Deletion rates are thought to be important factors in determining the genome size of organisms in nature. Although it is indisputable that deletions, and thus deletion rates, affect genome size, it is unclear how, or indeed if, genome size is regulated via the deletion rate. Here, we employ a mathematical model to determine the evolutionary fate of deletion rate mutants. Simulations are employed to explore the interactions between deletions, deletion rate mutants, and genome size. The results show that, in this model, the fate of deletion rate mutants will depend on the fraction of essential genomic material, on the frequency of sexual recombination, as well as on the population size of the organism.

We find that there is no optimal deletion rate in any state. However, at one critical coding density, all changes in deletion rate are neutral and the rate may drift either up or down. As a consequence, the coding density of the genome is expected to fluctuate around this critical density. Characteristic differences in the impact of deletion rate mutations on prokaryote and eukaryote genomes are described.

Key words: genome size, junk DNA, insertion–deletion dynamics, deletion bias, modifier evolution.

E-mail: otto.berg@ebc.uu.se.

doi:10.1093/molbev/msp054
Advance Access publication March 13, 2009

© The Author 2009. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org
they might experience selection for or against deletion during these cell generations.

Genome size is determined both by functional genes and by sequences that appear to have no function. There are differences in the amount of functional sequence between different organisms, for instance, eukaryotes have genes that are both longer and larger in number compared with those of Bacteria and Archaea. However, it is the amount of apparently nonfunctional sequence that causes the huge differences in genome size between these organisms. We assume that minimal genome size is a selected property. So, the present modeling applies only to organisms where this is the case. The general rationale for this assumption is that smaller genomes support a faster growing organism with its smaller cell size and lower metabolic demands for maintenance and growth. However, this selective advantage is counteracted by the damage which random deletions, like all mutations, might cause to essential genes with a consequent loss of fitness. The balance between selection and counterselection determines the net selective pressure for, or against, deletions in any given situation. This pressure is translated to some degree into selection for either reduced or increased deletion rates in any particular situation. Of course, the selective effect of deletions might be influenced by other factors as well. For example, if a genome is heavily burdened by active transposons, there will be additional selection in favor of frequent deletions that reduce the transposon load to confer a fitness benefit. However, in the present calculations, we have avoided the extra complications caused by transposon dynamics.

We describe the selection model outlined above in a formal framework, and we apply it to a virtual population of cells in order to examine the dynamics of indirect selection under different circumstances. The results of our simulations allow us to address general questions regarding the evolution and regulation of deletion rates and genome sizes. This study suggests that it is in general difficult to control genome size even when large sizes are strongly counterselected. For example, if sexual recombination is sufficiently frequent, increasing the deletion rate becomes favorable only when the coding density is very low. We use the results of our simulations to contrast the typical features of genomes of eukaryotes on the one hand with those of Archaea and Bacteria.

Models

Our model considers haploid cells in a Moran population with overlapping generations, where the cells grow and evolve in a population of constant size, with selection for growth as well as equal death rates for all cells. Each cell has a given genome size, which is the sum of two parts. The first part is the core genome, which is the same size for all cells, representing the set of essential genes. The second part is nonessential genetic material, which changes over time due to deletions and insertions and is potentially different from cell to cell. The population contains cells with different deletion rates, upon which they compete. For simplicity, there are just two competing deletion rates and one is fixed whereas the other one is lost. This is the basis for the estimates of fixation probability and, by inference, the effective selection for mutant alleles. Given the assumption that small genome size is favorable, deletions are selected and insertions are counterselected.

Parameters

The fundamental parameters are the population size \(N\), the wild-type and mutant deletion rates per nucleotide \(v_{\text{del}}^v\) and \(v_{\text{del}}^u\), respectively, the frequency of sexual recombination per genome replication \(P_{\text{rec}}\), and selection coefficient for a nonlethal deletion \(s\). Also included are the insertion rate \(v_{\text{ins}}\) per nucleotide, the length of an indel event \(l\) and the initial sizes of the essential core, and the nonessential part of the genome \(L_{\text{cod}}\) and \(L_{\text{non}}\), respectively as well as the number of allowed insertion points in the core genome \(g_{\text{cod}}\). This final parameter, which describes the short strings separating coding sequences, is relevant primarily in genomes with very little nonessential material; it determines the number of sites, where an insertion is acceptable but where a (larger) deletion is lethal.

Because fixation or loss will happen quickly unless the population is huge, simulation times are short both in terms of real time and generations. Therefore, the total number of indel events is small, and the influence of changes in the genome composition is also small. As long as these assumptions apply, the expected selection pressure for or against increased deletion rates will be constant during the simulation.

