Genome Analysis Of Vectorial Capacity In Major *Anopheles* Vectors Of Malaria Parasites

Submitted August 2008 by:
Nora J. Besansky on behalf of the *Anopheles* Genomes Cluster Committee[1]

Executive Summary

Malaria causes an estimated 300-500 million cases and kills three million children annually. Despite considerable emphasis on the development and deployment of control methods, the disease remains a significant threat. Mosquito control in particular has suffered from the development of resistance to insecticides. Of the ~500 anopheline species, only two dozen are important vectors of human malaria parasites. Why some members of the same anopheline species transmit malaria parasites while others do not, or are less efficient, is of intense interest to vector biologists. Developing a better understanding of this 'vectorial capacity' may enable its eventual manipulation in order to reduce disease burden.

This document proposes sequencing of 13 anopheline vector genomes (average size ~250 Mb), representing 26 billion base pairs, to complement and facilitate comparative analysis with the three other sequenced anophelines, *Anopheles gambiae* PEST, M and S forms. Using *An. gambiae* as the anchor and adopting a 'ladder-and-constellation' approach inspired by the successful 12 *Drosophila* genomes project, we propose deep sampling of species belonging to the *An. gambiae* sibling species complex (**Tier 1**), followed by sampling at increasing evolutionary distances within the three main *Anopheles* subgenera (**Tiers 2 and 3**), with particular emphasis on subgenus *Cellia* that contains *An. gambiae*. In addition to genomic sequencing, we propose EST sequencing for each species in support of genome annotation.

Generating genome sequence data using this scheme will allow inferences about both rapid and gradual evolutionary changes relevant to vector ability. This is necessary to determine, for example, the underlying genetic determinants of feeding preference, since these are unlikely to be conserved across large evolutionary distances and specialization on human blood feeding is likely to have been a very recent evolutionary event. It will also enable the development of powerful genomic tools that are the necessary foundation for identifying new approaches to the control of vectors whose biology is poorly understood, in contrast to genetic and evolutionary models such as *Drosophila*.

Summary of Proposed 13 Anopheline Genomes WGS Project

Anopheles Classification	Priority	No. Species at 8X coverage	Total Bases WGS (billions)	No. ESTs (thousand)
Subgenus Cellia				
Series Pyretophorus	Tier 1	4	8	800
Series Neocellia	Tier 2	2	4	400
Series Myzomyia	Tier 2	3	6	600
Series Neomyzomyia	Tier 2	2	4	400
Subgenus Anopheles	Tier 3	1	2	200
Subgenus Nyssorhynchus	Tier 3	1	2	200
Total	**3**	**13**	**26 billion**	**2.6 million**

This white paper has strong support from vector biologists and members of the malaria community, in addition to the interest and commitment of geneticists, evolutionary biologists and computational biologists whose contributions to this project will aid in the analysis of the data and quicken the pace of discovery.

1. Introduction:

Malaria kills an estimated three million people annually, mostly children in sub-Saharan Africa under the age of five. Human malaria parasites (genus *Plasmodium*) have a complex life cycle that requires development in the mosquito as well as the human host. Thus, human malaria transmission is critically dependent upon mosquito vectors. All malaria vectors are mosquitoes in the genus *Anopheles*, but of the ~500 anopheline species, only two dozen are important vectors of human malaria parasites. A measure of the ability of a mosquito to transmit malaria is the **vectorial capacity**, formally defined as the 'maximal potential force of malaria parasite transmission by a local population of *Anopheles* vectors'. Vectorial capacity is determined by a combination of four attributes: the mosquito's physiological ability to support parasite development (formally known as vector competence), average daily size of the host-seeking population, average adult longevity of female mosquitoes, and the proportion of the mosquito population that feeds on blood.

The broad goal of this proposal is to undertake comparative analysis of the genomes of a total of 16 different *Anopheles* species that have been selected for their vectorial capacity and genetic relatedness. These 16 genomes include the published *An. gambiae* PEST genome, its two incipient species (*An. gambiae* M and S forms) also recently completed, plus 13 genomes proposed in this project: no other anopheline species have been sequenced. At the center of this comparison is the African sibling species group known as the *An. gambiae* complex, which comprises seven formally recognized species that vary considerably in vectorial capacity, from the nominal *An. gambiae* considered as the world's most important malaria vector to its non-vector sibling *An. quadriannulatus*. Unlike the genomes of culicine mosquitoes such as *Aedes* and *Culex*, the relatively small size of anopheline genomes (~250 Mb) makes this entire project roughly equivalent to the sequencing effort for one mammalian genome. The insights gained from these comparative genomic studies will have several significant applications to the overall goal of malaria control, most importantly through a greater understanding of vectorial capacity and its eventual manipulation to reduce disease burden.

2. Background:

2.1. Only Anopheline mosquitoes transmit human malaria. Mosquitoes (Diptera: Culicidae) are an ancient monophyletic group of at least 4,500 species whose origin predates the Jurassic period (**Fig 1**). The first basal split of the ancestral mosquito lineage gave rise to two deeply diverged subfamilies, Culicinae (containing *Aedes aegypti* and *Culex pipiens quinquefasciatus*) and Anophelinae (containing *An. gambiae*), an estimated 145-200 million years ago (MYA) (Krzywinski *et al.*, 2006). Although subfamily Culicinae contains important vectors of arboviruses and filarial worms, only subfamily Anophelinae contains vectors of human malaria.

