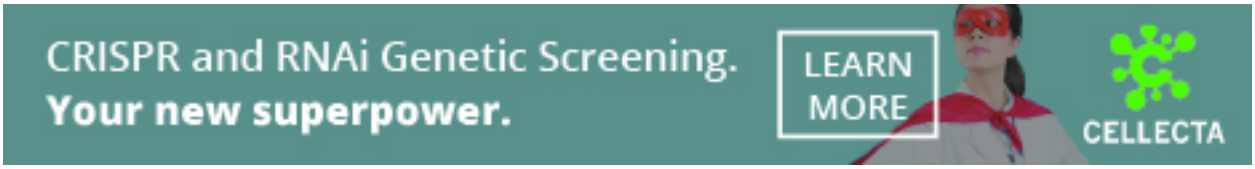
Retroposed Copies of the HMG Genes: A Window to Genome Dynamics

Liora Z. Strichman-Almashanu, Michael Bustin and David Landsman

Genome Res. 2003 13: 800-812
Access the most recent version at doi:10.1101/gr.893803



Letter

Retroposed Copies of the HMG Genes: A Window to Genome Dynamics

Liora Z. Strichman-Almashanu,[1] Michael Bustin,[2] and David Landsman[1,3]

[1]*Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA;* [2]*National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA*

Retroposed copies (RPCs) of genes are functional (intronless paralogs) or nonfunctional (processed pseudogenes) copies derived from mRNA through a process of retrotransposition. Previous studies found that gene families involved in mRNA translation or nuclear function were more likely to have large numbers of RPCs. Here we characterize RPCs of the few families coding for the abundant high-mobility-group (HMG) proteins in humans. Using an algorithm we developed, we identified and studied 219 HMG RPCs. For slightly more than 10% of these RPCs, we found evidence indicating expression. Furthermore, eight of these are potentially new members of the HMG families of proteins. For three RPCs, the evidence indicated expression as part of other transcripts; in all of these, we found the presence of alternative splicing or multiple polyadenylation signals. RPC distribution among the HMGs was not even, with 33–65 each for HMGB1, HMGB3, HMGN1, and HMGN2, and 0–6 each for HMGA1, HMGA2, HMGB2, and HMGN3. Analysis of the sequences flanking the RPCs revealed that the junction between the target site duplications and the 5′-flanking sequences exhibited the same TT/AAAA consensus found for the L1 endonuclease, supporting an L1-mediated retrotransposition mechanism. Finally, because our algorithm included aligning RPC flanking sequences with the corresponding HMG genomic sequence, we were able to identify transcribed regions of HMG genes that were not part of the published mRNA sequences.

Processed pseudogenes (PPSs) are nonfunctional genomic insertions that arise by a process of retrotransposition. This process involves reverse transcription of processed RNA polymerase II transcripts and integration into a random site in the genome (Vanin 1985; Weiner et al. 1986; Maestre et al. 1995). The resulting insertions have a structure similar to that of retroposons such as Alu and L1 elements in that they lack introns, have DNA-encoded 3′ poly(A) tails, and direct repeat sequences flanking the insertion, termed target site duplications or TSDs. These similarities between Alu and L1 elements led to the hypothesis that Alu element retrotransposition is mediated by the L1 reverse transcriptase (Boeke 1997). It has subsequently been shown that the length and composition of Alu element TSDs fit the L1 reverse transcriptase preference (Jurka 1997). Following the same logic, PPSs were also thought to be retroposed via the L1 reverse transcriptase machinery, a process that was carried out successfully in cultured cells (Esnault et al. 2000). However, the features of the TSDs and structure of the insertions were only studied in isolated cases.

Because the inserted sequences are promoterless and nonfunctional, they accumulate insertions, deletions, and single nucleotide substitutions that lead to frame shifts and premature stop codons (Ophir and Graur 1997). However, if the insertion site happens to be downstream from a promoter, the inserted sequence may become expressed (Rogalla et al. 2000; Birger et al. 2001). For this reason we prefer to use the term "retroposed copies" (RPCs) rather than "processed pseudogenes."

[3]Corresponding author.
E-MAIL landsman@ncbi.nlm.nih.gov; FAX (301) 480-2288.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.893803.

Earlier studies (Goncalves et al. 2000; Venter et al. 2001; Harrison et al. 2002) found that genes giving rise to RPCs generally have short transcripts, are widely expressed, GC-poor, and have nuclear or protein synthesis functions. A recent study (Zhang et al. 2002) analyzed RPCs of the ribosomal proteins genes and found them to be uniformly distributed throughout the genome, with a preference toward GC-intermediate regions. This analysis also found GC-poor genes to produce more RPCs. Another group of genes with a predicted large number of RPCs encodes the HMG (high-mobility group) proteins. These are abundant nonhistone nuclear proteins that bind to DNA or chromatin and affect DNA-related activities such as replication, transcription, and chromatin compaction (for review, see Bustin 1999). The HMGs are divided into three families: HMGB genes, which contain an HMG-box domain that bends DNA, affecting transcription, and include the genes HMGB1, HMGB2, and HMGB3 (formerly known as HMG1, HMG2, and HMG4, respectively); HMGN genes, which contain a nucleosomal binding domain affecting replication, chromatin structure, and transcription, and include HMGN1, HMGN2, and HMGN3 (formerly HMG14, HMG17, and Trip7, respectively); and HMGA genes, which contain an AT-hook domain affecting chromatin structure and transcription, and include the splice isoforms HMGA1a/b/c/, and HMGA2 (formerly HMGI, HMGY, HMGR, and HMGC, respectively, summarized in the HMG chromosomal proteins nomenclature home page at http://www.informatics.jax.org/mgihome/nomen/genefamilies/hmgfamily.shtml). HMG genes in both human and mouse had been suspected to have a large number of RPCs by Southern blot analyses (Landsman et al. 1986a,b; Srikantha et al. 1987; Johnson et al. 1988, 1989, 1992; Wen et al. 1989), and several HMG RPCs were identified and mapped in human

13:800–812 ©2003 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/03 $5.00; www.genome.org

(Srikantha et al. 1987; Stros and Dixon 1993; Rogalla et al. 1998, 2001; Dunham et al. 1999). Two cases of expressed HMG RPCs have been described. In one, the expressed sequence was mostly similar to the original HMG mRNA, with a similar open reading frame (ORF), and most likely a similar function (Birger et al. 2001). In the other, the retroposed copy sequence was expressed as an alternative exon of the SP100 gene, adding the HMGB DNA-binding domain to the SP100 nuclear antigen (Rogalla et al. 2000).

Retroposed copies of known genes present us with a unique tool to investigate genome dynamics and function, to learn about the process of retrotransposition, to use as complementary information to ESTs and cDNAs by representing a variety of transcripts that may or may not be represented in the present expression databases, and to learn about the creation of new genes. The anticipated large number of RPCs for the HMG genes in whole genome studies provided us with the data set for these analyses.

Previous whole genome studies of RPCs focused mainly on types of genes that were likely to generate them, excluded potentially expressed RPCs, and used criteria too general for an in-depth analysis (Goncalves et al. 2000; Venter et al. 2001). We devised a strategy aimed at the specific and accurate identification of HMG RPCs and their flanking sequences. We describe their genomic distribution, their flanking sequence characteristics, and their potential for being expressed, and we identify sequences included in RPCs that are not represented in the corresponding known mRNA sequences.

RESULTS

Retroposed Copies (RPCs) of HMG Genes

We used eight human HMG mRNA sequences representing HMGA1, HMGA2, HMGB1, HMGB2, HMGB3, HMGN1, HMGN2, and HMGN3 (Fig. 1; Table 1) in a BLASTN search (Altschul et al. 1990) against the human chromosome contig database at NCBI, which contains draft and finished sequence contigs that are part of the RefSeq database (Pruitt and Maglott 2001). In the first round, we classified genomic fragments corresponding to HMG parent genes as hits with ≥97% identities to the HMG mRNA sequence, and covering ≥70% of its length. We additionally required that an overlap or gap in the mRNA sequence between local alignments will be ≤4 bp, so these local alignments would correspond to exons, and the location of the edges of local alignments in the mRNA would correspond to splice junctions. RPCs were subsequently defined as hits in which the alignments with the HMG mRNA were uninterrupted through at least one identified splice junction, as defined above, evidence that splicing had occurred (Fig. 1). We identified 219 RPCs with identity to the HMG mRNA sequences ranging from 64% to 98%; 78% of them had ≥70% coverage of the mRNA (Table 1). An additional 37 hits were all confined to the last exon of the HMG gene, and as these hits did not cross splice junctions, we categorized them as PSs (for pseudogenes). Although they were not RPCs by our criteria, some had TSDs, and it is likely that these hits represented truncated RPCs resulting from an incomplete reverse transcription event. The PS hits were not included in the analyses of chromosomes, TSDs, repeats and GC content in flanking sequence, and insertion start/end points; however, we did analyze them individually. RPCs could be divided into two groups. The abundant group included RPCs of HMGB1, HMGB3, HMGN1, and HMGN2,