Indel and Fitness Dynamics

Deletions and insertions occur with rates that depend both on the intrinsic rates characteristic of the cell in question and its current genome configuration. The model uses the following expressions to calculate the probability of a deletion \(P_{\text{del}}\) or insertion \(P_{\text{ins}}\) in a cell with deletion rate \(v_{\text{del}}\), insertion rate \(v_{\text{ins}}\), carrying \(n_{\text{ins}}\) insertions, and \(n_{\text{del}}\) deletions:

\[
P_{\text{del}} = \frac{v_{\text{del}}[(n_{\text{ins}} - n_{\text{del}})l + L_{\text{cod}} + L_{\text{non}}]}{1 + v_{\text{del}}[(n_{\text{ins}} - n_{\text{del}})l + L_{\text{cod}} + L_{\text{non}}]},
\]

\[
P_{\text{ins}} = \frac{v_{\text{ins}}[(n_{\text{ins}} - n_{\text{del}})l + L_{\text{cod}} + L_{\text{non}}]}{1 + v_{\text{ins}}[(n_{\text{ins}} - n_{\text{del}})l + L_{\text{cod}} + L_{\text{non}}]}.
\] (1)

The rates of indel are both proportional to the size of the genome and therefore in agreement with the model of proportional evolution also proposed by Oliver et al. (2007). These expressions, equation (1), are only suitable for small probabilities and will lead to underestimates of the deletion rate if the expected number of events per division approaches one. The fitness of cell \(i\) \((w_i)\) is given by the difference between the number of deletions and insertions it carries, according to the expression below.

\[
w_i = (1 + s)^{(n_{\text{del}} - n_{\text{ins}})}.
\] (2)

Furthermore, any mutational event may inactivate an essential gene. This happens with a probability equal to the
relative size of the essential genome compared with the entire genome. If an essential gene is inactivated, \( w_i \) is set to 0 for that cell.

\[
\begin{align*}
\frac{p_{\text{del lethal}}}{P_{\text{rec lethal}}} &= \frac{L_{\text{cod}}}{(n_{\text{ins}} - n_{\text{del}})l + L_{\text{cod}} + L_{\text{non}}}, \\
\end{align*}
\]

The lethal deletion probability in equation (3) does not account for boundary effects, that is, situations where the deletion overlaps part of a nonessential region as well as a coding sequence. This will influence the results at very high coding density. Naturally, the fitness of each cell is interesting only in relation to the total fitness of the population \( w_{\text{tot}} \), which is simply:

\[
w_{\text{tot}} = \sum_{i=0}^{N} w_i.
\]

Recombination

A fraction \( (P_{\text{rec}}) \) of the cell divisions consists of meiosis, and this involves sexual recombination. In these events, indel mutations can move from one genome to the other. Transfer is done by counting the number of deletions in each genome, taking the difference of these numbers, and moving some part of this difference from the genome with the most deletions to the other. Then, the process is repeated for the insertions. With this recombination model, only a subset of the indel events is eligible for transfer in each meiosis event. The rationale for using this model is that indel mutations are not necessarily unique, due to the relatedness of the cells involved. We also tried, with similar results, a model in which recombination did not mix the individual indels but only allowed the linkage between the deletion rate allele and its associated indels to be broken.

Monte Carlo Simulations

The simulations proceed cell division by cell division according to the following scheme.

1. The dividing cell is picked from the population. A cell with fitness \( w_i \) has probability equal to \( w_i / w_{\text{tot}} \) to be chosen for division. This is the selection step.
2. The cell replicates, and may acquire additional deletions or insertions, with probabilities \( P_{\text{del}} \) and \( P_{\text{ins}} \), respectively. If a lethal event occurs, the cell is assigned a fitness of zero and that cell division is aborted.
3. The type of cell division, sexual or asexual, is chosen. Sexual divisions occur with probability \( P_{\text{rec}} \).
4. Asexual division. The newborn cell then replaces a cell in the population at random. This constitutes death in the system, which occurs without selection.

3.1. Sexual division. In this case, a mate to the dividing cell is picked at random, without considering fitness. The cells exchange genetic material, with each eligible indel mutation having a 50% probability of moving from one cell to the other.

3.2. Asexual division. The newborn cell then replaces a cell in the population at random. This constitutes death in the system, which occurs without selection.

4. The fitness values of the cells involved in the division are calculated and \( w_{\text{tot}} \) is updated, whereupon, the simulation proceeds to the next cell division.

Results

Fixation of a Deletion Rate Mutant

We obtain a measure of the effective selection pressure acting on a mutant allele from the simulations of the fixation probabilities of the deletion rate mutants in populations with varying frequencies of sexual recombination. As expected, the effective selection strength decreases with increasing sexual recombination, which is in line with other results concerning modifier genes (Johnson 1999). The same is true for increasing the density of essential genetic material. These two effects are particularly strong when they are combined. That is to say, increasing the deletion rate is strongly counterselected in genomes with frequent sexual recombination and high densities of essential coding sequences. The results are summarized in figure 1, which clearly shows that sexual recombination effectively disrupts positive selection for increased deletion rate, unless the coding sequences of the genome are very sparse. Only if sexual recombination is very uncommon (fig. 1A) does the deletion rate mutant remain favorable in even moderately dense genomes.

