Identification and analysis of Single Nucleotide Polymorphisms (SNPs) in the mosquito Anopheles funestus, malaria vector

Charles S Wondji*, Janet Hemingway and Hilary Ranson

Address: Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, UK
Email: Charles S Wondji* - c.s.wondji@liverpool.ac.uk; Janet Hemingway - hemingway@liverpool.ac.uk; Hilary Ranson - hranson@liverpool.ac.uk
* Corresponding author

Abstract

Background: Single nucleotide polymorphisms (SNPs) are the most common source of genetic variation in eukaryotic species and have become an important marker for genetic studies. The mosquito Anopheles funestus is one of the major malaria vectors in Africa and yet, prior to this study, no SNPs have been described for this species. Here we report a genome-wide set of SNP markers for use in genetic studies on this important human disease vector.

Results: DNA fragments from 50 genes were amplified and sequenced from 21 specimens of An. funestus. A third of specimens were field collected in Malawi, a third from a colony of Mozambican origin and a third form a colony of Angolan origin. A total of 494 SNPs including 303 within the coding regions of genes and 5 indels were identified. The physical positions of these SNPs in the genome are known. There were on average 7 SNPs per kilobase similar to that observed in An. gambiae and Drosophila melanogaster. Transitions outnumbered transversions, at a ratio of 2:1. The increased frequency of transition substitutions in coding regions is likely due to the structure of the genetic code and selective constraints. Synonymous sites within coding regions showed a higher polymorphism rate than non-coding introns or 3’ and 5’flanking DNA with most of the substitutions in coding regions being observed at the 3rd codon position. A positive correlation in the level of polymorphism was observed between coding and non-coding regions within a gene. By genotyping a subset of 30 SNPs, we confirmed the validity of the SNPs identified during this study.

Conclusion: This set of SNP markers represents a useful tool for genetic studies in An. funestus, and will be useful in identifying candidate genes that affect diverse ranges of phenotypes that impact on vector control, such as resistance insecticide, mosquito behavior and vector competence.

Background

Anopheles funestus and Anopheles gambiae are the major malaria vectors in Africa. Due to the difficulty of laboratory colonization, An. funestus has not received the same attention as An. gambiae and as a consequence there are few molecular markers for this species. However, the recent successful colonization of two strains of An. funestus [1] and the identification of a number of microsatellite markers [2,3] have facilitated more detailed studies of this species. Microsatellite markers particularly have been used to study population structure and gene flow between An. funestus populations [4-6] and a subset of these microsatellite markers were used to build the first linkage map of this species [7]. However, microsatellite markers...
are not evenly distributed across the genome, and their low number so far is an obstacle to the development of high resolution linkage maps needed for QTL mapping or association studies in *An. funestus*. Therefore, this study was initiated to increase the availability of characterized and mapped markers for *An. funestus*. Physically mapped ESTs were used to identify SNPs. Such ESTs have been used to study the genetic variability in a number of species such as *Aedes aegypti*, *Drosophila melanogaster* or *Homo sapiens* [8-10] and should be a source of DNA polymorphisms for *An. funestus* as well.

Single nucleotide polymorphisms (SNPs) are by far the most common type of molecular variation in all organisms. They are extremely abundant with an occurrence of about one SNP per kb in human [11] and about one SNP every 125 bp in *An. gambiae* [10]. Significant progress has been made in the development of tools for detection and genotyping of SNPs and they are now becoming the markers of choice for association studies, high-resolution linkage mapping and population genomics studies [12]. SNPs located in non-coding regions of the genome and synonymous SNPs (sSNPs) in coding regions, which have no impact on the phenotype, may provide useful markers for population genetics studies. Non-synonymous SNPs (nsSNPs) which alter the structure (change of amino acid sequence) and potentially the function of encoded proteins are useful markers for association studies to detect genetic variations linked with phenotypic traits.

Patterns of genetic diversity in *An. funestus* have not been studied to the same extent as in *An. gambiae* or *Drosophila* species. Nucleotide diversity in these species has been used to compare patterns of nucleotide variation, such as the relative occurrence of transitions/transversions in different regions of the genome [8,13]. These surveys have established codon usage and usage bias patterns in many species, with bias hypothesized to occur as a result of selection for efficient translation [14,15].

The sequencing of the 278 million base pairs (Mbp) constituting the *An. gambiae* genome has revealed more than 400,000 SNPs indicating a high level of polymorphism in mosquito species [16]. We hypothesize that by sequencing DNA fragments of different genes of *An. funestus*, a similar level of polymorphism should be encountered and will allow the identification of a significant set of SNPs.

Here, we describe the detection and characterization of a set of genome-wide SNP markers from 50 nuclear genes using two laboratory strains and field samples of *An. funestus*. We also examined patterns of polymorphism and nucleotide diversity in coding and non-coding regions of the genome and define the pattern of codon usage in *An. funestus*. The utility of the SNPs was assessed by genotyping a subset of these SNPs during a linkage mapping study.

**Results and discussion**

**Gene amplification**

In total, 70 primer pairs were tested by PCR, 55 of which gave reliable amplification with PCR products ranging from 194 to 1342 bp. Sequence data from a total of 21 specimens of *An. funestus* was obtained for 50 of these genes (see Table 1) from laboratory and field samples. Overall, we sequenced a total of 20,547 bp consisting of 14,671 bp of coding region and 5,876 bp of non-coding region. We identified 494 SNPs consisting of 303 coding SNPs (cSNPs) and 191 non-coding SNPs. Each gene contained at least one polymorphism with *BU73* having one and *BU88* having 29. The distribution of SNPs among the 50 genes is presented in Table 1. All information concerning the location and the nature of each individual SNP have been submitted to dbSNP, the SNP database of GenBank. These SNPs with their respective reference SNP number (rs) are publicly available in dbSNP Build N° 127. The NCBI ss (submitted SNP) numbers of these SNPs are ss65917063 to ss65917416.

