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 Dr., Atlanta, GA 30332, USA
Corresponding author: Yi, Soojin (soojin.yi@biology.gatech.edu).

Materials and Methods

Genome Size

The haploid genome size (C-value, in pg) of 1461 ray-finned fish species (Actinopterygii) was downloaded from the Animal Genome Size Database (www.genomesize.com). Unless otherwise stated, comparisons of genome sizes are performed on the C-values, which can be converted to number of base pairs by the following: number of base pairs = C-value × 0.978 × 10^{9}. Several species are represented more than once in this database, often with different estimates of genome size owing to differences in the methods of estimation and the cell types used. We collapsed such redundant cases to single data points by either taking the average (when there are two entries per species) or taking the median (if more than three entries observed for each species). Clear case of recent polyploidy (induced by hybridization or experimental manipulation) has been removed from our data set.

Habitat assignment

We separated fishes into two categories based on the habitat in which they breed (freshwater, FW or marine, M). This information was gathered from Nelson [1] and/or from FishBase (www.fishbase.org). In most cases, this distinction is straightforward. Anadromous fishes such as salmonids (live in oceans but breed in rivers) were assigned to the FW category, whereas catadromous fishes such as the American eel (which live in rivers but breed in the Sargasso Sea) were assigned to the M category. We removed 26 cases where the assignment was ambiguous.

Microsatellite heterozygosity

Data Collection: Data on microsatellite heterozygosity were collected from the literature (see Supplemental Table 1). Frequencies of marker types (di-, tri- and tetra-nucleotides) did not differ between marine or freshwater groupings (P>0.09). We excluded marker loci if the cloned allele was (i) monomorphic in sampled individuals, (ii) short (less than six repeat units) or (iii) complex and interrupted. We confined analysis to loci cloned from focal species to avoid ascertainment bias. In cases of known ancient polyploidy (e.g. sturgeon, paddlefish), only disomic loci were included.

Population sampling: In the majority of examples (n = 30), we report mean heterozygosity averaged over multiple loci from a single population. In Anguilla anguilla we used the average intra-population heterozygosity value (six loci) averaged over 11 populations, as given in the literature. For Danio rerio, we used the average intra-population heterozygosity value (six loci) averaged over four populations (SD = 0.036). In the case of Thymallus thymallus, we employed the grand mean of 17 loci from 15 populations (SD = 0.079).

Transformation: According to the stepwise mutation model [2], expected microsatellite heterozygosity (He) at equilibrium is equal to 1–(1+8Neue^−0.5). We take a transformed value of HR = (1–He)^−1, which is linearly related to Ne on a log–log scale for the range of He in our data set.

Phylogenetically independent contrasts: We used the 'Crunch' algorithm of CAIC v. 2.0.0 [3] without branch length information, to compute phylogenetically independent contrasts of transformed data,
using a phylogeny of taxa inferred from published molecular analyses [4,5]. The inferred phylogeny is presented in the Supplementary Figure.

**Multiple regression analyses**

To address the relationship between genome size and microsatellite heterozygosity, we further took into account the effects of body size and generation time. We recorded estimates of maximum body size (in mm) and generation time (in years), for each of the 33 species in Figure 2 in the main text, from FishBase (www.fishbase.org). All values are transformed by taking natural logarithms, to be approximately normally distributed before calculating phylogenetically independent contrasts. Qualitative conclusions are similar if the variables are not transformed or are always transformed (results not shown).

**References**


The Cyprinidae is one of the largest families of fishes (~190 genera, 2100 species). At various times, these species have been included in distinct sub-families.

A, B, C refer to distinct subfamilies

<table>
<thead>
<tr>
<th>Genomic Size</th>
<th>Genome Size</th>
<th>Genome Size</th>
<th>Genome Size</th>
<th>Genome Size</th>
<th>Genome Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6344</td>
<td>0.623</td>
<td>0.607</td>
<td>0.51</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>0.92</td>
<td>0.885</td>
<td>0.788</td>
<td>0.40</td>
<td>0.70</td>
<td></td>
</tr>
</tbody>
</table>

References: