Patterns of Selection in Anti-Malarial Immune Genes in Malaria Vectors: Evidence for Adaptive Evolution in LRIM1 in Anopheles arabiensis

Michel A. Slotman1*, Aristeidis Parmakelis1, Jonathon C. Marshall1++, Parfait H. Awono-Ambene2, Christophe Antonio-Nkondjo2, Frederic Simard2*, Adaligsa Caccone1, Jeffrey R. Powell1

1 Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut, United States of America, 2 Organisation de Coordination pour la Lutte Contre les Endémies en Afrique Centrale, Yaoundé, Cameroon, 3 Institute of Recherche pour le Développement, Yaoundé, Cameroon

Background. Co-evolution between Plasmodium species and its vectors may result in adaptive changes in genes that are crucial components of the vector’s defense against the pathogen. By analyzing which genes show evidence of positive selection in malaria vectors, but not in closely related non-vectors, we can identify genes that are crucial for the mosquito’s resistance against Plasmodium. Methodology/Principle Findings. We investigated genetic variation of three anti-malarial genes; CEC1, GNBP-B1 and LRIM1, in both vector and non-vector species of the Anopheles gambiae complex. Whereas little protein differentiation was observed between species in CEC1 and GNBP-B1, McDonald-Kreitman and maximum likelihood tests of positive selection show that LRIM1 underwent adaptive evolution in a primary malaria vector; An. arabiensis. In particular, two adjacent codons show clear signs of adaptation by having accumulated three out of four replacement substitutions. Furthermore, our data indicate that this LRIM1 allele has introgressed from An. arabiensis into the other main falciparum infection, Anopheles gambiae. Conclusions/Significance. Although no evidence exists to link the adaptation of LRIM1 to P. falciparum infection, an adaptive response of a known anti-malarial gene in a primary malaria vector is intriguing, and may suggest that this gene could play a role in Plasmodium resistance in An. arabiensis. If so, our data also predicts that LRIM1 alleles in An. gambiae vary in their level of resistance against P. falciparum.

INTRODUCTION

Despite ongoing control efforts during the last decades, malaria remains one of the most deadly infectious diseases. The vast majority of its burden is carried by people on the African continent, where 1 to 2 million people die annually from this disease [1]. Current malaria control efforts are hampered by the spread of insecticide and drug resistance, which has inspired research programs aimed at the development of insecticide and drug resistance, which has inspired research programs aimed at the development and eventual release of genetically altered mosquitoes that would be resistant to Plasmodium falciparum transmission. The need to identify refractory genes for this effort has focused much attention on the immune system of malaria’s main vector in Africa, An. gambiae. The completion of the An. gambiae genome [2] has greatly facilitated research in this direction, and various anti-malarial immunity genes have now been identified [e.g. 3–6]. Additionally, two recent studies provided many candidate anti-malarial immune genes that are up-regulated in response to Plasmodium infection [7,8].

So far little attention has been devoted to examining polymorphism of immunity genes in natural malaria vector populations [9,10]. It is known however that molecules that are involved in interactions with pathogens, such as immune genes, are one of the major types of proteins on which positive selection has been demonstrated [11,12]. Presumably, this is because such genes are involved in co-evolution between hosts and pathogens. In the case of malaria, if Plasmodium infection affects the mosquito’s fitness, we may expect the accumulation of adaptive amino acid substitutions in those anti-malarial genes that are crucial in specifically limiting Plasmodium infection in vector species, whereas such changes should not be found in closely related species that do not transmit malaria.

That An. gambiae has in fact undergone an adaptive response to P. falciparum infection is suggested by several lines of evidence. First of all, P. falciparum goes through severe bottlenecks during its life cycle in this mosquito [13], demonstrating that the mosquito immune system is limiting the Plasmodium infection. Furthermore, P. berghei, which is not transmitted naturally by An. gambiae, produces a much higher oocyst number in An. gambiae than its natural pathogen P. falciparum. In fact, in a review of studies estimating the fitness effect of Plasmodium infection on Anopheles species, reduced fitness was observed in 10 combinations of Plasmodium and Anopheles species that do not occur naturally, whereas in 10 natural combinations, including An. gambiae and P. falciparum, no fitness effects were observed [14]. This is an indication that Anopheles species have evolved to limit infections of the Plasmodium species they come into contact with. This is corroborated by the fact that the immune response to P. falciparum...
is specific, i.e. *An. gambiae* up-regulates different genes in response to infection with *P. falciparum* vs. *P. berghei* [5,7]. Salivary gland infection rates of *P. falciparum* in *An. gambiae* and *An. arabiensis* are typically low, ranging between 3–9% [15]. This raises the question whether selection pressures on the mosquito immune system are strong enough to result in an adaptive response to *P. falciparum* infection. However, it should be kept in mind that the data summarized above indicate that the rate and intensity of *P. falciparum* infection is likely to have been much higher when *Anopheles* mosquitoes first came into contact with this pathogen.

