Mammalian Genome 13, 388–398 (2002).
DOI: 10.1007/s00335-001-3059-x

Mammalian Genome
Incorporating Mouse Genome

Comparative sequence analysis of the *INS–IGF2–H19* gene cluster in pigs

Valérie Amarger,[1] Minh Nguyen,[2] Anne-Sophie Van Laere,[1] Martin Braunschweig,[1] Carine Nezer,[2] Michel Georges,[2] Leif Andersson[1]

[1]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 597, SE-751 24 Uppsala, Sweden
[2]Department of Genetics, Faculty of Veterinary Medicine, University of Liege (B43), 20, bd. de Colonster, 4000 Liege, Belgium

Received: 18 July 2001 / Accepted: 12 March 2002

Abstract. *IGF2* is the major candidate gene for a paternally expressed Quantitative Trait Locus (QTL) in the pig primarily affecting muscle development. Here we report two sequence contigs together comprising almost 90 kb containing the *INS–IGF2* and *H19* genes. A comparative sequence analysis of the pig, human, and mouse genomic sequences was conducted to identify the exon/intron organization, all promoters, and other evolutionarily conserved elements. RT-PCR analysis showed that *IGF2* transcripts originated from four different promoters and included various combinations of seven untranslated exons together with three coding exons, in agreement with previous findings in other mammals. The observed sequence similarity in intronic and intragenic regions among the three species is remarkable and is most likely explained by the complicated regulation of imprinting and expression of these genes. The general trend was, as expected, a higher sequence similarity between human and pig than between these species and the mouse, but a few exceptions to this rule were noted. This genomic region exhibits several striking features, including a very high GC content, many CpG islands, and a low amount of interspersed repeats. The high GC and CpG content were more pronounced in the pig than in the two other species. The results will facilitate the further characterization of this important QTL in the pig.

A paternally expressed QTL (Quantitative Trait Locus) affecting muscle mass in pig has been identified at the distal end of pig Chromosome (Chr) 2p. The QTL was mapped independently in a Large White/Pietrain intercross (Nezer et al. 1999), a Wild Boar/Large White intercross (Jeon et al. 1999), and later in an intercross between Landrace/Large White and Meishan pigs (de Koning et al. 2000). Pig Chr 2p1.7 shows conserved synteny with human chromosome 11p15, which is extensively studied because of the presence of a cluster of imprinted genes. Among them, the insulin-like growth factor II (*IGF2*) gene is paternally expressed and was identified as the major candidate gene for the QTL, because of its involvement in muscle growth and differentiation (Florini et al. 1995). The paternal expression of pig *IGF2* has been confirmed (Nezer et al. 1999).

IGF2 is flanked on its 5′ and 3′ sides by the insulin (*INS*) and *H19* genes, respectively, and these three genes cover a region of about 150 kb on human Chr 11p15 and mouse Chr 7 (Zemel et al. 1992; Onyango et al. 2000). These three genes seem to have a closely related regulation and have been extensively studied because of their involvement in several pathologies. The VNTR present in the 5′ region of the human *INS* gene is associated with susceptibility to insulin-dependent diabetes mellitus (IDDM; Bennett et al. 1995). This VNTR has an effect on *INS* mRNA levels (Pugliese et al. 1997; Vafiadis et al. 1998), and it also influences the expression of *IGF2* in human placenta *in vivo* (Paquette et al. 1998). However, this transcriptional effect is absent in leukocytes (Vafiadis et al. 1998), suggesting a tissue-specific regulation dependent on the particular promoter used for *IGF2* transcription. *IGF2* is a complex transcription unit that consists of 10 exons in human (Mineo et al. 2000). The first seven exons (denoted 1–6 and 4b) are non-coding leader exons, while exons 7–9 encode pre-pro IGF2 consisting of 180 amino acid residues. Exons 1, 4, 5, and 6 are preceded by distinct promoters (P1–P4), which give rise to a family of mRNA transcripts containing different leader exons but the same coding exons (Holthuizen et al. 1990). The different promoters confer a tissue-specific as well as a development-specific expression of the gene.

IGF2 and *H19* are expressed in a monoallelic fashion from the paternal and maternal chromosomes, respectively, and their imprinting is closely co-regulated. Over-expression of *IGF2*, with or without disruption of the imprinting pattern of itself and *H19* is implicated in several disorders in the human, mostly growth disorders and tumors. For instance, the Beckwith-Wiedemann syndrome shows evidence of both *H19*-dependent and *H19*-independent pathways affecting the *IGF2* imprinting status (Brown et al. 1996; Reik et al. 1995).

Considering the complex regulation of *IGF2* and the close interaction between *INS*, *IGF2*, and *H19*, our first step in understanding the molecular basis of the QTL effect was to sequence the region covering the three genes in pig. Because no difference in the coding sequence of *IGF2* was identified in animals carrying different QTL alleles, the causative mutation(s) may be regulatory (Nezer et al. 1999). Comparative sequencing is a powerful tool to identify functionally important sequences that are evolutionarily conserved even between distantly related organisms. Human and mouse comparative sequence analysis was used to identify new genes and potential regulatory elements in the human Chr 11 imprinted domain (Onyango et al. 2000; Ishihara et al. 2000). In this study, we used comparative sequence analysis of pig, human, and mouse to define the organization of these three genes in the pig and to identify potential regulatory elements that could be responsible for the QTL effect.

The genomic and cDNA sequence data reported in this paper have been submitted to GenBank and have been assigned the accession numbers AY044827–AY044828 and AF466293–AF466299.

V.A. and M.N. contributed equally to this work

Present address of V.A.: UMR1061 INRA/Université Limoges, 123 av. Albert Thomas, 87060 Limoges, France.

Correspondence to: L. Andersson; *e-mail:* Leif.Andersson@bmc.uu.se

Materials and methods

Human and mouse sequence data. The human sequence covering the *INS–IGF2* regions is a combination of two overlapping sequences available in GenBank; LI5440 (from 1 to 12348) and AC006408 (from 69001 to 42189, reverse complemented). The mouse sequence covering this region was taken from the sequence AC012382. The human *H19* sequence is from AC004556, and the mouse *H19* sequence from AP003182 and AF049091.

Restriction mapping of the BAC clones. BAC DNA was purified by using the QIAGEN plasmid midi kit (QIAGEN, Germany). Two microgram of BAC DNA was digested with 10 units of *Not*I restriction enzyme. The fragments were separated by PFGE. Gels were run at 4 V/min for 16 h at 14°C with the pulse times ramped from 0.1 to 2.5 sec. Following electrophoresis, gels were stained with ethidium bromide, and the fragments were visualized by exposure to UV light. *Not*I restriction fragments were subcloned into pNEB193 (New England Biolabs).

BAC sequencing. DNA from BAC 370 was purified by an alkaline lysis method followed by phenol/chloroform extraction. Twenty microgram of DNA was partially digested by 10 units *Sau*3AI (New England Biolabs) for 10 min at 37°C. Digested DNA was separated on 1% agarose gels, and fragments between 1.5 and 2.5 kb were excised and purified with the QIAEX II kit (QIAGEN, Germany). About 100 ng of purified DNA was ligated into 100 ng of *Bam*H1 restricted pUCI8 (Amersham-Pharmacia Biotech, Uppsala, Sweden) by using T4 DNA ligase (New England Biolabs) and was used to transform XL1 blue *E. coli*. Plasmid DNA was prepared by an alkaline lysis method and sequenced with universal M13 reverse and forward primers by using the Big Dye Terminator sequencing kit (Perkin Elmer Applied Biosystem). Sequences were run on an ABI 377 automatic sequencer (Perkin Elmer Applied Biosystem). Difficult templates with a very high GC content and long stretches of Gs or Cs were sequenced by using the dGTP Big Dye Terminator kit (Perkin Elmer Applied Biosystem).

*Not*I restriction fragments containing the *H19* region were sequenced using the EZ::TN™ Transposon Insertion System (Epicentre Technologies, Madison, WI). Transposon inserted recombinant plasmid DNA was sequenced as described above.

