A computational prediction of isochores based on hidden Markov models

Keywords: isochores, genome, Markov, model,

Christelle Melodelima[a,*], Laurent Guéguen[a], Didier Piau[b] and Christian Gautier[a]

[a]*Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS 5558, Université Claude Bernard – Lyon I, Lyon, France*
[b]*Institut Camille Jordan, UMR CNRS 5208, Université Claude Bernard – Lyon I, Lyon, France*

Address for correspondence:

Christelle Melodelima
Laboratoire de Biométrie et Biologie Évolutive,
UMR CNRS 5558,
Université Claude Bernard Lyon I, 43 bd. Du 11 Novembre 1918
69622 Villeurbanne Cedex
France
Tel: +33 (0)4 72 44 81 42
Fax: +33 (0)4 72 43 13 48
E-mail: melo@biomserv.univ-lyon1.fr

ABSTRACT

Mammalian genomes are organised into a mosaic of regions (in general more than 300 kb in length), with differing, relatively homogeneous *G+C* contents. The *G+C* content is the basic characteristic of isochores, but they have also been associated with many other biological properties. For instance, the genes are more compact and their density is highest in *G+C* rich isochores. Various ways of locating isochores in the human genome have been developed, but such methods use only the base composition of the DNA sequences. The present paper proposes a new method, based on a hidden Markov model, which takes into account several of the biological properties associated with the isochore structure of a genome. This method leads to good segmentation of the human genome into isochores, and also permits a new analysis of the known heterogeneity of *G+C* rich isochores: most (60%) of the *G+C* poor genes embedded in *G+C* rich isochores have UTR sequences characteristic of *G+C* rich genes. This genomic feature is discussed in the context of both evolution and genome function.

1. Introduction

Isochores were originally identified as a result of gradient density analysis of fragmented genomes (Macaya et al., 1976): mammalian genomes are a mosaic of regions (DNA segments on average more than 300 kb in length) with differing, homogeneous *G+C* contents. High, Medium and Low-density genomic segments are called H, M and L isochores in order of decreasing *G+C* content respectively. The isochore has been classified as a "fundamental level of genome organisation" (Eyre-Walker and Hurst, 2001) and this concept has increased our appreciation of the complexity and variability of the composition of eukaryotic genomes (Nekrutenko and Li, 2000). Many important biological properties have been associated with the isochore structure of genomes. In particular, the density of genes has been shown to be higher in H- than in L isochores (Mouchiroud et al., 1991). Genes in H isochores are more compact, with a smaller proportion of intronic sequences, and they code for shorter proteins than the genes in L isochores (Duret et al., 1995). The amino-acid content of proteins is also constrained by the isochore class: amino acids encoded by *G+C* rich codons (alanine, arginine….) being more frequent in H isochores (D'Onofrio et al. 1991, Clay et al., 1996). Moreover, the insertion process of repeated elements depends on the isochore regions. SINE (short-interspersed nuclear element) sequences, and particularly *Alu* sequences, tend to be found in H isochores, whereas LINE (long-interspersed nuclear element) sequences are preferentially found in L isochores (Jabbari and Bernardi, 1998).

The recent availability of the draft human genome sequence has allowed direct tests of the isochore model, and it was hoped that isochores could be identified at the sequence level. Since then, there has been considerable debate about the existence of isochores in the human genome. Different approaches have been developed for isochore prediction. A *G+C*-plot thus routinely accompanies the publication of every new genome sequence, the long-range patterns within the plots usually being identified only by eye. This has happened, for instance, with the isochores tentatively identified on human chromosomes 21 (Hattori et al., 2000) and 22 (Dunham et al. 1999). Other methods based on sliding windows use one model independently of the sequence type (exon, intron…) to test sequence homogeneity (Nekrutenko and Li, 2000). However, Häring and Kypr (Haring and Kypr, 2001) denied the existence of isochores in the human chromosomes 21 and 22, and Lander et al. (2001) concluded that isochores do not appear to deserve the prefix "iso". The methodological problem with these studies is that the nucleic bases were considered as independent from each other and local heterogeneity was neglected. This leads to the

conclusion that only highly repetitive DNA sequences are homogeneous. However, when the heterogeneity within isochore families was quantified (Cuny et al., 1981), it was shown that the homogeneity of isochores is only relative, leading to their definition as "relative homogeneous" regions (Bernardi, 2001).

An alternative tool for the analysis of genome heterogeneity is compositional segmentation. Windowless methods have been developed to calculate the *G+C* content, and some have been used to identify isochores in various genomes. These methods have also been called "DNA segmentation methods" (Bernaola-Galvan, 2001). Among them, we could mention the method of entropic segmentation (Li et al., 2003, Oliver et al., 2004) and the Z-curve method (cumulative *G+C* profile), which leads to a unique representation of DNA sequences (Zhang and Zhang, 2003). These windowless methods reveal isochore structures in the human genome.

These different methods use only the overall base composition of the DNA sequence to predict isochores. However, the statistical characteristics of the *G+C* content differ in the coding and non-coding regions of vertebrate genes. For example, Isofinder (Oliver et al., 2004) is a segmentation algorithm based only on the *G+C* content, and it ignores local heterogeneity due, for example, to differences between coding and non coding regions. Compositional domains are characterised, but the real isochore segmentations of the human genome are not, probably because the output of this program represents segments that are often less than 300 kb. The human genome consists of many nested structures (chromosome, isochore, gene …). Thus, for the analysis of the isochore organisation of genomes, in this study we propose a new method based on hidden Markov models, which takes genes as the local structure.

Hidden Markov models are a challenging approach to describing the compositional properties of chromosome-size DNA sequences. HMMs were first applied to the analysis of genetic data by Churchill (1989), who intended to analyse the compositional heterogeneity of natural DNA sequences. More recently, Peshkin (1999) has shown that HMMs can be used for both further structural analysis and direct biological interpretation. The objective of the present paper is therefore to propose a method, based on a hidden Markov model, which makes it possible to detect and analyse the isochore structure of the human genome within a reasonable period of time. To improve the isochore prediction capacity, we have introduced the idea of using an HMM that takes into account not only the *G+C* content of the DNA sequence, but also several biological properties associated

with the isochore structure of the genome (such as gene density, differences in the *G+C* content of different regions of the gene, lengths of exons and introns).

2. Materials and methods

Gene sequences were extracted from Hovergen (Homologous Vertebrate Genes Database, March 2003 release 43) (Duret et al. 1994) for the human genome. To ensure that the data concerning the intron/exon organisation was correct, we restricted our analysis to genes of which the RNA transcripts have been sequenced. To avoid distortion of the statistical analysis, redundancy was discarded. This procedure yielded a set of 5034 multi-exon genes and 817 single-exon (that is, intronless) genes. Three classes were defined based on the *G+C* frequencies at the third codon position ($G+C_3$). The limits were set so that the three classes contained approximately the same number of genes. This yielded classes H=[100%, 72%], M=]56%,72%[and L=[0%,56%], which were used to build a training set. These classes were roughly the same as those used by other authors (Mouchiroud et al., 1991, Zoubak et al. 1996). To test the model, data on all human chromosomes were retrieved from ENSEMBL.

We propose here a new method to detect isochores and analyse their structure along the human genome. To characterize the three isochore regions (H, L and M), three HMM models (*H, L* and *M*) were adjusted using the training sets, and then compared on all human chromosomes. In an HMM model, the duration of stay in each state follows a geometric law. The empirical length distributions of intergenic and intronic regions are geometric, but this is not the case for exons (Burge and Karlin, 1997), as shown by the bell-shaped histograms obtained. Thus, hidden Markov models cannot represent the exact length distribution of exons. To model the empirically-obtained, bell-shaped length distributions of the exons, we used sums of a variable number of geometric laws with equal or different parameters (Melodelima et al., 2004). Thus, each region (intergenic, intronic or exonic) is represented by a macro-state in the HMM (Figure 1). Exons consist of a succession of codons, and each of the three possible positions in a codon (1, 2, 3) has characteristic statistical properties. This implies the need to divide exons into three states (Burge et Karlin, 1998, Borodovsky and McIninch 1993). Moreover this model takes into account the direct and reverse strands of the DNA sequences. Exon states are separated into two categories, corresponding to the direct coding state and the reverse coding state. HMMs take into account the dependency between a base and its *n* preceding neighbours.

