

**Shotgun sequencing, assembly, and annotation of an *Anopheles funestus*
bacterial artificial chromosome (BAC)**

Cassone BJ¹, White BJ¹, Lobo N

University of Notre Dame - Center for Tropical Disease Research and Training

¹ Both authors contributed equally to this report

Abstract

Anopheles funestus is a malaria vector found on Sub-Saharan African, which has the potential to cause significant infection rates. Developing novel malarial control strategies will be considerably aided by completing a BAC library of *A. funestus*. In this study, we shotgun sequenced and assembled a BAC clone *aflb* of *A. funestus*. We employed *ab initio* gene prediction techniques to identify putative genes in the BAC sequence. In addition, we employed programs to annotate the putative genes based on comparisons with the genomes of *Anopheles gambiae*, *Drosophila melanogaster*, and *Homo sapiens*. We found a total of 12 predicted genes in our BAC sequence. *H. sapiens* had typically the most closely-matched genes relative to our putative genes, whereas *A. gambiae* had the lowest contrasting scores. This is likely because the *H. sapiens* genome is more thoroughly annotated. We were able to speculate function of only one of our predicted genes with reasonable accuracy. The gene scored a reasonably high match when contrasted against the *A. gambiae* genome, and is comprised of three putative conserved domains. Inferring function of only one of the 12 predicted genes may be indicative of poor prediction employing *ab initio* techniques. Based on our results, thorough annotation of the genome through comparative genomics will be a very difficult, and may have to rely more on ESTs and cDNAs in future studies.

Introduction

Completion of the *Anopheles gambiae* genome in 2002 gave the scientific community a vital tool to be used in the development of novel, vector-based control strategies of malaria in Africa (Holt et al. 2002). However, comprehensive malaria control in Africa requires not only high quality genomic data for *A. gambiae*, but also that of the other major African malaria vector *A. funestus*. Although *A. gambiae* continues to dominate the landscape of malaria research, recent studies have indicated that *A. funestus* may be just as lethal a vector as *A. gambiae*. In fact, Michel and colleagues (2005a) found *Plasmodium falciparum* infection rates up to 9.1% in geographically diverse *funestus* populations. These infection rates are significantly higher than any reported rates in *A. gambiae*.

A grant to sequence the genome of *An. funestus* has been submitted to the National Human Genomics Research Institute (NHGRI), and it appears that the genome will likely be sequenced in the coming years. Critical to the success of any genome project is a high quality Bacterial Artificial Chromosome (BAC) library. Other major projects, including the mapping of insecticide resistance in *A. funestus* will rely heavily on a complete BAC library. For these reasons, a high quality 10x BAC library has been constructed for *A. funestus*. Plans to physically map and end sequence 2,000 of these BACs have been developed and are currently being carried out. In this study, we report the full sequencing of BAC *af1b* from this BAC library. We also report results from our preliminary annotation of the BAC sequence using a variety of gene finding and annotating algorithms.

Fully sequenced BACs from *A. funestus* will allow us to predict the amount of genetic variation present in the colony (FUMOZ) from which the BACs were established. If there is too much variation in this colony, it may not be feasible for whole-genome sequencing. Previous results have indicated that the mitochondrial DNA (mtDNA) in the colony represents two distinct clades, corresponding to the two wild mtDNA haplotypes observed in population genetics studies (B. White, unpublished data; Michel et al 2005b). Once other BACs from the same physical area of the genome are sequenced, we will be able to predict whether or not this mtDNA variation extends into the nuclear DNA as well.

Further, experimental annotation of the fully sequenced BACs will allow us to determine which organism's genes (or combination of organisms' genes) produce the best algorithms for gene finding in *A. funestus*. Such knowledge will be crucial when annotation of the *A. funestus* genome is necessary. Also, full sequencing and annotation of BACs can be extrapolated to predict the gene and transposon content of the whole genome, knowledge that will be extremely useful in annotation of the entire genome.

