BMC Genomics

Research article

SNP discovery and molecular evolution in *Anopheles gambiae*, with special emphasis on innate immune system

Anna Cohuet*[1], Sujatha Krishnakumar[2], Frédéric Simard[1,3], Isabelle Morlais[1,3], Anastasios Koutsos[4,5], Didier Fontenille[1], Michael Mindrinos[2] and Fotis C Kafatos*[4]

Address: [1]Institut de Recherche pour le Développement, UR 016, BP 64501, 911 Avenue Agropolis, 34394 Montpellier Cedex 5, France, [2]Stanford Genome Technology Center, 855 California Avenue, Palo Alto CA 94304, USA, [3]Organisation de Coordination pour la lutte contre les Endémies en Afrique Centrale, Laboratoire de Recherche sur le Paludisme, BP 288, Yaounde, Cameroon, [4]Imperial College London, Division of Cell and Molecular Biology, Sir Alexander Fleming Building, South Kensington Campus, London, SW7 2AZ, UK and [5]BMC-series Journals, BioMed Central, Middlesex House, 34-42 Cleveland Street, London W1T 4LB, UK

Email: Anna Cohuet* - cohuet@mpl.ird.fr; Sujatha Krishnakumar - sujatha.kumar@gmail.com; Frédéric Simard - Frederic.Simard@ird.fr; Isabelle Morlais - morlais@ird.fr; Anastasios Koutsos - anastasios.koutsos@biomedcentral.com; Didier Fontenille - didier.fontenille@mpl.ird.fr; Michael Mindrinos - mindrinos@stanford.edu; Fotis C Kafatos* - f.kafatos@imperial.ac.uk

* Corresponding authors

Published: 19 May 2008

BMC Genomics 2008, **9**:227 doi:10.1186/1471-2164-9-227

Received: 19 June 2007
Accepted: 19 May 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/227

Abstract

Background: *Anopheles* innate immunity affects *Plasmodium* development and is a potential target of innovative malaria control strategies. The extent and distribution of nucleotide diversity in immunity genes might provide insights into the evolutionary forces that condition pathogen-vector interactions. The discovery of polymorphisms is an essential step towards association studies of susceptibility to infection.

Results: We sequenced coding fragments of 72 immune related genes in natural populations of *Anopheles gambiae* and of 37 randomly chosen genes to provide a background measure of genetic diversity across the genome. Mean nucleotide diversity (π) was 0.0092 in the *A. gambiae* S form, 0.0076 in the M form and 0.0064 in *A. arabiensis*. Within each species, no statistically significant differences in mean nucleotide diversity were detected between immune related and non immune related genes. Strong purifying selection was detected in genes of both categories, presumably reflecting strong functional constraints.

Conclusion: Our results suggest similar patterns and rates of molecular evolution in immune and non-immune genes in *A. gambiae*. The 3,214 Single Nucleotide Polymorphisms (SNPs) that we identified are the first large set of *Anopheles* SNPs from fresh, field-collected material and are relevant markers for future phenotype-association studies.

Background

Anopheles gambiae, the main vector of the human malaria parasite *Plasmodium falciparum* in SubSaharan Africa, is the most medically relevant insect in the world. Together these two organisms are responsible annually for more than a million of deaths in Africa, mostly young children. This epidemic is worsening [1], prompting the search for innovative strategies towards effective and efficient

malaria control. One approach aims to disrupt parasite development in the mosquito vector and thus alleviate malaria transmission intensity [2]. This strategy requires clear understanding of the intimate interactions between parasite and vector, and of the mechanisms that regulate the interaction. Large scale gene expression profiling in *A. gambiae* has revealed that the insect's innate immune system is stimulated following infection by *Plasmodium* parasites [3-6], highlighting this system as a primary candidate for interventions to control the infection. Extensive studies identified potentially relevant mechanisms of innate immune response, including a balance between positive and negative mosquito factors towards the parasite [7-10]. Genetic variation underlies the susceptibility of *A. gambiae* to *Plasmodium* infection: refractory mosquito strains have been selected [11,12] and QTLs identified for susceptibility/refractoriness to the model parasite, *P. cynomolgi* [13,14]. However, to date the mechanism(s) involved and the underlying genetic basis remain unknown [15]. Some studies conducted in field conditions with wild mosquito and parasite populations further demonstrated genetic variability in the mosquito's susceptibility to *P. falciparum* infection and identified putatively involved genes [16-18]; their contribution to the phenotype and relevance *in natura* still remain to be assessed fully.

The availability of the *A. gambiae* genome sequence [19] has opened new perspectives for exploratory genetic studies in this species. The activities and evolution of its immune system are now being studied intensively, and could provide insights into the past and present patterns of interaction with the pathogen. Initial studies revealed selective constraints of diverse nature acting on some immunity genes [20-24]. The current pattern of malaria transmission is relatively recent (less than 10,000 years), but has exerted strong selective pressure on human populations that led to the selection of resistance alleles, some of which are strongly deleterious [reviewed in [25]]. Signatures of selection were found on most genes implicated in *P. falciparum* resistance in humans [26] and conversely, population studies on *P. falciparum* have detected selection sweeps on its genome [27]. It is rational to expect that *P. falciparum* has also exerted selective pressure on *A. gambiae*. If so, evolutionary genetics of the immune system in natural *A. gambiae* populations should pinpoint resistance or susceptibility genes based on their peculiar molecular makeup.

Single nucleotide polymorphisms (SNPs) are the commonest mode of genetic variation in vertebrates and invertebrates [28-30]. As such, SNPs rapidly became the preferred and most useful molecular markers for association studies, high resolution linkage mapping and population genetics studies [26,31,32]. In coding regions, synonymous SNPs (sSNPs) that do not result in amino-acid change are likely neutral markers for population genetic studies. In contrast, non synonymous SNPs (nsSNPs) alter protein structure and could be retained by natural selection. The search for SNPs in *A. gambiae* immunity genes is an initial step towards genetic dissection of vector competence in the wild [21].

Here, we report coding region SNPs in a representative set of genes from different families associated with putative innate immune functions in *A. gambiae* [5]. Genes chosen randomly across the genome were included for comparison with the immunity data, such that the genome-wide effects of demographic history may be distinguished from gene-specific effects of selection. The study was carried out on natural populations of *A. gambiae s.s.* (both M and S molecular forms) and its sibling species, *A. arabiensis*, collected from field sites in Cameroon (Central Africa).

Results

We studied 72 immune related genes, representing innate immunity gene families and functions throughout the genome [5]. Additionally, 37 non-immune related genes randomly chosen along the genome were included in the study. The relative proportion of genes involved in the different stages of immune response [5] and their chromosomal location are shown in Figure 1. Target genes are listed in tables [see Additional files 1 and 2], with their Ensembl gene IDs, accession numbers in the EMBL database, their putative role in the immune response (functional class), primer sequences, gDNA or cDNA nature of the template, chromosomal location, length of the fragment analyzed, and the number of alleles sequenced.

Sequence Polymorphism

We analyzed a total of 2,608,472 nucleotides across 109 coding fragments of an average length of 524 base pairs. Nucleotide diversity indices and results of the tests for selection within populations are given in Additional files 3 and 4. A total of 3,214 SNPs were detected in our dataset: 2,026 were observed in immune related genes and 1,188 in non immune related genes, respectively. Correspondingly, 1,711 and 1,071 SNPs were newly identified polymorphisms (not previously reported). We also detected 54 indels, always as a multiple of 3 bases preserving the open reading frame (ORFs). In the populations we studied, 432 (78%) of the 554 SNPs previously reported in ENSEMBL were detected (data not shown). Nucleotide diversity along the chromosomes is presented Figure 2, including immune related genes and control genes. Mean nucleotide diversity (π) across all genes varied significantly between species and populations (Mann-Withney U test, $P < 0.05$): it was higher in the *A. gambiae* S form ($\pi = 0.0092$) than in either the M form ($\pi = 0.0076$) or in *A. arabiensis* ($\pi = 0.0064$). Within each species, however, no

