J Mol Evol (2004) 59:620–631
DOI: 10.1007/s00239-004-2654-3

JOURNAL OF MOLECULAR EVOLUTION

Evolution of the Exon–Intron Structure and Alternative Splicing of the MAGE-A Family of Cancer/Testis Antigens

Irena I. Artamonova,[1] Mikhail S. Gelfand[2,3]

[1] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Miklukho-Maklaya 16/10, Moscow 117997, Russia
[2] Institute for Problems of Information Transmission RAS, Bolshoi Karetny per. 19, Moscow 127994, Russia
[3] State Scientific Center GosNIIGenetika, 1st Dorozhny proezd 1, Moscow 117545, Russia

Received: 9 December 2003 / Accepted: 1 July 2004 [*Reviewing Editor:* Dr. Dmitri Petrov]

Abstract. Cancer/testis antigens (CT-antigens) are proteins that are predominantly expressed in cancer and testis and thus are possible targets for immunotherapy. Most of them form large multigene families. The evolution of the MAGE-A family of CT-antigens is characterized by four processes: (1) gene duplications; (2) duplications of the initial exon; (3) point mutations and short insertions/deletions inactivating splicing sites or creating new sites; and (4) deletions removing sites and creating chimeric exons. All this concerns the genomic regions upstream of the coding region, creating a wide diversity of isoforms with different 5′-untranslated regions. Many of these isoforms are gene-specific and have emerged due to point mutations in alternative and constitutive splicing sites. There are also examples of chimeric mRNAs, likely produced by splicing of read-through transcripts. Since there is consistent use of homologous sites for different genes and no random, indiscriminant use of preexisting cryptic sites, it is likely that most observed isoforms are functional, and do not result from relaxed control in transformed cells.

Key words: Multigene family — Alternative splicing — Gene organization — Exon–intron structure — MAGE-A family — CT-antigens

Introduction

Cancer/testis (CT-) antigens are a group of diverse proteins that are predominantly expressed in normal testis and in cancer cells. They form 20 families, of which MAGE-A is one of the largest and historically the first (reviewed by Scanlan et al. 2002). The interest in these genes is caused by their expression pattern, making them a promising target for immunotherapy. Several protein families contain a common domain, and thus the corresponding genes form a larger MAGE (for “melanoma antigen”) superfamily (Chomez et al. 2001).

Initially identified in human melanoma (van der Bruggen et al. 1991), MAGE-A genes were found in a variety of cancers, whereas among normal tissues their expression is largely limited to the germ line cells from testis, ovary, and placenta. Other CT-antigens, including non-MAGE ones, were identified by a number of experimental techniques including SEREX (serological expression cloning), differential mRNA expression analysis, and cytotoxic T lymphocyte epitope cloning, as well as mining EST databases for genes with cancer/testis expression patterns (Chomez et al. 2001; Scanlan et al. 2002). Genes from MAGE-A, MAGE-B, and MAGE-C families are expressed in germ line and cancer cells, whereas genes from the remaining MAGE families, most notably MAGE-D (Lucas et al. 1999; Pold et al. 1999) and NECDIN/MAGE-L2 (Jay et al. 1997; Boccaccio et al. 1999), are expressed ubiquitously. The latter are candidate genes for the Prader–Willi syndrome (Jay et al. 1997; Boccaccio et al. 1999).

Correspondence to: Mikhail S. Gelfand; *email:* gelfand@iitp.ru

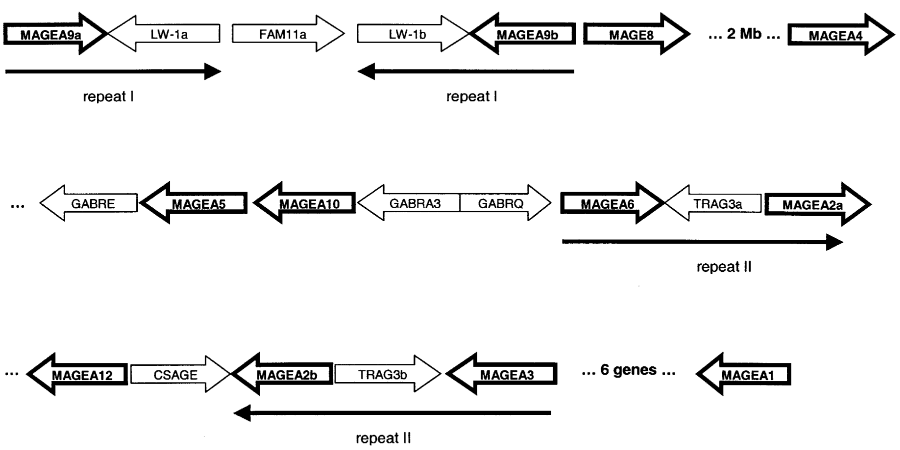


Fig. 1. Genomic organization of the MAGE-A locus. Boldface arrows: MAGE-A genes. Long arrows: inverted repeats.

All MAGE-A genes have one protein-coding exon, preceded by several noncoding exons (De Plaen et al. 1994; Rogner et al. 1995). The same holds for several other MAGE families (MAGE-B, MAGE-G, MAGE-H), whereas coding regions of genes from the MAGE-C and MAGE-D families are interrupted by introns (Lucas et al. 2000; Chomez et al. 2001). Some of these genes are known to be alternatively spliced with isoforms having different 3′-untranslated regions or even coding regions (De Plaen et al. 1997; Lucas et al. 2000).

Most MAGE genes map to the human chromosome X, with the largest families MAGE-A, forming a locus at q28 (De Plaen et al. 1994; Rogner et al. 1995), MAGE-B, at p21–p22 (Muscatelli et al. 1995; Dabovic et al. 1995), and MAGE-C, at q26–q27 (Lucas et al. 1998), although MAGE-F1 is encoded on chromosome 3 (Stone et al. 2001).

Based on the fact that most MAGE families contain no introns in protein-coding regions and show very narrow expression pattern, whereas MAGE-D genes contain numerous introns, are expressed ubiquitiously, and have nonmammalian orthologs, it was suggested that MAGE genes had been formed by retroposition of the ancestral MAGE-D gene (Chomez et al. 2001). If this is correct, the insertion of introns in the upstream region and appearance of alternative splicing of MAGE-A genes should have happened after the founder MAGE-A gene was introduced into the genome. The orthologous mouse family MAGE-A consisting of seven active genes maps to two loci at mouse chromosome X (De Plaen et al. 1999; Chomez et al. 2001). Like their human counterparts, these genes are transcribed in cancer cell lines and in testis. Human MAGE-A and mouse Mage-a proteins form two separate branches on the tree of all MAGE proteins (Chomez et al. 2001; Cannon and Young 2003), and thus it is likely that the multiplication of the ancestral gene occurred independently in these two genomes.

Alternative splicing was recently established as one of the main mechanisms of generating protein diversity in multicellular eukaryotes, and now at least half of human genes are believed to be alternatively spliced (Mironov et al. 1999; Brett et al. 2002; for a review see Modrek and Lee 2002). Moreover, recent comparisons of the human and mouse genomes demonstrated that about half of alternatively spliced genes have genome-specific isoforms (Modrek and Lee 2003; Nurtdinov et al. 2003; Thanaraj et al. 2003), although the functionality of these isoforms was questioned (Kan et al. 2002; Sorek et al. 2004). Creation of alternatively spliced exons is often associated with exon duplication (Kondrashov and Koonin 2001; Letunic et al. 2002) or Alu insertion (Sorek et al. 2002). It has been demonstrated that point mutations can influence the choice of acceptor splicing sites and the ratio of alternatively spliced isoforms (Lev-Maor et al. 2003). On the other hand, splicing-affecting mutations may account for at least 15% of human genetic diseases, and likely even more (Krawszak et al. 1992; Nakai and Sakamoto 1994; Faustino and Kooper 2003). However, evolution of splicing patterns within genomes has not been studied.

In this study we analyze alternative splicing of the MAGE-A genes using mapping of available ESTs to the genomic sequence. We demonstrate the existence of gene-specific isoforms and study the influence of point mutations on constitutive and alternative splicing. To our knowledge, this is the first attempt to reconstruct the evolution of alternative splicing in a family of recently duplicated paralogs. As such, it provides an additional level of resolution to large-scale comparative analyses of alternative splicing in human and mouse genomes.