*An. gambiae* belongs to a complex of closely related species that includes another primary African vector, *An. arabiensis*. Additionally it contains several species, i.e. *An. melas*, *An. merus* and *An. lavansae*, that occasionally transmit malaria locally, but do not have wide enough distributions to be considered important vectors. More importantly for the purpose of the present study, the *An. gambiae* complex also contains the highly zoophilic *An. quadriannulatus A* and *An. quadriannulatus B*, which are never or rarely exposed to the human-limited *P. falciparum*.

In this study, we investigated patterns of polymorphism in three anti-malarial genes, i.e. CEC1, GNB-B1 and LRIM1, in six species of the *An. gambiae* complex. CEC1 (ENSANG00000009468, www.ensembl.org/Anopheles_gambiae) is a cecropin gene whose expression in *An. gambiae* is induced by infection with bacteria and *Plasmodium berghei* [16]. Additionally, genetically modified *An. gambiae* that express CEC1 24 hours after a blood meal, showed a 60% reduction in the number of *P. berghei* oocysts [4]. GNB-B1 (ENSANGG00000015205) is a pattern recognition receptor whose expression is strongly upregulated in response to infection with both *P. berghei* [17], and *P. falciparum* [5]. LRIM1 (ENSANGG00000010552) is a leucine-rich repeat immune protein that is an important plasmodium antagonist. This protein is upregulated in response to infection with *P. berghei*, and silencing of this gene increases oocyst load 3.6-fold [6]. Furthermore, this gene has been implicated in the melanization reaction of parasites [18].

We performed various tests for positive selection on these anti-malarial genes in the two main vectors, *An. gambiae* and *An. arabiensis*, to examine if these genes show signs of an adaptive response that may implicate them in the co-evolution of the mosquito vector and *Plasmodium* pathogen. Whereas no evidence for positive selection was found in CEC1 and GNB-B1, our results clearly indicate that LRIM1 underwent an adaptive response in the *An. arabiensis* lineage. Additionally our data also indicate that LRIM1 has introgressed from *An. arabiensis* into *An. gambiae*.

### RESULTS

The complete CEC1 gene, consisting of 177 bp of coding sequence and two introns comprising a combined 90 bp, was amplified. A total of 186 alleles of this gene were obtained from six species of the *An. gambiae* complex, 66 of which were unique (Table 1, genbank accession nos EU073426–EU073462). All of these alleles were unique, but some polymorphisms were shared between species. *Dxy* ranged from 0.723 to 2.5 (per 100 bp) for the coding region. Very few fixed replacement substitutions were observed between species (Table 3). Two of the McDonald-Kreitman tests were significant. However, both indicated an excess of non-synonymous polymorphisms, and in all comparisons the ratio of non-synonymous to synonymous substitutions was higher for polymorphisms than for fixed differences. Between the non-vector species *An. quadriannulatus A* and the malaria vectors *An. arabiensis* and *An. gambiae*, only a single replacement substitution was observed.

For LRIM1 we sequenced 858 bp that were thought to represent a single exon constituting the entire gene. However, in the most recent release of the *Ensembl An. gambiae* genome (release 45) the annotation of this gene was altered, and it is now thought that these 858 bp represent about half of the coding sequence of LRIM1. We obtained a total of 138 alleles from six species, of which 108 were unique (Table 1, genbank accession nos EU073328–EU073397). As in the other two genes investigated here, polymorphisms were shared between species, but alleles were

### Table 1. Number of sampled alleles.

|        | CEC-A | GNB-B1 | LRIM1 |
|--------|-------|--------|-------|
| *gam*  | 49 (22) | 6 (6) | 28 (26) |
| *ara*  | 25 (15) | 8 (8) | 36 (35) |
| *qua*  | 21 (10) | 7 (7) | 25 (14) |
| *mer*  | 19 (11) | 6 (6) | 14 (13) |
| *mel*  | 57 (6) | 6 (6) | 22 (13) |
| *bwa*  | 15 (2) | 4 (4) | 13 (7) |

Table 1. Number of unique alleles is between brackets.

