

What controls the length of noncoding DNA?

Josep M Comeron

Several recent studies of genome evolution indicate that the rate of DNA loss exceeds that of DNA gain, leading to an underlying mutational pressure towards collapsing the length of noncoding DNA. That such a collapse is not observed suggests opposing mechanisms favoring longer noncoding regions. The presence of transposable elements alone also does not explain observed features of noncoding DNA. At present, a multidisciplinary approach – using population genetics techniques, large-scale genomic analyses, and *in silico* evolution – is beginning to provide new and valuable insights into the forces that shape the length of noncoding DNA and, ultimately, genome size. Recombination, in a broad sense, might be the missing key parameter for understanding the observed variation in length of noncoding DNA in eukaryotes.

Addresses

Department of Ecology and Evolution, University of Chicago,
1101 East 57th Street, Chicago, Illinois 60637, USA;
e-mail: jcomeron@midway.uchicago.edu

Current Opinion in Genetics & Development 2001, **11**:652–659

0959-437X/01/\$ – see front matter

© 2001 Elsevier Science Ltd. All rights reserved.

Abbreviations

DB	deletion bias
DOA	dead-on-arrival
Indel	insertion or deletion
IS	interference selection
ODB	overall deletion bias
PDB	polymorphic deletion bias
TE	transposable element

Introduction

The total amount of DNA in a haploid genome, known as the C-value, varies enormously among eukaryotic organisms, ranging from 1.2×10^7 bp (e.g. the yeast *Saccharomyces cerevisiae*) to $>6 \times 10^{11}$ bp (e.g. *Amoeba dubia*). More intriguingly, this irregularity is also frequently observed between closely related species (i.e. within the same taxonomic group or genus) with similar levels of complexity (morphological, developmental, behavioral, etc.), number of genes, and regulatory networks. As emblematic examples, the size of genomes of flowering plants varies 1000-fold, and the DNA content can change >200- and 100-fold among vertebrates and species of salamanders, respectively [1,2]. The human genome size is not exceptional among vertebrates, with a genome larger than that of birds but smaller than those of most fishes or amphibians; mammalian genomes vary by >40% in length [3]. *Drosophila* species provide another interesting case, where differences in genome size show no obvious relationship to rate of cell division or development time, nor do they show clear phylogenetic trends [4], indicating that the forces involved in genome size change may act quite rapidly. This lack of correspondence between genome size and biological complexity has been called the C-value paradox or enigma [1,5,6].

Eukaryotic genomes consist mostly of DNA sequences that are not part of coding regions, regulatory elements, or RNA genes, and it is this apparently superfluous DNA that contributes to the variation in C-value. Most euchromatic noncoding DNA comprises introns and intergenic regions, with repetitive sequences of different types — satellite and microsatellite DNA, and different classes of transposable elements (TEs). In this review, I begin by describing several genomic features of euchromatic noncoding DNA, including intron and intergenic length, and TE and microsatellite presence, focusing particularly on their relationship with recombination rates across genomes. Next, I detail results of studies on the mutational tendencies of small indels (insertions or deletions). Finally, I discuss selective forces that may be influencing the length of noncoding regions.

Genome features

Intron size

In an innovative study, Lengyel and Penman [7] showed that the size of hnRNA, but not mature mRNA, increases with genome size in dipterans. They reported significantly longer hnRNA in *Aedes* than in *Drosophila*, consistent with *Aedes*' 5–6-fold larger genome. This observation, dated before the discovery of the intervening sequences or introns in 1977, was the first indication of a positive relationship between genome size and total intron length. A significant, although weak, positive relationship between intron and genome size has now been established for many other eukaryotes [8–11], both on a large evolutionary time-scale [11] and between *Drosophila* species [9]. In all cases, however, the differences in intron size alone cannot fully account for the differences in euchromatic genome size, indicating that the differences in genome size are not easily explained by a single class of noncoding DNA. Variation in genome size among organisms is usually associated to congruent changes across different classes of noncoding DNA (e.g. introns and intergenic regions), suggesting that they may be responding to similar, or at least overlapping, evolutionary forces. Two factors that can influence the length of introns and intergenic regions are TE and, to a lesser degree, microsatellite presence.

Transposable elements

TEs are ubiquitous in all eukaryotes and can be a leading factor influencing the length of noncoding DNA and genome size in some species. For instance, in humans, recognizable TEs represent a fraction as large as 45% of the euchromatic DNA, and are abundantly present in introns [12,13••]. In other organisms, such as *Drosophila melanogaster*, TEs have a more limited influence on the overall size of noncoding sequences, with minor effects on intron length; only 0.4% of introns in genes with known full-length mRNAs have detectable remnants of TE sequences (JM Comeron, unpublished data). The rare detection of TEs in *Drosophila*

introns is not unexpected because they may often alter the accuracy of transcript processing, mostly with deleterious consequences on fitness. In organisms with much longer introns (e.g. humans), the insertion of TEs may have an attenuated effect on gene expression and would further increase the chance of successive insertions. An equivalent argument can be made for intergenic regions based on the consequence on gene expression of long terminal repeats in the proximity of a gene.

