J Mol Evol (2004) 59:72–79
DOI: 10.1007/s00239-004-2605-z

JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York, LLC 2004

Clustering of Genes Coding for DNA Binding Proteins in a Region of Atypical Evolution of the Human Genome

Jose Castresana,[1] Roderic Guigó,[1,2] M. Mar Albà[2]

[1] Centre de Regulació Genòmica (CRG), Programme of Bioinformatics and Genomics, Passeig Marítim 37-49, 08003 Barcelona, Spain
[2] Grup de Recerca en Informàtica Biomèdica, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Passeig Marítim 37-49, 08003 Barcelona, Spain

Received: 14 August 2003 / Accepted: 3 February 2004

Abstract. Comparison of the human and mouse genomes has revealed that significant variations in evolutionary rates exist among genomic regions and that a large part of this variation is interchromosomal. We confirm in this work, using a large collection of introns, that human chromosome 19 is the one that shows the highest divergence with respect to mouse. To search for other differences among chromosomes, we examine the distribution of gene functions in human and mouse chromosomes using the Gene Ontology definitions. We found by correspondence analysis that among the strongest clusterings of gene functions in human chromosomes is a group of genes coding for DNA binding proteins in chromosome 19. Interestingly, chromosome 19 also has a very high GC content, a feature that has been proposed to promote an opening of the chromatin, thereby facilitating binding of proteins to the DNA helix. In the mouse genome, however, a similar aggregation of genes coding for DNA binding proteins and high GC content cannot be found. This suggests that the distribution of genes coding for DNA binding proteins and the variations of the chromatin accessibility to these proteins are different in the human and mouse genomes. It is likely that the overall high synonymous and intron rates in chromosome 19 are a by-product of the high GC content of this chromosome.

Key words: Chromosome 19 — DNA binding proteins — Evolutionary rates — Introns — Isochores — Mammalian genome — Transcription factors

Introduction

Comparison of mammalian genes has shown that the synonymous rate at which genes evolve is different for different genes (Casane et al. 1997; Matassi et al. 1999; Wolfe et al. 1989). The analysis of large numbers of mouse and human orthologous genes confirmed that this variation occurs at the genomic level, that is, some genomic regions show neutral evolutionary rates higher than others (Castresana 2002b; Hardison et al. 2003; Lercher et al. 2001; Waterston et al. 2002). Similarly, a pattern of regional variation of synonymous rates was confirmed in the analysis of a set of human and chimpanzee genomic regions (Ebersberger et al. 2002). A major part of this variation is interchromosomal (Castresana 2002b; Lercher et al. 2001), and, in particular, the use of maximum likelihood distances shows that genes situated in chromosome 19 have synonymous divergences that are significantly faster than those of genes

Correspondence to: Jose Castresana, Department of Physiology and Molecular Biodiversity, Institut de Biologia Molecular de Barcelona, CSIC, Jordi Girona 18, 08034 Barcelona, Spain; *email:* jcvagr@ibmb.csic.es

in any other chromosome (Castresana 2002b; Waterston et al. 2002).

The large-scale differences in evolutionary rates could have a neutral explanation, such as differences in recombination rates (Hurst and Eyre-Walker 2000; Li et al. 2002), but could also reflect important functional differences in the organization of the human and mouse genomes. In order to try to explain this phenomenon we have analyzed the distribution of gene functions in human and mouse chromosomes. Several clusters of genes have been detected previously in the human and mouse genomes but most of them are due to recent multiple tandem duplications and are located in very specific regions of the genome (Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002). In addition, clustering of housekeeping genes (expressed in most tissues) has also been detected in the human genome (Lercher et al. 2002). However, no systematic attempt to detect the most significant clusters of genes with related molecular functions has been performed so far in the mammalian genomes. In this work, we used the Gene Ontology (GO) molecular functions (The Gene Ontology Consortium 2001) to analyze the distribution of gene functions in human and mouse chromosomes. We found that the most important clusters at the chromosome level in the human genome are a cluster of olfactory receptor genes in chromosome 11 and, most interestingly, a cluster of genes coding for DNA binding proteins in chromosome 19. However, no strong clustering of DNA binding protein coding genes is present in the mouse genome. It is known that, among human chromosomes, chromosome 19 is the one with the highest GC content, whereas the syntenic regions of the mouse genome have a more moderate GC content (Castresana 2002b; Saccone et al. 2002). The clustering of genes coding for DNA binding proteins and extremely high GC content seem to indicate the existence of a possible function at the level of gene expression and regulation in this human chromosome that is not present in the mouse genome.