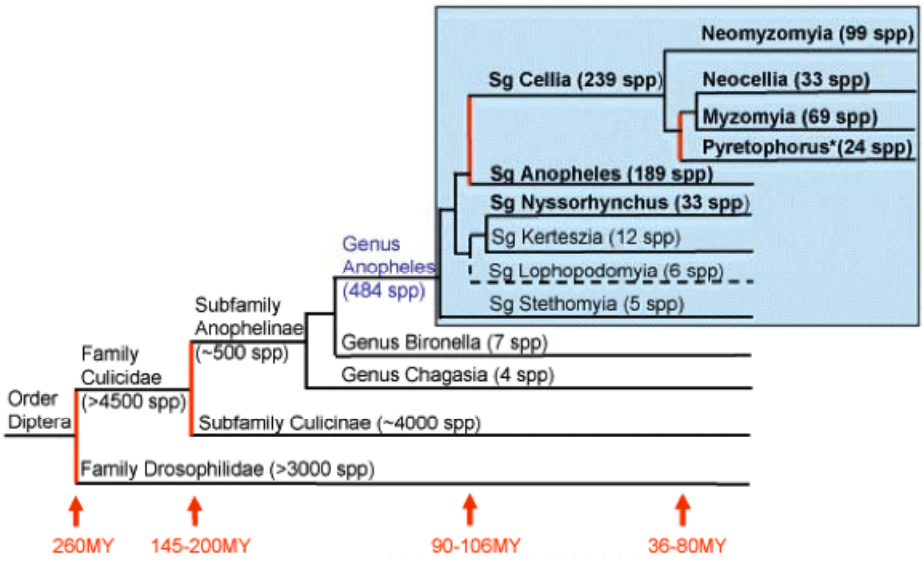


Fig 1. Cladogram of relationships within Culicidae, with emphasis on anophelines (blue box). Reference genome *Anopheles gambiae*) is classified in Subgenus (Sg) Cellia, within Series Pyretophorus (*). Branch lengths are not scaled by evolutionary time, however, maximum-likelihood estimates of select divergence times are indicated (from Krzywinski *et al.,* 2006). Subgenus Cellia is classified into six "Series", only the four most speciose are shown (those omitted comprise 14 additional species). Subgenera and Series to be sampled for WGS are denoted in **bold**.

2.2. Malaria vectors are a special subset of anopheline mosquitoes. Subfamily Anophelinae (~500 species) probably arose in the neotropics (Krzywinski *et al.*, 2001). It is divided into three genera, of which the largest by far -- genus *Anopheles* (484 species) -- contains the major malaria vectors. The genus *Anopheles* consists, in turn, of six subgenera whose relationships are largely uncertain due to their relatively ancient and rapid divergence (Krzywinski *et al.*, 2001). Divergence within the genus probably dates to events associated with the breakup of Pangaea, which began about 150 MYA and reached major continental separation by about 100 MYA. The two largest subgenera, the cosmopolitan *Anopheles* (189 species) and Old World *Cellia* (239 species) are sister taxa that together contain most of the important malaria vectors.

While the ability to transmit human malaria is uniquely present within the genus *Anopheles*, it is quite rare: only ~30 among nearly 500 species are major vectors (Collins, Paskewitz, 1995). Not only are good malaria vectors rare among anopheline mosquitoes, they are not closely related. This observation is surprising if it is assumed that the traits underlying high vectorial capacity arose only once in anopheline mosquitoes, and if it is also assumed that such traits are stably maintained. Under these assumptions, all significant malaria vectors would be expected to be close relatives, *i.e* cluster in a phylogenetic tree, to the exclusion of other anophelines. This is not the case, as illustrated by the plethora of anopheline sibling species complexes (groups of very closely related and morphologically indistinguishable species presumed to have arisen both recently and rapidly), some members of which are not involved in malaria transmission.

2.3. "Vector traits" are recurrent and rapidly evolving. A salient feature of these sibling species complexes, and the reason they are so important in malaria epidemiology, is that they typically contain both vector and non-vector species. For example, the two principal malaria vectors, *An. gambiae* and *An. funestus*, are embedded within different species complexes in subgenus *Cellia* -- the *An. gambiae* complex belonging to Pyretophorus Series, and the Funestus Subgroup belonging to Myzomyia Series. Each complex contains the non-vectors *An. quadriannulatus* and *An. vaneedeni*, respectively. This leads to two conclusions of fundamental importance to understanding the nature of traits underlying successful malaria transmission by anophelines. First, because vectors from different complexes are not closely related, at least some of the underlying vector traits arose independently multiple times in different lineages. Second, the presence of both vector and non-vector species in the same species complex implies either rapid loss or rapid gain of "vector traits". Thus, at least some of the genes that are associated with vectorial capacity -- whether involved in immunity, host-preference, or some other physiological or behavioral response -- are likely to be rapidly evolving rather than highly conserved over long evolutionary distances. In particular, genes associated with behaviors like preference for human blood meals, selection of anthropogenic breeding sites, or preference to rest inside human dwellings, all of which represent 'use of' the human environment, are likely to be very recent evolutionary adaptations that postdate human cultural innovations such as the development of agriculture and animal husbandry that enabled the human populations to reach high and stable (ie. non-nomadic) densities.

3. Rationale:

3.1. Multiple *Anopheles* genomes will provide a framework to illuminate the genetic basis of vectorial capacity. Three medically relevant questions (expanded upon below) in vector biology include: (1) Why do anopheline mosquitoes transmit human malaria parasites and not other genera? (2) Why do some members of the same anopheline species transmit malaria parasites while others do not? (3) Why are some anopheline species more efficient vectors than others? Answering these questions, particularly the latter two, requires an in-depth understanding of the key traits that determine vectorial capacity. Our goal of facilitating comparative analysis of the multiple *Anopheles* genomes outlined here would establish a rich datasource and generate a framework for gathering this information and answering such questions.

The traits that impact a mosquito's role in malaria transmission are known, in principle. They include susceptibility of the mosquito to the parasite throughout the entire sporogonic stage, and mosquito population density, longevity, and bloodfeeding behavior. Acquisition of genome assemblies for the mosquito species highlighted in this proposal is critical for understanding the genetic basis for these traits.

(1) Why do anopheline mosquitoes transmit human malaria parasites and not other genera? Culicine mosquitoes cannot transmit human (indeed, mammalian) *Plasmodium* species due to the fact that they are not susceptible to infection by these parasites. However, the molecular basis of this infection barrier, *i.e* whether due to the lack of specific receptors, or to variable immunity, or to other mechanisms, and the variation in infection barriers among non-vector species remain unknown. Specific answers to this question will depend largely on genome comparisons involving other mosquito species – including *Aedes* and *Culex* – that are being sequenced as part of ongoing efforts in other laboratories.