TE invasion and expansion can cause rapid changes in genome size, generating differences between very closely related species that might share most other forces that influence genome size. An extreme example is the genome of maize, which has doubled its size as a result of retrotransposon insertions, mostly in the last three million years [14]. In *Drosophila*, TE distribution can also vary among closely related species, or even among populations of the same species [15•]. The recent invasion by the *Penelope* TE of *D. virilis* [16•] is an interesting case in point.

Microsatellites

The main mutational mechanism causing changes in the number of microsatellite repeats is polymerase template slippage that is not corrected by the mismatch repair system [17,18]. In *Drosophila*, recent studies have revealed differences both in density and length of microsatellites among species [19•–21•]. These differences concur with the idea that larger genomes also tend to have more and longer microsatellites [22]. Among distant organisms, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *D. melanogaster*, and *Homo sapiens* show a congruent increase in overall microsatellite presence with genome size when both length and density of microsatellites are considered [23–25,26•,27]. This relationship, however, is not universal, with the pufferfish *Fugu rubripes* having a higher density and longer microsatellites than humans but with a genome eight times smaller [28]. Nevertheless, the general trend suggests that the forces acting on intron and intergenic length are not independent of those influencing abundance and length of microsatellites.

Noncoding regions and recombination rates across genomes

The study of characteristics of noncoding DNA in relation to recombination rates across a given genome is informative because population genetics models of selection predict that recombination increases the efficacy of selection [29,30]. Therefore, assuming that there is no bias in the indel-repair mechanism associated with recombination (see below), the observation of different patterns of noncoding DNA between regions of high and low recombination may reveal the varying action of selection.

Introns and intergenic regions

In humans, there is evidence [13•,31,32••] that introns are shorter in isochores with high recombination (i.e. G+C-rich isochores [33,34]) than in those isochores with low

recombination rates (i.e. G+C-poor isochores). Moreover, gene density and the length of intergenic regions also decrease with recombination [13••], suggesting qualitatively similar forces shaping the length of both types of noncoding DNA across genomes.

Further, the observed relationship between the length of noncoding DNA and recombination in the human genome can be only partially attributed to TE presence. For instance, both median intron length and gene density vary ~10-fold between GC-rich and GC-poor genomic regions but the proportion of the genome comprising TEs varies much less conspicuously between these regions [13••,35]. It is possible to argue that the limited variation of TE presence is caused by the fact that TEs in regions with low recombination are older than in regions with high recombination; they therefore tend to be no longer recognizable but still contribute to size. This possibility is unlikely in humans, however, where the opposite trend is observed, with TEs younger in GC-poor isochores [13••], ruling out a simple mutational scenario with rare TE insertions and frequent small deletions [36,37] to explain variation in length of noncoding DNA in humans.

In *Drosophila*, introns also tend to be longer in regions of low recombination [32••,38••]. This relationship is mostly caused by genes with long coding sequence (JM Comeron, unpublished data) and is not explained by TE presence alone (see above). Congruently, gene density also increases with recombination in *D. melanogaster* (JM Comeron, unpublished data), paralleling the relationship between the length of noncoding regions and recombination observed in the human genome. A rare case of very long introns can be found also in *Drosophila*, where genes located in the heterochromatic Y chromosome can have mega-introns (>1 Mb) with large clusters of satellite DNA [39,40•]. This result suggests either that recombination is required to prevent a one-way growth process for satellite DNA or that selection against very long introns is inefficient in this genomic region.

Transposable elements

Although TEs are not the only contributors to differences in length of noncoding DNA across a given genome, they may play some role. There are at least two reasons to suggest that TEs should accumulate in chromosomal regions with reduced recombination. First, non-homologous recombination between different elements has deleterious consequences on fitness because it induces chromosomal rearrangements [41]. Second, as indicated above, selection against the deleterious effects of TEs either in or near genes is more efficient in regions with high recombination. In *Drosophila*, this general pattern is observed when the frequency of a given element in the population is taken into account [41,42]. In addition, different TE elements might show characteristic distributions across genomes [15•], representing different stages in the temporal dynamics of TE invasion. Illustrative of this pattern is the

Penelope element. In *D. virilis*, the recently inserted *Penelope* sequences are restricted to recombining euchromatic regions whereas in other species of the virilis group, where the element has been long present, *Penelope* is mainly detected in heterochromatic regions [43].

Interestingly, recent studies have shown that recombination rates in particular genomic regions of *Drosophila* can change rapidly. For instance, the X-telomeric region in *D. melanogaster* has strongly reduced frequency of crossing over compared to other genomic regions. The comparison with other species of the melanogaster group shows that this reduced crossover frequency was recently acquired in the *D. melanogaster* lineage, after the evolutionary split between *D. melanogaster* and *D. yakuba* [44•]. Incidentally, this finding might suggest that TEs may not have yet reached the new equilibrium of presence and segregating frequency in this genomic region in *D. melanogaster*, which might explain the absence of high TE accumulation at the tip of the X chromosome [41,42].