A striking feature of the simulations is the presence of a critical coding ratio, which is particularly evident in genomes where recombination is frequent (fig. 1B and C). This critical ratio can be described as a switch point for selection on the deletion rate. At densities lower than the critical ratio, increasing the deletion rate is strongly favored; at the critical ratio, there is little or no selection acting on the deletion rate, and at higher densities, there is selection to reduce the deletion rate.

In order to explain the results in figure 1, it is useful to examine the effective selection coefficient \( \left( s^* \right) \) for a deletion rate mutant that confers a deletion rate of \( v_{\text{del}}^* \). \( s^* \) can be expressed as

\[
s^* = m_{\text{ex}} s - (v_{\text{del}}^* - v_{\text{del}}^\text{wt})L_{\text{cod}},
\]

where \( m_{\text{ex}} \) is the extra number of deletions in a genome that carries the mutant deletion rate allele, compared with a wild-type genome. Thus, the first term is the expected positive contribution from a decreased genome size. At large recombination rates, the linkage between the deletion rate mutant and the extra deletions it causes will be disrupted and \( m_{\text{ex}} \) is consequently smaller. The second term is the extra number of lethal mutations that are associated with a higher deletion rate. It can be expected that \( m_{\text{ex}} \) is proportional to \( L_{\text{non}} \) as well as to the difference in deletion rates:

\[
m_{\text{ex}} = C(v_{\text{del}}^* - v_{\text{del}}^\text{wt})L_{\text{non}},
\]

Here, \( C \) is a constant that depends on other parameters, including the recombination rate \( (P_{\text{rec}}) \). In the
density is calculated as

\[ \rho = \frac{m_{\text{mut}}}{P_{\text{fix}}} \]

where \( m_{\text{mut}} \) represents neutrality. Other parameters are given by equation (6). Using this result, \( s^* \) can be expressed as

\[ s^* = \frac{\left( v_{\text{del}} - v_{\text{wt}} \right) L_{\text{cod}} C_s L_{\text{non}}}{L_{\text{cod}} - 1} \]

which holds reasonably well in almost all cases we have tested (cf., fig. 1).

Equation (7a) is a powerful expression, and it can be used to describe the entire system. Using the simulations to estimate \( d_c \) and applying the initial condition of \( n_0 = 200 \) and \( N = 1,000 \), where \( n_0 \) is the initial number of individuals carrying the mutant allele, the fixation probability will be determined by the classical expression (Kimura 1962) which for a haploid genome takes the form

\[ P_{\text{fix}} = \frac{1 - e^{-ns^*}}{1 - e^{-Ns^*}}. \]

The curves in figure 2 are very similar to the simulations shown in figure 1, except at very low \( P_{\text{rec}} \) and very high deletion rate (cf., figs. 1A and 2A). With lower values of \( P_{\text{rec}} \), \( d_c \) increases and densely coded genomes are expected. When \( P_{\text{rec}} \) becomes too low, the critical coding ratio becomes less distinct (cf. figs. 1A and 2A). In this limit, mutants with very high deletion rates will switch from selected to counterselected at a coding ratio lower than the expected critical ratio. This probably results from the fact that \( m_{\text{ex}} \) is no longer accurately described by equation (6) for these mutants. When recombination is low and the coding density is high, it is reasonable to expect that \( m_{\text{ex}} \) will no longer be directly proportional to \( v_{\text{ex}} \) and that it will also be affected by the ratio between lethal and nonlethal deletions.

Accordingly, it is possible to estimate the probability of fixation for a single mutant once it has appeared (eq. 8 with \( n_0 = 1 \)). This is difficult to estimate accurately in simulations because the fixation probability is very small under most conditions. Nevertheless, setting \( n_0 \) to 1 in equation (8) yields equation (9), which is plotted in figure 3.

\[ P_{\text{fix}}^1 = \frac{1 - e^{-s^*}}{1 - e^{-Ns^*}}. \]
number also determines how quickly perturbations in genome size will relax to an ambient value.

**Genome Size Drift**

So far, the simulations describe the fate of changes in deletion rates. For any time that the relevant rates are constant, genome size will change under the combined influence of the prevailing rates of indels as given in equation (1). These rates correspond to proportional evolution where both additions and removals occur in proportion to the amount of noncoding material already present (Oliver et al. 2007). The average rate of change \( (R, \text{in base pairs per generation}) \) in genome size in this model is given by

\[
R = \frac{dL_{\text{non}}}{dt} = \frac{v_{\text{fix}}}{L_{\text{non}}} - \frac{v_{\text{del}}}{L_{\text{non}}}.
\]

