OPEN ACCESS Freely available online

PLoS GENETICS

Viewpoints

Statistical Inference on the Mechanisms of Genome Evolution

Michael Lynch*

Department of Biology, Indiana University, Bloomington, Indiana, United States of America

Introduction

In a series of publications, I and my colleagues have developed hypotheses for how the evolution of various aspects of genome architecture is expected to proceed under conditions in which the forces of random genetic drift and mutation predominate (e.g., [1–15]). These models, collectively referred to below as the mutational-hazard (hereafter, MH) hypothesis, are sometimes represented as neutral models [16,17], but this is not correct, as the key component of each model is the deleterious mutational consequence of excess DNA. The MH hypothesis is, however, a nonadaptational model, in that it yields expectations on the structure of genomes without invoking external selective forces.

It is likely that some aspects of these models will need to be changed as more is learned about the molecular consequences of various aspects of gene structure and the nature of mutation. Such modifications will not alter the need for baseline null hypotheses in attempts to defend adaptive explanations for variation in genomic architecture [9]. Nevertheless, any theory that strives to provide a unifying explanation for diverse sets of genomic observations must be scrutinized extensively from a variety of angles and interpreted in the context of well-established molecular and population-genetic processes. Although I will argue that a recent challenge to the MH hypothesis by Whitney and Garland ([18]; hereafter, WG) contains numerous problems, this exchange may help clarify more broadly misunderstood issues.

Errors in Statistical Logic and Analysis

Statistical theory provides a framework for rigorously testing hypotheses in biology, with two of the more dramatic examples being the formal theory of quantitative genetics [19] and phylogenetic inference [20]. Nevertheless, the utility of statistical methods for hypothesis testing depends critically on the extent to which the underlying model assumptions match the features of the system under investigation. Like an ill-defined verbal argument, overconfidence in an inappropriate quantitative analysis can lead to misleading interpretations.

Unfortunately, because large-scale changes in genomic architecture emerge on time scales of tens to hundreds of millions of years, tests of general theories of genome evolution are highly reliant on comparative data. This can raise issues regarding the significance of hypothesis tests when the underlying data share evolutionary history. Since Felsenstein [21] introduced the rationale for the phylogenetic comparative method, various derivative techniques have been developed, some by the author of this paper [22,23]. These approaches have been used broadly in evolutionary ecology, although not always with good justification (as emphasized in [24–26]). Using such methods, WG concluded that phylogenetic diversity of genomic features is unaffected by variation in the power of random genetic drift, challenging the MH hypothesis, but there are at least four classes of statistical problems associated with this study.

First, the analyses employed by WG are only justified when the characters under consideration have some possibility of shared evolutionary history among related taxa. The degree to which history is shared across related lineages is often unclear with phenotypic traits. However, the issues are well-understood for the central variable in the analyses of WG, the level of average nucleotide heterozygosity at silent sites (π_s), which has an expected value of $N_e u$ under mutation-drift equilibrium (where N_e is the effective population size, and u is the base-substitutional mutation rate per nucleotide site; ignoring, for simplification, the factor of 4 or 2 that should precede this expression in diploid versus haploid populations).

The expected coalescence time for a neutral gene genealogy, $4N_e$ generations in a diploid species, is dramatically less than the divergence time between even the most closely related species in WG's analysis (e.g., *Mus* and *Homo*, *Drosophila* and *Anopheles*, none of which share ancestral polymorphisms). Therefore, if any trait can be stated as having no shared phylogenetic history in the analyses of WG, it is the *estimator* of $N_e u$. Although all traces of ancestral π_s values have been erased many times over for the taxa in this study, one could perhaps still argue that some shared history remains with respect to the underlying population size and mutation rate determinants in some pairs of lineages, which might allow similar heterozygosity values to re-emerge. It is notable, however, that there is considerable turnover among lineages in the genes encoding for enzymes that dictate the mutation rate, with the replication polymerases in eukaryotes and eubacteria not even being orthologous, and the repair polymerases in numerous eukaryotic lineages being absent from others. In any event, this concern is dwarfed by other limitations, including the very high sampling variance associated with π_s estimates (the standard errors of estimates often being of the same order of magnitude as the estimates themselves), and the unknown element of temporal variation on time scales exceeding N_e generations. Because of such enormous sampling var-

Citation: Lynch M (2011) Statistical Inference on the Mechanisms of Genome Evolution. PLoS Genet 7(6): e1001389. doi:10.1371/journal.pgen.1001389

Editor: Nancy A. Moran, Yale University, United States of America

Published June 9, 2011

Funding: The author acknowledges the National Science Foundation, National Institutes of Health, and the Department of Defense for financial support. The funders had no role in the preparation of the article.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: milynch@indiana.edu

iation, this author has generally simply reported average estimates of π_s across wide phylogenetic groups (e.g., [5]). By deriving independent contrasts on π_s, WG greatly inflated the sampling variance of this parameter, and it can be shown that this problem alone will cause a ~30% decline in expected r^2 values involving correlations with other traits.