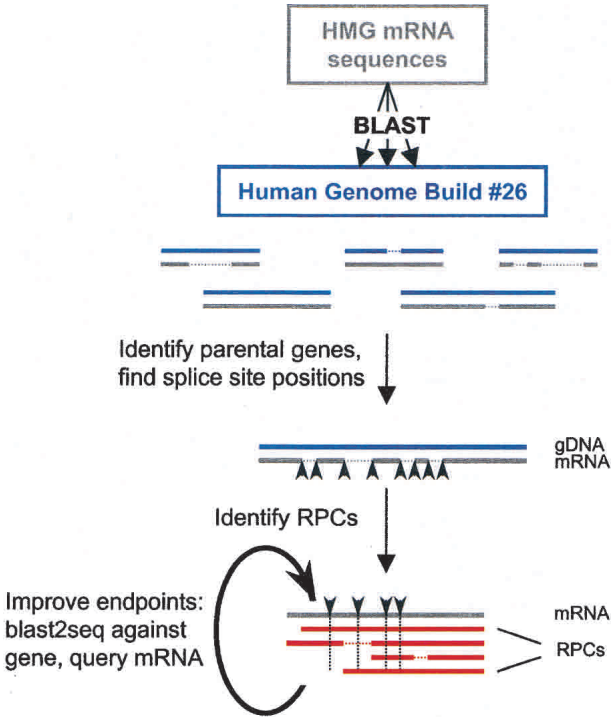


Figure 1 Identification of retroposed copies (RPCs) for HMG genes. HMG mRNA sequences (in gray) were soft-masked and used in BLAST searches against the Human Genome Build #26 chromosome files (in blue). HMG parental genes were first identified (see Methods), and splice junction positions in the mRNA sequences were defined from these alignments (black arrowheads) and were subsequently used as guides for identifying RPCs (in red). RPC endpoints were adjusted by conducting a BLAST2SEQ comparison against the corresponding HMG gene, as well as with less-stringent parameters (see Methods) against the HMG mRNA. Horizontal dashed lines represent gaps in the alignment, and vertical dashed lines represent the splice positions in the mRNA–RPC alignment.

which each had 33–65 copies; the sparse group included RPCs of HMGA1, HMGA2, HMGB2, and HMGN3, with each only 0–6 copies. Among the RPCs were the previously reported processed pseudogenes HMG1L1/3/4/5/6/8/9 (Stros and Dixon 1993; Rogalla et al. 1998), HMG17L1/3 (Dunham et al. 1999), clones 28H and 60H (Srikantha et al. 1987), and HMGIYL3 (Rogalla et al. 2001).

Characterization of Integration Sites

An analysis of pseudogene distribution on Chromosomes 21 and 22 indicated an uneven distribution across the chromosome length, with an excess of pseudogenes in the 5 Mb closest to the centromere (Harrison et al. 2002). To find potential hot spots for RPC integration, we looked at their chromosomal localization (Fig. 2). Although the RPCs seemed to be distributed throughout the genome, a region of higher density was found on Chromosome arm 15q. Conversely, no RPCs were found on Chromosome Y, and Chromosome 8 contained only HMGB1 RPCs. We did not find, however, an increased density of RPCs around centromeres.

To scan the integration sites for repeats and unique sequences, we used RepeatMasker (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker; A.F.A. Smit and P. Green, unpubl.) to identify and mask repetitive ele-

Table 1. RPCs of HMG Genes

HMG gene	Query mRNA	mRNA length (bp)	CDS (% GC)	Number of exons	Location	Total number of matches	Number of RPCs (% ID)	Number of PS (% ID)
HMGA1	NM_002131.1	1875	60	6	6p21.1b	8	6 (75–95)	2 (74–83)
HMGA2	NM_003483.3	4111	58	?	12q14.1b	0	0	
HMGB1	D63874.1	1194	44	5	13q12.3a	59	57 (64–97)	2 (73–88)
HMGB2	X62534.1	1301	47	5	4q34.1b	4	2 (86–90)	2 (80–90)
HMGB3	NM_005342.1	1633	46	5	Xq28d	45	33 (66–95)	12 (72–94)
HMGN1	NM_004965.1	1240	51	6	21q22.3a	66	55 (66–97)	11 (78–93)
HMGN2	NM_005517.1	1198	51	6	1p36.11c	73	65 (82–98)	8 (84–96)
HMGN3	AF274949.1	863	46	6	6q16.1b	1	1 (93)	0
All						256	219	37

Query mRNA sequences were taken from RefSeq whenever possible. (RPCs) Retroposed copies; (PS) hits that did not cross splice junctions, see text; (% ID) identity to the query mRNA.

ments in the 250 bp on either side of the RPCs, followed by BLASTN searches with the masked sequences. Forty-five percent of the combined 5′-flanking sequence and 39% of the combined 3′-flanking sequence were masked by RepeatMasker and labeled as interspersed repeats (Table 2). In comparison, 45% of the human genome is composed of interspersed repeats (Lander et al. 2001). Most of the difference between the 5′ and 3′ flanks was caused by an excess of SINE elements (specifically Alu elements) in the 5′-flanking sequence (18%) relative to the 3′-flanking sequence (13%). The genomic fraction of SINE elements is also 13%. This could be a result of the preference of RPCs to integrate in stretches of A, which occur more frequently in the 3′ ends of Alu elements.

Four RPCs were found to flank a gene or another putative RPC (Fig. 3). An HMGN1 RPC on Chromosome 10p13 was identified in a tail-to-tail configuration with a sequence 92% identical to the 3′ end of Ca-ATPase (M23114) mRNA located on 12q24 (Fig. 3A). The corresponding genomic sequence for Ca-ATPase did not contain introns, and we could not identify a poly(A) tract or a target site duplication to ascertain whether it was, indeed, an RPC. From the differences in similarity to their respective mRNAs located on separate chromosomes, as well as from their relative orientation, it is most likely that these two pseudogenes were generated in two separate events. An HMGN1 RPC on 2q23 overlaps the 5′ end of the KARP-1 gene (AF039597) in the opposite orientation, and encodes the first two exons of this gene, according to the GenBank gene record (Fig. 3B). An HMGN2 RPC on 2q13 was identified in intron 4 of the IL1 homolog (NM_019618), in the opposite orientation (Fig. 3C). An HMGN2 RPC on 17q23 was identified immediately upstream of a sequence 95% identical to ribosomal protein L12 (L06505) and containing a poly(A) tail and TSD, most likely also an RPC (Fig. 3D). It is unlikely for both of these RPCs to have been created in a single event because both contained distinguishable TSDs derived from the sequence between them, which was identified by RepeatMasker to be a fragment from a THE1-int LTR/MaLR element (see Fig. 3D).

The base composition was calculated for 500 bp flanking the RPCs and was found to be AT-rich (Table 2). The 3′-

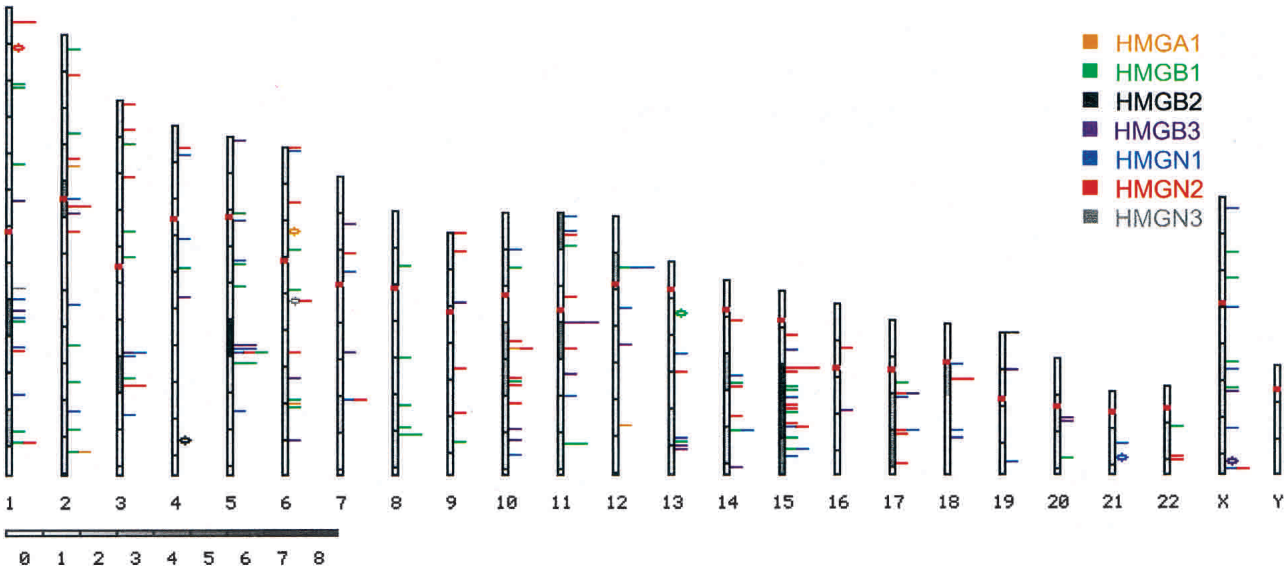


Figure 2 Genomic distribution of 198 mapped RPCs. The colored tick marks represent individual RPC locations according to the legend; a diamond shape represents the corresponding gene. The chromosome shading illustrates the number of RPCs in 20 megabase bins, according to the horizontal scale below; red regions in the chromosomes are centromeres.

flanking sequence average GC content was 37%, and the 5′-flanking sequence average GC content was 40%, both close to the genomic average of 41% (Lander et al. 2001). The slight decrease in GC content in the 3′ flank could be due to the presence of poly(A) tracts.

Expression of RPCs

Most RPCs are not expressed, and therefore are referred to as processed pseudogenes. However, the few that are expressed can potentially give rise to novel HMG genes, or to modular combinations of HMG domains with domains from other genes. We identified potentially expressed RPCs through BLASTN searches against mRNA and EST databases. A significant EST or mRNA hit was considered evidence for potential expression if it exhibited ≥99% similarity to the RPC but ≤97% to the HMG mRNA.