Methods

Intron Sequences. Human and mouse Reference Sequences (Pruitt and Maglott 2001) were mapped into the human and mouse genome sequences, versions NCBI 28 and MGSC3, respectively. The mapping was obtained from the University of California Santa Cruz (UCSC) Genome Browser database (Karolchik et al. 2003). The human and mouse orthologous genes were obtained from NCBI's Homologene database, which listed 3308 orthologous pairs from the RefSeq sequences. Pairs were discarded when the human and/or mouse RefSeq gene had multiple alternative mRNA sequences or when the human and mouse orthologous genes had a different number of exons. After excluding intronless genes, 1165 orthologous gene pairs with a clear one-to-one correspondence of introns remained. Repeats were eliminated from the introns with RepeatMasker (A.F. Smit and P. Green; http://repeatmasker.genome.washington.edu). Only intron pairs where both sequences had more than 150 positions and fewer than 18,000 positions were used in order to avoid alignments with poor information or excessively difficult. This left 7180 intron pairs with known chromosome position. Alignments were done with ClustalW (Thompson et al. 1994), which uses a dynamic programming algorithm for aligning pairs of sequences (Needleman and Wunsch 1970), and Gblocks 0.91 with default options was applied to extract the conserved parts of the intron alignments (Castresana 2000; Castresana 2002a). The distance in substitutions per site in each concatenated intron alignment was estimated by maximum likelihood with PAUP (Swofford 1998) using the HKY model of evolution (Hasegawa et al. 1985). This model takes differences in transition/transversion ratio and nucleotide composition (i.e., GC content) into account.

Retrieval of Gene Function Information and Statistical Analysis. Mouse and human gene descriptions, chromosome positions, InterPro domains (Mulder et al. 2003), and the gene ontology classification of the molecular functions assigned to the human genes (The Gene Ontology Consortium 2001) were retrieved from version 18 of the ENSEMBL database through the EnsMart facility (Clamp et al. 2003). Genes in unassigned chromosomes or in the Y chromosome, where there are too few genes, were not considered. For correspondence analysis (Greenacre 1984) we used human genes with an assigned GO function. Since mouse genes do not have a GO definition in the ENSEMBL database we transferred the GO definition of the human sequences to their mouse homologues as defined in ENSEMBL. These homologous sequences are defined, briefly, as pairs of reciprocal BLAST hits as well as pairs that have high similarity and conserved gene order (Clamp et al. 2003). According to this, most homologous pairs are likely to be orthologous. In addition, coming from closely related species it is very likely that all these pairs have the same GO function, which is in general very broadly defined. Some GO functions are very poorly represented to analyze their distribution in different chromosomes, so that only GO terms assigned to more than 300 genes were used. Correspondence analysis was performed with the JMP package (SAS Institute, Cary, NC).

Phylogenetic Analysis of Zinc-Finger Proteins. Amino acid sequences of all KRAB-ZFPs (KRAB-associated zinc-finger proteins) of the human and mouse genomes were retrieved from ENSEMBL. Since these sequences are quite divergent and very variable in length (Looman et al. 2002), only sequences between 400 and 800 amino acids were analyzed. After initial alignments and neighbor-joining trees performed with ClustalW (Thompson et al. 1994), the most divergent sequences (most probably not belonging to this family) were also eliminated. The final set of 182 human and 110 mouse sequences was realigned, and the non-phylogenetically informative positions were removed with Gblocks (Castresana 2000) using low-stringency conditions (Minimum Length Of A Block = 5 and allowing positions with gaps in half the number of sequences). The high divergence in these alignments made the use of these low-stringency conditions necessary in order to get enough number of positions. The final alignment contained 105 positions. Distance trees were constructed using PROTDIST with the JTT model (Jones et al. 1992) and NEIGHBOR of the PHYLIP package (Felsenstein 1993). In order to estimate the average divergence between orthologous human and mouse genes, a set of 12270 1:1 orthologous genes was retrieved from ENSEMBL (all genes with more than one orthologue in the other species were eliminated) and translated to amino acids. These pairs were aligned with ClustalW and filtered with Gblocks using the same conditions applied to the KRAB-ZFP genes. Genetic distances between human and mouse genes were equally calculated with PROTDIST using the JTT model.

