DOI: 10.1016/j.bbagrm.2016.03.014

REVISED MANUSCRIPT (text with changes Marked)

MIR retroposon exonization promotes evolutionary variability and generates species-specific expression of IGF-1 splice variants

Giosuè Annibalini[1]*†, Pamela Bielli[2]†, Mauro De Santi[1], Deborah Agostini[1], Michele Guescini[1], Davide Sisti[1], Serena Contarelli[1], Giorgio Brandi[1], ~~Claudio Sette[2],~~ Anna Villarini[3], Vilberto Stocchi[1], Claudio Sette[2] and Elena Barbieri[1]^

[1]Department of Biomolecular Sciences, University of Urbino Carlo Bo, 61029 Urbino, Italy.
[2]Department of Biomedicine and Prevention, University of Rome Tor Vergata, 00133 Rome, Italy; Laboratory of Neuroembryology, Fondazione Santa Lucia, 00133 Rome, Italy.
[3]Department of Preventive & Predictive Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori, 20133 Milan, Italy.
^IIM, Interuniversity Institute of Myology

* To whom correspondence should be addressed. Tel:+39 0722-303418 Fax:+39 0722-303401; Email: giosue.annibalini@uniurb.it

†Giosuè Annibalini and Pamela Bielli equally contributed to this work.

Abstract

Insulin-like growth factor-1 (IGF-1) is a pleiotropic hormone exerting mitogenic and anti-apoptotic effects. Inclusion or exclusion of exon 5 into the IGF-1 mRNA gives rise to three transcripts, IGF-1Ea, IGF-1Eb and IGF-1Ec, which yield three different C-terminal extensions called Ea, Eb and Ec peptides. The biological significance of the IGF-1 splice variants and how the E-peptides affect the actions of mature IGF-1 are largely unknown. In this study we investigated the origin and conservation of the IGF-1 E-peptides and we compared the pattern of expression of the IGF-1 isoforms *in vivo*, in nine mammalian species, and *in vitro* using human and mouse IGF-1 minigenes.

Our analysis showed that only IGF-1Ea is conserved among all vertebrates, whereas IGF-1Eb and IGF-1Ec are an evolutionary novelty originated from the exonization of a mammalian interspersed repetitive-b (MIR-b) element. Both IGF-1Eb and IGF-1Ec mRNAs were constitutively expressed in all mammalian species analyzed but their expression ratio varies greatly among species. Using IGF-1 minigenes we demonstrated that divergences in *cis*-acting regulatory elements between human and mouse conferred species-specific features to the exon 5 region. Finally, the protein-coding sequences of the exon 5 showed low rate of synonymous mutations and contain disorder-promoting amino acids, suggesting a regulatory role for these domains.

In conclusion, the exonization of a MIR-b element in the *IGF-1* gene determined gain of the exon 5 gain during mammalian evolution. Alternative splicing of this novel exon added new regulatory elements at the mRNA and protein level potentially able to regulate the mature IGF-1 across tissues and species.

Keywords

IGF-1 isoforms, alternative splicing, retroposon exonization, synonymous sites, intrinsically disordered regions

1. Introduction

The mammalian insulin-like growth factor (*IGF*) *-1* gene is a single copy gene composed of six exons and five introns, which gives rise to an immature IGF-1 peptide (IGF-1 pro-peptide) containing a signal peptide at the 5' end of the gene, a core region and an E-peptide at the 3' end (Fig.1 Figs. 1A and 1B). Both signal peptide and E-peptide are then removed by protease cleavage to form the 70 amino acid-long mature IGF-1 peptide (IGF-1 core), which displays growth-promoting and metabolic functions.

Four of the six *IGF-1* exons are subjected to alternative splicing (Fig.1 Figs. 1A and 1B) [1]. Exon 1 and exon 2 are mutually exclusive first exons and generate different signal peptides. Transcripts containing exon 1 are referred to as Class I transcripts whereas those containing exon 2 are referred to as Class II transcripts. The

Class II IGF-1 knockout mice indicated that class II isoforms are dispensable for fetal and postnatal growth, and the significance of the alternate signal peptides encoded by the first two IGF-1 exons was discussed in [2-3]. At the 3' end of the gene, alternative splicing t~~he 3' end of the pre-mRNA~~ yields three different mRNA transcripts, each encoding distinct carboxyl-terminal portions of E-peptide followed by the 3'-untranslated region (3'UTR) (~~Fig. 1~~ Figs. 1A and 1B). Thus, although alternative splicing generates different precursor peptides, it does not alter the sequence ~~structure~~ of the mature IGF-1 peptide.

Splicing of exon 4 with exon 6 yields the most common IGF-1Ea variant, which encodes the 35 amino acid-long Ea peptide. The first 16 amino acids of the Ea peptide are encoded by exon 4 and are common in all ~~of the three different~~ E-peptides, whereas the remaining 19 amino acids are encoded by exon 6 and are unique to this isoform. The IGF-1Eb variant is produced when exon 4 is spliced with exon 5 and encodes the Eb peptide, which contains the 16 common amino acids and 61 additional amino acids encoded by exon 5. IGF-1Eb transcript terminates in exon 5 and excludes exon 6 from the mRNA, hence it has a completely different 3'UTR compared to IGF-1Ea and IGF-1Ec. The third variant, named IGF-1Ec in humans, is generated by usage of a cryptic 5' splice site (c5'ss) (named IGF633) [4] present in exon 5, which is in turn spliced with exon 6. The ~~cryptic 5'ss~~ c5'ss of *IGF-1* exon 5 deviates from the vertebrate consensus and is commonly used in rodents and rabbits [5-7] but rarely and in a tissue-specific manner in human [1, 4]. Notably, although this variant is named IGF-1Eb in rodents, for clarity we will use ~~throughout this manuscript~~ the human IGF-1Ec nomenclature throughout this manuscript, regardless of the species. The human Ec peptide has a predicted length of 40 amino acids, with the common 16 amino acids deriving from the exon 4, 16 from exon 5 and the last 8 amino acids from exon 6. Since expression of IGF-1Ec was linked to mechanical stimuli in muscle and other mechanosensitive cells, this variant is sometime referred to as Mechano Growth Factor (MGF) [1].

Few studies have addressed the regulation of *IGF-1* alternative splicing. In human cells, splicing of exon 5 is regulated by the antagonistic activities of the serine-arginine protein splicing factor-1 (SRSF1) and heterogeneous nuclear ribonucleoproteins A1 (hnRNPA1) [8-9]. SRSF1 was proposed to increase splicing of the IGF-1Eb variant by binding to a purine-rich exonic splicing enhancer (ESE) in exon 5 and preventing the recruitment of hnRNPA1, which functions as a splicing repressor [8-9]. Notably, neither splicing factor was capable of regulating the IGF-1Ec variant in human IGF-1 minigenes [8-9]. The reason of the lack of usage of the ~~cryptic 5'ss~~ c5'ss, and hence for the lack of IGF-1Ec isoform production, reported in these studies~~,~~ has yet to be clarified. Similarly, whether or not the mouse ~~cryptic 5'ss~~ c5'ss, which is closer to the canonical splicing donor consensus, can be recognized in these experimental settings has not been tested. Furthermore, it is currently unknown whether SRSF1 and hnRNP A1 function specifically ~~on~~ with the human IGF-1 pre-mRNA or whether they also play a role in *IGF-1* alternative splicing in other species.

To date, the fate and the biological functions of ~~the~~ E-peptides are not entirely clear [1, 10-12]. Interestingly, gene structure comparison showed that ~~the~~ *IGF-1* exon 6 is conserved in all vertebrates, whereas exon 5 is conserved only among mammals [13-15]. In addition, comparative studies on mammalian *IGF-1* show that both the IGF-1 core and Ea peptide are subjected to strong purifying selection, whereas sequences of the Eb

and Ec peptides are more variable [16-17]. In particular it was shown that the ratio of non-synonymous (dN; amino-acid altering) to synonymous (dS; silent) substitutions is about 10-fold higher for Eb and Ec peptides compared to Ea peptide, suggesting that ~~the Eb and Ec peptides might have little~~ they have no specific roles or, at most, species-specific functions [17]. Indeed, it is usual to view poor protein sequence conservation, i.e. evolving under neutral or nearly neutral conditions, as evidence of reduced functional importance [17-18]. However, Lin and colleagues [19] recently demonstrated that the "dual-coding" DNA sequences of *IGF-1* exon 6, which is translated in two alternative reading frame to give rise to Ea and Ec peptides, showed a very low dS rate suggesting additional sequence constraints beyond those dictated by the amino acid sequence of the Ea and Ec peptides. Notably, the picture of largely neutral evolution of synonymous substitutions in mammals has been recently challenged [20], showing that selection may constrain not only dN~~,~~ to preserve amino acid sequence, but also dS~~,~~ to preserve regulatory elements in nucleotide sequences, such as ESE, RNA secondary structures and microRNA target sites [19-21]. However, whether the dS drop of the dual-coding exon 6 has contributed to the observed local increase in dN/dS ratio of E-peptides and hence to an inaccurate estimate of evolutionary rate is still unclear.

In this study, we have analyzed the conservation and the mRNA expression pattern of IGF-1 isoforms in different mammalian species and evaluated the evolutionary pressure within the different coding regions of the mammalian *IGF-1* gene. Moreover, we have used human and mouse minigene systems to compare the mechanisms regulating the *IGF-1* exon 5 splicing between these two species.

2. Materials and Methods

2.1 Sequences and databases

Orthologous of the human *IGF-1* gene were obtained from the UCSC Genome Browser Database (http://genome.ucsc.edu; Feb. 2009 GRCh37/hg19) and Ensembl website (http://www.ensembl.org). The "CDS FASTA alignment from multiple alignments" data, derived from the "multiz100way" alignment data prepared from 100 vertebrate genomes [22], were downloaded using the Table Browser tool of the UCSC Genome Browser. Sequences were subsequently realigned using MUSCLE [23] and protein coding sequences from 27 mammalian species were extracted from these alignment datasets. Supplementary file S1 contains the IGF-1 sequence of 27 mammalian species in FASTA format and the neighbor-joining tree used in the present study. Percent nucleotide and amino acid identities between sequences were computed using the CLUSTALW procedure [24]. The splice site scores were obtained for each 5' and 3' ss sequence using the Maximum Entropy scores [25]. For Transposable elements analysis, we used RepeatMasker (http://www.repeatmasker.org) and Repbase annotations [26-28].

2.2 Human and animal tissues

Freshly frozen normal human liver (2 males, mean age 54 +/- 9 years), adipose (2 males and 2 females, mean age 52 +/- 12 years) and muscle (3 males, mean age 25 +/- 6 years) samples were provided by the

complex structure of biomarkers (DOSMM) of National Cancer Institute of Milan. The macaque (*Macaca mulatta)* autoptic specimens were kindly provided by Elena Borra (Department of Neuroscience, University of Parma, Parma, Italy). Three 4-year olds macaques were tested in this study (2 female and 1 male). One-month old CD1 female mice (n=3) (*Mus Musculus*), 2-month old Young Sprague Dawley male rat (n=3) (*Rattus norvegicus*) and three 8 month-old New Zealand male rabbits (*Oryctolagus cuniculus*) (Charles River Laboratories, Milan, Italy) were housed with a 12-h light/dark cycle and free access to standard laboratory chow and water. Care and handling were in accordance with the *Guide for the Care and Use of Laboratory Animals* by Ministero della Sanità D.L. 116 (1992) and approved by the university committee for animal experiments. Animals were sacrificed with an overdose of anaesthetics (ketamine in combination with xylazine). Other tissues used in this study were collected at local slaughter during routine meat inspection: pig (*Sus scrofa*) (3 males), cow (*Bos taurus*) (3 females), sheep (*Ovis aries*) (3 female) and goat (*Capra hircus*) (2 males and 1 female) were tested in this study. The age of the animals ranged from 2 to 5 years. All tissues (about 30 mg) were immediately submerged in RNAlater stabilization solution (Qiagen, Milan, Italy) and left at least 10 min at room temperature. Then, tissues were stored at -80°C until RNA extraction.

2.3 RNA extraction and cDNA synthesis

The tissues were removed from RNAlater and transferred into a clean Eppendorf tube where the total RNA was extracted and purified using the Omega Bio-Tek E.Z.N.A.TM Total RNA kit (VWR International s.r.l., Milan, Italy) according to the manufacturer's instructions. The amount and quality of RNA were assessed with DU-640 UV Spectrophotometer (Beckman Coulter). After DNA digestion with DNase I enzyme (Qiagen, Milan, Italy) complementary DNA was synthesized from 1 µg of total RNA using Omniscript RT (Qiagen, Milan, Italy) and random hexamers or anchored oligo-dT primers where specified.

2.4 Real time PCR quantification of human, macaque and mouse IGF-1 isoforms

Serial dilution (1:4) of three recombinant plasmids containing the IGF-1Ea, IGF-1Eb and IGF-1Ec sequences of human and mouse were prepared in order to generate standard curves for plotting CT values against number of molecules. Molecules of each plasmid were calculated using the concentration of each plasmid, Avogadro's constant, the molecular weight of double-stranded DNA and the size of the target amplicon [29]. The copy number of genes was calculated in individual samples using a corresponding reference plasmid cDNA clone at known concentration. The amount of target transcripts was normalized to glyceraldehyde-3-phosphate dehydrogenase (GAPDH). Percentage of IGF-1 isoforms was calculated as [(mRNA copy number of single isoform/(mRNA copy numbers of IGF1Ea+IGF1Eb+IGF1Ec)*100]. Expression of IGF-1 isoforms was ~~were~~ compared using 2-ways analysis of variance with interactions: species and IGF-1 isoforms were used as predictive factors. In order to meet assumption of homoscedasticity percentage were arcsin radq transformed. Post-hoc analysis was performed using Bonferroni correction.

Real-time quantitative PCR was performed with two microliters of cDNA and 300 nM of of each primer in an Applied Biosystems StepOnePlus™ Real Time PCR System using SYBR Select Master Mix (Applied

Biosystems, Monza, Italy). The real-time PCR conditions were: 50°C for 2 min, 95°C for 2 min followed by 40 cycles of three-steps at 95°C for 15 sec, 60°C for 15 sec and 72°C for 30 sec. The specificity of the amplification products was confirmed by examining thermal denaturation plots and by sample separation in a 4% DNA agarose gel. The sequences and the annealing positions of the primers used to quantify the IGF-1 isoforms were shown in Supplementary Table S1.

