Update TRENDS in Genetics Vol.23 No.4

Research Focus

Ostreococcus tauri: seeing through the genes to the genome

Patrick J. Keeling

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

The marine green alga *Ostreococcus tauri* is the smallest-known free-living eukaryote. The recent sequencing of its genome extends this distinction, because it also has one of the smallest and most compact nuclear genomes. For other highly compacted genomes (e.g. those of microsporidian parasites and relic endosymbiont nucleomorphs), compaction is associated with severe gene loss. By contrast, *O. tauri* has retained a large complement of genes. Studying *O. tauri* should shed light on forces, other than parasitism and endosymbiosis, that result in densely packed genomes.

Genome compaction

Similar to the old adage about woods, sometimes it can be difficult to see the genome for the genes. Genomes are far more than the sum of the genes that they encode: without some level of order and control imparted by the genome, the genes would not 'survive'. Unfortunately, these genomic characteristics are more difficult to identify than the genes, especially in the nuclear genomes of eukaryotes, which are larger, more poorly sampled and apparently more loosely organized than those of prokaryotes. Sometimes, however, certain aspects of genomic organization are taken to extremes, and studying these genomes might make it easier to identify important principles. The recently sequenced genome of the marine green alga *Ostreococcus tauri*, reported by Derelle *et al.*, provides a good example [1]. When it was first discovered, *O. tauri* was immediately noted for its diminutive proportions [2], and its genome has lived up to this reputation. It is among the smallest of the known nuclear genomes, and its level of gene compaction is more akin to the hyper-compacted genomes of certain intracellular parasites and relic nuclei of endosymbiotic organelles [3–6] than to any other free-living cell. Because of these extremes, the *O. tauri* genome itself stands out against the backdrop of the genes encoded by it.

Characteristics of the *O. tauri* nuclear genome

The *O. tauri* genome [1] consists of 12.5 Mb distributed over 20 chromosomes. Altogether, 8166 protein-coding genes were identified, representing an impressive 81.6% of the genome. On average, these genes are separated by only 196 bases, giving an overall gene density of 1.54 kb per gene. Most gene families have been reduced in complexity compared with other green algae and plants. There are also several interesting features relating to the light-harvesting antennae, the carbon assimilation machinery and the possibility of C_4 photosynthesis. Overall, however, the major metabolic and information-processing pathways are fully represented, as expected for a free-living cell.

The *O. tauri* genome is not a homogeneous entity, and this feature is particularly exemplified by two regions: chromosome 19 and part of chromosome 2. Both of these regions contain fewer genes, encode a greater density of transposons and have a lower G+C content than the rest of the genome. Genome heterogeneity has been observed in other organisms, including those with small and compacted genomes, but it is exceptional in *O. tauri* because it is clearly associated with two discrete regions that differ in several respects. By reconstructing the phylogenetic relationships of the proteins encoded by these regions, Derelle *et al.* determined that most genes in the unusual region of chromosome 2 were closely related to homologues in other green algae. They concluded that chromosome 2 is of endogenous origin but evolved under different selective pressures from the rest of the genome, perhaps as a sex chromosome. By contrast, most genes on chromosome 19 do not have a clear and specific phylogenetic association with homologues in other green algae. This led to the suggestion that the unusual features of this chromosome indicate a recent and exogenous origin by horizontal chromosome transfer. The genes on this chromosome also tend to be poorly conserved compared with the rest of the genome [1], so an alternative possibility is that, similar to chromosome 2, it is an endogenous element that evolved under different selective pressures from the rest of the genome.

Comparisons with other reduced and compacted genomes

The genome sizes of eukaryotes differ markedly (Figure 1). The largest known genome is 1.8×10^6-fold larger than the smallest [7], and even within some groups of organisms (e.g. amoebae or plants), spectacular differences are seen between closely related organisms. Genome size is not, as is often assumed, linked to the 'complexity' of an organism or even to the number of genes in its genome [8], an observation known as the C-value paradox. Many attempts have been made to explain the C-value paradox through correlations with another feature of the genome or cell [9–12]; the best so far is cell size or nuclear size (although this does not always hold at the extremes) [7,10].

Corresponding author: Keeling, P.J. (pkeeling@interchange.ubc.ca).
Available online 28 February 2007.

