# 2 Genome Size and Organismal Complexity

**When the key insights of Crick, Franklin, Watson, and Wilkins** led to the elucidation of the structure of DNA, half a century ago, little was known about the molecular aspects of genome structure. We now know that chromosomes, the vehicles of DNA, are enormously versatile in terms of content and sequence malleability. Recurrent mutation generally ensures that most homologous chromosomes within a species are unique in multiple ways, and this variation provides the fuel for evolutionary divergence among species, as revealed in striking detail by whole-genome comparisons. For example, the 250 or so fully sequenced prokaryotic genomes contain between 350 and 8000 genes packed into 0.5–9.0 megabases (Mb), while all well-characterized genomes of animals and land plants contain more than 13,000 genes and are at least 100 Mb in size. As will be detailed in Chapter 3, most of the increase in genome size in multicellular eukaryotes is a consequence of the expansion of noncoding forms of DNA, including introns and mobile elements. The phylogenetic positions of animals and land plants suggest the independent emergence of this complex genomic syndrome in both lineages (Meyerowitz 2002), but what is cause and what is effect?

Across the entire domain of life, there is a moderate positive scaling between organism size and number of cell types (Bell and Mooers 1997; Bonner 2004). However, although specific genes play a central role in cellular differentiation, there is little evidence that a substantial increase in genome size or gene number is essential for the evolution of multicellularity. For example, numerous cyanobacteria (Meeks et al. 2002), myxobacteria (Goldman et al. 2006), streptomycetes (Bentley et al. 2002), methanogens (Galagan et al. 2002), and other prokaryotic lineages are capable of producing multiple cell types, despite having moderate numbers of genes (4000–8000) and relatively little noncoding DNA. A number of eukaryotes with complex cell structures and multiple cell types harbor 10,000 or fewer genes, whereas the genomes of some unicellular eukaryotes (e.g., *Paramecium*) harbor more genes than those of vertebrates.

This weak relationship between gene number and organismal complexity suggests that the increased structural innovation and developmental flexibility of the eukaryotic cell must largely be a consequence of the unique ways in which genes are deployed. But were new ways of expressing genes (such as complex spatial and temporal patterns of transcription regulation and alternative splicing) promoted as a direct response to selection for new cell types in large organisms? Or did the evolution of large size and/or multicellularity induce side effects that provoked nonadaptive changes in genomic architecture, which then secondarily paved the way for the adaptive origin of new cell functions? In the following chapters, the case will be made that the roots of many aspects of eukaryotic genomic complexity are likely to reside in nonadaptive processes (in particular, mutation pressure and random genetic drift) that are particularly potent in eukaryotes, especially in multicellular lineages. This chapter reviews some of the historical background leading up to this argument.

First, a broad phylogenetic survey will demonstrate the continuity of scaling of genome content with genome size across the transitions from prokaryotes to unicellular eukaryotes to multicellular species. This observation leads to the conclusion that aspects of cell structure and metabolism are not the central determinants of genomic architecture. Second, previous hypotheses for the evolution of genome size will be evaluated in this context and their limitations outlined. Although it is commonly argued that microbial genomes are kept streamlined by efficient selection against the negative metabolic costs of replicating excess DNA, there appear to be no data in support of this contention. Nor is there compelling evidence that an intrinsic bias toward deletion mutations is sufficient to prevent runaway genome growth. Finally, a brief verbal description of the mutational hazards of excess DNA will be given, setting the theme for many of the topics to be considered in subsequent chapters.

## Genome Size and Complexity

Many aspects of genomic architecture exhibit continuous transitions within and across all cellular domains of life, extending even to DNA viruses (Figure 2.1). In viruses and prokaryotes, the amount of coding DNA scales nearly linearly with total genome size, occupying 80%–95% of the latter, and a similar allocation is found in the smallest eukaryotic genomes. However, the expansion of coding DNA progressively slows in genomes with total sizes in excess of 10 Mb, eventually leveling off at about 100 Mb in vertebrates and land plants, in which 90%–98% of the genome is allocated to noncoding DNA. As a consequence, over the 10,000-fold range in total genome sizes for well-studied cellular species, there is only a 100-fold range in the amount of DNA devoted to protein coding.