(2) Why do some members of the same anopheline species transmit malaria parasites while others do not? This question acknowledges differences in vector ability even within the same species. Phenotypic differences in susceptibility to parasites are well-described, both among anopheline species and among individuals of the same species (Collins *et al.*, 1986; Vernick *et al.*, 2005). For example, *An. gambiae* is more susceptible to co-indigenous *Plasmodium falciparum* malaria parasites than those isolated from the New World or Asia (Collins *et al.*, 1986). This reflects the evolutionary arms race between malaria parasites and their anopheline vectors. On the vector side, this could involve genes controlling immunity or receptors required for invasion, and on the parasite side, this could involve surface ligands, proteases or chitinases. Insight will be gained through parallel comparative genomic analyses of multiple anopheline vectors and multiple *Plasmodium* species, especially now that genome sequencing of a large number of different *Plasmodium* species is ongoing (see the Comprehensive Sequencing Proposal for *Plasmodium* white paper at http://plasmodb.org/common/downloads/doc/PlasmodiumWhitePaperV5.pdf. The choice of *P. vivax* and *P. falciparum* strains for these recently approved sequencing projects have been based in part on compatibility with the *Anopheles* species in this proposal.) The genetic basis for one mechanism of refractoriness in natural populations of *An. gambiae* has recently been unraveled (Riehle *et al.*, 2006), an enterprise made practical only through the availability of the *An. gambiae* genome sequence. However, knowledge of refractory mechanisms is not nearly as advanced in other vector species, and the availability of additional genome sequence will open this very important avenue of investigation.

(3) Why are some anopheline species more efficient vectors than others? If individuals from the same species can differ in vector competence, it can be no surprise that different species -- even very closely related ones -- can differ as well. Thus, the answer to this question may be an obvious one, *i.e.* differences in parasite susceptibility. For example, in paired feeding experiments with malaria-infected blood involving the sibling species A, B and C of the *An. culicifacies* complex from India, species A was more susceptible (63%, <5%, and 26% infected individuals among species A, B and C, respectively; Adak *et al.*, 1999). Alternatively, differences in behavior or longevity may be responsible rather than differences in susceptibility. If a mosquito is susceptible to parasite infection, its blood-feeding behavior and longevity are two of the most important determinants in vector capacity. To transmit malaria efficiently, the mosquito must have a high probability of feeding on humans and must live long enough to allow the malaria parasite to complete extrinsic development. Average longevity among anopheline species in the tropics ranges widely, from 10 days to over one month (Sattabongkot *et al.*, 2004). Importantly, *P. falciparum* (the causative agent of malignant tertian malaria responsible for nearly all malaria deaths) develops more slowly in the mosquito than does *P. vivax* (causative agent of benign tertian malaria). For example, in experimental infections of the Asian vector *An. dirus*, *P. falciparum* required two more days to invade the salivary glands than did *P. vivax* (13 versus 11 days; Sattabongkot *et al.*, 2004). This difference is believed to limit the prevalence of *P. falciparum* to areas in which the vectors are sufficiently long-lived, and helps to explain the lower prevalence of this malignant parasite in Central and South America despite the presence of an anthropophilic vector, *An. darlingi*. The genetic basis for differences in longevity is likely highly complex and, despite its obvious importance in malaria transmission, has been difficult to study in the absence of genomic resources that would be generated by this project.

Average lifespan can account for differences in vectorial capacity among species, but it is certainly not the only factor, as illustrated by the *An. gambiae* complex. This complex includes seven species so closely related that their evolutionary relationships could not be resolved by DNA:DNA hybridization methods used with success in the *Drosophila melanogaster* group (N.J. Besansky and J.R. Powell, unpublished data). *An. gambiae* and its sibling species *An. quadriannulatus* represent an example of paired vector and non-vector species that differ profoundly in their roles in malaria transmission. The latter is not found naturally infected with malaria parasites, yet can be infected with cultured *P. falciparum* (albeit at lower infection prevalence compared to *An.*

gambiae; Takken *et al.*, 1999; Habtewold *et al.*, 2008). Its non-vector status in nature is due to the preference of *An. quadriannulatus* for feeding on animals; it very rarely feeds on humans. *An. gambiae* on the other hand, shows an overwhelming preference for human odor. A third sibling in the same species complex, the malaria vector *An. arabiensis*, is a more opportunistic feeder that feeds avidly on humans but whose bites can be diverted to nearby domestic animals. Yet another member of the complex, *An. merus*, is even more catholic in its blood feeding habits and thus is even less important as a vector where domestic or wild animals are abundant alternatives to people. Moreover, unlike the previous three members of the complex, the larvae of *An. merus* have the unusual capacity of being able to develop in brackish water, thus its distribution is limited largely to coastal east Africa.

This phenomenon -- the coexistence of vectors and non-vector species in the same species complex -- is the rule rather than the exception in *Anopheles*, implying that host preference can be quite labile. The ramifications of this lability are that the underlying genetic determinants of feeding preference are unlikely to be conserved across large evolutionary distances and that specialization on human blood feeding is likely to be a very recent evolutionary event, given that humans in Africa were probably abundant enough to warrant such specialization only with the onset of agriculture within the past few thousand years. Different feeding behaviors, likely mediated by differential responses to host odor (Besansky *et al.*, 2004), may depend upon the presence of particular gustatory or odorant receptors (*e.g.*, Hallem *et al.*, 2004) or other genes in the olfactory pathway that could be compared between different genomes for insights into the genetics of host preference. One of the best opportunities for understanding how the behavioral transition is made at the genetic level lies in comparing sibling species pairs of vector and non-vector mosquitoes. Moreover, given the close relatedness of species in the *An. gambiae* complex and the fertility of female interspecies hybrids, the use of genetic crosses to examine such phenotypes is possible.

3.2. Multiple *Anopheles* genomes and ESTs will enable (i) improved annotation of the *An. gambiae* PEST sequence, and (ii) identification of regulatory sequences underlying genes that contribute to "vector traits". Except where cDNAs are available as supporting evidence, automated annotation as practiced by *VectorBase* (www.vectorbase.org) relies on gene identification through sequence similarity and synteny in other organisms. Among the ~13,000 *An. gambiae* genes recognized in the latest database build (AgamP3.4, July 2007), only ~11% have well defined functions and almost 40% are considered as having unknown function. The conservative annotation approach followed by *VectorBase* (http://agambiae.vectorbase.org/Help/AgamP3.4/Notes_on_genebuild) clearly misses genes that are not supported by strong EST evidence or protein sequence similarity, or *ab initio* models with identifiable Pfam domains. Thus rapidly evolving genes and genes that are specific to mosquitoes or anophelines are clearly missed. For example, a previously unrecognized mosquito-specific protein family was recently discovered that includes candidate receptors for malaria sporozoite invasion of salivary glands (Korochkina *et al.*, 2006).

