Malaria: looking for selection signatures in the human PKLR gene region

Patricia Machado, Rui Pereira, Ana Mafalda Rocha, Licínio Anco, Natércia Fernandes, Juliana Miranda, M. Letícia Ribeiro, Virgílio E. Do Rosário, António Amorim, Leonor Gusmão, et al.

bjh british journal of haematology

Malaria: looking for selection signatures in the human PKLR gene region

Journal:	*British Journal of Haematology*
Manuscript ID:	BJH-2009-01764.R1
Manuscript Type:	Ordinary Papers
Date Submitted by the Author:	04-Feb-2010
Complete List of Authors:	Machado, Patricia; Centre for Malaria and Tropical Diseases, Instituto de Higiene e Medicina Tropical (CMDT.LA/IHMT), Malaria Pereira, Rui; Institute of Molecular Pathology and Immunology of University of Porto (IPATIMUP), Population Genetics; Universidade de Santiago de Compostela, Instituto de Medicina Legal Rocha, Ana; Institute of Molecular Pathology and Immunology of University of Porto (IPATIMUP), Sequencing Service anco, Licínio; Centre of Research in Anthropology in Health, Universidade de Coimbra, Anthropology; Centro Hospitalar de Coimbra, Haematology Fernandes, Natércia; Central Hospital of Maputo, Faculty of Medicine, Universidade Eduardo Mondlane Miranda, Juliana; Paediatric Hospital David Bernardino, Paediatrics Ribeiro, M. Letícia; Centro Hospitalar de Coimbra, Haematology do Rosário, Virgílio; Centre for Malaria and Tropical Diseases, Instituto de Higiene e Medicina Tropical (CMDT.LA/IHMT), Malaria Amorim, António; Institute of Molecular Pathology and Immunology of University of Porto (IPATIMUP), Population Genetics Gusmão, Leonor; Institute of Molecular Pathology and Immunology of University of Porto (IPATIMUP), Population Genetics Arez, Ana; Centre for Malaria and Tropical Diseases, Instituto de Higiene e Medicina Tropical (CMDT.LA/IHMT), Malaria
Key Words:	human malaria, pyruvate kinase-deficiency, selection signatures, PKLR, molecular markers

AUTHORS

Patrícia Machado[1*], Rui Pereira[2,6], Ana Mafalda Rocha[2], Licínio Manco[3,7], Natércia Fernandes[4], Juliana Miranda[5], Letícia Ribeiro[7], Virgílio E. do Rosário[1], António Amorim[2,8], Leonor Gusmão[2], Ana Paula Arez[1]

TITLE

Malaria: looking for selection signatures in the human *PKLR* gene region

RUNNING TITLE

Selection in the human *PKLR* gene region by malaria

AFFILIATION/ADRESSES

[1] Centre for Malaria and Tropical Diseases, Malaria Unit, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Rua da Junqueira, 100, 1349-008 Lisbon, Portugal

[2] Institute of Molecular Pathology and Immunology of University of Porto (IPATIMUP), Oporto, Portugal

[3] Centre of Research in Anthropology in Health / Department of Anthropology, Universidade de Coimbra, Portugal

[4] Central Hospital of Maputo and Faculty of Medicine, Universidade Eduardo Mondlane, Mozambique

[5] Paediatric Hospital David Bernardino, Luanda, Angola

[6] Institute of Legal Medicine, Universidade de Santiago de Compostela, Spain

[7] Haematology Department, Centro Hospitalar de Coimbra, Portugal

[8] Faculty of Sciences, Universidade do Porto, Portugal

CORRESPONDING AUTHOR

* Patrícia Machado

E-mail: pmachado@ihmt.unl.pt; Phone: +351 213659600; Fax: +351 213622458.

SUMMARY

The genetic component of susceptibility to malaria is both complex and multigenic and the better-known protective polymorphisms are those involving erythrocyte-specific structural proteins and enzymes. *in vivo* and *in vitro* data have suggested that pyruvate kinase deficiency, which causes a nonspherocytic haemolytic anaemia, could be protective against malaria severity in humans, but this hypothesis remains to be tested. In the present study, we conducted a combined analysis of Short Tandem Repeats (STRs) and Single Nucleotide Polymorphisms (SNPs) in the pyruvate kinase-encoding gene (*PKLR*) and adjacent regions (chromosome 1q21) to look for malaria selective signatures in two sub-Saharan African populations, from Angola and Mozambique, in several groups with different malaria infection outcome. A European population from Portugal, including a control and a pyruvate kinase-deficient group, was used for comparison. Data from STR and SNP loci spread along the *PKLR* gene region showed a considerably higher differentiation between African and Portuguese populations than that usually found for neutral markers. In addition, a wider region showing strong linkage disequilibrium was found in an uncomplicated malaria group, and a haplotype was found to be associated with this clinical group. Altogether, this data suggests that malaria selective pressure is acting in this genomic region.

KEYWORDS

Human malaria, selection signatures, pyruvate kinase-deficiency, *PKLR*, molecular markers.

INTRODUCTION

According to the World Malaria Report 2008 (World Health Organization, WHO, 2008), 109 countries are currently endemic for malaria, 45 of which are within the African region, and 247 million malaria cases were estimated among the 3.3 billion people at risk in 2006. These cases caused nearly a million deaths, mostly of children under 5 years old. Despite this disastrous picture, the current combination of tools and methods to combat malaria, including long-lasting insecticidal nets and artemisinin-based combination therapy (ACT), supported by indoor residual spraying of insecticide and intermittent preventive treatment in pregnancy, is leading to a significant reduction of cases in some countries, such as Gambia (Ceesay *et al*, 2008), Kenya (O'Meara *et al*, 2008) and São Tomé and Príncipe (Shaio *et al*, unpubl.). However, both *Anopheles* mosquito and *Plasmodium* parasite have developed resistance to insecticides (Anto *et al*, 2009) and new drugs (Noedl *et al*, 2008), which clearly shows that the fight against the disease continues to be a difficult challenge.

Malaria has been reported as one of the strongest known forces for evolutionary selection in the recent history of the human genome. The genetic component of susceptibility to malaria is complex and multigenic, with a variety of genetic polymorphisms reported to influence both pathogenesis and host response to infection (Kwiatkowski, 2005; Min-Oo & Gros, 2005; Williams, 2006). The identification of these variants might, therefore, help to improve the development of therapeutic and disease-prevention strategies.

The most common and best characterised malaria protective polymorphisms are those involving erythrocyte-specific structural proteins and enzymes, such as sickle cell disease and glucose-6-phosphate dehydrogenase (G6PD)-deficiency. More recently, pyruvate kinase (PK)-deficiency has also been reported as protective against malaria in murine models (Min-Oo *et al*, 2003) and two studies were published that reported the *in vitro* culturing of *P. falciparum* in PK-deficient blood with a significant decrease in parasite replication (Ayi *et al*, 2008; Durand & Coetzer, 2008). However, the possibility that PK-deficiency may affect susceptibility to malaria in humans remains to be confirmed.

Apart from results in murine models and *in vitro* cultures, there is no population data supporting a positive association between PK-deficiency and malaria protection. Given the differences in selection pressure that mice and humans have been exposed to over tens of millions of years, the major susceptibility genes in the two species are unlikely to be the same (Hill, 1998), and the possibility that any crucial insufficiency of the erythrocytes, besides PK-deficiency, may influence the development of the parasite make clear the need to perform additional studies to clarify this question. Moreover, until now, contrary to G6PD-deficiency or sickle cell disease, elevated frequencies of PK-deficiency have not been recorded in malaria endemic areas; however, a systematic

analysis has never been done and even the information about the frequency of PK-deficiency in African populations is clearly limited (Manco *et al*, 2001a; Mateu *et al*, 2002).

The first study including a population genetic approach concerning the possible association between the *PKLR* gene (PK-encoding gene) and malaria was carried out at the Island of Santiago, Cabo Verde (Alves *et al*, 2010). Although no association was then found between any *PKLR* polymorphism and infection status, a strong linkage between distant loci in the gene and adjacent regions was reported only in non-infected individuals. This linkage could mean that there is a more conserved gene region that is selected if protective against the infection and/or disease. The present study aimed to further analyse this previous preliminary result by looking at the *PKLR* gene and adjacent regions in individuals belonging to different population groups (from Angola and Mozambique, both malaria endemic countries, and from Portugal, a country with no malaria transmission) and to different malaria status (asymptomatic infection, mild and severe malaria), with the goal of identifying potential selection signatures in this genomic region imprinted by malaria.

MATERIAL AND METHODS

STUDY AREAS

Angola and Mozambique are both sub-Saharan countries. Angola (capital Luanda, 8° 50' 18" S, 13° 14' 4" E) is localised in south-western Africa and is bordered by the Atlantic Ocean to the west; Mozambique (capital Maputo, 25° 57' 55" S, 32° 35' 21" E) is in south-eastern Africa with its east coast on the Indian Ocean. Both have a tropical climate with two seasons, one wet and warm from September to May, and the other dry and cold from June to August. Malaria, predominantly caused by *Plasmodium falciparum*, is endemic (Cuamba *et al*, 2006; Mabunda *et al*, 2008). Portugal (39° 30' N, 8° 00' W) is in south-western Europe. Malaria transmission was interrupted in nearly all parts of the country by 1958 and eradication was confirmed by WHO in 1973 (Bruce-Chwatt, 1977).