Equilibrium obtains when on average, the increase and decrease of the genome size balance out, that is, when \( R = 0 \). This gives the expected amount of noncoding sequence

\[
L_{\text{eq}}^{\text{non}} = \frac{g_{\text{cod}}}{v_{\text{fix}}/C_0} - \frac{v_{\text{del}}L_{\text{non}}}{v_{\text{fix}}}. \tag{11}
\]

Here, the rates of fixation for an insertion (deletion) have been expressed as

\[
v_{\text{fix}} = v_{\text{ins}}N \frac{1 - e^{-s}}{1 - e^{-N s}} \quad \text{and} \quad v_{\text{del}} = v_{\text{del}}N \frac{1 - e^{-s}}{1 - e^{-N s}}. \tag{12a, b}
\]

These expressions are based on the assumption that individual events are fixed independently, which is not applicable when \( P_{\text{rec}} \) approaches zero while \( s \neq 0 \). Due to the relentless nature of the indel processes and the Hill–Robertson effect (McVean and Charlesworth 2000), the effective population size for fixation will be smaller than \( N \), with the consequence that the effect of selection is weakened. This is confirmed in our simulations.

The main result here is the general form of the equilibrium relation, where the equilibrium-coding ratio can be determined from

\[
d_{\text{eq}} = \frac{L_{\text{eq}}}{L_{\text{min}}} = \frac{L_{\text{eq}}}{(v_{\text{del}}/v_{\text{ins}})^{N_s} - 1). \tag{13}
\]

If the deletion bias is weak, that is, \( (v_{\text{del}}/v_{\text{ins}})^{N_s} < 1 \), no equilibrium is possible and the genomes will grow.

---

**Fig. 2.**—Predicted fixation probabilities calculated from equations (7a) and (8), plotted as a function of coding density \( d/(d + 1) \), using \( N = 1,000, n_0 = 200, v_{\text{del}} = 10^{-8} \), and \( L_{\text{cod}} = 2 \times 10^5 \). The four curves correspond to \( v_{\text{del}} \) values as indicated in the legends, and \( 10^{-8} \) from top to bottom on the left. Panel (A) is for \( d_c = 1 \), panel (B) for \( d_c = 0.1 \), and panel (C) is for \( d_c = 0.025 \), corresponding to the parameters used in figure 1A, B, and C, respectively.

**Fig. 3.**—Predicted fixation probability from equations (7a) and (9) with parameter values as in figure 2A.
without constraint. In contrast, with strong deletion bias, \( (v_{\text{del}}/v_{\text{ins}})e^{N_s} > 1 \), which leads to a coding sequence-dense genome. This probably corresponds to the situation for many microbes with large effective population sizes.

In contrast, a deletion balance with low coding density is possible only in a very narrow range where the deletion bias is larger than but very nearly equal to 1. For example, with \( L_{\text{cod}}/g_{\text{eq}} = 10^3 \) and \( d_{\text{eq}} = 0.1 \), equilibrium requires that the deletion bias equals 1.0001; here, a mere 0.01% decrease in the bias would lead to uncontrolled genome growth. Minute changes in the rates of indel will have enormous effects on the position of the equilibrium in this case. Furthermore, a stochastic formulation (Berg OG, unpublished data) of equation (10) leads to a size distribution that is a negative binomial with average as in equation (11) and variance

\[
\sigma_n^2 = L_{\text{non}}(1 + L_{\text{eq}}/g_{\text{cod}}).
\]

This corresponds to a very broad distribution, particularly for large \( L_{\text{non}} \). Thus, in this formulation, there is no well-defined equilibrium level for noncoding sequences in a genome with low coding density. That is to say, genome size may well drift randomly, as suggested by Oliver et al. (2007) from their study on eukaryotic genomes. However, large \( L_{\text{non}} \) and apparent random drift do require a deletion bias delicately poised just above the value 1. This “unpredictability” is only marginally stabilized when adaptation of deletion rates is included in the scheme, as will be shown next.

### Genome Size Drift Combined with Deletion Rate Adaptation

The dynamics of the system are determined by the relationship between the equilibrium ratio, \( d_{\text{eq}} \), and the critical ratio, \( d_c \). Here, \( d_c \) depends primarily on \( s/P_{\text{rec}} \) and is fairly independent of the other parameters. This is a fixed point around which everything moves. If the parameters are such that in equation (13) there is no equilibrium \( d_{\text{eq}} \) possible, that is, for weak deletion bias, the system will drift toward ever lower coding densities until the inequality \( L_{\text{cod}}/L_{\text{non}} < d_c \) is reached. At this point, mutations that increase the deletion rate will be favored. However, selection for an increased deletion rate will be effective only when \( Ns^* > 1 \) (eq. 7a), so that genome growth may continue until \( L_{\text{cod}}/L_{\text{non}} \approx d_c \), before deletion rate adaptation is likely to occur. In this way, the system can be pushed toward an equilibrium situation for which \( d_{\text{eq}} \approx d_c \). At this point, all changes in deletion rate are neutral but any change will again move \( d_{\text{eq}} \) away from \( d_c \). Although adaptation of the deletion rate can restore an equilibrium situation where \( d_{\text{eq}} \approx d_c \), it will also involve large fluctuations in noncoding sequence, particularly at low coding density.

