

A high frequency of overlapping gene expression in compacted eukaryotic genomes

Bryony A. P. Williams, Claudio H. Slamovits, Nicola J. Patron, Naomi M. Fast, and Patrick J. Keeling*

Canadian Institute for Advanced Research, Botany Department, University of British Columbia, 3529-6270 University Boulevard, Vancouver, BC, Canada V6T 1Z4

Edited by W. Ford Doolittle, Dalhousie University, Halifax, NS, Canada, and approved June 14, 2005 (received for review February 16, 2005)

The gene density of eukaryotic nuclear genomes is generally low relative to prokaryotes, but several eukaryotic lineages (many parasites or endosymbionts) have independently evolved highly compacted, gene-dense genomes. The best studied of these are the microsporidia, highly adapted fungal parasites, and the nucleomorphs, relict nuclei of endosymbiotic algae found in cryptomonads and chlorarachniophytes. These systems are now models for the effects of compaction on the form and dynamics of the nuclear genome. Here we report a large-scale investigation of gene expression from compacted eukaryotic genomes. We have conducted EST surveys of the microsporidian *Antonospora locustae* and nucleomorphs of the cryptomonad *Guillardia theta* and the chlorarachniophyte *Bigelowiella natans*. In all three systems we find a high frequency of mRNA molecules that encode sequence from more than one gene. There is no bias for these genes to be on the same strand, so it is unlikely that these mRNAs represent operons. Instead, compaction appears to have reduced the intergenic regions to such an extent that control elements like promoters and terminators have been forced into or beyond adjacent genes, resulting in long untranslated regions that encode other genes. Normally, transcriptional overlap can interfere with expression of a gene, but these genomes cope with high frequencies of overlap and with termination signals within expressed genes. These findings also point to serious practical difficulties in studying expression in compacted genomes, because many techniques, such as arrays or serial analysis of gene expression will be misleading.

genome compaction | microsporidia | nucleomorph | overlapping transcription

Eukaryotic genomes are generally considered to be relatively spacious compared to those of prokaryotes; however, there is a great deal of variability in both size and density of nuclear genomes. At one end of both of these spectra lie the genomes of microsporidian parasites and nucleomorphs. Microsporidia are obligate intracellular parasites related to fungi with genomes as small as 2.3 Mbp (1–3). The 2.9-Mbp genome of the microsporidian *Encephalitozoon cuniculi* has been fully sequenced, and it has a gene density of ≈ 0.97 genes per kilobase (1–3). Nucleomorphs, on the other hand, are relict nuclei of red and green algal endosymbionts that are found in cryptomonad and chlorarachniophyte algae, respectively (4). These are not free-living organisms, but hyperreduced organelles with genomes smaller than those of microsporidia. The completely sequenced nucleomorph genome of the cryptomonad *Guillardia theta* is 551 kbp and has a gene density of 1.02 genes per kilobase (5), and the nucleomorph of the chlorarachniophyte *Bigelowiella natans* is 380 kbp with a gene density of 0.88 genes per kilobase (4).

The reduction of these genomes is the result of the combined effect of several processes: a reduction in the total number of genes and the compaction of the remaining genes into a smaller space. The first process is relatively easy to understand because microsporidia are parasites and nucleomorphs are organelles, so both are highly dependent on their host. Compaction is harder to explain in general, but we can identify several distinct aspects of this process that have been found to operate in various

combinations in both microsporidia and nucleomorphs. “Non-essential” elements like mobile elements or introns may be lost or reduced in number or size, the average length of the remaining genes may decrease in length, and the noncoding intergenic regions may shrink substantially (4–9). This last process is probably the single greatest contributor to the increase in gene density in these genomes, because most eukaryotic genomes have large buffer regions that insulate individual genes from one another. Normally, intergenic regions encode essential regulatory elements, such as promoters and terminators, which direct the accurate initiation and termination of transcription and prevent the expression of one gene from interfering with that of neighboring genes, and in eukaryotes these regions can be large (10). In contrast, the mean intergenic distances in the genomes of the microsporidia *E. cuniculi* and *Antonospora locustae* are only 129 bp and 211 bp, respectively (6, 8), whereas nucleomorphs intergenic regions are further reduced (4, 5).