**Type of polymorphism**

For all sequenced DNA fragments, transition substitutions were more predominant than transversions (62 % vs 38%). Transitions C→T and A→G are over-represented with 35.4 and 27.2 % of the total substitutions respectively while the four transversion classes occurred at similar levels (Figure 1). The higher frequency of C→T and A→G SNPs is probably partly related to 5-methylcytosine deamination reactions that occur frequently, particularly at CpG dinucleotides [17]. The preponderance of transitions is more obvious for coding regions where out of the 303 SNPs identified, 201 were transitions (66.3%) and 102 were transversions (33.7%). The ratio of transitions/transversions observed here is close to the 2 : 1 ratio observed for *Drosophila* and humans [13,18]. For polymorphism in non-coding regions, transitions accounted for 55% (106) and transversion for 45% (86). The frequency of transitions between coding and non-coding regions were significantly different (66.3% vs 55% respectively; $\chi^2 = 5.86$, $P < 0.01$). This confirms that SNPs occur more frequently as transitions in coding regions than in non-coding regions. There is also a higher frequency of SNPs occurring at the third codon position (63.7%) than at the 1st or 2nd position (Table 1). Similar results have been observed for *Aedes aegypti* [10] and in three species of *Drosophila* [13]. The degeneracy of the genetic code and the selective pressure for gene conservation have been suggested as the main reasons for the preponderance of transitions over transversions [13]. Synonymous or silent substitutions are more often transitions than transver-
Table 1: Nucleotide polymorphism in An. funestus genes

| Gene   | nHap | L (bp) | Polymorphic sites | Nucleotide diversity | Polymorphic sites | Nucleotide diversity |
|--------|------|--------|-------------------|----------------------|-------------------|----------------------|
|        |      |        | Transition | Transversion |                  |                      |
|        |      |        | 1st 2nd 3rd Σ | 1st 2nd 3rd Σ | Syn | Rep | Σ | π | πo | Ks | Ka | L(bp) | T1 | T2 | Σ | π |
| 4G17   | 3    | 190    | 0 1 1 2 0 0 0 | 0 0 0 1 1 2 | 0.0052 | 0.0036 | 0.0200 | 0.0072 | 0 0 0 | 0.0000 |
| 9K1    | 23   | 199    | 3 1 3 7 1 0 1 2 | 5 4 9 | 0.0193 | 0.0107 | 0.1010 | 0.0270 | 0 0 0 | 0.0000 |
| 4J10   | 14   | 141    | 1 0 3 4 0 0 1 1 | 5 0 5 | 0.0116 | 0.0000 | 0.1300 | 0.0000 | 73 2 2 | 0.0157 |
| 6P3    | 18   | 0 0 0 | 0 0 0 0 0 0 | 0 0 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 179 6 2 | 0.0188 |
| 6P4    | 17   | 201    | 0 0 2 2 0 0 0 | 0 0 0 | 2 0 2 | 0.0041 | 0.0000 | 0.0450 | 0.0000 | 189 8 3 | 0.0233 |
| 6P5    | 5    | 189    | 0 0 1 1 0 0 0 | 1 0 1 | 0.0016 | 0.0000 | 0.0200 | 0.0000 | 112 2 0 | 0.0063 |
| 6Z1    | 14   | 444    | 1 0 5 6 0 0 5 5 | 8 3 | 0.0079 | 0.0022 | 0.0750 | 0.0088 | 29 1 0 | 0.0159 |
| 6Z3    | 14   | 453    | 1 0 13 14 1 0 | 5 6 | 18 2 20 | 0.0169 | 0.0011 | 0.1860 | 0.0057 | 0 0 0 | 0.0000 |
| 9J12   | 9    | 156    | 2 1 0 3 1 0 1 2 | 0 5 5 | 0.0117 | 0.0156 | 0.0000 | 0.0510 | 0 0 0 | 0.0000 |
| 9J14   | 14   | 174    | 0 3 3 2 0 0 | 0 2 3 2 5 | 0.0140 | 0.0077 | 0.0760 | 0.0150 | 0 0 0 | 0.0000 |
| BU01   | 2    | 162    | 0 0 1 1 0 0 | 0 0 1 0 1 | 0.0011 | 0.0000 | 0.0230 | 0.0000 | 26 0 0 | 0.0000 |
| BU08   | 12   | 394    | 0 0 2 2 2 | 1 2 3 5 4 3 | 7 | 0.0042 | 0.0026 | 0.0470 | 0.0098 | 16 0 0 | 0.0000 |
| Ache   | 18   | 249    | 0 1 7 8 | 0 1 2 3 8 3 | 11 | 0.0149 | 0.0034 | 0.1460 | 0.0160 | 85 2 2 | 0.0180 |
| BU10   | 16   | 294    | 1 0 3 4 0 0 | 2 2 5 1 6 | 0.0084 | 0.0024 | 0.0690 | 0.0045 | 245 3 1 | 0.0056 |
| BU11   | 4    | 177    | 0 1 0 1 0 | 0 0 0 0 1 | 0.0029 | 0.0039 | 0.0000 | 0.0074 | 53 0 1 | 0.0099 |
| BU12   | 7    | 504    | 1 1 3 0 1 1 | 2 3 2 5 | 0.0032 | 0.0011 | 0.0240 | 0.0053 | 38 0 0 | 0.0000 |
| BU13   | 19   | 363    | 2 0 1 3 0 0 | 0 0 1 2 3 | 0.0026 | 0.0029 | 0.0110 | 0.0073 | 148 2 5 | 0.0133 |
| BU19   | 22   | 526    | 3 1 4 | 8 0 | 1 2 6 4 10 | 0.0072 | 0.0035 | 0.0490 | 0.0100 | 0 0 0 | 0.0000 |
| BU21   | 16   | 417    | 0 1 2 3 1 0 | 2 3 4 2 6 | 0.0073 | 0.0018 | 0.0510 | 0.0077 | 81 4 1 | 0.0164 |
| BU25   | 2    | 234    | 0 0 0 | 0 0 0 | 0 0 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 111 2 0 | 0.0036 |
| BU29   | 11   | 258    | 0 0 2 2 0 | 0 2 2 | 3 1 | 4 | 0.0038 | 0.0008 | 0.0540 | 0.0049 | 152 0 2 | 0.0029 |
| BU34   | 11   | 231    | 0 0 2 2 0 2 | 2 4 4 2 6 | 0.0101 | 0.0051 | 0.0740 | 0.0113 | 28 0 0 | 0.0000 |
| BU35   | 12   | 264    | 0 0 7 7 0 | 0 0 1 1 | 8 0 8 | 0.0071 | 0.0000 | 0.1280 | 0.0000 | 108 1 1 | 0.0039 |
| BU40   | 14   | 255    | 2 1 1 | 4 1 0 | 0 1 1 4 5 | 0.0074 | 0.0076 | 0.0170 | 0.0200 | 415 3 2 | 0.0016 |
| BU56   | 4    | 156    | 0 1 2 3 | 1 0 | 2 3 3 3 3 | 6 | 0.0214 | 0.0132 | 0.0710 | 0.0263 | 126 5 2 | 0.0331 |
| BU58   | 7    | 261    | 1 0 2 3 | 0 1 1 2 3 2 5 | 0.0053 | 0.0019 | 0.0446 | 0.0103 | 20 0 0 | 0.0000 |
Table 1: Nucleotide polymorphism in *An. funestus* genes (Continued)