2.5 3' RACE-PCR

The mRNA splicing pattern of IGF-1 was assayed by 3-RACE-PCR using an anchored oligo-dT primer (5'-AAGCAGTGGTATCAACGCAGAGTACT$_{(30)}$NV-3') as reverse transcription primer. 1 µg of total liver RNA was reverse transcribed into cDNA. To amplify the 3'-end IGF-1 variants a reverse primer corresponding to the anchor sequence of the RT primers was used in combination with the following forward primers: human, macaque, rabbit, (5'-CCTCCTCGCATCTCTTCTACCTG-3'); mouse and rat (5'-GCTATGGCTCCAGCATTCG-3'); and pig (5'-CGTGGATGAGTGCTGCTTC-3'). The PCR reaction were as follows: 95° for 10 min; 35 cycles of three-steps at 95°C for 30 sec, 60°C for 30 sec, 72°C for 1 min and 30 sec followed by a final elongation cycle (72°C for 5 min). The PCR products were loaded on 4.0% agarose gel and DNA fragments were eluted from gel, subcloned and sequenced.

2.6 Conventional RT-PCR

RT-PCR was performed in 50 µl of reaction volume with 4 µl of cDNA, 800 nM of primers and 25 µl of 2X HotStartTaq mix (Qiagen, Milan, Italy). The sequences and the annealing positions of the primers used for RT-PCR were shown in Supplementary Table S1. RT-PCR conditions involved an initial denaturation step at 95°C for 10 min, followed by 35 cycles with denaturation step at 95°C for 30 sec, annealing at 60°C for 30 sec and extension at 72°C for 30 sec. PCR products were size-fractionated by electrophoresis on 4.0% agarose gels and visualized by ethidium bromide staining under UV light. Amplification products were purified with QIAquick gel extraction kit (Qiagen), cloned and sequenced.

2.7 Plasmid constructs

The 5' and 3'end of IGF-1 mouse minigene were amplified using primers #(1,2) and #(3,4), respectively, from C57 mice genomic DNA. After enzymatic digestion, the PCR products were cloned in KpnI/SalI and SalI/NotI restriction sites of pCI vector (Promega). The 5' and 3'end of the IGF-1 human minigene were amplified using primers #(5,6) and #(7,8), respectively, from HeLa cell genomic DNA and cloned in EcoRI/SalI and SalI/NotI restriction sites of pCI vector. The mGA*vs*TG and hTG*vs*GA IGF-1 mutant minigenes were constructed using the mega-primer strategy [30]. The mouse and human IGF-1 5' mutant ends were generated using primers #(1,9,2) or #(5,10,6) respectively. All oligonucleotide sequences are listed in Supplementary Table S1. After enzymatic digestion, the PCR products were subcloned in the corresponding wild-type minigene. PCR reactions were performed using Phusion Hot Start High-Fidelity DNA polymerase (Thermo Scientific) according to manufacturer's instruction. All plasmids were sequenced and validated.

2.8 Cell cultures and transfections and cell extract preparation

Cell cultures, transfections and sample preparation were carried out by standard methods as previously described [31]. Briefly, ~~HEK293T and primary mouse hepatocytes~~ human and mouse cell~~s~~ lines were transfected with various combination of vectors as indicated using Lipofectamine 2000 (Invitrogen). Twenty-four hours after transfection, cells were collected for RNA and protein isolation, as previously described [31]. For RNAi, cell were transfected twice with 60nM siRNAs using Lipofectamine RNAi Max according to manufacturer's instruction (Invitrogen). Sequences of siRNAs are listed in the Supplemental Table 1.

2.9 Splicing assay and PCR analyses

Twenty-four hours after transfection, cells were collected for RNA extraction using TRIzol (Ambion, Life Technologies) according to the manufacturer's instructions. After DNase digestion (Roche), 1 µg of total RNA was retrotranscribed using M-MLV reverse transcriptase (Promega). cDNA was used as template for conventional PCR reactions (GoTaq G2, Promega) in presence of specific primers and PCR products were analysed on agarose gel. In details, primers #(11,13) were used to amplify mIGF-1Eb isoform; primers #(11,12) to amplify mouse and human IGF-1Ea and IGF-1Ec isoforms; primers #(11,14) to amplify hIGF-1Eb; primers #(11,15) and #(11,16) to amplify total mouse and human IGF-1, respectively. In splicing assay human IGF-1Ea and IGF-1Eb were amplified using primers #(11,12,14) in the same PCR reaction. ~~All oligonucleotide sequences~~ The sequences and the annealing positions of the primers are showed ~~listed~~ in Supplementary Table S1.

2.10 IGF-1 evolutionary analysis

The MG94xREV_3x4 codon model with transition/transversion bias correction and nine base frequency parameters was fitted to sequence alignments by using the HyPhy 2.2.3 package [32]. A neighbor-joining tree was built using the IGF-1 core alignment of 27 mammalian species and the tree topology along with the alignment were used as input for IGF-1Ea, IGF-1Eb and IGF-1Ec sequence analysis (Supplementary file S1). For every branch in a phylogenetic tree, the expected number of synonymous substitutions per synonymous site (dS) and its counterpart for nonsynonymous substitutions (dN) were estimated by using maximum likelihood [33] allowing for independent dN and dS values for each branch (the local parameters option). Mean and 95% confidence interval for each dS, dN and dN-dS values are calculated using bootstrap approach with n=1000 resampling [34].

The web application of DisCons [35] was used to predict the protein disorder on the amino acid level, using IUPred long disorder prediction parameter [36].

3 Results

3.1 *IGF-1* exon 5 is conserved in Mammalia but not in other vertebrates

Previous studies demonstrated that the IGF-1Ea splice variant, which skips exon 5, is the most common E-peptide isoform expressed among vertebrates and is the predominant form in amphibians and birds [13-15]. To trace the evolutionary history of the IGF-1 isoforms that include exon 5 (i.e. IGF-1Eb and IGF-1Ec), we performed BLAST searches in mammalian and non-mammalian draft genomes using the Ensembl database and the UCSC vertebrate genome alignment provided at the UCSC Genome Browser site (Fe. 2009 GRCh37/hg19). ClustalX alignment of *IGF-1* sequences was performed along the exon ~~6~~ 5 and exon ~~5~~ 6 regions among 23 representative amniota vertebrates (Supplementary Fig. S1). We retrieved orthologs of the exon 5 sequence from all placental mammals (Eutheria), except Guinea pig (*Cavia aperea porcellus)*, probably because of insufficient sequencing coverage. In addition, we identified exon 5 sequences from the marsupial short-tailed opossums (*Didelphidae*) as well as from the monotremata egg-laying platypus (*Prototheria*). By contrast, we failed to identify exon 5 orthologs in all examined non-mammalian vertebrates (fish, amphibians, reptiles and birds), including turkey (*Meleagris gallopavo*), chicken (*Gallus gallus*), zebra finch (*Taeniopygia guttata*) and green anole lizard (*Anolis carolinensis*) (Supplementary Fig. S1). These data show that only exon 6 is conserved in all vertebrates and suggest that the Ea peptide represents the ancestral IGF-1 E-peptide, whereas the exon 5-encoded Eb and Ec peptides have been acquired during evolution in the mammalian lineage.

3.2 MIR-b retroposon exonization in the *IGF-1* gene explains the origin of the ~~Mammalia~~ alternatively spliced exon 5 in Mammalia

~~Recent studies suggest that~~ Transposon exonization is a major source of new exons in higher eukaryotes [37]. This process occurs when transposon-derived intronic sequences are recognized by the spliceosome and are exonized. Using the RepeatMasker web server, we searched for transposon elements in the genomic region encompassing *IGF-1* exon 5. No repetitive sequences were detected using the search tools “abblast”, “rmblast” or “cross-match” whereas the recently introduced “nhmmer” search engine ~~of the RepetMasker web server,~~ which uses the Dfam database, [27-28], highlighted a transposon belonging to the Mammalian Interspersed Repetitive-b (MIR-b) element (bit score 13.6; E-value 0.0031). The *IGF-1* exon 5 displays about 60% nucleotide identity to 103 nt of MIR-b consensus sequence from Repbase (Fig. 1C ~~2~~). The exonized MIR-b element is inserted in the antisense orientation relative to the *IGF-1* gene. Both the polypyrimidine tract and the *IGF-1* exon 5 splice sites are embedded within the MIR-b sequence (Fig. 1C ~~2~~). The 3'ss of *IGF-1* exon 5 is in close proximity to the ~~65-nt~~ 65 nt that form the highly conserved core region of MIR-b while the cryptic 5'ss lies ~~inside~~ within this region (Fig. 1C ~~2~~). This configuration is similar to those described for others antisense MIR-b exons [35-36] and ~~of~~ for disease-linked genes containing MIR pseudoexons [38].

3.3 Comparison of the MIR-derived *IGF-1* exon 5 sequence within mammalian species

Alignment of the *IGF-1* exon 5 sequence ~~nucleotides~~ originated from exonization of MIR-b element among ~~human, macaque, mouse, rat, rabbit, pig, cow, sheep and goat~~ mammalian orthologous showed a level of nucleotide identity ranging from 56 to 62% between MIR-b consensus sequence and *IGF-1* exon 5 coding for Ec peptide (49 bp in human and 52 bp in other species) ~~ranging from 56 to 62%~~ (Fig. 2A ~~3A~~). This analysis indicates that, after exonization, the exon 5 sequence coding for the Ec peptides has been fairly well conserved among species, even though the percent of identity of nucleotide sequence was lower in rodents and logomorpha than in others species (Fig. 2A ~~3A~~). Notably, the 3'ss of exon 5 is conserved and relatively strong, whereas the ~~cryptic 5'ss~~ c5'ss has only a limited match to the consensus, except in mouse and rat (Figs. 2A ~~3A~~ and 2B ~~3B~~). Interestingly, humans and murids showed a marked difference in c5'ss strength ~~of the cryptic 5'ss of exon 5~~. This splice site is very weak in humans; by contrast it appears strong in mouse and rat (Fig. 2B ~~3B~~). Further analysis of ~~the strength of~~ the *IGF-1* c5'ss strength ~~exon 5'ss~~ among 58 mammalian genomic sequence datasets available at the UCSC genome browser showed that most mammals (79%) ~~have a cryptic 5'ss with~~ display medium strength, while 16% have high strength ~~a strong 5'ss~~, including all members of the two rodent families Cricetidae: prairie vole (*Microtus ochrogaster*), chinese hamster (*Cricetulus barabensis griseus*), golden hamster (*Mesocricetus auratus*) and Muridae: mouse (*Mus musculus*) and rat (*Rattus norvegicus*) (Supplementary Fig. S2). These analyses suggest that processing of both IGF-1Eb and IGF-1Ec mRNA variants are expected in most mammals, whereas the weak nature of the ~~cryptic 5'ss~~ c5'ss of human exon 5 will likely prevent its recognition by the spliceosome, yielding predominantly the IGF-1Eb isoform. By contrast, the relatively strong c5'ss of exon 5 in the Muridae Family will favour processing of the IGF-1Ec. Therefore, our analyses suggest that, under normal conditions, the IGF-1 alternative splice variants are likely expressed in a species-specific manner.

3.4 Expression pattern of IGF-1 mRNA isoforms in mammals ~~within mammalian species~~

To test this hypothesis, we quantified by real-time PCR the IGF-1Ea, IGF-1Eb and IGF-1Ec mRNA variants in skeletal muscle, adipose tissues and liver of human (*Homo sapiens)*, macaque (*Macaca mulatta)* and mouse (*Mus musculus)*. These species represent examples of weak, intermediate and strong exon 5 c5'ss ~~cryptic 5'ss of exon 5~~, respectively (Fig. 2B ~~3B~~). As predicted by our analysis, the expression profile of these IGF-1 mRNA variants ~~changed~~ varies among species (Anova test; $p<0.01$) (Fig. 3A ~~4A~~). In particular the IGF-1Ea and IGF-1Eb isoforms are predominant in human skeletal muscle, adipose tissue and liver, whereas the IGF-1Ec isoform accounts only for 1-5% of total IGF-1 mRNAs (Fig. 3A ~~4A~~). Conversely, IGF-1Ea and IGF-1Ec isoforms predominate in mouse tissues, whereas the IGF-1Eb isoform was barely detectable (≤1% of total IGF-1 mRNAs) (Fig. 3A ~~4A~~), as also reported previously [39]. In macaque (*Macaca mulatta)*, the expression of IGF-1Eb was higher compared to IGF-1Ec in skeletal muscle and adipose tissue, whereas these splice variants were expressed at approximately equivalent level in the liver (Fig. 3A ~~4A~~). Noteworthy, the same PCR primers have been used to amplify human and macaque IGF-1 mRNA variants, ruling out differences in amplification efficiencies between species.

To ~~further~~ confirm the species-specific expression pattern of IGF-1 isoforms, we performed ~~a~~ 3' RACE-PCRs using liver-derived cDNA~~s~~ from human, macaque ~~and~~ or mouse. In addition, we ~~also~~ analyzed liver tissues from rat (*Rattus norvegicus*; strong c5'ss), rabbit (*Oryctolagus cuniculus*; intermediate c5'ss strength) and pig (*Sus scrofa*; intermediate c5'ss strength). IGF-1Ea, IGF-1Eb and IGF-1Ec mRNA products were amplified in macaque, rabbit and pig liver, which represent examples of intermediate ~~cryptic~~ c5'ss~~.~~ (Fig. 3B~~–4B~~ and Supplementary Figs. S3A, S3B and S3C for the complete nucleotide sequences and putative termination sites). On the contrary, the IGF-1Ec and IGF-1Eb splice variants were not detected in human and Muridae liver, respectively, probably because their expression level is too low for the sensitivity of the 3' RACE-PCR analysis. Notably, human, mouse, rabbit and pig IGF-1 isoforms show multiple termination sites (Fig. 3B and Supplementary Figs. S3A, S3B and S3C) ~~(Supplementary Fig. S3A)~~. Isoform-specific PCR analysis showed that both IGF-1Eb and IGF-1Ec mRNA variants were expressed in other species characterized by an intermediate c5'ss, such as cow (*Bos taurus*), sheep (*Ovis aries*) and goat (*Capra hircus*) (Supplementary Fig. S3D ~~S3B~~).

Collectively, these results document that ~~the~~ IGF-1Eb ~~isoform~~ is more widely expressed than previously reported and point out a sharp difference in the expression of IGF-1 isoforms ~~expression~~ between human and mouse.

