

# Predicted gene content of the 92-kb *A. funestus* BAC 1B assembled sequence

Andrew Sheehan, Nadine Shillingford, E.O. Stinson

University of Notre Dame

## Abstract

Shotgun sequencing and assembly of *Anopheles funestus* BAC 1B (AF1B) has provided a 92-kb region of DNA. Of approximately 20 predicted genes, 3 matched well to known genes in *A. gambiae* and *D. melanogaster*. The homologs are described as a vitelline protein homolog, a reverse transcriptase (possibly a transposon), and a polymerase homolog. The rest of the genes matched very poorly (if at all) to any known genes in NCBI's NR database, and it is likely that they are actually incorrect predictions.

---

As a carrier of malaria and bancroftian filariasis, *A. funestus* is an important study subject. Understanding the genome of this organism can lead to better control methods to reduce infection likelihood and save lives, much as with current projects for *A. gambiae* and *A. aegypti*. Analysis of BACs is an important part of the mapping of the genome prior to a full-scale sequencing project, and can provide regions and putative genes of interest for further study.

## Materials and methods

**Construction and sequencing of BAC region.** *A. funestus* BAC AF1B was constructed at TIGR. After sonication to fragment the genome, the fragments were separated by size on an agarose gel. Sequences of about 100kb were then excised and cloned into the pBeloBAC cloning vector. After cloning, one of the resulting BAC clones (AF1B) was picked for sequencing and partially digested. Fragments were then integrated into the pHOS2 cloning vector and amplified. The fragments were then further amplified using PCR and sequenced.

## Cleanup and assembly of BAC sequence.

Removal of the pBeloBAC sequence was done by N.F. Lobo, University of Notre Dame. As received by us, the sequence still contained pHOS2 vector sequence and potential contamination. Our first step was to remove any contaminating pHOS2 sequence at the ends of the sequence fragments. We used BLAST to search for high-identity matches to the pHOS2 sequence at the ends of the BAC sequence fragments. Rather than removing the entire sequence containing contamination, we clipped off the end containing the contamination and used the remaining sequence for the assembly process.

The assembly was done using PHRAP, which successfully generated 3 contigs, with only 7 singlet sequences left out of a total of 1756 sequence fragments. The contigs were of length 1662, 163, and 92234nt. In order to check for sequence contamination, the contigs were BLASTed against the NR database at NCBI. Both of the short sequences proved to be contamination; the 163nt sequence was bacterial sequence, and the 1662nt sequence was human contamination.

At this point our only remaining contig was the 92234nt sequence, which we assumed to be the majority of the sequence for the AF1B BAC.

**Gene prediction and finding.** We used GENSCAN and Fgenesh to predict genes. Fgenesh is a gene prediction program that has been trained with several organisms. We used the *D. melanogaster* and *A. gambiae* gene search models provided at the public SoftBerry installation of Fgenesh. Fgenesh seemed to provide fairly reasonable results; every gene we found with a homolog in *A. gambiae* or *D. melanogaster* was discovered by Fgenesh rather than GENSCAN. We used the GENSCAN Web

Server at MIT to scan for genes against vertebrates. GENSCAN predicted many hits on our sequence, most of which were fairly long. Most of the GENSCAN hits had poor results when those sequences were BLASTed against the NR database from NCBI.

All predicted genes were then tested against the NR database (using blastx) and InterPro (using InterProScan) to look for homologs and domains of interest, respectively. Genes matching the profile of transposons were tested at RepBase to determine the transposon type, if possible.

## Results

An overview of the annotation can be found in Figure 1. We found that only three of the genes predicted by Fgenesh and GENSCAN had homologs in *A. gambiae* and *D. melanogaster* with expect (e) values of  $1e-10$  or better. In fact, all of the predicted genes with homologs were from Fgenesh. The only GENSCAN-predicted genes with even partial homologs were those that overlapped the corresponding Fgenesh-predicted gene. Additionally, there was little correlation between the GENSCAN predicted genes and the Fgenesh predicted genes, although there was some overlap. In general, GENSCAN tended to predict longer genes, possibly because it is not well trained for the stop codons in mosquitoes.

One gene discovered, at position 1278-1747 on the assembled genome, corresponds to a vitelline membrane protein homolog in *A. gambiae*. This gene was predicted successfully by Fgenesh using both the *A. gambiae* and *D. melanogaster* models. There was an overlapping prediction by GENSCAN, but the structure was extremely dissimilar.

Another homologous gene was discovered at position 53846-54496. It hit against a pol-like protein in *A. gambiae*. There were also several weak matches to other organisms, which led us to believe that the gene might be a transposon. Using the RepBase Repeat Masker backed up our assumption, providing a good match against the ISR (interspersed repeat) family of transposon. Oddly, viewing the genomic region surrounding

this supposed transposon in Dotter seemed to indicate that there were no obvious repeat regions, which may indicate that this is a spurious find (or possibly a degraded transposon).

The last gene was discovered on the “backwards” strand at position 92974-82704. It matched very well against reverse transcriptase genes in both *A. gambiae* and *D. melanogaster*. Scanning with RepBase indicated that it was probably a HAT transcriptase. Viewing the surrounding genomic region in Dotter showed inverted repeats at both ends of the putative transposon, further confirming our assumptions. (See Figure 2.)

