be needed to assess this.
  - Our training dataset is missing some important regions (i.e. South Asia).

Acknowledgements

- LSHTM, UK
  - Prof. Taane Clark
  - Susana Campino
  - Prof. Cally Roper
  - Prof. Colin Sutherland
  - Jody Phelan
  - Monica Campos
  - The Hub
- Malaria Reference Lab, London, UK
  - Debbie Nolder
- University of São Paulo, Brasil
  - Prof. Claudio R. F. Marinho
  - Jamille G. Dombrowski
- Mahidol-Oxford Tropical Medicine Research Unit, Thailand
  - Prof. François Nosten
  - Kanlaya Sriprawat
- Harvard T.H. Chan School of Public Health, USA
  - Aimee Taylor
- Nuffield Department of Medicine, University of Oxford
  - James Watson