*gam* = *An. gambiae*, *ara* = *An. arabiensis*, *qua* = *An. quadriannulatus A*, *mer* = *An. merus*, *mel* = *An. melas*, *bwa* = *An. bwambae*.

doi:10.1371/journal.pone.0000793.t001

|        | Fixed | Polyomorph. |
|--------|-------|--------------|
|         | S     | NS   | S     | NS   | p-value |
| *gam-ara* | 2     | 0     | 6     | 6     | n.s.   |
| *gam-qua* | 2     | 0     | 5     | 5     | n.s.   |
| *gam-mel* | 0     | 0     | 4     | 6     | -      |
| *gam-mer* | 2     | 0     | 3     | 6     | n.s.   |
| *gam-bwa* | 2     | 1     | 4     | 5     | n.s.   |
| *ara-qua* | 0     | 0     | 6     | 1     | -      |
| *ara-mel* | 2     | 1     | 4     | 2     | n.s.   |
| *ara-mer* | 2     | 0     | 3     | 2     | n.s.   |
| *ara-bwa* | 0     | 1     | 4     | 1     | n.s.   |
| *qua-mel* | 2     | 1     | 4     | 1     | n.s.   |
| *qua-mer* | 2     | 0     | 3     | 1     | n.s.   |
| *qua-bwa* | 0     | 1     | 4     | 0     | n.s.   |
| *mel-mer* | 2     | 1     | 1     | 2     | n.s.   |
| *mel-bwa* | 2     | 2     | 2     | 1     | n.s.   |
| *mer-bwa* | 2     | 0     | 1     | 1     | n.s.   |

S = synonymous, NS = non-synonymous. Species names are abbreviated.

doi:10.1371/journal.pone.0000793.t002

| Fixed | Polyomorph. |
|-------|--------------|
|         | S     | NS   | S     | NS   | p-value |

Table 2. McDonald-Kreitman test on CEC1.
Table 3. MacDonald-Kreitman test on GNBP-B1.

|       | Fixed Polymorp. |
|-------|-----------------|
|       | S   | NS  | S   | NS  | p-value |
| gam-ara | 0   | 0   | 26  | 17  | -      |
| gam-qua | 17  | 1   | 32  | 12  | n.s.   |
| gam-mel | 13  | 0   | 20  | 12  | n.s.   |
| gam-bwa | 17  | 5   | 22  | 17  | n.s.   |
| ara-qua | 14  | 1   | 28  | 11  | n.s.   |
| ara-mel | 14  | 5   | 18  | 15  | n.s.   |
| ara-bwa | 0   | 0   | 17  | 12  | n.s.   |
| qua-mel | 13  | 1   | 8   | 5   | n.s.   |
| qua-bwa | 18  | 3   | 20  | 10  | n.s.   |
| quae | 0   | 0   | 17  | 12  | n.s.   |
| mel-bwa | 14  | 5   | 8   | 9   | n.s.   |
| mel-bwa | 16  | 3   | 10  | 7   | 0.010  |
| mel-bwa | 17  | 4   | 12  | 12  | n.s.   |

S = synonymous, NS = non-synonymous. Species names are abbreviated.

not. Dxy ranged from 1.03 to 3.06 (per 100 bp) between species. In contrast to CEC1 and GNBP-B1 however, McDonald-Kreitman tests of positive selection indicated a significant excess of fixed non-synonymous differences between An. arabiensis and An. quadriannulatus A, An. merus as well as An. bivittatus (Table 4). The fact that all three comparisons involve An. arabiensis, suggests this lineage underwent more non-synonymous substitutions than expected under a neutral model.

Surprisingly no fixed differences were present between An. gambiae and three other species; An. arabiensis, An. bivittatus and An.

Table 4. MacDonald-Kreitman test on LRIM1.

|       | Fixed Polymorp. |
|-------|-----------------|
|       | S   | NS  | S   | NS  | p-value |
| gam-ara | 0   | 0   | 26  | 17  | -      |
| gam-qua | 17  | 1   | 32  | 12  | n.s.   |
| gam-mel | 13  | 0   | 20  | 12  | n.s.   |
| gam-bwa | 17  | 5   | 22  | 17  | n.s.   |
| ara-qua | 14  | 1   | 28  | 11  | n.s.   |
| ara-mel | 14  | 5   | 18  | 15  | n.s.   |
| ara-bwa | 0   | 0   | 17  | 12  | n.s.   |
| qua-mel | 13  | 1   | 8   | 5   | n.s.   |
| qua-bwa | 18  | 3   | 20  | 10  | n.s.   |
| qua-bwa | 0   | 0   | 17  | 12  | n.s.   |
| mel-bwa | 14  | 5   | 8   | 9   | n.s.   |
| mel-bwa | 16  | 3   | 10  | 7   | 0.010  |
| mel-bwa | 17  | 4   | 12  | 12  | n.s.   |

S = synonymous, NS = non-synonymous. Species names are abbreviated.

