Genome Analysis

Fixation of conserved sequences shapes human intron size and influences transposon-insertion dynamics

Manuela Sironi[1], Giorgia Menozzi[1], Giacomo P. Comi[2], Nereo Bresolin[1,2], Rachele Cagliani[1] and Uberto Pozzoli[1]

[1]Scientific Institute IRCCS E, Medea, Via Don Luigi Monza 20, 23842 Bosisio Parini (LC), Italy
[2]Centro Dino Ferrari, Dipartimento di Scienze Neurologiche, Università di Milano, IRCCS Ospedale Maggiore Policlinico, Via Francesco Sforza 35, 20100 Milan, Italy

The basis for intron expansion in humans is largely unexplored. In this article, we demonstrate that intron expansion has primarily been determined by fixation of multispecies conserved sequences (MCSs) over time. The presence of MCSs has shaped intron features: the insertion of transposable elements (TEs) has been constrained as more MCSs were fixed. Analysis of TE and MCS distribution suggested an unprecedented estimate of information requirements for proper splicing of long introns with indication of sequence constraints extending up to >3 kb downstream 5′ splice sites.

Introduction

Increase in number and size of intronic sequences is a general feature during the evolution of eukaryotic genomes, with introns accounting for about one-quarter of genome size in humans [1]. Large intronic sequences are transcriptionally 'expensive' [2]; and they pose an extraordinary task to splicing efficiency, rendering the significance and origin of such expansion a tantalizing question. A considerable portion of the human genome (>45%) is composed of TEs [1] and the role of these sequences in shaping intron length has been matter of debate [3–7].

TE and MCS densities as a function of intron length

We created a human intron database by extracting intervening regions from NCBI RefSeq collection of genes (for more details, see supplementary data online). We limited our analysis to introns >1 kb because additional selective forces [2], and diverse splicing-regulation constraints [8], are thought to act on short introns. We discarded first introns because their increased content in regulatory sequences is already known [9]. After these selections, a total of 43 967 introns constituted the database.

We analyzed TE integration frequency (i.e. TE copy number ÷ intronic length) in human introns, after partition in intron length classes. An initial increase with intron size was observed, leading to a maximum corresponding to a length interval of 5314–9035 bp, whereas a steady decrease was observed for longer introns (Figure 1a); the mean TE length initially increased with intron length until reaching a stable value of ~300 bp (Figure 1d). In line with these findings, the portion of the intron that is free of TEs (residual) reaches a minimum for the 9036–14 906 bp length interval and then steadily increases (Figure 1c).

We speculated that this finding might be due to the need of preserving residual intron length from TE insertion. We therefore searched for intronic multispecies conserved sequences (MCSs); in particular, MCSs were retrieved using phastCons predictions [10], which are based on a phylogenetic hidden Markov model and derive from eight species multiz alignments. Because MCSs are searched for after masking repetitive sequences, their density (i.e. conserved bases per length) profile was calculated on residual intron size portion. The MCS density curve showed a steady increase with intron length (Figure 1b), with MCS mean length also constantly increasing (Figure 1d). These trends were not observed when human genomic random sequences were analyzed (Figure 1a–d). Indeed, analysis of single introns identified a significant positive association between MCS density and residual intron length (correlation coefficient = 0.117; $P < 10^{-12}$), whereas a negative correlation (correlation coefficient = −0.168, $P < 10^{-12}$) was detected by comparing TE frequency with MCS density.

MCS fixation is responsible for the increase in intron length

Overall, the data reported here suggest that (i) retropositions might be restrained from disrupting MCSs and; (ii) MSCs might have a role in intron growth.

To gain further insight, we studied MCS fixation over evolutionary times; in particular, we analyzed the density distribution of MCSs that have been fixed after or before the bird–mammailan split. In both cases, their density profile increases with present day intron length (Figure 2a), indicating that, independently from MCS fixation age, a direct relationship exists between the presence of MCSs and intron length.