Results

Interchromosomal Mutational Rate Variations Analyzed with Intron Sequences. It has been shown that genes situated in human chromosome 19 and their mouse orthologues have a very high synonymous divergence (Castresana 2002b; Hardison et al. 2003; Lercher et al. 2001; Waterston et al. 2002). However, analysis of the ancestral repeats of the mouse genome did not show the same pattern of divergence (see Fig. 29 in ref. Waterston et al. 2002), probably due to the difficulty in defining orthologous ancestral elements. To try to clarify this issue, we have analyzed here the levels of divergence using another genomic element, introns, in different chromosomes. In a previous study using a small set of intron sequences, we showed that intron genetic distances are correlated with synonymous distances measured from exons (Castresana 2002a). We confirm now, using a set of 7180 human and mouse orthologous introns derived from 1165 curated genes, that intron sequences also show the highest average genetic distance in chromosome 19 (Fig. 1). Specifically, the average intron distance is 0.771 substitution/position in chromosome 19 and 0.657 substitution/position in the rest of the genome. Very similar values were obtained when using the Tamura–Nei model of evolution instead of the HKY model in the maximum-likelihood estimation (not shown). Thus both introns and exons support the high divergence in human chromosome 19 (Castresana 2002b). However, unlike exons, the level of divergence in introns in chromosome 19 is not significantly different from chromosomes 16, 21, and 22, probably due to the more difficult alignments of introns (Castresana 2002a), which may saturate the highest distances.

Spatial Aggregation of Genes with Similar Molecular Functions in Human and Mouse Chromosomes. We retrieved from the ENSEMBL database (Clamp et al. 2003) all human genes with an assigned Gene Ontology (GO) molecular function and known chromosome position. In total, we used 6942 genes in which 9372 GO functions of wide distribution (present in >300 genes) had been mapped. Although different clusters may be detected in smaller genomic windows, we were interested in identifying clusters at the level of whole chromosomes in order to analyze any possible correlation with the interchromosomal differences found in evolutionary rates. Thus we calculated the number of GO functions per chromosome. Correspondence analysis (Greenacre 1984) of this contingency table allowed us to visualize the GO terms with the most biased distributions among chromosomes. A biased GO function and the chromosomes where this function is most overrepresented appear close in the plane delineated by the two principal axes (Greenacre 1984). Figure 2A shows that the GO function that produces the strongest chromosomal clustering is the "olfactory receptor activity." Of 405 genes coding for olfactory receptors which are annotated by GO, 190 are located in human chromosome 11. This cluster is due in part to recent multiple gene duplications (Glusman et al. 2001). Most interestingly, the second largest clustering in the human genome is due to the high proportion of genes coding for "nucleic acid binding activity" proteins in chromosome 19. One hundred seven of 479 genes with such GO function are present in chromosome 19. This cluster contains a large number of genes coding for proteins with zinc fingers, many of which belong to a family of transcription factors known as KRAB-ZFP (KRAB-associated zinc-finger proteins) (Dehal et al. 2001; Eichler et al. 1998; Shannon et al. 2003). Notably, although the mouse genome also contains a strong cluster of olfactory receptors (distributed in at least chromosomes 2, 7, and 9) as well as other different clusters, no cluster of "nucleic acid binding activity" is recognized by correspondence analysis (Fig. 2B).

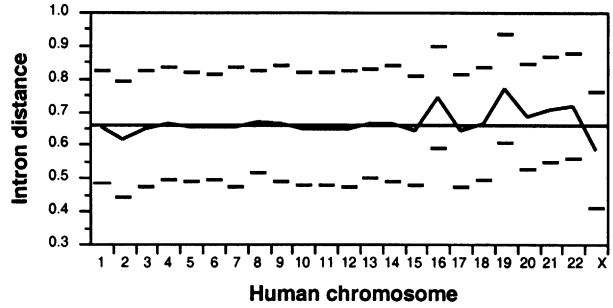

Fig. 1. Genetic distances in substitutions position measured in 7180 mouse and human intron alignments averaged in individual human chromosomes. The average and the standard deviation are shown. A line connects the mean values of each chromosome. The straight horizontal line represents the mean of all chromosomes.

"Nucleic acid binding activity" is a very generic function within the GO hierarchy, but as mentioned above, a large number of genes with this GO function in chromosome 19 contain zinc-finger domains (according to the InterPro descriptions), and therefore they code mostly for DNA binding proteins. In addition, two other GO functions that come under the heading "nucleic acid binding activity" in the GO molecular function hierarchy are more abundant than the average in chromosome 19, even if not so highly biased as to be detected by correspondence analysis. These are "DNA binding activity" and "transcription factor activity." On the other hand, "RNA binding activity" is underrepresented. Therefore the analysis of GO functions indicate that proteins that bind to DNA are located in high quantities in chromosome 19. To better appreciate the density of DNA binding proteins in different chromosomes, we have plotted the density of all genes with GO "nucleic acid binding