Bioinformatic analysis. Sequences were assembled by using the Phred/Phrap/Consed package (Ewing et al. 1998; Gordon et al. 1998). The assembled sequences were then analyzed with a variety of computer software programs. Sequence comparison with cDNA sequences was done with pairwise BLAST at http://www.ncbi.nlm.nih.gov. Repetitive elements were localized and identified by RepeatMasker (A.F.A. Smit & P. Green, unpublished; http://ftp.genome.washington.edu/index.html). In order to detect pig specific interspersed repeats, the mammalian library of repeats provided with the program was updated with a consensus pig SINE sequence and other pig-specific repeats. Sequence identity plots were obtained by using VISTA (Dubchak et al. 2000; Mayor et al. 2000) at http://www-gsd.lbl.gov. The comparison between pig and human sequences was done with Alfresco (Jareborg and Durbin 2000). Alfresco uses the program CpG (G. Micklem and R. Durbin, unpublished) for determining the presence of CpG islands. By default, a CpG island is defined as a DNA stretch at least 200 bp long with a GC content > 50% and an observed-to-expected ratio of CpG dinucleotides > 0.6 (Gardiner-Garden and Frommer 1987). Conserved elements were identified by using DBA (included in Alfresco) and pairwise BLAST, as said above.

RT-PCR analysis of IGF2 transcripts. Adult and fetal tissue samples were immediately frozen in liquid nitrogen and stored at −70°C or in RNAlater™ (Ambion) until total RNA was prepared by using TRIzol (GIBCO BRL) according to the manufacturer's protocol. The isolated RNA was DNAse I (Ambion) treated, which was subsequently inactivated by phenol–chloroform extraction. First-strand cDNA synthesis was done by using total RNA samples following the manufacturer's instructions (Amersham Pharmacia Biotech).

RT-PCR was carried out with the Advantage®-GC cDNA PCR kit (CLONTECH). The following primers were used to determine the usage of the four promoters (P1–P4): P1, forward primer IGF2EX1F 5′-GGTAGGCGGCTGGGATGAGTGG-3′ in exon 1 and reverse primer IGF2EX8R 5′-TGCCGGCCTGCTGAAGTAGAAG-3′ in the junction between exon 7 and 8; P2, forward primer IGF2EX4F 5′-TCCCTGGGTCTTCCAACGGACTGGGCGT-3′ in exon 4 and reverse primer IGF2EX7R 5′-CTCACTGGGGCGGTAAGCAGCA-TAGCAG-3′ in exon 7; P3, forward primer IGF2EX5F 5′-CGGCC-CGTCCTCCCCAAACAATCAGAC-3′ in exon 5 and reverse primer IGF2EX7R 5′-GGGCGGTAAGCAGCATAGCAGCACGAG-3′ in exon 7; and P4 forward primer IGF2EX6F 5′-GGCAGGCTCC-CAGCTTCCTCCTCCTCC-3′ in exon 6 and the reverse primer IGF2EX9R 5′-GGGCGGACTGCTTCCAGGTGTCATAGC-3′ in exon 9. The obtained PCR products were isolated from agarose gels and sequenced directly on a MegaBACE™ 1000 sequencing instrumently, using the DYEnamic™ ET dye terminator cycle sequencing kit (Amersham Pharmacia Biotech).

Results

Restriction mapping of the pig BAC clones. Two BAC clones (BAC 253 and 370) containing *IGF2* were isolated from a pig genomic library by using *IGF2* primers (Jeon et al. 1999). DNA from these two clones was digested with the restriction enzyme *Not*I, and the resulting restriction fragments were separated by conventional as well as pulsed field gel electrophoresis (PFGE). BAC 253 and 370 contained nine and seven *Not*I, restriction fragments, respectively, ranging in size from 1.2 to 63 kb. All these fragments (except the 63-kb fragment) were subcloned and sequenced from both ends. Outward-pointing primers were designed for all subclones and used for sequencing with the BAC DNA as template. Comparison of the obtained sequences and the end sequences of the subclones allowed unambiguous ordering of all *Not*I restriction fragments (Fig. 1A). The location of the *INS* and *H19* genes in BAC 370 and 253 was established by PCR amplification and sequencing. The tyrosine hydroxylase (*TH*) gene was identified during the sequencing process.

Sequencing of the INS–IGF2 and H19 regions. A shotgun library was constructed for BAC 370 containing the *INS–IGF2* region. One thousand clones were sequenced from both ends, giving approximately 1600 high-quality sequences. After assembly, three strategies were used to fill gaps. The first strategy involved a simple primer walk when a gap was found between the two ends of a shotgun clone. The second involved PCR amplification (with the BAC DNA as template) of a gap situated between two contigs that could be ordered and oriented after a comparative analysis by using the homologous human sequence. The third strategy involved subcloning of BAC restriction fragments covering a gap, followed by sequencing by primer walk. The 24-kb *Not*I fragment containing *H19* as well as the 1.2-, 1.5-, 6.1-, and 24-kb *Not*I fragments containing the *H19* upstream region were subcloned and sequenced with transposon insertions. A 32-kb contiguous sequence containing the last five exons of *TH* and the complete *INS* and *IGF2* genes and a 56-kb sequence containing the *H19* gene were determined. All the *Not*I sites of the restriction map were found in the sequence, allowing a precise localization of the genes on the restriction map (Fig. 1). The distances between the genes were determined as follows: *TH*–1.9 kb–*INS*–0.7 kb–*IGF2*–88.1 kb–*H19*.

The GC content of the two sequenced regions is significantly higher in pig than in the corresponding region in human (Table 1) and in both species much higher than the genome average. A large number of CpG islands was also identified and, in line with the difference in the GC content, the number and sizes are larger in pig than in human. This is most pronounced in the *H19* region, where the total length of CpG islands is about 10 times higher in pig than in human (Table 1; Fig. 1).