SAMPLING

A total of 417 DNA samples were analysed in this study. There were 316 collected from both uninfected and infected non-related children with a different malaria outcome: 166 from Luanda, Angola (**ANG**) [44 with severe malaria, 43 with uncomplicated malaria, 37 from asymptomatic infected individuals and 42 from healthy aparasitaemic individuals (uninfected)] and 150 from Maputo, Mozambique (**MOZ**) (51 with severe malaria and 99 with uncomplicated malaria). Two groups from Portugal were also

analysed: there were 80 samples from healthy individuals (control Portuguese group, **PT-C**) (described in Alves *et al*, 2007) and 21 belonging to individuals with PK-deficiency (**PT-PKD**) (described in Manco *et al*, 1999, 2000). The pooling of all samples from Angola (ANG) and Mozambique (MOZ) constituted the African group (**AFR**).

Malaria outcome was defined as follows: 1) Severe malaria (**SM**): slide positive for blood-stage asexual *P. falciparum* at any parasite density, fever (axillary temperature ≥ 37.5ºC), haemoglobin level of Hb≤ 5 g/dL and/or other symptoms such as coma, prostration or convulsions; 2) Uncomplicated malaria (**UM**): slide positive for blood-stage asexual *P. falciparum* at any parasite density, fever (axillary temperature ≥ 37.5ºC) and haemoglobin level of Hb> 5 g/dL; and 3) Asymptomatic infection (**AI**): slide positive for blood-stage asexual *P. falciparum* at any parasite density in the absence of fever or other symptoms of clinical illness. The additional group of uninfected children (**NI**) was defined as slide negative and the absence of fever or other symptoms of clinical illness. Slide negativity was afterwards confirmed by Polymerase Chain Reaction (PCR). The illness group (**ILL**) comprised all the individuals expressing clinical disease: SM plus UM.

BLOOD COLLECTION AND DNA EXTRACTION

Blood sample collections by finger-prick were carried out in Angola in August 2005 and in Mozambique during 2006 from children aged 3 months to 15 years who reported to the Emergency Services of the Paediatric Hospital David Bernardino, Luanda (Angola) or to the Paediatric Emergency Services of Central Hospital of Maputo, Health Centre of Bagamoyo or Health Centre of Boane (Mozambique). The blood was drawn after the clinician examination (malaria was considered to be the primary diagnosis if *Plasmodium* parasites were found in the peripheral blood and if other likely causes of the clinical presentation could be excluded at the admission) but before the administration of any anti-malarial therapeutics and/or blood transfusion. The registration of symptoms, axillary temperature, haemoglobin level and history of malaria was done for all individuals.

The investigation was approved by both the Ministry of Public Health of Angola and Mozambique and by the local Ethical Committees at the institutions involved in the study. Each individual and parent/tutor of the children was informed of the nature and aims of the study and told that participation was voluntary; informed consents were obtained from all individuals.

DNA was extracted using standard phenol-chloroform or chelex procedures from peripheral blood. In the case of infected individuals, human and *Plasmodium* DNA were extracted simultaneously.

GENOTYPING

A section of chromosome 1q21, including the *PKLR* gene and adjacent regions, with a total length of ≈ 95 Kb, was genotyped for 4 Short Tandem Repeats (STRs) and 15 Single Nucleotide Polymorphisms (SNPs). Samples were also genotyped for 32 Ancestry Informative Insertion/Deletion polymorphisms (AI-INDELs) distributed throughout the genome. The localisation of polymorphisms in chromosome 1 is represented in Fig 1.

STRs

The STRs used were IVS3 (in intron 3), IVS11 (intron 11), PKA (≈ 25 Kb upstream the *PKLR* gene) and PKV (≈ 65 Kb upstream the gene) and were genotyped after multiplex PCR as described in Alves *et al* (2010).

SNPs

SNPs localised in a region closer to *PKLR* than the abovementioned STRs were genotyped using a SNaPshot (Applied Biosystems) multiplex reaction.

The DNA sequence of chromosome 1q21, including the *PKLR* gene and flanking regions, was screened for SNPs in the HapMap database (www.hapmap.org). A total of 13 SNPs were selected in a region of 40,970 bp that spanned the *PKLR* gene (chr1:153515199..153556169; data source: HapMap Data Rel 22/phaseII Apr07, on NCBI B36 assembly, dbSNP b126), starting at 18,334 bp upstream and extending to 11,055 bp downstream of the gene. All the SNPs described for the *PKLR* gene were selected for genotyping, except rs3020781, which had amplification difficulties. SNPs outside of the gene that showed variation in the reference African population (Yoruba, Nigeria), with a minor allele frequency above 15% and distances between contiguous SNPs greater than 1,600 bp, were included in the study.

Two additional mutations were investigated in the *PKLR* gene: 1456C>T, because it is the most common mutation in South Europe, namely in Portugal (Manco *et al*, 2001b) and the only one described in PK-deficient Afro-American individuals (Beutler & Gelbart, 2000), and 1614A>T, identified in São Tome and Príncipe (Manco *et al*, 2009).

Primers were designed for the flanking regions of each of the 15 SNPs in the GenBank database sequence AY316591 with Primer 3 software v.0.4.0 (Rozen & Skaletsky, 2000; primer sequences in Supplementary Table I). Primers were first tested in singleplex and then multiplex reactions were carried out according to Goios *et al*, 2008, using Qiagen Multiplex PCR Kit (Qiagen).

For each SNP, an SBE-Primer was designed with Primer 3 software (Supplementary Table II). Amplified products were purified with ExoSAP-IT (Amersham Biosciences) and SNaPshot reactions were then performed using the SNaPshot Multiplex Kit (Applied Biosystems) in a reaction volume of 5 µL with primer concentrations as indicated, under the following conditions: 96ºC for 10 s, 55ºC for 5 s, and 60ºC for 30 s, repeated through 27 cycles. The final products were purified with SAP (Amersham Biosciences) and run in an ABI PRISM 3130 Genetic Analyzer. Allele assignment was performed using GeneMapper 4.0 (Applied Biosystems).

Ancestry Informative INDELs

The high levels of genetic substructure in Africa, even within small geographic regions, require the determination of individual ancestry and proper correction for substructure in association studies (Campbell & Tishkoff, 2008). To look into the structure of our African groups and to investigate if our PT-PKD group could have a relevant African genetic component, which would suggest that PK-deficiency could be frequent in that region, 32 INDEL polymorphic regions localised throughout the genome were genotyped as described in Santos *et al* (2010). In this work, we used only a subset of the original assay, comprising the INDELs that are especially informative of African and European ancestry. An additional reference Portuguese group (**PT-REF**) that was previously typed for these INDEL loci (Santos *et al*, 2010) was also used in this analysis.

STATISTICAL ANALYSIS

Analysis was performed by comparing population groups (ANG, MOZ, PT-C, PT-PKD) and malaria status groups (SM, UM, AM, NI, ILL). STR and SNP results were explored with Arlequin 3.1 (Excoffier *et al*, 2005): determination of the allele frequencies, expected and observed heterozygosity and population pairwise F_{ST} values, Hardy–Weinberg equilibrium tests, Linkage Disequilibrium (LD) tests, haplotype frequency estimation and analysis of molecular variance (AMOVA). When there were multiple tests, Bonferroni's correction was applied, dividing 0.05 by the number of tests to obtain the actual cut-off for significance. The allelic association of SNPs and STRs with malaria status groups was assessed by a Pearson's 2x2 contingency table χ^2 test using Simple Interactive Statistical Analysis (SISA, www.quantitativeskills.com/sisa/). Odds ratios (OR) and 95% confidence intervals (CI) were estimated using SISA. Allelic richness with rarefaction of private alleles was calculated with HP-Rare (Kalinowski, 2005). Bayesian clustering analysis as implemented by Structure 2.2 (Pritchard *et al*, 2000) was used to infer population substructure/ancestry from the INDEL data set, without prior information on sampling groups, under the admixture model with

correlated allele frequencies. Ten independent runs with 10^5 burn-in steps and 10^5 interactions were done for each value of *K* (*K*=1 to 5 clusters). For INDELs, Arlequin 3.1 (Excoffier *et al*, 2005) was also used for F_{ST} calculations.

RESULTS

STRs

The allele frequencies for the four STR loci found in ANG, MOZ, PT-C and PT-PKD are shown in Supplementary Table III. The IVS3 locus presented the greatest diversity indices in all groups, with the highest number of alleles and expected heterozygosity. In both African groups, the observed genotype frequencies were according to Hardy-Weinberg expectations for all loci except for IVS3, which revealed a heterozygosity significantly below the expected (P≤ 0.000). In Portuguese groups, all loci were in Hardy-Weinberg equilibrium in the control PT-C (P= 0.378 for IVS3) but not in the PT-PKD group, which showed a strong deviation from the expected values for IVS3 (P≤ 0.000) and IVS11 (P= 0.006).

