doi:10.1006/geno.2002.6692, available online at http://www.idealibrary.com on IDEAL

Candidate Genes Required for Embryonic Development: A Comparative Analysis of Distal Mouse Chromosome 14 and Human Chromosome 13q22

Laurie Jo Kurihara,[1,*] Ekaterina Semenova,[1,*] Webb Miller,[2] Robert S. Ingram,[1] Xiao-Juan Guan,[1] and Shirley M. Tilghman[1,†]

[1]*Howard Hughes Medical Institute and Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA*
[2]*Department of Computer Science and Engineering, Penn State University, University Park, Pennsylvania 16802, USA*

**These authors contributed equally to this work*

[†]*To whom correspondence and reprint requests should be addressed. Fax: (609) 258-3345. E-mail: stilghman@molbio.princeton.edu.*

Mice homozygous for the *Ednrb*$^{s\text{-}1Acrg}$ deletion arrest at embryonic day 8.5 from defects associated with mesoderm development. To determine the molecular basis of this phenotype, we initiated a positional cloning of the Acrg minimal region. This region was predicted to be gene-poor by several criteria. From comparative analysis with the syntenic human locus at 13q22 and gene prediction program analysis, we found a single cluster of four genes within the 1.4- to 2-Mb contig over the Acrg minimal region that is flanked by a gene desert. We also found 130 highly conserved nonexonic sequences that were distributed over the gene cluster and desert. The four genes encode the TBC (*Tre*-2, BUB2, CDC16) domain-containing protein KIAA0603, the ubiquitin carboxy-terminal hydrolase L3 (UCHL3), the F-box/PDZ/LIM domain protein LMO7, and a novel gene. On the basis of their expression profile during development, all four genes are candidates for the *Ednrb*$^{s\text{-}1Acrg}$ embryonic lethality. Because we determined that a mutant of *Uchl3* was viable, three candidate genes remain within the region.

Key Words: human chromosome 13, genome mapping, comparative study, UCHL3, LMO7, chromosome deletion, mutant mouse strain, embryology

INTRODUCTION

The sequence of the human genome is largely complete [1,2], and the mouse genome sequence is well underway. This information provides a precise transcript map, but does not, in many cases, reveal gene function, in that only 60% of human proteins have sequence similarity to proteins from organisms whose genomes have been sequenced. However, most human genes possess a mouse ortholog, making the mouse a model system for uncovering mammalian gene function. An efficient way to analyze genome function is by generating deletions. Because genes are not distributed uniformly in the genome, deletions will identify regions that contain essential genes and genes with more subtle roles as well as chromosomal segments that are devoid of function. Detailed analysis of overlapping sets of deletions has been used to assign functions to chromosomal regions essential for mouse development and survival [3]. Deletion complexes also provide a valuable tool for region-specific saturation mutagenesis screens for recessive point mutations [4].

In the mouse, deletions can be generated by targeted mutagenesis in embryonic stem cells [5] or by chemical and radiation mutagenesis of the germ line or embryonic stem cells [6,7]. The specific locus test (SLT) used chemical/radiation mutagenesis in the first genetic screen for induced mutations in the mouse [3,8]. This strategy allowed for the recovery of mutations at seven recessive "tester" loci chosen for their easily scored phenotypes. In this screen, mutagenized males were crossed to females homozygous for the seven tester loci, and resulting progeny carrying induced mutations were identified by uncovering one of the seven recessive phenotypes. Depending on the mutagen used, the resulting molecular lesions ranged in size from point mutations to large deletions spanning multiple centimorgans.