Methods

Genome and EST sequences were taken from the Human Genome Browser (Karolchik et al. 2003). The human genome assembly of April 2003 (UCSC version hg15) was used. The BLAT-generated alignments from the Human Genome Browser were additionally

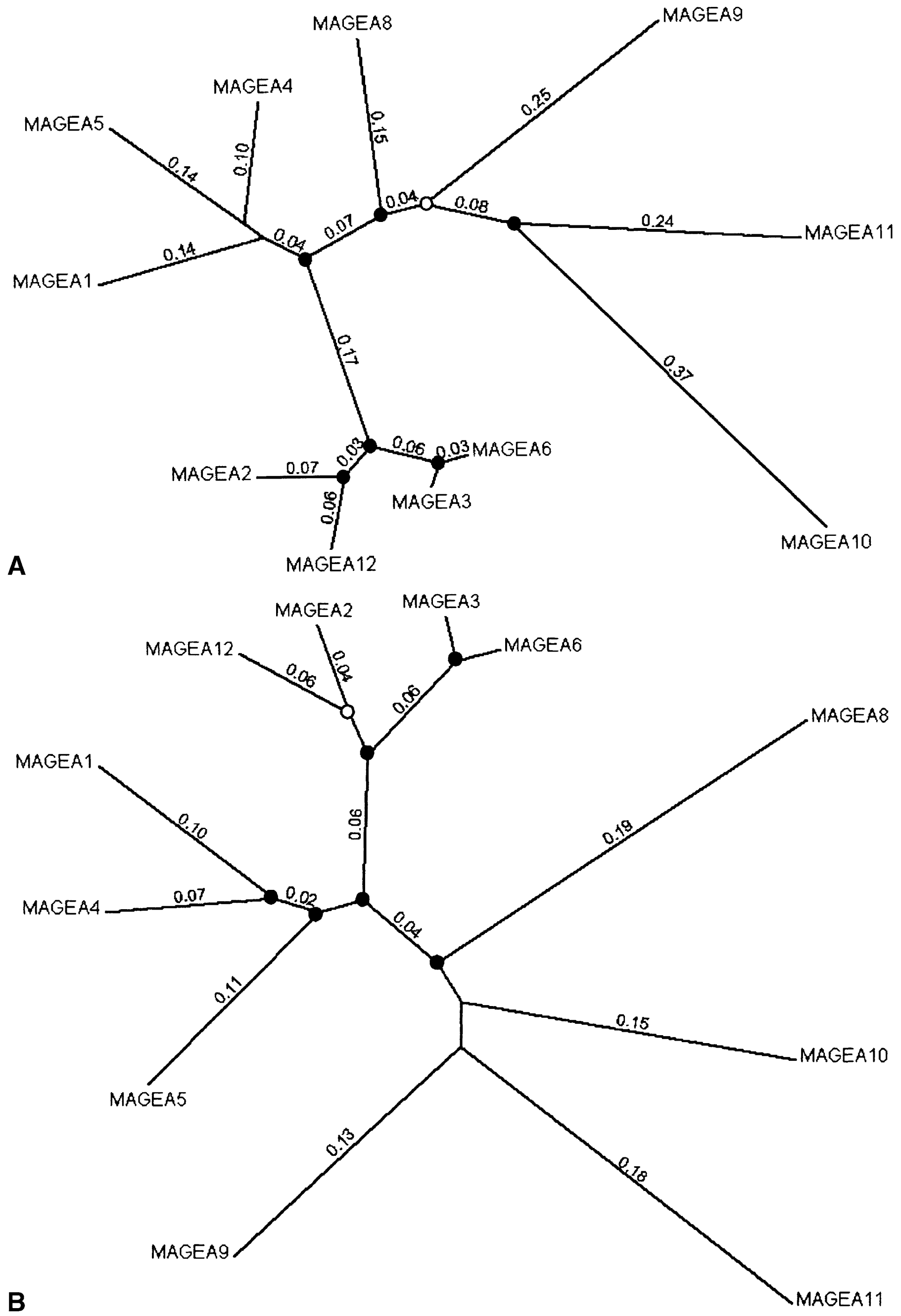


Fig. 2. Phylogenetic trees of the MAGE-A family. **a** Proteins. **b** Aligned ~1800-nucleotide regions upstream of the protein-coding regions of MAGE-A genes. Bootstrap support is shown by filled (>95%) and open (>75%) circles; unmarked nodes have support <55%. The units of branch length (shown by numbers) are the expected fraction of amino acids changed (for proteins, 0.01 is 1 PAM; shown if exceeding 0.02) and the expected nucleotide substitution per site (for DNA alignments, shown if exceeding 0.01).

verified using EST-to-genome alignment by Pro-EST (Mironov et al. 1999). Gene expression data were from Su et al. (2002), obtained via the Human Genome Browser.

Multiple alignment of protein and nucleic acid sequences was done using CLUSTALW with default parameters (Thompson et al. 1994). Phylogenetic trees were constructed using the maximum likelihood algorithm implemented in PHYLIP with default parameters (Felsenstein 1996). The trees were plotted using GeneMaster (Andrey A. Mironov, unpublished).

Alignment regions corresponding to splicing sites were analyzed manually. Mutations changing consensus nucleotides (c/a)AG/GTRAG in donor sites and polyY-NCAG/G in acceptor sites, as well as mutations creating AG dinucleotides upstream of acceptor sites, were considered as weakening the splicing sites, and mutations changing the invariant dinucleotides GT and AG in donor and acceptor sites, respectively, as completely disrupting splicing (Gelfand 1989; Iida 1990; Stenson et al. 2003). Conversely, mutations creating GT and AG dinucleotides or making a site closer to the consensus were considered as increasing the likelihood of splicing.

Results

Genes

The locus containing MAGE-A genes spans about 3 Mb of chromosome X (band Xq28, Fig. 1). There are two recent duplications: inverted duplication of MAGE-A9 and the adjacent region, forming sublo-

Table 1. Tissue specificity of MAGE-A gene expression

	MAGE-A									
Tissue (organ)	8	9	11	4	1	5	3	6	2	12
Testis			1.5	3	3	2.5	3	3	3	3
Chronic myelogenous leukemia (K562)				2	3	3	3	3	3	3
Thymus		1.5	1.5	1.5	1.5	1.5	2.5	2.5	1.5	2.5
Placenta	2		1.5	2	2		2.5	2.5	1.5	
Ovary				2	1.5	1.5	1.5	2	2	1.5
Pancreas	1.5		1.5				1.5	1.5	1.5	2
Brain (fetal, cortex, amygdala, etc.)	1.5		1.5				2	2	1.5	
Umbilical vein endothelium (HUVEC)							1.5	1.5		2.5
Cell line 10N							2.5	2.5		
Uterus				1.5	1.5	1.5				
Burkitt's lymphoma (DAUDI)									1.5	2.5
Prostate cancer? (PC4, PC6, PC8)							2	2		
Acute lymphoblastic leukemia (MOLT4)							2	2		
Salivary gland		1.5								1.5
Lung							1.5	1.5		
Heart									2	
Spleen										2

Note. The data are taken from http://web.gnf.org/ via the Human Genome Browser. Only expression levels exceeding the average by 1.5 standard deviations or more are shown.

cus I, and inverted duplication of the region between MAGE-A3/6 and MAGE-A2, forming sublocus II. In the second case the duplicated genes MAGE-A3 and MAGE-A6 are very similar (99% identity on the nucleotide level and 95% protein identity) but still less similar than the MAGE-A9 and MAGE-A2 (having only 8 substitutions of aligned 5800 nucleotides and 2 substitutions of 4000 aligned nucleotides, respectively). Following the standard nomenclature (Human Genome Browser; Chomez et al. 2001; Scanlan et al. 2002), we retain different names for MAGE-A3 and MAGE-A6 and consider them separately, whereas variants of MAGE-A2 and MAGE-A9, denoted "a" and "b," are not distinguished below.

The phylogenetic trees of the proteins (Fig. 2a) and of the nucleotide sequences immediately upstream of the coding regions (Fig. 2b) are similar. There are three branches in each tree. The first branch (called *subfamily I*) contains MAGE-A8, MAGE-A9, MAGE-A10, and MAGE-A11. The second one (*subfamily II*) contains two pairs, MAGE-A3/MAGE-A6 and MAGE-A2/MAGE-A12 (the former pair results from a recent duplication, whereas the latter pair is more diverged, with 91% nucleotide and 87% protein identity). The definition of these subfamilies is also supported by common duplication/insertion events (see below). The third branch, containing the three remaining genes, MAGE-A1, MAGE-A4, and MAGE-A5, is weakly supported by the bootstrap analysis.

