WHITE PAPER

Proposal for the Eight Genomes Cluster for Genus *Anopheles*

July 2004; revised March 2005

Nora J. Besansky on behalf of the Eight Genomes Cluster Committee[1]

I. Introduction

When the first draft of the *Anopheles gambiae* genome sequence was announced in October 2002 (Holt *et al.*, 2002), it was only the second insect to have its genome completely sequenced after *Drosophila melanogaster*. This species was chosen because of its undisputed biomedical importance as a vector of malaria, a disease that kills up to 3 million people a year, mainly in Africa where *An. gambiae* is its principal vector. As efforts to sequence the genomes of other arthropod vectors of diseases such as dengue, filariasis, West Nile and Lyme Disease are at various stages of planning or progress, it is pertinent to ask why funding agencies should invest more resources to sequence the genomes of additional anopheline malaria vectors. The answer lies in the power of comparative genomics, given that comparisons are made at the appropriate evolutionary depths. The argument developed in this proposal is twofold: comparative genomics of anopheline species is essential both for advancing the currently incomplete structural and functional annotation of the *An. gambiae* genome, and for advancing our knowledge about the modifications of genome sequence and organization that allowed a handful of anopheline species to evolve rapidly into the most dangerous vectors by specializing on humans.

II. Background

A. Most mosquitoes are not human malaria vectors: There are ~3500 species of mosquitoes (Diptera: Culicidae) classified into two subfamilies: Anophelinae and Culicinae (into which the previously recognized subfamily Toxorhynchitinae was sunk). Culicinae contains competent vectors of West Nile, dengue, filariasis, and even non-human malaria parasites, but only Anophelinae contains species capable of transmitting human malaria parasites. The absolute barrier to infection within culicine mosquitoes is entirely unknown, but is certainly physiological rather than behavioral. Despite repeated exposure and therefore evolutionary opportunity, not even culicine species with anthropophilic tendencies (*e.g.*, *Aedes aegypti* and members of the *Culex pipiens* complex) are competent vectors of *Plasmodium* parasites affecting humans. While ability to vector human malaria is uniquely present within genus *Anopheles*, it is quite rare: only ~30 of almost 500 species are major vectors of human malaria (Collins, Paskewitz, 1995). In the case of *Anopheles*, key differences between vectors and non-vectors are often behavioral rather than physiological (*e.g.*, Takken *et al.*, 1999).

B. Most malaria vectors are not closely related: If the ability to vector human malaria is conferred by a unique suite of traits that arose once in the distant past and if these traits are evolutionarily stable, one would expect that all malaria vectors would branch off from the same limb of a phylogenetic tree. Contrary to that pattern, malaria vectors within *Anopheles* are distributed across the entire tree, interspersed with non-vector species in four of the six *Anopheles* subgenera (Fig 1). Even considering individual subgenera, component malaria vectors are not phylogenetically clustered.

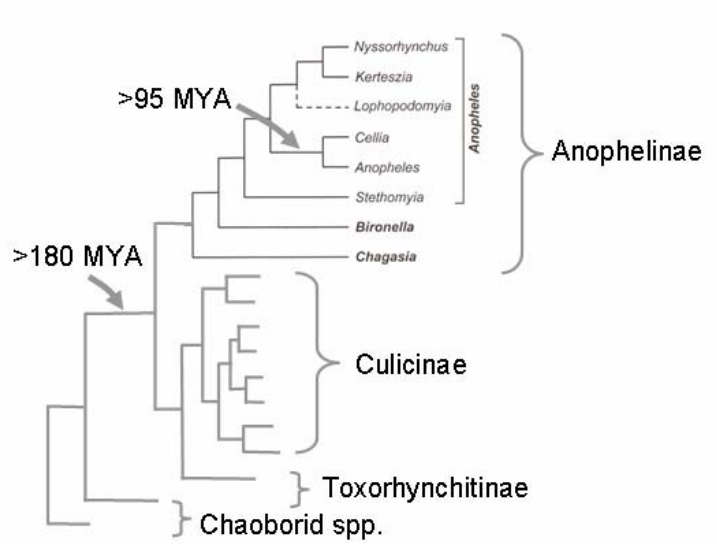

Figure 1. Composite phylogeny of Culicidae with chaoborids as an outgroup, after (Besansky, Fahey, 1997; Krzywinski *et al.*, 2001; Krzywinski, Besansky, 2003). The Culicidae stem lineage was present >180 MYA (Besansky, Fahey, 1997). Within Subfamily Anophelinae, the three genera (boldface type) and six *Anopheles* subgenera are shown. Of these subgenera, all contain vectors of human malaria except *Stethomyia* and *Lophopodomyia*. The position of *Lophopodomyia* remains unsupported.

C. Vector ability has evolved repeatedly and rapidly: Among the nearest relatives to most major malaria vectors are non-vectors. Each major vector typically belongs to a sibling species complex of very closely related and morphologically similar/identical members, most of which have negligible or no role in disease transmission. For example, *An. funestus*-- second only to *An. gambiae* in importance as a malaria vector (Gillies, De Meillon, 1968)-- is one of a group of closely related and recently diverged species known as the Funestus Subgroup (Harbach, 2004). None of the other species is significantly involved in malaria transmission, because most preferentially rest outdoors and all preferentially feed on animals. Among the seven sibling species comprising the *An. gambiae* complex, *An. quadriannulatus* is of no epidemiological importance, as it is never found naturally infected with malaria parasites. Yet when this species is brought into the laboratory and given an infected blood meal, it is permissive to the complete sporogonic development of *P. falciparum*, producing infective sporozoites (Takken *et al.*, 1999). This does not happen in nature, for behavioral rather than physiological reasons; *An. quadriannulatus* prefers to feed on animals. Thus, the status of *An. gambiae* and *An. funestus* as the predominant vectors of malaria results from the fact that these species, unique among their sibling species, spend nearly all of their adult life inside houses and feed almost exclusively on humans (Gillies, Coetzee, 1987; Gillies, De Meillon, 1968). With respect to each sibling species complex that contains a primary vector of malaria, anthropophily (specialization on humans) is most parsimoniously interpreted as a recently derived suite of behaviors, for two reasons. First, almost by definition, sibling species complexes represent bursts of recent speciation. Second, as a target for specialization by these ectoparasites, human populations did not achieve the necessary density and stability until relatively recently, with the rise of agricultural communities.

