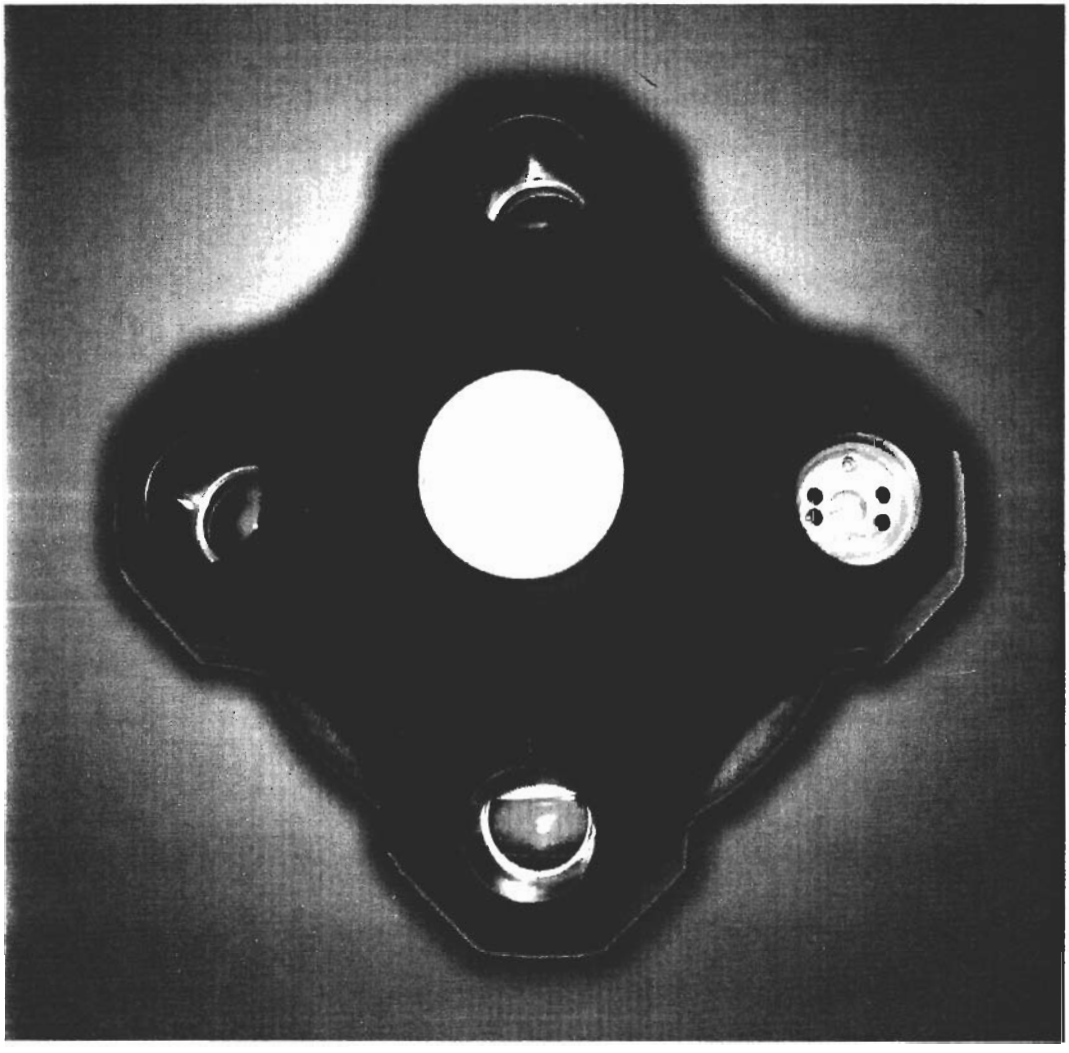
Edited by David J Scott, Stephen E Harding and Arthur J Rowe

Analytical Ultracentrifugation

Techniques and Methods



RSC Publishing

CHAPTER 5

Using Analytical Ultracentrifugation of DNA in CsCl Gradients to Explore Large-Scale Properties of Genomes

OLIVER CLAY, NICOLAS CARELS, CHRISTOPHE J. DOUADY AND GIORGIO BERNARDI

Introductory Remark

Analyses of absorbance profiles of DNA in CsCl density gradients at sedimentation equilibrium, obtained by analytical ultracentrifugation (AUC), played an important role in molecular genetics and genomics during almost half a century (1957–2004). They allowed accurate calculations of GC (base composition) distributions, GC mosaicism and gene densities in vertebrate genomes that have now been amply confirmed via completed sequences, including those of human, mouse, and pufferfish. We here review general principles guiding past and present uses of salt gradient AUC for exploring genomic DNA, and discuss open problems of AUC/CsCl inference that should become tractable with the aid of a few more entirely sequenced vertebrate genomes.

1 Historical Introduction

1.1 Measuring GC in an Ultracentrifuge

One of the most elementary properties of a DNA sequence is its base composition. If the orientations of the base pairs are neglected, the base composition is simply the sequence's GC: the molar fraction of guanine and cytosine or, equivalently, the proportion of the base pairs in the sequence that are GC rather than AT.

An early, elegant discovery opened the way to rapid experimental measurement of GC. In 1957, Meselson *et al.*[1] had introduced the methodology of CsCl gradient

ultracentrifugation: a salt (CsCl) gradient is built up during ultracentrifugation, and a biomolecule such as DNA eventually finds its equilibrium position in that gradient (see ref. 2 for a quantitative introduction to density gradient ultracentrifugation). Two years later, it was found that the GC levels of essentially all DNAs, except for heavily methylated or otherwise modified DNAs, were linearly related to their positions in the CsCl gradient at sedimentation equilibrium, *i.e.*, to the DNAs' buoyant densities in CsCl.[3–6] The GC of an organism's DNA could therefore be routinely determined by AUC at sedimentation equilibrium, which is usually attained in less than 24 h.

1.2 Measuring GC Distributions in an Ultracentrifuge

The CsCl/AUC method allows far more information to be gathered than just the total GC% of a genome, however. This important point went almost unnoticed at first. Indeed, one can fragment a genome in a number of ways (yielding different molar mass or molecular weight distributions, *i.e.*, different fragment length distributions), and then obtain, via AUC, the distribution or histogram of the GC levels of the fragments. Comparisons of such GC distributions within species (at different molar masses) and between species (preferably at similar molar masses) can lead to novel deductions about the large-scale structures of the species' genomes, their functional correlates, and their interspecific differences. The power of this approach is most obvious in the case of warm-blooded vertebrates, which have the widest GC distributions of all taxa, at fragment lengths above 10–30 kb (kilobase pairs; 30 kb of DNA correspond to about 20 MDa.)

Having said this, there were a few technical problems that needed to be overcome before a CsCl absorbance profile, *i.e.*, an AUC scan of a band of DNA at sedimentation equilibrium, could be routinely converted into its underlying GC distribution. At a first approximation, a CsCl absorbance profile is a convolution of the true GC distribution and a Gaussian-shaped diffusion broadening: when the molar mass of the DNA molecules or fragments is below about 50 kb, they will move appreciably around their equilibrium positions, thus widening the CsCl profile. The shorter the DNA fragments, the more the profile is widened. If one can reliably estimate this Gaussian diffusion broadening, one can recover the GC distribution from a DNA sample whose molar mass is as low as 10 kb (see ref. 7 and references therein).

