



HAL
open science

A long-term evolutionary pressure on the amount of noncoding DNA

Carole Knibbe, Antoine Coulon, Olivier Mazet, Jean-Michel Fayard,
Guillaume Beslon

► **To cite this version:**

Carole Knibbe, Antoine Coulon, Olivier Mazet, Jean-Michel Fayard, Guillaume Beslon. A long-term evolutionary pressure on the amount of noncoding DNA. *Molecular Biology and Evolution*, 2007, 24 (10), pp.2344-2353. 10.1093/molbev/msm165 . hal-00391447

HAL Id: hal-00391447

<https://hal.science/hal-00391447v1>

Submitted on 9 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Long-Term Evolutionary Pressure on the Amount of Noncoding DNA

Carole Knibbe,^{*†} Antoine Coulon,[†] Olivier Mazet,[‡] Jean-Michel Fayard,[§] and Guillaume Beslon[†]

^{*}Inserm, U571, Paris, France; [†]Laboratoire d'InfoRmatique en Images et Systèmes d'Information, UMR CNRS 5205, INSA-Lyon/ Université Claude Bernard Lyon 1, Villeurbanne, France; [‡]Laboratoire de Statistiques et Probabilités, INSA-Toulouse, Toulouse, France; and [§]Laboratoire de Biologie Fonctionnelle, Insectes et Interactions, UMR INRA/INSA 203 BF2I, INSA-Lyon, Villeurbanne, France

A significant part of eukaryotic noncoding DNA is viewed as the passive result of mutational processes, such as the proliferation of mobile elements. However, sequences lacking an immediate utility can nonetheless play a major role in the long-term evolvability of a lineage, for instance by promoting genomic rearrangements. They could thus be subject to an indirect selection. Yet, such a long-term effect is difficult to isolate either *in vivo* or *in vitro*. Here, by performing *in silico* experimental evolution, we demonstrate that, under low mutation rates, the indirect selection of variability promotes the accumulation of noncoding sequences: Even in the absence of self-replicating elements and mutational bias, noncoding sequences constituted an important fraction of the evolved genome because the indirectly selected genomes were those that were variable enough to discover beneficial mutations. On the other hand, high mutation rates lead to compact genomes, much like the viral ones, although no selective cost of genome size was applied: The indirectly selected genomes were those that were small enough for the genetic information to be reliably transmitted. Thus, the spontaneous evolution of the amount of noncoding DNA strongly depends on the mutation rate. Our results suggest the existence of an additional pressure on the amount of noncoding DNA, namely the indirect selection of an appropriate trade-off between the fidelity of the transmission of the genetic information and the exploration of the mutational neighborhood. Interestingly, this trade-off resulted robustly in the accumulation of noncoding DNA so that the best individual leaves one offspring without mutation (or only neutral ones) per generation.

Introduction

Eukaryotic genomes contain many sequences that are not translated into proteins. Although some of these sequences bear the hallmark of natural selection and are thus presumed to be functional (Duret et al. 1993; Frazer et al. 2001; Margulies et al. 2003; Bejerano et al. 2004; Andolfatto 2005; Dermitzakis et al. 2005; Keightley et al. 2005), many others seem to have no direct effect on the phenotype. Such sequences can be passively produced by mutational processes biased toward genome growth and driven to fixation by genetic drift (Lynch and Conery 2003). According to this view, a substantial amount of nonfunctional DNA can be maintained in a genome, depending on the balance between insertions and deletions (Petrov et al. 2000; Mira et al. 2001; Denver et al. 2004), the rate of proliferation of transposable elements (Kidwell 2002), the rate of retroposition of mRNAs (Maestre et al. 1995), and the population size (Lynch and Conery 2003).

Sequences acquired in such a nonadaptive way can then provide novel substrates for evolutionary innovations (Brosius and Gould 1992; Smit 1999; Lynch and Conery 2003). For instance, mRNA-derived retroposons can give rise to active genes (Brosius 2003). Furthermore, even when they remain nonfunctional, sequences present in several copies promote genomic rearrangements that can affect the phenotype (Hughes 1999; Kidwell 2002; Rocha 2003; Coghlan et al. 2005). Thus, sequences that are nonfunctional in a particular organism may nonetheless play a major role in the appearance of nonneutral mutations, leading to new phenotypes in the offspring of this organism.

Now the level of nonneutral genetic variation is a key element for the long-term evolutionary success of a lineage.

On the one hand, variability is a prerequisite for evolvability, the ability to innovate (Wagner and Altenberg 1996; Kirschner and Gerhart 1998; Radman et al. 1999; Burch and Chao 2000; Wagner 2005). On the other hand, the long-term evolutionary success also requires that a sufficient proportion of the offspring keep the ancestral phenotype by bearing no mutation or only neutral ones (Van Nimwegen et al. 1999; Wilke 2001a, 2001b; Wilke et al. 2001). Indeed, if the ancestral fitness cannot be retained from one generation to the next because deleterious mutations are too frequent, the lineage will face a heavy mutational burden that can lead to extinction. Taken together, these considerations imply that competing organisms need to achieve not only a high fitness but also an appropriate level of nonneutral genetic variation, reflecting a trade-off between the exploration of new phenotypes and the reliable transmission of the current one.

As nonfunctional sequences are not under immediate selection, their number can easily vary, which could be a way to reach the appropriate level of nonneutral variation. If this hypothesis is correct, the amount of nonfunctional sequences may not just be the passive result of mutational processes. The long-term selection of an appropriate variability may exert a selective pressure on the amount of nonfunctional DNA. This selective pressure would be indirect because varying the amount of nonfunctional DNA would not change the immediate fitness of the organism but would rather modify the chance that its offspring retain it.

This fairly simple hypothesis is however hard to test for many reasons. The first is the indirect, long-term nature of this selective pressure. Many generations would be necessary to reveal its effect. A second and perhaps more serious obstacle is the difficulty to isolate this effect from the other evolutionary pressures acting on the amount of nonfunctional DNA, including mutational biases and direct selective constraints on genome size. Finally, the long-term selection of an appropriate frequency of nonneutral mutations can also act at other levels than genome compactness. It can lead, for example, to more or less robust topologies for regulatory networks and metabolic pathways. Thus,

Key words: adaptive evolution, noncoding DNA, mutation rate, rearrangements, mutational variability, indirect selection.

E-mail: guillaume.beslon@insa-lyon.fr.