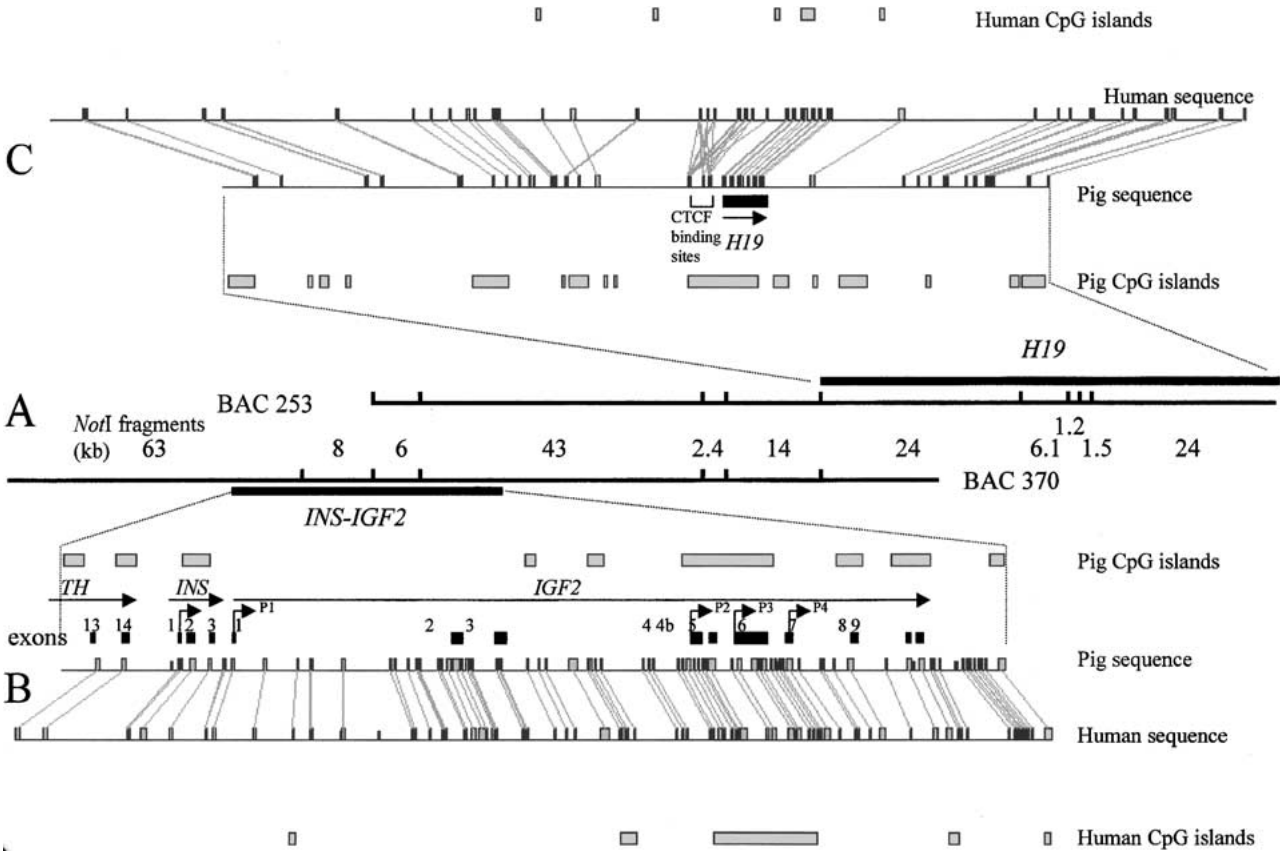

Fig. 1. **A**: *Not*I restriction map of two pig BAC clones (253 and 370) including the *INS–IGF2* and *H19* regions. **B** and **C**: Comparative sequence analysis of the pig and human sequences using Alfresco. Conserved elements are represented by two boxes linked by a line. Shaded and solid boxes are conserved elements identified by BLAST and DBA algorithms, respectively. Positions of exons, promoters, and CpG islands are indicated.

Table 1. Global sequence comparison of the human and pig *INS–IGF2* and *H19* regions.

	INS–IGF2		*H19*	
	Pig	Human[a]	Pig	Human[b]
Length (bp)	32,467	35,647	56,404	70,750
% GC	65.0	61.4	67.5	61.6
CpG islands	9	5	16	6
Total length of CpG islands (bp)	8,605	4,827	17,622	1,940
Total number of CpG dinucleotides	823	509	1,246	175
Interspersed repeats (% of total sequence)	0.47	1.41	1.66	8.21
Simple repeats (% of total sequence)	3.97	4.51	1.29	1.24
Total repeats (% of total sequence)	4.44	5.92	2.95	9.45

[a] The human *INS–IGF2* sequence is a combination of two overlapping sequences: L1 5440 (1 to 12348) and AC006408 (45702 to 69001, reverse complementary strand).
[b] GenBank sequence AC04556 (25967 to 96716, reverse complementary strand).

The sequences were screened for repetitive sequences by using RepeatMasker (Table 1). The total number of repeats is higher in human than in pigs. In the *H19* region, this difference is due only to the proportion of interspersed repeats, which is more than four times higher in human than in pig, whereas the number of simple repeats is similar. The *INS–IGF2* region is characterized by a surprisingly low amount of interspersed repeats and a high proportion of simple repeats in both species. The only interspersed repeat found in the pig *IGF2* gene was a Mammalian Interspersed Repeat (MIR), and its location was conserved between pig and human; MIR elements represent a class of short interspersed repeats (SINE) found in all mammals.

Structures of the pig INS, IGF2, and H19 genes. The exon/intron organization of the pig genes was deduced by aligning the genomic sequence with cDNA sequences from pig when available (*INS* mRNA AF064555; *IGF2* mRNA X56094 and RT-PCR products described below) or from human (*H19*). We identified 10 *IGF2* exons in the pig, and the corresponding ten exons of human *IGF2* were identified by a combination of several mRNA sequences obtained from different tissues (GenBank M22372, X06259, X03423, Y13633, X56539, X56540, and X03562). The *H19* exons in pig were identified by sequence similarity with the human mRNA sequence (GenBank M32053). Exon and intron sizes, and sequence identities between the human and pig genes are presented in Table 2. *IGF2* exons 7, 8, and 9 are coding and evolutionarily well conserved. However, the seven untranslated exons are also very well conserved, with sequence identities ranging from 74 to 91%. The insulin gene, although all exons are coding, is less conserved than *IGF2*, and the first exon, encoding the signal peptide, is only 61% identical between the two species. The

Table 2. Comparative exon/intron organization of the human and pig *INS*, *IGF2*, and *H19* genes.

Gene Exon/intron	Exon length (bp)		Pairwise align. score[a] (%)	Intron length (bp)		Pairwise align. score[a] (%)
	Pig	Human		Pig	Human	
INS						
1	43	42	61	162	179	70
2	203	204	81	391	787	<50
3	140	146	85			
IGF2						
1	112	117	83	7440	8922	<50
2	193	220	83	1260	1318	68
3	232	242	74	6157	6319	66
4	389	390	78	536	557	74
4b	165	165	90	648	692	69
5	1148	1164	89	859	941	73
6	115	100	90	1817	1658	<50
7	163	163	88	1909	1705	<50
8	145	145	91	258	293	<50
9	240	237	88			
H19						
1	1354	1328	76	88	96	<50
2	126	135	77	78	95	51
3	118	113	54	83	80	<50
4	128	123	82	82	81	56
5	550	614	70			

[a] Pairwise alignment scores were determined by using ClustalW at http://www.ebi.ac.uk; ClustalW could not perform correct alignments for sequence identities <50%.

Fig. 2. Sequence alignment of the human and pig promoter regions for the *INS, IGF2*, and *H19* genes. The lower case letters mark the transcription start. Human promoter sequences were found in the Eukaryotic Promoter Database (http://www.epd.isb-sib.ch). Identities to the human sequences are indicated by dots and alignment gaps are indicated by dashes.

structure of *H19* is conserved, but the level of sequence identity is on average lower than it is for *IGF2*. The level of sequence identity of intronic sequences shows considerable variation. Some introns are as conserved as exons, such as, for example, *INS* intron 1 (70%), *IGF2* introns 2, 3, 4, and 5 (68, 66, 74, and 73%, respectively).

The human promoter sequences for *INS*, *IGF2*, and *H19* are available in the Eukaryotic Promoter Database (http://www.epd.isb-sib.ch/; Perier et al. 2000). We have used these sequences to identify the pig promoters by sequence similarity search. The positions of the promoters are conserved, and they all show a high sequence identity to the human ones (Fig. 2).

Characterization of IGF2 transcripts and promoter usage in fetal and adult tissue. The *IGF2* transcripts and the promoter usage for different adult and fetal porcine tissues documented by RT-PCR analysis are compiled in Table 3. Promoter 1 (P1) usage was predominantly detected in adult liver, fetal ham, and fetal liver. In addition to the transcript documented in the previously reported *IGF2* cDNA sequence (X560940) containing exons 1,3 and 7–9, we found a P1 transcript without exon 3 and a P1 transcript including exon 2. The P1 usage in fetal ham but not in adult muscle indicates a developmental specific usage of P1. The same applies to the two P1 transcripts found in adult liver, but not in fetal liver. Promoter 2 to 4 (P2–P4) transcripts were detected in all investigated tissues. P2 usage resulted in two different transcripts, and the results showed that the transcript including exon 4b is much less abundant than the transcript without it. The sequencing results of the three pooled PCR products from adult muscle, liver, and kidney were consistent with the assumption that these transcripts containing exon 4b are identical. The finding of an *IGF2* exon 4b in pig agrees well with the results published by Ohlsen et al. (1994) for sheep, and by Mineo et al. (2000), who confirmed the existence of a 10th exon of *IGF2* in human.