Altogether, we found 22 potentially expressed RPCs and 3 potentially expressed PSs (Table 3), of which three had already been reported: HMG1L3, an RPC of HMGB1 that is expressed as an alternative exon of the SP100 gene (Rogalla et al. 2000); HMGN4, an intronless gene that consists of an RPC of HMGN2 surrounded by additional sequence (Birger et al. 2001); and HMG17L1, an RPC of HMGN2 reported as part of the Chromosome 22 project (Dunham et al. 1999).

Of the 22 novel expressed RPCs and PSs, 6 had an ORF similar to the corresponding HMG ORF (e.g., Fig. 4A,C,E); 6 had an ORF unrelated to the HMG ORF—a different reading frame, orientation, or location in the HMG transcript (e.g., Fig. 4C,D); and 3 were included as part of a transcript with a coding sequence (CDS) outside the pseudogene region (e.g., Fig. 4B,D,F). Interestingly, two of the three encoded alternative exons (Fig. 4B,D), and one was embedded in an alternative 3′-UTR region, thereby supplying an alternative polyadenylation signal. This last gene model was supported by human ESTs, as well as by mRNAs from monkey and mouse (Fig. 4F). Alternative splicing also plays a role in the only published example of an HMG RPC expressed as part of another gene, in which case an RPC of HMGB1 encodes an alternative exon of the SP100 gene, donating an HMG domain to the SP100 nuclear antigen to create the SP100–HMG splice variant (Ro-

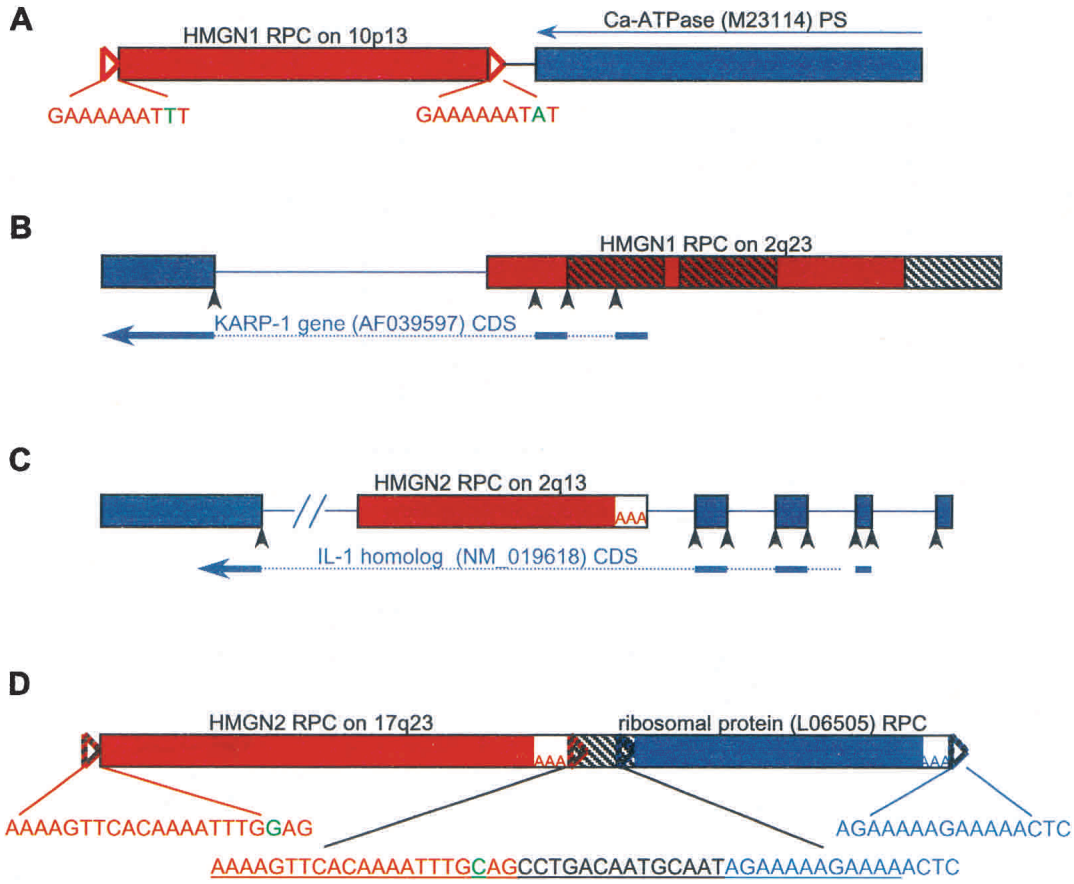


Figure 3 Genes and pseudogenes flanking HMG RPCs. HMG RPC (shown as red boxes) flanking sequences were masked and used in a BLAST search to find surrounding genes and pseudogenes (depicted as blue boxes). The orientation of all HMG RPCs is *left* to *right*. (*A*) A pseudogene of the Ca-ATPase (M23114, 92% identical) was found in the reverse orientation (denoted by a thin blue arrow) to an HMGN1 RPC (89% identical), at a distance of 40 bp (intervening black line). (*B*) An HMGN1 RPC (88% identical) in the 5′ flank of the KARP-1 gene contributing two exons. The three hatched boxes depict Alu elements. (*C*) An HMGN2 RPC (89% identical) in the fourth intron of the IL-1 homolog. (*D*) An HMGN2 RPC (93% identical) upstream and in the same orientation as an RPC of ribosomal protein L12 (95% identical; see text). The 47 bp between them is a fragment of a THE1 element (underlined in the middle insert), from which the TSDs for both RPCs are derived (colored nucleotides). (Open triangles) TSDs; TSD sequences appear in the inserts, and nucleotides in green represent ambiguity; (AAA) poly(A) tracts; (black arrowheads) splice sites; (hatched boxes) repetitive elements; (thick blue arrows) CDS; (dotted lines) position of introns in CDS; (lines between the boxes) intervening genomic DNA, or introns if colored; (–//–) a break artificially inserted into the long sequence for convenient display. Figure not drawn to scale.

Table 2. Composition and Repeat Content of Sequence Surrounding HMG RPCs

	Total (bp)	% GC	% Repeats	% LINEs	% SINEs	% Alu
5′ flank	54,718	40	45	16	18	16
3′ flank	54,478	37	39	16	13	12
Genomic		41	45	21	13	11

A total of 250 bp of sequence upstream or downstream of RPCs was analyzed for repeats and GC content. Genomic parameters were taken from Lander et al (2001). The % repeats and GC content are from RepeatMasker output.

galla et al. 2000). Six potentially expressed pseudogenes included putative splice sites (e.g., Fig. 4B,D,E), which was supported by the alignment of ESTs with the genomic sequence in some cases.

From RPCs Back to mRNA

Because RPCs are derived from mRNA molecules, a comparative analysis of their sequences may reveal the existence of variant mRNA species that are expressed only under specific conditions, are extinct, or are difficult to clone. Through our survey of RPCs derived from HMG mRNAs, we found RPCs that contained sequences present in the genomic sequence of the HMG gene but not in the published mRNA databases (Table 4).

The alignment of six HMGB1 RPCs from five different chromosomes to the HMGB1 genomic sequence did not end at the known 3′ end of the gene, but rather extended downstream for an additional 1 kb, including a putative polyadenylation signal (Fig. 5; Table 4). This is evidence that a transcript longer than the known HMGB1 mRNA was being created through the alternative use of poly(A) signals. The percent identity of these RPCs to HMGB1 ranged from 82% to 97%, implying that they were created in separate retrotransposition events and that this longer transcript represented ~10% of HMGB1 transcripts. The existence of this 3′-extended HMGB1 transcript was supported by several human ESTs and a cDNA clone from spleen (AK057120). Additionally, we found 40 RPCs of HMGB1 that extended up to 100 bp upstream of the mRNA start point (Fig. 5; Table 4), indicating a previously inaccurate identification of the correct transcription start site for this gene. Such 5′- or 3′-extended RPCs were found for other HMGs as well, in many cases without a corresponding transcript in expression databases (summarized in Table 4).

Evidence for alternative splicing of HMG genes was observed for HMGA1, which is known to have several alternative transcripts with the CDS starting at exon 3 (Johnson et al. 1989; Friedmann et al. 1993). Six RPCs were found for HMGA1 (Fig. 6). All of these RPC sequences begin within the first exon and correspond to a variety of spliced isoforms, which include HMGA1a (variant 2, NM_002131); HMGA1b, which has an insertion of 33 bp from intron 3 adding 11 amino acids to the CDS (variant 1, X14957); and another isoform that was the result of exon 2 skipping (variant 5, M23617). None of the RPCs had the structure of HMGA1c (AF176039) or included the alternative exons present in variants 3, 4, 6, or 7. A novel splice isoform with no corresponding mRNA was represented among the RPCs, resulting from a combination of exon 2 skipping and a 33-bp insertion. This finding demonstrates the utility of this method, enabling us to discover new variant transcripts.

