Characterization of a novel alternatively-spliced 5' exon in the human insulin-like growth factor I (IGF-I) gene, expressed in liver and some cancers

Article (Accepted Version)

Characterization of a novel alternatively-spliced 5' exon in the human insulin-like growth factor I (IGF-I) gene, expressed in liver and some cancers

Michael Wallis

Biochemistry and Biomedicine Group, School of Life Sciences, University of Sussex, Brighton BN1 9QG, UK

email: m.wallis@sussex.ac.uk

Running title: **A novel 5' exon in the human IGF-I gene**

This is the author's accepted version of an article published in
Growth Hormone & IGF Research, Volumes 46–47, June–August 2019, Pages 36-43
https://doi.org/10.1016/j.ghir.2019.06.002

Abstract

In mammals, the large IGF-I gene comprises 6 exons, which are subject to alternative splicing. All transcripts contain exons 3 and 4, encoding mature IGF-I, but the other exons are included in various combinations, giving at least 6 possible mature mRNAs. At the 5' end, exons 1 and 2 are spliced alternatively to exon 3, giving different leader/signal sequences. It is shown in this study that in human an additional exon (designated exon 0) is present, upstream of exon 1. This can be spliced directly to exon 3 or, less frequently, into exon 1. Exon 0 is utilised in liver, in about 24% of IGF-I transcripts, to a minor extent in prostate and endometrium (<1% of transcripts), but not in any of 29 other normal human tissues examined. The exon 0 sequence includes an in-frame ATG/AUG, potentially providing a translation start point giving an IGF-I precursor with a very long signal peptide. However, this ATG is very close to the 5' end, and may not be included in all transcripts; an in-frame ATG in exon 3 could provide an alternative start point. Utilization of exon 0 was detected in other apes, and to a small extent in Old World monkeys, but not in New World monkeys, prosimians or various non-primate mammals. Exon 0 was not expressed in most human tumours, but was utilized in many prostate tumours, at levels much greater than seen in normal prostate, and in liver tumours, at a lower level than in normal liver.

Keywords: IGF-I, exon 0, alternative splicing, liver, primates, prostate cancer

1. Introduction

Insulin-like growth factor I (IGF-I) is a small protein formed from a larger precursor, encoded by a very large gene extending over ~85 kb in human. In mammals this gene comprises at least 6 exons, with the sequence encoding mature IGF-I included in exons 3 and 4. Alternative splicing allows these two exons to be combined with various combinations of the others, leading to production of at least 6 different IGF-I precursors [1-3] all of which include the sequence of mature IGF-I, which is generally strongly conserved [4-6]. At the 5' end the alternative splicing leads to production of mRNAs and IGF-I precursors with differing leader/signal sequences while at the 3' end varying amino acid (aa) sequences are produced, which can be processed to a number of peptides with potential biological actions [2-5]. IGF-I is produced in many tissues, and has both endocrine and paracrine actions. It mediates the actions of pituitary growth hormone, forming the basis of the somatomedin hypothesis [7,8]. Initially this mediating role was thought to involve circulating IGF-I, but more recent evidence suggests that paracrine actions may be more important [8]. Nevertheless, most IGF-I in the body is found in the circulation, associated with IGF binding proteins [1]. The source of this circulating IGF-I appears to be largely the liver [9,10] though its function is unclear [8].

Alternative splicing at the 5' end of the IGF-I gene has been widely studied, especially in rodents and human, though its significance remains poorly understood. In mammals (but not other vertebrates, where there is no equivalent of mammalian exon 2 [11]) exon 3 can be spliced to either exon 1 or exon 2, giving mRNA variants that encode IGF-I precursors with signal peptides of respectively 48 or 32 amino acid residues (aas) preceding the mature IGF-I sequence. Studies using cell-free systems showed that with rat mRNAs, signal peptides of these lengths are indeed produced [12,13] although initiation of translation can also occur at an ATG/AUG within exon 3, giving a signal peptide of 22 aas [13]. Translational efficiency appeared to be much greater for the 32 and 22 aa signal peptide forms than for the 48 aa variant [13]. Further complications are introduced because transcription of both exon 1 and exon 2 is initiated at multiple sites, giving mRNAs with untranslated regions (utrs) of variable length [14,15]; for the shortest of these in exon 1, the potential initiating AUG is missing, so translation would presumably start in exon 3. Yet another complication in rat is the occurrence of a deletion of 186 nt within the 5' utr in some transcripts, apparently due to another splicing event [16]; this deletion

removes some of the potential transcription initiation sites. The significance of all this variation at the 5' end of the IGF-I mRNA (and potentially the N-terminus of the IGF-I precursor) is unclear. Use of alternative leader sequences, with different promoter sequences may allow differential regulation of expression, for example between tissues or developmental stages [17-20] or in response to growth hormone [21-23]. Initiation of translation at different points can occur with differing efficiency, as discussed. Use of different signal sequences could result in varying intracellular localization of IGF-I precursors and/or post-translational modification [24]. The physiological importance of these various possibilities remains to be determined.

In light of this, the variation of the alternative splicing observed for IGF-I between different tissues, developmental stages and species is clearly of interest. In primates, variation in splicing across tissues has been assessed to some extent in human, but has been little studied in other primate species. The availability of large amounts of data on gene transcription in the ncbi SRA database makes possible the assessment of IGF-I splicing in a number of primate species. Such a study is reported here, focussing particularly on splicing at the 5' end of the gene. It was found that splicing does vary considerably among different tissues and species. Most notably, a novel exon, encoding a third 5' leader sequence, was found to be used in human and some other primates, expressed primarily in liver and some tumours, and associated with a substantial proportion of IGF-I transcripts.

2. Methods

2.1. Sequence and transcriptomic data

The sequences of IGF-I genes for a range of primates and other mammals were obtained using the Ensembl genome browser (www.ensembl.org). Releases used were as follows: human (*Homo sapiens*) GRCH38, chromosome 12; chimpanzee (*Pan troglodytes*) Pan_tro_3.0, chromosome 12; gorilla (*Gorilla gorilla*) gorGor4, chromosome 12; orangutan (*Pongo pygmaeus*) PPYG2, chromosome 12; gibbon (*Nomascus leucogenys*) Nleu_3.0, chromosome 10; rhesus macaque (*Macaca mulatta*) Mmul_8.0.1, chromosome 11; African green monkey (vervet, *Chlorocebus sabaeus*) ChlSabl.1, chromosome 11; marmoset (*Callithrix jacchus*) ASM275486v1, contig NTIC01000002.1; squirrel monkey (*Saimiri boliviensis*) SaiBol1.0, scaffold JH378139; tarsier (*Tarsier syrichta* now renamed

Carlito syrichta) Tarsius_syrichta-2.0.1, scaffold KE941348.1; mouse lemur (*Microcebus murinus*) Mmur_3.0, chromosome 7; galago (*Otolemur garnettii*) OtoGar3, scaffold GL873526.1; tree shrew (*Tupaia belangeri*) tupBel1, scaffold GeneScaffold_475; rat (*Rattus norvegicus*) Rnor_6.0, chromosome 7; rabbit (*Oryctolagus cuniculus*) OryCun2.0, chromosome 4; dog (*Canis familiaris*) CanFam3.1, chromosome 15. Exons in these genes were identified by comparison with the human and rat genes; the gene structure for IGF-I is known to be well conserved across mammals [4].