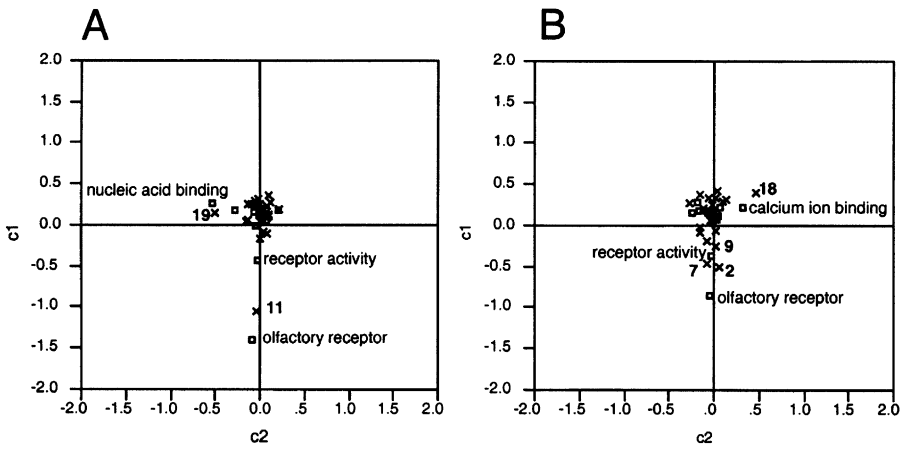

Fig. 2. Correspondence analysis of 9372 assigned GO molecular functions in human chromosomes (**A**) and of 10,431 assigned GO molecular functions in mouse chromosomes (**B**). The first two axes (c1 and c2) represent 77 and 76% of the total inertia for the human and mouse genomes, respectively. Only names of outliers are given.

activity," DNA binding activity," or "transcription factor activity" along their chromosome positions (Fig. 3). Fig. 3A shows that the density of this type of proteins is very high all along human chromosome 19 with the exception of the centromeric region, mostly unsequenced. In fact, the density of genes in chromosome 19 is approximately double than in the rest of chromosomes (Lander et al. 2001, Venter et al. 2001) but the density of genes coding for DNA binding proteins in this chromosome (with 262 genes out of 2104 in all chromosomes) is seven times the density in the rest of the genome. A binomial test corroborates that these proteins are more abundant than expected in chromosome 19. In addition, this test is also significant for the group of the three GO terms in chromosome 6, where there is a spatial cluster of 48 histones in a short region between 26 and 28 Mb. In the mouse genome there is no such large cluster of genes coding for DNA binding proteins (Fig. 3B). Only chromosome 13, which contains the same cluster of histones, has a significantly higher than expected proportion of DNA binding proteins. Chromosome 7 contains the largest syntenic region of human chromosome 19 (Dehal et al. 2001; Waterston et al. 2002), but the proportion of these proteins is not significantly higher than expected.

Phylogenomic Analysis of KRAB-Associated Zinc Finger Proteins. The family of KRAB-ZFP transcription factors constitutes an important fraction of this cluster in chromosome 19 (128 annotated KRAB-ZFPs of 262 DNA binding proteins). These genes are known to have suffered extensive gene duplications and losses in both the human and the mouse lineages (Dehal et al. 2001; Shannon et al. 2003). However, it is important to determine whether the majority of KRAB-ZFP genes have duplicated recently or whether this cluster predates the human-mouse split. We performed a phylogenomic analysis of 182 human and 110 mouse KRAB-ZFP sequences. In this large tree, KRAB-ZFP genes situated in human chromosome 19 (105 in total, after eliminating highly divergent genes, probably belonging to a different family; see Methods) tend to cluster in a few groups. Figure 4 represents nine clades of this large tree, six of them (A–F) containing most of the KRAB-ZFP genes of chromosome 19. Most probably, these genes have been present together in the same region of the genome since the occurrence of the multiple duplications that generated them, which may suggest that they are the product of recent gene duplications exclusive of the human lineage. However, the analysis of the branches separating KRAB-ZFP genes seems to give a different answer. The average divergence of two orthologous human and mouse genes (estimated from 12,270 1:1 orthologues with the same methods used for the analysis of KRAB-ZFP genes) is 0.205 substitution/position. This means that the average per lineage is 0.102 substitution/position. Comparing this distance with divergences of genes in chromosome 19 (see subtrees in Fig. 4 but there are still deeper branchings in the whole tree that joins these clades), it becomes clear that many of the duplications of these KRAB-ZFP genes occurred before the typical human–mouse orthologous separation.