When genomes reach this level of compaction, it is likely that fundamental processes like transcription are substantially affected. Indeed, the sequences of two transcripts from the *B. natans* nucleomorph have been shown to encode more than one gene, suggesting either that termination control is substantially altered or that nucleomorphs use polycistronic messages (7), like prokaryotes and, in some rare cases, eukaryotes (11, 12). Determining whether these are exceptional cases requires more data, but no systematic analysis of gene expression has been carried out in any highly compacted nuclear genome. Here we report EST surveys of three independently evolved compact genomes: the microsporidian *A. locustae* and the nucleomorphs of *B. natans* and *G. theta*. A large proportion of mRNAs from all three genomes encode multiple genes or gene fragments, sometimes as many as three additional genes apart from the one assumed to be the target of expression. Overall, transcript structure in these organisms suggests that promoter elements and termination signals may have been squeezed from shrinking intergenic regions and embedded in adjacent genes. Genome reduction may therefore result in paradoxically wasteful transcription systems that must at the same time cope with levels of transcriptional overlap that would probably not be tolerated in most genomes.

Materials and Methods

cDNA Library Construction and Sequencing. A total of 1×10^8 purified spores of *A. locustae* (ATCC 3086) from M&R Durango Biocontrol (Bayfield, CO) were ground under liquid nitrogen and RNA-purified by using TRIzol reagent (Invitrogen). mRNA was extracted from total RNA by using oligo dT cellulose powder. This mRNA was reverse-transcribed by using

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: *rpl24*, ribosomal protein L24.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. DQ057484–DQ057579 and DQ071178–DQ071262).

*To whom correspondence should be addressed. E-mail: pkeeling@interchange.ubc.ca.

© 2005 by The National Academy of Sciences of the USA

but also that frataxin sequence forms part of the 3' UTR of the upstream gene transcript. Results from neither method are consistent with contaminating DNA. Although these examples illustrate that *A. locustae* transcription patterns are unusual, transcripts of the most common single gene ESTs have a strong tendency to end at a single particular polyadenylation site (Table 2, which is published as supporting information on the PNAS web site), indicating that transcription termination/polyadenylation can be consistent.

The diversity and complexity in structure of *A. locustae* multigene transcripts raise questions about the large proportion of cDNAs that encode only the antisense of identifiable genes (Fig. 1A). Given the structure of many multigene transcripts and the fact that some cDNAs encode fragments of more than one antisense gene, it is possible that antisense transcripts may represent truncated cDNAs of multigene transcripts. These data suggest that the overall proportion of multigene transcripts may be higher than currently recognized: multigene plus antisense transcripts together represent approximately a quarter of *A. locustae* cDNAs with homologues in National Center for Biotechnology Information databases.

Microsporidian genomes are atypical, conspicuously in their highly reduced and compacted nature. The content and characteristics of these genomes have been studied in some detail (6, 8, 9), but whether this has any effects on genome function is not clear. The unusual nature of transcription in *A. locustae* indicates that compaction may indeed have an impact on expression. One of the obvious characteristics of compacted microsporidian genomes is the short intergenic regions: in *A. locustae* intergenic spaces average only 211 bp (8). If the pressure to shrink intergenic regions is strong enough, they could be reduced beyond the minimal size needed to encode the essential control regions for expression. The severe reduction of a 3' intergenic space could, for example, eliminate existing transcription termination signals and force transcription termination fortuitously within the next gene or, if it is in the same strand, using the downstream gene's existing termination signals. Consistent with this idea, the mean intergenic space between genes encoded on multigene transcript mRNAs is only 119 bp, just over half the average for this genome.

