OPINION

The evolution of isochores

Adam Eyre-Walker and Laurence D. Hurst

One of the most striking features of mammalian chromosomes is the variation in G+C content that occurs over scales of hundreds of kilobases to megabases, the so-called 'isochore' structure of the human genome. This variation in base composition affects both coding and non-coding sequences and seems to reflect a fundamental level of genome organization. However, although we have known about isochores for over 25 years, we still have a poor understanding of why they exist. In this article, we review the current evidence for the three main hypotheses.

With sequencing almost complete, it is tempting to forget how large the human genome is (~3.4×10^9 base pairs (bp) in size). However, only a small fraction of this sequence is known to have any function. It is estimated that there are ~30,000 genes in the human genome[1,2] that produce mRNAs that are on average 1,500 bp in length; so, less than 2% of the genome codes for proteins. A similar amount of DNA might be involved in gene regulation and chromosome structure[3], but most seems to have no function, and so has been called 'junk' DNA. However, this junk DNA is not without structure because base composition (that is, the proportions of A, C, T and G) varies along chromosomes over a large scale. For example, the telomeric 10 Mb of 17q is 50% G and C, whereas that of the adjacent 3.9 Mb of the chromosome is only 38% G and C[1]. This shows that genomes, like organisms, have an anatomy. But is this anatomy the consequence of selection, or is it a by-product of another cellular process?

This large-scale variation in base composition was discovered almost 30 years ago by Bernardi and colleagues[4]. They separated bovine genomic DNA, which had been sheared into large fragments, according to its G+C content, by ultracentrifugation, and found that there was substantial variation in its composition. Subsequent studies showed that compositional variation was a feature of the genomes of both mammals and birds, and that the G+C content of large (>300-kb) blocks of DNA varied from ~35 to 55% in a typical mammal[5]. It was originally thought that this variation in base composition was arranged in 'isochores', large blocks of DNA of homogeneous G+C content that were separated by borders of sharp transition (FIG. 1a). In reality, it seems that only some parts of human chromosomes fit this model. The human major histocompatibility (MHC) locus is a case in point; the MHC class II and III regions have consistent G+C contents of ~40 and ~52%, respectively, separated by a BOUNDARY of sharp transition (FIG. 1b). But this picture breaks down in the MHC class I region, in which G+C content varies between 52 and 42% with no obvious structure. So, although it is clear that much of the genome does not fit the classic isochore model, we use the term 'isochore' in this review to refer generally to large regions of the genome that contain local similarities in base content.

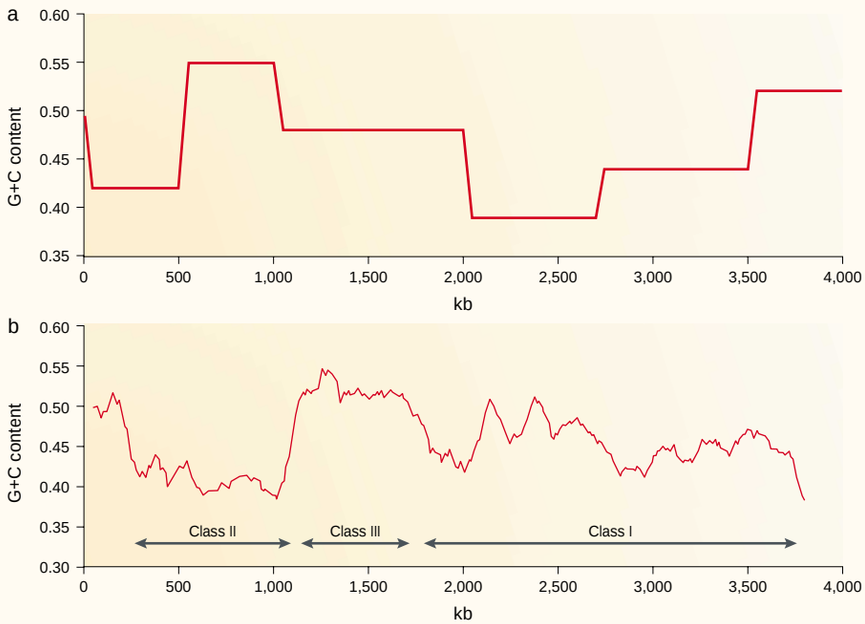

Figure 1 | **Large-scale variation in G+C content. a** | The classic isochore model. **b** | G+C content across the three classes of human major histocompatibility (MHC) region of chromosome 6 (data from GenBank); G+C content is plotted as a moving average, and the window size is 100 kb, advanced by 10 kb each step.

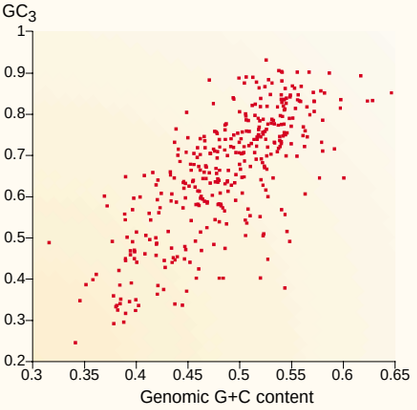

Figure 2 | **Correlation between G+C content of a gene and that of its surrounding region.** The correlation between GC_3 and the G+C content of the 50 kb that surrounds (~25 kb each side) each of 369 genes on human chromosomes 21 and 22 (data from GenBank). Only annotated genes that are multiples of three and that include a stop codon are included. The correlation coefficient is 0.73, which is highly significant ($p < 0.0001$).