One important application of the *An. gambiae* genome sequence is to locate and characterize those traits that confer high vectorial capacity, with the ultimate goal of altering or blocking their function. If the traits that we are interested in are rapidly

evolving and recently derived, comparative genomics approaches to their study will require levels of divergence far more shallow than *Anopheles-Drosophila* (~250MYA; Gaunt, Miles, 2002) or *Anopheles-Culex/Aedes* (~180MYA; Krzywinski *et al.*, 2001).

D. The sequenced *An. gambiae* genome is robust, but its utility can be greatly improved using comparative genomics:

Assembly: It is exceedingly difficult to adapt most anopheline species to laboratory culture, and is virtually impossible to derive isogenic lines, although the colonization process itself is inevitably associated with considerable inbreeding. The *An. gambiae* genome was sequenced from a laboratory strain named PEST (Pink-Eye STandard). Unlike any other available lab strain of this species, PEST had been selected to be monomorphic (homokaryotypic) for the standard arrangement of all known paracentric chromosomal inversions in *An. gambiae* (Mukabayire, Besansky, 1996). Because this was as close as possible to "inbred", and because the morphological marker (pink eyes) could serve as an indicator of contamination with other laboratory cultures, the PEST strain was chosen for sequencing. For reasons not appreciated at the time, it proved an unfortunate choice, as the subsequent scientific discovery of incipient species within *An. gambiae*, known as molecular forms M and S (della Torre *et al.*, 2002), has led to the realization that PEST is of mixed ancestry. This strain was derived from the crossing of M and S (Holt *et al*., 2002). It is likely that the hybrid background of the PEST strain accounts for the fact that the original genome assembly (8987 scaffolds spanning 278 Mbp) was significantly larger than the genome size of ~260 Mbp estimated from reassociation kinetics (Besansky, Powell, 1992). Many highly polymorphic regions may reflect the incorporation of two diverged haplotypes into adjacent locations of the assembly (Mongin *et al*., 2004).

An updated genome assembly, targeted for release in April 2005, will correct a number of errors and significantly increase the number of scaffolds whose genome location has been established by physical mapping of sequence-tagged clones (F. Collins, M. Hammond, R. Bruggner, M. Sharakhova, unpublished). In the original automated assembly, integrity was evaluated by analyzing the orientation and distance between the paired sequence reads from the opposite ends of the 10kb and 50kb plasmid clones. Distance and orientation violations accounted for 21 Mbp (~8%) of the total assembly. In the updated assembly, 261 scaffolds (~1.2 Mbp) with strong similarity to bacterial genomes will be removed. Additionally, 144 small (< ~300 kb) scaffolds (~8.2 Mbp) with high sequence similarity to larger scaffolds will be removed and displayed in an alternate haplotype track. Approximately 20 larger scaffolds (~3 Mbp) have been physically assigned map locations that were found to overlap other mapped scaffolds. Areas where adjacent regions of the genome appear to be tandemly duplicated will not be modified, but will be flagged in the Ensembl Genome Browser. A further 18 scaffolds (~6 Mbp) were newly mapped to specific chromosomal sites. Incorporating these changes, the new haploid genome assembly will be ~266 Mbp, of which ~234 Mbp (88%) is assigned to a genome location and unambiguously ordered and oriented by physical mapping of BAC clones. The remaining ~32 Mbp are currently assigned to an "unmapped" chromosome. The opportunity to sequence the individual genomes of molecular forms M and S will resolve and correct remaining aspects of the "haplotype" problem that resulted from the hybrid background of PEST.

An important feature of the *An. gambiae* genome is the direct correspondence of the physically mapped scaffolds in the genome with the polytene chromosome complement. More than 2,000 sequence-tagged clones, primarily BACs and cDNAs, have been physically mapped by *in situ* hybridization to specific bands on the Coluzzi map (Coluzzi *et al.*, 2002). The April 2005 genome update will include a field that physically associates landmark regions of the genome sequence with specific bands on the polytene map. This information will be especially valuable in the analysis of genome features such as polymorphic chromosomal inversions.

Annotation: The annotation of the *An. gambiae* genome as reported by Holt *et al*. (2002) was only a first approximation, as it depended almost entirely upon automated *ab initio* gene-finding algorithms known to be imperfect. No closely related organism has been annotated comprehensively, then or since. Not only is it likely that a significant fraction of the >14,000 predicted genes have inaccurately predicted structures, but it is estimated that 20% of the genes escaped prediction altogether (Holt *et al.*, 2002; Mongin *et al.*, 2004). While both manual curation and automated gene prediction will be aided by a set of >200,000 *An. gambiae* cDNAs sequenced to date, significant improvements in annotation will come only from more full-length cDNAs and comparative genomic data from additional anopheline genomes.

III. Benefits of Additional Anopheline Genomes

We propose the genomic sequencing of four species very closely related to *An. gambiae* (within the same sibling species complex, including M and S forms) as a first step, followed by four additional species representing each of the three subgenera of *Anopheles* that contain most of the major malaria vectors (*Cellia*, *Anopheles* and *Nyssorhynchus*; details provided in section IV). Among other important applications, the first step will resolve the haplotype issue within the existing *An. gambiae* genome assembly; later steps will be anchored on this verified *An. gambiae* genome assembly. The later steps will present no obvious chromosomal assignment problems during assembly, even for the most distantly related of the subgenera to be considered, *Nyssorhynchus* (see Fig 1). Various authors have studied linkage conservation in mosquitoes, and have concluded that it is highly conserved across the whole family (Matthews, Munsterman, 1994; Cornel, Collins, 2000; Sharakhov *et al*., 2002; Severson *et al*., 2004). “Genes travel together” (Matthews, Munsterman, 1994) because the integrity of whole arms is conserved; there is little if any evidence for partial arm translocations or pericentric inversions. Indeed, syntenic relationships were studied between *An. gambiae* and *An. albimanus* (subgenus *Nyssorhynchus*) by Cornel and Collins (2000), using *in situ* cross-hybridization to polytene chromosomes. Consistent with the conclusions of Matthews and Munsterman (1994), rearrangements were limited to whole arm translocations or-- much more commonly—changes in linear gene order in the form of numerous paracentric inversions. A caveat is that all these studies were performed at low resolution. Though we consider it unlikely that the conclusions will change with higher resolution, particularly in light of conserved synteny between *Ae. aegypti* and *An. gambiae* (Severson *et al*., 2004), a straightforward check on quality control is available for all candidate species through physical mapping to polytene

chromosomes. Each has polytene chromosomes favorable for spreading, from which photomaps and/or hand-drawn cytogenetic maps have been constructed.