Transcriptomes were accessed via the SRA database (trace.ncbi.nlm.nih.gov/Traces/sra). The SRA projects used in this study are summarized in Supplementary Table 1.

2.2. Transcriptome analysis

Each transcriptome database (experiment) was subjected to BLAST searches [25] on the SRA website, using IGF-I sequences from the appropriate species. For each transcriptome, sequences with exon 3 linked at the 5' end to exon 0, exon 1 and exon 2 respectively were used to determine the relative usage of these three exons. Full sequences of these Blast Queries are given in Supplementary Table 2. Expression of IGF-I in various tissues was determined by BLAST analysis with the sequence encoding mature IGF-I (since this is included in all IGF-I precursor transcripts) and expressed as BLAST hits/million reads.

2.3. Exon profiling

To obtain a graphical visualization of the relative usage of the exons at the 5' end of the IGF-I gene, an in-house perl programme was used. The perl script is given in Supplementary Method 1. Input for this programme was the Blast alignment output produced by searching the corresponding transcriptomic database with a section of the IGF-I gene extending from about 5000 nt upstream of exon 1 to about 2000 nt downstream of exon 3. The output from the programme gave the number of representations in the alignment for each nucleotide, which was converted to graphical form using Microsoft Excel. Identification of repetitive sequence elements was performed using RepeatMasker [26].

2.4. Statistical analysis

Statistical analysis was carried out using the R package (www.R-project.org).

3. Results

3.1. A novel 5' utr and leader sequence for human IGF-I (exon 0)

Investigation of sequences spliced to the 5' end of IGF-I exon 3 in human liver transcriptomes (initially in SRA project ERP003613) revealed not only sequences corresponding to exons 1 and 2, but additional sequence corresponding to a region about 1380 nucleotides (nt) upstream of the 3' end of exon 1. BLAST analysis of this region identified a new IGF-I leader exon extending over about 145 nt (Fig.1), though the majority of reads corresponded to a shorter sequence of 110 nt. This novel exon is referred to as exon 0. In the majority of transcripts (~90%) it is spliced to the 5' end of exon 3, but in a minority it is spliced into a site 82 nt upstream of the 3' end of exon 1. Exon 0 contains an in-frame ATG/AUG which would allow production of a very long signal peptide of 65 aas (Fig.1). However, this ATG is close to the apparent (putative) 5' end of the majority of transcripts, and has weak surrounding sequence for an initiating AUG [27], so it is doubtful whether initiation starts at this point. If not, initiation would be expected to start at the first in-frame ATG/AUG in exon 3, giving a signal peptide of 25 aas; the equivalent signal sequence in rat is known to be functional [13] though a little shorter (22 aas).

Exon profiling for the 5' end of the human IGF-I gene, including exons 0-3 and ~4750 nt upstream of exon 0 confirmed the expression of exon 0 in human liver, as well as exons 1 and 2 (Fig. 2). Minor peaks between exons 0 and 1, exons 1 and 2 and exons 2 and 3 may reflect traces of unspliced or undegraded introns; none of these showed evidence for splicing to exons. The small peak ~3800 nt upstream of exon 0 was identified by RepeatMasker as an L1-like repetitive element, with again no evidence for splicing to an exon. All the exon 3 sequences in this transcriptome that ran up from the 5' splice site were spliced to exon 0, exon 1 or exon 2. On the basis of the profile of Fig. 2, exon 1 is the major 5' leader sequence, but exon 0 appears to be more highly represented than exon 2.

Expression of exon 0 was observed in all normal human liver transcriptomes examined (from 11 individuals in 5 separate SRA projects). Exon 1 to exon 3 splices predominated (71% of all such

157 splices), with exon 0-exon 3 splices at 22%, exon 2-exon 3 6% and exon 0-exon 1-exon 3 2% (Fig. 3).
158 However, the range of values was large (Fig.3), probably reflecting the fact that liver samples were
159 obtained from a wide range of individuals with no controlling for sex, age or physiological state. In
160 many cases such information was not available. Where replicate values for individuals were available,
161 agreement was close.
162
163 *3.2. Expression of IGF-I and alternative splicing in normal human tissues*
164 Any one of three upstream exons can be used in the expression of the human IGF-I gene, with exon 0,
165 1 or 2 spliced to the 5' of exon 3 and in addition, a minor variant in which exon 0 is spliced into exon 1
166 (Fig. 1). Each of these variants gives a different leader sequence on the derived mRNA, and potentially
167 a different signal peptide on the IGF-I precursor. The patterns of expression of these variants in a wide
168 variety of human tissues was determined using transcriptomic data from SRA projects ERP003613 and
169 ERP006650 [28] and is summarized in Fig. 4. Fig. 4A shows the expression level of IGF-I for each
170 tissue, determined by a BLAST search using the coding sequence for mature IGF-I, and expressed as
171 BLAST hits/10^6 reads. Expression levels vary markedly between tissues, and for some tissues
172 (particularly endometrium, liver and smooth muscle) between individuals. Replicates for individuals
173 generally showed good agreement. The level seen for liver was similar to the mean for the 11
174 individuals of Fig. 3 (4.20 ± 0.92 BLAST hits/10^6 reads). Several tissues show an expression
175 level similar to that in liver. All 32 tissues studied showed some expression of IGF-I, though
176 levels in skin, bone marrow, skeletal muscle and pancreas were very low.
177
178 Fig. 4B indicates the relative proportions of splices of exons 0, 1 and 2 to exon 3 for each tissue. Use of
179 exon 0 was seen in liver and to a very minor extent (less than 1%) in prostate gland and endometrium,
180 but not in any other tissue. Use of exon 1 was predominant in all tissues. The percent utilization of
181 exon 2 varied between tissues, and this variation was shown to be significant by one-way analysis of
182 variance ($p<0.001$). High variability was seen for some tissues; this, and the failure to detect any exon
183 2-exon 3 splices in some tissues may have been due to the low number of exon 2-exon 3 BLAST hits
184 obtained for tissues with low overall expression of IGF-I.
185
186 *3.3. Expression and alternative splicing of IGF-I in non-human primates*

An alignment of exon 0-like sequences derived from genomic sequences of various primates and a few non-primate mammals is shown in Fig. 5. Strong similarity is seen between primate sequences, but is less clear cut for non-primates. Rat sequence was not included because a convincing alignment could not be produced. The positions of two potential initiating ATGs is shown. After splicing to exon 3 the more upstream of these (ATG-1) is in-frame with the IGF-I coding sequence for the human sequence. The chimpanzee sequence is identical to that of human. Gorilla, orangutan and gibbon sequences are very similar to human, but insertion of an additional base (A) in the long A-tract changes the reading frame so that the downstream ATG (ATG-2) rather than ATG-1 is in-frame. Sequences for two Old-World monkeys (macaque and vervet) differ more from human; insertion of 6 bases leaves ATG-1 in-frame. Sequences of two New-World monkeys (marmoset and squirrel monkey) show a 9 base deletion, compared with human, in the region of the long A-tract; this would potentially leave the reading frame unchanged, but mutation of ATG-1 to CTG would prevent translation initiation from this site. Sequences from prosimians (tarsier, mouse lemur and galago) diverge further from human, and include a number of insertions/deletions (indels). Sequences of non-primates are even more divergent, with convincing alignment difficult to establish at the 5' end.