To avoid the possibility that the rates estimated from all proteins are different from typical KRAB-ZFP genes, we have also attempted to do an internal calibration of the KRAB-ZFP tree. We have obtained from this tree all terminal pairs that include a human and a mouse sequence. There are in total 34 pairs (15 of them can be appreciated in Fig. 4). Many of these pairs—most likely those showing the shortest distances—probably represent real orthologues, whereas some other pairs may be paralogues if alternate orthologues are not in the tree. The average distance between the terminal pairs is 0.252 substitution/position (0.126 substitution/position/lineage). However, a better approximation to the typical divergence of orthologous KRAB-ZFP sequences is given by the modal class of the distribution of distances, which is situated between 0.15 and 0.2 substitution/position (the distribution is right-skewed

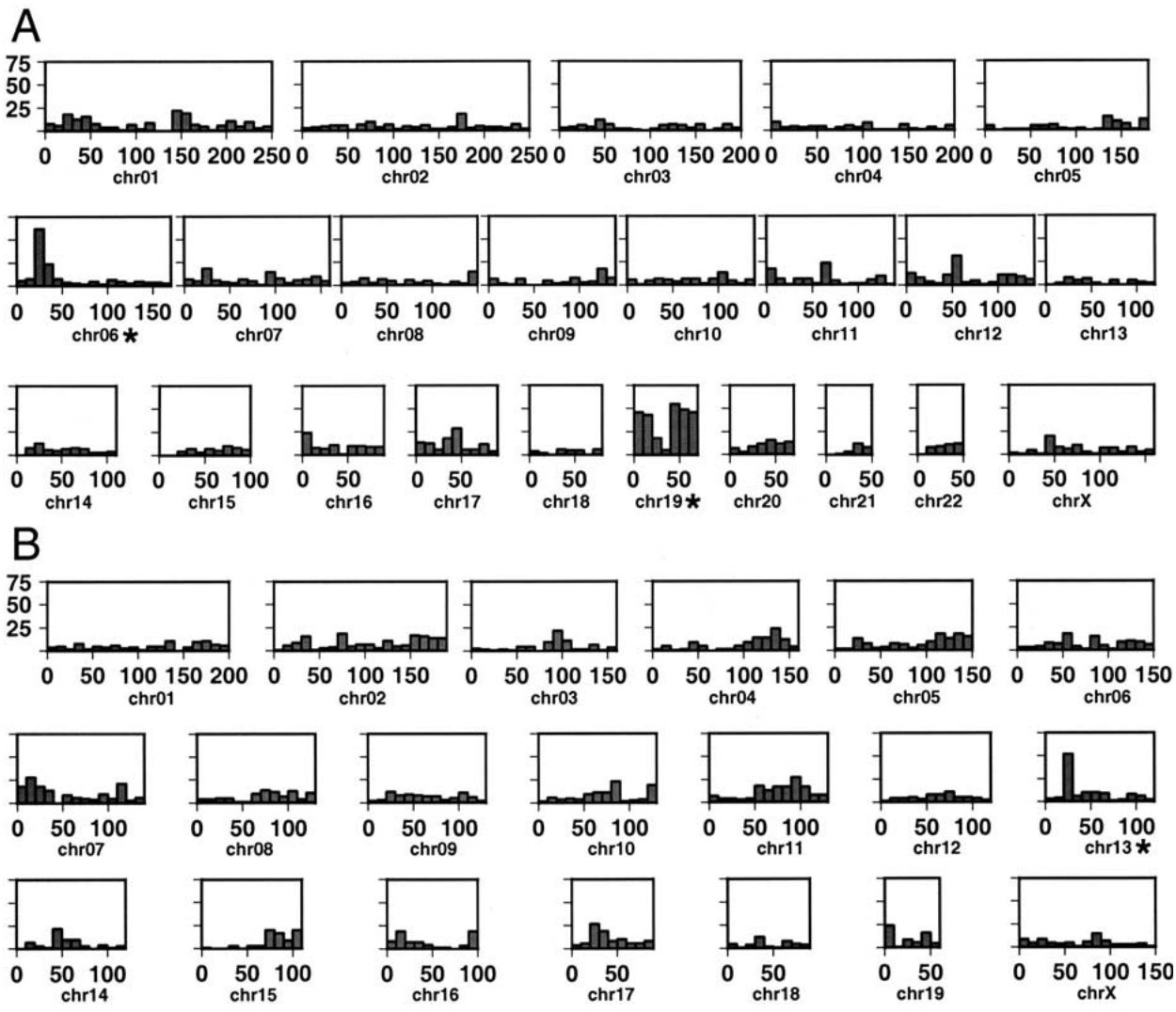

Fig. 3. Density of DNA binding proteins (including the GO terms: nucleic acid binding activity, DNA binding activity, and transcription factor activity) in human (**A**) and mouse (**B**) chromosomes. The *X* axis represents chromosome length in Mb and the *Y* axis number of genes. An asterisk indicates that the proportion of this sort of protein in the corresponding chromosome is higher than expected according to a one-tail binomial test ($p < 0.0001$).