Caenorhabditis elegans is an intriguing case also worth noting. *C. elegans* shows the pattern opposite to that in *Drosophila* or humans, with a positive relationship between TE density and recombination rate [45,46], although some miniature inverted-repeat TEs (MITEs) show different distribution profiles [47]. This pattern might suggest that selection against TEs is not an important factor explaining TE distribution in *C. elegans* [46]. It might also suggest that regions with high recombination in *C. elegans* are, overall, under reduced selective constraints, congruent with a low density of both highly expressed genes and conserved eukaryotic genes [45]. Alternatively, it could indicate a recent TE invasion, mostly targeting regions of high recombination, hence causing unusual and unstable genomic features.

Microsatellites

Genome analyses indicate that the differences in recombination rate across *D. melanogaster*'s genome has no effect on either microsatellite density or average repeat number [25]. This result is a first indication that there is no directional bias associated with recombination-dependent indel mismatch repair (see below).

Mutational tendencies of small indels

Two different approaches have been used to study the mutational rates and tendencies of indels: first, the analysis of indels in pseudogenes and DOA elements, and second, the study of polymorphic indels in noncoding regions.

Indels in pseudogenes and DOA elements

The study of pseudogenes and DOA elements is based on the assumption that these sequences are free of selective constraints for coding information and, therefore, the observed indel patterns are a faithful representation of the indel mutation process. Early studies, mostly of mammalian pseudogenes, showed that small deletions are more

frequent than insertions [48–50]. The study of 603 small indels in 156 processed pseudogenes from humans and murids shows a ratio of deletions to insertions (i.e. deletion bias [DB]) of 2.74 [50], evidence of a significant pressure towards DNA loss. Although DB is slightly larger in humans than in murids (2.9 versus 2.6, respectively), counter to expectations based on the smaller size of the mouse and rat genomes compared to human [3], insertions are longer and deletions are shorter in the human lineage, and the overall rate of DNA loss (overall deletion bias [ODB]) is lower in the human lineage (1.69) than in mouse and rat (2.63).

A novel contribution to understanding the role of indel evolution on genome size involved studies of indels in nontransposing copies of non-LTR retrotransposable elements (i.e. DOA elements). These studies first focused on the *Helena* element in *Drosophila* [51,52] and on the *Lau1* element in Hawaiian crickets of the genus *Laupala*, the latter having a genome 11 times larger than *Drosophila* [53•]. The rate of DNA loss in DOA elements differs between these insects, with a much higher rate in *Drosophila* (DB = 8.7 [52]) compared to *Laupala* (DB = 2.7 [53•]), in agreement with *Drosophila*'s more compact genome. Also in insects, the study of indels in 58 paralogous pseudogenes in *Podisma pedestris* [54••], which has a genome >100 times as large as that of *Drosophila*, shows a DB of 2.7 when hot spots for indel mutations are excluded from the analysis. ODB is also strongly affected by the size of indels in these species, with ODB of 74.7, 3.8 and 3.6, for *Drosophila*, *Laupala* and *Podisma*, respectively.

The study of indels in a large pseudogene family in *C. elegans* [55] also reveals an excess of deletions compared to insertions (DB = 3.8), with the detection of very large insertions and deletions. Intriguingly, the size of indels would generate an unexpected overall gain of DNA (ODB = 0.49 caused by two very long insertions) but no definitive conclusions can be made because the length of some long deletions may be gross underestimates [55].

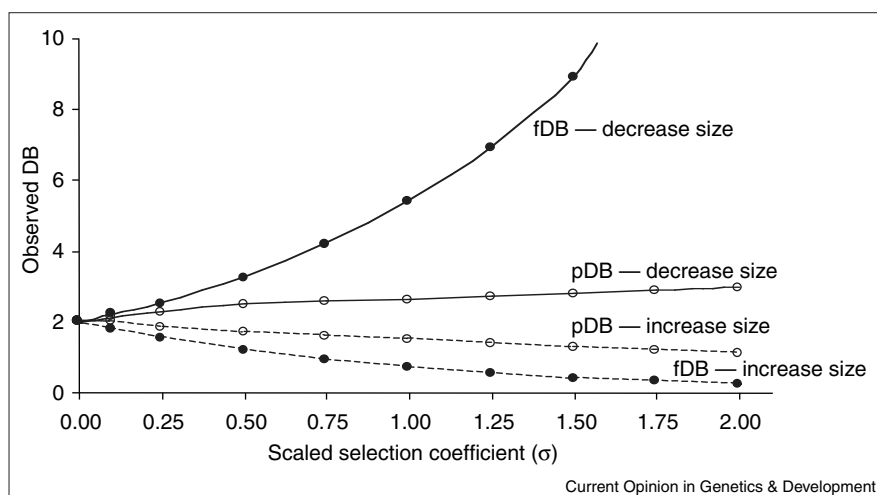
The conclusion of these studies based on indels in pseudogenes and between DOA elements is that, overall, the rate of DNA loss is greater than that of DNA gain, at least based on small indels. Moreover, this rate of overall DNA loss is greatest in *Drosophila* compared to *Laupala*, *Podisma*, and mammals, congruent with the smallest size of *Drosophila* genome compared to the other genomes. *C. elegans* depicts a contrasting observation, with a very compact genome that is smaller than that of *D. melanogaster* but with a lower rate of DNA loss.