Perhaps even more importantly, the availablity of multiple anopheline genome sequences will facilitate identification of functional non-coding elements, especially transcriptional regulatory elements. Among the most persuasive testaments to the power of comparative genomics for both improving genome annotation and identifying novel regulatory elements, and in particular the power of using multiple rather than pairwise genome comparisons (Boffelli *et al.*, 2003; Gumucio *et al.*, 1992; Tagle *et al.*, 1988), was the use of phylogenetic "footprinting" and "shadowing" approaches used for analysis of multiple *Saccharomyces cerevisiae* genomes (Cliften *et al.*, 2003; Kellis *et al.*, 2003). Phylogenetic footprinting involving 3-5 related yeast species affected the annotation of ~15% of the genes, resulting in the discovery of 43 previously unannotated genes and the elimination of ~500 previously annotated genes determined to be false positives. Furthermore, this analysis doubled the catalog of transcriptional regulatory elements and provided insights into their interactions (Kellis *et al.*, 2003).

4. Sequencing targets, priorities and considerations:

4.1. *Anopheles* genome size is relatively small. Based on estimates from four species in subgenus *Anopheles* and two from subgenus *Cellia*, genome size in anophelines is relatively constant and small (~230-284 Mb) in comparison to known culicine genomes (528 Mb-1.9 Gb; Rai, Black, 1999), and roughly

comparable to the *Drosophila* genomes. Thirteen anopheline genomes are roughly equivalent to one mammalian genome.

4.2. Isogenic lines are not required for WGS assembly. Anophelines are extremely difficult, if not impossible in many instances, to culture in the laboratory. Anopheline mosquitoes also experience severe bottlenecks during the process of laboratory adaptation and therefore existing colonies are somewhat inbred but are not isogenic. Generation of isogenic lines is very difficult, owing to the extensive labor and space-intensive husbandry that is required, as well as the mating behavior that normally requires swarming. In addition, experience suggests that heterozygosity in anopheline colonies does not pose an insurmountable problem for modern assembly algorithms, for several reasons: First, despite relatively high heterozygosity, the *An. gambiae* PEST genome was successfully assembled using the *Celera Assembler* algorithm (Holt *et al.,* 2002). A total of 87 scaffolds covering more than 88% of the total genome-- many larger than 10 Mb-- were assigned to chromosomal locations. Subsequent proof of principle was achieved with the independent assembly of the two genomes of *An. gambiae* incipient species M and S in 2007 (unpublished), providing further assurance that isogenic lines are not required, and that considerable levels of heterozygosity can be tolerated, given that appropriate assembly software and parameters are employed. To substantiate this claim, results from the ongoing *An. gambiae* M and S WGS project are provided next.

4.3. **de novo WGS assemblies of *An. gambiae* M and S genomes.** In June 2005, NHGRI approved the sequencing of *Anopheles gambiae* M and S form genomes and provided funding to the J. Craig Venter Institute (JCVI; S form) and the Washington University School of Medicine, Genome Sequencing Center (WUGSC; M form). The project has been coordinated by Besansky (Univ. Notre Dame). DNA samples or whole mosquitoes were provided by the University of Notre Dame and the Malaria Research and Reference Reagent Resource Center **(***MR4*; www.mr4.org**)** from sources described below. BAC libraries were provided by the Clemson University Genomics Institute (CUGI), and are available through CUGI or *MR4*. The combined whole genome shotgun (WGS) plasmid, fosmid and BAC end sequence reads for both genomes (~2.7 million traces from each genome, available in the NCBI Trace Archives) were assembled *de novo* by JCVI using the *Celera Assembler* and by WUGSC using the *Pcap* assembler. In the initial WUGSC assemblies based on the original *Pcap* algorithm, the M and S genome sizes were nearly twice the expected ~260 Mb (estimated from reassociation kinetic studies). This outcome was due to considerable numbers of high quality base discrepancies (polymorphisms), owing to the high allelic variation in the non-isogenic genome samples. Although a modification of *Pcap* (Pcap.rep.poly) resulted in more reasonable genome sizes, the algorithms implemented in the *Celera Assembler* to accommodate heterozygosity gave improved assemblies (**Table 1**). *Celera Assembler* is open source, and available for the assembly of the genomes proposed in this project (http://wgs-assembler.sourceforge.net/; Denisov *et al.,* 2008). Accordingly, both sequencing centers agreed on the JCVI assemblies submitted to GenBank as being the canonical assemblies for the M and S annotation and analysis (GenBank Accession numbers ABKP00000000 and ABKQ00000000, respectively). The WUGSC assemblies can also be found on the WUGSC website http://genome.wustl.edu/genome.cgi?GENOME=Anopheles%20gambiae%20M.

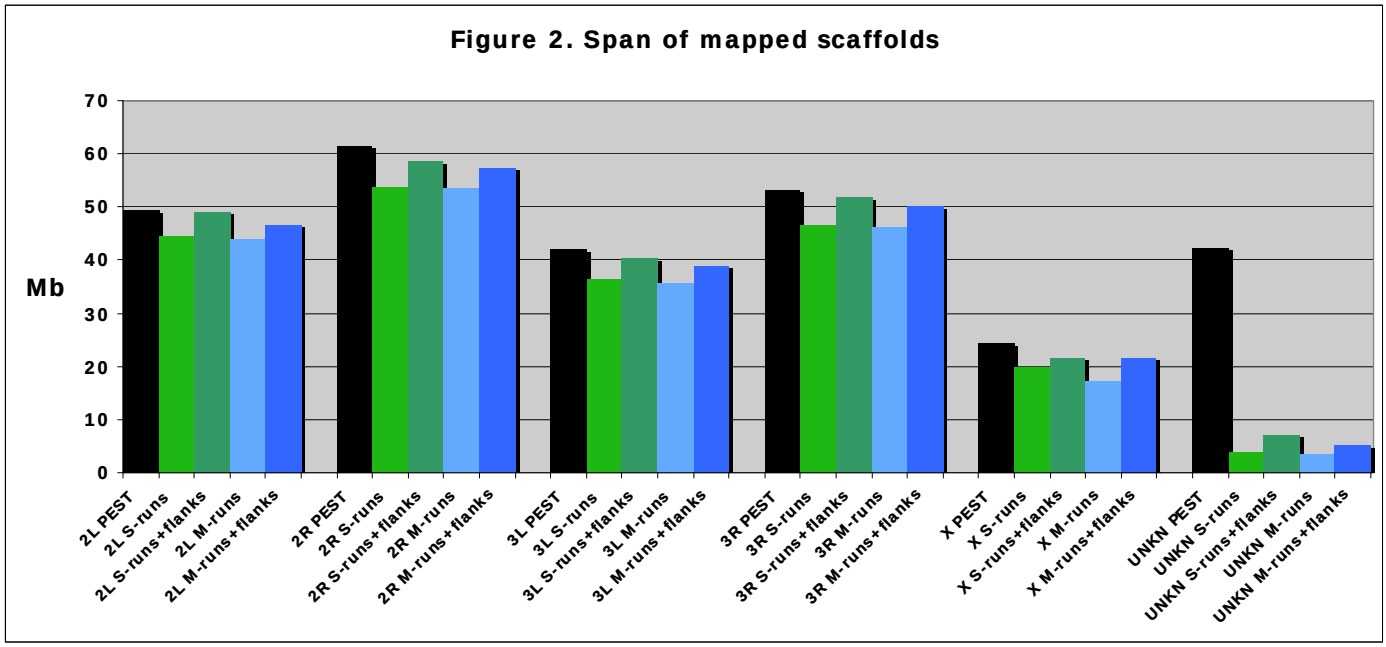
	S form (G4 assembly)	M form (M5 assembly)
No. scaffolds >200kb	77	79
Total span of scaffolds >200kb	207.5 Mb	192.2 Mb
No. scaffolds >2 kb	1,462	1,954
Total span of scaffolds >2kb	221.0 Mb	213.8 Mb
No. scaffolds (all lengths)	13,050	10,525
Total span of all scaffolds	236.4 Mb	224.5 Mb
Contig N50[1]	42,542	24,512
Scaffold N50[1]	3.9 Mb	4.4 Mb
Ave. read coverage of contigs	6.4X	6.0X