The piebald (*s*) locus on mouse chromosome 14 was one of the specific loci used in the SLT. The piebald gene encodes the endothelin B receptor (EDNRB), a G-protein-coupled seven-transmembrane receptor required for migration of two neural crest derivatives, melanocytes and enteric ganglia

0888-7543/02 $35.00

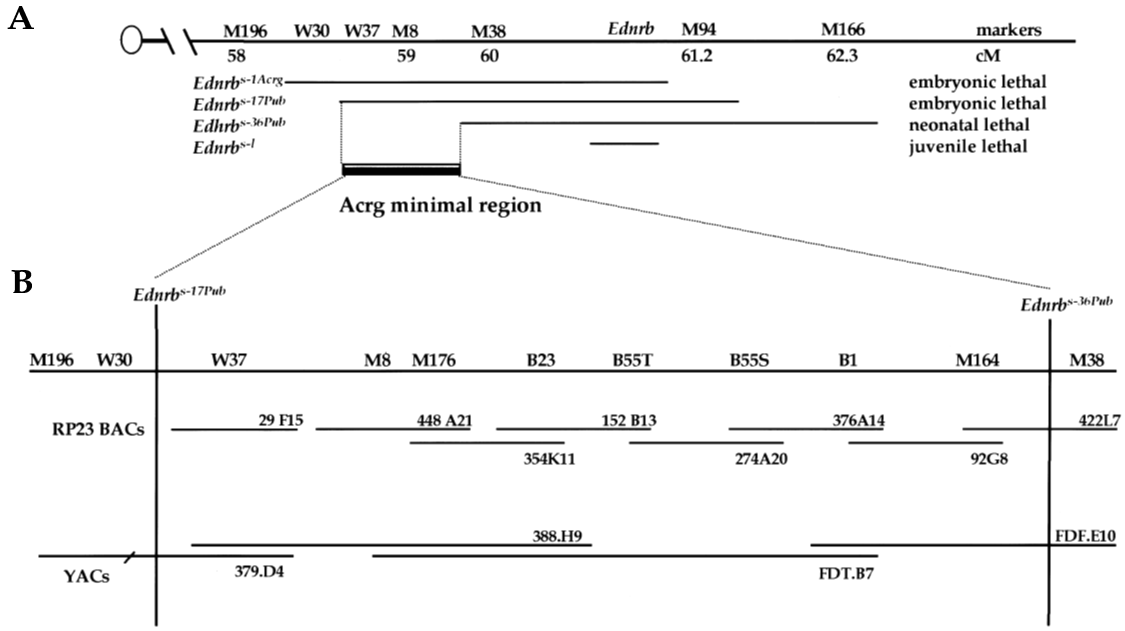


FIG. 1. The physical map of the Acrg minimal region. (A) *Ednrb* deletions that define the Acrg minimal region. Molecular markers are shown above the chromosome. M, *D14Mit* markers; W, Whitehead markers; W30, *30.MMHAP26FRB12.seq*; W37, *37.MMHAP25FLE4.seq*. Genetic map distances in centimorgans are shown below the chromosome. The extent of four deletions and their corresponding phenotypes are also depicted below the chromosome. The Acrg minimal region (shaded box) is defined by the proximal breakpoints of the $Ednrb^{s\text{-}17Pub}$ and $Ednrb^{s\text{-}36Pub}$ deletions. (B) BAC and YAC contigs that span the Acrg minimal region. Vertical lines correspond to the proximal breakpoints of the $Ednrb^{s\text{-}17Pub}$ and $Ednrb^{s\text{-}36Pub}$ deletions. Molecular markers are shown above the top line. B, BAC end markers. The RP23 C57Bl/6J mouse BAC contig and the YAC contig are depicted as solid lines. Corresponding clone numbers are indicated.

[9,10]. Homozygous null mutations in *Ednrb* result in juvenile lethality due to the loss of enteric neurons [11,12]. Many *Ednrb* alleles generated in the SLT are deletions that exhibit a more severe phenotype than loss of *Ednrb* alone, most likely due to the loss of linked essential genes [13]. Complementation analysis between deletion alleles combined with molecular mapping identified specific functional regions associated with embryonic lethality, neonatal lethality, skeletal defects, and CNS defects [14].