Mechanism of Generation

It has been suggested that RPCs are generated by the L1 reverse transcriptase (Ostertag and Kazazian Jr. 2001). In common with the end product of retrotransposition by an L1 reverse transcriptase, RPCs include a poly(A) tract at the 3′ end of the insertion and are flanked by direct repeats termed target site duplications (Vanin 1985). We scanned all the RPC flanking sequences for direct repeats that were ≥7 bp long and ≥90% identical, as well as for potential poly(A) tracts close to the 3′ end of the insertion. The distance between the direct repeats and the insertion ends was permitted to be up to 30 bp to allow for differences resulting from mutations and inaccurate identification of the ends. TSDs were found for 48% of all RPCs. However, a higher fraction of TSDs was found for subgroups in which the ends were defined with higher confidence. For example, 69% of RPCs that matched ≥70% of the mRNA length had TSDs. Furthermore, more recently retroposed RPCs that share a higher percentage of identity with an mRNA are more likely to contain intact TSDs (data not shown).

The L1 reverse transcriptase had been shown to have a sequence preference for integration (Feng et al. 1996). A corresponding preference of TT/AAA was found at the junctions between 5′ TSDs and 5′-flanking sequences of Alu elements that are thought to be retroposed via the L1 reverse transcriptase machinery (Boeke 1997; Jurka 1997). To find sequence patterns around the TSD–5′-flanking sequence junction (which represents the first nick in the L1 integration process), we aligned TSDs including 10 bp of 5′-flanking sequence justified at the 5′ end. We also aligned TSDs with 10 bp of 3′-flanking sequence justified at the 3′ end; this junction between TSD and the 3′-flanking sequence represents the second nick. The sequence preference found at the junction between the 5′ TSD and the 5′-flanking sequence was TT/AAA, which corresponds to the L1 preference as well as to the sequence at the 5′ junction of Alu elements (Fig. 7A). No consensus could be found at the 3′ junction. It is possible that the second nick is determined by the distance from the first nick rather than by the sequence at the site (Jurka 1997; data not shown). This distance corresponds to the length of the TSD, which is generally ~15 bp. We found an average length of 14 bp for TSDs (Fig. 7B), as well as false peaks at the shortest TSD size allowed (7 bp) and at 10 bp, the shortest in which mismatch is allowed. The size of these false peaks is reduced when considering only insertions with >90% identities to HMG or that cover >70% of its length (data not shown).

Because inversions/truncations are observed in 12% of L1 elements and are thought to be a feature of the L1 reverse transcriptase (Symer et al. 2002; Szak et al. 2002), we searched for and identified similar inversions/truncations in RPCs. One RPC and one PS were found to have an inverted and truncated 5′ region (Fig. 7 C,D), and because we could find a TSD and a poly(A) tract for the PS, it strengthens the likelihood that it was, indeed, a truncated RPC (Fig. 7D).

Taken together, the similarity of RPC structure to that of L1 and Alu elements, the similar length distribution and sequence composition of their TSDs, and the presence of inversions/truncations events among the RPCs, all support an L1 reverse transcriptase mechanism for the generation of RPCs. L1 is a mammalian retroelement—although all mammals were found to have sequences related to L1 elements, no such

Table 3. Potentially Expressed Pseudogenes

HMG gene	RPC gi (bp)	% ID[a]	Location	Number of ESTs[b]	ORF[c]	mRNA/gene structure[d]	Published RPCs
HMGA1	16163921 (1535901–1537678)	95	12q24.13	1	HMG	Figure 4A	HMGIYL3 (Rogalla et al. 2001)
HMGA1	16172601 (7397043–7398616)	76	6q24.1a	2 (BF514910, AW902237)	—	—	—
HMGA1	16180215 (564321–564891)[e]	76	9q34.13a	>4	non^{+0}	AS (Fig. 4B)	—
HMGB1	14724993 (27354–28521)	90	2q37.2a	4	HMG, non^{+0}	AS, AF056322 (Rogalla et al. 2000)	HMG1L3/AF076675 (Rogalla et al. 1998)
HMGB1	16165405 (251300–252448)	90	17q12a	1 (AA730118)	HMG	—	—
HMGB1	16192427 (955070–956344)	82	15q21.2c	1 (AI018785)	—	—	—
HMGB1	16194785 (21115834–21117043)	94	20q13.33	1	HMG, non^{-}	Figure 4C	HMG1L1/AF076674 (Rogalla et al. 1998)
HMGB1	16161490 (9591–10590)	70	8q24.12c	1 (AA053962)	non^{-}	—	—
HMGB3	16163451 (854636–857976)	91	Xq26.3a	1 (BF855795)	HMG	Intron	—
HMGB3	16180809 (2446695–2449023)	93	9p11.2a	4 (BF897138, BF919297, AI476295, BE677063)	—	—	—
HMGB3	16192031 (403162–405637)[e]	84	15q22.31	2 (BE930983, AA662247)	non^{-}, non^{+}	Intron	—
HMGN1	16162863 (320185–321432)	84	13q32.2a	1 (AA878378)	HMG	—	—
HMGN1	16164517 (140672–142564)	91	17q21.2b	2 (BI041905, BI057720)	—	—	—
HMGN1	16168858 (540504–541991)	85	1q23.1b	1 (AA282139)	—	—	—
HMGN1	16185437 (922703–923747)	82	12q13.12	3	non^{+}, non^{-0}	AS (Fig. 4D)	—
HMGN1	16164531 (3471415–3471654)[e]	93	17q25.1b	1 (AA484918)	—	—	—
HMGN2	10879979 (281050–282482)	89	22q13.31	0	HMG	NM_021024 (Dunham et al. 1999)	HMG17L1 (Dunham et al. 1999)
HMGN2	16156389 (265016–266479)	89	10q24.1b	2 (AW851196, AW851333)	non^{-}	—	—
HMGN2	16159856 (945234–946431)	91	5q23.1d	1 (AU185079)	—	—	—
HMGN2	16164517 (592249–593426)	94	17q21.2b	3	HMG	1/2 splice (Fig. 4E)	—
HMGN2	16173680 (237789–238927)	88	6p25.2b	1 (BF887233)	—	—	—
HMGN2	16177898 (6700982–6702095)	89	7q32.2a	1 (BG819180)	—	—	—
HMGN2	16190521 (11214161–11215280)	91	14q23.2b	5	non^{+0}	AA, AB037814 (Fig. 4F)	—
HMGN2	16191910 (1631498–1632678)	90	15q15.2a	2 (BG213025, AA627886)	non^{+}	Intron?	—
HMGN2	16195861 (4365380–4366052)	85	6p21.1b	>4	HMG	NM_006353, HMGN4 (Birger et al. 2001)	HMG17L3

[a]Relative to the query mRNA.
[b]The rest of the ESTs are noted in Figure 4 or belong to a published expressed +/−.
[c](HMG) ORF similar to HMG CDS; (non) ORF unrelated to HMG CDS; (non^{+}) orientation relative to RPC; (non^{0}) ORF outside RPC.
[d]mRNA/gene structure is derived from alignment of ESTs/cDNAs with RPC. (AS) Alternative splicing; (AA) alternative poly(A) signal.
[e]PS not defined as RPC, may have TSD, poly(A).

