

- Chapter One -

Introduction to the C-value enigma

- Chapter One - Introduction to the C-value enigma¹

Abstract

The study of genome size evolution began more than 50 years ago, before the molecular structure of DNA had been elucidated. In the time since, a great deal has been learned regarding the properties of genomes and their relationships to organismal characteristics. Most notably, it has been known since the earliest days of genome size investigation that nuclear DNA contents bear no relationship to intuitive notions of organismal complexity. And indeed, haploid genome sizes (“C-values”) do not correlate with the number of coding genes, an observation that became known as the “C-value paradox” in the early 1970s. The present chapter provides a short overview of the history of genome size study and briefly describes the major theoretical issues involved in understanding the process of C-value diversification. The outmoded concept of the “C-value paradox” is replaced by the more clearly delineated set of puzzles known as the “C-value enigma”. Other concepts such as “junk DNA” and “selfish DNA” are also dealt with briefly, as are proposals relating to the phenotypic effects of non-coding DNA including the “nucleoskeletal” and “nucleotypic” theories.

¹

This chapter contains material from Gregory (2001a,b) and a more detailed review article currently in preparation.

“The evolution of the large-scale features of the genome ... is perhaps the most difficult question in evolutionary biology.”
- John Maynard Smith, 1982

Introduction

The term “genome” has been part of the biological lexicon for more than 80 years, yet both its historical origin and exact definition remain somewhat ambiguous.

According to the *Oxford English Dictionary*, it was coined by the German botanist Hans Winkler in 1920 as a portmanteau of *gene* and *chromosome*. This assertion has been challenged recently by Lederberg and McCray (2001), who suggest that Winkler probably merged *gene* with the generalized suffix ‘*ome* (referring to “the entire collectivity of units”), and not ‘*some* (“body”) from *chromosome*. In either case, Winkler’s intent was to “propose the expression *Genom* for the haploid chromosome set, which, together with the pertinent protoplasm, specifies the material foundations of the species” (translation as in Lederberg and McCray 2001). Based on this initial formulation, “genome” could be taken to simultaneously signify the full set of chromosomes and all the genes contained therein, and indeed both definitions are still in use today. However, as light was shed on the structures and functions of both genes and chromosomes, it became increasingly apparent that one could not have it both ways – total chromosome content and number of genes are not interchangeable. This deceptively simple observation underlies one of the longest-running puzzles in evolutionary biology.

The issue in question is the evolution of genome *size*. That is, the amount of DNA contained within a haploid copy of the genome, measured as the number of base pairs or mass (in picograms, pg). The disconnect between genome size and gene number, and the specific patterns of distribution in genome size among taxa, have remained the

subjects of intense debate for several decades, yet for the most part the questions remain unresolved. Today, as previously, two opposite but equally important concepts form the basis of the study of genome size evolution: constancy and variation.

Genome size constancy and the C-value concept

It is interesting to note that the study of genome size predates the advent of modern molecular biology, which is often taken to have begun with the elucidation of DNA's chemical structure by Watson and Crick in 1953. As part of their pioneering work in quantifying nuclear DNA contents of cattle tissues, Roger and Colette Vendrely (1948) reported "a remarkable constancy in the nuclear DNA content of all the cells in all the individuals within a given animal species" [my translation]. Shortly thereafter, these findings were extended to several other mammals and to a few birds and fishes (Vendrely and Vendrely 1949, 1950; Davidson et al. 1950; Mandel et al. 1951). At a time when the subject was still a matter of some controversy, this constancy was taken as strong evidence that DNA, and not protein, was the chemical substance of heredity. As Vendrely (1955) would later write: "From the theoretical point of view the discovery of the constancy of the amount of DNA per nucleus in all tissues of the same animal and the fact that the sperm contains half the DNA content of somatic cells is confirmation of the theory that DNA is an important component of the gene".

