transcription-factor-binding site also correlates with expression rather weakly. 'It is widely accepted that knowledge of transcription-factor-binding motifs is not in itself adequate to elucidate transcriptional control mechanisms' [33]. Therefore, the facility of chromatin (de)condensation and B–Z-transition in the gene and around the promoter, which is tightly coupled (up to $r = 0.98$) with variation in GC content and CpG pattern, can determine the expression characteristics of a given gene in a synergistic (mutually enhancing) interplay with transcription-factor-binding sites.

Acknowledgements

I thank three anonymous reviewers and the editor for helpful comments. This work was supported by the Russian Foundation for Basic Research (RFBR) and by the Programme of the Presidium of the Russian Academy of Sciences 'Molecular and Cellular Biology' (MCB RAS).

References

1 Mouchiroud, D. *et al*. (1987) Compositional compartmentalization and gene composition in the genome of vertebrates. *J. Mol. Evol.* 26, 198–204
2 Larsen, F. *et al*. (1992) CpG islands as gene markers in the human genome. *Genomics* 13, 1095–1107
3 Goncalves, I. *et al*. (2000) Nature and structure of human genes that generate retropseudogenes. *Genome Res.* 10, 672–678
4 Ponger, L. *et al*. (2001) Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res.* 11, 1854–1860
5 D'Onofrio, G. (2002) Expression patterns and gene distribution in the human genome. *Gene* 300, 155–160
6 Vinogradov, A.E. (2003) Isochores and tissue-specificity. *Nucleic Acids Res.* 31, 5212–5220
7 Arhondakis, S. *et al*. (2004) Base composition and expression level of human genes. *Gene* 325, 165–169
8 Yamashita, R. *et al*. (2005) Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene* 350, 129–136
9 Robinson, P.N. *et al*. (2004) Gene-Ontology analysis reveals association of tissue-specific 5′ CpG-island genes with development and embryogenesis. *Hum. Mol. Genet.* 13, 1969–1978
10 Zhang, L. *et al*. (2004) GC/AT-content spikes as genomic punctuation marks. *Proc. Natl. Acad. Sci. U. S. A.* 101, 16855–16860
11 Semon, M. *et al*. (2005) Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. *Hum. Mol. Genet.* 14, 421–427
12 Vinogradov, A.E. (2005) Noncoding DNA, isochores and gene expression: nucleosome formation potential. *Nucleic Acids Res.* 33, 559–563
13 Bernardi, G. (2000) The compositional evolution of vertebrate genomes. *Gene* 259, 31–43
14 Vinogradov, A.E. (2001) Bendable genes of warm-blooded vertebrates. *Mol. Biol. Evol.* 18, 2195–2200
15 Turker, M.S. (2002) Gene silencing in mammalian cells and the spread of DNA methylation. *Oncogene* 21, 5388–5393
16 Yates, P.A. *et al*. (2003) Silencing of mouse Aprt is a gradual process in differentiated cells. *Mol. Cell. Biol.* 23, 4461–4470
17 Fazzari, M.J. and Greally, J.M. (2004) Epigenomics: beyond CpG islands. *Nat. Rev. Genet.* 5, 446–455
18 Vinogradov, A.E. (2003) DNA helix: the importance of being GC-rich. *Nucleic Acids Res.* 31, 1838–1844
19 Gilbert, N. *et al*. (2004) Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* 118, 555–566
20 Su, A.I. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 6062–6067
21 Suzuki, Y. *et al*. (2004) DBTSS, DataBase of transcriptional start sites: progress report 2004. *Nucleic Acids Res.* 32, D78–D81
22 Vinogradov, A.E. (2004) Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* 20, 248–253
23 Levitsky, V.G. (2004) RECON: a program for prediction of nucleosome formation potential. *Nucleic Acids Res.* 32, W346–W349
24 Levitsky, V.G. *et al*. (2001) Nucleosome formation potential of eukaryotic DNA: calculation and promoters analysis. *Bioinformatics* 17, 998–1010
25 Jordan, I.K. *et al*. (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* 19, 68–72
26 Oei, S.L. *et al*. (2004) Clusters of regulatory signals for RNA polymerase II transcription associated with *Alu* family repeats and CpG islands in human promoters. *Genomics* 83, 873–882
27 Aguilera, A. (2002) The connection between transcription and genomic instability. *EMBO J.* 21, 195–201
28 Horn, P.J. and Peterson, C.L. (2002) Molecular biology. Chromatin higher order folding – wrapping up transcription. *Science* 297, 1824–1827
29 Jenuwein, T. and Allis, C.D. (2001) Translating the histone code. *Science* 293, 1074–1080
30 Herbert, A. and Rich, A. (1999) Left-handed Z-DNA, structure and function. *Genetica* 106, 37–47
31 Herbert, A. and Rich, A. (1996) The biology of left-handed Z-DNA. *J. Biol. Chem.* 271, 11595–11598
32 Wray, G.A. *et al*. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* 20, 1377–1419
33 Frith, M.C. *et al*. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.* 32, 1372–1381

