Isochores Exhibit Evidence of Genes Interacting with the Large-scale Genomic Environment

William H. Press*,[1] and Harlan Robins†

*Los Alamos National Laboratory, Los Alamos, NM 87545
†Institute for Advanced Study, Princeton, NJ 08540

[1] *Corresponding author:* D-1 Group, MS F-600, Los Alamos National Laboratory, Los Alamos, NM 87545. E-mail: wpress@lanl.gov

Manuscript received December 8, 2005.

ABSTRACT

The genomes of mammals and birds can be partitioned into megabase-long regions, termed isochores, with consistently high, or low, average $C+G$ content. Isochores with high CG contain a mixture of CG-rich and AT-rich genes, while high AT isochores contain predominantly AT-rich genes. The two gene populations in the high CG isochores are functionally distinguishable by statistical analysis of their Gene Ontology categories. However, the aggregate of the two populations in CG isochores is not statistically distinct from AT-rich genes in AT isochores. Genes tend to be located at local extrema of composition within the isochores, indicating that the $C+G$ enriching mechanism acted differently when near to genes. On the other hand, maximum likelihood reconstruction of molecular phylogenetic trees shows that branch lengths (evolutionary distances) for CG-rich gene third codon positions are not substantially larger than those for AT-rich genes. In the context of neutral mutation theory this argues against any strong positive selection. Disparate features of isochores might be explained by a model in which about half of all genes functionally require AT-richness, while, in warm-blooded organisms, about half the genome (in large coherent blocks) acquired a strong bias for mutations to CG. Using mutations in CG-rich genes as convenient indicators, we show that $\approx 20\%$ of amino acids in proteins are broadly substitutable, without regard to chemical similarity.

Isochores, so named by Bernardi (Bernardi *et al.* 1985; Bernardi 2000), are large regions in the human genome, as long as tens of megabases, that are anomalously rich in C and G nucleotides. Isochores analogous to human are found in the genomes of all mammals and birds (Bernardi 2000), plus a small number of additional reptiles such as the Nile crocodile (Hughes *et al.* 1999). Invertebrates, and almost all cold-blooded vertebrates, do not manifest isochore structure in their genomes. The putative common ancestor in which isochores originated is thus an amniote in the Carboniferous period ($\approx$ 300 Ma b.p.), although it was not until after the Permian-Triassic extinction ($\approx$ 250 Ma b.p.) that the carriers of isochores, namely archosaurs, birds, and mammals, proliferated.

Isochores are by no means subtle features in the genome (IHGSC 2001). By way of example, Figure 1 shows the $A+T$ (complement of $C+G$) content of three human, and three zebrafish, chromosomes, plotted on a common scale. The nucleotide counts are shown as bars in 300 kb bins, with the base of the bars at $A+T=0.58$, an arbitrary value that approximately divides CG isochores from AT isochores (as we will refer to regions that are not CG-rich).

[Figure 1 about here.]

It is not a settled issue whether isochore formation continues today, that is, whether CG isochores are continuing to form from AT isochores. However, a body of recent evidence suggests that, on the contrary, isochores are gradually disappearing from mammalian genomes (Duret *et al.* 2002; Belle *et al.* 2004). If so, then we may view isochores as fossils of a unique period in our past during which a strong mutational pressure first appeared and then disappeared. Apart from the obvious question as to what caused this to happen, we may also hope to learn from the isochore-forming event something about the interaction of genes, as primary carriers of functional information, with the much larger genome that they inhabit.

It has proved surprisingly difficult to find functional relationships between isochores and the genes inside them (Vinogradov 2003; IHGSC 2001). By default, the more conservative view has been that isochores are predominantly the result of the accumulation of selection-neutral changes caused by (evidently spatially nonuniform) mutation or repair biases. One currently favored model is biased gene conversion (BGC) during homologous recombination (Eyre-Walker and Hurst 2001). If isochore evolution is selection-neutral, then genes are passive riders on the isochore background. That is, their noncritical elements, such as synonymous bases in 3rd codon positions and nonfunctional bases in their $3'$ and $5'$ untranslated regions (UTRs), should evolve towards CG richness along with the rest of an isochore. Indeed, it is well established (Bernardi *et al.* 1985; Clay *et al.* 1996; Hamada *et al.* 2003), and easy to show, that the CG content of 3rd codon positions and $3'$ and $5'$ UTRs are all strongly correlated with the CG content of the flanking genomic region.

Less conservative, but also longstanding, is the hypothesis that the evolution of isochores was favored by positive natural selection, for example selection in warm-blooded vertebrates for DNA that is stable at higher temperature (Bernardi 2000; Smith and Eyre-Walker 2001). However, several such hypotheses notwithstanding, the nature of the selection pressure remains obscure (Belle *et al.* 2002; Eyre-Walker and Hurst 2001).

If isochore formation was predominantly selection-neutral, then there should not be statistically significant functional differences between genes in an AT versus CG isochore, since during isochore formation the (pre-existing) population of genes are simply hitchhikers. However, without reference to isochores, we have previously shown (Robins and Press 2005) that AT-rich and CG-rich genes are readily distinguishable, statistically, by gene functionality. In particular, AT-rich genes are preferentially associated with one set of biological processes, centered on transcription and mRNA processing, while CG-rich genes are associated with another set, centered on signal transduction, receptors, and signaling cascades. Can this finding be reconciled with a selection-neutral model for isochores?

We will see below that the answer is yes, but with an important caveat. As one would expect, AT- and CG-rich genes are associated with the corresponding AT and CG isochores. But the association is not one-to-one: Genes in AT isochores are predominantly AT-rich, while genes in CG isochores can be either AT-rich or CG-rich, resulting in a complex landscape of AT-rich intrusions into what are otherwise CG isochores. The two groups of genes in CG isochores, AT-rich and CG-rich, are in fact statistically distinguishable by function. However, when one aggregates both groups of genes in CG isochores, one obtains a mixture that is not functionally distinguishable from the genes in AT isochores, consistent with the neutral model.

Thus, while the correlation of function with CG richness in CG isochores is clear evidence of selection, it may indicate only negative selection. That is, the evolutionary pressure towards CG richness could be entirely the result of neutral mutations; but in that case some genes, correlated by function, felt negative fitness pressure to resist the neutral mutations and remained AT-rich genes within a CG isochore.

On the other hand, and arguing for isochore formation by positive selection, we find evidence that genes that became CG-rich (in CG isochores) are far from passive passengers: They are *more* CG-rich than their surroundings. That genes are at special locations of composition is already suggested visually, at least for CG-rich genes, if one simply looks at the position and composition of genes relative to window counts (Figure 2), where an unexpected number of CG-rich genes seem to occur in bins that are extrema. We give a more quantitative test below.

[Figure 2 about here.]

METHODS

Defining Gene Populations and Large-Scale Isochores: We use $A + T$ and $C - G$ counts in a gene's $3'$ UTR to determine whether it is an AT-rich or CG-rich gene, applying the algorithm given in Robins and Press (2005), equation [1], to get a probability. This method was shown to yield the cleanest separation of the two gene populations.

Since isochores are not homogeneous (IHGSC 2001, and cf. Figure 1), a precise definition is perforce somewhat arbitrary. However, if one plots the above AT- versus CG-rich probability for each gene along the genome, as in Figure 3, a clear pattern emerges: Some regions extending over many megabases contain predominantly AT-rich genes, while other regions contain a more equal mixture of AT- and CG-rich genes. There are few, if any, large regions containing predominantly CG-rich genes, which is consistent with previous evidence (Pavlicek *et al.* 2002) that CG isochores have larger compositional variances than do AT isochores.

[Figure 3 about here.]

We can therefore define isochore boundaries by a Markov model that alternates between two states, AT-dominant and mixed. In the AT-dominant state, the respective probabilities of an AT-rich and CG-rich gene are taken as $(0.9, 0.1)$, while in the mixed state they are taken as $(0.5, 0.5)$. The state transition probability between any two consecutive genes is taken as 0.001 (that is, 0.999 chance of remaining in the same isochore state). We then use the standard forward-backward method to find the probability, at each gene, of its being in the AT-dominant state (which we now term an AT isochore) or the mixed state (which we call a CG isochore). We find that this classification is quite insensitive to varying all of the parameters above. In particular the transition probability can be varied over orders of magnitude, because the multiplicative probabilities of a relatively small number of genes can easily force a state transition, even if its *a priori* probability is unrealistically small. Results are shown in Figure 4.

[Figure 4 about here.]