When F_{ST} values were calculated, no significant differentiation was obtained for the pair ANG *vs.* MOZ (F_{ST}= 0.002; P= 0.189). When Portuguese groups were compared, significant values were obtained, as expected: F_{ST}(PT-C *vs.* PT-PKD)= 0.025; P≤ 0.000. Since no differentiation was found between Angola and Mozambique, a single group was formed for all of the African samples (AFR) and it was compared to Portuguese groups to investigate if African and Portuguese PK-deficient individuals were genetically closer in this genomic region than African and Portuguese controls. If so, we could hypothesise that PK-deficiency could be frequent in Africa (because of some kind of selective advantage conferred by the disease). The F_{ST} values obtained were as follows: F_{ST}(AFR *vs.* PT-C)= 0.102 and F_{ST}(AFR *vs.* PT-PKD)= 0.153 (P≤ 0.000 for both tests).

No significant differentiation was found between the several malaria status groups, whether considering each of the four STR loci separately or all together. Since F_{ST} was not significant when comparing ANG and MOZ, UM and SM samples from both countries were pooled into two larger groups, but still no significant values were found between these groups. No STR or SNP allele was associated with any malaria status group (P> 0.05) and OR values were non-significant for all groups. Moreover, when STR allelic private richness was calculated (considering 42 genes for all groups as PT-PKD only includes 21 samples), private alleles were not identified, supporting the previous result. However, allele 16 of locus IVS11 (χ^2= 10.918; P< 0.001 and OR= 6.200 with 95% CI 1.858-20.685) and allele 36.2 of locus IVS3 (χ^2= 13.265; P< 0.001 and OR= 5.961 with

95% CI 2.072-17.154) were significantly associated only with PT-PKD. These two specific alleles were not associated with any particular malaria status group.

The African groups ANG and MOZ showed a marked LD for all pairs of loci (P≤ 0.000). Conversely, the group PT-C only showed LD for the closer loci (PKV/PKA and PKA/IVS11), while the PT-PKD group only evidenced LD for PKV/IVS11. However, when the African malaria status groups were analysed separately, only UM sets from both Angola and Mozambique had significant results for all pairs of loci (P≤ 0.008), i.e., significant LD for a region spanning ≈ 75 Kb (IVS3 was not considered for this test as it was not in Hardy-Weinberg equilibrium). Furthermore, when UM samples from Angola and Mozambique were pooled in one single larger group, the previous result was reinforced: P ≤ 0.000 for all LD tests between locus pairs. Therefore, we searched for a haplotype (PKV/PKA/IVS11/IVS3) that could be associated with this larger UM group and 9/11/13/34 revealed this association, although it was borderline (χ^2= 5.898, P= 0.015; OR= 5.267; 95% CI: 1.188-23.355).

Concerning the population groups studied, they all revealed a large number of low frequency inferred haplotypes. The most common haplotypes were: in ANG, 10/14/12/38, 11/12/15/35, 11/11/17/35 and 10/13/12/34, with an approximate frequency of 3% each; in MOZ, haplotype 9/11/13/34 was prominent (6.3%, from which 5.5% were in UM) and four additional haplotypes were also frequent (≈ 3%): 10/13/14/35, 11/9/17/37.2, 10/13/12/35 and 10/14/12/38; in PT-C, the most frequent haplotype was 9/9/14/40.2 (5.6%), followed by 10/9/14/38.2, 10/9/14/39.2 and 9/9/14/37.2 (about 4%); and in PT-PKD, the most frequent haplotypes were 10/9/14/38.2 (23.8%), 9/9/15/36.2 (19.0%) and 9/9/16/38.2 (11.9%). These last two were not detected in PT-C and 9/9/15/36.2 was exclusively found in PT-PKD.

An AMOVA that considered these four loci for comparison in the follow three populations, Africa (NI, AM, UM and SM from Angola, UM and SM from Mozambique), Portugal - control (PT-C) and Portugal - PK-deficiency (PT-PKD), resulted in a significant percentage of variation between the three populations (10.92%, P= 0.000) and within each group (88.97%, P= 0.000). A non-significant value was obtained between groups within each population (0.12%, P= 0.512).

SNPs

Overall, 15 SNPs were analysed in this study: 13 were identified in the HapMap database and 2 were mutations previously described as associated with PK-deficiency. These mutations were not identified in any of the African groups studied or in the control Portuguese individuals. Mutation 1456C>T was identified in eight Portuguese PK-deficient individuals, two of whom were homozygous for the T allele (Manco *et al*,

1999, 2000). The allele frequencies found in the studied population groups are shown in Supplementary Table IV.

No significant differentiation was found between ANG and MOZ or between PT-C and PT-PKD, whether considering all 13 loci simultaneously or separately. A significant differentiation was found between African and Portuguese groups: F_{ST}(AFR *vs.* PT-C)= 0.239, F_{ST}(AFR *vs.* PT-PKD)= 0.341, P≤ 0.000 for both tests.

Comparing NI, AI, SM and UM from Angola and Mozambique, F_{ST} values were not significant for any pairs of groups tested. Since there were no differences between the two African populations, UM and SM from both countries were pooled into larger groups for comparison, but still no differences were found. The same result was obtained when these groups were compared to NI and AI.

The observed heterozygosity was according to the Hardy-Weinberg expected frequencies in all population groups but, strikingly, when performing an analysis on the malaria status groups from Angola, all loci in UM and SM that were localised in exon 12 (pk_177, pk_176 and pk_972) or downstream (pk_276, pk_184, pk_352 and pk_355) had a deviation from Hardy-Weinberg equilibrium (P< 0.050) with an excess of heterozygotes (as seen in Fig 2). However, when Bonferroni's correction was applied (P< 0.004 for significance), none of these results were statistically significant. However, when individuals of SM and UM were combined into the single ILL group, the deviation was significant even under Bonferroni's correction. These results were not obtained for the Mozambican groups, where the observed heterozygosity was similar to expectation.

African populations showed higher haplotype diversity than the Portuguese. The five main inferred haplotypes (pk_276/pk_184/.../pk_361, ordered as in Fig 1) were identified in both ANG and MOZ and also observed in the malaria status groups from each country. No specific haplotype was associated with any group. In PT-C, two main haplotypes, already identified in the African groups, were observed: G/G/T/C/G/A/G/T/C/G/A/C/A/T/A (frequency of 76%) and A/A/C/G/A/G/T/T/C/C/A/G/C/C/C (frequency of 18%). In PT-PKD, two main haplotypes were identified: one was the most common in PT-C, whereas the other was exclusive to this group, because of the mutation 1456T (G/G/T/C/G/A/G/T/**T**/G/A/C/A/T/A), which was in complete LD with all adjacent loci (Fig 3). When we looked for selective sweeps in African groups in this genomic segment, they were not found: in a general way, the expected heterozygosity in loci from ANG and MOZ was higher but followed the trend observed in PT-C and PT-PKD (Fig 2).

Similarly to AMOVA using the STRs, AMOVA using all of the SNP loci resulted in significant percentages of variation between the populations [Africa (NI, AM, UM and

SM from Angola, UM and SM from Mozambique), Portugal – control (PT-C) and Portugal – PK-deficiency (PT-PKD)] and within each group (25.47%, P= 0.000 and 74.52%, P= 0.000, respectively). The percentage of variation between groups within each population was not significant (≤ 0.00%, P= 0.481).

A combined analysis was performed using all STR and SNP loci, and the results supported those reported above: significant F_{ST} values were obtained when African groups were compared to Portuguese groups. A significant differentiation was also obtained between the two Portuguese groups, PT-C and PT-PKD.

Ancestry Informative INDELs

The structure of African and Portuguese (PT-PKD and PT-REF) groups was examined through the genotyping of 32 INDELs. K=2 was, undoubtedly, the most likely number of clusters, corresponding to the African and Portuguese samples. Even when K=3 to K=5 were tested, the division between African and Portuguese clusters was obvious (Fig 4). A clear differentiation was achieved between African and PT-REF (F_{ST}= 0.392; P≤ 0.000) and African and PT-PKD (F_{ST}= 0.423; P≤ 0.000) groups. MOZ and ANG could be slightly differentiated (F_{ST}= 0.003; P≤ 0.000) by genetic distance analysis but not when using Structure 2.2 software, even when only the two African groups were considered (data not shown). No differentiation was achieved between PT-REF and PT-PKD, or between malaria status groups within MOZ or within ANG under any circumstance.

DISCUSSION

A combined analysis with STR and SNP data was used to search for malaria selection signatures in the *PKLR* gene region. Two different approaches were performed: inter-population analysis, opposing two populations from malaria endemic regions (Angola and Mozambique) to a Portuguese population with no malaria, and an intra-population analysis, comparing malaria status groups within populations.

STR and SNP allelic frequencies in ANG and MOZ were similar and quite different from PT-C and PT-PKD, reflecting structural differences. In fact, when sample structure was tested using ancestry informative INDEL markers, two clusters were clearly formed: one with all ANG and MOZ samples and one including all PT-PKD and PT-REF samples.