0168-9525/$ - see front matter
doi:10.1016/j.tig.2005.09.002

Genome size is negatively correlated with effective population size in ray-finned fish

Soojin Yi and J. Todd Streelman

School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, GA 30332, USA

A recent theory suggesting that genome size and complexity can increase as a passive consequence of small effective population size has generated much controversy. In this article, we demonstrate that freshwater fish species, which have smaller effective population sizes than marine fish species, have larger

Corresponding author: Yi, S. (soojin.yi@biology.gatech.edu).
Available online 5 October 2005

genomes. We show that genome size is negatively correlated with genetic variability, independent of phylogeny, body size and generation time. Genome duplication is also observed predominantly in freshwater fish. These results suggest that the raw materials of complexity originate under conditions of reduced selection efficiency.

Introduction

Genome size varies dramatically across the tree of life. In general, larger genomes are more complex than smaller genomes: they have more genes, more introns, larger intergenic regions and a higher proportion of mobile genetic elements [1,2]. Increases in genome size might have triggered the evolution of biological complexity, through regulatory diversification [3]. Therefore, it is a major goal of modern biology to understand the factors that influence genome size and complexity. Lynch and Conery [4] have postulated that variation in genome complexity is explained by differences in effective population size among species. The effective population size (N_e, the number of independent individuals contributing genes to the next generation) governs the efficiency of natural selection against a mutation with a given selective disadvantage (s). If the product of the effective population size and the selective disadvantage ($N_e s$) is sufficiently small, natural selection is inefficient and deleterious mutations can reach fixation by genetic drift [5].

Lynch and Conery [4] argue that mutations that increase genome size are deleterious, and that the small N_e of many eukaryotic species facilitates the fixation of those mutations. Increasing genome size and complexity can then evolve as a non-adaptive consequence of small N_e [4,6]. This proposal has been controversial [7–10]. It has been argued that the model is not supported for closely related groups [7], and that fundamental differences between prokaryotes and eukaryotes can compromise its predictions [8]. Most significantly, it is often difficult to disentangle the effects of confounding factors on genome size and N_e [9]. For example, genome size is negatively correlated with developmental rate and hence negatively correlated with body size, which is also correlated with N_e [9,11]. This might reflect a relationship between genome size, metabolic rate and/or other mechanisms that govern these properties. Furthermore, biologists rarely have access to true values of N_e. Instead, it is usually estimated from the level of neutral genetic variability, which is the product of N_e and the mutation rate (commonly designated as u) at equilibrium [4,5]. Because u can also vary between species, using $N_e u$ to infer effective population size can introduce an additional source of inaccuracy [8]. Mutation rate itself can co-vary with other traits such as body size, metabolic rate and generation time [12]. Finally, it is important to analyze well-defined species such that each taxonomic unit represents an independent realization of population genetic processes [7,8].

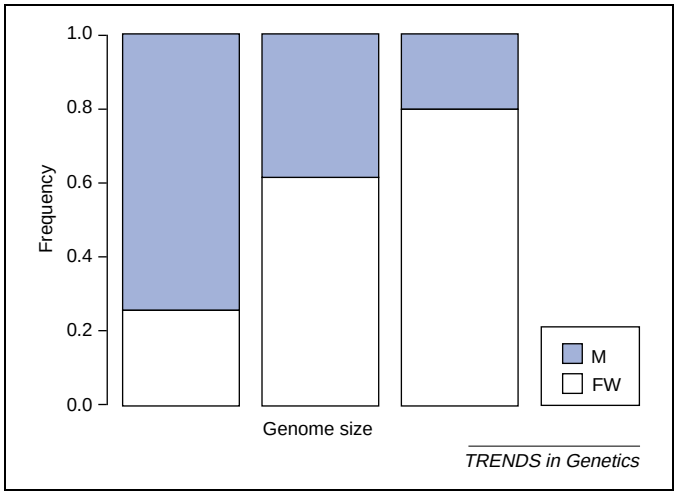

Figure 1. Marine species (M) have smaller genomes than freshwater species (FW). We divided 1043 taxa into three approximately equal bins, according to the distribution of genome sizes. Bins correspond to C-value $\leq$ 0.92 (353 species), 0.92 < C-value $\leq$ 1.26 (339 species) and 1.26 < C-value < 6.58 (351 species). Within each bin, the relative frequencies of M (blue) versus FW (white) species are shown.