Comparing Figures 1 and 4, one sees that the above Markov model largely captures one's visual impressions of large scale structure, but now objectively (at least up to choice of model parameters). We can also validate the gene-based model by comparing it to a similar Markov model that uses raw 300 kb window counts instead of genes, shown as the red line in Figure 4. In this model, we assign a 300 kb window to the high state if its count of A+T exceeds 0.565, a not untypical value in the isochore literature (Pavlicek *et al.* 2002). An AT isochore is taken to have high or low windows with respective probabilities $(0.75, 0.25)$. A CG isochore has $(0.5, 0.5)$, again reflecting its relatively larger variances. The transition probability is 0.001, as before. The results of this model are shown as the red line in Figure 4, and are insensitive to the adopted parameters. Our gene-based and window-based models for isochore identification agree in 93% of all locations in the human genome.

In characterizing variations on large, megabase scales, we necessarily miss smaller scale features, predominantly AT-rich intrusions into CG isochores. These show up as an increase in the observed variance. It is a matter of semantics whether or not to to regard these features as small isochores (Cohen *et al.* 2005).

Assessing Two Gene Populations by GO Score: In previous work (Robins and Press 2005) using Gene Ontology (GO) keyword counts, we characterized results by their statistical significance (t- and p- values). Here, we will want something more like a linear scale, so that a mixture of two populations will have a score that lies proportionally between the scores for the populations individually.

Using results from Robins and Press (2005), we define a set of "Population N" (for "nuclear") indicator words as the following: nucleic-acid, nucleus, transition-metal, zinc, bound, ZNF*, RNA, mRNA, DNA, nucleobase, nucleoside, translation. We define a set of "Population S" (for "signaling") indicator words as: signal-transduction, signaling cascade, receptor, transducer, communication, signal, transmembrane, channel, immune, pore. It is an important point that we did not choose these populations or words arbitrarily; rather, they emerged uniquely from the data as the word sets that most statistically significantly distinguish AT-rich and CG-rich genes (without regard to their locations in isochores).

Let N_N be the total number of occurrences of Population N words across the genome (e.g., in RefSeq genes), and N_S be the corresponding number for Population S words. Define $r_{NS} \equiv N_N/N_S$. (For the RefSeq genes we have $N_N = 31406$, $N_S = 16585$, and $r_{NS} = 1.89$.)

Now suppose that we have a large, probabilistically known, set of genes, meaning that we can assign a probability p_i of gene i's being in the set, and $\Sigma_i p_i \gg 10^2$ (say). Then we define that set's "Signaling Minus Nuclear Score"

(SMNS) by

$$\text{SMNS} \equiv \frac{r_{NS} \sum_i \sum_{j \in S} p_i \delta_{ij} - \sum_i \sum_{j \in N} p_i \delta_{ij}}{r_{NS} \sum_i \sum_{j \in S} p_i \delta_{ij} + \sum_i \sum_{j \in N} p_i \delta_{ij}} \tag{1}$$

Here S is the set of Population S words, N is the set of Population N words, and δij is 1 if word j occurs for gene i, zero otherwise. By construction, SMNS of the whole genome is zero. It is 1 for a set of genes that have no Population N words, and -1 for a set of genes that have no Population S words.

Usefully, we can also estimate the error for the SMNS:

$$\sigma(\text{SMNS}) \approx \frac{\sqrt{r_{SN}^2 \sum_i \sum_{j \in S} p_i^2 \delta_{ij} + \sum_i \sum_{j \in N} p_i^2 \delta_{ij}}}{r_{SN} \sum_i \sum_{j \in S} p_i \delta_{ij} + \sum_i \sum_{j \in N} p_i \delta_{ij}} \tag{2}$$

The approximation made is to ignore the error in the denominator of equation (1) as compared to that of the numerator. This is because (with foresight) it will turn out that the SMNS score is never larger than a few tenths.

Equation (2) allows us to compare different sets of genes for statistically significantly different SMNS's.

Determining Whether Genes Are More Or Less Compositionally Extreme: As discussed above, it is important to have an objective measure of whether genes are more or less extreme in $C+G$ or $A+T$ than their immediate surroundings. One measure of this tendency is to compare $A+T$ at a gene's location with $A+T$ at the midpoint of the intergenic region between the gene and its next neighbor. Referring to Figure 5, if genes are more compositionally extreme (as shown in panel A) we should get a different correlation between gene and intergene than if genes are less compositionally extreme (as shown in panel B). Panel D shows the two cases schematically.

A difference between the variance of genes and that of intergenes due to any other effect can confound the proposed measurement. For example, if genes had a smaller variance in their $A+T$ composition for functional reasons, this would bias the measurement toward panel B. Or, if the measurement accuracy of $A+T$ were poorer for genes (due to a smaller counting length) than for intergenes, then panel A would be erroneously favored. To mitigate these kinds of systematic errors, we adopt the strategy shown in Figure 5, Panel C: We characterize a gene's $A+T$ exclusively by its introns, which should have the least functional constraints; and we make intergenic counts with exactly the same window pattern as the gene to which they are being compared. If there are residual systematic biases in the introns (which do contain some functionality), we expect them to show up as a systematic shift in $A+T$, not a change in the variance. (In fact, below, we will see such small shifts.) The signature of genes that are compositionally more extreme than their surroundings is a positive correlation between gene and gene-minus-intergene. The signature of genes that are less compositionally extreme is a negative correlation.

[Figure 5 about here.]

Maximum Likelihood Phylogeny and Branch Lengths: Below, we will construct phylogenetic trees by aligning orthologous genes in human, chicken (G. gallus), and frog (X. tropicalis), with fish (D. rerio) as an outgroup. We use only genes with orthologs in all four organisms, as reported by the Ensembl data base (Birney *et al.* 2006). (We have also checked that similar results are obtained if this constraint is relaxed.) We can construct independent trees for any particular population of genes, e.g., AT-rich, or CG-rich in CG isochores. In most cases (identified below) we use only 4-fold degenerate third-codon positions, though, as we will see, interesting results are also obtained for *non*-synonymous first and second codon positions.

The reconstruction method is the standard maximum likelihood (ML) method (Felsenstein 1981; Felsenstein 2004), based on a Markov evolutionary model along each branch. We assume the established tree topology among the four species. We allow completely general transition matrices (e.g., not necessarily having the time reversible GTR form), and solve for a different transition matrix along each branch. This generality is possible because of the large amount of data available, yielding negligible statistical errors in the reconstruction. Errors are thus dominated by modeling errors, for example violation of the Markov model assumption or non-i.i.d. of individual base positions; these modeling errors are, of course, difficult to assess quantitatively. The maximum likelihood reconstruction is found iteratively by the EM method (Dempster *et al.* 1977), alternating between the calculation of node probabilities separately for each base position and the reestimation of the common (across base positions) set of transition matrices.

Because we do not assume time-reversibility, the ML method is in principle capable of producing a rooted tree, that is, the "pulley principle" (Felsenstein 1981) does not strictly apply. We find, however, that the location of the four-species common ancestor root is rather poorly determined by the data, indicating that deviations from time reversibility are small, at least along the path between the quadruped common ancestor and fish. We therefore use fish only as an outgroup and show, below, only the quadruped ancestor tree, which is accurately rooted (at least statistically).

Having obtained the transition (that is, base substitution) matrix $\mathbf{A}$ for an edge, we resolve it into an infinitesimal generator matrix $\mathbf{G}$ and a branch length μ, such that

$$\mathbf{A} = \exp(\mu \mathbf{G}) \tag{3}$$

with **G** having zero row sums, zero or negative diagonal, and zero or positive off-diagonal elements. Since **G** can absorb any constant factor from μ, it needs a normalization convention. A convenient one is

$$\mathrm{tr}(\mathbf{G}) = -4 \tag{4}$$

Then μ is the evolutionary distance measured in mean changes per base for a (standardized) uniform nucleotide distribution, essentially equivalent to the standard log-det distance (STEEL 1994; LOCKHART *et al.* 1994) and closely related to the paralinear distance (LAKE 1994). (See GU and LI (1996) for a comparison of these distance measures.)

The generator matrix **G** usefully encodes the mutational biases of individual mutation events. In the context of this paper's interest in mutations from AT to CG (or vice versa), two useful summary values are the sums of all off-diagonal elements corresponding to transitions in one direction (AT→CG) or the other (CG→AT). Below, we will refer to these values as "propensities" for each direction.