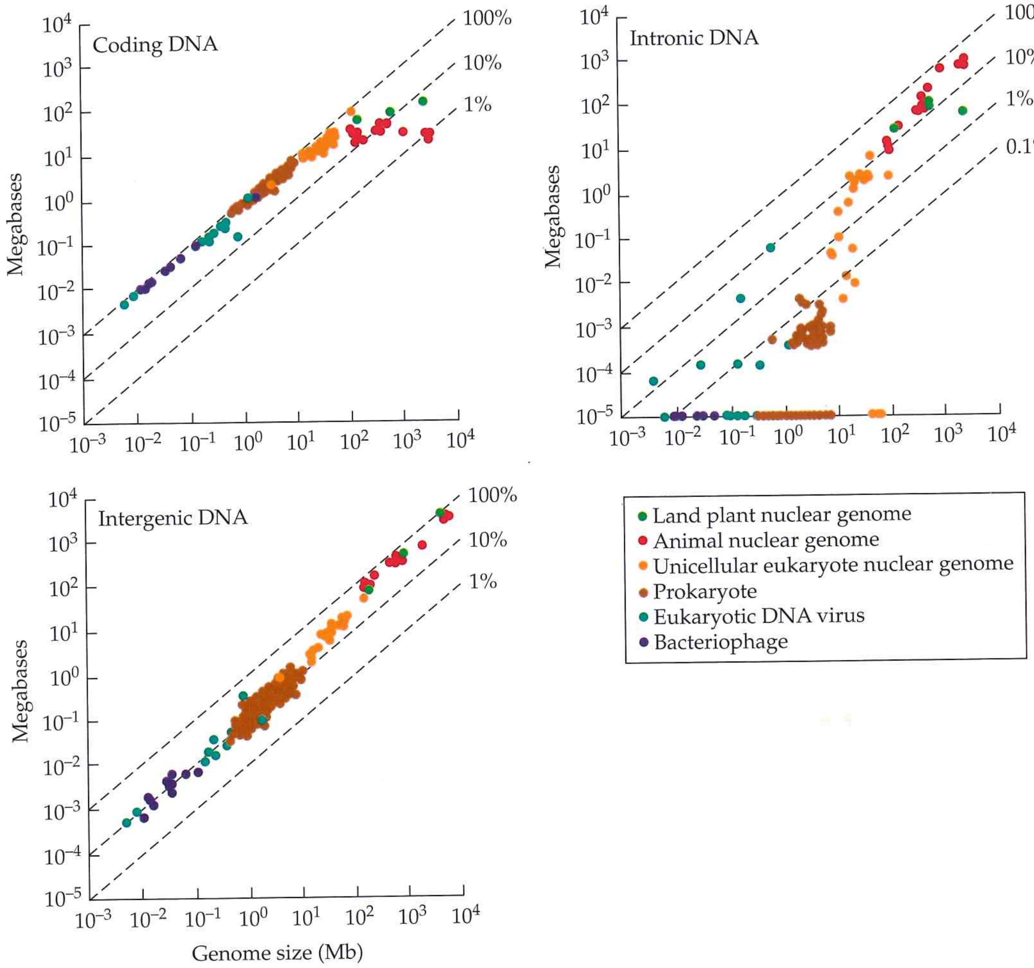

**Figure 2.1** The contributions of protein-coding, intronic, and intergenic DNA to total genome size in various organismal groups. The diagonal lines define points of equal proportional contribution to total genome size. The intronic DNA depicted here does not include introns in noncoding exons (UTRs). Intergenic DNA comprises all of the genome exclusive of spans between translation initiation and termination points for protein-coding genes. Points residing on the *x* axis denote situations in which the component contribution is zero. (Modified from Lynch 2006b.)

The two most easily identified classes of noncoding DNA, introns and mobile elements, scale similarly with genome size. Although spliceosomal introns are confined to eukaryotes, prokaryotes sometimes harbor small numbers of self-splicing introns (see Chapter 9), but never at levels exceeding 0.2% of the entire genome (see Figure 2.1). The smallest eukaryotic genomes also contain very little intronic DNA, but above total genome sizes of 10–20 Mb, there is an abrupt and progressive increase in investment in introns. Genomes that are about 100 Mb in size (all of which are eukaryotes) have nearly equal amounts of DNA allocated to introns and exons, whereas about 95% of the total length of protein-coding genes is intronic in the large (>2500 Mb) mammalian genomes. A qualitatively similar transition is seen with the fraction of the genome occupied by intergenic DNA: typically less than 20% for genomes smaller than 1 Mb and progressively increasing to more than 80% for genomes beyond 10 Mb. Residing within intergenic regions are sequences involved in transcription, chromatin packaging, and replication initiation (see Chapters 3, 5, and 10), but in species with large genomes, the majority of intergenic DNA consists of active mobile elements (transposons and retrotransposons) and other debris associated with their past activities.

Two general conclusions emerge from the enormous phylogenetic breadth of the patterns in Figure 2.1. First, the common assertion that there is essentially no correlation between genome size and organismal complexity (Thomas 1971; Cavalier-Smith 1978; Gregory 2005a), appears to derive from a focus on extreme outliers rather than on measures of central tendency. Although there is considerable variation in genomic features among species with similar levels of cellular/organismal complexity, there is a clear ranking from viruses to prokaryotes to unicellular eukaryotes to multicellular eukaryotes in terms of genome size, gene number, mobile element number, intron number and size, size of intergenic spacer DNA, and complexity of regulatory regions (Lynch and Conery 2003b; Lynch 2006a). Second, despite this gradient, there are no abrupt discontinuities in the scaling of genome content with genome size across radically different groups of organisms. This smooth transition in patterns of genome content scaling across all forms of life provide compelling evidence that the primary forces influencing the evolution of genomic architecture are unlikely to be direct consequences of organismal differences in cell structures or physiologies.