Thus, small variations in parameter values can lead to huge variations in genome size. Because of this, a small value of \( d_c \) could push the deletion bias toward a value just above 1 as required (eq. 13) to achieve \( d_{\text{eq}} \approx d_c \). Such erratic behavior resembles the apparent random drift observed by Oliver et al. (2007) for the eukaryotic genomes.

In dense genomes where \( d_{\text{eq}} > d_c \), mutations that decrease the deletion rate are favored (fig. 1), but effectively only as long as \( v_{\text{del}}^{\text{eq}} L_{\text{cod}} (1 - d_c/d_{\text{eq}}) N > 1 \) (cf. eq. 7a and fig. 3); otherwise, they will also be nearly neutral. When \( v_{\text{del}} \) is so small that this inequality is violated, further decreases are not favored. In contrast, when \( v_{\text{del}}^{\text{eq}} L_{\text{cod}} N > 1 \), which is not unlikely, the neutral density where \( d_{\text{eq}} \approx d_c \) can be reached. Thus, \( d_c \) is an attractor for the coding density in genomes with sexual recombination. Deletion rate mutations that lead to an overshoot in coding density will make favorable those mutations that tend to restore the density to a neutral value of \( d_c \).

### Selection in Inverse Proportion to Genome Size

The results discussed above are based on the assumption that counterselection is the same for all insertions independently of how many are already present. A reasonable alternative is to assume that counterselection is proportional to the fraction of the total genome size to which the insertion corresponds. Thus,

\[
s = h/(L_{\text{cod}} + L_{\text{non}}),
\]

where \( h \) is a proportionality constant. In the simplest picture, \( h = 1 \) is just the length of the deletion. From equation (7a) it can be seen that selection to increase the deletion rate is positive when

\[
L_{\text{non}} > L_{\text{eq}}^2 h/P_{\text{rec}} L_{\text{cod}} > 0.
\]

Hence, there will be no selection for an increased deletion rate if \( h < P_{\text{rec}} L_{\text{cod}} \), in which case unhindered genome growth may ensue. If \( h > P_{\text{rec}} L_{\text{cod}} \), there can be positive selection for an increased deletion rate only if \( L_{\text{non}} \) is sufficiently large.

The critical coding ratio for which the deletion rate selection changes sign is

\[
d_c = \frac{h}{P_{\text{rec}} L_{\text{cod}}} - 1.
\]

Introducing a selection coefficient that scales inversely with the genome size can lead to a high coding ratio only if \( h \approx P_{\text{rec}} L_{\text{cod}} \). This condition may be difficult to satisfy except possibly for organisms with very little recombination and short coding regions (as e.g., bacterial symbionts). In this way, the critical coding ratio depends crucially on both the recombination intensity and the size of the coding region.

The coding density will drift toward a steady-state value \( d_{\text{eq}} \) (eq. 13) under the pressure of the prevalent rates of indels as well as selection (s) against maintaining junk DNA. Assuming that the selection is as given by equation (15), the stationary state level from equation (11) is determined by

\[
L_{\text{eq}}^{\text{non}} = \frac{v_{\text{del}}^{\text{eq}} L_{\text{cod}}}{v_{\text{ins}}^{\text{eq}} \exp(N h L_{\text{cod}}^{\text{eq}} L_{\text{eq}}^{\text{non}})} - 1.
\]

The selection against junk DNA becomes weaker as the amount increases (eq. 15). Therefore, there can be
a stable stationary state only for very small amounts of junk, and this is possible only when

\[
\frac{v_{\text{del}}}{v_{\text{ins}}} \exp(N \frac{h}{L_{\text{cod}}}) > 1.
\]  

(19)

If this expression is much larger than 1, a stable state can be reached with virtually zero amount of junk. If this expression is larger than, but close to 1, the stationary state in equation (18) would allow a large amount of junk, but this state is not stable if \( L_{\text{non}} \) increases. Thus, for the conditions in which equation (15) is valid, a stationary state with low coding density is not possible. Moreover, a stationary state with high coding density may not be stable either if decreases in deletion rate are possible. This is because decreasing deletion rates are under positive selection when \( d_{\text{eq}} > d_c \) and \( d_c \) is expected to be small (eq. 17).

**Discussion**

The model presented in this paper was developed to explore constraints on the evolution of deletion rates. A particular issue is whether or not adjustments of deletion rates can be selected to dispose of junk DNA. The impacts of selection intensity, recombination intensity, and lengths of coding and noncoding sequences were identified as the principal parameters constraining the evolution of deletion rates. The basic question is to what extent adaptation of deletion rates can control genome size in the large populations characteristic of Archaea and Bacteria as well as in the small populations characteristic of multicellular eukaryotes.