A detailed theoretical treatment of the CsCl equilibrium profile was provided in a series of articles by Vinograd, Hearst and Schmid in the 1960s and early 1970s. If one demands the highest possible precision, one is faced by a multitude of factors that should, in principle, influence the CsCl profile. Such factors include pressure effects, charges on the DNA molecules, DNA methylation (typically present in small amounts), and light bending (see ref. 7 and its online supplement for references). When one is just trying to reconstruct a reasonably accurate GC distribution of a genome, many of these corrections become almost negligible, or almost cancel (*e.g.*, in the case of pressure corrections, if one uses a marker to calculate GC). One of these published corrections that can notably influence the calculations is the correction for virial effects. Another effect arises because DNA, CsCl salt, and water do not form an ideal solution. The solvation (hydration) of the

DNA/CsCl solute can be estimated, and this leads to a corrected expression for the broadening of a CsCl profile as a function of molecular weight and DNA concentration. As a check, the resulting expression was used to estimate the molar masses of intact phages from their CsCl profiles:[8,9] the estimates agreed well with independent calculations of these phages' genomic lengths. For vertebrates, the dependence of the broader CsCl profiles' widths on DNA concentration is more complicated (see Section 6).

1.3 Using the Ultracentrifuge to Probe Vertebrate Genomes: Discounting Diffusion and Repetitive DNA

In the mid-1970s began the first dedicated quantitative analyses of vertebrate profiles. The relative compositional heterogeneity of vertebrates (compared, for example, with bacteria or phages, at scales >10 kb) implies that their CsCl profiles encode non-trivial information, on the large-scale structure of the vertebrates' chromosomes, which can be resolved via AUC.[10–13] A few raw CsCl profiles of calf and some other mammals had been published earlier, but without the diffusion discounting that would have been necessary, given their relatively low molar masses; furthermore, in one or two cases so much DNA was loaded that the maximal absorbance was no longer proportional to the amount of DNA present (saturation). Such conditions did not yet allow quantitative interspecies comparisons.

Another condition made it difficult to progress: the presence of large amounts of highly repetitive DNA, or 'satellite' DNA, in some well-studied species. The species included calf, whose thymus DNA had been a paradigm for mammalian DNA since the early 1950s. The name 'satellite' was motivated by the ultracentrifugation metaphor: such repetitive DNA bands are often (but not always) found outside the main part of a species' CsCl profile (see refs. 11 and 14 for example). This satellite DNA is prone to rapidly contract or expand, on an evolutionary timescale. Indeed, tracts of repetitive DNA can be shrunk or extended either by slippage during replication, in which the template strand and its copy become shifted relative to each other (so that a sequence can afterwards be present in tandem duplicate or not at all), and/or by ectopic homologous recombination between sequences that happen to be similar but are not at corresponding chromosomal positions (so that there is a looping out of intervening DNA, leading to different lengths for the two homologous chromosomes). By eliminating satellite DNA from a GC distribution, we are left with a desired picture of evolutionarily stable DNA, which can be used to characterize large-scale genomic differences between different mammalian or vertebrate orders.

Filipski *et al.*[10] first quantified the contributions of satellite DNA to the bovine genome: it turned out that one-quarter of the genome consisted of such highly repetitive DNA. The large proportion explained why the CsCl profile of calf was broader than those of many other mammals, such as human. It later turned out that calf, other ruminants, and some geomyoid rodents are not typical mammals in this respect: most mammals have far less satellite DNA. The highly repetitive satellite DNA also sometimes bands anomalously, *i.e.*, it can form peaks at positions corresponding to 'wrong' GC levels.[15,16] The satellite peaks are typically narrow

compared to the total profile (see Figure 1 for example), because their DNA has almost constant base composition over long tracts, and because satellite DNA tends to aggregate[12] (see also Section 6). The satellite peaks can differ markedly among closely related species, or even among populations of the same species, because of the instability (rapid contraction, expansion and/or extinction) of the tracts. In a genome that has only few, isolated satellite peaks, the differences will affect the relative heights of the peaks, whereas in a genome that has many or closely overlapping satellite peaks the overall shape of the CsCl profile can change, so that it becomes difficult to resolve the peaks.

The method for experimentally quantifying satellite contributions, used with success for the bovine, mouse and guinea pig genomes,[10,15] was then also applied to other vertebrate genomes. The GC distributions of different taxa could now be compared after satellites had been taken into account.[11,12]

GC distributions measured at different molecular weights (fragment lengths) yield information on the organization of base composition along the chromosomes of a

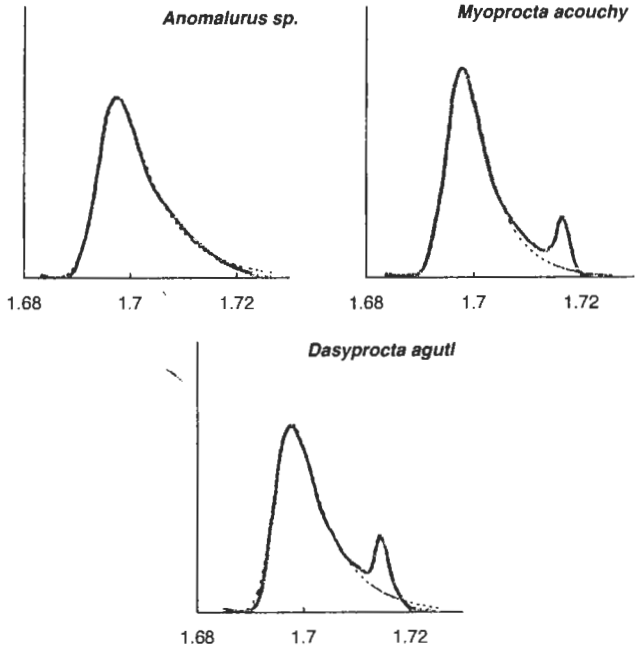


Figure 1 *Equilibrium CsCl absorbance profiles of DNA from three rodent species, Anomalurus sp. (scaly-tailed flying squirrel), Myoprocta acouchy (green acouchi), and Dasyprocta aguti (agouti). All three samples had molar masses above 50 kb, so that the CsCl profiles are essentially the GC distributions of the species' genomes. The horizontal axis is calibrated in buoyant density (g cm^{-3}; see Section 2); the vertical axis shows area-normalized UV absorbance readings at 260 nm, i.e., relative amounts of DNA. Buoyant densities of 1.68, 1.70 and 1.72 correspond to GC levels of 20.4, 40.8, and 61.2%, respectively. The narrow peaks are due to highly repetitive satellite DNA. Dashed lines show fits to a unimodal, asymmetric curve (Gaussian-broadened truncated exponential). Data are from ref. 14*