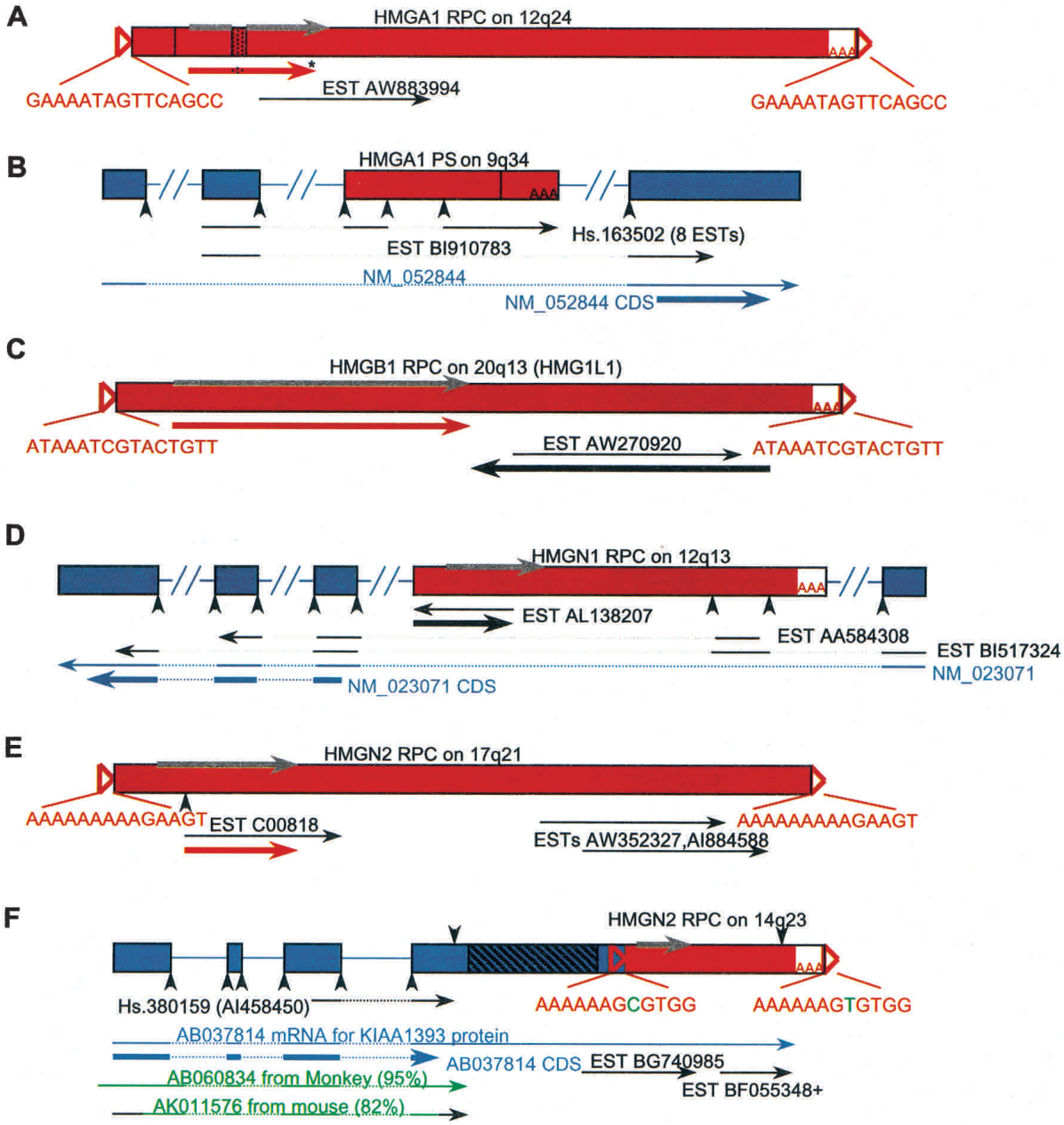


Figure 4 Potentially expressed RPCs. RPCs were masked and used in a BLAST search against the mRNA and EST databases. (*A*) HMGA1 RPC lacking exon 2 (vertical line) and with a 33-bp insertion (stippled box) harbors an ORF (red arrow) similar to HMGA1 CDS with an insertion (embedded stippled line) and a premature stop codon (asterisk), as well as one EST (represented by a thin black arrow). (Gray arrow) The CDS of HMGA1. (*B*) An HMGA1 PS with a 500-bp deletion (vertical line) encodes exons shared with an alternatively spliced mRNA (NM_052844; thin blue arrow) through ESTs (thin black arrows); (dotted lines) introns in transcribed sequence. The HMG insertion ends with poly(A); however, it is internal to the HMGA1 transcript. (*C*) HMGB1 RPC has an ORF (thick black arrow) in an opposite orientation to EST AW270920, and another ORF (thick red arrow) similar to HMG CDS. (*D*) An HMGN1 RPC has an ORF within EST AL138207 (thick black arrow) that is not similar to the HMGN1 CDS, and encodes an exon shared with an alternatively spliced mRNA (NM_023071; thin blue arrow) through ESTs. EST AA584308 is part of UniGene cluster Hs.152982. (*E*) An HMGN2 RPC has an ORF similar to HMG CDS (thick red arrow) and three ESTs in the same orientation. (*F*) An HMGN2 RPC in the 3′-UTR of cDNA AB037814 provides an alternative poly(A) signal (downward-pointing arrow). ESTs illustrate the use of both signals; a + after EST BF055348 stands for more ESTs at the same position. The gene structure of AB037814 was derived from an alignment with genomic DNA. (Green arrows) Monkey and mouse mRNA sequences similar to AB037814 that do not include an HMG sequence; the black part of the mouse DNA represents nonaligning sequence; (hatched box) a sequence of repetitive elements; (open triangles) TSDs, TSD sequences appear in the inserts, nucleotides in green represent ambiguity; (AAA) poly(A) tracts; (downward-pointing arrowheads) poly(A) signals; (upward-pointing arrowheads) splice sites; (thick arrows) ORFs or CDSs: The position of the HMG CDS is depicted in gray; ORFs similar to HMG CDS are red; ORFs not similar to HMG CDS are black, and CDSs outside the RPC region are blue. (Dotted lines) Intron positions in transcripts; (lines between the boxes) introns in genomic DNA; 3′ EST orientation is reversed to presumed sense; not drawn to scale.

Table 4. Possible Extensions to mRNA Ends

HMG gene	Query mRNA	← 5′ extensions →			← 3′ extensions →		
		RPCs	cDNA	ESTs	RPCs	cDNA	ESTs
HMGA1	NM_002131.1	19 bp (1)	—	—	—		
HMGB1	NM_002128.2	≤100 bp (40)	40 bp (1)	≤80 bp (several)	≤1 kb (6)	1 kb (1)	≤2 kb (several)
HMGB3	NM_005342.1	10 bp (10)	—	—		—	—
HMGN1	NM_004965.1	20–62 bp (6)	—	20–120 bp (several)	10–20 bp (25)	—	≤10 bp (several)
HMGN2	NM_005517.1	24 bp (1)	—	≤66 bp (several)		—	—

RPCs with parental gene alignments that extend beyond the known transcript start and end points indicate the existence of longer transcripts. Numbers in parentheses are the number of sequences in the sequence category that corresponds to the base pair extension. HMGB1 query sequence NM_002128.2 is an updated version of the similar sequence D53874.1 used in Table 1.

sequences could be found in nonmammals such as duck and sea urchin (Hutchison et al. 1989). If L1 is, indeed, the agent retroposing RPCs, it is not surprising that HMG RPCs could not be found in chicken (Landsman et al. 1988; Srikantha et al. 1989; Lum et al. 2000).

DISCUSSION

RPC Data as Supplement for Expression Databases

Several studies and genomic surveys of retroposed copies of genes have been carried out, focusing mainly on the characteristics of genes that generate RPCs (Goncalves et al. 2000; Venter et al. 2001; Harrison et al. 2002). In this work we chose a different perspective and studied the insertions themselves to better characterize each RPC and gain insight into the mechanism of their generation.

Because RPCs are derived from mRNAs, they provide a source of transcript information that is independent of expression sequence databases. RPCs reflect active transcription of genes at a certain time and in a certain tissue, and are free of cloning artifacts. If aligned directly to HMG genes, RPCs can reveal the usage of certain transcripts or isoforms that no longer exist or are rarely expressed, as well as sequences that have eluded cloning. We found deviations from published mRNA sequences for several HMG genes. The most instructive example comes from HMGB1. We found that 10% of HMGB1 RPCs with a wide range of identities with HMGB1 mRNA included an additional 1 kb of sequence immediately 3′ to the end of the gene. EST sequences and one cDNA clone isolated from spleen were found matching the extended sequence (Fig. 5), implying that this transcript presently exists. Because the known HMGB1 mRNA, the RPCs, and the 3′ ESTs starting from the respective endpoints all ended with poly(A) tails, we concluded that this longer transcript is the result of an alternative downstream polyadenylation signal. Both the long and the short transcripts encode the same protein, thus the additional sequence could serve a regulatory role. The existence of alternative transcripts for HMGB1 of the sizes corresponding to the known and the new transcription endpoint was previously observed on a Northern blot (Wen et al. 1989). Interestingly, this longer 3′-UTR was included in the only mRNA from pig found in GenBank, and together with the shorter isoform in mouse, indicating that it is evolutionarily conserved. At the 5′ end of HMGB1, 66% of RPCs included additional sequence of up to 100 bp upstream. ESTs were found to extend as far as 80 bp upstream, and a cDNA was found that

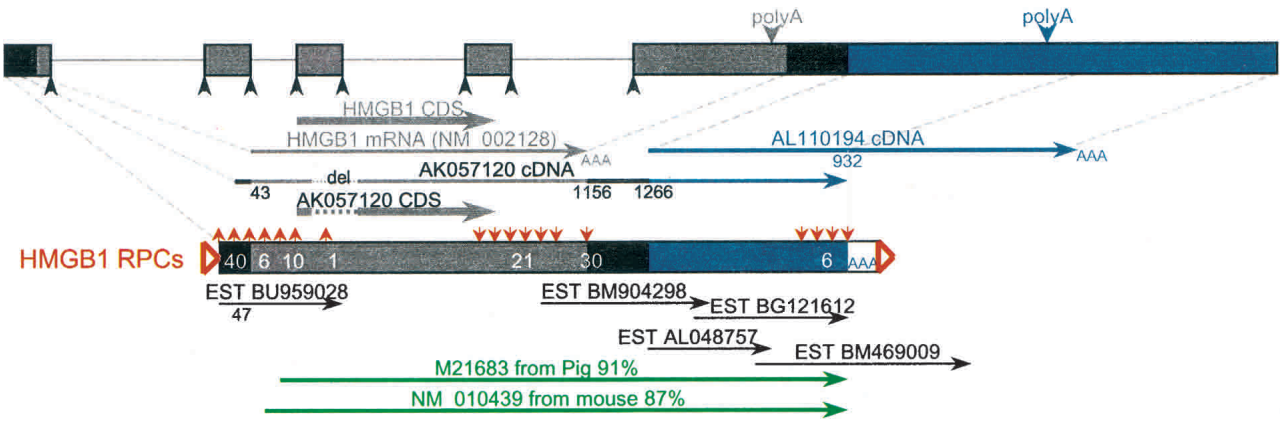


Figure 5 Transcripts of the human HMGB1 gene: alternative 3′ end, extended 5′ end. RPCs aligned directly to the HMGB1 genomic DNA (gray boxes) were found to have sequences extending upstream and downstream (black and blue boxes) relative to the HMGB1 RefSeq mRNA NM_002128. The middle structure is a schematic representing all RPCs derived from HMGB1, with gray structures corresponding to NM_002128 sequence, blue structures corresponding to a cDNA downstream from NM_002128 (AL110194), and black boxes are genomic sequences outside these cDNAs. The numbers within the boxes show how many RPCs start (with upward-pointing red arrows) or end (with downward-pointing red arrows) within this region. ESTs shown belong to UniGene cluster Hs.337757, and illustrate the alternative use of poly(A) signals (downward-pointing blue and gray arrowheads), as well as the existence of transcription upstream and downstream to NM_002128, also supported by a cDNA (AK057120). (Green arrows) Pig and mouse HMGB1 mRNAs, which also extend downstream of NM_002128. (Open triangles) TSDs; (AAA) poly(A) tracts; (upward-pointing black arrowheads) splice sites. The dotted lines in AK057120 depict a 70-bp deletion, and 3′ EST orientations are reversed.