Comparative analysis of the putative Nctc1/Rhit1 and Ihit1 transcription units. A mouse transcription unit mapping in the *Mrp123–H19* interval (~17 kb downstream of *H19*) was previously reported by Ishihara et al. (1998) on the basis of ho-

Table 3. Characterization of porcine *IGF2* transcripts and promoter (P1–P4) usage in adult and fetal tissues by using RT-PCR.[a]

	IGF2 promoter usage						
	P1			P2		P3	P4
Tissues	Exons 1,7–9	Exons 1,3,7–9	Exons 1–3,7–9	Exons 4,7–9	Exons 4,4b,7–9	Exons 5,7–9	Exons 6,7–9
Adult muscle		(+)		+	{+}	+	+
Adult liver	+	+	+	+	{+}	+	+
Adult kidney		(+)	(+)	+	{+}	+	+
Fetal ham	(+)		+	+	(+)	+	+
Fetal liver			+	+	(+)	+	+
Fetal kidney				+	(+)	+	+
Fetal heart				+	(+)	+	+
Fetal brain		(+)	(+)	+	(+)	+	+
Fetal placenta		(+)		+	(+)	+	+
Fetal lung				+	(+)	+	+

[a] The reverse PCR primers were located at the junction of exon 7 and 8 for the P1 amplicon, in exon 7 for the P2 and the P3 amplicons, and in exon 9 for the P4 amplicon. We assume that all transcripts contain the coding exons 7–9.
(+) = very faint PCR product of the corresponding size was obtained but not sequenced.
{+} = PCR products were pooled together and sequenced.

mologies with ESTs that were subsequently confirmed by Northern blot and RACE analyses. It was referred to as *Nctc1* by Ishihara et al. (1998) and subsequently as *Rhit1* by Onyango et al. (2000). The corresponding transcripts are non-imprinted, non-coding, and primarily expressed in skeletal muscle. The *Nctc1* transcription unit is characterized by an upstream exon of about 200 bp separated by an intron of 2235 bp from a downstream exon of at least 2474 bp. The 3′ end of this exon, however, does not exhibit a consensus polyadenylation signal and is bounded in the mouse by a genomic poly-A track, which is predicted to have primed the reverse transcription step. It is, therefore, unlikely to correspond to the genuine 3′ end of these transcripts.

The 5′ end of *Nctc1* corresponds to an evolutionary footprint conserved between mouse, human, and pig (no. 93–97 in Table 4). However, with the exception of a human EST (AF313096) of unknown origin reported by Onyango et al. (2000), no other human or pig EST mapping to this region could be identified. The AF313096 EST spans the exon 1–intron 1 boundary and would, therefore, correspond to an unspliced message. It is noteworthy in this regard that the corresponding donor splice site is not conserved in the human. BLAST searches performed with the AF313096 EST revealed a very significant, perfect 144-bp match with exon 13 of a gene predicted *in silico* to be transcribed from the *Nctc1* antisense strand. This putative XM_073653 gene comprises 14 exons spanning about 20 kb between *H19* and *Mrp123*. Alignment of the corresponding exons with the corresponding human and pig sequences did not reveal any evidence for conservation of the corresponding intron–exon boundaries or open reading frame. We therefore suspect that XM_073653 is a false-positive gene prediction.

Onyango et al. (2000) also identified a 282-bp open reading frame about 21 kb upstream of *H19*. The corresponding murine DNA sequence was reported to reveal strong 1.0-kb transcripts in murine and human liver, placenta, and brain (human), by Northern blot analysis. It was referred to as *Ihit1*. The corresponding sequence is, however, conserved neither in the human nor in the pig. Genescan analysis of the corresponding regions in human, mouse, and pig did not provide evidence for statistically well-supported, evolutionarily conserved exons. *Ihit1* is, therefore, considered as a false«positive gene prediction as well.

Highly conserved non-coding elements. Global sequence identity plots of the *INS–IGF2* and *HI9* regions between pig, human, and mouse show that the level of identity between human and pig is higher than that between human and mouse, or pig and mouse (Fig. 3). The exons are highly conserved, but a large number of intronic and intergenic regions are also remarkably well conserved among the three species. The Alfresco program (Jareborg and Durbin 2000) was used to perform a comparative analysis between the pig and human sequences and to identify evolutionarily conserved elements (Fig. 1). The sequence of every conserved element was then compared with the mouse sequence by pairwise BLAST (http://www.ncbi.nlm.nih.gov), with the following parameters: word size 7; penalty for a mismatch −1. The elements that were found to be conserved also in mouse are indicated in Table 4. Alfresco was also used to compare the pig and mouse sequence, and it gave similar results (data not shown). Several human regulatory elements already known were, as expected, found in the pig sequence, together with many other conserved elements whose functions remain to be determined. Alfresco uses an algorithm named DBA (DNA Block Aligner; Jareborg et al. 1999), which was designed to identify small conserved motifs in non-coding sequences that are difficult to align. When comparing the human and pig sequences, we found that the results of DBA were identical to the results obtained with the pairwise BLAST by using defaults parameters (word size 11, penalty for a mismatch −2).

We have focused primarily on regulatory elements that are assumed to be involved in the regulation of *IGF2* expression. Two conserved elements upstream of the insulin promoter (no. 1 and 2 in Table 4) are probably involved in the regulation of transcription. There was no evidence a corresponding pig VNTR sequence as present about 500 bp upstream of the human *INS* gene (Bennett et al. 1995). The human VNTR consensus unit is ACAGGGGTGTGGGG, which creates a very GC-rich region with strong strand disequilibrium of G and C nucleotides. However, the two motifs AGGGG and TGGGG are found five and seven times, respectively, together with several similar motifs (for instance CGGGG, TGGGT, AGGGT, AGGGA, and AGGGC) in the corresponding pig sequence, but they are not organized in a tandem repeat. There is no strand disequilibrium in the pig because of the presence of several stretches of Cs.

A conserved CpG island of 650 bp is found about 3 kb upstream of *IGF2* exon 4, harboring 87 and 70 CpG dinucleotides in pig and human, respectively. Several short sequence elements within this CpG island are very well conserved (no. 27–31 in Table 4). This region corresponds to a suggested control element, denoted differentially methylated region 1 (DMR1) in mouse, which shows a differential methylation

Table 4. Conserved elements (outside exons, promoters and simple repeats) found in the *INS–IGF2* (1 to 59) and *H19* (60 to 97) regions. Positions of *INS*, *IGF2*, and *H19* exons are indicated by arrows.

Conserved elements		Position in pig sequence	Length bp	Sequence Identity %	Comment
INS–IGF2 region					
	1	3997–4014	18	100	mouse[c]
	2	4070–4116	47	89	mouse[c]
INS →	3	5536–5557	22	90	
IGF2	4	5574–5621	48	89	
ex1→	5	6945–7066	122	87	
	6	8066–8099	34	94	mouse[c]
	7	8531–8553	23	91	mouse[c]
	8	8600–8652	53	94	mouse[c]
	9	9660–9749	90	89	MIR repeat
	10	11303–11350	48	89	
	11	11449–11487	39	94	
	12	11869–11904	36	91	
	13	12185–12206	22	95	
	14	12220–12236	17	100	
	15	12279–12295	17	100	
	16	12955–12984	30	87	
	17	13038–13068	31	90	
	18	13221–13379	159	85	mouse[c]
ex2 →	19	13710–13771	62	83	
	20	13897–13976	80	90	
	21	14038–14059	22	91	Similar to mouse anti–sense transc. AB030734
	22	14089–14131	43	88	
ex3 →	23	15994–16023	30	90	
	24	16358–16398	41	95	
	25	16640–16667	28	89	
	26	17406–17715	308	84	mouse[c]
	27	18069–18092	24	91	
	28	18149–18263	115	84	
	29	18326–18345	20	95	CpG island
	30	18523–18553	31	97	(mouse DMR1)[c]
	31	19975–19997	23	100	
	32	20196–21231	36	89	
ex4 →	33	20346–20371	26	92	
	34	21171–21209	39	90	mouse[c]
	35	21884–21923	40	87	
	36	22041–22093	53	100	CpG island
	37	22119–22135	17	100	*IGF2* intron 4
ex4b →					
ex5 →	38	24363–24429	67	85	
	39	24500–24540	41	93	
	40	24659–24695	37	89	*IGF2* intron 5
	41	24756–24793	38	95	
	42	24907–24938	32	92	
ex6 →	43	25339–25373	35	91	
	44	26059–26090	32	91	
	45	26162–26183	22	95	
	46	26536–26588	53	89	mouse[c]
ex7 →	47	28303–28367	65	84	
ex8,9 →	48	29880–29909	30	93	
	49	29947–29978	32	94	
	50	30135–30173	39	92	
	51	30945–30966	22	100	
	52	31123–31160	38	95	
	53	31203–31229	27	96	
	54	31333–31349	17	100	
	55	31377–31394	18	94	*IGF2* 3′ UTR
	56	31508–31526	19	95	
	57	31584–31655	72	88	
	58	31736–31782	47	83	
	59	32180–32438	259	85	
	H19 region				
	60	2099–2119	21	100	
	61	2258–2326	69	86	
	62	3997–4074	78	95	
	63	9333–9798	65	86	
	64	9858–9893	36	94	
	65	10847–10875	29	96	
	66	10933–11029	97	91	
	67	16095–16139	45	89	
	68	16215–16244	30	90	
	69	18473–18510	38	90	
	70	19276–19371	96	84	Similar AF313051[a]
	71	20266–20270	45	84	
	72	20899–21030	124	92	Similar AF313050[a]
	73	21232–21255	24	96	
	74	22416–22498	83	88	
	75	22585–22609	25	92	
	76	22701–22744	44	93	
	77	24280–24359	80	90	
	78	25472–25758	287	86	
	H19 → 79	34222–34252	31	93	
	80	40176–40487	312	86	Enhancer I[c]
	81	42616–42771	156	81	Enhancer II[c]
	82	46454–46525	72	86	CS5[b,c]
	83	47476–47528	53	87	mouse[c]
	84	48223–48259	37	95	
	85	49281–49316	36	97	
	86	49351–49397	47	91	CS6[b,c]
	87	49495–49557	63	84	
	88	50758–50794	37	92	
	89	51368–51398	31	97	CS7[b,c]
	90	51416–51475	60	89	
	91	52142–52196	55	95	
	92	52194–52257	64	95	CS8[b,c]
	93	52445–52581	137	82	Similar AF313096[a]
	94	52617–52635	19	100	
	95	54991–55021	31	87	mouse[c]
	96	55070–55115	46	87	
	97	56348–56386	39	87	