Figure 1. Bayesian tree (unrooted) of LRIM1 from five species of the An. gambiae complex. Posterior probabilities ≥0.99 are indicated by *. Number of non-synonymous/synonymous substitutions are indicated above or on the left side of the branches. Estimated v values are placed below or on the right side of the branches and are underlined. The foreground branch for the maximum likelihood tests of positive selection is indicated by a double line. For a more detailed phylogeny, including all posterior probabilities above 50% and sample names, see Figure S2 (supporting materials).

Adaptive Evolution in LRIM1.
The clustering of Figure 2. Shared polymorphism of LRIM1 in An. gambiae (TGG), An. gambiae (AAA), and An. arabiensis. (AAA) alleles also have the (AAA) arrangement. Positions 416 through 718, a 302 bp stretch, cluster these An. gambiae (AAA) alleles with An. arabiensis (Figure 2).

Only positions below 328 and above 767 contain polymorphisms that group some of the An. gambiae (AAA) alleles with the rest of An. gambiae. Additionally, no fixed differences were found between An. gambiae and An. arabiensis anywhere in this gene.

LRIM1 is located inside the 2La inversion. Therefore we determined the karyotype of our An. gambiae samples with respect to the 2La arrangement. Out of 32 specimens from Cameroon, three were 2La/+ heterozygotes, all of which were also heterozygous for the (AAA)/(TGG) alleles. The karyotype of one (AAA)/(TGG) heterozygote was not clear, as it produced a second band of unexpected size. The remaining four (AAA)/(TGG) heterozygotes, as well as all (TGG) homozygotes, carried the 2La/+ karyotype. These findings confirm that the (AAA) allele is at very high frequency or even fixed in the 2La inversion in Cameroon, and is present at very low frequency in the 2La+ arrangement (≈ 7%).

To examine if LRIM1 in An. arabiensis showed signs of a recent selective sweep, a HKA test was performed by comparing the polymorphism/fixed differences ratio of LRIM1 to CEC1 and GNBP-B1. A selective sweep reduces the amount of standing genetic variation within a species, as indicated by a relatively low ratio. However, no significant differences were found between the genes, and in fact, this ratio was considerably higher for LRIM1 (20/20.2) as compared to CEC1 and GNBP-B1 (17/28.1).

DISCUSSION
CEC1 and GNBP-B1 did not show any signs of positive selection, and in particular, showed little or no differentiation between malaria vectors and the non-vector species, indicating that these genes are largely subject to purifying selection. In two of the species comparisons GNBP-B1 showed a significant excess of non-synonymous polymorphisms. Some cloning error is expected to be present in the GNBP-B1 data set. Since a majority of possible mutations are non-synonymous, random errors will bias the observed number of non-synonymous polymorphisms upward. However, the number of PCR errors in the data is not nearly high enough to explain the difference. Therefore, most likely purifying selection is responsible, with numerous slightly deleterious substitutions present at low frequency in populations, but which are prevented from going to fixation. In contrast, LRIM1, a gene

![Figure 2. Shared polymorphism of LRIM1 in An. gambiae (TGG), An. gambiae (AAA), and An. arabiensis.](image-url)

Only sites beyond position 324 that favor the clustering of An. gambiae (AAA) alleles with either An. gambiae (TGG) or An. arabiensis are included.

doi:10.1371/journal.pone.0000793.g002
with no known homologue in other organisms, shows clear signs of positive selection in *An. arabiensis*.

As pointed out by MacDonald and Kreitman, an excess of non-synonymous fixed differences between species may result from a much smaller population size in the past [21]. This would have allowed slightly deleterious mutations to go to fixation by drift, whereas in the current larger population most of these are removed by purifying selection. Evidence for population expansion has been reported for both *An. gambiae* and *An. arabiensis* [22,23]. It is unlikely however that this could explain the excess of non-synonymous substitutions observed in *LRIM1* in *An. arabiensis*. First of all, a demographic explanation should affect all genes. As noted before, *GNBP-B1* has a higher ratio of non-synonymous/synonymous polymorphisms than fixed differences in most or all populations, including *An. arabiensis*. This indicates the presence of a relatively large number of slightly deleterious alleles in this gene, few or none of which became fixed in ancestral populations. This is contrary to the demographic explanation. More importantly however, it is extremely unlikely that three out of four fixed amino acid changes would occur in two adjacent positions in a 283 amino acid protein, if the random process of genetic drift were responsible.