In this case, the order of the model is *n*. For our study, *n* was taken to be equal to 5, as in the studies of Borodovsky and Burge (Burge et Karlin, 1998, Borodovsky and McIninch 1993). The emission probabilities of the HMM were therefore estimated from the frequencies of 6-letter words in the different regions (intron, initial exon, internal exons and terminal exon) that made up the training set. We found that the initial exons exhibited a very specific pattern, because half of them contained a peptide signal (Melodelima et al. 2004). Thus, the macro-state initial exon was split into two states. The three models were trained and adjusted separately using the H, L and M sets.

Our HMM method was used to identify isochores within the human genome. We divided the DNA of each human chromosome into overlapping, 100-kb segments. Two successive segments overlapped by half their length. For each segment and for each model (*H*, *L* and *M*), the probability P[*Mod* |*S*] was computed as follows, where *Mod* is *H*, *L* or *M*, and *S* the segment that is being tested. This probability is obtained by a Bayesian approach:

$$P(Mod|S) = \frac{P(S|Mod) \times P(Mod)}{\sum_{m \in \{H,M,L\}} P(S|m) \times P(m)},$$

P(S|Mod) is computed by the forward algorithm, and *P(Mod)* is estimated by the frequency of genes according to their $G+C_3$ content (due to our definition of $G+C_3$ limits we get P(*H*) ≈ P(*M*) ≈ P(*L*) ≈ 1/3, and so our Bayesian approach is numerically very close to a maximum likelihood approach). To be consistent with the preceding definition, we assumed an isochore to be a region consisting of at least 5 consecutive windows associated by the method with the same isochore class. This ensured that all estimated isochore lengths were greater than 300 kb, but meant that some windows are not associated with any isochore class.

The *G+C* rich regions are known to display high variability. For instance, some of the genes annotated by Ensembl have a low $G+C_3$, but were still classified as being H isochores. In order to investigate our isochore classes in detail, we extracted all the genes annotated by Ensembl that were contained in our isochores. The different sequences (CDS, gene, introns) that compose these genes were analysed separately in two steps. First, sub-models representing the CDS, introns and gene (HMM "CDS", HMM "intron" and HMM "gene" respectively) were extracted from the HMMs *H* and *L*. Second, for each type of sequence (CDS, gene, introns), the predictions of the *H* and *L* sub-models were compared using the maximum likelihood procedure. These findings were compared to the $G+C_3$ content of the gene, thus making it possible to investigate isochore heterogeneity.

3. Results

3.1 Isochore chromosome map

The quality of discrimination between isochore classes for a window can be measured by $\max_{H,L,M}$(P(Mod/S)). For each of the 56417 windows in the human genome this max value was greater than 0.75, leading to a very clear association between each window and a unique isochore class. A second important criterion was the isochore length, since the method imposes a minimum length of 300 kb, resulting in some unaffected windows. In the human genome, unallocated windows constitute only 5% of the total number of windows, and these windows were not considered to constitute an isochore. These 5% reveal the presence of a strong structure in the genome. If the genome were not structured, the number of unallocated windows would be 74%.

The human genome segmentation we obtained is shown in Figure 2. This segmentation correlates well with the *G+C* content. The mean *G+C* contents were 0.515 (σ=0.035), 0.450 (σ=0.012) and 0.395 (σ=0.017) respectively for the H, M and L isochore classes as defined by our HMM method. The Kruskal-Wallis non-parametric test was significant (p-value<10^{-8}), as were the pairwise comparisons using the Wilcoxon test (p-value = 10^{-4}, 5.10^{-4}, 3.10^{-2} for the H/L, M/L, H/M comparisons, respectively). Our approach is also fully compatible with the view of the human genome as a mosaic of regions with relatively homogeneous *G+C* contents (Bernardi, 2000, Li et al. 2003). Our data contradict the suggestion of Eyre-Walker and Hurst (2001) that the isochore structure accounts for only some parts of the genome, and confirm the findings of Oliver et al. (Oliver et al., 2002). These results are also confirmed by the close correlation between the *G+C* of the isochore, and the $G+C_3$ content of the genes contained in the isochore (Figure 3), with $R^2 = 0.43$.

The length of the isochores detected by their *G+C* content is known to depend on the isochore class, with L > M > H (De Sario et al., 1996). This is what we found here. Although the minimum length of isochores that can be predicted by our method was 300 kb, we were also able to predict much longer isochores as shown in Figure 4. The average length for L isochores was 7.71 Mb, whereas the average length for the H and M isochores was 2.93 Mb and 5.74 Mb respectively. These lengths were significantly different (Kruskal-Wallis p-value<10^{-8}). Figure 4 shows the length distribution of all the isochores found in the human genome by our method.

3.2 Gene characteristics and isochore classes

Figure 2 shows the well-known relationships between isochore class and gene density (Mouchiroud et al., 1991, Zoubak et al., 1996, Bernardi, 2000). For all chromosomes, we found that the isochore structure closely parallels the gene density distribution along the chromosome. The gene density in the H regions (15 genes per Mb) was higher than that in the L regions (3.7 genes per Mb), leading to a significant Wilcoxon test (p-value = $4.77.10^{-5}$). The same difference was observed when we compared the characteristics of the M region (7.5 Mb) with those of the H (p-value = 6.10^{-8}) and L regions (p-value = 2.10^{-3}). The Kruskal-Wallis non-parametric tests were also performed on other biological properties ($G+C_3$ of CDS and the $G+C$, the length and the number of introns, see Table 1).

3.3 Analysis of isochore heterogeneity

Although we did not have all the criteria required to validate our segmentation or the definition of isochore classes, we have demonstrated that "our" isochore classes satisfy all the usual properties of isochore classes. In the remainder of the text we use the word "isochores" to mean the homogeneous segments found by the HMM method. The characteristic of our approach is the homogeneity, at the 100 kb scale, of isochores. Due to their great variability it may be difficult to recognize $G+C$ rich regions by a simple $G+C$-plot of several chromosomes. We have shown that genes with $G+C_3 > 0.72$, $0.72 > G+C_3 > 0.56$ and $G+C_3 < 0.56$ have relative frequencies of 48%, 35% and 17% respectively in the H isochores.

The presence of genes annotated by Ensembl as having low $G+C_3$ content in H isochores could result from i) heterogeneity of the H isochores; ii) an error in our estimation of isochore limits; iii) an imperfect correlation between $G+C_3$ content and isochore class. The last possibility refers to the existence of characteristics, other than $G+C_3$, that could be shared by genes located within H isochores. Our approach provides a way of carrying out a detailed analysis of any genes of which the $G+C_3$ content is incoherent with their isochore class. To carry out such an analysis, the likelihood of H and L HMM models has been computed for all the genes annotated by Ensembl as having a $C+G_3$ content either greater than 0.72 or less than 0.56. These likelihoods have also been computed for all the subparts (CDS, introns, 5'UTR, 3'UTR) of these genes.

This first analysis was performed in the H isochores predicted by our method. 82% of the genes annotated by Ensembl as having a $G+C_3$ greater than 0.72 were classified as

belonging to the H class. Thus, the HMM "gene" *H* correctly describes the genes with a high $G+C_3$ content. However, 60% of the genes with a $G+C_3$ of less than 0.56 were assigned by our method to the H class (Figure 5). Our method identified two types of gene with a $G+C_3$ of less than 0.56: genes are recognized by the HMM "gene" *L* (40%) and those recognized by the HMM "gene" *H* (60%). This finding indicates that something other than the $G+C_3$ content could help us to characterise these genes. Figure 5 shows the influence of several regions (CDS, introns, UTRs, intergenic regions) on the gene predictions.