due to the likely presence of several paralogues). This means that the maximum of divergences occurred at approximately 0.175 substitution/position (midpoint of the modal class) or 0.087 substitution/position per lineage, a value that probably approximates the typical separation between two orthologous KRAB-ZFP genes. This value is smaller than the one obtained from orthologues derived from all types of proteins. The examination of gene duplications in the subtrees in Fig. 4 in relation to this distance indicates more clearly that a large number of the gene duplications happened before the human–mouse split. Thus some clades of only human genes are too divergent compared to this distance to have originated recently. In conclusion, at least part of the cluster of KRAB-ZFP genes of chromosome 19, rather than having originated in the human lineage, would have been present before the human–mouse split, in the common ancestor of human and mouse, and some of the genes would have been later lost in the mouse lineage. This does not exclude the existence of other more recent, lineage-specific, gene duplications and losses in specific clusters (Dehal et al. 2001; Shannon et al. 2003).

Discussion

We have shown with statistical confidence the existence of a strong aggregation of DNA binding protein coding genes in chromosome 19 and we have also shown that this aggregation of genes is not present in the mouse genome (Figs. 2 and 3). In addition, a phylogenetic analysis of a major part of this cluster (mainly formed by KRAB-ZFP genes) indicates that part of this cluster probably originated previously to the separation of the human and mouse lineages (Fig. 4), and therefore it was disrupted in the mouse lineage due to gene loss and rearrangements. This

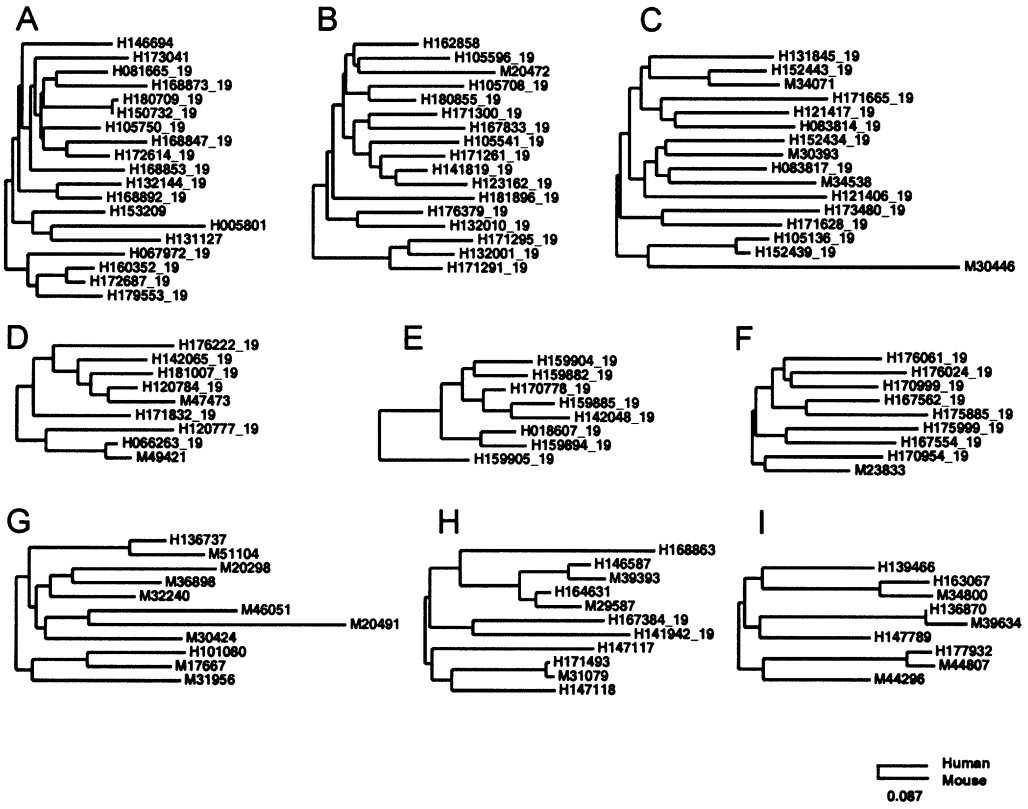

Fig. 4. Nine representative clades extracted from a phylogenetic tree of 292 KRAB-ZFP genes of the human and mouse genomes. Six of them (**A--F**) contain a large proportion of KRAB-ZFP genes in human chromosome 19. Human sequences start with H, and mouse sequences with M. Those sequences present in human chromosome 19 are also indicated. The scale-tree at the lower right indicates the typical distance (midpoint of the modal class) between putative KRAB-ZFP orthologues extracted from the whole tree (=0.175 substitution/position or 0.087 substitution/position/lineage).