As a check on the ML reconstruction, we used all sets of pairwise (only) alignments among the four species. It is well known (CHANG 1996; LAKE 1997; BAAKE 1998) that the full transition matrices cannot be obtained from pairwise data alone. However, it is easy to get branch lengths from pairwise data. All the pairwise paralinear distances (which are additive both up and down the tree) give an overdetermined set of linear relations among the individual branch lengths. We solve for the best solution in the least squares sense. Reassuringly, the lengths obtained by this method are almost identical to those obtained by ML reconstruction.

RESULTS

Characterizing the Three Gene Groups:

[Table 1 about here.]

With the above methods, we can assign to each gene a probability of being in the AT-rich (versus CG-rich) population, and, separately, a probability of being in an AT (versus CG) isochore. The results are shown in Table 1. We adopt the notation iAT and iCG as denoting isochores, AT and CG as denoting genes, so that the three principal populations are iAT/AT, iCG/CG, and iCG/AT. Although there are undoubtedly some genuine iAT/CG genes, many or most genes that we classify as iAT/CG are probably the result of misidentified isochore boundaries. Therefore we will often restrict our attention to the three principal groups mentioned above. It is previously known that CG-rich regions have higher gene density and smaller gene lengths (IHGSC 2001).

The $A+T$ fraction of genes classed as iAT/AT is significantly higher than those classed as iCG/AT, 53.6% versus 46.0% (3rd codon position counts). Part of this difference is likely due to false positives from the larger number of iCG/CG genes, since the AT-rich and CG-rich gene populations are overlapping distributions. For iCG/CG genes, the $A+T$ fraction is 30.1%.

GO Signature Is Strong in CG Isochores, Weak or Absent in AT Isochores: The SMNS score was defined above to be zero over the average gene population, positive for groups of genes with Population S GO keywords (like "signal transduction") and negative for groups of genes with Population N GO keywords (like "nucleic acid"). Scores, and uncertainties, for the four gene groups are as follows: $0.102 \pm .006$ for iCG/CG; $-0.239 \pm .009$ for iCG/AT; $-0.010 \pm .009$ for iAT/AT; and $0.019 \pm .018$ for iAT/CG (the larger uncertainty from the smaller population size).

What is remarkable is that the largest positive and negative scores, by far, are for genes in CG isochores, while genes in AT isochores have SMNS scores consistent with zero. In other words, AT-rich genes in CG isochores are functionally more extreme (Population N) than AT-rich genes in AT isochores, even as their nearby neighbors on the genome, the iCG/CG genes, tend strongly to Population S functionality. This effect is not a correlation with AT richness – indeed, it has the opposite sign – since iCG/AT genes are markedly less AT-rich than iAT/AT genes. The observed effect is likewise opposite to what would be expected from any misclassifying iCG/CG genes as iCG/AT.

The average SMNS scores for genes in CG and AT isochores are, respectively, $0.003 \pm .006$ and $-0.006 \pm .008$, that is, statistically zero. It is striking that the CG isochores are so accurately zero, since that value is obtained only by averaging a large positive (iCG/CG) and even larger negative (iCG/AT) value in just the right proportions.

These data suggest that AT and CG isochores in fact contain the same mixtures of functionality (average SMNS zero), but that only in CG isochores have these differences been made visible as differences in gene AT richness. This is evidence that whatever "marked" large contiguous regions of the genome as incipient CG isochores did so without reference to the gene content within those regions. It is consistent with a scenario in which genes in AT isochores never experienced the pressure that created the isochores (whether neutral or selection), while genes in CG isochores were thus challenged, but with dramatically different (and functionally correlated) responses, varying from gene to gene.

Human Genes Are More Compositionally Extreme Than Their Surroundings: Figure 6 shows the result of applying the methodology described above (and in Figure 5) to the human genome. A significant positive correlation between gene and gene-minus-intergene counts is seen for all three gene populations, most strongly for iCG/CG genes. This indicates that all genes have some tendency to be more extreme than their flanking sequence with respect to (depending on the gene) CG- or AT-richness. The tendency is by far strongest for CG-rich genes. As plotted, Figure 6 does not exclude repeating elements, but the results are not significantly different if we exclude either (i) all elements identified by RepeatMasker, or (ii) only the most common LINE and SINE elements.

[Figure 6 about here.]

Phylogenetic Reconstruction Shows Mutation Bias But Not Strong Positive Selection:

[Figure 7 about here.]

Figure 7 shows the result of constructing ML molecular phylogenies, separately for genes that are iAT/AT and iCG/CG in human. The phylogenies are based on the alignment of orthologous genes, as described in Methods, above. Figure 7 shows results for 4-fold degenerate 3rd codon positions. The red and blue arrows have areas proportional to the propensities for transitions in the direction AT→CG (blue) or CG→AT (red), as determined from the generator matrices $\mathbf{G}$ on each branch (see Methods). The branch lengths are determined with a statistical accuracy of about $\Delta\mu \approx 0.008$ (1-σ), as determined by resampling, so all the differences shown in the figure are highly statistically significant. (High statistical accuracy is obtainable because the amount of data is huge.)

In the human iAT/AT genes, one sees a high degree of consistency on all branches. The greater propensity towards AT results in an AT-rich equilibrium for the genes in all three species.

In the human iCG/CG genes, one sees for the branch between the common ancestor and frog about this same balance of propensities. For human and (to a lesser extent) chicken, however, one sees a strong mutational bias towards CG. Such a bias is not unexpected since we have, of course, selected this sample for CG richness – and it had to come from somewhere.

What is most interesting in Figure 7 is what is *not* seen, namely any large disproportionate elongation for iCG/CG of the human and chicken branches relative to frog. Third codon positions are generally accepted as being governed dominantly by the neutral theory of molecular evolution (Kimura 1983), albeit balanced by a moderate positive selection favoring the “major codons” for each amino acid (Akashi 1994; Akashi 1996). Under the neutral theory, any branches on which positive isochore selection operates should be lengthened by a factor $\max(N_e s, 1)$, where N_e is the effective population size and s is the positive selective advantage of a mutation per generation. On the other hand, there is no particular reason to think that $N_e s$ due to codon usage effects, and their effects on translation rates (Levy *et al.* 1996; Zolotukhin *et al.* 1996; Wells *et al.* 1999), should be very different on different branches of the quadruped ancestor tree.

While a small elongation of the human and chicken branches, on the order of ≈ 1.3, may be present in the data, it would require a remarkable numerical coincidence among unrelated quantities, namely $s \approx 1/N_e$, to interpret this as positive selection. Rather, barring subtle competing effects, one might reasonably have expected positive selection to manifest itself as a lengthening of, say, one or more powers of ten.

[Figure 8 about here.]

We can do the same molecular phylogenetic reconstruction on first and second codon positions, where mutations will (in general) result in protein amino acid changes, and which should therefore be functionally conserved. Results are shown in Figure 8. For both iCG/CG and iAT/AT genes, the trees for first and second codon positions are nearly identical to the trees for third codon positions, but scaled by a factor $f \approx 1/6$, which (in Kimura's language) we can identify as the functional constraint, that is, the fraction of mutations that are approximately neutral. In Figure 8, the branch lengths are determined with a statistical accuracy of about $\Delta\mu \approx 0.0003$ (1-σ), so all the differences shown in the figure are again highly statistically significant.

Since functional selection on first and second codon positions occurs at the protein level, quite different from functional selection on third codon positions due to codon usage bias, the consistency, up to a factor f, between Figures 7 and 8 is reassuring. It argues that the signature of positive selection for mutations to CG in iCG/CG genes is not being confounded by other functional effects. Such a signature, at least of any significant magnitude $N_e s \gg 1$, is simply not there.

Many Protein Amino Acid Changes Are Remarkably Neutral: If isochore formation is indeed dominated by neutral mutation, as Figures 7 and 8 suggest, then isochores, and iCG/CG genes in particular, provide an interesting window into the question of the neutrality (or lack thereof) of amino acid-changing mutations. The leaf-to-leaf phylogenetic distances shown in Figure 8 for orthologous genes, functional in all the species compared, immediately show that $\gtrsim 20\%$ of all amino acids can be mutated. However, by itself, this does not exclude such possibilities as, (i) the mutations are under positive selection and reflect divergences in gene function, or (ii) the mutations, while neutral, are only allowed between chemically similar amino acids.