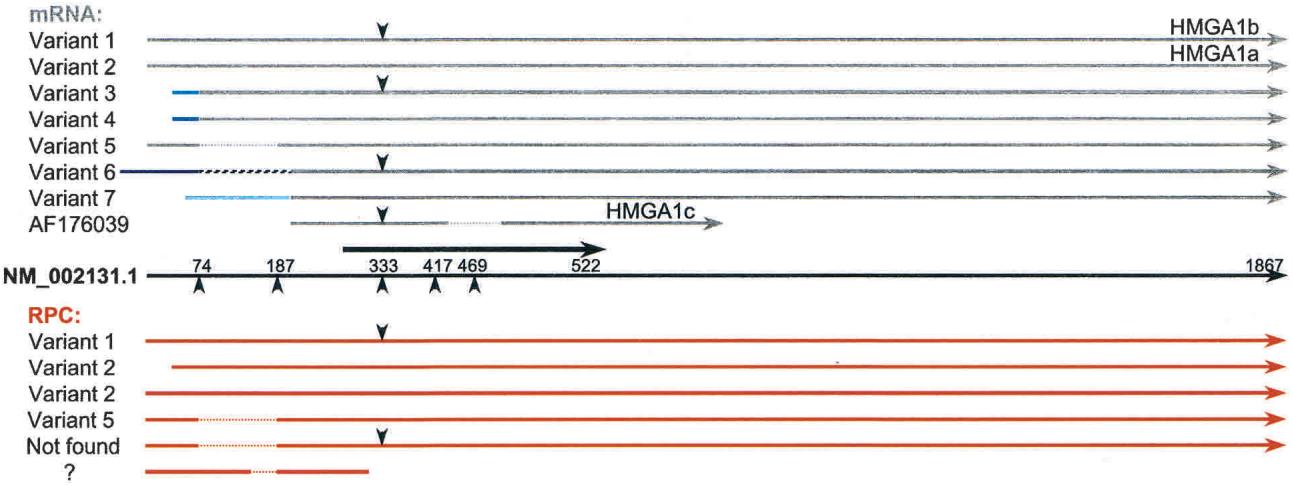


Figure 6 HMGA1 RPCs aligned with mRNA splice isoforms. (Black arrow in the middle) The NM_002131.1 RefSeq mRNA; (thick black arrow) its CDS position. Above are transcript variants in gray, below are RPCs in red; different shades of blue lines represent different alternative exons, and the hatched line in variant 6 represents an Alu element. (Downward-pointing arrowheads) The position of an additional 33 bp between exons 3 and 4 in some of the transcripts and RPCs. (Upward-pointing arrowheads) Splice positions; (dotted lines) deletions or exon skipping.

extended 40 bp upstream (Fig. 5; Table 4). RPCs of other HMGs also included 5′-extended sequences that were not found in the mRNA or EST databases (Table 4). Inaccurate 5′-end assignment of mRNA is a known artifact of cDNA library production methods. The study of RPCs circumvents this problem and enables an improved accuracy of 5′-end definition.

HMGA1 was the only HMG gene previously known to have alternatively spliced isoforms, and although only six RPCs were found for this gene, they represented three known spliced isoforms as well as at least one that does not exist in GenBank (Fig. 6). This implies that there are species of HMGA1 cDNAs that are not represented in expression databases, and that a variety of HMGA1 isoforms is expressed in the germ line.

The uneven distribution of RPCs among the HMG genes could be partly explained by differences in expression levels. HMGB1/2 are the most highly expressed followed by HMGN1/2, whereas the least expressed is HMGA1 (Bustin 1999). Also, if we set aside HMGB3 and HMGN3, the three shortest transcripts generated the most RPCs. Both expression levels and transcript size were among the features previously found to distinguish genes that produce RPCs (Goncalves et al. 2000; Venter et al. 2001). However, these features alone do not explain why HMGB2 generated very few RPCs, HMGN1/2 generated more RPCs than HMGB1, and the shortest transcript, HMGN3, only generated one RPC. In a recently published analysis of the ribosomal protein RPCs (Zhang et al. 2002), the authors find an inverse correlation between the GC content of the CDS and the number of RPCs. However, no similar properties seem apparent from our data (see Table 1).

One prerequisite for retrotransposition is expression in germ-line cells, both because in this tissue L1 reverse transcriptase is expressed (Ostertag and Kazazian Jr. 2001), and because germ-line expression is essential for vertical transmission. Studies of a transgenic mouse model showed that L1 expression occurs in late-meiotic and post-meiotic male germ cells (Ostertag et al. 2000). On the other hand, HMGB2 expression in mice testes was found to be restricted to primary spermatocytes, which occur just prior to the first meiotic prophase (Ronfani et al. 2001). This timing difference could explain why there are so few HMGB2 RPCs relative to HMGB1, in spite of the high similarity in structure and function between these two genes. HMGB3 (previously HMG4) has been shown to be expressed in early development (Vaccari et al. 1998). However, from the abundance of HMGB3 RPCs, it is most likely also expressed in germ-line cells. HMGN3 was found to be expressed in testes (West et al. 2001), which does not explain why only one RPC was found for this gene.

Another factor possibly determining the number of RPCs generated is the location of the mRNA in the cell. It has been hypothesized that Alu elements are successful retroposons in spite of their lack of coding regions, because they are derived from the 7SL RNA gene, which is part of the signal recognition particle localized to the endoplasmic reticulum (Boeke 1997). Alu elements can bind proteins of the signal recognition particle and thus be retained in the endoplasmic reticulum, in position to be retroposed by the L1 reverse transcriptase (Mighell et al. 1997).

Potentially Expressed RPCs Shed Light on Genome Evolution

Perhaps one of the most interesting contributions of our work is the discovery of potentially expressed sequences. Two groups were identified as most informative of the novel expressed RPCs. The first group consisted of six RPCs with an ORF similar to that of the corresponding canonical HMG (Table 3; Fig. 4). These could represent potential intronless paralogs of the respective genes. Indeed, the recently discovered HMGN4 gene is comprised mostly of an expressed RPC of HMGN2 (Birger et al. 2001). The second group consisted of three pseudogenes that were included as part of transcripts with a CDS from non-HMG genes (Table 3; Fig. 4). Interestingly, these HMG exons were all part of alternatively spliced transcripts or an altered poly(A) signal usage. In the only published example of an HMG RPC expressed as part of another gene, it was in an alternative exon of the SP100 nuclear antigen splice variant termed SP100-HMG, an HMG-box gene (Rogalla et al. 2000). HMG-box genes, as opposed to the ca-