| BU62 | 22 | 213 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.0013 | 0.0000 | 0.0207 | 0.0000 | 260 | 7 | 5 | 12 | 0.0143 |
| BU66 | 25 | 282 | 9 | 2 | 0 | 11 | 1 | 2 | 0 | 3 | 3 | 11 | 14 | 0.0174 | 0.0194 | 0.0503 | 0.0494 | 179 | 4 | 4 | 8 | 0.0099 |
| BU70 | 5 | 228 | 0 | 0 | 2 | 2 | 0 | 0 | 2 | 2 | 2 | 4 | 0.0066 | 0.0015 | 0.0506 | 0.0059 | 62 | 0 | 1 | 1 | 0.0069 |
| BU71 | 9 | 573 | 0 | 0 | 6 | 6 | 1 | 1 | 2 | 4 | 8 | 2 | 10 | 0.0044 | 0.0005 | 0.0615 | 0.0045 | 98 | 1 | 0 | 1 | 0.0051 |
| BU72 | 20 | 354 | 2 | 1 | 2 | 5 | 2 | 1 | 1 | 4 | 3 | 6 | 9 | 0.0082 | 0.0062 | 0.0368 | 0.0220 | 195 | 1 | 1 | 2 | 0.0052 |
| BU73 | 5 | 277 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 3 | 0 | 3 | 3 | 0.0039 | 0.0036 | 0.0000 | 0.0135 | 0 | 0 | 0 | 0 | 0.0000 |
| BU76 | 15 | 330 | 0 | 0 | 7 | 7 | 1 | 1 | 1 | 3 | 9 | 1 | 10 | 0.0087 | 0.0008 | 0.1104 | 0.0041 | 12 | 3 | 3 | 6 | 0.0095 |
| BU77 | 18 | 261 | 3 | 0 | 1 | 4 | 1 | 2 | 0 | 3 | 6 | 1 | 7 | 0.0125 | 0.0136 | 0.0166 | 0.0298 | 198 | 9 | 7 | 16 | 0.0345 |
| BU82 | 13 | 228 | 0 | 0 | 2 | 2 | 0 | 2 | 0 | 2 | 2 | 2 | 4 | 0.0059 | 0.0025 | 0.0402 | 0.0112 | 88 | 2 | 2 | 4 | 0.0156 |
| BU85 | 14 | 299 | 2 | 1 | 8 | 11 | 0 | 0 | 1 | 1 | 1 | 1 | 12 | 0.0108 | 0.0363 | 0.1155 | 0.0033 | 121 | 2 | 0 | 2 | 0.0087 |
| BU88 | 16 | 441 | 2 | 0 | 12 | 14 | 0 | 0 | 2 | 2 | 16 | 0 | 16 | 0.0135 | 0.0000 | 0.1601 | 0.0000 | 212 | 9 | 4 | 13 | 0.0179 |
| BU90 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 261 | 1 | 2 | 3 | 0.0021 |
| BU92 | 11 | 345 | 3 | 1 | 3 | 7 | 1 | 1 | 1 | 3 | 5 | 5 | 10 | 0.0075 | 0.0031 | 0.0650 | 0.0186 | 110 | 5 | 3 | 8 | 0.0226 |
| BU93 | 8 | 510 | 0 | 0 | 5 | 5 | 0 | 0 | 3 | 3 | 7 | 1 | 8 | 0.0035 | 0.0004 | 0.0626 | 0.0025 | 137 | 0 | 2 | 2 | 0.0023 |
| BU98 | 4 | 369 | 0 | 0 | 5 | 5 | 0 | 0 | 3 | 2 | 7 | 0 | 7 | 0.0101 | 0.0000 | 0.0843 | 0.0000 | 158 | 2 | 6 | 8 | 0.0257 |
| BU883 | 21 | 354 | 1 | 0 | 5 | 6 | 2 | 1 | 2 | 5 | 7 | 4 | 11 | 0.0120 | 0.0052 | 0.0913 | 0.0145 | 188 | 8 | 3 | 11 | 0.0189 |
| BU897 | 19 | 303 | 0 | 0 | 2 | 2 | 3 | 1 | 2 | 6 | 3 | 5 | 8 | 0.0080 | 0.0078 | 0.0458 | 0.0210 | 167 | 4 | 3 | 7 | 0.0180 |
| BU901 | 12 | 231 | 0 | 0 | 3 | 3 | 0 | 0 | 0 | 3 | 0 | 3 | 0 | 0.0035 | 0.0000 | 0.0552 | 0.0000 | 229 | 2 | 2 | 4 | 0.0065 |
| BU973 | 5 | 471 | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 0.0015 | 0.0019 | 0.0000 | 0.0054 | 67 | 1 | 0 | 1 | 0.0028 |
| BU974 | 3 | 459 | 0 | 0 | 2 | 2 | 0 | 0 | 1 | 1 | 3 | 0 | 3 | 0.0036 | 0.0000 | 0.0265 | 0.0000 | 52 | 1 | 0 | 1 | 0.0096 |
| BU982 | 4 | 240 | 1 | 1 | 2 | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 0.0030 | 0.0016 | 0.0190 | 0.0053 | 51 | 1 | 0 | 1 | 0.0107 |
| BU996 | 18 | 543 | 4 | 0 | 2 | 6 | 1 | 1 | 1 | 3 | 6 | 3 | 9 | 0.0073 | 0.0019 | 0.0518 | 0.0070 | 87 | 6 | 6 | 12 | 0.0312 |
| Kdr | 13 | 201 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 2 | 0 | 2 | 2 | 0.0020 | 0.0026 | 0.0000 | 0.0127 | 501 | 2 | 5 | 7 | 0.0039 |