A. Advancing the assembly and annotation of the *An. gambiae* genome: The sequenced PEST genome is an undefined mosaic of molecular forms M and S, a fact that undoubtedly contributed to elevated levels of polymorphism in all or part of the genome. Sequencing M and S genomes individually will correct this problem and allow discrimination between allelic variation within species/incipient species versus haplotype variation between them. Not only will this deliver higher quality end-products, but also the comparison of these two genomes is expected to reveal changes between M and S forms potentially involved in their differentiation.

Toward the goal of accurate genome annotation, full-length cDNAs are an important resource. However, on their own, full-length cDNAs from *An. gambiae* will not be sufficient to complete the process of gene discovery, nor can they address the challenge of functional annotation. Not only will the use of comparative genomics help bridge this gap, it also offers additional benefits that cDNA resources alone cannot: identification of functional noncoding elements, especially regulatory elements. More distant phylogenetic comparisons among subgenera within the monophyletic and relatively ancient (95 MY?) genus *Anopheles* will serve this end. Among the most eloquent testaments to the power of comparative genomics for improving genome annotation, in particular the power of multiple rather than pairwise genome comparisons (phylogenetic "footprinting" and "shadowing", Boffelli *et al.*, 2003; Gumucio *et al.*, 1992; Tagle *et al.*, 1988), was the impact of this approach on one of the best studied and simplest of eukaryotic genomes, *Saccharomyces cerevisiae* (Cliften *et al.*, 2003; Kellis *et al.*, 2003). Phylogenetic footprinting involving an additional 3-5 related yeast species ultimately affected the annotation of up to 15% of all genes and resulted in the discovery of 43 previously unannotated genes and the elimination of ~500 previously annotated genes determined to be false positives. Furthermore, this analysis roughly doubled the catalog of regulatory elements and provided insights into their interactions (Kellis *et al.*, 2003). Nor has this message been lost on the community of scientists interpreting the human genome sequence through comparative genomics of multiple vertebrate species (*e.g.*, Thomas *et al.*, 2003). Identification and study of regulatory elements may be especially important in understanding how *An. gambiae* and *An. funestus* acquired high vectorial capacity while their sibling species did not. Evidence suggests that phenotypic changes between closely related species often result from altered pattern and level of gene expression rather than changes in coding sequence per se (Barbash *et al.*, 2003; Michalak, Noor, 2003; Stern, 1998; Sucena, Stern, 2000).

B. Identification of genetic mechanisms responsible for differences in vectorial capacity: Vectorial capacity can be defined as the average number of potentially infective bites per person per day in the host population (Garrett-Jones, 1964; Garrett-Jones, Shidrawi, 1969). It depends upon vector competence, longevity, host preference, and density in relation to the host. The component traits of vectorial capacity all have genetic bases, and as seen from the examples below, it is likely that these traits represent adaptive differences between species or incipient species. Any adaptive differences at the behavioral or physiological level should be reflected in corresponding

differences at the sequence level. Bioinformatics and statistical tools to scan genomes for these differences can be used to identify the underlying genes, and will lay the groundwork for dissecting the mechanisms responsible for differences in vectorial capacity and ultimately for devising targeted strategies to reduce it.

Genes responsible for adaptive differences should show evidence of diversifying selection (adaptive molecular evolution). It is possible to detect the signature of adaptive molecular evolution in pairwise comparisons, but methods that analyze sequences from multiple species within a phylogenetic framework are far more powerful (Yang, Nielsen, 2002). Genes responsible for adaptive differences also are known to evolve very rapidly (e.g., Swanson *et al.*, 2001a; Swanson *et al.*, 2001b), and therefore detecting adaptive molecular evolution is facilitated by comparisons over relatively short divergence times.

Ecological adaptations and vectorial capacity. From the point of view of vector population biology, the ecotypic differentiation characteristic of vector species is a mechanism which not only exploits the environment more fully but also reduces competition for limiting resources, which should increase the fitness parameters of vectorial capacity: longevity and population density (Coluzzi *et al.*, 2002). The strategy of ecotypic differentiation has been used to great effect in *An. gambiae*, which is not a generalist species but rather a collection of specialists (M. Coluzzi, pers. comm). A good example may be the putative habitat shift by the M form, which has clearly increased malaria transmission across time and space because the M form efficiently exploits a novel resource-- ricefields and other irrigated agricultural zones-- relative to the traditional rain-dependent breeding sites of the S form. No candidate genes have been identified a priori, but it is expected that whole genome comparisons will identify unusually divergent sequences that could underlie this ecological specialization (*e.g.*, Stump *et al*., 2005).

Host preference and vectorial capacity. Adult female anopheline mosquitoes transmit malaria while bloodfeeding. Evidence from field studies suggests that species-specific differences in bloodfeeding behavior are mediated by differential responses to host odors (reviewed in Filchak *et al*., 2005). The importance of a given species as a malaria vector depends upon its affinity and specificity for human hosts. In choice tests of host preference between human and calf, carried out in the field using odor-baited entry traps, both *An. gambiae* and *An. funestus* show an overwhelming preference for human odor. These vectors are rare examples that have restricted their host range to one species: humans. Yet *An. gambiae's* closest relatives have very different host preferences. For example, its sibling species *An. arabiensis* is an opportunistic feeder which feeds avidly on humans but whose bites can be readily diverted to nearby domestic animals. Even more surprising, its sibling species *An. quadriannulatus* prefers animals and does not normally bite humans. These differential responses to host odor, from attraction at a distance to landing and biting, may depend upon the presence of particular gustatory or odorant receptors (*e.g.*, Hallem *et al.*, 2004) or other olfactory genes that can be compared between different genomes for insights into the genetics of host preference.

Vector competence and vectorial capacity. Genetic variation in mosquito susceptibility to malaria parasite species and strains is well-known (Collins *et al.*, 1986), though the

mechanisms controlling susceptibility are poorly understood. Variation in susceptibility is present among different mosquito species as well as among individual mosquitoes of the same species. For example, in paired feeding experiments with malaria-infected blood involving the sibling species A, B and C of *An. culicifacies* from India, species A was far more susceptible (63%, <5%, and 26% infected among fed mosquitoes; Adak *et al.*, 1999). Not only do different individuals and species exhibit varying degrees of susceptibility to the same parasite species, but also a given mosquito species has varying degrees of susceptibility to different species and isolates of malaria parasites. For example, *An. gambiae* is more susceptible to co-indigenous *P. falciparum* malaria parasites than those isolated from the New World or Asia (Collins *et al.*, 1986). This emphasizes the evolutionary arms race between malaria parasites and their anopheline vectors. In the mosquito, this could involve not only genes controlling immunity, but also gut and salivary gland receptors. In the parasite, this could involve surface ligands as well as proteases and chitinases. Some history of this arms race can be probed through parallel comparative genome analyses of multiple anopheline vectors and species of Plasmodium (PlasmoDB; http://plasmodb.org).

