The Evolution of Genome Size

Edited by
T. CAVALIER-SMITH

The Evolution of Genome Size

Edited by

T. CAVALIER-SMITH

Department of Biophysics
King's College London
University of London

A Wiley-Interscience Publication

JOHN WILEY & SONS
Chichester · New York · Brisbane · Toronto · Singapore

Library of Congress Cataloging in Publication Data:
Main entry under title:

The Evolution of genome size.

'A Wiley–Interscience publication.'
Includes bibliographies and index.
1. Genomes. 2. Evolution. I. Cavalier-Smith, T.
QH447.E96 1985 574.87'3282 84–25659

ISBN 0 471 10272 5

British Library Cataloguing in Publication Data:

The Evolution of genome size.
1. Chromosomes
I. Cavalier-Smith, T.
574.87'322 QH600

ISBN 0 471 10272 5

Typeset by Input Typesetting Ltd., London SW19 8DR.
Printed and bound in Great Britain by The Bath Press, Bath, Avon.

Contributors

AMMERMANN, D. Lehrstuhl Zoologie, Abteilung Zellbiologie, Institut fur Biologie III, Universität Tübingen, D 7400 Tübingen 1, Auf der Morgenstelle 28, Federal Republic of Germany.

BACHMANN, K. Hugo de Vries Laboratorium University of Amsterdam, Kruislaan 318, 1098 SM Amsterdam, The Netherlands.

BOYNTON, J. E. Department of Botany, Duke University, Durham, North Carolina 27706, USA.

CAVALIER-SMITH, T. Department of Biophysics, King's College London, 26–29 Drury Lane, London WC2B 5RL, England.

CHAMBERS, K. L. Herbarium, Department of Botany, Oregon State University, Corvallis, Oregon 97331, USA.

CHARLESWORTH, B. Department of Biology, University of Chicago, 1103 E. 57th Street, Chicago, Illinois 60637, USA.

CLARK-WALKER, D. Research School of Biological Sciences, Australian National University, Box 475, P. O., Canberra City, A. C. T. 2601, Australia.

CULLIS, C. A. John Innes Institute, Colney Lane, Norwich NR4 7UH, England.

DOOLITTLE, W. F. Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H7.

DYSON, P. Department of Genetics, University of Glasgow, Glasgow G11 5JS, Scotland.

GILLHAM, N. W. Department of Zoology, Duke University, Durham, North Carolina 27706, USA.

HARRIS, E. H. Department of Botany, Duke University, Durham, North Carolina 27706, USA.

HERDMAN, M. Unité de Physiologie Microbienne, Département de Biochimie et Génétique Microbienne, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris, France.

JONES, R. N. Department of Agricultural Botany, School of Agricultural Sciences, Penglais, Aberystwyth, Dyfed SY23 3DD, Wales.

NURSE, P. Imperial Cancer Research Fund, Lincoln's Inn Fields, London WC2, England.

PRICE, H. J. Genetics, Department of Plant Sciences, Texas A and M University, College Station, Texas 77843, USA.

SHERRATT, D. Department of Genetics, University of Glasgow, Glasgow G11 5JS, Scotland.

Preface

The greatest advance in evolutionary biology since Darwin and Wallace's 1858 papers came in the period 1922–1932 with the formulation of the genetical theory of evolution by Muller, Haldane, Chetverikov, Fisher and Wright. Since then molecular genetics has progressed so dramatically that major new insights into evolution are probable during the next decade, as the structure of genes, and mechanisms of mutation, recombination and gene expression, become well understood. New discoveries are coming thick and fast but they will increase evolutionary understanding only if they are synthesized with existing knowledge and with new findings in all areas of biology.

The purpose of this book is to contribute to this synthesis by focusing on one relatively neglected but fundamental problem: the evolution of genome size. Bacterial genome size is probably directly related to the number of genes, and is a good measure of their genetic and organismic complexity. But in eukaryotes this is not so, and the existence of vast amounts of apparently non-genic DNA has long been a major puzzle, now known as the DNA C-value paradox. Does all this DNA, sometimes as much as 50 000-fold in excess of protein-coding requirements, have a function? Or is it merely useless 'junk' or a 'selfish' parasite? Attempts to answer these questions soon lead into a great variety of fascinating questions: the mechanisms of DNA replication, duplication and transposition; the significance of repetitive DNA, split genes, heterochromatin, and chromatin elimination; the genetic control and evolutionary significance of cell and nuclear volume and cell growth rates; and many others. Therefore a number of cell and molecular biologists and molecular, population, and cyto-geneticists, were invited to make their distinctive contributions to this book, which is the first to concentrate on the evolution of genome size. It presents important new arguments in favour of the skeletal DNA theory of the evolution of genome size, and against the junk and selfish DNA explanations of the C-value paradox.

