

# Assembly and annotation of a BAC clone from a newly sequenced genome of *Anopheles funestus*

Faruck Morcos\*

Department of Computer Science and Engineering  
University of Notre Dame

Jeffrey R. Spies\*

Department of Psychology  
Department of Computer Science and Engineering  
University of Notre Dame

Irene Kasumba\*

Department of Biological Sciences  
University of Notre Dame

December 8, 2005

Whole genome sequencing of disease vectors such as *Aedes aegypti*, *Anopheles gambiae*, and recently *Anopheles funestus* are enabling scientists to identify and investigate regions involved in the genetic plasticity of these vectors in order to formulate new disease control strategies. In the current study, sequence data from a single *Anopheles funestus* BAC clone is examined in order to predict and annotate genes in this region. After cloning, cleaning contaminating sequences, and assembling a consensus BAC sequence, several prediction algorithms provided the data then used to refine the predictions by selecting only overlapping predicted genes. These genes and the assembled BAC sequence were compared to the *Drosophila melanogaster* and *Anopheles gambiae* annotated genomes found in the Ensembl database. Two genes of the *Anopheles funestus* BAC were found with high similarity in *Drosophila melanogaster*. Furthermore, these genes showed synteny across *Anopheles funestus* and *Drosophila melanogaster* by appearing in clusters in both organisms. A high degree of conservation of one of these genes was discovered when compared with *Anopheles gambiae*; this gene presented a 100% similarity with respect to the BAC sequence. Putative genes belonging to the KQT Potassium Voltage Gated Channel Subfamily were found to belong to a syntenic cluster conserved both in *Drosophila melanogaster* and in *Anopheles gambiae*.

## Introduction

Efforts to sequence complex vertebrate genomes have increased since the announcement in the 1990s by Celera and the IHGSC of their intention to sequence the human genome and ultimately extended genomic research both in molecular and computational biology. Data generated from mapping, sequencing and annotating genomes, especially of model research organisms (i.e. Mouse, *Drosophila melanogaster*), provide unprecedented opportunities to study and understand fundamental human health issues, improve agricultural output (i.e. cow, pig and chicken genomes), study animal health, and gain insight into disease vectors (mosquitoes) or viruses/*E. coli*. This data that is generated is subjected to computer-based analysis to obtain information regarding the genetic map of the genes. The genomes are scanned to identify protein coding and non-coding regions and organize them into functional groups. They are also scanned for junk sequences—contaminating DNA—such as vector sequences from the plasmid used for the cloning and sequence primer binding step. As much as possible, repeat sequences from

repetitive transposable elements are identified and masked from the genome sequence before genome assembly and annotation.

The general goal of gene assembly and annotation projects is to identify disease causing candidate genes or genes responsible for disease phenotypes and coordinate this information with other agencies such as drug companies (Elstein, Comis, Suszkiw, & Flores, 2005). Comparisons of assembled sequences with annotated genomes in public databases (NCBI, EBL, etc.) identify exons, introns or proteins, detect whole sequence similarities in other organisms that may indicate genes, and reveal sequence conservation in various organisms. The identity of similarly conserved sequences and percentage similarity, especially in closely related organisms, is used to guide gene identification, reveal possible gene function, and aid in studying evolutionary relationship between organisms. Annotation of genomes “beyond conserved genes” using the *ab initio* prediction method has been reportedly accomplished using the Exofish (EXOn Finding by Sequence Homology) program to reevaluate whole genome annotations of *Drosophila melanogaster* and *Anopheles gambiae* (Jaillon et al., 2003). Jaillon et al. found significantly more exons using this prediction method within previously annotated regions and more outside these

---

\* All three authors contributed equally to this research work

genome regions, indicating the need to reevaluate gene models used for the initial prediction.

