Available online at www.sciencedirect.com
SCIENCE DIRECT®
Insect Biochemistry and Molecular Biology 35 (2005) 799–814
Insect Biochemistry and Molecular Biology
www.elsevier.com/locate/ibmb

Comparative analysis of BAC and whole genome shotgun sequences from an *Anopheles gambiae* region related to *Plasmodium* encapsulation

Karin Eiglmeier[a], Patrick Wincker[b], Laurence Cattolico[b], Veronique Anthouard[b], Inge Holm[a], Ralph Eckenberg[b,1], Hadi Quesneville[c], Olivier Jaillon[b], Frank H. Collins[d], Jean Weissenbach[b], Paul T. Brey[a], Charles W. Roth[a,*]

[a]*Biochimie et Biologie Moléculaire des Insectes, CNRS-FRE 2849, Institut Pasteur, 25, rue du Dr. Roux, F-75724 Paris Cedex 15, France*
[b]*GENOSCOPE and CNRS UMR 8030, Centre National de Séquençage, 2, rue Gaston Crémieux, F-91000 Evry Cedex, France*
[c]*Laboratoire Dynamique du Génome et Evolution, Institut Jacques Monod, 2, place Jussieu, F-75251 Paris Cedex 05, France*
[d]*Center for Tropical Disease Research and Training, Department of Biological Science, University of Notre Dame, Notre Dame, IN 46556-0369, USA*

Received 8 December 2004; received in revised form 10 February 2005; accepted 15 February 2005

Abstract

The only natural mechanism of malaria transmission in sub-Saharan Africa is the mosquito, generally *Anopheles gambiae*. Blocking malaria parasite transmission by stopping the development of *Plasmodium* in the insect vector would provide a useful alternative to the current methods of malaria control. Toward this end, it is important to understand the molecular basis of the malaria parasite refractory phenotype in *An. gambiae* mosquito strains. We have selected and sequenced six bacterial artificial chromosome (BAC) clones from the *Pen-1* region that is the major quantitative trait locus involved in *Plasmodium* encapsulation. The sequence and the annotation of five overlapping BAC clones plus one adjacent, but not contiguous clone, totaling 585 kb of genomic sequence from the centromeric end of the *Pen-1* region of the PEST strain were compared to that of the genome sequence of the same strain produced by the whole genome shotgun technique. This project identified 23 putative mosquito genes plus putative copies of the retrotransposable elements BEL12 and TRANSIBN1_AG in the six BAC clones. Nineteen of the predicted genes are most similar to their *Drosophila melanogaster* homologs while one is more closely related to vertebrate genes. Comparison of these new BAC sequences plus previously published BAC sequences to the cognate region of the assembled genome sequence identified three retrotransposons present in one sequence version but not the other. One of these elements, Indy, has not been previously described. These observations provide evidence for the recent active transposition of these elements and demonstrate the plasticity of the *Anopheles* genome. The BAC sequences strongly support the public whole genome shotgun assembly and automatic annotation while also demonstrating the benefit of complementary genome sequences and of human curation. Importantly, the data demonstrate the differences in the genome sequence of an individual mosquito compared to that of a hypothetical, average genome sequence generated by whole genome shotgun assembly.

Keywords: *Pen-1*; Genome; BAC end sequence; Annotation; BEL; Indy

1. Introduction

The approximately 280 Mb genome of *Anopheles gambiae*, one of the most important malaria vectors in Africa, has recently been sequenced (Holt et al., 2002). This genome is composed of three chromosomes pairs in

Abbreviations: *Pen–Plasmodium* encapsulation trait; BAC–bacterial artificial chromosome; Mb–megabase; Kb–kilobase; LTR–long terminal repeat.

*Corresponding author. Tel.: +33 1 40 61 36 25; fax: +33 1 40 61 34 71.
E-mail address: croth@pasteur.fr (C.W. Roth).
[1]Current address: Genomining, 4, rue René Barthélémy, F-92120 Montrouge, France.

