Sequence and Comparative Analysis of the Mouse 1-Megabase Region Orthologous to the Human 11p15 Imprinted Domain

Patrick Onyango,[1,2] Webb Miller,[3] Jessica Lehoczky,[4] Cheuk T. Leung,[1,5] Bruce Birren,[4] Sarah Wheelan,[5,7] Ken Dewar,[4] and Andrew P. Feinberg[1,2,5,6,8]

[1]Institute of Genetic Medicine and [2]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA; [3]Department of Computer Science and Engineering, Pennsylvania State University, University Park, Pennsylvania 16802, USA; [4]Whitehead Institute/MIT Center for Genome Research, Cambridge, Massachusetts 02141, USA; [5]Department of Molecular Biology and Genetics and [6]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA; [7]Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA

A major barrier to conceptual advances in understanding the mechanisms and regulation of imprinting of a genomic region is our relatively poor understanding of the overall organization of genes and of the potentially important *cis*-acting regulatory sequences that lie in the nonexonic segments that make up 97% of the genome. Interspecies sequence comparison offers an effective approach to identify sequence from conserved functional elements. In this article we describe the successful use of this approach in comparing a ~1-Mb imprinted genomic domain on mouse chromosome 7 to its orthologous region on human 11p15.5. Within the region, we identified 112 exons of known genes as well as a novel gene identified uniquely in the mouse region, termed *Msuit*, that was found to be imprinted. In addition to these coding elements, we identified 33 CpG islands and 49 orthologous nonexonic, nonisland sequences that met our criteria as being conserved, and making up 4.1% of the total sequence. These conserved noncoding sequence elements were generally clustered near imprinted genes and the majority were between *Igf2* and *H19* or within *Kvlqt1*. Finally, the location of CpG islands provided evidence that suggested a two-island rule for imprinted genes. This study provides the first global view of the architecture of an entire imprinted domain and provides candidate sequence elements for subsequent functional analyses.

[The sequence data described in this paper have been submitted to the GenBank data library under accession nos. AF313042 to AF313150.]

Genomic imprinting is an epigenetic modification of the gamete or zygote that leads to preferential expression of a specific parental allele in somatic cells of the offspring. The mechanism of imprinting is unknown but it is thought to involve CpG island methylation (Sapienza et al. 1987; Sutcliffe et al. 1994), antisense transcripts (Wutz et al. 1997), short repeat elements (Szebenyi and Rotwein 1994), and/or *trans*-acting binding proteins that may interact with one or more of these sequences (Bell and Felsenfeld 2000; Hark et al. 2000; Srivastava et al. 2000). One of the most surprising recent discoveries in the study of genomic imprinting is that imprinted genes are grouped in large multigene domains (Lee et al. 1997; Ainscough et al. 1998; Feinberg 1999). In particular, we and others have found that human chromosomal band 11p15 contains at least eight imprinted genes concentrated in an ~1-Mb domain, of which six are expressed from the maternal allele and two are expressed from the paternal allele (Feinberg 1999). The organization of this domain is somewhat complicated in that we have identified two separate subdomains that are imprinted, separated by a region of three genes that appear to escape imprinting (Lee et al. 1998, 1999). The boundaries of the overall 11p15 imprinted domain are known at both centromeric and telomeric ends because of the presence of at least eight nonimprinted genes that extend beyond the imprinted domain, including *NAP2* and *NUP98* on the centromeric side, and *L23MRP* and *CTDS* on the telomeric side (Rachmilewitz et al. 1993; Tsang et al. 1995; Hu et al. 1996, 1997; Zubair et al. 1997). Thus it is likely that both local and regional *cis*-acting elements are involved in the regulation of genomic imprinting. However, almost nothing is known about the identity or location of such regulatory elements, with the notable exception of a region that has been intensively studied upstream of and downstream from the *H19* gene (Thorvaldsen et al. 1998; Bell and Felsenfeld 2000; Hark et al. 2000; Srivastava et al. 2000).

[8]Corresponding author.
E-MAIL afeinberg@jhu.edu; FAX (410) 614-9819.
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.161800.

Understanding the genomic organization of this domain is also critical to the study of the disorder Beckwith-Wiedemann syndrome (BWS), which causes prenatal overgrowth, birth defects, and predisposition to a wide variety of childhood cancers, most commonly Wilms tumor (Feinberg 1999). We have found that BWS can involve altered imprinting of either of the two subdomains within the 11p15 imprinted domain, one including *H19* and *IGF2* and the other including the maternally expressed genes $p57^{KIP2}$, K_VLQT1 and paternally expressed *LIT1*, an antisense orientation transcript within K_VLQT1 (Weksberg et al. 1993; Steenman et al. 1994; Lee et al. 1997,1999).

A powerful approach to identifying functionally important sequences is by aligning of orthologous genomic regions. Evolutionarily conserved genes often have similar structure and function and important regulatory elements may be conserved even between distantly related organisms whose genomes may have little or no similarity overall (Elgar 1996; Hardison et al. 1997). When comparing the mouse and human genomes, the average size of syntenic segments is estimated to be 7.1–15 Mb (O'Brien et al. 1999). The mouse ortholog of the entire human 11p15 imprinted domain is contained in a single syntenic block on mouse chromosome 7 (Blake et al. 2000).

We have taken a comparative genomics approach to identify novel genes and potential regulatory elements within the 11p15 imprinted domain. We identified 87 overlapping BACs spanning ~1 Mb of mouse chromosome 7 that includes the entire imprinted domain and flanking nonimprinted genes. Draft sequence was obtained from a minimal tiling path of five BACs and this sequence could be ordered by comparison with the publicly available human sequence. Deeper coverage mouse sequence was obtained for the region corresponding to an estimated 250-kb gap remaining within the Human Genome Project sequence, so that ~95% of the sequence across the entire domain could be ordered and analyzed. This work represents the largest ordered and oriented sequence comparison between mouse and human to date and the first comparative sequence analysis of an entire imprinted domain.

RESULTS

Construction of a BAC Contig across the Entire Orthologous Mouse Imprinted Domain

Forty-five overgo probes (Table 1) were pooled and used for hybridization screening (Ross et al. 1999) of high-density BAC clone filters of the 11.2 × genome equivalent RPCI-23 female mouse C57 BL/6J library. Single-colony isolates were recovered from all addresses identified in the primary screen, then rearrayed and replicated onto sets of filters. In a second round of screening, individual copies of the arrayed clones were tested with individual overgo probes to establish the clone–marker relationships (Table 2). The BAC contig was estimated to span 1.2 Mb and includes the entire orthologous mouse imprinted domain, flanked by the *NAP2* gene at the centromeric end and the *L23MRP* gene at the telomeric end (determined by subsequent sequence analysis). Overall, BAC clones contained an average of 7.8 probes per clone, and each probe tested positive against an average of 9.3 redundantly identified clones (data not shown). Marker density across the region, recovered clone depth, and the marker–clone relationships indicated that the entire region had been captured in an overlapping set of clones. A minimal path of clones for genomic sequencing was selected using combined knowledge of marker content and restriction enzyme digestion fingerprint analysis (Marra et al. 1997). A restriction enzyme map was constructed for *Hin*dIII (data not shown), which allowed a more refined interpretation of clone order and overlaps. From this a set of five overlapping clones collectively spanning the region were selected for genomic sequencing (Fig. 1).

For four of the BAC clones (RP23–209o22, RP23–366m16, RP23–101n20, and RP23–124b2) draft quality sequencing and assembly were performed to 5× depth sequence coverage based on NotI/pulsed field gel estimates of clone size (data not shown). Draft assemblies at this level of coverage contain the vast majority of the clone sequence (>90%), with the remaining sequence gaps being small (<1 kb; Bouck et al. 1998). Although the outcome of a draft assembly is a series of sequence contigs of unknown order and orientation, sequence alignments to references (other genomic sequences, genes, etc.) can be used to determine the correct positioning of the draft sequence contigs. Deeper coverage sequencing (10–12×) and assembly, especially using paired forward/reverse reads from sequencing subclones, further reduces the gap number and can generate self-ordered contig sets (Bouck et al. 1998). For the RP23–92l23 clone, deeper coverage sequencing and finishing was performed. This corresponds to the portion of the human genome that has not been sequenced.