agrees with the overall conservation of the synteny of human chromosome 19 in different orders of mammals (Dehal et al. 2001; Tanabe et al. 2002) and even in birds (Smith et al. 2002a), but not in rodents (Dehal et al. 2001). The maintenance of an aggregation of genes with a similar function for a long time might indicate that this cluster is favored by selection. Accordingly, some type of selective force may have favored the maintenance of the cluster of genes coding for DNA binding proteins in human chromosome 19, whereas this force is not acting in the rodent genome.

Another exceptional feature of chromosome 19 is that it has the highest GC content of all human chromosomes (75% in the fourfold degenerate positions of exons versus an average of 55%) (Castresana 2002b; Saccone et al. 2002). Isochores with a high GC content have recently been shown to increase DNA bendability and consequently ease the transition from nucleosome-wrapped to extended DNA (Anselmi et al. 2000; Vinogradov 2003). This in turn would favor the interaction of DNA binding proteins, such as transcription factors, with the DNA helix and therefore an active transcription of genes. In fact, it is known that chromosome 19 has the highest levels of transcription of the genome (Caron et al. 2001). In rodents there is not such a large region with a high GC content. The human-like isochoric organization, highly fluctuant between very low and very high GC content, seems to be ancestral in mammals (Bernardi 2000a,b; Galtier and Mouchiroud 1998). Thus, it is likely that a reduction of GC content occurred in this part of the mouse genome rather than an increase in GC in certain regions of the human genome. In this respect, it is also possible to think that the reduction of GC content occurred concomitantly with the disruption of the cluster of DNA binding proteins in the mouse genome, since both the gene cluster and the extremely high GC content affect the same genomic regions. Meanwhile, the primate lineage has maintained both the cluster of genes coding for DNA binding proteins and the high accessibility to this type of proteins in chromosome 19 mediated by the high GC. This raises the question whether active transcription concentrated in certain large genomic regions is more important for primates than for rodents, but more data would be necessary to study this possibility.

Finally, another intriguing feature of human chromosome 19 is the high rate of synonymous evolution of its genes (Castresana 2002b; Hardison et al.

2003; Lercher et al. 2001; Waterston et al. 2002), corroborated in this work by the analysis of introns (Fig. 1). In a previous work we favored the idea that the accumulation of synonymous substitution occurred mainly in the mouse lineage as a consequence of the alteration of its isochoric organization (Castresana 2002b). However, a recent comparison of 7645 human and chimpanzee genes (supplementary informationin in Clark et al. 2003) indicates that genes situated in chromosome 19 also show the highest synonymous divergence of these two genomes. Thus it is likely that the primate lineage experienced many of the substitutions measured between mouse and human. Our data indicate that the synonymous rate of genes coding for DNA binding proteins in chromosome 19 is not higher than that of other genes in this chromosome (not shown), and therefore the cause of the high rates cannot be attributed to them. Another possibility is that, if the isochoric organization has remained mainly unchanged in the primate lineage (Bernardi 2000a; Bernardi 2000b; Galtier and Mouchiroud 1998), the extremely high GC content of human chromosome 19 (rather than its alteration) is the cause of the high rates of mutation in synonymous and intron positions observed in genes of this chromosome. This in turn may be due partly to the base composition not being at equilibrium and partly to enhanced mutagenesis in high GC regions (Smith et al. 2002b). Although the evolutionary pathway that led to the peculiar features of human chromosome 19 remains speculative, the analysis of the human and mouse genomes is showing important differences in the organization of both genomes. Other mammalian genomes will be necessary to study how these differences arose.

Acknowledgments. J.C. and M.A. are recipients of a Ramón y Cajal contract of the Spanish Ministerio de Ciencia y Tecnología (MCYT) and are supported by grant numbers BIO2002-04426-C02-02 and BIO2002-04426-C02-01, respectively, from the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I) of the MCYT, cofinanced with FEDER funds. We thank Alison Shaw, Xavier Estivill, Arcadi Navarro, and Martin Lercher for carefully reading the manuscript and making useful suggestions.