The classification of the genes with a $G+C_3$ of less than 0.56 does not depend on the characteristics of the CDS or of the introns, but only on those of the 5'UTR and 3'UTR regions. To confirm this hypothesis, we analysed the correlation between the predictions of the genes and the 5'UTR and 3'UTR regions by their respective HMMs. In 80% of cases (Table 2: sum of lines 1 and 4), the decision based on the HMM "gene" and on the HMM "5'UTR" was similar (the same results were obtained when the genes and the 3'UTR were compared). The UTR regions predicted in the *H* isochores had a *G+C* content (0.510 ± 0.0041) that was higher than that of the UTR region predicted in the L isochores (0.429 ± 0.0008). Therefore, when the HMM model finds an isochore in the human genome, two facts make it possible to classify the genes with $G+C_3$ <0.56 to the *H* isochore. First, the *G+C* content in the UTR regions influence the predictions of the HMM models (60% of cases). Second, there is a smoothing effect (40% of cases), *i.e.* the window around the gene influences the choice of the model (particularly the influence of the intergenic regions).

A similar analysis was performed for the L isochore class. The *G+C* poor regions were more homogeneous than the *G+C* rich regions. Thus, the distribution of genes allocated by Ensembl to the L isochores was 6%, 19% and 75% for genes in which $G+C_3$ was >0.72, between 0.56 and 0.72 and <0.56 respectively. Figure 6 shows that the classifications of the genes with $G+C_3 > 0.72$ do not depend on the characteristics of the CDS or of the introns, but only on the 5'UTR and 3'UTR regions, as for the H isochores. This hypothesis is confirmed by the correlation found between the prediction of the genes and the 5'UTR and 3'UTR regions by the models in 77% of the cases (see Table 3). These findings show the importance of regions outside the initiator and stop codon. Thus, 60% of the 17% of genes with $G+C_3$ <0.56 and classified as H isochores were classified by our method as belonging to the H class, and consisted mainly of UTRs and not introns. Similar results were obtained for genes with $G+C_3$ >0.72, and classified as L isochores (Figure 6).

3.4 Statistical correlations between the isochore class and the different regions of the genes.

In this part, we compare the biological properties of the intronic and UTRs regions to their classification in isochore class by our model. We can see that the length and $G+C$ content of these regions correspond to the choice of the model. Thus, the low $G+C_3$ genes inside H isochores and defined as H gene by our HMM were more precisely studied. The length of the introns was significantly shorter (p-value=5.10^{-4}) than the length of the introns of gene associated with a low $G+C_3$ content and predicted to belong to L isochores. Furthermore, the $G+C$ content of these introns was significantly higher (p-value=6.10^{-8}) than $G+C$ content of the introns of genes associated with a low $G+C_3$ content and predicted to belong to L isochores. Similarly, the length of the UTR was significantly shorter (p-value=7.10^{-24}) than the length of the UTR of genes associated with a low $G+C_3$ content and predicted to belong to L isochores. The $G+C$ content of these UTRs is significantly higher (p-value=8.10^{-84}) than $G+C$ content of the UTR of genes associated with a low $G+C_3$ content and predicted to belong to L isochores.

4. Discussion

The use of Markov models for the purpose of data exploration has been underestimated in genome analysis. Our study shows that simple hidden Markov models could be used to model human genome organisation, and to identify new biological structures. For example, it has previously been shown that the initial exons exhibit a very specific pattern, since half of them contain a peptide signal (Melodelima et al. 2004). We have also previously observed that the average duration of stay in the first state of the macro-state "initial exon" was 80 bases in length (Melodelima et al. 2004). This is consistent with biological knowledge about such signals, which are 45 to 90 bases long. Initial exons with no peptide signal, and the second parts of the initial exons, which do have a peptide signal, are statistically similar to internal exons and terminal exons, respectively.

The statistical characteristics of the coding and noncoding regions of vertebrates differ dramatically between the different isochore classes (Bernardi et al. 1985). The clarification of the isochore structure is a key to understanding the organisation and biological function of the human genome. We found that hidden Markov models were appropriate for each isochore class. The number of basic states for each isochore class (without taking the frame

and coding strand into account) in our model is only 7: first, internal and terminal introns and exons and "intergenic regions", the last state being used to model all non-coding regions of the genome and the introns inserted between two coding exons. This small number of states has made it possible to conduct a complete human genome analysis with reasonable CPU cost.

This method has clearly confirmed the finding (Macaya et al. 1976) that an isochore structure exists in the human genome (see Figure 2). The distribution of gene density along a chromosome closely matches the isochore structure identified here: the higher gene density regions are located in the *G+C*-isochores with the highest *G+C* content. Moreover, the relationship between isochore class and gene structure was clearly shown by our HMM approach. A very surprising finding emerged from the individual analysis of genes with a low $G+C_3$ content embedded in H isochores. The model recognized most of them (60%) as "H genes", despite their low $G+C_3$ content. This discrimination is based on the statistical properties of the non-coding regions of the primary transcript located before the beginning of the coding region and after the stop codon (referred to here as 5'UTR and 3'UTR respectively). One hypothesis would be that these genes have recently been recruited into the H isochore, and that only regions with low functional constraints have undergone sufficient mutations to adapt to their new genomic environment. But, in this case, introns would have evolved at least as fast as UTRs, which is not the case as introns are clearly of the L type. A more precise analysis of the history of these genes is necessary to enable us to understand the characteristics of low $G+C_3$ content H genes. Strong functional constraints linked to expression (transcription, chromatin organisation, …) may pattern non-coding regions in the neighbourhood of genes.

When the draft human genome sequence was made available, Lander et al. (2001) looked for isochores, but failed to find any. The existence of isochores in human chromosome 22 is questionable, and based only on a sequence analysis. The reason for the dispute is due to the lack of a sequence-based definition of isochores. The concept of an isochore is related to the concept of homogeneous domains extending over large scales (of hundreds of kilobases) within genomes, in which the *G+C* content can be considered to be relatively homogeneous. In the present study, the isochore structure was predicted by HMMs that are not based on *G+C* content alone. This method provides better prediction of isochores than classical methods. Three hidden Markov models were adjusted to each isochore class in order to take into account other biological properties associated with isochore classes *H*, *L* and *M* (such as the different lengths of the exons or introns, and the gene density, both of

which depend on the *G+C* content). These properties were ignored by conventional methods. In our case, this supplementary information allowed us to determine the precise boundary of the isochores.

Last year, a large number of genomes were sequenced. This huge amount of data makes it impossible to analyse patterns in order to provide a biological interpretation "by hand", and so mathematical and computational methods have to be used. Our approach, using HMMs, looks like a very promising method for analysing the organisation of genomes.

In conclusion, we have developed a computational method to predict isochores in the whole human genome using an HMM. This method is able to predict isochores of 300 kb, and clearly reveals a mosaic structure of the human genome. The isochores identified were separated into three classes classified as heavy, light and medium isochores, depending on their *G+C* content.

ACKNOWLEDGEMENTS

The computations were made at the PRABI and at the IN2P3 computer centre using a large computer farm (more than 1000 cpu). We are grateful to Dr Sagot, Dr. Duret and Dr Mouchiroud for their helpful comments about the manuscript.

REFERENCES

Bernaola-Galvan, P., Carpena, P., Roman-Roldon, R., Oliver, J.L. 2001. Mapping isochores by entropic segmentation of long genome sequences. In: Sankoff D, Lengauer T, RECOMB Proceedings of the Fifth Annual International Conference on Computational Biology, Montreal, Canada, ACM Press, New York, 217-218.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. Gene. **241(1)**, 3-17. Review.

Bernardi, G. 2001. Misunderstandings about isochores. Part 1. Gene. **276(1-2)**, 3-13. Review.