Global Comparison of the Mouse and Human Orthologous Imprinted Domain

We used the program `PipMaker` (Schwartz et al. 2000) to perform a detailed comparison between mouse and human genomic sequences. This analysis is shown graphically in the percent identity plot (PIP) in Figure 2. The reference sequence is mouse and it is oriented from centromere to telomere (the human domain is oriented oppositely). We have used both geometric figures and coloring to annotate the PIP. Structural features in the mouse, including exons, repeats, and CpG

Table 1. STSs Used to Identify BACs within the Mouse Imprinted Domain Orthologous to Human 11p15

Overgo probe	Oligo A sequence	Oligo B sequence	GenBank accession
nap114.1	GTTCTGGTTTACCATCTTCAGA	AGCATGTCCACATTTCTGAAGA	AC001228
nap114.3	TACAGTGTTTTCAATTTGTACA	TATTCTAACTATTCTGTACAAA	AC001228
pac1.2	ATATCTAGGCCTCCATGCCTTC	CCTCGTGTGTTTGTGAAGGCAT	AF093714
pac1.1	CCTGGTCATCCCTCTGGAATCC	ATAGAGGGATGGGAGGATTCCA	AF093714
KvL-18.1	TCCTTGGGCTTGGGCACCACGC	TTGATTTCTGTGAGGCGTGGTG	U70068
KvL-17.1	GCACACCCACCTGGTTCTCACC	CAGGCCTTCAAGGGGGTGAGAA	U70068
KvL-16.1	CATCACAGACATGCTCCACCAG	TGCATGGACAGCAGCTGGTGGA	U70068
KvL-15.1	GAGCAAAGACCGTGGCAGTAAC	CGGGCACCGATGGTGTTACTGC	U70068
KvL-14.1	CAGTCCATTGGGAAGCCATCTT	GATGGGGATGAACAAAGATGGC	U70068
KvL-13.1	CGTGCGAGATGTCATCGAGCAG	TGGCCCTGGGAGTACTGCTCGA	U70068
KvL-12.2	GGTCATCAGGCGCATGCAGTAC	TTCTTGGCTACAAAGTACTGCA	U70068
KvL-11.2	TGGACCTGGAAGGGGAGACACT	TGATGGGGGTCAGCAGTGTCTC	U70068
KvL-11.1	TGCTGGAATTAAGCACACCCCA	TTGTTCTCAAGAAATGGGGTGT	U70068
KvL-10.2	TCCCCCAGAGGATAGGAGGCCA	ATGGAGAAATGGTCTGGCCTCC	U70068
KvL-10.1	GTAAAGAAGAAGAAGTTCAAGC	ATTATCCTTATCCAGCTTGAAC	U70068
KvL-9.1	AAGCCTGCCGGAGTCACACGC	GCTGGGGGACAGAAGCGTGTGA	U70068
KvL-8.1	GCAGAAGCACTTCAACCGGCAG	GCTGCAGCTGGGATCTGCCGGT	U70068
KvL-7.1	ACCTCAGACGTGGGTTGGGAAG	CAGGAGGCGATGGTCTTCCCAA	U70068
KVLQT1.1	GAGTTTGGCAGCTACGCAGATG	CCCCCACCACAGAGCATCTGCG	U70068
KvL-6.1	GGCCGCATCGAGTTTGGCAGCT	CAGAGCATCTGCGTAGCTGCCA	U70068
KvL-5.1	CTTCAGATCCTGCGGATGCTGC	CTGGCGATCGACATGCAGCATC	U70068
KvL-4.1	TGCGTGGGTTCCAAAGGACAAG	TGATGTGGCGAACACTTGTCCT	U70068
KvLQT.1	GTACGTGGGCATCTGGGGCCGG	CGGGCAAAACGTAGCCGGCCCC	U70068
KvL-3.1	AAGTACGTGGGCATCTGGGGCC	GGCAAAACGTAGCCGGCCCCAG	U70068
KvL-2.2	GTATGCCGCTCTGGCCACCGGG	ATCCAGAAGAGGGTCCCGGTGG	U70068
KvL-2.1	TCCTCATTGTTCTGGTCTGCCT	GGACACTGAAGATGAGGCAGAC	U70068
KVLQT1.3	TGGCGCGCACCCACATCCAGGG	AGTTGTAGACTCGGCCCTGGAT	U70068
KvL-1.2	GTGAGCCTTGACCCGCGGGTCT	CGCACTGTAGATGGAGACCCGC	U70068
tssc4.1	CACCCTGAGCGTTGGACCAAAT	ATCCTCCAGACTGTATTTGGTC	AA241958
tapa1.2	ATTCTGAGCATGGTGCTGTGCT	GTTCCGGATGCCACAGCACAGC	X59047
pac2.2	GAGTTTTGTCTGGCATTGCTTG	CATCCAAAAACGGCCAAGCAAT	AF093715
tapa1.1	CGCCAAGGATGTGAAGCAGTTC	AGGGCCTGGTCATAGAACTGCT	X59047
pac2.1	GAGGAGCCTTCAGTCTTCCTTG	GAGCCTGTTCTTCTCAAGGAAG	AF093715
tssc6.2	GTGGCCTTCTGGAGATTCTACA	CACCTGGGTGGGGTTGTAGAAT	AA200225
tssc6.3	CTATGTGGGGATCAGCCTAGCG	AGGCTCAGGAGCCCCGCTAGGC	AA200225
mash2.1	CAGCTTCTTGTTGGCGCCGCCG	CAGCACGTGCCGCACGGCGGCG	U77628
mash2.2	GGAAGCCCAAGTTTACCAGCTT	AGCGCAACCGCGTAAAGCTGGT	U77628
pac3.1	ATCAGGCCAGTTACTTCTGGAC	AGCAGCACGTCTGTGTCCAGAA	AF093708
th.3	GCAGAGTCTCATCGAGGATGCC	TCCCGCTCCTTGCGGGCATCCT	M69200
th.1	CAAGGAAAGTGTCAGAGTTGGA	GGTGGTGACACTTATCCAACTC	M69200
th.2	GCCAGTCTGGCCTTCCGTGTGT	CTGTGTGCACTGAAACACACGG	M69200
igf2.1	GAAGCACCAACATCGACTTCCC	TGGGGATCCCAGTGGGGAAGTC	M14951
igf2.2	TGTCATATTGGAAGAACTTGCC	CCAGATACCCCGTGGGCAAGTT	M14951
h19.1	CAAGTCCACTGTGGGCCCTTTC	CAGGGACTGGTGCGGAAAGGGC	X07201
h19.2	GGATTCAAAGGCCCAGACATCA	TGGTCCTACCCAGCTGATGTCT	X07201

islands, are shown above the top line. Evolutionarily conserved elements were identified by PIP analysis. Segments between consecutive gaps in a `PipMaker` alignment and having ≥50% nucleotide identity are displayed in Figure 2 as short horizontal lines. Exons are considered to be conserved (Fig. 2, green) if they are completely spanned by `PipMaker` alignments. To determine conserved CpG islands (Fig. 2, orange), we used `BLAST2` to identify segments having ≥50% nucleotide identity. Sequences that do not appear to be an exon, a CpG island, or part of an interspersed repeat identified by `RepeatMasker` are considered to be conserved (Fig. 2, blue) if they align without a gap for ≥100 bp in the `PipMaker` alignment with ≥70% nucleotide identity. This criterion, although arbitrary, was used by Loots et al. (2000). Other authors (e.g., Lund et al. 2000; Mallon et al. 2000) have adopted different thresholds. In our analysis, there were eight instances in which a cluster of nearby segments, each meeting this criterion, was merged and considered to be a single conserved region. Novel exons identified by `Genscan`, `GRAIL`, or EST identity and confirmed by RT-PCR or Northern blot analysis are also depicted (Fig. 2, red), whether or not they are conserved.

In all our comparisons, it should be noted that ~250 kb of the human imprinted domain has not yet been completed (Figs. 1, 2) and that the mouse reference sequence was constructed from draft sequences

Table 2. BACs and STSs Marker Content of the Mouse Imprinted Domain Orthologous to Human 11p15