[a] Human / mouse conserved elements (Onyango et al. 2000).
[b] Conserved putative enhancers (Ishihara et al. 2000).
[c] Human / mouse conserved elements (this study).

pattern, and the unmethylated allele is associated with a silencer function in several mesodermal tissues (Constancia et al. 2000). The transcription factor GCF2 (GC binding factor) was suggested to repress the transcription through binding the unmethylated allele. Methylation would prevent binding of the repressor on the paternal allele (Eden et al. 2001). However, the imprinting status of *Igf2* in mouse skeletal muscle is independent of the DMR1 methylation, suggesting an important, tissue-specific regulation of the imprinting (Weber et al. 2001). The 54-bp core region of differentially methylated region 2 (DMR2) present in mouse *Igf2* exon 6 (Murrell et al. 2001) is also well conserved in the pig sequence (89% identity).

A large CpG island of about 3.5 kb containing the region from *IGF2* exon 4 to a part of intron 5 is found in both pigs and human. A large number of conserved sequences (no. 35–42 in Table 4) are found in this region and in a part of intron 5 outside the CpG island. Introns 4 and 5 harbor several highly conserved elements, which are likely to be involved in the regulation of transcription from promoters P3 and P4. Among these elements, the SP1 binding motif GGGGGCGGGGC-GAGG upstream of the P3 promoter is perfectly conserved as well as the P3–4 element described by Rietveld et al. (1997). Many other conserved elements with a length varying from 17 to 158 bp and whose function is still unknown are spread all over the *INS–IGF2* region (Table 4). The *IGF2* 3′UTR harbors a large number of conserved elements. These include several simple repeats, together with a group of non-repetitive elements (no. 49–59 in Table 4).

A complex pattern of conserved elements was found upstream of *H19*. A conserved motif repeated seven times in the human sequence was found three times in the pig (Fig. 4) be-

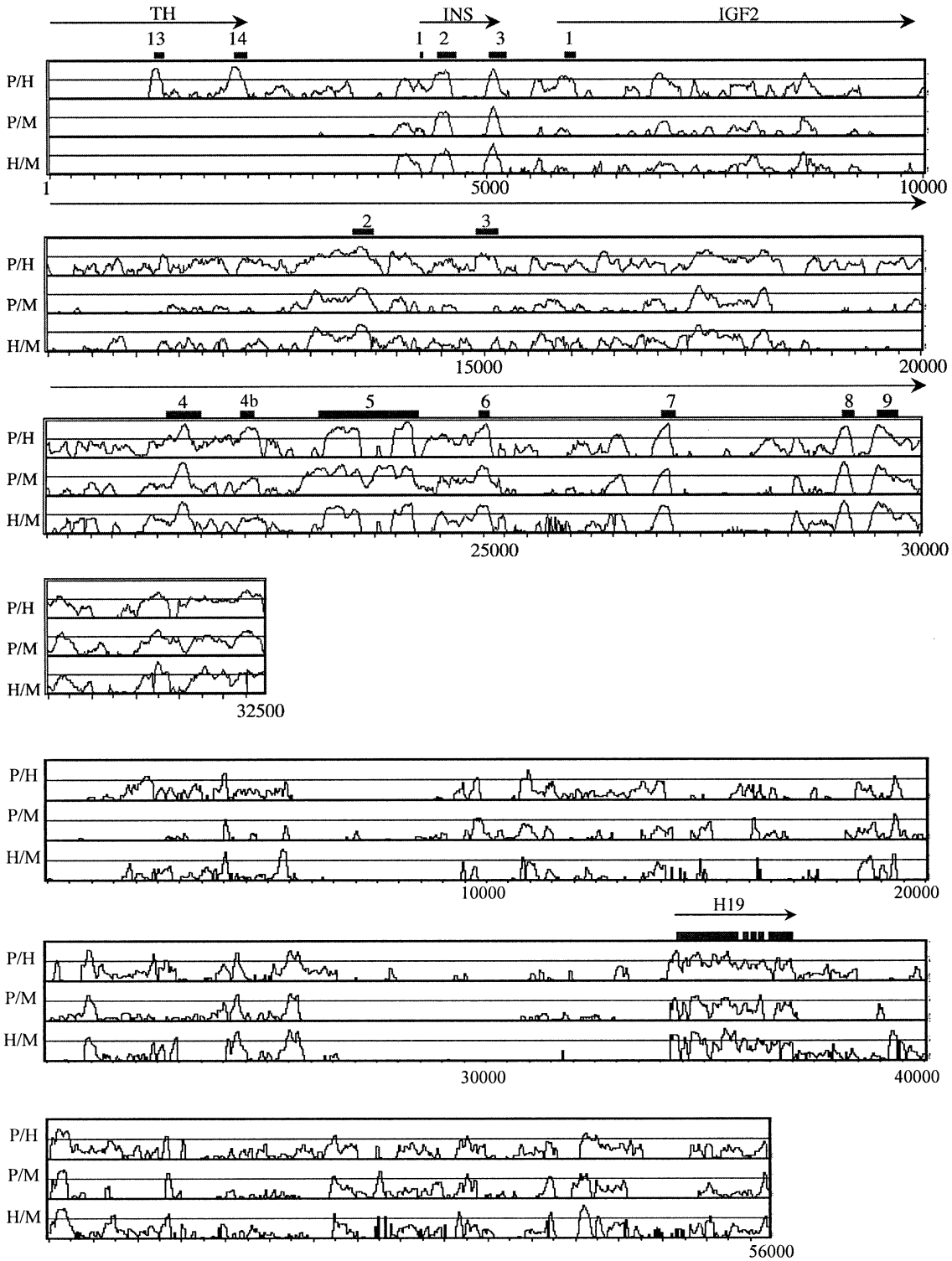

Fig. 3. Sequence identity plots for pairwise comparisons of the pig, human, and mouse *IGF2*, and *H19* genomic regions. P/H: pig/human, P/M: pig/mouse, H/M: human/mouse (percentage identity calculated with a window length of 100 bp). Conserved sequences are shown relative to their positions in the pig sequence (horizontal axes), and the percentage of identities (50%–100%) are indicated on the vertical axes. The locations of exons are shown above the profile.

tween −1.1 and −2.6 kb from the transcription start of *H19*. These repeats, previously identified in human, mouse, and rat (Frevel et al. 1999; Bell and Felsenfeld 2000), contain a 12-bp consensus sequence that is a binding site for CTCF (CCCTC-binding factor). The sequence identity of the repeats between human and pig extends outside the 12-bp repeat in a motif covering 95 bp (Fig. 4). CTCF is a vertebrate regulation factor able to form a large variety of DNA–protein complexes involved in distinct functions, including gene activation, repression, silencing, and chromatin insulation (Ohlsson et al. 2001).