Although MCSs comprise, at most, 9% of residual intron length (Figure 1b), the hypothesis that their preservation over time is primarily responsible for present day intron size is substantiated by a couple of observations. First, indels are known to be biased in favor of deletions in vertebrates [11], but the emergence of conserved sequences might proportionally constrain deletion frequencies in those introns where they become fixed.

Corresponding author: Pozzoli, U. (upozzoli@bp.lnf.it).
Available online 6 July 2005

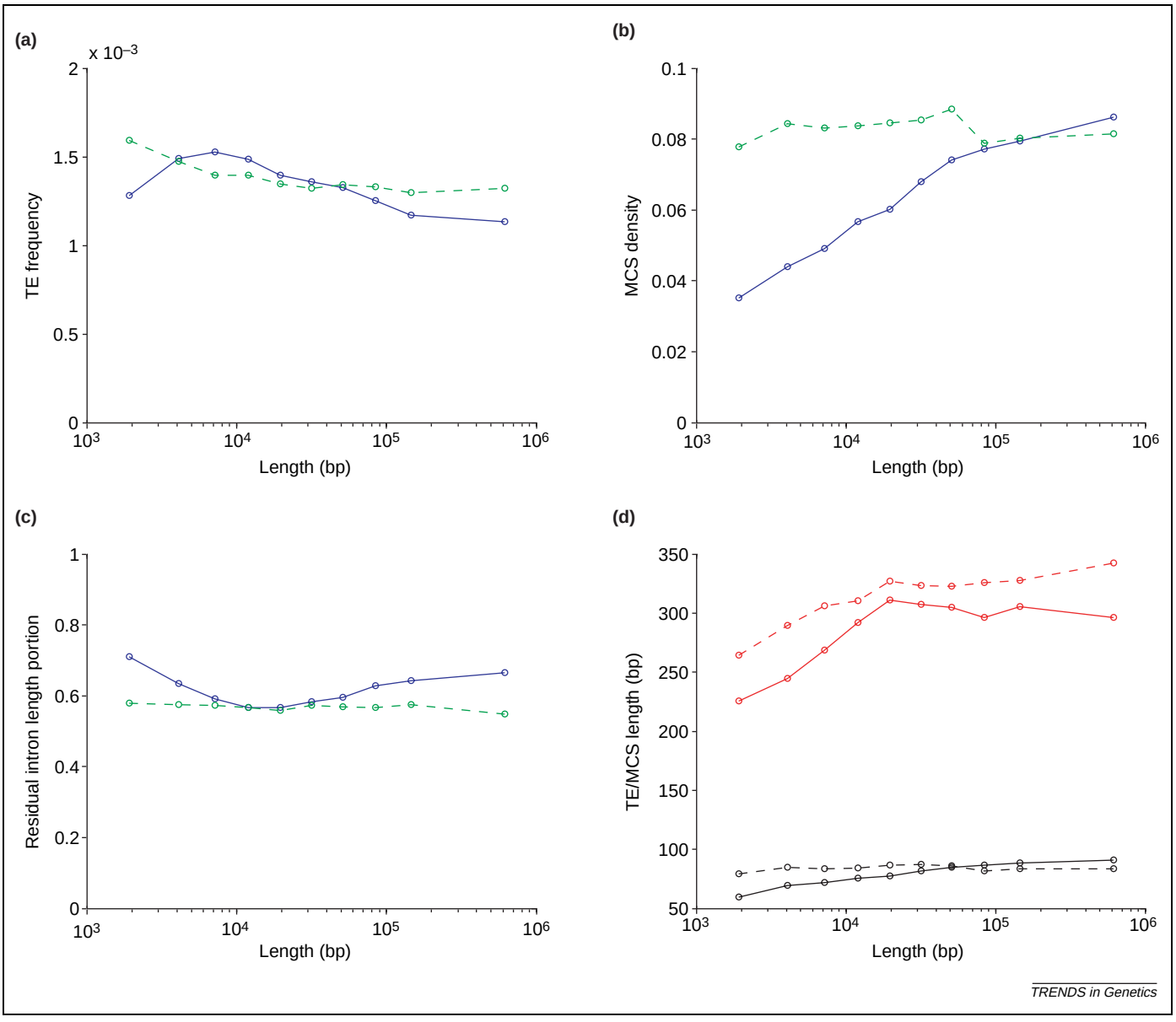

Figure 1. Analysis of MCS and TE distribution in human introns and in random sequences. To analyze MCS and TE distribution as a function of intron size, we divided intron sequences into ten length classes. The results obtained using intron sequences were checked against an equal number of independent sequences, randomly chosen for each intron from the same chromosome and having the same length. All methods concerning these analyses are described in the online supplementary material. **(a–c)** The TE frequency, MCS density and residual length portion in human introns. The results using human introns are shown in blue and those using random sequences are shown in green. **(d)** The mean TE and MCS length in introns and random sequences. TEs are in red, MSCs in black, the results of those using introns are depicted by unbroken lines and those using random sequences by broken lines.