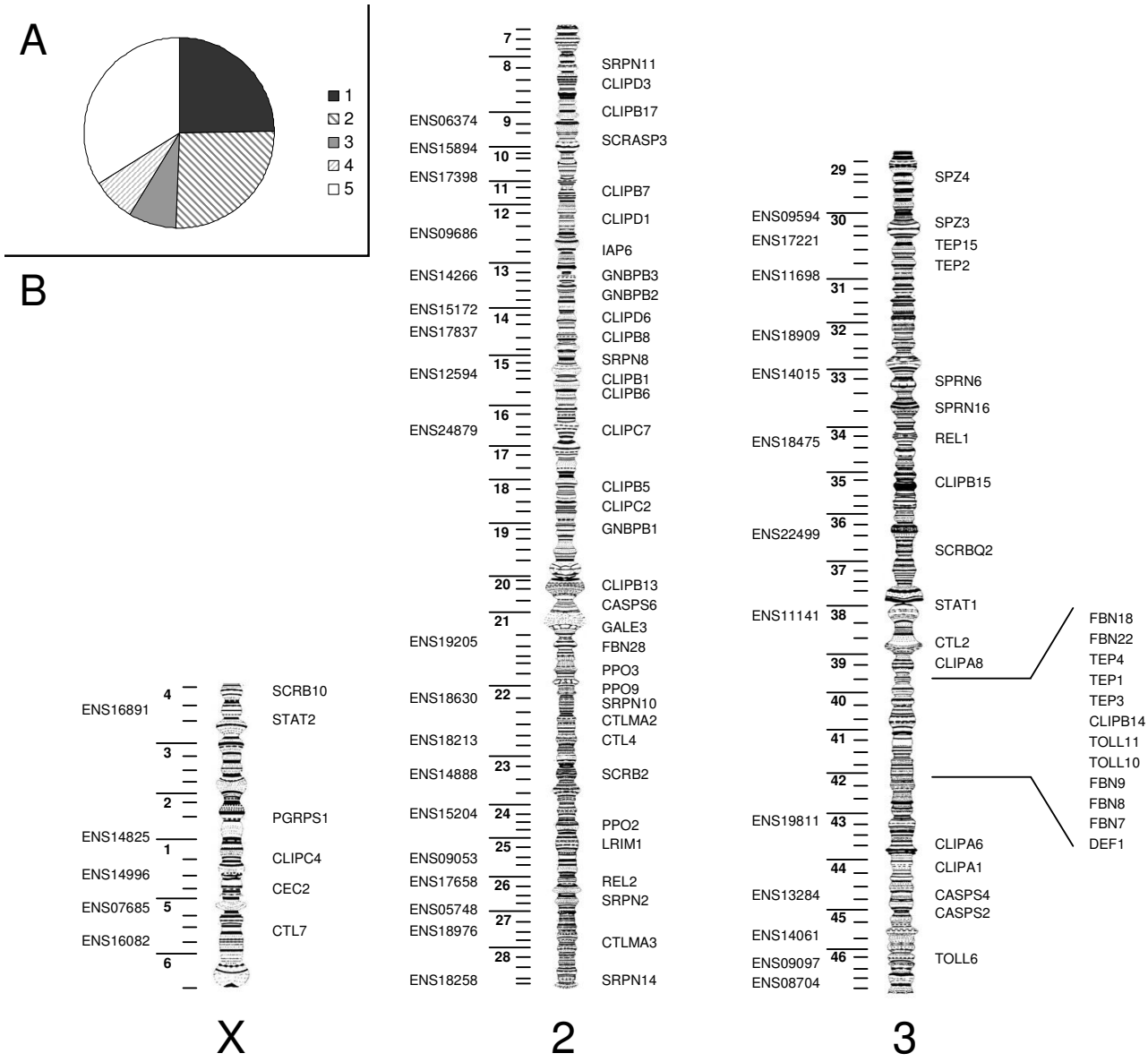

Figure 1

Studied genes. A: Proportion of studied genes putatively involved in (1) Recognition, (2) Modulation, (3) Signal transduction and (4) Effectors molecules of immune response and (5) Non imnnune related genes. B: Relative position of studied genes on the *Anopheles gambiae* genome. Immune related genes are indicated at the right of each chromosome, non immune related genes on the left by using the ENSEMBL ID reduced to the three first letters and five last numbers.

statistically significant differences in mean nucleotide diversity was detected between immune related and non immune related genes, nor between different functional class within immune related genes ($P > 0.05$). Similarly, nucleotide diversity was evenly spread on the four autosomal arms in the three populations. However, even after correction for lower effective population size on the X chromosome by multiplying estimates by 4/3 (see Methods), significantly lower genetic diversity ($P < 10^{-3}$) was observed on the X chromosome in each population (mean π: 0.00131 for *A. arabiensis*, 0.00319 for the M form and 0.00358 for the S form). Noticeably however,

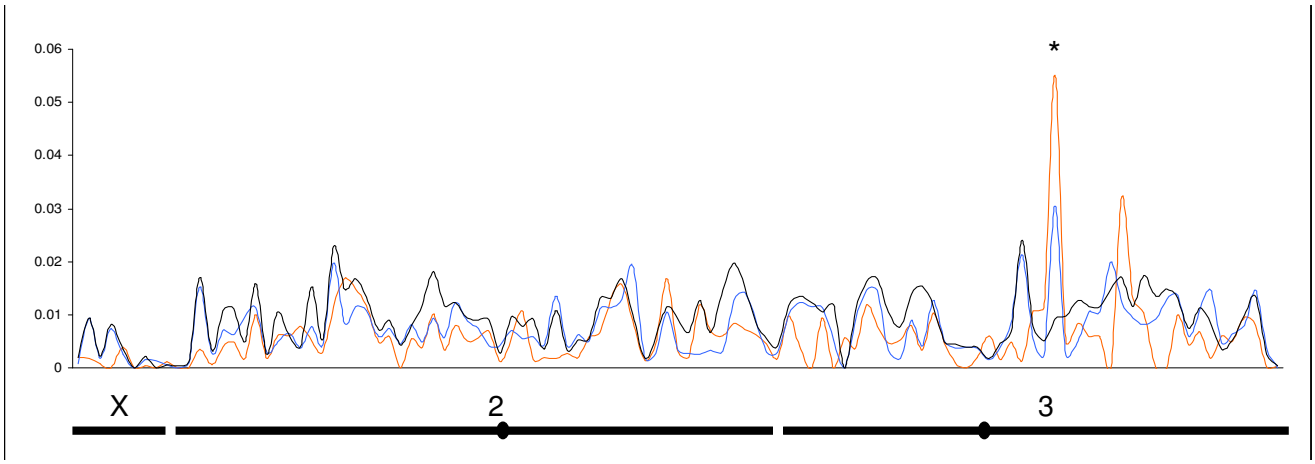

Figure 2
Nucleotide diversity. Nucleotide diversity along the chromosomes for *A. arabiensis* (red line), *A. gambiae* M form (blue line) and *A. gambiae* S form (black line). Data from immune related genes and control genes are included. Abscise represents position of the genes along the genome. Chromosomes and centromeres are represented below. The asterisk shows the position of TEP1 gene.

the gene TEP1 on the third chromosome showed much higher genetic diversity than all the other genes we investigated (Figure 2).

Species divergence and population differentiation

Pairwise estimates of genetic differentiation (Fst) between all population pairs are given on Figure 3, together with their statistical significance, for each of the 109 genes. Similar levels of genetic differentiation were detected by immune related and non immune related genes, as well as between different functional classes within immune related genes (Mann-Whitney U test on Fst estimates, $P > 0.05$). Hence, data from all categories of genes were pooled for further analyses. Average Fst estimates between the M and S molecular forms of *A. gambiae* was similar across all 3 chromosomes (Mann-Whitney U test on Fst estimates, $P > 0.05$), with an overall Fst = 0.1377 ($P < 10^{-3}$). Allelic frequencies differed among the incipient species but only one sSNP found in a non-immune related gene in section 5D on the X chromosome (ENSANGG00000016082) segregated between the M and S forms. Mean genetic divergence between *A. arabiensis* and either the M or S form were similar (Fst = 0.5268 and Fst = 0.4729 respectively, Mann-Whitney U test $P > 0.05$), and were significantly higher than between the two molecular forms ($P < 10^{-3}$). Between *A. arabiensis* and the M or S form, 175 and 125 SNPs, respectively were fixed and mean genetic differentiation estimates were significantly higher on the X chromosome ($P < 0.01$).