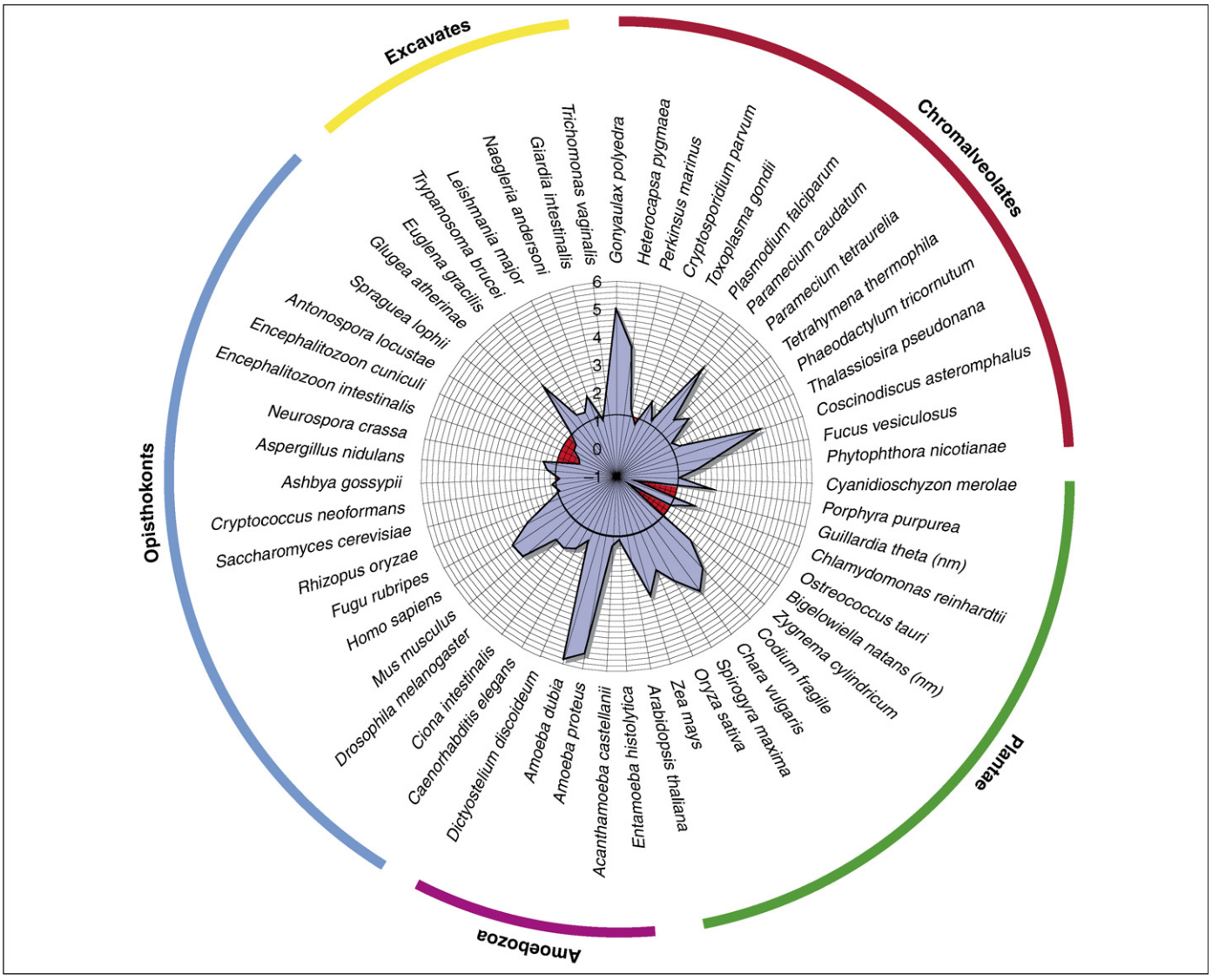

Figure 1. Nuclear genome size variation. Genome size estimates are plotted on a log scale in blue at the centre, and corresponding organism names and the 'supergroup' to which they belong [20] denoted at the periphery. Some genome size figures are based on complete genome sequences, but some are estimates based on other methods and may not be as accurate, in particular the very large genomes as they are poorly studied. The *Ostreococcus tauri* genome size is indicated by a circle so that smaller genomes are indicated by red areas. Abbreviation: nm, nucleomorph.