C. Understanding the role of inversions in adaptation and speciation:
Chromosome number is conserved across Culicidae (2N=6 with one known exception), as is synteny. By contrast, linear gene order varies owing to chromosomal inversions, especially in anophelines (Cornel, Collins, 2000; Severson *et al.*, 2001). A comparison of synteny, linear gene order and sequence conservation among orthologous genes in *An. gambiae* and *An. funestus* revealed levels of chromosomal rearrangements and silent site substitution that were comparable to estimates from Drosophila, assuming a divergence time of 5MY between these anophelines (see supporting online material in Sharakhov *et al.*, 2002 for divergence time calculations). In Drosophila and *Anopheles*, the number of inversions/Mb/MY (half the number of disruptions/Mb/MY) is estimated at 0.026 and 0.031, respectively (Gonzalez *et al.*, 2002; Sharakhov *et al.*, 2002), and the average synonymous rate (K_s) between *D. melanogaster/D. erecta* and *An. gambiae/An. funestus* is 0.36 and 0.53, respectively (Bergman *et al.*, 2002; M. Hillenmeyer, unpublished, based on a maximum likelihood method using 213 *An. funestus* ESTs and their putative *An. gambiae* orthologs). These data suggest that a high rate of chromosomal rearrangement may be correlated with a high rate of nucleotide divergence, as found in vertebrates and nematodes (Burt *et al.*, 1999; Coghlan, Wolfe, 2002). Indeed, the average nucleotide divergence based on 213 orthologous pairs from *An. gambiae* and *An. funestus* was ~16.9% (M. Hillenmeyer, unpublished). Both the high rate of chromosomal rearrangement and nucleotide divergence reinforce the notion that to identify regions of the genome that account for traits unique to anophelines, especially traits unique to human malaria vectors, sequence comparisons should involve relatively closely related species.

Given the high number of inversions fixed between *An. gambiae* and *An. funestus* since their divergence, perhaps it is not surprising that intraspecific chromosomal inversion polymorphism is a prominent feature of several anopheline species, notably *An. funestus*, *An. gambiae*, and its sibling species *An. arabiensis*. Compelling evidence from these three species suggests that inversion polymorphism is associated with ecotypic differentiation, epidemiological differentiation, and speciation (Coluzzi, 1982; Coluzzi *et*

al., 2002; Coluzzi *et al.*, 1979; Costantini *et al.*, 1999; della Torre *et al.*, 2002; Petrarca, Beier, 1992; Powell *et al.*, 1999). If inversions serve as engines driving adaptation, diversification, and speciation—with negative epidemiological consequences—it serves more than purely academic interest to advance our understanding of how and why. Comparative genomics provides the most efficient means to answer these questions, and arguably the most economical means as well, if emphasis is placed on species very closely related to *An. gambiae*.

IV. Choice of anophelines for genome sequencing

Table 1. Eight Genomes Cluster for Genus *Anopheles* (~0.25 pg/genome)

	Species	Source of DNA	Resources Available
Phase I: *An. gambiae* complex (n=4)			
Subgenus Cellia, Series Pyretophorus	*An. gambiae M*	SUA2La (MR4)	PEST genome sequence, 200K cDNAs, 2 BAC libraries
	An. gambiae S	S-GA-KIS (MR4)	
	An. arabiensis	KGB (MR4)	
	An. quadriannulatus	SKUQUA (MR4)	BAC library
Phase II: Other anophelines (n=4)			
Subgenus Cellia, Series Myzomyia	*An. funestus*	FUMOZ	1K cDNAs (157 physically mapped), BAC library, cytogenetic map
Subgenus Cellia, Series Neomyzomyia	*An. farauti*	FAR1 (MR4)	
Subgenus Anopheles	*An. quadrimaculatus*	ORLANDO (MR4)	Genetic, cytogenetic maps
Subgenus Nyssorhynchus	*An. albimanus*	STECLA (MR4)	Genetic, cytogenetic maps

Genome size, as estimated for four species in subgenus *Anopheles* and two from subgenus *Cellia*, is relatively constant and reasonably small (0.23-0.29 pg), less than culicine genomes (0.54-1.9 pg; Rai, Black, 1999) and roughly comparable to Drosophila. Eight anopheline genomes together represent a fraction of one typical mammalian genome.

Major considerations for choice of species relate to (1) availability of well-established colonies, (2) vector status, and (3) degree of evolutionary divergence from *An. gambiae*, the anchor of this project. Within this framework, we propose to sequence four species and incipient species very closely related to *An. gambiae* (in the same sibling species complex), two additional species within the same subgenus *Cellia*, one species from the sister subgenus *Anopheles*, and one from a more distant subgenus *Nyssorhynchus* (Fig 1; Table 1). Further, we propose that sequencing proceed from the species most closely related to *An. gambiae* (Phase I) to those most distantly related (Phase II). Thus, species within the *An. gambiae* complex will be sequenced first, followed in succession by species in *Cellia*, *Anopheles* and *Nyssorhynchus.*

Our plan is loosely modeled after the "ladder and constellation" approach adopted by the Drosophila Comparative Genomics project. This approach encompasses species representing different lineages that branch off at increasing evolutionary depths from a reference species (steps on the ladder), as well as clusters of species representing the same lineage (a constellation from a single step). By offering multiple points of comparison at a given phylogenetic depth, as well as multiple depths, this approach is flexible enough to allow the inference of recent, rapid mutational changes as well as more ancient ones, with the highest possible level of resolution.

Malaria Research and Reference Reagent Resource Center (MR4). MR4 was developed by NIAID in response to the need for improved community access to parasite, vector, and

human reagents; and standardization of assays using well-characterized and renewable reagents (www.malaria.mr4.org). ATCC currently holds the contract, with a subcontract to CDC (Mark Benedict) for the provision of anopheline reagents, including living laboratory colonies. All species proposed for genome sequencing, with the exception of *An. funestus*, are-- or soon will be-- registered with and maintained by MR4. This mechanism is intended to guard against extinction of colonies that were the source of genome sequencing projects, as happened with PEST. Added insurance against loss of colonies is provided by the use (where possible) of very well-established colonies that have been maintained stably in the laboratory for decades.