The formation of iCG isochores in effect “labels” a set of mutations, identifiable at least statistically in iCG/CG genes, whose origin is unrelated to the function of any particular gene. We can then look at patterns of amino acid substitution across the reconstructed molecular phylogeny. Particularly interesting are substitutions that correspond to *net* changes in amino acid usage, because these indicate broad trends, not gene-specific optimizations. We have done this comparison between human and fish. D. rerio was analyzed simultaneously with the species shown in Figures 7 and 8. Although, as an outgroup, it cannot be rooted, it is available for pairwise comparison between leaf (extant) taxa. There is of course no imputed direction of time in this pairwise comparison.

We have examined the aligned sequences of all human iCG/CG genes and their known zebrafish orthologs, and counted the frequency with which amino acids are substituted. The resulting 20 × 20 table of counts may be looked at from two viewpoints: From a biochemical perspective, we may ask whether the substitution patterns “make sense” in favoring substitutions that are close in chemical property. Or, from a genomic perspective, we may ask whether the substitutions seem driven by an exogenous pressure to increase $C + G$.

A first observation is that amino acid usage differs very significantly between human and fish coding regions. The difference is greatest between iCG/CG human genes and their fish orthologs, and less for iAT/AT and iCG/AT. For example (Table 2), for iCG/CG genes, proline, alanine, and glycine usage is respectively 20%, 19%, and 13% higher in human than in fish, while asparagine, isoleucine, and methionine usage is respectively 21%, 18%, and 17% lower. One notices immediately that the former have exclusively C or G in the first and second codon positions, while the latter have A and T.

[Figure 9 about here.]

[Table 2 about here.]

Figure 9 shows a biochemical perspective. For each of the eight amino acids with the greatest positive or negative changes in usage between fish and human, arrows are shown indicating the four most frequent substitutions. The underlying diagram, after Betts and Russell (2003), puts closely substitutable amino acids close to each other. One sees that only a few of the most frequent observed substitutions make biochemical sense, e.g., Lys (K) to Arg (R), or Ile (I) to Leu (L), while many others are, figuratively and literally, a stretch: Lys (K) to Ser (S), Leu (L) to Arg (R), Leu (L) to Pro (P), Thr

(T) to Pro (P), etc. It is hard to imagine that there would not be significant functional consequences in making these kinds of substitutions in $\approx 20\%$ of particular amino acids, unless something like this fraction of amino acid positions in proteins are close to universally substitutable.

If the data in Figure 9 do not make sense biochemically, they do make sense when mapped into the genetic code. As shown in Table 2, of the 32 largest substitutions, 27 can be explained as single mutations in the first or second codon that change an A or T into a C or G. The remaining five are all CG-neutral. None of 32 are codon changes favoring A or T.

DISCUSSION

To be viable as an explanation for isochores, a theory must be consistent with all of the following observations, from this work and the previous literature:

(i) Role of genes. Gene locations are special in isochores. Genes, both AT-rich and CG-rich, have more extreme compositions than their immediate intergenic flanking regions. A theory must explain Figure 6.

(ii) Composition asymmetry. CG isochores contain many AT-rich genes, while AT isochores contain few CG-rich genes. Not unrelated, CG isochores have a larger compositional variance on all scales than do AT isochores.

(iii) Gene functional correlations. Genes in CG isochores show a correlation between AT-richness and GO function. Genes in AT isochores don't show such a correlation. On average, however, genes in the two isochores appear to have the same mixture of GO functions.

(iv) Spatial broken symmetry. How was any specific large region selected to become a CG isochore, or not so selected?

(v) Evolutionary distance. There was not much more molecular evolution in the phylogenetic tree of iCG/CG genes than there was in iAT/AT genes. Rather, the branch leading to iCG/CG genes shows a strong mutational bias, not seen for iAT/AT genes. In the context of Kimura's theory of neutral evolution, this argues against positive selection pressure.

Properties (i) and (v), both seemingly strongly supported by the data, seem contradictory. If the evolution of isochores is entirely neutral, then why are the genes in special locations? This question would be answered by a hypothetical (germ line) mutational mechanism or repair process that acts differently in the vicinity of a gene than it does in a typical intergenic region: Either mutation rates near genes are higher, or else they are (in CG isochores) more biased towards C and G.

Properties (ii) and (iii), on the other hand, seem quite explainable. We start with a genome in a natural state of relative AT richness, as is seen in almost all animals except warm-blooded vertebrates. We suppose that a population of genes, about half, depend critically on regulatory or other functional mechanisms that depend on this "universal" AT richness. An example of such a mechanism may be regulation by microRNAs (Robins and Press 2005). Genes that do, or don't, require AT richness are randomly distributed on the genome. Something now happens, as posited by property (iv): A dramatic mutational bias towards CG occurs in large genomic regions. Genes that are not dependent on AT-rich machinery become CG-rich, that is, become iCG/CG genes. Those that are dependent experience purifying (negative) selection, and remain AT-rich, that is, become iCG/AT genes.

Property (iv), requiring an explanation of what originally "painted" the propensity towards CG mutations onto large, coherent parts of the ancestral genome, is thus seen to be fulcrum on which any explanation of isochores may tip.

Sequence data used in this paper, including alignments, is available at http://www.nr.com/bio/IsochoreSuppMat.html.

Acknowledgments: The authors thank Arnold Levine and Alistair McGregor for useful discussions. David Rand, as editor, provided substantive and important input. This work was supported in part by the James Simons Foundation, the Ambrose Monell Foundation, and the Shelby White and Leon Levy Initiatives Fund.

References

Akashi, H., 1994 Synonymous Codon Usage in Drosophila melanogaster: Natural Selection and Translational Accuracy. Genetics **136**: 927–935.

Akashi, H., 1996 Molecular Evolution Between Drosophila melanogaster and D. simulans: Reduced Codon Bias, Faster Rates of Amino Acid Substitution, and Larger Proteins in D. melanogaster. Genetics **144**: 1297–1307.

Baake, 1998 What can and what cannot be inferred from pairwise sequence comparisons? Mathematical Biosciences **154**: 1–21.

Belle, E., L. Duret, N. Galtier and A. Eyre-Walker, 2004 The Decline of Isochores in Mammals: An Assessment of the GC Content Variation Along the Mammalian Phylogeny. J Mol Evol **58**: 653–660.

Belle, E., N. Smith and A. Eyre-Walker, 2002 Analysis of the Phylogenetic Distribution of Isochores in Vertebrates and a Test of the Thermal Stability Hypothesis. J Mol Evol **55**: 356–363.

Bernardi, G., 2000 Isochores and the evolutionary genomics of vertebrates. Gene **241**: 3–17.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas *et al.*, 1985 The mosaic genome of warm-blooded vertebrates. Science **228**: 953–958.

Betts, M. J. and R. B. Russell, 2003 Amino Acid Properties and Consequences of Substitution. In M. R. Barnes and I. C. Gray, editors, *Bioinformatics for Geneticists*, Wiley, New York, 289–316.

Birney, E., D. Andrews, M. Caccamo, Y. Chen, L. Clarke *et al.*, 2006 Ensembl 2006. Nucleic Acids Res. **34**: D556–D561.

Chang, J., 1996 Full Reconstruction of Markov Models on Evolutionary Trees: Identifiability and Consistency. Mathematical Biosciences **137**: 51–73.

Clay, O., S. Caccio, Z. Zoubak, D. Mouchiroud and G. Bernardi, 1996 Human coding and noncoding DNA: compositional correlations. Mol Phyl Evol **5**: 2–12.

Cohen, N., T. Dagan, L. Stone and D. Graur, 2005 GC composition of the human genome: in search of isochores. Mol Biol Evol **22**: 1260–1272.

Dempster, A., N. Laird and D. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc., Series B **39**: 1–38.

Duret, L., M. Semon, G. Piganeau and D. Mouchiroud, 2002 Vanishing GC-rich isochores in mammalian genomes. Genetics **162**: 1837–1847.

Eyre-Walker, A. and L. D. Hurst, 2001 The evolution of isochores. Nat Rev Gen **2**: 549–555.

Felsenstein, J., 1981 Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. J. Mol. Evol. **17**: 368–376.

Felsenstein, J., 2004 *Inferring Phylogenies*. Sinauer Associates.

Gu, X. and W. Li, 1996 Bias-Corrected Paralinear and LogDet Distances and Tests of Molecular Clocks and Phylogenies Under Nonstationary Nucleotide Frequencies. Mol. Biol. Evol. **13**: 1375–1383.

Hamada, K., H. Tokumasa, H. Ota, K. Mizuno and T. Shinozawa, 2003 Presence of isochore structure in reptile genomes suggested by the relationship between GC contents of intron regions and those of coding regions. Genes Genet. Syst. **78**: 195–198.