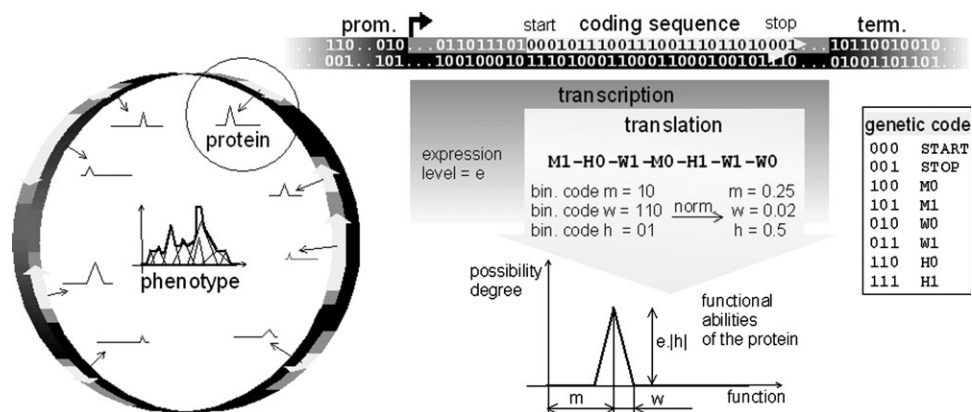


FIG. 1.—Intergenic sequences, coding sequences, and proteins are the central concepts of the model. Coding sequences are detected using signal sequences (promoter, terminator, start, and stop) and translated using an artificial genetic code. This translation step characterizes the functional abilities of the protein (by determining the subset of functions it can activate or inhibit, among an abstract set of possible functions). The phenotype, describing the global functional abilities of the organism, results from the combined actions of all the proteins. Possibility distributions are used to describe the functional abilities of the proteins and of the organism as a whole.

testing the hypothesis of an indirect selective pressure on nonfunctional DNA requires a specific approach, allowing us to isolate its effects. In silico experimental evolution of simple “organisms” is particularly useful in this context (Adami 2006). Direct selective pressures are controlled and mutational biases can be turned-off. Moreover, the exact knowledge of lineages, ancestral sequences, and fixed mutations allows for a detailed analysis of the evolutionary mechanisms.

However, in previous in silico experiments designed to study long-term evolutionary forces, the effects on the genomic structure could not be predicted because only one gene was modeled (Eigen 1971) or because the genome representation did not explicitly include the notions of gene, gene product, and intergenic sequences (Eigen 1971; Wilke 2001b; Wilke et al. 2001). In other models involving a more realistic genome architecture (Wu and Lindsay 1996; Burke et al. 1998), the complexity of the phenotype was not allowed to evolve with the complexity of the genotype. As a consequence, unrealistic heuristics were used in the transition from genotype to phenotype, which introduced artificial effects on the evolution of genome size.

Here, we study the evolution of artificial organisms where the genomic structure is biologically interpretable and where the complexity of the phenotype is allowed to evolve. This allows us to investigate the spontaneous evolution of genome size, that is, without either direct selection on genome size, or mutational biases, or self-replication of selfish elements. We show that in these conditions, the amount of noncoding sequences maintained in the genome, far from being random, is determined by the long-term selection of an appropriate level of nonneutral variation. This indirect selective pressure is at the origin of a strong relationship between the mutation rate and the amount of nonfunctional DNA contained in the artificial genomes.

Materials and Methods

These in silico experiments were performed on the “aevol” platform (Knibbe et al. 2007), version 4.5. The

source code, as well as the configuration files used here, is available on request.

General Principles

The simulated organisms have circular, double-strand binary genomes containing both coding and noncoding sequences (fig. 1). Each coding sequence encodes a “protein,” able to either activate or inhibit a number of functions. The phenotype is defined as the set of functional abilities of the organism, resulting from the combination of all its proteins. Adaptation is then measured by comparing the functions the organism can achieve to the functions to be performed and to be avoided in the environment. During replication, genomes can undergo not only point mutations, small insertions and deletions but also genomic rearrangements, consisting of duplications, deletions, translocations, and inversions.

Detection of the Coding Sequences

Promoter and terminator signals define the boundaries of the transcribed regions. Within them, start and stop signals delimit the coding sequences. Promoters are sequences whose Hamming distance with a predefined 28-bp consensus sequence is $d \leq d_{\max}$, with $d_{\max} = 4$ in this study. Terminator signals are sequences able to form a stem-loop secondary structure: $abcd**\bar{d}\bar{c}\bar{b}\bar{a}$. The expression level of a transcribed region is defined as $e = 1 - \frac{d}{d_{\max} + 1}$. Note that this modulation of the expression level models only (in a simplified way) the basal interaction of the RNA polymerase with the promoter without additional regulation. The purpose here is not to accurately model the regulation of gene expression but rather to provide duplicated genes a way to reduce temporarily their phenotypic contribution while diverging toward other functions. Inside transcribed regions, the start signal for the translation is made up of a Shine-Dalgarno-like sequence followed by the start codon (011011***000), whereas the stop signal is simply the stop codon (001, see the artificial genetic code in fig. 1). Overlapping coding sequences are allowed.

Translation and Phenotype Computation

A global set of feasible functions is defined as the real interval $\Omega=[0, 1]$. The functional abilities of each gene product are represented by a fuzzy subset $[m - w, m + w] \subset \Omega$. The possibility distributions of these subsets are piecewise linear with “triangular” shapes, with a maximal possibility degree $H=e|h|$ for the function m (fig. 1). The 3 real parameters m , w , and h are encoded by the coding sequence. Each coding sequence is read codon by codon using the genetic code shown in figure 1. This genetic code is not degenerated in order to prevent robustness at this level interfering with the effect of the noncoding sequences. The run of codons m_0 and m_1 (respectively w_0 and w_1 , h_0 and h_1) forms a gray encoding of m (respectively w , h). The sign of h determines whether the gene product activates or inhibits the functions $[m - w, m + w]$. The functional abilities of the organism as a whole is the fuzzy set of functions that are activated and not inhibited by its proteins: $P=(\cup_i A_i) \cap (\overline{\cup_j I_j})$, where A_i is the subset of the i th activator protein and I_j the subset of the j th inhibitor protein. Lukasiewicz fuzzy operators are used to compute the possibility distribution $P(x)$ of this set, which represents the phenotype of the organism.

Adaptation Measure

The abilities required to survive in the environment are also modeled by a fuzzy set E , whose possibility distribution $E(x)$ can be seen on figure 2. Adaptation is then measured by the gap $g=\int_{\Omega}|E(x) - P(x)|dx$ between the possibility distributions $E(x)$ and $P(x)$. Note that although this adaptation measure penalizes both the under- and the overrealized functions, it does not prevent increases in gene number. There is indeed a constant need for new activator and inhibitory genes to refine the phenotypic distribution $P(x)$.