species. For example, a random sequence of GCs and ATs would yield a GC distribution that narrows rapidly as the molecular weight increases (binomial distribution). In contrast, the GC distributions of mammals such as human and mouse remained invariant – with no narrowing – for all molecular weights above 70 kb that could be analyzed at that time, which extended to well over 300 kb. This was an altogether unexpected result: it implied that mammalian chromosomes are mosaics of long (typically ≫300 kb) regions of fairly homogeneous base composition, *i.e.*, of long GC-rich regions alternating with long GC-poor regions.[12] It should be pointed out that this result, which concerns properties of the genome sequence, was derived via AUC alone, well before DNA sequences were available. The AUC/CsCl analysis of DNA, at different molecular weights, and its interpretation in terms of a mosaic genome structure, were however motivated by the results after fractionating numerous eukaryotic genomes in cesium sulfate gradients in the presence of sequence-specific ligands, and decomposing the fractions' CsCl profiles into Gaussian components.[11,12] The organization of mammalian chromosomes into the long regions, later called isochores,[13] as well as the invariance of the GC distribution in the 100–300 kb range, have now been confirmed, a quarter of a century later, by the full human and mouse sequences.[17,7] Figure 2 shows the molecular weight dependence of human, bacterial, and random DNA, calculated from sequence data; similar curves were obtained much earlier via AUC.[12,13] The lower bound of ≫300 kb for the average isochore length in mammals (which corresponded to experimental limits for measuring molar masses at the time of the original study) remains correct. In fact, many isochores are far longer and extend over several megabases.

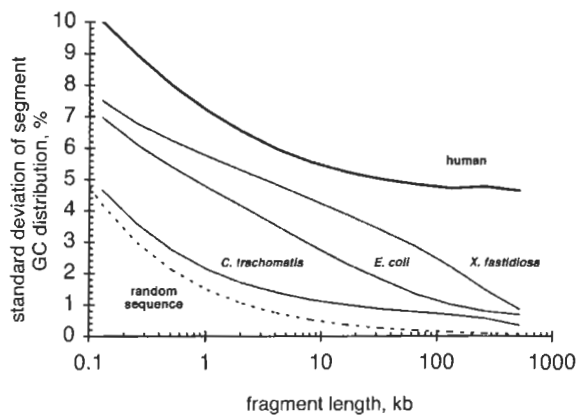


Figure 2 *Standard deviations (widths) of the GC distributions of different genomes, plotted against fragment length on a logarithmic scale. Solid curves, from top to bottom: human, with plateau due to isochore structure; an unusually heterogeneous bacterium, Xylella fastidiosa; a bacterium with intermediate heterogeneity, Escherichia coli; and a very homogeneous bacterium, Chlamydia trachomatis. The dashed curve at bottom left illustrates that no known species is as homogeneous as a random DNA sequence consisting of independent, identically distributed nucleotides. For large fragment sizes, human and other vertebrates have much wider GC distributions than bacteria. Data are from ref. 61*

The invariance of a typical mammal's GC distribution, in the 70–300 kb range, has also a practical implication. One then needs only one CsCl profile in this range, in order to predict the mammal's profiles at all other molar masses in the range (invariance zone): they are essentially identical to the observed profile.

On the topic of repetitive DNA, a final point should be mentioned. In addition to the highly repetitive satellite DNA, there is also another class of repetitive DNA, the so-called interspersed repeats or middle-repetitive DNA. These repeats include the shorter SINES (*e.g.* Alu elements in human) and the longer LINES (*e.g.* L1). Studies of the entire genome sequences have now shown that repeats account for as much as 45% of the human DNA and 35% of the mouse DNA. Since the repeats are interspersed, they have often coevolved with their flanking DNA over quite long evolutionary time spans, they are relatively stable, and they show GC levels approaching those of their local environment. One would therefore not feel authorized to simply 'remove' such repeats from GC distributions, for phylogenetic or genome comparative studies. Instead, it seems most appropriate to regard the interspersed repeats as an integral part of the genomic DNA that coevolves with the unique DNA, and to retain the GC distribution of satellite-free DNA (*i.e.* excluding highly repetitive DNA but including interspersed repeats) as the trait to be analyzed and compared among species. The AUC alone then gives us all the information we need in order to extract the relevant GC distributions.

2 Equations for CsCl Gradients

We now give equations to quantify some of the points made above.

The buoyant density ρ of a molecule in CsCl is defined as the density of the salt solution where that molecule is found, at sedimentation equilibrium. This buoyant density is essentially a linear function of the molecule's radial distance r from the ultracentrifuge axis,[18]

$$\rho = \rho_m + \omega^2(r^2 - r_m^2)/(2\beta_B) \tag{1}$$

where ρ_m and r_m are the buoyant density and radial position of a suitable marker DNA (such as phage 2C of *Bacillus subtilis*, which has $\rho_m = 1.7420$ g cm^{-3}), ω is the angular speed (usually we work at 44 000 rpm) and β_B is $1.190–1.195 \times 10^9$ cgs units for Beckman models E and XL-A under standard conditions; note that $r^2 - r_m^2 \approx 2r_m(r - r_m)$, an expression that is linear in r. In turn, the GC level of a DNA molecule is linearly related to the molecule's buoyant density in CsCl,[6]

$$GC\% = \frac{100\% \times (\rho - 1.660 \text{ g cm}^{-3})}{0.098} \tag{2}$$

so GC is again a linear function of the radial distance r.

Diffusion broadening contributes to the total variance of a CsCl absorbance profile. We formally re-express distances r or buoyant densities ρ as equivalent GC percentages x, specified by Equation (2) above. We can then write the absorbance

profile, *i.e.*, the frequency distribution (probability density) of the DNA's radial position in the AUC cell, as the convolution

$$f(x) = h*\varphi(x) \tag{3}$$

where $h(x)$ is the GC distribution, $\varphi(x)$ is a Gaussian point spread function describing the diffusion broadening, and * denotes the convolution integral. Thus the variances σ^2 add

$$\sigma_f^2 = \sigma_h^2 + \sigma_\varphi^2 \tag{4}$$

The standard deviation σ_h of the GC distribution is often denoted H, and is called the compositional heterogeneity of the DNA. A reasonable estimate of the diffusion contribution to the profile variance is given, at 25 °C and other standard conditions, by

$$\sigma_\varphi^2 \approx 44.5 \text{ kb}/l \tag{5}$$

where l is the average fragment length in kilobases (see ref. 7, and for the full formulae ref. 8). There is actually a slight GC dependence, the formula (5) being intended for use at around 50% GC; at 30% GC the right-hand side should be 44.0 kb/l, at 70% GC, it should be 45.0 kb/l. There is also a concentration dependence, which is however difficult to estimate (see Section 6). We see that when l is 50 kb or higher, the diffusion broadening (standard deviation of the spread function) corresponds to less than 1% GC. Thus, for a vertebrate profile with a heterogeneity of 4.3% GC, the proportion of the total profile variance explained by diffusion is about 0.04.