Overgo probe	BAC clones
nap114.1	175c14, 369k9, 257o11, 477n6, 6i17
nap114.3	175c14, 369k9, 257o11, 6i17, 344f10, 346117
pac1.2	344f10, 346117, **124b2**, 257o11, 477n6
pac1.1	175c14, 369k9, 257o11, 344f10, 346117, **124b2**
KvL-18.1	**124b2**, 111a21, 35i20, 36a17, 111e23
KvL-17.1	400c11, 111e23, **124b2**, 111a21, 35i20, 36a17
KvL-16.1	400c11, 111e23, **124b2**, 111a21, 35i20, 36a17
KvL-15.1	**124b2**, 111a21, 35i20, 36a17
KvL-14.1	400c11, 161j24, 118h24, 24116, 469p12, **124b2**, 111a21, 35i20, 36a17, 111e23
KvL-13.1	400c11, 161j24, 24116, 469p12, **124b2**, 35i20, 36a17
KvL-12.2	161j24, 118h24, **124b2**, 111a21
KvL-11.2	**101n20**, 161j24, 118h24, 207g7, 35i20
KvL-11.1	**101n20**, 161j24, 118h24, 207g7, 35i20, 111a21, 111e23
KvL-10.2	**101n20**, 207g7, 424120, 296b22, 101a11, 421g3, 437o5, 172e1, 200g3, 95m15, 374o15
KvL-10.1	**101n20**, 207g7, 424120, 296b22, 101a11, 421g3, 437o5, 172e1, 200g3, 95m15, 374o15, 417b4
KvL-9.1	**101n20**, 207g7, 424120, 296b22, 101a11, 421g3, 437o5, 172e1, 200g3, 95m15, 417b4, 388121,
KvL-8.1	**101n20**, 207g7, 424120, 296b22, 101a11, 421g3, 172e1, 200g3, 95m15
KvL-7.1	**101n20**, 207g7, 424120, 296b22, 101a11, 421g3, 437o5, 172e1, 200g3, 95m15, 417b4, 388121 281p7, 373d8, 374o15
KVLQT1.1	**101n20**, 207g7, 424120, 296b22, 101a11, 421g3, 437o5, 172e1, 200g3, 95m15, 417b4, 388121 373d8, 374o15
KvL-6.1	**101n20**, 207g7, 424120, 296b22, 101a11, 421g3, 437o5, 172e1, 200g3, 95m15, 417b4, 388121, 281p7, 373d8, 374o15
KvL-5.1	**101n20**, 207g7, 296b22, 101a11, 172e1, 200g3, 95m15
KvL-4.1	**101n20**, 207g7, 296b22, 101a11, 172e1, 200g3, 95m15, 437o5
KvLQT.1	**101n20**, 207g7, 296b22, 101a11, 200g3, 437o5, 417b4, 424120, 421g3, 374o15
KvL-3.1	**101n20**, 207g7, 424120, 296b22, 101a11, 421g3, 437o5, 172e1, 200g3, 95m15, 417b4, 388121 281p7, 373d8, 374o15
KvL-2.2	**101n20**, 207g7, 424120, 296b22, 101a11, 437o5, 172e1, 200g3, 95m15, 281p7, 374o15, 366n2
KvL-2.1	**101n20**, 207g7, 424120, 296b22, 101a11, 421g3, 437o5, 172e1, 200g3, 95m15, 417b4, 388121 281p7, 374o15, 366n2, 17n3, 299i6, 51j21, 50n22
KVLQT1.3	**101n20**, 207g7, 424120, 296b22, 101a11, 421g3, 172e1, 200g3, 95m15, 417b4, 388121, 373d8
KvL-1.2	**101n20**, 424120, 296b22, 101a11, 421g3, 437o5, 172e1, 200g3, 95m15, 417b4, 388121 281p7, 17n3, 299i6, 51j21, 50n22, 405o8, 319p9
tssc4.1	424120, 296b22, 101a11, 437o5, 172e1, 200g3, 95m15, **366m16**, 319p9, 421g3, 417b4, 405o8
tapa1.2	424120, 296b22, 101a11, 437o5, 172e1, 200g3, **366m16**, 319p9, 417b4, 405o8, 421g3, 95m15
pac2.2	424120, 296b22, 101a11, 437o5, 172e1, 200g3, **366m16**, 319p9, 417b4, 405o8, 421g3
tapa1.1	424120, 296b22, 101a11, 437o5, 172e1, 200g3, **366m16**, 319p9, 417b4, 405o8, 454a2, 421g3, 95m15, 388121
pac2.1	424120, 296b22, 101a11, 437o5, 172e1, 200g3, **366m16**, 319p9, 405o8, 454a2, 421g3, 417b4
tssc6.2	424120, 296b22, 101a11, **366m16**, 405o8, 454a2, 71o21, 365f7
tssc6.3	119e20, 365f7, **366m16**, 71o21, 405o8, 454a2, 473n24, 319p9
mash2.1	119e20, 365f7, **366m16**, 473n24, 71o21, 405o8, 319p9, 295n5, 334j1, 350e13
mash2.2	119e20, 365f7, **366m16**, 473n24, 71o21, 405o8, 319p9, 295n5, 334j1, 350e13, 295e16, 113b24
pac3.1	119e20, 365f7, **366m16**, 71o21, 405o8, 319p9, 295n5, 334j1, 350e13, 295e16, 113b14
th.3	119e20, 365f7, 473n24, 71o21, 295n5, 334j1, 350e13, 295e16, **92123**, 113b14
th.1	119e20, 365f7, 473n24, 71o21, 295n5, 334j1, 350e13, 295e16, **92123**, 113b14
th.2	119e20, 365f7, 295n5, 334j1, 350e13, 295e16, **92123**, 113b24
igf2.1	299i6, 50n22, **209o22**, 51j21, 473m23, 17n3
igf2.2	299i6, 50n22, **209o22**, 51j21, 473m23, 17n3
h19.1	73d4, 299i6, 50n22, **209o22**
h19.2	73d4, 299i6, 50n22, **209o22**

All BACs were recovered from the RPCI-23 female mouse C57 BL/6J library. BAC clones selected for sequencing are indicated in bold.

for four of the five mouse clones spanning this region. As the sequencing efforts of both species give rise to fully accurate and complete data, many of our observations will become more refined, especially with regard to precise physical distances between features. Nonetheless, the accuracy and comprehensiveness of the existing sequences have provided an important resource for the identification of new candidate genes and regulatory sequences.

A global comparison of the human and mouse sequence revealed the presence of 16 known genes: *Rl23mrp*, *H19*, *Igf2*, *Ins*, *Th*, *Mash2*, *Tssc6*, *Tapa1*, *Tssc4*, *Trpc5l*, *Kvlqt1*, *Lit1*, *p57^{KIP2}*, *Tssc5*, *Tssc3*, and *Nap2* (Fig. 2; Table 3). The genomic organization of these genes is,

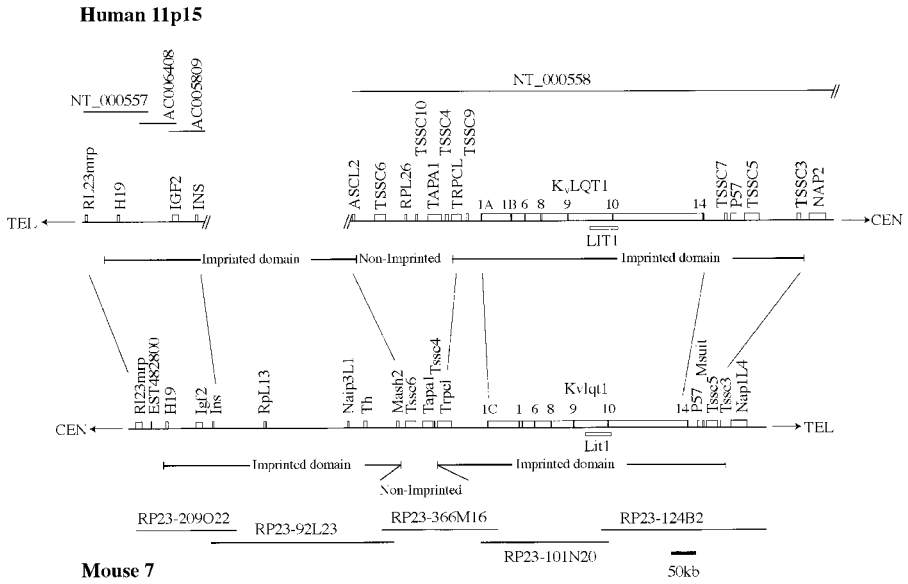

Figure 1 Overview of the imprinted gene domain on human 11p15 and mouse chromosome 7. The organization of the human and mouse domains is depicted, including the locations of the two imprinted subdomains within the region, the locations of the mouse BAC clones that were sequenced and analyzed, and the sources of human sequence for comparison.

for the most part, comparable between the two species. The total number of exons of known genes is 119 in the human and 112 in the mouse. Of these exons, 110 were conserved. However, some exons were present in the imprinted domain of one species and not the other. For example, mouse Igf2 consists of eight exons whereas the human gene contains one additional exon, and the single-exon encoded ribosomal proteins L26 and L13 were only present in the human and mouse, respectively (Table 4).

To assess the level of background sequence similarity between human and mouse, we determined the fraction of noncoding, nonrepetitive mouse sequence that can be aligned to the human sequence using the protocol of Endrizzi et al. (1999) and Zhang et al. (1999). The imprinted domain between *Trpc51* and *Tssc3* and the nonimprinted domain from *Tssc6* to *Tssc4* showed a similar fraction of aligned positions (19.6% and 18.8%, respectively). In contrast, the imprinted domain between *H19* and *Mash2* showed approximately twice the degree of alignment (35.8%), which indicates either that it contains a larger fraction of functional DNA or that neutral mutations are being fixed at a lower rate. Although variable, these numbers are in the range (6.4%–78.1%) observed using the same technique in nine other genomic regions (see Endrizzi et al. 1999, Table 3).

The GC content of the entire domain was less in mouse (47.8%) than that in the human (54.7%). Thirty-three CpG islands were conserved between the two species, and there were approximately twice as many CpG islands in human as there were in the mouse (119 vs. 65). There were an additional 49 conserved nonisland intergenic or intragenic sequences (Tables 3 and 5). Some of these conserved sequences may represent previously unrecognized exons of genes, based on their location, for example, conserved sequences at 67609–67753 (145 nt, 79%) and 82671–82887 (217 nt, 86%) located between *H19* and *Igf2* (Fig. 2; Table 5). However, 39 of the 49 conserved sequences are unlikely to be part of the coding sequence of genes because they did not have high coding potentials following predictions with `Genscan` or `GRAIL`. The total sequence represented by all of the nonexonic conserved elements combined was ~27 kb or 4.1% of the total genomic sequence analyzed.