## The Selfish-DNA and Bulk-DNA Hypotheses

The early idea that genome sizes vary wildly among organisms with similar levels of cellular and developmental complexity became known as the C-value paradox (where the C value denotes the total amount of DNA in a haploid genome). Depending on one's point of view, the puzzle was either solved or deepened as it became clear that a substantial fraction of many eukaryotic genomes consists of noncoding and putatively nonfunctional DNA. Two general classes of hypotheses emerged to explain this odd set of observations.

On the one hand, Doolittle and Sapienza (1980) and Orgel and Crick (1980) promoted the idea that a good deal of noncoding DNA consists of "selfish" elements capable of proliferating until the cost to host fitness becomes so prohibitive that natural selection prevents their further spread.

of 10–20 Mb, there is an abrupt and progressive increase in investment in introns. Genomes that are about 100 Mb in size (all of which are eukaryotes) have nearly equal amounts of DNA allocated to introns and exons, whereas about 95% of the total length of protein-coding genes is intronic in the large (>2500 Mb) mammalian genomes. A qualitatively similar transition is seen with the fraction of the genome occupied by intergenic DNA: typically less than 20% for genomes smaller than 1 Mb and progressively increasing to more than 80% for genomes beyond 10 Mb. Residing within intergenic regions are sequences involved in transcription, chromatin packaging, and replication initiation (see Chapters 3, 5, and 10), but in species with large genomes, the majority of intergenic DNA consists of active mobile elements (transposons and retrotransposons) and other debris associated with their past activities.

Two general conclusions emerge from the enormous phylogenetic breadth of the patterns in Figure 2.1. First, the common assertion that there is essentially no correlation between genome size and organismal complexity (Thomas 1971; Cavalier-Smith 1978; Gregory 2005a), appears to derive from a focus on extreme outliers rather than on measures of central tendency. Although there is considerable variation in genomic features among species with similar levels of cellular/organismal complexity, there is a clear ranking from viruses to prokaryotes to unicellular eukaryotes to multicellular eukaryotes in terms of genome size, gene number, mobile element number, intron number and size, size of intergenic spacer DNA, and complexity of regulatory regions (Lynch and Conery 2003b; Lynch 2006a). Second, despite this gradient, there are no abrupt discontinuities in the scaling of genome content with genome size across radically different groups of organisms. This smooth transition in patterns of genome content scaling across all forms of life provide compelling evidence that the primary forces influencing the evolution of genomic architecture are unlikely to be direct consequences of organismal differences in cell structures or physiologies.

## The Selfish-DNA and Bulk-DNA Hypotheses

The early idea that genome sizes vary wildly among organisms with similar levels of cellular and developmental complexity became known as the C-value paradox (where the C value denotes the total amount of DNA in a haploid genome). Depending on one's point of view, the puzzle was either solved or deepened as it became clear that a substantial fraction of many eukaryotic genomes consists of noncoding and putatively nonfunctional DNA. Two general classes of hypotheses emerged to explain this odd set of observations.

On the one hand, Doolittle and Sapienza (1980) and Orgel and Crick (1980) promoted the idea that a good deal of noncoding DNA consists of "selfish" elements capable of proliferating until the cost to host fitness becomes so prohibitive that natural selection prevents their further spread.

This selfish-DNA hypothesis, under which genome size expansion is a simple pathological response to internal genomic upheaval, draws support from the ubiquity of mobile elements across the eukaryotic domain (see Chapter 7). However, other major contributors to genome size, such as spliceosomal introns, small repetitive DNAs, and random insertions, are not self-replicable and hence not subject to selection for proliferative ability within host genomes. Thus, a central challenge for the selfish-DNA hypothesis is the need to explain why *all* types of excess DNA mutually expand (or contract) in some genomes and not in others.

In striking contrast to the view that much of noncoding DNA is expendable junk, Commoner (1964), Bennett (1972), and Cavalier-Smith (1978) had argued earlier that the total content of the noncoding DNA within a genome (independent of its information content) is a direct product of natural selection. This bulk-DNA hypothesis postulates that genome size has a direct effect on nuclear volume, cell size, and cell division rate, all of which in turn influence life history features such as developmental rate and size at maturity. The supporters of this hypothesis have pointed out an impressive number of correlations between genome size and cell properties in a diversity of phylogenetic groups (Figure 2.2 contains two examples), although the evolutionary mechanisms responsible for such statistical relationships are unclear. Cavalier-Smith (1978) suggested that the evolution of large cell size imposes secondary selection on nuclear genome size as a physical mechanism for modulating the area of the nuclear envelope and hence regulat-

