

Karin Eiglmeier

Unité de Biochimie et Biologie Moléculaire des Insectes
Institut Pasteur
25, rue du Dr Roux
75015 Paris
France
Tel: 33-1-40 61 36 25
Email: kei@pasteur.fr

Full-length enriched cDNA libraries from *Anopheles gambiae*

**Karin Eiglmeier¹, Shawn Gomez¹, Beatrice Segurens², Inge Holm¹, Corinne Dasilva²,
Betina M. Porcel², Pierre Dehoux¹, Patrick Wincker², Paul Brey¹, Vincent Schaechter²,
Jean Weissenbach² and Charles Roth¹**

1) Unité de Biochimie et Biologie Moléculaire des Insectes. Institut Pasteur, 75015 Paris, France.

2) Génoscope, Centre National de Séquençage, CP 5706, 91057 Evry, France

The first draft of the *Anopheles gambiae* genome sequence, generated by an international consortium of institutions using the shotgun sequencing technique, was released in October 2002, accompanied by its preliminary annotation. One of the primary tasks in genome annotation is to elucidate the location and structure of all protein-coding genes; however, the available prediction tools still have difficulties predicting complete gene structures and defining their boundaries. Deciphering the *An. gambiae* genome for which only a very limited number of genes had been experimentally characterized, turned out to be no less a challenge than the sequencing effort itself. The utilisation of computer based methods by Celera Genomics as well as by the European Bioinformatics Institute (EBI), including the integration of available EST and cDNA data, permitted the prediction and characterization of approximately 15000 genes in the initial annotation.

With the objective of improving the genome annotation and generating useful tools for functional genomic approaches, our laboratory in collaboration with the Genoscope initiated a program to construct and sequence *An. gambiae* full-length enriched cDNA libraries. A collection of these cDNAs represents an important resource for gene identification, for determining their intron-exon structure, including splice variants, as well as providing a

resource for the rapid production of the complete protein corresponding to the gene. In an initial pilot project, the importance of this approach was validated with a library from adult, uninfected female mosquitoes, prepared using the RNA oligo capping technique (K. Marujama et S. Sugano, *Gene* 1994, 138:171). The inserts of 34,000 non-normalized clones were sequenced from both ends. Analysis of the sequences identified nearly 3700 genes of which about 17%, 650 genes, are apparently new genes, absent from the initial genome annotation (Gomez *et al.* *Genome Biology* 2005, 6:R39). The cDNAs also improved the predicted models by extending the boundaries for 85% of the predicted genes covered by the cDNAs. To increase the number of genes covered by full-length cDNAs, we have constructed new, additional libraries from two developmental stages (embryos and larvae). These libraries are currently being analysed and will be discussed at the meeting.