Isochores reflect a level of genome organization. This is because gene density and that of short interspersed repetitive DNA elements (SINEs), as well as recombination frequency, are higher in the (G+C)-rich parts of the genome, whereas long interspersed repetitive DNA elements (LINEs) are almost exclusively restricted to the (G+C)-poor parts of the genome[1,5]. Furthermore, the two isochore boundaries that have been studied in detail seem to represent a boundary in more than composition: the isochore boundary between the MHC class II and III regions is reflected in the different times at which these regions replicate (with the (G+C)-rich region replicating earlier)[6]; the boundary at the neurofibromatosis (NF1) region is reflected in differing recombination rates (with the (G+C)-rich region showing higher recombination levels)[7].

The G+C content of a gene is highly correlated to the G+C content of the region of the genome in which it is found[8,9]. This is particularly evident at the largely SILENT, third-codon position (denoted GC_3, to indicate the proportion of codons that end in G or C) (FIG. 2), but is also evident at the first two codon positions (denoted GC_{12})[10]. This shows that systematic variation in the pattern of base SUBSTITUTION exists across the genome, as almost all substitutions in coding sequences are base substitutions; insertion and deletion mutations occur, but they rarely become FIXED. It is this variation in the pattern of base substitution that we focus on in this review, and as a consequence we concentrate most of our attention on explaining the variation in third-position G+C content (GC_3). It should be noted that although the variation in GC_3 is much greater than the variation in isochore G+C content (~30–90% for GC_3, versus ~35–60% for isochores[9]), the correlation is very strong (FIG. 2). This indicates that whatever causes the variation in GC_3 is the main determinant of the variation in isochore G+C. It seems likely that isochore G+C content is less variable than GC_3 because of the integration of repetitive DNA elements in non-coding DNA (see below).

The proposed causes

There has been considerable interest over the past 15 years in explaining why there is large-scale variation in base composition along chromosomes. It has been suggested that the variation could be a consequence of three processes: mutation bias[11–13], natural selection[14–16], or BIASED GENE CONVERSION (BGC)[17,18]. These hypotheses can be grouped into two categories: those that involve natural selection and those that do not. However, the effects of BGC, when mathematically modelled, are considered to be equivalent to those of weak DIRECTIONAL SELECTION[19], so natural selection and BGC are often grouped together.

These hypotheses are not mutually exclusive — two or more of the processes could be acting together, or selection could be acting on the pattern of mutation bias or the process of BGC. However, for simplicity we assume that isochores are a consequence of one process, and that the process is responsible for both their formation and subsequent maintenance.

"There has been considerable interest over the past 15 years in explaining why there is large-scale variation in base composition along chromosomes."

Mutation bias. Mutation bias is probably the most appealing of these hypotheses; it offers a very simple explanation for why the composition of coding sequences is correlated to that of the region of the genome in which the gene is found, and there are at least three simple molecular mechanisms by which variation in base composition can arise. The most elegant of these was suggested by Wolfe and colleagues[13]. They noted that others had made three discoveries: first, that the pattern of base misincorporation during DNA replication is affected by the concentrations of the free nucleotides (for example, G and C nucleotides are preferentially misincorporated into DNA if the DNA is replicated in a pool of free nucleotides rich in G and C (REFS 20–22); second, that free nucleotide concentrations vary during the cell cycle[23,24]; and last, that some parts of the genome are generally replicated early, whereas others are replicated late[25]. So, regions of the genome that replicate at different times should have different mutation patterns and, therefore, different compositions. However, these observations were made in somatic cell lines and not in the germ line (only mutations in the germ line contribute to evolution), and direct evidence for a relationship between G+C content and replication time is ambiguous. Most of the (G+C)-rich chromosome bands replicate in the first half of S phase[26], and the class III region of the MHC, replicates before the class II region, which is less (G+C)-rich[6]. However, the class III region around the *TNFα* (tumour-necrosis factor-α) gene replicates before the region around the *TNFX* gene[6], despite having a slightly lower G+C content[27] than the *TNFX* region. Furthermore, there is no clear, general relationship between GC_3 (or isochore G+C content) and replication time in humans and in mice[28].

Filipski[11] has suggested that variation in the efficiency of DNA repair might be responsible for the formation and maintenance of isochores, as the efficiency of certain types of DNA repair is known to vary across the genome[29], and because some types of repair are biased. For example, base mismatches introduced into human cell lines are preferentially repaired to GC (REF. 30). So, variation in repair efficiency should cause variation in the pattern of mutation. However, theoretical analyses show that variation in base composition is limited according to this model under most conditions[31], and repair has never been shown to vary over the scales needed to generate isochores.

Recently, Fryxell and Zuckerkandl[32] suggested that isochores are a consequence of CYTOSINE DEAMINATION. The deamination of methyl-cytosine and cytosine (that is, C→T and C→U, respectively) is expected to occur more readily in (A+T)-rich DNA because (A+T)-rich DNA is more unstable than (G+C)-rich DNA (see below). This could then lead to isochores through a positive feedback loop; if a sequence becomes (G+C)-rich for some reason (for example, because it codes for a protein), this could lead to a reduction in cytosine deamination and an increase in the G+C content in surrounding areas. This might lead to an isochore structure, but there is at least one problem with this theory. Using this model, isochores would be expected to grow, but

there is no evidence of that; the mammalian isochore structure seems to have been stable since the mammalian radiation, except in rodents[33].