Polymorphisms are shared between species in all three genes we examined, and there is no doubt some of this is due to the retention of ancestral polymorphism. However, the pattern of polymorphism observed in *LRIM1* provides strong evidence that the presence of “*arabiensis*-like” alleles in *An. gambiae* is caused by introgression and not by the retention of ancestral polymorphism. It is unlikely that recombination would not have broken down a linkage group of at least 352 bp, if it were maintained in the population for long time. Furthermore, the introgression hypothesis is supported by the complete absence of fixed differences between these species anywhere in the *LRIM1* gene.

Based on the shared polymorphisms between *An. gambiae* (AAA) and *An. arabiensis* (positions 416, 624 and 701), introgression has occurred multiple times, after which, according to the shared polymorphism between the *An. gambiae* (AAA) and (TGG) alleles (positions 324 and 327), these alleles recombined between position 327 and 416. The introgression of *LRIM1* from *An. arabiensis* into *An. gambiae* is also consistent with previous studies that have shown that introgression between these two species has occurred in the past [24,25]. Additionally, it has been shown through sequence analyses [25], as well as crossing experiments [26], that different chromosomes vary in their capacity for horizontal transfer between these two species. The 2*Chromosome*, on which *LRIM1* is located, has been shown to transfer most readily between *An. gambiae* and *An. arabiensis*. [26].

Mosquitoes, like other organisms, encounter numerous pathogens during their life cycle, all of which could potentially exert selection pressure on the immune system. In fact, molecules that play a role in infection were best at infecting all of a set of iso-female *An.gambiae* lines [29]. Additionally, *LRIM1* is located inside the 2La inversion. While *An. arabiensis* is fixed for this 2La arrangement, *An. gambiae* is *2La/+* polymorphic. Since *LRIM1* alleles introgressed from *An. arabiensis* into *An. gambiae*, we may expect that these *An. gambiae* (AAA) alleles are mostly found in the 2La arrangement. Interestingly, the mosquitoes that failed to show an effect of *LRIM1* knockdown on *P. falciparum* infection [28] all carried the standard chromosome arrangement (2La+). Our molecular karyotyping of the 2La inversion in our *An. gambiae* specimens shows that the (AAA) allele is indeed found at very high frequency in 2La inversions, whereas it is present in very low frequency in 2La+ (~7%).

Pathogen-host co-evolution has mainly been considered in terms of an evolutionary arms race [30]. Under this model, the host continuously evolves to limit infection with the pathogen, which in turns evades to evade host defenses. This is expected to lead to repeated selective sweeps, which leave a signature in the selected genes in the form of a low level of standing genetic variation. A comparison of the polymorphism to divergence ratio in *LRIM1* vs *GNBP-B1/CEC1* did not show a relatively low level of genetic variation in *LRIM1* in *An. arabiensis*. In fact, the relative level of polymorphism was higher in this gene than in CEC1 and *GNBP-B1*. Therefore we have no indication that *LRIM1* in *An. arabiensis* is currently involved in an evolutionary arms race. This also implies that possible selective sweeps occurred long enough ago to allow mutation to regenerate polymorphism.

The data presented here indicate that the anti-malarial gene *LRIM1* has undergone adaptive evolution in a primary malaria vector. This could be because this gene has evolved in response to *P. falciparum* infection in this species. If so, *LRIM1* is expected to play a role in the resistance of *An. arabiensis* against *P. falciparum*. So far the immune system of this mosquito species has not yet been investigated, and our data suggest the possibility that a knockdown of *LRIM1* will enhance infections of *P. falciparum* in *An. arabiensis*. If *LRIM1* did indeed evolve in response to *P. falciparum* infection in *An. arabiensis*, this gene also deserves further study in *An. gambiae*, in particular with respect to potential variation in resistance between the two major alleles found in this species.

**MATERIALS AND METHODS**

**Mosquito sampling**

Adult females of *An. gambiae* were collected from the villages of Mbehê and N’abessa, Cameroon in Dec. 2005. *An. gambiae* from Mali were collected from Banambani in 2000. Adult *An. arabiensis* females from Cameroon were collected from Kousseri in Dec 2005. Adult *An. melas* were collected in Ipoko, Cameroon, Dec. 2005. Larvae of *An. gambiae*, *An. arabiensis* and *An. bwambae* from Kwambo district, Uganda (2004) were kindly provided by Ralph Harbach. DNA extractions of *An. merus* from Furvela, Mozambique (2001 and 2003) were kindly provided by David O’Brochta. *An. quadrimaculatus A* from Kruger National Park, South Africa,
were kindly provided by Anton Cornel. Sample sizes for each gene and species are represented in Table 1.