Polymorphic indels

Weak selection has considerably less influence on the presence of mutations as polymorphisms in a population than it does on their probability of fixation [56] (Figure 1). Therefore, studies based on polymorphism presence may give us a pattern very close to the mutational tendencies,

Figure 1

Observed deletion bias (DB) based on polymorphic (pDB) and fixed (fDB) indels under weak selection and the infinitely many sites model with free recombination and semidominance. A real mutational DB of 2.0 is assumed. The scaled selection coefficient for diploid organisms (σ) is $4 N_e s$, where N_e is the effective population size and s is the selection coefficient. Results for pDB are based on the study of 10 alleles. Continuous lines indicate a selective scenario favoring shorter sequences (favorable deletions and deleterious insertions) and dashed lines indicate selection favoring longer sequences (deleterious deletions and favorable insertions). In both selective scenarios, pDB shows a value close to the real DB, even when indels are under weak selection.



even if mutations are under weak selection. A study of 256 polymorphic indel events in *D. melanogaster* [32••] shows a significant excess of deletions compared to insertions in noncoding (introns and intergenic) regions, with a polymorphic deletion bias (pDB) of 1.4 and ODB of 1.8. Several lines of reasoning lead to the conclusion that the DB observed using this method is, if anything, only a small underestimation of the true rate. First, beyond essential splicing signals and regulatory elements, many introns and flanking intergenic regions are fast-evolving [57,58], indicating paucity of selective constraints against new mutations. Second, small indels are commonly detected as polymorphisms, suggesting that many of these indels are not subject to strong selection. Finally, strong selection acting on indels does not appear to be common because similar pDBs are seen for introns and intergenic regions. In addition, long deletions, possible targets of strong selection, are infrequent (only 8% and 27% of deletions are >100 bp and 30 bp, respectively, based on DOA elements [52]).

Moreover, this study in *D. melanogaster* [32••] found equivalent pDBs in genomic regions with high and low recombination, suggesting that the observation of different lengths of noncoding regions in genomic regions with different recombination rates is not caused by a DB that changes with recombination rate. Conversely, the discrepancy between genomic patterns and indel polymorphisms unveils a possible role of selection in shaping the length of noncoding regions, at least in *D. melanogaster*. Future studies differentiating euchromatic versus heterochromatic, high versus low recombination regions, and fixed versus polymorphic indels, in DOA elements, pseudogenes and other noncoding regions will be useful for completing the picture of the rate of spontaneous DNA loss across the *Drosophila* genome.

Directionality in indel repair mechanism

Subtle differences in both the efficacy and the directionality of the indel mismatch repair system can generate significant

differences in DB and in the length of noncoding DNA between species. Moreover, directionality in the recombination-dependent indel mismatch repair can generate patterns of noncoding length correlated with recombination rates across genomes. In yeast, however, there is no clear directionality in indel repair. The repair of mispaired loops in heteroduplex DNA shows diverse biases, efficiencies, and repair pathways, for short and long indels, for different genomic regions, and for nicked and continuous DNA (see references in [32••]). On the other hand, a generally biased repair towards deletions associated with recombination is not observed in *D. melanogaster*'s genome on the basis of available data of indel polymorphisms and their frequency in natural populations [32••], and microsatellite distribution [25]. Nonetheless, more definitive studies are required to appraise its actual contribution.

Natural selection on the length of noncoding regions

Irrespective of the precise rate towards DNA loss, this tendency will lead to the collapse of noncoding DNA, and TE insertion alone cannot explain the differences in length of noncoding regions between species or across genomes. Therefore, mutational explanations based on 'junk' [59] or 'selfish' [60] DNA don't provide a sufficient explanation. Several selective hypotheses have been proposed to explain observed relationships between genome size and cellular and developmental traits — some observed only in particular lineages — although one cannot immediately discern between causal, indirect, and coincidental relationships. Typical of this class of selective hypotheses, the nucleotypic [61] and nucleoskeletal [62] theories propose structural functions to noncoding DNA, setting the minimum size attainable by a nucleus, or an indirect (or coevolutionary) response to genetically-controlled nuclear size.

Other selective explanations for the presence of noncoding DNA relate to its role in regulating gene expression [63].