An equally substantial problem is associated with the strict interpretation of π_s as a measure (or linear correlate) of $N_e u$ across all of life. Most notably, many prokaryotes appear to approach the maximum level of N_e (and minimum level of u) dictated by the effects of selection on linked genes [7,15], in which case, the independent contrasts of true values of $N_e u$ between such species pairs will be essentially randomly distributed around zero. This problem is compounded by the downward bias in π_s-based estimates of $N_e u$ in unicellular species that results from selection on silent sites [5,7,27,28]. Even if we can be confident that $N_e u$ is much higher in prokaryotes than in vertebrates, the estimates based on π_s may be off by more than an order of magnitude [7].

Owing to the long time scale on which genomic alterations accrue, the concern for shared evolutionary history in such attributes might in some cases be more justified. However, for the lineages evaluated by WG, such phylogenetic inertia is overshadowed by other evolutionary effects. For example, for the two most closely related species included in the WG analysis, mouse and human (and most other eutherian mammals), numerous shared features of genome architecture are a consequence of convergent evolution, not shared ancestry [29]; the same is true of the ancestral species leading to the land-plant and metazoan lineages [7]. The complete turnover of various mobile-element families among eukaryotic lineages provides additional compelling evidence for the absence of strong phylogenetic effects among the taxa examined by WG. Thus, as in the case of factors influencing the mutation rate, it is unclear whether the aspects of shared biological history that are the targets of the WG analysis are any more meaningful than applying a similar strategy in combined study of bat, bird, and insect wings.

Second, use of a phylogenetic tree with questionable branch lengths will further obfuscate any phylogenetic analysis, as branch-length scaling must yield uniform sampling variances of the contrast data for downstream hypothesis tests to be valid. In an attempt to remove such issues, WG standardized all branch lengths to unit length, although there are no obvious evolutionary models that would produce the desired behavior for the characters examined. The relevant time scale for evolutionary processes is the number of generations per branch, whereas phylogenetic trees are simply based on net accumulations of nucleotide substitutions. Under the assumption that the molecular sites on which a tree is based are neutral (which can be questioned), the rate of mutation accumulation would be proportional to the product of the per-generation mutation rate and the number of generations elapsed. The first quantity varies by approximately two orders of magnitude among the species in this study [15], and the generation length varies by more than five orders of magnitude (from <1 hour to ~20 years). Thus, at the very least, the consequences of the arbitrary scaling to equal branch lengths are obscure.

A more significant issue is the validity of the topology of the phylogenetic tree employed. WG appear to have simply spliced together subtrees from several independent studies, many aspects of which continue to be highly debated. These include the issues of whether echinoderms and tunicates are monophyletic, and whether nematodes and arthropods are united in the ecdysozoa. Most phylogeneticists agree that the deep branching positions of all of the major eukaryotic lineages other than animals, fungi, and slime molds are highly uncertain. Thus, although some phylogenetic nonindependence may have been removed in the analyses of WG, numerous spurious internal relationships were also likely created, rendering the analysis much less rigorous than the authors imply.

Third, perhaps the most fundamental issue of the analysis of WG is the very nature of the hypothesis test that was carried out. Although the authors assumed that various measures of genome architecture will be linearly related to π_s on a logarithmic scale under the MH hypothesis, this is not what the theory predicts. Rather, the theory predicts a threshold response to $N_e u$ (or N_e) for many aspects of genome architecture, and such scaling can be seen in many genomic contexts, ranging from intron investment to mobile-element contributions to genome size itself [7]. Failure to account for this feature naturally eliminates any obvious scaling with $N_e u$ when independent contrasts are employed. For example, if most pairs of species reside to the right or left of a threshold, which is certainly the case with the taxa examined by WG, an independent-contrast analysis will produce a situation in which nearly all contrasts have expected values equal to zero, yielding a near-zero correlation (and removing all positional information with respect to the threshold). Thus, rather than being a contradiction of the MH hypothesis, a substantial reduction in the correlation of genomic attributes with the independent contrasts of π_s employed by WG is completely consistent with theoretical expectations.