**Average**

| 14671 | 45 | 17 | 139 | 201 | 25 | 23 | 54 | 102 | 206 | 97 | 303 | 5876 | 106 | 85 | 191 |
| 0.0072 | 0.0041 | 0.0537 | 0.0097 | 0.0123 |

nHap, number of haplotypes; L, length of the nucleotide sequence; \( \sum \), total; Syn, synonymous substitutions; Rep, replacement substitutions; \( \pi \), average number of nucleotide substitution per site; \( \pi_n \), average number of non-synonymous nucleotide substitution per site; \( K_s \), average number of nucleotide substitution per synonymous site; \( K_a \), per non-synonymous site; \( T_s \), transitions; \( T_v \), transversions.
sions and there is a stronger selection against replacement substitutions than against synonymous, leading to an increase of the relative frequency of transitions [13]. For fourfold degenerate codons, selection should be neutral, since no amino acid change is induced by a nucleotide substitution at the third position, and each of the 4 codons will produce the same amino acid. We tested this hypothesis by comparing the proportions of transversions at fourfold degenerate codon positions and at non-coding positions for all the 50 genes (Table 2). The result shows that there is no significant difference between the frequency of transversions at fourfold degenerate codon positions (36.8%) and at non-coding regions (44.5%) ($\chi^2 = 2.55$, $P = 0.11$), while this difference is significant between coding and non-coding regions ($\chi^2 = 5.33; P = 0.021$). The fact that fourfold degenerate sites have a similar ratio of transitions/transversions to non-coding regions is consistent with an hypothesis that the structure of the genetic code and selection against replacement polymorphisms accounts for the preponderance of transition substitutions in coding regions.

Five insertion/deletion polymorphisms (indels) were observed in four genes ranging from 1 to 4 bp in coding, intronic and 5'UTR regions (Table 3). Two indels of 2 and 4 bp were observed in the BU10 intron. The frequency of indels (8% for 4/50) is lower than that reported in Ae aegypti of 24% [10] or 25% in An. gambiae [19]. Only one indel, in the BU93 gene, was located in a coding region. This indel was a triplet that did not cause a frame shift. The four indels identified can serve as molecular makers for mapping studies.

Approximately 2/3 (206) of the 303 cSNPs were synonymous substitutions (no modification in amino acid) while around 1/3 (97) were non-synonymous or replacement SNPs leading to a change of amino acid. As approximately two-thirds of random coding substitutions change an amino acid, the fact that only 1/3 of cSNPs are non-synonymous implies strong selection against changes that alter amino acid. This ratio of synonymous and replacement cSNPs is similar to that observed in An. gambiae [19] and Ae. aegypti [10].

### Genetic diversity

We estimated the nucleotide diversity for each of the 50 genes in coding and non-coding regions (Table 1). The average nucleotide diversity per gene in coding regions was $7.2 \times 10^{-3}$ or around 1 SNP every 138 bp similar to that observed in An. gambiae (1 SNP every 125 bp) [19] but much higher than the frequency of 1 SNP/kb observed in humans [20]. SNPs were observed in non-coding regions at a frequency of 1 SNP per 100 bp, corresponding to $\pi = 10 \times 10^{-3}$. Figure 2 shows that there is a positive correlation in the level of polymorphism between coding and non coding regions of An. funestus genome within a gene ($r = 0.48$, $P < 0.01$). This positive correlation may be the consequence of many factors notably the correlated genealogies existing between coding regions and their surrounding non-coding regions. This correlation may also be strengthened by the presence of indirect selection (hitchhiking or background selection) and probably by variable recombination rate, as it is the case in Drosophila [21]. Mutational effect of recombination or biased gene conversion can also operate, but this needs to be confirmed as even in Drosophila, the effect of biased gene conversion is only suspected but unwarranted [22,23]. The average nucleotide diversity in non-coding DNA (0.010) was lower than in synonymous sites of the coding regions (0.0207), $P < 0.01$. This pattern was also observed in An gambiae, Ae. aegypti and Drosophila species [10,13,19]. This is an indication that non-coding regions are under greater purifying selection than synonymous sites within coding regions. This is not surprising, given that non-coding regions may be involved in gene regulation. The non-coding 5'-flanking sequence of a gene may contain regulatory elements such as the promoter that control the expression of that gene, and single-base mutations can affect essential structures for splicing and processing [24].