Figure 2

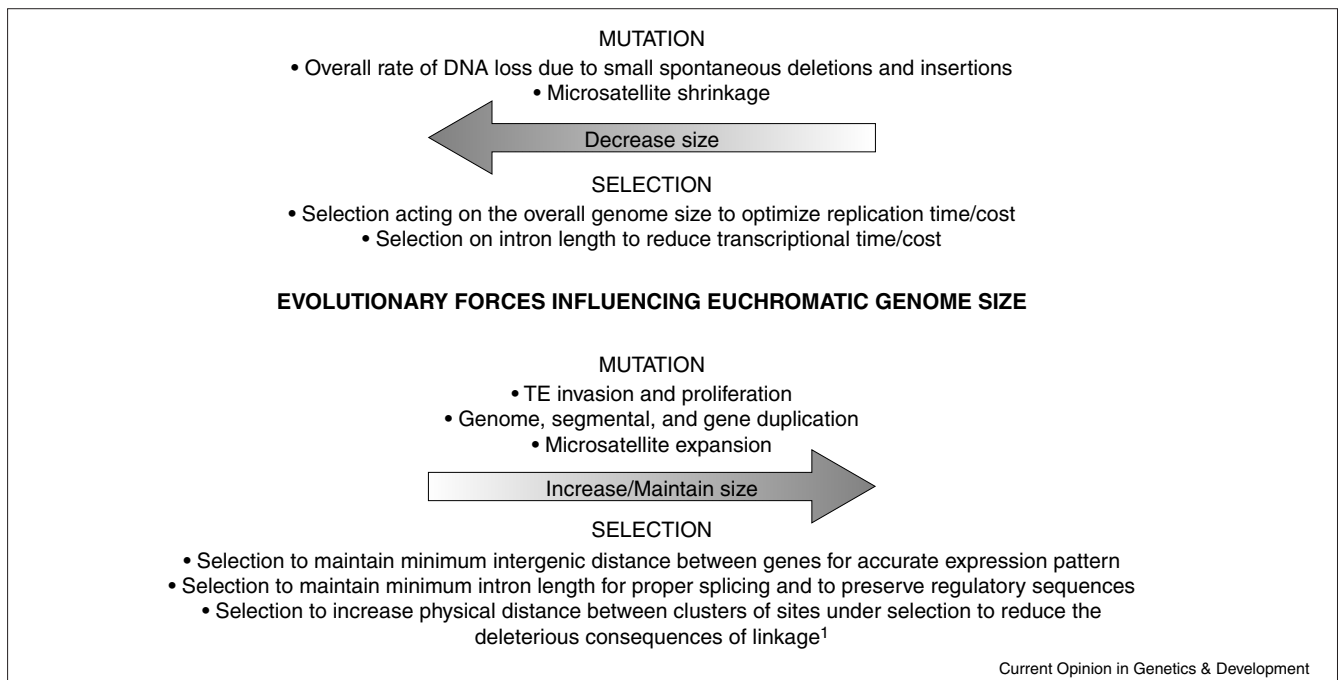


Diagram of mutational and selective evolutionary forces that influence the noncoding euchromatic genome size. The equilibrium size is expected to vary among species and be highly dynamic according to species specific mutational and selective tendencies. ¹Only in recombining genomes.

Greater developmental and/or morphological complexity may require more elaborate regulatory systems, which in turn require larger amounts of noncoding DNA. Consistent with this idea, the comparison between the number of selectively constrained nucleotides per intergenic interval is distinctly higher in mammals than in *C. elegans* [64]. The increment in noncoding constrained sites, however, can explain only a small fraction of the difference in the length of noncoding regions between these organisms. This observation is another indication that the length of noncoding DNA is not solely determined by functional requirements.

Noncoding DNA as an enhancer of recombination

Theoretical and simulation studies [65–67] have shown that mutations that raise the recombination rate between strongly selected loci have an increased chance of fixation because they reduce the detrimental consequences of linkage between selected mutations. Results of *in silico* evolution (i.e. forward simulations) [68,69,70,71] have shown that many weakly selected mutations interfere with each others' fixation probability, also reducing the efficacy of selection relative to expectations for independently evolving mutations (also known as the Hill–Robertson effect [72,73] or interference selection [IS]). IS increases with the number of sites under weak selection and is reduced with recombination.

Weakly selected mutations are numerous in natural populations and are physically clustered across genomes — mostly in

genes, exons, and in regulatory regions — creating genomic regions with a high density of sites under selection, with limited recombination between them. This physical distribution of mutations causes IS to have a measurable effect in most eukaryotic genomes, and this raises the possibility that indels between clusters of selected sites (genes or exons) might be subject to natural selection as modifiers of recombination [32••]. Mutations that increase the distance (hence recombination) between mutations under selection will be favored by selection under this scenario because, ultimately, they increase the effectiveness of selection acting on these flanking sites. Longer introns and lower gene density will be favored in regions of low recombination. This tendency for selection to favor insertions and to oppose deletions will counter-balance the mutational bias towards DNA loss.

Genomic patterns in *D. melanogaster* and humans cannot be explained by either TE presence or small indels alone (see above). The fact that both intron length and intergenic sizes negatively correlate with recombination is consistent with the action of IS. *In silico* evolution results support detectable effects of small indels on the efficacy of selection in adjacent regions (JM Comeron, M Kreitman, unpublished data). Therefore, IS in recombining genomes might be influencing genomic features [32••] such as gene structure, intron length and gene density, and it might be key to the C-value paradox. Under this scenario, the length of noncoding regions represent a highly dynamic mutation–selection–drift equilibrium responding to changes in effective population size and/or recombination rates (per physical unit) as well as

to variation of species-specific developmental requirements and mutational tendencies (Figure 2).