Expression of IGF-I was seen in livers of three apes (chimpanzee, gorilla and orangutan; no liver transcriptomic data is available in the SRA for gibbon) and in the various other primates and non-primates studied (Fig. 6A). Notably the expression level in human liver was lower than that in any other species except chimpanzee (though for gorilla, orangutan and squirrel monkey data from only two individuals were available, so this conclusion must be tentative). The level of IGF-I mRNA in rat liver was about 25-fold greater than that in human.

Utilization of exon 0 was seen in liver transcriptomes from the apes for which data is available (chimpanzee, gorilla and orangutan) though in all cases it was lower than in human. Exon 0 was also utilized, but to a very minor extent (<1%) in the Old-World monkeys (Fig. 6B). In none of the other species examined was there any utilization of an exon corresponding to human exon 0, with the 5' of exon 3 being spliced only to exon 1 or exon 2, except for the mouse lemur where a sequence between exons 1 and 2 was utilized to a minor extent (~3%).

The relative importance of exons 1 and 2 varies markedly between species, with exon 1 predominating in some species, including the apes and mouse lemur, but close to equal usage of exons 1 and 2 in some cases (e.g. galago and rat) (Fig. 6B). Notably, in the two New-World monkeys use of exon 2 predominates (Fig. 6B); the IGF-I system is unusual in this group in several other ways, including accelerated evolution of IGF-I and of its receptor [29].

3.4. IGF-I alternative splicing in human tumours

As indicated, in normal human tissues, utilization of exon 0 is confined largely to liver. The possibility that this exon is expressed in tumours was explored by examining SRA transcriptomic databases for a number of human cancers. In a number of tumour types examined, including breast, colorectal and uterine cancer and Ewing sarcoma (SRA projects SRP63460, SRP042620, SRP029880, SRP060016) utilization of exon 0 was not detected. In endometrial cancer (SRP040442) very low expression of exon 1 was detected in 2 of the 9 individuals studied, a level similar to that seen in normal endometrium. Low expression of exon 1 was also detected infrequently in lung cancer (in 2 of 40 individuals examined; project ERP001058).

The situation in prostate cancer is markedly different, as seen in three SRA projects (ERP000550, ERP017433, ERP006077). Study ERP006077 [30] included the largest number of cases, and results derived from this are shown in Fig. 7; similar results were obtained for the other two studies. Expression levels of IGF-I (Fig. 7A) in most tumours were similar to those in normal prostate (Fig. 4), but in some they were much lower and in one case (T23) markedly higher. Many of these tumours showed utilization of exon 0, with 17/25 showing such utilization at >1% of all 5' splices to exon 3, 6/25 utilization at >10%, and in one case >25% (Fig. 7B). Utilization of exon 0 in normal prostate is very low (<1% in the study of Fig. 4; no expression detected in two other studies, ERP076277, SRP163292). For 5 of the patients included in the project illustrated in Fig. 7 apparently normal tissue adjacent to the prostate tumour was also investigated; utilization of exon 0 was very low (<1%) in 3 of these cases, but it was higher (1.7% and 9.0%) in the other 2.

Many liver tumours express IGF-I at levels similar to those seen in normal liver (SRA studies SRP030040, SRP064431, SRP007560). Utilization of exon 0 in these tumours was variable, but

generally rather lower than that seen in normal liver (Fig. 3) or, in most cases, apparently normal tissue adjacent to the tumours. Fig. 8 summarises data derived from study SRP064431.

4. Discussion

Alternative splicing of the IGF-I gene has been recognized for many years, with, in mammals, the 6 exons of this complex gene differentially spliced at both the 5' and 3' ends, resulting in at least 6 different mRNAs and IGF-I precursors, all including the mature IGF-I sequence [1-3,6]. At the 5' end of the gene, two exons have been identified, exons 1 and 2 (Fig. 1), either of which can be spliced to exon 3. The relative usage of these 5' exons in various tissues, developmental stages and disease conditions has been widely studied, but their significance remains unclear. Usage of exon 1 is generally greater than that of exon 2, although the ratio between the two varies and relative exon 2 usage is increased by growth hormone treatment [19]. For both these exons the transcription start points can vary, giving mRNAs with 5' utrs of differing length [14,15]. Initiation of translation can occur within both exon 1 and exon 2, giving signal peptides of 48 or 32 aas respectively, of which the C-terminal 27 residues are encoded by exon 3 and therefore identical in the two signal peptides [13]. Initiation can also occur within exon 3, giving a signal peptide of 22 aas in rat (25 aas in human); exons 1 and 2 would then be untranslated. Initiation within exons 2 or 3 is more efficient than within exon 1 [13], so the higher usage of exon 1 may not necessarily mean that it is of more importance in overall IGF-I production. The 48-residue signal peptide encoded when initiation occurs within exon 1 is exceptionally long, and may have a specific role, such as directing the protein to a particular subcellular compartment.

It is shown here that in human the situation is further complicated by utilization of an additional upstream exon (exon 0) in the IGF-I gene. Alternative splicing of this gives rise to two novel leader sequences; in the predominant one exon 0 is spliced directly to exon 3, while less frequently exon 0 is spliced into exon 1. Utilization of exon 0 in human IGF-I has not been generally recognized previously and is not shown in the GTEx database (https://gtexportal.org/home/gene/IGF1), though it is predicted in three entries in the ncbi nucleotide database (www.ncbi.nlm.nih.gov/nuccore/), produced by

automated computer analysis (Variant X1, XM_017019259; Variant X3, XM_017019262; Variant X4, XM_017019263). In normal human tissues utilization of exon 0 is confined mostly to liver, where it comprises about 24% of splices to exon 3 (Fig. 3, Fig. 4). An AUG/ATG codon 114 nt upstream of the splice site would be in-frame with the IGF-I mature sequence when exon 0 is spliced to exon 3, giving a very long putative signal peptide of 65 aas. Whether translation does in fact initiate at this site is not clear. Analysis of transcriptomic data cannot define precisely the transcription start point, but few transcripts extend up 114 nt from the splice site, and those that do mostly extend only a short distance upstream of this point (Fig. 5). There does not appear to be a TATA box or CCAAT element upstream of the exon 0 sequence, suggesting that as for exon 1 multiple transcription start points may apply, so many of the transcripts may be too short to allow translation initiation at the in-frame AUG/ATG. Translation would then probably start at the AUG 7 nt from the 5' end of exon 3-derived sequence, giving a signal peptide of 25 aas. In the exon 0-exon 1-exon 3 variant, the AUG/ATG 114 nt upstream of the exon 0 splice site would not be in-frame, but an AUG/ATG 22 nt downstream of this would, potentially allowing translation of a precursor with a very long signal peptide of 85 aas.