Second, in the presence of MCSs, large deletions can be expected to be selected against more often than large insertions because deletions are expected to eliminate all MCSs located between the breakpoints, whereas insertions are predicted to only change the relative position of MCSs with respect to one another (or at most to disrupt one at the breakpoint). In line with these observations, it was previously demonstrated that, in *Drosophila*, the presence of functional elements within introns affects the insertion–deletion spectrum in favor of insertions and against large deletions [12].

Indeed, analysis of chimpanzee deletion distribution confirms this view. In particular, we selected those human regions that were deleted in *Pan troglodites* and that, in humans, contained $<50\%$ (by length) of repetitive elements so as to exclude those differences that are exclusively the result of known human-specific TE insertions. Deletion frequency showed a decreasing trend for longer human introns (Figure 2b). Moreover, the mean MCS density was calculated for all introns that displayed no deletions and for those showing at least one deletion, and a significant difference was observed (mean densities $=0.0445$ and 0.0387, respectively; t-Test, $P<10^{-3}$). Finally, deletion frequency was negatively correlated to MCS density when single introns displaying at least one deletion and one MCS were considered (Figure 2c). For further confirmation, we analyzed human and mouse orthologous introns (6050 pairs). On the basis of the abovementioned observations, introns with greater MCS densities (i.e. the longer intorns) would be predicted to be proportionally more conserved in size. Our data indicate that this is true: a significant negative correlation was observed between