By the late 1940s it had "become a textbook generalization that all cells of an organism have the same chromosomal constitution" (Schrader and Leuchtenberger 1949), and although it was challenged by several authors in the 1950s and '60s, the DNA constancy hypothesis has remained important to the present day. Indeed, the first measurements of broad interspecific genome size variation by Vendrely and Vendrely

(1949, 1950) and the monumental survey of Mirsky and Ris (1951) were based, methodologically and theoretically, on the notion of DNA constancy across tissues and within species. The same is true, whether recognized or not, of each of the hundreds of studies conducted in the past 50 years which have utilized cells from a species of “known” DNA content as a “standard” to calculate absolute DNA contents in a series of unknowns from reassociation kinetic, densitometric, or fluorescence data (see Chapter 5).

The continuity of the DNA constancy hypothesis is clearly reflected in the persistence of the term “C-value” to describe the haploid DNA content of a species. Contrary to almost universal belief, “C-value” does not signify “constant” or “characteristic”, but it does relate directly to the notion of DNA constancy. Specifically, “C-value” was coined by Hewson Swift (1950) in reference to the haploid, or “1C”, *class* of DNA as part of an early empirical defence of the DNA constancy hypothesis.

Genome size variation and the C-value paradox

The constancy of DNA content within chromosome sets may make the quantification of genome sizes possible, but it is the staggering variation among them that makes such measurements interesting. As shown in Figure 1.1, the genome sizes of eukaryotes vary over more than five orders of magnitude. Even within animals the range is more than 3,500-fold, and in vertebrates it is greater than 350-fold. The distribution is clearly non-random, with some groups exhibiting large and highly variable C-values, and others appearing constrained to relatively small ranges in genome size. For example, among the vertebrates, mammals, birds, reptiles, and teleost fishes all display narrow ranges in genome size, whereas the C-values of cartilaginous fishes, lungfishes, and amphibians (especially salamanders) are considerably larger (Fig. 1.1). A few key groups

of animals will be singled out for more detailed discussion in this thesis, as outlined below.

Perhaps the most obvious feature of the genome size distribution shown in Figure 1.1 is the total lack of association between genome size and any intuitive notions of organismal complexity, which is usually taken as a very rough proxy for the number of coding genes. As Comings (1972) put it (and rather bluntly at that),

Being a little chauvinistic toward our own species, we like to think that man is surely one of the most complicated species on earth and thus needs just about the maximum number of genes. However, the lowly liverwort has 18 times as much DNA as we, and the slimy, dull salamander known as *Amphiuma* has 26 times our complement of DNA. To further add to the insult, the unicellular *Euglena* has almost as much DNA as man.

As discussed above, the very notion of DNA constancy (and hence, “C-values”) upon which the study of genome size is based was derived from the view that DNA was the stuff of genes, and yet it is obvious from the observed genome size distribution that genome size and gene number are entirely unrelated. In 1971, C.A. Thomas summarized the situation as follows: “It was argued that mammals display a greater developmental complexity than primitive fish, therefore, they must have more genes, yet why should the lower forms have more DNA, if DNA is the chemical basis of the gene?”. To this seemingly impossible situation Thomas (1971) gave the name “C-value paradox”, a term which continues to enjoy widespread use.

As is now well known, eukaryotic genomes represent congeries of various types of DNA sequences, the majority of which exhibit no protein-coding function whatsoever.

As a prime example, the recently sequenced human genome (which is actually rather unimpressive in terms of its size) contains roughly 98.5% non-coding DNA (Fig. 1.2). It was by the simple discovery of such non-coding DNA that the C-value paradox was solved once and for all:

The C-value paradox is the observation that genome size does not correspond to the amount of DNA needed for protein-coding functions. This observation is a paradox only under the expectation that genome size should be equal or proportional to gene number and should therefore increase with 'organismal complexity'. This paradox has literally disappeared with the discovery that genomes contain 'excess' (largely repetitive) DNA that is not transcribed into functional products. Thus it is no longer mysterious that salamanders (for example) have larger genomes than humans. The origin and precise function of the 'excess' DNA (which may constitute more than 99% of the genomic DNA) remains an unsolved problem, but it is not a paradox. (Sessions 1986).