The origin of exon 0 has been clarified by analysing sequences from various primates. A sequence upstream of exon 1 similar to that of human exon 0 was identifiable in all primates, and in several non-primate mammals, but not rat, where homologous sequence may be absent or just unrecognizable due to divergent evolution. Utilization of exon 0, judged by its appearance in liver transcripts, is only seen in Old World monkeys and apes, and in the former is very low (< 1% of all splices to exon 3; Fig. 6). It would thus appear that exon 0 arose by exaption of an existing sequence present upstream of the IGF-I gene, possibly only becoming fully functional in apes. Also notable from the comparative study of Fig. 6 are (1) the relatively low expression of IGF-I in human liver compared with many other species, the contrast being particularly marked in the case of rat, and (2) the rather low utilization of exon 2 in human and other apes, compared with most other primates and non-primate mammals.

Expression of exon 0 was largely confined to liver and not detected in 29 other human tissues examined, though very low expression was detected in prostate and endometrium. Notably, for many of these tissues, expression of IGF-I was similar to, or higher than, that in liver, though variation between individuals was high. Unlike exon 0, expression of exon 2 was seen in most human tissues,but always

less than that of exon 1, though the ratio between the two varied considerably. ANOVA showed that such variation was significant, indicating independent regulation of the expression of these two exons.

The potential involvement of IGF-I in cancer has been discussed extensively [31-37], and the possibility of aberrant expression of exon 0 in cancer was investigated. In most cancers examined expression of exon 0 was absent (breast, colorectal, uterus, Ewing sarcoma) or very low (endometrium, lung). A marked exception was prostate cancer, where a high proportion of patients showed substantial expression of exon 0. The significance of this is unclear, but the abnormal expression of a potent growth factor could well play a part in tumour growth and progression. Exon 0 expression was also seen in liver cancer (Fig. 8), but at a rather lower level than seen in normal liver tissue.

In conclusion, the study described here shows that in human and apes, the IGF-I gene includes a novel exon, at the 5' end. Alternative splicing allows either exon 0, exon 1 or exon 2 to be spliced to exon 3, potentially giving 3 IGF-I precursors with different signal peptides. A further minor variant can also be produced, resulting from splicing of exon 0 into exon 1. Since alternative splicing also occurs at the 3' end of the IGF-I gene the total number of potential variants of the IGF-I precursor becomes quite large. The significance of the alternative leader sequences for human IGF-I remains unclear, but it is notable that unlike exons 1 or 2 utilization of exon 0 is very tissue specific, for normal tissues being confined largely to liver. In mice prevention of utilization of exon 2 has no apparent effect on development, growth or circulating IGF-I levels [38]. Whether the same might be true for exon 0 in humans is not clear, but its utilization in many prostate tumours does suggest that it may play a role in human disease.

5. Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

6. Declaration of interest

None.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version, at

References

[1] W.H. Daughaday, P. Rotwein, Insulin-like growth factors I and II. Peptide, messenger ribonucleic acid and gene structures, serum, and tissue concentrations, Endocr. Rev. 10 (1989) 68-91.

[2] A. Philippou, M. Maridaki, S. Pneumaticos, M. Koutsilieris, The complexity of the *IGF1* gene splicing, postranslational modification and bioactivity, Mol. Med. 20 (2014) 202-214.

[3] G. Vassilakos, E.R. Barton, Insulin-like growth factor 1 regulation and its actions in skeletal muscle, Compr. Physiol. 9 (2019) 413-438.

[4] P. Rotwein, Diversification of the insulin-like growth factor 1 gene in mammals, PLoS ONE 12 (2017) e0189642.

[5] P. Rotwein, Variation in the insulin-like growth factor 1 gene in primates, Endocrinology 158 (2017) 804-814.

[6] 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.

[7] W.D. Salmon, W.H. Daughaday, A hormonally controlled serum factor which stimulates sulfate incorporation by cartilage in vitro, J. Lab. Clin. Med. 49 (1957) 825–836.

[8] D. LeRoith, C. Bondy, S. Yakar, J.-L. Liu, A. Butler, The somatomedin hypothesis: 2001, Endocr. Rev. 22 (2001) 53-74.

[9] K. Sjögren, J.-L. Liu, K. Blad, S. Skrtic, O. Vidal, V. Wallenius, D. LeRoith, J. Törnell, O.G. Isaksson, J.-O. Jansson, C. Ohlsson, Liver-derived insulin-like growth factor I (IGF-I) is the principal source of IGF-I in blood but is not required for postnatal body growth in mice, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 7088–7092.

[10] S. Yakar, J.-L. Liu, B. Stannard, A. Butler, D. Accili, B. Sauer, D. LeRoith, Normal growth and development in the absence of hepatic insulin-like growth factor I, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 7324–7329.

[11] P. Rotwein, Insulinlike growth factor 1 gene variation in vertebrates, Endocrinology 159 (2018) 2288-2305.

[12] P. Rotwein, R.J. Folz, J.I. Gordon, Biosynthesis of human insulin-like growth factor I (IGF-I). The primary translation product of IGF-I mRNA contains an unusual 48-amino acid signal peptide, J. Biol. Chem. 262 (1987) 11807-11812.

[13] H. Yang, M.L. Adamo, A.P. Koval, M.C. McGuinness, H. Ben-Hur, Y. Yang, D. LeRoith, C.T. Roberts, Alternative leader sequences in insulin-like growth factor I mRNAs modulate translational efficiency and encode multiple signal peptides, Mol. Endocrinol. 9 (1995) 1380-1395.

[14] M.L. Adamo, H. Ben-Hur, D. LeRoith, C.T. Roberts, Transcription initiation in the two leader exons of the rat IGF-I gene occurs from disperse versus localized sites, Biochem. Biophys. Res. Commun. 176 (1991) 887-893.

[15] E. Jansen, P.H. Steenbergh, D. LeRoith, C.T. Roberts, J.S. Sussenbach, Identification of multiple

transcription start sites in the human insulin-like growth factor-I gene, Mol. Cell. Endocrinol. 78 (1991) 115-125.

[16] A. Shimatsu, P. Rotwein, Sequence of two insulin-like growth factor I mRNAs differing within the 5' untranslated region, Nuc. Acids Res. 15 (1987) 7196.

[17] M.L. Adamo, H. Ben-Hur, C.T. Roberts, D.LeRoith, Regulation of start site usage in the leader exons of the rat insulin-like growth factor-I gene by development, fasting, and diabetes, Mol. Endocrinol. 5 (1991) 1677-1686.

[18] E. Jansen, P.H. Steenbergh, F.M.A. van Schaik, J.S. Sussenbach, The human IGF-I gene contains two cell type-specifically regulated promoters, Biochem. Biophys. Res. Commun. 187 (1992) 1219-1226.

[19] W.L. Lowe, C.T. Roberts, S.R. Lasky, D. LeRoith, Differential expression of alternative 5' untranslated regions in mRNAs encoding rat insulin-like growth factor I, Proc. Natl. Acad. Sci. U.S.A. 84 (1987) 8946-8950.