**DNA methods**

DNA was extracted using the DNeasy tissue kit (Qiagen). Species and molecular form diagnostics were performed following Fanello et al. [31] and Besansky et al. [32]. All *An. gambiae* specimens belonged to the S molecular form. Molecular identification of 2La karyotypes was performed following White et al. [33]. Primers to amplify CEC1, GNBP-B1 and LRIM1 were designed using Primer3 [34] based on the *An. gambiae* genome and anneal to the flanking or non-transcribed regions of the genes [3]. PCR of CEC-A and LRIM1 was performed using Amplitaq Gold polymerase (Perkin Elmer) using respectively the following primer pairs CECin1 (GGTACCGAGCGGTGTGTTTG/CECin12 ACAGTCGGTTCAAGGGTTG) and LRIM1m6 (AGGTAAACGGACAGCAGCCTA/LRIM1in9 GTCCGGTACTGCTCTTGAAG). The following program was used for PCR amplification of CEC1 and LRIM1; 2 min at 94°, 35 cycles of 30 sec at 94°, 30 sec at 52° and 45 sec at 72°, followed by 20 min at 72°. PCR products were excised from an agarose gel and purified using the Gel Purification Kit (Qiagen) and submitted for direct sequencing. A subset of the sequences from individuals heterogeneous for two or more positions were amplified again and cloned using the TOPO-TA cloning kit (Invitrogen). Individuals were selected for cloning such that all observed polymorphic sites were represented in the final data set. From each individual a single colony was sequenced. PCR of GNBP-B1 was performed using Platinum High Fidelity Taq (Invitrogen) with the primer pair GNBP-B1 (GGGTCTTAGGGCATGGATGA) /GNBP-B1 (GCCGTTTCACTGTGGTTGTTTG). The following program was used: 2 min at 94°, 35 cycles of 30 sec at 94°, 30 sec at 52° and 45 sec at 72°, followed by 20 min at 72°. Direct sequencing of the PCR product of GNBP-B1 was not possible in many cases because of the presence of indels. Therefore, PCR products of this gene were cloned and sequenced as outlined above. However, nine sequences were produced through direct sequencing, allowing for an estimation of the PCR/cloning error. Sequencing was performed on an ABI 3730 Genetic Analyzer using Big Dye v 3.1 (Applied Biosystems).

**PCR error**

Based on a comparison between direct sequencing and plasmid sequencing, the PCR/cloning error using Amplitaq Gold was estimated to be approximately 1.5 per 1000 bp. However, because all LRIM1 and CEC1 samples that were cloned were also sequenced directly, we were able to derive both alleles from each individual while removing PCR/cloning errors. The PCR error in the GNBP-B1 sequences amplified using the proof-reading polymerase was estimated to be 0.625 per 1000 bp. Therefore, each 1100 bp GNBP-B1 allele for which no direct sequence was available is expected to have an average of 0.74 errors.

**Data analysis**

All sequences were aligned using MEGA3.1 [35] and alignments were improved manually. Introns were included in the phylogeny reconstructions. For all other analyses the coding region was used. Dxy values were calculated using DnaSP 4.0 [36]. This software was also used to perform McDonald-Kreitman tests [21], using Fisher’s exact test. The McDonald-Kreitman test compares the dN/dS ratio between species to within species and is based on the idea that substitutions under positive selection will go to fixation rapidly, and are therefore rarely observed as polymorphisms. However, they are present as fixed differences between species and an excess of replacement fixed differences is therefore an indication of positive selection.

Since few or no fixed differences were observed in CEC1 and GNBP-B1, subsequent analyses were limited to LRIM1. Aimed at reducing the computational effort, a reduced LRIM1 data set, containing 70 sequences, was used for phylogenetic analyses and maximum likelihood tests of positive selection. This data set was compiled in such a way that at every observed polymorphism and fixed difference, *i.e.* the relevant information for tests for maximum likelihood tests of positive selection, was retained. This reduced data set was used to construct 50% majority-rule consensus trees with MrBayes 3.1.2 [37]. Modeltest 3.7 [38] was used to determine the most appropriate nucleotide substitution model for our data set.