3.5 The mouse- and human-specific IGF-1 isoforms pattern is not dependent on the cell context

In order to analyze the species-specific splicing of exon 5 in human and mouse, we constructed minigene systems consisting of exon 4, intron 4, exon 5, part of intron 5 and exon 6 of the corresponding *IGF-1* genes (Fig. 4A~~-5A~~). Minigene splicing assay performed in four human ~~HEK293T cells (Fig. 5B)~~ and three ~~in primary~~ mouse cell lines ~~hepatocytes (Fig. 5C)~~ showed that alternative splicing of the human minigene (hIGF-1) did not generate the IGF-1Ec variant (~~Figs. 5B and 5C~~ Fig. 4B, *left panel* and Supplementary Fig. S4). ~~This result is~~ These results are in line with the barely detectable amounts of this mRNA found by quantitative real-time PCR in human tissues (Fig. 3A~~-4A~~). By contrast, the mouse minigene (mIGF-1) yielded mainly the IGF-1Ea and IGF-1Ec variants in all cell types analyzed (~~Figs. 5B and 5C~~ Fig. 4B, *right panel* and Supplementary Fig. S4), similarly to what observed in mouse tissues (Fig. 3A~~–4A~~), with splicing ~~in primary mouse hepatocytes~~ mostly shifted toward the IGF-1Ec isoform in mouse cell types ~~(Fig. 5C)~~. We suppose that these little differences might be due to the lack of part of intron 5 and/or of chromatin context, or to the presence of a different promoter in the minigene [40-41]. Collectively, these results indicate that the mouse and human IGF-1 minigenes mostly recapitulate the splicing pattern observed in tissues.

Since the ~~m~~hIGF-1 and ~~h~~mIGF-1 splicing patterns were mostly preserved regardless of the cell context, it is likely that splicing of exon 5 is not regulated by a specific human or mouse trans-acting environment. To validate this hypothesis, we forced the expression of selected splicing factors to determine their effect on the mouse and human IGF-1 minigenes. Human *IGF-1* exon 5 ~~alternative~~ splicing is regulated by competition between SRSF1 and hnRNPA1 [8-9]. Thus, we tested whether or not the differences in human and mouse exon 5 ~~alternative~~ splicing could be influenced by modulation of the expression of SRSF1, hnRNP A1 and

other similar splicing factors (Fig. 5 ~~6A and 6B~~ and Supplementary Fig. S5). As expected [8-9], overexpression of hnRNP A1 promoted splicing of the IGF-1Ea variant by the hIGF-1 minigene, whereas overexpression of SRSF1 favo~~u~~red splicing of the IGF-1Eb variant (Fig. 5B, *left panel* ~~6A~~). However, overexpression of neither splicing factor was capable of inducing the IGF-1Ec variant from the hIGF-1 minigene (Fig. 5B, *left panel* ~~6A~~). Moreover, splicing assays in presence of several other hnRNPs (A2, F, G, H, I and K), SR (SRSF3 and 7) and SR-like (TRA2α and 2β) proteins showed that none of them was capable of inducing the IGF-1Ec variant from the hIGF-1 minigene in HEK293T cells (Fig. 5B, *left panel* ~~6A~~), even though most of these splicing factors influenced the IGF-1Eb/IGF-1Ea ratio (Fig. 5B, *left panel* ~~6A~~). Similarly, depletion of hnRNPA1, SRSF1 or TRA2β did not promote splicing of IGF-1Ec isoform by the endogenous human gene (Fig. 5D, *left panel* and Supplementary Fig. S5B) and hIGF-1 minigene (Supplementary Fig. S5C). Since, splicing of IGF-1Ec is promoted by local muscle tissues damage in rodent [7], we tested whether oxidative stress condition in a human cell line (LNCaP) may promote splicing of this isoform. Notably, we observed modulation of IGF-1Eb/IGF-1Ea ratio but not splicing of IGF-1Ec (Supplementary Fig. S5D). These results strongly suggest that lack or low expression of the IGF-1Ec variant is unlikely due to deficient expression of trans-acting factors in human cells. ~~Similar conclusions could be drawn by experiments performed with the mIGF-1 minigene (Fig. 6B). In this case,~~ In the case of the mIGF-1 minigene, overexpression (Fig. 5B, *right panel* ~~5C~~) or depletion (Supplementary Fig. S5C) of most splicing factors modulated the IGF-1Ec/IGF-1Ea ratio between the two canonical variants ~~expressed by the mIGF-1 minigene~~, but none of them affected the low levels of basal expression of the IGF-1Eb variant produced by ~~this~~ the mIGF-1 minigene. Accordingly, depletion of hnRNPA1, SRSF1 or TRA2β did not promote splicing of IGF-1Eb isoform from the endogenous mouse gene (Fig. 5D, *right panel* and Supplementary Fig. S5B). These results argue against a key role played by trans-acting factors in the alternative processing of *IGF-1* exon 5 observed in both human and mouse.

3.6 The relative strength of the cryptic 5'ss of *IGF-1* exon 5 contributes to the species-specific production of the IGF-1Ec isoform in human and mouse

Given the marked difference in strength between the mouse and human ~~cryptic~~ c5'ss (Fig. 2B ~~3B~~), we hypothesized that evolutionary modifications of this splice site explains its different usage in the two species. To verify this possibility, we replaced by site-directed mutagenesis the weak human c5'ss with the strong mouse splice site (hMUT: TGvsGA) (Fig. 6A ~~7A~~). Notably, strengthening of human c5'ss partially recovered splicing of IGF-1Ec (Fig. 6B ~~7B~~). By contrast, weakening of the mouse c5'ss by reverting it to the human sequence (mMUT: GAvsTG) (Fig. 6A ~~7A~~) prevented its usage and favored selection of another cryptic 5'ss located downstream the canonical one (IGF-1Ec* in Figs. 6A ~~7A~~ and 6C ~~7C~~). These results strongly suggest that the strength of the ~~cryptic~~ c5'ss in exon 5 is a crucial factor in determining the species-specific expression of the IGF-1Ec splice variant.

We noticed that humanization of the mouse c5'ss (mMUT: GAvsTG) was not sufficient to enhance splicing of the IGF-1Eb variant from the mouse minigene (Fig. 6C ~~7C~~). Moreover, splicing was not affected by

overexpression of SRSF1, which promotes the IGF-1Eb variant from the human minigene (Supplementary Fig. S6A ~~S4B and S4C~~). Similarly, unlike with the mouse minigene, overexpression of SRSF1 did not decrease the IGF-1Ec/IGF-1Ea ratio in presence of hMUT (TGvsGA) minigene (Supplementary Fig. S6A ~~S4B~~). Altogether, these results suggest that other sequence elements, in addition to the ~~cryptic~~ c5'ss, contribute to species-specific variations of *IGF-1* alternative splicing, possibly by conferring sensitivity to trans-acting factors.

3.7 Molecular evolution of mammalian IGF-1 isoforms and prediction of IGF-1 domain structural properties

The selective pressure acting on IGF-1 domains is currently largely unknown [17, 19]. The higher variability of E-peptide sequences compared to core region sequences among mammals might reflect a less stringent functional requirement for this portion of the IGF-1 protein [17]. On the other hand, the extremely low synonymous mutation rates found on the dual-coding exon 6 across 29 mammalian species [19] suggests an enrichment of conserved regulatory elements in this region of the gene [18].

To better estimate the evolutionary pressure acting on IGF-1 domains we compared the non-synonymous (*d*N) and synonymous (*d*S) substitutions among 27 mammalian protein-coding sequences of *IGF-1* [32]. As shown in Table 1, there was substantial variation in the substitution rate within the gene regions encoding the different domains. Specifically, the Ea, Eb and Ec peptides showed a marked reduction of *d*S and only a slight increase of *d*N (significant only for Eb peptide) compared to the core region. Thus, the increase of the E-peptides dN/dS ratio was only marginally caused by an increase of dN, rather, it is likely due to a significantly smaller dS.

This result confirms and extends ~~the~~ previous findings ~~of Lin M.F. et al.~~ [19], as ~~since~~ we demonstrated strong reduction in the rate of on synonymous substitution both in the dual-coding region of exon 6 and in the single-coding region of exon 5. The strong evolutionary constraint on these exons is consistent with previous observations on alternatively spliced exons and is a signature of enrichment of functional elements, such as splicing enhancer or ~~exclusion~~ silencer elements [18, 21]. Accordingly, ~~the~~ exon 5 contains binding sites for SRSF1 and hnRNP A1 [8-9] and ~~the~~ exon 5 alternative splicing is regulated by multiple splicing factors (Fig. 5 ~~6~~). Intriguingly, recent studies demonstrated that protein-coding regions with low synonymous mutation rate are significantly enriched in Intrinsically Disordered Regions (IDRs) [42-43]. IDRs are polypeptide chains lacking a stable tertiary structure and the relaxed protein structural constraint provide an advantage when a protein-coding region simultaneously contains additional regulatory sites such as splicing enhancer and repressor sequences [43]. In order to analyze the structural properties of IGF-1 domains, we used DisCons [35], a sequence analysis tool able to identify protein disorder segments in an evolutionary context. As expected, among 27 mammalian species analyzed, the IGF-1 core region was predicted to be structured for the most part (Fig. 7 ~~8~~). Conversely all the E-peptides were predicted to be almost completely “constrained disorder” (Fig. 7 ~~8~~). Namely, they were enriched in disorder-promoting residues and there was conservation of the property of disorder across mammals. These results do not depend on the disorder predictor algorithm because core results were qualitatively replicated using PONDR® VSL2 instead of IUPred (data not shown).

Thus, during the course of mammalian evolution, the structure of IGF-1 protein has been ~~was~~ strongly conserved to maintain both the folded structure of IGF-1 core and its intrinsically disordered E-peptide tails.

4. Discussion

In the present work we provided evolutionary evidence supporting the emergence of a functional alternative exon 5 by ~~an~~ exonization of a MIR-b element in mammalian *IGF-1* intron 4 (Fig. 1C). The exon 5 gain leads to the generation of two new IGF-1 E-peptides in the mammalian lineage: Eb and Ec. Our work indicates that the Ea peptide represents the ancestral E-peptide, common to all vertebrates, whereas Eb and Ec peptides are an evolutionary novelty appeared in Mammalia after the exonization of MIR-b transposon by exaptation [44]. The MIR element ~~transposon~~ belongs to a family of transposable elements, which were actively propagating prior to ~~the~~ mammalian radiation [45]. The MIR family, with a conservation rate between 60 and 70%, is one of the most diverged transposable element families identifiable in the human genome and only recent molecular approaches have improved their chances to be identified [27, 46]. For example, MIRs have a 95% chance of being missed by RepeatMasker if they are 50 bp in length, and 50% chance of being missed if 100 bp long [46]. In our study, we ~~were~~ successfully identified ~~in identifying~~ and aligned ~~aligning~~ the MIR-b element to 103 nucleotides of the human *IGF-1* exon 5 sequence only by using the recently introduced “hmmer” search engine of RepeatMasker, which uses the hidden Markov model search tool nhmmer and the Dfam library [27-28].

Several genome-wide studies have focused on the exaptation and exonization of MIR elements and found over 100 exonized MIR elements in the human genome, located both in the UTRs and the CDSs of annotated genes [47-48]. These studies showed that exonization of retroposed sequences can occur at any time following their insertion [49-50]. Notably, both DNA sequence analyses performed in our study and mRNA analyses conducted in birds [13], fish [15] and reptiles [14] showed that *IGF-1* exon 5 is absent in these vertebrates. Therefore, the emergence of *IGF-1* exon 5 may have taken place around the mammalian diversification, which coincides with the timing of integration of the MIR elements [51].

The fact that MIR-b integrated in the *IGF-1* exon 5 in its antisense orientation and contributed the ~~to~~ splice sites and the oligopyrimidine tract may have facilitated its exonization immediately after integration [50]. Moreover, the 3'ss of *IGF-1* exon 5 is located near the highly conserved core region of MIR-b, a feature that is present in many of the exonized-MIR described so far [38, 50] and that may have also contributed to its early exonization [50]. The conservation and the relative strength of the 3'ss of *IGF-1* exon 5, together with the maintenance of open reading frame in all ~~the major~~ main mammalian branches, suggest that during evolution the creation of a functional 3'ss inside the MIR-b integrated in the *IGF-1* intron 4 was essential for successful exonization of this element [52].

Exonization of MIR-b in the mammalian *IGF-1* gene also created a ~~cryptic~~ c5'ss that allows splicing of the IGF-1Ec variant in some species. Our analyses highlighted that, in contrast to the 3'ss, the mammalian ~~cryptic~~ c5'ss shows a considerable variation in term of the number of matching nucleotides to the vertebrate

5'ss consensus, with the most marked difference observed between humans and Muridae (Figs. 2A, ~~3A~~ and 2B ~~3B~~ and Supplementary Fig. S2). This led us to hypothesize that the spliceosome ability to recognize the ~~cryptic~~ c5'ss present in ~~*IGF-1*~~ exon 5 may vary among mammalian species. Hence, the relative quantity of the IGF-1Eb and IGF-1Ec isoforms might also differ among ~~between humans, Muridae and others~~ mammals. Accordingly, we observed a differential expression pattern of IGF-1 isoforms across 9 different mammalian species analyzed (Fig. 3 ~~4A 4B~~ and Supplementary Fig. S3 ~~S3A and S3B~~). In particular, we found a marked difference in *IGF-1* splicing between humans and Muridae, with a prevalence of IGF-1Eb in humans and IGF-1Ec in Muridae. In other mammalian species ~~analysed~~, including the primate macaque, ~~the~~ expression of ~~Eb~~ the IGF-1Eb isoform was slightly higher ~~than~~ or equal to the ~~IGF1-Ec~~ IGF-1Ec. Hence, the expression of IGF-1Eb is not restricted to humans, and this splice variant should no longer be considered human-specific [1, 39].