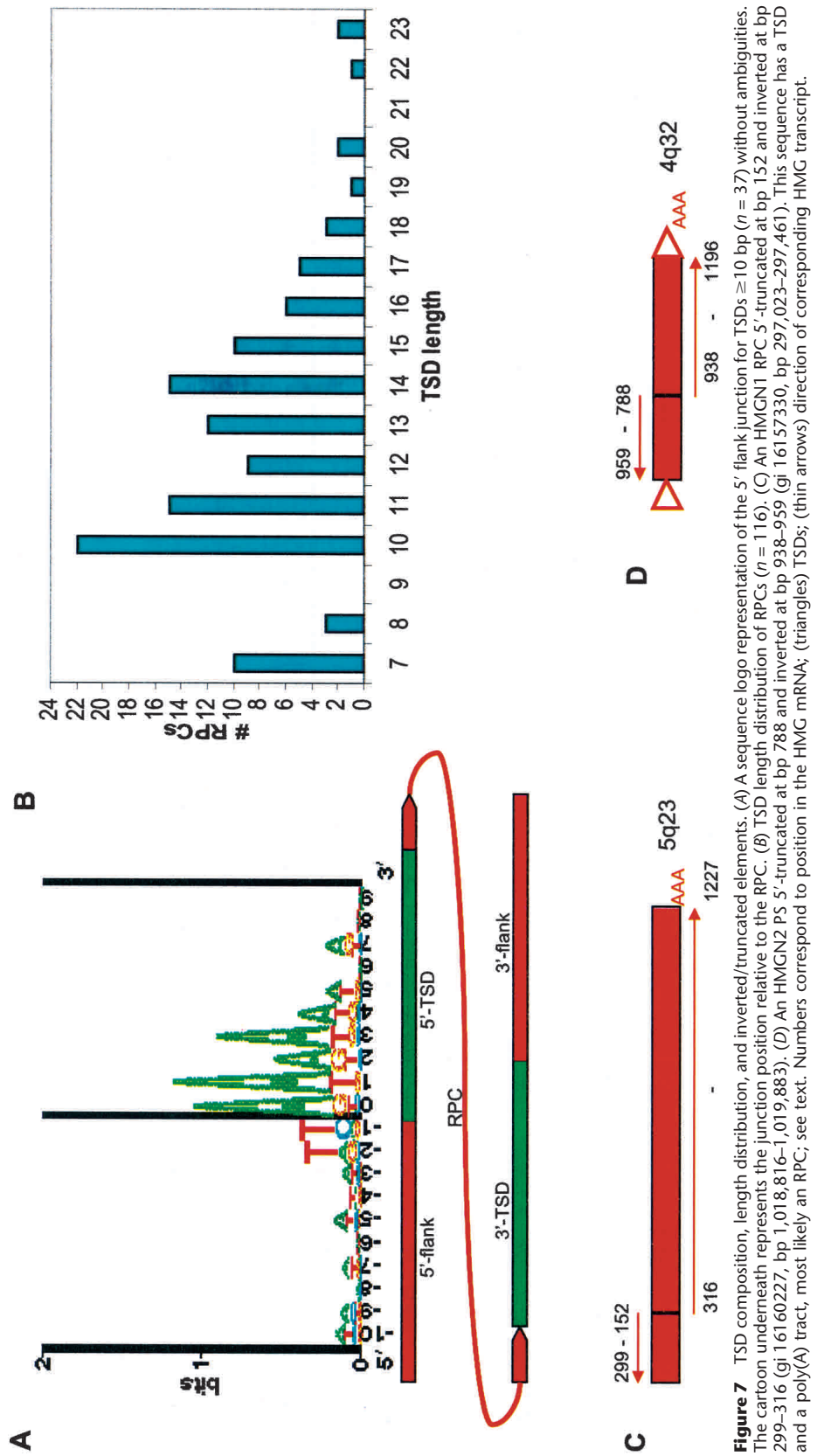


Figure 7 TSD composition, length distribution, and inverted/truncated elements. (*A*) A sequence logo representation of the 5′ flank junction for TSDs ≥10 bp ($n = 37$) without ambiguities. The cartoon underneath represents the junction position relative to the RPC. (*B*) TSD length distribution of RPCs ($n = 116$). (*C*) An HMGN1 RPC 5′-truncated at bp 152 and inverted at bp 299–316 (gi 16160227, bp 1,018,816–1,019,883). (*D*) An HMGN2 PS 5′-truncated at bp 788 and inverted at bp 938–959 (gi 16157330, bp 297,023–297,461). This sequence has a TSD and a poly(A) tract, most likely an RPC; see text. Numbers correspond to position in the HMG mRNA; (triangles) TSDs; (thin arrows) direction of corresponding HMG transcript.

nonical HMG genes, are genes that have other specificities or functions in addition to the HMG-box (Landsman and Bustin 1993). Therefore, retrotransposition could be one mechanism through which the HMG-box proteins are generated. This second group of transcribed RPCs is especially interesting because it can potentially give rise to new genes or domain combinations. The significance of the new exons being part of alternative transcripts may be that insertion into a gene can disrupt its function and be selected against, whereas expression as an alternative transcript preserves the original transcript. Conceivably, the new exon could either be selected for to replace an original exon, or to develop a specialized function as an alternative exon.

Sequence Preference for Integration

We found the chromosomal distribution of RPCs to be only slightly skewed, with none on the Y-chromosome and the highest density on Chromosome 15. The fact that the sequence of the Y-chromosome is only ~60% complete could potentially explain the apparent lack of RPCs; nevertheless, it is clear that there is a paucity of HMG RPCs on this chromosome. From genomic annotations, RPCs in general are scarce on the Y-chromosome, yet this chromosome does harbor both L1 and Alu elements, and L1 has been shown to be expressed in testes (Ostertag et al. 2000). The lack of RPCs on the Y-chromosome could be a phenomenon specific to HMG genes, or a more general one. A possible explanation could be a lack of HMG expression at a crucial stage in the male germline development, causing transmission of HMG RPCs to be female-specific. Another explanation could be the lack of transcripts only in gametes carrying the Y-chromosome, because L1 has been shown to be active in late-meiotic and post-meiotic male germ cells (Ostertag et al. 2000). A similar finding was observed in the recent analysis of ribosomal proteins' RPCs (Zhang et al. 2002), where the overall distribution of RPCs was quite uniform, but the Y-chromosome was also underrepresented. Because RPCs are thought to be generated by the L1 reverse transcriptase, one might expect an increased concentration on the X-chromosome, as L1s are twofold overrepresented on this chromosome (Ostertag and Kazazian Jr. 2001). We do not see a similar overrepresentation of HMG RPCs on the X-chromosome, which could mean that the sequence being retroposed affects the target site, or that HMG RPCs on the sex chromosomes may be selected against, as has been suggested to explain the difference between the distribution of Alu elements and that of L1s (Lander et al. 2001).

RPCs as Genomic Indicators

The human genome assembly and annotation projects encounter many hurdles, some of which stem from the repetitive nature of the human genome, as well as from the difficulty in accurately predicting genes in genomic sequence. Some of our findings will help alleviate these problems by providing independent information regarding potential redundancies, segmental duplications, and gene predictions. Of 256 total entries, 48 shared ≥96% identities with at least one other entry, including the RPC flanking sequence (data not shown). These similarities could be due either to segmental duplications or assembly flaws. Our analysis of expressed sequences yielded several matches to mRNA models of the GenBank accession type "XM_" (data not shown). These RPCs were typically highly similar to the HMG mRNA and still maintained an ORF. However, no ESTs could be found that were more similar to these RPCs than to the original HMGs, providing no real support for their presumed expression. More likely, these were recently created RPCs, and they were identified as model intronless genes because they have not yet lost their ORF owing to random mutations, and still exhibited a high degree of similarity to ESTs that were derived from the HMG genes.

METHODS

RPCs of HMG Genes

A set of eight mRNA sequences (RefSeq were used when possible) representing HMGA1/2, HMGB1–3, and HMGN1–3 (Fig. 1; Table 1) was soft-masked by RepeatMasker (to mask the 6% of low-complexity sequences present; http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) and used as a query in BLASTN searches (Altschul et al. 1990) against two databases: the human genomic division of the nonredundant database in GenBank (ftp://ftp.ncbi.nlm.nih.gov/genbank), and the genome build 26 (10.18.01) chromosome files (ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens). These files contain draft and finished sequence contigs that are part of the RefSeq database (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html) and have been assembled as explained at http://www.ncbi.nlm.nih.gov/genome/guide/build.html.

The results were processed by adding up the local alignment endpoints for each mRNA–genomic location pair to find the full alignment endpoints, orientation, total match length, and overall percent identities, while retaining the local alignment endpoints, and allowing each genomic location only one mRNA match (generally the longest). Alignments under 100 bp were removed from the analysis. Genomic fragments corresponding to HMG parental genes were identified among the genomic hits according to the following criteria: ≥97% identities to mRNA, ≥70% mRNA length, and ≤4 bp overlap or gap in mRNA query sequence between local alignments, corresponding to splice junctions. The identified splice junctions were stored for each mRNA, as well as the number of exons found. As a supplement to the chromosome database, the hits to the nonredundant database were used only to identify HMG parent genes. RPCs were defined in the chromosome files sequences as hits in which the alignment to the mRNA spanned an identified splice junction, as defined above. The rest of the hits were then aligned to the mRNA using BLAST2SEQ (Tatusova and Madden 1999) with lower stringency parameters (-W7 -G3 -E1 -q1 -FF) to find additional RPCs, and the hits not defined as RPCs were termed PSs.

To define the RPC endpoints independently from the mRNA used for the query, each RPC sequence with 1000 or 2000 bp from both the 5′ and 3′ flanks was extracted from the databases and aligned to the previously identified HMG parent genomic sequence with 1000 or 2000 bp of flanking sequence on both sides, using BLAST2SEQ (Tatusova and Madden 1999). The fetched RPC sequences were also aligned with less-stringent parameters to the corresponding HMG mRNA. The combination of endpoints that encompassed the longest genomic sequence was used.

Integration Sites Analysis

Chromosomal localization was determined for RPCs by finding the physical location of the matching NT contig subsequence in the NCBI file seq_contig.md (ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens), then looking in the NCBI file ISCN800_abc for the cytogenetic location of mapped contigs (ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/maps/mapview). The base-pair composition was calculated for 250 bp flanking the RPC on either side. RepeatMasker was used to find repetitive elements in these flanking sequences, as well as

for masking and subsequent use as queries in MEGABLAST searches to look for unique hits in the nonredundant database (Zhang et al. 2000).

Expressed RPCs

RPC sequences were soft-masked and used in BLASTN queries against the human mRNA section of the nonredundant database and in MEGABLAST queries against human ESTs from dbEST. Each mRNA/EST hit of ≥99% identity, EST match length of ≥300 bp, was then compared using BLAST2SEQ to the corresponding HMG mRNA and considered as evidence for RPC expression only if it was ≤97% identical to the HMG mRNA and therefore was not likely to be its transcript.

Expressed RPCs were aligned with the corresponding mRNA/ESTs as well as the HMG mRNA using CLUSTAL W (Thompson et al. 1994). Alignments were edited using Gene Doc (Nicholas et al. 1997; http://www.psc.edu/biomed/genedoc/), which was also used to find ORFs. Splice sites were determined near the ends of a match with an expressed sequence by looking for the consensus GT–intron–AG in immediately adjacent genomic sequence.

Target Site Duplications (TSDs)

For TSDs, 250 bp of sequence flanking the endpoints of RPCs were fetched from each end. Sequences in the 3′ end included the last 10 bp of the alignment, in case the poly(A) tail was included in the mRNA. The flanking sequences were scanned for ≥90% identical matches ≥7 bp long. A poly(A) tract was defined as ≥6 As out of a moving window of 10 bp. We selected all matches found within 30 bp from both the 5′ alignment endpoint and the end of the poly(A) tract, or the 3′ alignment endpoint if no poly(A) tract was found. Of the selected matches, we defined the two matching sequences closest to the alignment ends as the TSD. TSDs ≥10 bp were aligned in two ways: the 5′-most 10 bp from each TSD and 10 bp of the sequence immediately 5′ to it, or the 3′-most 10 bp from each TSD and 10 bp of the sequence immediately 3′, were aligned and analyzed using SeqLogo (Schneider and Stephens 1990; http://www.lecb.ncifcrf.gov/~toms/logoprograms.html). TSDs with ambiguities in this region were eliminated from this analysis (Fig. 7A).