Hughes, S., D. Zelus and D. Mouchiroud, 1999 Warm-blooded isochore structure in Nile crocodile and turtle. Mol Biol Evol **16**: 1521–1527.

IHGSC, 2001 International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. Nature **409**: 860–921.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press.

Lake, J., 1994 Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. PNAS **91**: 1455–1459.

Lake, J., 1997 Phylogenetic Inference: How Much Evolutionary History is Knowable? Mol. Biol. Evol. **14**: 213–219.

Levy, J., R. R. M. S. Zolotukhin and C. J. Link, 1996 Retroviral transfer and expression of a humanized, red-shifted green fluorescent protein gene into human tumor cells. Nat. Biotechnol. **14**: 610–614.

Lockhart, P., M. Steel, M. Hendy and D. Penny, 1994 Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**: 605–612.

Pavlicek, A., J. Paces, O. Clay and G. Bernardi, 2002 A compact view of isochores in the draft human genome sequence. FEBS Letters **511**: 165–169.

Robins, H. and W. H. Press, 2005 Human microRNAs target a functionally distinct population of genes with AT-rich $3'$ UTRs. PNAS **102**: 15557–15562.

Smith, N. and A. Eyre-Walker, 2001 Synonymous Codon Bias Is Not Caused By Mutation Bias in G+C-Rich Genes in Humans. Mol Biol Evol **18**: 982–986.

Steel, M., 1994 Recovering a tree from the leaf colourations it generates under a Markov model. Appl. Math. Lett. **7**: 19–24.

Vinogradov, A. E., 2003 Isochores and tissue-specificity. Nucleic Acids Res **31**: 5212–5220.

WELLS, K., J. FOSTER, K. MOORE, V. PURSEL and R. WALL, 1999 Codon optimization, genetic insulation, and an rtTA reporter improve performance of the tetracycline switch. Transgenic Res. **8**: 371–381.

ZOLOTUKHIN, S., M. POTTER, W. HAUSWIRTH, J. GUY and N. MUZYCZKA, 1996 A 'humanized' green fluorescent protein cDNA adapted for high-level expression in mammalian cells. J. Virology **70**: 4646–4654.

Table 1: RefSeq Genes by Gene and Isochore AT- or CG-richness

	Gene type		
Isochore type	AT	CG	Total
iAT	28%	$\leq$ 7%[1]	35%
iCG	19%	46%	65%
Total	47%	53%	100%

[1] iAT/CG genes are likely to be overcounted (see text).

Table 2: Amino acid usage changes, fish orthologs to human CG-rich genes

Amino acid	% change	Relevant codons	Largest four "came from" (when +%), or "went to" (when −%), and codon change			
Pro	+20		Ser	Leu	Ala	Thr
		ccn	(t→c)cn	c(t→c)n	*(g→c)cn*	(a→c)cn
Ala	+19		Ser	Thr	Val	Gly
		gcn	(t→g)cn	(a→g)cn	g(t→c)n	*g(g→c)n*
Gly	+13		Ser	Ala	Glu	Arg
		ggn	(a→g)gn	*g(c→g)n*	g(a→g)n	(a→g)gn
Arg	+6	cgn	Lys	Gln	Ser	Leu
		agn	a(a→g)n	c(a→g)n	(tc→cg)n	c(t→g)n
			. . .			
Lys	−13		Arg	Gln	Glu	Ser
		aan	a(a→g)n	(a→c)an	(a→g)an	*complex*
Met	−17		Leu	Val	Ile	Thr
		atg	(a→c)tg	(a→g)tg	*at(g→n)*	a(t→c)g
Ile	−18		Val	Leu	Thr	Ala
		atn	(a→g)tn	(a→c)tn	a(t→c)n	(at→gc)n
Asn	−21		Ser	Asp	Gly	Thr
		aan	a(a→g)n	(a→g)an	(aa→gg)n	a(a→c)n

All codon changes increase $C + G$ except five in *italics*, which are neutral.

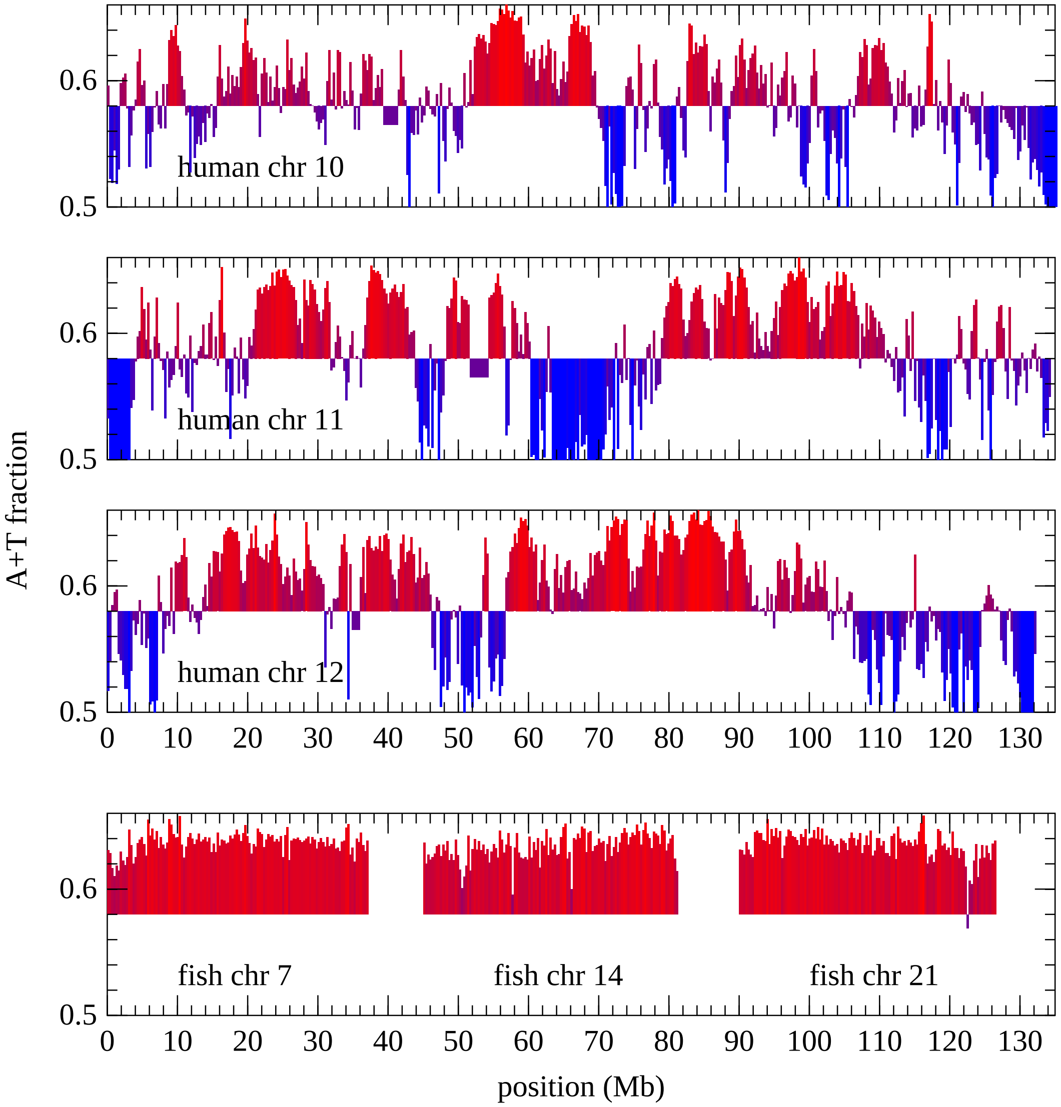

Figure 1: Local A+T fraction of typical human and zebrafish chromosomes. Counts are shown in nonoverlapping 300 kb windows.