Inbred lines. The ability to generate isogenic lines is curtailed, owing in part to the mating behavior of anopheline species. In nature, they mate in swarms at dusk and therefore resist adaptation to laboratory cages and unnatural lighting conditions. Genome sequencing cannot be contemplated for any species that has not been colonized, and few have been. While all anopheline colonies experience severe bottlenecks during the process of laboratory adaptation and are therefore considerably inbred, they can resist brute-force attempts to make them isogenic or even homokaryotypic because of a high frequency of lethals. Indeed, a peculiar feature of even some long-term (*e.g.*, 50 year-old) laboratory colonies is the persistence of chromosomal inversion polymorphisms, some of which appear overdominant (Seawright *et al*., 1991). However, single pair matings are possible by two methods. The painstaking and time-consuming method of forced copulation entails human-choreographed forced matings between anesthetized females and decapitated males. Alternatively (though with a lower proportion of inseminations), single males can be caged with multiple females, in the presence of-- but without access to-- a cup of additional males (A. della Torre, pers. comm).

Evolutionary spacing of distant relatives. Identifying distant relatives of *An. gambiae* with the appropriate evolutionary spacing will be based, at least initially, on two considerations: available colonies, and phylogenetic relationships within genus *Anopheles*. *Anopheles* systematics is not well-developed. Although we can be quite certain of the relationships among the three subgenera that contain the most important vectors of malaria: ((*Cellia* + *Anopheles*) *Nyssorhynchus*), the very limited and young fossil record for mosquitoes provides no clues for dating the divergence of subgenera of *Anopheles*. The estimate of ~95 MY for the earliest branching events within subgenus *Anopheles* (Fig. 1) is based on a biogeographic hypothesis of Anophelinae history, in which the earliest branching events were inferred to have taken place before the loss of the land connection between Africa and South America (Krzywinski *et al.*, 2001). All available evidence suggests a rapid diversification of basal subgeneric lineages of *Anopheles*, possibly associated with the breakup of Gondwanaland (Krzywinski *et al.*, 2001), and therefore all subgenera may have emerged at roughly the same time. Given that we want to sample each of three subgenera, we have proposed to include *An. funestus* and *An. farauti* (*Cellia*), *An. quadrimaculatus* (*Anopheles*) and *An. albimanus* (*Nyssorhynchus*). Very limited prior sequence information exists from these taxa, but divergence from *An. gambiae* based on nucleotide p-distance from 759 sites of the single-copy nuclear gene *white* ranges from 0.123 to 0.145 to 0.148 in *Cellia, Anopheles* and *Nyssorhynchus*, respectively (Krzywinski *et al.*, 2001). This can be compared to the

white gene divergence between *D. melanogaster* and *D. virilis*, giving a nucleotide p-distance of 0.173 across 711 sites. Based on these uncorrected *white* gene estimates, it appears that the proposed anopheline species are spaced similarly to species in the Drosophila cluster. We propose that remaining concerns about the appropriateness of evolutionary spacing can be addressed more thoroughly by exploratory sequencing of BAC clones in advance of a genome-sequencing effort (see below).

A. The *Anopheles gambiae* complex (four species): Here, we employ a constellation of very close relatives, among which at least one is not a vector, to examine the set of evolutionary events that led to bursts of speciation and the development of a major malaria vector, and to resolve the confounding M and S sequences present in the current PEST genome. This constellation will consist of: (1) Molecular form M, an incipient species of *An. gambiae*, (2) Molecular form S, an incipient species of *An. gambiae*, (3) *An. arabiensis*, (4) *An. quadriannulatus* (Table 1). As emerging species, M and S provide one of the best windows on the process of adaptive divergence and speciation. *An. arabiensis* is a major vector that is of interest because of its more catholic host preferences, and because of introgression with *An. gambiae* in some—but not all—regions of the genome (Besansky *et al*., 2003). *An. quadriannulatus*, though capable of experimental infection with *Plasmodium falciparum* in the laboratory, is not involved in transmission in nature because it rarely feeds on humans. It is also considered to be most closely related to an ancestral species from which the *An. gambiae* complex arose.

The four proposed taxa are so closely related to each other and to the sequenced PEST genome that we anticipate no difficulty with assembly at 3x coverage, using the backbone of the existing *An. gambiae* genome sequence. However, in consideration of the fact that either M or S (or both) will represent the new *An. gambiae* reference genome, more investment (5-6x) may be needed to ensure a high enough level of quality and contiguity that the assembly can be used as a reference not only for studies of protein-coding genes but also for cis-regulation, positive selection, and even repetitive DNA distribution.

All taxa have been colonized, and most are maintained in multiple locations. If they prove chromosomally polymorphic, we will attempt to select for a homokaryotypic strain. ***An. arabiensis* KGB** strain is maintained by MR4 and by Maureen Coetzee (South Africa); it may be polymorphic for inversion *b* on the right arm of chromosome 2 (designated 2*Rb*) and if so will be selected for the standard (or inverted) arrangement. ***An. quadriannulatus* SKUQUA** colony is also maintained by MR4 and M. Coetzee. It is homokaryotypic and carries the standard arrangements. ***An. gambiae* M (Sua2La)** and S (**S-GA-KIS**) strains are maintained by MR4 and A. della Torre (Italy). These strains are both homokaryotypic for the standard arrangement on 2R; SUA is fixed for 2La while S-GA-KIS currently is polymorphic for this inversion. We will select for 2La homozygotes.

B. Outside of the *An. gambiae* complex (4 species). To reap the full benefits of comparative genomics, we require independent assemblies with a minimum of small gaps. The rationale for this was expressed in the Drosophila Comparative Genomics White Paper (flybase.net/.data/docs/CommunityWhitePapers/GenomesWP2003.pdf):
"With less than 5x shotgun sequencing, one is forced into mapping reads to a known genome in order to

assemble the data. In such an approach most of the information about translocations, duplications, and deletions is lost. While individuals who wish to view the genome as an unordered sack of proteins may be content with such a result, anyone interested in genomic evolution, the diaspora of retrotransposons, or cis-regulation, to name but a few topics, will find the results inadequate. With 5-6x one can assemble independently, but one will invariably be missing a portion of a gene such as a bit of an exon, core promoter, or enhancer region. If we truly want to provide a dataset that is high value resource to both experimental and computational biologists, it seems wise to let the sequencing pipelines run to 8x".