In the current investigation, we obtained a BAC clone that had been sequenced using the shotgun method developed by Sanger in 1977 (Sanger, Nicklen, & Coulsen, 1977) and utilized since 1982 (Sanger, Coulsen, Hong, Hill, & Peterson, 1982). The BAC clone represented a contig of the newly sequenced *Anopheles funestus* genome. We performed gene prediction and annotation on the contig; the prediction software used included GENSCAN (Burge & Karlin, 1997), Fgenesh (Solovyev, Salamov, & Lawrence, 1994) and Geneid (Parra, Blanco, & Guigó, 2000). An analysis by Burge and Karlin (1997) found a significant difference in the sensitivity and specificity of these packages in their ability to predict genes, indicating the need to use multiple programs for gene prediction; they also found their software, GENSCAN, significantly more accurate in its prediction than other programs. The prediction programs produced a total of 32 potential genes which had considerable overlapping regions. These genes were visually aligned in Wormbase (Stein, Sternberg, Durbin, Thierry-Mieg, & Spieth, 2001) to determine the degree of overlap. The areas of greatest overlap (in the same strand and direction) in each of the three programs were selected for gene annotation. Afterwards, these genes were compared to the *Drosophila melanogaster*, human and *Anopheles gambiae* annotated genomes found in the Ensembl database (Hubbard et al., 2005). Putative genes were found to belong to a syntenic cluster that is conserved both in *Drosophila melanogaster* and in *Anopheles gambiae*. Interestingly, all of the putative genes belonged to the KQT Potassium Voltage Gated Channel Subfamily. There was a strong degree of evolutionary conservation of this protein subfamily across (insect) species, which firmly indicates its integral cellular function.

## Methods

### *Anopheles funestus* BAC clone sequencing

The BAC clone was subjected to a high frequency sound wave shearing (sonicated) technique to obtain overlapping fragments (8-10 times). The resulting fragments (100-200kb) were cloned into a plasmid and the flanking ends were sequenced with the plasmid primers.

### BAC Assembly

After the BAC clone sequence was cleaned, it was assembled using PHRAP. Several sequences of varying length were obtained; each was analyzed using the BLAST (nr) search engine to eliminate non-mosquito sequences. These contaminating sequences were identified and discarded from the resulting assembled sequences. Finally, a consensus sequence representing an assembled contig of *Anopheles funestus* was created. A 92Kb sequence was selected as our possible consensus assembled BAC sequence for gene prediction and annotation.

## Gene Prediction

Three of the main gene prediction software packages were used for this task: GENSCAN, Fgenesh and Geneid. Wormbase was used to visualize the overlap between the predicted genes and extract for further analysis in Ensembl. In total, the three prediction packages produced 32 potential genes. Regional similarities between these genes were examined in order to produce a high quality subset that could lead us to identify real genes in *Anopheles funestus*.

## Results

### Predictions

GENSCAN, developed at Stanford University by Burge et al., uses a probabilistic model of transcriptional, translational and splicing signals in order to predict potential genes in a sequence. It was used on the current study's 92Kbp BAC sequence. Figure 1 shows a graphical representation of the predicted genes in our BAC sequence; seven genes were predicted. Figure 2 shows the graphical representation of the prediction results using Fgenesh. Figure 3 shows a sample of the predicted genes using GeneId.

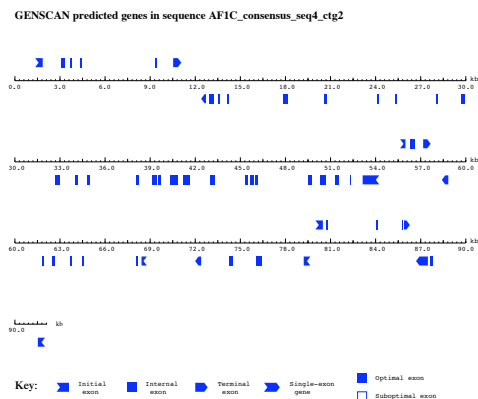


Figure 1. GENSCAN predictions of the *Anopheles funestus* BAC clone.