Mutations

Every time a genome is replicated, it can undergo point mutations, small indels (1–6 bp), inversions, translocations, large deletions, and duplications. The mutation algorithm proceeds as follows. When a genome of length L is replicated, we first draw the 4 numbers of rearrangements it will undergo. These 4 numbers all follow the binomial law $B(L, u_{\text{rearr}})$, where u_{rearr} is the per-base rate for the 4 types of rearrangement. Hence, the genome undergoes on average $u_{\text{rearr}} L$ inversions, $u_{\text{rearr}} L$ translocations, $u_{\text{rearr}} L$ large deletions, and $u_{\text{rearr}} L$ duplications (the fact that larger genomes undergo more rearrangements per replication aims at simply taking into account the fact that they contain more repeated sequences, while avoiding a time-consuming similarity search). Then, all these rearrangements are performed in a random order. To perform, for instance, a large deletion, 2 breakpoints p_1 and p_2 are chosen randomly (uniformly) on the chromosome, and the segment ranging from p_1 to p_2 in the clockwise sense is excised. In a similar manner, the boundaries of the duplicated, in-

verted, and translocated segments, as well as the reinsertion points for the translocated and duplicated segments, are also chosen uniformly on the chromosome. Once all the rearrangements have been performed, the new chromosome length is called L' and we draw the 3 numbers of local mutations (point mutations, small insertions, and small deletions). They all follow the binomial law $B(L', u_{\text{loc}})$, where u_{loc} is the per-base rate for the 3 types of local mutations. We finally perform all the local events in a random order, the affected positions being again randomly chosen. All the mutation rates were first adjusted to a same per-base pair value, $u_{\text{loc}} = u_{\text{rearr}} = u$ (with $u = 5.10^{-6}$, 10^{-5} , 2.10^{-5} , 5.10^{-5} , 10^{-4} , or 2.10^{-4}), in order not to give a priori importance to a specific category of genetic change. Then, we ran additional simulations where the rate of the local events, u_{loc} , was either smaller or larger than the rate of large-scale rearrangements, u_{rearr} .

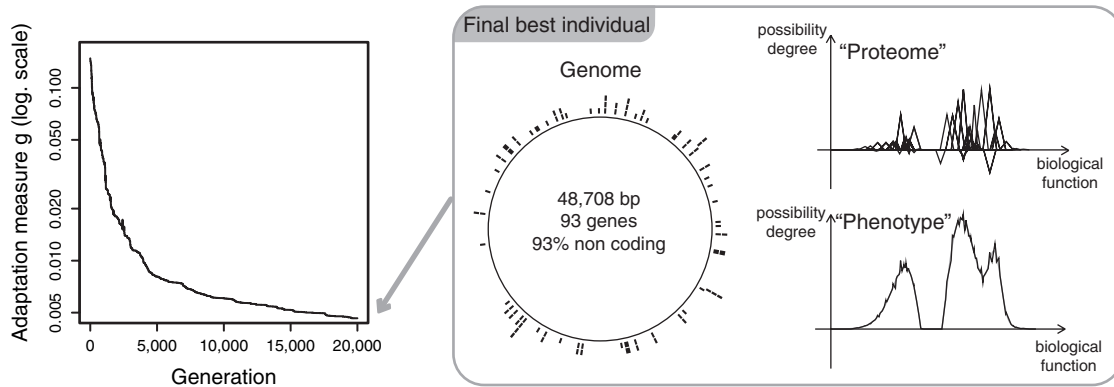
Initialization

To initialize each population, random genomes of 5,000 bp were tested until one was found whose phenotype narrows the gap g , due to at least one beneficial gene. The whole population was seeded with that single genome. The number of trials required to get a suitable genome can be used to estimate the probability to find by chance a functional gene in a random sequence. On average, 610 genomes of 5,000 bp were tested before getting a suitable one, which means that a functional gene is found every 3,050,000 bp (on average) in a random sequence. This shows, albeit indirectly, that local mutations in the intergenic sequences have a low probability to create new genes ex nihilo.

Evolution of the Population

The population size, N , is fixed and organisms reproduce asexually, according to their adaptation. To control the selective pressure, and to keep it constant throughout the evolution period (Whitley 1989), we used an exponential ranking selection scheme (Blickle and Thiele 1996): The expected number of offspring of a given organism is an exponential function of its rank in the population. Thus, at each generation, the N organisms were sorted from the least adapted to the best adapted. Their expected numbers of offspring then followed the multinomial law with $N = 1,000$ trials and reproduction probabilities $w_r = \frac{c-1}{c^N-1} c^{N-r}$, where r is the rank of the organism. The parameter $c \in]0, 1[$ is the curvature of the relationship between the rank and the probability of reproduction; hence, it controls the efficiency of the selection. The closer c is to 1, the less efficient the selection. This selection scheme allows us to test various selection efficiencies while keeping the population size tractable. We tested 4 values for c (0.9900, 0.9950, 0.9980, and 0.9995). For each combination of u and c , we tested 3 populations of $N = 1,000$ organisms. Supplementary text S2 (Supplementary Material online) presents additional experiments that were performed under a more classical selection scheme, where the probability

A. Typical evolution under a low mutation rate ($u = 5.10^{-6}$)



B. Typical evolution under a high mutation rate ($u = 2.10^{-4}$)

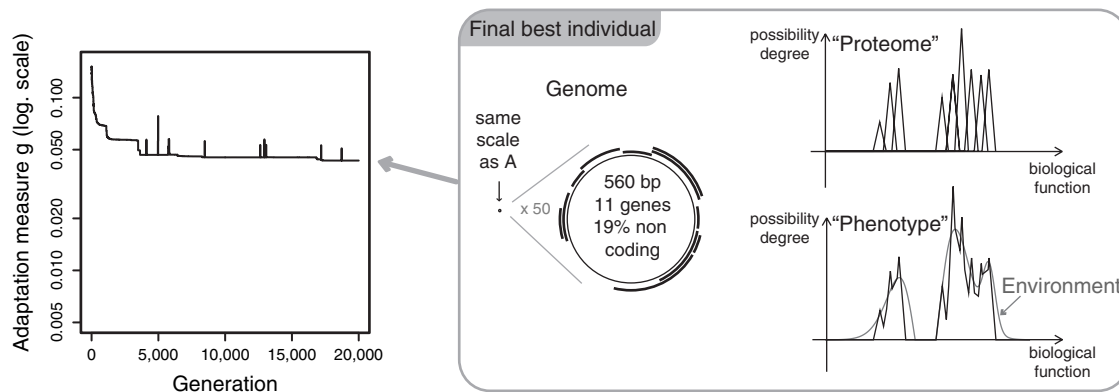


FIG. 2.—The evolved genomic structure strongly depends on the mutation rate. The figure shows the final fittest organism for 2 typical runs as well as the evolution of the adaptation measure (g) on its line of descent. For both runs, the initial genome size was 5,000 bp and the selection efficiency was $c = 0.9980$. Genome: The circle represents the chromosome, and the arcs represent the coding sequences. Proteome: Superimposition of the possibility distributions of the proteins encoded in the genome. Phenotype: Global possibility distribution resulting from the combination of all the proteins. Under a low mutation rate, the genome contains much more genes and a higher proportion of noncoding material than under a high mutation rate. The organism is also much better adapted to the environment.

of reproduction of an individual directly depends on its adaptation measure g rather than on its rank in the population.

evolve during 20,000 generations under various mutation rates combined with various selection efficiencies.