Same subgenus *Cellia* (*An. funestus, An. farauti*): Within Phase II of this project, we set the highest priority on the sequencing of two species within the same subgenus *Cellia*. *An. funestus* is an African malaria vector as anthropophilic or even more anthropophilic than *An. gambiae*, only ~5MY diverged from *An. gambiae* (Sharakhov *et al.*, 2002). The colony (**FUMOZ**), maintained by Coetzee for >4 years, is polymorphic for 4 inversions at the following frequencies (2a+, 0.85; 3a+, 0.98; 3b+, 0.67; 5a+, 0.81). Due to the difficulty in colonizing and maintaining *An. funestus* in the laboratory, it is not feasible to select for a homokaryotypic strain. Significant resources are available in support of a genome sequencing project for this species. A polytene chromosome map with numbered divisions and lettered subdivisions (Sharakhov *et al.*, 2001b), a BAC and a cDNA library from multiple tissues and developmental stages, physical mapping of 157 cDNAs to the *An. funestus* polytene chromosomes (Sharakhov *et al.*, 2002), 1,019 high quality EST sequences (A. Serazin, A. Dana, N. Lobo, B. Harker, M. Hillenmeyer, M. Kern, I. Sharakhov, M. Coetzee, F.H. Collins, and N. Besansky, in prep), and a total of 70 microsatellite markers developed, of which 32 are physically mapped (Cohuet *et al.*, 2002; Schemerhorn *et al.*, 2003; Sharakhov *et al.*, 2004; Sharakhov *et al.*, 2001a; Sinkins *et al.*, 2000). Importantly, NIH funding to J. Hemingway and collaborators is supporting genetic mapping of the determinant(s) of resistance to pyrethroid insecticides that has emerged in South African *An. funestus* populations. As part of this project, the ends of all BAC clones in a 10X coverage library will be sequenced, and 2,000 of these clones will be physically mapped to the *An. funestus* polytene chromosomes. These data will provide a useful framework for a genome sequencing project.

Subgenus *Cellia* is divided into six informal taxonomic categories called "Series". *An. gambiae* is in the Pyretophorus Series, *An. funestus* in Myzomyia. Within the Neomyzomyia Series is an Australasian species, *An. farauti*. This species and a close relative, *An. punctulatus*, are almost as anthropophilic as *An. gambiae* and *An. funestus*, which explains why malaria transmission in Papua New Guinea, Irian Jaya and the Solomon Islands is almost as intense as in Africa. A robust colony of ***An. farauti* FAR1** is maintained by MR4.

The addition of these two species in the *Anopheles* Genome Cluster project will give three points of comparison within subgenus Cellia, the minimum required for phylogenetic shadowing.

Subgenera *Anopheles* (*An. quadrimaculatus species A* ORLANDO) & *Nyssorhynchus* (*An. albimanus* STECLA): Aside from vector status (past or present) and taxonomic placement, the choice of *An. quadrimaculatus* and *An. albimanus* was dictated principally by very robust colonies that have been in the laboratory for decades. Both are maintained by MR4 and elsewhere. There is a long history of classical genetic studies and a more recent history of population genetic studies of these species, which are or were major vectors in the US or Central America, the Caribbean and the northern coasts of South America. It has been reported (Seawright *et al*., 1991) that *An.*

quadrimaculatus colonies-- including ORLANDO-- as well as natural populations exhibited temporally stable overdominance associated with dimorphic forms of chromosome 3L. If this remains the case for ORLANDO, we will attempt to select for homokaryotypes. STECLA does not carry any polymorphic chromosomal inversions.

V. Quality control and value-added resources for the community: Adequate resources are already available for the species in the *An. gambiae* complex, including ESTs, cDNA and BAC libraries from *An. gambiae* as well as a BAC library from *An. quadriannulatus* (F. Collins and J. Tu, pers. comm). In addition, a 10x BAC library is already available for *An. funestus* (Table 1). For the other 3 candidate species outside of the *An. gambiae* complex, BAC libraries should be constructed, with copies deposited in MR4. Exploratory sequencing of BACs, perhaps centered around the *white* gene, can be a first step prior to genome sequencing, to evaluate evolutionary spacing of candidate species. Through directed end-sequencing and *in situ* hybridization (*i.e.*, as needed where problems arise), BACs will serve as one means of quality control. A second means of quality control would be the construction of one normalized cDNA library per species outside of the *An. gambiae* complex, each prepared from pooled RNA from different developmental stages (4-8h eggs, 24h eggs, $3^{rd}/4^{th}$ instar larvae, pupae, males, females, 6h/24h/48h blood-fed females, 9d malaria-infected females), and the sequencing of 25,000 ESTs per each normalized library. Evidence from the *An. gambiae* normalized pooled library from complex mixtures (every part/every life-stage/infected/uninfected) suggests that sequencing of 25,000 ESTs leads to a contig rate of ~0.4 (i.e., 10,000-12,000 individual contigs) after removal of mitochondrial/transposon/vector contamination (N. Lobo and F. Collins, unpublished).

VI. Leveraging parallel initiatives in support of the *Anopheles* genome cluster

Drosophila Genome Cluster (NHGRI): The scientific community will soon benefit from the complete genome sequences of 10 more species in addition to those of *D. melanogaster* and *D. pseudoobscura*. The availability of these data will drive the development of algorithms and software designed for (1) genome assembly and (2) comparative genome analysis and phylogenetic footprinting at multiple evolutionary depths. Such applications, having been tested and perfected in Drosophila, will empower the assembly and analysis of multiple anopheline genomes.

VectorBase (NIAID): Just as the additional Drosophila genome sequences and other data are being incorporated into FlyBase, so these additional *Anopheles* genomes and related data will be incorporated into a centralized relational database and web interface, VectorBase, that will enable the research community to search and analyze genomic data and related data types that are produced for invertebrate vectors of human pathogens. VectorBase has recently been funded by NIAID for an initial five-year period, through a contract to Collins (PI). It embraces the Genome Model Organism Database Construction set (GMOD) design goals, implementing the chado generic genome database schema in common with FlyBase. A web services interface will be provided to allow others to directly query the VectorBase data and tools without requiring that users install the full VectorBase package. The central feature of VectorBase will be adapted from the Ensembl genome display and analysis tool developed and managed by the European Bioinformatics Institute (EBI) and the Sanger

Institute. Other key components of VectorBase will be provided by the European Molecular Biology Laboratory in Heidelberg, Germany, the Institute of Molecular Biology and Biotechnology (IMBB) in Heraklion, Crete, the FlyBase group at Harvard University, and the Center for Tropical Disease Research and Training at the University of Notre Dame. VectorBase will manage, display, and analyze data for all invertebrate vectors for which genome level data sets are developed (genome sequences, extensive EST sequence sets, other large scale genome-derived data sets, or data sets based on functional analysis of the genome). Moreover, VectorBase will assume responsibilities for developing, maintaining and updating internationally recognized Reference Data Sets for the organisms included.