Compaction of the genome is a related issue but differs in several crucial ways. Most fundamentally, genome size is a measure of quantity of DNA, whereas compaction is a measure of density. Therefore, it is possible to have a genome that is both large and compact, but interestingly, this does not seem to occur often.

Moreover, the extent of compaction might not be entirely uniform across a genome. Many nuclear genomes have compacted to ~2 kb per gene [e.g. small fungal genomes such as those of *Saccharomyces cerevisiae* and *Ashbya gossypii* (also known as *Eremothecium gossypii*)], but few have exceeded this. Those few, the 'hypercompacted' genomes [6], have until now been restricted to microsporidians and nucleomorphs. Microsporidia are obligate intracellular parasites and have a highly reduced gene complement, probably because they can rely on their host for most of their nutritional and energy requirements [13,14]. Nucleomorphs, by contrast, are relic nuclei of endosymbiotic algae found in two lineages: a green algal endosymbiont in chlorarachniophytes and a red algal endosymbiont in cryptomonads [15,16]. Nucleomorphs are even more reduced than microsporidians because they are genetically integrated with the host: they rely on their host not only for energy and nutrients but also for products of genes that moved to the host nucleus [17]. In nucleomorph lineages and microsporidians, gene numbers have been reduced markedly, but the genomes also compacted to about twice the density of the *S. cerevisiae* genome. These genomes are perhaps the most interesting points of comparison with the *O. tauri* genome, because the *O. tauri* genome has compacted to nearly the same degree: the density of *O. tauri* is 1.54 kb per gene, compared with 1.25 kb per gene for the microsporidian *Encephalitozoon cuniculi*, 1.21 kb per gene for the nucleomorph of the chlorarachniophyte *Bigelowiella natans*, and 1.13 kb per gene for the nucleomorph of the cryptomonad *Guillardia theta* (all completely sequenced genomes) [13,15,16]. This is interesting because the *O. tauri* genome has compacted

without the concomitant loss of genes. Indeed, the loss of genes in parasites and endosymbionts is expected because of their host dependence, but why should the genome also compact? In microsporidians and nucleomorphs, compaction is frequently associated with their status as a parasite or endosymbiont, but there has never been a particularly compelling explanation for why these life strategies should lead to genome compaction. For example, the argument that replicating the genome more quickly is advantageous does not apply to nucleomorphs, because there is no advantage to replicating faster than the large host genome. From *O. tauri*, it is evident that the genome of a free-living cell can compact to almost the same degree as these extreme cases, bringing us back to the idea that compaction is a result of cell size [7].

Derelle *et al.* did not include either microsporidians or nucleomorphs in their comparative analyses. This is unfortunate, because we might learn more by comparing genomes that independently converged on similar states than we could from most other comparisons. In fact, the *O. tauri* genome is structurally similar to those of nucleomorphs in many ways. However, similar to other compacted genomes, it has reached its present state through a unique combination of factors (Box 1), and it lacks some characteristics that evolved independently in both nucleomorphs and microsporidians (e.g. rRNA operons as subtelomeric repeats).

Ostreococcus and comparative genomics

The *O. tauri* nuclear genome is a starting point for comparative analyses at many levels. One level might be an intracellular comparison: *O. tauri* also contains mitochondrial and plastid genomes, and one immediately wonders if all genomes in a cell evolved under any of the same selective pressures. The mitochondrial and plastid genomes have been completely sequenced (GenBank Accession No.s CR954200 and CR954199; and H. Moreau, personal communication) and are relatively small and gene dense compared with the organelle genomes of other green algae. The plastid gene density is similar to that of the parasitic green alga *Helicosporidium* [18], reminiscent of the similarities between the nuclear genome of *O. tauri* and that of microsporidians. Intraspecific and interspecific comparisons will be possible soon, because genome-sequencing projects are underway for another *O. tauri* strain and for the high-light-adapted species *Ostreococcus lucimarinus* (see the US Department of Energy Joint Genome Institute website; http://genome.jgi-psf.org). These organisms probably have the same overall characteristics, but identifying specific similarities and differences will be interesting: for example, do chromosomes 2 and 19 have similar abnormalities? Finally, intergeneric comparisons will be informative in several ways. Identifying closely related green algae with less compacted genomes might shed light on the *O. tauri* genome before it compacted. But comparisons with other hypercompacted genomes might be the most interesting, because these will reveal whether genome form has affected function in similar ways. For example, it will be interesting to determine whether *O. tauri* has the high levels of overlapping transcription that have been observed in microsporidians and nucleomorphs [19] and, if so, to determine how these different systems cope with this situation. For characteristics such as compaction and its effects on function, genomes such as that of *O. tauri* are similar to fun-house mirrors: they take otherwise rare processes and amplify them to an extreme degree. But they offer more than just a surprise, because understanding how these processes are affected in organisms with extreme genomes offers a unique way to understand how they work in normal genomes.