[20] C.A. West, T.R. Arnett, S.M. Farrow, Expression of insulin-like growth factor I (IGF-I) mRNA variants in rat bone, Bone 19 (1996) 41-46.

[21] D.J. Chia, J.J. Young, A.R. Mertens, P. Rotwein, Distinct alterations in chromatin organization of the two IGF-I promoters precede growth hormone-induced activation of IGF-I gene transcription, Mol. Endocrinol. 24 (2010) 779-789.

[22] H.L.Foyt, F. Lanau, M. Woloschak, D. LeRoith, C.T. Roberts, Effect of growth hormone on levels of differentially processed insulin-like growth factor I mRNAs in total and polysomal mRNA populations, Mol. Endocrinol. 6 (1992) 1881-1888.

[23] J.M.Pell, J.C.Saunders, R.S.Gilmour, Differential regulation of transcription initiation from

insulin-like growth factor-I (IGF-I) leader exons and of tissue IGF-I expression in response to changed growth hormone and nutritional status in sheep, Endocrinology 132 (1993) 1797-1807.

[24] J.G. Simmons, J.J. van Wyk, E.C. Hoyt, P.K.Lund, Multiple transcription start sites in the rat insulin-like growth factor-I gene give rise to IGF-I messenger-RNAs that encode different IGF-I precursors and are processed differently in-vitro, Growth Factors 9 (1993) 205-221.

[25] S.F. Altschul, W. Gish, W. Miller, E.W.Myers, D.J. Lipman, Basic Local Alignment Tool, J. Mol. Biol. 215 (1990) 403-410.

[26] A.F.A. Smit, R. Hubley, P. Green, *RepeatMasker Open-4.0*, (2013-2015) <http://www.repeatmasker.org>.

[27] M. Kozak, At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells, J. Mol. Biol 196 (1987) 947-950.

[28] L. Fagerberg, B.M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund, A. Asplund, E. Sjöstedt, E. Lundberg, C.A-K. Szigyarto, M. Skogs, J.O. Takanen, H. Berling, H. Tegel, J. Mulder, P. Nilsson, J.M. Schwenk, C. Lindskog, F. Danielsson, A. Mardinoglu, Å. Sivertsson, K. von Feilitzen, M. Forsberg, M. Zwahlen, I. Olsson, S. Navani, M. Huss, J. Nielsen, F. Ponten, M. Uhlén, Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics, Mol. Cell. Proteomics 13 (2014) 397-406.

[29] M. Wallis, Coevolution of insulin-like growth factors, insulin and their receptors and binding proteins in New World Monkeys, Growth Horm. IGF Res. 25 (2015) 158-167.

[30] A.W. Wyatt, F. Mo, K. Wang, B. McConeghy, S. Brahmbhatt, L. Jong, D.M. Mitchell, R.L. Johnston, A. Haegert, E. Li, J. Liew, J. Yeung, R. Shrestha, A.V. Lapuk, A. McPherson, R. Shukin, R.H. Bell, S. Anderson, J. Bishop, A. Hurtado-Coll, H. Xiao, A.M. Chinnaiyan, R.

Mehra, D. Lin, Y. Wang, L. Fazli, M.E. Gleave, S.V. Volik, C.C. Collins, Heterogeneity in the inter-tumor transcriptome of high risk prostate cancer, Genome Biol. 15 (2014) art. no. 426.

[31] P.F. Christopoulos, P. Msaouel, M. Koutsilieris, The role of the insulin-like growth factor-1 system in breast cancer, Mol. Cancer 14 (2015) art. no. 43.

[32] E.J. Gallagher, D. LeRoith, The proliferating role of insulin and insulin-like growth factors in cancer, Trends Endocrinol. Metab. 21 (2010) 610-618.

[33] E.J. Gallagher, D. LeRoith, Minireview:IGF, insulin and cancer, Endocrinology 152 (2011) 2546-2551.

[34] C. Gennigens, C. Menetrier-Caux, J.P. Droz, Insulin-like growth factor (IGF) family and prostate cancer, Crit. Rev. Oncol. Hematol. 58 (2006) 124-145.

[35] C. Kang, D. LeRoith, E.J. Gallagher, Diabetes, obesity, and breast cancer, Endocrinology 159 (2018) 3801-3812.

[36] A. Kasprzak, W. Kwasniewski, A. Adamek, A. Gozdzicka-Jozefiak, Insulin-like growth factor (IGF) axis in cancerogenesis, Mut. Res. 772 (2017) 78-104.

[37] P.G. Vigneri, E.Tirro, M.S. Pennisi, M. Massimino, S. Stella, C. Romano, L. Manzella, The insulin/IGF system in colorectal cancer development and resistance to therapy, Front. Oncol. 5 (2015) art. no. 230.

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

Legends for Figures

Fig. 1. The human IGF-I gene. A. Organization of the gene showing the 6 previously-recognized exons (solid rectangles), the novel exon 0 (open rectangle), and the possible alternative splices at the 5' end, the minor exon 0-exon 1-exon 3 variant being shown with dotted lines. Introns 3 and 5 are shown truncated; their full lengths are approx. 55.9 kb (intron 3) and 14.9 kb (intron 5). B. Sequence of exon 0. The start of the exon at the 5' end is poorly defined and there may be multiple transcription start points; the region under the dashes is represented in few transcripts. Flanking sequences are shown in light grey. C. The potential signal peptide derived from exon 0 spliced to exon 3. Solid arrows indicate possible translation initiation sites; which of these is actually used is not clear (see text). The N-terminal end of mature IGF-I is also included; the broken arrow indicates the cleavage site for the signal peptide.

Fig. 2. Exon profile for the 5' end of IGF-I gene in human liver transcriptome (SRA project ERP003613; 5 experiments combined). Arrows indicate positions of in-frame translation initiation sites (ATG/AUG).

Fig. 3. Utilization of exons 0, 1 and 2 in human liver. Values are means ± SEM for 11 individuals from 5 SRA projects.

Fig. 4. Expression level and 5' exon utilization for IGF-I in human tissues. Based on data in SRA projects ERP003613 and ERP006650. A. Expression level (BLAST hits/10^6 reads), based on BLAST searching with the sequence encoding mature human IGF-I. B. Utilization of exons 0, 1 and 2 in each tissue, expressed as a percentage of all splices to exon 3. The exon 0-exon 3 and exon 0-exon 1-exon 3 splices were combined. Error bars show S.E.M. for 2-7 individuals.