Nucleotide diversity varies greatly from one gene to another (Table 1) and this is likely related to individual gene function and potentially to differences in selective constraints. However, non-synonymous diversities need to be compared in order to definitely estimate the influence of differences in selective constraints. Among the most polymorphic genes sequenced were cytochrome P450 genes, lysozyme, translation initiation factor and ubiquitin conjugating genes. The non-synonymous nucleotide diversity of these genes varied from 14 to $36.3 \times 10^{-3}$. Most of these genes are involved in specific mechanisms that evolve very rapidly, such as detoxification of xenobi-
otics for cytochrome P450s or defense mechanisms against bacteria like lysozyme. For example, P450s present a high level of redundancy with less genetic constraints and therefore more polymorphism. In contrast some genes showed very low level of variation particularly those involved in transcriptional or translational regulation (BU973 and BU25, BU93) or in signaling processes (BU01, BU08, BU13). Examples of selective constraints have been observed as well in Drosophila spp where substitution rate between conservative genes and fast evolving genes differ by around 10-fold [25]. Nucleotide diversity was not statistically different between laboratory strains and field collected mosquitoes (7.4 × 10^{-3} and 6.9 × 10^{-3}; P = 0.21 by Student's t-test), despite an apparent low level of heterozygosity (fewer heterozygote SNPs) observed in the two laboratory strains compared to the field sample. This result could be due to the fact that FUMOZ and FANG (the two laboratory strains used in this study), were only recently colonized in laboratory and therefore still largely retain the polymorphism of natural populations of An. funestus.

The ratio of synonymous to non-synonymous changes (Ka/Ks) gives an indication of the magnitude of the purifying selection against deleterious mutations in a species. The rate of non-synonymous nucleotide substitution per non-synonymous site (Ka) is generally expected to be much lower than the rate of synonymous substitution per synonymous site (Ks), because random amino acid changes are usually deleterious, whereas synonymous changes are likely to be neutral or nearly so [26]. Thus, the expectation is Ka << Ks, except when positive selection is involved favouring particular amino acid replacements, in which case Ka will increase. For An. funestus the Ka/Ks ratio was equal to 0.181 and is similar to the ratio of 0.192 observed in An. gambiae [19] or 0.204 in Ae. aegypti [10] but, higher than the ratio of 0.115 reported in D. melanogaster [13]. This result indicates that the purifying selection against deleterious mutations is acting in An. funestus. Indeed species with large effective population size such Drosophila or Anopheles species are generally more effective at purging deleterious mutations [26].

### Clustering pattern of the SNPs

We analyzed the distribution of SNPs identified in this study. We found 16 clusters of two directly neighboring SNPs, one cluster of 3 consecutive SNPs and 13 clusters of two SNPs separated by just 1 bp. For some SNP genotyping methods based on allele-specific amplification, ligation or single base extension principles for which primers need to be designed immediately adjacent to the SNP, it is important that the SNPs are not too close together to prevent primer designing. The presence of a polymorphism within approximately 20 bp will limit the possibilities for designing a robust primer. Most of the 494 SNPs identified in this study do not have a SNP within 20 bp on either side thus, and should be easily genotyped by one of these methods.

### Genomic position of the SNPs

Among the 50 genes amplified for SNP detection in this study, 45 are already physically mapped to the An. funestus genome by in situ hybridization [27], and the remaining 5 genes were genetically located to their respective chromosome by linkage mapping [7]. Overall, 29 SNPs were located on the X chromosome, 334 on chromosome 2 and 131 on chromosome 3. The higher number of SNPs observed on chromosome 2 is also a consequence of the fact that most of the studied genes are located on that chromosome. Table 4 gives the chromosomal location of the 50 genes across the genome of An. funestus.

### Polymorphism reliability

To assess the validity of the SNPs identified in this study, 30 SNPs were tested for segregation in isofemale lines. These SNPs were tested using different methods (pyrose-

---

**Table 2: Transition (Ts) and transversion (Tv) polymorphisms for different classes of DNA**

| Polymorphism                  | Probability | Coding Region | 3rd coding position | Fourfold |
|-------------------------------|-------------|---------------|---------------------|---------|
| Non coding regions            |             |               |                     |         |
| Ts                            | 106         | 85            | 44.5                | P = 0.021 |
| Tv                            | 85          | 85            | 100                 | P = 0.0  |
| %Tv                           |             |               |                     |         |
| Coding regions (Cd-R)         |             |               |                     |         |
| Ts                            | 201         | 102           | 33.6                | P = 0.204 |
| Tv                            | 102         | 102           | 100                 | P = 0.507 |
| %Tv                           |             |               |                     |         |
| Third coding position         |             |               |                     |         |
| Ts                            | 139         | 54            | 27.9                | P = 0.047 |
| Tv                            | 54          | 54            | 100                 | P = 0.0  |
| %Tv                           |             |               |                     |         |
| Fourfold degenerate sites     |             |               |                     |         |
| Ts                            | 60          | 35            | 58.8                |         |
| Tv                            | 35          | 35            | 100                 |         |
| %Tv                           |             |               |                     |         |

---

**Table 3: Indel polymorphism**

| Gene   | Coding region | Introns | 5' UTR | 3' UTR |
|--------|---------------|---------|--------|--------|
| 6P5    |               | 4 bp    |        |        |
| BU10   |               | 2 bp, 4 bp |      | 1 bp   |
| BU66   |               |         |        |        |
| BU93   |               | 3 bp    |        |        |
quencing, HOLA, SBE and AS-PCR) [7,28]. The Mendelian segregation ratio of each of these SNP loci at F0, F1 and F2 generations was examined in four families from reciprocal crosses between a pyrethroid resistant strain (FUMOZ-R) and a susceptible strain (FANG). Homozygous and heterozygous genotypes for each of these SNPs were observed. Importantly, the expected Mendelian ratio of 1:2:1 was respected in 27 of these 30 SNPs [7], confirming the polymorphism observed at these different positions. We can conclude from this result that the SNPs described in this study are then likely to be true polymorphisms rather than sequence artifacts and our scoring results indicate that they are suitable for use as genetic markers.