`RepeatMasker` identified a significantly greater number of repetitive elements in the human sequence than in the mouse (Table 3). Most of this difference was because of the nearly twofold higher fraction of long interspersed nuclear elements in the human sequence (Table 3). In addition, there were threefold more DNA transposon fossils belonging to the medium reiterated repeats (MER) and mariner families. Finally, a VNTR-like repeat, $[TGTGAATA(C/T)GCTC(A/G)G]_N$ was located between human *NAP2* and *TSSC3* (i.e., at the centromeric end of the imprinted domain) but was not conserved in the mouse. In addition, there were 17.9 tandem copies of a 27-bp motif at mouse positions 126926–127409, upstream of *Igf2*. A very prominent

Figure 2 Comparison of mouse and human sequence of the imprinted gene domain. Percent Identity Plot (PIP) showing order and alignment of the entire imprinted domain on mouse chromosome 7 as compared with the orthologous region on human 11p15.5. The mouse sequence is the reference sequence and the short horizontal lines correspond to segments of sequence conservation. Conserved features are color coded as follows: Conserved exons, green; conserved CpG islands, orange; conserved nonexonic sequences not obviously within one of these categories, blue (see text for criteria). Novel genes are shown in red. Where two features apply, two colors are used. The white area is the portion of the human genome sequence that is incomplete but for which mouse sequence was obtained. Vertical black lines show the position of the remaining gaps within the mouse draft assembly sequences. The sequences within these gaps are expected to be <10% (Bouck et al. 1998) of the overall region. Where there is disagreement about nomenclature, exons are numbered arbitrarily (e.g., *Igf2*).

Table 3. Global Sequence Comparison of Human 11p15 and the Orthologous Mouse Domain

Feature	Mouse	Human
Total sequence	915699 bp	900050 bp
Known genes	16	16
Exons	112	119
Content	28710 bp	27890 bp
Novel transcripts	6	7
GC content	47.76%	54.70%
CpG islands	65	119
Conserved CpG islands	33	33
Conserved content	17600 bp	17600 bp
Conserved nonexonic, nonisland sequences	49	49
Conserved content	10007 bp	10007 bp
Total interspersed repeats	219 kb (23.96%)	264 kb (29.38%)
SINEs	367 (5.24%)	215 (5.25%)
LINEs	114 (8.08%)	192 (16.37%)
LTR elements	170 (7.68%)	104 (6.37%)
MERs, Mariners	20 (0.41%)	53 (1.4%)
Total simple repeats	294 (2.07%)	153 (1.92%)
Small RNA repeats	3 (0.02%)	2 (0.02%)
Low complexity repeats	69 (0.48%)	63 (0.54%)

The percent of conserved sequence (4.2%) is calculated by dividing the sum of the conserved CpG island content and non-exonic non-island content, by the aligned mouse sequence excluding the human sequence not yet completed.

feature was found at 144–350kb. The region, when masked for interspersed repeats and low-complexity regions using `RepeatMasker`, shows a striking pattern of alignments between different parts of the region, while having no matches with other genomic sequences in the NCBI databases. Overall, the human imprinted domain has a greater physical size than the orthologous region in mouse (900 kb plus a gap estimated at 250 kb in the human vs. 916 kb in the mouse). This size difference may be partially explained by the increased presence of retroposons. The completion of the human and mouse sequences, in addition to permitting even more refined analyses of the genomic features associated with imprinting, will also be informative in showing how the regions of the two species have been evolving since the time of the mammalian radiation.

Msuit, a Novel Imprinted Transcript Present in Mouse but not Human

Although our primary focus was the identification of conserved sequences, we also observed that several predicted transcripts were present in one species but not the other. For example, by searching dbEST we found that nucleotides 862814 to 864030, approximately 1.9

Table 4. Novel Genes in the Imprinted Domain

Location	Name	GenBank accession	Species
36798–37130	*Rhit1*	AF313043	Mouse
30111–30452	*RHIT1*	AF313096	Human*
76583–76864	*Ihit*	AF313044	Mouse
216446–217129	Ribosomal protein L13	NM_016738	Mouse
267824–268473	Ribosomal protein L26	NM_016093	Human
308887–309057	*TSSC11*	AF313097	Human
337171–337476	*Naip3LI*	AF313045	Mouse
395979–396370	*TSSC10*	AF313098	Human
573501–574207	*TSSC9*	AF313099	Human
660496–663300	*Tssc8*	AF313046	Mouse
585546–588434	*Tssc8*	AF313100	Human
802282–803751	*TSSC7*	AF313101	Human
862814–864030	*Msuit1*	AF313042	Mouse

*Expression of this gene not yet confirmed in human.

Table 5. Conserved Non-exonic Non-CpG Island Sequences

Mouse locus	Human locus	% Identity	% GC	Nucleotides
40543–40648	34847–34952	72	60	106
43834–43953	38487–38606	74	60	120
63886–64008	69875–69997	76	48	123
65700–65811	71856–71967	75	54	112
67609–67753	74146–74290	79	49	145
82671–82887	90514–90730	86	61	217
84008–84146	92141–92279	79	55	139
91169–91295	111355–111481	97	47	127
101002–101123	127060–127181	88	45	122
116781–116886	152244–152349	77	59	106
119507–119614	156490–156597	72	56	108
413083–413186	200922–201025	71	58	104
431704–431807	242780–242883	74	56	104
597519–597629	472592–472702	80	51	111
603553–603665	483759–483871	70	39	113
612048–612186	492154–492292	73	57	139
624417–624534	507831–507948	82	73	118
625643–625801	509051–509209	77	46	159
628927–629080	515637–515790	73	60	154
658796–659542	583755–584504	91	47	747
670930–671231	596556–596858	88	48	302
672398–673099	598134–598892	75	54	702
680099–681020	605934–606878	82	49	922
708399–708530	637159–637290	83	59	132
711877–712293	640732–641156	88	56	417
715006–715128	643506–643628	77	45	123
716450–716564	644863–644977	72	36	115
717171–717335	645533–645697	88	46	165
724163–724313	655959–656109	85	49	151
726026–726300	657680–657955	86	55	275
730769–730948	664410–664589	96	60	180
732850–732987	666829–666966	72	43	138
736659–736781	672156–672278	82	36	123
738780–738891	674435–674546	76	55	112
746195–746685	678699–679189	93	53	491
759476–759603	697942–698069	72	39	128
765600–765699	705854–705953	71	47	100
768252–768367	710501–710616	74	42	116
769119–769229	711972–712082	70	68	111
776638–776925	720398–720684	85	51	288
779020–779154	723073–723207	81	55	135
779723–779857	723685–723819	84	55	135
785790–786017	730301–730528	95	33	228
793474–793611	739674–739811	75	59	138
793893–794010	740103–740220	78	52	118
824072–824279	775347–775554	82	44	208
829984–830328	781964–782307	88	58	345
843477–843740	797825–798088	83	41	264
876879–877049	841281–841451	77	61	171

GenBank accession nos. AF313047–AF313095 and AF31302–AF313150. Details are available at http://www.hopkinsmedicine.org/imprinting.

kb upstream of the mouse p57KIP2 gene, matched EST1179335 (accession no. AA717997; Fig. 2, red). RT-PCR and Northern blot analysis of this EST revealed expression in all fetal and adult tissues, but low stringency Southern blots did not show conservation in human (Fig. 4 and data not shown). Given the location of this sequence between *p57^{KIP2}* and *Tssc5*, we thought the transcript might be imprinted despite its lack of conservation. To test this hypothesis, we used a G/C transcribed polymorphism that distinguishes *Mus musculus castaneus* from *Mus musculus musculus*, at nucleotide 247 of the EST (Fig. 4). RT-PCR analysis of fetal and adult tissues revealed monoallelic expression, with preferential expression from the maternal allele in all tissues analyzed, indicating that the gene is imprinted (Fig. 4). Based on this result, we designated the gene *Msuit1*, for mouse-specific ubiquitously imprinted transcript 1.

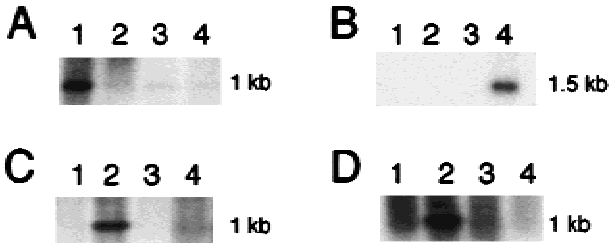

Figure 3 Expression analysis of novel transcripts in the imprinted gene domain. Human and mouse Northern blots were hybridized with expressed sequence tag (EST) probes. (*A*) Mouse Northern blot hybridized with EST670599 (accession no. AA221972): *1*, heart; *2*, brain; *3*, spleen; *4*, lung. (*B*) Human Northern blot hybridized with EST1422939 (GenBank accession no. AI732937): *1*, spleen; *2*, lung; *3*, prostate; *4*, testes. (*C*) Human Northern blot hybridized with *Ihit*, a Genscan-predicted cDNA located between *H19* and *Igf2*: *1*, heart; *2*, brain; *3*, spleen; *4*, kidney. (*D*) Human fetal Northern blot hybridized with *Ihit*, *1*, kidney; *2*, liver; *3*, lung; *4*, brain.