Fig. 5. Sequences of IGF-I exon 0 region for a number of primates and other mammals. The human sequence is given in full. For other sequences, a dot (.) indicates identity to human and a dash (-) indicates a gap. The exon 0 region is included in the large box, which is open at the left hand side because the 5' end is not defined. ATG-1 and ATG-2 indicate the positions of two potential translation start sites, in-frame in some sequences (see text). a and b indicate the 5' limits seen in the human

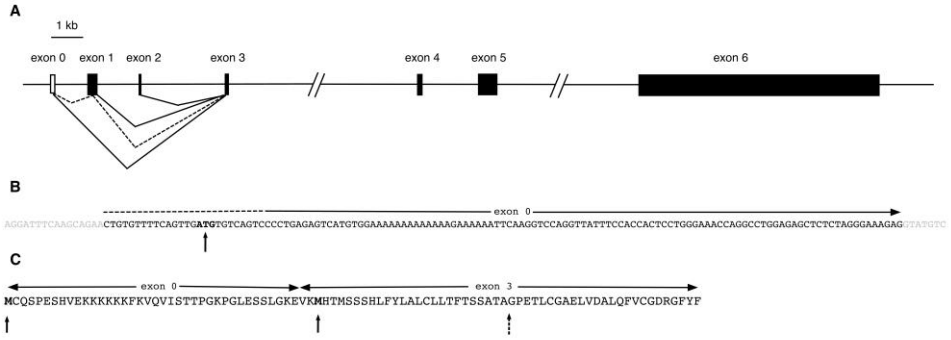
normal liver transcriptomes, b being reached in all 5 projects studied (including that illustrated in Fig. 2), a in only two projects. Abbreviations of species names are as follows: Hsa, *Homo sapiens* (human), Ptr, *Pan troglodytes* (chimpanzee), Ggo, *Gorilla gorilla* (gorilla), Ppy, *Pongo pygmaeus* (orangutan), Nle, *Nomascus leucogenys* (gibbon), Mmu, *Macaca mulatta* (rhesus macaque), Csa, *Chlorocebus sabaeus* (African green monkey), Cja, *Callithrix jacchus* (marmoset), Sbo, *Saimiri boliviensis* (squirrel monkey), Tsy, *Tarsius syrichta* (tarsier), Mmur, *Microcebus murinus* (mouse lemur), Oga, *Otolemur garnettii* (galago), Tbe, *Tupaia belangeri* (tree shrew), Ocu, *Oryctolagus cuniculus* (rabbit), Cfa, *Canis familiaris* (dog).

Fig. 6. Expression level and 5' exon utilization for IGF-I in liver of various primates plus tree shrew, rat, rabbit and dog. A. Expression level (BLAST hits/10^6 reads), based on BLAST searching with the sequence encoding mature IGF-I for the appropriate species. Numbers in brackets indicate the number of individuals in the sample. B. Utilization of exons 0 (white), 1 (grey) and 2 (black) in each tissue, expressed as a percentage of all splices to exon 3. The exon 0-exon 3 and exon 0-exon 1-exon 3 spliced forms were combined. Error bars show ± S.E.M. except where n<3, where individual values are shown (x). Species abbreviations as for Fig. 5 plus Rno, *Rattus norvegicus* (rat).

Fig. 7. Expression level and 5' exon utilization for IGF-I in prostate cancers from 25 individual patients. Based on data in SRA project ERP006077. A. Expression level (BLAST hits/10^6 reads), based on BLAST searching with the sequence encoding mature IGF-I. B. Utilization of exons 0 (light grey), 1 (grey) and 2 (black) in each tumour, expressed as a percentage of all splices to exon 3. The exon 0-exon 3 and exon 0-exon 1-exon 3 spliced forms were combined. Error bars show ± S.E.M. except where n<3. n = 3, except where individual points are shown (x, n = 1 or 2).

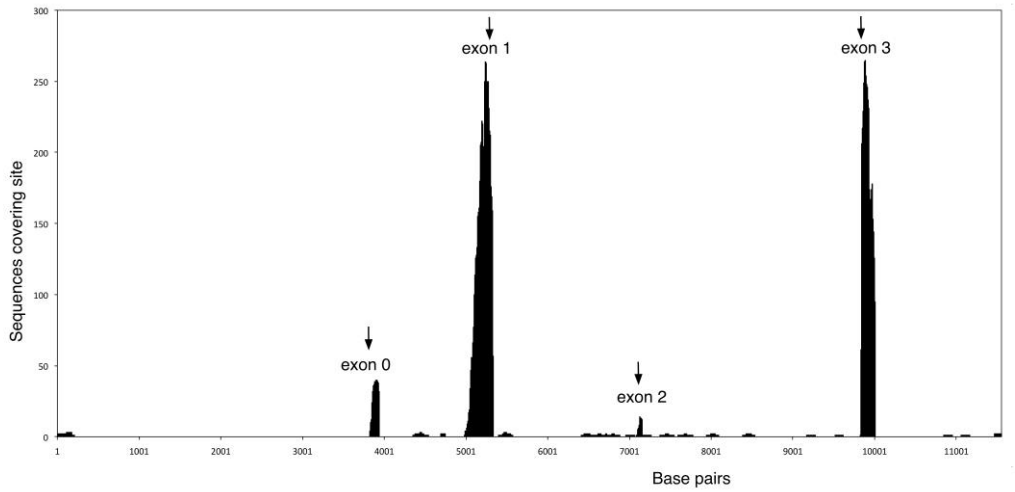
Fig 8. Expression level and 5' exon utilization for IGF-I in liver cancers from 4 individual patients. Based on data in SRA project SRP064431. A. Expression level (BLAST hits/10^6 reads), based on BLAST searching with the sequence encoding mature human IGF-I for each tumour (T) or adjacent apparently normal tissue (N). B. Utilization of exons 0 (white), 1 (grey) and 2 (black) in each tumour (T) or adjacent apparently normal tissue (N), expressed as a percentage of all splices to exon 3. The exon 0-exon 3 and exon 0-exon 1-exon 3 spliced forms were combined.