Nearly half a century ago Darlington very clearly expressed the problems inherent in a study of this kind in his preface to the second edition of his *Recent Advances in Cytology* (1937, Churchill, London): 'Finding out *why* things happen in the cell is an entirely different matter from finding out how they happen. The one problem is a matter of skill and common sense. The

other takes us into a new element. We are plunged into inferences, often speculative inferences, which connect mechanics, physiology and genetics. We find that the cell is part of an interlocking system of growth and reproduction, heredity and variation. Everything that happens in the cell is related to everything else that happens in the organism, or indeed has happened in its ancestors. It is impossible at one and the same time to deal with all these dialectical relationships. I can describe only those that seem to me most important at the moment.' Too often such evolutionary aspects of cell biology have been relegated to brief and often superficial speculations at the ends of papers primarily concerned with reporting new facts. To avoid such superficiality and to carry further the integrative approach pioneered by Darlington, and which must be central to evolutionary studies, it is best to consider in detail a single major evolutionary problem, and explore many different aspects of it. This book therefore does *not* attempt to deal with all aspects of genome evolution. It should interest not only evolutionary biologists but also geneticists and molecular and cell biologists concerned primarily with basic phenomena such as DNA replication and the cell cycle, since comparative evolutionary studies can also give valuable clues about fundamental mechanisms.

I warmly thank all the contributors, and apologise to them for the varying delays between receipt of their manuscripts and going to press. I much appreciate the encouragement given to my work on genome size by Professors H. G. Callan and M. H. F. Wilkins and thank Bridget Sarsby and Patricia Collins for the typing.

T. Cavalier-Smith

Units of measurement of genome size

Nuclear genome size of eukaryotes is usually measured in picograms (pg) of DNA (1 pg = 10^{-12}g).

The smaller prokaryotic genomes, consisting of single molecules, are more commonly measured in daltons – the unit of relative atomic and molecular mass.

1 dalton $\simeq 1.66 \times 10^{-12}$pg
10^9 daltons $\simeq 0.00166$ pg
1 pg $\simeq 6.02 \times 10^{11}$ daltons

Plasmid genome size is more often expressed as the number of base pairs per molecule or as the number of kilobases or kilobase pairs (abbreviated kb or kbp).

For double stranded DNA 1 kb = 1000 base pairs $\simeq 6.18 \times 10^5$ daltons.
1 pg $\simeq 0.98 \times 10^6$ kb.

Contents

The Evolution of Genome Size
Edited by T. Cavalier-Smith

CHAPTER 1

Introduction: the Evolutionary Significance of Genome Size

T. CAVALIER-SMITH

THE DNA C-VALUE PARADOX

Molecular genetics began with the discovery of DNA (Miescher, 1871), and the formulation of four key ideas:

1. The idea of the exact replication of molecular genetic determinants (Roux, 1883; Nägeli, 1884).
2. The idea that these consist of chromosomal DNA (Sachs, 1882; Hertwig, 1884).
3. The idea that genes may be active or dormant (Darwin, 1868).
4. The idea that genes act by the passage of short-lived copies into the cytoplasm where they control biosynthesis and cell structure (de Vries, 1889).

A century of work has firmly established these basic qualitative ideas and clothed them with remarkable molecular detail, but a fundamental quantitative problem remains: why should the cellular DNA content of differing species vary by over a million-fold? This book attempts to answer this question.

The mass of DNA in an unreplicated haploid genome, such as that of sperm nucleus, is called the genome size (Hinegardner, 1976) or, alternatively, the C-value (Swift, 1950) because it is usually constant in any one species. Hardly had this constancy been established when Mirsky and Ris (1951) demonstrated huge interspecific variation in C-values that bore no relationship to differences in organismic complexity or to the likely number of different genes in the species studied. Molecular biologists immediately saw the constancy of the C-value within a species, and the fact that unreplicated diploid nuclei contained twice this amount of DNA (Boivin *et al.*, 1948; Vendrely and Vendrely, 1948), as support for the idea that genes consisted

of DNA. But for years the problem posed by the great variability in genome size in different species was widely ignored.