F_{ST} among human populations from major geographical regions, based on more than 370 STRs, was estimated to be 0.05 (Rosenberg *et al*, 2002), and it was estimated to be 0.10 when based on 600,000 SNPs (Li *et al*, 2008). Moreover, an AMOVA using the same STR loci (Rosenberg *et al*, 2002) showed 3.6% to 5.2% variation between major regions of the world and 3.1% variation between populations within Africa. In this

study, F_{ST} values obtained between African and Portuguese groups were considerably higher, varying between 0.102 and 0.153 for STRs and between 0.239 and 0.341 for SNPs. In addition, an AMOVA for STR loci had a significant outcome of 10.92% variation between Africans and Portuguese, whereas variation between groups within each population was 0.12%. In a typical multilocus sample, it is reasonable to assume that all autosomal loci have experienced the same demographic history and the same rates and patterns of migration. Loci showing unusually large amounts of differentiation may indicate regions of the genome that have been subject to diversifying selection (Holsinger & Weir, 2009) of which malaria could have been the cause. The AMOVA results show that, whereas variation between Africa and Portugal more than doubled in this study, the opposite occurred in the degree of variation between groups within populations, suggesting that some (selective) force is homogenising this genomic fragment in African regions and, at the same time, extending the differences between Africa and other global areas. Curiously, the F_{ST} value for Africans *vs.* PT-PKD was higher than for Africans *vs.* PT-C, suggesting that, even if PK-deficiency is frequent in sub-Saharan Africa, mutations should be different from those found in the Portuguese.

Concerning the Portuguese groups, differentiation was only significant when STR data was used, which may be explained by the different molecular resolution of SNPs and STRs: in humans, the average nucleotide mutation rate is assumed to be 2.5×10^{-8} and the STR mutation rate has been estimated to be 10^{-2}-10^{-5} per generation (Tishkoff & Verreli, 2003). Thus, SNPs are best used for inferring human evolutionary history over longer time scales and STRs can be used to trace recent demographic events (Agrafioti & Stumpf, 2007). Therefore, we can presume that Portuguese PK-deficiency variants have emerged recently, which is supported by the lower diversity found within this group.

No differentiation was ever obtained between malaria status groups, either using SNPs or STRs, although insufficient sampling of each group may be influencing this result. Of all the STR loci, IVS3 in the *PKLR* gene was the only one with frequencies that were out of Hardy-Weinberg equilibrium in the African groups, with a significant excess of homozygotes. This was already observed in a previous study with African samples from Cabo Verde (Alves *et al*, 2010). Conversely, as expected, the control group PT-C, had a heterozygosity that was similar to that expected. These data suggest that IVS3 homozygosity is being promoted in some manner. Possible causes for the Hardy-Weinberg equilibrium deviation include admixture and substructure or non-random mating patterns. However, as this deviation was observed in several African populations, it is possible that it is caused by the impact of selection pressures from environmental conditions (e.g., infectious diseases like malaria). IVS3 is in intron 3, a critical functional location as it is where the splicing of exon 2 occurs for the production of PKL mRNA, and as it is not a simple polymorphic locus (it includes eight contiguous variation regions), it should be carefully analysed.

The LD test for the STRs showed a significant LD along the entire studied region for UM. This is interesting as suggests an association between this conserved genomic block and a mild malaria outcome. Moreover, this LD emphasises the result previously found in Cabo Verde, where an LD test revealed an association of these same loci but in non-infected individuals (Alves *et al*, 2010). Additionally, this LD outcome is not expected under neutrality, which also supports our results: several datasets show differences in haplotype structure between African and non-African samples, where blocks are significantly smaller in African samples and extend longer and are less diverse in non-Africans (Tishkoff & Verrelli, 2003). Reinforcing the LD result, a haplotype was identified as associated with this group: 9/11/13/34. This association must be further analysed since it is not robust (P= 0.015), but we believe that insufficient sampling may be the cause for this lack as this association was identified only when UM and SM samples from both Angola and Mozambique were pooled together in a larger group.

The LD test for the SNPs had a significant result in all groups and populations for all pairs of SNP loci in exon 12 and upstream (between loci pk_276 and pk_176). Curiously, the ILL group from Angola had a significant SNP heterozygote excess exactly in the same region. Three of these loci are located in exon 12 of *PKLR*, and the remaining are in the *HCN3* gene. This gene, coding for a hyperpolarisation-activated cyclic nucleotide-gated potassium channel 3, is a voltage-gated channel performing ionic, potassium and sodium transport (Uniprot database/Swiss-Prot Q9P1Z3) and is highly expressed in early erythroid cells (Su *et al*, 2004), which produce mature erythrocytes. Heterozygosity in this genomic fragment seems to be associated with clinical malaria in Angola but not in Mozambique, suggesting that, additionally to malaria, some geographic factor may be involved in this scenario.

Five main inferred SNP haplotypes were identified in ANG and MOZ and only two in PT-C (contained within those five) and two in PT-PKD. These results were expected as African populations are older and have maintained a larger N whereas non-African populations have experienced a bottleneck event during the expansion of modern humans out of Africa within the past 100,000 years (Tishkoff & Verreli, 2003). The high mutation rate of STRs explains why the same STR haplotype diversity is present in both African and non-African regions. Haplotype 6 was exclusive to PT-PKD, differing only from haplotype 3 (the most common in PT-C) at the pk_1456 locus. As a result of its strong LD, the segment between pk_276 and pk_176 was extremely well-conserved in all haplotypes, with only two possible allelic combinations. The remaining segment revealed strong recombination. Neither of the two mutations that were potentially associated with PK-deficiency in Africa (as indicated in previous reports) were identified in our African samples.

Previous studies have also examined this particular genomic fragment, seeking for other disease-associated variants. Multiple studies in populations from diverse origins have shown linkage of type 2 diabetes (T2D) to chromosome 1q over a broad region and the *PKLR* gene arises as the first candidate (Wang *et al*, 2002; Das & Elbein 2007; Wang *et al*, 2009). A search for prevalence of T2D in the African continent revealed that Afro-Americans have a twofold increase in risk for T2D compared to other populations in the United States, but its prevalence is lower in Africa (1-2%) than among people of African descendant in industrialised nations (11-13%) (Rotimi *et al*, 2004). In addition, this region includes the *GBA* gene, which codes for the housekeeping enzyme beta-glucocerebrosidase, which has mutations causing Gaucher disease; however, especially high frequencies of this disease have not been detected in Africa (Goldblatt & Beighton, 1979). Therefore, the probability that these diseases would be selectively acting on this genomic region is lower than it is for malaria, denying the possibility of relevant selective confounding factors.

In summary, in this study, several results were obtained supporting the hypothesis that malaria is acting as a selective force in the *PKLR* gene region. Firstly, F_{ST} values between African and Portuguese populations using STR and SNP data from this specific fragment were considerably higher than those found using STR and SNP neutral markers, and the same was observed with AMOVA, revealing that this genomic section is under selection; secondly, the LD block included a more extensive region in the mild malaria group and a haplotype was found to be associated with this clinical group, suggesting that this conserved genomic block is associated with some protection against malaria severity. Thus, the output of this work, using human population data, seems to be in agreement with the results previously obtained with murine models and *in vitro Plasmodium* culturing. For future work, a larger number of samples from malaria status sets should be used and locus IVS3 should be carefully analysed. A more extensive field work with deeper phenotype discrimination and identification of PK abnormal alleles is currently under way.

ACKNOWLEDGMENTS

We thank all individuals and parents/tutors of children who agreed to participate in this study and to all health technicians working at Emergency Services of the Paediatric Hospital David Bernardino (Luanda, Angola), Paediatrics Department of Central Hospital of Maputo, Health Centres of Bagamoyo and Boane (Maputo, Mozambique) for all technical support.

This study was supported by "Financiamento Programático do Laboratório Associado CMDT.LA/IHMT" and POCI/SAU-ESP/55110/2004 (Fundação para a Ciência e Tecnologia/Ministério da Ciência, Tecnologia e Ensino Superior, FCT/MCTES, Portugal).

P. Machado, R. Pereira and A.P. Arez were funded by FCT/MCTES Portugal (SFRH/BD/28236/2006, SFRH/BD/30039/2006 and SFRH/BPD/1624/2000—until 2007, respectively).

REFERENCES

Agrafioti, I. & Stumpf, M.P. (2007) SNPSTR: a database of compound microsatellite-SNP markers. *Nucleic Acids Research*, **35** (Database issue), D71-D75.

Anto, F., Asoala, V., Anyorigiya, T., Oduro, A., Adjuik, M., Owusu-Agyei, S., Dery, D., Bimi, L. & Hodgson, A. (2009) Insecticide resistance profiles for malaria vectors in the Kassena-Nankana district of Ghana. *Malaria Journal*, **8**, 81.

Alves, C., Gomes, V., Prata, M.J., Amorim, A., Gusmão, L. (2007) Population data for Y-chromosome haplotypes defined by 17 STRs (AmpFlSTR YFiler) in Portugal. *Forensic Science International*, **171**, 250-255.

Alves, J., Machado, P., Silva, J., Gonçalves, N., Ribeiro, L., Faustino, P., do Rosário, V.E., Manco, L., Gusmão, L., Amorim, A. & Arez, A.P. (2010) Analysis of malaria associated genetic traits in Cabo Verde, a melting pot of European and sub Saharan settlers. *Blood Cells, Molecules and Diseases*, **44**, 62-68.

Ayi, K., Min-Oo, G., Serghides, L., Crockett, M., Kirby-Allen, M., Quirt, I., Gros, P. & Kain, K.C. (2008) Pyruvate kinase deficiency and malaria. The *New England Journal of Medicine*, **358**, 1805-1810.

Beutler, E. & Gelbart T. (2000) Estimating the prevalence of pyruvate kinase deficiency from gene frequency in the general white population. *Blood*, **95**, 3585-3588.

Bruce-Chwatt, L.J. (1977) Malaria eradication in Portugal. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **71**, 232-240.