Bernardi, G. Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F. 1985. The mosaic genome of warm-blooded vertabrates. Science, **228(4702)**, 953-958.

Borodovsky, M., McIninch, J. 1993. Recognition of genes in DNA sequence with ambiguities. Biosystems. **30(1-3)**, 161-71.

Burge, C., Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. J Mol Biol. **268(1)**, 78-94.

Burge, C., Karlin, S. 1998. Finding the genes in genomic DNA. *Curr Opin Struct Biol.* 346-54. Review.

Churchill, G.A. 1989. Stochastic Models for heterogeneous DNA Sequences. Bull. Mathematical Biology. **51**, 79-94.

Clay, O., Caccio, S., Zoubak, S., Mouchiroud, D., Bernardi, G. 1996. Human coding and non coding DNA: compositional correlations. Mol Phyl Evol. **1**, 2-12.

Cuny, G., Soriano, P., Macaya, G., Bernardi, G. 1981. The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. Eur J Biochem. **115(2)**, 227-33.

De Sario, A., Geigl, E.M., Palmieri, G., D'Urso, M., Bernardi, G. 1996. A compositional map of human chromosome band Xq28. Proc Natl Acad Sci U S A. **93(3)**, 1298-302.

D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., Bernardi, G. 1991. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J. Mol. Evol. **32**, 504-510.

Dunham, I., et al. 1999. The DNA sequence of human chromosome 22. Nature. **402(6761)**, 489-95.

Duret, L., Mouchiroud, D., Gouy, M. 1994. HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res. **22(12)**, 2360-5.

Duret, L., Mouchiroud, D., Gautier, C. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J. Mol. Evol. **40**, 308-317.

Eyre-Walker, A., Hurst, L.D. 2001. The evolution of isochores. Nat Rev Genet. **2(7)**, 549-55. Review.

Haring, D., Kypr, J. 2001. Mosaic structure of the DNA molecules of the human chromosomes 21 and 22. Mol Biol Rep. **28(1)**, 9-17.

Hattori, M., et al. 2000. The DNA sequence of human chromosome 21. Nature. 405(6784),311-9. Erratum in: Nature. **407(6800)**, 110.

Jabbari, K., Bernardi, G. 1998. CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. Gene. **224**(1-2), 123-7.

Lander, E.S. et al. 2001. Initial sequencing and analysis of the human genome. Nature. **409(6822)**, 860-921. Erratum in: Nature.

Li, W., Bernaola-Galvan, P., Carpena, P., Oliver, J.L. 2003. Isochores merit the prefix 'iso'. Comput Biol Chem. **27(1)**, 5-10.

Macaya, G., Thiery, J.P, Bernardi, G. 1976. An approach to the organization of eukaryotic genomes at a macromolecular level. J Mol Biol., 108(1), 237-54.

Melodelima, C., Guéguen, L., Piau, D., Gautier, C. 2001. Modelling the length distribution of exons by sums of geometric laws. Analysis of the structure of genes and *G+C* influence , JOBIM, Montréal.

Mouchiroud, D., D'Onofrio, G., Aissani, B., Macaya, G., Gautier, C., Bernardi, G. 1991 The distribution of genes in the human genome. Gene. **100**, 181-7.

Nekrutenko, A., Li, W.H. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res. **10(12)**, 1986-95.

Oliver, J.L., Carpena, P., Roman-Roldan, R., Mata-Balaguer,T., Mejias-Romero,A., Hackenberg,M., Bernaola-Galvan,P. 2002. Isochore chromosome maps of the human genome. Gene. **300(1-2)**, 117-27.

Oliver, J.L., Carpena, P. Hackenberg, M., Bernaola-Galvan, P. 2004. IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Res.* **32**(Web Server issue):W287-92.

Peshkin, L., Gelfand, M.S. 1999. Segmentation of yeast DNA using hidden Markov models. Bioinformatics. **15(12)**, 980-6.

Zhang, C.T, Zhang, R. 2003. An isochore map of the human genome based on the Z curve method. Gene. **317(1-2)**, 127-35.

FIGURES CAPTIONS

Figure 1: An HMM is composed of states corresponding to different regions within the genome (exons, introns…). Each state emits DNA nucleotides (A, C, G and T) with specific emission probabilities. These states are interconnected by state transition probabilities. Many mathematical characteristics are known for HMMs, three of them are used here, for a given sequence and a given model: 1) the most probable path of the states can be exactly determined, 2) the probability of the observation sequence can be computed and 3) the probability that a state is the active state at a given position can be determined. The active state, S_n, for a position n depends on the state S_{n-1}, which is active at position $n-1$, and on the probabilities of transition from S_{n-1} to all other potentially possible states When the data were well defined in the training set, the model parameters would be estimated by the maximum likelihood method. States and transitions between states are represented by rectangles and arrows respectively.

Figure 2: Distribution of isochores along human chromosomes obtained by our method. The detected H, L and M isochores appear respectively in red, green and blue. To check the coherence of isochore prediction, the graph is shown with two other plots: a plot of the distribution of the gene density, and a plot of the $G+C$ content along the chromosome. (a) Chromosomes 1 to 6, (b) chromosomes 7 to 12, (c) chromosomes 13 to 18 and (d) chromosomes 19 to Y.

Figure 3: Density graph of the $G+C$ content of the isochore, and the $G+C_3$ content of the genes contained in the isochore ($R^2 = 0.43$).

Figure 4: The length distribution histogram of all the human isochores found by our method.

Figure 5: Relationship between the $G+C_3$ content of the genes in H isochores and their classification in isochore class by the model.

Due to the window size, genes with L statistical properties may occur inside an H isochore. The isochore class of a gene may be determined either by the classical method using $G+C_3$, or by a Bayesian approach using the models discussed here. In the latter case, either the whole gene or parts of it were taken

into account to elucidate the specificities of the gene inside H isochores. The figure shows the relative frequencies of sequences recognized by the model as H isochores. The crucial fact is that genes with low $G+C_3$ were recognized as an H sequence in 60% of cases for models based upon the whole gene sequences and sequences from UTR.

Figure 6: Relationship between the $G+C_3$ content of genes in H isochores and their classification in isochore class by the model.

As in Figure 5, but the result is clearer: genes with high G+C3 are recognized as L genes when they are located within an L isochore.

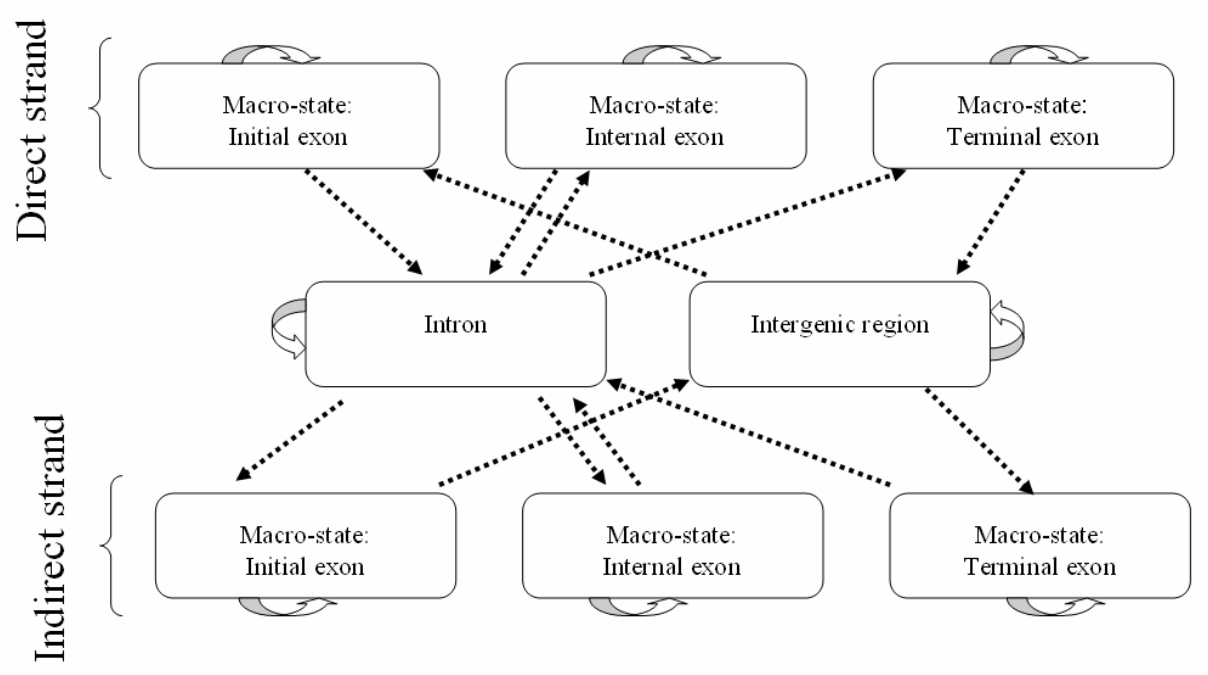
Direct strand
Macro-state:
Initial exon
Macro-state:
Internal exon
Macro-state:
Terminal exon
Intron
Intergenic region
Indirect strand
Macro-state:
Initial exon
Macro-state:
Internal exon
Macro-state:
Terminal exon