548 Fig. 1.



549

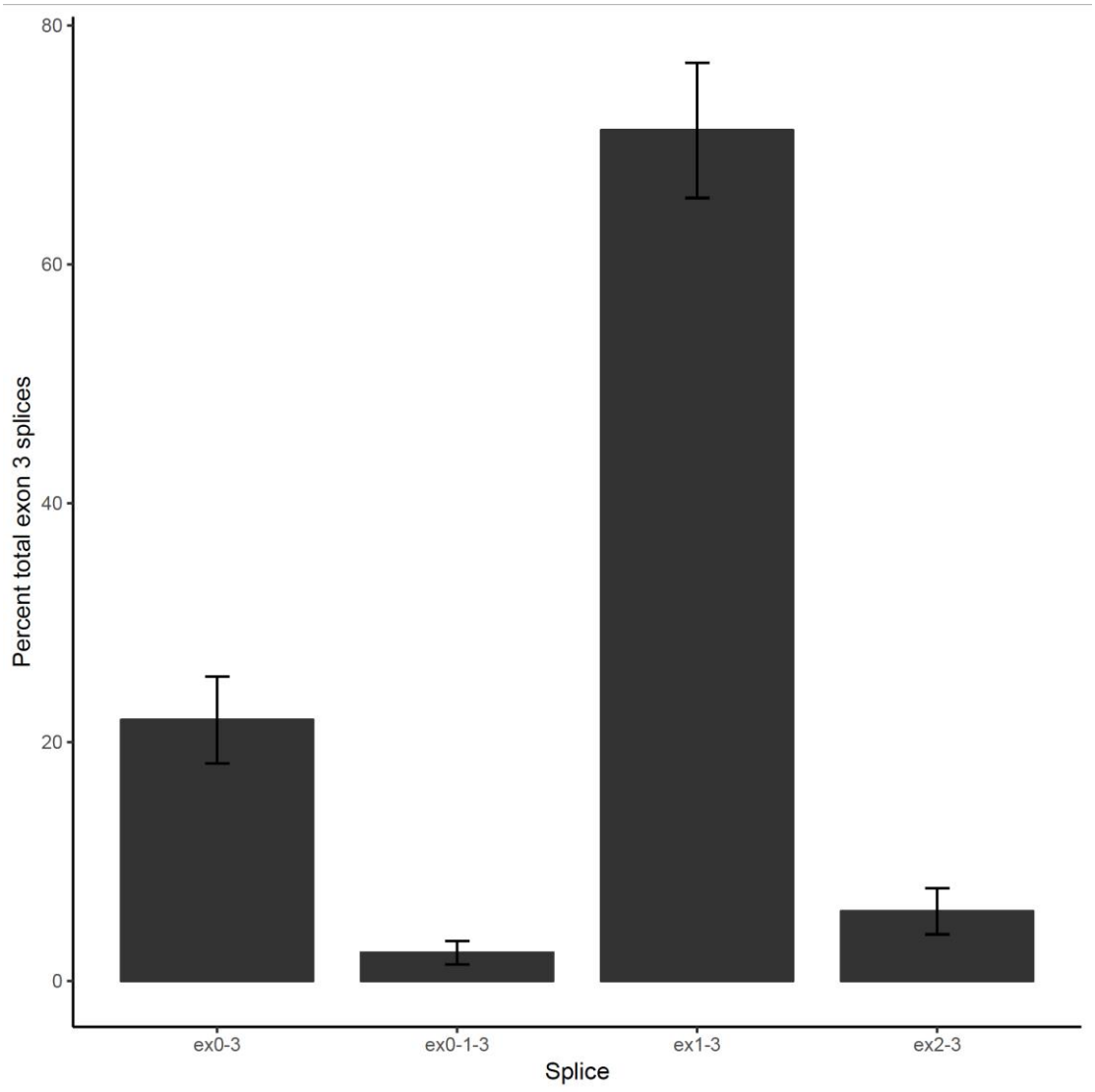
550 Fig. 2.



551

552

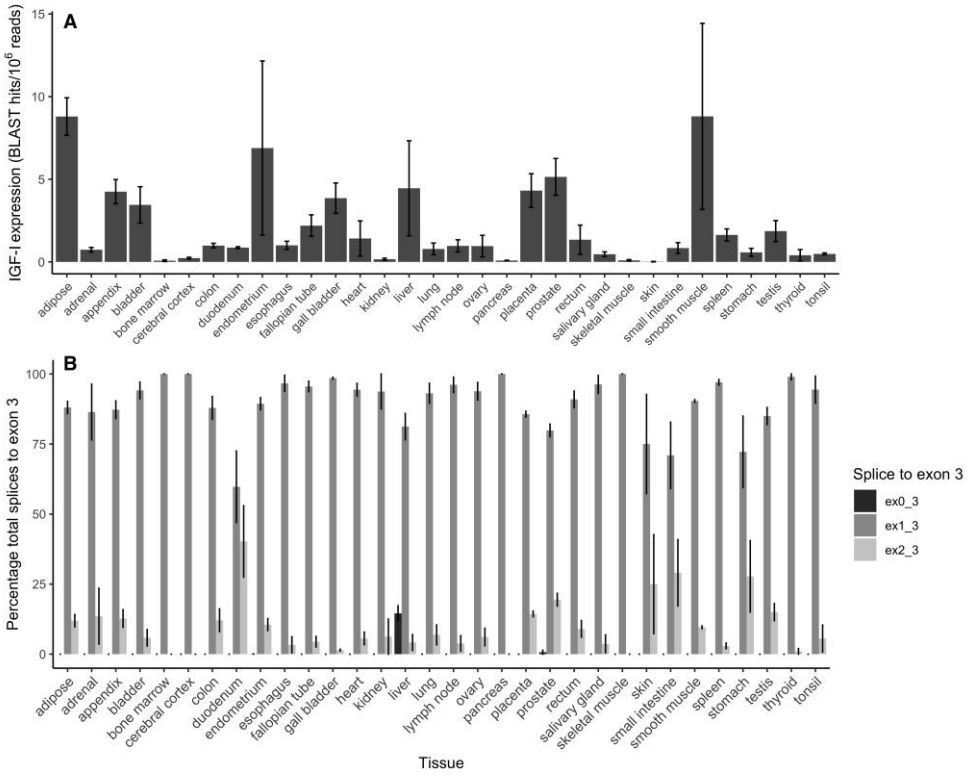
553 Fig. 3.



554

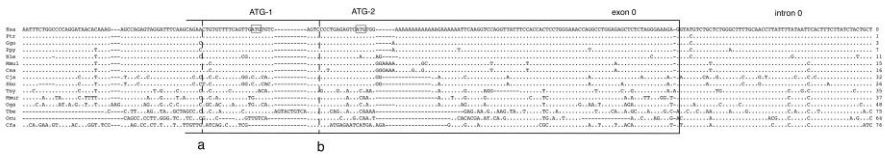
555

556 Fig. 4.



557

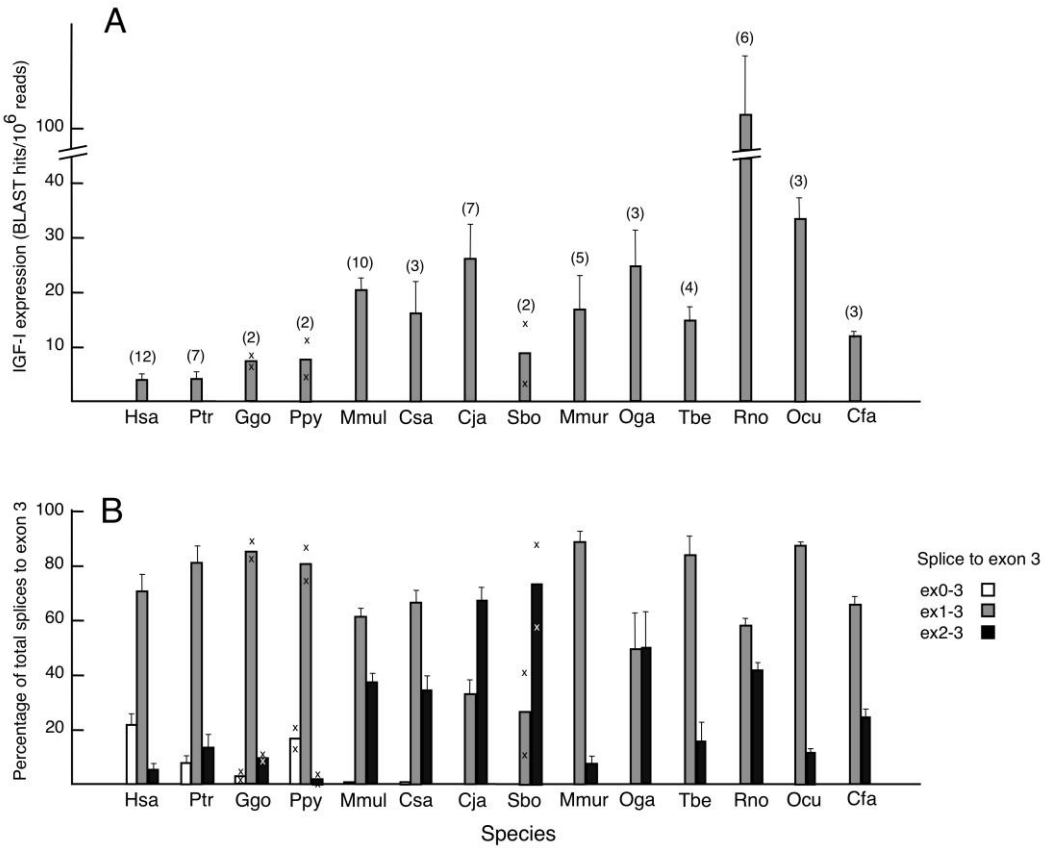
558 Fig. 5.



