Differential evolution of noncoding DNA across eukaryotes and its close relationship with complex multicellularity

Irma Lozada-Chávez[1]*, Peter F. Stadler[1-6] and Sonja J. Prohaska[1]

[1]Evo-Devo & Bioinformatics Group, Department of Computer Science – IZBI, University of Leipzig, Härtelstrasse 15-18, D-04107 Leipzig, Germany
[2] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
[3] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Competence Center for Scalable Data Services and Solutions, and Leipzig Research Center for Civilization Diseases, University Leipzig, Germany
[4] Fraunhofer Institute for Cell Therapy and Immunology, Perlickstrasse 1, D-04103 Leipzig, Germany
[5] Center for RNA in Technology and Health, University of Copenhagen, Grønnegardsvej 3, Frederiksberg C, Denmark
[6] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM, 87501, USA

* ilozada@bioinf.uni-leipzig.de

Background: Genome size enigmatically intersects genotype and phenotype. Despite several assessments have shown the functional relevance of some forms of non-protein-coding DNA (ncDNA), the distribution of different types of ncDNA across eukaryotes, its relationship with genome size and its influence on the emergence of complex multicellularity (CM) are still highly unclear and controversial.

Results: Here, we analyzed the distribution and genome content contribution of four types of ncDNA: spliceosomal introns, repeats, pseudogenes and unique ncDNA across 500 complete sequenced genomes distributed among the major eukaryotic groups. After accounting for phylogenetic signal, we found that genome size correlates weakly with intron and pseudogene features at the broadest phylogenetic scale. However, genome size correlates strongly and positively with repeat and unique ncDNA features, and negatively with several exon features. Our regressions are significantly robust to variation in phylogenetic topologies and genome size estimations. We also found that most intron features are moderately stable and show a minor contribution of repeats within the clades of Fungi and Viridiplantae, whereas introns features vary dramatically and have a major contribution of repeats in most metazoans and protists. Furthermore, the distribution of pseudogenes classes (*i.e.*, processed, duplicated and fragmented) are lineage specific and usually show a minor contribution from repeats. Interestingly, there is no correlation between the number of paralogous members within a protein family and its pseudogene complement. Strikingly, high non-repetitive intron and pseudogene contents are observed in several species with small genomes. After defining simple *versus* complex multicellularity, PCAs corrected for phylogenetic biases underscore a significant relationship between CM and both intron-richness and unique ncDNA.

Conclusions: In contrast to previous estimations [1,2], our findings show that several ncDNA classes have been decoupled from genome size evolution at the broadest phylogenetic scale in Eukarya. These findings also support the hypothesis that ncDNA has been the major genetic source for the exploration of independent paths to complex multicellularity despite its mainly non-adaptive origins across eukaryotes.

References:

[1] Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302:1401

[2] Lynch M, et al. (2011) The repatterning of eukaryotic genomes by random genetic drift. *Annu Rev Genomics Hum Genet.* 12:347