ACKNOWLEDGMENTS

We thank Suzanne Szak and Shlomo Almashanu for many helpful discussions; and Deanna Church, John Anderson, Wataru Fujibuchi, and Steven Sullivan for their comments and help.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Birger, Y., Ito, Y., West, K.L., Landsman, D., and Bustin, M. 2001. HMGN4, a newly discovered nucleosome-binding protein encoded by an intronless gene. *DNA Cell. Biol.* **20:** 257–264.

Boeke, J.D. 1997. LINEs and Alus—The polyA connection. *Nat. Genet.* **16:** 6–7.

Bustin M. 1999. Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins. *Mol. Cell. Biol.* **19:** 5237–5246.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al. 1999. The DNA sequence of human Chromosome 22. *Nature* **402:** 489–495.

Esnault, C., Maestre, J., and Heidmann, T. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* **24:** 363–367.

Feng, Q., Moran, J.V., Kazazian Jr., H.H., and Boeke, J.D. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87:** 905–916.

Friedmann, M., Holth, L.T., Zoghbi, H.Y., and Reeves, R. 1993. Organization, inducible-expression and chromosome localization of the human HMG-I(Y) nonhistone protein gene. *Nucleic Acids. Res.* **21:** 4259–4267.

Goncalves, I., Duret, L., and Mouchiroud, D. 2000. Nature and structure of human genes that generate retropseudogenes. *Genome Res.* **10:** 672–678.

Harrison, P.M., Hegyi, H., Balasubramanian, S., Luscombe, N.M., Bertone, P., Echols, N., Johnson, T., and Gerstein, M. 2002. Molecular fossils in the human genome: Identification and analysis of the pseudogenes in Chromosomes 21 and 22. *Genome Res.* **12:** 272–280.

Hutchison, C.A., Hardies, S.C., Loeb, D.D., Shehee, W.R., and Edgell, M.H. 1989. LINEs and related retroposons: Long interspersed repeated sequences in the eukaryotic genome. In *Mobile DNA* (eds. D.E. Berg and M.M. Howe), pp. 593–617. American Society for Microbiology, Washington, DC.

Johnson, K.R., Lehn, D.A., Elton, T.S., Barr, P.J., and Reeves, R. 1988. Complete murine cDNA sequence, genomic structure, and tissue expression of the high mobility group protein HMG-I(Y). *J. Biol. Chem.* **263:** 18338–18342.

Johnson, K.R., Lehn, D.A., and Reeves, R. 1989. Alternative processing of mRNAs encoding mammalian chromosomal high-mobility-group proteins HMG-I and HMG-Y. *Mol. Cell. Biol.* **9:** 2114–2123.

Johnson, K.R., Cook, S.A., and Davisson, M.T. 1992. Chromosomal localization of the murine gene and two related sequences encoding high-mobility-group I and Y proteins. *Genomics* **12:** 503–509.

Jurka, J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci.* **94:** 1872–1877.

Lander, E.S, Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Landsman, D. and Bustin, M. 1993. A signature for the HMG-1 box DNA-binding proteins. *BioEssays* **15:** 539–546.

Landsman, D., Soares, N., Gonzalez, F.J., and Bustin, M. 1986a. Chromosomal protein HMG-17. Complete human cDNA sequence and evidence for a multigene family. *J. Biol. Chem.* **261:** 7479–7484.

Landsman, D., Srikantha, T., Westermann, R., and Bustin, M. 1986b. Chromosomal protein HMG-14. Complete human cDNA sequence and evidence for a multigene family. *J. Biol. Chem.* **261:** 16082–16086.

Landsman, D., Srikantha, T., and Bustin, M. 1988. Single copy gene for the chicken non-histone chromosomal protein HMG-17. *J. Biol. Chem.* **263:** 3917–3923.

Lum, H.K., Lee, K.D., and Yu, G. 2000. The chicken genome contains no HMG1 retropseudogenes but a functional HMG1 gene with long introns. *Biochim. Biophys. Acta* **1493:** 64–72.

Maestre, J., Tchenio, T., Dhellin, O., and Heidmann, T. 1995. mRNA retroposition in human cells: Processed pseudogene formation. *EMBO J.* **14:** 6333–6338.

Mighell, A.J., Markham, A.F., and Robinson, P.A. 1997. Alu sequences. *FEBS Lett.* **417:** 1–5.

Nicholas, K.B., Nicholas Jr., H.B., and Deerfield II, D.W. 1997. GeneDoc: Analysis and visualization of genetic variation. *EMBNEW.NEWS* **4:** 14.

Ophir, R. and Graur, D. 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* **205:** 191–202.

Ostertag, E.M. and Kazazian Jr., H.H. 2001. Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35:** 501–538.

Ostertag, E.M., DeBerardinis, R.J., Kim, K.S., Gerton, G., and Kazazian Jr., H.H. 2000. Human L1 retrotransposition in germ cells of transgenic mice. *Am. J. Hum. Genet.* **67:** A102 (Abstr.).

Pruitt, K.D. and Maglott, D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137–140.

Rogalla, P., Borda, Z., Meyer-Bolte, K., Tran, K.H., Hauke, S., Nimzyk, R., and Bullerdiek, J. 1998. Mapping and molecular characterization of five HMG1-related DNA sequences. *Cytogenet. Cell. Genet.* **83:** 124–129.

Rogalla, P., Kazmierczak, B., Flohr, A.M., Hauke, S., and Bullerdiek, J. 2000. Back to the roots of a new exon—The molecular archaeology of a SP100 splice variant. *Genomics* **63:** 117–122.

Rogalla, P., Blank, C., Helbig, R., Wosniok, W., and Bullerdiek, J. 2001. Significant correlation between the breakpoints of rare clonal aberrations in benign solid tumors and the assignment of HMGIY retropseudogenes. *Cancer Genet. Cytogenet.* **130:** 51–56.
Ronfani, L., Ferraguti, M., Croci, L., Ovitt, C.E., Scholer, H.R., Consalez, G.G., and Bianchi, M.E. 2001. Reduced fertility and spermatogenesis defects in mice lacking chromosomal protein Hmgb2. *Development* **128:** 1265–1273.
Schneider, T.D. and Stephens, R.M. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* **18:** 6097–6100.
Srikantha, T., Landsman, D., and Bustin, M. 1987. Retropseudogenes for human chromosomal protein HMG-17. *J. Mol. Biol.* **197:** 405–413.
———. 1989. Cloning of the chicken chromosomal protein HMG-14 cDNA reveals a unique protein with a conserved DNA binding domain. *J. Biol. Chem.* **263:** 13500–13503.
Stros, M. and Dixon, G.H. 1993. A retropseudogene for non-histone chromosomal protein HMG-1. *Biochim. Biophys. Acta* **1172:** 231–235.
Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. 2002. Human L1 retrotransposition is associated with genetic instability in vivo. *Cell* **110:** 327–338.
Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., and Boeke, J.D. 2002. Molecular archeology of L1 insertions in the human genome. *Genome Biol.* **3:** research0052.1–0052.18.
Tatusova, T.A. and Madden, L.T. 1999. BLAST 2 sequences—A new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174:** 247–250.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Vaccari, T., Beltrame, M., Ferrari, S., and Bianchi, M.E. 1998. Hmg4, a new member of the Hmg1/2 gene family. *Genomics* **49:** 247–252.
Vanin, E.F. 1985. Processed pseudogenes: Characteristics and evolution. *Annu. Rev. Genet.* **19:** 253–272.
Venter, C.J., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Weiner, A.M., Deininger, P.L., and Efstratiadis, A. 1986. Nonviral retroposons: Genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* **55:** 631–661.
Wen, L., Huang, J.K., Johnson, B.H., and Reeck, G.R. 1989. A human placental DNA that encodes nonhistone chromosomal protein HMG-1. *Nucleic Acids Res.* **17:** 1197–1214.
West, K.L., Ito, Y., Birger, Y., Postnikov, Y., Shirakawa, H., and Bustin, M. 2001. HMGN3a and HMGN3b, two protein isoforms with a tissue-specific expression pattern, expand the cellular repertoire of nucleosome-binding proteins. *J. Biol. Chem.* **276:** 25959–25969.
Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. 2000. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7:** 203–214.
Zhang, Z., Harrison, P., and Gerstein, M. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.* **12:** 1466–1482.

WEB SITE REFERENCES

ftp://ftp.ncbi.nlm.nih.gov/genbank; human genomic division of the nonredundant database in GenBank.
ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens; NCBI human genome build 26 (10.18.01) chromosome files.
ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/maps/mapview; NCBI file ISCN800_abc for the cytogenetic location of mapped contigs.
http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker; RepeatMasker.
http://www.informatics.jax.org/mgihome/nomen/genefamilies/hmgfamily.shtml; HMG chromosomal proteins nomenclature home page.
http://www.lecb.ncifcrf.gov/~toms/logoprograms.html; SeqLogo.
http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html; RefSeq database.
http://www.psc.edu/biomed/genedoc/; GeneDoc.

Received October 9, 2002; accepted in revised form February 25, 2003.