References

Anselmi C, Bocchinfuso G, De Santis P, Savino M, Scipioni A (2000) A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability. Biophys J 79:601–613

Bernardi G (2000a) The compositional evolution of vertebrate genomes. Gene 259:31–43

Bernardi G (2000b) Isochores and the evolutionary genomics of vertebrates. Gene 241:3–17

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, Versteeg R (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291:1289–1292

Casane D, Boissinot S, Chang BH, Shimmin LC, Li W (1997) Mutation pattern variation among regions of the primate genome. J Mol Evol 45:216–226

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 7:540–552

Castresana J (2002a) Estimation of genetic distances from human and mouse introns. Genome Biol 3:research0028.0021-0028.0027

Castresana J (2002b) Genes on human chromosome 19 show extreme divergence from the mouse orthologues and a high GC content. Nucleic Acids Res 30:1751–1756

Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Hubbard T, Kasprzyk A, Keefe D, Lehvaslaiho H, Iyer V, Melsopp C, Mongin E, Pettett R, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Birney E (2003) Ensembl 2002: Accommodating comparative genomics. Nucleic Acids Res 31:38–42

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963

Dehal P, Predki P, Olsen AS, Kobayashi A, Folta P, Lucas S, Land M, Terry A, Ecale Zhou CL, Rash S, Zhang Q, Gordon L, Kim J, Elkin C, Pollard MJ, Richardson P, Rokhsar D, Uberbacher E, Hawkins T, Branscomb E, Stubbs L (2001) Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. Science 293:104–111

Ebersberger I, Metzler D, Schwarz C, Pääbo S (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet 70:1490–1497

Eichler EE, Hoffman SM, Adamson AA, Gordon LA, McCready P, Lamerdin JE, Mohrenweiser HW (1998) Complex β-satellite repeat structures and the expansion of the zinc finger gene cluster in 19p12. Genome Res 8:791–808

Felsenstein J (1993) PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle

Galtier N, Mouchiroud D (1998) Isochore evolution in mammals: A human-like ancestral structure. Genetics 150:1577–1584

Glusman G, Yanai I, Rubin I, Lancet D (2001) The complete human olfactory subgenome. Genome Res 11:685–702

Greenacre MJ (1984) Theory and applications of correspondence analysis. Academic Press, London

Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, Schwartz S, Furey TS, Whelan S, Goldman N, Smit A, Miller W, Chiaromonte F, Haussler D (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res 13:13–26

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Hurst LD, Eyre-Walker A (2000) Evolutionary genomics: Reading the bands. Bioessays 22:105–107

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ (2003) The UCSC Genome Browser Database. Nucleic Acids Res. 31:51–54

Lander ES, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lercher MJ, Williams EJ, Hurst LD (2001) Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. Mol Biol Evol 18:2032–2039

Lercher MJ, Urrutia AO, Hurst LD (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet 31:180–183

Li WH, Yi S, Makova K (2002) Male-driven evolution. Curr Opin Genet Dev 12:650–656

Looman C, Abrink M, Mark C, Hellman L (2002) KRAB zinc finger proteins: An analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. Mol Biol Evol 19:2118–2130

Matassi G, Sharp PM, Gautier C (1999) Chromosomal location effects on gene sequence evolution in mammals. Curr Biol 9:786–791

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31:315–318

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Pruitt KD, Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 29:137–140

Saccone S, Federico C, Bernardi G (2002) Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. Gene 300:169–178

Shannon M, Hamilton AT, Gordon L, Branscomb E, Stubbs L (2003) Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. Genome Res 13:1097–1110

Smith J, Paton IR, Murray F, Crooijmans RP, Groenen MA, Burt DW (2002a) Comparative mapping of human chromosome 19 with the chicken shows conserved synteny and gives an insight into chromosomal evolution. Mamm Genome 13:310–315

Smith NG, Webster MT, Ellegren H (2002b) Deterministic mutation rate variation in the human genome. Genome Res 12:1350–1356

Swofford DL (1998) PAUP*: Phylogenetic analysis using parsimony (*and other methods).

Tanabe H, Muller S, Neusser M, von Hase J, Calcagno E, Cremer M, Solovei I, Cremer C, Cremer T (2002) Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. Proc Natl Acad Sci USA 99:4424–4429

The Gene Ontology Consortium (2001) Creating the gene ontology resource: Design and implementation. Genome Res 11:1425–1433

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Venter JC, et al. (2001) The sequence of the human genome. Science 291:1304–1351

Vinogradov AE (2003) DNA helix: The importance of being GC-rich. Nucleic Acids Res 31:1838–1844

Waterston RH, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Wolfe KH, Sharp PM, Li WH (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285