Box 1. How to shrink a genome

A variety of eukaryotic lineages are known to have compacted nuclear genomes. Although no two genomes have been compared using the same methods or criteria, it is still clear that several factors might operate when a genome shrinks [6]. In different genomes, different factors have contributed to different degrees, and in combination, these factors have contributed to the 'gene density' or the number of genes packed into a given space of sequence. The following are some of the factors that have made considerable contributions in one or more compacted nuclear genomes.

Gene content
Reducing the number of genes does not compact the genome (compaction refers to density not quantity), but compaction is often associated with a reduction in gene number. Indeed, the most compact genomes are those of the relic endosymbiont nucleomorphs and microsporidian parasites, all of which also have severely reduced gene content [13,15,16]. By contrast, *Ostreococcus tauri* has a compacted genome with a relatively large number of genes, which is a rare occurrence.

Introns
Some genomes have lost all but a few of their introns (e.g. microsporidian genomes). By contrast, curiously, others have retained large numbers of introns, but each one is severely reduced in size (e.g. chlorarachniophyte nucleomorph genomes). *O. tauri* has a moderate number of introns of normal size.

Transposons and repeats
Selfish elements and other repeated sequences are often dispensable and, accordingly, are rare in most highly compact genomes. Nevertheless, some regions of these genomes can vary in the density of such elements, as is illustrated by *O. tauri* chromosomes 2 and 19.

Intergenic spaces
Reducing the quantity of intergenic sequence has the greatest potential effect on the compactness of a genome. Intergenic spaces are defined by exclusion and therefore are anything but homogeneous sequence: many regions are crucial for survival, whereas others are filler, or 'junk'. The loss of junk is not difficult to understand; however, at some stage, the reduction in intergenic spaces starts to affect the control the cell has over important activities, in particular transcription. The point at which this line is crossed is unclear, but with average intergenic spaces of $<$200 bases, it is likely to have had an effect in *O. tauri*.

Gene size
One of the more unexpected findings to emerge from the sequencing of highly compacted genomes is that the genes themselves tend to be shorter on average than their homologues in other genomes. This characteristic might be due to a compacting force, but it might also be associated with the complexity of protein interactions in the cell. To distinguish these, it will be necessary to analyse a compact but gene-rich genome, such as that of *O. tauri*.

Acknowledgements

I thank H. Moreau for providing an unpublished manuscript on *O. tauri* organelle genomes and H. Moreau and C. Slamovits for critical comments. Genome evolution research in the laboratory of P.J.K. is supported by the Natural Sciences and Engineering Research Council of

Canada and the Tula Foundation. P.J.K. is a Fellow of the Canadian Institute for Advanced Research and a Michael Smith Foundation for Health Research Senior Investigator.

