Correspondence

Isochore conservation between MHC regions on human chromosome 6 and mouse chromosome 17

Adam Pavlíček[a,b], Oliver Clay[c], Kamel Jabbari[b], Jan Pačes[a,d], Giorgio Bernardi[b,c,*]

First published online 2 January 2002

The major histocompatibility (MHC) region of mammalian genomes embeds a series of highly polymorphic genes that are expressed on the surface of several cell types, including the T and B lymphocytes that form an integral part of the adaptive immune system. In human, the MHC (or HLA) locus covers roughly 4 Mb of chromosome 6, while in mouse the MHC (or H2) locus is located in a largely syntenic region on chromosome 17. The locus has long been of intrinsic interest in view of its central role in conferring immunity, its polymorphism, and its paralogous loci on chromosomes 1, 9 and 19 in human [1,2]. More recently, it has provided a paradigmatic example of the sharp boundaries that can exist between mammalian isochores, the long regions of fairly homogeneous GC, extending over several hundred kb or even over Mb scales, into which mammalian genomes are organized. The MHC region in human has also provided the first examples of a precise link between isochore boundaries and replication timing switchpoints at the sequence level. The central, GC-poor isochore (Fig. 1, blue and yellow landscape for human) is replicated late, the two GC-rich isochores that flank it (orange and red) are replicated early, and the sharp boundaries of the central isochore both coincide with the switches in replication timing ([2,3] and references therein). This region therefore provides a particularly clear illustration of a genome-wide tendency of GC-rich regions (and the bands in which they are located) to replicate earlier than GC-poor regions [4].

The recent sequencing of the mouse MHC locus offers a first opportunity to compare a contiguously sequenced, nearly syntenic region spanning three clearly delimited isochores in human and mouse. Such a comparison is of particular interest because human and mouse represent the two principal variants, or patterns, of base compositional organization in eutherians: mouse and rat belong to a subset of rodents (coinciding essentially with the myomorphs) that differ from other eutherians in having less dramatic contrasts in GC level among and within their isochores (reviewed in [5]). This phenomenon also affects the compositionally prominent, typically unmethylated CpG islands, which surround the promoters of nearly all housekeeping and many tissue-specific genes: in mouse they are lower, shorter and less abundant than in human.

As can be seen in Fig. 1, the isochore structure of the locus is conserved between human and mouse, not only in the regions showing sequence homology (as estimated by standard similarity criteria), but also in the intervening regions (represented by dull colors and the absence of diagonals in the dot plot). The intervening regions between the bright homology stripes include the H2-K class I subregion in mouse, which is located between the two most proximal homology regions shown in Fig. 1 ($\sim$0.25 Mb from the proximal beginning of the sequence), and which is thought to have been transposed from a distant class I region near the distal end of the sequence. The H2-K insertion locally disrupts the synteny of genes, the otherwise strict partitioning of the classes into different regions and the sequence similarity between human and mouse, but it does not interrupt the homogeneity of the isochore or the compositional synteny of the locus.

The conservation of the number and extent of the isochores is not altogether trivial, since human and mouse span the largest differences in compositional genome organization that exist among eutherians [5]. Such a conservation could, however, be expected from the GC_3 levels of the genes in this locus, from the strong compositional correlation that exists between the GC_3 of genes and the GC levels of the much longer regions of DNA that embed them, and from the correlation between genic GC_3 levels of orthologous genes in human and mouse. Furthermore, it is of interest that the general correlation between the GC levels of corresponding regions in human and mouse loci remains clearly visible (from the similar landscapes of the GC scans) even where local sequence similarity is weak or insignificant.

The GC-rich proximal parts (left for human, upper for mouse) correspond to the extended class II MHC region, the central GC-poor parts in both species correspond to the 'classical' (immunological) class II region, and the distal parts represent the class III region and the beginning of the class I region in human [6]. Although class III and extended class II are both GC-rich, they do not have the same degree of polymorphism: the polymorphism of the extended class II in human, although not as high as that of the classical class II, is closer to it than to the low polymorphism in class III. In line with the observations on polymorphism in human, homology between human and mouse is again typically well conserved in MHC class III, while the level of conservation is lower for the extended and especially for the very GC-poor classical class II MHC, where homology is limited to a few short regions. The MHC locus thus shows on the whole a higher level of evolutionary conservation of the GC-rich, and gene-rich, regions, and less conservation of the long GC-poor segment. Moreover, the lengths of the GC-rich class III and extended class II regions are similar for both human and mouse, while the GC-poor classical class II region is significantly shorter in mouse than in human. This difference, which corresponds to a loss of GC-poor DNA in an ancestor of mouse, is in agreement with a general propensity of more compact vertebrate genomes to contract their DNA preferentially in the GC-poorer, and typically gene-poorer, regions [7].