The existence of a critical coding ratio \( (d_c \approx s/P_{\text{rec}}, \text{eq. 7b}) \) shows that there is some potential for regulation, even in eukaryotes. However, the results also show that large fluctuations are expected, especially for genomes with low coding densities. For these genomes, regulation can occur only if the deletion rate is continually changing, which in turn requires that both rate-increasing and rate-decreasing mutations can occur at all times. If this is the case, the deletion rate will tend to increase if the actual coding ratio \( (d = L_{\text{cod}}/L_{\text{non}}) \) is \(<d_c\), and it will tend to decrease if \( >d_c\). However, in the absence of deletion rate mutants or if the effective selection \( (sN, \text{eq. 13}) \) against insertions is strong enough to prevent genome expansion, genome size cannot be regulated via the deletion rate.

**Behavior of the System**

The system responds to two selection coefficients, which are linked but must still be considered individually. One of these describes selection \( (s) \) for the removal of junk DNA, which depends on the cost of maintaining useless genetic material in the cell. The other describes selection for enhanced deletion rates \( (s' \text{ in eq. 7a}) \), which depends both on the magnitude of \( s \) and on the coding density of the genome. Both tendencies are at work here: First, there is the change in genome size determined by the prevailing frequencies of indel events and the selection \( s \). This drift could lead to an equilibrium coding density, \( d_{\text{eq}} \) of equation (13), if the deletion bias is sufficiently strong. Second, there is the appearance and fixation of deletion rate mutations determined by \( s' \), which will change the dynamics of genome growth and drive \( d_{\text{eq}} \) toward \( d_c \). However, at the critical coding ratio \( d_c \), effective selection is lost and it becomes more likely that a mutant which pushes the genome away from \( d_c \) will be fixed. Such a perturbation will reestablish selection for a deletion rate that restores the trend toward \( d_c \). Thus, the regulation will be imprecise and allow for large and erratic fluctuations in deletion rates, coding density, and genome size.

The significant result is that even if there is a strong selection against junk DNA, the genome will not necessarily approach a minimal size with little junk DNA. Instead, the conflicting requirements for an optimal deletion rate to remove junk balanced against the necessity to avoid excessive rates of lethal deletions, as in equation (5), can easily lead to genomes bloated with junk DNA. The fact that lethal mutations are not affected by recombination ensures that counterselection against deletions, and thus against increased deletion rate, will be strong in dense genomes, irrespective of the recombination rate.

The model shows that the selection for a deletion rate modifier will be more effective in systems with low rates of sexual recombination per genome replication. The critical coding ratio is large for small \( P_{\text{rec}} \) and, therefore, the genome size is expected to be smaller with a low frequency of sexual recombination. On the surface, this appears to contradict the general observation that genome size correlates inversely with recombination rate (e.g., Lynch 2006b). However, the recombination rate used in those comparisons corresponds to the probability of a crossover per base pair per genome replication. Furthermore, in those cases, the number of crossovers per chromosome is roughly the same (one or two), whereas chromosome number is independent of genome size, which explains the main trend in the observed correlations (Lynch 2006b).

In contrast, our model deals with the probability for intergenome exchange, \( P_{\text{rec}} \) per replication. In a sexual species, \( P_{\text{rec}} \approx 1 \), and in a strictly asexual (clonal) population, \( P_{\text{rec}} = 0 \). Prokaryotes that often exchange DNA through conjugation, transformation, or transduction will have intermediate values. For instance, if the recombination rate is \( c = 10^{-9} \) per bp per genome replication (Feil et al. 2001; Lynch 2006b), the genome size is \( 2 \times 10^9 \) bp and the number of crossovers per event is 2, then \( P_{\text{rec}} \approx 10^{-3} \). Because \( P_{\text{rec}} \) enters the model primarily through the critical coding ratio \( d_c \approx s/P_{\text{rec}} \), \( P_{\text{rec}} \ll s \) is required for selection of a deletion rate that effectively produces a high coding density genome.

**Coding Density in Various Groups of Organisms**

There are striking differences in coding density between different groups of organisms: from the generally small, streamlined genomes of Archaea and Bacteria to the bloated genomes of multicellular eukaryotes. Some of these differences have been pointed out above. Here, we describe three possible regimes for the evolution of
coding density that depend on which processes dominate genome size evolution. We note that genome size control through deletion rate adaptation could be at work in at least two of them.

(i) Multicellular eukaryotes have in general small values for $d_e$ because of frequent sexual recombination: $P_{rec} \approx 1$, and $s \ll 1$ gives $d_e \ll 1$. Furthermore, their effective population sizes ($N_e$) are generally small, on the order of $10^3$–$10^5$. Accordingly, the effective selection ($N_e s^2$) against genome expansion is weak and $N_e s^2 < 1$ is not unlikely. However, the selection in favor of a deletion rate mutation, as described by $N_e s^2$ from equation (7a), could still be strong, particularly in this group where the amount of coding sequence ($L_{cod}$) is large. Also, the selection for an increased deletion rate will become significant when the coding density has drifted sufficiently far below the critical $d_c$. Even if an equilibrium size with low coding density (eq. 13 or 18) can be found where $d_{eq} \approx d_e \ll 1$, huge fluctuations in genome size are expected. Although deletion rate adaptation in this way could keep genome size within the bounds determined by $d_e$, this constraint would be extremely soft in such organisms.