Table 1. Assembly statistics for M and S forms of *An. gambiae* generated by *Celera Assembler*.

[1]N50 contig or scaffold size defines the size above which 50% of the assembly is found

The DNA source for the S form was the *An. gambiae* Pimperena colony (available through *MR4*). This was established from mosquitoes collected in Pimperena, Mali in November 2005, by combining ~5 isofemale families (~20 genome-equivalents) molecularly identified as *An. gambiae* S. The DNA source for the M form was *An. gambiae* Mali-NIH colony (available through *MR4*). This derived from mosquitoes collected in a village near Niono, Mali in June 2005 and was established by combining ~80 isofemale families molecularly identified as *An. gambiae* M. Both colonies have been maintained at moderately high numbers per generation, with no specific effort at inbreeding to reduce genetic variation; each carries considerable polymorphism. In these characteristics, the Pimperena and Mali-NIH colonies are like the other *Anopheles* colonies proposed for sequencing. Thus, their assembly results are likely to be predictive of the results that will be obtained from the other species.

In comparison with the *An. gambiae* PEST genome, there is a high level of concordance for both S and M genomes, with ~99% of annotated PEST genes being identified in each of the latter two assemblies. Moreover, in 1-to-1 mappings of the assembly-to-assembly comparisons to PEST (*i.e.*, each M or S scaffold mapped to only one PEST chromosome; see below), the summed lengths of M (or S) scaffolds that map to PEST chromosomes 2L, 2R, 3L and 3R are 93-100% of the PEST chromosome lengths (**Fig 2**). For X, owing to the lower sequence coverage, the summed lengths are only 89% for both M and S.



These preliminary analyses of the M and S assemblies suggest that despite relatively high levels of variation in the genomic DNA of the two colonies, assembly has not been compromised.

4.4. Choice of species (Table 2). Three criteria were considered when selecting species for genome sequencing: (1) availability of colonies, (2) vector status, and (3) degree of evolutionary divergence from *An. gambiae* (see Fig 3, below), the reference genome and anchor of this project. **The single overriding criterion for selection of species for sequencing is the availability of colonies**, as genome sequencing cannot be contemplated for species that have not been colonized. This paramount consideration explains why important vectors and otherwise obvious choices such as *An. darlingi* are not included, as this and many other species have not been successfully colonized (although not for a lack of effort). Of the few anopheline species for which colonies are available, vector status relative to *P. falciparum* and *P. vivax* was considered (**Table 2**). Represented in this project is a range of vectors, from those of major importance (*e.g.*, *An. funestus*, *An. gambiae* and *An. arabiensis* in Africa and *An. farauti* in PNG) to lesser vectors, including a non-vector species in the *An. gambiae* complex.

Table 2. Proposed Thirteen Genomes Cluster for Genus *Anopheles*. Each genome is estimated as ~250Mb. Abbreviations: *MR4*, malaria repository; *P.f.* and *P.v., Plasmodium falciparum* and *P. vivax*, respectively.

Priority	Species	Classification	Colony (Source)	Parasites	Resources
		Subgenus Cellia			
TIER 1		**Series**			
	1. *An. arabiensis*	**Pyretophorus** (Gambiae complex)	Dongola (MR4)	Vector *P.f.* in Africa	
	2. *An. quadriannulatus*	(Gambiae complex)	SKUQUA (MR4)	Non-vector	BAC library
	3. *An. merus*	(Gambiae complex)	OPHANSI (MR4)	Minor vector *P.f.* in Africa	
	4. *An. epiroticus* (formerly *An. sundaicus* species A)		(Dr. Ho Dinh Trung, National Institute of Malariology, Parasitology & Entomology, Hanoi VIETNAM)	Moderate vector P.f and P.v. in Asia	
TIER 2	5. *An. stephensi*	**Series Neocellia**	STE2 (*MR4*)	Vector *P.f.* and *P.v* in Indian subcontinent	BAC library
	6. *An. maculatus* (species B)		(Dr. LEE HAN LIM, Head, Medical Entomology Unit, Infectious Diseases Res Ctr, WHO Collaborating Ctr for Vectors, Dean, DAP&E School, Institute for Medical Research, Kuala Lumpur)	*Vector P.f.* and *P.v* in Asia	
	7. *An. funestus*	**Series Myzomyia**	FUMOZ (Dr. Maureen Coetzee, Vector Control Reference Unit, South Africa)	Vector of *P.f.* in Africa	BAC library, cDNAs
	8. *An. minimus s.s.*(species A)		MINIMUS1 (MR4)	Moderate vector *P.f.* and *P.v.* in central and east Asia	
	9. *An. culicifacies* A		(Pasteur Institute of Iran, Tehran, Iran)	Vector of *P.f.* and *P.v.* in Indian subcontinent	
	10. *An. farauti* 1	**Series Neomyzomyia**	FAR1 (*MR4*)	Vector of *P.f.* and *P.v.* in PNG	
	11. *An. dirus s.s.* (species A)		WRAIR2 (*MR4*)	Vector of P.f. and P.v. in East Asia	
TIER 3	12. *An. atroparvus*	**Subgenus Anopheles**	EBRO (*MR4*)	Former vector in Europe	
	13. *An. albimanus*	**Subgenus Nyssorhynchus**	STECLA (*MR4*)	Minor vector of *P.v.* and *P.f.* in Latin America	