Tests of selection in immune related and non-immune related genes

The Tajima D statistic was computed for each gene in all populations [see Additional files 1 and 4]. At equilibrium between random genetic drift and mutation, the expected value of D for neutral markers is close to zero. This statistic detected only few genes with significant departure from neutrality, reflecting locus-specific selection rather than the effect of demographic instability. Noteworthy, TEP1 showed a highly significant positive value of D in *A. arabiensis* ($P < 0.01$). This departure from neutrality was due to the coexistence, at high frequency in this population, of two highly diverged alleles: the previously described and widespread TEP1r and TEP1s [7] were observed at a relative frequency of 37.5% and 62.5%, respectively, in *A. arabiensis*. In *A. gambiae*, TEP1s was the most frequent allele and TEP1r was found at low frequency (12.5% in the M form).

The Z test of selection revealed a very high proportion of genes that are under selection across populations: only 3 genes among the immune related genes (3/72 = 4.1%) and one among the non immune related genes (1/37 = 2.7%) showed the ratio of sSNPs/nsSNPs expected under the hypothesis of neutrality. All other genes are strongly deficient in non-synonymous changes, suggesting prevalence of purifying selection. At the population level, the test was less often significant, most likely reflecting the lack of statistical power when sequences are not sufficiently polymorphic [33].

For each pair of populations, Ka/Ks ratios are shown in Additional files 5 and 6 for immune and non immune related genes, respectively. Ka/Ks ratios are expected to equal 1 if the genes under scrutiny behave neutrally. In most case, however, and for both immune and non immune related genes, the Ka/Ks ratios were much lower than 1. Such a pattern implies that a mutation changing the encoded amino acid sequence is much less likely to be different between two species than one which is silent. This is consistent with the results of the Z test of selection and suggests that purifying selection is a major force driving the evolution of both immune and non immune related genes in all populations.

Lack of fixed mutations precluded the implementation of the McDonald-Kreitman test between *A. gambiae* molecular forms. For the comparison of *A. arabiensis* with the two *A. gambiae* molecular forms, the test was computed across groups of genes in order to encompass a sufficient number of fixed variations for building the contingency tables. However, combining data from different genes for McDonald-Kreitman tests has established drawbacks which can lead to spurious cases of positive selection [34], thus significant tests need to be carefully interpreted. Table 1 shows the total numbers of replacement and silent polymorphisms that are shared or fixed between pairs of species, across immunity and non immunity related genes. For the latter group, the test revealed adequacy with neutral expectations (Two tailed Fisher exact test: P = 0.426 and P = 0.622 for each pair of species, respectively). Deviations from neutrality in immune related genes were at the limit of the significance threshold for both species pairs (Fisher exact test: P = 0.038 and P = 0.081, respectively) leading to inconclusive results as a reflection of limited statistical power in our dataset.

Discussion

This study provides the most extensive set of data on molecular polymorphisms in immune and non immune related genes in *A. gambiae* and *A. arabiensis*. We observed a high level of nucleotide diversity (π>0.006) in coding regions within members of the *A. gambiae* complex. This is approximately ten fold higher than the level of nucleotide diversity observed in coding regions of the human genome [29] but is comparable to the estimates reported from *Drosophila* or in previous studies on *Anopheles* [20,21,24,35-38]. Nucleotide diversity is a product of both mutation rate and effective population size. Mutation rates were shown to be similar in humans and *Drosophila* [39,40], so the ten fold difference in average

nucleotide diversity between species was attributed to the approximately ten fold larger long term effective population size in *Drosophila* than in humans (Ne = 300,000 versus 20,000 respectively) [41,42]. The effective population size of *A. gambiae* is estimated at several hundred thousand [43], which falls within the range of *Drosophila* estimates and might explain similarities in levels of genetic diversity in both species. We also observed a slightly reduced diversity in *A. arabiensis* compared to *A. gambiae* and in the M form of *A. gambiae* compared to the S form. Reduced genetic diversity in *A. arabiensis* was already observed previously [20,44], but colonization and further maintenance of *A. arabiensis* for several generations in an insectary might have added to this trend through increased genetic drift. Alternatively, long term effective population sizes differences might be expected between these species, in light of their distinct bionomics and distribution across Africa [45,46].

Patterns of genetic diversity can be influenced by differences in recombination rate across the genome, with a higher genetic diversity and a faster evolution rate expected in high recombination regions [47,48]. Our results must therefore be interpreted taking into account this potential heterogeneity along the *A. gambiae* genome. The known factors influencing recombination rate in *A. gambiae* are the chromosomal inversions and the proximity to centromeres and telomeres [44,49]. The distribution of immune related genes on the genome did not show any aggregation in particular regions (Figure 1), therefore we did not expect that the diversity pattern observed in immune related genes would be a consequence of heterogeneity of recombination rates along the genome. However, considering genes position on the genome sheds light on some evolutionary processes in *A. gambiae*. Paracentric chromosomal inversions are very abundant in the *A. gambiae* complex. More than 120 polymorphic inversions have been detected in natural populations [49,50]. Ten inversions are fixed in the different species of the complex and can be used to differentiate individual specimens. Several lines of evidence suggest that these chromosomal arrangements are incidental to ecotypic adaptation and may be directly involved in the past and current speciation processes occurring within this species complex, mainly because recombination suppression between alternative chromosomal arrangements might protect arrays of co-adapted genes [46,51-56]. Mathematical models for inversion-based local adaptation and/or speciation predict lower genetic diversity within species in the chromosomal region involved in speciation, and higher divergence between species [57,58]. In agreement with these models, we found evidence for reduced variability and higher genetic divergence between species on the X-chromosome. Indeed, *A. arabiensis* and *A. gambiae* are differentiated by a fixed chromosomal arrangement (Xag) that inverts a large part of the X chromosome [49]. Reduced effective population size for the X chromosome is not sufficient to explain the observed pattern which is therefore in agreement with previous studies demonstrat-