Campbell, M.C. & Tishkoff, S.A. (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annual Review of Genomics and Human Genetics*, **9**, 403-433.

Ceesay, S.J., Casals-Pascual, C., Erskine, J., Anya, S.E., Duah, N.O., Fulford, A.J.C., Sesay, S.S.S., Abubakar, I., Dunyo, S., Sey, O., Palmer, A., Fofana, M., Corrah, T., Bojang, K.A., Whittle, H.C., Greenwood, B.M. & Conway, D.J. (2008) Changes in malaria indices between 1999 and 2007 in The Gambia: a retrospective analysis. *Lancet*, **372**, 1545–1554.

Cuamba, N., Choi, K.S. & Townson, H. (2006) Malaria vectors in Angola: distribution of species and molecular forms of the Anopheles gambiae complex, their pyrethroid insecticide knockdown resistance (kdr) status and Plasmodium falciparum sporozoite rates. *Malaria Journal*, **5**, 2.

Das, S.K. & Elbein, S.C. (2007) The search for type 2 diabetes susceptibility loci: the chromosome 1q story. *Current Diabetes Reports*, **7**, 154-164.

Durand, P.M. & Coetzer, T.L. (2008) Pyruvate kinase deficiency protects against malaria in humans. *Haematologica*, **93**, 939-940.

Excoffier, L., Laval, G. & Schneider, S. (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.

Goios, A., Gusmão, L., Rocha, A.M., Fonseca, A., Pereira, L., Bogue, M. & Amorim, A. (2008) Identification of mouse inbred strains through mitochondrial DNA single-nucleotide extension. *Electrophoresis*, **29**, 4795-4802.

Goldblatt, J. & Beighton, P. (1979) Gaucher disease in the Afrikaner population of South Africa. *South African Medical Journal*, **55**, 209-210.

Hill, A.V. (1998) Host genetics of infectious diseases: old and new approaches converge. *Emerging Infectious Disease*, **4**, 695-697.

Holsinger, K.E. & Weir, B.S. (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews Genetics*, **10**, 639-650.

Kalinowski, S.T. (2005) HP-Rare: a computer program for performing rarefaction on measures of allelic diversity. *Molecular Ecology Notes*, **5**, 187-189.

Kwiatkowski, D.P. (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. *The American Journal of Human Genetics*, **77**, 171-192.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L. & Myers, R.M. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100-1104.

Mabunda, S., Casimiro, S., Quinto, L. & Alonso, P. (2008) A country-wide malaria survey in Mozambique. I. Plasmodium falciparum infection in children in different epidemiological settings. *Malaria Journal*, **7**, 216.

Manco, L. & Abade, A. (2001b) Pyruvate kinase deficiency: prevalence of the 1456C-->T mutation in the Portuguese population. *Clinical Genetics*, **60**, 472-473.

Manco, L., Oliveira, A.L., Gomes, C., Granjo, A., Trovoada, M., Ribeiro, M.L., Abade, A. & Amorim, A. (2001a) Population genetics of four PKLR intragenic polymorphisms in Portugal and São Tomé e Princípe (Gulf of Guinea). *Human Biology*, **73**, 467-474.

Manco, L., Ribeiro, M.L., Almeida, H., Freitas, O., Abade, A. & Tamagnini, G. (1999) PK-LR gene mutations in pyruvate kinase deficient Portuguese patients. *British Journal of Haematology*, **105**, 591-595.

Manco, L., Ribeiro, M.L., Máximo, V., Almeida, H., Costa, A., Freitas, O., Barbot, J., Abade, A. & Tamagnini, G. (2000) A new PKLR gene mutation in the R-type promoter region affects the gene transcription causing pyruvate kinase deficiency. *British Journal of Haematology*, **110**, 993-997.

Manco, L., Trovoada, M.J. & Ribeiro M.L. (2009) Novel Human Pathological Mutations. Gene Symbol : PKLR. Disease : Pyruvate kinase deficiency. *Human Genetics*, **125**, 340.

Mateu, E., Perez-Lezaun, A., Martinez-Arias, R., Andres, A., Vallés, M., Bertranpetit, J. & Calafell, F. (2002) PKLR-GBA region shows almost complete linkage disequilibrium over 70 kb in a set of worldwide populations. *Human Genetics*, **110**, 532-544.

Min-Oo, G., Fortin, A., Tam, M.F., Nantel, A., Stevenson, M.M. & Gros, P. (2003) Pyruvate kinase deficiency in mice protects against malaria. *Nature Genetics*, **35**, 357-362.

Min-Oo, G. & Gros, P. (2005) Erythrocyte variants and the nature of their malaria protective effect. *Cellular Microbiology*, **7**, 753-763.

Noedl, H., Se, Y., Schaecher, K., Smith, B.L., Socheat, D., Fukuda, M.M. (2008) Evidence of artemisinin-resistant malaria in western Cambodia. *The New England Journal of Medicine*, **359**, 2619-2620.

O'Meara, W.P, Bejon, P., Mwangi, T.W., Okiro, E.A., Peshu, N., Snow, R.W., Newton, C.R. & Marsh, K. (2008) Effect of a fall in malaria transmission on morbidity and mortality in Kilifi , Kenya. *Lancet*, **372**, 1555–1562.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. & Feldman, M.W. (2002) Genetic structure of human populations. *Science*, **298**, 2381-2385.

Rotimi, C.N., Chen, G., Adeyemo, A.A., Furbert-Harris, P., Parish-Gause, D., Zhou, J., Berg, K., Adegoke, O., Amoah, A., Owusu, S., Acheampong, J., Agyenim-Boateng, K., Eghan, B.A. Jr, Oli, J., Okafor, G., Ofoegbu, E., Osotimehin, B., Abbiyesuku, F., Johnson, T., Rufus, T., Fasanmade, O., Kittles, R., Daniel, H., Chen, Y., Dunston, G. & Collins, F.S. (2004) A genome-wide search for type 2 diabetes susceptibility genes in West Africans: the Africa America Diabetes Mellitus (AADM) Study. *Diabetes*, **53**, 838-841. Erratum in: *Diabetes*, **53**, 1404.

Rozen, S. & Skaletsky, H.J. (2000) Primer 3 on the WWW for general users and for biologist programmers. In: *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, edited by Krawetz, S. & Misener, S., Humana Press Inc., 365-386.

Sanchez, J.J., Borsting, C. & Morling, N. (2005) Typing of Y chromosome SNPs with multiplex PCR methods (cap.15). In: *Methods in Molecular Biology*, **297** (Forensic DNA Typing Protocols), edited by Carracedo, A., Humana Press Inc.

Santos, N.P., Ribeiro-Rodrigues, E.M., Ribeiro-dos-Santos, A.K., Pereira, R., Gusmão, L., Amorim, A., Gerreiro, J.F., Zago, M.A., Matte, C., Hutz, M.H. & Santos, S.E. (2010) Assessing individual interethnic admixture and population substructure using a 48 insertion-deletion ancestry-informative marker panel. *Human Mutation*, **31**, 184-190.

Su, A., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M.P., Walker, J.R. & Hogenesch, J.B. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences*, **101**, 6062-6067.

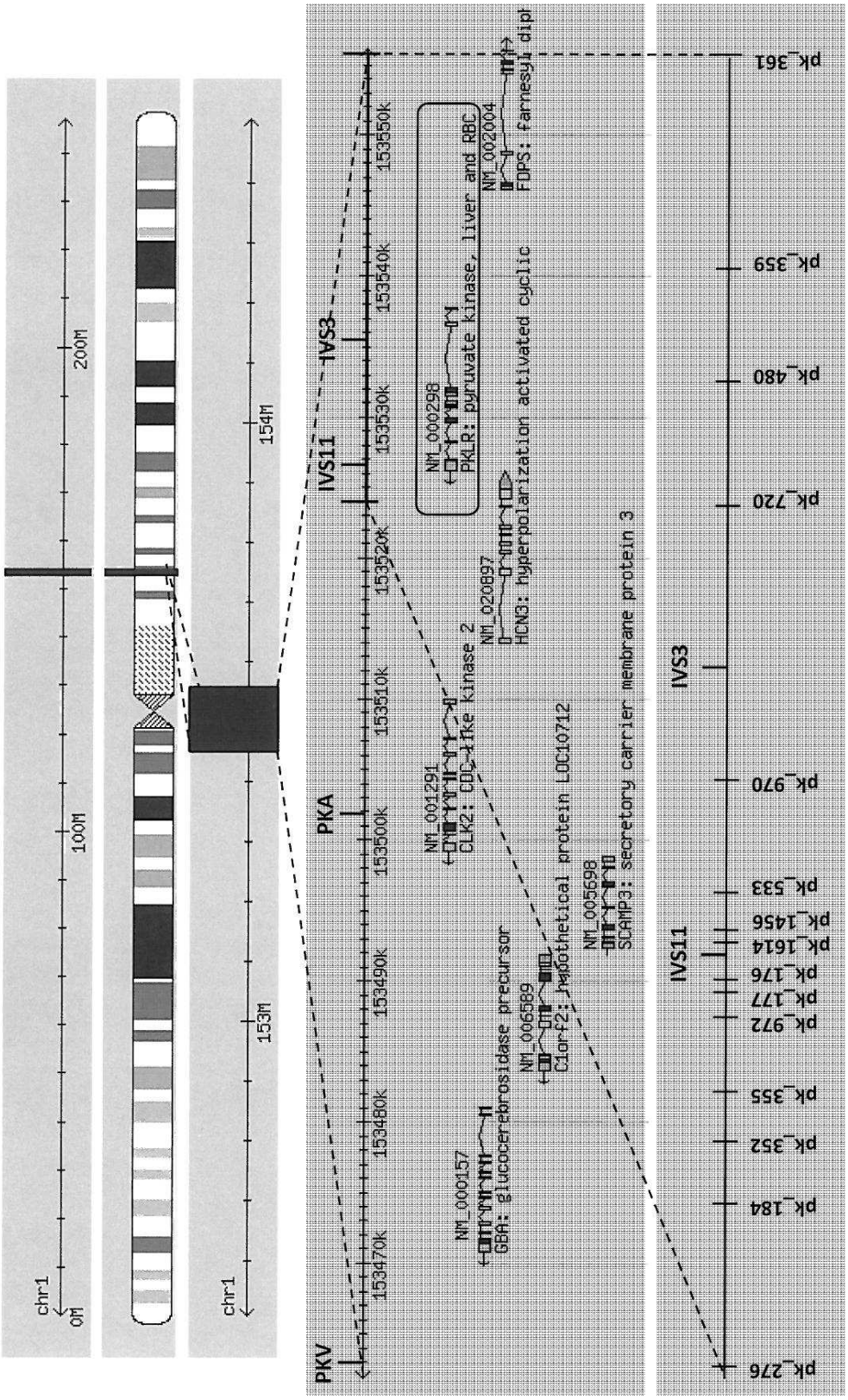
Tishkoff, S.A. & Verrelli, B.C. (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annual Review of Genomics and Human Genetics*, **4**, 293–340.