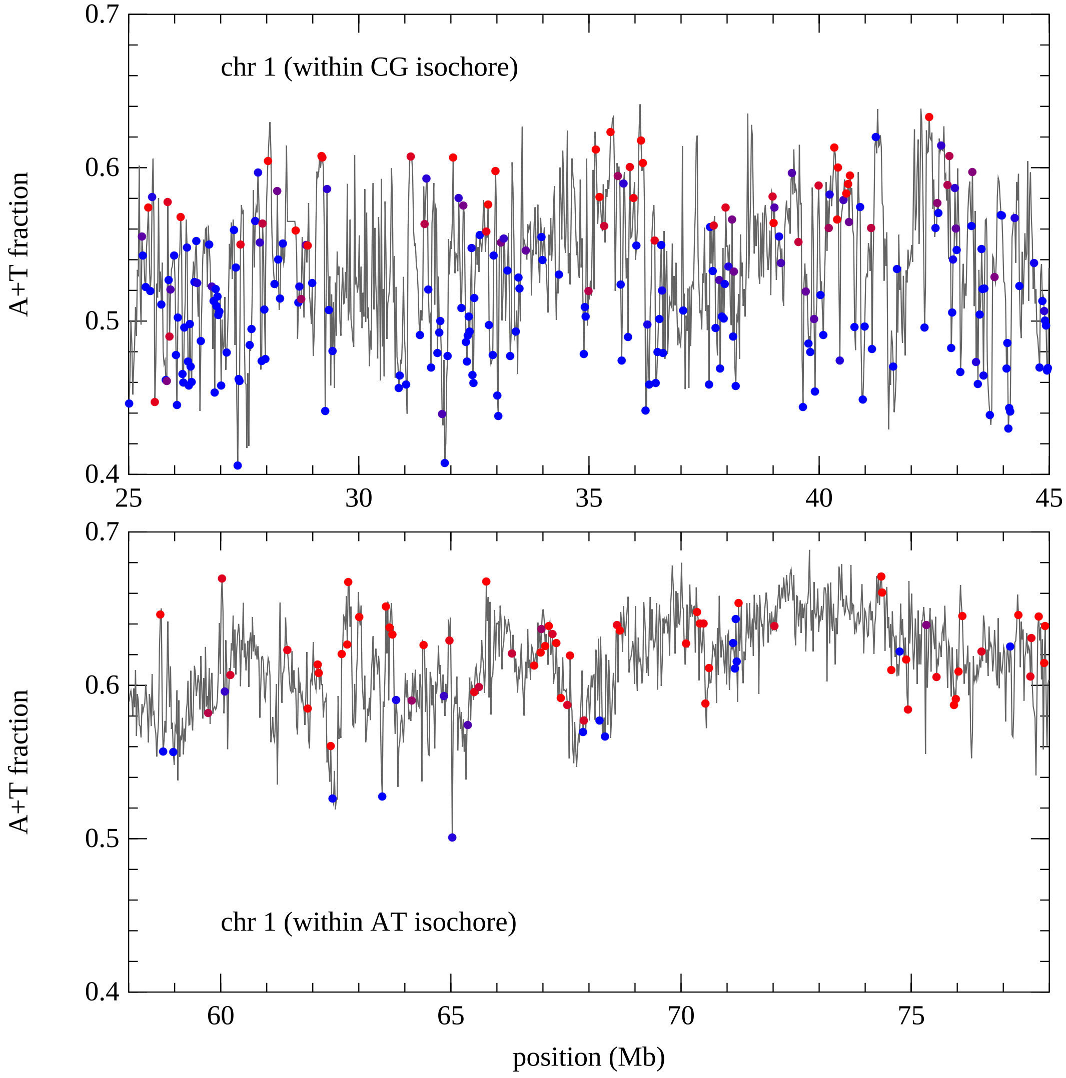

Figure 2: Two regions of human chromosome 1, plotting $A + T$ counts in 20 kb windows, and showing the location of all RefSeq genes. Genes are plotted at the $A + T$ value of the window in which they occur, but with their color continuously varying from red (AT-rich gene) to blue (CG-rich gene). Genes in a CG isochore (upper panel), notably CG-rich genes, tend to be more extreme than their flanking regions; there is less such tendency in an AT isochore (lower panel).

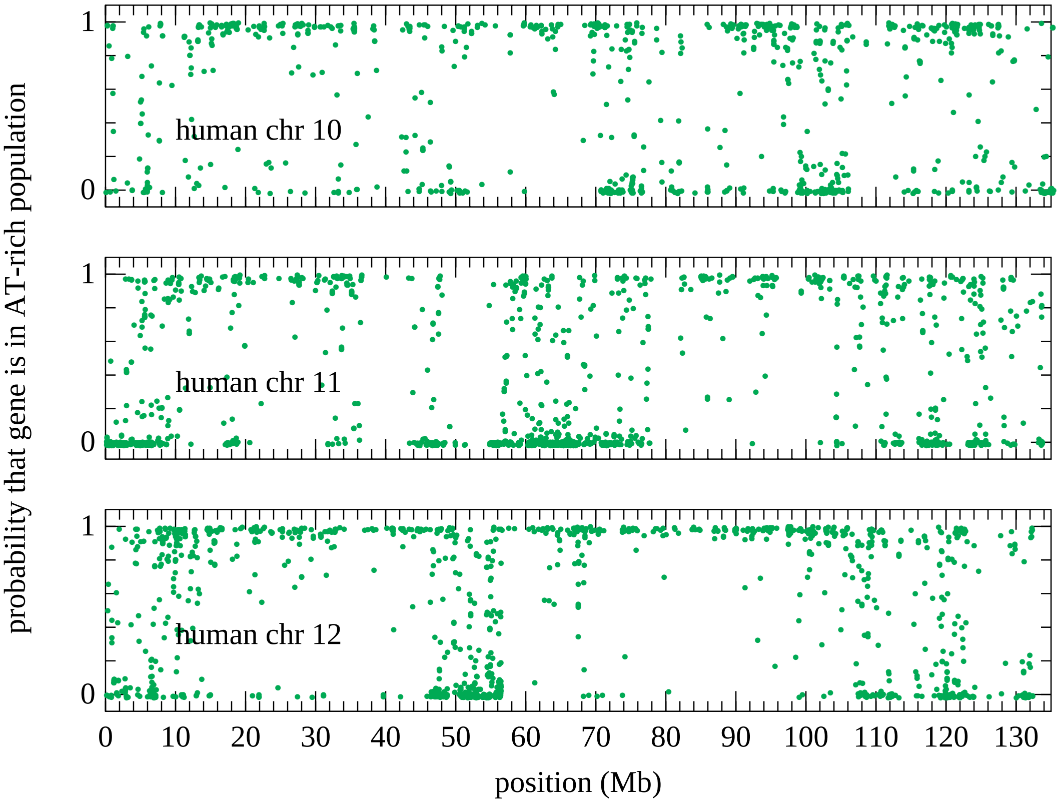

Figure 3: RefSeq genes plotted according to their probability of being in the AT-rich population, for three typical chromosomes. Large regions of AT-rich genes, and of mixed AT- and CG-rich genes, are evident. Large regions of CG-rich genes are conspicuously absent.

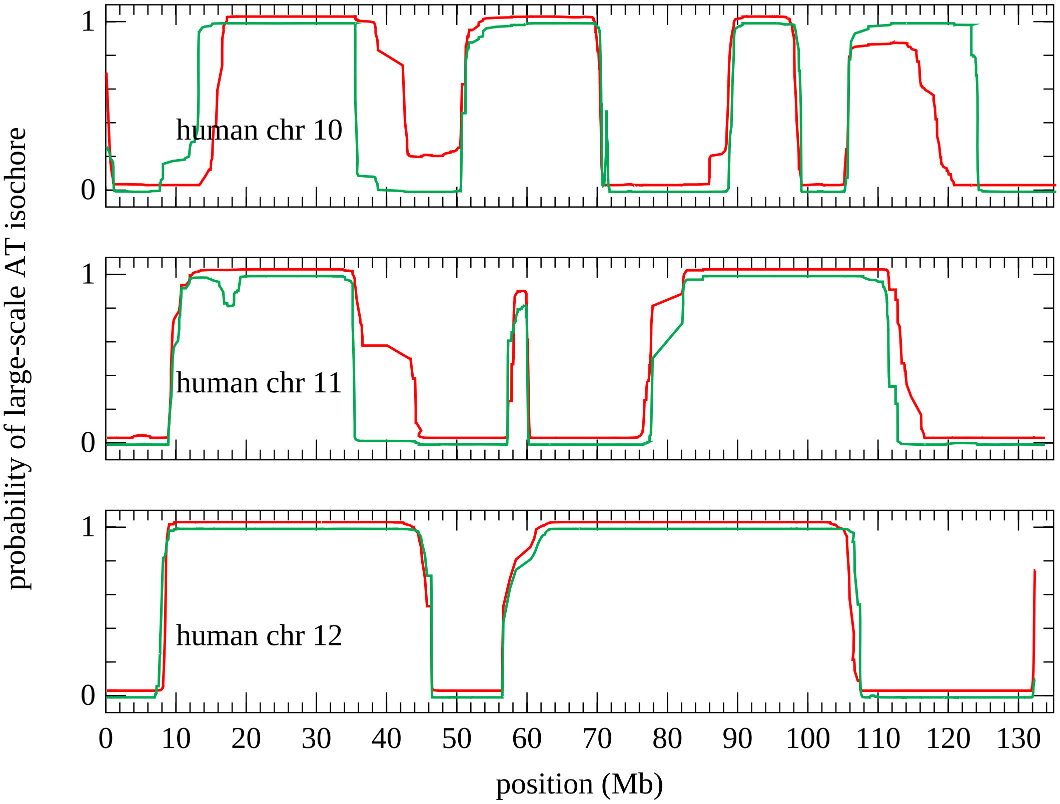

Figure 4: Green line: Isochore boundaries obtained by applying a Markov model with two states: "AT-rich genes" and "mixed genes". Red line: Isochore boundaries obtained by a similar model using raw counts in 300 kb windows.