These trees are slightly different from the tree of Chomez et al. (2001), where subfamily I does not form a cluster, MAGE-A8 is clustered with MAGE-A4 and MAGE-A1, and MAGE-A5 is missing altogether. The remaining differences can be explained by placing the root of the tree of Chomez et al. (2001) within our subfamily I.

The order of genes in the MAGE-A locus weakly correlates with the phylogenetic trees. MAGE-A12 is located between two copies of MAGE-A2, which in turn are framed by MAGE-A3 and MAGE-A6 (Fig. 1). Similarly, MAGE-A11 is located between two copies of MAGE-A9, and MAGE-A8 is immediately downstream of the duplicated area; this sublocus is separated by about 2 Mb from the second sublocus containing the remaining genes (including MAGE-A10 from the same subfamily).

The coding region of each gene is contained in the last exon, preceded by several alternatively spliced untranslated exons. The coding region of MAGE-A5 is interrupted by a premature stop codon that results from a CGA → TAA mutation and likely is a recently inactivated, transcribed pseudogene (or encodes a shorter protein of 120 amino acids). MAGE-A7 seems to be a ghost or a pseudogene, as it could not be found either in GenBank or in the UCSC Human Genome Browser; indeed, initially it had been reported to be nontranscribed (De Plaen et al. 1994). MAGE-A10 is extended by about 30 amino acids at the C terminus, whereas close to the N terminus, it contains a serine-rich insert of about 20 amino acids. The length of the remaining proteins is approximately 300 amino acids.

Expression array analysis showed that MAGE-A1, -A2, -A3, -A4, -A5, -A6, and -A12 were highly expressed in testis and cancer and, to a lesser extent, in thymus, placenta, and ovary, followed by pancreas and brain (Table 1) (Su et al. 2002). Expression of the

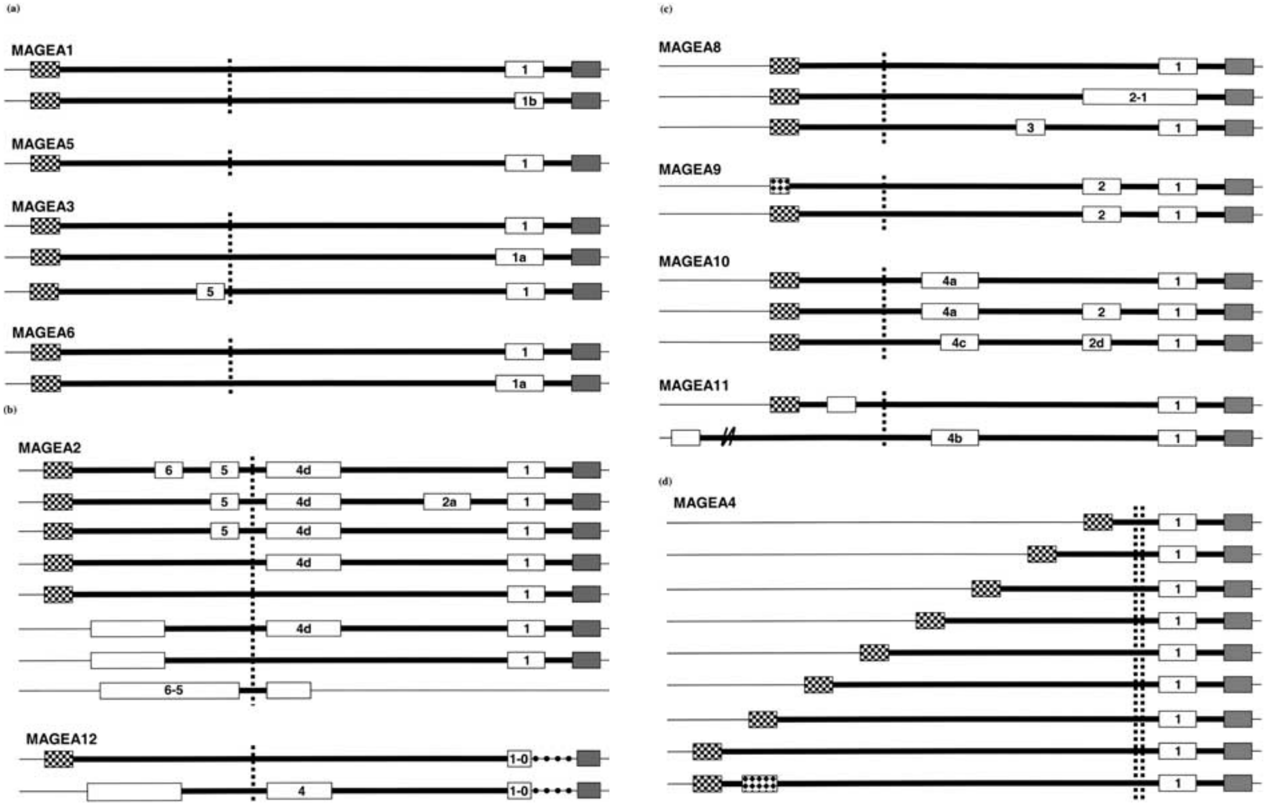


Fig. 3. Schematic representation of the exon–intron structure of the MAGE-A genes. Boxes: exons. Thick lines: introns. Gray boxes: protein-coding exon 0. Checkered boxes: homologous initial exons. Dotted vertical line: boundary of the well-alignable region. The major isoform supported by multiple ESTs is the first one in all cases excluding MAGE-A2 (no predominant isoform) and MAGE-A4 (two main isoforms, the second and fourth ones from the top; see the text). **a** Simple cases. **b** Remaining genes of subfamily II (dotted horizontal line in MAGE-A12: deletion removing parts of exons 1 and 0 and the intron between these exons). **c** Subfamily I. **d** Multiplication of the initial exon in MAGE-A4 (double-dotted vertical line in MAGE-A4: the area between the initial exon and coding exon 0 contains no exons and is not shown).

subfamily I genes was weak in all tissues. On the other hand, RT-PCR demonstrated that, in addition to cancer and testis, MAGE-A3, -A4, -A8, -A9, -A10, -A11, and -A12 were expressed in placenta, whereas MAGE-A1, -A2, -A6, and -A12 were not (De Plaen et al. 1994).

Analysis of EST data demonstrated that most genes have one predominant isoform and several isoforms supported by only one or two ESTs. The main isoform is always shown as the first one in Fig. 3, with two exceptions: There is no predominant isoform of MAGE-A2, where all isoforms are supported by one or two ESTs, and there are two main isoforms of MAGE-A4, corresponding to the second and the fourth variants of the initial exon (counting from the right). Most ESTs were derived from cancer cell line libraries with the following exceptions: All MAGE-A5 ESTs were from normal placenta; the main MAGE-A6 isoform and two major MAGE-A4 isoforms were observed in cancer cell lines as well as normal testis and brain (medulla); the minor MAGE-A9 isoform was observed in testis; and, finally, multiple ESTs derived from the MAGE-A4 coding exon common to all isoforms were seen in placenta.

About 1800 bp upstream of the coding region could be aligned in all representatives of the family. This region contains four groups of noncoding exons with slightly different or alternative splicing sites (Fig. 3). This region also contains several deletions or insertions that are specific for branches of the phylogenetic trees and thus support our definitions of subfamilies: a deletion in MAGE-A3 and MAGE-A6 and two likely insertions, a long one common to subfamily I and a short one common to subfamily II. One more deletion is specific to MAGE-A12. All deletions and/or insertions are flanked by short (4–7 nt) direct or inverted repeats. By the way, an alignment excluding the deleted/inserted regions produces the same phylogenetic tree as the long, complete alignment.

Upstream of this region, genes of subfamily I cannot be aligned with the remaining genes, and farther upstream there are multiple duplications of the starting exon in most genes (see below) and thus alignment makes no sense.