Transcript Structure in Nucleomorphs. If multigene transcripts in *A. locustae* are the result of its compacted genome, then we might expect transcription in other gene-dense nuclear genomes to share similar characteristics; therefore, we carried out EST surveys of *G. theta* and *B. natans* to examine transcription of nucleomorph genes. Not only are these nucleomorph genomes the smallest and most compact of any nuclear genomes, but the *G. theta* and *B. natans* nucleomorphs have evolved in parallel from a red and green alga, respectively (4), so they give us two independent points of comparison. Two multigene transcripts from the *B. natans* nucleomorph have been characterized previously (7); however, with only two examples it is not clear whether these messages are representative or not. Because in one case the coding regions characterized were in the sense strand, they could be interpreted as polycistronic or processed messages (7). A larger sample from both nucleomorphs would discern these possibilities from multigene transcripts as observed in microsporidia.

A total of 2,125 and 3,448 ESTs were sequenced from *G. theta* and *B. natans*, respectively, and transcripts derived from nucleomorph genes were identified by comparison with genomic sequence resulting in 52 and 38 nucleomorph loci, respectively (the vast majority of transcripts being from host nuclear genes because nucleomorph transcription does not form a large proportion of expression in the cell). Once completely sequenced, 19 and 3 of these loci, respectively, appeared to represent spurious products that do not clearly correspond to mRNAs of

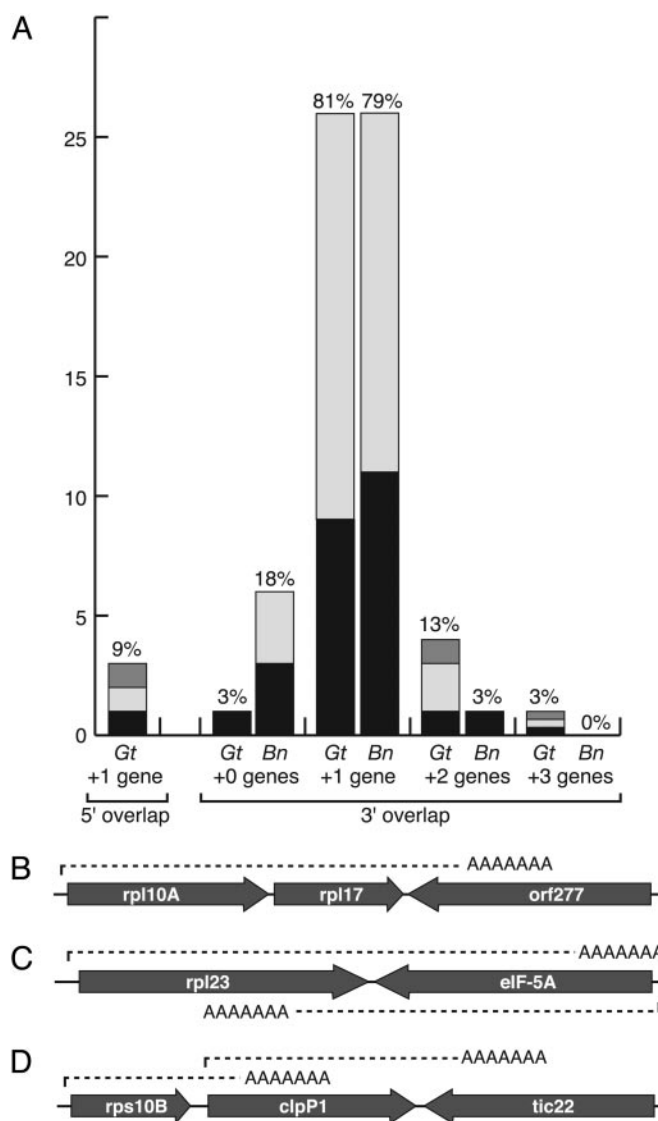


Fig. 4. Summary of nucleomorph transcripts from *G. theta* and *B. natans*. (A) Graph showing frequency of overlap with adjacent genes. The y axis shows (left to right) overlap with another gene at the 5' end (nearly all *B. natans* cDNAs were 5' truncated, so only *G. theta* data are shown), no overlap at the 3' end (i.e., a single gene transcript), or 3' overlap of one, two, or three additional genes. The x axis represents the total occurrence of each class of cDNA, and the bars are further subdivided to indicate whether the adjacent genes are protein-coding genes in the same strand as the target gene (black), in the opposite strand (light gray), or tRNA genes (gray). The percentage of the total number of cDNAs is shown above each bar. (B–D) Examples of *G. theta* multigene transcripts including cases where two genes are encoded in the same strand (B), where transcripts for two convergent adjacent genes overlap (C), and where transcripts from two parallel adjacent genes overlap (D). The last case is of particular interest because it shows that transcripts can read through termination signals for upstream transcripts.