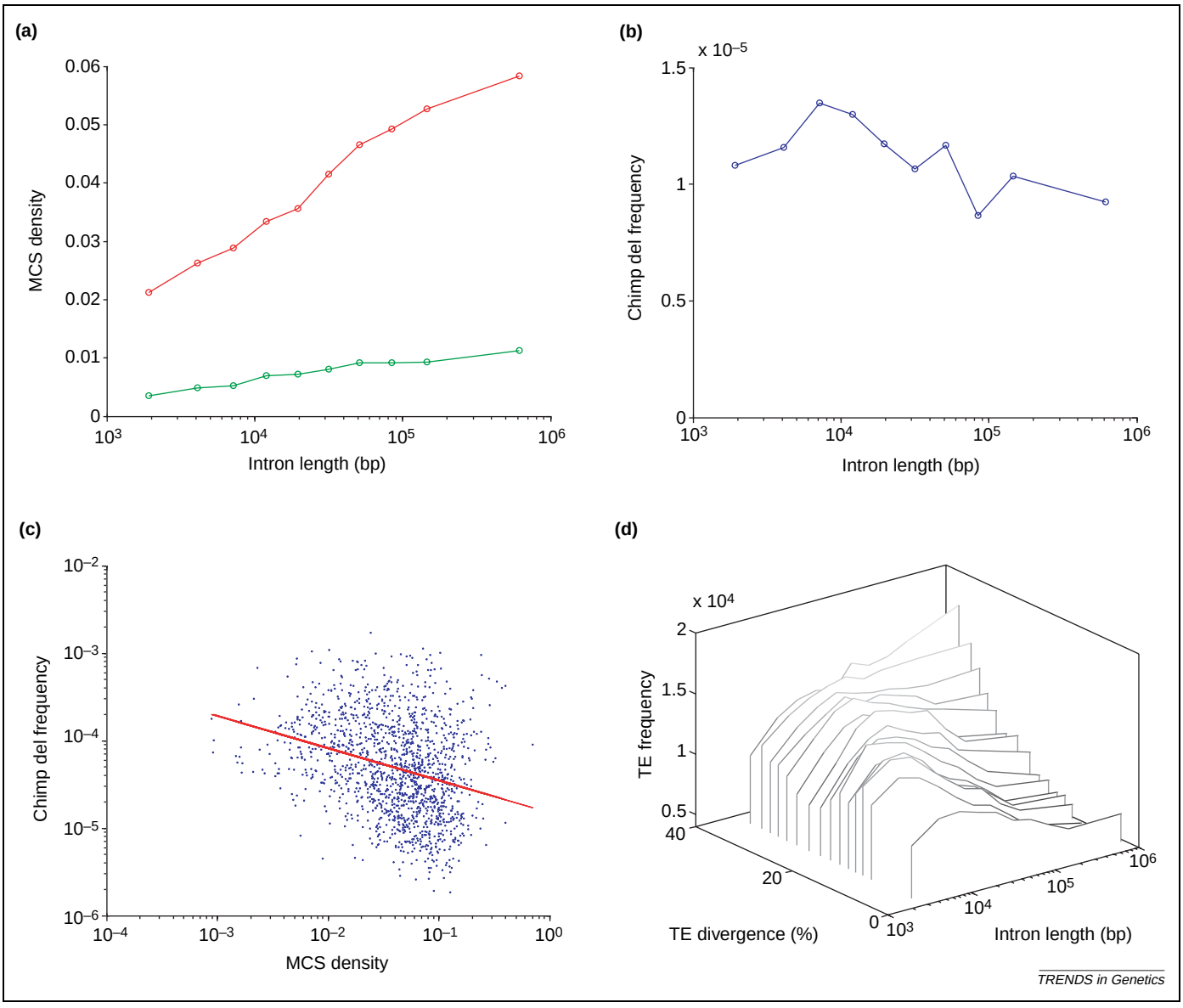

Figure 2. Analysis of MCS-fixation age, chimpanzee-deletion frequency and TE-integration dynamics. **(a)** MCS fixation at different evolutionary ages was estimated by mapping human phastCons elements onto pairwise BlastZ alignments [19] (see supplementary material online) between human genomic sequences and a second organism (chicken and mouse). Among MCS that have been preserved in humans, we identified those that have been fixed before (red line) or after (green line) the bird–mammalian split. **(b)** Frequency of chimpanzee deletions in different human intron length classes. Regions of the human genome assembly that are deleted in the chimpanzee draft assembly were retrieved through the UCSC website [chimpanzee (chimp) deletions track]. These sequences are accounted for by regions of between 80 and 12 000 bases selected so as to exclude assembly and alignment artifacts. These deletions were mapped onto the hg16 (July 2003) human genome assembly and we relocated them on the new assembly (hg17). **(c)** Regression of human and chimpanzee deletion frequency on MCS density ($\log_{10}$ transformed values) for introns displaying at least one MCS and one deletion event: R = −0.295 ($P<10^{-6}$), b = −0.368; c.i. [−0.448, −0.288] sig = 0.01; red line. **(d)** Analysis of TE integration over time. TE integration frequency was calculated, for each divergence class with respect to the inferred intron length at the time of insertion. TE divergence assignment was performed using UCSC annotation tables.

normalized Δlength [absolute values of (human – mouse length) ÷ (human + mouse length)] and MCS densities [regression of MCS density (log10 transformed values) on Δlengths: R = −0.235 ($P<10^{-6}$), b = −0.741; c.i. (−0.888, −0.594) sig = 0.01; Figure S1].