Several Additional Nonconserved Transcripts Unique to the Mouse or Human

Within this region, we identified two transcripts (Fig. 2, red; Table 4) that were unique to the mouse: Ribosomal protein *L13* (GenBank accession no. NM_016738) located 78 kb telomeric to *Ins*; and EST670599 (GenBank accession no. AA221972), located 14 kb centromeric to *Th* in the mouse. We also identified five transcripts that were unique to the human (Table 4): Ribosomal protein L26 (accession no. NM_016093) located 15 kb centromeric to *TSSC6*; EST7905961 (GenBank accession no. AW812967) located upstream of *Kvlqt1*; EST1100208 (GenBank accession no. AA584837) located 42 kb telomeric to K_vLQT1; EST42127 (GenBank accession no. AA337385) located 3 kb telomeric to *TAPA1*; and EST1422939 (GenBank accession no. AI732937) located 15 kb telomeric of $p57^{KIP2}$. Northern blot hybridization and RT-PCR confirmed that all of these were genuine transcripts (Fig. 3; data not shown). Except for the ribosomal proteins and EST670599, which was homologous with the neuronal apoptosis inhibitory protein 3 (*Naip3*) gene (and thus designated *Naip3L1*), none of the other five human sequences showed similarity to any sequence in the public databases. Based on the location of these five human transcripts within the minimal region defined by a tumor-suppressing subchromosomal fragment that suppresses the growth of RD cells (Koi et al. 1993), we designated these transcripts tumor-suppressing subchromosomal fragment cDNAs 7, 9, 10, and 11 (*Tssc7*, *Tssc9*, *Tssc10*, and *Tssc11*; *TSSC8* is described below) in accordance with our previously established nomenclature (Fig. 1; Table 4)

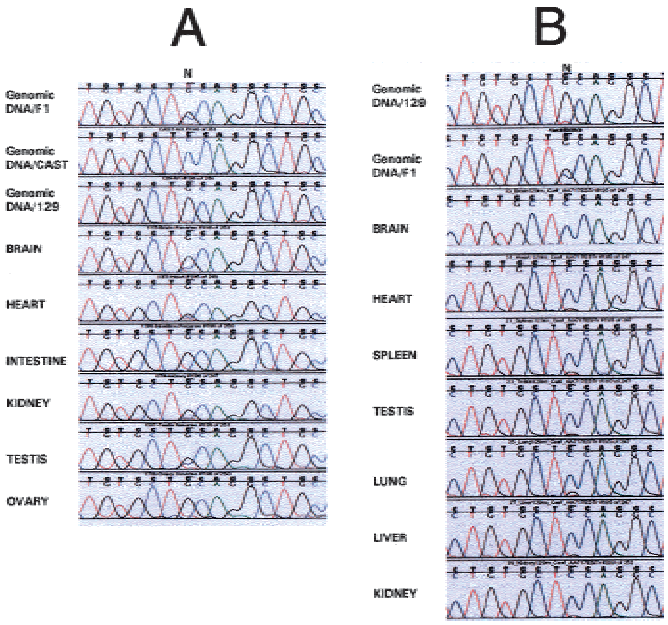

Figure 4 Imprinting analysis of *Msuit*. F1 cDNA derived from fetal and adult tissues was sequenced from bidirectional crosses of *Mus musculus musculus* (129/Sv) and *Mus musculus castaneus* (CAST). A G/C (129/CAST) transcribed polymorphism identified in the genomic DNA at nucleotide 247 was used to distinguish the two alleles. (*A*) Expression analysis of *Msuit* in the brain, heart, intestine, kidney, testis, and ovary of F1 obtained from a cross of 129 (mother) and CAST (father). Genomic DNA sequences from each parent and from F1 are included. (*B*) Expression of *Msuit* in the brain, heart, spleen, testis, lung, liver, and kidney of F1 from the reciprocal cross. Genomic DNA sequences from paternal parent (129) and F1 are included.

Conserved Novel Transcripts

By using PIP matches to search dbEST, we identified a sequence of 332 nt in mouse at nucleotides 660496 to 663300 with 85% identity to human sequence that corresponded to mouse ESTJ1011C10 (accession no. AU041933), as well as to human EST2466762 (accession no. AI933351). This conserved sequence was located 5 kb telomeric to exon 10 of *Kvlqt1* (Table 4; Fig. 2, red) and was designated *Tssc8*. RT-PCR with gene-specific primers showed a transcript in all tissues examined, with transcriptional orientation opposite to *Lit1*, even though *Tssc8* lies within *Lit1* (data not shown). The ESTs do not contain an obvious ORF, nor do they show homology with any known transcripts. Similarly, we identified a mouse EST482800 (accession no. AI594936) located between *H19* and *Rl23mrp* that showed 88% sequence identity to human sequence (Table 4, Fig. 2, red). Because this transcript is immediately telomeric to *H19*, elucidation of its imprinting status may further delimit the telomeric imprinted–nonimprinted subdomain boundary. We designated this transcript *Rhit1* (*R123mrp-H19* interval transcript −1).

```
  1 ATGCCTGCTGCCCACTGGCCCATGCCAGGAAGTGAGTTCTCTGCCAAAGAA
  1 M  P  A  A  H  W  P  M  P  G  S  E  F  S  A  K  E

 52 GCCCAGGCTAAAGTGCCTGTCCTGGCGGCTGCTGAAACAGCCCAGTGTGTC
 18 A  Q  A  K  V  P  V  L  A  A  A  E  T  A  Q  C  V

103 AGCCATGTGTCCAGAACGAGGCACACTGGACTGGTGTCTGCCCCAGATTTG
 35 S  H  V  S  R  T  R  H  T  G  L  V  S  A  P  D  L

154 TGTCCGCAAAACCGAGTTAGTAAGACCTCATTGGAGCCTTGCAAACCCTAC
 52 C  P  Q  N  R  V  S  K  T  S  L  E  P  C  K  P  Y

205 AAAATGGCCCATGCTCTGACTGCTGTGCATGGAGAAGCCATTCCCTTTGCC
 69 K  M  A  H  A  L  T  A  V  H  G  E  A  I  P  F  A

256 TTGCCATGGAGTTTTGTGAGCAACTGA
 86 L  P  W  S  F  V  S  N  *
```

Figure 5 Genscan-predicted nucleotide and amino acid sequence of *Ihit*. The transcript is located between *H19* and *Igf2* in the mouse.

Conserved Intergenic Sequences and a Nonconserved Transcript between *IGF2* and *H19*

The *IGF2* and *H19* genes have attracted great interest as a model for imprinting studies (Wolffe 2000), and both genes can undergo loss of imprinting in cancer (Rainier et al. 1993; for review, see Feinberg 1999). Comparison of mouse and human sequence allowed us to order the region from *Ins* to *L23mrp*, which existed previously only as draft assembly sequence in the Human Genome Project (Bentley 2000). This analysis revealed the presence and location of several previously unrecognized conserved sequences. These include two CpG islands between *Igf2* and *Ins* and two CpG islands located downstream from *H19* (Fig. 2, orange).

In addition, we observed seven conserved nonexonic, nonisland sequences between *Igf2* and *H19* (Fig. 2, blue; Table 5). RT-PCR did not reveal a product in mouse fetal and adult tissues and there were no matches to EST sequences, which indicates that these may represent conserved regulatory sequences. Consistent with this possibility, the conserved sequences are within the region shown in functional complementation experiments to be necessary to maintain normal imprinting of a transgenic YAC containing both *Igf2* and *H19* (Ainscough et al. 1997). Finally, Genscan and GRAIL analysis of the mouse sequence between *Igf2* and *H19* revealed several predicted exons that were not previously known. For one of these predicted exons (nucleotides 76583–76864), we detected a strong 1-kb signal on Northern blots derived from both mouse and human RNA from fetal and adult liver and from placenta (Figs. 3 and 5; data not shown). In addition, a similarly sized transcript was apparent in the human brain (Fig. 3). The predicted protein sequence showed no homology with any known sequences and we designated the gene *Ihit1* (Igf2-H19 interval transcript-1). Northern blot hybridization indicated that the sequence is conserved in human. However, the precise localization must await the completion of the human sequence between *H19* and *IGF2*.

A Two-Island Rule for Imprinted Genes

CpG islands are defined as sequences of ≥200 bp with a GC content (i.e., [G + C]/N > 0.5) and an observed-to-expected CpG dinucleotide content (i.e., [CpG × N]/[C × G] > 0.6; Gardiner-Garden and Frommer 1987). CpG islands are normally unmethylated, but allele-specific methylation of CpG islands appears to mark both the inactive X chromosome (Yen et al. 1984) and many imprinted genes, for example, *H19*, *Snrpn*, and *Igf2r* (Brandeis et al. 1993; Shemer et al. 1997; Wutz et al. 1997). In addition, GC-rich sequences that are not CpG islands (i.e., they meet the first, but not the second criterion above) may also be differentially methylated (termed a differentially methylated region) in the vicinity of imprinted genes, for example *Igf2* (Sullivan et al. 1999) and a second site 2–4 kb upstream of the *H19* CpG island (Thorvaldsen et al. 1998). Therefore, one of our goals was to identify conserved CpG islands and GC-rich sequences that might serve as a substrate for future experiments to investigate allele-specific methylation.