There has been a recurrent tendency to regard DNA C-values as related simply to the number of genes, and to equate evolutionary accumulation of DNA with increases in gene number (e.g. Kimura, 1961). This view of the significance of genome size appears to be true only for bacteria (see Chapter 2) and viruses. But ever since the first measurements (Mirsky and Ris, 1951) the evidence has been strongly against it for eukaryotes. In Chapter 3 I show that variation in eukaryote genome size is over a thousand-fold greater than can be accounted for by variations in gene number and that variation in eukaryote genome size bears no relationship to organismic complexity.

As it became increasingly firmly established that genes consist of DNA this puzzling lack of relationship between estimates of the number of genes and the DNA C-values of different eukaryotes became more and more of a problem, now usually called the C-value paradox (Thomas, 1971).

Six quite different kinds of solution have been proposed:

1. The whole idea of particulate heredity and of a definite number of discrete genes is false (Goldschmidt, 1955).
2. Genes do not consist of DNA, which must have some other role, perhaps structural (Darlington, 1956).
3. Mendel's idea that there is only one copy of each gene per gamete is false; instead the number of copies vary with C-value, either because each chromosome contains a variable number of longitudinal DNA strands (the multistrand theory: Darlington, 1955; Martin and Shanks, 1966; Rothfels *et al.*, 1966) or because of tandem duplications along its length (Callan and Lloyd, 1960; Callan, 1967; Ohno, 1970). To reconcile his postulate of many copies of a gene per genome with the genetic evidence for a single copy, Callan (1967) proposed his ingenious master-slave theory.
4. Much DNA is genetically or physiologically inert (Muller and Painter, 1932; Darlington, 1937) and is useless junk (Ohno, 1972) carried passively by the chromosome merely because of its linkage to functional genes (Rees, 1972).
5. Much DNA is a functionless parasite (Östergren, 1945; 'selfish DNA' *sensu* Doolittle and Sapienza, 1980) or 'genetic symbiont' (Cavalier-Smith, 1983a) that accumulates and is actively maintained by intragenomic selection (Cavalier-Smith, 1980a).
6. DNA has quantitative non-genic, or 'nucleotypic' (Bennett, 1971) functions, in addition to its qualitative genic functions (Commoner, 1964; Stebbins, 1966; Bennett, 1971, 1972, 1973; Rees, 1972; Cavalier-Smith, 1978, 1980a, 1982a).

The first two possibilities were soon decisively rejected by the elucidation of the genetic code (Crick, 1967). But at about the same time the powerful new

method of DNA renaturation kinetics (Bolton *et al.*, 1965) led to the discovery that most eukaryotes, unlike bacteria, contained large amounts of repetitive DNA (Britten and Kohne, 1968). This at first appeared to support the third idea that there might be several copies of protein-coding genes in each genome. However, further studies using renaturation kinetics soon showed that eukaryote chromosomes were not multistranded (Laird, 1971) and that protein-coding genes were mostly present in only one copy per genome, even in organisms with very large C-values (Rosbash *et al.*, 1974). This led to the rejection of the third possibility, and to the present position where we have to accept not only that genes do consist of DNA but also that in many eukaryotes the major part of the DNA does not consist of functional cellular genes. A convenient non-committal name for this extra DNA, which occurs in very large amounts in high C-value eukaryotes, is secondary DNA (Hinegardner, 1976). Why this DNA should be there, and the relative merit of the last three possibilities, is the fundamental problem considered in this book. Secondary DNA must either have a non-genic function or be the incidental result of mutational and intragenomic selective forces of no positive significance to the organism. Since this book is intended not only for molecular biologists but also for population geneticists, cell biologists and general evolutionary biologists, who may be less familiar with the arguments from DNA renaturation kinetics that have led to this conclusion, I shall start with a brief outline of them; a more detailed and documented explanation of renaturation kinetics is given by Lewin (1980, ch. 18, 19) and Bouchard (1982).