Figure 1

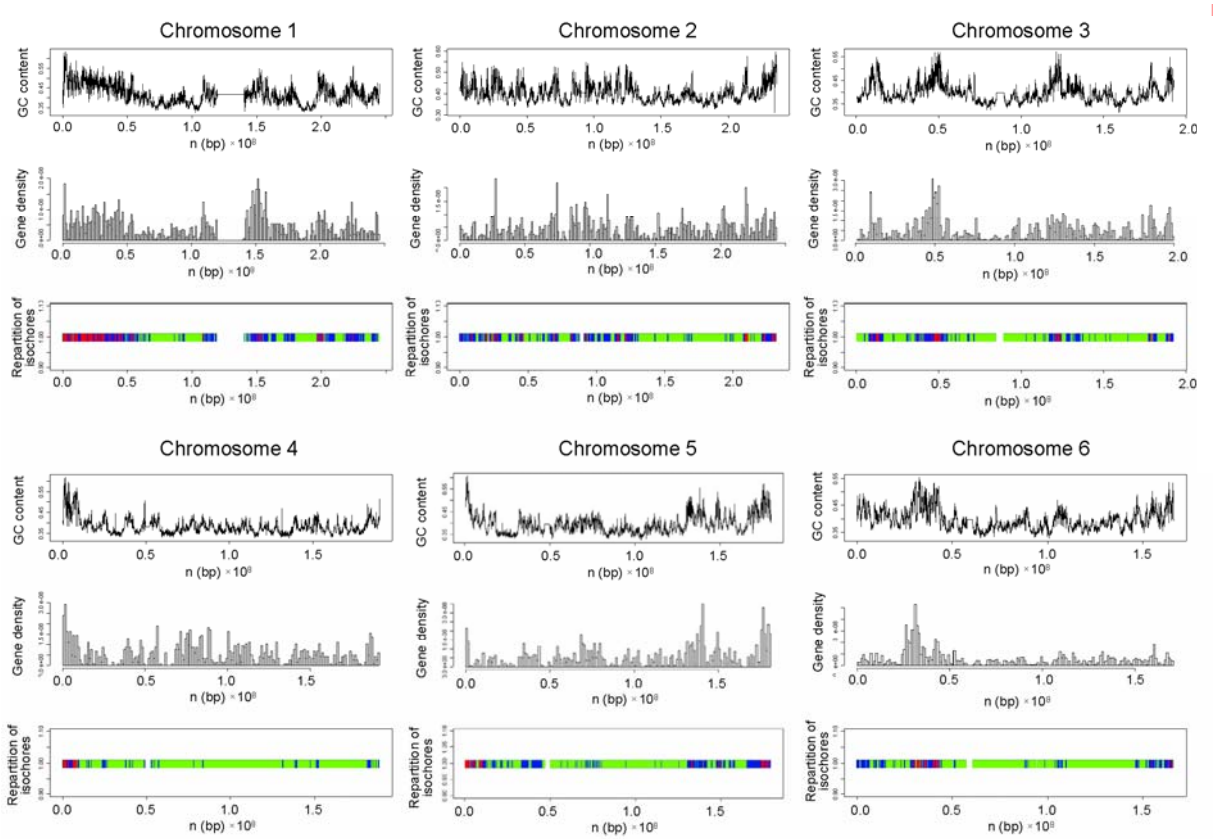

Figure 2a

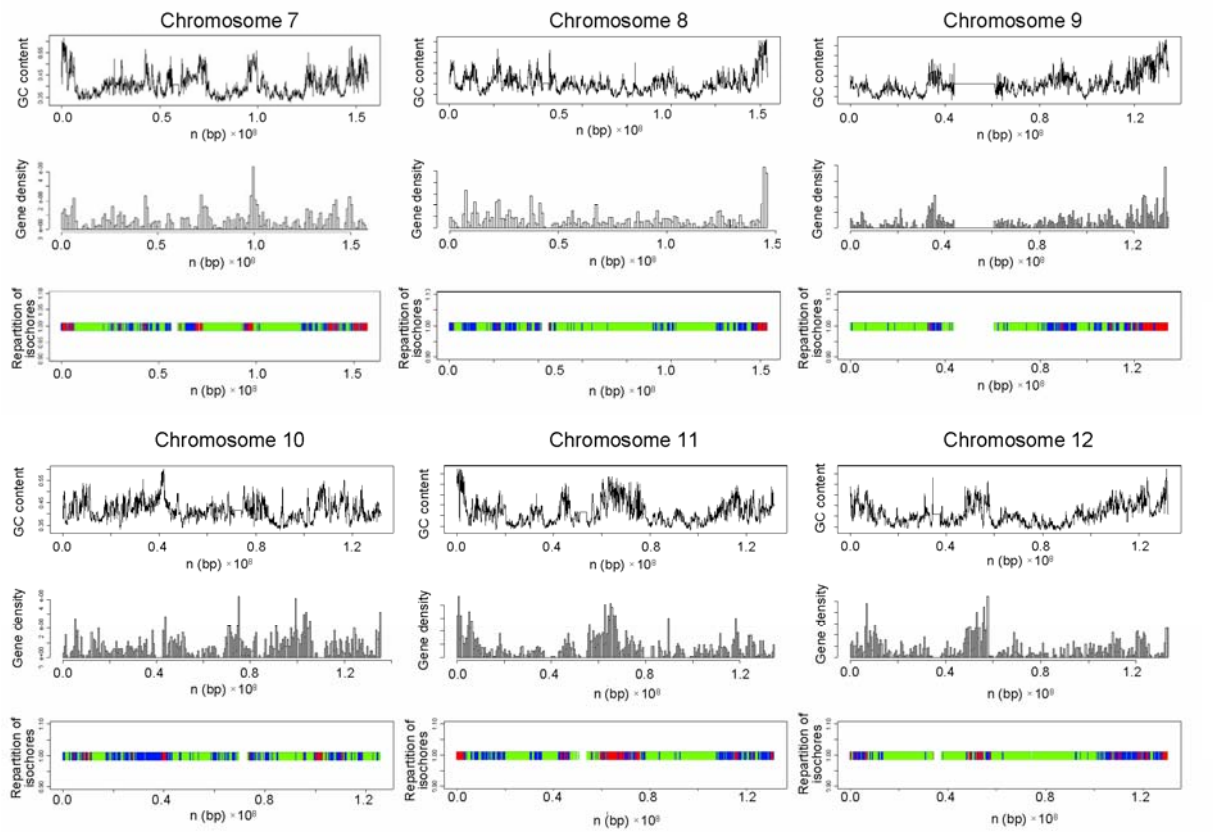

Figure 2b

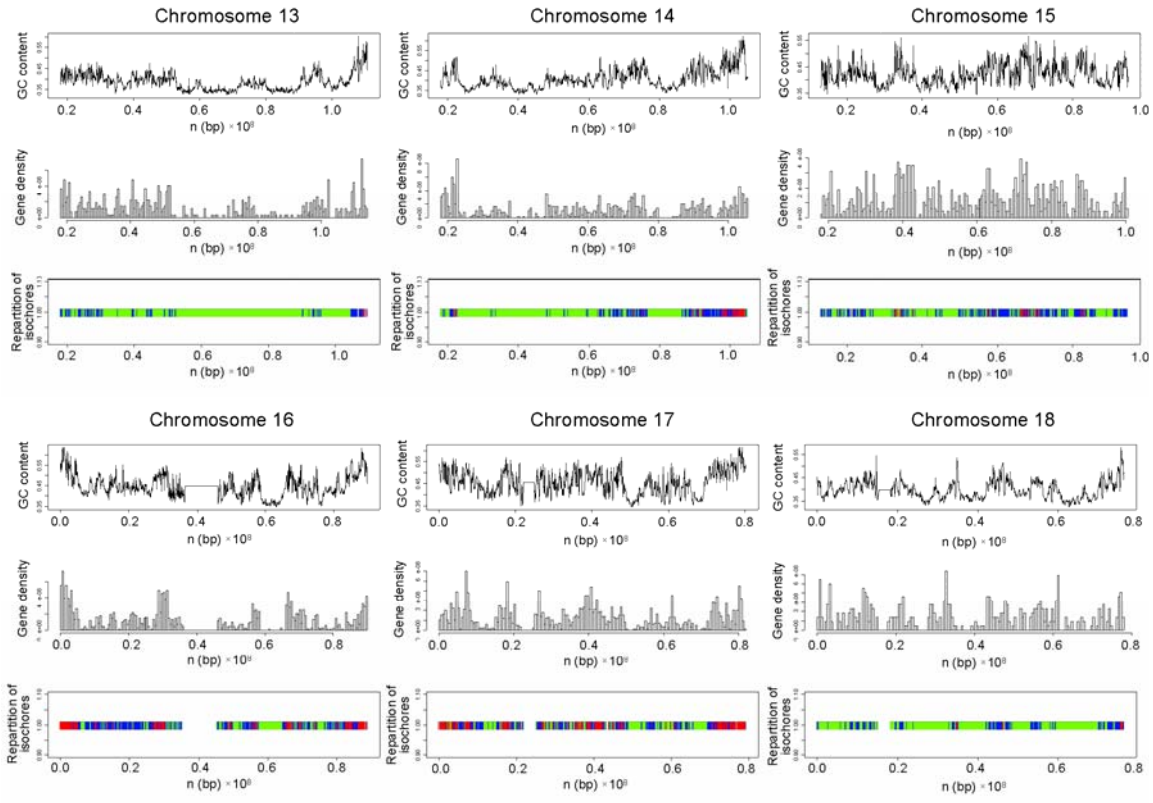

Figure 2c

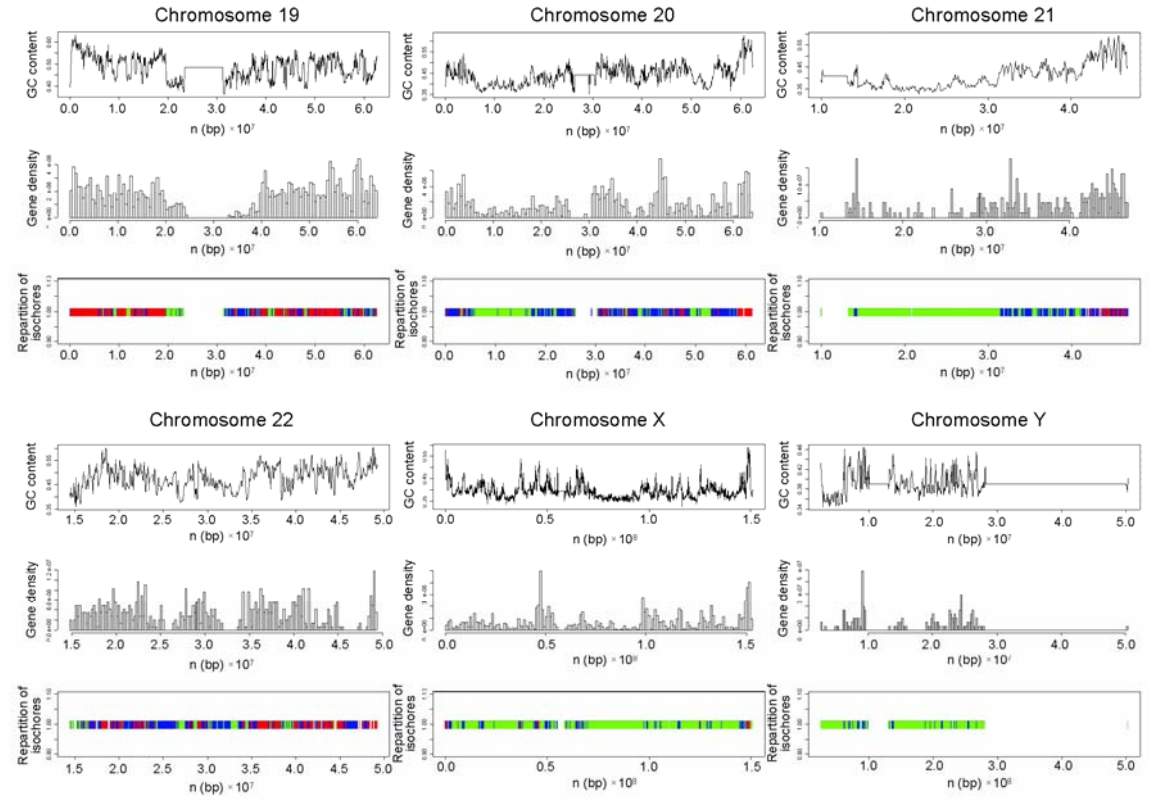

Figure 2d