Overall, these results are consistent with the idea that fixation of MCSs in introns favors length increase by counter selecting deletion events.

TE integration frequency varies with time

To analyze their integration dynamics, we divided TEs in 15 percentile divergence (i.e. age) classes and calculated their frequency with respect to the inferred intron length at insertion time (i.e. the percentage of residual intron length plus the percentage of the size accounted for by older TEs). A gradual overall decrease in integration frequency with time was observed (Figure 2d). Remarkably, a progressive exclusion in extremely long introns was observed for lower divergence classes (i.e. for younger TEs), despite the lower average CG-content of long introns [13] and the reported preference of younger, long interspersed nuclear elements (LINE-1 or L1s) and Alus for AT-rich regions [1]. Conversely, older TEs are abundant in long introns, an unexpected finding given that MCS fixation has been stronger in those introns at any time. However, in humans, old TEs are mainly accounted for by mammalian-wide interspersed

repeats (MIRs) and LINE-2 (L2s) [1], and these sequences have been shown [14] to be over-represented within human–mouse conserved intergenic segments, suggesting that they have been subjected to negative selection preventing their divergence beyond recognition.

MCS and TE frequencies vary across intron sequences

Finally, to gain some insight into the possible functions of MCSs, we analyzed their distribution across intron sequences. Regions close to splice sites displayed an increased MCS normalized frequency (Figure 3a,b) and fewer TEs (Figure 3c,d); the most striking feature is the fact that these MCS-rich, TE-poor regions, which increase in length with intron size, reaching up to >3 kb downstream of 5′ splice sites when extremely long introns are analyzed. Distribution differences upstream of 3′ splice sites are much less pronounced although qualitatively similar. MCS and TE-frequency distribution calculated around central intron positions for each length class (Figure 3e) do not differ substantially from those reported in Figure 1a,b; thus, TE and MCS distribution in complete introns are partially accounted for, but not fully explained, by regions downstream of 5′ splice sites.

However, MCS and TE distribution across introns suggests that information requirements relevant for splicing processes might extend well beyond previous estimates [9]. This observation, although breaking commonly held views, is not surprising if, as previously demonstrated [15], a role for intron size in affecting removal kinetics is taken into account. In line with this view, a positive