This analysis revealed 33 conserved CpG islands (Fig. 2, orange), and 28 conserved GC-rich (>50%) sequences (Table 5). Remarkably, eight of nine conserved imprinted genes within the entire domain showed two or more conserved CpG islands upstream of or within the gene (Table 6), but all of the six nonimprinted genes were associated with no or one CpG island (Table 6). This difference was statistically significant ($p < 0.01$, Fisher's exact test). Generally, one conserved CpG island associated with each imprinted gene was located <2 kb upstream of the gene and, in some cases, overlapped the first exon, for example, *H19*, *Igf2*, *Mash2*, *Kvlqt1*, *p57^{KIP2}*, *Msuit1*, *Tssc5*, and *Tssc3*. Additional conserved CpG islands associated with the imprinted genes were generally located within an intron and often extended into one or both of the adjacent exons.

Nonisland Conserved Sequences

We identified 49 nonisland conserved sequences that did not correspond to known exons (Fig. 2, blue; Table 5). These sequences were clustered predominantly around imprinted genes. In particular, within the imprinted gene subdomain that extends from *Mash2* to *H19* we identified 10 conserved nonisland sequences, seven of which were located between *H19* and *Igf2* (Fig. 2), and two that were within *Igf2*. Two additional such sequences were located within 14 kb downstream from *H19*. Of the remaining 37 nonisland conserved se-

Table 6. CpG Island Organization and Allelic Expression

Gene	Expressed allele[a]	CpG island 1 location[b]	CpG island 2 location	CpG island 3 location[c]
Rl23mrp	Biallelic	Exon 1 (−100 to +300)	None	None
H19	Maternal	Upstream (−600 to −250)	Intron 1 (+800 to +1300)	Downstream (+6500 to +6700)
Igf2	Paternal	Upstream (−1960 to −2760)	Exon 1 (+500 to +1800)	Intron 2 (+2400 to +2800)
Ins	Biallelic	None	None	None
Th	Biallelic	None	None	None
Mash2	Maternal	Exon 1 (−500 to +900)	Downstream (+1100 to +1600)	None
Tssc6	Biallelic	Intron 7 (+10500 to +10900)	None	None
Tapa1	Unknown	Upstream (−1500 to −600)	None	None
Tssc4	Biallelic	Exon 1,2 (−300 to +400)	None	None
Trpc5l	Biallelic	None	None	None
Kvlqt1	Maternal	Exon 1 (−400 to +300)	Intron 10	Intron 15
Lit1	Paternal	Upstream (−1600 to −1200)	None	None
p57 KIP2	Maternal	Upstream (−2300 to −2700)	Exons 1,2,3	None
Msuit	Maternal	Upstream (−1000 to −3500)	Downstream (+2000 to +2800)	None
Tssc5	Maternal	Upstream (−1100 to −900)	Intron 6	None
Tssc3	Maternal	Upstream (−1300 to −1100)	Exon 1 (−500 to +300)	None
Nap1l4	Biallelic	None	None	None

[a]Considered imprinted if imprinted in the predominantly expressing tissues at some stage of development.
[b]Locations are calculated with respect to the start site of transcription.
[c]More than 3 CpG islands are not listed.

quences, 36 were located within the imprinted gene subdomain that extends from *Tssc3* to *Kvlqt1*, and 33 of these were within *Kvlqt1* itself. Interestingly, 12 of these conserved sequences were located within 44 kb upstream of the *Lit1* CpG island (Fig. 2), and six of these are GC rich, even though they did not meet the full definition of a CpG island. It will be of interest to determine whether any of these conserved GC-rich sequences are differentially methylated between the two parental chromosomes, given that the CpG island immediately upstream of *Lit1* is not conserved between human and mouse.

DISCUSSION

In this report, we have described the first sequencing and comparative analysis of an entire imprinted gene domain between human and mouse. If one excludes a gap that remains within the human genome sequence, which we have sequenced in the mouse, and smaller gaps within the mouse sequence, this analysis includes 915 kb of mouse and 900 kb of human, the largest comparative sequencing analysis of a single ordered and oriented domain to date. The majority of the mouse sequence analyzed in this study reflects draft sequence assemblies (Collins et al. 1998). The value of the draft sequence, which is anticipated to provide >90% coverage (Bouck et al. 1998), has been greatly enhanced through the availability of sequence from an orthologous region of a second species.

The order and orientation of the mouse sequence contigs could be established through alignment with respect to the human sequence, allowing us to clearly establish positional information for the conserved sequence elements. In this case, the available human sequence was finished, but for organisms for which the evolutionary distance is similar to that between human and mouse, comparable utility can be obtained when each of the sequences is draft (K. Dewar and W. Miller, unpubl.).

We found 16 conserved known human genes that were made up of 119 exons in the human and 112 in the mouse. Of these, 110 (98%) were conserved. There were also several transcripts present in this region in one species but not the other, including ribosomal protein *L26* in human, ribosomal protein L13 and a homolog of *Naip3* (*Naip3L1*) in mouse, and several ESTs unique to one species or the other. We showed that one of the sequences unique to the mouse was imprinted, and we designated it *Msuit1*, for Mouse-specific ubiquitously imprinted transcript-1. An intriguing potential mechanistic explanation for the imprinting of *Msuit1* is that the location of a gene within this domain may subject it to long-range *cis*-acting regulatory sequences that are responsible for allele-specific silencing, such as chromatin alterations acting at a distance, similar to telomere silencing in yeast or to position effect variegation in *Drosophila*.

One of the most striking conclusions of this analysis is that the number of conserved sequences outside the known coding exons and interspersed repeats is small. There were 82 such sequences, with an average length of 337 bp, thus making up ~4.1% of the total noncoding sequence throughout the domain. The sequence analysis of Loots et al. (2000) found 91 conserved sequences (each ≥100 bp of 70% identity) distributed >900 kb of noncontiguous draft assembly se-

quence, although the fraction of sequence this represents was not reported. Conservation of 1% of noncoding sequence was also reported over a relatively short interval (92 kb; Jang et al. 1999). Thus comparative sequencing may be a powerful strategy for identifying the critical nonexonic regulatory sequences that would be difficult to determine by analysis of a single genome.

Of these 82 sequences, 33 (42%) were CpG islands and 28 were GC-rich sequences in both species. Thus 61 of 82 (74%) of the conserved nonexonic sequences were either GC rich or were true CpG islands. This provides further evidence of an important role for DNA methylation in the regulation of genes throughout this domain. Consistent with this idea, at least some of these sequences appear to show partial methylation in genomic DNA (P. Onyango and A.P. Feinberg, unpubl.), including the CpG islands, which are normally unmethylated except for the inactive X-chromosome and imprinted genes (Yen et al. 1984; Brandeis et al. 1993; Shemer et al. 1997; Wutz et al. 1997). We are currently determining which of these sequences might show allele-specific methylation.

The location of these conserved sequences is also of particular interest in that they are not randomly distributed. We had previously shown that the imprinted domain is itself divided into two imprinted subdomains in human (*TSSC3* to K_VLQT1, and *ASCL2* to *H19*), with a region of little or no imprinting between them (*TSSC4* to *TSSC6*) (Lee et al. 1998; Feinberg 1999). All but one of the conserved sequences fell within one of the two imprinted subdomains. This observation provides further support for a role of these sequences in the regulation of genomic imprinting.

Curiously, we found that the imprinted genes tended to be associated with two or more CpG islands. This also appears to be true for imprinted genes on other chromosomes (Yen et al. 1984; Brandeis et al. 1993; Shemer et al. 1997; Wutz et al. 1997), although, to our knowledge, this has not been commented on in the literature, likely because interspecies global sequence comparisons have not been possible. We suggest that there may be a two-island rule for imprinting, that is, in most cases more than one CpG island is required to maintain normal imprinting. Perhaps the additional CpG island is related to a second methylation mark or, alternatively, to the presence of antisense transcripts associated with these genes. The latter appears to be the case for *Kvlqt1*, *Igf2r*, and *Igf2*.

This analysis also revealed that a CpG island upstream of the human *Lit1* antisense RNA is in fact not conserved in the mouse, even though it shows differences in allele-specific methylation and alterations in BWS. However, we identified several GC-rich sequences, 5–44 kb upstream of this CpG island that are >70% conserved between human and mouse. Preliminary analysis suggests that at least one of these sequences also shows allele-specific methylation (P. Onyango and A.P. Feinberg, unpubl.) and thus it might be important in normal imprint regulation or disease. Another potentially important sequence is a 75% conserved CpG island 4 kb upstream of $p57^{KIP2}$. In contrast to the CpG island within $p57^{KIP2}$, which is unmethylated in humans, this newly identified sequence is partially methylated in humans (P. Onyango and A.P. Feinberg, unpubl.).