Selection. Giorgio Bernardi — one of the original discoverers of isochores[4,34,35] — has argued, using various evidence, that isochores are the consequence of natural selection. Although there is little evidence against this possibility, it raises the difficult question of why natural selection should be acting on millions of base pairs of non-coding DNA. The leading hypothesis is that selection is acting upon the thermal stability of DNA, because (G+C)-rich DNA tends to be more thermally stable than (A+T)-rich DNA, and the two main groups of organisms that have (G+C)-rich isochores — birds and mammals — are HOMEOTHERMS with high body temperatures[5]. However, the precise relationship between body temperature and genomic base composition has not been studied in detail. Although most fish, amphibians and reptiles show little evidence of large-scale variation in composition[36], some reptiles, such as the Nile crocodile and the red-eared slider turtle, which represent two highly diverged reptile lineages, show marked variation in GC_3 values[37]. Furthermore, the GC_3 values of these reptiles are correlated to those of homologous genes in chicken, which indicates that they are likely to have isochores, just as birds do, because there is a correlation between GC_3 and isochore G+C content in birds[5]. Interestingly, genomic G+C content and mean GC_3 are not correlated to optimal growth temperature in bacteria (when the relationships between bacteria are accounted for), which shows that high G+C content is not a prerequisite for coping with high temperature[38,39].

Bernardi and colleagues have argued that isochores are a consequence of selection because genes in the (G+C)-rich isochores yield proteins with different amino-acid compositions[10] and hydropathies[40] to those in the (G+C)-poor isochores; both features seem to be a consequence of the correlation between isochore G+C content and GC_{12}. The case for selection could be argued in two ways. First, isochores could be present to influence amino-acid composition; however, this seems unlikely because if a certain amino-acid composition is advantageous then it will change, but this should not affect the G+C content of the sequence that surrounds the gene. Second, it could be argued that isochores must be a consequence of selection, if all the amino-acid sites are under selection, otherwise isochore G+C content would not affect amino-acid composition. This seems logical, but proves nothing, as some amino-acid-altering mutations are likely to be NEUTRAL.

Biased gene conversion. BGC is thought to arise during HOMOLOGOUS RECOMBINATION through the formation of HETERODUPLEX DNA. This leads to a base mismatch if the heteroduplex extends across a heterozygous site. These base mismatches are sometimes repaired by the DNA-repair machinery, but this process tends to be biased, leading to an excess of one allele in gametes. For example, base mismatches tend to be repaired to GC in mammalian cell lines[30]. So, the variation in recombination rate across a genome will cause variation in G+C content if the rate of BGC is sufficiently high.

The suggestion that BGC might cause isochore formation comes from two observations indicating that a correlation exists between the rate of recombination and G+C content. First, there is correlation between the frequency of recombination and G+C content both between and within human chromosomes[1,18,41,42]. Second, sequences that have stopped recombining are either declining in G+C content, or have a lower G+C content than their recombining PARALOGUES[18].

Glossary

ALU ELEMENT
A dispersed, intermediately repetitive, 300-bp DNA sequence, ~1,000,000 copies of which exist in the human genome.

BIASED GENE CONVERSION
(BGC). Gene conversion is a non-reciprocal recombination process that causes one sequence to be converted into the other. BGC is when the two possible directions occur with unequal probabilities.

BOUNDARY (ISOCHORE)
A genomic region in which base composition changes markedly between regions of homogeneous composition.

CODON USAGE BIAS
Unequal frequencies, in a protein-coding sequence of DNA, of the alternative codons that specify the same amino acid.

CYTOSINE DEAMINATION
The reaction of a water molecule with the amino-group on position 4 of the pyrimidine ring of cytosine, which results in the conversion of cytosine to uracil. The deamination of methyl-cytosine converts cytosine to thymine.

DIRECTIONAL SELECTION
Natural selection that acts to promote the establishment of a particular mutation.

EFFECTIVE POPULATION SIZE
(N_e). The size of a population as determined by the number of individuals who contribute to the next generation. N_e is related to, but never exceeds, the actual population size (N).

FIXATION (ALLELE)
When an allele replaces all other alleles in a population, so that its frequency is equal to one (100%).

GENETIC DRIFT
The random fluctuation that occurs in allele frequencies as genes are transmitted from one generation to the next. This is because allele frequencies in any sample of gametes perpetuating the population might not represent those of the adults in the previous generation.

HETERODUPLEX DNA
A double-stranded DNA molecule (or DNA–RNA hybrid), in which each strand is of a different origin.

HOMEOTHERM
An organism that uses cellular metabolism to stabilize its own body temperature.

HOMOLOGOUS RECOMBINATION
The process by which segments of DNA are exchanged between two DNA duplexes that share high sequence similarity.

LINE ELEMENT
Long, interspersed sequences, such as *L1*, generated by retrotransposition.

NEUTRAL MUTATION
A mutation that is selectively equivalent to the allele from which it arose.

PARALOGUE
A locus that is homologous to another in the same genome.

PSEUDOGENE
A DNA sequence originally derived from a functional protein-coding gene that has lost its function owing to the presence of one or more inactivating mutations.

SILENT CODON POSITION
One at which a nucleotide change is not accompanied by an amino-acid change in the translation product.

SINE ELEMENT
Short, interspersed, repetitive sequences, such as *Alu* elements, generated by retrotransposition.