## Conclusions

BAC AF1B represents a fairly empty section on the genome, although it does contain two putative transposons which may be worth examining for purposes of cross-species comparisons. The results we got from GENSCAN, while they occasionally led us towards a gene, were poor. Only one of the genes predicted by GENSCAN actually turned out to be noteworthy. On the whole, Fgenesh turned out to be a much more reliable program, and most of its predicted genes turned out to have hits that were relatively good compared to the those generated by GENSCAN. This is primarily because GENSCAN has relatively few options when it comes to what type of organism it should be compared against, while Fgenesh gives the user the option to select both *D. melanogaster* and *A. gambiae*, both of which are much more similar species to *A. funestus*. GENSCAN only has options for Vertebrate, Arabidopsis, or Maize.

At this point, further examination of the genes found could be done by collecting mRNA samples from *funestus* and searching for either active versions of the transposons (unlikely) or mRNA for the vitelline membrane protein.

# Figures

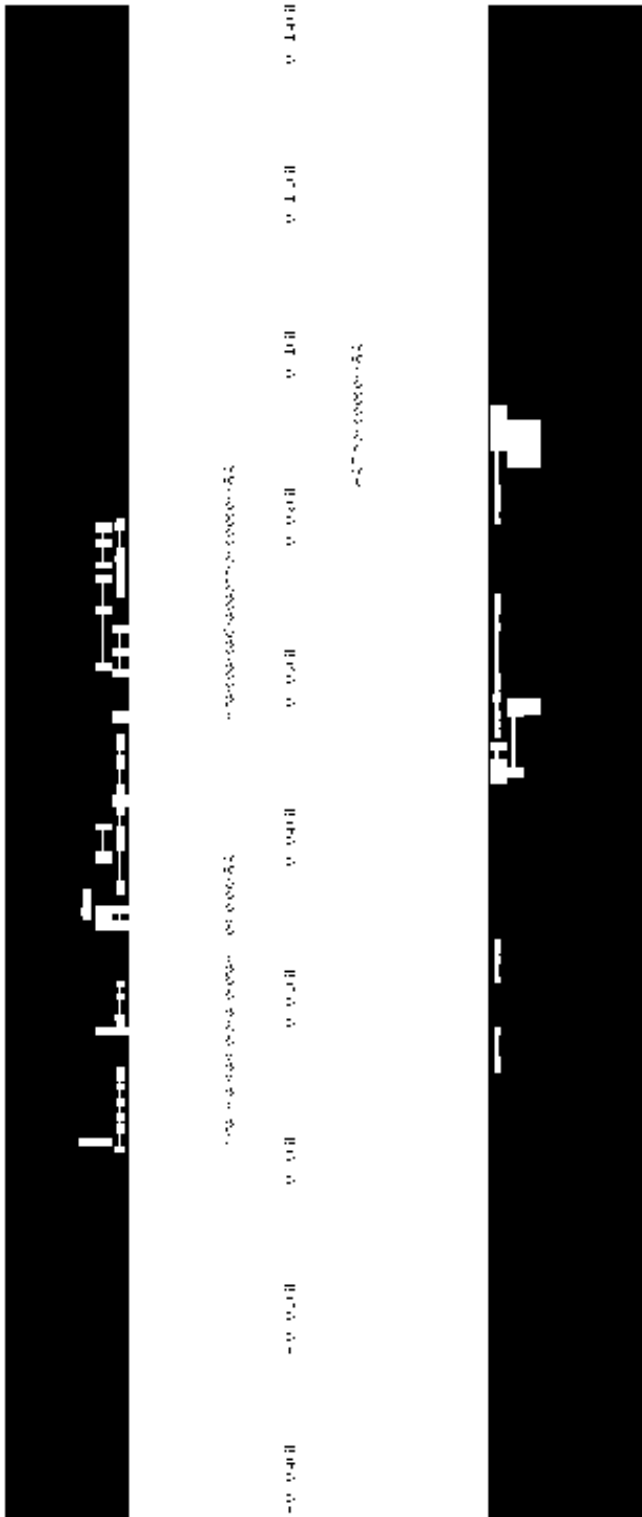


Figure 1: Apollo view of annotated BAC AF1

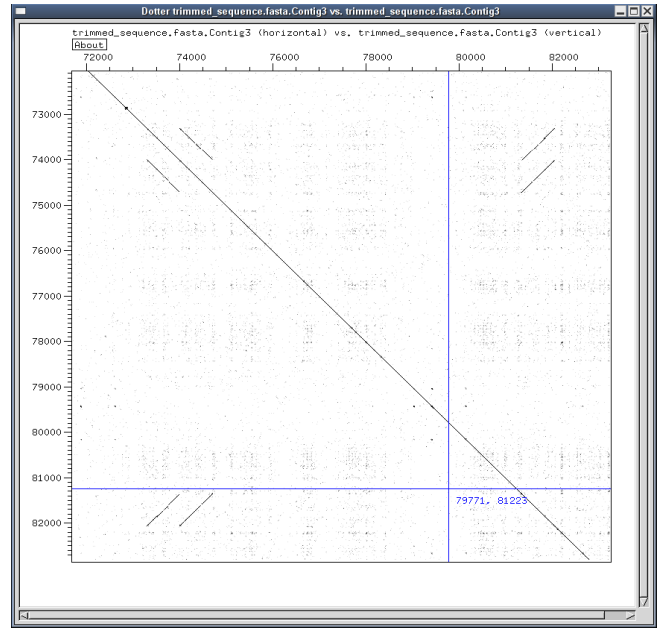


Figure 2: Transposon repeats in Dotter