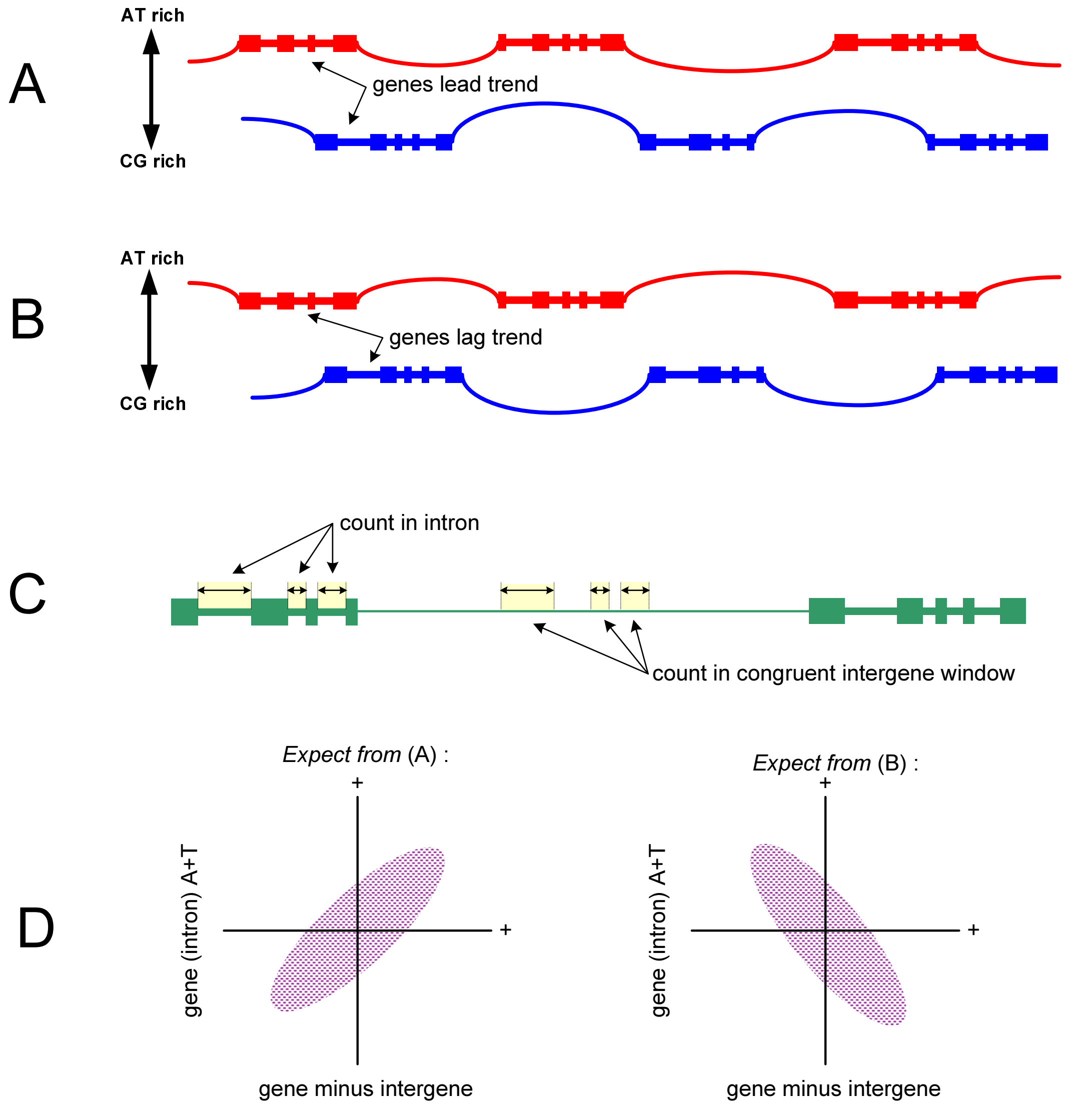

Figure 5: Strategy for measuring whether genes are more (A) or less (B) compositionally extreme than their immediate surroundings. The gene's composition is measured by counts in its introns only (C). Counts in the adjacent intergenic region are made with an identical window function. We expect (D) a positive correlation between gene and gene-minus-intergene if genes are compositionally more extreme, a negative correlation if less extreme.

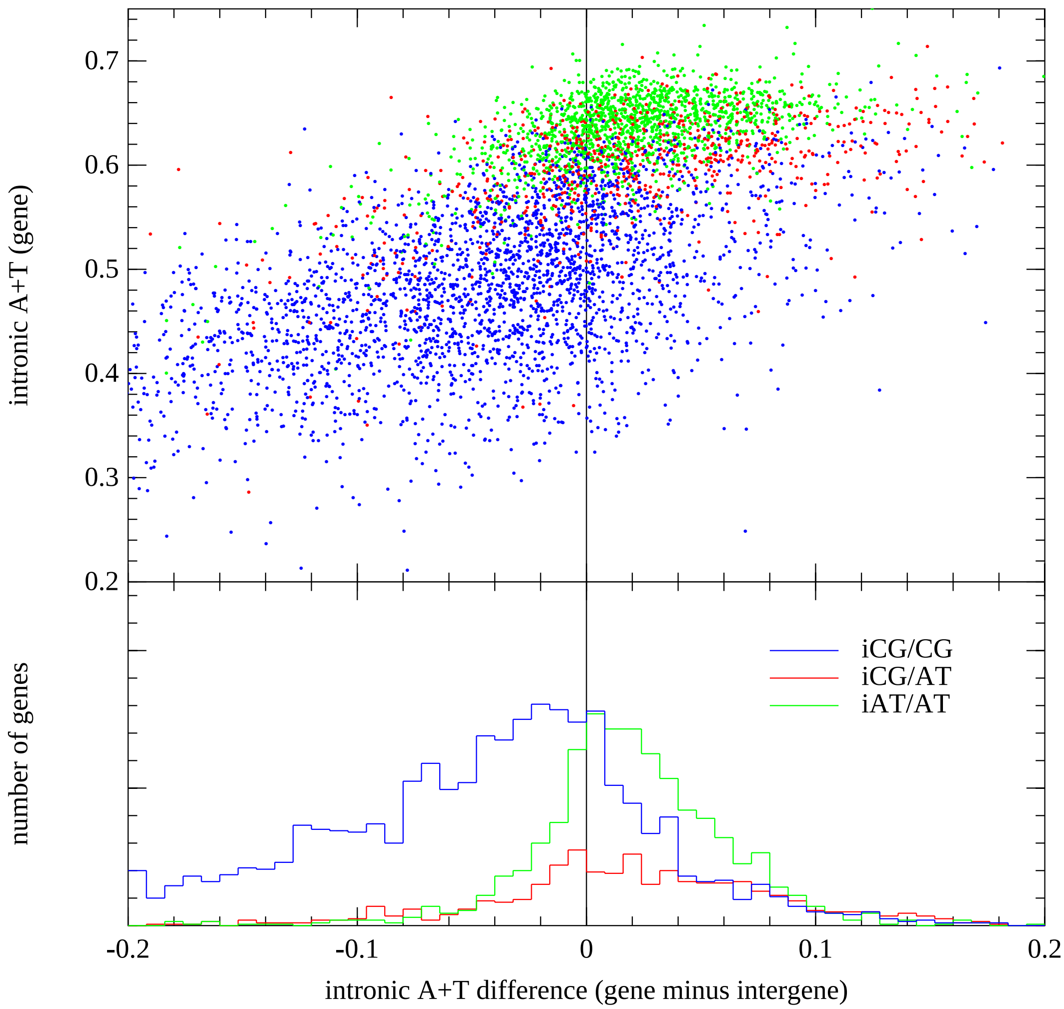

Figure 6: Results of testing whether genes are more or less compositionally extreme than their surroundings. Blue, red, and green denote respectively CG genes in CG isochores (iCG/CG), AT genes in CG isochores (iCG/AT), and AT genes in AT isochores (iAT/AT). All three gene types tend to be more extreme than their immediate surroundings, most strongly for iCG/CG (compare Figure 5D).