RENATURATION KINETICS, CHROMOSOME UNINEMY AND THE NUMBER OF COPIES OF EACH GENE

To measure the kinetics of renaturation, a sample of purified DNA broken into fragments a few hundred nucleotides in length (i.e. shorter than most genes) is first denatured by heating the DNA solution to force apart the complementary strands. If the denatured sample is placed at a temperature about 20°C less than was needed to cause complete denaturation, the separated single-stranded fragments will in time meet and pair with complementary molecules; the rate of this renaturation is determined by measuring the fraction of fragments that has become double-stranded as time progresses. Under standard conditions this rate depends only on the concentration and nature of the DNA. If the concentration and base composition of the DNA is kept constant the rate of renaturation under standard conditions depends only on the number of copies of each DNA sequence in the solution. The larger the number of copies of each sequence the greater the chance that complementary sequences will meet and reanneal by zipping together, so the faster will be the observed rate of renaturation.

Nuclear DNA from most eukaryotes proved to be a mixture of DNA that renatures rapidly and DNA that renatures much more slowly (Britten and Kohne, 1968). The rapidly renaturing DNA fragments therefore consist of sequences that are repeated many more times in the genome than the slowly renaturing ones. Almost all the rapidly renaturing sequences become double-stranded before a detectable fraction of the slowly renaturing sequences have begun to reanneal: therefore, by separating double and single-stranded DNA from each other at this stage of the renaturation reaction by hydroxyapatite chromatography, one can obtain relatively pure samples of the rapidly and slowly renaturing DNA for separate study. Because both the multistrand and the master-slave theories predicted many copies per genome of each gene, and that the average number of copies should be greater in high C-value species than in low C-value species, it was important to test this prediction.

This was first done by Laird (1971) who compared the renaturation rates of the slowly renaturing DNA fraction in mice, the fruit fly *Drosophila* and the sea squirt *Ciona*, with each other and with the renaturation rates of DNA from the bacteria *Escherichia coli* and *Bacillus subtilis* and the phage T_4. Unlike eukaryote DNA, neither bacterial nor viral DNA could be subdivided into fractions renaturing at different rates. However, the viral DNA renatured 18 times faster than the *E. coli* DNA. This was expected because the viral genome is 18-fold smaller than the bacterial genome; so long as both kinds of DNA contain only one copy of each gene per genome a solution of viral DNA would contain 18 times as many copies of each sequence per unit volume as a solution containing the same concentration of bacterial DNA (where concentration is measured in moles of nucleotide per litre) and will therefore renature proportionally faster. Laird found that the time taken for renaturation of the slowly renaturing DNA from the three eukaryotes was also directly proportional to the mass of this DNA per genome. Moreover, when the eukaryote and prokaryote data were plotted on the same graph (Fig. 1.1) a single straight line passed through all the experimental points. Thus, if phage T_4 and *E. coli* had only one copy of each sequence per genome, the same must be true for the slowly annealing DNA of each of the three eukaryotes even though they differed from each other by 22-fold in C-value. The slowly annealing fraction of eukaryote DNA is therefore now referred to as non-repeated, unique or single-copy DNA. Since single-copy DNA makes up well over half the genome in these and many other species with vastly differing C-values, it is clear that the bulk of the variation in genome size reflects variation in the amount of single-copy DNA and therefore cannot be explained either by a multistrand model, the master-slave hypothesis, or any modifications of them.

Three additional independent methods: relaxation viscometry (Lauer and Klotz, 1975; Kavenoff and Zim, 1973), ultracentrifugation (Petes and Fangman, 1973) and electron microscopy (Petes *et al.*, 1973) have now shown