From paradox to puzzle

As with most questions in science, the solution to the "C-value paradox" generated a series of quandaries of its own. Granted that genome sizes vary primarily as a result of non-coding sequences, there is still the puzzle of why these sequences are not found in equal amounts in all organisms. To put it mildly, genome size variation is still as puzzling as ever, despite no longer qualifying as paradoxical. The questions inherent in this puzzle are also more complex, thereby making it immune to unifactorial explanations (unlike the former paradox). These include: Whence this non-coding DNA? What mechanisms account for its spread and/or loss over time? Why do some groups have so much of it, while others have very little? Which types of non-coding sequences

predominate in genomes? What impacts, if any, does this non-genic DNA have on the cellular and organismal phenotype? In combination, these questions form the much larger and more sophisticated puzzle recently recast as the “C-value enigma” (Gregory 2001a, 2002a).

The shift from “paradox” to “enigma” is not trivial, as it finally allows a full appreciation for the complexity of the puzzle. As long as the notion persists that a single “paradox” awaits resolution, one-dimensional explanations will continue to be offered. The traditional approaches to the issue of genome size have usually been classified into two different categories and considered mutually exclusive. These include mutation pressure theories and optimal DNA theories, as described below. However, because the various questions inherent in the C-value enigma are largely independent of one another, it is clearly a false dichotomy to pit these classes of theories against one another, since they deal with different aspects of the puzzle. (But note that they do come into conflict on some specific issues, like the explanation for why genome size should be related to cell size; see Chapter 2).

Mutation pressure theories

Junk DNA

Two main problems struck Susumu Ohno as particularly important in his work on the genetics of evolutionary diversification. The first was the standard “C-value paradox”, which was a prominent topic of discussion in the early 1970s: “If we take the simplistic assumption that the number of genes contained is proportional to the genome size, we would have to conclude that 3 million or so genes are contained in our genome. The falseness of such an assumption becomes clear when we realize that the genome of

the lowly lungfish and salamanders can be 36 times greater than our own.” (Ohno 1972a).

The second problem related to the conservative force of purifying selection and the limitations it places on the diversification of species.

Ohno attempted to kill both of these vexatious birds with a single conceptual stone:

The points I wish to make are: 1) Natural selection is an extremely conservative force. So long as a particular function is assigned to a single gene locus in the genome, natural selection only permits trivial mutations of that locus to accompany evolution. 2) Only a redundant copy of a gene can escape from natural selection and while being ignored by natural selection can accumulate meaningful mutation to emerge as a new gene locus with a new function. Thus, evolution has been heavily dependent upon the mechanism of gene duplication. 3) The probability of a redundant copy of an old gene emerging as a new gene, however, is quite small. The more likely fate of a base sequence which is not policed by natural selection is to become degenerate. My estimate is that for every new gene locus created about 10 redundant copies must join the ranks of functionless DNA base sequence. 4) As a consequence, the mammalian genome is loaded with functionless DNA. (Ohno 1973).

The corpulent genomes of dipnoans and urodele amphibians were similarly thus accounted for: “Lungfish and salamanders clearly show the tragic consequences of exclusive dependence upon tandem duplication” (Ohno 1970, p.96; see Chapter 4 for a different view).

To Ohno, this situation not only permitted, but also paralleled, the evolution of life at large. As he put it, “The earth is strewn with fossil remains of extinct species; is it

any wonder that our genome too is filled with the remains of extinct genes?” (Ohno 1972a). The primary outcome of this gene duplication mechanism would not be the generation of new genes (as important as this is), but the deactivation of redundant copies – just as extinction has been the fate of more than 99% of species that have ever lived (Raup 1991). Once purifying selection ceased to shelter gene sequences from change, they would be free to mutate and in most cases “would join the ranks of ‘garbage DNA’” (Ohno 1970, p.62).