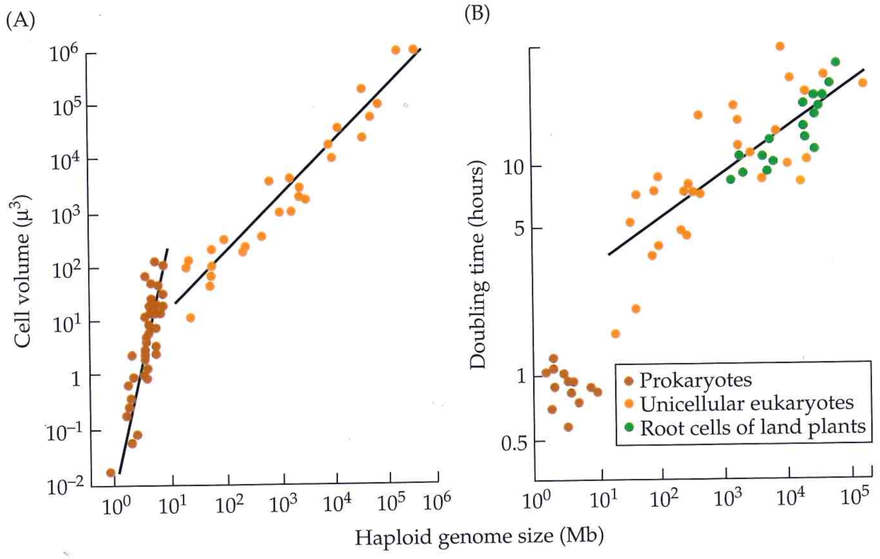

**Figure 2.2** Correlations of two cell biological features, (A) cell volume and (B) doubling time, with genome size. Doubling times are measured for cells at 23°C. (Modified from Shuter et al. 1983.)

ing the flow of transcripts to the cytoplasm. One concern with this view is that many additional mechanisms for achieving elevated transcript numbers exist (e.g., increases in nuclear membrane porosity, gene copy number, ribosome number, and transcript longevity), all of which appear to be less risky than the expansion of bulk DNA. Indeed, the most striking challenge for any adaptive hypothesis for the expansion of excess DNA is the nature of the filler material itself, predominantly mobile elements, which are known to impose a heavy mutational burden in eukaryotes (see Chapter 7). Remarkably, there is also a strong positive correlation between cell size and genome size in prokaryotes (see Figure 2.2), which cannot be a consequence of either cytoskeletal effects (given the absence of nuclear membranes) or of noncoding DNA expansion (given its near constant proportion; see Figure 2.1).

Because evolution is a population-level process, any evolutionary explanation for a pattern of variation must be consistent with basic population genetic mechanisms (e.g., mutation, random genetic drift, recombination, and natural selection), but the failure of prior studies to directly confront these issues in a quantitative manner has been a major impediment to sorting out cause and effect in genome size evolution. For example, the logic underlying the bulk-DNA hypothesis will remain unconvincing until it is demonstrated that: (1) heritable within-population variation in genome size significantly covaries with cellular features that are mechanistically associated with individual fitness, and (2) mobile element proliferation is an easy means of achieving such variation with minimal negative side effects. The absence of population-level thinking from much of the ongoing debate about genome size evolution has fostered the impression that unknown evolutionary mechanisms remain to be discovered, leading some to invoke undefined "macroevolutionary" phenomena (Gregory 2005c). However, the logical problems with arguments that abandon established microevolutionary principles are well known (e.g., Charlesworth et al. 1980), and a central goal of this book is to demonstrate that there are very few, if any, aspects of genomic evolution that cannot be explained with well-accepted population genetic mechanisms.

## The Metabolic Cost of DNA

Microbial species pose a special challenge for both the bulk-DNA and selfish-DNA hypotheses. With its adherence to adaptive arguments, the bulk-DNA hypothesis invokes a premium on energetic efficiency as an explanation for the diminutive genomes of prokaryotes (Cavalier-Smith 2005), whereas adherents to the selfish-DNA hypothesis have argued that small genomes are products of strong selection for high replication rates (Doolittle and Sapienza 1980; Orgel and Crick 1980; Pagel and Johnstone 1992). With both competing hypotheses conceptually aligned on at least this one matter, the metabolic expense of DNA is widely cited as the explanation for the streamlining of microbial genomes (e.g., Rogozin et al. 2002; Giovannoni et al. 2005; Ranea et al. 2005).

Is the cost of maintaining and replicating an additional DNA segment of a few base pairs (the typical size of an intergenic insertion/deletion; see below) significant enough to be perceived by natural selection? Because the large population sizes of unicellular species magnify the efficiency of natural selection (see Chapter 4), this possibility cannot be ruled out entirely. However, there is no direct evidence that cell replication is ever limited by DNA metabolism, and there are several reasons to think otherwise. First, within and among prokaryotic species, there is no correlation between cell division rate and genome size (Bergthorsson and Ochman 1998; Mira et al. 2001). Second, during rapid growth phases, prokaryotic chromosomes are often present in a nested series of replication stages (Casjens 1998), with some species harboring tens to hundreds of chromosomal copies at various stages of the life cycle (e.g., Maldonado et al. 1994; Komaki and Ishikawa 2000). Third, in *E. coli* and other eubacteria, DNA replication forks progress 10–20 times faster than mRNA elongation rates (Bremer and Dennis 1996; Cox 2004; French 1992). Fourth, DNA constitutes 2%–5% of the total dry weight of a typical prokaryotic cell (Cox 2003, 2004), and the estimated cost of genomic replication relative to a cell's entire energy budget is even smaller (Ingraham et al. 1983). Similar conclusions emerge for eukaryotic cells (Rolfe and Brown 1997).

