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 paternally expressed 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 (Mino 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 AJ044827–AJ044828 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. DNA was purified by using the QIAGEN plasmid midi kit (QIAGEN, Germany). Two microgram of BAC DNA was digested with 10 units of NotI 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. NotI 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 Sau3AI (New England Biolabs) for 10 min at 57°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 BamHI restricted pUC18 (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 automated 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).

NotI restriction fragments containing the H19 region were sequenced using the EZ-TN™ Transposon Insertion System (Epiconcept 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. Repeatitive elements were localized and identified by RepeatMasker (A.F.A. Smit & P. Green, unpublished: http://ftp.genome.washing- ton.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 (Dobchak et al. 2000; Mayor et al. 2000) at http://www.sgd.iub.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 purified 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 Bio- tech).

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 IGF2E2XSF

5′-GTAGCGCGCTGAGTGGA′-3′ in exon 1 and reverse primer IGF2E2XMR 5′-CTCCCTGCTACGTAGAAG′-3′ in the junction between exon 7 and 8; P2, forward primer IGF2E2XFR 5′-CTCCTGCTGCTACGTAGAAG′-3′ in exon 7 and reverse primer IGF2E2XTR 5′-CTCCTGCTGCTACGTAGAAG′-3′ in exon 7; P3, forward primer IGF2E2XSF 5′-CTCCTGCTGCTACGTAGAAG′-3′ in exon 7 and reverse primer IGF2E2XTR 5′-CTCCTGCTGCTACGTAGAAG′-3′ in exon 7; and P4 forward primer IGF2E2XFR 5′-GGGCGCTCAGTACGACGGAG′-3′ in exon 7. The obtained PCR products were isolated from agarose gels and sequenced directly on a MegaBACE™ 1000 sequencing instrument, 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 NotI, 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 NotI, 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 NotI restriction fragments (Fig. 1A). The location of the INS and H19 genes in BAC 253 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 NotI fragment containing H19 as well as the 1.2-, 1.5-, 6.1-, and 24-kb NotI 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 NotI 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).