Freshwater fish have larger genomes than marine fish

Here, we report analyses of the relationship between genome size and N_e in different species of ray-finned fish, addressing the concerns mentioned earlier. Ray-finned fish (Actinopterygii) comprise a monophyletic assemblage that diverged from other vertebrate lineages ~425 million years ago (Mya) [13]. We analyzed a non-redundant data set of 1043 ray-finned fish species (from ~190 families) whose genome sizes and habitat information are available (supplementary material online). Haploid genome size in ray-finned fish varies ~20 fold, from 0.37×10^9 bp (guineafowl pufferfish, *Arothron meleagris*) to 6.44×10^9 bp (shortnose sturgeon, *Acipenser brevirostrum*). Previous analyses with allozyme and microsatellite markers suggest that freshwater fish species have lower levels of genetic variability than marine species and hence a lesser N_e on average, because freshwater species are confined to smaller geographic areas for breeding [14–16]. Based on this expectation, we hypothesized that freshwater species should have larger genomes than marine species. Freshwater species indeed have larger genomes than marine species (mean $\pm$ SE = 1.480pg $\pm$ 0.069 versus 0.976pg $\pm$ 0.039, respectively; Mann-Whitney, $P < 10^{-3}$; Figure 1). We further analyzed the median genome sizes per genus (n = 593) and found the same pattern (mean $\pm$ SE = 1.343pg $\pm$ 0.035 versus 0.976pg $\pm$ 0.023, respectively, Mann-Whitney, $P < 10^{-3}$), suggesting that this result is not caused by a few predominant genera.

Genome size is negatively correlated with neutral variability in ray-finned fish

Next, we directly examined the relationship between genome size and the level of neutral genetic variability in a subset of species. For this analysis, we compiled estimates of expected heterozygosities from microsatellite markers. The expected heterozygosity is a function of $N_e u$ at equilibrium following the stepwise mutation model [17]. We used a transformed value of expected heterozygosity, referred to as 'heterozygosity' or H_T in the supplementary

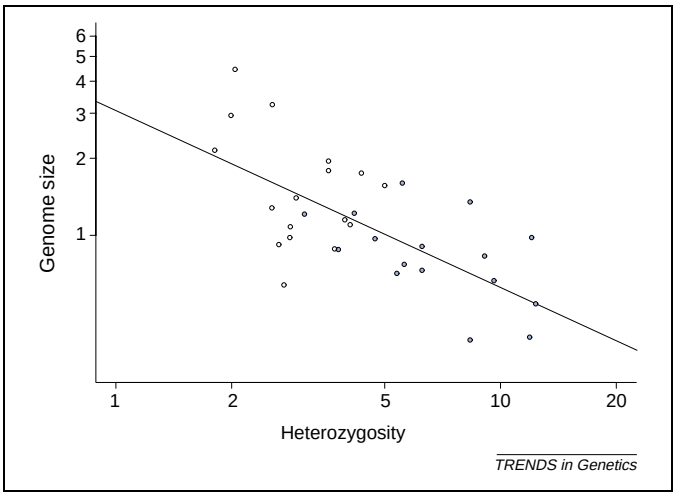

Figure 2. Heterozygosity (H_T) is negatively correlated with genome size in ray-finned fish. Blue circles indicate marine species, and the white circles indicate freshwater species. The regression between natural log-transformed variables is statistically significant ($P<10^{-4}$), with an intercept of 1.11 ± 0.22 (SE), a slope of -0.67 ± 0.14, and adjusted $r^2=0.42$, df=31.

material online. Our data set comprised 33 species (16 marine, 17 freshwater; from 24 families in 14 orders), about which we had information from four to 28 microsatellite loci (298 loci total) in population samples of 10–672 individuals (supplementary material online). We found a strong negative correlation between genome size and heterozygosity (Kendall's $\tau=-0.426$, $P<10^{-3}$, Figure 2). In other words, species with smaller effective population sizes tend to have larger genomes.

We further assessed the effects of other traits that might co-vary with N_eu, namely body size and generation time (supplementary material online). Body size and generation time are significantly correlated with each other in our data set (Kendall's $\tau=0.504$, $P<10^{-3}$), consistent with previous studies (e.g. [12]). To tease apart the relative contributions of these factors to estimates of genome size, we performed a multiple regression analysis with heterozygosity, body size and generation time as explanatory variables, and genome size as the dependent variable. We found that neither body size nor generation time had a significant effect on genome size, whereas heterozygosity had a significant effect ($F[1,31]=23.88$, $P<10^{-4}$).