In Ohno’s usage, as in the vernacular, “garbage” refers to both the *loss* of function and the *lack* of any further utility (it was once useful, but now it isn’t). “Garbage DNA” proved to be an unsuccessful meme, but its essence remains in the wildly popular term coined by Ohno two years later – “junk DNA”. Thus, as Ohno (1972b) stated, “at least 90% of our genomic DNA is ‘junk’ or ‘garbage’ of various sorts”. Interestingly, Ohno mentioned “junk DNA” only in the titles of two of his papers (1972a, 1973), and used the term only once in passing in a third (1972b). Comings (1972), on the other hand, gave what must be considered the first explicit discussion of the nature of “junk DNA”, and was the first to apply the term to *all* non-coding DNA.

Although extinct gene duplicates are now better known as “pseudogenes”, the term “junk DNA” has persisted as “a catch-all phrase for chromosomal sequences with no apparent function” (Moore 1996). In reference to the C-value enigma, the “junk DNA theory” is one in which “the nuclear genomes of eukaryotes accumulate junk DNA until the costs to the organism of replicating it become too great” (Pagel and Johnstone 1992). Organisms with large, slowly dividing cells are seen as being capable of tolerating this accumulation to a greater degree, thereby explaining the differences in genome size

among species.

Selfish DNA

While the junk DNA theory invokes natural selection only as a constraining force on random genomic expansion, a second mutation pressure theory employs selection directly (at two different levels) to explain the C-value enigma. The first example of this approach came with Östergren's (1945) description of non-coding B chromosomes as "parasitic" elements. A more thorough explication of this notion of "selfish DNA" as it relates to the C-value enigma was advanced independently by Doolittle and Sapienza (1980) and Orgel and Crick (1980). These authors criticized the tendency of genome evolutionists to seek adaptive, organism-level functions for each and every component of the genome. As an alternative, they argued that non-coding DNA could exist and expand solely for its own benefit. Citing such examples as transposable elements, bacterial plasmids, and retroviruses, they described a dynamic evolutionary process within the genome, with selfish, parasitic genomic elements competing amongst each other for maximum representation (by what was termed "non-phenotypic selection", a process more aptly dubbed "intragenomic selection" by Cavalier-Smith [1980a]). In its most rigorous formulation, the selfish DNA theory refers only to those elements which are actively undergoing replicative transposition (since inactive replicators correspond more closely to the definition of "junk" than to that of "selfish").

The inevitable expansion of efficient self-replicators, it is argued, would produce a constant pressure for increased genome size (e.g., Maynard Smith and Szathmáry 1995). In similar fashion to junk DNA theories, the selfish DNA view assumes that this expansion would be constrained only when this added genetic baggage became too

unwieldy for the host cell, and that there exists a greater tolerance for this superfluous DNA in organisms with large cells and slow cell division rates.

Optimal DNA theories

The next chapter provides an in depth discussion of the positive relationship between DNA content and cell and nucleus sizes, what Cavalier-Smith (1982) calls “the most reliably established fact about genome evolution”. As noted above (and discussed critically in more detail in Chapter 2), both mutation pressure theories attempt to explain this relationship based on the differential tolerance for DNA accumulation in cells of varying sizes. A different interpretation is offered by the two “optimal DNA theories”, which place emphasis on the fitness consequences of these associations with cellular parameters.

The nucleoskeletal theory

In 1978, Tom Cavalier-Smith suggested that variation in DNA content could be explained as the result of a compromise between selection for rapid growth and efficient metabolism – in other words, between the inversely related parameters of cell division rate and cell size. Under Cavalier-Smith’s “nucleoskeletal theory”, cell size is adjusted adaptively in response to these selective pressures, and these changes enjoin correlated shifts in nucleus size, which are in turn carried out by modulating the amount of “skeletal” DNA in the nucleus. The proposed reasons for this necessary correspondence between nucleus and cell size have changed considerably over the years (see Chapter 2), but the fundamental emphasis on the adaptive manipulation of cell size followed by a coevolutionary adjustment of genome size has persisted throughout. Under this view, there is an optimal amount of DNA for each type of organism determined by the specific

compromise between rapid cell division and efficient cellular metabolism. The mechanisms of DNA content change are inconsequential under the nucleoskeletal theory. A critique of this explanation for the genome size-cell size relationship and the C-value enigma in general is provided in the next chapter.