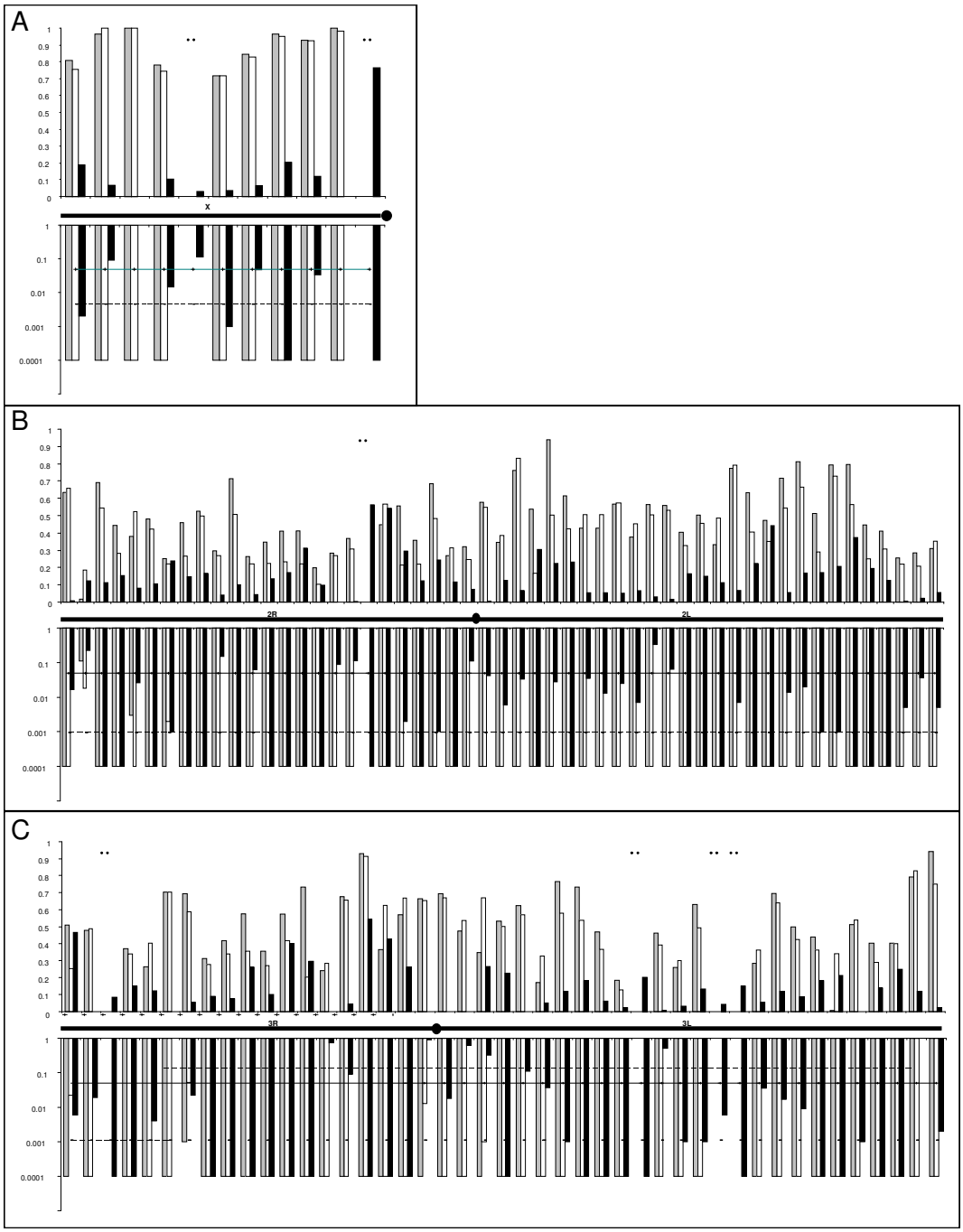

Figure 3

Genetic differentiation between populations. Genetic distance between pairs of populations on chromosomes X (A), 2 (B) and 3 (C). Fst estimates are shown in the top part of each graph and the corresponding P-values are shown below, in grey for *A. arabiensis*/*A. gambiae* M form comparisons, white for *A. arabiensis*/*A. gambiae* S form and in black for *A. gambiae* M form/*A. gambiae* S form. Data from immune related genes and control genes are included. Horizontal straight lines represent the significance threshold at P= 0.05, dashed lines: P = 0.05 after correction for multiple tests (Bonferroni sequential procedure). Missing data are indicated with a dot. Negative Fst values that are always not significant were represented as equal to zero. Highly significant P values ($<10^{-4}$) were represented as P = 0.0001 to be shown on a logarithmic scale.

Table 1: McDonald and Kreitman test results in immune and non immune related genes

A: Between *A. arabiensis* and M form for immune related genes			
	Non synonymous	Synonymous	Ratios
Interspecific fixed divergences	36	63	0.571
Intraspecific polymorphisms	344	905	0.380
Ratios	0.104	0.070	P = 0.081
B: Between *A. arabiensis* and S form for immune related genes			
	Non synonymous	Synonymous	Ratios
Interspecific fixed divergences	24	37	0.649
Intraspecific polymorphisms	421	1171	0.360
Ratios	0.057	0.032	P = 0.038
C: Between *A. arabiensis* and M form for non immune related genes			
	Non synonymous	Synonymous	Ratios
Interspecific fixed divergences	16	55	0.290
Intraspecific polymorphisms	151	675	0.224
Ratios	0.106	0.081	P = 0.426
D: Between *A. arabiensis* and S form for non immune related genes			
	Non synonymous	Synonymous	Ratios
Interspecific fixed divergences	14	49	0.286
Intraspecific polymorphisms	192	798	0.241
Ratios	0.073	0.061	P = 0.622

SNPs were counted for the groups of genes (immune related or non immune related genes) within species (Intraspecific polymorphisms) or between species (Interspecific fixed divergences). The probability (P) is calculated using a 2-tailed Fisher's exact test.

ing a "large × effect" on differentiation between these sibling species [44,59]. The "large × effect" hypothesis assumes the existence of speciation genes on the X chromosome responsible for ecological and/or behavioral adaptations that affect interspecific mating and/or hybrid fitness.

No X chromosome inversions, detectable at the cytogenetic level, distinguish the M and S forms of *A. gambiae*. Moreover, our samples were collected from an area of South Cameroon where *A. gambiae* M and S are known to be homosequential for the standard karyotype on all autosomes [60-62]. Accordingly, levels of genetic differentiation (single gene Fst estimates) were generally much lower between the M and S forms of *A. gambiae* than they were between these populations and *A. arabiensis*, and only occasionally did reach statistical significance. However, consistent with recent evidence for increased differentiation due to reduced recombination in the centromeric region of the X chromosome of *A. gambiae* [44,62-64], the only fixed SNP we observed between M and S maps to the proximal region of the X chromosome (X5D), a region that is considered as a "speciation island" between the M and S forms of *A. gambiae* [62]. Except for the X chromosome, we observed comparable genetic diversity across the entire genome. We did not detect any centromere or telomere effect, but a higher density of genetic markers would be necessary to draw firm conclusions about reduced diversity in these regions. The constant distribution of genetic diversity on autosomal chromosomes observed in the present study is in contrast with results of the *A. gambiae* genome project, where a highly variable distribution was observed [19]. The genome sequencing project utilized the PEST strain that was established from a mix of several natural populations of the M and S forms of *A. gambiae*, maintained under insectary conditions for several years and exposed to bottlenecks and selections [65]. It is likely that the uneven distribution of diversity in the PEST strain resulted from maintenance in the insectary, and that diversity in natural populations is more evenly distributed. It must be kept in mind, however that the present study focuses on populations from only one location for each species and might reflect only a portion of the natural genetic diversity of the species. Moreover, Cameroon might be an area where the M and S forms of *A. gambiae* have achieved one of the highest level of genetic differentiation observable throughout the species range, as was first described by

Wondji et al. [61] using microsatellite data and further expanded by Turner et al.[62,66], using sequence data. Overall, our estimates of Fst between species/molecular forms are in strong agreement with those of Wondji et al. [61] and Turner et al. [62,66], who detected one of the highest and most significant level of genetic differentiation between the M and S forms of *A. gambiae* observable throughout the species range.

Tests for departure from neutrality are based on the assumption of mutation-migration-drift equilibrium. However, evidence for recent population expansion and radiation has been found in both *A. gambiae* and *A. arabiensis* [67,68]. Unstable demographic history can produce patterns of genetic variation indistinguishable from those of selection [69,70]. However, demographic history affects similarly the entire genome, while selection is locus-specific. For example, rapid population expansion is expected to result in highly negative Tajima D values, as a consequence of the rapid increase in number of polymorphic sites (S), and excess of low frequency alleles that have little effect on π [71]. This pattern was not observed in our dataset, where most computed Tajima D values were negative but not statistically significant. The Z test of selection and Ka/Ks ratios demonstrated a deficit of non synonymous mutations in most of the genes (immune related or not). The influence of population size changes on tests based on synonymous and non synonymous variations appears to be weak, even if not fully understood [35]. Therefore, it is most likely that the deficit of non synonymous mutations is due to generalized purifying selection acting on *A. gambiae* ORFs, probably reflecting functional constraints on the encoded proteins.

In *Drosophila*, several immunity genes revealed directional selection [22,72,73] and a broad comparison of immune system and non immunity genes supported the hypothesis that pathogens exert a selective pressure on the immune system [74]. Evidence for directional selection driving evolution of the immune system in *Drosophila* was consistent with the "arms race" model of co-evolution [75]. In this model, the pathogen constantly evolves to escape the host's immune response and, in turn, the hosts' immune system evolves to better control infections. Such dynamic iterative interactions would promote rapid evolution in the genes involved in pathogen-host interactions with rapid rise in frequency of selectively advantageous alleles and high turn-over between alleles leaving insufficient time for the accumulation of neutral polymorphism [76,77]. In contrast with data in *Drosophila*, our results did not detect a pattern of directional selection in innate immunity genes of *A. gambiae*. Although it is likely that different evolutionary forces are at play in these organisms, our inconclusive results probably reflect the limited statistical power of our dataset. Indeed, evidence for a higher rate of evolution in immune related genes compared to housekeeping genes in *Drosophila* was generated through a powerful sequence dataset and the results of the tests of selection was right below the 5% statistical significance thresholds [74]. In our data, the statistical power was limited by the small number of fixed mutations detected between species, pointing towards the necessity of using a more distant outgroup than *A. arabiensis* for evolutionary studies in *A. gambiae* [24]. Suitable outgroup species should diverge sufficiently to allow powerful selection tests without reaching mutation saturation. There is evidence that members of the *A. gambiae* complex are so closely related to each other that their low level of divergence would limit the statistical power of any conventional test of selection. Moreover, genetic introgression between well established species within the complex further overshadows their phylogenetic relationships and reduces divergence time [78]. In contrast, species of the Pyretophorus series other than *A. gambiae* complex members appeared too divergent to represent appropriate outgroups [24]. The identification of a suitable outgroup for comprehensive and powerful evolutionary studies in the *An. gambiae* complex is still pending but it is likely that ongoing whole-genome sequencing efforts and increased interest in this burgeoning field will soon provide appropriate candidate species and allow revision of previously inconclusive inferences.