Material & Methods

Shotgun Sequence BAC Assembly

Anopheles funestus BAC *af1b* was constructed by The Institute for Genomic Research (TIGR; <http://www.tigr.org/>) using standard protocols in preparation of a BAC library for this species. Genetic material was obtained from pupae of the *An. funestus* colony FUM0Z established by Maureen Coetzee and currently maintained by both Coetzee and Frank Collins. The BAC was cloned and then fragmented using sonication techniques into 1756 shorter, random sequences (includes forward and reverse strands). The DNA fragments were cloned into plasmid vectors (*pHos2*), and these vectors were subsequently shotgun sequenced using standard big dye 3.1 chemistry on ABI DNA Analyzers (various models). The shotgun reads were trimmed using Seqman (<http://www.dnastar.com/web/r13.php>), which automatically incises known vector sequences from the ends of shotgun reads.

After the sequence reads were trimmed, the shotgun sequences were assembled using the computer program Phrap (P. Green unpublished; <http://www.genome.washington.edu/UWGC/analysistools/Phrap.cfm>). This program was chosen since: 1) it allows use of the entire read not just the highest quality part; 2) uses a combination of user-supplied and internally computed data quality information to improve accuracy of assembly in the presence of repeats, and; 3) constructs contig sequence as a mosaic of the highest quality parts of reads, rather than a consensus. A large contig was generated, which corresponded to the sequence of the BAC. Many smaller contigs were also produced, which represented contamination from either *Homo sapiens* or bacterial DNA. To verify, the contamination sequences (either human or bacterial respectively) were compared against the *H. sapiens* genome at the Ensembl Genome Browser (www.ensembl.org) and the UniVec database (<http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html>) using BlastN for both. The contaminant contigs were omitted from further analysis.

Gene Prediction

Locating the positions of all genes and determining their structures is the first step toward deciphering the functions of a sequenced genome (Yao et al. 2005). Gene prediction software programs were employed to locate genes in the assembled shotgun sequenced BAC clone from the *Anopheles funestus* genome. We implemented the *ab initio* programs Genscan version 1.0 (Burge and Karlin 1997) and Genemark.hmm version 2.2 (Lukashin and Borodovsky 1998). We chose these gene predicting programs because they have been extensively used and presumed to be amongst the most accurate (Tabaska et al. 2001). *Ab initio* gene prediction uses statistical and computational methods to detect coding regions, splice sites, and start and stop codons in the genomic sequence (Yao et al. 2005). Both programs predict gene models that can be single or multiple in a genomic sequence and a predicted exon is indicated as initial (starting with the initiation codon and ending with a donor site), internal (starting with an acceptor site and ending with the donor site), terminal (starting with an acceptor site and ending with the stop codon), or single (starting with the initiation codon and ending with the stop codon) exon in the output. Furthermore, Genemark.hmm and Genscan use explicit state

duration HMM, which is often used in gene-finding programs (e.g. Genie). The optimal gene candidates selected by the HMM and dynamic programming are further processed by a ribosomal binding site recognition algorithm. Both programs were run through their websites (Genscan: <http://genes.mit.edu/GENSCAN.html>; Genemark.hmm: <http://opal.biology.gatech.edu/GeneMark/>) by using their organism specific default parameters (GenScan: vertebrate; Genemark.hmm: *Anopheles gambiae*, *Drosophila melanogaster*, and *Homo sapiens*) to obtain the prediction results for the *A. funestus* sequence. The programs automatically translate the submitted nucleotide sequence into its amino acid sequence and accounts for all six open reading frames.

Gene Annotation

Putative genes were annotated by implementing a variety of techniques. The BlastP (<http://www.ncbi.nlm.nih.gov/blast/>) and Pfam version 18.0 (<http://pfam.wustl.edu/>) software programs were run through their websites to identify putative conserved domains within the predicted genes. The predicted genes were subsequently annotated by employing the Ensembl Mosquito BlastView version 35 (Stalker et al. 2004; http://www.ensembl.org/Anopheles_gambiae/blastview) computer program. In this step, the amino acid sequences of predicted genes determined by Genscan and GeneMark.hmm were compared to the peptide database of *A. gambiae*, *D. melanogaster*, and *H. sapiens* to identify the protein family each gene is most likely associated with. *A. gambiae* was chosen since it is the most closely-related species to *A. funestus* of the genomes provided in Ensembl. *D. melanogaster*, and *H. sapiens* were selected because both have well annotated genomes relative to most species. The BlastP and Swiss-Prot (<http://www.expasy.org/tools/blast/>) programs were also used to annotate the putative genes. Both programs compare the given amino acid sequences to those of a diverse group of taxa to determine the most similar gene, and provides the function of that gene. In each program, a statistical score and an expect value parameter (*E-val*) is assigned to a match between two sequences. The *E-val* describes the number of hits one can "expect" to see just by chance when searching a database of a particular size, and this value decreases exponentially with the score. Finally, we implemented the RepBase software program (Jurka 1998; <http://www.girinst.org/replib/update/index.html>), which is a database of prototypic sequences representing repetitive DNA from different eukaryotic species. This was done in order to determine the putative transposon content in our BAC sequence.