The nucleotypic theory

The notion that DNA content can directly influence cell size has existed for over a century (see Chapter 2), but was not formally codified until Michael Bennett (1971, 1972) developed the concept of the “nucleotype” to describe “those conditions of the nucleus [especially its size] that affect the [cellular] phenotype independently of the informational content of the DNA”. As applied to the C-value enigma, the nucleotypic theory also emphasizes the relationship with cell size, and is therefore an optimal DNA theory. But unlike the nucleoskeletal view, the simplest version of the nucleotypic theory attempts to explain variation in DNA content as the result of direct selection for cell and genome size (versus indirect selection for cell, then nucleus, and ultimately genome size). A more complex view of genome size evolution will be presented over the course of the following chapters, but the basic notion of a causal influence of bulk DNA on the cellular phenotype remains fundamental.

Outline of the thesis

As a truly comprehensive treatment of animal genome size is clearly not possible within the confines of a thesis format, the following chapters discuss only a subset of the relevant issues. Chapter 2 provides a detailed discussion of the relationship between nuclear DNA content and cell size, focussing primarily on vertebrate erythrocytes. A case is made for a causal (nucleotypic) interpretation of these correlations, and a

mechanistic model is proposed to account for them. In Chapter 3, mammals and birds are singled out for a discussion of the patterns and implications of variation in genome size. These animals have relatively constrained C-values and are unique in having a homeothermic lifestyle. The fact that they have been well-studied from physiological and developmental perspectives allows some important ideas about genome size constraints to be tested in detail for the first time. Chapter 4 deals with a group with opposite features, namely the amphibians, which have highly variable genome sizes and very different physiologies. Again, these animals have been well-studied from a developmental perspective and therefore provide an important test case for ideas relating genome size to the organismal phenotype. Chapters 2 through 4 make use of previously published data which have been collated from a large number of sources and presented in the appendices of Volume Two. The entire *Animal Genome Size Database* (Gregory 2001c) is provided for use as a resource, although not all the groups covered in the appendices are discussed in this thesis. Chapter 5 presents the new method of Feulgen image analysis densitometry used in the present study for the collection of new genome size data. Some historical and technical background information is provided, as is a guideline for the preparation of specimens from all the major animal groups. Chapter 6 presents 230 new insect C-values, effectively doubling the current dataset for this group. These new estimates are combined with previously published values to give an order-by-order review of the patterns of variation in insects. More than 110 spider genome sizes are also presented, which represents the first survey for these animals. Finally, Chapter 7 outlines the arguments in favour of a hierarchical macroevolutionary theory presented by palaeontologists, and then translates the major concepts into the language of genome size

evolution. The importance of maintaining a hierarchical view for understanding the C-value enigma is emphasized, and some intriguing ways in which genome-level processes have influenced the major transitions in evolution are discussed.

Concluding Remarks

Many of the concepts developed during the past 50 years of genome size study remain important today, although several have clearly outlived their utility and are in need of revision or deletion. The former “C-value paradox”, for example, has dissolved and in its place has emerged the complex and multifaceted “C-value enigma”. The two main approaches to the C-value enigma – mutation pressure and optimal DNA theories – deal with different aspects of this puzzle, and are therefore not mutually exclusive (so long as a hierarchical view is maintained; see Chapter 7).

The junk DNA theory, which is based on the passive accumulation of non-coding sequences (especially pseudogenes), and the selfish DNA theory, which emphasizes the active multiplication of transposable elements, both deal with the mechanisms by which non-coding DNA may originate and spread. The optimal DNA theories focus instead on the phenotypic impacts of bulk DNA on the cell, the nucleoskeletal theory favouring a coevolutionary interpretation and the nucleotypic theory a causative one (Gregory 2001a). Importantly, these optimal DNA theories make no mention of the mechanisms by which DNA content might be altered. The main (perhaps only) source of antagonism between all these theories is in their views on the significance and cause of the correlation between genome size and cell size – the topic of the next chapter.