In *A. gambiae* as well as in *Drosophila*, the level of genetic diversity appeared to be similar between immune related genes and control genes and between functional categories of genes involved in immunity. Balancing selection does not drive the evolution of the immune system (the specific case of TEP1 in *A. gambiae* is discussed below). This is contrasting with the pattern of selection observed in vertebrates, in which genes involved in defense mechanisms are under balancing selection in addition to directional selection. In vertebrates, the system of recognition of acquired immunity requires a large diversity of major histocompatibility complex genes to bind large diversity of antigens [79]. As such, the pattern of variability in *A. gambiae* and *Drosophila* immune systems is consistent with the recognition of relatively few motifs conserved across broad ranges of pathogens. However, the recent discovery of hypervariable immunoglobulin domain-encoding genes, Dscam, capable of producing pathogen-specific splice [80,81] opens new insights into recognition system in insects suggesting specific recognition of a spectrum of pathogens. Future investigations of molecular evolution will determine the selective forces at play on such genes and will help understanding their role in pathogen recognition.

One of the immune-related genes, TEP1, showed a unique pattern of variation. It displayed the highest level of

genetic diversity among the genes we investigated and a significant positive value of the Tajima D statistic suggesting maintenance of divergent alleles. A previous study [7] revealed two highly differentiated TEP1 alleles that were initially mistaken as distinct genes, in the first version of the genome assembly. Crosses between laboratory strains showed Mendelian inheritance of these allelic forms, TEP1s and TEP1r, which are associated with two *A. gambiae* strains susceptible and refractory to *P. berghei*, respectively [7]; it was hypothesized that the alternative alleles are causally related to these phenotypes. The diversity we observed in TEP1 can be the result of balancing selection and would be reminiscent of selection for diversity in acquired immunity. However, gene conversion between different genes in the TEP family might result in similar patterns of diversity and further investigation is needed to disentangle these hypotheses. Nonetheless, our results emphasize the importance of the TEP1s and TEP1r alleles, and demonstrate these to segregate in natural populations. However, their role in determining the susceptibility of *A. gambiae* to *Plasmodium* infection remains to be established, as their segregation in laboratory strains could be due to increased genetic drift at the onset and throughout the colonization process.

Conclusion

Through the sequencing of 109 fragments of genes in *A. gambiae*, we identified 3,214 SNPs that are relevant markers for future phenotype-association studies. The pattern of genetic variability showed little evidence for maintenance of protein variation by balancing selection in *A. gambiae* immune system. It revealed strong purifying selection as the main force driving evolution of the *A. gambiae* genome, probably as a result of functional constraints for protein integrity and activity. TEP1 showed a unique pattern of genetic diversity that could be the consequence of balancing selection or gene conversion.

Methods

Mosquito populations

A. gambiae s.s. larvae were collected in Simbock (03°51'N, 11°30'E), a South Cameroon village near Yaoundé, where both molecular forms M and S are sympatric [82]. The fact that the M and S populations were collected in a single village allows measuring genetic differentiation without bias due to geographical distance between collection sites. Larvae were reared in an insectary until adult emergence. *A. arabiensis* larvae were collected in Pitoa (09°24'N, 13°30'E), in North Cameroon [83] and the offspring were maintained in the insectary for approximately 10 generations, in 26–27 degrees Celsius, relative humidity 70–80% with 12 h/12 h light dark cycle. The number of *A. arabiensis* at each generation was always more than 100 individuals, avoiding strong bottleneck and genetic drift for the given number of generations. Anophelines were identified as members of the *A. gambiae* complex using morphological keys [45,84]. Species were identified using species-specific PCR [85] and the molecular forms of *A. gambiae* were distinguished by the PCR assay of Favia et al. [86]. Eight M molecular form females, 9 of the S molecular form and 8 *A. arabiensis* were used for sequence analysis.

DNA/RNA isolation and sequencing

Coding regions are especially informative in evolutionary genetics and allow tests of selection based on comparison of synonymous (sSNPs) and non-synonymous (nsSNPs) mutations. Depending on the distribution of introns and exons, specific PCR assays for each gene were developed from coding regions of genomic DNA (gDNA) or complementary DNA (cDNA).

DNA was isolated from legs of adult females as described [21] and amplified with the Genomiphi kit (GE Healthcare, UK). This procedure conserves DNA polymorphism and does not alter SNP detection [87]. RNA was isolated from the same individuals (entire mosquitoes minus legs) by Trizol reagent (Invitrogen). After DNase I treatment, total RNA was reverse transcribed using the Superscript II kit (Life Technologies).

PCR assays were developed to amplify all or part of the coding regions of studied genes. Data on DNA sequence, genomic position and known polymorphism were obtained from the ENSEMBL website [88]. PCR primers were designed using Primer3 [89]. PCR reactions were performed in 50 μl solution containing 20 pmol of each primer, each dNTP at 0.2 mM, 2.5 mM $MgCl_2$, 10 mM Tris-HCl (pH 8.3), 50 mM potassium chloride (KCl), 2 units of Taq polymerase and approximately 10 ng of template DNA. Amplification conditions included an initial 5 min 94°C denaturation, followed by 12 cycles at 94°C for 30 s, 65°C for 30 s, with a decrease of one degree per cycle, and finally 72°C for 1 min 30 s. They were followed by 25 cycles of 94°C for 30 s, 56°C for 30 s, and 72°C for 1 min 30 s. A final 72°C extension step lasted 10 min. The excess dNTPs were digested with Shrimp Alkaline Phopshatase and primers with ExonucleaseI (United State Biochemicals). Both strands were sequenced using the Bigdye terminator v3.1 cycle sequencing kit (Applied Biosystems) and an Applied Biosystems 3730 sequencer. Sequences were assembled and verified using SeqScape (Applied Biosystems).

Data analysis

Sequence alignments were performed using the ClustalW included in MEGA 3.1 [90]. Non-coding regions were removed from analysis. Calculations were carried out after elimination of alignment gaps. Polymorphism analyses

and molecular population genetic test statistics were calculated using DnaSP 4.10 [91] and MEGA 3.1.

For each population (hereafter *A. arabiensis*, *A. gambiae* M, *A. gambiae* S), we calculated the numbers of segregating sites (SS), informative segregating sites (ISS: polymorphisms found more than once in the dataset), and nsSNPs. Nucleotide diversity was estimated as the average pairwise nucleotide difference per site, considering all sites (π), synomymous sites only (πs) or non-synonymous sites only (πns) [92] and from the proportion of segregating sites θw [93]. Genetic diversity estimates were compared between groups of genes using the Mann-Whitney U test. Estimates derived from X-linked genes were adjusted for their lower effective population size by multiplying estimates by 4/3 because male mosquitoes carry only one copy of the X chromosome [e.g. [29]].

Divergence between species and genetic differentiation between *A. gambiae* molecular forms was assessed by sequence-based *F* statistics (Fst) analogous to Wright *F* statistics [94], calculated according to Hudson *et al* [95]. Levels of genetic divergence in groups of genes were compared using the Mann-Whitney U test. P values of average Fst across several loci were calculated by Fisher's method.

Molecular signatures of selection were searched using various statistical tests. The Tajima's D statistics [71] compares two estimators of genetic diversity, one based on the average number of differences between all pairs of sequences sampled (θ_π) and the other based on the total number of polymorphic sites observed (θ_S). If the population is at mutation-drift equilibrium and polymorphism is neutral, both estimators should be equal and the test statistic is zero. However, under selection or non-equilibrium, the two estimators will differ, and this difference reflects the mode of selection or the direction of change in population size. While θ_S is only influenced by the number of segregating sites in the dataset, π is sensitive to allele frequencies at segregating sites, such that alleles at intermediate frequencies contribute more than alleles at low frequencies. Consequently, if a sample has an excess of rare variants (as a consequence of purifying selection or of population growth), θ_π would be less than θ_S and the statistic is negative. In contrast, if there is an excess of alleles at intermediate frequency (e.g. balancing selection or population bottleneck), Tajima's D statistic will be positive [70,71]. Comparison of the pattern observed across multiple independent genes allows distinguish locus-specific effects of selection from genome-wide patterns attributable to demographic changes. To avoid a possible bias due to mildly deleterious alleles towards low frequency variants, Tajima'D tests were computed using silent sites only.