any particular gene or are truncated at A-tracks in the genome, and these were not analyzed further. The structure of the cDNAs for the remaining loci were examined for all genes and gene fragments encoded within them (Tables 3 and 4, which are published as supporting information on the PNAS web site). The overall picture from these data are that multigene transcripts are more common in nucleomorphs than in microsporidia (Fig. 4A), but they are simpler in some respects. Identifying the likely expressed gene was not difficult, because virtually every cDNA

complete genome sequence of the apicomplexan parasite *Cryptosporidium* has shown it to be reduced and somewhat compacted (34). We examined a small number of short *Cryptosporidium* EST sequences from public databases (www.cryptodb.org) and found at least 10 of 576 instances where cDNAs encoded fragments of more than one gene. Whether these all correspond to mRNA is not verifiable, and, because the sequences are short, these data are not appropriate to estimate the frequency of multigene transcripts in this organism. Nevertheless, observations from microsporidia and nucleomorphs suggest that this process should be sought in *Cryptosporidium*. Last, there is no reason to expect that multigene transcripts are restricted to cases where whole genomes are compacted. Transcription of genes within compacted regions of otherwise normal genomes may follow similar patterns, as indeed has been shown in one region of the yeast genome (35).

A high frequency of multigene transcripts also has important practical implications. Several of the eukaryotes with compact genomes (such as microsporidia or *Cryptosporidium*) are parasites of medical or commercial importance. Because of their requirement for a host, studying these parasites and their interactions with their hosts can be challenging. One method to circumvent these problems when a complete genome is available is to examine expression profiles at various stages of infection

using methods such as arrays or serial analysis of gene expression. However, these methods will be misleading in organisms with high levels of multigene transcripts, because “gene” sequences will be encoded in untranslated regions of other genes (some methods distinguish strands, and others do not). Indeed, in *A. locustae* we have many instances of completely sequenced cDNAs, and it is still not possible to conclusively state which gene is being expressed. By using less direct methods, the chance for error is clear. It will be important to determine the frequency of multigene transcripts in other microsporidia where transcription profiles are desirable (e.g., the human parasite *E. cuniculi*) and to confirm whether they are also abundant in *Cryptosporidium* and other organisms with genomic characteristics to suggest that they may be prevalent.

We thank G. I. McFadden and P. R. Gilson for providing the *B. natans* cDNA library and for use of the nucleomorph sequence. This work was supported by a grant from the Canadian Institutes for Health Research (CIHR) and a New Investigator Award from the Burroughs-Wellcome Fund. *A. locustae* and *G. theta* EST sequencing was supported by the Protist EST Program of Genome Canada/Genome Atlantic, and *B. natans* EST sequencing was supported by a grant from the Natural Sciences and Engineering Research Council. P.J.K. is a scholar of the Canadian Institute for Advanced Research and a New Investigator of the CIHR and Michael Smith Foundation for Health Research.