Estimates of the Fraction of Neutral Offspring

The theoretical estimates of the fraction F_v of neutral offspring were computed for the final fittest organism by considering the transcribed regions—including their promoters and terminators—as the coding units, overlapping regions being merged into a single unit (see supplementary text S1, Supplementary Material online). Empirical estimates were obtained by generating 1,000 offspring for each final fittest organism, with the same mutation rate, u , as during the evolution period and by counting the number of offspring that retained the same gap g .

Results

To study the spontaneous evolution of the amount of nonfunctional DNA, we allowed 72 asexual populations to

Relation between the Mutation Rate and Genome Compactness

The initial genomes contained only one gene. In all cases, the very first generations were characterized by duplication-divergence events, allowing the organisms to acquire new functional capabilities and to reduce the gap g with the environment (fig. 2). Then, after a few thousands of generations, both the gene number and the amount of noncoding sequences reached equilibrium (fig. 3A). The equilibrium values were independent of the initial genome size (data not shown) but strongly dependent on the mutation rate (figs. 2 and 3B). It has been suggested that as most mutations are deleterious, the per-base pair mutation rate can impose an upper limit to the number of genes (Eigen 1971; Maynard-Smith 1983; Hurst 1995; Pal and Hurst 2000) and this is indeed what happened here. The higher the mutation rate, the lower the number of genes at

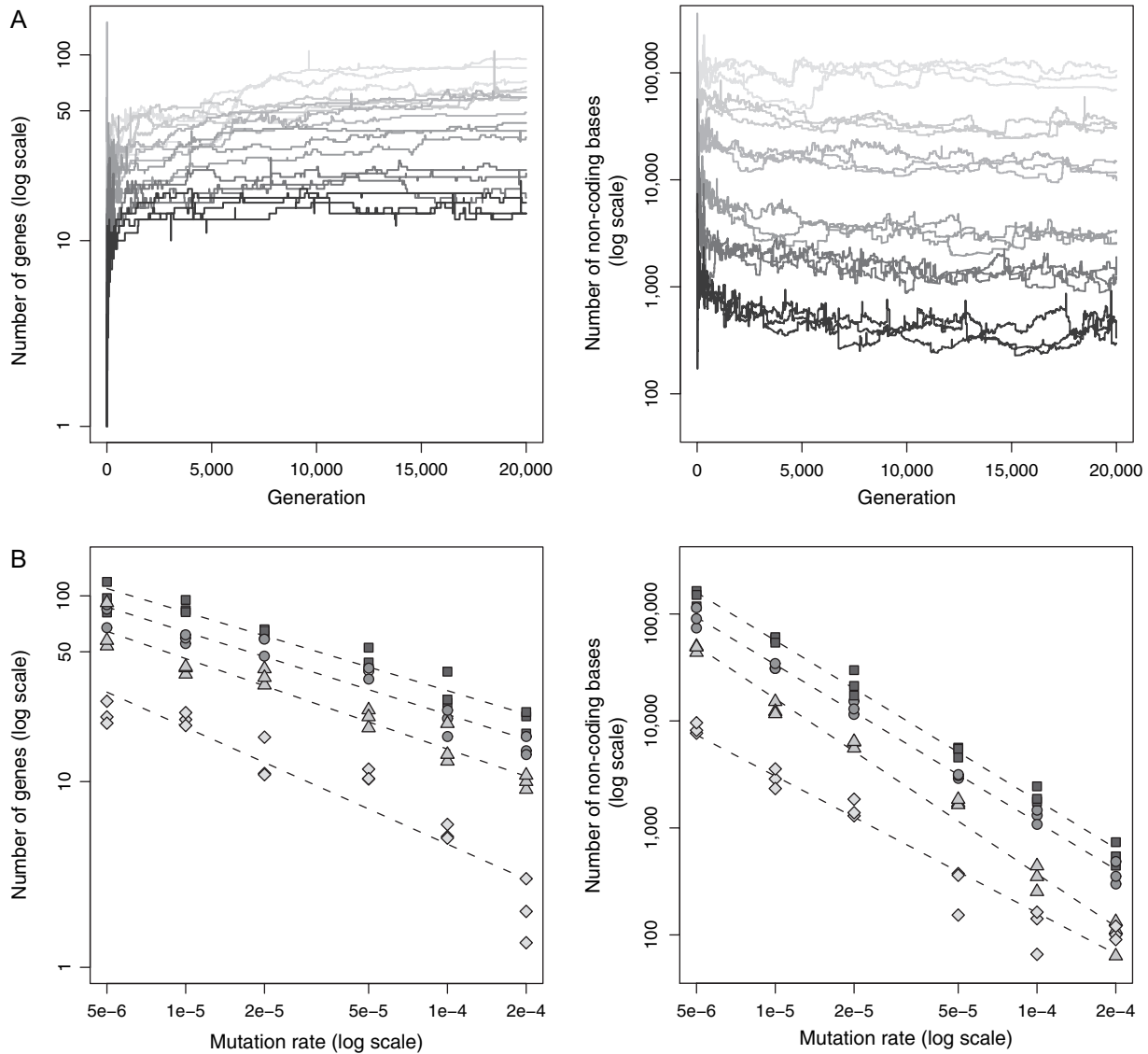


FIG. 3.—The lower the mutation rate, the higher both the gene number and the quantity of noncoding sequences. (A) On the line of descent of the final fittest organism, both the gene number (left) and the number of noncoding bases (right) reach equilibrium. Data shown are for a same selection efficiency ($c = 0.9950$) and a mutation rate ranging from 5.10^{-6} (light gray) to 2.10^{-4} (black) per base pair. (B) Both the gene number (left) and the number of noncoding bases (right) at equilibrium depend on the mutation rate, for the 4 selection efficiencies tested ($c = 0.9900$: squares, 0.9950 : circles, 0.9980 : triangles, and 0.9995 : diamonds). Equilibrium is estimated from the mean of the last 5,000 generations. Each log-log regression of gene number on mutation rate is significant ($n = 18$; $r^2 = 0.94, 0.93, 0.94, 0.87$ and $P = 5.10^{-11}, 2.10^{-10}, 5.10^{-11}, 2.10^{-8}$ for $c = 0.9900, 0.9950, 0.9980, 0.9995$, respectively), as is each log-log regression of the number of noncoding bases on the mutation rate ($n = 18$; $r^2 = 0.99, 0.99, 0.98, 0.95$ and $P = < 2.10^{-16}, < 2.10^{-16}, 2.10^{-15}, 1.10^{-11}$).

equilibrium (figs. 2 and 3B) and the higher the gap with the target (supplementary fig. S1, Supplementary Material online). However, more surprisingly, our experiments show that the mutational pressure also acted on the amount of noncoding sequences (figs. 2 and 3B). Under high mutation rates, the evolved genomes resembled viral ones, with overlapping genes and almost no noncoding sequences (fig. 2B). Under low mutation rates, the genomes contained high proportions of noncoding sequences (fig. 2A), up to 97% of the genome here. This implies that during adaptive evolution, large amounts of noncoding sequences can accumulate in the absence of self-replicating elements and without a predominance of the insertions on the deletions, if the per-base pair mutation rate is low. To further test this strong relation-

ship between the mutation rate and the architecture of the genome, we changed the per-base pair mutation rate after 10,000 generations. This caused the genomes to evolve quickly toward the size corresponding to the new mutation rate. We observed this tight coupling for the 4 selection efficiencies tested, the genomes being globally larger when the selection strengthens (fig. 3B).