Several LRIM1 alleles from *An. gambiae* clustered within *An. arabiensis* (Figure S1 supp. mat.). Therefore, phylogeny reconstruction was also performed excluding *An. gambiae* sequences. This inferred tree was used for maximum likelihood tests of positive selection along the branch leading to *An. arabiensis* in PAML 3.15 and to estimate ω (i.e. dN/dS) along the major branches of the tree. Under the neutral model the relative number of synonymous and non-synonymous substitutions is expected to be 1. Under positive selection, amino acid substitutions are favored and ω>1, whereas under purifying selection amino acid substitutions are prevented and ω<1. The *An. arabiensis* lineage was designated as the foreground branch, i.e. the branch of interest, and model $v_2\omega$ was compared to model 0 to test if ω along the foreground branch was significantly larger compared to the ω along the background branches, i.e. all other branches. Model $2\omega v_1$ with the ω value fixed at 1 along foreground branch was compared to model $2\omega v_2\omega$ to test if ω along the foreground branch was significantly larger than 1. Model 1 was used to estimate ω along the central branches of the tree (Figure 1) and to infer the number of substitutions along these branches of the phylogeny. As an additional test for positive selection along the foreground branch, we used the more powerful branch-site test 2 by comparing model A and Model A1 [19]. Bayes Empirical Bayes (BEB) analysis was used to identify positively selected codons in the foreground branch [20].

To test for a reduction in the polymorphism of LRIM1 in *An. arabiensis*, an HKA test was performed in DnaSp 4.0, using eight *An. arabiensis* alleles for CEC1 and LRIM1, as well as all eight GNBP-B1 alleles from this species. The CEC1 and LRIM1 alleles were from the same individuals as the GNBP-B1 sequences if possible, otherwise were randomly chosen from the same population. Seven *An. quadriannulatus* A alleles from each gene were used to calculate inter-specific divergence.

**Supporting Information**

**Figure S1** Bayesian tree (unrooted) of LRIM1 from six species of the *An. gambiae* complex. Posterior probabilities are indicated along branches. *An. gambiae* samples and *An. arabiensis* samples from Uganda, Madagascar and Mali are indicated by UG, MAD, and MAL respectively, remaining samples of these two species are from Cameroon. Found at: doi:10.1371/journal.pone.0000793.s001 (1.19 MB TIF)

**Figure S2** Bayesian tree (unrooted) of LRIM1 in five species of the *An. gambiae* complex. Posterior probabilities are indicated along branches. *An. arabiensis* samples from Uganda, Madagascar and Mali are indicated by UG, MAD, and MAL respectively, with all remaining *An. arabiensis* samples originating from Cameroon. Found at: doi:10.1371/journal.pone.0000793.s002 (0.72 MB TIF)
ACKNOWLEDGMENTS

We are very grateful to Anton Cornel, Ralph Harbach and David O’Brochta for providing An. quadrimaculatus A, An. bicolor and An. merus specimens, respectively. Additionally, we are grateful to the anonymous reviewers who provided many useful comments for improving earlier versions of the manuscript.