Noncoding exons of human MAGE-A genes could not be aligned with noncoding exons of mouse Mage-a genes (data not shown). Indeed, as mentioned in the Introduction, the human MAGE-A genes were

(a) Main acceptor site and alternative site 1a in MAGE-A1.

```
Ancestral:  GCC--CTCTCACTTCCTCCTTCAGGTTCTsAGGGG-ACAGGCTGAC
MAGE-A3     GCC--CTCTCACTTCCTCCTTCAGGTTCTGAGGGG-ACAGGCTGAC
MAGE-A6     GCC--CTCTCACTTCCTCCTTCAGGTTCTGAGGGG-ACAGGCTGAC
MAGE-A2     GCC--CTCCCACTTCCTCCTTCAGGTTCTGAGGGGGACAGGCTGAC
MAGE-A12    GCC--CTCTCACTTGTTCCTTCAGGTTCTGAGGAG-ACAGGC----
MAGE-A1     ACC--CTCTCACTTCCTCCTTCAGGTTTTCAGGGG-ACAGGCCAAC
MAGE-A4     GCT--CTCTCACTTCCTCCTTCAGGTTCTGAGCAG-ACAGGCCAAC
MAGE-A5     GCC--CTCTCACTTTTTCCTTCAGGTTCTCAGGGG-ACAGGCTGAC
MAGE-A8     GCC--CTCTCAATTTCTCCTTCAGGTTCGCAGAGA-ACAGGCCAGC
MAGE-A9     GCC--CTCTCACTTCCTCCCTCAGGTTCTC--GGG-ACAGGCTAAC
MAGE-A10    GCCATCTCTCACTTCCTTCTTCAGGTTCTCAGGGG-ACAGGGAGAG
MAGE-A11    GCC--TTCTCACTTCCTTTTTCAGGTTTTTAGAGA-ACAGGCCAAC
```

(b) Acceptor site 1b in MAGE-A3 and MAGE-A6.

```
Ancestral:  TGGCCTCACCCTCCCTACTGTCAGTCCTGCAGCC
MAGE-A3     TGGCCTCACCCTCAATACTTTCAGTCCTGCAGCC
MAGE-A6     TGGCCTCACCCTCAATACTTTCAGTCCTGCAGCC
MAGE-A2     TGGCCTCACCCTCCCTATTCTCAGTCCTGCAGCC
MAGE-A12    TGGCCTCACCCTCCCTACTGTCATTCCTTCAGCC
MAGE-A1     TGGCCTCACC-TCCCTACTGTCAGTCCTGTAGAA
MAGE-A4     TGGCCTCACC-TCCCTACCATCAATCCTGCAGAA
MAGE-A5     TGGCCTCACCCTCTCTACTGTCAGTCCTGCAGAA
MAGE-A8     TGGCCCTACCAGCCCTTTTGTCAGTCCTGGAGCC
MAGE-A9     TGGCCCCATTCCCCCTACTGTCACTCATAGAGCC
MAGE-A10    CAGCTCCACCTACCCTACTGTCAGTCCTGGAGCC
MAGE-A11    CAACCTCACCTGCCCTACC-TCAGTCCTGGAGCC
```

Fig. 4. Alternative and constitutive acceptor sites of exon 1. Underlined: exons. Wavy underlined: alternative regions (belonging to the intron or exon dependent on the choice of an alternative site). Upper line: deduced ancestral sequence. Boldface: nucleotides, conforming to the consensus splicing signals. Italics: nucleotides not conforming to the consensus splicing signals. Boldface italics: nucleotides that are consensus for one site and nonconsensus for the other site (only for observed sites and ancestral sequences). Shaded: mutations that could have changed site functionality (see text).

duplicated after the divergence of the human and mouse genomes. Thus the mouse genes were not considered in this study.

Exons and Splicing Sites

As there is no universal correspondence between the upstream exons for different genes, for consistency the exon groups are numbered in the 3′-to-5′ order, thus the only coding exon is numbered 0, the preceding exon is numbered 1, etc. The region well alignable throughout the MAGE-A family contains four groups of noncoding exons and the coding exon, by the above convention numbered 4 through 0 (if counting in the standard 5′-to-3′ direction). Not all exons are present in all genes; there are also alternative sites (specified by lowercase letters, e.g., exon 1a, acceptor site 2b, etc.). The isoforms are shown in Fig. 3 and discussed in detail below.

Exon E1. Exon E1 is present in all isoforms of all genes, excluding MAGE-A12, where two deletions remove the intron and the adjacent parts of exons 1 and 0, creating a chimeric exon, 1–0. There are alternative acceptor sites: a site upstream of the main one in MAGE-A3 and MAGE-A6 and a site downstream of the main one in MAGE-A1. The latter is only 8 nt downstream of the main site and, thus, has a poor splicing context (interfering AG immediately upstream). Indeed, it is used in only one EST of five MAGE-A1 ESTs, and the corresponding sites in other genes are inactive, although in some genes the sequence is almost the same (Fig. 4a). The upstream site in MAGE-A3 and MAGE-A6 has been created by an activating mutation G → T at position −5 that has enhanced the polypyrimidine tract. Inactive sites in other genes have G or C there. Farther upstream, these genes have AA dinucleotides that weaken the polypyrimidine tract but, on the other hand, might serve as a branch point.

Cassette Exon E2. Cassette exon E2 was observed in MAGE-A2, MAGE-A9, MAGE-A10, and MAGE-A8, although potentially it could be incorporated into mRNA in some other genes as well, as the splicing sites are conserved. The only gene where this exon is included in the major isoform is MAGE-A9.

It has two alternative donor sites, the upstream one and the downstream one. Neither site is used in MAGE-A8, and in one isoform the intron between exon 2 and exon 1 is retained, producing a long exon, 2-1. Both sites are used in MAGE-A10, where the use of the upstream site leads to an exon that is only 19 nt long. This site is activated by mutation GCA → GAG in the exon positions of the site, making it closer to the consensus (Fig. 5).

The history of the acceptor sites of exon 2 is somewhat complicated. Exons denoted 2a and 2 are not really alternative. In genes from subfamily II, a 7-nt insertion flanked by a direct repeat CAG/GA (overlapping with the intron–exon junction) created two possible sites (Fig. 5). In MAGE-A2 the upstream site is used, as the downstream site is in a weak context, namely, an upstream AG, G at position −3, and a weak polypyrimidine tract.

Cassette Exon E3. Cassette exon E3 is specific to MAGE-A8, where it was observed in only one EST. The corresponding region has been created by a long insertion in subfamily I. The insertion is flanked by a direct repeat of TGAGGAC. It contains the donor site of exon E3. The corresponding positions in other members of the family do not contain the GT dinucleotide, which is destroyed by point mutations or short deletions (not shown). The acceptor site in MAGE-A8 corresponds to a well-alignable region, and it has been created by a short gene-specific insertion that created the AG dinucleotide and the polypyrimidine tract (Fig. 6).

Cassette Exon E4. Cassette exon E4 was observed in a pair of genes from subfamily I, MAGE-A10 and MAGE-A11, and a pair of genes from subfamily II,

```
Subfamily II: GGCCCGTGCATTCCTCTCCCAGGAATCAGGAGCTCCAGGAACAAGGCAGTGAGGCCTTGGTCTG
MAGE-A3       GGCCCGTGGATTCCTCTCCCAGGAATCAGGAGCTCCAGGAACAAGGCAGTGAGGACTTGGTCTG
MAGE-A6       GGCCCGTGGATTCCTCTCCCAGGAATCAGGAGCTCCAGGAACAAGGCAGTGAGGACTTGGTCTG
MAGE-A2       GGCCCGTGCATTCCTCTCCCAGGAATCAGGAGCTCCAGGAACCAGGCAGTGAGGCCTTGGTCTG
MAGE-A12      GGCCCATGGATTCCTCTCCCAGGAATCAGGAGTTCCAAGAACAAGGCAGTGAGGCCTTGGTCTG

Other genes:  GGCCCCTCGATTCCTCTTCCAG-------GAGCTCCAGGAACCAGGCAGTGAGGCCTTGGTCTG
MAGE-A1       GGCCCGTGGATTCCTCTTCCTG-------GAGCTCCAGGAACCAGGCAGTGAGGCCTTGGTCTG
MAGE-A4       GGCCCATGGATTCCTCTCCTAG-------GAGCTCCAGGAACAAGGCAGTGAGGCCTTGGTCTG
MAGE-A5       GGCCCATGCATTCCTCTTCCAG-------GAGCTCCAGGAAACAGACACTGAGGCCTTGGTCTG
MAGE-A8       TTCCCCTCGATTTATCTTCTAG-------GAGCTCCAGGAACCAGGCTGTGAGGTCTTGGTCTG
MAGE-A9       GGCCCCTCGATTCCTCTTCCAG-------GAGCTCCAGGAA----GCAGGCAGGCCTTGGTCTG
MAGE-A10      GGCCCCTCAATTCCTCTTTCAG-------GAGCTCCAGGGACTGTGAGGTGAGGCCTTGGTCTA
MAGE-A11      GGCCCCATGATTCCTCTTCTAG-------AAGCTCCAAAAACTGAGCAGTGAGGCCTTGGTCTC
```