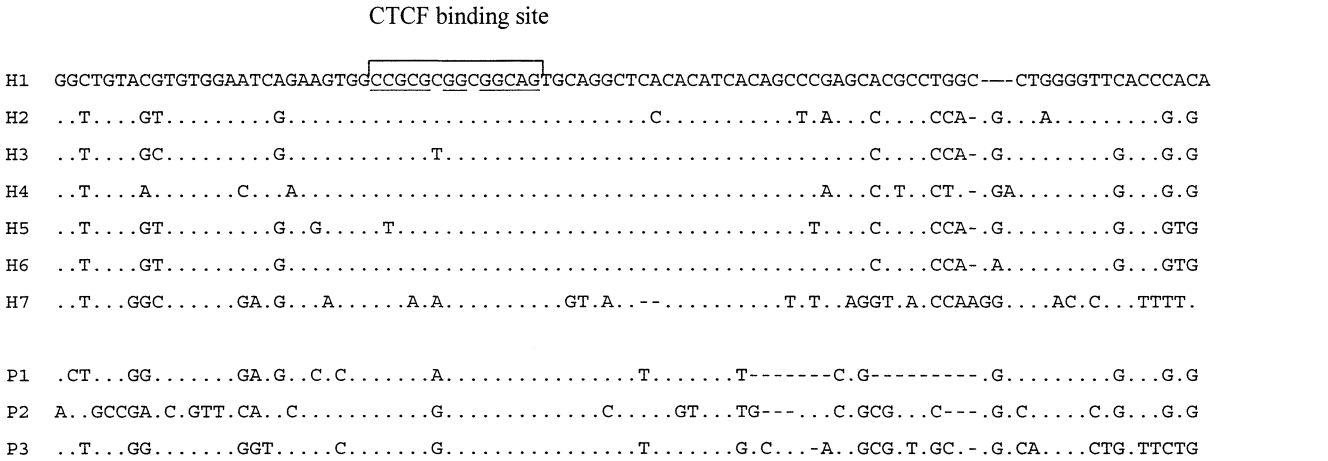

Fig. 4. Sequence alignment of the human (H1 to H7) and pig (P1 to P3) repeats upstream *H19* and containing the CTCF factor binding site. Identities to the human sequences are indicated by dots and alignment gaps are indicated by dashes.

The 5′ region of *H19* contains an imprinting mark characterized by paternal allele-specific methylation (Thorvaldsen et al. 1998). CTCF binding sites in this region are found in human, mouse, rat, and pig as well as in other imprinted domains having a similar regulation (Wylie et al. 2000) and were shown to contribute to imprint regulation of neighboring genes by the formation of chromatin boundaries (Bell and Felsenfeld 2000; Kaffer et al. 2000).

Several enhancer elements previously characterized in human and/or mouse (Webber et al. 1998; Ishihara et al. 2000) were found upstream and downstream of *H19*. In the pig, the two endoderm-specific enhancers are situated at 3.1 and 5.6 kb downstream *H19* (no. 80 and 81 in Table 4). Four other conserved sequences (CS) identified by Ishihara et al. (2000) between mouse and human were found in the same region in pig (no. 82, 85 + 86 + 87, 89 + 90, and 91 + 92 in Table 4, similar to CS5, 6, 7, and 8, respectively). CS5, 6, and 7 exhibit a tissue-specific enhancer activity, and CS6 is particularly interesting because it is primarily active in skeletal muscle (Ishihara et al. 2000). We have also found several other elements that were previously identified to be conserved between human and mouse, but their functions are still unknown (no. 70, 72, and 93 + 94 in Table 4).

Discussion

We present here a comparative sequence analysis of about 90 kb from the region containing the *INS–IGF2* and *H19* genes in pig, human, and mouse. The divergence between the human and pig lineage is estimated at 70 million years, whereas human and mice diverged approximately 100 million years ago (Andersson et al. 1996). The sequence similarity is high not only in exons, but also in introns and intergenic regions. However, it is likely that the identity scores are somewhat inflated owing to the high GC content and the many CpG islands in this region. The sequence identity plots for the three pairwise species comparisons were remarkably similar (Fig. 3). The general trend was a higher sequence similarity between pig and human than between these species and mouse as expected from the phylogenetic relationship. There are some interesting discrepancies between species that may very well be functionally important. The intergenic distance between *TH* and *INS* is much larger in the mouse (200 kb; Onyango et al. 2000) than it is in pig and human (2–3 kb). *IGF2* exon 3 is well conserved between pig and human (74%), but not in mouse (< 50% identity compared with human or pig, Fig. 3). The 5′ flanking region of *IGF2* exon 5 as well as part of this exon is well conserved between pig and mouse (>75%), but not in humans (<50%). These cases imply a faster substitution rate in one of these lineages suggesting an altered function for the actual region. The results illustrate how a comparative analysis based on three or more species adds power to the interpretation of genome evolution.

The programs (DBA and BLAST) that were used to identify conserved elements between human and pig are designed to identify short conserved elements inside large genomic regions that cannot be aligned. The DBA algorithm (applied in Alfresco) is supposed to be more sensitive than BLAST when conserved motifs are very short (Jareborg et al. 1999). In our case, the short, highly conserved elements identified by BLAST and DBA correspond to the top of the peaks displayed on the sequence identity plots, and most of them are conserved between the three species. We identified several 17-bp elements that are 100% conserved between pig and human (Table 4). We believe that this approach is useful to detect short regulatory elements like transcription factor binding sites. A conserved sequence at a conserved position is likely to be functionally important. Other parameters were used by Onyango et al. (2000) when comparing the l-Mb sequence covering the mouse and human imprinted domain. Their definition of a conserved elements was >100 bp with >70% nucleotide identity. All the elements we identified show >80%, identity, and several of them could be merged and considered as a single region with a lower degree of similarity, as illustrated on the identity plots. This would increase the proportion of aligned sequences (outside genes and promoters), in our case 9.7% and 4.6% in the *INS–IGF2* and *H19* regions, respectively, to a level closer to the 35.8% estimated for the human/mouse comparison of the *H19* and *Mash2* region (Onyango et al. 2000). When the same programs and same parameters were used to compare the human *H19* region with the pig *INS–IGF2* region, no conserved sequences were identified between these two unrelated regions. This shows that the majority of the conserved regions reported in this study are not random sequence similarities.

Our sequence data revealed the exon/intron organization, promoters, and other potential regulatory regions of the *INS*, *IGF2*, and *H19* genes in pig. The three genes are very well conserved between human and pig. The ten exons of human *IGF2* and the four promoters were identified in pig by comparative sequencing. Our RT-PCR analysis confirmed that all ten *IGF2* exons and the four promoters are used in the pig. It is worth noticing that the number of *IGF2* exons in humans are

generally considered to be nine, but our observation of a minor transcript containing a tenth exon (denoted 4b) is in perfect agreement with Mineo et al. (2000).

A very high GC content and an exceptional concentration of CpG islands characterize this genomic region in both human and pig. This tendency is significantly stronger in pig than in human, in which it was already shown to be higher than in mouse (Onyango et al. 2000). The small amount of pig genomic sequence publicly available is too limited to know whether this higher GC content is a general trend of the pig genome compared with the human genome or whether this is specific to this region. However, an independent comparative study of a 130-kb genomic region on pig Chr 15 and human Chr 2 (V. Amarger, in preparation) shows a similar GC content in both species (45.6% and 45.4% in pig and human, respectively). The presence and importance of CpG islands in the vicinity of imprinted genes is now well established. Although their action is not clearly demonstrated, it appears that CpG islands are necessary to establish and maintain imprinting patterns (Paulsen et al. 2000; Engemann et al. 2000). The role of CpG islands in gene regulation is strongly supported by comparison of the *Impact* gene that is imprinted in mouse but not in humans. The two homologs display a striking difference in the CpG island pattern in their upstream promoter region (Okamura et al. 2000). CpG islands are often associated with imprinting control regions (ICR) harboring a differential allelic methylation. The way methylation of these ICRs regulates expression involves methylation-sensitive DNA binding factors that can act as insulators (Holmgren et al. 2001) or repressors (Eden et al. 2001).