Wang, H., Chu, W., Das, S.K., Ren, Q., Hasstedt, S.J. & Elbein, S.C (2002) Liver pyruvate kinase polymorphisms are associated with type 2 diabetes in northern European Caucasians. *Diabetes*, **51**, 2861-5.

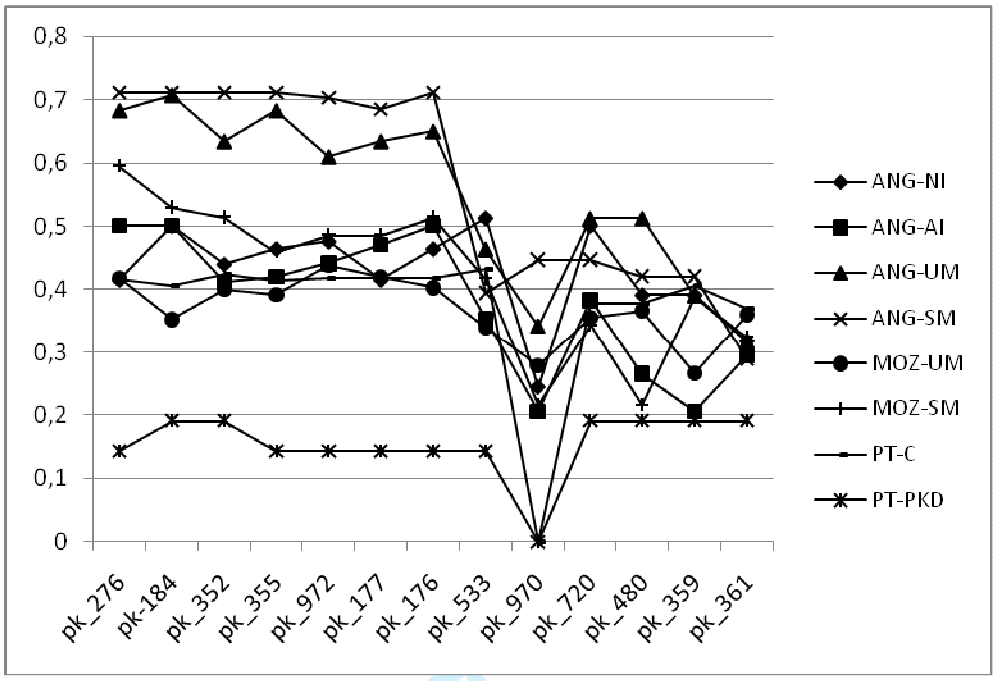
Wang, H., Hays, N.P., Das, S.K., Craig, R.L., Chu, W.S., Sharma, N. & Elbein, S.C. (2009) Phenotypic and molecular evaluation of a chromosome 1q region with linkage and association to type 2 diabetes in humans. *Journal of Clinical Endocrinology & Metabolism*, **94**, 1401-1408.

Williams, T.N. (2006) Red blood cell defects and malaria. *Molecular and Biochemical Parasitology*, **149**, 121-127.

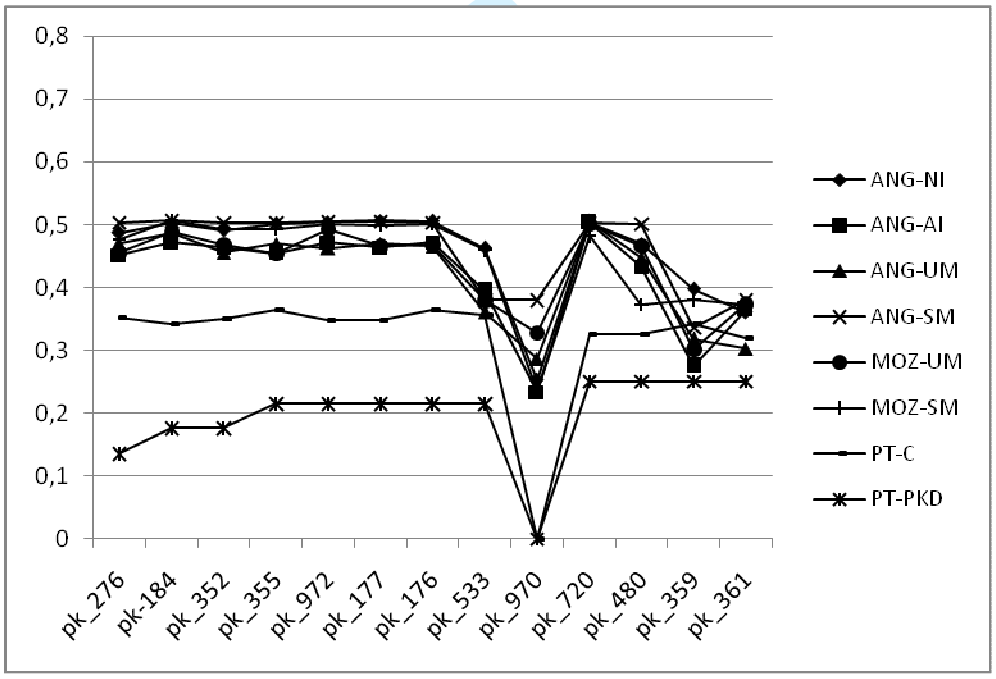
World Malaria Report 2008, World Health Organization (WHO) 2008: http://apps.who.int/malaria/wmr2008/malaria2008.pdf

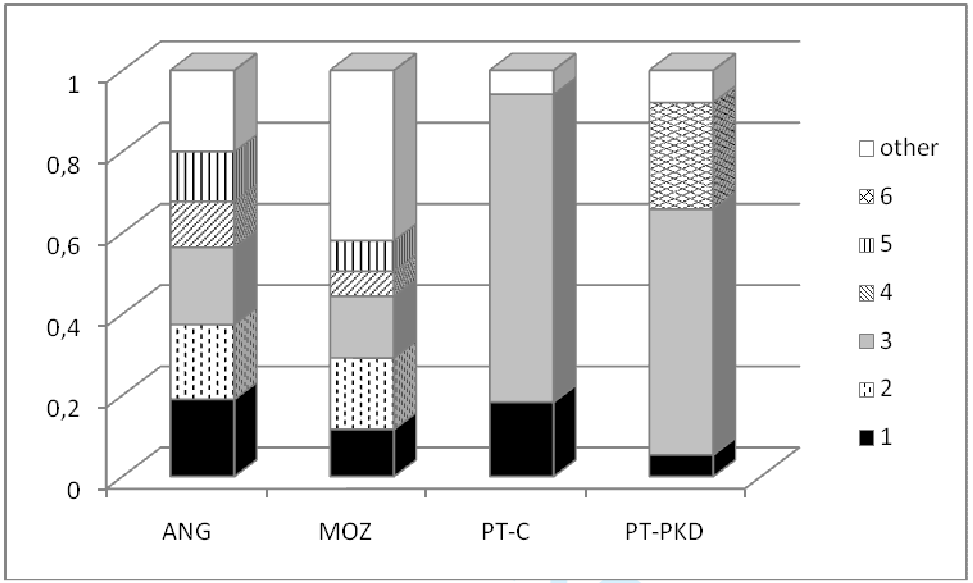

A.
chr1
0M
100M
200M
153M
154M
chr1
B.
PKV
PKA
IVS11
IVS3
153470k
153480k
153490k
153500k
153510k
153520k
153530k
153540k
153550k
NM_000298
PKLR: pyruvate kinase, liver and RBC
NM_002004
FDPS: farnesyl dipl
NM_001291
CLK2: CDC-like kinase 2
NM_020897
HCN3: hyperpolarization activated cyclic
NM_000157
GBA: glucocerebrosidase precursor
NM_006589
C1orf2: hypothetical protein LOC10712
NM_005698
SCAMP3: secretory carrier membrane protein 3
C.
IVS3
IVS11
pk_361
pk_359
pk_480
pk_720
pk_970
pk_533
pk_1456
pk_1614
pk_176
pk_177
pk_972
pk_355
pk_352
pk_184
pk_276

A.