Results

The assembled shotgun sequenced BAC contig from the *Anopheles funestus* genome was comprised of 92,285 base pairs. Nucleotide frequencies of the data set were moderately A and T biased (frequencies: A = 25.2%, C = 20.0%, G = 20.6, T = 29.8%).

Putative Genes

GenScan identified 12 putative genes, whereas GeneMark.hmm predicted 8, 10, and 26 genes when contrasted with *Anopheles gambiae*, *Drosophila melanogaster*, and *Homo sapiens*, respectively. In all cases, the Ensemble statistical scores of the putative genes that were identified by both GenScan and GeneMark.hmm were higher using the former program; thus annotation focused exclusively on the predicted genes of GenScan. The gene location, exon base pair length, and most plausible annotations of all 12 possible genes are displayed in Table 1.

In general, putative genes identified by GenScan had the higher Ensemble statistical scores when compared with *H. sapiens* than with *A. gambiae*, and *D. melanogaster*, whereas *A. gambiae* typically had the lowest *e* values. All Ensemble comparisons had considerably higher scores than those of BlastP. This is likely because even though Ensemble genome browser contains of significantly less species for comparisons than BlastP, it has a much more complete genome record of the species it does contain.

Gene 1 had the highest scores of all 12 genes regardless of which species was compared. The *A. gambiae* contrast had the highest statistical score, at 794 (*E*-val = 4.6 e-69); whereas, the scores for *D. melanogaster*, and *H. sapiens* were also considerable (227, *E*-val = 2.1e-11; 332, *E*-val = 8.0e-30, respectively). The consensus annotation from *A. gambiae* was RNA directed DNA polymerase from mobile element jockey EC_2.7.7.49 reverse transcriptase (PB B PA A EG: BAC 5 P and reverse transcriptase were identified for *D. melanogaster*, and *H. sapiens*, respectively). Furthermore, this gene was identified at multiple locations in the genome sequence of all three species comparisons. Three putative conserved domains were detected within gene 1: an endonuclease/exonuclease/phosphatase, reverse transcriptase, and RNaseH (see Figure 1).

The annotations of all other genes were much more ambiguous and have statistical scores less than those of all species contrasts of gene 1. Moreover, no putative conserved domains were detected for all other putative genes. Gene 3 had the second highest score at 207 (*E*-val = 9.6e-12), when compared with *D. melanogaster*, as well as the second highest score for *H. sapiens* at 203 (*E*-val = 1.3e-10). The consensus annotation of this possible gene was a PB B PA A P papillote for *D. melanogaster*, and unknown function for *H. sapiens*. Gene 5 had the second highest scores for *A. gambiae* at 171 (*E*-val = 4.7e-08), and was consensually annotated as an abnormal spindle microcephaly associated homolog. Gene 11 was the least likely to be a genuine peptide, as all three species comparisons had scores under 105 and *E*-vals over 0.005.

By implementing RepBase on the entire BAC sequence, we identified a total of 14 putative transposons. Moreover, one of the transposons had inverted repeats characteristic of a type II transposon.