**Relevance of the SNPs**

The set of SNPs identified in this study provide a very useful tool for future genetic studies in *An. funestus*. These markers are of immediate use for association and QTL mapping studies. Some of these SNPs have been used for linkage mapping and identification of QTL involved in pyrethroid resistance in *An. funestus* [7]. This set of SNPs can be used as tools for population genetic studies in *An. funestus*. Genotyping large number of SNP markers will facilitate the study of genetic structure of natural populations and provide independent estimates of gene flow. It may provide additional markers to study the speciation process observed between the Folonzo and Kiribina chromosomal forms of *An. funestus* [29]. These markers may also be invaluable in monitoring insecticide resistance genes or genes involved in vector competence.

**Conclusion**

Through the sequencing of DNA fragments from 50 genes of *An. funestus*, we identified a set of 494 SNP markers and studied the pattern of genetic variability in this species. The distribution of SNPs in *An. funestus* was not neutral but under the influence of regional factors such as recom-
bination, the degeneracy of the genetic code and selective constraints for gene conservation. The SNP markers described constitute an important resource for more genetic studies in this important malaria vector.

Methods

Mosquito samples used for polymorphism discovery
We used adult female specimens of An. funestus from two laboratory strains, FANG and FUMOZ-R (seven specimens for each strain) as well as seven field specimens. FANG is a pyrethroid susceptible strain from Caluque, southern Angola and FUMOZ-R is a pyrethroid resistant strain from southern Mozambique [1]. Field specimens of An. funestus were collected from Kela village in Chikwawa district in southern Malawi.

Selection of gene sequences for SNP identification
Target genes were selected among cytochrome P450 genes for their putative involvement in insecticide resistance [30] or among genes of a broad range of functions that had been physically mapped to An. funestus polytene chromosomes [27] (see Table 4; Figure 3). They were also chosen to be distributed across the genome of An. funestus. The sequences of the physically mapped cDNAs were retrieved from Genbank. Determination of coding sequence, UTRs and intronic regions were done using the BLAST procedure through NCBI.

Gene amplification and sequencing
Genomic DNA was extracted using the LIVAK method as described previously [31]. Primers were designed using Primer3 software [32] to flank putative intron sites to maximize the chance of SNP identification. Genomic DNA from 21 individuals (7 from FUMOZ-R, 7 from FANG and 7 from Kela) was amplified for each gene. PCR was performed with 10 ng of genomic DNA in a final volume of 25 μl containing, 2.5 μl Taq buffer, 0.2 mM of dNTPs, 10 pmoles of each primer, 2.5 mM of MgCl2, 0.2 unit of Taq polymerase (Qiagen). Amplification was performed with the following conditions: 1 cycle at 94°C for 3 min; 35 cycles of 94°C for 30 s, 57°C for 30 s and elongation at 72°C for 30 s; followed by 1 cycle at 72°C for 10 min. The annealing temperature was optimized for each primer pair and varied between 53°C to 62°C.

PCR products were purified using the Qiaquick PCR purification kit (Qiagen) and directly sequenced on both strands using a Beckman CEQ 8000 and a pyrosequencing method [7].

SNP validation
Many of the SNPs discovered in this study were validated by different methods. As a part of an effort to construct a genetic map and to identify QTL involved in pyrethroid resistance, 30 SNP loci were genotyped in several families generated from a cross between FANG and FUMOZ-R strains of An. funestus. These SNPs were scored using a HOLA (Hot Oligonucleotides Ligation Assay) method [36], single base extension (SBE) using Beckman CEQ8000 and a pyrosequencing method [7].