Vertebrate GC distributions, or raw absorbance profiles, can be modeled by superpositions of Gaussians, although typically this means fitting many parameters. Alternately, after removing narrow satellite peaks one can often model the underlying GC distribution $h(x)$ or absorbance profile $f(x)$ by a convolution of a left-truncated exponential $g(x)$ and a Gaussian,

$$h(x) = g*\psi(x)$$

$$f(x) = h*\varphi(x) = (g*\psi)*\varphi(x) = g*(\psi*\varphi)(x) \tag{6}$$

Note that the convolution of two Gaussians is again a Gaussian; as usual the variances add. This means, for example, that we can estimate the GC distribution $h(x)$ from a fit of the raw CsCl profile: we convolve the fitted $g(x)$ with a Gaussian whose variance is the variance of the fitted Gaussian $\psi*\varphi$ minus the diffusion variance (5). Explicit formulae for g, h, and f in terms of the truncation or cutoff point x_c, exponential rate k, and Gaussian widths can be found in refs. 14 and 7.

3 Anatomy of a CsCl Absorbance Profile

Figure 1 shows, schematically, three CsCl absorbance profiles from a rodent study.[14] The molar mass was above 50 kb for all samples: the diffusion can therefore be

practically neglected and the absorbance profiles faithfully represent the GC distributions. As we have seen, the GC distributions are obtained by a simple linear recalibration of the horizontal axis; only the GC of the narrow peaks, formed by satellites, may be unreliable.

When the satellite peaks are discounted, the underlying profiles of different species of mammals, birds, and many cold-blooded vertebrates have a similar broad, positively asymmetric shape, or functional form. The conserved shape can be characterized to high accuracy by an exponential decrease broadened by a Gaussian. In other words, this functional form is a convolution of a left-truncated exponential function and a Gaussian point spread function (dashed curve in Figure 1). The Gaussian spread function is, in turn, the convolution (Equation 6) of two narrower Gaussians: one of them represents the diffusion, while the other is inherent to the GC distribution. In summary, what differs among mammalian species – generally more so among distantly related species, and less so among closely related species, as one would expect for a phylogenetically informative trait – is not the functional form of the profile, but for example its position along the GC axis and its width.

Given the conservation of the basic profile shape, two or three essential parameters suffice to capture most of the compositional variation among genomes of different species (or higher-level taxonomic groups such as families, suborders, or orders). For many purposes, one position parameter (mode or mean) and one width parameter (standard deviation or asymmetry, *i.e.*, mode–mean difference) are enough. The functional form mentioned above gives two different width parameters, one characterizing an asymmetric (exponential-like) broadening and the other the symmetric (Gaussian) broadening of the GC distribution; the total variance is the sum of these two variances.[14,7]

How does one rid a GC distribution of its satellite DNA components? At the species level and above, such components or peaks typically just represent annoying phylogenetic noise. One approach, which is still the best but quite laborious, is to characterize the satellites experimentally. This involves several steps. In one method,[11,12,15,19] a mercury- or silver-based, oligonucleotide-specific DNA ligand is first added to the DNA. After preparative ultracentrifugation in a cesium sulfate gradient one fractionates the DNA, removes the ligand from the fractions and re-centrifuges them in analytical CsCl gradients, so that one can then observe the behavior of the peaks as one passes from one fraction to the next. Usually this method detects and quantifies all but the most stubbornly cryptic satellites. Subtraction of the satellite components leaves us the GC distribution of the satellite-free DNA.

A simpler expedient, which often gives very satisfactory results, is to simply postulate that the usual, basic shape of vertebrates' satellite-free GC distributions, or CsCl profiles, will be conserved for the taxa that are being examined. The automatic fitting of this special shape to the raw CsCl profile (or, alternatively, to a GC distribution of an entirely sequenced vertebrate) yields estimates of three fit parameters: an exponential decline rate, a position parameter, and a Gaussian spreading parameter. (Details are given in ref. 7 and its online supplement.) Satellite peaks are bypassed by the fitting program (see Figure 1), except in very rare cases (geomyoid rodents, some ruminants), where satellites are so abundant that they form a continuum rather than isolated peaks, and thus distort the profile beyond recognition.

Standard deviation, mean, mode, and so on of the satellite-free GC distribution can be calculated directly from the fit parameters.

4 The Biological Meaning of GC Distributions

4.1 GC-Rich DNA and GC-Poor DNA are Functionally Different

After the technical points discussed so far, we now make an excursion to illustrate why extracting and comparing genomic GC distributions is a biologically relevant pursuit.

Of prime interest in AUC-assisted genomics is the understanding of nuclear genomes, how they are organized, how their organization evolved or changed in different taxa, and how this organization is related to the functional activities of the genomes and their genes. The analysis of GC distributions alone can give deep insights into the workings of vertebrate genomes and the differences among them.

The mosaic structure of mammalian and avian genomes, *i.e.*, the alternating of GC-rich and GC-poor isochores along their chromosomes, has a number of important biological correlates. GC-rich DNA above, say, 50% GC is rare (accounting for not more than about 10% of human DNA, for example). This is seen in the positively asymmetric CsCl profiles of mammals and birds, which have their modes near 40% GC.

In contrast to the scarcity of GC-rich DNA is the abundance of GC-rich genes. For example, the histogram of GC levels of human genes (rather than of human DNA) is almost flat-topped, with even a slight negative asymmetry. Consequently, the gene density in GC-rich regions is much higher than in GC-poor regions. Proper alignment of the two distributions can be achieved using the orthogonal regression line (major axis) of a bivariate landscape or scatterplot, showing GC levels of genes' third codon positions along the vertical axis and GC levels of the surrounding DNA along the horizontal axis. For human genes, this line has a slope close to 3: GC levels of genes residing in GC-rich isochores stand out, compositionally, against their environment (see refs. 20 and 21 and the figures and references therein). GC of third codon positions is a particularly sensitive monitor of the GC level of the DNA surrounding the gene. Indeed, third positions are only mildly constrained by the protein that the gene must encode: in almost all codons the third nucleotide can be changed from G to C, or from A to T, without changing the encoded amino acid.

In human, the gene density was estimated earlier, by aligning the distributions for genes and DNA, to be about 17 times higher in the GC-richest regions of the genome than in the GC-poorest regions.[20,22] This ratio was confirmed by Lander *et al.*,[23] a decade later, using the entire human genome sequence. The GC-rich isochores could thus be viewed as a genome core, and the GC-poorest isochores as gene deserts.