Fig. 5. Splicing sites of exon 2. Acceptor sites and the downstream donor site are shown. Double-underlined: cassette exons. Other notation as in Fig. 4.

```
MAGE-A3     TTGAGGGTACC-----------CCTGGGA---CAGAATGCGGA
MAGE-A6     TTGAGGGTACC-----------CCTGGGA---CAGAATGCGGA
MAGE-A2     TTGAGGGTACT-----------CCTGGGC---CAGAATGCAGA
MAGE-A12    TTGAGGGTACC-----------CCTGGGC---CAGAACGCTGA
MAGE-A1     CTGAGGGTACC-----------CCAGGAC---CAGAACACTGA
MAGE-A4     TTGAGGGTACC-----------ACAGGGC---CAGAACGCAGA
MAGE-A5     TTGAGGGCACC-----------CTTGGGC---CAGAACACAGA
MAGE-A8     TTGAGGGTACCCTCGATGGTTCTCCTAGCAGGCAAAAAACAGA
MAGE-A9     TCGAGGGTACC-----------TCCAGGC---CAGAGAAACTC
MAGE-A10    CTGAGGGTACC-----------CCCAGCC---CATAACACAGA
MAGE-A11    TTGAGGGTTCC-----------TCCTGGC---CAGAACACAGA
```

Fig. 6. Donor splicing site of exon 3 in MAGE-A8. Notation as in Figs. 4 and 5.

```
(a)

Ancestral: GCCAAGmGTCCAGGTGAGGAACCGGAGGGAGGATTGAGGGTACC
MAGE-A3    GCCAAGCATCCAGGTGAAGAGACTGAGGGAGGATTGAGGGTACC
MAGE-A6    GCCAAGCATCCAGGTGAAGAGACTGAGGGAGGATTGAGGGTACC
MAGE-A2    GCCAAGCATCCAGGTGGAGAGCCTGAGGTAGGATTGAGGGTACT
MAGE-A12   ACCAAGCATCCAGGTGAGAAGCCTGAGGTAGGATTGAGGGTACC
MAGE-A1    GCCATGCGTTCGGGTGAGGAACATGAGGGAGGACTGAGGGTACC
MAGE-A4    GCCAAGAGTCCTGGTGAGGAATGTGAGGGAGGATTGAGGGTACC
MAGE-A5    GTCAGTAGTTCCGGTGAGGAACATGAGGGACGATTGAGGGCACC
MAGE-A8    ACCAAGAGTCTAGGTGACAACACTGAGGGAAGATTGAGGGTACC
MAGE-A9    GAGAGCAGTCCAGGTGAGGAACCTAAGGGAGGATCGAGGGTACC
MAGE-A10   GACAAGAGTCCAGGTAAGGAACCTGAGGGAAATCTGAGGGTACC
MAGE-A11   GCCAAGAGTCCAGGTGAGAAACCTGAGGGAGGATTGAGGGTTCC

(b)

Ancestral: CTCCTCCTCACTTCTGTTTCCAGATCTCAGGGAGGTGAGGACCT
MAGE-A2    CCCCTCCTCACTTCTGTTTCCAGATCTCAGGGAGTTGATGACCT
MAGE-A12   CTCCTCCTCACTTCTGTTTCCAGATCTCAGGGAGTTGAGGACCT
MAGE-A1    CTTCTTTTCACTCCTGTTTCCAGATCTGGGGCAGGTGAGGACCT
MAGE-A4    CTCCTTCTCATTTCTGATTCCAGATCTCAGTGAGGTGAGGACCT
MAGE-A5    CTCCATCTCATTCCTGTTTTCAGATCTCGGGGAGGTGAGGAACT
MAGE-A8    TGGCTCCTCATTTCTCTCTTGAGATCTCAGGGAAGTGAGGACCT
MAGE-A9    GGCCTCCTCACCTCTGTTTCTGGATCTCAGGGAGGTGAGGACTT
MAGE-A10   CTCCTTCTTACTTTTGTTTTGGAATCTCAGGGAGGTGAGAACCT
MAGE-A11   GGCCC-CTTACTTCTGTTTTGGAATCTTGGGCAGGTGAGCACTA

(c)

Ancestral: AGGGGCCCCCATGTGGTCGACAGACACAGTGG
MAGE-A3    AGGGGCCCCTATGTGGTGGACAGATGCAGTGG
MAGE-A6    AGGGGCCCCTATGTGGTGGACAGATGCAGTGG
MAGE-A2    AGGGGCCCCCATCTGGTCGACAGATGCAGTGG
MAGE-A12   AGGGGCCCCCATGTAGTCGACAGACACAGTGG
MAGE-A1    AGGGACCCCCATCTGGTCTAAAGACAGAGCGG
MAGE-A4    AGGGACCCCCATCTGGTCTACAGACACAGTGG
MAGE-A5    AGGGGCCCCCATCTGGTGGATAGACAGAGTGG
MAGE-A8    AGGGACCCCCATGTGGGCAACAGACTCAGTGG
MAGE-A9    AGGGAGGCCC-TGTGTTCGACAGACACAGTGG
MAGE-A10   AGGGAACCCC-TCTTTCTACAGACACAGTGG
MAGE-A11   AAAGAGCCCCATATGGTCCACAACTACAGTGG
```

Fig. 7. Donor and acceptor splicing sites of exon 4. **a** Donor sites. **b**, **c** Acceptor sites. Notation as in Figs. 4 and 5.

MAGE-A2 and MAGE-A12, although potentially it could exist also at least in MAGE-A1 (despite a slightly weaker donor site) and MAGE-A9 (which is basically indistinguishable in the site regions from MAGE-A10); on the other hand, the acceptor site region is covered by a long deletion in MAGE-A3 and MAGE-A6, and this deletion is flanked by an inverted repeat CCCCT–AGGGG. In MAGE-A12 it was observed in only one EST, whereas in the other three genes it was included in the majority of ESTs.

The dynamics of the site choice in this exon is rather clear. In MAGE-A2, the ancestral (upstream) donor site was weakened by mutations AG → GA at positions 4–5, and this led to the use of the downstream site that extends the exon by 14 nt; the latter was created by a G → T mutation that produced the canonical GT dinucleotide (Fig. 7a). The relative timing of these events is unknown, but not very important in this context. The use of very close acceptor sites in MAGE-A2, MAGE-A12, MAGE-A10, and MAGE-A11 (exons 4, 4a, 4b) can be explained by a simple rule: The upstream AG is used, and if it is inactivated by a point mutation, the next one is used; in MAGE-A11 the downstream acceptor site is enhanced by a G → C mutation in position −3 (Fig. 7b). Finally, a MAGE-A10-specific alternative acceptor site (exon 4c) was created by a series of upstream transversions and deletions that removed upstream purines and thus created a perfect polypyrimidine tract (Fig. 7c).