Discussion

Annotation of *A. funestus* BAC *af1b* was attempted using gene prediction algorithms based off of the genomes of *Drosophila melanogaster*, *Homo sapiens*, and *Anopheles gambiae*. Surprisingly, the predictions with the highest confidence, or score, were generally obtained when using the algorithms built from the *H. sapiens* genome (Table 1). The organism that typically produced the second highest confidence in gene prediction was *D. melanogaster*, while the organism that often scored the least confidence was *A. gambiae*. Upon first look these results make little sense since one would expect the algorithms based off of more closely-related organisms (*A. gambiae* and *D. melanogaster*) to achieve more plausible than those algorithms based off more evolutionarily distant organisms (*Homo sapiens*). However, these results are not surprising given the current state of annotation in the respective organisms. The annotation of the *An. gambiae* genome remains very poor, with many false positive and false negatives through out the genome (Kriventseva et al 2005). On the other hand, annotation of the *D. melanogaster* genome is more thorough, while the human genome represents the gold standard in annotation. In other words, we have found that more thoroughly annotated genomes from distant organisms work better as a foundation for gene predictions than more poorly annotated genomes from more closely related organisms.

We were able to speculate a function of only one predicted gene with reasonable accuracy (gene 1; Table 1). This gene scored a reasonably high match when contrasted against the *A. gambiae* genome. It is comprised of three putative conserved domains, found in multiple locations in the *A. gambiae* genome, and its consensus annotation is a reverse transcriptase (i.e. transposon). Inferring function of only one of the 12 predicted genes is surprising; when annotating the other 11 genes we found no putative conserved domains or closely-related genes in comparison with the other three species. This is perhaps the result of poor prediction employing *ab initio* techniques. We may be trying to find homologs of genes that are incorrectly predicted. This would obviously limit our success in determining the function of these genes. Based on this BAC alone it appears that the gene content of the *A. funestus* genome will be similar to that of the *A. gambiae* genome. However, based on our results, it appears that a thorough annotation of the genome through comparative genomics will be a very difficult task. Indeed, the *A. funestus* genome annotation may have to rely more on ESTs and cDNAs than comparative genomics. On the other hand, we were able to identify 14 transposons or pieces of transposons with reasonable accuracy. It appears that the *A. funestus* genome may be rich in transposons and that their identification will be laborious, but entirely feasible.

Sequencing of other BACs from the same physical region as our BAC will provide great insight into the genetic variability that exists in the FUMOZ colony of *An. funestus*. Such sequencing is a prudent step before the initiation of any genome project. Such information will lead to an informed decision on whether or not to sequence the genome of *funestus* using this particular colony.

Literature Cited

- Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* **268**, 78-94.
- Jurka, J (1998). Repeats in genomic DNA: mining and meaning. *Curr Opin Struct Biol.* **8**, 333-337.
- Holt, RA et al (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129.
- Kriventseva, EV (2005) AnoEST: Toward *A. gambiae* functional genomics. *Gen. Res.* **15**, 893-899.
- Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. *Nucl Acids.* **26**, 1107-1115.
- Michel, AP et al (2005a) Molecular differentiation between chromosomally defined incipient species of *Anopheles funestus*. *Ins. Mol. Bio.* **14**, 375-387.
- Michel, AP et al (2005b) Rangewide population genetic structure of the African malaria vector *Anopheles funestus*. *Mol. Ecol.* **14**, 4325-4248.
- Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz H-R, Cox A (2004) The Ensembl website: mechanics of genome browser. *J Mol Biol.* **337**, 307-317.
- Tabaska JE, Davuluri RV, Zhang MQ (2004) Identifying the 3'-terminal exon in human DNA. *Bioinformatics.* **17**, 602-607.
- Yao H, Ling G, Yan, F, Borsuk LA et al. (2005) Evaluation of five ab initio gene prediction programs for the discovery of maize genes. *Pl Mol Biol.* **57**, 445-460.

Fig. 1. Putative conserved domains that have been detected for the putative gene 1 found in the BAC sequence of *Anopheles gambiae*.



Table 1. Gene location, base pair length, consensus annotation, and statistical scores of the 12 predicted genes in the *Anopheles funestus* BAC sequence. The statistical score is indicated for each consensus annotation of all three species contrasts (the E-val of each comparison is indicated in parentheses)

^a Base pair length represents only the exon regions of the gene

^b Represents the function with the highest statistical score for each of the three species comparisons (*Ag* = *Anopheles gambiae*; *Dm* = *Drosophila melanogaster*; *Hs* = *Homo sapiens*).