To detect positive Darwinian selection (directional selection), we compared the number of synonymous substitutions per synonymous site (dS) and the number of non-synonymous substitutions per non-synonymous site (dN) [76] using bootstrapping in MEGA 3.1 (Z-test of selection). We took alternatively as the null hypothesis dS = dN (neutral hypothesis), dS<dN (positive selection) and dS>dN (purifying selection).

The Ka/Ks ratio compares the number of replacement substitutions per site (nsSNPs) and silent substitutions per site (sSNPs) among different populations [96,97]. This ratio is higher for genes under selection for beneficial amino acid changes. Ka/Ks ratios were calculated for each gene and for each pair of populations.

Under neutral evolution, the ratio of replacement to silent mutations that are fixed between species should equal the ratio of replacement to silent polymorphisms within species [98]. The MacDonald-Kreitman test uses a 2 × 2 contingency table to test differences in these ratios. This test could not be performed with each gene separately because, in most cases, the number of fixed polymorphisms was too low for contingency table computation. Therefore, the test was performed with the sum of fixed/polymorphic sSNPs and nsSNPs across all genes for each pair of populations, by using the 2 × 2 test of independence in DnaSP. Summing mutations across genes can lead to spurious cases of positive selection [34] the results must therefore be carefully interpreted

Authors' contributions

AC contributed to the design of the study, carried out the experiments, analyzed the data and wrote the manuscript. SK and MM, participated in the sequencing. FS and IM helped in analysis and to draft the manuscript. AK participated in the sequence alignments. DF and FCK conceived of the study, participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Characteristics of locus amplified for SNP detection in coding regions of immune related genes. **: 1: Recognition, 2: Modulation, 3: Signal transduction, 4: Effector molecules according to Christophides* et al *[5]*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-227-S1.xls]

Additional file 2

Characteristics of locus amplified for SNP detection in coding regions of non immune related genes.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-227-S2.xls]

Additional file 3

***Nucleotide polymorphism in immune related genes**. SS: number of segregating sites. ISS: number of informative sites (polymorphisms found more than once in the dataset). nsSNP: number of non synonymous SNPs. π: nucleotide diversity considering all sites. πs: nucleotide diversity considering synomymous sites only. πns: nucleotide diversity considering non-synomymous sites only. θw: the proportion of segregating site. Tajima's D test: D value, significant values are in bold characters if $P < 0.05$, and bold and underlined characters if $P < 0.01$. Z test of selection: test for purifying selection (Nei and Gojobori method). _: data not available due to a failure of sequencing or lack of polymorphism.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-227-S3.xls]

Additional file 4

***Nucleotide polymorphism in non immune related genes**. SS: number of segregating sites. ISS: number of informative sites (polymorphisms found more than once in the dataset). nsSNP: number of non synonymous SNPs. π: nucleotide diversity considering all sites. πs: nucleotide diversity considering synomymous sites only. πns: nucleotide diversity considering non-synomymous sites only. θw: the proportion of segregating site. Tajima's D test: D value, significant values are in bold characters if $P < 0.05$, and bold and underlined characters if $P < 0.01$. Z test of selection: test for purifying selection (Nei and Gojobori method). _: data not available due to a failure of sequencing or lack of polymorphism.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-227-S4.xls]

Additional file 5

***Divergence between populations detected on immune related genes**. Polymorphic sites: number of polymorphic sites within populations. Fixed sites: number of fixed divergence between populations. _: data not available due to a failure of sequencing.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-227-S5.xls]

Additional file 6

***Divergence between populations detected on non immune related genes**. Polymorphic sites: number of polymorphic sites within populations. Fixed sites: number of fixed divergence between populations. _: data not available due to a failure of sequencing.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-9-227-S6.xls]

Acknowledgements

We thank George Christophides for stimulating discussions. This investigation received financial support from the UNICEF/UNDP/World Bank/WHO Special Program for Research and Training in Tropical Diseases (TDR), grant AE0241.

References

1. WHO: **2005 world malaria report.** [http://www.rollbackmalaria.org/wmr2005/].
2. Christophides GK, Vlachou D, Kafatos FC: **Comparative and functional genomics of the innate immune system in the malaria vector Anopheles gambiae.** *Immunol Rev* 2004, **198:**127-148.
3. Vlachou D, Schlegelmilch T, Christophides GK, Kafatos FC: **Functional genomic analysis of midgut epithelial responses in Anopheles during Plasmodium invasion.** *Curr Biol* 2005, **15(13):**1185-1195.
4. Oduol F, Xu J, Niare O, Natarajan R, Vernick KD: **Genes identified by an expression screen of the vector mosquito Anopheles gambiae display differential molecular immune response to malaria parasites and bacteria.** *Proc Natl Acad Sci U S A* 2000, **97(21):**11397-11402.
5. Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S, Blass C, Brey PT, Collins FH, Danielli A, Dimopoulos G, Hetru C, Hoa NT, Hoffmann JA, Kanzok SM, Letunic I, Levashina EA, Loukeris TG, Lycett G, Meister S, Michel K, Moita LF, Muller HM, Osta MA, Paskewitz SM, Reichhart JM, Rzhetsky A, Troxler L, Vernick KD, Vlachou D, Volz J, von Mering C, Xu J, Zheng L, Bork P, Kafatos FC: **Immunity-related genes and gene families in Anopheles gambiae.** *Science* 2002, **298(5591):**159-165.
6. Dimopoulos G, Christophides GK, Meister S, Schultz J, White KP, Barillas-Mury C, Kafatos FC: **Genome expression analysis of Anopheles gambiae: responses to injury, bacterial challenge, and malaria infection.** *Proc Natl Acad Sci U S A* 2002, **99(13):**8814-8819.
7. Blandin S, Shiao SH, Moita LF, Janse CJ, Waters AP, Kafatos FC, Levashina EA: **Complement-like protein TEP1 is a determinant of vectorial capacity in the malaria vector Anopheles gambiae.** *Cell* 2004, **116(5):**661-670.
8. Osta MA, Christophides GK, Kafatos FC: **Effects of mosquito genes on Plasmodium development.** *Science* 2004, **303(5666):**2030-2032.
9. Frolet C, Thoma M, Blandin S, Hoffmann JA, Levashina EA: **Boosting NF-kappaB-dependent basal immunity of Anopheles gambiae aborts development of Plasmodium berghei.** *Immunity* 2006, **25(4):**677-685.
10. Volz J, Muller HM, Zdanowicz A, Kafatos FC, Osta MA: **A genetic module regulates the melanization response of Anopheles to Plasmodium.** *Cell Microbiol* 2006, **8(9):**1392-1405.
11. Collins FH, Sakai RK, Vernick KD, Paskewitz S, Seeley DC, Miller LH, Collins WE, Campbell CC, Gwadz RW: **Genetic selection of a Plasmodium-refractory strain of the malaria vector Anopheles gambiae.** *Science* 1986, **234(4776):**607-610.
12. Vernick KD, Fujioka H, Seeley DC, Tandler B, Aikawa M, Miller LH: **Plasmodium gallinaceum: a refractory mechanism of ookinete killing in the mosquito, Anopheles gambiae.** *Exp Parasitol* 1995, **80(4):**583-595.
13. Zheng L, Cornel AJ, Wang R, Erfle H, Voss H, Ansorge W, Kafatos FC, Collins FH: **Quantitative trait loci for refractoriness of Anopheles gambiae to Plasmodium cynomolgi B.** *Science* 1997, **276(5311):**425-428.
14. Zheng L, Wang S, Romans P, Zhao H, Luna C, Benedict MQ: **Quantitative trait loci in Anopheles gambiae controlling the encapsulation response against Plasmodium cynomolgi Ceylon.** *BMC Genet* 2003, **4:**16.
15. Thomasova D, Ton LQ, Copley RR, Zdobnov EM, Wang X, Hong YS, Sim C, Bork P, Kafatos FC, Collins FH: **Comparative genomic analysis in the region of a major Plasmodium-refractoriness locus of Anopheles gambiae.** *Proc Natl Acad Sci U S A* 2002, **99(12):**8179-8184.
16. Niare O, Markianos K, Volz J, Oduol F, Toure A, Bagayoko M, Sangare D, Traore SF, Wang R, Blass C, Dolo G, Bouare M, Kafatos FC, Kruglyak L, Toure YT, Vernick KD: **Genetic loci affecting resistance to human malaria parasites in a West African mosquito vector population.** *Science* 2002, **298(5591):**213-216.
17. Riehle MM, Markianos K, Niare O, Xu J, Li J, Toure AM, Podiougou B, Oduol F, Diawara S, Diallo M, Coulibaly B, Ouatara A, Kruglyak L, Traore SF, Vernick KD: **Natural malaria infection in Anopheles gambiae is regulated by a single genomic control region.** *Science* 2006, **312(5773):**577-579.
18. Luckhart S, Li K, Dunton R, Lewis EE, Crampton AL, Ryan JR, Rosenberg R: **Anopheles gambiae immune gene variants associated**