Divergence of alternative splicing represents one of the major driving forces to shape phenotypic differences across species [53-57]. Such changes could arise from ~~the~~ divergence~~s~~ in *cis*-regulatory elements and/or *trans*-acting splicing factors [58]. Therefore, we next focused on the mechanism underlying the diversification of *IGF-1* alternative splicing between human and mouse. For this purpose we constructed minigene-based systems of the ~~mouse and human *IGF-1*~~ alternatively spliced region~~s~~. Both minigenes recapitulated the splicing pattern observed in tissues, as ~~indeed~~ the human and mouse minigenes produced very low levels of IGF-1Ec and IGF-1Eb isoforms, respectively (~~Figs. 5A and 5B~~ Fig. 4 and Supplementary Fig. S4). Nevertheless, our minigene splicing assays also showed some small differences with respect to mouse tissues. ~~For instance, alternative splicing of the mouse minigene in mouse primary hepatocytes is mostly shifted toward the IGF-1Ec isoform. We suppose that these differences might be due to alterations of the genomic and epigenetic structure of the *IGF-1* gene with respect to the minigene.~~ In particular, mouse minigene produces a small amount of the IGF-1Eb variant, whereas this isoform is minimally expressed in mouse tissues (Figs. 3A and 4B and Supplementary Fig. S4). Moreover, alternative splicing in the mouse cell lines is mostly shifted toward the IGF-1Ec isoform (Fig. 4B and Supplementary Fig. S4). We suppose that these differences might be due to alterations of the genomic and epigenetic structure of the *IGF-1* gene with respect to the minigene and/or to the different promoter present in the minigenes, as all these features have been shown to modulate the outcome of pre-mRNA splicing [40-41]. ~~For instance, the minigene lacks part of intron 5 and it might miss elements that modulate splicing of the endogenous gene. Moreover, it is known that the chromatin context, epigenetic modifications within the transcription unit and the presence of different promoters can influence alternative splicing events. Thus, it is likely that the small differences observed are due to alterations of the chromatin context in the minigene.~~ Nevertheless, our splicing assays indicate that the IGF-1 minigenes are reliable tools for the analysis of the splicing regulation of the corresponding endogenous genes. By using these minigenes, we tested whether *cis*-acting elements and/or *trans*-acting factors were responsible for the different splicing patterns observed *in vivo* between mouse and human IGF-1 mRNA. Alternative splicing outcome is determined by competition between splice sites in different exons and is influenced by *trans*-acting splicing factors, which in turn influence splice site

recognition by the spliceosome [59]. As a consequence, both changes in the sequence of *cis*-acting elements and in the expression levels of splicing factors can modulate usage of specific exons [58, 60]. Our results argue against changes in expression of specific splicing factors as a key determinant between the splicing differences across species. Indeed, we found that splicing of the mouse and human IGF-1 minigenes is mostly preserved across different cell type contexts (~~Figs. 5A, 5B and 5C~~ Fig. 4B and Supplementary Figure S4). Moreover, by ~~forcing~~ modulating the expression of specific splicing factors it was not possible to promote splicing of the IGF-1Ec variant from the human minigene ~~(hIGF-1)~~ (Fig. 5B, *left panel* and Supplementary Fig. S5C) or endogenous human gene (Fig. 5D, *left panel* and Supplementary Fig. S5B). Similarly, neither the overexpression (Fig. 5B ~~5C~~, *right panel*) nor the siRNA-mediated depletion (Fig. 5D, *right panel* and Supplementary Figs. S5B and S5C) of splicing factors affected the expression of mouse IGF-1Eb isoform. ~~nor that of IGF-1Eb from the mouse minigene (mIGF-1)~~. We tested several hnRNPs and SR proteins, including SRSF1 and hnRNP A1 that were previously shown to ~~modulate alternative~~ regulate splicing of human *IGF-1* [8-9]. Most of these *trans*-acting factors modulated the human IGF-1Eb/IGF-1Ea and mouse IGF-1Ec/IGF-1Ea ratio ~~between the variants produced by the minigene~~, but none of them promoted the variant typical of the other species. Thus, our results suggest that the mechanism regulating this process does not rely on differential expression of species-specific *trans*-acting factors.

An alternative possibility is the presence of *cis*-acting elements conferring species-specific features to the exon 5 region. This hypothesis is in line with a recent study showing that most vertebrate specie-specific splicing patterns are primarily *cis*-directed [54, 58]. Accordingly, our analysis revealed that the consensus bases within the ~~cryptic~~ c5'ss are highly divergent among mammalian species (Fig. 2A ~~3A~~ and Supplementary Fig. S2). This observation led us to hypothesize that variation in the ~~cryptic~~ c5'ss strength might represent a potential source of species-specific splicing pattern. In support of this hypothesis, ~~we found that~~ strengthening the human c5'ss by replacing it with the mouse one ~~5'ss~~ (hMUT: TG*vs*GA) was sufficient to promote usage of this splice site and to yield IGF-1Ec variant (Figs. 6A and 6B ~~7A and 7B~~). On the other hand, weakening the mouse c5'ss by swapping it with the human sequence (mMUT: GA*vs*TG) completely prevented its usage (Figs. 6A and 6C ~~7A and 7C~~). Nevertheless, our results also indicate that other *cis*-acting elements likely contribute to the splicing pattern typically observed in tissues. First, we found that “murinization” of the human ~~cryptic~~ c5’ss only partially promoted splicing of the IGF-1Ec variant, but did not restore the IGF-1Ec/IGF-1Ea ratio observed with the mouse minigene and in mouse tissues (Figs. 3A and 6B ~~7B~~). ~~This result suggests that additional sequence elements differing between human and mouse exon 5 might also contribute to its alternative splicing, perhaps by recruiting more or less efficiently specific splicing factors. In support of the presence of additional *cis*-acting elements is also the analysis of the mutant mouse IGF-1 minigene.~~ Furthermore, “humanization” of the mouse ~~cryptic~~ c5'ss (mMUT: GA*vs*TG) prevented splicing of the IGF-1Ec variant, but did not promote IGF-1Eb variant (Figs. 6A and 6C ~~7A and 7C~~). By contrast, we observed the selection of another cryptic splice site (*5'ss) located 38 nt downstream of the canonical one (Figs. 6A and 6C ~~7A and 7C~~). Notably, this cryptic splice site is not conserved in the human gene, thus explaining the generation of the unusual IGF-1 mRNA variant instead of IGF-1Eb from the mMUT

(GAvsTG) minigene. Moreover, overexpression of SRSF1 did not promote splicing of the IGF-1Eb variant from the mMUT (GA*vs*TG) minigene (Supplementary Fig. S6A ~~S4B~~). SRSF1-dependent splicing of human IGF-1Eb requires its binding to a purine-rich enhancer located in exon 5 [8-9]. Sequence alignment of human and mouse exon 5 revealed that the SRSF1 binding site is not conserved in the mouse (Supplementary Fig. S6B ~~S4D~~), possibly explaining the different behavior of SRSF1 in presence of the mouse minigene. These results strongly indicate that the strength of the ~~cryptic~~ c5'ss in exon 5 is a prerequisite for *IGF-1* alternative splicing diversification between species, but also highlight that the co-evolution of additional ~~*cis*-acting~~ exonic and intronic *cis*-regulatory elements contribute to such diversification [40-41]. Notably, our phylogenetic analysis of the cryptic 5'ss indicates that this splice site is relative strong not only in mouse but also in rat and all members of Cricetidae family (Fig. 2B~~–3B~~ and Supplementary Fig. S2). It is possible that strengthening of this cryptic c5'ss in a common ancestor of Muridae and Cricetidae, as well as the emergence of other splicing regulatory motifs, have determined the marked shift toward production of the IGF-1Ec isoform in these species and possibly a ~~the acquisition of~~ rodent-specific gain of functional properties. ~~Accordingly~~ Interestingly, a recent study comparing the IGF-1 isoform expression in both mouse and human muscle samples at different ages, showed that the age-related change of IGF-1 splice variants is species-dependent~~,~~ and, unlike the mouse, only the human IGF-1Eb isoform was regulated during ageing in skeletal muscles [61]. Thus, the results obtained with murine models on IGF-1Ec/mechano growth factor regulation must be interpreted with ~~more than usual~~ caution and additional studies comparing IGF-1 expression levels across species are needed to clarify the functional role of the IGF-1 isoforms.

~~The regulation of IGF-1 is a critically important factor in several conditions such as the pathophysiology of several cancers, or the mitogenic and myogenic processes during muscle development, regeneration or hypertrophy .~~ Compared with mature IGF-1 relatively little is known about the mechanism of action of the different E peptides [1, 10-12]. ~~How the E-peptides affect the actions of IGF-1 is still under debate.~~ From an evolutionary point of view the unchanged persistence over long evolutionary periods of MIRb-derived *IGF-1* exon 5 implies its functional relevance [52, 62]. Moreover, E-peptides are protein-coding regions in which synonymous mutation rates are extremely low compared to IGF-1 core (Table 1), indicating additional sequence constraints beyond those dictated by the structure and function of the proteins. These additional constraints probably stem from the demands of regulatory sites involved in transcript splicing (Fig. 5) and [8-9, 18], nevertheless the presence of other regulatory sites such as specific RNA secondary structures and microRNA targets [20-21] cannot be excluded. The systematic analysis of the role of synonymous variants and the comparative splicing evaluation of mammalian sequence divergences [63-64] will help to characterize the functional roles of these regulatory elements.

Analysis of structural properties of IGF-1 domains adds a new layer of complexity to the function of E-peptides. Indeed, our work represents the first evidence that E-peptides contain disorder-promoting amino acids and that there is substantial evolutionary pressure to keep the different E-peptides as intrinsically disordered regions (IDRs) (Fig. 7~~–8~~). There is a growing interest on IDRs since they are usually enriched in posttranslational modification sites and may exert a number of regulatory functions on their “host” protein

[65-66]. Intriguingly, we and others recently demonstrated that the IGF-1 protein retaining C-terminal E peptides are the predominant forms produced intracellularly, instead of mature IGF-1, and are subjected to extensive post-translational modifications [67-68], further hinting to their functional relevance. ~~Intriguingly, in murine skeletal muscle, the IGF-1 protein retaining the C-terminal Ea peptide was the predominant form, instead of mature IGF-1, and exists in both non-glycosylated and glycosylated forms. Moreover overexpression of IGF-1 lacking any E-peptide did not promote muscle hypertrophy in young mice, further hinting to its functional relevance.~~

Exonization of previously non-coding sequences, together with creation of novel domain combinations, has been directly related to the increase of organismal complexity [66, 69-71]. Accordingly, we propose that MIR-b exonization during mammalian evolution determined the IGF-1 exon 5 gain and hence the addition of two new disordered tails to IGF-1: the Eb and Ec-tails. Thus, novel exon and E-peptide combinations may have created new layers of regulation to mature IGF-1 in mammalian species. Targeting these regulatory elements may represent a new strategy to control IGF-1 bioavailability in different physiological/pathological conditions, with particular attention to possible differences between species.

Acknowledgement

We wish to thank Dr. Franco Berrino of the Department of Predictive & Preventive Medicine, National Institute of Cancer, Milan and Dr. Giorgio Arnaldi, Department of Internal Medicine, Polytechnic University of Marche Region, Ancona, Italy for a critical reading of the manuscript. This research was supported by the Italian Ministry of Health, "Ricerca finalizzata 2009" (grant number: RF-2009-1532789) and “Ricerca Finalizzata 2011 (grant number: GR-2011-02348423); by the Associazione Italiana Ricerca sul Cancro (AIRC IG14581) and by Telethon (CGP14095).

References

[1] E.R. Barton, The ABCs of IGF-I isoforms: impact on muscle hypertrophy and implications for repair, Appl Physiol Nutr Metab 31 (2006) 791-797.

[2] L. Temmerman, E. Slonimsky, N. Rosenthal, Class 2 IGF-1 isoforms are dispensable for viability, growth and maintenance of IGF-1 serum levels, Growth Horm IGF Res 20 (2010) 255-263.

[3] A.M. Oberbauer, The Regulation of IGF-1 Gene Transcription and Splicing during Development and Aging, Front Endocrinol (Lausanne) 4 (2013) 39.

[4] S.L. Chew, P. Lavender, A.J. Clark, R.J. Ross, An alternatively spliced human insulin-like growth factor-I transcript with hepatic tissue expression that diverts away from the mitogenic IBE1 peptide, Endocrinology 136 (1995) 1939-1944.

[5] C.T. Roberts, Jr., S.R. Lasky, W.L. Lowe, Jr., W.T. Seaman, D. LeRoith, Molecular cloning of rat insulin-like growth factor I complementary deoxyribonucleic acids: differential messenger ribonucleic acid processing and regulation by growth hormone in extrahepatic tissues, Mol Endocrinol 1 (1987) 243-248.

[6] S. Yang, M. Alnaqeeb, H. Simpson, G. Goldspink, Cloning and characterization of an IGF-1 isoform expressed in skeletal muscle subjected to stretch, J Muscle Res Cell Motil 17 (1996) 487-495.

[7] M. Hill, G. Goldspink, Expression and splicing of the insulin-like growth factor gene in rodent muscle is associated with muscle satellite (stem) cell activation following local tissue damage, J Physiol 549 (2003) 409-418.

[8] P.J. Smith, E.L. Spurrell, J. Coakley, C.J. Hinds, R.J. Ross, A.R. Krainer, S.L. Chew, An exonic splicing enhancer in human IGF-I pre-mRNA mediates recognition of alternative exon 5 by the serine-arginine protein splicing factor-2/alternative splicing factor, Endocrinology 143 (2002) 146-154.

[9] A. Clery, R. Sinha, O. Anczukow, A. Corrionero, A. Moursy, G.M. Daubner, J. Valcarcel, A.R. Krainer, F.H. Allain, Isolated pseudo-RNA-recognition motifs of SR proteins can regulate splicing using a noncanonical mode of RNA recognition, Proc Natl Acad Sci U S A 110 (2013) E2802-2811.

[10] P. Rotwein, Editorial: the fall of mechanogrowth factor?, Mol Endocrinol 28 (2014) 155-156.

[11] A. Armakolas, M. Kaparelou, A. Dimakakos, E. Papageorgiou, N. Armakolas, A. Antonopoulos, C. Petraki, M. Lekarakou, P. Lelovas, M. Stathaki, C. Psarros, I. Donta, P.S. Galanos, P. Msaouel, V.G. Gorgoulis, M. Koutsilieris, Oncogenic Role of the Ec Peptide of the IGF-1Ec Isoform in Prostate Cancer, Mol Med 21 (2015) 167-179.

[12] R.W. Matheny, Jr., B.C. Nindl, M.L. Adamo, Minireview: Mechano-growth factor: a putative product of IGF-I gene expression involved in tissue repair and regeneration, Endocrinology 151 (2010) 865-875.

[13] Y. Kajimoto, P. Rotwein, Structure of the chicken insulin-like growth factor I gene reveals conserved promoter elements, J Biol Chem 266 (1991) 9724-9731.

[14] A.M. Sparkman, T.S. Schwartz, J.A. Madden, S.E. Boyken, N.B. Ford, J.M. Serb, A.M. Bronikowski, Rates of molecular evolution vary in vertebrates for insulin-like growth factor-1 (IGF-1), a pleiotropic locus that regulates life history traits, Gen Comp Endocrinol 178 (2012) 164-173.

[15] D.M. Tiago, V. Laize, M.L. Cancela, Alternatively spliced transcripts of Sparus aurata insulin-like growth factor 1 are differentially expressed in adult tissues and during early development, Gen Comp Endocrinol 157 (2008) 107-115.

[16] M. Reinecke, C. Collet, The phylogeny of the insulin-like growth factors, Int Rev Cytol 183 (1998) 1-94.