Conclusion

The constituency that would benefit from additional anopheline genomes is the entire community of vector biologists and parasitologists who are seeking novel solutions for controlling malaria. *An. gambiae* is the model organism for malaria vectors, and rightly so. The ultimate goal is to extract from the genome information that will lead us to develop new control tools, be they chemical or genetic, aimed at altering vectorial capacity. Two problems stand in the way of our ability to fully exploit this resource. First, the genome is not fully annotated. In coding sequence there are false positives, false negatives, and misannotations; in noncoding sequence, regulatory elements are very poorly known and notoriously difficult to identify. Second, even if the genome were completely annotated, it is difficult to discover the targets of interest. Focused studies on *An. gambiae* remain a must, but comparative genomics is a powerful ally. Analyzing the genome sequences of multiple related species in a phylogenetic framework, under the dual rationales that "what is important is conserved" (Gibbs, Nelson, 2003) and "what is adaptive is unusually highly diverged", is an extremely efficient technique for discovering those targets.

[1]Eight Genomes Cluster Committee:

Nora J. Besansky (Chair), University of Notre Dame, USA
John Adams, Scientific Advisor for Malaria Research and Reference Repository (MR4), University of Notre Dame, USA
Michael Ashburner, Visiting Group Leader and former Joint-Head, European Bioinformatics Institute, UK; Development and Oversight, FlyBase
Mark Benedict, MR4 subcontractor for vectors, CDC, USA
Jane Carlton, PI of *Plasmodium* genome sequencing and comparative genomics projects, The Institute for Genomic Research (TIGR), USA
Maureen Coetzee, Head, Vector Control Reference Unit, National Institute for Communicable Diseases, SOUTH AFRICA
Frank H. Collins, Scientific Advisor for MR4 and PI of *An. gambiae* genome project, University of Notre Dame, USA; Development and Oversight, VectorBase
Alessandra della Torre, University of Rome, ITALY
Janet Hemingway, Director, Liverpool School of Tropical Medicine, UK
David S. Roos, Merriam Professor of Biology; Director, Genomics Institute; Development and Oversight, PlasmoDB; University of Pennsylvania, USA

Yeya Touré, Manager, Molecular Entomology Committee and Malaria Research Coordinator, WHO/TDR, SWITZERLAND
Rick Wilkerson, Manager, Walter Reed Biosystematics Unit, USA

References

Adak, T, Kaur, S, Singh, OP (1999) Comparative susceptibility of different members of the *Anopheles culicifacies* complex to *Plasmodium vivax*. *Trans Royal Soc Trop Med Hyg* **93**, 573-577.

Barbash DA, Siino DF, Tarone AM, Roote J (2003) A rapidly evolving MYB-related protein causes species isolation in Drosophila. *Proc Natl Acad Sci U S A* **100**, 5302-5307.

Bergman CM, Pfeiffer BD, Rincon-Limas DE, *et al.* (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Biol* **3**, RESEARCH0086.

Besansky NJ, Krzywinski J, Lehmann T, *et al.* (2003) Semipermeable species boundaries between *Anopheles gambiae* and *Anopheles arabiensis*: evidence from multilocus DNA sequence variation. *Proc Natl Acad Sci USA* **100**, 10818-10823.

Besansky NJ, Fahey GT (1997) Utility of the white gene in estimating phylogenetic relationships among mosquitoes (Diptera: Culicidae). *Mol Biol Evol* **14**, 442-454.

Besansky NJ, Powell JR (1992) Reassociation kinetics of Anopheles gambiae (Diptera: Culicidae) DNA. *J Med Entomol* **29**, 125-128.

Boffelli D, McAuliffe J, Ovcharenko D, *et al.* (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**, 1391-1394.

Burt DW, Bruley C, Dunn IC, *et al.* (1999) The dynamics of chromosome evolution in birds and mammals. *Nature* **402**, 411-413.

Cliften P, Sudarsanam P, Desikan A, *et al.* (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* **301**, 71-76.

Coghlan A, Wolfe KH (2002) Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. *Genome Res* **12**, 857-867.

Cohuet A, Simard F, Berthomieu A, *et al.* (2002) Isolation and characterization of microsatellite DNA markers in the malaria vector *Anopheles funestus*. *Mol Ecol Notes* **2**, 498-500.

Collins FH, Paskewitz SM (1995) Malaria: current and future prospects for control. *Annual Review of Entomology* **40**, 195-219.

Collins FH, Sakai RK, Vernick KD, *et al.* (1986) Genetic selection of a Plasmodium-refractory strain of the malaria vector *Anopheles gambiae*. *Science* **234**, 607-610.

Coluzzi M (1982) Spatial distribution of chromosomal inversions and speciation in anopheline mosquitoes. In: *Mechanisms of Speciation*, pp. 143-153. Alan R. Liss, Inc., New York.

Coluzzi M, Sabatini A, Della Torre A, Di Deco MA, Petrarca V (2002) A Polytene Chromosome Analysis of the Anopheles gambiae Species Complex. *Science* **3**, 3.