Another striking feature is the low proportion of interspersed repeats in the close vicinity of *INS*, *IGF2*, and *H19*, particularly in the pig (Table 1). It is possible that the pig estimates are slightly biased, since interspersed repeats are more studied in the human, but this cannot explain the observation of the large difference in interspersed repeats between the pig and human *H19* regions, 1.7% vs 8.2%. Furthermore, this difference is accompanied by a similar difference in GC content that should be unbiased. This remarkably low proportion of interspersed repeats seems to be specific to the *INS–IGF2–H19* region, since 30% of the human 1-Mb region on Chr 11p15 is composed of interspersed repeats (Onyango et al. 2000), and the average frequency for the human genome is about 45% (Lander et al. 2001). Thus, the *INS–IGF2* region in pigs and human and the *H19* region in pigs is as devoid of interspersed repeats as the *HOX* gene clusters, which recently were identified as being some of the most repeat-poor regions in the human genome (Lander et al. 2001). A low amount of interspersed repeats was also observed in the imprinted domain spanning from the *KCNQ1* to *CARS* genes together with a remarkably uneven distribution of these repeats, which appears to be conserved between human and mouse (Engemann et al. 2000). The reason that very few interspersed repeats are present in the *INS–IGF2–H19* region is not clear. However, the phenomenon may be related to the complicated regulation of gene expression and imprinting of these genes, and introduction of foreign sequences may disturb essential *cis*-regulatory mechanisms. In contrast, the frequency of simple repeats in the *INS/IGF2* region is slightly higher than the genome average (~4% vs ~3%). It has previously been suggested that simple repeats may play a role in the regulation of imprinted genes (Shibata et al. 1998), but no clear mechanism has been characterized so far. Among these repeats, a complex microsatellite element (several CA stretches with interruptions) in the 3′UTR of *IGF2* is conserved in human, pig, horse, and mouse (Jeon et al. 1999).

The VNTR upstream of the human insulin gene evidently has a functional role, but it was not found in the pig sequence. However, several 5-bp motifs present in the VNTR unit were found interspersed in the corresponding region in the pig. It is not clear how this VNTR affects insulin expression, but it may influence the chromatin structure, modulating the accessibility of transcription factors. It could also be a transcription factor-binding site. The fact that this VNTR is not conserved between two quite closely related species like human and pig, while many other regulatory elements are very well conserved, suggests that the function of this sequence might not be directly related to its repetitive structure. VNTRs are predominantly present in subtelomeric regions in the human genome (Amarger et al. 1998).

Several repeats are present in the 5′ region of *H19*. This region has been shown to contain an epigenetic mark required for the imprinting of both *H19* and *IGF2* (Tremblay et al. 1995; Thorvaldsen et al. 1998). It acts as an insulator (Kaffer et al. 2000), and its activity is dependent upon the vertebrate enhancer-blocking protein CTCF (Bell and Felsenfeld 2000). The methylation status of this region seems to be the major mechanism by which the insulator activity is modulated (Holmgren et al. 2001; Reed et al. 2001), and CTCF function marks the *Igf2/H19* expression domain in a parent-of-origin-dependent manner (Ohlsson et al. 2001). The loss of imprinting of *IGF2* correlated with biallelic hypermethylation of this region was suggested to give a predisposition to colorectal cancer (Nakagawa et al. 2001). However, aberrant methylation is not always associated with abnormal imprinting (Cui et al. 2001). In the pig, three copies of this binding site are found in a conserved larger repeated element situated −1.1 to −2.6 kb upstream of *H19*. A silencer element involved in the regulation of the imprinting of the paternal allele in a methylation-unsensitive manner overlaps this insulator region in the mouse (Drewell et al. 2000: Ferguson-Smith 2000). However, we did not find any evidence of conservation of this region in the pig. Furthermore, a skeletal muscle-specific element involved in the silencing of the *Igf2* maternal allele was identified in the murine *Igf2-H19* intergenic region (Ainscough et al. 2000), but our pig sequence does not completely cover this region. Because of its tissue-specific function, this element is of potential interest for further studies of the *IGF2*-linked QTL in the pig. The 3′ region of *H19* harbors several enhancer elements that affect expression of both *IGF2* and *H19* (Webber et al. 1998). Besides the two well-characterized, endoderm-specific enhancers, other putative enhancer elements that could have a tissue-specific action are present in this region.

This study revealed a large number of conserved elements that might have a functional role in regulating *IGF2* expression. Several approaches can now be used to identify the molecular basis for the paternally expressed QTL affecting muscle development in the pig. From the sequence information presented here, a number of genetic markers can be developed and used to define haplotypes associated with different QTL alleles. This strategy would allow us to narrow the candidate region. A search for polymorphisms associated with the phenotype would then be conducted among the conserved elements in the defined region. However, the results of the present study show that this will still be a challenge owing to the large number of conserved elements potentially influencing *IGF2* function, and we cannot exclude the possibility that some QTL alleles may be due to the combined effect of multiple substitutions. An important topic for future research is also to study *IGF2* expression, in particular in skeletal muscle. The sequence information provided here will facilitate these studies.

Acknowledgments. This work was supported by the Swedish Research Council for Forestry and Agriculture, Swedish Foundation for Strategic Research, Seghers Genetics, and a grant from the Belgian Ministry of Agriculture (D1/2–5795A). The authors thank R. Erlandsson

and B. Amini for sequencing a part of the BAC370 at the Genome Center, Royal Institute of Technology, Stockholm; Niclas Jareborg for his help with Alfresco; Erik Bongcam-Rudloff for bioinformatic assistance; and Göran Andersson for comments on the paper.

References

Ainscough JFX, John RM, Barton SC, Surani MA (2000) A skeletal muscle-specific mouse *Igf2* repressor lies 40 kb downstream of the gene. Development 127, 3923–3930

Amarger V, Gauguier D, Yerle M, Apiou F, Pinton P, et al. (1998) Analysis of distribution in the human, pig, and rat genomes points toward a general subtelomeric origin of minisatellite structures. Genomics 52, 62–71

Andersson L, Archibald A, Ashburner M, Audun S, Barendse W et al. (1996) Comparative genome organization of vertebrates. The First International Workshop on Comparative Genome Organization. Mamm Genome 7, 717–734

Bell AC, Felsenfeld G (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. Nature 405, 482–485

Bennett ST, Lucassen AM, Goug, SC, Powell EE, Undlien DE et al. (1995) Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. Nat Genet 9, 284–292

Brown KW, Villar AJ, Bickmore W, Clayton-Smith J, Catchpoole D et al. (1996) Imprinting mutation in the Beckwith-Wiedemann syndrome leads to biallelic *IGF2* expression through an *H19*-independent pathway. Hum Mol Genet 5, 2027–2032

Constancia M, Dean W, Lopes S, Moore T, Kelsey G et al. (2000) Deletion of a silencer element in *Igf2* results in loss of imprinting independent *H19*. Nat Genet 26, 203–206

Cui H, Niemitz EL, Ravenel JD, Onyango P, Brandenburg SA et al. (2001) Loss of imprinting of insulin-like growth factor-II in Wilms' tumor commonly involves altered methylation but not mutations of CTCF or its binding site. Cancer Res 61, 4947–4950

de Koning DJ, Rattink AP, Harlizius B, van Arendonk JA, Brascamp EW et al. (2000) Genome-wide scan for body composition in pigs reveals important role of imprinting. Proc Natl Acad Sci USA 97, 7947–7950

Drewell RA, Brenton JD, Ainscough JFX, Barton SC, Hilton KJ et al. (2000) Deletion of a silencer element disrupts *H19* imprinting independently of a DNA methylation epigenetic switch. Development 127, 3419–3428