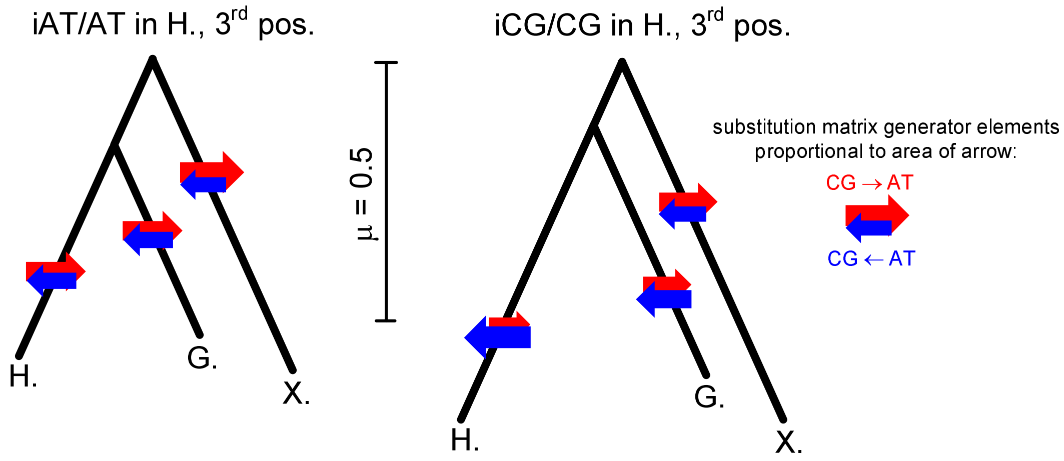

Figure 7: Maximum likelihood reconstruction of branch lengths and transition matrices derived from synonymous 3rd codon positions for orthologous genes that are iAT/AT (left tree) and iCG/CG (right tree) in human. Shown are ancestral branches of human, chicken (G. gallus), and frog (X. tropicalis). Red and blue arrows summarize mutational biases towards AT or CG richness, respectively. The expected bias towards CG is seen in iCG/CG genes on the human and chicken branches. Not seen, however, is any large increase in apparent mutation rate (branch length), as would result from significant positive selection. Branch lengths μ are in units of mutations per base position. The trees are rooted by using an additional species (D. rerio) as an outgroup.

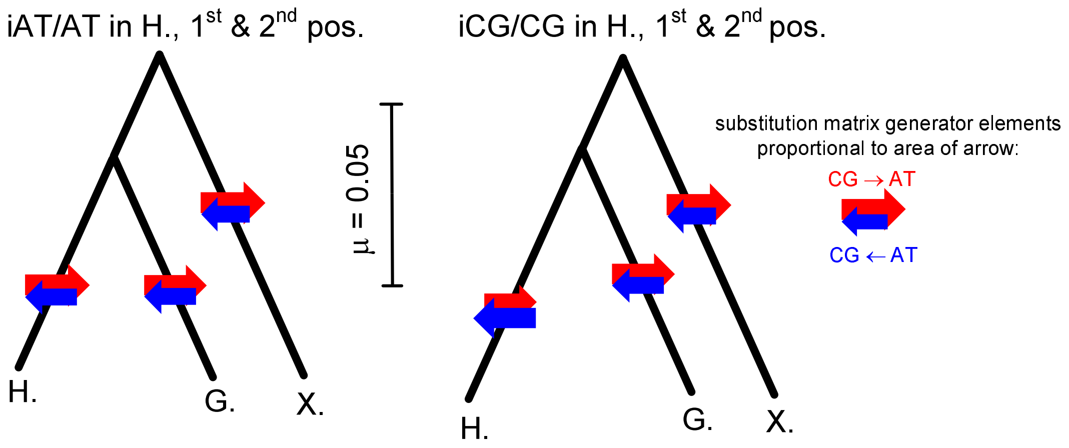
iAT/AT in H., 1st & 2nd pos.
H.
G.
X.
μ = 0.05
iCG/CG in H., 1st & 2nd pos.
H.
G.
X.
substitution matrix generator elements proportional to area of arrow:
CG → AT
CG ← AT

Figure 8: Same as Figure 7, but now for 1st and 2nd codon positions, where mutations cause protein amino acid changes. Note change of scale from Figure 7. Evolutionary distances are found to be reduced by a factor of ≈ 6, but the Figure is otherwise almost a scaled copy of Figure 7.

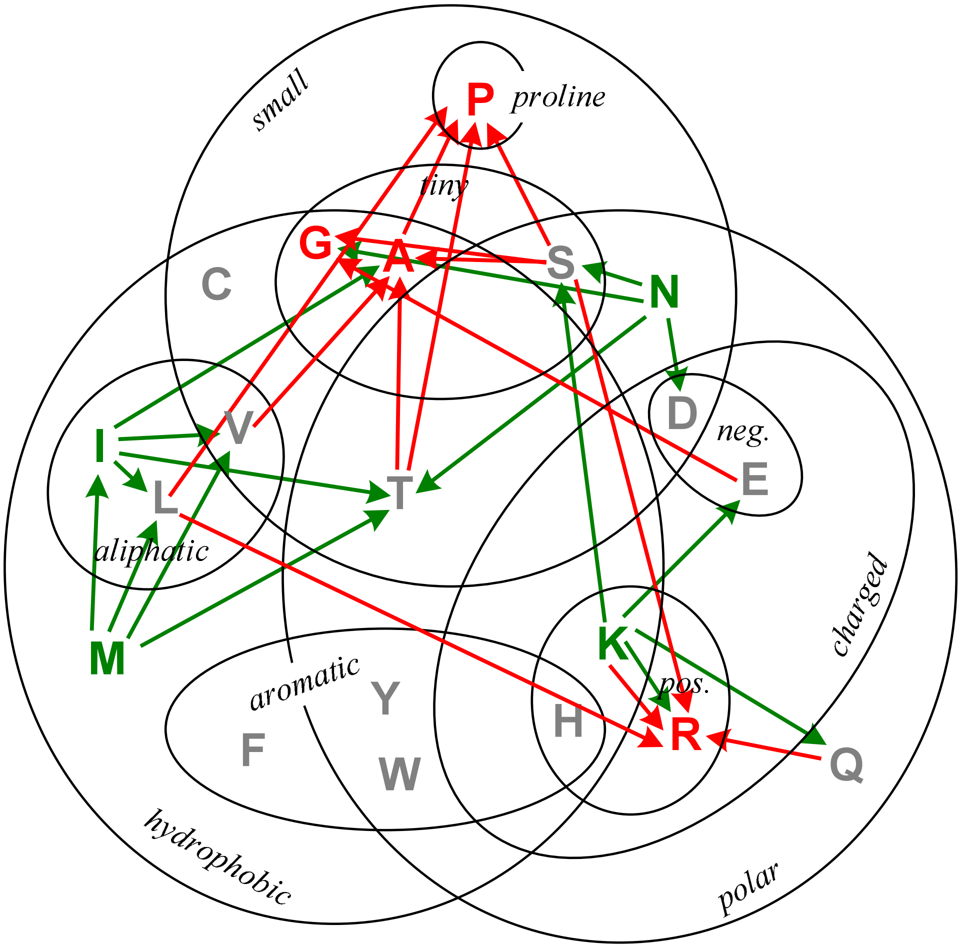

Figure 9: Principal amino acid usage differences between fish and human orthologs, for CG-rich human genes in CG isochores. Green denotes fractionally most decreasing, red most increasing, amino acids. For these, the four most frequent substitutions are shown. The direction of all arrows is a comparison from fish to human. (This is a graphical convention, not an arrow of time.) The observed trends make little sense biochemically, but can all be explained by a strong preference for amino acids with CG-rich codons in human. (Underlying diagram after Betts and Russell 2003.)