Role of the Noncoding Sequences in the Mutational Variability of the Phenotype

To test whether the indirect selection of a specific level of variability could underlie this coupling, we investigated

the role of genome compactness in the mutational variability of phenotype. One indicator of this variability is the fraction of “neutral offspring” (Ofria et al. 2003), that is, the fraction of offspring without mutation or with only neutral mutations. Here, it can be inferred from genomic parameters, if we assume that a mutation is neutral if it does not affect any functional region, a functional region being defined as a transcribed region (promoter and terminator included) containing at least one coding sequence. This hypothesis is simplistic for a real organism but quite accurate for the artificial organisms that were evolved. In these conditions, the fraction of neutral offspring F_v can be approximately calculated using the probability that no functional region mutates during replication

$$F_v \approx \prod_{\substack{j=\text{point mut.}, \\ \text{small ins.}, \text{ small del.}}} (1 - u_{\text{loc}}(1 - \tilde{v}_j))^L \cdot \prod_{\substack{j=\text{inv.}, \text{ transloc.}, \\ \text{duplic.}, \text{ large del.}}} (1 - u_{\text{rearr}}(1 - \tilde{v}_j))^L, \quad (1)$$

where u_{loc} is the rate of local mutations, u_{rearr} the rate of rearrangements, L the genome length, and \tilde{v}_j the probability that a random mutation of type j does not affect any transcribed region. This probability can be computed for each type of mutation:

$$\left\{ \begin{array}{l} \tilde{v}_{\text{point mut.}} = \tilde{v}_{\text{small ins.}} = \tilde{v}_{\text{small del.}} = 1 - \frac{l}{L} \\ \tilde{v}_{\text{inv.}} = \left(1 - \frac{l}{L}\right)^2 \\ \tilde{v}_{\text{transloc.}} = \left(1 - \frac{l}{L}\right)^3 \\ \tilde{v}_{\text{large del.}} = \frac{1}{2L^2} \sum_{i=1}^{N_G} \lambda_i (\lambda_i + 1) \\ \tilde{v}_{\text{duplic.}} = \frac{1}{2L^2} \left(1 - \frac{l}{L}\right) \sum_{i=1}^{N_G} \lambda_i (\lambda_i + 1) \end{array} \right., \quad (2)$$

with l the total length of functional regions, N_G the number of functional regions, and λ_i the length of the intergenic sequence between the functional regions i and $i + 1$ (see supplementary text S1, Supplementary Material online for the details of this derivation). These equations show that, for a given mutation rate, longer intergenic sequences lower the fraction of neutral offspring (fig. 4) and hence promote the exploration of new phenotypes. There are 2 reasons for this. The first is that when new noncoding bases are acquired, the genome undergoes more mutational events. The second is that longer intergenic sequences do not make duplications and large deletions more neutral (eq. 2 and fig. 4). Indeed, contrary to the other types of mutation, their deleterious effects are not concentrated on a few points. Here, the average length of the rearranged segments increases with genome length, which implies that a duplication or a large deletion is not more likely to be neutral when intergenic sequences grow. Longer genomes undergo, however, more duplications and deletions per replication. The net effect of longer intergenic sequences is that genes have a higher probability to be deleted or duplicated at each replication. In short, intergenic sequences promoting large deletions and duplications are

mutagenic for the genes they surround. Thus, longer intergenic sequences tend to enhance the level of nonneutral variation, that is, the mutational variability of the phenotype.

Indirect Selection of a Constant Level of Mutational Variability

In our model, noncoding sequences are not under direct selection; hence, their sizes can be easily increased and compensate for a low mutation rate or, conversely, decreased and compensate for a high mutation rate. To test whether this is what actually happened in our experiments, we calculated, for each run, the fraction F_v of neutral off-

spring of all ancestors of the final best individual, using equations 1 and 2. As shown in supplementary figure S3 (Supplementary Material online), F_v stabilizes quickly, after less than 5,000 generations. The final values are shown in figure 5A. We also computed empirical estimations of the final F_v (by simulating 1,000 independent replications of the final fittest individual, see Materials and Methods). These empirical values are shown in figure 5B. Both methods agree well and show that for a given selection efficiency, the evolved organisms exhibit roughly the same fraction of neutral offspring whatever the mutation rate be (fig. 5A and B). Thus, for each of the 4 selection efficiencies, a same level of mutational variability was indirectly selected. Under a low (respectively high) mutation rate, the organisms that exhibited the selected level of variability were those with a large (respectively compact) genome. Hence, the indirect selection of 4 specific levels of variability drove the evolution of genome compactness in the 4 data sets and underlies the 4 observed relationships between the mutation rate and the amount of nonfunctional DNA.

To test the generality of this principle, we ran additional experiments, where the rate of local mutations u_{loc} and the rate of rearrangements u_{rearr} were allowed to differ (all these experiments were run under a same selection intensity, $c = 0.9980$). As shown in table 1, the evolved fraction of neutral offspring is of the same order when $u_{\text{loc}} > u_{\text{rearr}}$, when $u_{\text{loc}} < u_{\text{rearr}}$, and when $u_{\text{loc}} = u_{\text{rearr}}$. This suggests that the fraction of neutral offspring is a general criterion driving the spontaneous evolution of genome compactness.