Apart from gene density there are numerous other, functionally important differences between GC-rich and GC-poor isochores. For example, genes residing in GC-rich DNA have shorter introns than genes in GC-poor DNA;[24] repetitive DNA categorically remains at low levels in the GC-richest DNA, but not in the GC-poorest DNA;[25,26] GC-rich DNA has preferentially an open chromatin structure whereas GC-poor DNA is typically in closed, compact chromatin; in GC-rich DNA, CpG

islands (regions having unusually high densities of unmethylated CpG dinucleotides) are frequent and often cover the entire gene in GC-rich DNA, while in GC-poor DNA they are much rarer and typically do not cover much more than the promoter of the gene;[27,28] GC-rich DNA apparently recombines more frequently than GC-poor DNA, in eukaryotes from yeast to human ($R \approx 0.4$),[29,30] so that impressively long haplotype blocks[31] would be unlikely to occur in gene-dense regions; GC-rich isochores replicate early in S phase whereas GC-poor isochores replicate late;[32] the GC-richest DNA tends to loop far into the nucleus during interphase, while the GC-poorest DNA typically stays close to the periphery of the nucleus.[33]

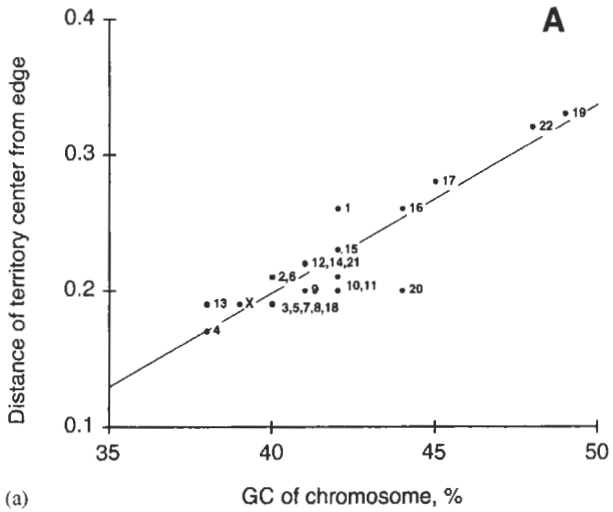
4.2 The Territories of the GC-Richest Isochores are in the Interior of the Interphase Nucleus

This last correlate of GC is a recent discovery that deserves some comment.

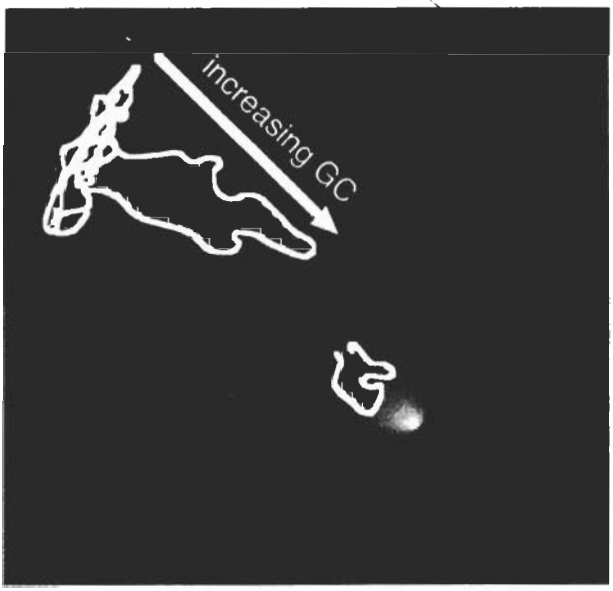
The general topic of chromosomal or sub-chromosomal territories – preferred regions within the cell's nucleus, around which the DNA diffuses – has been discussed in a number of articles in the last few years; we cite here, with their references, Saccone *et al.*,[33] Zink *et al.*,[34] and Boyle *et al.*[35] The latter group have calculated the average position of each human chromosome (centroid) and its distance from the edge of the nucleus, during interphase. We have plotted these distances against the GC of the chromosomes and obtain an impressively linear relation (Figure 3(a); $R=0.91$). At sub-chromosomal scales, the same tendency has been shown directly, by fluorescent *in situ* hybridization (FISH) experiments: mammalian and avian sub-chromosomal regions or bands that hybridize with the GC-richest DNA tend to localize in the interior of the nucleus, while those that hybridize with the GC-poorest DNA localize at the periphery[33] [see Figure 3(b)]. As a caricature, an ultracentrifuge cell could thus be viewed as a direct physical model of the biological cell: the distance into the AUC cell where one finds a large region's DNA at sedimentation equilibrium tells us the average distance that region extends into the interior of the nucleus at interphase.

We still need, however, to clarify how large a sub-chromosomal region must be for its position to depend on GC, a problem that is still largely open. For example, if a 100 kb region having 30% GC were immediately adjacent to a 100 kb region having 60% GC, one would not expect to find the GC-poor region at the nuclear edge, and all of the immediately adjacent GC-rich region far away in the nuclear interior. More generally, what region around a gene best predicts that gene's position (territory) in the nucleus: is it the isochore, a chromosomal band, or a symmetric window of 50, 100, or 1000 kb, for example, that should represent the local GC?

Chromosomal and sub-chromosomal territories have recently received attention in the context of translocations that can lead to medical syndromes or disease, notably cancer. One would intuitively expect that two regions, located on two different chromosomes, are more likely to experience an accidental translocation if they are in close physical proximity during relevant phase(s) of the cell cycle.[36] The data obtained so far for interphase are, however, slightly disappointing. Although known translocation-prone pairs of sub-chromosomal regions in human (such as *MYC:IGH*, involved in Burkitt's lymphoma) are significantly more often close to each other in interphase



(a)



(b)

Figure 3 *(a) Distance of human chromosomes' centroids (territory centers) from the edge of the nucleus at interphase correlates with the chromosomes' average GC level (R=0.91). We have taken the distances from Boyle et al.,[35] where they are normalized for nuclear size, and plotted them against the GC of the chromosomes (labeled; here the data of ref. 62 are used, more recent sequence releases give essentially the same plot). (b) Caricature of GC-rich mammalian or avian DNA looping into the interior of a cell's nucleus during interphase, while the GC-poor DNA remains close to the periphery. The GC-rich DNA for ribosomal RNA genes (rDNA) is an exception, since it loops into separate nucleoli. Adapted from ref. 33*

than pairs that do not undergo translocations (such as *MYC:TGFBR2*), when one looks at the numbers the difference is small (32.7 *vs.* 22.2%, in the observations made by Roix *et al.*[37]). This suggests that physical proximity during interphase is likely to be only one of the several factors determining translocation probabilities. Modified criteria, and/or the inclusion of additional factors, may improve the predictive power of interphase distances – and thus possibly of GC differences – for translocations or other functionally relevant interactions. Such interactions might also include contacts in *trans* that could facilitate transcriptional activation during interphase (*cf.* ref. 38).

4.3 Changes in GC Distributions During Vertebrate Evolution

Homeothermy evolved twice, independently, in the lineages leading to the present-day mammals and birds. Similarly, the GC-richest DNA is found only in mammals and birds. In cold-blooded vertebrates (except some reptiles[39]), the corresponding isochores are clearly GC-poorer, and the contrast between GC-rich and GC-poor isochores did not evolve to the dramatic differences seen in mammals and birds [Figure 4(a)]. In other words, the GC-enrichment in the gene-densest regions of the genome evolved twice, independently, in precisely the two lineages leading to the present-day warm-blooded vertebrates. This and other evidence (see ref. 40 for detailed discussions) suggests that the GC-richest regions of the ancestral genomes became GC-richer in response to a need for higher thermal stability of the DNA (and possibly also RNA or proteins) as the body temperature increased. It therefore appears that primarily the GC-richest DNA loops into the interior of the nucleus at interphase (see above) and is thus 'exposed', while the GC-poorest DNA remains in a more compact form, and in a more closed chromatin structure, near the edge of the nucleus and therefore does not have the same need to be stabilized.