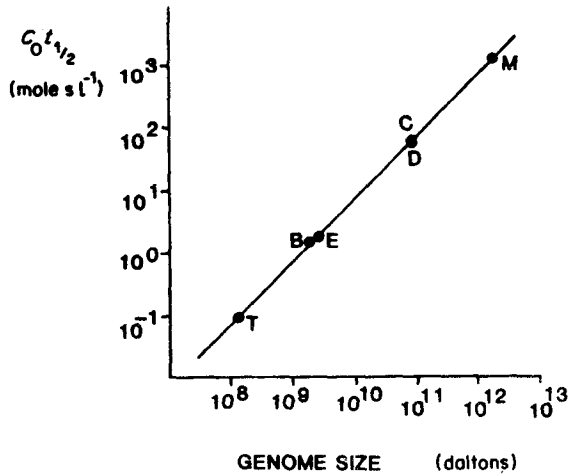

Fig. 1.1. Correlation between genome size and the time taken for the DNA to renature. The time taken for the renaturation of half the DNA at a concentration C_0 was measured for the total DNA of phage T_4 (T) and the bacteria *Bacillus subtilis* (B) and *Escherichia coli* (E); and for the major slowly-renaturing fraction of the DNA of three animals (the fruit fly *Drosophila melanogaster* (D); the seasquirt *Ciona intestinalis* (C) and the mouse *Mus musculus* (M)). The product of the half-time for renaturation ($t_{1/2}$) and the DNA concentration, C_0, is referred to as the C_0t value and is constant for any one species; the linearity and unitary slope of the curve shows that it depends only on the (haploid) genome size. Data from Laird (1971).

that the chromosomes of the yeast *Saccharomyces cerevisiae* and *Drosophila melanogaster*, which differ 20-fold in C-value, each consist of one single DNA molecule. Furthermore, the chromatids of the lampbrush chromosomes of *Triturus cristatus*, whose C-value is 3142 times that of yeast and 157 times that of *Drosophila*, showed the same kinetics of digestion by DNase as a single double-helical DNA molecule (Gall, 1963), and therefore could not be multistranded. Experiments stretching *Chironomus* polytene chromosomes to breakpoint (Gruzder and Reznik, 1981) confirm the uninemy of chromatids. There is therefore no longer reason to doubt that all unreplicated eukaryotic chromosomes consist of single linear DNA molecules, to which are attached numerous protein molecules.

Further evidence against explanations for the C-value paradox in terms of variable numbers of identical copies of genes lies in the demonstration by renaturation kinetics that most messenger RNA molecules are transcribed from unique sequences, present only once per haploid genome. This is true in adult tissues and for cells in tissue culture (Greenberg and Perry, 1971; Lewin, 1980: chapter 24) and in amphibian ovaries in which all the messen-

gers required for early embryonic development are being synthesized and stored in oocytes for later use (Rosbash *et al.*, 1974). In the latter case it was shown that the kinetic complexity (a measure by renaturation kinetics of the number of different sequences) of the messenger RNA fraction was the same in the amphibians *Xenopus* and *Triturus* which differ 7-fold in genome size. If this could also be established for every stage of the life history, it would prove that most of the extra DNA in organisms with larger genomes does not code for proteins. Unfortunately, this is not strictly possible using present techniques which could not detect extremely rare messengers that might be present only transiently in a rare somatic cell type. But even in the absence of definitive proof of this the evidence for it outlined in Chapter 3 is very strong. The extra DNA in organisms with high C-values must therefore either have a non-coding (perhaps structural) function or be non-functional and simply produced as the net result of non-adaptive evolutionary forces.

CORRELATIONS BETWEEN EUKARYOTE GENOME SIZE AND QUANTITATIVE CHARACTERS

Many authors (e.g. Lewin, 1980) who have discussed the C-value paradox have appeared to be unaware of the extensive evidence for positive correlations between eukaryote genome size and a variety of quantitative characters, notably:

1. the total volume and mass of the metaphase chromosomes;
2. interphase nuclear volume and dry mass;
3. cell volume and weight;
4. nucleolar volume and dry mass;
5. centromere volume;
6. nuclear RNA content;
7. nuclear protein content;
8. seed weight in herbaceous flowering plants;
9. length of the cell cycle;
10. the duration of meiosis;
11. the length of the DNA synthesis or S-phase of the cell cycle;
12. pollen maturation time;
13. minimum generation time in flowering plants;
14. time taken for embryogenesis from fertilization to the hatching tadpole in amphibians.

Since many of these quantitative characters are of considerable adaptive and functional significance to the organisms possessing them, many biologists (e.g. Stebbins, 1966; Szarski, 1970, 1983; Bennett, 1973; Cavalier-Smith,

1978, 1982a) have reasonably assumed that natural selection acting on such characters has played a major role in the evolution of different genome sizes.

