Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 333 (2004) 183–186
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

Comparative genomics of *Anopheles gambiae* and *Drosophila melanogaster*

Kamel Jabbari[a,1], Giorgio Bernardi[a,b,*]

[a] *Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, F-75005 Paris, France*
[b] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Naples, Italy*

Received 17 December 2003; accepted 10 February 2004

Available online 23 April 2004

Abstract

A sequence analysis of the genomes of *Anopheles gambiae* and *Drosophila melanogaster* reveals that *Anopheles* DNA is more heterogeneous and GC-richer than *Drosophila* DNA. The gene concentration across the *Anopheles* genome is characterized by low levels in the GC-poor part of the genome and a 3-fold increase in the GC-richest part; this gene density gradient is approximately half that of *Drosophila*. GC levels of introns and flanking sequences are correlated with GC_3 values (GC levels of third codon positions) of the corresponding genes with slopes much lower than unity; in other words, most introns and intergenic sequences are less GC-rich than the corresponding GC_3 values. These findings, which describe a compositional shift within Diptera, is of interest because of their parallels in the well studied major shift in vertebrates.

Keywords: Genome size; Evolutionary genomics; Insects

1. Introduction

Eukaryotic genomes exhibit a number of characteristic features. The compositional (GC) heterogeneity is different in different organisms and different gene density gradients are observed in different species. The human genome, for instance, covers a 30% GC range at an average size of 50 kb (Bernardi et al., 1985; Zoubak et al., 1996; Jabbari and Bernardi, 1998; see Bernardi, 2004, for a review); whereas, at the same average size, the *Drosophila* genome only covers a 10% GC range (Jabbari and Bernardi, 2000), and the *Arabidopsis* genome an 8% GC range (Barakat et al., 1998; Carels and Bernardi, 2000).

Gene concentration increases, as a general rule, from GC-poor to GC-rich regions of eukaryotic genomes (Bernardi et al., 1985; Mouchiroud et al., 1991; Zoubak et al., 1996). The gene density gradient is, however, very steep in the human genome, since it covers a 20-fold range, but much less so in *Drosophila*, where the range is only 6-fold (Jabbari and Bernardi, 2000; Adams et al., 2000), and in *Arabidopsis* where the range is only 2-fold. There are also differences in the slopes of the regression lines of plots of GC levels in introns and flanking sequences vs. the GC_3 values (GC levels in third codon positions) of the corresponding genes. In the human genome, the slope of the orthogonal regression line is about 8 (Zoubak et al., 1996), whereas in *Drosophila* it is about 3 and in *Arabidopsis* only 2.

In the present work, we performed a comparative analysis on the compositional features of large DNA segments from *Anopheles* and *Drosophila*. We also analysed the gene distributions, as well as the compositional correlations between GC_3 and GC levels of introns and flanking sequences.

2. Materials and methods

Genomic sequences of *Anopheles* and *Drosophila* were downloaded from ftp://ftp.ensembl.org/pub/Anopheles-7.1a/data/golden_path and Celera and BDGP's, respectively. These sequences were analysed using nonoverlapping 50 kb windows. To obtain an estimate of the compositional

* Corresponding author. Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Naples, Italy. Tel.: +39-81-583-3215; fax: +39-81-245-5807.
E-mail address: bernardi@szn.it (G. Bernardi).
[1] Present address: ENS/CNRS FRE 2433, Organismes Photosynthétiques et Environnement, Département de Biologie, Ecole Normale Supérieure, 46 Rue D'Ulm, 75230 Paris Cedex 05, France.

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.02.038

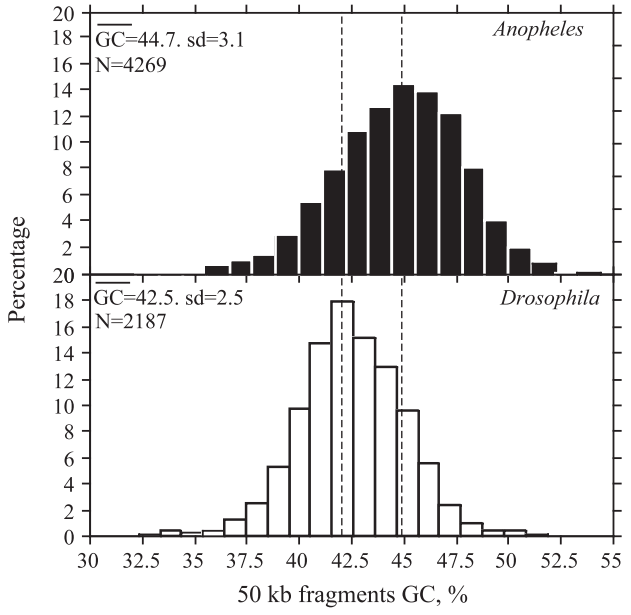

Fig. 1. GC histograms of large (50 kb) DNA sequences from *Anopheles* and *Drosophila*. Dashed lines correspond to the modal GC levels. Average GC levels, standard deviations and sample sizes are indicated.

homogeneity along large sequences, we used the so-called phase plots (Ruelle, 1989; Jabbari and Bernardi, 2000; see also Clay, 2001, and references therein), in which the GC level of each window (GC_n) is plotted against that of the following one (GC_{n+1}). Correlation coefficients of such plots are good indicators of the GC level homogeneity across the compositional spectrum of each genome.