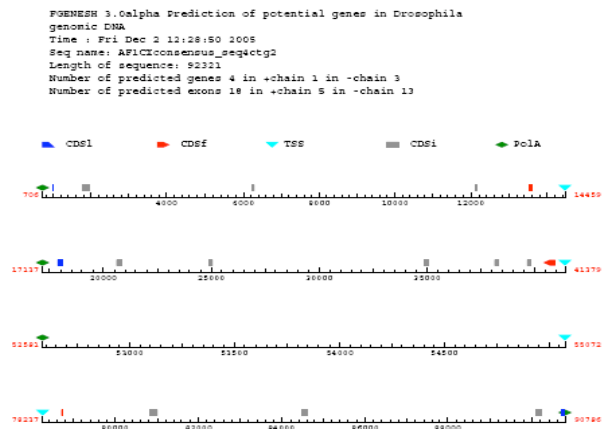


Figure 2. Fgenesh predictions using the *Drosophila*

*melanogaster* genome visually mapped onto the BAC sequence.

```
## source-version: geneid v 1.2 -- geneid@mim.es
# Sequence AF1CXconsensus_seq4ctg2 - Length = 92321 bps
# Optimal Gene Structure. 2 genes. Score = 11.42

# Gene 1 (Reverse). 2 exons. 113 aa. Score = 1.88
AF1CXconsensus_seq4ctg2 geneid_v1.2 Internal 34855 35030 4.87
AF1CXconsensus_seq4ctg2 geneid_v1.2 First 40742 40904 -2.98

# Gene 2 (Reverse). 1 exons. 294 aa. Score = 9.54
AF1CXconsensus_seq4ctg2 geneid_v1.2 Single 53246 54127 9.54
```

Figure 3. Output of Geneid predictions of the *Anopheles funestus* BAC clone.

### Comparative Analysis

The results of the three prediction packages were processed in such a way that it was possible to compare visually the resulting predicted genes and easily find overlapping predictions that in turn provided more accurate results. For this purpose, the Wormbase gene visualization tool. The output from Wormbase is shown in Figure 4. The shaded box in this figure shows an example of the overlapping area that was used for further analysis.

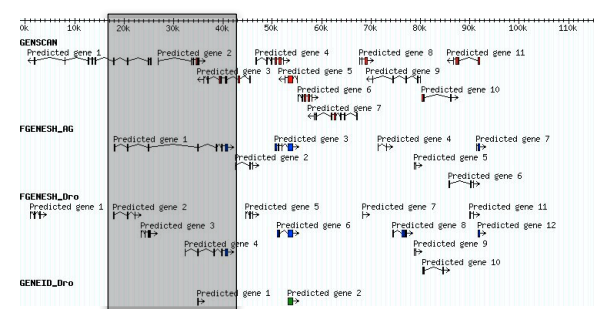


Figure 4. Comparison of gene predictions using the Wormbase visualization tool. The shaded box depicts an example area of interest for further analysis.

### Gene Annotation

The overlapping regions from the gene predictions in the BAC sequence were blasted in Ensembl against two closely related species: *Anopheles gambiae* and *Drosophila melanogaster*. The example selection shown in Figure 4 showed hits with both high sequence homology and very small e-values. Table 1 summarizes the findings. The two genes CG33135-PA and CG33135-PB in chromosome 2R of *Drosophila melanogaster* are two isoform genes; they were also detected in *Anopheles funestus*. Both genes are found in the same neighborhood in both species, hence showing that synteny is conserved between these two organisms. These genes produce proteins working as KCNQ potassium channels. Moreover, when comparing the sequences in *Anopheles funestus* with *Anopheles gambiae*, there was 100% homology with the gene ENSANGP00000004228 in chromosome 2L and ENSANGP00000008384 in chromosome 3L which are closely

related with the genes CG33135-PA(B). These proteins are part of the *KQT Potassium Voltage Gated Channel Subfamily*. This suggests a strong degree of conservation of these protein families in the three organisms.

Table 1

Gene homology found in two species when compared against the *Anopheles funestus* BAC sequence.

Organism	Gene	E-value	Id. (%)
<i>Drosophila melanogaster</i>	CG33135-PA	7.2e-156	96.61
<i>Drosophila melanogaster</i>	CG33135-PB	1.6e-111	96.6
<i>Anopheles gambiae</i>	ENSANGP00000004228	1.3e-177	100
<i>Anopheles gambiae</i>	ENSANGP00000008384	3.2e-6	64.29

### Discussion

The current study concentrated on the process of BAC assembly, gene prediction, and gene annotation of the newly sequenced genome of *Anopheles funestus*. In this particular case, *Anopheles funestus* was the organism analyzed, however the methods presented here serve as a basis to analyze any other newly sequenced organism. This investigation also demonstrates the importance of bioinformatic tools in this process. Although these tools offer a great deal in aiding discovery, it is important to recognize that they are not flawless. For example, gene prediction in its current state should only be considered exploratory given its less than ideal accuracy. It is important to stress the need for biological insight and rigorous analysis to achieve proper conclusions.