## Directional Mutation Pressures on Genome Size

Genome size evolution ultimately depends on two factors: the relative rates of mutational production of insertions and deletions, and the ability of natural selection to promote or eliminate such changes. Thus, if the energetic consequences of noncoding DNA are not great enough to be perceived by natural selection, species with small genomes must be subject to unusual deletional mutation pressures, and/or excess DNA must be disadvantageous in some other way.

Several types of mutational activity encourage genome size expansion. For example, mobile elements are capable of self-replicating and inserting copies elsewhere in the genome at high rates (in excess of $10^{-5}$ per element per generation; see Chapter 7), and their activities also result in the insertion of pseudogenes (dead-on-arrival copies of otherwise normally functioning genes) (see Chapter 3). In addition, segmental duplications involving stretches of hundreds to thousands of kilobases (kb) are universal among eukaryotes (see Chapter 8), and strand slippage during replication can also lead to small-scale insertions (Chen et al. 2005).

Double-strand breaks of chromosomes are another common source of insertions and deletions. Such breaks occur spontaneously in nonreplicating cells and are also produced when a replication fork encounters a single-strand nick, severing the entire chromosome. In mammals, 5%–10% of somatic cells acquire at least one double-strand break per cell division

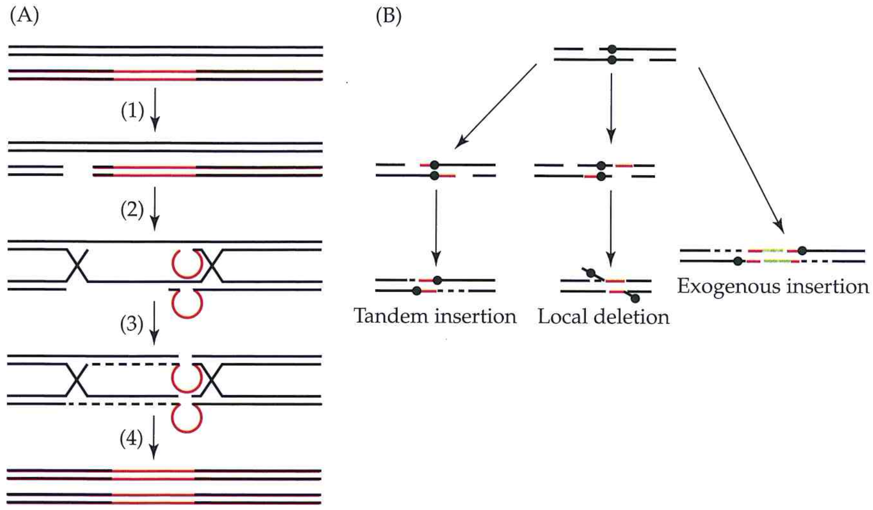

**Figure 2.3** Some ways in which insertions and deletions can be created at double-strand breaks. (A) Repair by homologous recombination can result in a local conversion of one chromosome type to the other. (1) A break appears in one chromatid adjacent to an insertion (red) that is absent from the homologous chromosome. (2) The free ends of the lower chromosome invade the upper chromosome to initiate formation of a recombination junction. Complementary DNA from the two chromatids aligns, leaving loops in each recombination intermediate. (3) This mismatch is resolved by cutting the non-looped strands. (4) Synthesis of the missing complementary DNA results in the conversion of the top chromatid to the insertion-bearing form. (If the loop-containing strands were cut instead, the invading strands could be converted to the insertion-free form). (B) Repair by nonhomologous end joining proceeds in the absence of a homologous chromosome. Such repair requires small regions of microhomology (illustrated in red), which are relatively free to align at multiple sites. Black dots serve as points of reference, and the starting point is a staggered cut. On the left and in the center, the selected regions of homology on the two strands (red) determine whether there is a local duplication or deletion. Occasionally, double-strand breaks capture foreign DNA (green lines), as shown on the far right.

(Lieber et al. 2003), and germ line cells that have experienced multiple divisions in the progression toward gamete production can be expected to incur even more. To maintain cell lineage viability, such breaks must be stitched back together by one of two mechanisms.

First, if a homologue (in a diploid species) or a sister chromosome (in the early stages of mitosis or meiosis) is available as a template, homologous recombination can restore the original state of the broken chromosome, provided the regions involved have near complete sequence identity. However, if in the region of the break, one of the chromosomes happens to have an insertion that is absent from the partner chromosome, recombination can alter the state of one of the chromosomes to that of the other by the process of gene conversion (Figure 2.3A). Some evidence suggests that insertions are retained more often than lost (Lamb 1985), although the degree to which this is the case depends on the size and structure of the insertion (Bill et al. 2001). Such biased gene conversion, which is purely a physical process, can encourage the spread of insertions throughout a population in a manner that is indistinguishable from positive selection (Nagylaki 1983; Walsh 1983).

Second, in nondividing cells, where templates are less likely to be available, the error-prone process of nonhomologous end joining, which directly ligates the two edges of a break back together, must be relied on (Moore and Haber 1996a; Heidenreich et al. 2003; Daley et al. 2005; Puchta 2005). Nonhomologous end joining is initiated by complementary base pairing in regions of microhomology (at least 2–3 bp), and the way in which this is done can lead to small insertions or deletions (Figure 2.3B). Double-strand break repairs can also be accompanied by the capture of exogenous DNA derived from the mitochondrial genome, retrotransposons, or microsatellites (small repetitive DNAs, such as dinucleotide repeats) (Moore and Haber 1996b; Teng et al. 1996; Ricchetti et al. 1999; Yu and Gabriel 1999; Lin and Waldman 2001b; Decottignies 2005).