The mouse *Igf2* and *H19* genes have attracted a great deal of interest, but the sequence between them has been previously unknown. The human sequence between these genes has been reported by the Human Genome Project in six unordered fragments. We were able to order the human interval between *IGF2* and *H19* by comparison to mouse sequence. This analysis revealed 10 conserved sequences in this interval, including three CpG islands. A novel gene termed *Ihit* also lies within this interval, at least in the mouse.

Finally, an intriguing concept in the study of genomic imprinting is the idea of a large genomic domain that might be regulated hierarchically, with some local elements regulating individual genes and other elements having more global effects. Such an idea is consistent with the imprinting center deletions observed in Prader-Willi and Angelman syndromes, which disrupt imprinting over several megabases. Similarly, we have observed patients with BWS and loss of imprinting affecting either *LIT1* or *IGF2* but not both, and others with loss of imprinting in both gene regions (Lee et al. 1999; DeBaun et al., in prep.). It will thus be of interest to examine the conserved sequences identified here not only in normal tissues, but also in disease tissues, to gain insight into their potential role as more global *cis*-acting regulators of gene expression.

METHODS

Isolation of a 10×-Depth BAC Contig from Mouse Chromosome 7, Identification of a Minimal Tiling Path, and Sequencing of the Mouse Contig

An overgo hybridization protocol (Ross et al. 1999) was used for probes generated from gene sequences of the imprinted region. Forty-five overgos were pooled and screened against high-density BAC clone filters of a 11.2× genomic equivalent female mouse C57 BL/6J genomic library (RPCI-23; BAC/PAC Resources, Oakland CA; www.chori.org/bacpac/). Single-colony isolates were recovered from all positive well addresses, rearrayed into a 384-well microtitre plate, and then duplicated onto a series of filters (HybondN+, Amersham). Each overgo probe was tested against a rearrayed copy to establish the marker and clone relationships. Using marker-clone content and *Hin*dIII fingerprint information (Marra et al. 1997) a set of five minimally overlapping clones were selected for sequencing (GenBank accessions nos. AC013548, AC012382, AC015800, AC012540, and AC023248). Draft sequence assembly of all the clones was performed by ligating

mechanically sheared 2-kb fragments of BAC DNA into an m13 sequencing vector, followed by random shotgun sequencing at 5× coverage of the estimated clone size, and then assembly. To increase sequence contiguity and establish the order and orientation of the sequence within AC012382, an additional subclone library of 4-kb fragment size was prepared and sequenced in a plasmid sequencing vector. Plasmid subclones were sequenced from both ends to an additional 5× coverage and integrated into the assembly. Sequence gaps and ambiguities were subsequently resolved using standard finishing techniques (Wilson and Mardis 1997). We were able to order and align the mouse draft sequences with the human by performing both a PIP comparison and an analysis using a novel NCBI toolkit termed Alignment Construction Utility and Tools Environment (ACUTE). ACUTE is capable of generating, viewing, and analyzing discontinuous or overlapping sequence alignments. The mouse draft assembled sequence, although multipass and >99.9% accurate, was in unordered fragments, and the human sequence was in three large pieces, with gaps of unreported size. The initial set of mouse–human alignments was used to order and orient the mouse draft sequence. Approximately 95% of the sequence could be unambiguously ordered this way to generate an ordered and oriented sequence spanning the entire imprinted region. Similarly, the human sequences could be ordered, oriented, and concatenated. The sequences used in our analysis can be obtained at http://www.jhmi.edu/feinberg_lab or http://bio.cse.psu.edu/. A gap remains in the human sequence spanning the *TH* gene. Therefore, in this area, deeper coverage mouse sequence was obtained. Thus comprehensive sequence was generated over the entire imprinted domain and comparison between mouse and human could be performed over all but the portion not yet completed by the Human Genome Project.

Global Comparison of the Mouse and Human Sequences

To compare the mouse and human sequences over the entire imprinted domain we used `PipMaker` (http://bio.cse.psu.edu/PipMaker/). The program was run in a manner constraining matches to be both conserved and colinear between the two species. Matches of a desired minimum length and percent identity lying between consecutive gaps in a `PipMaker` alignment were found with a program called strong_hits, which can be downloaded from the `PipMaker` site. The human sequences were retrieved from GenBank (accession nos. NT_000558, NT_000557, and AC006408). We used the concatenated mouse sequence as the reference sequence in `PipMaker` analysis. To eliminate spurious matches resulting solely from low and high complexity repeats, we masked the mouse sequence using `RepeatMasker` (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) before performing the `PipMaker` analysis. `RepeatMasker` was also used to deduce the repeat content for the sequences from each species. Tandem repeats were identified with the program `Tandem Repeats Finder` (http://c3.biomath.mssm.edu/trf.html; Benson 1999).

Gene Prediction

To identify potential genes in both the mouse and the human sequences we used a four-step approach. First, we masked the sequences for high complexity repeats using `RepeatMasker`. Second, repeat-masked sequences were analyzed for exon content using `Genscan` (http://ccr-081.mit.edu/Genscan.html), `GRAIL` (http://grail.lsd.ornl.gov/Grail-1.3) and `PipMaker`. Third, we used all the predicted coding sequences or highly conserved sequences from step one to search GenBank databases. The fourth step involved direct `BLAST` database searches using fragments of either the mouse or human sequences

Identification of Conserved Sequences

CpG islands were found by a simple program, written in `C`, that looks in 200-residue windows for regions that meet the definition of Gardiner-Garden and Frommer (1987). Conserved sequences were identified as described in the text.

Imprinting Analysis

Mice were purchased from Jackson Laboratory. We crossed inbred *Mus musculus* (129/Sv) to inbred *Mus musculus castaneus* (CAST/Ei) to obtain F1 mice with polymorphic genotype. To identify polymorphisms we amplified by PCR and sequenced genomic DNA from F1, 129Sv, and CAST/Ei. PCR conditions were as follows: 2 min at 95°C; then 40 cycles each of 1 min at 95°C, 30 sec at 60°C, 1 min at 72°C; then 9 min at 72°C. RNA was extracted from tissues of F1 animals derived from crosses from both directions using the protocols outlined below. Total RNA was isolated using RNeasy minikit from Qiagen. To eliminate DNA contamination from RNA preparations, samples were treated with preamplification-grade DNase I (GIBCO) according to supplied protocols. RT-PCR was performed using the Superscript II preamplification system (GIBCO) and was performed for each sample in the presence and absence (negative controls) of RT. Samples were sequenced only when no bands were obtained with the negative controls. The primers used for the imprinting analysis were ESTAA7179-F: 5′-AAGCAAGTGATGCAAGCATCC-3′ and ESTAA7179-R: 5′-ACTCCACACTTATTTGTGACC-3′. DNA and cDNA sequencing was run on an ABI-377 automated sequencer following protocols recommended by the manufacturer (Perkin-Elmer).

Northern Blots

Multiple-tissue Northern blots were purchased from Clontech. Hybridization and washes were performed according to manufacturer's recommendations. Blots were exposed to X-Ray films for 1–14 days.

ACKNOWLEDGMENTS

We thank Eric S. Lander for encouragement and support, members of the WI/MIT Center for Genome Research and UTSW Genome Science and Technology Center for genomic sequencing of the mouse and human regions, respectively, and the members of the Feinberg laboratory for helpful discussions and technical assistance. This work was supported by grants from the National Institutes of Health to A.P.F., W.M., and E.S.L.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Ainscough, J.F., John, R.M., and Surani, M.A. 1998. Mechanism of imprinting on mouse distal chromosome 7. *Genet. Res.* **72:** 237–245.

Ainscough, J.F., Koide, T., Tada, M., Barton, S., and Surani, M.A. 1997. Imprinting of *Igf2* and *H19* from a 130 kb YAC transgene. *Develop.* **124:** 3621–3632.

Bell, A.C. and Felsenfeld, G. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405:** 482–485.

Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27:** 573–580.

Bentley, D.R. 2000. The Human Genome Project - an overview. *Med. Res. Rev.* **20:** 189–196.

Blake, J.A., Eppig, J.T., Richardson, J.E., and Davisson, M.T. 2000. The Mouse Genome Database (MGD): Expanding genetic and genomic resources for the laboratory mouse. The mouse genome database group. *Nucl. Acids Res.* **28:** 108–111.

Bouck, J., Miller, W., Gorrell, J.H., Muzny, D., and Gibbs, R.A. 1998. Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Res.* **8:** 1074–1084.

Brandeis, M., Kafri, T., Ariel, M., Chaillet, J.R., McCarrey, J., Razin, A., and Cedar, H. 1993. The ontogeny of allele-specific methylation associated with imprinted genes in the mouse. *EMBO J.* **12:** 3669–3677.

Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. 1998 New goals for the U.S. Human Genome Project: 1998–2003. *Science* **282:** 682–689.

Elgar, G. 1996. Quality not quantity: The pufferfish genome. *Hum. Mol. Genet.* **5:** 1437–1442.

Endrizzi, M., Huang, S., Scharf, J.M., Kelter, A.R., Wirth, B., Kunkel, L.M., Miller, W., and Dietrich, W.F. 1999.Comparative sequence analysis of the mouse and human Lgn1/SMA interval. *Genomics* **60:** 137–151.