SUBSTITUTION
A mutation that has become fixed and, therefore, shows a sequence difference between orthologous sites in different species.

SYNONYMOUS CODON
One at which a nucleotide change does not alter the amino acid encoded.

SYNONYMOUS SUBSTITUTION RATE
The number of synonymous changes per synonymous site.

TRANSITION
A point mutation in which a purine base (A or G) is substituted for a different purine base, and a pyrimidine base (C or T) is substituted for a different pyrimidine base; for example, an AT→GC transition.

TRANSVERSION
A point mutation in which a purine base is substituted for a pyrimidine base and vice versa; for example, an AT→CG transversion.

But these observations do not establish causation and there are at least two problems with the BGC hypothesis: parameter sensitivity and the GC_3 values of some Y-linked genes. The effect of BGC on base composition is highly dependent on the EFFECTIVE POPULATION SIZE (N_e); if two species differ by as little as tenfold in their effective population sizes, the effect of BGC could be undetectable in the species with the lower N_e, owing to the effects of GENETIC DRIFT, but could convert every site not under selection to G or C in the species with the larger N_e. However, although mammalian effective population sizes seem to vary by at least an order of magnitude (P. Keightley and A.E.-W., unpublished data), there is little apparent variation in isochore structure; almost all mammals, except rodents[43], have isochores of similar composition, a pattern that seems not to have altered since the principal mammalian groups diverged from one another[33]. Second, the GC_3 values of some human genes that have been Y-linked (and therefore not recombining) for over 100 million years are high. For example, the human *SRY* (sex-determining region Y) gene has a GC_3 value of 60%, which is close to the average GC_3 value of human genes (~61%)[9]. This indicates that BGC is the sole cause of neither SYNONYMOUS CODON bias in humans nor, by inference, base composition bias.

"…genomes, like organisms, have an anatomy. But is this anatomy the consequence of selection, or is it a by-product of another cellular process?"

The observations

So far, we have considered the three main hypotheses proposed to explain the formation of isochores and the evidence that is pertinent to them. We now turn to several observations that seem to have the potential to differentiate between these three hypotheses, but that have generally failed to do so.

Pattern of substitution. It has been shown that PSEUDOGENES and repetitive DNA elements have a substitution pattern that matches the composition of the genomic region in which they appear[1,44–46]. For example, Francino and Ochman[46] recently showed that an α-globin pseudogene located in a (G+C)-rich part of the genome had a (G+C)-biased substitution pattern, whereas a β-globin pseudogene, located in a (G+C)-poor region of the genome, had an (A+T)-biased substitution pattern. It has been argued that, as pseudogenes are not subject to selection, variation in the pattern of substitution reflects variation in the pattern of mutation and, therefore, that isochores are a consequence of mutation bias[45,46]. However, this analysis only shows that the data are consistent with mutation bias; the authors assume that pseudogenes are not subject to selection (or to BGC) to deduce that isochores are not subject to selection (or to BGC). Yet pseudogenes are in isochores, so if isochores are under selection (or BGC), so are pseudogenes. In fact, the variation in the pattern of substitution in pseudogenes is consistent with mutation bias, selection and BGC; (G+C)-rich parts of the genome must have a (G+C)-biased substitution pattern or they will become (G+C)-poor, just as (A+T)-rich parts of the genome must have an (A+T)-biased substitution pattern.

Single nucleotide polymorphisms. Although the pattern of substitution does not yield any insight into the evolution of isochores (except to establish that they are at equilibrium), the pattern of mutations that segregate in a species is much more informative and can be investigated using the large amount of single nucleotide polymorphism (SNP) data that have recently become available. In mammalian MHC genes and human genes in general, it has been shown that an excess of new A or T mutations segregate at sites that were ancestrally G or C (GC→AT mutations), compared to new G or C mutations that segregate at sites that were ancestrally A or T (AT→GC), at both synonymous sites or in intron sequences[15,47]. This is inconsistent with the mutation bias hypothesis because if mutation bias is the only factor that affects the composition of a sequence, we expect the number of GC→AT mutations to be equal to the number of AT→GC mutations, irrespective of the G+C content of the sequence[15,48]. This seems counter-intuitive, but the reasoning is as follows: if the G+C content of a sequence is stationary, as the substitution data indicates it to be[15,47], the number of GC→AT substitutions is equal to the number of AT→GC substitutions, irrespective of the forces acting on the sequence. If there is no selection or BGC, the pattern of substitution is a reflection of the pattern of mutation; so, there must be equal numbers of GC→AT and AT→GC mutations entering the population, according to the mutation bias hypothesis, if the composition of the sequence is stationary. However, the excess of GC→AT mutations that segregate in populations is consistent with either selection or BGC acting to increase G+C content. If there is no mutation bias but selection or BGC has elevated the G+C content of a sequence to 80%, then 80% of the new mutations will be