Dubchak I, Brudn M, Loots GG, Mayor C, Pachter L et al. (2000) Active conservation of noncoding sequences revealed by 3-way species comparisons. Genome Res 10, 130–136

Eden S, Constancia M, Hashimshony T, Dean W, Goldstein B et al. (2001) An upstream repressor element plays a role in *Igf2* imprinting. EMBO J 20, 3518–3525

Engemann S, Strödicke M, Paulsen M, Franck O, Reinhardt R et al. (2000) Sequence and functional comparison in the Beckwith-Wiedemann region: implications for a novel imprinting center and extended imprinting. Hum Mol Genet 9, 2691–2706

Ewing B, Hillier L, Wendl M, Green P (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8, 175–185

Ferguson-Smith AC (2000) Genetic imprinting: silencing elements have their say. Curr Biol 10, R872–R875

Florini JR, Ewton DZ, Me Wade FJ (1995) IGFs, muscle growth, and myogenesis. Diabetes Rev 3, 73–92

Frevel MA, Sowerby SJ, Petersen GB, Reeve AE (1999) Methylation sequencing analysis refines the region of *H19* epimutation in Wilms tumor. J Biol Chem 274, 29331–29340

Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196, 261–282

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8, 195–202

Holmgren C, Kanduri C, Dell G, Ward A, Mukhopadhya R et al. (2001) CpG methylation regulates the *Igf2*/*H19* insulator. Curr Biol 11, 1128–1130

Holthuizen P, van der Lee FM, Ikejiri K, Yamamoto M, Sussenbach JS (1990) Identification and initial characterization of a fourth leader exon and promoter of the human IGF-II gene. Biochim Biophys Acta 1087, 341–343

Ishihara K, Kato R, Furuumi H, Zubair M, Sasaki H (1998) Sequence of a 42-kb mouse region containing the imprinted H19 locus: identification of a novel muscle-specific transcription unit showing biallelic expression. Mamm Genome 9, 775–777

Ishihara K, Hatano N, Furuumi H, Kato R, Iwaki T et al. (2000) Comparative genomic sequencing identifies novel tissue-specific enhancers and sequence elements for methylation-sensitive factors implicated in *Igf2*/*H19* imprinting. Genome Res 10, 664–671

Jareborg N, Durbin R (2000) Alfresco—a workbench for comparative genomic sequence analysis. Genome Res 10, 1148–1157

Jareborg N, Birney E, Durbin R (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res 9, 815–824

Jeon JT, Carlborg Ö, Törnsten A, Giuffra E, Amarger V et al. (1999) A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the *IGF2* locus. Nat Genet 21, 157–158

Kaffer CR, Srivastava M, Park KY, Ives E, Hsieh S et al. (2000) Transcriptional insulator at the imprinted *H19*/*Igf2* locus. Genes Dev 14, 1908–1919

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC et al. (2001) Initial sequencing and analysis of the human genome. Nature 409, 860–921

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM et al. (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16, 1046–1047

Mineo R, Fichera E, Liang S-J, Fujita-Yamaguchi Y (2000) Promoter usage for insulin-like growth factor-II in cancerous and benign human breast, prostata, and bladder tissues, and confirmation of a 10th exon. Biochem Biophys Res Commun 268, 886–892

Murrell A, Heeson S, Bowden L, Constancia M, Dean W et al. (2001) An intragenic methylated region in the imprinted Igf2 gene augments transcription. EMBO Rep 2, 1101–1106

Nakagawa H, Chadwick RB, Peltomaki P, Plass C, Nakamura Y et al. (2001) Loss of imprinting of the insulin-like growth factor II gene occurs by biallelic methylation in a core region of *H19*-associated CTCF-binding sites in colorectal cancer. Proc Natl Acad Sci USA 98, 591–596

Nezer C, Moreau L, Brouwers B, Coppieters W, Detilleux J et al. (1999) An imprinted QTL with major effect on muscle mass and fat deposition maps to *IGF2* locus in pigs. Nat Genet 21, 155–156

Ohlsen SM, Lugenbeel KA, Wong EA (1994) Characterization of the linked ovine insulin and insulin-like growth factor-II genes. DNA Cell Biol 13, 377–388

Ohlsson R, Renkawitz R, Lobanenkov V (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. Trends Genet 17, 520–527

Okamura K, Hagiwara-Takeuchi Y, Li T, Vu TH, Hirai M et al. (2000) Comparative genome analysis of the mouse imprinted gene *Impact* and its nonimprinted human homolog *IMPACT*: toward the structural basis for species-specific imprinting. Genome Res 10, 1878–1889

Onyango P, Miller W, Lehoczky J, Leung CT, Birren B et al. (2000) Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain. Genome Res 10, 1697–1710

Paquette J, Giannoukakis N, Polychronakos C, Vafiadis P, Deal C (1998) The *INS* 5′ variable number of tandem repeats is associated with *IGF2* expression in humans. J Biol Chem 273, 14158–14164

Paulsen M, El-Maarri O, Engemann S, Strödicke M, Franck O et al. (2000) Sequence conservation and variability of imprinting in the Beckwith-Wiedemann syndrome gene cluster in human and mouse. Hum Mol Genet 9, 1829–1841

Perier RC, Praz V, Junier T, Bonnard C, Bucher P (2000) The eukaryotic promoter database (EPD). Nucleic Acids Res 28, 302–303

Pugliese A, Zeller M, Fernandez Jr A, Zalcberg LJ, Bartlett RJ et al. (1997) The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. Nat Genet 15, 293–297

Reed MR, Huang CF, Riggs AD, Mann JR (2001) A complex duplication created by gene targeting at the imprinted *H19* locus results in two classes of methylation and correlated *Igf2* expression phenotypes. Genomics 74, 186–196

Reik W, Brown KW, Schneid H, Le Bouc Y, Bickmore W et al. (1995) Imprinting mutations in the Beckwith-Wiedemann syndrome suggested by altered imprinting pattern in the IGF2-H19 domain. Hum Mol Genet 4, 2379–2385

Rietveld LE, Holthuizen PE, Sussenbach JS (1997) Identification of a key regulatory element for the basal activity of the human insulin-like growth factor II gene promoter P3. Biochem J 327, 689–697

Shibata H, Yoda Y, Kato R, Ueda T, Kamiya M et al. (1998) A methylation imprint mark in the mouse imprinted gene *Grf1*/*Cdc25Mm* locus shares a common feature with the *U2afbp-rs* gene: an association with a short tandem repeat and a hypermethylated region. Genomics 49, 30–37

Thorvaldsen JL, Duran KL, Bartolomei MS (1998) Deletion of the *H19* differentially methylated domain results in loss of imprinted expression of *H19* and *Igf2*. Genes Dev 12, 3693–3702

Tremblay KD, Saam JR, Ingram RS, Tilghman SM, Bartolomei MS (1995) A paternal-specific methylation imprint marks the alleles of the mouse *H19* gene. Nat Genet 9, 407–413

Vafiadis P, Bennett ST, Todd JA, Grabs R, Polychronakos C (1998) Divergence between genetic determinants of *IGF2* transcription levels in leukocytes and of IDDM2-encoded susceptibility to type 1 diabetes. J Clin Endocrinol Metab 83, 2933–2939

Webber AL, Ingram RS, Levorse JM, Tilghman SM (1998) Location of enhancers is essential for the imprinting of *H19* and *Igf2* genes. Nature 391, 711–715

Weber M, Milligan L, Delalbre L, Antoine E, Brunel C et al. (2001) Extensive tissue-specific variation of allelic methylation in the *Igf2* gene during mouse fetal development: relation to expression and imprinting. Mech Dev 101, 133–141

Wylie AA, Murphy SK, Orton TC, Jirtle RL (2000) Novel imprinted *DLK1*/*GTL2* domain on human chromosome 14 contains motifs that mimic those implicated in *IGF2*/*H19* regulation. Genome Res 10, 1711–1718

Zemel S, Bartolomei MS, Tilghman SM (1992) Physical linkage of two mammalian imprinted genes, H19 and insulin-like growth factor 2. Nat Genet 2, 61–65