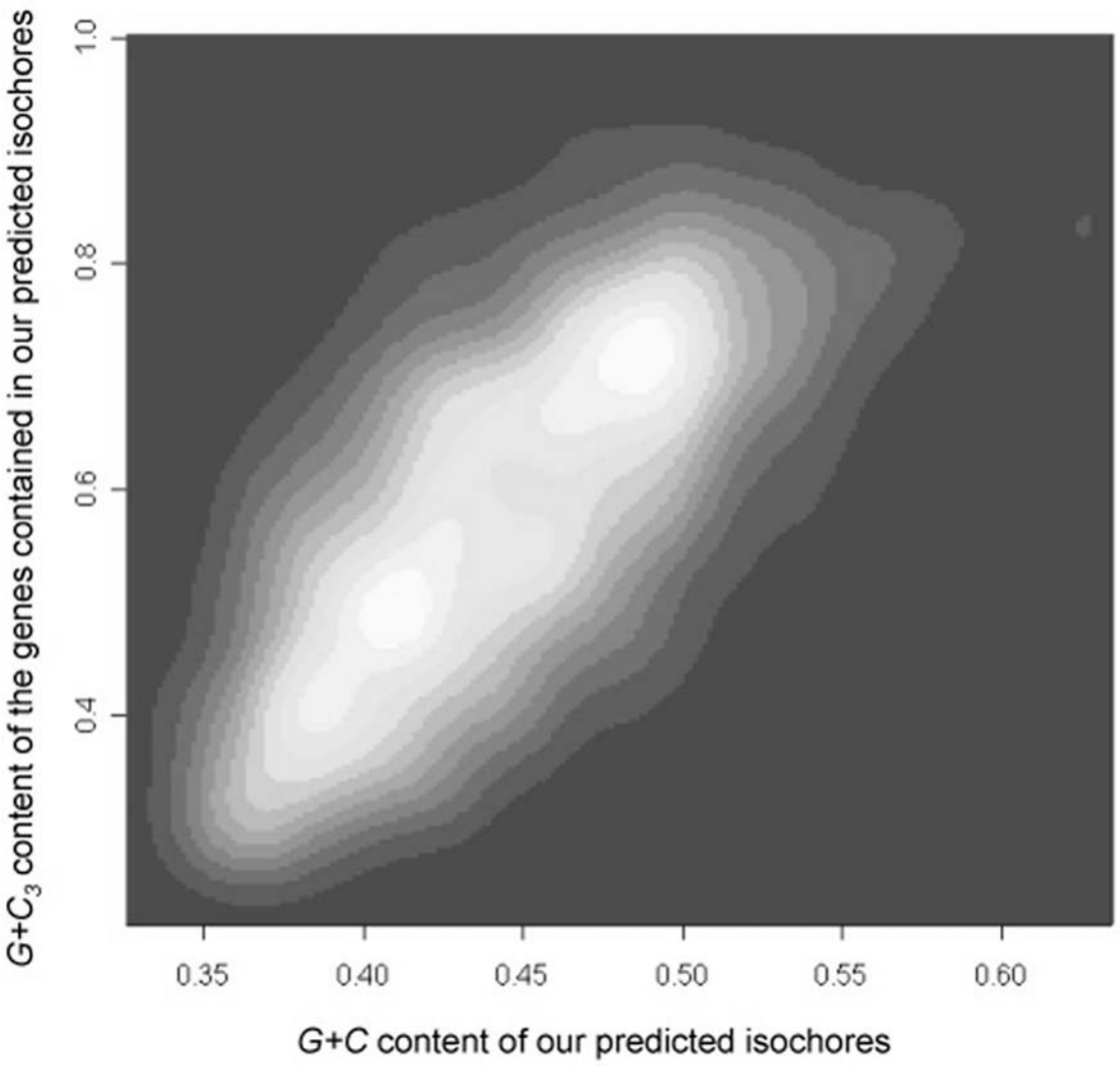
G+C3 content of the genes contained in our predicted isochores
1.0
0.8
0.6
0.4
0.35
0.40
0.45
0.50
0.55
0.60
G+C content of our predicted isochores

Figure 3

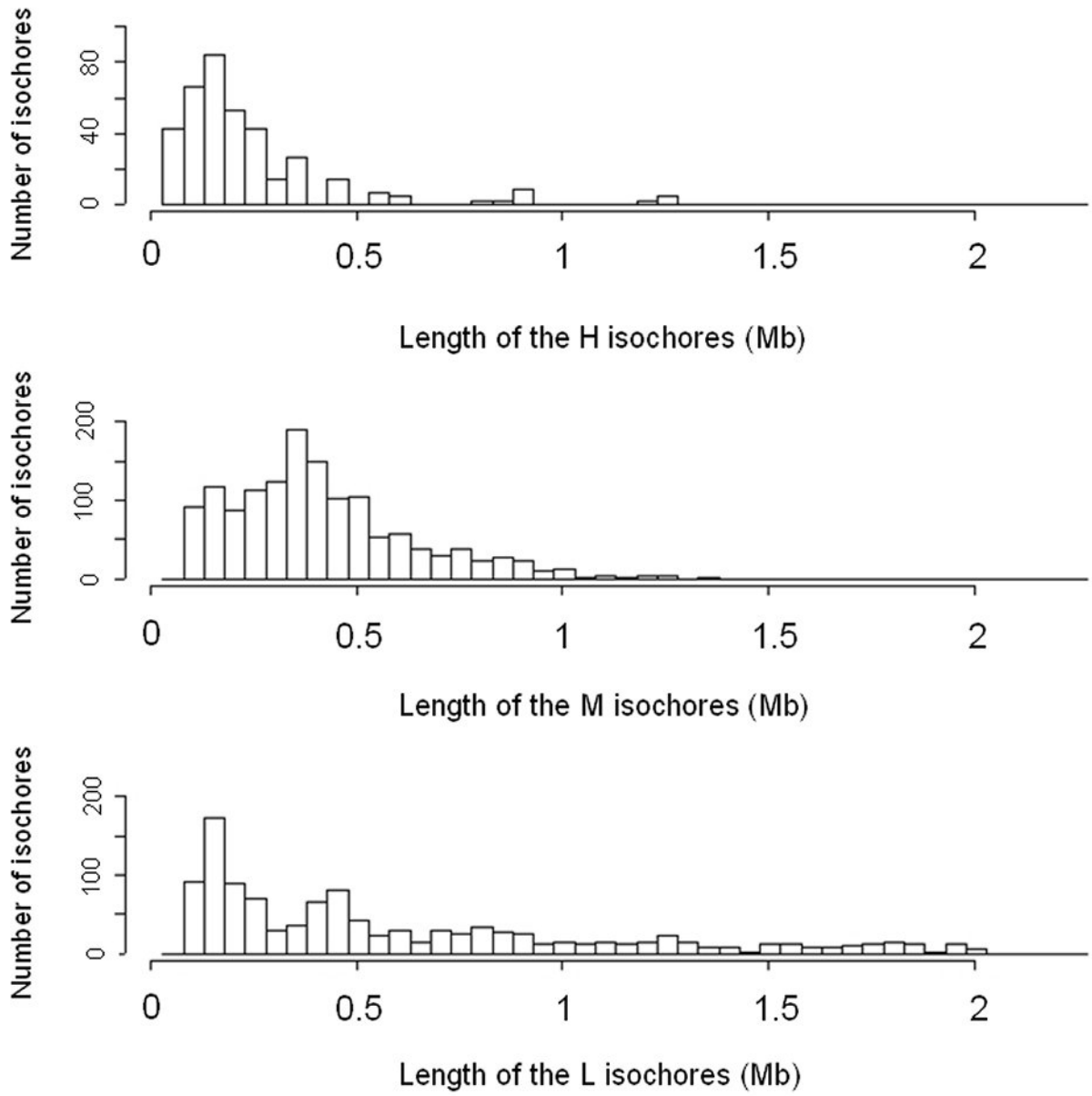
Number of isochores
0
40
80
0
0.5
1
1.5
2
Length of the H isochores (Mb)
Number of isochores
0
100
200
0
0.5
1
1.5
2
Length of the M isochores (Mb)
Number of isochores
0
100
200
0
0.5
1
1.5
2
Length of the L isochores (Mb)