REFERENCES

1. Breman JC, Egan A, Teusch GT (2001) The intolerable burden of malaria: A new look at the numbers. Am J Trop Med Hyg 64: iv–vii.
2. Holt R, Subramanian GM, Halpern A, Sutton G, Charlab R, et al. (2002) The genome sequence of the malaria mosquito Anopheles gambiae. Science 298: 129–140.
3. Riehle MM, Markkino K, Niare O, Xu J, Li J, et al. (2006) Natural malaria infection in Anopheles gambiae is regulated by a single genomic control region. Science 312: 577–579.
4. Kim W, Koo H, Richman AM, Seeley D, Vizoli J, et al. (2004) Ectopic expression of a cecropin transgene in the human malaria vector mosquito Anopheles gambiae (Diptera: Culicidae): Effects on susceptibility to Plasmodium. J Med Entomol 41: 447–455.
5. Tahar R, Boudin C, Thiery I, Bourgouin C (2002) Immune response of Anopheles gambiae to the early sporogonic stages of the human malaria parasite Plasmodium falciparum. The EMBO journal 21: 6673–6680.
6. Ost MA, Christophides GK, Kafatos FC (2004) Effects of mosquito genes on Plasmodium development. Science 303: 2030–2032.
7. Dong Y, Aguilar R, Xi Z, Warr E, Mongin E, et al. (2006) Anopheles gambiae immune responses to human and rodent Plasmodium parasite species. PLoS Pathol 2: 513–525.
8. Vlachou D, Schlegelmilch T, Christophides GK, Kafatos FC (2005) Functional genomic analysis of midgut epithelial responses in Anopheles during Plasmodium invasion. Current Biol 15: 1185–1193.
9. Little TJ, Cobbe N (2005) The evolution of immune-related genes from disease carrying mosquitoes: diversity in a peptidoglycan- and a thioredoxin-like-encoding gene. Insect Mol Biol 14: 599–606.
10. Simard F, Licht M, Besansky NJ, Lehmann T (2006) Polymorphism at the major histocompatibility complex: testing different selection hypotheses. Infect Genet Evol 7: 285–292.
11. Yang Z (2006) Computational Molecular Evolution. New York: Oxford University Press. pp 357.
12. Schlenke TA, Begun DJ (2003) Natural selection drives Drosophila immune system evolution. Genetics 165: 1471–1480.
13. Collins FH, Zavala F, Graves PM, Cochaune AH, Goadsby RW, et al. (1984) First field trial of an immunoradiometric assay for the detection of malaria sporozoites in mosquitoes. Am J Trop Med Hyg 33: 538–543.
14. Ferguson HM, Read AF (2002) Why is the effect of malaria parasites on mosquito survival still unresolved? Trends Parasitol 18: 256–261.
15. Hay SI, Rogers DJ, Toomer JF, Snow RW (2000) Annual Plasmodium falciparum entomological inoculation rates (EIR) across Africa: literature survey, internet access and review. Trans R Soc Trop Med Hyg 94: 113–127.
16. Vizoli J, Bulet P, Charlet M, Lowenberger C, Blass C, et al. (2000) Cloning and expression of a cecropin transgene in the human malaria vector mosquito, Anopheles gambiae. EMBO Reports 7: 1285–1289.
17. Dimopoulos G, Richman AM, Bourgouin C, Dimopoulos G (2006) Anopheles gambiae immune response to Sephadex beads: involvement of anti-Plasmodium factors in regulating melanization. Insect Biochem Mol Biol 36: 769–778.
18. Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472–2479.
19. Yang Z, Wong WS, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22: 1107–1118.
20. MacDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652–654.
21. Ronald JY, Licht MC, Lehmann T (2001) Evidence for recent population expansion in the evolutionary history of the malaria vectors Anopheles arabiensis and Anopheles gambiae. Mol Biol Evol 18: 1353–1364.
22. Osterfjeld DV, Conn JR (2000) Population structure of the malaria mosquito Anopheles arabiensis across Nigeria suggests range expansion. Mol Ecol 10: 2577–2591.
23. Besansky NJ, Kryszwinski J, Lehman T, Simard F, Kern M, et al. (2003) Sibling species boundaries between Anopheles gambiae and Anopheles arabiensis: Evidence from multilocus DNA sequence variation. Proc Natl Acad Sci U S A 19: 10818–10823.
24. Besansky NJ, Powell JR, Caccone A, Hamm DM, Scott JA, et al. (1994) Molecular phylogeny of the Anopheles gambiae complex suggests genetic introgression between principle malaria vectors. Proc Natl Acad Sci U S A 91: 6815–6818.
25. Slomann MA, della Torre A, Calzetta M, Powell JR (2003) Differential introgression of chromosomal regions between Anopheles gambiae and An. arabiensis. Am J Trop Med Hyg 73: 326–335.
26. Enkhbayar P, Kamiya M, Osaki M, Matsumoto T, Matsushima N (2004) Structural principles of Leucine-Rich Repeat (LRR) Proteins. Proteins 54: 409–446.
27. Cohuet A, Osta M, Morlais I, Avondo-Ambehe PH, Michel K, et al. (2006) Anopheles and Plasmodium: from laboratory models to natural systems in the field. EMBO Reports 7: 1285–1289.
28. Lambrechts L, Hallett J, Durand P, Gossagna LG, Koella JC (2005) Host genotype by parasite genotype interactions underlying the resistance of anopheine mosquitoes to Plasmodium falciparum. Malar J 4: 3.
29. Dawkins R, Krebs JR (1979) Arms races between and within species. Proc R Soc Lond B Biol Sci 205: 489–511.
30. Fanzillo C, Santolamazza F, della Torre A (2002) Simultaneous identification of species and molecular forms of the Anopheles gambiae complex by PCR-RFLP. Med Vet Entomol 16: 461–464.
31. Besansky NJ, Collins FH, Towom H (2006) A species-specific PCR for the identification of the malaria vector Anopheles gambiae. Ann Trop Med Parasitol 100: 277–280.
32. White BJ, Santolamazza F, Kamau L, Pombi M, Grushko O, et al. (2007) Molecular karyotyping of the 2La inversion in Anopheles gambiae. Am J Trop Med Hyg 76: 334–339.
33. Rouen S, Skalsky HI (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S, eds. Bioinformatics Methods and Protocols: Methods in Molecular Biology. Totowa: Humana Press. pp 307–315.
34. Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings in Bioinformatics 3: 130–163.
35. Roux J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP. DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496–2497.
36. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1573.
37. Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. Bioinformatics 14: 817–818.