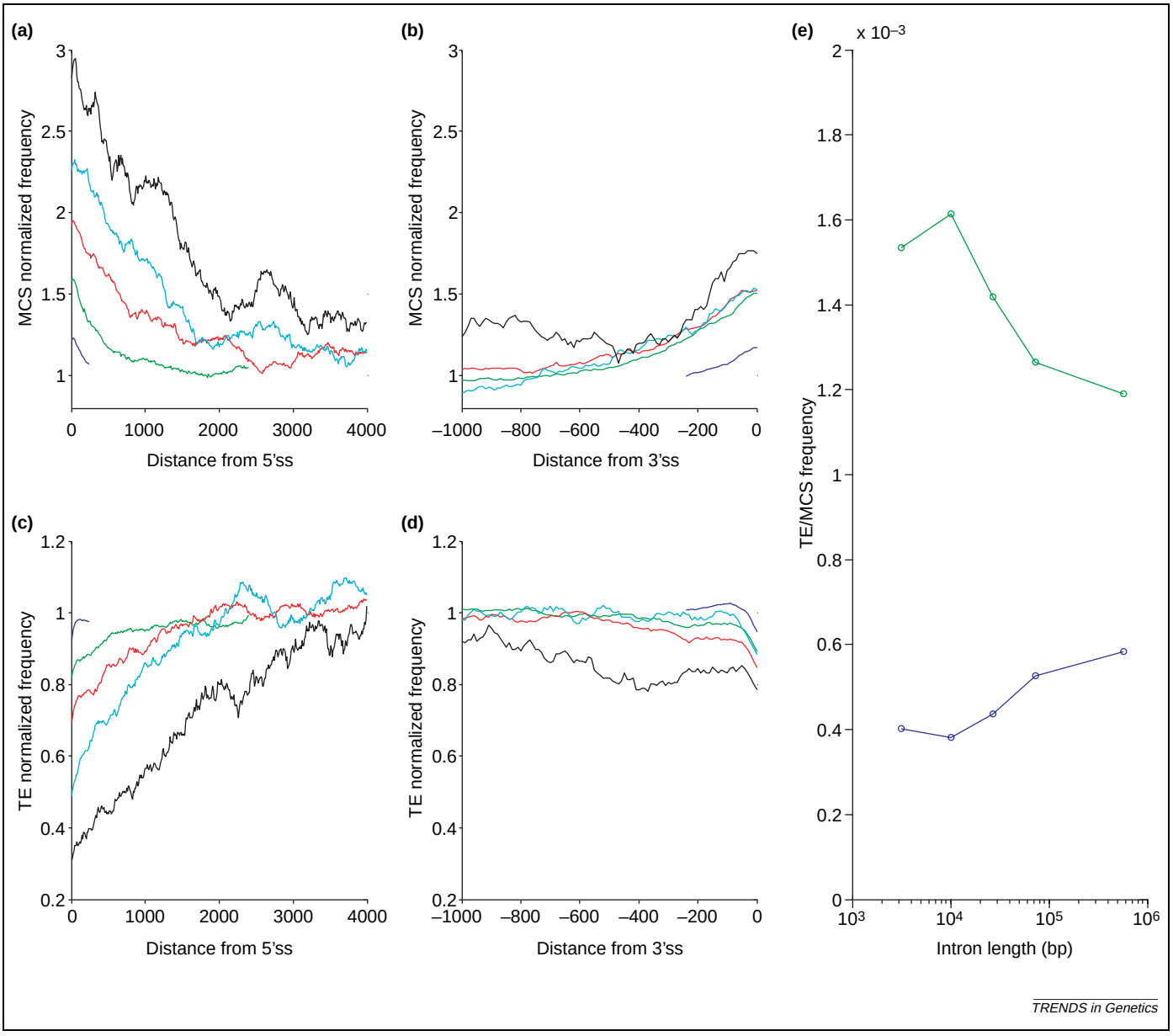

Figure 3. For analysis of MCS and TE frequency across intron sequences, introns were divided in five length classes; MCS or TE frequencies were determined at each position from splice sites (using 500-bp sliding windows moving by 10 bp) and normalized to the average frequency of the length class being analyzed calculated on 500-bp windows surrounding the middle position of each intron (see supplementary materials online for more details). MCS normalized frequencies are shown for **(a)** 5′ and **(b)** 3′ splice sites and TE normalized frequencies [shown in **(c)** and **(d)** for 5′ and 3′ splice sites, respectively]. The length intervals (in bp) are as follows: 1000–5312 (blue); 5313–14905 (green); 14906–39105 (red); 39106–105919 (cyan); 105920–1043912 (black). (**e**) MCS (blue) and TE (green) frequencies. These were calculated, for the same length classes, on 500-bp windows surrounding the middle position of each intron.

correlation between intron length and splice-site information content has been shown in *Drosophila* [16]. However, in the absence of experimental validations, the possibility that other constraints (e.g. the presence of a gradient of transcription-factor-binding sites) might account for MCS distribution across introns cannot be ruled out.

Concluding remarks

The reason(s) for the emergence and preservation of long introns in the human genome has been matter of debate. Although a role for these sequences was hypothesized, [3,5,17,18], our group [6] and others [3] had previously thought of long introns as nearly empty containers to be filled by TEs, with TEs being the major cause of disproportionate elongation and unique sequences being of little (or unknown) relevance. Our data indicate that, in long introns, unique sequences have a major impact on size and probably represent the underlying determinant of length expansion, notwithstanding their function being still largely unknown. A wealth of data is emerging on the role and relevance of non-coding conserved sequence elements in our genome; here we have shown that these elements might have influenced large-scale genomic features, such as intron size and TE-integration dynamics.