(ii) Free-living Archaea and Bacteria usually have very large effective population sizes, on the order of $10^8$–$10^9$, which can make the effective selection ($N_e s^2$) against genome expansion strong, even if the selection coefficient $s$ is very small. At the same time, with sufficiently frequent intergenome recombination in some of these organisms (Feil et al. 2001; Lynch 2006b), the critical ratio, $d_c \approx s P_{rec}$, is expected to be small unless the selection $s$ is unexpectedly large. Therefore, mutants with lower deletion rates would in principle be favored, thus allowing drift toward lower coding density. However, it may not be possible to decrease the deletion rate to the point where $d_{eq}$ approaches $d_c$. Thus, with $N_e s > 1$, it is necessary that $v_{del} << v_{ins}$ (see eq. 13 or 19) in order to decrease the coding density. Such low deletion rates may be mechanistically unattainable. Thus, coding density in this group would remain high even though $d_c$ cannot be determined by adaptation of the deletion rate. Indeed, the effective selection coefficient ($N_e s$) associated with different population sizes may be the main distinguishing character between bloated versus dense genomes in groups (i) and (ii) above (Lynch 2006a, 2006b).

(iii) Symbionts usually have effective population sizes that are smaller than their free-living cousins. Therefore, effective selection for a small genome may be weaker in such populations. However, they have much lower intergenome recombination rates, and therefore, the critical coding ratio is expected to be high. Thus, with the help of the strong linkage between deletion rate mutants and their effects on genomes in this group, coding density could be kept high through deletion rate adaptation. If so, the difference between groups (i) and (iii), which are both under weak selection due to relatively small effective population size, could be determined by the critical densities based on their very different recombination probabilities. However, if there is a total absence of intergenome recombination, the critical coding ratio is undefined and the genome dynamics will be dominated by Muller’s ratchet (Pettersson and Berg 2007).

If selection for a deletion of noncoding DNA is proportional to its size relative to the total genome (eq. 15), it becomes more difficult to constrain the genome size. There is no critical coding ratio if $P_{rec} L_{cod}$ is large (eq. 17) and, consequently, there will be no selection for an increased deletion rate even in very large genomes. This may be the case particularly for genomes in group (i) (eukaryotes) described above. Probably, it will also be the case in group (ii). However, due to the large effective population sizes in this second group, selection for high coding density may be sufficiently strong (eq. 19) to prevent genome bloating. The third group (symbionts) could still be subject to deletion rate adaptation because both $P_{rec}$ and $L_{cod}$ are expected to be small (eq. 17).

In addition to changes of deletion rates, there are of course potential effects of changes in insertion rates. For example, changes in sequence duplication rates or transposon invasion rates would influence the deletion bias and therefore also the genome size. Symbionts are shielded against foreign DNA uptake and transposon invasion, which may support high genome density. However, insertion rates and deletion rates are not reciprocal, and they do not follow the same dynamics. In our model, increasing $v_{ins}$ is always counterselected because some nonlethal insertions also lead to loss of fitness. But the insertion–deletion bias will determine the coding density when the rates are constant (eq. 13). For this reason, the insertion rates influence the dynamics of deletion rate mutants.

Oliver et al. (2007) suggest that genome size could be regulated without selection against larger genomes. They point to the fact that larger genomes are less stable than small ones because the absolute rate of change accelerates with increasing genome size. However, to achieve an essentially random, though upwards-constrained, drift in genome size would require a indel bias that is delicately poised at a value larger than but very nearly equal to 1 (eq. 13). Furthermore, in this limit, genome size would be unstable even with minute changes in the deletion bias. If eukaryotic genome sizes are indeed drifting randomly, it may be that their size distribution is determined and constrained by variations in the deletion bias. This could conceivably be achieved by deletion rate adaptation as described in our model.

The Selective Effect of Deletions

Although it is the central issue, the strength of selection acting on deletions is difficult to estimate. The most obvious benefit of deleting unneeded genetic material is the decrease in the amount of DNA that needs to be replicated. However, given the relative cheapness of DNA replication, the fitness gain from this is most likely minute for all but the most massive deletions. Thus, if there is significant selection in favor of deletions the reason must be sought elsewhere. One possibility is that selection is expressed through inactivation of open reading frames,
transposons, and other expressible genetic elements. By disrupting the expression of proteins that do not contribute (significantly) to fitness, a deletion might be selected more intensely than expected from its size as a fraction of the total genome. Similarly, destroying an active transposon could mean a significant contribution to the fitness of the organism. To some extent, these arguments are valid also for insertions, but due to their larger footprint on the genome, deletions are more likely to cause disruption. These arguments also provide some basis for not maintaining \( s \) proportional to the inverse of the genome length, as would be expected if the benefit was proportional to the actual reduction in DNA amount alone.