Authors’ contributions
CSW (corresponding author) carried out the experiments; analyzed the data and wrote the manuscript. JH is the PI of the program that funded the work and contributed to the critical review of the draft manuscript. HR contributed to the design of the study and critical review of the draft manuscript. All authors read and approved the final manuscript.
| Genes | Chromosomal Location | Accession no. | Function | Forward primer | Reverse primer | Product length | No of SNPs |
|-------|----------------------|---------------|----------|----------------|----------------|----------------|-----------|
| 4G21  | X                    | AY648704      | Cytochrome P450 | GGCGTATAGCAAACGTAAAGC | CGCGGTAAACCCGTAATAGC | 303            | 2         |
| 9K1   | X                    | AY987362      | Cytochrome P450 | GTACGAGCTTGCCGTAATAC | CTTTCTTCTAGCTGACCTTGT | 243            | 9         |
| 4J12  | 3R                   | AY648706      | Cytochrome P450 | CCAACAAATCAGTTCCATGC | TTGAAAAGTGTCAAATAGC | 270            | 9         |
| 6P9   | 2R: 9A-12C           | AY779261      | Cytochrome P450 | GGCGCTTAGAACAAGAGATCA | AAGGAAGTAGTTCTCTT | 350            | 8         |
| 6P4   | 2R: 9A-12C           | AY987359      | Cytochrome P450 | GTACGAGCTTGCCGTAATAC | AAGGAAGTAGTTCTCTT | 430            | 13        |
| 6P5   | 2R: 9A-12C           | AY987360      | Cytochrome P450 | CTGCCCTTTAGAACTCTCCTC | AGATACACGTTAGGAATGC | 550            | 3         |
| 6Z1   | 2L: 25A-27D          | AY779263      | Cytochrome P450 | ACGATCGTTCTCGATCATG | TCGGTAGAAGCTCCAG | 550            | 12        |
| 6Z2   | 2L: 25A-27D          | AY779265      | Cytochrome P450 | BUC038901 type II transforming growth factor-beta receptor | TCGGTAGAAGCTCCAG | 525            | 1         |
| 9J12  | 3L                   | AY779263      | Cytochrome P450 | TACCGCGTGCAGCCTTGA | CTTTGGCAGCGAAGTAA | 195            | 4         |
| 9J14  | 3L                   | AY779265      | Cytochrome P450 | BUC038901 Acetylcholinesterase | GGGTACGGGACACATGAGCCG | 1050           | 15        |
| BU01  | X:2B                 | BU039001      | Type II transforming growth factor-beta receptor | GTGTTGTTGCTGAGGATTT | GGCATCGGTAAACCCGAGA | 500            | 7         |
| BU08  | 2R:7C-10B            | BU038908      | Rhodopsin | CATTGGTTGACCCACATTTT | GTCATTGTGTTTACCCGAGA | 500            | 7         |
| Ache  | 2R:9C-12C            | D0534335      | Acetylcholinesterase | GGGTACGGGACACATGAGCCG | GGTAAAAGTGTGACCTTGA | 1050           | 15        |
| BU10  | 2L:28A               | BU039010      | Cyt-c-p1 | AAGCACTGTTAAACCCCTT | ACCCTGAAGAATCAGGACAT | 650            | 10        |
| BU11  | 3L:43B               | BU038911      | Protein transporter | ATCTGCTTTCCCTGAGCTTG | ATCGGGAATCCCTGAGC | 670            | 10        |
| BU12  | 2R:7B                | BU038912      | Alpha tubulin | AAGCTCGAGTCGCGACAT | AGGTAAGCGAGTGCAGT | 800            | 5         |
| BU13  | 2R:15C               | BU038913      | Signal sequence receptor | ACCCTGAAGAATCAGGACAT | GGTAAAGTGTGACCTTGA | 630            | 10        |
| BU19  | 2R:12B               | BU038919      | Chitinase | CTGTTGCTTCTGCTACATC | CGGTCACGTACAAATGCT | 670            | 10        |
| BU021 | 3L:38C-40B           | BU039021      | Tubulin beta-3 chain | GAGTTGTTGATGCGTGTG | GTCGGGAAAACACATGCT | 400            | 8         |
| BU21  | X:3A                 | BU038921      | Phosphoribosylaminomimidazole-carboxylase | TTTCAAGGTTAACCGGTTGA | CCATCAAGATGACGACCA | 475            | 11        |
| BU25  | 2R:12B               | BU038925      | Ferritin heavy chain-like protein | GCGTAAGCTGCTGCTTCTC | ATTCCTACGCTAGGAGCTTTATGCT | 1200           | 2         |
| BU29  | 2L:27B               | BU038929      | Sensory appendage protein | CACCAACGCTGCGTTGCAATC | AGGAAAATGTTGACCGT | 410            | 6         |
| BU34  | X:1C                 | BU038934      | NADH dehydrogenase | GGCAGCTGACGCAAGTGTC | CAGTACACGACCCGACACAC | 400            | 6         |
| BU35  | 2R:12B               | BU038935      | CG6846 gene product | TCCAGAACACACGTTGCTG | ACTTGCTCGATCCTGGAATGTC | 400            | 9         |
| BU40  | 2R:14B               | BU038940      | Glutathion peroxidase | AGGCAAAATCAATTTTTGGA | CGTAAACCAGGACACACAC | 1150           | 10        |
| BUS6  | 2R:7B                | BU038956      | Novel An. gambiae salivary protein | AATCTGAGATGCGGACAGA | AATCTGAGGTGGCGTGG | 1350           | 13        |
| Gene ID | Position | Protein Description |
|---------|-----------|---------------------|
| BU58    | 2R:12D    | Translation initiation factor |
| BU62    | 2L:23A    | cAMP responsive element binding protein |
| BU66    | 3R:30A    | Lysozyme |
| BU70    | 2R:7C     | Ubiquitin fusion 80 |
| BU71    | 3L:39A    | Structural protein of peritrophic membrane |
| BU72    | 2R:12B    | RHO small monomeric GTPase |
| BU73    | 2R:7A     | Actin binding |
| BU76    | 2R:10B    | Translation initiation factor |
| BU77    | 2R:10C    | Ubiquitin conjugating enzyme |
| BU82    | 2R:14D    | Unknown |
| BU85    | 2R:12E    | Phosphoglycerate mutase |
| BU88    | 2R:11B    | Translation initiation factor |
| BU90    | 3R:35B    | Gustatory receptor |
| BU92    | 3L:39A    | Microtubule binding |
| BU93    | 2R:7C-10C | Prefoldin subunit 2 |
| BU98    | 3L:46B    | CG7630 gene product |
| BU83    | 3R:32B    | Peripherin |
| BU97    | 3R:36C    | NADH dehydrogenase (ubiquinone) |
| BU91    | 2L:20C    | CG18997 gene product |
| BU973   | 3R:36F    | PolyA-binding protein II |
| BU974   | 3L:40A    | Serine-type peptidase |
| BU982   | 2R:12B    | Ferritin 2 light chain homologue |
| BU996   | 2R:8D     | Vacuolar hydrogen-transporting ATPase |
| Kdr     | 3R:36A-37E| Sodium channel gene |

**Table 4: Characteristics of genes amplified for SNP detection (Continued)**
Relative location of studied genes on the *An. funestus* genome. For definitions of genes, see Table 4. This figure was adapted from [37].
References