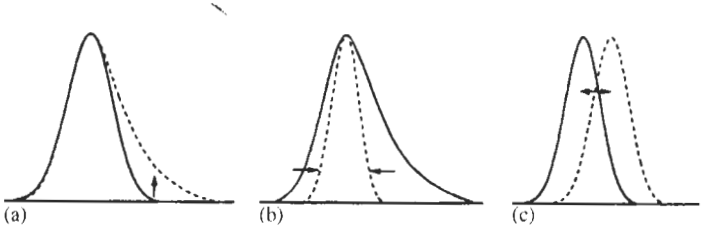


Figure 4 *Schematic portrayal of three elementary compositional shifts that characterize vertebrate evolution. A compositional shift is a relatively rapid, concerted change in GC during evolution, affecting part or all of a taxon's GC distribution. (a) The increase in GC affects mainly the genome's GC-richest regions; such a shift occurred during the transition from cold- to warm-blooded vertebrates (major shift; upward arrow). (b) Both GC-rich and GC-poor DNA is 'eroded' from the genome; such a shift occurred in a group of myomorph rodents that include mouse and rat (minor shift; inward arrows). (c) Genome-wide increase or decrease in GC occurred, for example, during fish evolution (horizontal shift; lateral arrows). Shifts are often composed of two or more of the elementary events depicted here. Conversely, many mammalian orders have witnessed no substantial shifts (conservative mode). Modified from ref. 63*

Mouse, and a number of other myomorph rodents, underwent a 'reverse' shift, well after the GC-rich isochores were already established in mammals (see ref. 41 and the publications cited therein). The mouse has a narrower GC distribution than other eutherian mammals: the GC-poorest and especially the GC-richest DNA is missing in mouse [Figure 4(b)].[41,14]

The murid example illustrates how biological findings can be deduced or predicted by studying and comparing AUC profiles. In view of the biological correlates listed above, the loss of the GC-rich tail in murid profiles suggests that CpG islands should be very different in mouse than in human, and one does indeed find striking differences when one compares the (orthologous) CpG islands of these two species. In mouse the CpG islands are severely 'eroded', covering less of the gene, exhibiting less contrast in CpG density with the surrounding DNA, and in some cases being absent altogether.[27,28,42,43] Such differences tie in well with other observations that have been made for mouse, compared to human and other mammals: its DNA repair is less meticulous, its inactive X chromosome is more often accidentally activated, its methylation-free or hypomethylated genes are more often accidentally methylated, and its substitution rate is higher.[44,45]

Fishes present another type of change. Not only are their GC distributions narrower and more symmetric than, for example, mammals, birds, and some reptiles, but the positions (modes or means) of these distributions on the GC axis are often far from those of most terrestrial vertebrates, ranging from about 37 to 49% GC.[46,47] Fishes have therefore apparently undergone several 'horizontal' shifts: the shape and width of the CsCl profile is often well conserved while its position shifts substantially [Figure 4(c)].

In the context of gene density calculations in human, we discussed the linear relation between the GC of human genes, or their third positions, and the GC of their surrounding DNA. How well are such relations conserved across different species, or how quickly do they change during evolution? If such relations and their equations were perfectly conserved throughout a taxonomic group, the GC level of a sequenced GC-rich gene, *i.e.*, of a 'compositional marker' sequence, would be predictable from the species' AUC profile and would add little new information: a profile with a GC-rich tail would predict a correspondingly high GC for the sequenced marker gene, and *vice versa*. A preliminary analysis,[7] of 24 species/genera from seven eutherian orders for which both the AUC profile and a GC-rich marker sequence were available (exon 28 of the gene for von Willebrand factor, vWF; see ref. 48), suggests that the linear relation linking their GC levels does change, but slowly, in the sense that the relation is typically well conserved within orders but can differ among orders.

5 Comments on Molar Mass Dependence

Molar mass distributions of vertebrate DNA samples are often quite narrow. For our purposes, we can then treat the samples as if they were monodisperse, *i.e.*, as if all the fragments had the same length (the mean). Here, as elsewhere, we assume that the mean molar mass of each sample is at least 10 kb. Extracted DNA with lower mean molar masses will often be polydisperse, may have been degraded and/or partly denatured, and can be difficult to use for comparative AUC studies.

For many species, particularly if they are rare or endangered, we will not have the luxury of high (>50 kb) molecular weight DNA. Our sample may instead have fragments of, say, 15 kb. For another species, such as human or mouse, we may be able to prepare (and inject into the AUC cell using the usual protocol) DNA having a much higher molecular weight, *e.g.*, 90 kb. Such high molecular weight DNA provides more information on the large-scale properties of the genome. Can we then compare the two species? At 15 kb a mammalian GC distribution still broadens as one decreases the molecular weight, whereas at 90 kb it is almost invariant (see Figure 2). From 90 kb fragments, we could obtain the profile for 15 kb fragments by chopping or shearing the 90 kb fragments, but we could not reconstruct 90 kb fragments from 15 kb fragments even in principle, since we would need adjacency information that we do not have. So what can we do? More generally, how can we compare 50 or 200 species, each one represented by a sample that has a different molar mass?

As an expedient, we can simply plot the raw data for the estimated GC distributions, for example with standard deviation on the vertical axis (after using Equation (5) to correct for diffusion) and mean fragment length on the horizontal axis (see ref. 7 for an example). Although we cannot extrapolate through individual points, the general shape of the scatterplot does give a rough idea of how a given GC distribution's width should change as a function of fragment length.

A more satisfactory solution may be possible in future, after studying individual genomes from different orders and comparing, for each species, its CsCl profile widths at different molecular weights. Whereas earlier studies, such as the mouse study of Macaya *et al.*,[12] could only be done experimentally, the sequencing of entire vertebrate genomes now allows accurate studies to be conducted quickly on the computer (see, *e.g.*, ref. 49 for an in-depth study of the human chromosomes). One spin-off of such studies is an indication of conserved properties, shapes, or regions in the standard deviation *vs.* molar mass plots or their equivalent representations, namely correlograms and Fourier spectra (discussed in ref. 49). Progress along such lines should allow results from a few sequenced model genomes to be extrapolated to the many vertebrate taxa for which we have no genome sequence. Indeed, for many known species we will have at most an AUC profile and possibly one or two sequenced nuclear genes, for example from gene-specific sequencing projects that were designed to cover a wide range of taxa (see, *e.g.*, refs. 48 and 50). At the time of writing, entire genome sequences of vertebrates are available for human, chimpanzee, mouse, rat, and two species of pufferfish, and for parts of the chicken, dog, and zebrafish genomes; a few sequenced contigs exist even for platypus. The systematic study of such long genomic sequences may soon tell us where and how we could normalize GC distributions for molecular weight.

6 Comments on Concentration Dependence

Concentration dependence of vertebrate CsCl profiles is a more complicated problem than molecular weight dependence, for several reasons. Luckily its effects are less pronounced, and much of what follows should be unnecessary if one just wants to extract a reasonable estimate of a species' GC distribution.

If we load twice as much DNA into the ultracentrifuge cell, from the same sample, we might expect the absorbance profile to be twice as high, at all parts of the band that forms at sedimentation equilibrium.