The general reduction of compositional contrast in mouse is visible also in the MHC region: the difference in GC level between the central and flanking isochores is much less pronounced than in human. The precise localization of isochore boundaries in mouse, and the prediction of corresponding replication timing switchpoints, could thus profit from comparisons with orthologous human sequences. Indeed, com-

PII: S0014-5793(01)03282-3

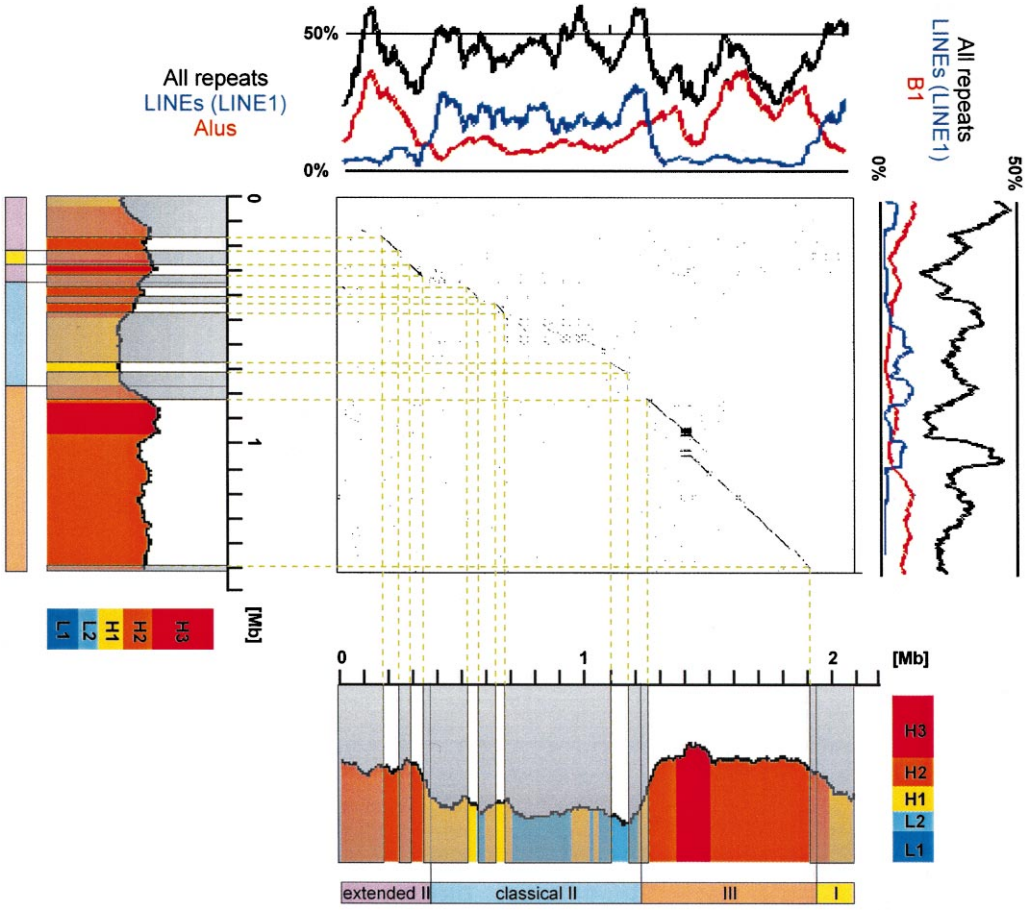

Fig. 1. Color-coded GC scans, sequence similarity dot plot and repetitive element contents for homologous parts of human and mouse MHC loci. In both species the scale increases from the proximal (centromeric, upper/left) beginning of the sequence to its distal (telomeric) end. GC levels were plotted using overlapping 100 kb sliding windows and partitioned into five color-coded intervals, representing the five isochore families in human: their limits, indicated by horizontal dashed lines, are as in [11]. Bright colors correspond to dot plot diagonals and show extended sequence homology, dull colors show regions with no or only local homology. In both species, the locus spans three distinct isochores, characterized in human by red and orange for the two GC-rich isochores, and by blue and yellow for the central GC-poor isochore. The MHC regions are indicated by bars below the GC scans. These regions correspond to the classes of MHC genes [2,6], and are blockwise syntenic between human (HLA) and mouse (H2), with the exception of the H2-K subregion in mouse (short yellow bar near the proximal end). Opposite each GC scan are the corresponding contributions of repetitive elements, which closely follow expectations from genome-wide analyses: e.g. in human the LINE1 elements are more frequent in L1 and L2 DNA, while Alus are most frequent in H2 DNA ([11] and references therein). Details: The sequence of the mouse MHC locus (NT_002588) was extracted from the mouse genome sequencing resources at NCBI (http://www.ncbi.nlm.nih.gov/genome/seq/MmHome.html). The human sequence was selected from chromosome 6 in the GoldenPath assembly of the human genome draft (D. Haussler et al., http://genome.ucsc.edu). For the similarity analysis, both sequences were first masked for repeated elements using the RepeatMasker program (A.F.A. Smit and P. Green), and the masked sequences were then compared by the dotmatcher program in the EMBOSS package (http://www.uk.embnet.org/Software/EMBOSS/index.html), with window size set at 150 bp and threshold set at 50.