1. Peyretailade, E., Biderre, C., Peyret, P., Duffieux, F., Méténier, G., Gouy, M., Michot, B. & Vivarès, C. P. (1998) *Nucleic Acids Res.* **26**, 3513–3520.
2. Keeling, P. J. & Fast, N. M. (2002) *Annu. Rev. Microbiol.* **56**, 93–116.
3. Biderre, C., Pagès, M., Méténier, G., Canning, E. U. & Vivarès, C. P. (1995) *Mol. Biochem. Parasitol.* **74**, 229–231.
4. Gilson, P. R. & McFadden, G. I. (2002) *Genetica* **115**, 13–28.
5. Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L. T., Wu, X., Reith, M., Cavalier-Smith, T. & Maier, U. G. (2001) *Nature* **410**, 1091–1096.
6. Katinka, M. D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prenier, G., Barbe, V., Peyretailade, E., Brottier, P., Wincker, P., et al. (2001) *Nature* **414**, 450–453.
7. Gilson, P. R. & McFadden, G. I. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 7737–7742.
8. Slamovits, C. H., Fast, N. M., Law, J. S. & Keeling, P. J. (2004) *Curr. Biol.* **14**, 891–896.
9. Vivarès, C. P., Gouy, M., Thomarat, F. & Méténier, G. (2002) *Curr. Opin. Microbiol.* **5**, 499–505.
10. Nelson, C. E., Hersh, B. M. & Carroll, S. B. (2004) *Genome Biol.* **5**, R25.
11. Jacob, F. & Monod, J. (1961) *J. Mol. Biol.* **3**, 318–356.
12. Blumenthal, T. (1998) *BioEssays* **20**, 480–487.
13. Fast, N. M., Law, J. S., Williams, B. A. & Keeling, P. J. (2003) *Eukaryot. Cell* **2**, 1069–1075.
14. Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S. & Hayashizaki, Y. (2003) *Genome Res.* **13**, 1324–1334.
15. Vanhee-Brossollet, C. & Vaquero, C. (1998) *Gene* **211**, 1–9.
16. Hurowitz, E. H. & Brown, P. O. (2003) *Genome Biol.* **5**, R2.
17. Hansen, K., Birse, C. E. & Proudfoot, N. J. (1998) *EMBO J.* **17**, 3066–3077.
18. Gerads, M. & Ernst, J. F. (1998) *Nucleic Acids Res.* **26**, 5061–5066.
19. Peterson, J. A. & Myers, A. M. (1993) *Nucleic Acids Res.* **21**, 5500–5508.
20. Slamovits, C. H. & Keeling, P. J. (2004) *J. Mol. Biol.* **341**, 713–721.
21. Cavalier-Smith, T. (2005) *Ann. Bot. (London)* **95**, 147–175.
22. Van'T Hof, J. & Sparrow, A. H. (1963) *Proc. Natl. Acad. Sci. USA* **49**, 897–902.
23. Hurst, L. D., Williams, E. J. & Pal, C. (2002) *Trends Genet.* **18**, 604–606.
24. Sameshima, J. H., Wek, R. C. & Hatfield, G. W. (1989) *J. Biol. Chem.* **264**, 1224–1231.
25. Prescott, E. M., Proudfoot, N. J., Furger, A., Dye, M. J. & Greger, I. H. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 8796–8801.
26. Springer, C., Valerius, O., Strittmatter, A. & Braus, G. H. (1997) *J. Biol. Chem.* **272**, 26318–26324.
27. Martens, J. A., Laprade, L. & Winston, F. (2004) *Nature* **429**, 571–574.
28. Kruglyak, S. & Tang, H. (2000) *Trends Genet.* **16**, 109–111.
29. Proudfoot, N. (2004) *Curr. Opin. Cell Biol.* **16**, 272–278.
30. Elmendorf, H. G., Singer, S. M. & Nash, T. E. (2001) *Nucleic Acids Res.* **29**, 4674–4683.
31. Chen, J., Sun, M., Kent, W. J., Huang, X., Xie, H., Wang, W., Zhou, G., Shi, R. Z. & Rowley, J. D. (2004) *Nucleic Acids Res.* **32**, 4812–4820.
32. Dahary, D., Elroy-Stein, O. & Sorek, R. (2005) *Genome Res.* **15**, 364–368.
33. Courties, C., Perasso, R., Chretiennot-Dinet, M. J., Gouy, M., Guillou, L. & Troussellier, M. (1998) *J. Phycol.* **34**, 844–849.
34. Abrahamsen, M. S., Templeton, T. J., Enomoto, S., Abrahante, J. E., Zhu, G., Lancto, C. A., Deng, M., Liu, C., Widmer, G., Tzipori, S., et al. (2004) *Science* **304**, 441–445.
35. Puig, S., Perez-Ortin, J. E. & Matallana, E. (1999) *Curr. Microbiol.* **39**, 369–373.