Finally, it should be noted that when the features of the underlying data do not violate the assumptions of a statistical model (which is not the case in the WG study), ordinary least-squares correlations are, on average, unbiased with respect to the true underlying parameter, i.e., species sampling simply leads to greater noise among individual samples, but does not alter the average outcome [23,26,30]. Consequently, unlike the aberrant behavior observed by WG, relationships that evolve in a double-diffusion-like process generally yield similar correlations whether or not shared phylogenetic history is accounted for [24].

To improve the quality of future work in comparative genomics, WG advocate an even broader use of phylogenetic methods. However, unless a model more relevant to the tempo and time scale of evolution of the components of genomic evolution is incorporated, unless unbiased estimators of $N_e u$ can be procured, and unless appropriate metrics and topologies of the underlying phylogenies can be obtained, it appears that the methods being promoted by WG will be no more informative than ordinary least squares and may even continue to be misleading.

Biological Misinterpretations

To strengthen their argument that drift has little influence on genome architecture, WG claim that three other sets of observations are inconsistent with the MH hypothesis. For example, they note that Whitney et al. [31] found a low correlation of genome size with estimates of N_e derived from measures of allozyme heterozygosity in a wide variety of plants. Contrary to the authors' arguments, such estimates of N_e are quite problematic. First, because allozymes are functions of protein-sequence variation, they are much less reliable surrogates of neutral variation than silent sites. There is no theoretical basis for a positive correlation between allozyme variation and $N_e u$, and if there is substantial selection on allozymes, the relationship could even be negative. Second, although the authors extrapolated estimates of N_e by dividing levels of allozyme heterozygosity by a mutation rate of $u = 10^5$ per allele per generation (the basis of which is unclear), even if the assumption of neutrality were correct, this is an inappropriate manipulation. Per-generation mutation rates vary substantially across species in such a way that the very strong negative correlation between N_e and u results in π_s scaling only weakly with N_e [15]. Thus, although the observations in [31] are again superficially consistent with the MH hypothesis, no confident conclusions can be drawn from the results.

WG also suggest that the tendency for microbial genome sizes to decline with decreasing N_e [32] is inconsistent with the MH hypothesis. In fact, the opposite is true—the theory predicts that with increasing power of random genetic drift, effectively neutral genomic features will diverge in the direction of mutation bias. Because there is a deletion bias in bacteria, the observation of Kuo et al. [32] actually provides compelling support for the MH hypothesis, in that a pattern different from that in eukaryotes (where there is an insertion bias due to a strong contribution from mobile-element insertions) is both predicted and observed. Notably, this shift in the direction of mutation pressure is also a striking violation of the underlying assumption of a constant background pattern of stochastic evolution in the linear independent-contrasts methods employed by WG.

In advocating the need for better estimators for N_e, WG emphasize the utility of the K_a/K_s ratio of nonsynonomous to synonymous divergence, which is often used as a measure of the efficiency of selection. However, this overlooks two significant issues. First, the theoretical expectations of the MH hypothesis are not a simple function of N_e but of the product $N_e u$, which is the ratio of the power of mutation to the power of drift. Thus, the criticism that an estimator of $N_e u$ is a poor proxy for N_e is misplaced, as it is the former that is critical to testing the MH hypothesis, whereas the latter is insufficient. Fortunately, it is easier (although, as noted above, not easy) to estimate $N_e u$ than N_e. Second, the K_a/K_s index at best provides an estimate of the average efficiency of selection operating on amino acid substitutions, whereas the MH hypothesis is focused on the vulnerability of gene/genome-structural modifications to mutation pressure. There is no theoretical or empirical basis for expecting K_a/K_s to covary with $N_e u$. Although commonly used, it is not even clear that K_a/K_s scales appropriately with the efficiency of selection in populations of large size. If, for example, N_e is sufficiently large that nearly all nonsynonomous changes involve neutral substitutions, any further increase in N_e will have no effect on K_a while reducing K_s, and hence reducing K_a/K_s (contrary to the assumption that low K_a/K_s implies large N_e).