In flowering plants there is also clear evidence for an association between genome size and climate (Bennett, 1976; Levin and Funderberg, 1979), and for herbaceous perennials in the temperate zone between genome size and temperature at the season of maximal growth (Grime, 1983). Similar correlations between climate and chromosome size in plants have long been known to botanists (Heitz, 1926, 1927; Avdulov, 1931; Stebbins, 1971, 1976), which may be why they have been less ready than zoologists to regard variations in genome size as non-adaptive.

Those aware of these facts have tended to assume, like Bachmann *et al.* in Chapter 9, that they are incompatible with the idea that extra DNA in larger genomes is merely useless 'junk' DNA or 'selfish' DNA and that they show instead that genome size *itself* genetically controls most or all of these quantitative characters and is subject to selection so as to optimize them. Certainly the mere postulate of the existence of junk DNA and parasitic DNA does not solve the problem of the evolution of genome size, since it does not explain why these correlations exist and why they have such numerical values. But on the other hand the mere existence of these correlations does not prove that genome size exerts a direct nucleotypic control over these phenotypic characters, or that selection acting on such characters directly favours an increase in genome size; while this could be the case, it is in principle possible that none of these characters except chromosome volume is nucleotypically determined, and that secondary DNA has no positive function. My own view is that the majority of the quantitative characters that correlate with genome size are not directly and mechanistically determined by genome size, and that the observed correlations are mostly the result of indirect developmental and/or evolutionary correlations; but two phenotypic characters, namely chromosome volume and nuclear volume, do seem to be directly and causally determined by a combination of the sheer amount of DNA together with the tightness or looseness of its folding. Chapter 4 argues that this joint nucleotypic and genic control of nuclear volume is the fundamental reason for the observed correlations, but that the crucial evolutionary factor is cell volume, not genome size itself; evolutionary variations in cell volume probably lead to the selection of correlated variations in the other characters, including nuclear volume and therefore genome size which partly determines it. In order to substantiate or refute this thesis we need a better understanding of the genetic control of the cellular and organismic characters in question; several aspects of the problem are reviewed in Chapters 4 and 5.

Two other features of genome size variation that have been ignored until

recently (Cavalier-Smith, 1978) require explanation and also hold important clues to the solution of the C-value paradox. One is the fact that the three amniote classes (mammals, birds, reptiles) are exceptional among eukaryotes in their small variation in genome size (2 to 4-fold). Most extensively studied eukaryotes, whether plant, animal or microorganism, show a much greater variation in genome size. Protozoa show a 1000 and unicellular algae a 3000-fold range, whereas most classes of multicellular animals and plants show a variation of the order of 10 to 100-fold. Great variation in C-value is therefore not the exception, but the general rule in most eukaryote groups: it is the near constancy in mammals and in birds and reptiles that is the exception and which requires a special explanation. The second neglected fact is that unicellular eukaryotes in general show a greater range in C-values than multicellular ones. Both the smallest, *and* the largest, eukaryote genome sizes are found in unicellular species; from yeast to Amoeba the range is 80 000-fold, but from sponges to humans only 55-fold! Multicellular genomes also cluster in the middle range (from 0.5–10 pg).

Therefore, whatever are the evolutionary forces that prevent large reductions or increases in genome size, they appear to be stronger in multicellular eukaryotes than in unicellular ones, stronger in animals than in plants and strongest of all in amniotes; one might argue conversely that the evolutionary forces causing large changes in genome size are strongest in unicellular eukaryotes.

An ideal solution to the C-value paradox would be a theory that could simultaneously explain:

1. these differences in the range of C-values in different eukaryote groups;
2. the positive correlations between genome size and so many quantitative characters;
3. the absence of any correlation between genome size and gene numbers.

My hypothesis that genic and non-genic DNA together have a universally important function as a nuclear skeleton that determines the volume of nuclei (Cavalier-Smith, 1978, 1982a) still seems to achieve this more satisfactorily than any competing hypothesis.

THE SKELETAL DNA HYPOTHESIS

'My basic argument is that natural selection acts powerfully on organisms to determine their cell size and developmental rates (which are inversely related). The mean cell volume of an organism is the result of an evolutionary compromise between conflicting selection for large cell size and for rapid developmental rates: the particular compromise reached for a particular species will depend on its ecological niche and organismic properties. Since larger cells require larger nuclei, selection for a particular cell volume will secondarily select for a corresponding nuclear volume, producing a close correlation between cell and nuclear