Initial Exons. In region 5′ of exon 4, subfamily I and the remaining genes need to be aligned separately. There is also no correspondence between exons of different genes. However, the main initial exons are homologous (Table 2). In MAGE-A2, MAGE-A12, MAGE-A8, and MAGE-A10 these exons are unique, whereas in the remaining genes the corresponding region was duplicated several times. The most dramatic expansion of this region happened in MAGE-A4 (Fig. 3d), where this region forms a tandem repeat of nine copies; in addition, there are six more partial copies. The core of the repeated region is about 100 nt, covering the donor site of the initial exon. The region corresponding to

Table 2. Region of the donor site of the main initial exon

(1) Gene	(2) No.	(3) Position	(4) Repeat from	(5) Repeat to	(6) Donor site consensus, aagGTgaga
MAGE-A8	1*	−2963	−164	+119	agggtctgtgaggaggcaaggtgaga
MAGE-A9	1	−3187	−167	+115	aggctctgtgaggaggcaagatgaga
	2	−3627	−167	+119	aggctctgtgaggaggcaaggtgaga
	3*	−4186	−165	+100	aggctctgtgaggagtcaaggtgagg
MAGE-A10	1*	−2856	−154	+118	aggctctgtgaggaggcaaggtgaga
MAGE-A11	1	−3222	−88	+120	aggctctgtgaggaggcaagatgaga
	2*	−3656	−167	+115	aggctccatgaggaggcaaggtgaga
MAGE-A4	1	−2319	−59	+119	aagctctgtggcgaggcaaggtgaga
	2	−2812	−166	+47	aggctttgtgaagaggcaaagtgaga
	3	−3379	−172	+199	aggctctgtgaggtggcatagtgaga
	4	−3943	−172	+265	aggccctgtgaggagtcaaggtgaga
Init. ex. i1	5*	−4627	−141	+117	aggccctgtgaggagtcaaggtgaga
Init. ex. i2	6*	−5309	−155	+182	aggccctgtgaggagtcaaggtgaga
Init. ex. i3	7*	−5984	−155	+182	aggccctgtgaggagtcaaggtgaga
Init. ex. i4	8*	−6663	−155	+176	aggccctgtgaggagtcaaggtgaga
Init. ex. i5	9*	−7345	−141	+117	aggccctgtgaggagtcaaggtgaga
	10	−8028	−155	+182	aggccctgtgaggagtcaaggtgaga
Init. ex. i6	11*	−8707	−155	+176	aggccctgtgaggagtcaaggtgaga
	12	−9389	−141	+117	aggccctgtgaggagtcaaggtgaga
	13	−10077	−172	+178	aggccctgtgaggagtcaaggtgaga
Init. ex. i7	14*	−10696	−172	+199	aggctctataaggagacaaggtgaga
Init. ex. i8	15*	−11123	−45	+3	aggctctgtgaggaggcaaggtgggg
MAGE-A1	1	−2341	−167	+96	aggctctgtgaggaggcaaggtgaga
	2*	−3026	−122	+118	aggctctgtgaggaggcaaggtgaga
	3	−3620	−170	+119	aggctctgtaaggaggcaaggtgaca
MAGE-A5	1*	−2380	−172	+116	gggctctgtgaggaggcaaggtgaga
	2	−2921	−167	+119	tggctctgtgaggatgcaaggtgaga
	3	−3551	−164	+119	aggctctgtgaagaggcaaggtgaga
	4	−4130	−172	+119	aggctctgtgaggaggcaaggtgagg
	5	−4709	−168	+83	aggttctgtgaggggttaaggtgaga
MAGE-A3	1*	−1998	−41	+94	aggctcggtgaggaggcaaggtaaga
	2	−21035	−165	+120	gggctgtgtgtgaaggaaaggtgagg
	3	−21603	−79	+119	aggctctgtgagaagacaaggtgaga
MAGE-A6	1*	−2004	−41	+94	aggctcggtgaggaggcaaggtaaga
	2	−21032	−165	+120	gggctgtgtgtgaaggaaaggtgagg
	3	−21600	−79	+119	aggctctgtgagaagacaaggtgaga
MAGE-A2	1*	−2425	−172	+114	aggctccgtgaggaggcaaggtaaga
MAGE-A12	1*	−2308	−172	+114	aggctccgtgaggaggcaaggtaaga

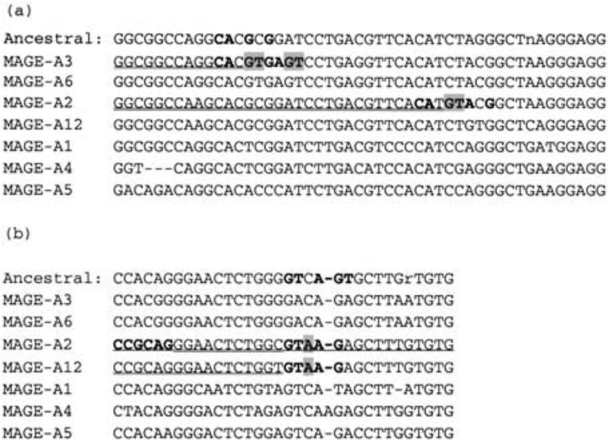
Note. The neighborhood of the donor site and homologous positions is shown. Active donor sites are marked with asterisks. Column (2): The copies of the repeat are numbered in the 3′-to-5′ direction (that is, from the coding exon). Column (3): Approximate position of the donor site aligned positions relative to the start codon. Columns (4) and (5): Positions of the aligned region relative to the donor site (or the corresponding point).

the initial exon can be aligned well in all genes from subfamily I, but it is expressed only in MAGE-A8 and MAGE-A10, and not in MAGE-A9 and MAGE-A11.

As usual, the exact boundaries of the repeated region are difficult to define. Sometimes the aligned region corresponding to the donor site is rather narrow, whereas in other cases several hundred nucleotides may be aligned.

In some genes there are additional or alternative start exons. In particular, alternative initial exons may be used in MAGE-A9 (where a shortened version of the main initial exon is preferred), in MAGE-A11 (an alternative start exon 22 kb upstream of the main one), in MAGE-A3 (an alternative start exon downstream of the main one), and in MAGE-A12 (an alternative start exon downstream of the main one). The donor site of the downstream start exon of MAGE-A3 lies within the region corresponding to exon 5 of MAGE-A2 (see below); it has been created by a C → T mutation that produced a GT dinucleotide (Fig. 8a). Similarly, the donor site of exon 5 of MAGE-A3 has been created by a C → G mutation that produced a GT dinucleotide (Fig. 8a). The donor site of the downstream start exon of MAGE-A12 is the same as the donor site of exon 6 of MAGE-A2 (see below); it has been created by a C → A mutation that improved the match to the consensus (Fig. 8b).

In MAGE-A4, of 15 candidate initial exons formed by the expanded repeat, 8 are used; 6 of these are within the long tandem repeat (three candidate exons in the repeat region are not used). However, as



```
(a)

Ancestral: GGCGGCCAGGCACGCGGATCCTGACGTTCACATCTAGGGCTnAGGGAGG
MAGE-A3    GGCGGCCAGGCACGTGAGTCCTGAGGTTCACATCTACGGCTAAGGGAGG
MAGE-A6    GGCGGCCAGGCACGTGAGTCCTGAGGTTCACATCTACGGCTAAGGGAGG
MAGE-A2    GGCGGCCAAGCACGCGGATCCTGACGTTCACATGTACGGCTAAGGGAGG
MAGE-A12   GGCGGCCAAGCACGCGGATCCTGACGTTCACATCTGTGGCTCAGGGAGG
MAGE-A1    GGCGGCCAGGCACTCGGATCTTGACGTCCCCATCCAGGGCTGATGGAGG
MAGE-A4    GGT---CAGGCACTCGGATCTTGACATCCACATCGAGGGCTGAAGGAGG
MAGE-A5    GACAGACAGGCACACCCATTCTGACGTCCACATCCAGGGCTGAAGGAGG

(b)

Ancestral: CCACAGGGAACTCTGGGGTCA-GTGCTTGrTGTG
MAGE-A3    CCACGGGGAACTCTGGGGACA-GAGCTTAATGTG
MAGE-A6    CCACGGGGAACTCTGGGGACA-GAGCTTAATGTG
MAGE-A2    CCGCAGGGAACTCTGGCGTAA-GAGCTTTGTGTG
MAGE-A12   CCGCAGGGAACTCTGGTGTAA-GAGCTTTGTGTG
MAGE-A1    CCACAGGGCAATCTGTAGTCA-TAGCTT-ATGTG
MAGE-A4    CTACAGGGGACTCTAGAGTCAAGAGCTTGGTGTG
MAGE-A5    CCACAAGGGACTCTGGAGTCA-GACCTTGGTGTG
```

Fig. 8. Donor sites of downstream alternative start exons. **a** MAGE-A2 and MAGE-A3. **b** MAGE-A2 and MAGE-A12. Notation as in Figs. 4 and 5.

most isoforms are observed only once or twice, all statements about the absence of isoforms are very preliminary; in fact, it is likely that most candidate exons will be observed once more ESTs for this gene are sequenced. In addition, there is an isoform starting with initial exon 8 and using an internal exon bounded by an acceptor site upstream of the initial exon 7 and the donor site of initial exon 7. Similarly, in MAGE-A11 there is an additional internal exon following the main initial one.