0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
pk_276
pk-184
pk_352
pk_355
pk_972
pk_177
pk_176
pk_533
pk_970
pk_720
pk_480
pk_359
pk_361
ANG-NI
ANG-AI
ANG-UM
ANG-SM
MOZ-UM
MOZ-SM
PT-C
PT-PKD

B.


0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
pk_276
pk-184
pk_352
pk_355
pk_972
pk_177
pk_176
pk_533
pk_970
pk_720
pk_480
pk_359
pk_361
ANG-NI
ANG-AI
ANG-UM
ANG-SM
MOZ-UM
MOZ-SM
PT-C
PT-PKD



Haplotype	pk_276	pk_184	pk_352	pk_355	pk_972	pk_177	pk_176	pk_1614	pk_1456	pk_533	pk_970	pk_720	pk_480	pk_359	pk_361
	G/A	**G/A**	**T/C**	**C/G**	**G/A**	**A/G**	**G/T**	**T>A**	**C>T**	**G/C**	**A/G**	**C/G**	**A/C**	**T/C**	**A/C**
1	A	A	C	G	A	G	T	T	C	C	A	G	C	C	C
2	A	A	C	G	A	G	T	T	C	C	A	C	A	C	A
3	G	G	T	C	G	A	G	T	C	G	A	C	A	T	A
4	A	A	C	G	A	G	T	T	C	C	A	G	A	C	A
5	G	G	T	C	G	A	G	T	C	C	G	G	C	C	A
6	G	G	T	C	G	A	G	T	**T**	G	A	C	A	T	A

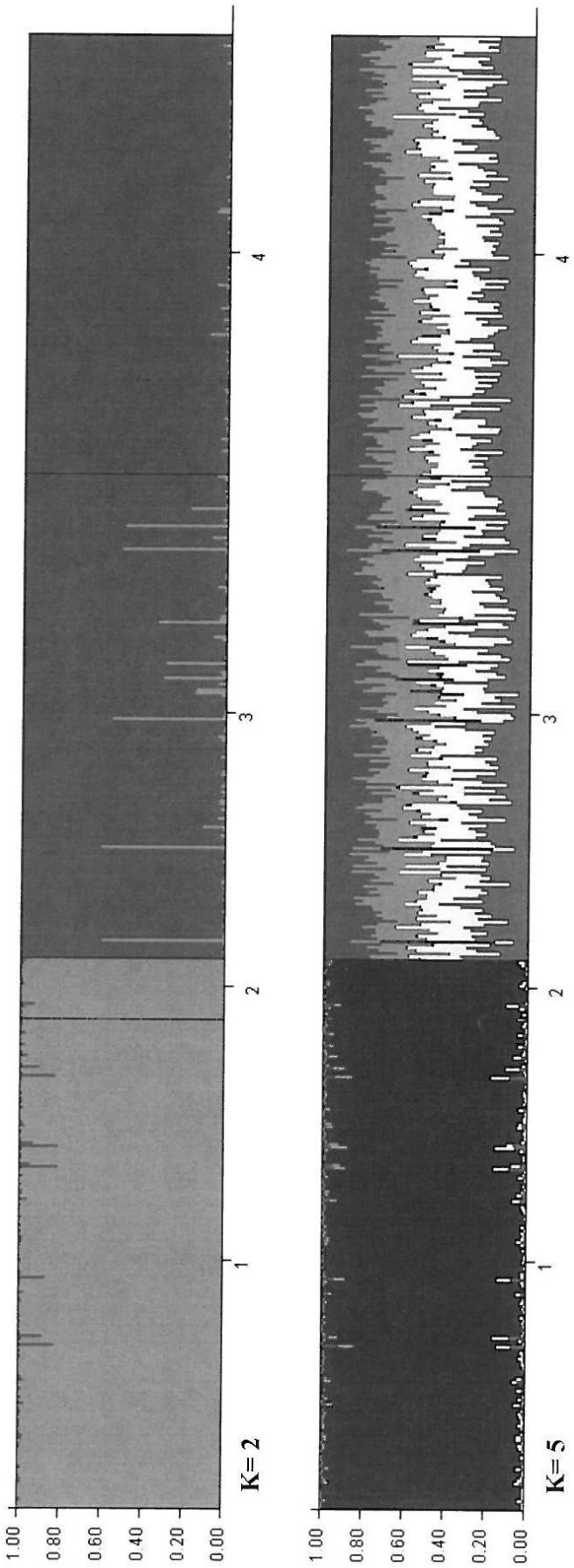

1.00
0.80
0.60
0.40
0.20
0.00
1
2
3
4
K= 2
K= 5

Fig 1. The 95 Kb fragment analysed in this study, including *PKLR* gene. **A.** Localisation in chromosome 1q21; **B.** The 4 STR loci (PKV, PKA, IVS11 and IVS3) genotyped in the present study and genes near *PKLR*; **C.** The 15 SNP loci analysed spread along a region closer to the gene *PKLR*. *Adapted from* www.hapmap.org.

Fig 2. Observed heterozygosity (**A.**) and expected heterozygosity (**B.**) of the SNP loci in Portuguese groups and malaria status groups from both Angola and Mozambique. ANG-UM and ANG-SM revealed a heterozygote excess for all loci included between pk_276 and pk_176. ANG-NI: Angola - non-infected; ANG-AI: Angola - asymptomatic infection; ANG-UM: Angola - uncomplicated malaria; ANG-SM: Angola - severe malaria; MOZ-UM: Mozambique - uncomplicated malaria; MOZ-SM: Mozambique - severe malaria; PT-C: Portugal - control group; PT-PKD: Portugal - PK-deficiency group.

Fig 3. Estimated frequencies of inferred haplotypes in the studied population groups. ANG: Angola; MOZ: Mozambique; PT-C: control Portuguese; PT-PKD: Portuguese with PK-deficiency. The segment between pk_276 and pk_176 was extremely conserved in all haplotypes with only two possible allelic combinations (two different grey shades in the table).

Fig 4. Estimated population structure determined with Structure 2.2. (no prior information of sampling groups, under the admixture model with correlated allele frequencies; ten independent runs with 10^5 burn-in steps and 10^5 interactions). Each bar represents a single individual and is partitioned into K different grey shaded segments that represent the individual's estimated coefficients of ancestry. **K= 2** is the most suitable division, with clusters corresponding to the Portuguese (mainly light grey) and African (mainly dark grey) samples. **1- PT-REF** [reference group from Portugal (Santos *et al*, 2010)]; **2- PT-PKD** (individuals with PK-deficiency from Portugal) **3- ANG** (Angola); **4-MOZ** (Mozambique).

Supplementary Table I. SNP loci selected for analysis (ordered according to localization), allelic frequencies and primers used for multiplex PCR.

SNP (along 40970 bp)	**Allelic Frequency**		**Primer**	**Product (bp)**	**Primers Sequence (5'-3')**
11055 bp after TGA					
refSNP rs7549276 – pk_276 – gene *HCN3*	G	A	pk_276	**442**	GCTGTCCCTAGTGCTGAAGG
chr1:153515199..153515199	0.500	0.500			GACTAGAAAAGGCGCACTGG
(5008 bp)					
refSNP rs7520184 – pk_184 – gene *HCN3*	G	A	pk_184	**413**	CTGCACCCACTAACTCGTCA
chr1:153520207..153520207	0.583	0.417			CAGCCTGGCAAATTCTCTTC
(2254 bp)					
refSNP rs11264352 – pk_352 – gene *HCN3*	T	C	pk_352	**127**	ATCCTACTTTGGGGGTCAGC
chr1:153522461..153522461	0.542	0.458			GGCTGGAGCTCTGTGATTCT
(1655 bp)					
refSNP rs11264355 – pk_355 – gene *HCN3*	C	G	pk_355	**393**	TGAGTACCAGTCCCCTGACC
chr1:153524116..153524116	0.569	0.431			GTACCAGTGGCTCCCACAGT
(2604 bp)					
chr1: 153526254 – *pkLR* gene TGA					
refSNP rs932972 - pk_972 - EXON 12	C	T			
chr1:153526720..153526720	0.583	0.417			
(254 bp)					
refSNP rs1052177 – pk_177 - EXON 12	T	C	pk_972_177_176	**406**	CTGGTGATTGTGGTGACAGG
chr1:153526974..153526974	0.585	0.415			AACCAGCCAAACTGGGATTA
(33 bp)					
refSNP rs1052176 – pk_176 - EXON 12	C	A			
chr1:153527007..153527007	0.583	0.417			
(1168 bp)					
1614A>T – pk_1614 - EXON 11					
chr1:153528175..153528175					
(158 bp)		Mutations associated to PK-deficiency	pk_mut	**372**	TGACACCTGGAACTGGAACA
1456C>T – pk_1456 - EXON 11					GACCACAGGAGAGAGGCAAG
chr1:153528333..153528333					
(904 bp)					