From the gene annotation in this study, two of the genes in *Anopheles funestus* interestingly showed synteny (syntenic clusters) with their orthologs in *Drosophila melanogaster*. The genes annotated belonged to the KQT Potassium Voltage Gated Channel family. The KQT subfamily is an integral component of the plasma membrane and has been reported in several organisms ranging from E. coli to human to mouse. Mutations occurring in the critical region of the novel KQT-like potassium channel was linked to several diseases in man including idiopathic epilepsy (Charlier et al., 1998). The strong degree of conservation of this protein family across (insect) species is a firm indicator that it has an integral function and is evolutionarily conserved. The molecular function includes activity in regulating ion transportation across the plasma membrane. The GO lists the KQT subfamily as involved in cellular physiological processes, and work by Charlier et al. (1998) and others have found direct involvement in neuronal functioning. It was also reported that KQT subfamily played a vital role in perception, learning, and locomotion (i.e. nervous system) that was observed in metazoans and insects (McCormack, 2003).

Clearly, it is interesting that two of the genes in *Anopheles funestus* with orthologs in *Drosophila melanogaster* showed synteny. In addition, this analysis showed that synteny was also conserved in *Anopheles Gambiae*, thus leading to the conclusion that the KQT subfamily is evolutionarily con-

served in all three insect species. McCormack (2003) observed: “[T]he *Anopheles* genome encodes more voltage-gated and inward rectifier K<sup>+</sup>-channel genes than that of *Drosophila* ...[but] despite the conservation of intron-exon boundaries, orthologs of genes flanking K<sup>+</sup>-channel genes in *Drosophila* are generally not found adjacent to the *Anopheles* K<sup>+</sup>-channel orthologs.” McCormack then concluded that this may have resulted from large gene translocations from years of species divergence. Whether this is the case with the putative genes in this study (i.e. whether it is syntenic or microsyntenic clustering), we are unable to conclude at this time, but research on the KQT subfamily (of disease vectors) could lead to the development of new insecticides or other strategies to control mosquitoes.

Finally, the need to develop new and more efficient tools to automate the process of gene finding and annotation cannot be emphasized enough. Although computers are of great utility, this process is still slow and inefficient, stemming from the lack of inter-operability between genomic databases, algorithms and analysis tools. Although as much of an art as a science at times, there are many occasions when proper automation and report generation could accelerate the process greatly.

## References

- Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268, 78–94.
- Charlier, C., Singh, N. A., Ryan, S. G., Lewis, T. B., Reus, B. E., Leach, R. J., & Leppert, M. (1998). A pore mutation in a novel KQT-like potassium channel gene in an idiopathic epilepsy family. *Natural Genetics*, 18, 53–55.
- Elstein, D., Comis, D., Suszkiw, J., & Flores, A. (2005). An agency effort to sequence genomes: unraveling the genome of the honey bee, pig, cow, and chicken. *Agricultural Research*.
- Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinski, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., G. Proctor, M. R., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., & Birney, E. (2005). Ensembl 2005. *Nucleic Acids Res.*, 33.
- Jaillon, O., Dossat, C., Eckenberg, R., Eiglmeier, K., Segurens, B., Aury, J., Roth, C., Scarpelli, P., Weissenbach, J., & Wincker, P. (2003). Assessing the *Drosophila melanogaster* and *Anopheles gambiae* genome annotations using genome-wide sequence comparisons. *Genome Research*, 13, 1595–1599.
- McCormack, T. J. (2003). Comparison of K<sup>+</sup>-channel genes within the genomes of *Anopheles gambiae* and *Drosophila melanogaster*. *Genome Biology*, 4, R58.
- Parra, G., Blanco, E., & Guigó, R. (2000). Geneid in *Drosophila*. *Genome Research*, 10(4), 511–515.
- Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., & Peterson, G. B. (1982). Nucleotide sequence of bacteriophage DNA. *J. Mol. Biol.*, 162, 729–773.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain terminating inhibitors. *Proc. Natl. Acad. Sci.*, 74, 5463–5467.
- Solovyev, V. V., Salamov, A. A., & Lawrence, C. B. (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of sliceable open reading frames. *Nucl. Acid. Res.*, 22, 5156–5163.
- Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., & Spieth, J. (2001). Wormbase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Research*, 29(82–86).