The most complicated situation is that in MAGE-A2 (Fig. 3b). There are two cassette exons, 6 and 5, and two variants of the start exon, the standard initial exon and a longer exon that starts downstream of the main one. The donor site of the longer exon is downstream of the acceptor site of exon 6, and thus there is a short overlap (11 nt) between these two exons (double-underlined in Fig. 8b). In addition, one EST contains exon 6-5, which spans exons 6 and 5, the retained intron between them, and the sequence upstream of exon 6. This exon is spliced to the acceptor site of exon E4, but after that the EST is incomplete.

Alternatively Spliced and Chimeric Isoforms

The balance between isoforms is usually rather uneven. In most cases one isoform clearly dominates (the only exclusion seems to be MAGE-A4, where two isoforms using initial exons i2 and i4 are prevalent, and MAGE-A2, where each of the multiple isoforms is supported by one or two ESTs). Exon 4 tends to be used when the gene has sufficiently strong splicing sites (it is always used in MAGE-A10 and MAGE-A11, used in most cases in MAGE-A2, and rarely used in MAGE-A12). On the contrary, exon 2 is constitutive only in MAGE-A9, whereas it is seen only in minor isoforms in MAGE-A2 and MAGE-A10. Exon 3 of MAGE-A8 is supported by only one EST. Of the start exons, the main one (or its duplicates in MAGE-A4) is preferred by all genes but MAGE-A9, where the major isoform uses the shorter version.

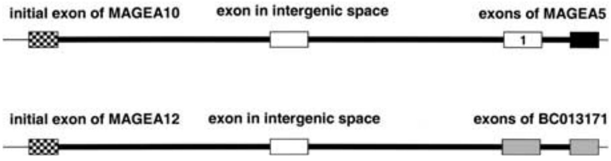


Fig. 9. Chimeric ESTs. Notation as in Fig. 3.

Finally, there are several chimeric ESTs containing exons of different genes (Fig. 9). One EST splices the initial exon of MAGE-A12 to the candidate downstream gene BC013171. There is also one EST splicing the initial exon of MAGE-A10 to exons 1 and 0 of MAGE-A5; there is one short exon in the intervening region, supporting the hypothesis that these isoforms are produced by splicing of read-through transcripts.

Discussion

Inactivation of splicing sites due to mutations is a well-known phenomenon, extensively studied in the context of human genetic disease (reviewed, e.g., by Faustino and Cooper 2003; Stoilov et al. 2002). In most cases this leads to exon skipping or activation of cryptic sites: Numerous examples were observed in many different genes (Nakai and Sakamoto 1994; O'Neill et al. 1998; Tuffery-Giraud et al. 1999; Stenson et al. 2003). On the other hand, there are cases of creation of new splicing sites by point mutations (Nelson and Green 1990; O'Neill et al. 1998; Bagnall et al. 1999). The use of specific sites is difficult to predict, although there is some correlation between site choice and its closeness to the consensus (Iida 1990; Ketterling et al. 1999; Lev-Maor et al. 2003). In most cases considered here, the creation of new sites or the choice between alternative sites indeed can be explained by mutations improving match to the consensus or, vice versa, changing functionally important nucleotides.

One specific case is the choice between AG dinucleotides in acceptor sites. We have observed several cases when the most upstream AG is used in a group of several AG dinucleotides. Inactivation of the upstream AG by mutation leads to activation of the (cryptic) AG downstream. Indeed, the avoidance of upstream AG in acceptor sites is a well-known phenomenon (Gelfand 1989). Another feature of functional acceptor sites is C (or, to a lesser extent, T) at position −3, immediately preceding the AG, and again, we have observed mutations that activate sites by changing the nucleotide at this position to C. This agrees with the results of Lev-Maor et al. (2003), where the choice of acceptor splicing sites in alternative exons created by insertions of Alu repeats was studied.

Exon duplication is a well-known mechanism of molecular evolution, in many cases mediated by alternative splicing (Kondrashov and Koonin 2001; Letunic et al. 2002). In most studied cases it concerns internal protein-coding exons, although it might be due to limitations of the applied computational techniques and the fact that protein-coding regions are more conserved than noncoding ones. Indeed, alternative splicing is often associated with differential choice of promoters (Mironov et al. 1999; Tasic et al. 2002). In MAGE-A it is not clear whether this mechanism is implicated since all genes of this family seem to have the same tissue specificity.

Finally, chimeric mRNAs are often a likely product of splicing of read-through transcripts (Romani et al. 2003).

Two remaining questions are whether our scenario of evolution of splicing sites is correct and whether the observed isoforms are functional. As regards the former question, in all cases we invoked the most parsimonious explanation assuming the smallest number of evolutionary events. Indeed, in most cases the observed mutations are gene- or lineage-specific.

The problem of functionality is more complicated. It is well known that transformed cells produce numerous aberrant mRNAs, possibly due to relaxation of control mechanisms. In particular, many alternative isoforms, especially those supported by unique ESTs, might be nonfunctional (Kan et al. 2002). So, one possible explanation for the observed diversity of spliced isoforms of MAGE-A genes might be that they represent rare mis-splicing events. We believe that this is not the case.

Indeed, in at least some cases several full-length mRNA isoforms have been observed (De Smet et al. 1994; De Plaen et al. 1994; Rogner et al. 1995); in other cases, alternative exons were found in different tissue samples (Ali Osmay Güre, personal communication). This means that the fraction of the corresponding isoforms is large enough to be detectable. Further, although in many cases the observed isoforms are supported by unique ESTs, alternative splicing sites occur in several different isoforms. Moreover, a more recent version hg16 of the Human Genome Browser contained only a few new isoforms and all of them were generated by the sites considered here and conformed to the observed patterns (data not shown).

There is also remarkable consistency in isoforms of different MAGE-A genes. It is well known that genomic sequences contain a large number of cryptic splicing sites (e.g., Thanaraj 2000). One could expect that aberrant splicing events would indiscriminantly use these sites. Instead, the same alternative sites and exons are utilized by several genes, whereas the differences in splicing of MAGE-A genes in most cases can be explained by mutations in splicing sites.

Thus the observed isoforms are real in the sense that they represent naturally occurring events. Their functionality in the sense of differential regulation, specific properties, expression patterns, etc., remains an open issue for experimental analysis.

From the evolutionary point of view, it seems that we witness an early stage of gene diversification. A plausible scenario seems to be retroposition of an ancestral MAGE-D gene (Chomez et al. 2001) in the common ancestor of human and mouse, accompanied by loss of a part of the coding region. Thus created pre-MAGE-A was subject to several independent duplications in each of these genomes. Duplications in the human genome seem to be continuing in the sense that some of them are very recent.

Diversification on the level of genes was accompanied by diversification on the level of alternatively spliced isoforms. Alternative splicing did not influence the coding region but generated different 3′-untranslated regions in mRNAs. This process was shaped by two types of events: deactivating mutations in splicing sites, leading to exon loss or intron retention, and birth of new sites by point mutations or insertions creating GT or AG dinucleotides in a proper context, leading to emergence of new alternatively spliced exons. Both types of events could cause exon truncation or extension due to the use of preexisting cryptic sites or newly generated sites.

Analysis of the Mage-a genes in mouse and other families of MAGE/Mage genes will show how common the observed situation is. It should be particularly interesting to analyze the MAGE-B family, where the human and mouse genes are intermixed in the phylogenetic tree, and the MAGE-D family, whose members are highly conserved in the human and mouse genomes and can be found in more distant genomes as well.

Acknowledgments. We are grateful to Andrew Simpson for the suggestion to analyze alternative splicing of CT antigens, to Andrey Mironov for useful discussion, and to Ali Osmay Güre for critical reading of the manuscript. We appreciate unpublished data on expression of alternatively spliced isoforms kindly provided by Matthew Scanlan and Ali Osmay Güre. This study was partially supported by grants from the Ludwig Institute of Cancer Research (CRDF RBO-1268), Howard Hughes Medical Institute (55000309), Russian Fund of Basic Research (04-04-49440), and Program in Molecular and Cellular Biology of the Russian Academy of Sciences.