Acknowledgements

We are grateful to Roberto Giorda for useful discussions. We thank an anonymous reviewer for helpful comments and suggestions.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2005.06.009

References

1 Venter, J.C. *et al*. (2001) The sequence of the human genome. *Science* 291, 1304–1351
2 Castillo-Davis, C.I. *et al*. (2002) Selection for short introns in highly expressed genes. *Nat. Genet.* 31, 415–418
3 Vinogradov, A.E. (2002) Growth and decline of introns. *Trends Genet.* 18, 232–236
4 Cardazzo, B. *et al*. (2003) Intervening sequences in paralogous genes: a comparative genomic approach to study the evolution of X chromosome introns. *Mol. Biol. Evol.* 20, 2034–2041
5 Pozzoli, U. *et al*. (2003) Comparative analysis of vertebrate dystrophin loci indicate intron gigantism as a common feature. *Genome Res.* 13, 764–772
6 Pozzoli, U. *et al*. (2002) Comparative analysis of the human dystrophin and utrophin gene structures. *Genetics* 160, 793–798
7 McNaughton, J.C. *et al*. (1997) The evolution of an intron: analysis of a long, deletion-prone intron in the human dystrophin gene. *Genomics* 40, 294–304
8 McCullough, A.J. and Berget, S.M. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol. Cell. Biol.* 17, 4562–4571
9 Majewski, J. and Ott, J. (2002) Distribution and characterization of regulatory elements in the human genome. *Genome Res.* 12, 1827–1836
10 Siepel, A. and Haussler, D. (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* 11, 413–428
11 Petrov, D.A. (2002) Mutational equilibrium model of genome size evolution. *Theor. Popul. Biol.* 61, 531–544
12 Ptak, S.E. and Petrov, D.A. (2002) How intron splicing affects the deletion and insertion profile in *Drosophila melanogaster*. *Genetics* 162, 1233–1244
13 Duret, L. *et al*. (1995) Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* 40, 308–317
14 Silva, J.C. *et al*. (2003) Conserved fragments of transposable elements in intergenic regions: evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet. Res.* 82, 1–18
15 Bell, M.V. *et al*. (1998) Influence of intron length on alternative splicing of CD44. *Mol. Cell. Biol.* 18, 5930–5941
16 Weir, M. and Rice, M. (2004) Ordered partitioning reveals extended splice-site consensus information. *Genome Res.* 14, 67–78
17 Mattick, J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.* 2, 986–991
18 Vinogradov, A.E. (2004) Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* 20, 248–253
19 Schwartz, S. *et al*. (2003) Human-mouse alignments with BLASTZ. *Genome Res.* 13, 103–107

0168-9525/$ - see front matter
doi:10.1016/j.tig.2005.06.009

Chicken orthologues of mammalian imprinted genes are clustered on macrochromosomes and replicate asynchronously

Ulrich Dünzinger[1], Indrajit Nanda[2], Michael Schmid[2], Thomas Haaf[1] and Ulrich Zechner[1]

[1]Institute of Human Genetics, Johannes Gutenberg University Mainz, Germany
[2]Institute of Human Genetics, University of Würzburg, Germany

In the chicken genome, most orthologues of mouse imprinted genes are clustered on macrochromosomes. Only a few orthologues are located in the microchromosome complement. Macrochromosomal and, to a lesser extent, microchromosomal regions containing imprinted gene orthologues exhibit asynchronous DNA replication. We conclude that highly conserved arrays of imprinted gene orthologues were selected during vertebrate evolution, long before these genes were recruited for parent-specific gene expression by genomic

Corresponding author: Zechner, U. (zechner@humgen.klinik.uni-mainz.de).
Available online 21 July 2005