It is essential to grasp two main points if one is to understand and perhaps resolve the C-value enigma. The first is that the puzzle is complex and multifaceted, and

therefore immune to any one-dimensional solutions (unlike the paradox of yore). The second is that the enigma falls under the purview of several biological disciplines, and that features of genomes cannot be understood without reference to features of cells and organisms. An acceptance of this complexity and a recognition that a pluralistic approach capable of treating the various component questions of the puzzle individually and in an appropriate context represent the first crucial steps along the path to understanding the C-value enigma.

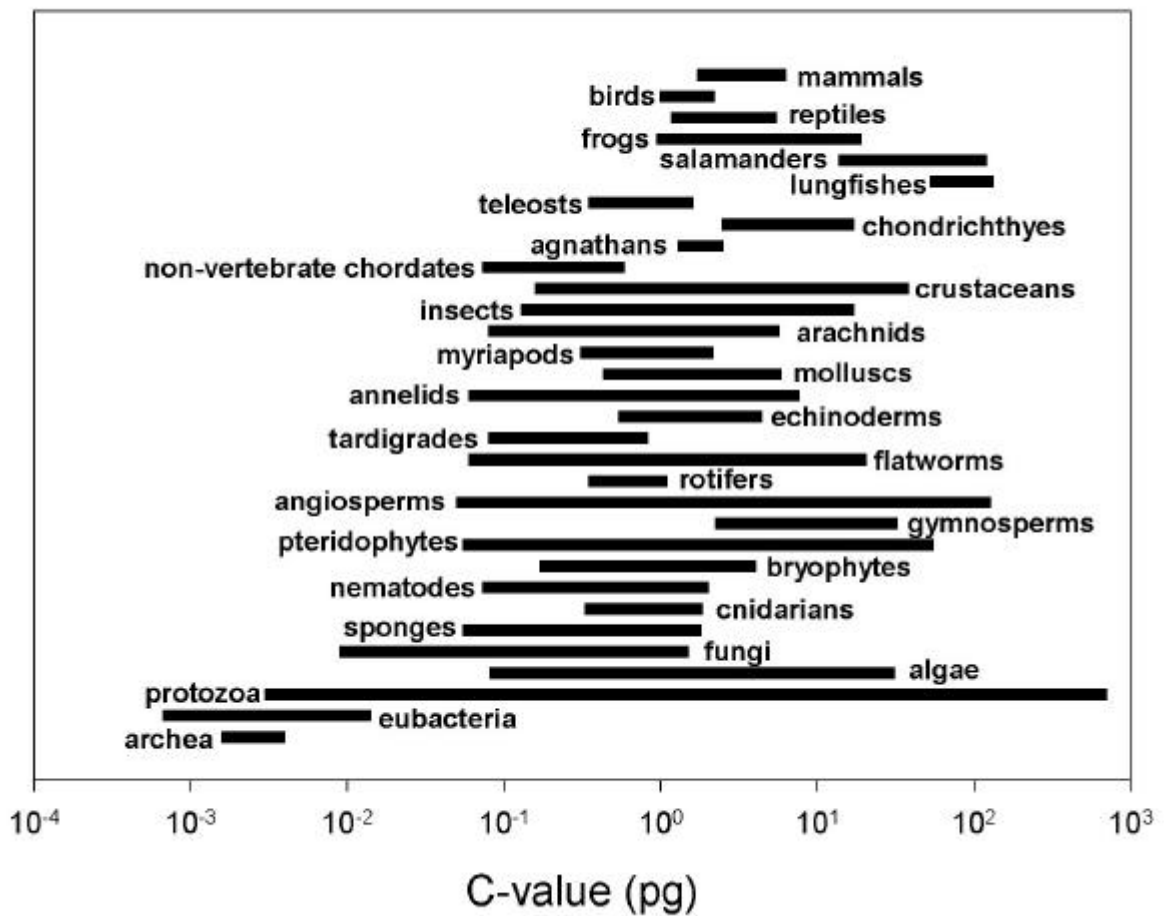


Figure 1.1. The ranges in C-values (log-transformed) so far observed in different groups of organisms, arranged according to the sort of *scala naturae* still clung to by many people. Genome size is clearly unrelated to