To analyse the gene distribution across the genome, we partitioned the *Anopheles* chromosomes into 1 Mb segments and counted the coding sequences (CDSs), as annotated in the *Anopheles* genome release of Ensembl (Release 9.1a.1; 2-12-2002 and 19.2a.1; 29-09-2003).

3. Results and discussion

3.1. Compositional heterogeneity

As shown in Fig. 1, the GC distributions of large DNA sequences (50 kb) from *Drosophila* and *Anopheles* are different. The genome of *Anopheles* is GC-richer and more heterogeneous (GC% = 44.7 ± 3.1) compared to *Drosophila* (GC% = 42.5 ± 2.5). We note in passing that the 35.2% GC of *Anopheles* reported by Holt et al. (2002) is grossly incorrect.

Fig. 2 shows plots of GC levels of 50, 100 and 200 kb segments from both *Drosophila* and *Anopheles* DNAs (GC_n) against those of the following segments (GC_{n+1}). These are the so-called phase-plot (Ruelle, 1989; see also Clay, 2001). At 50 kb, correlation coefficients were 0.61 for *Drosophila* and 0.70 for *Anopheles*, indicating a larger short-range compositional heterogeneity in *Drosophila* compared to *Anopheles*. The differences were smaller for 100 and 200 kb segments, the slopes approaching the diagonal with increasing segment size.

3.2. The compositional correlation of third codon positions with introns and with long DNA sequences

Fig. 3a and b (bottom) shows a plot of GC_3 levels of coding sequences (only sequences starting with an ATG were considered) against the GC levels of the corresponding long sequences (20 kb on each side of

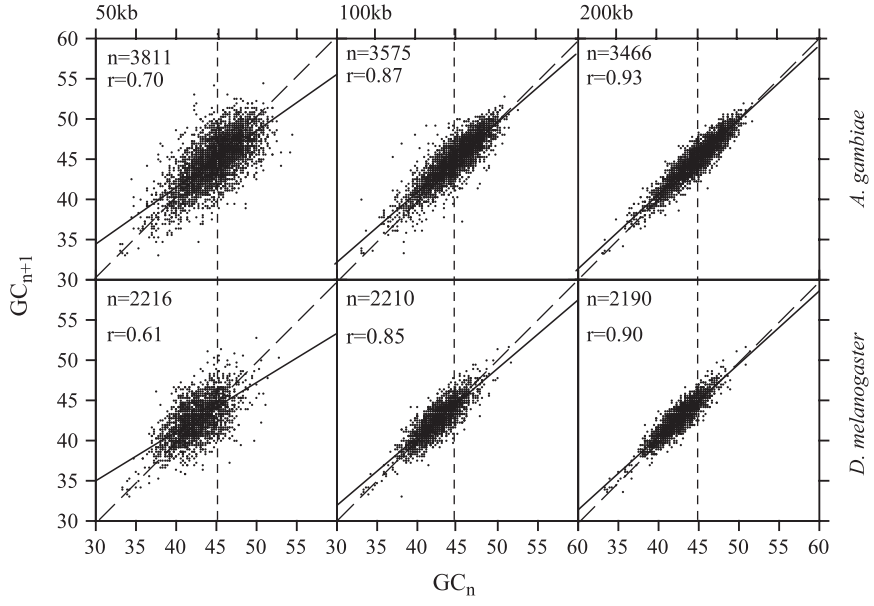

Fig. 2. GC levels of *Anopheles* and *Drosophila* DNA sequences of 50, 100 and 200 kb are plotted against the GC levels of the following (adjacent) sequences of the same size. *r* is the correlation coefficient, *n* the sample size. The main diagonal is shown.