with natural Plasmodium infection. *Mol Biochem Parasitol* 2003, **128(1):**83-86.
19. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O'Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, Zhao S, Zhu SC, Zhimulev I, Coluzzi M, della Torre A, 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 malaria mosquito Anopheles gambiae.** *Science* 2002, **298(5591):**129-149.
20. Simard F, Licht M, Besansky NJ, Lehmann T: **Polymorphism at the defensin gene in the Anopheles gambiae complex: testing different selection hypotheses.** *Infect Genet Evol* 2007, **7:**285-292.
21. Morlais I, Poncon N, Simard F, Cohuet A, Fontenille D: **Intraspecific nucleotide variation in Anopheles gambiae: new insights into the biology of malaria vectors.** *Am J Trop Med Hyg* 2004, **71(6):**795-802.
22. Little TJ, Cobbe N: **The evolution of immune-related genes from disease carrying mosquitoes: diversity in a peptidoglycan- and a thioester-recognizing protein.** *Insect Mol Biol* 2005, **14(6):**599-605.
23. Dassanayake RS, Silva Gunawardene YI, Tobe SS: **Evolutionary selective trends of insect/mosquito antimicrobial defensin peptides containing cysteine-stabilized alpha/beta motifs.** *Peptides* 2007, **28(1):**62-75.
24. Obbard DJ, Linton YM, Jiggins FM, Yan G, Little TJ: **Population genetics of Plasmodium resistance genes in Anopheles gambiae: no evidence for strong selection.** *Mol Ecol* 2007, **16(16):**3497-3510.
25. Hume JC, Lyons EJ, Day KP: **Human migration, mosquitoes and the evolution of Plasmodium falciparum.** *Trends Parasitol* 2003, **19(3):**144-149.
26. Tishkoff SA, Verrelli BC: **Patterns of human genetic diversity: implications for human evolutionary history and disease.** *Annu Rev Genomics Hum Genet* 2003, **4:**293-340.
27. Volkman SK, Sabeti PC, DeCaprio D, Neafsey DE, Schaffner SF, Milner DA Jr., Daily JP, Sarr O, Ndiaye D, Ndir O, Mboup S, Duraisingh MT, Lukens A, Derr A, Stange-Thomann N, Waggoner S, Onofrio R, Ziaugra L, Mauceli E, Gnerre S, Jaffe DB, Zainoun J, Wiegand RC, Birren BW, Hartl DL, Galagan JE, Lander ES, Wirth DF: **A genome-wide map of diversity in Plasmodium falciparum.** *Nat Genet* 2007, **39(1):**113-119.
28. Feuk L, Carson AR, Scherer SW: **Structural variation in the human genome.** *Nat Rev Genet* 2006, **7(2):**85-97.
29. **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms: Nature.** *Volume 409.* ; 2001:928-933.
30. Black WC, Baer CF, Antolin MF, DuTeau NM: **Population genomics: genome-wide sampling of insect populations.** *Annu Rev Entomol* 2001, **46:**441-469.
31. Berger J, Suzuki T, Senti KA, Stubbs J, Schaffner G, Dickson BJ: **Genetic mapping with SNP markers in Drosophila.** *Nat Genet* 2001, **29(4):**475-481.
32. Lazarus R, Vercelli D, Palmer LJ, Klimecki WJ, Silverman EK, Richter B, Riva A, Ramoni M, Martinez FD, Weiss ST, Kwiatkowski DJ: **Single nucleotide polymorphisms in innate immunity genes: abundant variation and potential role in complex human disease.** *Immunol Rev* 2002, **190:**9-25.
33. Nei M, Kumar S: **Molecular Evolution and Phylogenetics.** Oxford University Press; 2000.
34. Bierne N, Eyre-Walker A: **The genomic rate of adaptive amino acid substitution in Drosophila.** *Mol Biol Evol* 2004, **21(7):**1350-1360.
35. Andolfatto P: **Contrasting patterns of X-linked and autosomal nucleotide variation in Drosophila melanogaster and Drosophila simulans.** *Mol Biol Evol* 2001, **18(3):**279-290.
36. Begun DJ, Whitley P: **Reduced X-linked nucleotide polymorphism in Drosophila simulans.** *Proc Natl Acad Sci U S A* 2000, **97(11):**5960-5965.
37. Jiggins FM, Hurst GD: **The evolution of parasite recognition genes in the innate immune system: purifying selection on Drosophila melanogaster peptidoglycan recognition proteins.** *J Mol Evol* 2003, **57(5):**598-605.
38. Moriyama EN, Powell JR: **Intraspecific nuclear DNA variation in Drosophila.** *Mol Biol Evol* 1996, **13(1):**261-277.
39. Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Charlesworth B, Keightley PD: **Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila.** *Nature* 2007, **445(7123):**82-85.
40. Agrafioti I, Stumpf MP: **SNPSTR: a database of compound microsatellite-SNP markers.** *Nucleic Acids Res* 2007, **35(Database issue):**D71-5.
41. Nachman MW, Bauer VL, Crowell SL, Aquadro CF: **DNA variability and recombination rates at X-linked loci in humans.** *Genetics* 1998, **150(3):**1133-1141.
42. Schug MD, Hutter CM, Wetterstrand KA, Gaudette MS, Mackay TF, Aquadro CF: **The mutation rates of di-, tri- and tetranucleotide repeats in Drosophila melanogaster.** *Mol Biol Evol* 1998, **15(12):**1751-1760.
43. Lehmann T, Hawley WA, Grebert H, Collins FH: **The effective population size of Anopheles gambiae in Kenya: implications for population structure.** *Mol Biol Evol* 1998, **15(3):**264-276.
44. Stump AD, Fitzpatrick MC, Lobo NF, Traore S, Sagnon N, Costantini C, Collins FH, Besansky NJ: **Centromere-proximal differentiation and speciation in Anopheles gambiae.** *Proc Natl Acad Sci U S A* 2005, **102(44):**15930-15935.
45. Gillies MT, De Meillon B: **The Anophelinae of Africa South of the Sahara (Ethiopian zoogeographical region).** 2° Ed. edition. Johannesburg, South Africa , The South African Institute for Medical Research; 1968:343.
46. della Torre A, Tu Z, Petrarca V: **On the distribution and genetic differentiation of Anopheles gambiae s.s. molecular forms.** *Insect Biochem Mol Biol* 2005, **35(7):**755-769.
47. Marais G, Charlesworth B: **Genome evolution: recombination speeds up adaptive evolution.** *Curr Biol* 2003, **13(2):**R68-70.
48. Lercher MJ, Hurst LD: **Human SNP variability and mutation rate are higher in regions of high recombination.** *Trends Genet* 2002, **18(7):**337-340.
49. Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V: **A polytene chromosome analysis of the Anopheles gambiae species complex.** *Science* 2002, **298(5597):**1415-1418.
50. Coluzzi M, Sabatini A, Petrarca V, Di Deco MA: **Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex.** *Trans R Soc Trop Med Hyg* 1979, **73(5):**483-497.
51. Coluzzi M: **Spatial distribution of chromosomal inversions and speciation in Anopheline mosquitoes.** *Prog Clin Biol Res* 1982, **96:**143-153.
52. della Torre A, Costantini C, Besansky NJ, Caccone A, Petrarca V, Powell JR, Coluzzi M: **Speciation within Anopheles gambiae--the glass is half full.** *Science* 2002, **298(5591):**115-117.
53. Noor MA, Grams KL, Bertucci LA, Reiland J: **Chromosomal inversions and the reproductive isolation of species.** *Proc Natl Acad Sci U S A* 2001, **98(21):**12084-12088.
54. Rieseberg LH: **Chromosomal rearrangements and speciation.** *Trends Ecol Evol* 2001, **16(7):**351-358.
55. Kirkpatrick M, Barton N: **Chromosome inversions, local adaptation and speciation.** *Genetics* 2006, **173(1):**419-434.
56. Coluzzi M, Di Deco MA, Petrarca V: **Chromosomal inversion intergradation and incipient speciation in Anophleles gambiae.** *Boll Zool* 1985, **52:**45-63.
57. Navarro A, Barton NH: **Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation.** *Evolution Int J Org Evolution* 2003, **57(3):**447-459.