The choice of anopheline species also takes into account what is known of anopheline phylogeny and the evolutionary range of divergences (**Fig 1** and **Fig 3**, below), though it should be emphasized that this is not yet as well defined as within genus *Drosophila*. Available data support the monophyly of the six *Anopheles* subgenera (Krzywinski, Besansky, 2003, and refs within). Both morphological and molecular data support a sister-group relationship between subgenera *Nyssorhynchus* and *Kerteszia*, and molecular data suggest the same for subgenera *Cellia* and *Anopheles* (**Fig 1**). Given the evolutionary relationships within subgenus *Cellia* (**Fig 3**), there are some clear similarities between the 13 *Anopheles* species proposed and the 12 Drosophila species that have been sequenced (*Drosophila* 12 Genomes Consortium, 2007). Most notable is the focus on closely related species in the *An. gambiae* complex, which is analogous to the *D. melanogaster* group of species, and the selection of additional *Anopheles* species that step progressively further out within the subgenus *Cellia* and beyond, from the core species complex. Sampling encompasses species outside of the complex but in the same taxonomic Series (*Pyretophorus*), as well as representatives of three other Series within *Cellia* (*Myzomyia*, *Neocellia* and *Neomyzomyia*), and finally includes representatives of two different subgenera, *Anopheles* and *Nyssorhynchus*. (Further information concerning taxonomic classification within genus *Anopheles* can be found in Harbach 1994, 2004). One major difference between the *Anopheles* and

Drosophila clusters is that the deepest *Anopheles* branch in the cluster may extend to more than 100 MY, while it is probably only ~60 MY for the *Drosophila* cluster.

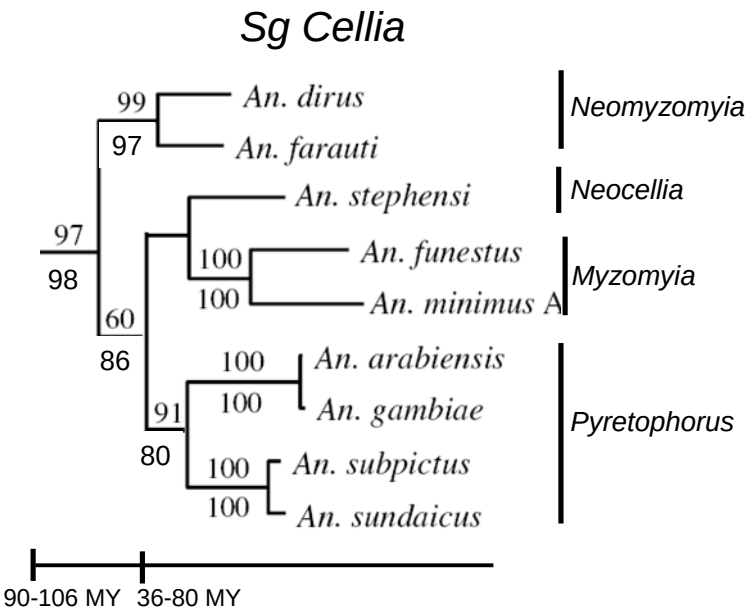


Figure 3. Evolutionary relationships within Subgenus *Cellia* inferred from maximum likelihood (ML) and maximum parsimony (MP) analyses of combined rDNA (18S and 28S) and mtDNA (COI and COII). Numbers above and below branches indicate ML and MP bootstrap support, respectively. Adapted from Sallum *et al*., 2002.

4.5. Suggested sequencing strategy and priorities. We propose 8x coverage of each genome, sufficient to allow (1) generation of a high quality assembly and (2) comparative genomics of non-protein coding control regions. We propose three tiers of sequencing priorities.

- The first tier (blue in **Table 2**) includes the species most closely related to *An. gambiae* (*An. arabiensis, An. merus*, and *An. quadriannulatus* and *An. epiroticus*). Of particular interest in this grouping are rapidly evolving differences between the non-vector *An. quadriannulatus* and vectors of intermediate and high importance within and outside of the species complex.

- The second tier (yellow in **Table 2**) includes a sample of seven other species representing three additional taxonomic Series within *Cellia*. Together with the species in Tier 1, Tier 2 species represent most of the evolutionary diversity present within subgenus *Cellia*, and divergence times vary from ca. 10,000 years (*An. gambiae*) up to 40-80 MY.

- The third tier (orange **in Table 2**) provides one species from each of the two largest remaining subgenera in genus *Anopheles* after *Cellia*: subgenus *Anopheles* and subgenus *Nyssorhynchus*. Comparisons between these Tier 3 species and those within *Cellia* in Tiers 1 and 2 are important because of the extended evolutionary divergence times represented, up to ~100 MY. Although these last two species also are (or were) important vectors of human malaria, they are outgroups with respect to divergence inside *Cellia*. Issues that can be addressed across this broad time-frame include maintenance of orthology in olfactory receptors that may be involved in host-seeking, as well as maintenance of orthology in the immune system.

Overall, the 13 proposed species will provide information about how the genomic determinants of vectorial capacity evolve over evolutionary time spanning four orders of magnitude, from 10,000 years up to 100 MY.

4.6. EST Sequencing. EST sequencing is essential for accurate annotation of the proposed genomes, particularly those outside of the *An. gambiae* complex and distantly related to the reference genome. We propose to sequence at least 200,000 ESTs per species based on normalized cDNA pools comprised of multiple developmental stages and malaria-infected or uninfected tissues.

5. Community input and support:

As documented in the **Appendix**, more than 70 scientists form the arthropod community have given enthusiastic and positive feedback for this proposal.