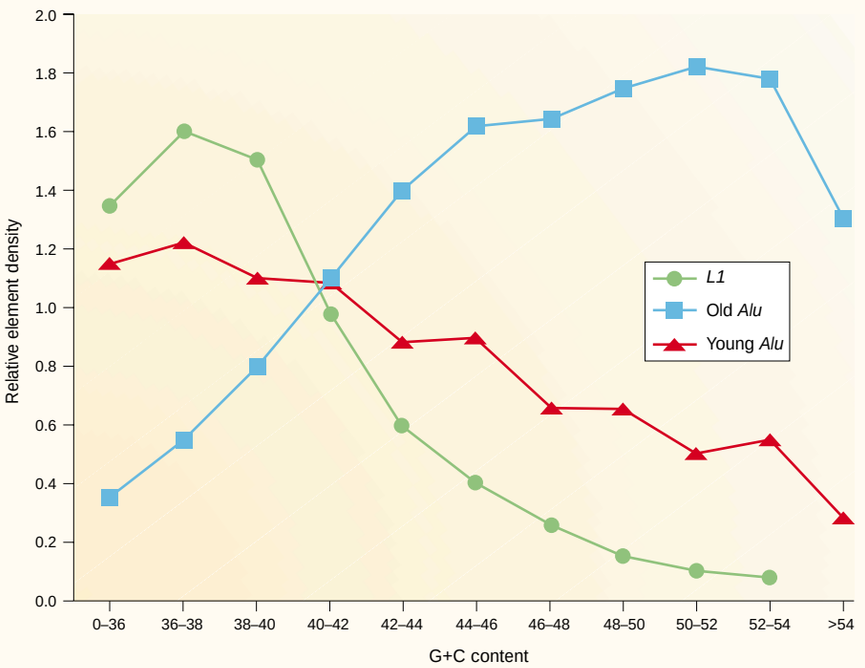

Figure 3 | **The density of *Alu* and *L1* elements in the human genome.** The density of young *Alu* (<1 million years (Myr) old), old *Alu* (60–100 Myr old) and *L1* (<65 Myr old) elements in the human genome, relative to the average density of *Alu* and *L1* elements, separated according to the G+C content of the 50 kb that surrounds each element. Data derived from REF. 1.

Table 1 | Three models for the evolution of isochores*

Model	For	Against	Overturned, previously supportive, evidence	Overturned, previously contradictory, evidence
Neutral				
Mutation bias	Known biases in DNA repair can give rise to compositional biases in bacteria	Frequency of GC→AT SNPs is not equal to frequency of AT→GC SNPs	Claim that GC→AT substitutions are as common as AT→GC substitutions proves only that sequences are at equilibrium, not that mutational bias is responsible	Claim that K_s does not covary with GC seems to be wrong (but whether the pattern is consistent with mutational bias is unclear)
	Theoretical: isochore evolution does not require selective death‡	G+C content at non-synonymous sites also correlates with GC_3 and GC_I, indicating that the force affecting GC can overcome weak purifying selection		
Biased gene conversion	G+C content and recombination rate covary	K_s positively covaries with G+C content		Claim that GC→AT substitutions are as common as AT→GC substitutions proves only that sequences are at equilibrium, not that mutational bias is responsible.
	Number of GC→AT SNPs is not equal to the number of AT→GC SNPs	There are ancient Y-linked genes that are (G+C)-rich		
	Theoretical: isochore evolution does not require selective death	Theoretical: model is highly parameter sensitive		
Selection				
Non-specific cause	Number of GC→AT SNPs is not equal to the number of AT→GC SNPs	K_s positively covaries with G+C content	GC_3>GC_I is owing partly to transposable elements in introns; it is not necessary to invoke local compensatory selection	Claim that GC→AT substitutions are as common as AT→GC substitutions proves only that sequences are at equilibrium, not that mutational bias is responsible
	Distribution of *Alu* elements different to *L1* elements	Theoretical: if selection acts on each GC→AT mutation, the high number of selective deaths could be much higher than a mammalian population could tolerate		
Thermostability hypothesis		GC_3 does not covary with optimal growth temperature in bacteria		
		Isochore evolution might precede the evolution of homeothermy		

*Evidence for and against each model is shown, as well as previous evidence that has been overturned. ‡Selective death is the failure to survive or breed owing to natural selection.
GC_I, G+C content of sequence that surrounds a gene; GC_I, G+C content of an intron; K_s, synonymous substitution rate; SNP, single nucleotide polymorphism.

GC→AT, and 20% will be AT→GC. It has so far proved impossible to differentiate between selection and BGC using SNP data, but if we could gather sufficient SNPs from a (G+C)-rich part of the Y chromosome, then we could test the BGC hypothesis.

Substitution rate and G+C content. The relationship between the SYNONYMOUS SUBSTITUTION RATE (K_s) and GC_3 is also potentially informative about isochore evolution, as whatever is driving isochore evolution would be expected to affect the rate of nucleotide substitution. The latest statistical analyses, using recent protocols that allow for CODON USAGE BIAS, as well as for TRANSITION/TRANSVERSION bias, suggest that K_s is positively correlated to GC_3 in most mammals[49–51]. The correlation is such that the least (G+C)-rich genes have a substitution rate that is less than half that of the most (G+C)-rich genes[49–51]. A similar pattern is also evident in pseudogenes[46].

The positive correlation between GC_3 and K_s seems to be inconsistent with both selection and BGC (REF. 51). This is because the available evidence indicates that if selection or BGC are acting, then they are increasing G+C content (see SNP section above) and we would expect sequences under greater selective constraint to evolve more slowly. Although selection does not always reduce the substitution rate[52], it is difficult to see how selection or BGC could explain the fairly pronounced increase of K_s with GC_3, unless there is also correlated variation in the mutation rate.