With large-scale insertion events operating on a recurrent basis, the prevention of runaway genome expansion requires direct selection to prevent the fixation of insertions at the population level and/or mutational mechanisms for their subsequent deletion. To evaluate whether deletion mutations alone are capable of putting a cap on genome size without any assistance from selection, Petrov and colleagues (Petrov 2001, 2002a,b; Petrov et al. 1996, 2000) have performed comparative surveys of the numbers and sizes of insertions and deletions in various types of pseudogenes in insects. Their studies and others (Table 2.1) suggest that the rate of small-scale nucleotide losses exceeds that of gains, yielding a net erosion in the length of large inserts of nonfunctional DNA over time. Taken at face value, these data imply a half-life of nonfunctional DNA in the nematode *Caenorhabditis* and the fruit fly *Drosophila* on the order of the time required for neutral DNA to acquire 0.15–0.25 substitutions per site (10 million years or so), whereas that for orthopterans (grasshoppers and crickets), mammals, and birds (all of whose genomes are much larger) is 15–50 times longer (see Table 2.1). Mira et al. (2001) also document substantially higher rates of small-scale nucleotide losses than gains from pseudogenes in a variety of prokaryotes. However, a dramatically different picture emerges in rice (*Oryza*), where the rate of nucleotide gain by pseudogenes exceeds that of loss by a factor of 16 (Noutsos et al. 2005).

These kinds of observations have encouraged the view that interspecific variation in the mutational tendency to delete excess DNA is a primary determinant of genome size, with species with the highest rates of deletion having the smallest genome sizes (Petrov et al. 2000; Mira et al. 2001; Ochman and Davalos 2006). However, a number of uncertainties remain. Why, for example, should insertion/deletion rates differ so dramatically among animals, given the high degree of conservation of their DNA repair

**TABLE 2.1 Rates and average sizes of deletions and insertions derived from observations of nonfunctional DNA in various animals**

| | RATE | | SIZE (BP) | | NET | HALF- |
|---|---|---|---|---|---|---|
| | DELETION | INSERTION | DELETION | INSERTION | CHANGE | LIFE |
| *Caenorhabditis* | 0.034 | 0.019 | 166 | 151 | –2.8 | 0.25 |
| *Drosophila* | 0.115 | 0.028 | 42 | 12 | –4.5 | 0.15 |
| *Laupala* | 0.070 | 0.020 | 7 | 7 | 0.0 | — |
| *Podisma* | 0.060 | 0.030 | 2 | 1 | –0.1 | 6.93 |
| Birds | 0.043 | 0.007 | 12 | 4 | –0.3 | 2.31 |
| Mammals | 0.033 | 0.017 | 5 | 6 | –0.1 | 6.93 |

*Sources*: *C. elegans*, Witherspoon and Robertson 2003; *Drosophila*, average from Blumenstiel et al. 2002 and Petrov 2002b; *Laupala* (Hawaiian cricket) and *Podisma* (grasshopper), Petrov 2002b; birds (pigeons and doves), Johnson 2004; mammals (mouse, rat, and human), average from Ophir and Graur 1997 and Zhang and Gerstein 2003.

*Note*: All rates are given relative to the time required for the accumulation of one nucleotide substitution per silent site. Net change is defined to be the difference between (insertion rate × size) and (deletion rate × size), so, for example, –2.8 for *C. elegans* implies that by the time an average surviving nucleotide site has acquired a single substitution, an average net loss of 2.8 nucleotides per site is expected to occur. Half-life is the number of substitutions per silent site that are expected to accrue by the time a nonfunctional stretch of DNA experiences a 50% erosion in length, assuming exponential decay.

machinery (Eisen and Hanawalt 1999)? And if species with small genomes have evolved increased deletion rates, as Lawrence et al. (2001) have suggested as an adaptive mechanism for the streamlining of prokaryotic genomes, how is the increased burden on coding DNA avoided?

Central to these issues is the matter of whether the long-term behavior of pseudogenes provides an unbiased view of the de novo mutation spectrum or whether deletions and insertions in pseudogenes are subject to selection (Charlesworth 1996a). This is a concern because insertion-associated disadvantages and/or deletion-associated benefits will tilt the observed spectrum of effects toward deletions relative to the mutational distribution, in which case the negative association between observed net deletion rates and genome size could simply reflect interspecific variation in the efficiency of selection rather than intrinsic differences in mutational properties. As noted above, it is unclear whether the energetic advantages of small deletions are ever substantial enough to cause perceptible fitness differences, but as will be discussed in the following section, excess DNA can impose additional disadvantages.

