

VECTOR GENOME ANNOTATION & COMPARATIVE GENOME ANALYSIS PROJECTS

Gelbart Lab / FlyBase / VectorBase
 Dept. of Molecular & Cellular Biology
 Harvard University
 16 Divinity Ave.
 Cambridge, MA 02138
 USA

Arjun Bhuktar^{1,2}, Susan Russo¹, Temple Smith², Doug Smith³ & William Gelbart¹. *Genome architecture in the genus Drosophila*. (¹FlyBase, Harvard University, ²Boston University, ³Agencourt Corp.) By this fall, assembled genome sequences for each of 12 species of *Drosophila* should be publicly available. Based on our lab's previous experience in defining syntenic relationships between *D. melanogaster* and *D. pseudoobscura*, we have been developing software to fully automate the process of defining orthologies and syntenic organization between *D. melanogaster* and each of the other species. The basic approach, leverages the much higher quality of the *D. melanogaster* genome sequence and annotations. Briefly, we run translational blast analysis (TBLASTN) of the protein set from *D. melanogaster* against the 6-frame translations of the WGS sequence assemblies of each of the other species, identify the strongest hits, filter the hit set based on conservation of chromosome arm location and other criteria to produce a well-anchored set of orthologies and syntenic blocks. We will discuss the advantages of this approach over direct comparisons of gene prediction sets between any two species. We will also discuss some of the initial observations that we have made from an analysis of preliminary assemblies of *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis*. Some observations of relevance to the analysis of vector genomes are:

- evaluating multiple species provides a much more robust picture of gene content of any one species, since the gaps in any one genome are often covered in related species.
- comparisons of multiple genomes at evolutionary distances is essential, because different genes are diverging at very different rates.
- the vast majority of genes within genus *Drosophila* are conserved and remain on the same chromosome arm.
- within each arm, gene order has been extensively rearranged throughout the genus.
- interspecific comparisons can, to some degree, help determine the gene content and gene order for the founding common ancestor of the genus *Drosophila*, providing a better reference genome for comparisons to insect vector genomes.
- a few hundred genes have shifted arms within the genus *Drosophila*. It is likely that retrotransposition is a major contributing mechanism. Such events need to be considered in evaluating syntenic relationships between related species.

Kathryn Campbell¹ and William Gelbart¹. Contributions of expert manual annotation to understanding the gene sets of *Anopheles gambiae* and *Aedes aegypti*. (¹VectorBase, Harvard University) Current best practice toward a full description of the gene products encoded by a vector species is to combine state-of-the-art gene prediction approaches (in our case, the Ensembl pipeline) with expert annotator evaluation of the Ensembl predictions and the supporting experimental and computational evidence. We have been prototyping the application of expert manual annotation to *Anopheles gambiae* and *Aedes aegypti* by annotation of about 1.5-2.0 megabases of each of these two genomes. We find that about 50% of Ensembl annotations of *A. gambiae* require major restructuring (new annotation, elimination of unsupported annotations, major splits, merges, etc.). The state of the *A. aegypti* genome and its predicted gene models is quite preliminary, with less supporting evidence available at this time. Even so, it is already clear that manual annotation will be at least as important for a high quality *A. aegypti* annotation set. Based on a very limited set of comparisons, it is likely that conserved gene order and orientation (synteny) will be an important asset in mosquito annotation.