6. Data Release and relevant repositories for strains and sequence data:

Sequences emerging from this project will be rapidly released to the public via GenBank and other archival repositories, in accordance with NHGRI (http://www.genome.gov/25521732, http://www.genome.gov/25521732) and NIAID (http://www.niaid.nih.gov/dmid/genomes/mscs/data_release.htm) policies. The following two NIAID-funded contracts should play key roles in colony and data management:

6.1. 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, as well as standardized assays using well-characterized and renewable reagents (www.malaria.mr4.org). ATCC currently holds the contract, with a subcontract to CDC Foundation (Robert Wirtz) for the provision of anopheline reagents, including living laboratory colonies. All except three species proposed for genome sequencing are 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 unfortunately happened with *An. gambiae* PEST.

6.2. VectorBase. *VectorBase* (http://www.vectorbase.org/) is a web-based resource supported by NIAID and serving the scientific community. It houses, manages and displays invertebrate vector genomes, their predicted gene sets, and associated information such as microarray data, and provides for browsing and data-mining (Lawson *et al.*, 2007). *VectorBase* provides first-pass annotation of new genome sequences and re-annotation of existing genome sequences. The database manages, displays, and analyzes 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* assumes responsibilities for developing, maintaining and updating internationally recognized Reference Data Sets for the organisms included. It will display and manage all anopheline genomes proposed for sequencing here.

7. Conclusion:

The entire community of vector biologists and parasitologists who are seeking novel solutions for controlling malaria will benefit from the availability of additional anopheline genome sequence data. *An. gambiae* represents *the* model organism for the study of malaria vector biology and control. Our ultimate goal is to extract from its genome information that will expedite development of new malaria control methods, both chemical and genetic, that will alter vectorial capacity. Two problems stand in the way of our ability to fully exploit this genomic resource. First, the *An. gambiae* genome is not fully annotated. Within coding sequences, there are false positive gene calls, false negative gene calls, and mis-annotations; within non-coding sequences, regulatory elements are very poorly defined and notoriously difficult to identify. Second, a lack of comparative data prevents the discovery of targets of interest. Focused studies on *An. gambiae* remain essential, but comparative genomics of related species is a powerful toll for enhanced annotation and identification of targets of interest. Analyzing the genome sequences of multiple related species within 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]***Anopheles* Genomes Cluster Committee:**

Nora J. Besansky (Chair), University of Notre Dame, USA
Michael Ashburner, Visiting Group Leader and former Joint-Head, European Bioinformatics Institute, UK; Development and Oversight, FlyBase; Cambridge University, UK.

Jane M. Carlton, Co-chair, joint NHGRI/NIAID Eukaryotic Pathogens and Disease Vectors Target Selection Working Group, Department of Medical Parasitology, New York University Langone Medical Center, USA
Maureen Coetzee, Head, Vector Control Reference Unit, National Institute for Communicable Diseases, South Africa
Frank H. Collins, PI of VectorBase, University of Notre Dame, USA
Alessandra della Torre, University of Rome, Italy
Ralph E. Harbach, Head, Mosquitoes Programme, The Natural History Museum, London
Janet Hemingway, Director, Liverpool School of Tropical Medicine, UK
Jeffrey R. Powell, Professor, Ecology and Evolutionary Biology, Yale University, USA
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
Robert Wirtz, *MR4* subcontractor for vectors, CDC, USA

Acknowledgements: The *Anopheles* white paper has been reviewed by the Eukaryotic Pathogens and Disease Vectors Target Selection Working Group; we would like to thank this group for constructive criticism. In particular, we acknowledge the following representatives for their major input: Bruce Birren, Frank Collins, Dan Hartl, and Jim Mullikin.

REFERENCES

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

Besansky NJ, Hill CA, Costantini C (2004) No accounting for taste: host preference in malaria vectors. *Trends Parasitol* **20**, 249-251.

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.

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

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.

Denisov, G., B. Walenz, A.L. Halpern, J. Miller, N. Axelrod, S. Levy, and G. Sutton. 2008. Consensus generation and variant detection by Celera Assembler. *Bioinformatics* **24:** 1035-1040.

Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*. **450**, 203-118.

Gibbs RA, Nelson DL (2003) Human genetics. Primate shadow play. *Science* **299**, 1331-1333.

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.

Habtewold, T., M. Povelones, A.M. Blagborough, and G.K. Christophides. 2008. Transmission blocking immunity in the malaria non-vector mosquito Anopheles quadriannulatus species A. *PLoS Pathog* **4:** e1000070.

Hallem EA, Nicole Fox A, Zwiebel LJ, Carlson JR (2004) Olfaction: mosquito receptor for human-sweat odorant. *Nature* **427**, 212-213.

Harbach RE (1994) Review of the internal classification of the genus *Anopheles* (Diptera: Culicidae): the foundation for comparative systematics and phylogenetic research. *Bulletin of Entomological Research* **84**, 331-342.

Harbach RE (2004) The classification of genus *Anopheles* (Diptera: Culicidae): a working hypothesis of phylogenetic relationships. *Bull Entomol Res* **94**, 537-553.

Harbach RE, Kitching IJ (2005) Reconsideration of anopheline mosquito phylogeny (Diptera: Culicidae: Anophelinae) based on morphological data. *Systematics and Biodiversity* **3**, 345-374.

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.

Korochkina S, Barreau C, Pradel G*, et al.* (2006) A mosquito-specific protein family includes candidate receptors for malaria sporozoite invasion of salivary glands. *Cell Microbiol* **8**, 163-175.

Krzywinski J, Besansky NJ (2003) Molecular systematics of Anopheles: from subgenera to subpopulations. *Annu Rev Entomol* **48**, 111-139.

Krzywinski J, Grushko OG, Besansky NJ (2006) Analysis of the complete mitochondrial DNA from Anopheles funestus: an improved dipteran mitochondrial genome annotation and a temporal dimension of mosquito evolution. *Mol Phylogenet Evol* **39**, 417-423.

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.

Lawson D, Arensburger P, Atkinson P*, et al.* (2007) VectorBase: A home for invertebrate vectors of human pathogens. *Nucleic Acids Res* **35**,D503-D505.

Molineaux, L. and G. Grammicia. 1980. *The Garki Project. Research on the Epidemiology and Control of Malaria in the Sudan Savanna of West Africa*. World Health Organization, Geneva, Switzerland.

Pombi M, Stump AD, Della Torre A, Besansky NJ (2006) Variation in recombination rate across the X chromosome of Anopheles gambiae. *Am J Trop Med Hyg* **75**, 901-903.

Rai KS, Black WC (1999) Mosquito genomes: structure, organization, and evolution. *Adv Genet* **41**, 1-33.