[17] M. Wallis, New insulin-like growth factor (IGF)-precursor sequences from mammalian genomes: the molecular evolution of IGFs and associated peptides in primates, Growth Horm IGF Res 19 (2009) 12-23.
[18] Y. Xing, C. Lee, Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences, Proc Natl Acad Sci U S A 102 (2005) 13526-13531.
[19] M.F. Lin, P. Kheradpour, S. Washietl, B.J. Parker, J.S. Pedersen, M. Kellis, Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes, Genome Res 21 (2011) 1916-1928.
[20] J.V. Chamary, J.L. Parmley, L.D. Hurst, Hearing silence: non-neutral evolution at synonymous sites in mammals, Nat Rev Genet 7 (2006) 98-108.
[21] Y. Xing, C. Lee, Can RNA selection pressure distort the measurement of Ka/Ks?, Gene 370 (2006) 1-5.
[22] M. Blanchette, W.J. Kent, C. Riemer, L. Elnitski, A.F. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green, D. Haussler, W. Miller, Aligning multiple genomic sequences with the threaded blockset aligner, Genome Res 14 (2004) 708-715.
[23] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res 32 (2004) 1792-1797.
[24] J.D. Thompson, T.J. Gibson, D.G. Higgins, Multiple sequence alignment using ClustalW and ClustalX, Curr Protoc Bioinformatics Chapter 2 (2002) Unit 2 3.
[25] G. Yeo, C.B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, J Comput Biol 11 (2004) 377-394.
[26] J. Jurka, Repbase update: a database and an electronic journal of repetitive elements, Trends Genet 16 (2000) 418-420.
[27] T.J. Wheeler, J. Clements, S.R. Eddy, R. Hubley, T.A. Jones, J. Jurka, A.F. Smit, R.D. Finn, Dfam: a database of repetitive DNA based on profile hidden Markov models, Nucleic Acids Res 41 (2013) D70-82.
[28] T.J. Wheeler, S.R. Eddy, nhmmer: DNA homology search with profile HMMs, Bioinformatics 29 (2013) 2487-2489.
[29] G. Annibalini, M. Guescini, D. Agostini, R.D. Matteis, P. Sestili, P. Tibollo, M. Mantuano, C. Martinelli, V. Stocchi, The expression analysis of mouse interleukin-6 splice variants argued against their biological relevance, BMB Rep 45 (2012) 32-37.
[30] S. Barik, Site-directed mutagenesis in vitro by megaprimer PCR, Methods Mol Biol 57 (1996) 203-215.
[31] P. Bielli, R. Busa, S.M. Di Stasi, M.J. Munoz, F. Botti, A.R. Kornblihtt, C. Sette, The transcription factor FBI-1 inhibits SAM68-mediated BCL-X alternative splicing and apoptosis, EMBO Rep 15 (2014) 419-427.
[32] S.L. Pond, S.D. Frost, S.V. Muse, HyPhy: hypothesis testing using phylogenies, Bioinformatics 21 (2005) 676-679.
[33] S.V. Muse, Estimating synonymous and nonsynonymous substitution rates, Mol Biol Evol 13 (1996) 105-114.
[34] R. Szklarczyk, J. Heringa, S.K. Pond, A. Nekrutenko, Rapid asymmetric evolution of a dual-coding tumor suppressor INK4a/ARF locus contradicts its function, Proc Natl Acad Sci U S A 104 (2007) 12807-12812.
[35] M. Varadi, M. Guharoy, F. Zsolyomi, P. Tompa, DisCons: a novel tool to quantify and classify evolutionary conservation of intrinsic protein disorder, BMC Bioinformatics 16 (2015) 153.
[36] Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, Bioinformatics 21 (2005) 3433-3434.

[37] R. Sorek, The birth of new exons: mechanisms and evolutionary consequences, RNA 13 (2007) 1603-1608.
[38] I. Vorechovsky, Transposable elements in disease-associated cryptic exons, Hum Genet 127 (2010) 135-154.
[39] A.M. Oberbauer, J.M. Belanger, G. Rincon, A. Canovas, A. Islas-Trejo, R. Gularte-Merida, M.G. Thomas, J.F. Medrano, Bovine and murine tissue expression of insulin like growth factor-I, Gene 535 (2014) 101-105.
[40] S. Naftelberg, I.E. Schor, G. Ast, A.R. Kornblihtt, Regulation of alternative splicing through coupling with transcription and chromatin structure, Annu Rev Biochem 84 (2015) 165-198.
[41] N.N. Singh, M.N. Lawler, E.W. Ottesen, D. Upreti, J.R. Kaczynski, R.N. Singh, An intronic structure enabled by a long-distance interaction serves as a novel target for splicing correction in spinal muscular atrophy, Nucleic Acids Res 41 (2013) 8144-8165.
[42] E. Kovacs, P. Tompa, K. Liliom, L. Kalmar, Dual coding in alternative reading frames correlates with intrinsic protein disorder, Proc Natl Acad Sci U S A 107 (2010) 5429-5434.
[43] M. Macossay-Castillo, S. Kosol, P. Tompa, R. Pancsa, Synonymous constraint elements show a tendency to encode intrinsically disordered protein segments, PLoS Comput Biol 10 (2014) e1003607.
[44] J. Brosius, S.J. Gould, On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA", Proc Natl Acad Sci U S A 89 (1992) 10706-10710.
[45] L. Lin, P. Jiang, S. Shen, S. Sato, B.L. Davidson, Y. Xing, Large-scale analysis of exonized mammalian-wide interspersed repeats in primate genomes, Hum Mol Genet 18 (2009) 2204-2214.
[46] A.P. de Koning, W. Gu, T.A. Castoe, M.A. Batzer, D.D. Pollock, Repetitive elements may comprise over two-thirds of the human genome, PLoS Genet 7 (2011) e1002384.
[47] B. Mersch, N. Sela, G. Ast, S. Suhai, A. Hotz-Wagenblatt, SERpredict: detection of tissue- or tumor-specific isoforms generated through exonization of transposable elements, BMC Genet 8 (2007) 78.
[48] N. Sela, B. Mersch, N. Gal-Mark, G. Lev-Maor, A. Hotz-Wagenblatt, G. Ast, Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome, Genome Biol 8 (2007) R127.
[49] R. Baertsch, M. Diekhans, W.J. Kent, D. Haussler, J. Brosius, Retrocopy contributions to the evolution of the human genome, BMC Genomics 9 (2008) 466.
[50] M. Krull, M. Petrusma, W. Makalowski, J. Brosius, J. Schmitz, Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs), Genome Res 17 (2007) 1139-1145.
[51] F.S. de Souza, L.F. Franchini, M. Rubinstein, Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong?, Mol Biol Evol 30 (2013) 1239-1251.
[52] J. Schmitz, J. Brosius, Exonization of transposed elements: A challenge and opportunity for evolution, Biochimie 93 (2011) 1928-1934.
[53] H. Keren, G. Lev-Maor, G. Ast, Alternative splicing and evolution: diversification, exon definition and function, Nat Rev Genet 11 (2010) 345-355.
[54] N.L. Barbosa-Morais, M. Irimia, Q. Pan, H.Y. Xiong, S. Gueroussov, L.J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C.M. Misquitta-Ali, M.D. Wilson, P.M. Kim, D.T. Odom, B.J. Frey, B.J. Blencowe, The evolutionary landscape of alternative splicing in vertebrate species, Science 338 (2012) 1587-1593.
[55] J. Merkin, C. Russell, P. Chen, C.B. Burge, Evolutionary dynamics of gene and isoform regulation in Mammalian tissues, Science 338 (2012) 1593-1599.
[56] T. Lappalainen, M. Sammeth, M.R. Friedlander, P.A. t Hoen, J. Monlong, M.A. Rivas, M. Gonzalez-Porta, N. Kurbatova, T. Griebel, P.G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlof, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M.

Sultan, G. Bertier, D.G. MacArthur, M. Lek, E. Lizano, H.P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S.B. Montgomery, P. Donnelly, M.I. McCarthy, P. Flicek, T.M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S.E. Antonarakis, R. Hasler, A.C. Syvanen, G.J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigo, I.G. Gut, X. Estivill, E.T. Dermitzakis, Transcriptome and genome sequencing uncovers functional variation in humans, Nature 501 (2013) 506-511.

[57] A. Necsulea, H. Kaessmann, Evolutionary dynamics of coding and non-coding transcriptomes, Nat Rev Genet 15 (2014) 734-748.

[58] Q. Gao, W. Sun, M. Ballegeer, C. Libert, W. Chen, Predominant contribution of cis-regulatory divergence in the evolution of mouse alternative splicing, Mol Syst Biol 11 (2015) 816.

[59] D.L. Black, Mechanisms of alternative pre-messenger RNA splicing, Annu Rev Biochem 72 (2003) 291-336.

[60] M. Chen, J.L. Manley, Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches, Nat Rev Mol Cell Biol 10 (2009) 741-754.

[61] M. Sandri, L. Barberi, A.Y. Bijlsma, B. Blaauw, K.A. Dyar, G. Milan, C. Mammucari, C.G. Meskers, G. Pallafacchina, A. Paoli, D. Pion, M. Roceri, V. Romanello, A.L. Serrano, L. Toniolo, L. Larsson, A.B. Maier, P. Munoz-Canoves, A. Musaro, M. Pende, C. Reggiani, R. Rizzuto, S. Schiaffino, Signalling pathways regulating muscle mass in ageing skeletal muscle: the role of the IGF1-Akt-mTOR-FoxO pathway, Biogerontology 14 (2013) 303-323.

[62] V. Gotea, W. Makalowski, Do transposable elements really contribute to proteomes?, Trends Genet 22 (2006) 260-267.

[63] F. Pagani, M. Raponi, F.E. Baralle, Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution, Proc Natl Acad Sci U S A 102 (2005) 6368-6372.

[64] Z.E. Sauna, C. Kimchi-Sarfaty, Understanding the contribution of synonymous mutations to human disease, Nat Rev Genet 12 (2011) 683-691.

[65] M. Buljan, G. Chalancon, S. Eustermann, G.P. Wagner, M. Fuxreiter, A. Bateman, M.M. Babu, Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks, Mol Cell 46 (2012) 871-883.

[66] G. Thieulin-Pardo, L. Avilan, M. Kojadinovic, B. Gontero, Fairy "tails": flexibility and function of intrinsically disordered extensions in the photosynthetic world, Front Mol Biosci 2 (2015) 23.

[67] J. Durzynska, A. Philippou, B.K. Brisson, M. Nguyen-McCarty, E.R. Barton, The pro-forms of insulin-like growth factor I (IGF-I) are predominant in skeletal muscle and alter IGF-I receptor activation, Endocrinology 154 (2013) 1215-1224.

[68] M. De Santi, G. Annibalini, E. Barbieri, A. Villarini, L. Vallorani, S. Contarelli, F. Berrino, V. Stocchi, G. Brandi, Human IGF1 pro-forms induce breast cancer cell proliferation via the IGF1 receptor, Cell Oncol (Dordr) (2015).

[69] M. Buljan, A. Frankish, A. Bateman, Quantifying the mechanisms of domain gain in animal proteins, Genome Biol 11 (2010) R74.

[70] P.R. Romero, S. Zaidi, Y.Y. Fang, V.N. Uversky, P. Radivojac, C.J. Oldfield, M.S. Cortese, M. Sickmeier, T. LeGall, Z. Obradovic, A.K. Dunker, Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms, Proc Natl Acad Sci U S A 103 (2006) 8390-8395.

[71] V.N. Uversky, The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini, FEBS Lett 587 (2013) 1891-1901.

Figure captions

Figure 1. Schematic representation of the *IGF-1* gene (A), its splice variants (B) and the exonization of MIR-b in the *IGF-1* gene (C). (A) Map of the *IGF-1* gene showing exons (*boxes*), introns (*solid lines*), splicing options (*dashed lines*), cryptic 5' splice site (*c5'ss*) in exon 5 and poly(A) sites (*pA*). The graph is not drawn to scale. (B) Splice variants of the *IGF-1* gene; human *IGF-1* NCBI RefSeq transcripts: class I IGF-1Ea (NM_000618), class I IGF-1Eb (NM_001111285), class I IGF-1Ec (NM_001111283) and class II IGF-1Ea (NM_001111284). (C) Pairwise alignment of the human *IGF-1* gene sequence and the MIR-b inverse consensus sequence created by RepeatMasker using "nhmmer" search engine (intronic nucleotides in lower-case letters, exonic nucleotides in upper-case letters). The polypyrimidine tract and exon 5 splice sites are shown in bold in *IGF-1* sequence; the bold region of MIR-b sequence corresponds to the 65-nt conserved central domain of MIR-b sequence. The middle line shows exact matches (spaces), gaps (dashes) and mismatches (i=transition; v=transversion).

~~**Figure 1. Schematic representation of the *IGF-1* gene and its splice variants.** Exons (*boxes*), introns (*lines*), cryptic 5' splice site (c5'ss) in exon 5 (*upper panel*) and splice variants of the *IGF-1* gene (*lower panel*) are indicated.~~

~~**Figure 2. *IGF-1* exon 5 is an exapted retroposon of the MIR-b family.** Pairwise alignment of the human *IGF-1* exon 5 sequence and the MIR-b inverse consensus sequence created by RepeatMasker using "nhmmer" search engine (intronic nucleotides in lower-case letters, exonic nucleotides in upper-case letters). The polypyrimidine tract and exon 5 splice sites are shown in bold in *IGF-1* sequence; the bold region of MIR-b sequence correspond to the 65-nt conserved central domain of MIR-b sequence. The middle line shows exact matches (spaces), gaps (dashes) and mismatches (i=transition; v=transversion).~~

~~**Figure 3. (A) Multiple alignment of MIR-derived exon 5 of *IGF-1* gene among mammals.**~~ **Figure 2. Comparative analyses of exon 5 splice sites among mammals.** (A) Multiple alignment of MIR-derived exon 5 of *IGF-1* gene among mammals. Partial intron 4 and exon 5 sequences is shown (intronic nucleotides in lower-case letters, exonic nucleotides in upper-case letters). The nucleotides that did not match the 3'ss and 5'ss consensus sequences are highlighted in grey (y=purine; r=pyrimidine). Nucleotides conservation is marked at the lower edge with asterisks indicating full conservation relative to the MIR-b consensus sequence (lower row). The percentage of nucleotide identity between mammalian sequences and MIR-b consensus sequence is indicated under "MIR-b identity". (B) Splice site score of the 3'ss and c5'ss of *IGF-1* exon 5 was calculated using the MaxEntScan algorithm [25].

Figure 3. Expression pattern of IGF-1 mRNA isoforms within mammalian species. (A) Quantification of IGF-1Ea, IGF-1Eb and IGF-1Ec isoforms by real time RT-PCR on skeletal muscle, adipose tissue and liver of Human, Macaque and Mouse. Percentage mean (+/-SD) of IGF-1 isoforms was calculated as: [(mRNA of single isoform/(mRNAs of IGF1Ea + IGF1Eb + IGF1Ec)*100]. n=2-4. (B) Agarose gel of 3' RACE-PCR analysis performed in liver tissues from different mammalian species. PCR products were eluted from gel, subcloned and sequenced (See Supplementary Figs. S3A, S3B and S3C). A schematic representation of IGF-1 gene and the relative position of forward primers used in the 3'RACE-PCR is also shown. See Materials and Methods section for universal reverse primer information. Diagrammatic representation of splice variants is given on the left of the gel; their sizes (bp) are indicated on the right.