Respective Roles of the Local Mutations and the Rearrangements

Table 1 also shows that u_{loc} and u_{rearr} both influence genome compactness and that they do so in the same direction. A higher rate of local mutations leads to more compact

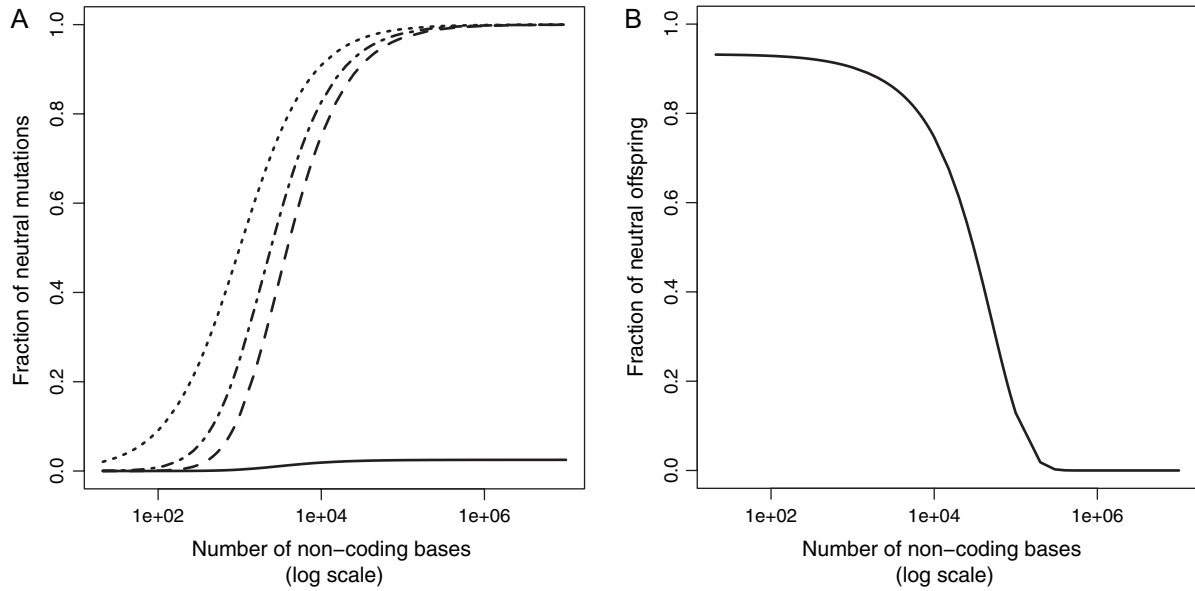


FIG. 4.—Noncoding sequences promote variability. (A) When noncoding bases are added into a genome, point mutations ($\bar{v}_{\text{pointmut.}}$, dotted line), inversions ($\bar{v}_{\text{inv.}}$, dotted-dashed line), and translocations ($\bar{v}_{\text{transloc.}}$, dashed line) have a higher probability to be neutral. On the contrary, the proportions of neutral duplications and deletions ($\bar{v}_{\text{dup.}}$ and $\bar{v}_{\text{del.}}$, solid line) do not increase. (B) As a result, the theoretical fraction of neutral offspring F_V decreases when the genome contains more noncoding bases. All these theoretical curves were calculated for $u = 10^{-5}$, with 20 genes, each of length 50 bp, regularly distributed on the chromosome.

genomes and so does a higher rate of rearrangements. Conversely, either a lower rate of local mutations or a lower rate of rearrangements leads to a larger genome. This means that if either u_{loc} or u_{rearr} is changed, this is compensated for by changing the number of genes, N_G , and the lengths of the intergenic sequences, λ_i (see eqs. 1 and 2). These additional experiments raise an interesting point: Although the existence of duplications and large deletions is indispensable for the effect to take place (without them, changing the λ_i 's would have no effect on F_V), it is not mandatory to change their own rate to get an effect on the amount of non-coding DNA. Changing only the local mutation rate suffices to induce an effect. By changing u_{loc} , the left term in equation 1 is modified and it is compensated for in the right term by changing the λ_i 's. This effect is however smaller

than the effect induced by changing directly the rate of rearrangements table 1.

What Determines the Selected Level of Variability?

There remains an important question: What determines the value of the selected level of variability and why does this value depend on the selection intensity? The intensity of the selection, c , sets the relative probability of reproduction of the best individuals compared with the least adapted. When $c = 0.9900$ (efficient selection), the best individual gets an average of $W = 10$ reproductive trials. Now, the lineage can persist only if at least one of these offspring retains the ancestral phenotype, that is, if $F_V W \geq 1$. Hence, F_V must be greater than $1/W$, which

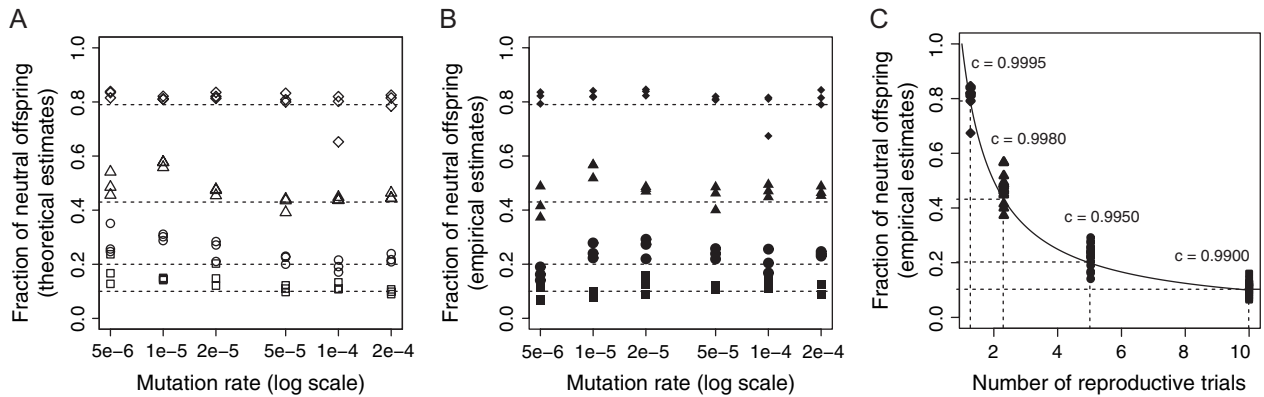


FIG. 5.—The intensity of the selection sets the appropriate level of variability. For each run, we estimated the fraction F_V of neutral offspring of the final fittest organism, both theoretically (using eqs. 1 and 2) and empirically (by simulating 1,000 independent replications, see Materials and Methods). (A and B) For a given selection intensity ($c = 0.9900$: squares, 0.9950: circles, 0.9980: triangles, 0.9995: diamonds), this evolved F_V is roughly the same for the 6 mutation rates tested. The evolved F_V is close to the value that would ensure an average of one neutral offspring to the best individual, namely $\frac{1}{W} = \frac{c^N - 1}{N(c-1)}$ (dotted horizontal lines). (C) The evolved F_V (all mutation rates together) as a function of W , the number of reproductive trials of the individual. They are indeed close to $1/W$ (black curve).