The relationship between genome size and genetic variability is independent of phylogeny

The taxa in our data set are not statistically independent because of shared evolutionary history. For instance, owing to data availability, fish from orders Salmonidae and Cyprinidae are over-represented in the analysis (Supplementary Table 1). To test the possibility that such bias can confound our results, we computed phylogenetically independent contrasts between character values (supplementary material online). Using recent molecular phylogenies of ray-finned fish as the reference, we calculated 23 independent contrasts in our data set. Genome size and H_T are significantly correlated among independent contrasts (Kendall's $\tau=-0.31$, $P=0.04$). As above, heterozygosity is the only variable with a significant effect on genome size in multiple regression analysis of contrasts ($F[1,21]=5.20$, $P=0.03$).

Concluding remarks

Therefore, we have shown that genome size in ray-finned fish is negatively correlated with N_e, and this relationship is independent of phylogeny, body size and generation time. Our results lend strong support to the idea that reduced N_e underlies the evolution of larger and more complex genomes. Another observation that can be explained by this notion is the distribution of polyploidization in ray-finned fish. All recorded cases of polyploidy (~27 instances in eight orders; see Table 1 in Ref. [18]), after the genome duplication in the ancestor of ray-finned fish [19], are observed in freshwater lineages. Reduced N_e in freshwater fish might have permitted fixation of otherwise deleterious [20] mutations that led to genome duplications. Future research will address whether passive increases in genome size have in fact been co-opted for the adaptive evolution of complexity in fish and other lineages.

Acknowledgements

We thank R. Blanton for literature survey; B. Charlesworth, M. Lynch, T. Kocher, E. Vigoda and M. Roberts for discussions and/or comments on the article. S.Y. and J.T.S. are supported by start-up funds from the Georgia Institute of Technology.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2005.09.003

References

1 Carroll, S.B. (2001) Chance and necessity: the evolution of morphological complexity and diversity. *Nature* 409, 1102–1109
2 Rubin, G.M. *et al*. (2000) Comparative genomics of the eukaryotes. *Science* 287, 2204–2215
3 Levine, M. and Tjian, R. (2003) Transcription regulation and animal diversity. *Nature* 424, 147–151
4 Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science* 302, 1401–1404
5 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, UK
6 Lynch, M. (2002) Intron evolution as a population genetic process. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6118–6123
7 Vinogradov, A.E. (2004) Testing genome complexity. *Science* 304, 389–390
8 Daubin, V. and Moran, N.A. (2004) *Comment on 'the origins of genome complexity.' Science* 306, 978
9 Charlesworth, B. and Barton, N. (2004) Genome size: does bigger mean worse? *Curr. Biol.* 14, R233–R235
10 Vinogradov, A.E. (2004) Evolution of genome size: multilevel selection, mutation bias or dynamical chaos? *Curr. Opin. Genet. Dev.* 14, 620–626
11 Cavalier-Smith, T. (1985) *The Evolution of Genome Size*, John Wiley, Chichester, UK
12 Martin, A.P. and Palumbi, S.R. (1993) Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. U.S.A.* 90, 4087–4091
13 Kikugawa, K. *et al.* (2004) Basal jawed vertebrate phylogeny inferred from multiple nuclear DNA-coded genes. *BMC Biology* DOI: 10.1186/1741-7007-2-3 1741–7007/2/3 (http://www.biomedcentral.com/1741-7007/2/3).

14 DeWoody, J.A. and Avise, J.C. (2000) Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. *J. Fish Biol.* 56, 461–473
15 Gyllensten, U. (1985) The genetic structure of fish: differences in the intraspecific distribution of biochemical genetic variation between marine, anadromous, and freshwater fishes. *J. Fish Biol.* 26, 691–699
16 Ward, R.D. *et al*. (1994) A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. *J. Fish Biol.* 44, 213–232
17 Ota, T. and Kimura, M. (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.* 22, 201–204
18 Le Comber, S.C. and Smith, C. (2004) Polyploidy in fishes: patterns and processes. *Biol. J. Linn. Soc.* 92, 431–442
19 Jaillon, O. *et al*. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431, 946–957
20 Andalis, A.A. *et al*. (2004) Defects arising from whole-genome duplications in *Saccharomyces cerevisiae*. *Genetics* 167, 1109–1121

0168-9525/$ - see front matter
doi:10.1016/j.tig.2005.09.003