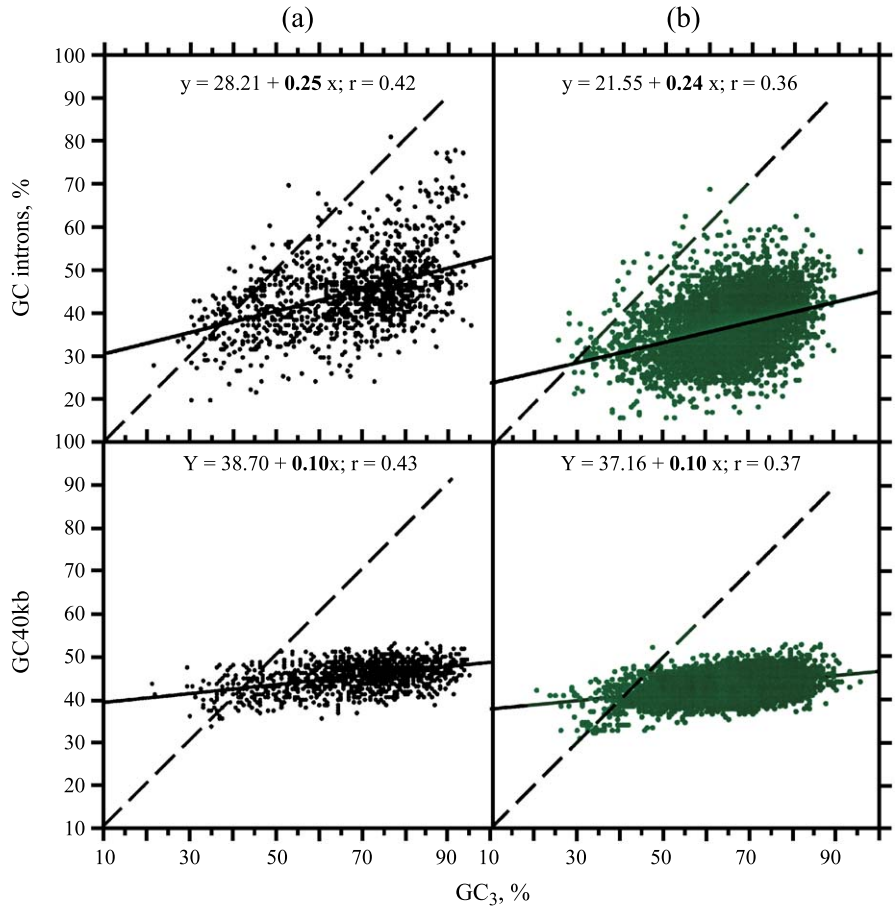

Fig. 3. Plot of GC_3 levels of *Anopheles* and *Drosophila* coding sequences against GC levels (a, b; top panels) of introns and (a, b; bottom panels) large (20 kb on each side of the CDS) DNA sequences. The equations of the linear regression lines are shown.

the coding sequences). The correlation coefficients r were 0.43 ($p < 0.0001$) for *Anopheles* and 0.37 ($p < 0.0001$) in *Drosophila*.

Fig. 3a and b (top) shows correlations between intron GC levels and the GC_3 levels of the corresponding coding sequences. The correlation coefficients are 0.42 in the case of *Anopheles* and 0.36 in the case of *Drosophila* ($p < 0.0001$). Interestingly, the slopes are almost identical (0.25 vs. 0.24).

We notice that in both cases (introns and flanking sequences) GC-poor genes tend to be closer to the diagonal compared to those of GC-rich genes. Such a difference was also observed in mammals and birds (Aïssani et al., 1991; Clay et al., 1996; Musto et al., 1999), as well as in the compact genome of *Fugu* and the large genome of zebra fish (paper in preparation). The fact that the slopes are almost identical in the GC-intron vs. GC_3 plots and that *Drosophila* intron (or genome) size is about half that of *Anopheles* (Zdobnov et al., 2002) is in contrast with the idea (Duret and Hurst, 2001) that the slope of the regression line relating GC_3 and intron GC levels is dependent upon transposon and repeat content of introns. It has also been noted that the correlation between GC_3 and intron GC, with a slope which is very different from unity, is incompatible

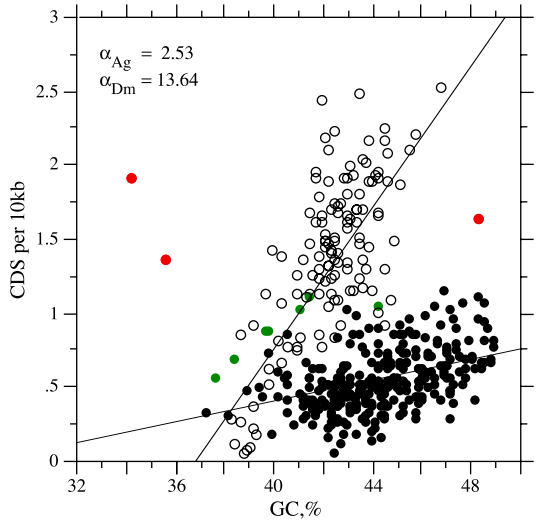

Fig. 4. Comparison of gene density in *Anopheles* (filled circles, potential outliers are in green) and *Drosophila* (empty circles, potential outliers are in red). The regression equations exclusion outliers are $y = -1.023 + 0.036x$, $r = 0.39$ and $y = -8.84 + 0.24x$; $r = 0.72$, respectively. Orthogonal angles are indicated. Note that no significant change was noticed after exclusion of potential outliers: 3 (red circles) out of 126 points in the case of *D. melanogaster* and 7 (green circles) out of 191 points in the case of *A. gambiae*.

with mutational models of compositional heterogeneity, because in that case all sequences are subject to the same changes (Eyre-Walker, 1999).

