Supplementary Material

The geographical origin of the two Plasmodium falciparum isolates detected.

DNA extraction.

≥10ml of venous blood in EDTA was obtained from each case and centrifuged at 1500 rpm for 5 min. Plasma was removed and the red blood cell pellet resuspended in PBS at 10% hematocrit (Htc). White blood cells were removed using cellulose columns as described previously [1]. DNA was extracted with QIAmp DNA mini blood kit (Qiagen) (using 200 µl 50 % Htc per case) following the manufacturer’s instructions. Extracted DNA was eluted in 50 µL of elution buffer and 856 ng and 111 ng of DNA of the male and female cases respectively, was sent to GenomeScan (Leiden, the Netherlands) for sequencing.

Analysis of Belgian samples

Paired sequencing reads of 2x150 bp were generated on the Illumina NovaSeq platform. Reads were aligned to the 3D7 P. falciparum reference genome as available on the PlasmoDB website (version 48) [2, 3] using BWA (version 0.7.17) with seed length set to 50. Quality control was performed using FastQC [4] and FastQ Screen [5] and showed only a minor contamination with human DNA (lower than 3 % of the reads for both samples). Only properly paired reads with a mapping quality higher than 30 were selected using SAMtools [6]. Duplicate reads were removed using the RemoveDuplicates command in the Picard software [7]. SNP calling was performed using the GAKT pipeline [8]. Only those positions were retained for which a genetic variant is described in the Plasmodium falciparum database obtained from MalariaGEN.net.

Database of Plasmodium falciparum

A catalogue (version 6.0) of genetic variants was downloaded from the Plasmodium falciparum Community Project (MalariaGEN.net), which contains genotype calls of 7113 strains, each of them genotyped using the GATK pipeline resulting in 1,045,910 polymorphic coding SNPs across the
genome. For the samples originating from Gabon, which are not included in this catalogue, sequencing data were downloaded from the NCBI Sequence Read Archive based on the accession numbers provided in [9], and consecutively subjected to the same pipeline as described above for the Belgian samples. All genetic variants were combined using the merge command of the BCFtools. Biallelic SNP variants in the core genome region only [10] were filtered and subjected to LD pruning using the ‘locate_unlinked’ function from the python package scikitallel [11]. LD-pruning was performed in several iterations until no more variants were removed.

**Data analysis**

The PCA was applied to two data sets: (1) *Plasmodium* samples from all countries which passed Quality Control; and (2) isolates from West and Central Africa, and South America. While PCA clusters the different samples without foreknowledge, Discriminant Analysis of Principal Components (DAPC) [12] calculates the discriminant components using the country information in such a way that samples originating from the same country will group together, while it tries simultaneously to maximize the distance with samples from other countries. DAPC was performed using the R library Adegenet on the latter data set, excluding the South American countries and Belgian isolates using 200 principal components (PCs) and 12 discriminants (determined through cross-validation). Based on those DAPC results, posterior probabilities for the Belgian samples were calculated, which gives for every country the probability whether either of the Belgian samples has a similar genetic profile as the isolates from those countries. For visualization DAPC was also conducted including the Belgian isolates with 200 PCs and 13 discriminants in the analysis.
Supplementary material. Figure 1. Principle component analysis (PCA) of biallelic variants in the core genome region of isolates from A) across the whole world, and B) Western, Central and South-Central African countries. The first two components are depicted on the left, components 3 and 4 on the right of the figure. Isolates collected in Belgium are indicated as red dots, and cluster with isolates from Central Africa. WAF = West Africa, EAF = East Africa, WSEA = West South-East Asia, ESEA = East South-East Asia, OCE = Oceania, SAM = South America, SAS = South Asia, CAF = Central Africa.

A.

B.
References.

1. Sriprawat K, Kaewponsri S, Suwanaruks R, Leimanis ML, Lek-Uthai U, Phyo AP, et al. Effective and cheap removal of leukocytes and platelets from Plasmodium vivax infected blood. Malar J. 2009;8:115.

2. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 2009;37(Database issue):D539-43.

3. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.

4. Andrews S, Krueger F, Segonds-Pichon A, Biggs L, Krueger C, Wingett S, et al. Babraham Bioinformatics 2021 [Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

5. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000 Research. 2018;7:1338.

6. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987-93.

7. Picard. Picard 2021 [Available from: http://broadinstitute.github.io/picard/.

8. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-8.

9. Amambua-Ngwa A, Amenga-Etego L, Kamau E, Amato R, Ghansah A, Golassa L, et al. Major subpopulations of Plasmodium falciparum in sub-Saharan Africa. Science (New York, NY. 2019;365(6455):813-6.

10. Otto TD, Bohme U, Sanders M, Reid A, Bruske EI, Duffy CW, et al. Long read assemblies of geographically dispersed Plasmodium falciparum isolates reveal highly structured subtelomeres. Wellcome Open Res. 2018;3:52.

11. Zenodo. Zenodo. cggh/scikit-allel: v1.3.2 2020 [Available from: https://zenodo.org/record/3976233#.YD5KqmhKtIkJK.

12. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010;11:94.