Figure 4

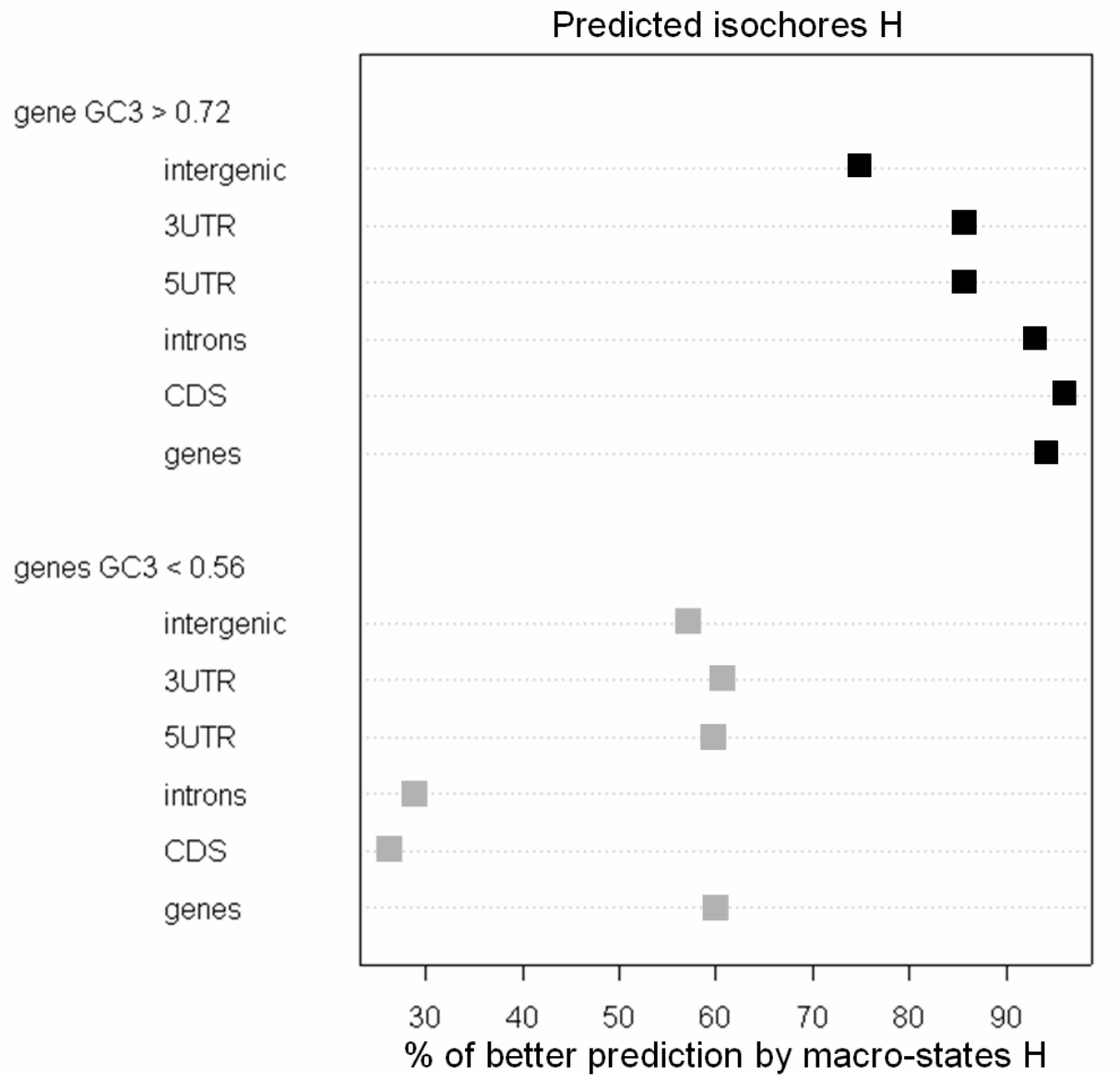
Predicted isochores H
gene GC3 > 0.72
intergenic
3UTR
5UTR
introns
CDS
genes
genes GC3 < 0.56
intergenic
3UTR
5UTR
introns
CDS
genes
30
40
50
60
70
80
90
% of better prediction by macro-states H

Figure 5

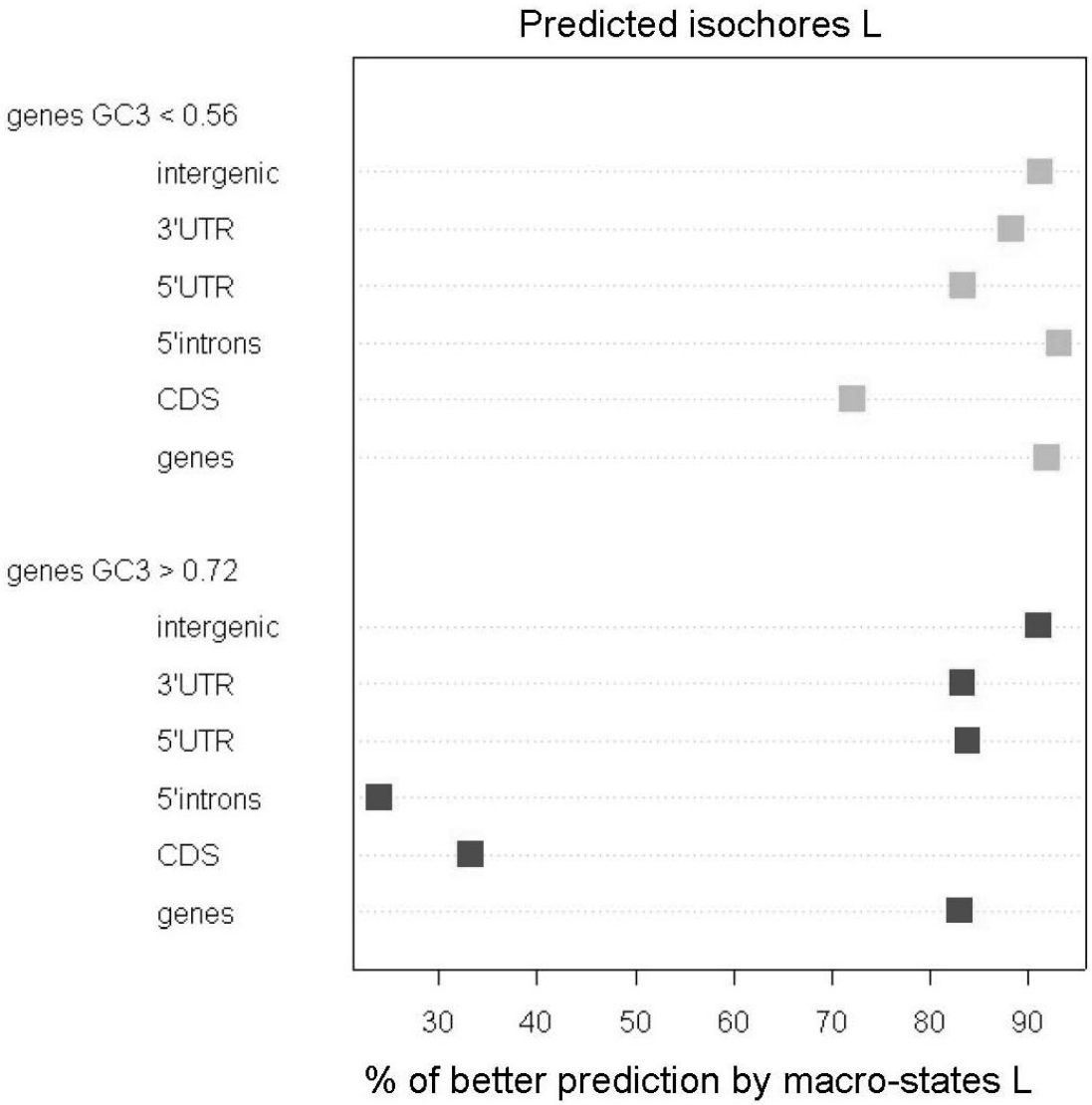
Predicted isochores L
genes GC3 < 0.56
intergenic
3'UTR
5'UTR
5'introns
CDS
genes
genes GC3 > 0.72
intergenic
3'UTR
5'UTR
5'introns
CDS
genes
30
40
50
60
70
80
90
% of better prediction by macro-states L

Figure 6

Table 1. Kruskal-Wallis non-parametric tests performed on the *G+C3* of CDS, and on the *G+C*, length and number of introns

Biological property	Type of isochore			p-value
	H	*M*	*L*	
G+C3 of CDS	80	64	43	2.10^{-21}
G+C of introns	59	51	38	3.10^{-264}
Length of introns	1275	1809	3117	10^{-130}
Number of exons	8.93	10.76	12.31	2.10^{-66}

Table 2. Comparison of the isochore H predictions of genes with a $G+C_3$ of less than 0.56 and 5'UTR.

Isochore prediction of the HMM "genes"	Isochore prediction of the HMM "5'UTR"	% of genes in this configuration
H	H	50%
H	L	10%
L	H	10%
L	L	30%

Table 3. Comparison of the genes with $G+C_3 > 0.72$ and 5'UTR predictions in isochore *L*.

Isochore prediction of the HMM "genes"	Isochore prediction of the HMM "5'UTR"	% of genes in this configuration
H	H	0%
H	L	7%
L	H	16%
L	L	77%