The model assumes that selective pressure is affecting all generations, not just sexual generations. For eukaryotes such as yeast, this is most likely the case, but for multicellular organisms, it may be more questionable. If there is little or no selection on the nonsexual generations, organisms will not behave uniformly throughout their life cycle as assumed in our model. Discontinuities in selection and counterselection will influence the effective deletion rate averaged over the life cycle of the organism’s germ line.

Our basic model uses a constant \( s \) value that is independent of the actual size of the genome. This should be considered as a “local” value in the region where changes are considered. That is, the addition of new noncoding material is assumed to be counterselected by \( s \), whereas the amount already present may be considered as settled, or de-activated, and therefore much less costly. Otherwise the total loss of fitness accrued in a bloated genome would be insupportable.

Mechanisms Involved in Determining the Deletion Rate

The mechanism to adapt genome length proposed here depends on the possibility to change the deletion rate, both up and down, by means of mutations. However, it is difficult to pinpoint which sorts of mutations are most effective in adapting the deletion rate. In *E. coli*, it has been shown that deletions occur via RecA-independent recombination and that the deletion rate can increase if the regulation of recombination is disrupted (Lovett et al. 1993; Bzymek and Lovett 2001). There is also evidence-linking replication to the occurrence of small deletions via slippage (Petrov 2002).

It is not obvious in which direction random mutations would tend to push the rates of indel occurrence. It may be that mechanistically, the deletion bias is likely to be larger than 1 and that random change in the rates therefore is likely to lead to constrained genome sizes even in the absence of selection.

Alternative Solutions and Extensions of the Model

The present model is fairly simplistic because it was designed to identify the most important aspects of the deletion rate dynamics. It could be extended in several ways. Although a more complicated model is not desirable in itself, there are relevant factors that are not covered by the present formulation. In particular, the recombination dynamics and selection model could be extended, in order to address additional issues.

The description of the recombination dynamics could be refined, by tracking each indel event individually. This would enable the precise determination of unique events that are eligible for transfer, when two genomes recombine. It would also make it possible to account for the boundary effects where a deletion spans from a noncoding into a coding region. Although such a detailed description is beyond the scope of this study, we have performed some simulations with an alternative recombination model that allows for each indel event to change genome location during a meiotic event. The results of these simulations show no significant difference in the fate of deletion rate mutants, although the accumulation of the actual indel mutations was affected. In other words, the details of the recombination description are of importance only when studying long-term changes in genome size. We also tested a model in which each recombination event completely separated the deletion allele from all its associated deletions, but here too, we obtained no significant differences in outcome.

The selection model could be made to include variable selection coefficients for indel events, both by introducing a distribution of lengths of indel events and by making the selection coefficient scale proportionally to the ambient genome size (eq. 15). We carried out some preliminary studies where indel events were followed individually and each was assigned an \( s \) value sampled from a normal distribution with mean \( \bar{s} \) and variance \( \sigma^2 \). This produced an outcome very similar to the one obtained when all deletions use \( s = \bar{s} \).

Frequency and Size

Increased frequency means increased risk of lethal deletion. If the strength of selection is related to the size of the deletion, long, but rare, deletions will be strongly favored over short, but frequent, ones. This tendency is amplified by recombination because the increased risk of lethal mutations cannot be mitigated by moving the deleterious deletion to another cell. The obvious conclusion seems to be that mutations conferring increased deletion rates in the form of very frequent but small deletions will have a lower critical coding ratio, due to the lower selective benefit of the individual deletions.

It is expected that organisms with large fractions of essential genes should be constrained to selectively minimize the number of deletion events because the higher deletion rates increase the probabilities of lethal deletions. Given that bacterial genomes are in general densely packed with essential sequences, they should also experience bias toward larger average size of deletions. Equation (7a) clearly indicates that large deletions are preferable to small ones, given that the value of \( s \) is proportional to deletion size.

Conclusions

The calculations presented above suggest that small genomes may be constrained primarily by selection against junk DNA, which is strengthened by large effective
population sizes and low recombination rates. In addition, a small size of the coding genome also facilitates high deletion rates. Large (bloated) genomes are associated with small effective population sizes, where selection is much weaker. However, the model suggests that the sizes of bloated genomes still may be constrained by deletion rate adaptation. However, for these genomes, we expect very large fluctuations that set in when the genomes become excessively large.

Acknowledgments

This work was funded by the Swedish Research Council. Institution where the research was performed: Department of Molecular Evolution, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

Literature Cited


Dan Graur, Associate Editor

Accepted March 13, 2009