refSNP rs4620533 – pk_533 - INTRON 10	C	G	pk_533	**180**	TCCTGTTAATCCTGCCAACC
chr1:153529237..153529237	0.517	0.483			GCTCAGAGGCAAGTCCATTC
(3048 bp)					
refSNP rs8177970 – pk_970 - INTRON 3	A	G	pk_970	**151**	AGGGAAGGGGAGTCTGTGAT
chr1:153532285..153532285	0.892	0.108			TCACGTTCAGACAACGTTCC
(9299 bp)					
chr1: 153537835 – ATG of *pkLR* gene					
refSNP rs12032720 – pk_720			pk_720	**321**	GGCACCCATAGGAGATGAGA
chr1:153541584..153541584	G	C			CTCCACTATCTGGGCCTGAA
(4522 bp)	0.700	0.300			
refSNP rs2297480 – pk_480 - gene *FDPS*	A	C	pk_480	**357**	GAAGACCCCCACAGATCTCA
chr1:153546106..153546106	0.783	0.217			TCCTTTCAGCCCCTAATCCT
(3347 bp)					
refSNP rs11264359 – pk_359 - gene *FDPS*	A	G	pk_359	**206**	TCCAAAGGCTATTCAGAAGCA
chr1:153549453..153549453	0.375	0.625			GCAGAAGTTGCATCCACTCA
(6716 bp)					
refSNP rs11264361 – pk_361 - gene *FDPS*	T	G	pk_361	**212**	CACCAGCTTCACTCCTCCTC
chr1:153556169..153556169	0.817	0.183			GGCACCTTCAGGATCTGGTA
(5227 bp)					
18 334 bp before ATG					

(**refSNP rs...– pk_"nr"– "x"** - SNP reference in HapMap – SNP designation in the study – localization; chr1: ... – localization in chromosome 1; **("nr" bp)** – distance between adjacent SNPs; **Allelic Frequency** – allelic frequencies in Nigerian population, African populations reference in Hapmap; **ATG** – *pkLR* gene start codon; **TGA** – STOP codon of *pkLR* gene.

Supplementary Table II. Single Base Extension (SBE) primers used for SNaPshot reaction.

Target region	SNP	Mutation	Detection	Conc. (µM)	SBE-primer sequence
pk_972_177_176	pk_177		A>G	0.4	**GTAGGCTGGGCCAGAGG**
pk_352	pk_352		T>C	0.4	GTCTGACAA**GCTCTGGGTCCCTGCC**
pk_972_177_176	pk_972		G>A	1.22	TCTGACAA**CTGAGCAGATTGGATGCAG**
pk_184	pk_184		G>A	0.4	**CCTATCTATAAGATGAGAGAAATAAGAAACT**
pk_276	pk_276		G>A	0.4	GTGAAAGTCTGACAA**CCCATTGTTCCTTTCACTCCT**
pk_355	pk_355		C>G	1.22	GCCACGTCGTGAAAGTCTGACAA**CCCACCCCATCCTGATA**
pk_720	pk_720		C>G	0.4	AGGTGCCACGTCGTGAAAGTCTGACAA**GGGCAAGGGTGTTGGTAAA**
pk_mut	pk_1614		T>A	0.2	GCCACGTCGTGAAAGTCTGACAA**GAAGGTCTAGGTAGCTCACCACT**
pk_480	pk_480		A>C	0.4	AAACTAGGTGCCACGTCGTGAAAGTCTGACAA**CCAGATAACTCCCACCCC**
pk_972_177_176	pk_176		G>T	0.4	GACTAAACTAGGTGCCACGTCGTGAAAGTCTGACAA**CAGGATATGCTTAGCACCC**
pk_361	pk_361		A>C	1.22	TGACTAAACTAGGTGCCACGTCGTGAAAGTCTGACAA**CAGCAAAAGAGGAAGGATG**
pk_mut	pk_1456		C>T	1.22	CAACTGACTAAACTAGGTGCCACGTCGTGAAAGTCTGACAA**CTCAGCCCAGCTTCTGTCT**
pk_970	pk_970		A>G	0.4	CAACTGACTAAACTAGGTGCCACGTCGTGAAAGTCTGACAA**GGTTGCATCAGGGAATAAAG**
pk_359	pk_359		T>C	0.4	CCCCCAACTGACTAAACTAGGTGCCACGTCGTGAAAGTCTGACAA**AGTGAGCTGCCAGTTTTCAAT**
pk_533	pk_533		G>C	0.12	CCCCCCCCCCAACTGACTAAACTAGGTGCCACGTCGTGAAAGTCTGACAA**AGAAATGTAGCTCTATTAGCCTGCT**

Target region – Multiplex PCR product; **Detection** – alternative alleles detected; **Conc. (µM) –** concentration in the SBE-primer mix; **bold nucleotides in SBE-primer sequence** – target sequence of the SBE-primes; **nucleotides not in bold** – neutral sequence as described in Sanchez *et al,* 2005.

Supplementary Table III. STR loci allele frequencies found in Angola (ANG), Mozambique (MOZ), control Portuguese (PT-C) and PK-deficient Portuguese (PT-PKD).

Loci	Allele	ANG	MOZ	PT-C	PT-PKD
IVS11	**7**	0.007	0.012	0.000	0.000
	9	0.000	0.000	0.006	0.000
	10	0.058	0.067	0.006	0.000
	11	0.000	0.000	0.000	0.000
	12	0.273	0.287	0.156	0.071
	13	0.054	0.146	0.063	0.000
	14	0.115	0.063	0.506	0.452
	15	0.155	0.071	0.188	0.262
	16	0.119	0.118	0.0313	0.167
	17	0.209	0.197	0.038	0.048
	18	0.011	0.039	0.006	0.000
PKV	**8**	0.011	0.008	0.025	0.024
	9	0.162	0.197	0.406	0.452
	10	0.428	0.433	0.481	0.476
	11	0.381	0.354	0.075	0.048
	12	0.018	0.008	0.013	0.000
PKA	**8**	0.004	0.008	0.000	0.000
	9	0.248	0.252	0.688	0.929
	10	0.054	0.047	0.075	0.024
	11	0.216	0.244	0.019	0.000
	12	0.151	0.079	0.013	0.000
	13	0.162	0.193	0.044	0.000
	14	0.104	0.134	0.069	0.024
	15	0.043	0.016	0.075	0.000
	16	0.018	0.028	0.019	0.000
	17	0.000	0.000	0.000	0.024
IVS3	**30**	0.004	0.000	0.006	0.000
	31	0.000	0.004	0.013	0.000
	31.2	0.000	0.004	0.000	0.000
	32	0.025	0.043	0.013	0.000
	32.2	0.000	0.000	0.013	0.000
	33	0.061	0.067	0.031	0.000
	34	0.176	0.220	0.056	0.024
	34.2	0.007	0.000	0.031	0.048
	35	0.198	0.177	0.031	0.024
	35.2	0.036	0.008	0.050	0.024
	36	0.097	0.087	0.056	0.000
	36.2	0.032	0.016	0.044	0.214
	37	0.061	0.047	0.056	0.024
	37.2	0.076	0.083	0.163	0.048
	38	0.050	0.031	0.019	0.024
	38.2	0.054	0.075	0.194	0.381
	39	0.022	0.024	0.025	0.000
	39.2	0.040	0.051	0.100	0.167
	40	0.004	0.008	0.013	0.024
	40.2	0.014	0.031	0.088	0.000
	41	0.004	0.000	0.000	0.000
	41.2	0.040	0.024	0.000	0.000

Supplementary Table IV. SNP loci allelic frequencies observed in Angola, Mozambique and Portuguese groups.

SNP loci	Allele	Population groups			
		ANG	MOZ	PT-C	PT – PKD
pk_276	**A**	0.610	0.646	0.222	0.053
	G	0.390	0.354	0.778	0.947
pk_184	**A**	0.566	0.549	0.213	0.079
	G	0.434	0.451	0.788	0.921
pk_352	**C**	0.610	0.612	0.220	0.079
	T	0.390	0.388	0.780	0.921
pk_355	**C**	0.404	0.373	0.768	0.895
	G	0.596	0.627	0.232	0.105
pk_972	**A**	0.588	0.566	0.219	0.105
	G	0.412	0.434	0.781	0.895
pk_177	**A**	0.423	0.393	0.781	0.895
	G	0.577	0.607	0.219	0.105
pk_176	**G**	0.414	0.399	0.769	0.895
	T	0.586	0.601	0.231	0.105
pk_1614	**A**	1.000	1.000	1.000	1.000
	T	0.000	0.000	0.000	0.000
pk_1456	**C**	1.000	1.000	1.000	0.737
	T	0.000	0.000	0.000	0.263
pk_533	**C**	0.726	0.715	0.225	0.105
	G	0.274	0.285	0.775	0.895
pk_970	**A**	0.826	0.818	1.000	1.000
	G	0.174	0.182	0.000	0.000
pk_720	**C**	0.515	0.565	0.800	0.868
	G	0.485	0.435	0.200	0.132
pk_480	**A**	0.632	0.680	0.800	0.868
	C	0.368	0.320	0.200	0.132
pk_359	**C**	0.790	0.794	0.219	0.132
	T	0.210	0.206	0.781	0.868
pk_361	**A**	0.771	0.758	0.805	0.868
	C	0.229	0.242	0.195	0.132