0965-1748/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ibmb.2005.02.020

female mosquitoes: two individual autosome pairs, chromosomes 2 and 3, plus a pair of sex (X) chromosomes while the males have the same autosomes plus a single X complemented by a Y chromosome. The importance of *An. gambiae* as a malaria parasite vector was a major stimulus leading to an international effort to sequence this mosquito's complete genome. The first draft of the *An. gambiae* Pink eye strain (PEST) genome sequence was produced with the shotgun sequencing technique (Holt et al., 2002). Automated annotation procedures predicted about 14,000 genes but further analysis of the proposed genes reduced the probable gene number to about 13,000 (Zdobnov et al., 2002). However, additional detailed gene searches using other techniques are continually identifying new genes and it is reasonable to assume that the total number of genes will rapidly rebound to or beyond the original estimate. Bacterial artificial chromosome (BAC) mapping has generally confirmed the genome assembly; nonetheless, many gaps remain in the sequence and a large number of small sequence scaffolds have not been assigned to chromosomal locations and are grouped into an "unknown chromosome". Despite these latter problems, the overall assembly has been accepted as an accurate version of the *An. gambiae* PEST genome.

An accurate genomic sequence is important for understanding the mosquito genes involved in the interaction between the mosquito and the protozoan malaria parasite it hosts. Collins et al. (1986) made the observation that *An. gambiae* variants, which block oocyst development by ookinete encapsulation, can be isolated and maintained in the laboratory. Several subsequent projects were directed toward identifying the genetic basis of this refractoriness or *Plasmodium* encapsulation (*Pen*) trait (Zheng et al., 1997). These projects led to the quantitative trait loci (QTL) analysis of the *Plasmodium* refractoriness trait and revealed that three loci are involved in blocking *Plasmodium* development with *Pen-1*, the most important locus, located on arm 2R region 8C–D (Collins et al., 1997). Determining which mosquito gene(s) in this 3 Mb *Pen-1* region are important for the refractory phenotype remains a challenge. The genome sequence is playing a central role for studies of the genes in the *Pen-1* region and it is important to reassess the genomic features in this particular region.

Thomasova and colleagues (Thomasova et al., 2002) sequenced and analyzed six BAC clones covering a 528 kb region at one end of the *Pen-1* locus. This group was the first to observe the large polymorphism in the *An. gambiae* PEST strain genome sequence and they suggested that it plays a role in the refractory phenotype. We report here the sequence of six additional BACs from the same library but representing another section of *Pen-1* (Fig. 1A). The six BAC clones including five overlapping and one adjacent clone were sequenced and annotated by a combination of ab initio and manual techniques. The sequence of a BAC clone can be considered as the actual chromosome sequence from one mosquito within the population whereas an assembled shotgun sequence represents an average sequence of the entire population. This latter may be correct but it does not reflect the variation in individuals. In this light, the new BAC sequences generally agree with the shotgun assembly although there are numerous local differences. Notably, there are extensive regions of high and low single nucleotide polymorphism when comparing the genome sequence to that of a single BAC sequence, reflecting the distribution of polymorphism within the population of the PEST colony. The most visible difference between the genome assembly and the BAC sequences is the presence of transposon-like sequences in only one of the two.

Our gene predictions for this region correspond well to those in the Ensembl database and while most of the predicted genes have homologs in the *Drosophila*, one of these genes was more similar to several vertebrate homologs than to any *Drosophila* gene and two other genes (excluding the partial D2.14 gene) are unique to *Anopheles*. The 448 kb BAC Contig D1 contains a new member of the transposon BEL12_AG family. The sequence of this transposon, as well as that for another BAC encoded BEL-type transposon (Thomasova et al., 2002) is missing from the assembled genome. In contrast, the shotgun assembly was found to contain an unusual class I transposon, tentatively named Indy, that is not represented in the BAC clones corresponding to the cognate genomic region. In all three cases, these putative transposons appear to be recent additions to the genome and contain long open reading frames. Our BAC sequence and annotation give strong support to the previous automatic annotation procedures while bringing out the importance of individual sequence differences that can be found within the mosquito population.

2. Material and methods

2.1. End sequencing of the BAC clones from the An. gambiae PEST strain

An. gambiae PEST strain (Mukabayire and Besansky, 1996) genomic DNA for BAC library construction was prepared from half gravid mosquito ovaries rich in polytene chromosomes (Hong et al., 2003). The insert ends of 12,288 BAC clones were sequenced as previously described (Roest Crollius et al., 2000a,b).

2.2. Identification and sequencing of BAC clones in the Pen-1 region

BAC clones of interest were initially identified by microdissection of the region of the polytene