bined GC scan/dot plot alignments such as the one shown in Fig. 1 could help reveal isochore boundaries: for example, the proximal boundary of the GC-poor isochore in mouse is compositionally less sharp than the distal boundary in mouse, or than the two boundaries in human. In the case of this central MHC isochore in mouse, all comparisons (homology of sequences, homology of corresponding syntenic genes, corresponding changes in GC level) consistently point to the same location for its proximal boundary, namely within a short region ($\ll$100 kb) delimited by two homology stripes, located approximately 350 kb into the mouse sequence, and indicated by a division in Fig. 1. More detailed analyses by experiments (cf. [3]) and segmentation algorithms (cf. [8]) should help to confirm its precise position.

The repeat content of the human MHC sequence in Fig. 1 is 45.5%, Alus comprising 16.1% and LINE1 elements 15.6% of the sequence. The mouse sequence contains 6% B1 sequences (the most recent mouse SINE class, analogous to Alus in human) and 3.1% LINE1 elements. These numbers are very close to previous genome-wide estimates, at corresponding GC levels, of repeat densities in mouse and human. It is interesting that, in spite of independent insertions of these classes of retroelements, their isochore distribution is similar in both species. Human and mouse LINE1 elements are, as expected, more dense in the GC-poor part, namely classical class II, while SINE repeats, human Alus and mouse B1s are more frequent in the GC-rich parts of the locus (extended class II and class III). Conservation of this pattern is important in the light of recent reports that both Alus and LINEs preferentially integrate into the GC-poor regions of the genome, but that Alus are unstable there and are preferentially excluded from such regions [9,10]. We have proposed that

negative selection on compositionally non-matching elements is the main factor influencing the higher loss of Alu DNA from GC-poor regions, and that there is a strong selection on the conservation of isochore GC level even in GC-poor DNA [10]. The GC-rich bias of the distribution of B1 sequences is another argument in favor of the negative selection hypothesis. The situation in mouse is analogous to that in the human genome: GC-poor LINEs are more dense in the GC-poor part, but GC-rich B1 elements (about 60% GC for the consensus) are more frequent in the GC-rich part. From other repeat classes we should mention long terminal repeat (LTR) elements (endogenous retroviruses and non-autonomous LTR retrotransposons), which contribute 11.5% of the DNA in the mouse MHC sequence but only 7.8% in the human sequence.

The functional meaning of the strict regional partitioning of MHC gene classes into different isochores, and of the conservation of this partitioning despite high recombinogenicity during the evolution of the locus, remains to be elucidated further. The observations that can be made on the MHC locus speak, however, for a strong selection that maintains the precise isochore structure of this region.

References

[1] Endo, T., Imanishi, T., Gojobori, T. and Inoko, H. (1997) Gene 205, 19–27.
[2] The MHC sequencing consortium (1999) Nature 401, 921–923.
[3] Tenzen, T., Yamagata, T., Fukagawa, T., Sugaya, K., Ando, A., Inoko, H., Gojobori, T., Fujiyama, A., Okumura, K. and Ikemura, T. (1997) Mol. Cell. Biol. 17, 4043–4050.
[4] Federico, C., Andreozzi, L., Saccone, S. and Bernardi, G. (2000) Chromosome Res. 8, 737–746.
[5] Bernardi, G. (2000) Gene 259, 31–43.
[6] Stephens, R., Horton, R., Humphray, S., Rowen, L., Trowsdale, J. and Beck, S. (1999) J. Mol. Biol. 291, 789–799.
[7] Bernardi, G. (1995) Annu. Rev. Genet. 29, 445–476.
[8] Oliver, J.L., Bernaola-Galván, P., Carpena, P. and Román-Roldán, R. (2001) Gene 276, 47–56.
[9] IHGSC (International Human Genome Sequencing Consortium) (2001) Nature 409, 860–921.
[10] Pavlíček, A., Jabbari, K., Pačes, J., Pačes, V., Hejnar, J. and Bernardi, G. (2001) Gene 276, 39–45.
[11] Bernardi, G. (2001) Gene 276, 3–13.

*Corresponding author. Fax: (39)-081-245 5807.
E-mail address: bernardi@alpha.szn.it (G. Bernardi).

[a] *Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Flemingovo 2, CZ-16637 Prague, Czech Republic*
[b] *Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 place Jussieu, 75005 Paris, France*
[c] *Laboratorio di Evoluzione Molecolare, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy*
[d] *Center for Integrated Genomics, Flemingovo 2, CZ-16637 Prague, Czech Republic*

PII: S0014-5793(01)03282-3