58. Navarro A, Barbadilla A, Ruiz A: **Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in Drosophila.** *Genetics* 2000, **155(2):**685-698.
59. Slotman M, Della Torre A, Powell JR: **The genetics of inviability and male sterility in hybrids between Anopheles gambiae and An. arabiensis.** *Genetics* 2004, **167(1):**275-287.
60. Wondji C, Frederic S, Petrarca V, Etang J, Santolamazza F, Della Torre A, Fontenille D: **Species and populations of the Anopheles gambiae complex in Cameroon with special emphasis on chromosomal and molecular forms of Anopheles gambiae s.s.** *J Med Entomol* 2005, **42(6):**998-1005.
61. Wondji C, Simard F, Fontenille D: **Evidence for genetic differentiation between the molecular forms M and S within the Forest chromosomal form of Anopheles gambiae in an area of sympatry.** *Insect Mol Biol* 2002, **11(1):**11-19.
62. Turner TL, Hahn MW, Nuzhdin SV: **Genomic islands of speciation in Anopheles gambiae.** *PLoS Biol* 2005, **3(9):**e285.
63. Stump AD, Shoener JA, Costantini C, Sagnon N, Besansky NJ: **Sex-linked differentiation between incipient species of Anopheles gambiae.** *Genetics* 2005, **169(3):**1509-1519.
64. Pombi M, Stump AD, Della Torre A, Besansky NJ: **Variation in recombination rate across the X chromosome of Anopheles gambiae.** *Am J Trop Med Hyg* 2006, **75(5):**901-903.
65. Mukabayire O, Besansky NJ: **Distribution of T1, Q, Pegasus and mariner transposable elements on the polytene chromosomes of PEST, a standard strain of Anopheles gambiae.** *Chromosoma* 1996, **104(8):**585-595.
66. Turner TL, Hahn MW: **Locus- and population-specific selection and differentiation between incipient species of Anopheles gambiae.** *Mol Biol Evol* 2007, **24(9):**2132-2138.
67. Donnelly MJ, Licht MC, Lehmann T: **Evidence for recent population expansion in the evolutionary history of the malaria vectors Anopheles arabiensis and Anopheles gambiae.** *Mol Biol Evol* 2001, **18(7):**1353-1364.
68. Donnelly MJ, Simard F, Lehmann T: **Evolutionary studies of malaria vectors.** *Trends Parasitol* 2002, **18(2):**75-80.
69. Fu YX, Li WH: **Statistical tests of neutrality of mutations.** *Genetics* 1993, **133(3):**693-709.
70. Tajima F: **The effect of change in population size on DNA polymorphism.** *Genetics* 1989, **123(3):**597-601.
71. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123(3):**585-595.
72. Jiggins FM, Kim KW: **A screen for immunity genes evolving under positive selection in Drosophila.** *J Evol Biol* 2007, **20(3):**965-970.
73. Lazzaro BP, Clark AG: **Molecular population genetics of inducible antibacterial peptide genes in Drosophila melanogaster.** *Mol Biol Evol* 2003, **20(6):**914-923.
74. Schlenke TA, Begun DJ: **Natural selection drives Drosophila immune system evolution.** *Genetics* 2003, **164(4):**1471-1480.
75. Dawkins R, Krebs JR: **Arms races between and within species.** *Proc R Soc Lond B Biol Sci* 1979, **205(1161):**489-511.
76. Yang Z, Bielawski JP: **Statistical methods for detecting molecular adaptation.** *Trends in Ecology and Evolution* 2000, **15(12):**496-503.
77. Ford MJ: **Applications of selective neutrality tests to molecular ecology.** *Mol Ecol* 2002, **11(8):**1245-1262.
78. Besansky NJ, Krzywinski J, Lehmann T, Simard F, Kern M, Mukabayire O, Fontenille D, Toure Y, Sagnon N: **Semipermeable species boundaries between Anopheles gambiae and Anopheles arabiensis: evidence from multilocus DNA sequence variation.** *Proc Natl Acad Sci U S A* 2003, **100(19):**10818-10823.
79. Hughes AL, Yeager M: **Natural selection at major histocompatibility complex loci of vertebrates.** *Annu Rev Genet* 1998, **32:**415-435.
80. Dong Y, Taylor HE, Dimopoulos G: **AgDscam, a hypervariable immunoglobulin domain-containing receptor of the Anopheles gambiae innate immune system.** *PLoS Biol* 2006, **4(7):**e229.
81. Watson FL, Puttmann-Holgado R, Thomas F, Lamar DL, Hughes M, Kondo M, Rebel VI, Schmucker D: **Extensive diversity of Ig-superfamily proteins in the immune system of insects.** *Science* 2005, **309(5742):**1874-1878.
82. Antonio-Nkondjio C, Awono-Ambene P, Toto JC, Meunier JY, Zebaze-Kemleu S, Nyambam R, Wondji CS, Tchuinkam T, Fontenille D: **High malaria transmission intensity in a village close to Yaounde, the capital city of Cameroon.** *J Med Entomol* 2002, **39(2):**350-355.
83. Chouaibou M, Simard F, Chandre F, Etang J, Darriet F, Hougard JM: **Efficacy of bifenthrin-impregnated bednets against Anopheles funestus and pyrethroid-resistant Anopheles gambiae in North Cameroon.** *Malar J* 2006, **5:**77.
84. Gillies MT, Coetzee M: **A supplement to the Anophelinae of Africa south of the Sahara.** Johannesburg , The South African Institute for Medical Research; 1987:143.
85. Scott JA, Brogdon WG, Collins FH: **Identification of single specimens of the Anopheles gambiae complex by the polymerase chain reaction.** *Am J Trop Med Hyg* 1993, **49(4):**520-529.
86. Favia G, Lanfrancotti A, Spanos L, Siden-Kiamos I, Louis C: **Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of Anopheles gambiae s.s.** *Insect Mol Biol* 2001, **10(1):**19-23.
87. Holbrook JF, Stabley D, Sol-Church K: **Exploring whole genome amplification as a DNA recovery tool for molecular genetic studies.** *J Biomol Tech* 2005, **16(2):**125-133.
88. **ENSEMBL** [http://www.ensembl.org/Anopheles_gambiae/]
89. **Primer3** [http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi]
90. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.** *Brief Bioinform* 2004, **5(2):**150-163.
91. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R: **DnaSP, DNA polymorphism analyses by the coalescent and other methods.** *Bioinformatics* 2003, **19(18):**2496-2497.
92. Nei M: **Molecular evolutionary genetics.** New York , Columbia University Press; 1987.
93. Watterson GA: **On the number of segregating sites in genetical models without recombination.** *Theor Popul Biol* 1975, **7(2):**256-276.
94. Wright S: **Evolution and the Genetics of Populations: Variability Within and Among natural Populations.** Chicago , University Press; 1978.
95. Hudson RR, Slatkin M, Maddison WP: **Estimation of levels of gene flow from DNA sequence data.** *Genetics* 1992, **132(2):**583-589.
96. Li WH, Wu CI, Luo CC: **A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes.** *Mol Biol Evol* 1985, **2(2):**150-174.
97. Hurst LD: **The Ka/Ks ratio: diagnosing the form of sequence evolution.** *Trends Genet* 2002, **18(9):**486-487.
98. McDonald JH, Kreitman M: **Adaptive protein evolution at the Adh locus in Drosophila.** *Nature* 1991, **351(6328):**652-654.