Conclusions

Indels are beginning to be studied using both population genetics and genomic techniques. As is the case for coding sequence evolution, these techniques and corresponding statistical tests should allow discrimination of the relative importance of selection, mutational tendencies, and random genetic drift in influencing the length of noncoding regions and genome size. This genomics-meets-population genetics approach can be implemented with the study of *in silico* evolution to capture the expected outcome of complex evolutionary interactions, including weak selection and drift. TE invasion and expansion, and a general mutational tendency towards DNA loss are two mutational forces that inescapably influence the length of noncoding regions in most eukaryotes. These mutational tendencies alone, however, cannot explain genomic features or indel polymorphism data, indicating that natural selection might be acting on indels to influence the length of noncoding regions. A recent proposal [32**] suggests that noncoding regions (introns and intergenic regions) can be viewed as modifiers of recombination, and thus have a selective role. Based on this suggestion, a better understanding of the C-value paradox might be attained when recombination is entered into the consideration, both as the population recombination parameter per site (ironically, often also noted by *C* [74]), and as the physical distance between sites under selection.

Acknowledgements

I thank Marty Kreitman, Ana Llopart, and Chris Toomajian for helpful comments on the manuscript and the members of the Kreitman lab for many fruitful discussions.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
 - of outstanding interest
1. Cavalier-Smith T: *The Evolution of Genome Size*. New York: John Wiley; 1985.
 2. Li WH: *Molecular Evolution*. Sunderland, MA: Sinauer Associates; 1997
 3. Vinogradov AE: **Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship**. *Cytometry* 1998, **31**:100-109.
 4. Powell JR: *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. New York: Oxford University Press; 1997.
 5. Mirsky AE, Ris H: **The desoxyribonucleic acid content of animal cells and its evolutionary significance**. *J Gen Physiol* 1951, **34**:451-462.
 6. Thomas CA: **The genetic organization of chromosomes**. *Annu Rev Genet* 1971, **5**:237-256.
 7. Lengyel J, Penman S: **hnRNA size and processing as related to different DNA content in two dipterans: *Drosophila* and *Aedes***. *Cell* 1975, **5**:281-290.
 8. Hughes AL, Hughes MK: **Small genomes for better flyers**. *Nature* 1995, **377**:391.
 9. Moriyama EN, Petrov DA, Hartl DL: **Genome size and intron size in *Drosophila***. *Mol Biol Evol* 1998, **15**:770-773.
 10. Deutsch M, Long M: **Intron-exon structures of eukaryotic model organisms**. *Nucleic Acids Res* 1999, **27**:3219-3228.
 11. Vinogradov AE: **Intron-Genome size relationship on a large evolutionary scale**. *J Mol Evol* 1999, **49**:376-384.
 12. Smit AF: **Interspersed repeats and other mementos of transposable elements in mammalian genomes**. *Curr Opin Genet Dev* 1999, **9**:657-663.
 13. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921. This is another example of a recent invasion of a *Drosophila* species by a TE, in this case *Penelope*. Populations showing several copies of *Penelope* per genome were free of this element only thirty years ago!
 14. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: **The paleontology of intergene retrotransposons of maize**. *Nat Genet* 1998, **20**:43-45.
 15. Biémont C, Cizeron G: **Distribution of transposable elements in *Drosophila* species**. *Genetica* 1999, **105**:43-62.
 16. Evgen'ev M, Zelentsova H, Mnjoian L, Poluectova H, Kidwell MG: **Invasion of *Drosophila virilis* by the *Penelope* transposable element**. *Chromosoma* 2000, **109**:350-357.
 17. Schlotterer C, Tautz D: **Slippage synthesis of simple sequence DNA**. *Nucleic Acids Res* 1992, **20**:211-215.
 18. Strand M, Prolla TA, Liskay RM, Petes TD: **Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair**. *Nature* 1993, **365**:274-276.
 19. Pascual M, Schug MD, Aquadro CF: **High density of long dinucleotide microsatellites in *Drosophila subobscura***. *Mol Biol Evol* 2000, **17**:1259-1267. See annotation [21*].
 20. Schlotterer C, Harr B: ***Drosophila virilis* has long and highly polymorphic microsatellites**. *Mol Biol Evol* 2000, **17**:1641-1646. See annotation [21*].
 21. Warner RD, Noor MA: **High frequency of microsatellites in *Drosophila pseudoobscura***. *Genes Genet Syst* 2000, **75**:115-118. This study and [19*,20*] show that *Drosophila* species with genomes larger than *D. melanogaster* and *D. simulans* tend to have more and longer microsatellites.
 22. Hancock JM: **Simple sequences and the expanding genome**. *Bioessays* 1996, **18**:421-425.
 23. Kruglyak S, Durrett RT, Schug MD, Aquadro CF: **Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations**. *Proc Natl Acad Sci USA* 1998, **95**:10774-10778.
 24. Schug MD, Wetterstrand KA, Gaudette MS, Lim RH, Hutter CM, Aquadro CF: **The distribution and frequency of microsatellite loci in *Drosophila melanogaster***. *Mol Ecol* 1998, **7**:57-70.
 25. Bachtrog D, Weiss S, Zangerl B, Brem G, Schlotterer C: **Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome**. *Mol Biol Evol* 1999, **16**:602-610.
 26. Harr B, Schlotterer C: **Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation**. *Genetics* 2000, **155**:1213-1220. This study of mutation-accumulation lines of *D. melanogaster* indicates that long microsatellites have an overall tendency to lose repeat units. Moreover, intra- and interspecific analyses reveal that these long microsatellites in *D. melanogaster* represent the youngest class of alleles and have short persistence times. The authors propose a species-specific critical length at which the mutational behavior of a microsatellite changes to account for the observed differences in microsatellite length between species.
 27. Katti MV, Ranjekar PK, Gupta VS: **Differential distribution of simple sequence repeats in eukaryotic genome sequences**. *Mol Biol Evol* 2001, **18**:1161-1167.
 28. Elgar G, Clark MS, Meek S, Smith S, Warner S, Edwards YJ, Bouchireb N, Cottage A, Yeo GS, Umrana Y *et al.*: **Generation and analysis of 25 Mb of genomic DNA from the pufferfish *Fugu rubripes* by sequence scanning**. *Genome Res* 1999, **9**:960-971.
 29. Maynard Smith J, Haigh J: **The hitch-hiking effect of a favorable gene**. *Genet Res* 1974, **23**:23-35.