Evidence that deletions may not outnumber insertions at the mutational level derives from observed excesses of insertions over deletions in several laboratory experiments. In *Drosophila melanogaster*, spontaneous insertions greater than 4 kb in length are fourfold more abundant than deletions (Yang et al. 2001), and reporter construct experiments in the yeast *Saccharomyces cerevisiae* suggest a similar insertion/deletion disparity (Kunz et al. 1998; Ohnishi et al. 2004; Hawk et al. 2005). However, although these direct assays imply an innate mutational tendency for genome size *expansion*, above and beyond that caused by mobile element activity and segmental duplications, such a bias may not be universal. For example, estimates of the human mutational spectrum derived from de novo mutations for genetic disorders suggest that microdeletions are 2.5 times more common than microinsertions, with both exhibiting very similar size distributions (Kondrashov 2003; Ball et al. 2005). Studies involving reporter constructs in *E. coli* also reveal a deletion bias (Schaaper and Dunn 1991; Sargentini and Smith 1994).

In principle, these indirect assays could be biased if deletions and insertions are not equally likely to produce a detectable phenotype, and the only truly unambiguous way to ascertain the insertion/deletion spectrum is to randomly sequence genomic regions after a long period of complete relaxation of selection. Such a study has been performed with the nematode *Caenorhabditis elegans* by using long-term mutation accumulation lines taken through single-individual bottlenecks each generation to eliminate the effectiveness of natural selection against all mutations except those causing complete lethality or sterility (Denver et al. 2004). This study revealed a 15:4 insertion/deletion ratio (both types of mutations were of similar size, and none were associated with mobile element activity), a dramatically different pattern from the 1:1 ratio derived from phylogenetic analysis (see Table 2.1).

Thus, despite the clear need for more data of the type procured for *C. elegans*, these observations, along with the enormous half-life estimates in Table 2.1, raise significant questions as to whether mutational deletion processes are *universally* sufficient to prevent the runaway growth of genome size. If they are not, then some form of natural selection is necessary for genome size stabilization, and lineages with a reduced ability to selectively promote deletions and/or purge insertions can be expected to experience nonadaptive expansions in genome size. Bennetzen and Kellogg (1997) refer to species in this kind of evolutionary situation as having acquired a "one-way ticket to genome obesity." However, arguments presented in the following chapters suggest that genomic expansion and contraction is really a two-way street, with the prevailing direction of traffic depending on the current population genetic conditions. A key question that remains to be resolved is whether the large genomes of multicellular eukaryotes are still in active phases of expansion.

## Population Size and the Mutational Hazard of Excess DNA

Although DNA without a function is often assumed to be neutral, this view ignores a fundamental genetic observation: that the operation of every gene depends on its local physical environment. Thus, even if inert spacer DNA is immune to selection against loss-of-function mutations—i.e., is totally expendable—it need not be immune to harmful gain-of-function mutations. Many lines of evidence support this view. First, noncoding regions are known to be depauperate in short motifs with the potential for generating inappropriate transcription factor binding (Hahn et al. 2003), posttranscriptional silencing (Farh et al. 2005), and translation initiation (Rogozin et al. 2001; Lynch et al. 2005). Selection against mutations causing inappropriate gene expression is the likely cause of the maintenance of such sequences below levels expected by chance. A dramatic example of this point is a human blood disorder in which a single nucleotide substitution creates a novel regulatory element in an otherwise inert segment of intergenic DNA (De Gobbi et al. 2006). Second, insertions of mobile elements into coding exons will virtually always inactivate a gene, whereas those in noncoding regions can influence the regulation of adjacent genes (Sorek et al. 2002; Lev-Maor et al. 2003; Kreahling and Graveley 2004; Shankar et al. 2004). Third, introns are a mutational burden for their host genes, as the splicing of each intron requires a specific set of local sequences for proper spliceosome recognition (Lynch 2002b). Fourth, the fact that the majority of eukaryotic genomic DNA may be transcribed (Cawley et al. 2004; Kampa et al. 2004), at least at low levels, raises the question as to whether any segment of nonfunctional DNA is truly neutral.

All of these issues will be explored in further detail in the following chapters. The central point to be understood here is that a primary cost of excess DNA is its mutational liability (Lynch 2002b, 2006a; Lynch and Conery 2003b). Each embellishment of the structure of a gene or of its surrounding area increases the risk that the gene will be rendered defective by subsequent mutational processes.

This matter becomes important in the context of comparative genomics because the mutational burden associated with most excess DNA is quite small, but not so small as to be effectively neutral in all phylogenetic contexts. A key theme that will appear repeatedly in the following pages (particularly in Chapter 4) is that population size is a central determinant of the efficiency of natural selection: by magnifying the power of random genetic drift, fluctuations in allele frequencies caused by small population size can overwhelm the ability of natural selection to influence the dynamics of mutations of small effect. A second key point is that although random genetic drift is often viewed as simple noise that causes variation in evolutionary outcomes around expectations under selection alone, this is a false caricature. The size of a population specifically defines the kinds of genomic evolution that can and cannot proceed, with small population size facilitating the accumulation of deleterious mutations and inhibiting the promotion of beneficial changes. Finally, the tendency for mutationally hazardous DNA to accumulate depends on both the population size and the mutation rate: the latter defines the burden of excess DNA, while the former defines the ability of natural selection to eradicate it. These simple ideas provide a potentially unifying explanation for a wide range of observations on phylogenetic variation in gene structure and genomic composition.