Instead, we observe slight, systematic distortions. In other words, the area-normalized CsCl profile changes as a function of the total DNA loaded or, correspondingly, as a function of the peak concentration or maximum absorbance. This is true even when maximum concentrations remain within the recommended range, *i.e.*, giving a maximum absorbance (optical density) between about 0.3 and 1.0. One would like to predict how the profile will change, *i.e.*, if and how one can normalize for concentration in general, but this is still a largely open problem. Unlike the molar mass problem, this problem is entirely an AUC problem, since obviously the true GC distribution will not depend on the amount of the same DNA that one analyzes.

We begin by considering a sample of identical DNA molecules, so that its GC distribution is, in principle, an infinitely sharp peak (delta function). What will happen to the sample's CsCl profile as we increase the amount of DNA loaded, beginning at zero (infinite dilution)? When the amount loaded is low, we are presumably in the domain of classical virial effects. The DNA and CsCl in water will not be an ideal solution, and there will be solvation (hydration). Increasing the total concentration will broaden the profile; typically the width will increase exponentially with increasing concentration.[51,9] As the amount of DNA loaded increases, however, there is a possibility that some of the molecules will aggregate, which would increase the effective molar mass and thus narrow the profile. The conditions for aggregation are still poorly understood, but it is likely to occur, for example, in satellite DNA (or when molecules/fragments have 'sticky' ends).[12]

Vertebrate profiles have relatively broad, and often asymmetric GC distributions, which compound our problem. A peak region (high concentration) may then be narrowed by aggregation at the same time as the GC-rich tail (low concentration) is broadened by virial effects, to different extents at different parts of the tail. In technical terms: if the two opposing effects of concentration could be modeled, we would afterwards face not just a simple convolution problem, but a complicated, general folding problem. Indeed, whereas molar mass does not depend on GC in a typical preparation of DNA (no correlation between molar mass and GC),[52] local concentration obviously does. In summary, changing the total concentration can lead to local, non-monotonous horizontal contractions and expansions of the profile. These may, in turn, slightly shift the mode in seemingly unpredictable ways. One way to approach the problem is by experimentally obtaining plots of profile parameters (mode, standard deviation) *vs.* maximum absorbance (profile height at the mode). When this is done for bacteriophages, the plots of log(standard deviation) *vs.* maximum absorbance have slopes that differ from one phage species to another,[51,9,8] suggesting that the slope can depend on the genomes' sequences. Even for these essentially infinitely narrow GC distributions, the only way to gauge the effect of concentration was to obtain profiles for several concentrations and then extrapolate to infinite dilution. Comparing the GC distributions of entirely sequenced vertebrate or other genomes with their AUC profiles may now give better insight and show how concentration-related profile distortions are related to GC and/or sequence features other than GC.

7 Conclusion and Perspectives

We have discussed some of the applications of CsCl gradient AUC in genomics. In particular, we have described large-scale GC variations and other GC-related genomic features that can be deduced from CsCl profiles, either alone or in conjunction with sequence data (where sparse sequence data often suffice).

CsCl gradient ultracentrifugation also serves other uses. Early investigations employed the labeling of DNA and/or RNA. Two well-known experiments of this type are that of Meselson and Stahl,[53] who showed that DNA replication is semi-conservative, and that of Hall and Spiegelman[54] (discussed in ref. 2), who showed the existence *in vivo* of RNA molecules, now called messenger RNAs (mRNAs), which are complements of one strand of a DNA genome. In the first experiment, DNA that had incorporated heavy nitrogen (N^{15}) had a shifted AUC peak; in the second experiment DNA was labeled with H^3, RNA with P^{32}, and preparative ultracentrifugation in CsCl was followed by fractionation and assessment of the levels of labeled DNA and RNA in each fraction.

CsCl profiles and their parameters (*e.g.*, mode, standard deviation) can also be used to fine-tune phylogenies, or to highlight regions of accepted phylogenies that may benefit from closer attention. We have found that AUC profile data often confirm established phylogenies, but differ from them in some regions that are widely regarded as tentative or controversial (see, *e.g.*, ref. 14). While it might be tempting to propose profile parameters as phylogenetically faithful traits, such reliance can mislead. If only CsCl data are available for a group of species, compositional shifts (see above) could be confounded with drifts that would indicate large distances. A more prudent course is to use AUC profile data together with phylogenies obtained from independently gathered sequence or other data. One should however keep in mind that nuclear protein-coding genes will be subject to the same GC shifts as the species' CsCl profiles, so third and often first codon positions must be checked for interspecific stationarity before they are used for phylogenetic purposes.[55]

Another application of CsCl gradient AUC is in analyzing mixtures of genomes of naturally associating (symbiotic or parasitic) species. By locating the peaks of the resulting CsCl landscape and inferring their GC levels, one can screen possible candidate species, unambiguously identify the species by follow-up studies, and quantify their relative abundances in the mixture or community to which they belong *in vivo*.[56,57] The extrapolation of this principle to diagnostics, *e.g.*, for quick GC-based identification of the microbial species in a sample, seems straightforward and promising. Indeed, the GC levels of prokaryotes span a wide range, their profiles are typically narrow compared to this range (see Figure 2), and different species are therefore likely to be resolved by AUC.

A use of the CsCl method that has recently been revived is for detecting, quantifying, and monitoring protein–DNA clamp formations or other salt-stable complexes or interactions (*e.g.*, refs. 58–60). The application has so far been largely limited to studies of topoisomerase and salt-stable clamps that are associated with it, but the principle is general: well-spaced peaks tell the relative abundances of salt-stable complexes between DNA and 1, 2, 3, ... protein molecules.

It seems fair to conclude that CsCl gradient AUC of DNA remains a powerful tool in functional, comparative, and evolutionary genomics, as well as in molecular biology. It can be hoped that increasingly varied and creative applications will be devised in the near future, as more researchers reconsider this method when designing experiments.

Acknowledgements

We thank Salvatore Bocchetti for the technical AUC work that made our recent studies possible. We also thank Kamel Jabbari and Salvatore Saccone for helpful discussions, Maria Costantini for sharing data, and Laura Giangiacomo for literature references.