Riehle MM, Markianos K, Niare O*, et al.* (2006) Natural malaria infection in Anopheles gambiae is regulated by a single genomic control region. *Science* **312**, 577-579.
Sallum MAM, Schultz TR, Wilkerson RC (2000) Phylogeny of Anophelinae (Diptera Culicidae) based on morphological characters. *Ann. Entomol. Soc. Am.* **93**, 745-775.
Sallum MAM, Schultz TR, Foster PG*, et al.* (2002) Phylogeny of Anophelinae (Diptera: Culicidae) basedon nuclear ribosomal and mitochondrial DNA sequences. *Syst. Entomol.,* **27**, 361-382.
Sattabongkot J, Tsuboi T, Zollner GE, Sirichaisinthop J, Cui L (2004) Plasmodium vivax transmission: chances for control? *Trends Parasitol* **20**, 192-198.
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.
Vernick KD, Oduol F, Lazzaro BP*, et al.* (2005) Molecular genetics of mosquito resistance to malaria parasites. *Curr Top Microbiol Immunol* **295**, 383-415.

Appendix

Date: Wed, 23 May 2007 16:16:30 -0400
To: announcements@vectorbase.org
From: Nora Besansky <nbesansk@nd.edu>(by way of Katie Cybulski <kmerz@nd.edu>) Subject: [VectorBase Announcement] 12 Anopheles Genomes White Paper

To: Scientific community members with an interest in vectors of human disease and the pathogens they transmit

Re: White paper proposal for sequencing 12 additional anopheline genomes

Dear Scientists:

NIH (NHGRI and NIAID) are evaluating proposals for whole genome sequencing (WGS) of additional vectors and pathogens. There is growing recognition of the power of comparative analysis in identifying targets of biological and epidemiological significance. Among other vectors that are being considered and that will eventually enter the pipeline are a set of 12 additional anopheline species as outlined in the attached white paper. This paper is co-authored by members of the Anopheles Genomes Cluster Committee (see attached document), and has been endorsed by members of the Eukaryotic Pathogens and Disease Vectors Target Selection Working Group, but we are soliciting the endorsement of the wider scientific community including you and and any of your colleagues who we may have inadvertantly left off of the mailing list. In particular, we are trying to assess the size and strength of the community of potential users of these sequence data. Will these data be useful to you in your research? If so, we would like your feedback.

A word on the choice of 12 species: in the white paper, you will find that choice of species was seriously constrained by what is available in colony as well as considerations of evolutionary depths. As for the number of species (n=12 plus A. gambiae M and S), Jeffrey Powell has emphasized the following arguments. Thirteen is a good number to do comparative genomics. It allows (1) outgroups at different divergence times that makes reconstruction of ancestral states much more rigorous; inferring ancestral states in non-vectors is crucial in identifying the derived states associated with disease transmission (human blood choice, association with human habitats, ability to support development of a pathogen, etc.). (2) Having multiple focal species (vectors) allows for replicates in discerning derived traits associated with disease transmission. Three proposed species regularly transmit falciparum (gambiae, arabiensis, and funestus, plus others?) and others transmit primarily vivax. In addition, there are close relatives that don't transmit. So independent parallel genomic changes allows one to eliminate possible false positives, or at least allow one to ascertain whether the genetic changes in traits associated with disease transmission have occurred multiple times or whether each time something independent has occurred. This allows evaluation of whether a genetic control strategy that works in one vector can be transferred to others, or whether each vector is unique in acquiring vector-associated traits and need to be dealt with separately. Having 12 (or 14?) taxa with which to do the comparative genomics is the minimum that would allow rigorous analysis. In fact, one could imagine that the complete genomes of these 12/14 allows one to narrow the search and many more species (based on field samples) could be selectively sequenced for particular parts of the genome to fill in holes. But the 12/14 complete species genomes are a minimum on which to build framework of genome evolution in Anopheles that can be elaborated on selectively to narrow in on particular traits.

On behalf of the Anopheles Genomes Cluster Committee and the Eukaryotic Pathogens and Disease Vectors Target Selection Working Group, I thank you in advance for your time and your feedback.

Best wishes,

Nora Besansky

NORA J. BESANSKY, Professor
Center for Global Health and Infectious Diseases
Dept. Biological Sciences
317 Galvin Life Sciences Bldg
University of Notre Dame
Notre Dame IN 46556-0369
Tel: 574-631-9321

Fax: 574-631-3996
e-mail: nbesansk@nd.edu

List of Respondents as of 6 June 2007. Most of the more than 70 people on the list below are scientists who work with arthropod vectors of human pathogens. A small percent of them work on vector-borne pathogens, in the area of genomics, or in the field of evolutionary biology (e.g., the consortium listed at the end).

Diabate Abdoulaye
Serap Aksoy
Peter Atkinson
Carolina Barillas-Mury
Jesus Martinez Barnetche
John Beier
Harald Biessmann
Daniel Boakye
Henk Braig
William Brogdon
Gisella Caccone
Don Champagne
Anton Cornel
Mamadou Coulibaly
John Dame
Jose de la Fuente
David Denlinger
George Dimopoulos
Mike Eisen
Paul Eggleston
Ann Fallon
Mike Ferdig
Desmond Foley
Sarjeet Gill
John Gimnig
Katia Gondim
Fred Gould
Robert Gwadz
William Hawley
Janet Hemingway
Stephen Higgs
Catherine Hill
Robert Holt
A.A. James
J. Spencer Johnston
Brian Kay
Jonathan Kayondo
Mark Klowden
Jaroslaw Krzywinski
Robert Lane
Angela Lange
Brian Lazzaro
Kenneth Linthicum
Christos Louis
Lucian Moreira
Roger Nasci
Doug Norris
Ian Orchard
Susan Paskewitz
Barry Pittendrigh
Mihai Pop
Hilary Ranson
Jose M.C. Ribeiro
Hugh Robertson
Patricia Romans
Dick Sakai
Abhimanyu Sarkar
Tom Scott
Maria Sharakhova
Igor Sharakhov
Cheolho Sim
Frédéric Simard
Steven Sinkins
Daniel Sonenshine
Sarala Subbarao
Antonio R.L. Teixeira
Richard Titus
Jake Tu
Ken Vernick
John Vulule
Judy Willis
Mark Yandell
Larry Zweibel

Consortium: Mike Eisen and his group; Tom Little; Darren Obbard; Matt Hahn; Andy Clark; Brian Lazzaro