3.3. Gene distribution in the genome of Anopheles

June 2003 releases (ftp://ftp.ensembl.org) of *Anopheles* and *Drosophila* were analysed; CDS sequences were counted in nonoverlapping chromosome segments of 1 Mb. In Fig. 4, the correlation coefficients of gene density vs. GC level are strong and statistically very significant ($p < 0.0001$); the angles of the orthogonal regression lines are very different in *Anopheles* (2.5°) and *Drosophila* (13.6°); and no significant change was noticed after exclusion of outliers (3 out of 126 points in the case of *Drosophila* and 7 out of 191 points in the case of *Anopheles*). The genome of *Anopheles* shows an increase in gene density from the GC-poorest to the GC-richest regions, an almost 3-fold enrichment in genes when considering the GC-poorest and the GC-richest DNA segments, i.e., a 2-fold lower increase compared to *Drosophila*.

Furthermore, a lesser variation of gene density is observed in the GC-poor compared to the GC-rich part of the genome. This gene density gradient is shared with the human genome (Jabbari and Bernardi, 2000). Note that this is also the case of *Takifugu rubripes*, as indicated by the whole genome sequence analysis (Aparicio et al., 2002; Jabbari and Bernardi, in preparation). Similarly to *Anopheles* and *Drosophila*, the well characterized genome of *Arabidopsis* also shows a gradient in gene distribution which, however, only shows a 2-fold range (Barakat et al., 1998; Carels and Bernardi, 2000). The lower gene density in *Anopheles* is understandable if one takes into account the approximately 2-fold difference in genome size (Holt et al., 2002) between these genomes (260 and 170 Mb, using Cot analysis and 278 and 122 Mb using genome assemblies).

The higher GC levels attained by long DNA sequences and coding sequences of *Anopheles* compared to *Drosophila* might be related to higher body temperature of the former. We have not found, however, indications on this point in the literature.

References

Adams, M.D., et al., 2000. The genome sequence of *Drosophila*. Science 24, 2185–2195.

Aïssani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K., Gautier, C., Bernardi, G., 1991. The compositional properties of human genome. J. Mol. Evol. 32, 493–503.

Aparicio, S., et al., 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297, 1301–1310.

Barakat, A., Matassi, G., Bernardi, G., 1998. Distribution of genes in the genome of *Arabidopsis thaliana* and its implications for the genome organization of plants. Proc. Natl. Acad. Sci. U. S. A. 95, 10044–10049.

Bernardi, G., 2004. Structural and evolutionary genomics. Natural Selection in Genome Evolution. Elsevier, Amsterdam.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of warm-blooded vertebrates. Science 24, 953–958.

Carels, N., Bernardi, G., 2000. The compositional organization and the expression of the *Arabidopsis* genome. FEBS Lett. 472, 302–306.

Clay, O., 2001. Standard deviations and correlations of GC levels in DNA sequences. Gene 276, 33–38.

Clay, O., Caccio, S., Zoubak, S., Mouchiroud, D., Bernardi, G., 1996. Human coding and noncoding DNA: compositional correlations. Mol. Phylogenet. Evol. 5, 2–12.

Duret, L., Hurst, L.D., 2001. The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. Mol. Biol. Evol. 18, 757–762.

Eyre-Walker, A., 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics 152, 675–683.

Holt, R.A., et al., 2002. The genome sequence of the malaria mosquito *Anopheles*. Science 289, 129–149.

Jabbari, K., Bernardi, G., 1998. CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. Gene 11, 123–127.

Jabbari, K., Bernardi, G., 2000. The distribution of genes in the *Drosophila* genome. Gene 247, 287–292.

Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C., Bernardi, G., 1991. The distribution of genes in the human genome. Gene 100, 181–187.

Musto, H., Romero, H., Zavala, A., Bernardi, G., 1999. Compositional correlations in the chicken genome. J. Mol. Evol. 49, 325–329.

Ruelle, D., 1989. Chaotic Evolution and Strange Attractors: The Statistical Analysis of Time Series for Deterministic Nonlinear Systems. Lezioni Lincee. Cambridge Univ. Press, Cambridge (UK), p. 28.

Zdobnov, E.M., et al., 2002. Comparative genome and proteome analysis of *Anopheles* and *Drosophila*. Science 298, 149–159.

Zoubak, S., Clay, O., Bernardi, G., 1996. The gene distribution of the human genome. Gene 174, 95–102.