~~**Figure 4. Expression pattern of IGF-1 mRNA isoforms within mammalian species.** (A) Quantification of IGF-1Ea, IGF-1Eb and IGF-1Ec isoforms by real time RT-PCR on skeletal muscle, adipose tissue and liver of Human, Macaque and Mouse. Percentage of IGF-1 isoforms was calculated as: [(mRNA of single isoform/(mRNAs of IGF1Ea + IGF1Eb + IGF1Ec)*100]. The percentage mean (+/- S.D.) of IGF-1Ea is shown. (B) The 3' RACE-PCR products obtained from liver tissues of mammalian species. The 3' RACE-PCR products were loaded on 4.0% agarose gel and DNA fragments were eluted from gel, subcloned and sequenced. Dashes mark: 1, PCR fragment correspondent to IGF-1Ea isoform; 2, PCR fragment correspondent to IGF-1Eb isoform; and 3, PCR fragment correspondent to IGF-1Ec isoform.~~

Figure 4. The expression pattern of human and mouse IGF-1 minigenes. (A) Schematic representation of IGF-1 minigenes. Exons (*boxes*), introns (*lines*), and cryptic 5' splice site (c5'ss) in exon 5 are indicated (*left panel*). A diagram of primers position used for RT-PCR analysis and the expected PCR products are also shown (right panel). A minigene-specific forward primer (P1) was used to amplify minigene-derived IGF1 isoforms. Primers P1-P4 were used to amplify a constant region of minigene-derived IGF-1 mRNA; primers P1-P3 were used to amplify IGF-1Ea and IGF-1Ec isoforms;

primers P1-P2 were used to amplify IGF-1Eb isoform. All oligonucleotide sequences are listed in Supplementary Table S1. (B) RT-PCR analysis of splicing assay performed in human HEK293T cells (*left panel*) and primary mouse hepatocytes (*right panel*) in presence of human (hIGF-1) or mouse (mIGF-1) minigenes. Untransfected cells (-) or cells transfected with the empty vector (vect.) were used as a PCR control. L34 was used as loading control.

~~**Figure 5. The expression pattern of human and mouse IGF-1 minigenes.** (A) Schematic representation of IGF-1 minigenes. Exons (*boxes*), introns (*lines*), and cryptic 5' splice site (c5'ss) in exon 5 are indicated (*upper panel*). A diagram of primers position used for RT-PCR analysis is also shown (*lower panel*): primers P1-P3 were used to amplify mouse and human IGF-1Ea and IGF-1Ec isoforms; primers P1-P2 were used to amplify human and mouse IGF-1Eb isoform; primers P1-P2-P3 were used to amplify human IGF-1Ea and IGF-1Eb isoforms. All oligonucleotide sequences are listed in Supplementary Table S1. (B and C) RT-PCR analysis of *in vivo* splicing assay performed in human HEK293T cells (B) and primary mouse hepatocytes (C) in presence of human (hIGF-1) or mouse (mIGF-1) minigenes.~~

Figure 5. The IGF-1 splicing pattern of human and mouse minigenes after the splicing factors overexpression (A,B) and effect of siRNA-mediated depletion of splicing factors on endogenous IGF-1 isoforms expression in human and mouse cells (C,D). (A) Diagram of primers position used to amplify minigene-derived IGF1 isoforms. (B) RT-PCR analysis of splicing assay performed in HEK293T cell line in presence of human (*left panel*) or mouse (*right panel*) IGF-1 minigenes and the indicated splicing factors. The bar graph shows the ratio of densitometric analysis between IGF-1Eb and IGF-1Ea (*left panel*) or IGF-1Ec and IGF-1Ea (*right panel*) isoforms (mean ± SD, n=3). C) Diagram of primers position used to amplify endogenous gene-derived IGF-1 isoforms. (D) RT-PCR analysis of splicing assay of endogenous IGF-1 gene performed in human LNCaP cells (left panel) and in mouse NIH/3T3 cells (*left panel*) depleted with the indicated splicing factors.

~~**Figure 6. The IGF-1 splicing pattern of human and mouse minigenes after the splicing factors overexpression.** A,B) RT-PCR analysis of splicing assay of human (A) or mouse (B) IGF-1 minigenes performed in HEK293T cell line in presence of indicated splicing factors. Primers P1-P3 were used to amplify mouse and human IGF-1Ea and IGF-1Ec isoforms (A,B). Primers P1-P2-P3 were used to amplify human IGF-1Ea and IGF-1Eb isoforms (A); primers P1-P2 were used to amplify mouse IGF-1Eb isoform (B); primers P1-P4 were used to amplify total human and mouse *IGF-1*. The primer positions are indicated in Fig. 5. The bar graph shows the ratio of densitometric analysis between IGF-1Eb and IGF-1Ea (A) or IGF-1Ec and IGF-1Ea (B) isoforms (mean ± SD, n=3).~~

~~**Figure 7.**~~ **Figure 6. Comparative analysis of human and mouse IGF-1 alternative splicing on wild-type and mutated minigenes.** (A) Scheme of human (hMUT: TG*vs*GA) and mouse (mMUT: GA*vs*TG) mutant minigenes. The mutated bases are underlined. Position of the canonical (c5'ss) and non-canonical (*5'ss) cryptic 5' splice site used in mMUT minigene are also shown. (B and C) RT-PCR analysis of splicing assay performed in human HEK293T cells in presence of human (B) and mouse (C) IGF-1 mutant minigenes. RT-PCR was performed using primers as indicated in Fig. 4.

Figure 7 ~~**8.**~~ **The sliding-window plot of the disorder conservation score of IGF-1 domains.** The x axis represents the codon positions and the y axis shows disorder conservation profiles for IGF-1 core (black line), Ea (dashed line), Eb (dotted line) and Ec (gray line) peptides. The disorder conservation score for IGF-1 domains was calculated using the Disorder Conservation (DisCons) software with default parameters [35]. Low values correspond to order, high values to disorder.

Supplementary data captions

Supplementary Figure S1. Partial sequence alignment of IGF-1 genomic sequences among 23 amniota vertebrata retrieved from Ensembl database (23 amniota vertebrates Pecan). (A) The nucleotide sequences surrounding the alternatively spliced exon 6 were shown. The exon 6 sequences are in upper case letters and highlighted in gray, the intron sequences are in lower case letters. The 3'splice site of exon 6 is indicated by bold letters. Sequence gaps are indicated by dashed lines. Human chromosomal location:GRCh38:12:102537063:102537363. (B) The nucleotide sequences surrounding the alternatively spliced exon 5 were shown. The exon 5 sequences are in upper case letters and highlighted in black and gray: the black nucleotides designated a part of exon 5 common to IGF-1Eb and IGF-1Ec variants while the grey region is included only in the IGF-1Eb isoform; the intron sequences are in lower case letters .The 3' splice site and the cryptic 5' splice site (c5'ss) of exon 5 are shown in bold. Sequence gaps are indicated by dashed lines. Human chromosomal location:GRCh38: 12:102458404:102458924. Ensembl Genomes: Human (*Homo sapiens*) GRCh38.p5; Mouse (*Mus musculus*) GRCm38.p4; Rat (*Rattus norvegicus*) Rnor_6.0; Rabbit (*Oryctolagus cuniculus*) OryCun2.0; Chimpanzee (*Pan troglodytes*) CHIMP2.1.4 ; Gorilla (*Gorilla gorilla gorilla*) gorGor3.1; Orangutan (*Pongo abelii*) PPYG2; Macaque (*Macaca mulatta*) MMUL 1.0 ; Marmoset (*Callithrix jacchus*) C_jacchus3.2.1; Cat (*Felis catus*) Felis_catus_6.2; Dog (*Canis lupus familiaris*) CanFam3.1; Horse (*Equus caballus*) Equ Cab 2; Cow (*Bos taurus*) UMD3.1; Sheep (*Ovis aries*) Oar_v3.1; Pig (*Sus scrofa*) Sscrofa10.2; Opossum (*Monodelphis domestica*) monDom5 ; Platypus (*Ornithorhynchus anatinus*) OANA5; Turkey (*Meleagris gallopavo*) Turkey_2.01; Chicken (*Gallus gallus*) Galgal4; Zebra Finch (*Taeniopygia guttata*) taeGut3.2.4; Anole lizard (*Anolis carolinensis*) AnoCar2.0.

~~**Supplementary Figure S1. ClustalX alignment of IGF-1 sequences along the IGF-1 exon 6 and exon 5 regions among 23 representative amniota vertebrates.** > Alignment of IGF-1 sequences along the **exon 6** region of 23 amniota vertebrata (chromosome:GRCh38:12:102537063:102537363: -1) >Alignment of IGF-1 sequences along the **exon 5** region of 23 amniota vertebrata (chromosome:GRCh38:12:102458404:102458924:-1)~~

Supplementary Figure S2. Analyses of the 3′ and the 5′ splice-site signals of the IGF-1 exon 5 among 58 mammalian genomic sequence datasets available at the UCSC genome browser. Evolutionary tree of 58 organisms and the reconstructed ancestral state was adapted from the UCSC Genome Browser. Strength of 3' and 5'ss in exon5 obtained using the Maximum Entropy scores is shown for each species [Yeo G. and Burge CB. Comput. Biol. (2004);11:377-394]. Species with relative strong exon 5 5'ss are boxed. n.d.= not determined.

~~**Supplementary Figure S2. The strength of the IGF-1 exon 5 splice sites among 58 mammalian genomic sequence datasets available at the UCSC genome browser.** The phylogenetic tree for the 58 organisms was adapted from the UCSC Genome Browser. The splice site scores were obtained for each 5′ and 3′ splice site sequence using the Maximum Entropy scores [Yeo G. and Burge CB. Comput. Biol. (2004);11:377-394]. Species with relative strong exon 5 5'ss are in highlighted bold in the table and boxed in the phylogenetic tree.~~

Supplementary Figure S3. Species-Specific differences in expression pattern of the IGF-1 isoforms among mammals. (A, B and C) 3' RACE-PCR sequences obtained from liver tissues of mammalian species (see Figure 3B). 3' RACE-PCR products were sequenced and the nucleotide sequences corresponding to IGF-1Ea (A), IGF-1-1Eb (B) and IGF-1Ec (C) were aligned using Clustal W [Thompson J.D. et al. Curr Protoc Bioinformatics Chapter 2 (2002) Unit 23]. Poly(A) Signal Miner (http://dnafsminer.bic.nus.edu.sg/PolyA.html) was used to predict poly(A) signal and, if detected in the sequences, highlighted in bold. PCR product sizes (nucleotides) for each sequence are indicated in parentheses. The nucleotide sequences matched to the following NCBI Reference Sequence (RefSeq): IGF-1Ea (Pig 350 and 283: NM_214256; Rabbit 537: XM_008256720/ XM_008256719; Macaque 537: NM_001260726; Human 627 and 534: NM_000618/ NM_001111284; Mouse 391: NM_001314010/ NM_001111275/ NM_001111276; Rat 381: NM_178866/ NM_001082479); IGF-1Eb (Pig 548 and 490: XM_005664196/ XM_005664195; Rabbit 725 and 678: XM_008256717/ XM_008256716; Macaque 685: NC_027903; Human 734, 688 and 491: NM_001111285); IGF-1Ec (Rabbit 589: XM_008256718; Macaque 589: XM_015152532/ XM_015152534; Pig 401: XM_005664197; Rat 466: NM_001082477/ NM_001082478; Mouse 539 and 450: NM_010512/ NM_001111274). (D) Amplification of IGF-1Eb and IGF-1Ec isoforms in liver tissue of goat, sheep, pig and cow with isoform-specific PCR strategy. Map of IGF-1 gene with the relative position of primers used to separately amplify IGF-1 isoforms is shown on the left. The IGF-1F forward primer is common to all mammals (5'-CGTGGATGAGTGCTGCTTC-3'); the IGF-1R1 reverse primer is specific for IGF-1Eb amplification in goat, sheep and cow (5'-TCCTTCTGTTCCCCTCCTGG-3'); the IGF-1R2 reverse primer is specific for

pig IGF-1Eb isoform (5'-CCCTCCTGGGTGTTTCTTTG-3'); the IGF-1R3 reverse primer is specific for IGF-1Ec amplification in all mammals (5'-CTTCAAATGTACTTCCTTTCC-3'). The conventional PCR was carried out as described in Material and Methods. The PCR products were loaded on 4.0% agarose gel and DNA fragments were eluted from gel, subcloned and sequenced. The nucleotide sequences matched to the following NCBI Reference Sequence (RefSeq): IGF-1Eb (Goat: XM_013963970/ XM_013963971/ XM_005680531/ XM_005680532/ XM_005680533/ XM_005680534/ XM_005680535; Sheep: NC_019460; Cow; XM_015471061/ XM_015471062/ XM_015471063/ XM_015471064; Pig: XM_005664196/XM_005664195); IGF-1Ec (Goat: NC_022297; Sheep: NC_019460; Cow: XM_005206490; Pig: XM_005664197).

~~**Supplementary Figure S3. The IGF-1 isoform expression pattern of several mammalian species analysed by 3' RACE-PCR (Figure S3A) and isoform-specific PCR strategy (Figure S3B). Supplementary Figure S3B.** Expression analysis of IGF-1Eb and IGF-1Ec splice variants in liver tissue of goat, sheep, pig and cow. The schematic representation of the exon/intron structure of IGF-1 gene and position of primers used to separately amplify IGF-1 isoforms were shown for IGF-1Eb (left) and IGF-1Ec (right). The PCR products were loaded on 4.0% agarose gel and DNA fragments were eluted from gel, subcloned and sequenced. The following primers were used to amplified the IGF-1Eb and IGF-1Ec isoforms in liver tissues of different mammalian species: IGF-1F (forward primer common to all mammals): 5'-CGTGGATGAGTGCTGCTTC-3'; and the following reverse primers for IGF-1Eb amplification: IGF-1R1 (goat, sheep and cow): 5'-TCCTTCTGTTCCCCTCCTGG-3' and IGF-1R2 (pig): 5'-CCCTCCTGGGTGTTTCTTTG-3'; and for IGF-1Ec amplification: IGF-1R3 (common to all mammals): 5'-CTTCAAATGTACTTCCTTTCC-3'.~~

Supplementary Figure S4. The expression pattern of human and mouse IGF-1 minigenes in different human and mouse cell lines. RT-PCR analysis of splicing assay of human (hIGF-1; *upper panel*) and mouse (mIGF-1, *lower panel*) IGF-1 minigenes performed in the indicated human and mouse cell lines. Untransfected cells (-) or cells transfected with the empty vector (vect.) were used as a PCR control. L34 was used as loading control. A diagram of primers position used for RT-PCR analysis is also shown.

