

Assembly and Annotation of *Anopheles Funestus* BAC Clone AF1B

Robert Bruggner, Ryan Butler, Alec Pauling

December 6, 2005

Bioinformatics

Introduction

The task of shotgun sequencing an organism's genome is a formidable task whose workload is proportional to the length of the genome. For many eukaryotic organisms, the amount of sequence data produced and workload entailed during assembly is so great that it inherently makes the task of whole genome shotgun sequencing and assembly unreasonable for small and medium sized labs. However, in most cases, researchers are not interested in the entire sequence of an organism's genome – only a small subsection. By dividing up the organism's genome into smaller parts, replicating these individual parts in a cloning vector, and then sequencing each of these pieces, the daunting job of entire genome shotgun sequencing subdivided into smaller, more manageable parts. This technique is more cost and labor efficient than and is often used to assess the viability of an entire genome shotgun project for an organism or to sequence genomic DNA surrounding particular genes of interest. Here we present a 100Kb chunk of genomic DNA from *Anopheles funestus* that has been replicated in the pHOS2 cloning vector, assembled, and had preliminary annotation inferred via novel gene prediction and sequence alignment to *Anopheles gambiae* and *Drosophila melanogaster*.

Sequence Cloning

Large amounts of genomic sequence from *Anopheles funestus* were fragmented via sonication and then run

through an agarose gel to separate the pieces according to size. DNA fragments of approximately 100,000bp were extracted from the gel, ligated with end adapters, and then inserted into a Bacterial Artificial Chromosome (pBeloBAC) for replication. After replication, each of our BAC vectors and their genomic sequence were fragmented again by way of a partial restriction digest. Those new pieces were separated in a gel and fragments ranging in size from 2kb-3kb were inserted into a pHOS2 plasmid vector for individual replication. After replication, each of the plasmids was sequenced using the insert site sequence of the plasmid as templates for primers. After sequencing, the sequence data was ready to be cleaned.

Sequence Post-Processing

Once the sequence data had been obtained from the lab, we searched for and removed sections of the BAC cloning vectors that made it into the sequence data. We then assembled the sequence and searched for contamination. Once any contamination was removed, we began searching for genes.

Ideally, when the sequence data were extracted from the BAC cloning vector, the entire fragment was extracted with no portion of the vector; however, in reality, errors will occur. There are two main types of error: (1) a sequence data containing a portion of the vector at one or both ends and (2) a sequence fragment containing a portion of the vector in the center. Since the sequence

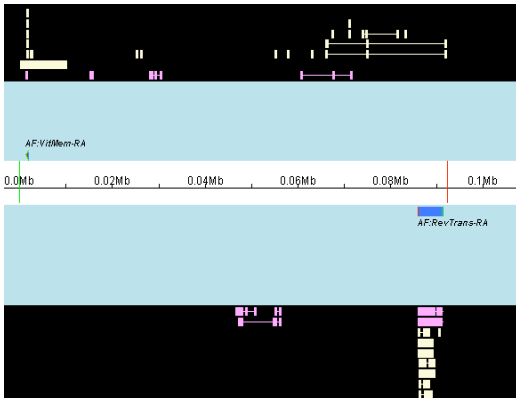


Figure 2 Tiers of Evidence as Shown in Apollo Annotation Editor

Based on our results we believe we have found one gene and one transposon. Based on hits to previously annotated proteins and presence of specific domains, we have putatively identified the gene as a Vitelline Membrane protein. We believe the other sequence displaying gene-like structure is a retro-transposon. BLAST results indicate that the putatively produced protein is a reverse transcriptase. Using Dotter, we also identified inverted repeats flanking the reverse transcriptase producing gene, thus providing strong evidence supporting our theory that the sequence is a transposon. Further strengthening this argument is the presence of several reverse transcriptase domains within the predicted peptide sequence. These were detected using an InterPro scan of the peptide sequence.

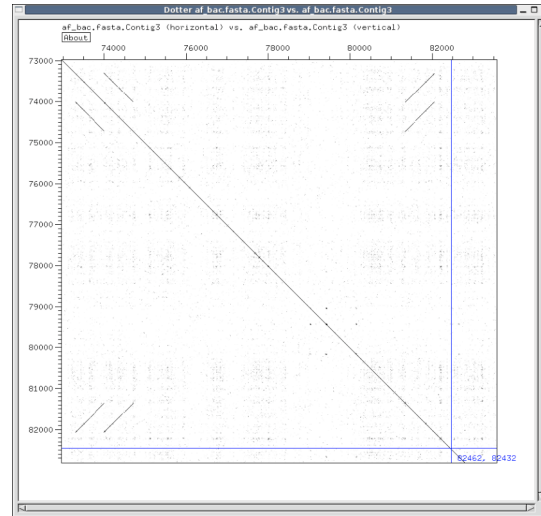


Figure 3 Inverted Repeats at Tail of Putative Transposon

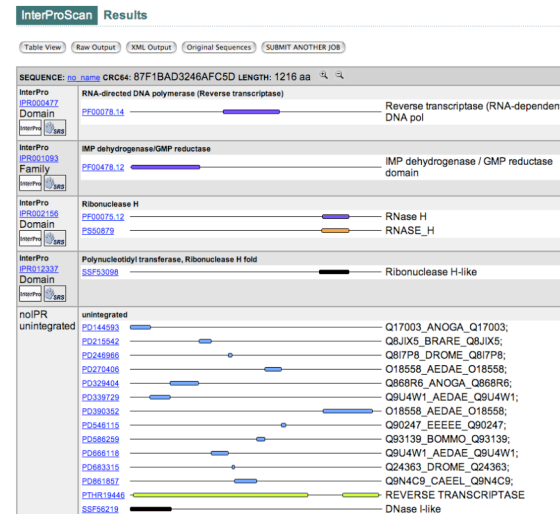


Figure 4 InterPro Detection of Reverse Transcriptase Domains

Conclusions

While the process of sequence cleaning and assembly was fairly straight-forward and scriptable, the annotation of our final sequence was not as easily automated. After observing results produced by novel gene

prediction programs, it's clear that identification of putative genes is greatly enhanced by human inspection. Through visual inspection, we were able to assign putative function to Vitelline Membrane gene simply based on BLAST results. However, without further investigation it's easy to understand how automatic gene prediction pipelines would not be able to detect and classify our found transposon without a fairly complex set of logic involving various other methods of sequence examination beyond simple alignments. We suspect that there are many other gene-like sequences beyond transposons that probably are mis-annotated due to lack of human inspection.