Moving Forward

In questioning the role of drift, and apparently mutation (based on their treatment of it as a nuisance parameter), in the evolution of genomic attributes, WG provide no alternative explanations for the numerous patterns of genomic structural variation known to exist within and among prokaryotes and eukaryotes. In contrast, the MH hypothesis provides a potential solution to the problem of why various aspects of animal and plant genomes evolve in opposite directions within organelles while converging within the nucleus; that the explanation is related to variation in u rather than N_e further demonstrates the difficulty of focusing solely on N_e, in accordance with the dual nature of the proposed process [13]. The MH hypothesis provides a plausible explanation for the expansion but near constancy of average UTR lengths in eukaryotes [12], for various aspects of intron evolution [4,33], and for numerous features in nonrecombining chromosomal regions [7]. The model expectations are also consistent with the genomic modifications incurred by endosymbiotic bacteria, and with the remarkable convergence of the features of integrated polydnaviral genomes on those of their insect host chromosomes. Finally, the hypothesis provides an explanation for the parallel contraction in numbers of retrotransposons, pseudogenes, and insertions of mitochondrial DNA into the nuclear genomes of independent mammalian lineages following the post-KT geographic expansion of mammals [29]. In short, the evidence that excess DNA is associated with weak mutational disadvantages is compelling, and by invoking the inability of selection to oppose such changes in populations of sufficiently small size, the MH hypothesis provides a potentially unifying explanation for a diversity of previously disconnected observations.

Given its broad phylogenetic perspective across species with widely different features, the MH hypothesis is admittedly difficult to test with comparative data. However, the general theory is based on fundamental principles of population genetics that transcend species boundaries and are readily evaluated with modern-day organisms. For example, the deleterious nature of introns has recently been demonstrated in at least two ways (e.g., [33,34]), and suggestions have been made as to how models on duplicate-gene evolution might be tested with information on within-species polymorphisms [35]. Nonetheless, legitimate questions about the breadth of applicability of the theory remain to be answered [36,37]. The hypothesis cannot explain the precise gene content of species, which must be molded to a large extent by the environment. Nor can it explain all aspects of "noncoding DNA," as some of this territory has positive functions. Additional complications arise from the fact that some modern-day genomes have structures that are out of equilibrium with current effective population sizes (e.g., [29]), a factor that may explain the apparently complex genome of the ancestral eukaryote and the continuing loss of such complexity in many of today's unicellular lineages [7,38,39].

Future observations on key phylogenetic lineages varying in significant ways with respect to long-term intensities of mutation, drift, and recombination will provide the observations on which the MH hypothesis will stand or fall. Improvements are already possible, now that mutation rates can be directly measured in a wide variety of genomes with high-throughput sequencing [15]. Unfortunately, the procurement of direct estimates of N_e remains dauntingly difficult [40], and until this problem is solved, it will remain difficult to obtain unbiased estimates of the key parameter $N_e u$. However, there is no justification for rejecting a theory based on its accessibility to formal hypothesis testing. It can be tempting to invoke observations on single genomes as being in support or conflict with the MH hypothesis [41–44], but due to the stochastic nature of evolutionary processes, the full domain of applicability of the model will only be known after the accumulation of many such observations. Well-reasoned applications of statistics will surely play a role, but the real advances will come from an enhanced understanding of genome biology.

Acknowledgments

I am grateful to F. Catania, J. Felsenstein, J. F. Gout, J. I. Lucas, N. Moran, and M. Nachman for comments.