30. Charlesworth B, Morgan MT, Charlesworth D: **The effect of deleterious mutations on neutral molecular variation.** *Genetics* 1993, **134**:1289-1303.
31. Duret L, Mouchiroud D, Gautier C: **Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores.** *J Mol Evol* 1995, **40**:308-317.
32. Comeron JM, Kreitman M: **The correlation between intron length and recombination in *Drosophila*. Dynamic equilibrium between mutational and selective forces.** *Genetics* 2000, **156**:1175-1190.
A study of patterns of indel mutations detected at the polymorphic level in *D. melanogaster* showing a significant but weak excess of deletions compared to insertions. Equivalent indel patterns are observed in regions of high and low recombination, and in intergenic regions and introns. The authors propose that introns and intergenic sequences can be viewed as modifiers of recombination, hence with a selective role. See also annotation [38**].
33. Eyre-Walker A: **Recombination and mammalian genome evolution.** *Proc R Soc Lond B Biol Sci* 1993, **252**:237-243.
34. Fullerton SM, Bernardo Carvalho A, Clark AG: **Local rates of recombination are positively correlated with gc content in the human genome.** *Mol Biol Evol* 2001, **18**:1139-1142.
35. Gu Z, Wang H, Nekrutenko A, Li WH: **Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence.** *Gene* 2000, **259**:81-88.
36. Bennetzen JL, Kellogg EA: **Do plants have a one-way ticket to genomic obesity?** *Plant Cell* 1997, **9**:1509-1514.
37. Duret L: **Why do genes have introns? Recombination might add a new piece to the puzzle.** *Trends Genet* 2001, **17**:172-175.
38. Carvalho AB, Clark AG: **Intron size and natural selection.** *Nature* 1999, **401**:344
This article and [32**] show that genes located in regions of low recombination in *D. melanogaster* tend to have longer introns, and offer selective hypotheses to account for the observation.
39. Reugels AM, Kurek R, Lammermann U, Bunemann H: **Mega-introns in the dynein gene DhDhc7(Y) on the heterochromatic Y chromosome give rise to the giant threads loops in primary spermatocytes of *Drosophila hydei*.** *Genetics* 2000, **154**:759-769.
40. Carvalho AB, Lazzaro BP, Clark AG: **Y chromosomal fertility factors kl-2 and kl-3 of *Drosophila melanogaster* encode dynein heavy chain polypeptides.** *Proc Natl Acad Sci USA* 2000, **97**:13239-13244.
Using an iterative BLAST search against unmapped *D. melanogaster* scaffolds, the authors identify three new Y chromosome genes, two of them fertility factors. The authors illustrate the bias that library construction and cloning methods can introduce in regions with repetitive DNA in heterochromatin. Genes with introns long enough to contain heterochromatin will be underrepresented in whole-genome shotgun projects.
41. Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B: **On the role of unequal exchange in the containment of transposable element copy number.** *Genet Res* 1988, **52**:223-235.
42. Charlesworth B, Langley CH, Sniegowski PD: **Transposable element distributions in *Drosophila*.** *Genetics* 1997, **147**:1993-1995.
43. Zelentsova H, Poluectova H, Mnjoian L, Lyozin G, Veleikodvorskaja V, Zhivotovsky L, Kidwell MG, Evgen'ev MB: **Distribution and evolution of mobile elements in the virilis species group of *Drosophila*.** *Chromosoma* 1999, **108**:443-456.
44. Takano-Shimizu T: **Local changes in gc/at substitution biases and in crossover frequencies on *Drosophila* chromosomes.** *Mol Biol Evol* 2001, **18**:606-619.
This paper shows that the frequency of crossover can vary substantially among closely related species of *Drosophila*. The author describes a 20-fold reduction of crossover frequency in the telomeric region of the X chromosome, which occurred in the ancestral population of *D. melanogaster* after the *D. melanogaster/D. yakuba* divergence.
45. Wilson RK: **How the worm was won. The *C. elegans* genome sequencing project.** *Trends Genet* 1999, **15**:51-58.
46. Duret L, Marais G, Biemont C: **Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*.** *Genetics* 2000, **156**:1661-1669.
47. Surzycki SA, Belknap WR: **Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes.** *Proc Natl Acad Sci USA* 2000, **97**:245-249.
48. Graur D, Shuali Y, Li WH: **Deletions in processed pseudogenes accumulate faster in rodents than in humans.** *J Mol Evol* 1989, **28**:279-285.