organismal complexity. Based on Raff and Kauffman (1983), using the most recent C-value data available (from Biderre et al. 1995; Renzaglia et al. 1995; Li 1997; Baumann et al. 1998; Bennett et al. 2000; Voglmayr 2000; Bennett and Leitch 2001a,b; Gregory 2001c; Leitch et al. 2001), along with some unpublished estimates presented here for the first time (Chapter 6 and appendices).

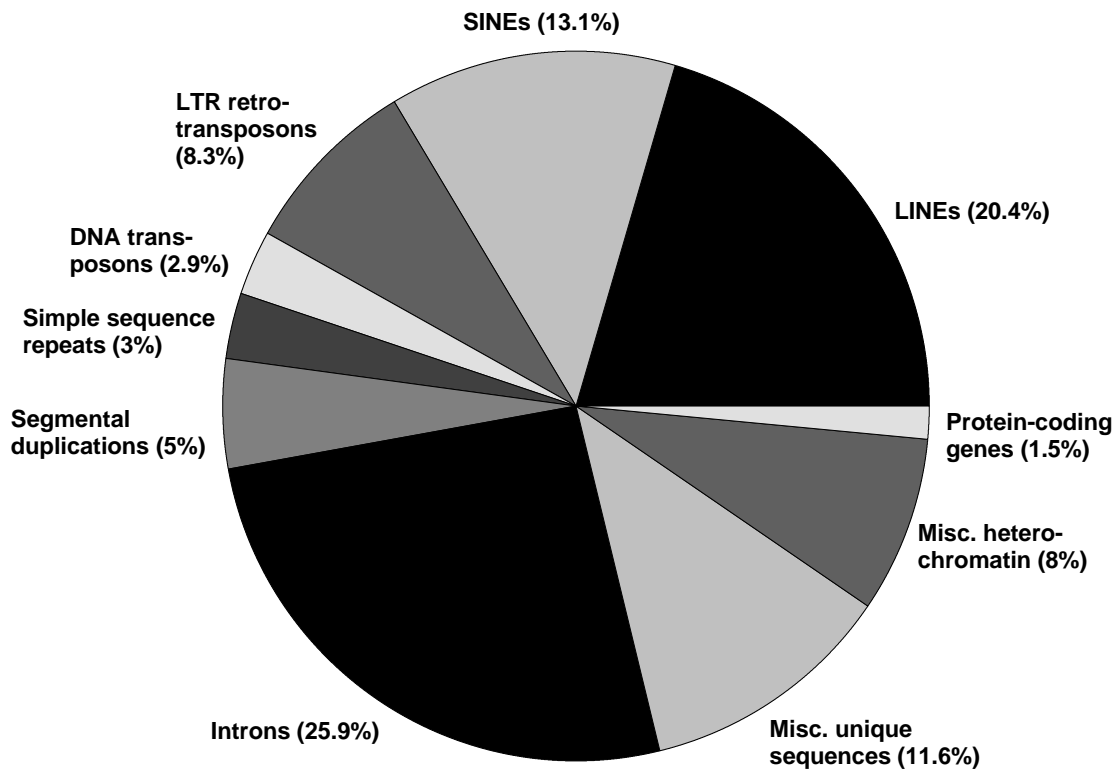


Figure 1.2. Summary of the different components of the human genome. Only about 1.5% of the genome consists of protein-coding sequences, whereas about 45% of it is composed of transposable elements of various types. The existence of non-coding DNA, which is present in differing amounts among various eukaryotes, explains the C-value paradox but raises several new questions as part of the C-value enigma. Data from the International Human Genome Sequencing Consortium (2001) and The Wellcome Trust (<http://www.wellcome.ac.uk/en/genome/thggag.htm>).