Coluzzi M, Sabatini A, Petrarca V, Di Deco MA (1979) Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. *Trans R Soc Trop Med Hyg* **73**, 483-497.
Cornel AJ, Collins FH (2000) Maintenance of chromosome arm integrity between two Anopheles mosquito subgenera. *J Hered* **91**, 364-370.
Costantini C, Sagnon NF, Ilboudo-Sanogo E, Coluzzi M, Boccolini D (1999) Chromosomal and bionomic heterogeneities suggest incipient speciation in *Anopheles funestus* from Burkina Faso. *Parassitologia* **41**, 595-611.
della Torre A, Costantini C, Besansky NJ, *et al.* (2002) Speciation within Anopheles gambiae--the glass is half full. *Science* **298**, 115-117.
Filchak KE, Etges WJ, Besansky NJ, Feder JL (2005) The ecological genetics of host use in the Diptera. In: *The Evolutionary Biology of Flies* (eds. Wiegman BM, Yeates DK). Columbia University Press, New York.
Garrett-Jones C (1964) Prognosis for interruption of malaria transmission through assessment of the mosquito's vectorial capacity. *Nature* **204,**1173-1175.
Garrett-Jones C, Shidrawi GR (1969) Malaria vectorial capacity of a population of Anopheles gambiae: an exercise in epidemiological entomology *Bull WHO* **40**, 531-545.
Gaunt MW, Miles MA (2002) An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Mol Biol Evol* **19**, 748-761.
Gibbs RA, Nelson DL (2003) Human genetics. Primate shadow play. *Science* **299**, 1331-1333.
Gillies MT, Coetzee M (1987) *A Supplement to the Anophelinae of Africa South of the Sahara* The South African Institute for Medical Research, Johannesburg.
Gillies MT, De Meillon B (1968) *The Anophelinae of Africa South of the Sahara*, 2nd edn. South African Institute for Medical Research, Johannesburg.
Gonzalez J, Ranz JM, Ruiz A (2002) Chromosomal elements evolve at different rates in the Drosophila genome. *Genetics* **161**, 1137-1154.
Gumucio DL, Heilstedt-Williamson H, Gray TA, *et al.* (1992) Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol Cell Biol* **12**, 4919-4929.
Hallem, EA, Nicole Fox, A, Zwiebel, LJ, Carlson, JR (2004) Olfaction: mosquito receptor for human-sweat odorant. *Nature* **427**, 212-213.
Harbach RE (2004) The classification of genus *Anopheles* (Diptera: Culicidae): a working hypothesis of phylogenetic relationships. *Bull Entomol Res* **94**, 537-553.
Holt RA, Subramanian GM, Halpern A, *et al.* (2002) The genome sequence of the malaria mosquito Anopheles gambiae. *Science* **298**, 129-149.
Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-254.
Krzywinski J, Besansky NJ (2003) Molecular systematics of Anopheles: from subgenera to subpopulations. *Annu Rev Entomol* **48**, 111-139.
Krzywinski J, Wilkerson RC, Besansky NJ (2001) Toward understanding Anophelinae (Diptera, Culicidae) phylogeny: insights from nuclear single-copy genes and the weight of evidence. *Syst. Biol.* **50**, 540-556.

Matthews TC, Munstermann LE (1994) Chromosomal repatterning and linkage group conservation in mosquito karyotypic evolution. *Evolution* **48**, 146-154.
Michalak P, Noor MA (2003) Genome-wide patterns of expression in Drosophila pure species and hybrid males. *Mol Biol Evol* **20**, 1070-1076.
Mongin E, Louis C, Holt RA, Birney E, Collins FH (2004) The Anopheles gambiae genome: an update. *Trends Parasitol* **20**, 49-52.
Mukabayire O, Besansky NJ (1996) Distribution of T1, Q, Pegasus and mariner transposable elements on the polytene chromosomes of PEST, a standard strain of Anopheles gambiae. *Chromosoma* **104**, 585-595.
Petrarca V, Beier JC (1992) Intraspecific chromosomal polymorphism in the Anopheles gambiae complex as a factor affecting malaria transmission in the Kisumu area of Kenya. *Am J Trop Med Hyg* **46**, 229-237.
Powell JR, Petrarca V, della Torre A, Caccone A, Coluzzi M (1999) Population structure, speciation, and introgression in the *Anopheles gambiae* complex. *Parassitologia* **41**, 101-113.
Rai KS, Black WCt (1999) Mosquito genomes: structure, organization, and evolution. *Adv Genet* **41**, 1-33.
Schemerhorn BJ, Greeman S, Banks M, Vulule J, Sagnon N'F, Costantini C, Besansky NJ (2003) Dinucleotide microsatellite markers from *Anopheles funestus*. *Mol Ecol Notes* **3**, 505-507.
Seawright, JA, Kaiser, PE, Narang, SK (1991) A unique chromosomal dimorphism in species A and B of the *Anophels quadrimaculatus* complex. *J Hered* **82**, 221-227.
Severson DW, DeBruyn B, Lovin DD, *et al.* (2004) Comparative genome analysis of the yellow fever mosquito *Aedes aegypti* with *Drosophila melanogaster* and the malaria vector mosquito *Anopheles gambiae*. *J Hered* **95**, 103-113.
Severson DW, Brown SE, Knudson DL (2001) Genetic and physical mapping in mosquitoes: molecular approaches. *Annu Rev Entomol* **46**, 183-219.
Sharakhov I, Braginets O, Grushko O, Cohuet A, Guelbeogo WM, Boccolini D, Weill M, Costantini C, Sagnon N'F, Fontenille D, Yan G, Besansky NJ (2004) A microsatellite physical map of the African human malaria vector *Anopheles funestus*. *J Hered* **95**, 29-34.
Sharakhov IV, Serazin AC, Grushko OG, Dana A, Lobo N, Hillenmeyer ME, Romero-Severson J, Costantini C, Sagnon N'F, Collins FH, Besansky NJ (2002) Inversions and gene order shuffling in *Anopheles gambiae* and *A. funestus*. *Science* **298**, 182-185.
Sharakhov IV, Braginets O, Mbogo CN, Yan G (2001a) Isolation and characterization of trinucleotide microsatellites in African malaria mosquito *Anopheles funestus*. *Mol Ecol Notes* **1**, 289-292.
Sharakhov IV, Sharakhova MV, Mbogo CM, Koekemoer LL, Yan G (2001b) Linear and spatial organization of polytene chromosomes of the African malaria mosquito A*nopheles funestus*. *Genetics* **159**, 211-218.
Sinkins SP, Hackett BJ, Costantini C, Vulule J, Ling YY, Collins FH, Besansky NJ (2000) Isolation of polymorphic microsatellite loci from the malaria vector *Anopheles funestus*. *Mol Ecol* **9**, 490-492.
Stern DL (1998) A role of Ultrabithorax in morphological differences between Drosophila species. *Nature* **396**, 463-466.

Stump A, Shoener JA, Costantini C, Sagnon N'F, Besansky NJ (2005) Sex-linked differentiation between incipient species of *Anopheles gambiae*. *Genetics*, **in press.**

Sucena E, Stern DL (2000) Divergence of larval morphology between Drosophila sechellia and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. *Proc Natl Acad Sci U S A* **97**, 4530-4534.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001a) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. *Proc Natl Acad Sci U S A* **98**, 7375-7379.

Swanson WJ, Yang Z, Wolfner MF, Aquadro CF (2001b) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc Natl Acad Sci U S A* **98**, 2509-2514.

Tagle DA, Koop BF, Goodman M, *et al.* (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* **203**, 439-455.

Takken W, Eling W, Hooghof J, *et al.* (1999) Susceptibility of Anopheles quadriannulatus Theobald (Diptera: Culicidae) to Plasmodium falciparum. *Trans R Soc Trop Med Hyg* **93**, 578-580.

Thomas JW, Touchman JW, Blakesley RW, *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788-793.

Yang Z, Nielsen R (2002) Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. *Mol Biol Evol* **19**, 908-917.