References

1. Force A, Cresko W, Pickett FB, Proulx S, Amemiya, Lynch M (2005) The origin of gene subfunctions and modular gene regulation. Genetics 170: 433–446.
2. Force A, Lynch M, Pickett FB, Amores A, Yan YL, et al. (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.
3. Hong X, Scofield DG, Lynch M (2006) Intron size, abundance, and distribution within untranslated regions of genes. Mol Biol Evol 23: 2392–2404.
4. Lynch M (2002) Intron evolution as a population-genetic process. Proc Natl Acad Sci U S A 99: 6118–6123.
5. Lynch M (2006) The origins of eukaryotic gene structure. Mol Biol Evol 23: 450–468.
6. Lynch M (2006) Streamlining and simplification of microbial genome architecture. Ann Rev Microbiol 60: 327–349.
7. Lynch M (2007) The Origins of Genome Architecture. Sunderland: Sinauer Associates, Inc.. pp 340.
8. Lynch M (2007) The evolution of genetic networks by nonadaptive processes. Nat Rev Genet 8: 803–813.
9. Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A 104(Suppl): 8597–8604.
10. Lynch M, Conery JS (2003) The origins of genome complexity. Science 302: 1401–1404.
11. Lynch M, O'Hely M, Walsh B, Force A (2001) The probability of fixation of a newly arisen gene duplicate. Genetics 159: 1789–1804.
12. Lynch M, Scofield DG, Hong X (2005) The evolution of transcription-initiation sites. Mol Biol Evol 22: 1137–1146.
13. Lynch M, Koskella B, Schaack S (2006) Mutation pressure and the evolution of organelle genome architecture. Science 311: 1727–1730.
14. Lynch M (2008) The cellular, developmental, and population-genetic determinants of mutation-rate evolution. Genetics 180: 933–943.
15. Lynch M (2010) Evolution of the mutation rate. Trends Genet 26: 345–352.
16. Pigliucci M (2007) Postgenomic musings. Science 317: 1172–1173.
17. Gregory TR, Witt JD (2008) Population size and genome size in fishes: a closer look. Genome 51: 309–313.
18. Whitney KD, Garland T (2010) Did genetic drift drive increases in genome complexity? PLoS Genet 6: e1001080. doi:10.1371/journal.pgen.1001080.
19. Lynch M, Walsh JB (1998) Genetics and Analysis of Quantitative Traits. Sunderland: Sinauer Assocs, Inc.. pp 980.
20. Felsenstein J (2004) Inferring phylogenies. Sunderland: Sinauer Associates, Inc. 664 p.
21. Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125: 1–15.
22. Lynch M (1991) Methods for the analysis of comparative data in evolutionary biology. Evolution 45: 1065–1080.
23. Housworth E, Martins E, Lynch M (2003) The phylogenetic mixed model. Am Nat 163: 84–96.
24. Ricklefs RE, Starck JM (1996) Applications of phylogenetically independent contrasts: a mixed progress report. Oikos 77: 167–172.
25. Björklund M (1997) Are 'comparative methods' always necessary? Oikos 80: 607–612.
26. Rohlf FJ (2006) A comment on phylogenetic correction. Evolution 60: 1509–1515.
27. Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6: e1001115. doi:10.1371/journal.pgen.1001115.
28. Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of selection upon genomic GC-content in bacteria. PLoS Genet 6: e1001107. doi:10.1371/journal.pgen.1001107.
29. Rho M, Zhou M, Gao X, Kim S, Tang H, et al. (2009) Independent mammalian genome contractions following the KT boundary. Genome Biol Evol 2009: 2-12.
30. Martins EP, Garland, T (1991) Phylogenetic analyses of correlated evolution of continuous characters: a simulation study. Evolution 45: 534–557.
31. Whitney KD, Baack EJ, Hamrick JL, Godt MJ, et al. (2010) A role for nonadaptive processes in plant genome size evolution? Evolution 64: 2097–2109.
32. Kuo CH, Moran NA, Ochman H (2009) The consequences of genetic drift for bacterial genome complexity. Genome Res 19: 1450–1454.
33. Li W, Tucker AE, Sung W, Thomas WK, Lynch M (2009) Extensive, recent intron gains in *Daphnia* populations. Science 326: 1260–1262.
34. Lynch M (2009) Rate, molecular spectrum, and consequences of spontaneous mutations in man. Proc Natl Acad Sci U S A 107: 961–968.
35. Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11: 97–108.
36. Charlesworth B (2008) The origin of genomes – not by natural selection? Curr Biol 18: R140.
37. Charlesworth D (2008) Book review: The Origins of Genome Architecture. Genet Res 90: 217–219.
38. Koonin EV (2009) Evolution of genome architecture. Int J Biochem Cell Biol 41: 298–306.
39. Koonin EV (2009) Intron-dominated genomes of early ancestors of eukaryotes. J Hered 100: 618–623.
40. Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10: 195–205.
41. Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, et al. (2010) Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol Biol Evol 27: 1436–1448.
42. Sloan DB, MacQueen AH, Alverson AJ, Palmer JD, Taylor DR (2010) Extensive loss of RNA editing sites in rapidly evolving silene mitochondrial genomes: selection vs. retroprocessing as the driving force. Genetics 185: 1369–1380.
43. Smith DR, Lee RW (2009) Nucleotide diversity of the *Chlamydomonas reinhardtii* plastid genome: addressing the mutational-hazard hypothesis. BMC Evol Biol 9: 120.
44. Smith DR, Lee RW (2010) Low nucleotide diversity for the expanded organelle and nuclear genomes of *Volvox carteri* supports the mutational-hazard hypothesis. Mol Biol Evol 27: 2244–2256.