Supplementary Figure S5. Cis-acting elements contribute to species-specific variation of IGF-1 alternative splicing. (A,B) Representative Western-blot and RT-PCR analysis showing the expression level of hnRNP, SR and SR-like proteins in the splicing assay experiments. β-actin was used as loading control. C) RT-PCR analysis of splicing assay of human (hIGF-1) and mouse (mIGF-1) IGF-1 minigenes performed in HEK293T cells silenced with the indicated siRNAs. A representative western-blot showing the expression level of the silenced proteins is also shown. β-actin was used as loading control. D) RT-PCR analysis of splicing assay of endogenous human IGF-1 gene performed in LNCaP treated for 3hrs with 100μM Sodium Arsenite (Na_3AsO_3).

Supplementary Figure S6. (A) RT-PCR analysis of splicing assay performed in HEK293T cell line of mouse and human wild-type (WT), hMUT (TG*vs*GA) (*left panel*) and mMUT (GA*vs*TG) (*right panel*) IGF-1 minigenes in presence or not of SRSF1. The ratio between IGF-1Ec/IGF-1Ea (*left panel*) and IGF-1Eb/IGF-1tot (*right panel*) is also indicated. n.d.= not determined. A diagram of primers position used for RT-PCR analysis is also shown. (B) Representation of SRSF1 binding site (black box) in IGF-1 exon 5 identified by Smith PJ. et al. [Endocrinology (2002) 143:146-154] and Cléry A. et al. [PNAS (2013) 110:E2802-E2811]. Canonical cryptic 5' splice site (c5'ss) is also indicated.

Supplementary file S1. IGF-1 isoform nucleotide sequences of the 27 mammalian species used for IGF-1 evolutionary analysis. The "CDS FASTA alignment from multiple alignments" data, derived from the "multiz100way" alignment data prepared from 100 vertebrate genomes [Blanchette M. et al. Genome Res (2004), 14, 708-715], were downloaded using the Table Browser tool of the UCSC Genome Browser. Sequences were subsequently realigned using MUSCLE [Edgar R.C. Nucleic Acids Res (2004), 32, 1792-1797] and protein coding sequences from 27 mammalian species were extracted from these alignment datasets. The phylogenetic tree in Newick format was constructed using the IGF-1 core alignment of 27 mammalian species and is shown at the end of supplementary file S1.

Table 1. Substitution rates of IGF-1 domains among 27 mammalian species

IGF-1 domain	Average length (codons)	dS	dN	*dN-dS*
IGF-1 core	70.0	0.33 (0.23;0.44)[a]	0.01 (0.01;0.02) [a]	-0.31 (-0.42;-0.22) [a] *
Ea peptide	35.0	0.13 (0.09;0.18) [b]	0.02 (0.01;0.02) [a]	-0.12 (-0.16;-0.08) [b] *
Eb peptide	63.3	0.09 (0.06;0.12) [b]	0.05 (0.03;0.06) [b]	-0.04 (-0.06;-0.02) [c] *
Ec peptide	41.1	0.06 (0.04;0.09) [b]	0.02 (0.02;0.03) [a]	-0.04 (-0.07;-0.02) [c] *

All rates are estimated by maximum likelihood using the MG94xREV_3x4 (local) codon substitution model. Mean (95% confidence intervals) and *P* values for the neutrality tests (dN-dS expected value=0) are estimated by using parametric bootstrap based on 1000 replicates. *P<0.05. Multiple comparisons among IGF-1 domains were performed with Analysis of variance boostrap based. Values in the same column bearing the same letter are not significantly different. The difference dN-dS was used in place of a more common ratio dN/dS, to avoid numerical issues when dS is zero, which is possible because HyPhy permits synonymous substitution rates to vary from site to site.

FIGURE 1

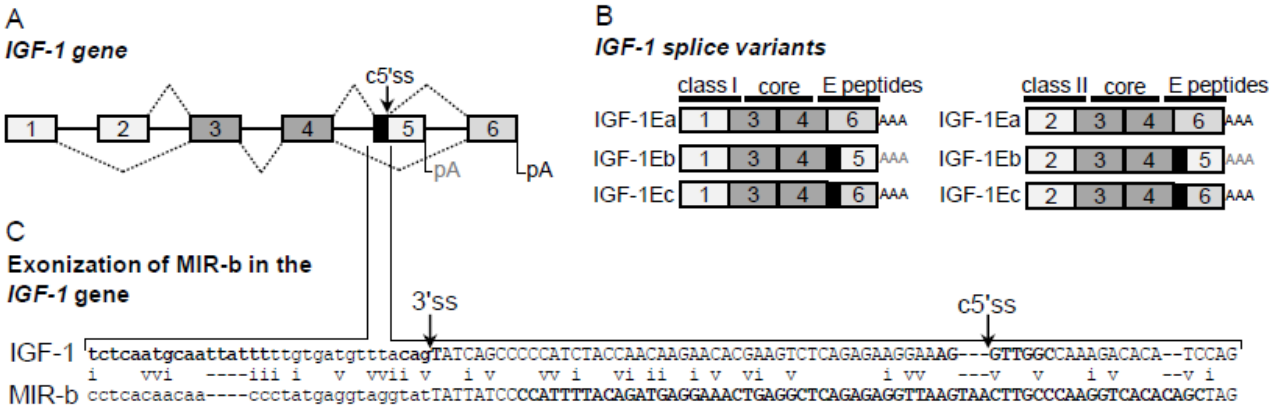


FIGURE 2

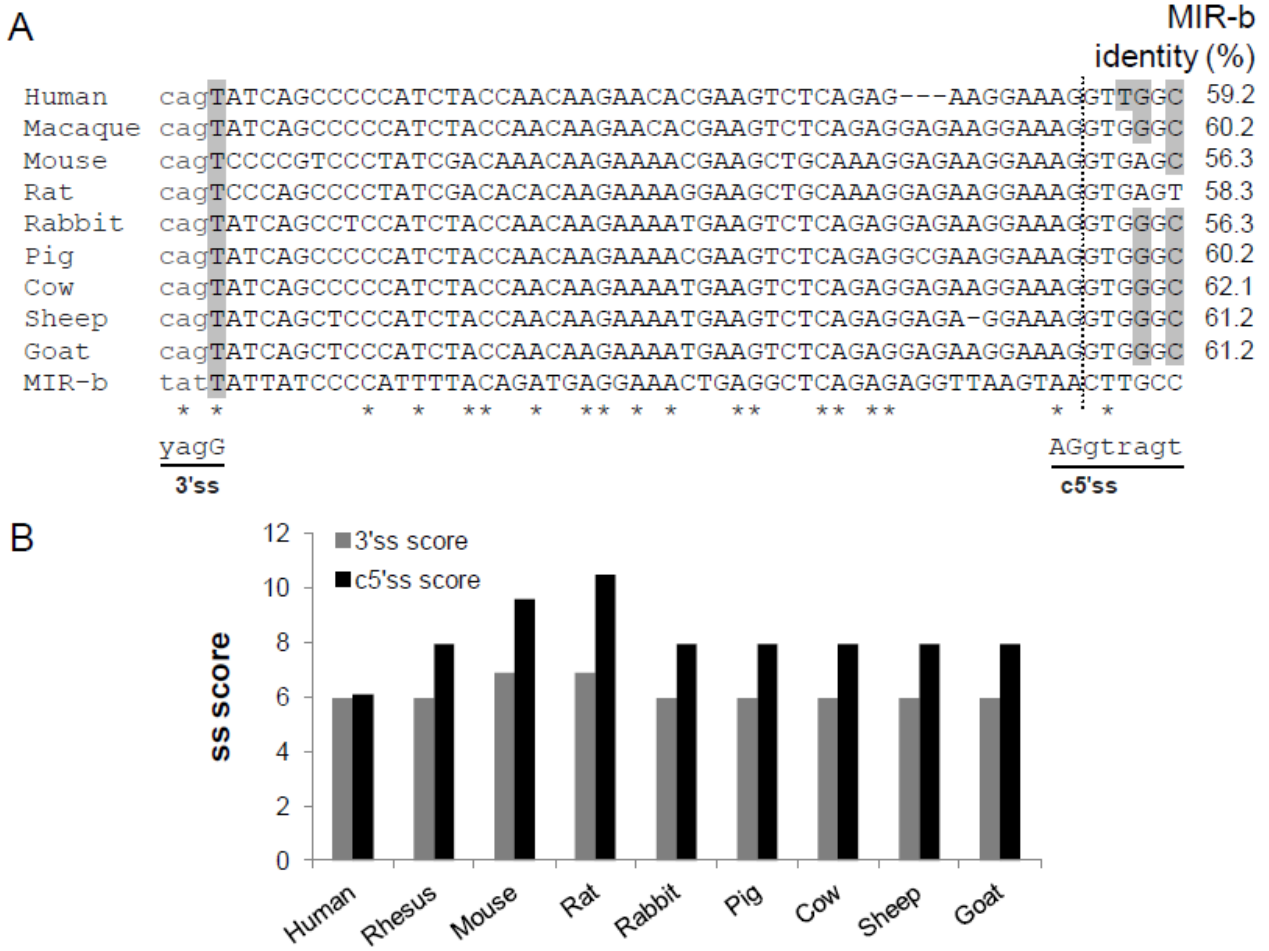
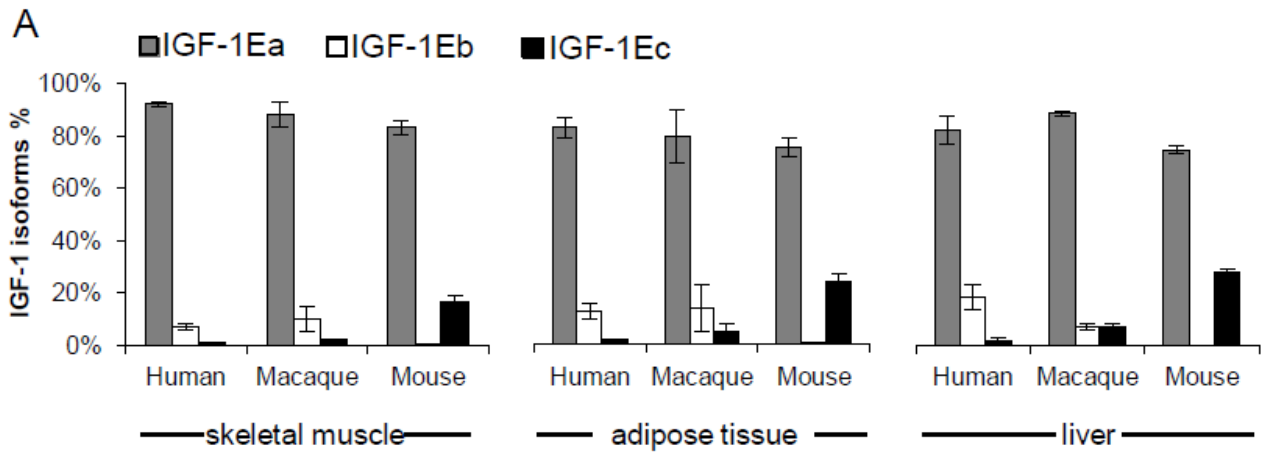


FIGURE 3


A
IGF-1Ea
IGF-1Eb
IGF-1Ec
IGF-1 isoforms %
100%
80%
60%
40%
20%
0%
Human
Macaque
Mouse
skeletal muscle
adipose tissue
liver

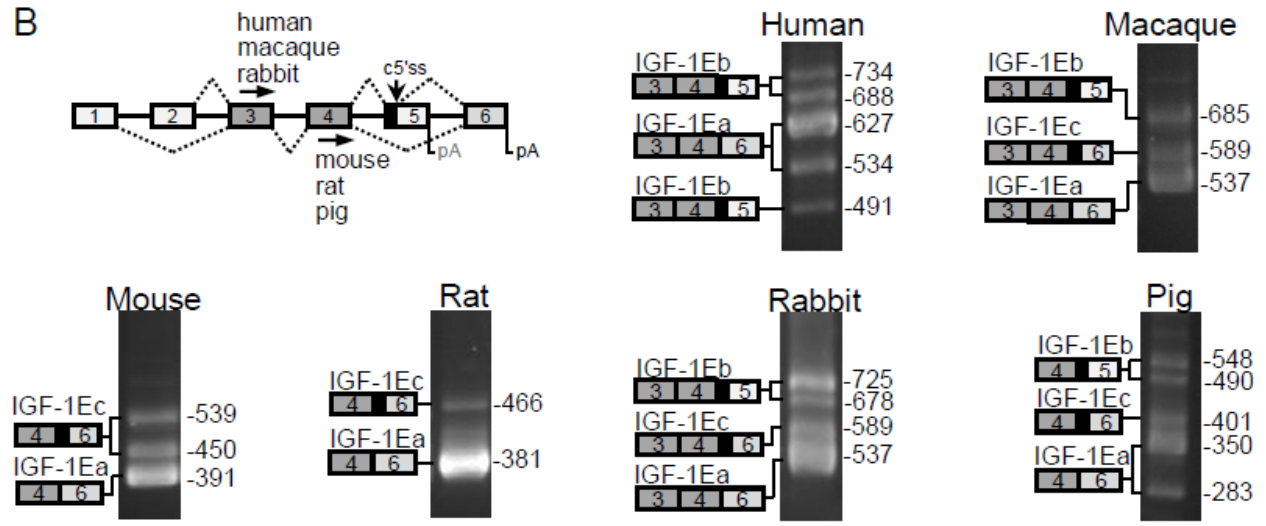

B
human
macaque
rabbit
c5'ss
mouse
rat
pig
pA
Human
IGF-1Eb
IGF-1Ea
IGF-1Eb
-734
-688
-627
-534
-491
Macaque
IGF-1Eb
IGF-1Ec
IGF-1Ea
-685
-589
-537
Mouse
IGF-1Ec
IGF-1Ea
-539
-450
-391
Rat
IGF-1Ec
IGF-1Ea
-466
-381
Rabbit
IGF-1Eb
IGF-1Ec
IGF-1Ea
-725
-678
-589
-537
Pig
IGF-1Eb
IGF-1Ec
IGF-1Ea
-548
-490
-401
-350
-283