Table 1
Respective Influence of the Local Mutation Rate and the Rearrangement Rate

Experiment	Local Mutation Rate (u_{loc})	Rearrangement Rate (u_{rearr})	Gene Number ^a	Number of Noncoding Positions ^a	Fraction of Neutral Offspring (F_v) ^b
Reference	2.10^{-5}	2.10^{-5}	36.6 ± 3.7	$6\ 106 \pm 470$	0.48 ± 0.01
Lower u_{rearr}	2.10^{-5}	5.10^{-6}	42.2 ± 0.3	$21\ 546 \pm 4,563$	0.59 ± 0.05
Higher u_{rearr}	2.10^{-5}	2.10^{-4}	14.3 ± 5.1	289 ± 23	0.45 ± 0.04
Lower u_{loc}	5.10^{-6}	2.10^{-5}	45.7 ± 4.5	$7\ 500 \pm 1,168$	0.50 ± 0.05
Higher u_{loc}	2.10^{-4}	2.10^{-5}	15.9 ± 0.5	2197 ± 527	0.46 ± 0.02

^a Value at equilibrium (estimated as in fig. 3), mean \pm standard deviation (SD) on $n = 3$ repetitions.

^b Empirical estimate for the final best organism (see Materials and Methods), mean \pm SD on $n = 3$ repetitions.

means here that at least 10% of the offspring must bear no mutation or only neutral ones. Let us consider now a weaker selection, where the best individuals do not get many more reproductive trials than the worst ones. If $c = 0.9995$, for example, the expected number of reproductive trials of the best individual is as low as $W = 1.27$; hence, F_v must be greater than 79%. These examples show that the intensity of the selection indirectly determines a lower bound for F_v , and hence—for a given mutation rate—an upper bound for the number of genes and for the amount of noncoding sequences (eqs. 1 and 2).

It is harder to explain why the evolved F_v is actually always almost equal to its lower bound, $1/W$ (fig. 5). This means that all the 72 successful organisms share one property: When they reproduce, their progeny contains one neutral offspring, that is, the minimum ensuring the persistence in the following generation, but not more. In other words, for each of them, the genome is as large as possible given 1) the maximal number of reproductive trials he can expect and 2) the per-base pair mutation rate he undergoes. To make sure that this is not due to a hidden mutational bias toward genome growth, we monitored the evolution of genome size without any selection. In all cases, the genomes lost all their genes and shrank to less than 100 bp (supplementary fig. S4, Supplementary Material online), which suggests that in the standard runs the genome size is actively maintained by the selective pressure. This could basically be the direct selective pressure to close the gap g , which may tend to favor genomes with many genes and hence with low F_v . This could also reflect the indirect selection of the lineages that were variable enough to explore new phenotypes and to sometimes discover fitter ones (or rediscover a fit phenotype after a deleterious mutation): in the lineage of the final best individual, beneficial mutations keep occurring even in the last 1,000 generations (data not shown). Such a pressure would not only favor high gene numbers, but could also favor high amounts of noncoding sequences (figs. 2 and 4). Thus, the empirical “rule” of one neutral offspring as a key of the long-term evolutionary success most likely reflects a trade-off between a sufficient fidelity of the transmission of the phenotype, a sufficient ability to explore new phenotypes, and a sufficient fitness.

To test whether this principle still applies under a more realistic selection scheme, we performed all experiments again under a “fitness-proportionate” selection scheme. In these experiments, the probability of reproduction of an individual directly depended on the absolute value of its gap g with environment, rather than on its rank in the

population (see Materials and Methods). The evolved F_v were again of the order of $1/W$, which means that the successful individuals were again those who produce one neutral offspring when they reproduce (see supplementary text S2, Supplementary Material online for more details). Besides, we obtained the same type of relationship between the mutation rate and the amount of noncoding positions (supplementary text S2, Supplementary Material online). This data set confirms that the previous results are not an artifact of the ranking selection scheme.

Discussion

Taken together, our experiments and the mathematical analysis show that a specific level of mutational variability is indirectly selected, which in turn induces the selection of a specific amount of noncoding sequences, depending on the mutation rate and the selection efficiency. This does not require the evolutionary process to be farsighted. Nor does it require that selection acts on a group level. In our experiments, selection acted only on the individuals. Individuals whose phenotypes are not robust enough undergo deleterious mutations and disappear, whereas individuals whose phenotypes are not variable enough are outcompeted by those that were able to discover innovations. In our experiments, the long-term evolutionary success requires that one of the offspring produced at each generation retains the phenotype of its progenitor, which reflects this trade-off between the exploration of new phenotypes and the reliable transmission of the current one.

What are the consequences of the selection of a specific variability level on genome compactness? It depends on the contribution of nonfunctional DNA to the variability level. Here, the simple mutational patterns we used allowed us to describe the relationship between genome structure and variability by simple equations (see eqs. 1 and 2 and fig. 4). In the tested situation, the fraction of neutral offspring decreases when additional noncoding bases are acquired because 1) more rearrangements occur and 2) the average size of the rearranged segments increases. The former effect is plausible if the number of repeated elements increases with genome size, which seems plausible for both bacterial and eukaryote genomes (Achaz et al. 2001; Achaz et al. 2002). The latter is a consequence of the uniform distribution we assumed here for the size of the spontaneous rearrangements. Is such a distribution relevant for living species? Although comparative genomics approaches can reveal the size distribution of the fixed rearrangements, it

is extremely difficult to assess the size distribution of all the spontaneous rearrangements that occur in evolving populations. Indeed, in living organisms, illegitimate recombination, site-specific recombination, general homologous recombination, gene amplification by retroposition, and whole-genome duplications all contribute to genome dynamics at different levels (Hughes 1999; Rocha 2003; Cannon et al. 2004; Dujon et al. 2004). Hence, one can expect that each species, depending on its mutational patterns, exhibits its own complex, probably multimodal, size distribution. Our choice of a uniform distribution basically reflects the lack of knowledge in this area. However, we expect that in qualitative terms, the dynamics of noncoding DNA does not depend on the specific distribution of segment size, provided that the average segment size increases with genome size.

Aside from genome structure and the variety of mutational patterns, many other factors can influence the level of mutational variability of a living organism. From the robustness of protein folding to the robustness of developmental pathways, a multitude of mechanisms modulate the fraction of neutral mutations (Wagner 2005). Recombination also plays a major role in variation and may have its own effect on the length of the noncoding sequences (Cameron 2001). Thus, the relationship between genome size and the fraction of neutral offspring can be more complex in a living species than in our experiments.

As a consequence, the link we underscore here between the mutation rate and genome compactness may be difficult to reveal experimentally in living species. It is, however, noteworthy that the relationship we obtained is qualitatively consistent with Drake's (1991) data, gathered for several microbial species from phage to fungi. Our results may provide an explanation for the constant genome-wide mutation rate he observed. It may reflect the indirect selection of the genome structure that allows for the best trade-off between a reliable transmission of the genetic information and the exploration of the mutational neighborhood. If the tested species share roughly the same mutational patterns, the same selective pressure and similar mutational robustness at other levels than the genome, then we can indeed expect them to appear on the same line on the log-log plot of genome size versus mutation rate (fig. 3). This pattern cannot be seen when unicellular and multicellular species are mixed (Lynch 2006), probably because the transition to multicellularity has introduced fundamentally new mechanisms of mutational robustness, like a cellular selection in the germ line.