References

1. M. Meselson, F. W. Stahl and J. Vinograd, *Proc. Natl. Acad. Sci. USA*, 1957, **43**, 581.
2. K. E. van Holde, W. C. Johnson and P. S. Ho, *Principles of Physical Biochemistry*, Prentice-Hall, Upper Saddle River, NJ, 1998.
3. N. Sueoka, J. Marmur and P. Doty, *Nature*, 1959, **183**, 1429.
4. J. Marmur and P. Doty, *Nature*, 1959, **183**, 1427.
5. R. Rolfe and M. Meselson, *Proc. Nat. Acad. Sci. USA*, 1959, **45**, 1039.
6. C. L. Schildkraut, J. Marmur and P. Doty, *J. Mol. Biol.*, 1962, **4**, 430.
7. O. Clay, C. Douady, N. Carels, S. Hughes, G. Bucciarelli and G. Bernardi, *Eur. Biophys. J.*, 2003, **32**, 418.
8. C. W. Schmid and J. E. Hearst, *Biopolymers*, 1972, **11**, 1913.
9. J. E. Hearst and C. W. Schmid, *Meth. Enzymol.*, 1973, **27**, 111.
10. J. Filipski, J. P. Thiery and G. Bernardi, *J. Mol. Biol.*, 1973, **80**, 177.
11. J. P. Thiery, G. Macaya and G. Bernardi, *J. Mol. Biol.*, 1976, **108**, 219.
12. G. Macaya, J. P. Thiery and G. Bernardi, *J. Mol. Biol.*, 1976, **108**, 237.
13. G. Cuny, P. Soriano, G. Macaya, and G. Bernardi, *Eur. J. Biochem.*, 1981, **115**, 227.
14. C. Douady, N. Carels, O. Clay, F. Catzeflis and G. Bernardi, *Mol. Phylogenet. Evol.*, 2000, **17**, 219.
15. G. Corneo, E. Ginelli, C. Soave and G. Bernardi, *Biochemistry*, 1968, **7**, 4373.
16. R. D. Wells and J. E. Blair, *J. Mol. Biol.*, 1967, **27**, 273.
17. A. Pavlícek, J. Paces, O. Clay and G. Bernardi, *FEBS Lett.*, 2002, **511**, 165.
18. J. B. Ifft, D. M. Voet and J. Vinograd, *J. Phys. Chem.*, 1961, **65**, 1138.
19. J. Cortadas, G. Macaya and G. Bernardi, *Eur. J. Biochem.*, 1977, **76**, 13.
20. S. Zoubak, O. Clay and G. Bernardi, *Gene*, 1996, **174**, 95.
21. S. Cruveiller, K. Jabbari, O. Clay and G. Bernardi, *Brief. Bioinform.*, 2003, **4**, 43.
22. D. Mouchiroud, G. D'Onofrio, B. Aïssani, G. Macaya, C. Gautier and G. Bernardi, *Gene*, 1991, **100**, 181.
23. E. S. Lander *et al.*, *Nature*, 2001, **409**, 860.
24. L. Duret, D. Mouchiroud and C. Gautier, *J. Mol. Evol.*, 1995, **40**, 308.
25. J. Paces, R. Zíka, V. Paces, A. Pavlícek, O. Clay and G. Bernardi, *Gene*, 2004, **333**, 135.
26. A. Pavlícek, K. Jabbari, J. Paces, V. Paces, J. Hejnar and G. Bernardi, *Gene*, 2001, **276**, 39.
27. B. Aïssani and G. Bernardi, *Gene*, 1991, **106**, 173.
28. B. Aïssani and G. Bernardi, *Gene*, 1991, **106**, 185.
29. J. L. Gerton, J. DeRisi, R. Schroff, M. Lichten, P. O. Brown and T. D. Petes, *Proc. Natl. Acad. Sci. USA*, 2000, **97**, 11383.

30. S. M. Fullerton, A. Bernardo Carvalho and A. G. Clark, *Mol. Biol. Evol.*, 2001, **18**, 1139.
31. M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander, *Nat. Genet.*, 2001, **29**, 229.
32. C. Federico, S. Saccone, and G. Bernardi, *Cytogenet. Cell Genet.*, 1998, **80**, 83; C. Federico, L. Andreozzi, S. Saccone and G. Bernardi, *Chromosome Res.*, 2000, **8**, 737.
33. S. Saccone, C. Federico and G. Bernardi, *Gene*, 2002, **300**, 169.
34. D. Zink, A. Bolzer, C. Mayr, W. Hofmann, N. Sadoni and K. Überla, *Gene Ther. Mol. Biol.*, 2001, **6**, 1.
35. S. Boyle, S. Gilchrist, J. M. Bridger, N. L. Mahy, J. A. Ellis and W. A. Bickmore, *Hum. Mol. Genet.*, 2001, **10**, 211.
36. T. Pederson, *Nat. Genet.*, 2003, **34**, 242.
37. J. J. Roix, P. G. McQueen, P. J. Munson, L. A. Parada and T. Misteli, *Nat. Genet.*, 2003, **34**, 287.
38. H.-P. Müller and W. Schaffner, *Trends Genet.*, 1990, **6**, 300.
39. S. Hughes, O. Clay and G. Bernardi, *Gene*, 2002, **295**, 323.
40. G. Bernardi, *Structural and Evolutionary Genomics: Natural Selection in Genome Evolution*, Elsevier Science, Amsterdam, 2004.
41. G. Bernardi, *Gene*, 2000, **259**, 31.
42. K. Matsuo, O. Clay, T. Takahashi, J. Silke and W. Schaffner, *Somat. Cell Mol. Genet.*, 1993, **19**, 543.
43. F. Antequera and A. Bird, *Proc. Natl. Acad. Sci. USA*, 1993, **90**, 11995.
44. R. Holliday, *Understanding Ageing*, Cambridge University Press, Cambridge, UK, 1995.
45. W. H. Li, *Molecular Evolution*, Sinauer, Sunderland, MA, 1997.
46. G. Bernardi and G. Bernardi, *J. Mol. Evol.*, 1990, **31**, 282.
47. G. Bucciarelli, G. Bernardi and G. Bernardi, *Gene*, 2002, **295**, 513.
48. O. Madsen, M. Scally, C. J. Douady, D. J. Kao, R. W. DeBryk, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong and M. S. Springer, *Nature*, 2001, **409**, 610.
49. W. Li and D. Holste, *Phys. Rev. E*, 2005, **71**, 041910.
50. W. J. Murphy, E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. J. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. De Jong, and M. S. Springer, *Science*, 2001, **294,** 2348.
51. C. W. Schmid and J.E. Hearst, *J. Mol. Biol.*, 1969, **44**, 143.
52. N. Sueoka, *Proc. Natl. Acad. Sci. USA*, 1959, **45**, 1480.
53. M. Meselson and F.W. Stahl, *Proc. Natl. Acad. Sci. USA*, 1958, **44**, 671.
54. B. D. Hall and S. Spiegelman, *Proc. Natl. Acad. Sci. USA*, 1961, **47**, 137.
55. C. Saccone, C. Lanave, G. Pesole and G. Preparata, *Meth. Enzymol.*, 1990, **183**, 570.
56. M. Costantini, B. Lafay and G. Matassi, *Boll. Mus. Ist. Biol. Univ. Genova*, 2002, **66-67**, 48.
57. M. Costantini, *Gene*, 2004, **342**, 321.
58. R. E. Depew, L. F. Liu and J. C. Wang, *J. Biol. Chem.*, 1978, **253**, 511.
59. T. Hu, S. Chang and T. Hsieh, *J. Biol. Chem.*, 1998, **273**, 9586.
60. V. H. Oestergaard, L. Bjergbaek, C. Skouboe, L. Giangiacomo, B.R. Knudsen and A.H. Andersen, *J. Biol. Chem.*, 2004, **279**, 1684.
61. P. Bernaola-Galvàn, J. L. Oliver, P. Carpena, O. Clay and G. Bernardi, *Gene*, 2004, **333**, 121.
62. J. C. Venter et al., *Science*, 2001, **291**, 1304.
63. G. Bernardi, S. Hughes and D. Mouchiroud, *J. Mol. Evol.*, 1997, **44** (Suppl 1), S44.