Unfortunately, the mutation bias hypothesis does not make any strong predictions about the relationship expected between the mutation rate and G+C content. Theoretical analyses of DNA-replication models, in which variation in composition is generated by variation in free nucleotide concentrations, indicate

that mutation rates are expected to reach a maximum at intermediate G+C contents when the overall concentration of free nucleotides is constant[53–55]. However, experimental studies indicate that the mutation rate depends on the overall concentration of free nucleotides[20–22] and this effect might dominate the relationship between K_s and GC_3 (REF. 53). So the DNA-replication hypothesis can, in theory, generate almost any relationship between the substitution rate and G+C content. By contrast, the DNA-repair hypothesis makes a clear prediction: increasing repair efficiency generally decreases the mutation rate[31], but we do not know whether DNA repair is likely to increase or decrease G+C content; the one repair pathway for which we have information — the repair of spontaneously arising base mismatches — is G+C biased[30], but DNA repair is complex and other pathways might have different biases.

Distribution of repetitive DNA. Although GC_3 is strongly correlated to isochore G+C, it is often greater than it (FIG. 2); much the same relationship holds for GC_3 and intron G+C content (GC_i)[9]. It has been indicated that this is inconsistent with the mutation bias hypothesis[15,16]. How can the pattern of mutation bias be different in exons and introns? It would also seem to be inconsistent with BGC for the same reason. However, it is clear[56] that we need to take into account repetitive DNA elements in any model that seeks to explain the relationship between GC_3 and isochore G+C, because at least 40% of mammalian DNA comprises repetitive DNA elements or their remains[57]. The consequences for the mutation bias model are relatively straightforward: the integration of repetitive DNA elements will tend to ameliorate the G+C content of any region with an extreme substitution bias because the G+C contents of the two main families of mammalian repetitive DNA, ALU and *L1* elements, are modest at 52 and 37%, respectively[56]. Such a model correctly predicts that short introns, which are expected to contain less transposable-element-derived sequence, should be more similar to GC_3 (REF. 56). However, whether transposable elements explain every difference between GC_3 and GC_i is unknown, not least because it is impossible to identify old transposable elements. The consequences of repetitive DNA integration for the selective and BGC hypotheses is more difficult to predict: in the case of selection, this is because selection will act on the element itself; and in the case of BGC, because we do not know whether BGC will act to increase or decrease the frequency of an element in the population.

"The question of why there is large-scale variation in base composition along mammalian and avian chromosomes is far from resolved because none of the available hypotheses adequately explains all the data."

Given that a large proportion of the human genome is repetitive DNA in various states of decay, the evolution of these sequences and their genomic localization should provide some information about the forces that give rise to isochores. The mutation bias and BGC hypotheses do not make clear predictions; however, the selective hypothesis predicts that the pattern of repetitive DNA integration and fixation should be different if selection is acting on base composition. This prediction seems to be met by the distribution of *Alu* and *L1* elements; these elements are believed to share a common integration pathway with a preference for (A+T)-rich integration sites[58–60], but the distribution of *Alu* elements is biased towards (G+C)-rich DNA compared to that of *L1* elements[1,57,61] (FIG. 3). This pattern is particularly evident for old *Alu* elements, which might have become enriched in the (G+C)-rich regions through higher rates of deletion and decay in the (G+C)-poor regions of the genome[1,57,61]. However, there is a clear difference in the distribution of even the youngest *Alu* elements compared to *L1* elements[1].

Concluding remarks

The question of why there is large-scale variation in base composition along mammalian and avian chromosomes is far from resolved because none of the available hypotheses adequately explains all the data (see TABLE 1 for a summary of this evidence). The excess of GC→AT mutations that segregate in mammalian populations seems to be evidence against all mutation bias hypotheses, whereas the high GC_3 values of some Y-linked genes and the positive correlation between K_s and GC_3 seems to be evidence against the BGC hypothesis. The BGC model is also parameter sensitive. Yet it is unclear whether selection can explain the relationship between K_s and GC_3, and it leaves some important questions to be answered. What is selection acting on? Could selection be strong enough to act in organisms, such as humans, that have very small effective population sizes? And how would mammals cope with the mutation load imposed by selection on large numbers of sites, especially when the deleterious mutation rate in their protein-coding sequences is so high[62,63]? Given these unanswered questions and the evidence against the mutation bias hypothesis from the SNP data, we believe that BGC is probably the most likely cause of isochores.

Despite the limited progress towards an understanding of isochore evolution, we now have a set of well-established observations that any hypothesis must explain: the excess of GC→AT mutations, the positive correlation between K_s and GC_3, and the distribution of repetitive DNA. Furthermore, the collection and analysis of several types of new data is likely to be very informative; for example, SNP data from the Y-chromosome will allow us to test the BGC hypothesis, whereas information about the G+C content of sequences that flank polymorphic and fixed repetitive DNA elements will allow us to test the selection hypothesis.