References

Bagnall RD, Waseem NH, Green PM, Colvin B, Lee C, Giannelli F (1999) Creation of a novel donor splice site in intron 1 of the factor VIII gene leads to activation of a 191 bp cryptic exon in two haemophilia A patients. Br J Haematol 107:766–771

Boccaccio I, Glatt-Deeley H, Watrin F, Roeckel N, Laland M, Muscatelli F (1999) The human MAGEL2 gene and its mouse homologue are paternally expressed and mapped to the Prader-Willi region. Hum Mol Genet 8:2497–2505

Brett D, Pospisil H, Valcarcel J, Reich J, Bork P (2002) Alternative splicing and genome complexity. Nat Genet 30:29–30

Cannon SB, Young ND (2003) OrthoParaMap: Distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. BMC Bioinformatics 4:35

Chomez P, De Backer O, Bertrand M, De Plaen M, Boon T, Lucas S (2001) An overview of the MAGE gene family with the identification of all human members of the family. Cancer Res 61:5544–5551

Dabovic B, Zanaria E, Bardoni B, Lisa A, Bordignon C, Russo V, Matessi C, Traversari C, Camerino G (1995) A family of rapidly evolving genes from the sex reversal critical region in Xp21. Mamm Genome 6:571–580

De Plaen E, Arden K, Traversari C, Gaforio JJ, Szikora JP, De Smet C, Brasseur F, van der Bruggen P, Lethe B, Lurquin C, Brasseur R, Chomez P, De Backer O, Cawenee W, Boon T (1994) Structure, chromosomal localisation, and expression of 12 genes of the MAGE family. Immunogenetics 40:360–369

De Plaen E, Naerhuyzen B, De Smet C, Szikora J-P, Boon T (1997) Alternative promoters of gene MAGE-4a. Genomics 40:305–313

De Plaen E, De Backer O, Arnaud D, Bonjean B, Chomez P, Martelange V, Avner P, Baldacci P, Babinet C, Hwang S-Y, Knowles B, Boon T (1999) A new family of mouse genes homologous to the human *MAGE* genes. Genomics 55:176–184

De Smet C, Lurquin C, van der Bruggen P, De Plaen E, Brasseur F, Boon T (1994) Sequence and expression pattern of the human MAGE2 gene. Immunogenetics 39:121–129

Faustino NA, Cooper TA (2003) Pre-mRNA splicing and human disease. Genes Dev 17:419–437

Felsenstein J (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol 266:418–427

Gelfand MS (1989) Statistical analysis of mammalian pre-mRNA splicing sites. Nucleic Acids Res 17:6369–6382

Iida Y (1990) Quantification analysis of 5′-splice signal sequences in mRNA precursors. Mutations in 5′-splice signal sequence of human beta-globin gene and beta-thalassemia. J Theor Biol 145:523–533

Jay P, Rougeulle C, Massacrier A, Moncla A, Mattei M-G, Malzac P, Roëckel N, Taviaux S, Bergé Lefranc J-L, Cau P, Berta P, Lalande M, Muscatelli F (1997) The human necdin gene, NDN, is maternally imprinted and located in the Prader-Willi syndrome chromosomal region. Nat Genet 17:357–361

Kan Z, States D, Gish W (2002) Selecting for functional alternative splices in ESTs. Genome Res 12:1837–1845

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

Ketterling RP, Drost JB, Scaringe WA, Liao DZ, Liu JZ, Kasper CK, Sommer SS (1999) Reported in vivo splice-site mutations in the factor IX gene: severity of splicing defects and a hypothesis for predicting deleterious splice donor mutations. Hum Mutat 13:221–231

Kondrashov FA, Koonin EV (2001) Origin of alternative splicing by tandem exon duplication. Hum Mol Genet 10:2661–2669

Krawczak M, Reiss J, Cooper DN (1992) The mutational spectrum of base-pair substitutions in messenger RNA splice junctions of human genes—Causes and consequences. Hum Genet 90:41–54

Letunic I, Copley RR, Bork P (2002) Common exon duplication in animals and its role in alternative splicing. Hum Mol Genet 11:1561–1567

Lev-Maor G, Sorek R, Shomron N, Ast G (2003) The birth of an alternatively spliced exon: 3′ Splice-site selection in Alu exons. Science 300:1288–1291

Lucas S, De Smet C, Arden KC, Viars CS, Lethe B, Lurquin C, Boon T (1998) Identification of a new MAGE gene with tumor-specific expression by representational difference analysis. Cancer Res 58:743–752

Lucas S, Brasseur F, Boon T (1999) A new MAGE gene with ubiquitous expression does not code for known MAGE antigens recognized by T cells. Cancer Res 59:4100–4103

Lucas S, De Plaen E, Boon T (2000) MAGE-B5, MAGE-B6, MAGE-C2, and MAGE-C3: Four new members of the MAGE family with tumor-specific expression. Int J Cancer 87:55–60

Mironov AA, Fickett JW, Gelfand MS (1999) Frequent alternative splicing of human genes. Genome Res 9:1288–1293

Modrek B, Lee CJ (2002) A genomic view of alternative splicing. Nat Genet 30:13–19

Modrek B, Lee CJ (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet 34:117–180

Muscatelli F, Walker AP, De Plaen E, Stafford AN, Monaco AP (1995) Isolation and characterization of a new MAGE gene family in the Xp21.3 region. Proc Natl Acad Sci USA 92:4987–4991

Nakai K, Sakamoto H (1994) Construction of a novel database containing aberrant splicing mutations of mammalian genes. Gene 141:171–177

Nelson KK, Green MR (1990) Mechanism for cryptic splice site activation during pre-mRNA splicing. Proc Natl Acad Sci USA 87:6253–6257

O'Neill JP, Rogan PK, Cariello N, Nicklas JA (1998) Mutations that alter RNA splicing of the human HPRT gene: a review of the spectrum. Mutat Res 411:179–214

Nurtdinov RN, Artamonova II, Mironov AA, Gelfand MS (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. Hum Mol Genet 12:1313–1320

Pold M, Zhou J, Chen GL, Hall JM, Vescio RA, Berenson JR (1999) Identification of a new, unorthodox member of the MAGE gene family. Genomics 59:161–167

Rogner UC, Wilke K, Steck E, Korn B, Poustka A (1995) The melanoma antigen gene (MAGE) family is clustered in the chromosomal band Xq28. Genomics 29:725–731

Romani A, Guerra E, Trerotola M, Alberti S (2003) Detection and analysis of spliced chimeric mRNAs in sequence databanks. Nucleic Acids Res 31:e17

Scanlan MJ, Gure AO, Jungbluth AA, Old LJ, Chen Y-T (2002) Cancer/testis antigens: An expanding family of targets for cancer immunotherapy. Immunol Rev 188:22–32

Sorek R, Ast G, Graur D (2002) Alu-containing exons are alternatively spliced. Genome Res 12:1060–1067

Sorek R, Shamir R, Ast G (2004) How prevalent is functional alternative splicing in the human genome? Trends Genet 20:68–71

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN (2003) Human Gene Mutation Database (HGMD(R)): 2003 update. Hum Mutat 21:577–581

Stoilov P, Meshorer E, Gencheva M, Glick D, Soreq H, Stamm S (2002) Defects in pre-mRNA processing as causes of and predisposition to diseases. DNA Cell Biol 21:803–818

Stone B, Schummer M, Paley PJ, Crawford M, Ford M, Urban N, Nelson BH (2001) MAGE-F1, a novel ubiquitously expressed member of the MAGE superfamily. Gene 267:173–182

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB (2002) Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci USA 99:4465–4410

Tasic B, Nabholz CE, Baldwin KK, Kim Y, Rueckert EH, Ribich SA, Cramer P, Wu Q, Axel R, Maniatis T (2002) Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. Mol Cell 10:21–33

Thanaraj TA (2000) Positional characterisation of false positives from computational prediction of human splice sites. Nucleic Acids Res 28:744–754

Thanaraj TA, Clark F, Muilu J (2003) Conservation of human alternative splice events in mouse. Nucleic Acids Res 31:2544–2552

Thompson JD, Higgins DG, Gibson TJ (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

Tuffery-Giraud S, Chambert S, Demaille J, Claustres M (1999) Point mutations in the dystrophin gene: Evidence for frequent use of cryptic splice sites as a result of splicing defects. Hum Mutat 14:359–368

van der Bruggen P, Traversari C, Chomez P, Lurquin C, De Plaen E, Van den Eynde B, Knuth A, Boon T (1991) A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. Science 254:1643–1647