Gene	Gene Location (bps)	Base Pairs	Consensus Annotation ^b	Statistical Score
1	1188 - 6910	5130	<i>Ag</i> – Reverse transcriptase <i>Dm</i> – PB B PA A EG:BAC 5 P <i>Hs</i> - Reverse transcriptase	<i>Ag</i> – 794 (4.8e-69) <i>Dm</i> -227 (2.1e-11) <i>Hs</i> – 332 (8.0e-30)
2	12007 - 14667	382	<i>Ag</i> – Zinc finger <i>Dm</i> - Suppressor of TY 6 homolog chromatin <i>Hs</i> - Gas2 growth arrest specific	<i>Ag</i> – 113 (0.00052) <i>Dm</i> – 133(0.00010) <i>Hs</i> – 137 (9.8e-06)
3	14777 - 22770	1534	<i>Ag</i> – Peptidyl prolyl <i>cis trans</i> isomerase <i>Dm</i> - PB B PA A P Papillote <i>Hs</i> - Unknown	<i>Ag</i> – 150 (9.5e-06) <i>Dm</i> – 207 (9.6e-12) <i>Hs</i> – 203 (1.3e-10)
4	22859 - 24655	814	<i>Ag</i> – F box/WD repeat <i>Dm</i> - F box/WD repeat <i>Hs</i> – Sodium/hydrogen exchanger	<i>Ag</i> – 127 (6.0e-05) <i>Dm</i> – 144 (5.7e-06) <i>Hs</i> – 164 (5.0e-08)
5	25019 - 41891	1960	<i>Ag</i> – Abnormal spindle microcephaly associated homolog <i>Dm</i> – Myosin muscle <i>Hs</i> – Angiotensin converting enzyme precursor	<i>Ag</i> – 171 (4.7e-08) <i>Dm</i> – 172 (8.0e-08) <i>Hs</i> – 187 (2.4e-09)
6	42495 - 49223	1135	<i>Ag</i> – Sorting nexin 25 <i>Dm</i> - Homeobox <i>Hs</i> - TP53 target gene 5	<i>Ag</i> – 139 (2.0e-05) <i>Dm</i> – 173 (1.1e-08) <i>Hs</i> – 171 (5e-08)
7	49287 - 61457	1246	<i>Ag</i> – Retrovirus related polyprotein from transposon <i>Dm</i> – Zinc finger <i>Hs</i> – Adenylosuccinate synthetase	<i>Ag</i> – 137 (5.1e-05) <i>Dm</i> – 149 (3.9e-06) <i>Hs</i> – 143 (1.1e-05)
8	61600 - 64735	1060	<i>Ag</i> - E3 Ubiquitin ligase <i>Dm</i> – Longitudinals lacking <i>Hs</i> – Zinc finger	<i>Ag</i> – 155 (7.8e-07) <i>Dm</i> – 169 (2.5e-08) <i>Hs</i> – 159 (1.8e-07)
9	67886 - 70167	655	<i>Ag</i> - Phosphatidylinositol 4 kinase alpha <i>Dm</i> -1 3 glucan binding precursor <i>Hs</i> - MYB	<i>Ag</i> – 134 (2.3e-06) <i>Dm</i> – 158 (1.9e-07) <i>Hs</i> – 162 (2.1e-07)
10	73091 -	541	<i>Ag</i> – Transcription factor	<i>Ag</i> – 127 (4.9e-05)

	78056		<i>Dm</i> - Antigen	<i>Dm</i> - 137 (.00010)
			<i>Hs</i> - Plasminogen	<i>Hs</i> - 140 (1.7e-06)
11	79068 -	220	<i>Ag</i> - Paired box pax	<i>Ag</i> - 87 (0.068)
	83378		<i>Dm</i> - Metal regulatory transcription factor 1	<i>Dm</i> - 102 (0.0074)
			<i>Hs</i> - Alpha actinin	<i>Hs</i> - 96 (0.25)
12	83837 -	1414	<i>Ag</i> - Copia GAG Int Pol	<i>Ag</i> - 142 (2.8e-05)
	90704		<i>Dm</i> - Lethal 3 malignant brain tumor	<i>Dm</i> - 125 (0.0010)
			<i>Hs</i> - Myosin	<i>Hs</i> - 118 (0.0085)