Adam Eyre-Walker is at the Centre for the Study of Evolution and School of Biological Sciences, University of Sussex, Brighton BN1 9QG, UK.
Laurence Hurst is at the Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK.
Correspondence to A.E.-W.
e-mail: a.c.eyre-walker@sussex.ac.uk

Links

DATABASE LINKS MHC class II | MHC class I | neurofibromatosis | *TNFα* | *SRY*
FURTHER INFORMATION Human major histocompatibility (MHC) locus | Nile crocodile | Red-eared slider turtle | Human SNP database | Codon usage database | Adam Eyre-Walker's lab | Laurence Hurst's lab

1. International Human Genome Sequencing Consortium (IHGSC). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Shabalina, S. A. & Kondrashov, A. S. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet. Res.* **74**, 23–30 (1999).
4. Filipski, J., Thiery, J. P. & Bernardi, G. An analysis of the bovine genome by Cs_2SO_4-Ag^+ density centrifugation. *J. Mol. Biol.* **80**, 177–197 (1973).
5. Bernardi, G. Isochores and the evolutionary genomics of vertebrates. *Gene* **241**, 3–17 (2000).
6. Tenzen, T. *et al.* Precise switching of DNA replication timing in the GC content transition area of the human major histocompatibility complex. *Mol. Cell. Biol.* **17**, 4043–4050 (1997).
7. Eisenbarth, I., Vogel, G., Krone, W., Vogel, W. & Assum, G. An isochore transition in the *Nf1* gene region coincides with a switch in the extent of linkage disequilibrium. *Am. J. Hum. Genet.* **67**, 873–880 (2000).
8. Bernardi, G. *et al.* The mosaic genome of warm blooded vertebrates. *Science* **228**, 953–958 (1985).
9. Clay, O., Caccio, S., Zoubak, Z., Mouchiroud, D. & Bernardi, G. Human coding and noncoding DNA: compositional correlations. *Mol. Phyl. Evol.* **5**, 2–12