The basic argument can be summarized with the following example. Suppose that natural selection favors an increase in body size in a particular lineage. A general observation from population ecology is that an increase in body size results in a reduction in the number of individuals within a species (see Chapter 4). By reducing the efficiency of natural selection, diminished population size magnifies the tendency for mildly deleterious insertions to accumulate in a genome, while also reducing the ability of selection to promote advantageous deletions. Thus, genome size is expected to expand in organisms of increasing body size, not necessarily because of an intrinsic tolerance of excess DNA or because of an increased need for bulk DNA, but because of a reduced ability to eradicate it. In contrast, genome size contraction can be expected in organisms selected for small size, again not because of direct selection for rapid genome replication, but because purifying selection more efficiently eliminates deleterious genomic elements from large populations. Thus, contrary to the assertion that "the limited variation in prokaryotic genome sizes … presents an important puzzle in its own right … just as important a question as the enormous genome size variation in eukaryotes" (Gregory 2005c), from a population genetic perspective, the uniformly simple genomes of prokaryotes are not surprising at all—they are the expectation.

Before we proceed, three fundamental points need to be emphasized. First, an increase in body size is not the only mechanism that results in population size reductions. Thus, the theory to be presented below does not rule out the possibility of a large genome in a unicellular species, provided that the latter has experienced an unusually high level of random genetic drift for reasons associated with the breeding system, degree of linkage in the genome, or historical long-term population bottlenecks (see Chapter 4). Random genetic drift is a fundamental attribute of all species, which in effect operates like a rheostat for modulating genomic growth versus contraction. Once this is understood, it is relatively straightforward to move beyond the rather vague description of excess DNA as being selfish or inert junk to a more mechanistic theory of genomic evolution capable of explaining why specific phylogenetic lineages are prone to the establishment of certain kinds of genetic elements and gene structures.

Second, because evolution is a stochastic process, subject to numerous probabilistic events, we should not expect to find tight deterministic relationships between all genomic attributes in all lineages. Indeed, when random genetic drift is a prominent evolutionary force, substantial deviations around overall patterns are *expected*. For a particular genome size, there can often be an order-of-magnitude range of variation among species with respect to genomic composition, and near the threshold for intron/mobile element expansion, the dispersion can be even greater (see Figure 2.1). Thus, although there is a tendency for those embroiled in the debate over genome size evolution to treat every deviation from an overall pattern as a definitive argument for or against a particular point of view (e.g., Cavalier-Smith 2005; Gregory 2005b), this strategy can be quite misleading. Returning to Figure 2.1, it is fairly easy to pick out two or three points that appear contrary to the overall trend, but much more difficult to do so with five to ten points. Consistent patterns are observed both within and among major phylogenetic groups, and it is this level of variation that merits explanation. This is not to say that outliers to an overall trend are uninteresting. Indeed, as will be seen in the following chapters, once the key biological features of such oddities are understood, they often provide deeper insights into evolution than might otherwise be possible.

One final point on the fundamental mechanisms driving genomic evolution merits further attention. As noted above, striking correlations exist between genome size and various aspects of cell size, metabolism, and division rate. In particular, within many phylogenetic groups, organisms with larger cells and lower metabolic rates generally have larger genomes (e.g., Gregory 2001, 2002a,b; Vinogradov and Anatskaya 2006). Is there any way to reconcile such patterns with the hypothesis that drift- and mutation-associated phenomena are the primary drivers of genomic evolution? The tentative answer is yes, in that a plausible case can be made that the evolved correlations between genome size and cytological features may be by-products of the shared involvement of intervening factors, rather than outcomes of direct causal connections.

As noted above (and argued more formally in Chapter 4), small population sizes and low mutation rates independently encourage the expansion of genome complexity. It is noteworthy that both population size and mutation rates tend to decline with increasing cell size. There is, of course, a one-to-one relationship between cell size and organism size in unicellular species. However, increases in cell size often accompany selectively driven increases in body size in laboratory experiments with multicellular species (Falconer et al. 1978; Riska and Atchley 1985; Stevenson et al. 1995; Calboli et al. 2003), and such associations also exist within a variety of phylogenetic groups, e.g., nematodes (Wang et al. 2002; Watanabe et al. 2005; Lozano et al. 2006), and sea anemones (Francis 2004). Thus, given the nearly universal inverse relationship between population size and body size (see Chapter 4), a negative correlation between cell size and population size is expected within many major taxonomic groups, with the causal intermediary being body size. In addition, metabolic rates tend to increase with decreasing cell size (Kozlowski et al. 2003), and because mutation rates scale positively with metabolic rates (Martin and Palumbi 1993; Gillooly et al. 2005), we can expect a negative association between cell size and mutation rate within broad taxonomic groups, driven by the intermediate factor of metabolic rate. Thus, although population size and mutation rate need not be inversely related to cell size in all phylogenetic groups, such scaling is likely for at least one of these parameters in most groups. This implies that although correlations between genome size and cytological features appear to support the causal connections postulated by the bulk-DNA hypothesis, they are also compatible with the mutational-hazard hypothesis.