559

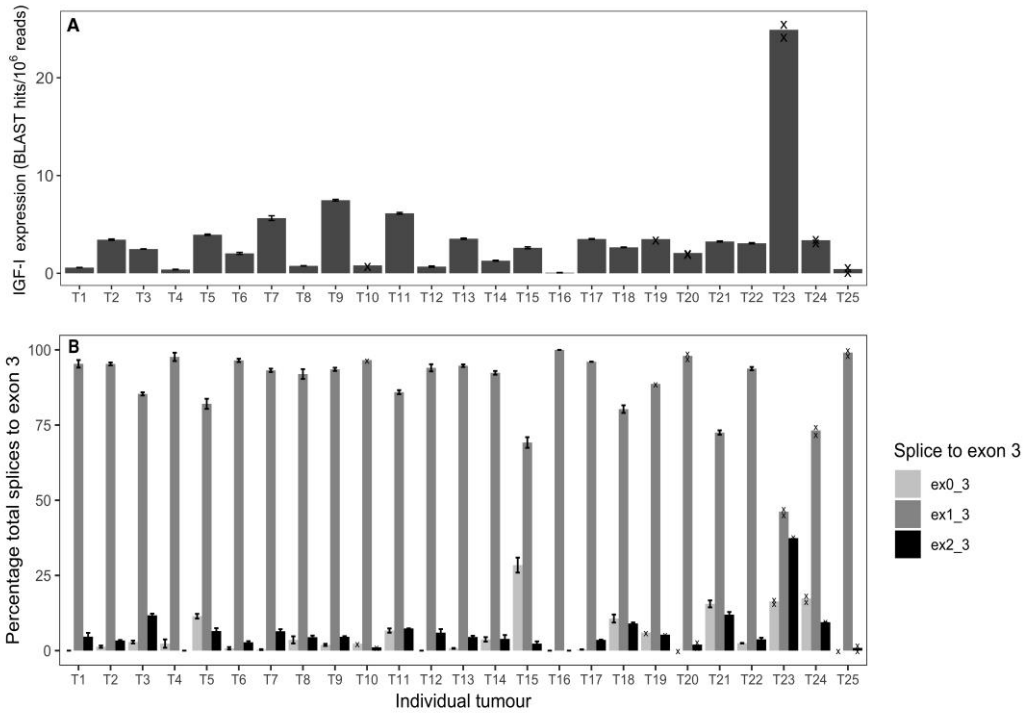
560

561 Fig. 6.



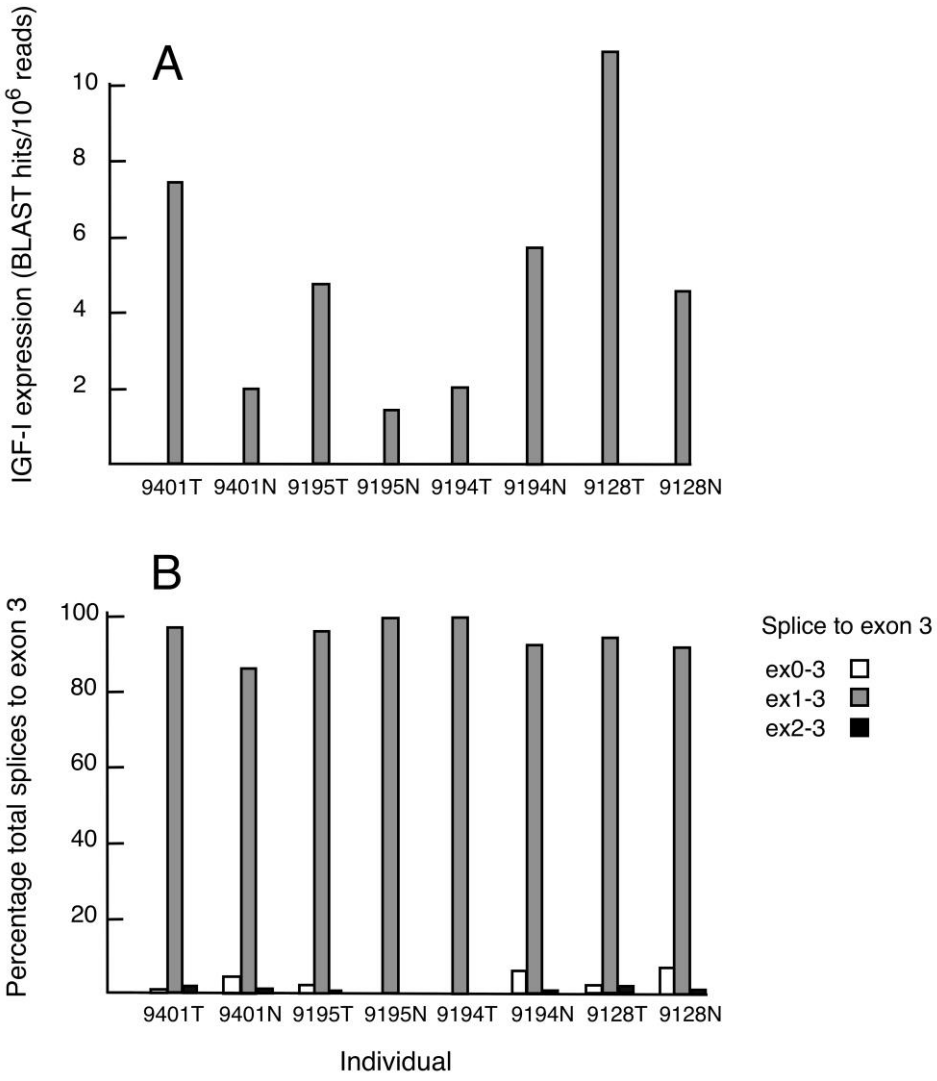
562

563 Fig. 7.



564

565 Fig. 8.


A
IGF-I expression (BLAST hits/10^6 reads)
10
8
6
4
2
9401T 9401N 9195T 9195N 9194T 9194N 9128T 9128N
B
Percentage total splices to exon 3
100
80
60
40
20
9401T 9401N 9195T 9195N 9194T 9194N 9128T 9128N
Individual
Splice to exon 3
ex0-3
ex1-3
ex2-3

566

567 Fig. 8.

568