1. Hunt RH, Brooke BD, Pillay C, Koekemoer LL, Coetzee M. Laboratory selection for and characteristics of pyrethroid resistance in the major vector Anopheles funestus. *Med Vet Entomol* 2005, 19(3):271-275.
2. Sinkins SP, Hackett BJ, Costantini C, Vulule J, Ling YY, Collins FH, Besansky NJ. Isolation of polymorphic microsatellite loci from the malaria vector Anopheles funestus. *Mol Ecol* 2000, 9(4):490-492.
3. Cohu et al. Simard F, Berthomieu A, Raymond M, Fontenille D, Weill M. Isolation and characterization of microsatellite DNA markers in the malaria vector Anopheles funestus. *Molecular Ecology Notes* 2002, 2(4):490-492.
4. Entman AL, I, Simard F, Raymond M, Fontenille D. Population structure of the malaria vector Anopheles funestus in Senegal based on microsatellite and cytogenetic data. *Insect Mol Biol* 2004, 13(3):251-258.
5. Cohuet et al. Diaz I, Simard F, Raymond M, Rosset F, Antonio-Nkondjo C, Awono-Ambene. *Phil.* Wondji CS, Fontenille D. Gene flow between chromosomal forms of the malaria vector Anopheles funestus in Cameroon, Central Africa, and its relevance in malaria fighting. *Genetics* 2005, 169(1):301-311.
6. Michel AP, Ingrasci MJ, Schernhammer BJ, Kern M, Le Goff G, Coetzee M, Elissa N, Loeffer S, Coetzee C, Besansky NJ. Rangewide population genetic structure of the African malaria vector Anopheles funestus. *Mol Ecol* 2005, 14(14):4235-4248.
7. Wondji CS, Hunt RH, Pignatelli P, Steen K, Coetzee M, Besansky NJ, Lobo N, Collins FH, Hemingway J, Coetzee C. Besansky NJ. An integrated genetic and physical map for the malaria vector Anopheles funestus. *Genetics* 2005, 171(4):1779-1787.
8. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Paul N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Young LM, Wells T, Kopczynski C, Ellis MC, Byrne N, Gierasch LJ, Ward MR, Kolb H, Batzer MA, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen LX, Basilion JP, Stanton VP Jr., Baker BS, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Mardis ER, Wincker P, Aken KB, Searle SM, Paten B, Hickey RJ, Prigozhin I, Yang H, Ruan Y, Chen Y, Jiang H, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jailion O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malea JA, Mcloosh TC, Meister S, Miller J, Mobarbry C, Monig E, Siroky D, O’Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharonkova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Under MF, Walfinz B, Wang A, Wang J, Wang M, Wirth FX, Woodford KJ, Worton MR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, Zhao S, Zhu SC, Zhimulev I, Coluzzi M, della Torre U, Roth CW, Louis C, Kalush F, Mural RJ, Myers EW, Adams MD, Smith HO, Broder S, Gardner MJ, Fraser CM, Birney E, Bork P, Brey PT, Venter JC, Weissenbach J, Kafatos FC, Collins FH, Hoffman SL. The genome sequence of the mosquito Anopheles gambiae. *Science* 2002, 298(5591):129-149.
17. Holliday R, Grigg GW. DNA methylation and mutation. *Mutat Res* 1993, 285(1):61-67.
18. Brooks AJ: The essence of SNPs. *Gen* 1999, 234(2):177-186.
19. Liconon P, Cohuet A, Fontenille D. Intraspecific nucleotide variation in Anopheles gambiae: new insights into the biology of the malaria vectors. *Am J Trop Med Hyg* 2004, 71(6):795-802.
20. Aquadro CF, Bauer DuMont V, Reed FA: Genome-wide variation in the human and fruitfly: a comparison. *Curr Opin Genet Dev* 2001, 11(6):627-634.
21. Begun DJ, Aquadro CF: Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* 1992, 356(6369):519-520.
22. Kreitman AD, Begun DJ: Polymorphism and divergence from noncoding sequences of Drosophila melanogaster and D. simulans: evidence for nonequilibrium processes. *Mol Biol Evol* 2005, 22(1):51-62.
23. Galtier N, Bazin E, Biem N: GC-biased segregation of noncoding polymorphisms in Drosophila. *Genetics* 2006, 172(1):221-228.
24. Shen LX, Basilion JP, Stanton VP Jr.: Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc Natl Acad Sci U S A* 1999, 96(14):7871-7876.
25. Schmid KJ, Tautz D: A screen for fast evolving genes from Drosophila. *Proc Natl Acad Sci U S A* 1997, 94(18):9746-9750.
26. Kimura M, Takahata N: Selective constraint in protein polymorphism: study of the effectively neutral mutation model by using an improved pseudosampling method. *Proc Natl Acad Sci U S A* 1983, 80(4):1048-1052.
27. Sharakhov IV, Serazin AC, Grushko OG, Dana A, Lobo N, Hillenmeyer ME, Westerman R, Romero-Severon J, Costantini C, Sagnon N, Collins FH, Besansky NJ: Inversions and gene order shuffling in Anopholes gambiae and A. funestus. *Science* 2002, 298(5591):182-185.
28. Lynd A, Ranson H, McCaill PJ, Randle NP, Black WC, Walker ED, Donnelly MJ: A simplified high-throughput method for pyrethroid knock-down resistance (kdr) detection in Anopheles gambiae. *Malar J* 2003, 2(16):1-6.
29. Costantini C, Sagnon N, Ibboudo-Sanogo E, Coluzzi M, Boccolini D: Chromosomal and bionomic heterogeneities suggest incipient speciation in Anopheles funestus from Burkina Faso.* Parasitologia* 1999, 41(4):595-611.
30. Ameyia DA, Koekemoer LL, Vaughan A, Morgan JC, Brooke BD, Hunt RH, Ranson H, Hemingway J, Coetzee M: Isolation and sequence analysis of P450 genes from a pyrethroid resistant colony of the major malaria vector Anopheles funestus. *DNA Seq* 2005, 16(6):437-445.
31. Collins FH, Menez MA, Rasmussen MO, Meahaffy PC, Besansky NJ, Finneroy V: A ribosomal RNA gene probe differentiates members species of the Anopheles gambiae complex. *Am J Trop Med Hyg* 1987, 37(1):37-41.
32. Rozen S, Skaltsky H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 2000, 132:365-386.
33. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994, 22(22):4673-4680.
34. Rozas J, Sanchez-DeBarrio JC, Messegue X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 2003, 19(18):2476-2479.
35. Nei M, Gojobori T: Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 1986, 3(5):418-426.

36. Black WC, Gorrochotegui-Escalante N, Duteau NM: Heated oligonucleotide ligation assay (HOLA): an affordable single nucleotide polymorphism assay. *J Med Entomol* 2006, 43(2):238-247.

37. Sharakhov I, Bragina O, Grushko O, Cohuet A, Guelbeogo WM, Boccolini D, Weil M, Costantini C, Sagnon N, Fontenille D, Yan G, Besansky NJ: A microsatellite map of the African human malaria vector *Anopheles funestus*. *J Hered* 2004, 95(1):29-34.