Feinberg, A.P. 1999. Imprinting of a genomic domain of 11p15 and loss of imprinting in cancer: An introduction. *Cancer Res.* **59:** 1743–1746.

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

Hardison, R.C., Oeltjen, J., and Miller, W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7:** 959–966.

Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M., and Tilghman, S.M. 2000. CTCF mediates methylation-sensitive enhancer-blocking activity at the *H19/Igf2* locus. *Nature* **405:** 486–489.

Hu, R.J., Lee, M.P., Connors, T.D., Johnson, L.A., Burn, T.C., Su, K., Landes, G.M., and Feinberg, A.P. 1997. A 2.5-Mb transcript map of a tumor-suppressing subchromosomal transferable fragment from 11p15.5, and isolation and sequence analysis of three novel genes. *Genomics* **46:** 9–17.

Hu, R.J., Lee, M.P., Johnson, L.A., and Feinberg, A.P. 1996. A novel human homolog of yeast nucleosome assembly protein, 65 kb centromeric to the $p57^{KIP2}$ gene, is biallelically expressed in fetal and adult tissues. *Hum. Mol. Genet.* **5:** 1743–1748.

Jang, W., Hua, A., Spilson, S.V., Miller, W., Roe, B.A., and Meisler, M.H. 1999. Comparative sequence of human and mouse BAC clones from the mnd2 region of chromosome 2p13. *Genome Res.* **9:** 53–61.

Koi, M., Johnson, L.A., Kalikin, L.M., Little, P.F.R., Nakamura, Y., and Feinberg, A.P. 1993. Tumor cell growth arrest caused by subchromosomal transferable DNA fragments from human chromosome 11. *Science* **260:** 361–364.

Lee, M.P., Brandenburg, S., Landes, G.M., Adams, M., Miller, G., and Feinberg, A.P. 1998. Two novel genes in the center of the 11p15 imprinted domain escape genomic imprinting. *Hum. Mol. Genet.* **8:** 683–690.

Lee, M.P., DeBaun, M.R., Mitsuya, K., Galonek, H.L., Brandenburg, S., Oshimura, M., and Feinberg, A.P. 1999. Loss of imprinting of a paternally expressed transcript, with antisense orientation to K_VLQT1, occurs frequently in Beckwith-Wiedemann syndrome and is independent of insulin-like growth factor II imprinting. *Proc. Natl. Acad. Sci.* **96:** 5203–5208.

Lee, M.P., DeBaun, M., Randhawa, G.S., Reichard, B.A., and Feinberg, A.P. 1997. Low frequency of $p57^{KIP2}$ mutation in Beckwith-Wiedemann syndrome. *Am. J. Hum. Genet.* **61:** 304–309.

Lee, M.P., Hu, R.J., Johnson, L.A., and Feinberg, A.P. 1997. Human KVLQT1 gene shows tissue-specific imprinting and encompasses Beckwith-Wiedemann syndrome chromosomal rearrangements. *Nat. Genet.* **15:** 181–185.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.

Lund, J., Chen, F., Hua, A., Roe, B., Budarf, M., Emanuel, B.S., and Reeves, R.H. 2000. Comparative sequence analysis of 634 kb of the mouse chromosome 16 region of conserved synteny with the human velocardiofacial syndrome region on chromosome 22q11.2. *Genomics* **63:** 374–383.

Mallon, A.M., Platzer, M., Bate, R., Gloeckner, G., Botcherby, M.R., Nordsiek, G., Strivens, M.A., Kioschis, P., Dangel, A., Cunningham, D., et al. 2000. Comparative genome sequence analysis of the Bpa/Str region in mouse and Man. *Genome Res.* **10:** 758–775.

Marra, M.A., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D., and Waterston, R.H. 1997. High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7:** 1072–1084.

O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J., Nash, W.G., Wienberg, J., Stanyon, R., Copeland, N.G., Jenkins, N.A., Womack, J.E., and Marshall Graves, J.A. 1999. The promise of comparative genomics in mammals. *Science* **286:** 458–481.

Rachmilewitz, J., Gonik, B., Goshen, R., Ariel, I., Schneider, T., de Groot, N., and Hochberg, A. 1993. Use of a novel system for defining a gene imprinting region. *Biochem. Biophys. Res. Commun.* **196:** 659–664.

Rainier, S., Johnson, L.A., Dobry, C.J., Ping, A.J., Grundy, P.E., Feinberg, A.P. 1993. Relaxation of imprinted genes in human cancer. *Nature* **362:** 747–749.

Ross, M.T., LaBrie, S., McPherson, J., and Stanton, V.P. 1999. Screening large-insert libraries by hybridization. In *Current protocols in human genetics* (eds. N.C. Dracopoli, J.L. Haines, B.R. Korf, D.T. Moir, C.C. Morton, C.E. Seidman, J.G. Seidman, and D.R. Smith), pp. 5.6.1–5.6.52. J. Wiley, New York.

Sapienza, C., Peterson, A.C., Rossant, J., and Balling, R. 1987. Degree of methylation of transgenes is dependent on gamete of origin. *Nature* **328:** 251–254.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. 2000. PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10:** 577–586.

Shemer, R., Birger, Y., Riggs, A.D., and Razin, A. 1997. Structure of the imprinted mouse *Snrpn* gene and establishment of its parental-specific methylation pattern. *Proc. Natl. Acad. Sci.* **94:** 10267–10272.

Srivastava, M., Hsieh, S., Grinberg, A., Williams-Simons, L., Huang, S.P., and Pfeifer, K. 2000. *H19* and *Igf2* monoallelic expression is regulated in two distinct ways by a shared *cis*-acting regulatory region upstream of *H19*. *Genes & Dev.* **14:** 1186–1195.

Steenman, M.J.C., Rainier, S., Dobry, C.J., Grundy, P., Horon, I.L., and Feinberg, A.P. 1994. Loss of imprinting of *IGF2* is linked to reduced expression and abnormal methylation of *H19* in Wilms' tumor. *Nat. Genet.* **7:** 433–439.

Sullivan, M.J., Taniguchi, T., Jhee, A., Kerr, N., and Reeve, A.E. 1999. Relaxation of *IGF2* imprinting in Wilms tumours associated with specific changes in *IGF2* methylation. *Oncogene* **18:** 7527–7534.

Sutcliffe, J.S., Nakao, M., Christian, S., Orstavik, K.H., Tommerup, N., Ledbetter, D.H., and Beaudet, A.L. 1994. Deletions of a differentially methylated CpG island at the *SNRPN* gene define a putative imprinting control region. *Nat. Genet.* **8:** 52–58.

Szebenyi, G. and Rotwein, P. 1994. The mouse insulin-like growth factor II/cation-independent mannose 6– phosphate (*IGF-II/MPR*) receptor gene: Molecular cloning and genomic organization. *Genomics* **19:** 120–129.

Thorvaldsen, J.L., Duran, K.L., and Bartolomei, M.S. 1998. Deletion

of the *H19* differentially methylated domain results in loss of imprinted expression of *H19* and *Igf2*. *Genes & Dev*. **12:** 3693–3702.

Tsang, P., Gilles, F., Yuan, L., Kuo, Y-H., Lupu, F., Samara, G., Moosikasuwan, J., Goye, A., Zelenetz, A.D., Selleri, L., and Tycko, B. 1995. A novel *L23*-related gene 40 kb downstream of the imprinted *H19* gene is biallelically expressed in mid-fetal and adult human tissues. *Hum. Mol. Genet.* **4:** 1499–1507.

Weksberg, R., Shen, D.R., Fei, Y.L., Song, Q.L., and Squire, J. 1993. Disruption of insulin-like growth factor 2 imprinting in Beckwith-Weidemann syndrome. *Nat. Genet.* **5:** 143–150.

Wilson, R.K. and Mardis, E.R. 1997. Shotgun sequencing. In *Genome analysis: A laboratory manual* (eds. B. Birren, E.D. Green, S. Klapholz, R.M. Myers, and J. Roskams), pp. 397–454. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Wolffe, A.P. 2000. Imprinting insulation. *Curr. Biol.* **10:** 463–465.

Wutz, A., Smrzka, O.W., Schweifer, N., Schellander, K., Wagner, E.F., and Barlow, D.P. 1997. Imprinted expression of the *Igf2r* gene depends on an intronic CpG island. *Nature* **389:** 745–749.

Yen, P.H., Patel, P., Chinault, A.C., Mohandas, T., and Shapiro, L.J. 1984. Differential methylation of hypoxanthine phosphoribosyltransferase genes on active and inactive human X chromosomes. *Proc. Natl. Acad. Sci.* **81:** 1759–1763.

Zhang, Z., Berman, P., Wiehe, T., and Miller W. 1999. Post-processing long pairwise alignments. *Bioinform.* **15:** 1012–1019.

Zubair, M., Hilton, K., Saam, J.R., Surani, M.A., Tilghman, S.M., and Sasaki, H. 1997. Structure and expression of the mouse *L23mrp* gene downstream of the imprinted *H19* gene: Biallelic expression and lack of interaction with the *H19* enhancers. *Genomics* **45:** 290–296.

Received August 24, 2000; accepted in revised form September 19, 2000.