49. Gu X, Li WH: **The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment.** *J Mol Evol* 1995, **40**:464-473.
50. Ophir R, Graur D: **Patterns and rates of indel evolution in processed pseudogenes from humans and murids.** *Gene* 1997, **205**:191-202.
51. Petrov DA, Lozovskaya ER, Hartl DL: **High intrinsic rate of DNA loss in *Drosophila*.** *Nature* 1996, **384**:346-349.
52. Petrov DA, Hartl DL: **High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups.** *Mol Biol Evol* 1998, **15**:293-302.
53. Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL: **Evidence for DNA loss as a determinant of genome size.** *Science* 2000, **287**:1060-1062.
This article shows that the excess of deletions compared to insertions is not only observed in mammals and *Drosophila* but also in Hawaiian crickets. The overall rate of DNA loss based on indels in DOA elements is lower in *Laupala* than in *Drosophila*, in agreement with *Laupala*'s larger genome size.
54. Bensasson D, Petrov DA, Zhang DX, Hartl DL, Hewitt GM: **Genomic gigantism: DNA loss is slow in mountain grasshoppers.** *Mol Biol Evol* 2001, **18**:246-253.
The study of indels in nuclear pseudogenes derived from mitochondrial DNA in a grasshopper with a very large genome unveils a relatively slow rate of DNA loss compared to insects with smaller genomes. This analysis reveals the presence of indel hot spots within the pseudogene sequence.
55. Robertson HM: **The large *srh* family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses.** *Genome Res* 2000, **10**:192-203
56. Crow JF, Kimura M: *An Introduction to Population Genetics Theory*. Edina, MN: Alpha Editions; 1970.
57. Li WH, Graur D: *Fundamentals of Molecular Evolution*. Massachusetts: Sinauer, Sunderland; 1991.
58. Hughes AL, Yeager M: **Comparative evolutionary rates of introns and exons in murine rodents.** *J Mol Evol* 1997, **45**:125-130.
59. Ohno S: **So much 'junk' DNA in our genome.** In *Evolution of Genetic Systems*. Edited by Smith HH. New York: Gordon and Breach; 1972:366-370.
60. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution.** *Nature* 1980, **284**:601-603.
61. Bennett MD: **The duration of meiosis.** *Proc R Soc Lond B Biol Sci* 1971, **178**:259-275.
62. Cavalier-Smith T: **Skeletal DNA and the evolution of genome size.** *Annu Rev Biophys Bioeng* 1982, **11**:273-302.
63. Zuckerkandl E: **Gene control in eukaryotes and the c-value paradox 'excess' DNA as an impediment to transcription of coding sequences.** *J Mol Evol* 1976, **9**:73-104.
64. Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS: **Selective constraint in intergenic regions of human and mouse genomes.** *Trends Genet* 2001, **17**:373-376.
65. Barton NH: **Linkage and the limits to natural selection.** *Genetics* 1995, **140**:821-841.
66. Otto SP, Barton NH: **The evolution of recombination: removing the limits to natural selection.** *Genetics* 1997, **147**:879-906.
67. Hey J: **Selfish genes, pleiotropy and the origin of recombination.** *Genetics* 1998, **149**:2089-2097.
68. Li WH: **Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons.** *J Mol Evol* 1987, **24**:337-345.
69. Comeron JM, Kreitman M, Aguade M: **Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*.** *Genetics* 1999, **151**:239-249.
This article and [70*] show that IS, or the Hill-Robertson effect, caused by the interaction between linkage and selection, is detectable with many weakly selected mutations, not only for total linkage but also with recombination rates observed in most eukaryotes. Therefore, IS is expected to have manifest consequences in most eukaryotic genes and genomes.

70. McVean GA, Charlesworth B: **The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation.** *Genetics* 2000, **155**:929-944.

Interference between many weakly selected mutations reduces codon bias and levels of polymorphism, and alters the frequency distribution of segregating mutations, particularly in regions of low recombination. See also annotation [69*].

71. Tachida H: **Molecular evolution in a multisite nearly neutral mutation model.** *J Mol Evol* 2000, **50**:69-81.

72. Hill WG, Robertson A: **The effect of linkage on limits to artificial selection.** *Genet Res* 1966, **8**:269-294.

73. Felsenstein J: **The evolutionary advantage of recombination.** *Genetics* 1974, **78**:737-756.

74. Hudson RR: **Estimating the recombination parameter of a finite population model without selection.** *Genet Res* 1987, **50**:245-250.