FIGURE 4

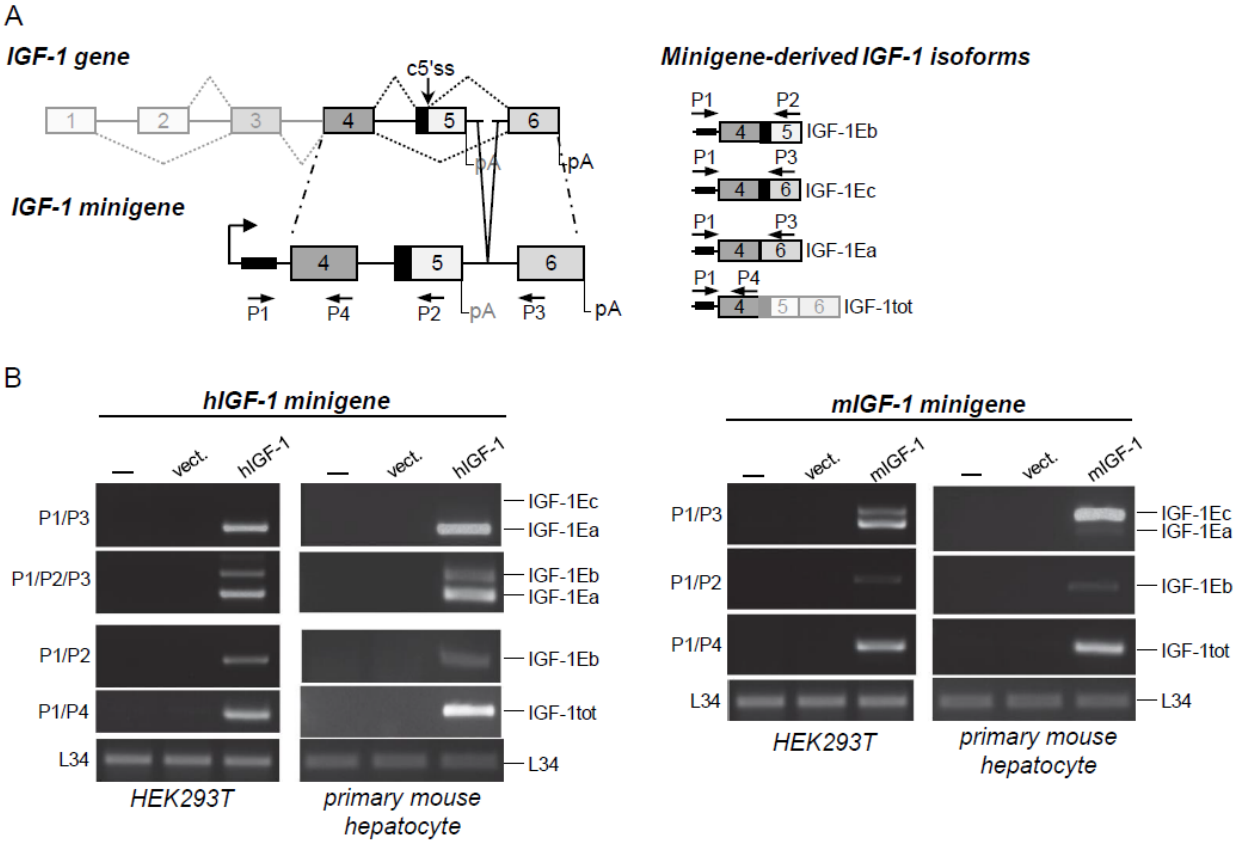

A
IGF-1 gene
c5'ss
1
2
3
4
5
6
pA
pA
IGF-1 minigene
4
5
6
P1
P4
P2
pA
P3
pA
Minigene-derived IGF-1 isoforms
P1
P2
4
5
IGF-1Eb
P1
P3
4
6
IGF-1Ec
P1
P3
4
6
IGF-1Ea
P1
P4
4
5
6
IGF-1tot
B
hIGF-1 minigene
—
vect.
hIGF-1
—
vect.
hIGF-1
P1/P3
P1/P2/P3
P1/P2
P1/P4
L34
IGF-1Ec
IGF-1Ea
IGF-1Eb
IGF-1Ea
IGF-1Eb
IGF-1tot
L34
HEK293T
primary mouse hepatocyte
mIGF-1 minigene
—
vect.
mIGF-1
—
vect.
mIGF-1
P1/P3
P1/P2
P1/P4
L34
IGF-1Ec
IGF-1Ea
IGF-1Eb
IGF-1tot
L34
HEK293T
primary mouse hepatocyte

FIGURE 5

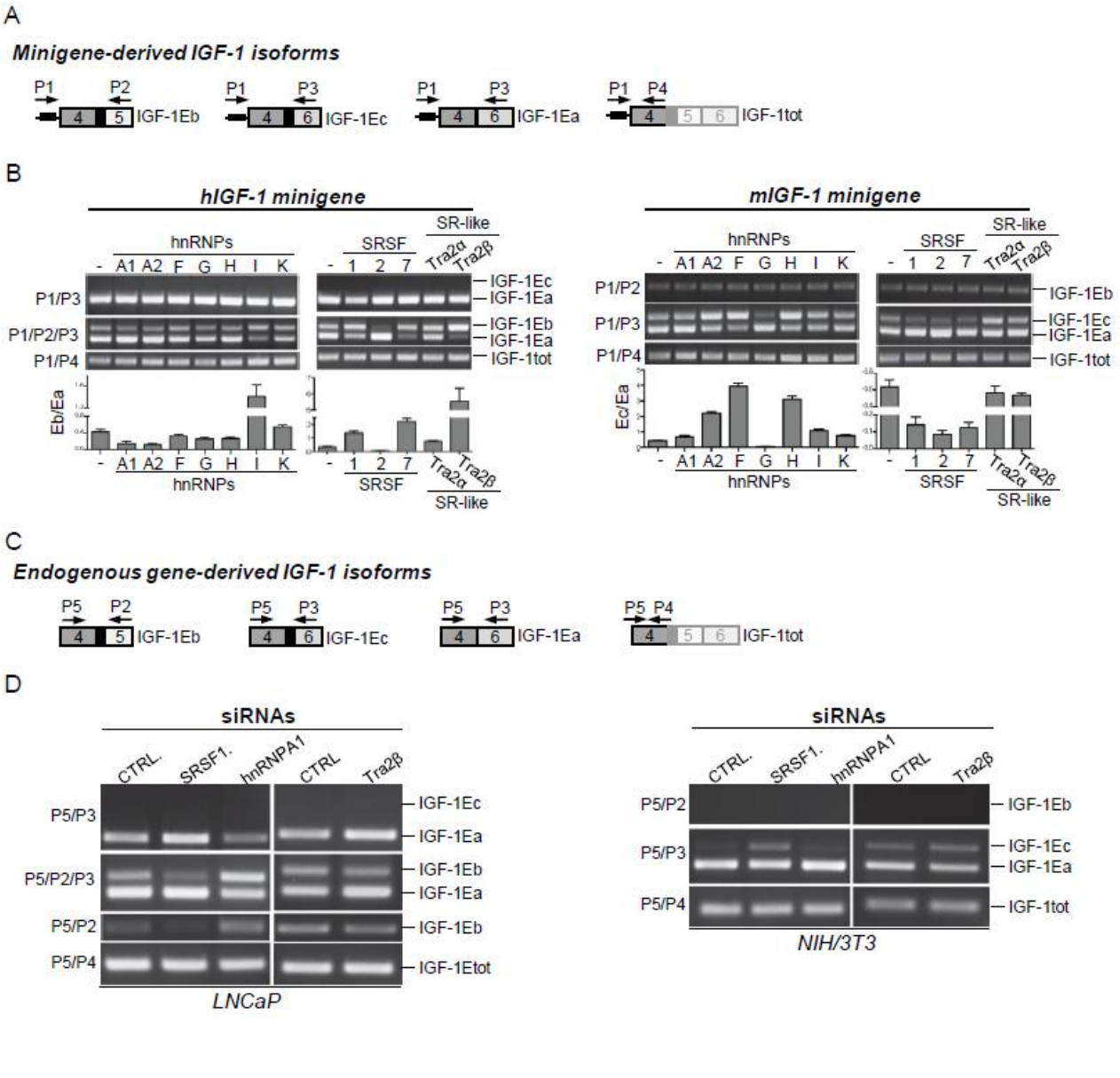

A
Minigene-derived IGF-1 isoforms
P1
P2
P3
P4
4
5
6
IGF-1Eb
IGF-1Ec
IGF-1Ea
IGF-1tot
B
hIGF-1 minigene
mIGF-1 minigene
hnRNPs
SRSF
SR-like
Tra2α
Tra2β
- A1 A2 F G H I K
- 1 2 7
P1/P3
P1/P2/P3
P1/P4
P1/P2
Eb/Ea
Ec/Ea
C
Endogenous gene-derived IGF-1 isoforms
P5
D
siRNAs
CTRL.
SRSF1.
hnRNPA1
CTRL
Tra2β
P5/P3
P5/P2/P3
P5/P2
P5/P4
IGF-1Etot
LNCaP
NIH/3T3

FIGURE 6

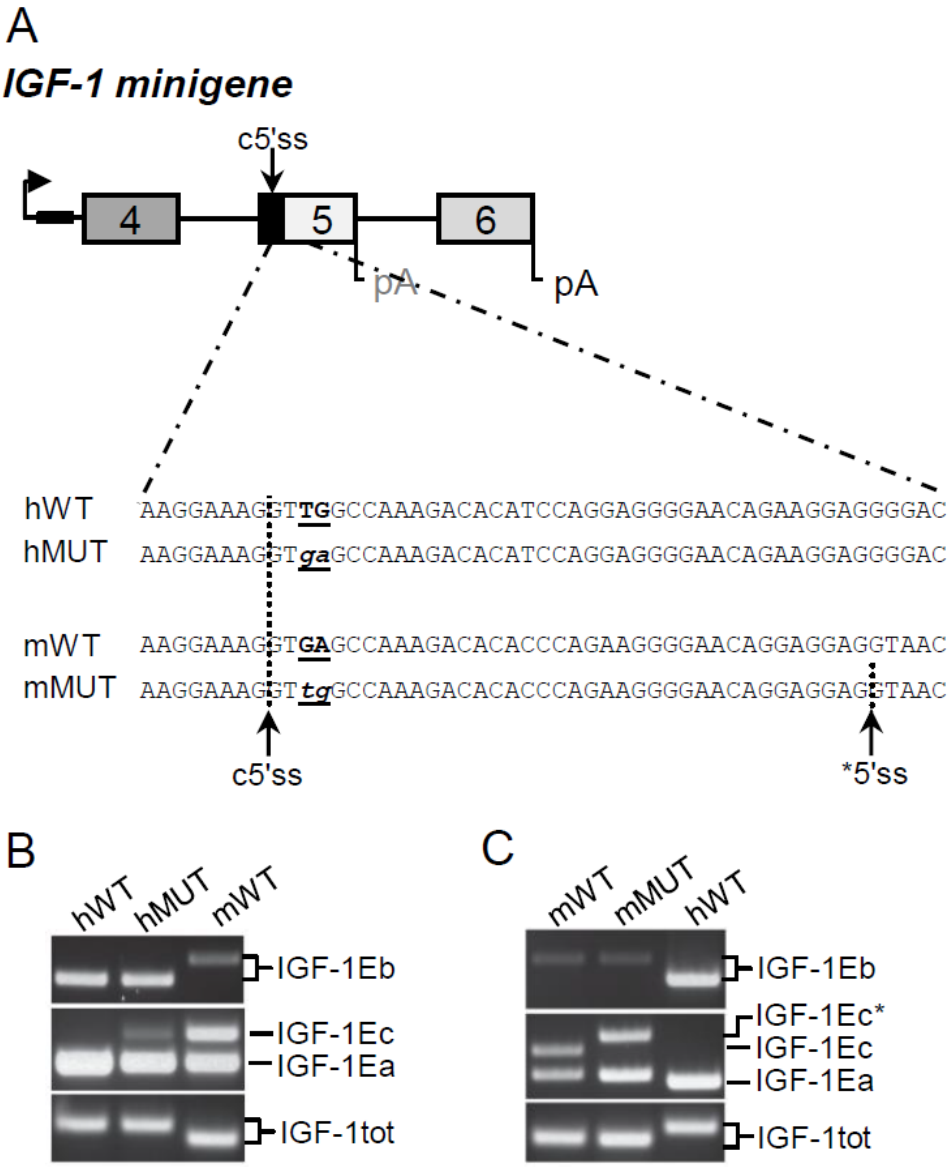
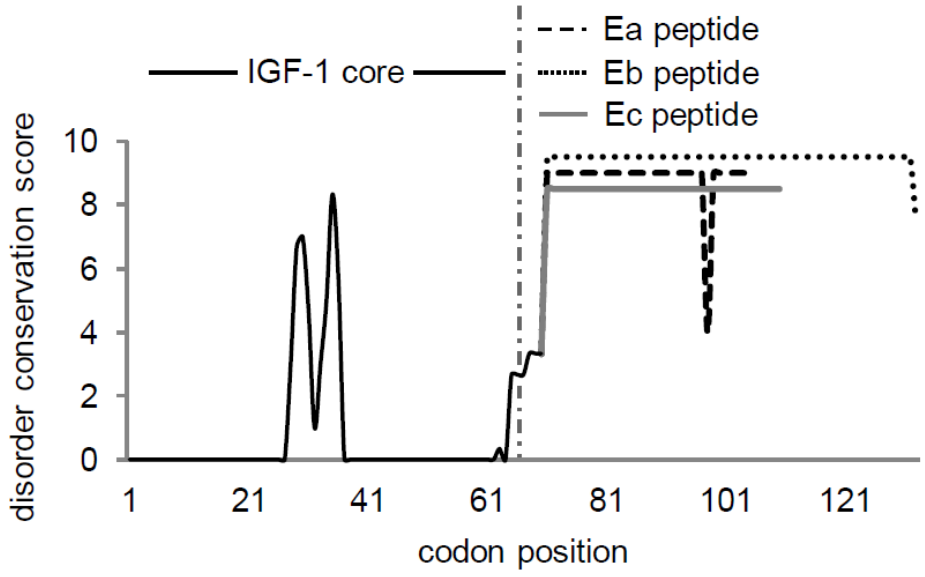

A
IGF-1 minigene
c5'ss
4
5
6
pA
pA
hWT AAGGAAAGGTTGGCCAAAGACACATCCAGGAGGGGAACAGAAGGAGGGGAC
hMUT AAGGAAAGGTgaGCCAAAGACACATCCAGGAGGGGAACAGAAGGAGGGGAC
mWT AAGGAAAGGTGAGCCAAAGACACACCCAGAAGGGGAACAGGAGGAGGTAAC
mMUT AAGGAAAGGTtgGCCAAAGACACACCCAGAAGGGGAACAGGAGGAGGTAAC
c5'ss
*5'ss
B
hWT
hMUT
mWT
IGF-1Eb
IGF-1Ec
IGF-1Ea
IGF-1tot
C
mWT
mMUT
hWT
IGF-1Eb
IGF-1Ec*
IGF-1Ec
IGF-1Ea
IGF-1tot

FIGURE 7


IGF-1 core
Ea peptide
Eb peptide
Ec peptide
disorder conservation score
10
8
6
4
2
0
1
21
41
61
81
101
121
codon position