(1996).
10. D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C. & Bernardi, G. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.* **32**, 504–510 (1991).
11. Filipski, J. Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Lett.* **217**, 184–186 (1987).
12. Suoeka, N. Directional mutation pressure and neutral molecular evolution. *Proc. Natl Acad. Sci. USA* **85**, 2653–2657 (1988).
13. Wolfe, K. H., Sharp, P. M. & Li, W.-H. Mutation rates differ among regions of the mammalian genome. *Nature* **337**, 283–285 (1989).
14. Bernardi, G. & Bernardi, G. Compositional constraints and genome evolution. *J. Mol. Evol.* **24**, 1–11 (1986).
15. Eyre-Walker, A. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* **152**, 675–683 (1999).
16. Hughes, A. L. & Yeager, M. Comparative evolutionary rates of introns and exons in murine rodents. *J. Mol. Evol.* **45**, 125–130 (1997).
17. Holmquist, G. P. Chromosome bands, their chromatin flavors and their functional features. *Am. J. Hum. Genet.* **51**, 17–37 (1992).
18. Eyre-Walker, A. Recombination and mammalian genome evolution. *Proc. R. Soc. Lond.* B **252**, 237–243 (1993).
19. Nagylaki, T. Evolution of a finite population under gene conversion. *Proc. Natl Acad. Sci. USA* **80**, 6278–6281 (1983).
20. Meuth, M. The molecular basis of mutations induced by deoxyribonucleoside triphosphate imbalances in human cells. *Exp. Cell Res.* **181**, 305–316 (1989).
21. Phear, G. & Meuth, M. A novel pathway for the transversion mutation induced by dCTP misincorporation in a mutator strain CHO cells. *Mol. Cell. Biol.* **9**, 1810–1812 (1989).
22. Phear, G. & Meuth, M. The genetic consequences of DNA precursor pool imbalance: sequence analysis of mutations induced by excess thymidine at the hamster aprt locus. *Mutat. Res.* **214**, 201–206 (1989).
23. Leeds, J. M., Slabaugh, M. B. & Matthews, C. K. DNA precursor pools and ribonucleotide reductase activity: distribution between the nucleus and the cytoplasm of mammalian cells. *Mol. Cell. Biol.* **5**, 3443–3450 (1985).
24. McCormick, P. J., Danhauser, L. L., Rustim, Y. M. & Bertram, J. S. Changes in ribo- and deoxyribonucleoside triphosphate pools within the cell cycle of a synchronised mouse fibroblast cell line. *Biochim. Biophys. Acta* **755**, 36–40 (1983).
25. Holmquist, G. P. Evolution of chromosome bands: molecular ecology of noncoding DNA. *J. Mol. Evol.* **28**, 469–486 (1989).
26. Federico, C., Saccone, S. & Bernardi, G. The gene-richest bands of human chromosomes replicate at the onset of the S-phase. *Cytogenet. Cell Genet.* **80**, 83–88 (1998).
27. Consortium, M. S. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401**, 921–923 (1999).
28. Eyre-Walker, A. Evidence that both G+C-rich and G+C-poor isochores replicate early and late in the cell cycle. *Nucleic Acids Res.* **20**, 1497–1501 (1992).
29. Boulikas, T. Evolutionary consequences of nonrandom damage and repair of chromatin domains. *J. Mol. Evol.* **35**, 156–180 (1992).
30. Brown, T. C. & Jiricny, J. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* **54**, 705–711 (1988).
31. Eyre-Walker, A. DNA mismatch repair and synonymous codon evolution in mammals. *Mol. Biol. Evol.* **11**, 88–98 (1994).
32. Fryxell, K. & Zuckerkandl, E. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.* **17**, 1371–1383 (2000).
33. Galtier, N. & Mouchiroud, D. Isochore evolution in mammals: a human-like ancestral structure. *Genetics* **150**, 1577–1584 (1998).
34. Macaya, G., Thiery, J. P. & Bernardi, G. An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.* **108**, 237–254 (1976).
35. Thiery, J. P., Macaya, G. & Bernardi, G. An analysis of eukaryotic genomes by density gradient centrifugation. *J. Mol. Biol.* **108**, 219–235 (1976).
36. Bernardi, G. & Bernardi, G. Compositional patterns in the nuclear genomes of cold-blooded vertebrates. *J. Mol. Evol.* **31**, 265–281 (1990).
37. Hughes, S., Zelus, D. & Mouchiroud, D. Warm-blooded isochore structure in Nile crocodile and turtle. *Mol. Biol. Evol.* **16**, 1521–1527 (1999).
38. Galtier, N. & Lobry, J. Relationships between genomic G+C content, RNA secondary structures and optimal growth temperature in prokaryotes. *J. Mol. Evol.* **44**, 632–636 (1997).
39. Hurst, L. D. & Merchant, A. R. High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proc. R. Soc. Lond.* B **268**, 493–497 (2001).
40. D'Onofrio, G., Jabbari, K., Musto, H. & Bernardi, G. The correlation of protein hydropathy with the composition of codin sequences. *Gene* **238**, 3–14 (1999).
41. Ikemura, T. & Wada, K.-N. Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers: relation between nucleotide sequence data and cytogenetic data. *Nucleic Acids Res.* **16**, 4333–4339 (1991).
42. Fullerton, S. M., Bernardo Carvalho, A. & Clark, A. G. Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* **18**, 1139–1142.
43. Mouchiroud, D., Gautier, C. & Bernardi, G. The compositional distribution of coding sequences and DNA molecules in humans and murids. *J. Mol. Evol.* **27**, 311–320 (1988).
44. Filipski, J. Chromosome localization-dependent compositional bias of point mutations in *Alu* repetitive sequences. *J. Mol. Biol.* **206**, 563–566 (1989).
45. Casane, D., Boissinot, S., Chang, B. H. J., Shimmin, L. C. & Li, W.-H. Mutation pattern variation among regions of the primate genome. *J. Mol. Evol.* **45**, 216–226 (1997).
46. Francino, P. & Ochman, H. Isochores result from mutation not selection. *Nature* **400**, 30–31 (1999).
47. Smith, N. G. C. & Eyre-Walker, A. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. *Mol. Biol. Evol.* **18**, 982–986.
48. Eyre-Walker, A. Differentiating selection and mutation bias. *Genetics* **147**, 1983–1987 (1997).
49. Smith, N. G. C. & Hurst, L. D. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153**, 1395–1402 (1999).
50. Bielawski, J. P., Dunn, K. A. & Yang, Z. Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. *Genetics* **156**, 1299–1308 (2000).
51. Hurst, L. D. & Williams, E. J. B. GC content and the silent site substitution rate do covary in rodents: implications for methodology and for the evolution of isochores. *Gene* **261**, 107–114 (2001).
52. Eyre-Walker, A. The effect of constraint on the rate of evolution in neutral models with biased mutation. *Genetics* **131**, 233–234 (1992).
53. Eyre-Walker, A. The role of DNA replication and isochores in generating mutation and silent substitution rate variance in mammals. *Genet. Res.* **60**, 61–67 (1992).
54. Gu, X. & Li, W.-H. A model for the correlation of mutation rate with GC content and the origin of GC-rich isochores. *J. Mol. Evol.* **38**, 468–475 (1994).
55. Wolfe, K. Mammalian DNA replication: mutation biases and the mutation rate. *Theor. Biol.* **149**, 441–451 (1991).
56. Duret, L. & Hurst, L. D. The elevated G and C content at exonic third sites is not evidence against neutralist models of isochore evolution. *Mol. Biol. Evol.* **18**, 757–762.
57. Smit, A. Interspersed repeats and the other mementos of transposable elements in the mammalian genomes. *Curr. Opin. Genet. Dev.* **9**, 657–663 (1999).
58. Feng, Q., Moran, J. V., Kazazian, H. H. & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905–916 (1996).
59. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl Acad. Sci. USA* **94**, 1872–1877 (1997).
60. Toda, Y., Saito, R. & Tomita, M. Characteristic sequence pattern in the 5- to 20-bp upstream region of primate *Alu* elements. *J. Mol. Evol.* **50**, 232–237 (2000).
61. Gu, Z., Wang, H., Nekrutenko, A. & Li, W.-H. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene* **259**, 81–88 (2000).
62. Eyre-Walker, A. & Keightley, P. D. High genomic deleterious mutation rates in hominids. *Nature* **397**, 344–347 (1999).
63. Keightley, P. D. & Eyre–Walker, A. Deleterious mutations and the evolution of sex. *Science* **290**, 331–333 (2000).

Acknowledgements
Many thanks to M. Lercher, N. Smith, E. and A. Urrutia. Both A.E.-W. and L.D.H. are supported by the Royal Society, to whom they are grateful.

SCIENCE AND SOCIETY

Molecular metaphors: the gene in popular discourse

Dorothy Nelkin

Geneticists deploy a striking range of metaphors to communicate their science, to promote its value and to suggest its social meaning to the public. So too, critics of science and special interest groups use metaphorical constructs to express their concerns about the implications of the 'genetic revolution'. Through metaphors, genetics can seem a source of salvation or a means of exploitation, a boon to health or a source of risk. This paper is a critical review of the metaphors used to communicate genetic information to the public.

"There's a metaphor contest going on"[1]
Harold Varmus, former Director of the National Institutes of Health

The human genome is "like the torn pages of a giant novel, written in an unknown language, blowing about helter skelter in an air-conditioned, enclosed space such as Houston's Astrodome"[2]. The scientists involved in mapping and sequencing the genome, so this extended metaphor implies, will capture all these pages, put them in proper order and analyse the meaning of the resulting text.