Conclusion

These *in silico* experiments shed light on a long-term evolutionary pressure that can drive the loss or the accumulation of noncoding sequences. These results show that under low mutation rates, a large amount of noncoding sequences can be maintained despite the absence of mutational biases or proliferation of "selfish" elements. A forthcoming challenge is the design of *in vitro* or *in vivo* experiments that could assess the strength of this spontaneous, long-term evolutionary dynamics compared with more immediate pressures like the self-replication of transposable elements. Furthermore, with the evidence that indirect

selective pressures shape genome structure *in silico*, it is relevant to search for the hallmark of such pressures at all levels between genotype and phenotype, from protein sequence to gene networks and developmental pathways.

Supplementary Material

Supplementary texts S1 and S2 and figures S1, S2, S3, and S4 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

We thank F. Taddéi, E. Rocha, C. Adami, L. Duret, V. Daubin, J. Lobry, S. Mousset, H. Charles, and H. Soula for comments on the manuscript. This work is part of the Biologie des Systèmes et Modélisation Cellulaire project. It is supported by the Rhône-Alpes region, the Bioinformatics Program of INSA Lyon, and the Rhône-Alpes Complex Systems Institute.

Literature Cited

- Achaz G, Netter P, Coissac E. 2001. Study of intrachromosomal duplications among the eukaryote genomes. *Mol Biol Evol.* 18:2280–2288.
- Achaz G, Rocha EPC, Netter P, Coissac E. 2002. Origin and fate of repeats in bacteria. *Nucleic Acids Res.* 30:2987–2994.
- Adami C. 2006. Digital genetics: unravelling the genetic basis of evolution. *Nat Rev Genet.* 7:109–118.
- Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature.* 437:1149–1152.
- Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science.* 304:1321–1325.
- Blickle T, Thiele L. 1996. A comparison of selection schemes used in evolutionary algorithms. *Evol Comput.* 4:361–394.
- Brosius J. 2003. How significant is 98.5% 'junk' in mammalian genomes? *Bioinformatics.* 19:ii35.
- Brosius J, Gould SJ. 1992. On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proc Natl Acad Sci USA.* 89:10706–10710.
- Burch CL, Chao L. 2000. Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature.* 406:625–628.
- Burke DS, De Jong KA, Grefenstette JJ, Ramsey CL, Wu AS. 1998. Putting more genetics into genetic algorithms. *Evol Comput.* 6:387–410.
- Cannon SB, Mitra A, Baumgarten A, Young ND, May G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* 1:4–10.
- Coghlan AG, Eichler EE, Oliver SG, Paterson AH, Stein L. 2005. Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends Genet.* 21:673–682.
- Cameron JM. 2001. What controls the length of noncoding DNA? *Curr Opin Genet Dev.* 11:652–659.
- Denver DR, Morris K, Lynch M, Thomas WK. 2004. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature.* 430:679–682.
- Dermitzakis ET, Reymond A, Antonarakis SE. 2005. Conserved non-genic sequences—an unexpected feature of mammalian genomes. *Nat Rev Genet.* 6:151–157.
- Drake JW. 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci USA.* 88:7160–7164.
- Dujon B, Sherman D, Fischer G, et al. (67 co-authors). 2004. Genome evolution in yeasts. *Nature.* 430:35–44.

- Duret L, Dorkeld F, Gautier C. 1993. Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Res.* 21:2315–2322.
- Eigen M. 1971. Selforganization of matter and evolution of biological macromolecules. *Naturwissenschaften.* 58:465–523.
- Frazer KA, Sheehan JB, Stokowski RP, Chen X, Hosseini R, Cheng JF, Fodor SPA, Cox DR, Patil N. 2001. Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* 11:1651–1659.
- Hughes D. 1999. Impact of homologous recombination on genome organization and stability. In: Charlebois RL, editor. *Organization of the Prokaryotic Genome.* Washington (DC): ASM Press. p. 109–128.
- Hurst LD. 1995. The silence of the genes. *Curr Biol.* 4:459–461.
- Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ. 2005. Evolutionary constraints in conserved nongenic sequences of mammals. *Genome Res.* 15:1373–1378.
- Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica.* 115:49–63.
- Kirschner M, Gerhart J. 1998. Evolvability. *Proc Natl Acad Sci USA.* 95:8420–8427.
- Knibbe C, Mazet O, Chaudier F, Fayard J-M, Beslon G. 2007. Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences. *J Theor Biol.* 244:621–630.
- Lynch M. 2006. The origins of eukaryotic gene structure. *Mol Biol Evol.* 23:450–468.
- Lynch M, Conery JS. 2003. The origins of genome complexity. *Science.* 302:1401–1404.
- Maestre J, Tchenio T, Dhellin O, Heidmann T. 1995. mRNA retroposition in human cells: processed pseudogene formation. *EMBO J.* 14:6333–6338.
- Margulies EH, Blanchette M, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* 13:2507–2518.
- Maynard-Smith J. 1983. Models of evolution. *Proc R Soc Lond B Biol Sci.* 219:315–325.
- Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17:589–596.
- Ofria C, Adami C, Collier TC. 2003. Selective pressures on genomes in molecular evolution. *J Theor Biol.* 222:477–483.
- Pal C, Hurst LD. 2000. The evolution of gene number: are heritable and non-heritable errors equally important? *Heredity.* 84:393–400.
- Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA loss as a determinant of genome size. *Science.* 287:1060–1062.
- Radman M, Matic I, Taddei F. 1999. Evolution of evolvability. *Ann N Y Acad Sci.* 870:146–155.
- Rocha EPC. 2003. An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. *Genome Res.* 13:1123–1132.
- Smit AFA. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev.* 9:657–663.
- Van Nimwegen E, Crutchfield JP, Huynen M. 1999. Neutral evolution of mutational robustness. *Proc Natl Acad Sci USA.* 96:9716–9720.
- Wagner A. 2005. *Robustness and evolvability in living systems.* Princeton (NJ): Princeton University Press.
- Wagner GP, Altenberg L. 1996. Complex adaptations and the evolution of evolvability. *Evolution.* 50:967–976.
- Whitley D. 1989. The GENITOR algorithm and selection pressure: why rank-based allocation of reproductive trials is best. In: Schaffer JD, editor. *Proceedings of the 3rd International Conference on Genetic Algorithms.* San Mateo (CA): Morgan Kaufmann. p. 116–121.
- Wilke CO. 2001a. Adaptive evolution on neutral networks. *Bull Math Biol.* 63:715–730.
- Wilke CO. 2001b. Selection for fitness versus selection for robustness in RNA secondary structure folding. *Evolution.* 55:2412–2420.
- Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C. 2001. Evolution of digital organisms at high mutation rates leads to the survival of the flattest. *Nature.* 412:331–333.
- Wu AS, Lindsay RK. 1996. A comparison of the fixed and floating building block representation in the genetic algorithm. *Evol Comput.* 4:169–193.