References

1 Derelle, E. *et al.* (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. U. S. A.* 103, 11647–11652
2 Courties, C. *et al.* (1994) Smallest eukaryotic organism. *Nature* 370, 255
3 McFadden, G.I. *et al.* (1997) Bonsai genomics: sequencing the smallest eukaryotic genomes. *Trends Genet.* 13, 46–49
4 Gilson, P.R. and McFadden, G.I. (2002) Jam packed genomes – a preliminary, comparative analysis of nucleomorphs. *Genetica* 115, 13–28
5 Vivarès, C.P. *et al.* (2002) Functional and evolutionary analysis of a eukaryotic parasitic genome. *Curr. Opin. Microbiol.* 5, 499–505
6 Keeling, P.J. and Slamovits, C.H. (2005) Causes and effects of nuclear genome reduction. *Curr. Opin. Genet. Dev.* 15, 601–608
7 Cavalier-Smith, T. (2005) Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann. Bot.* 5, 147–175
8 Mirsky, A.E. and Ris, H. (1951) The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J. Gen. Physiol.* 34, 451–462
9 Cavalier-Smith, T. (1978) Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *J. Cell Sci.* 34, 247–278
10 Gregory, T.R. (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev. Camb. Philos. Soc.* 76, 65–101
11 Vinogradov, A.E. (2004) Evolution of genome size: multilevel selection, mutation bias or dynamical chaos? *Curr. Opin. Genet. Dev.* 14, 620–626
12 Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science* 302, 1401–1404
13 Katinka, M.D. *et al.* (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414, 450–453
14 Méténier, G. and Vivarès, C.P. (2001) Molecular characteristics and physiology of microsporidia. *Microbes Infect.* 3, 407–415
15 Douglas, S. *et al.* (2001) The highly reduced genome of an enslaved algal nucleus. *Nature* 410, 1091–1096
16 Gilson, P.R. *et al.* (2006) Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. *Proc. Natl. Acad. Sci. U. S. A.* 103, 9566–9571
17 Gould, S.B. *et al.* (2006) Protein targeting into the complex plastid of cryptophytes. *J. Mol. Evol.* 62, 674–681
18 de Koning, A.P. and Keeling, P.J. (2006) The complete plastid genome sequence of the parasitic green alga, *Helicosporidium* sp. is highly reduced and structured. *BMC Biol.*4, 12 DOI: 10.1186/1741-7007-4-12 (www.biomedcentral.com/bmcbiol)
19 Williams, B.A.P. *et al.* (2005) A high frequency of overlapping gene expression in compacted eukaryotic genomes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 10936–10941
20 Keeling, P.J. *et al.* (2005) The tree of eukaryotes. *Trends Ecol. Evol.* 20, 670–676

0168-9525/$ – see front matter
doi:10.1016/j.tig.2007.02.008

LOH-proficient embryonic stem cells: a model of cancer progenitor cells?

Jason H. Bielas*, Ranga N. Venkatesan* and Lawrence A. Loeb

Gottstein Memorial Cancer Research Laboratory, Department of Pathology, University of Washington School of Medicine, Seattle, WA 98195-7705, USA

Cancers are thought to originate in stem cells through the accumulation of multiple mutations. Some of these mutations result in a loss of heterozygosity (LOH). A recent report demonstrates that exposure of mouse embryonic stem cells to nontoxic amounts of mutagens triggers a marked increase in the frequency of LOH. Thus, mutagen induction of LOH in embryonic stem cells suggests a new pathway to account for the multiple homozygous mutations in human tumors. This induction could mimic early mutagenic events that generate cancers in human tissue stem cells.

Mutations generate human cancers

Cancers are thought to arise in pluripotential stem cells, and, when clinically detected, these stem cells contain numerous mutations. Cancer cell genomes are frequently aneuploid (see Glossary), epigenetically altered and 'peppered' with mutations [1–3]. This raises the following important questions: (i) how are the mutations generated?; (ii) how are the mutations selected?; and (iii) what is the biological significance of the various mutations for tumorigenesis? The origins of mutations in cancer cells are unknown but include random events that damage DNA, such as attack by environmental carcinogens and reactive cellular metabolites. Mutations in oncogenes (e.g. *K-ras* and *myc*) and tumor-suppressor genes (e.g. *RB* and *APC*) can impart growth advantages to malignant cells, resulting in clonal selection [4–6]. Other, non-clonal mutations occur randomly throughout the genome and contribute to the characteristic heterogeneity of malignant cells within a tumor [7]. Various methods have been established to detect clonal mutations, including loss of heterozygosity (LOH) in

Glossary

Aneuploid: having a chromosome number that is not an exact multiple of the normal diploid number, with either fewer or more than the normal number of chromosomes in the cell.
Loss of heterozygosity (LOH): loss of the contribution of one parent to the genome of the cell.
Mutator phenotype hypothesis: that normal rates of mutation in somatic cells are insufficient to account for the multiple mutations observed in cancer cells; therefore, an increase in the mutation frequency is necessary to account for the large number of genetic changes observed in human tumors.

Corresponding author: Loeb, L.A. (laloeb@u.washington.edu).
* Authors contributed equally.
Available online 27 February 2007.