Embryos homozygous for one such deletion, $Ednrb^{s\text{-}1Acrg}$, arrest at embryonic day 8.5 and exhibit defects in the primitive streak, node, notochord, and somites [15]. In addition, embryos display randomized direction of embryonic turning, arrested heart looping morphogenesis, vascular defects, and incomplete neural tube closure. Welsh and O'Brien [15] proposed that $Ednrb^{s\text{-}1Acrg}$ disrupts a morphogenetic pathway important for the development of streak-derived posterior mesoderm and lateral morphogenesis. Although the $Ednrb^{s\text{-}1Acrg}$ phenotype is complex and may be multigenic, several single-gene mutations lead to arrest at embryonic day 8.5 with a similar phenotype [16]. Therefore, the $Ednrb^{s\text{-}1Acrg}$ phenotype could result from the loss of a single gene that is essential during development.

Complementation analysis indicated that only the proximal portion of the $Ednrb^{s\text{-}1Acrg}$ deletion is responsible for the embryonic lethality [14]. We refer to this as the Acrg minimal region (Fig. 1). To determine the molecular basis of the embryonic lethality, we identified and analyzed the genes within the minimal region.

RESULTS

Mapping of the Acrg Minimal Region

The Acrg minimal region was originally defined as lying between the proximal breakpoint of the $Ednrb^{s\text{-}1Acrg}$ deletion and the proximal breakpoint of the $Ednrb^{s\text{-}36Pub}$ deletion, on the basis of the latter's ability to complement the embryonic lethality of $Ednrb^{s\text{-}1Acrg}$ [14]. Homozygous embryos from a third deletion, $Ednrb^{s\text{-}17Pub}$, showed a developmental arrest point and gross morphology that was indistinguishable from that of $Ednrb^{s\text{-}1Acrg}$ homozygotes (data not shown). Mapping with additional molecular markers determined that the position of the proximal breakpoint of $Ednrb^{s\text{-}17Pub}$ is distal to that of $Ednrb^{s\text{-}1Acrg}$. Markers W30 and W37 are both absent from $Ednrb^{s\text{-}1Acrg}$, but marker W30 is present in $Ednrb^{s\text{-}17Pub}$ (Fig. 1). This result effectively reduced the size of the Acrg minimal region.

To identify genes within the Acrg minimal region, a contig of P1, BAC, and YAC clones was isolated using a combination of *D14Mit* markers and P1 and BAC end probes within the region (Fig. 1). The Acrg region is estimated to be a maximum size of 2 Mb on the basis of the sum of the YAC sizes covering the region. However, because the positions of markers W30 and W37 are not known on YAC 379.D4, the entire length of this YAC was included in our estimate, even though this YAC contains the $Ednrb^{s\text{-}17Pub}$ breakpoint and more proximal markers such as *D14Mit196*. Therefore, the Acrg region may be as small as 1.4 Mb. The mouse RPCI-23 (C57BL/6J)

doi:10.1006/geno.2002.6692, available online at http://www.idealibrary.com on IDEAL

TABLE 1: Cross-hybridization of syntenic human ESTs to the mouse *Ednrb*$^{s\text{-}1Acrg}$ contig

cR	STS	Acrg contig	Gene
210.67	stsG3741	–	
211.28	WI-17043	–	
213.61	WI-13803	–p	
213.71	Cda16b10	–	
214.01	WI-11923	–	
215.64	R26283	–	
217.58	stSG51977	354K11 / 448A21	*AK000009* (AA425275)
219.10	WI-19625	274A20	*LMO7* (U90654)
219.19	stSG50944	–	
219.19	WI-16413	–^	
219.19	WI-17550	–	
219.31	SGC30014	354K11 / 152B13	*UCHL3* (R42127)
219.60	stSG15664	274A20	EST (R74041)
220.75	A006Y41	–d	
226.84	sts-N92658	–	
226.94	stSG63191	–	
226.94	D13162	nd	*EDNRB*
228.56	SGC30285	–d	

The "cR" column indicates radiation hybrid map position from Genemap 99; the "STS" column denotes human clones analyzed for cross-hybridization; the "Acrg contig" column indicates hybridization results to mouse RP23 BACs; the "Gene" column lists gene names and corresponding cDNA clones used for hybridization. In the "Acrg contig" column, -p and -d refer to clones that hybridized to YACs proximal and distal to the Acrg contig, respectively; -^ refers to a 3′ UTR *Lmo7* clone that failed to hybridize as a result of insufficient cross-species homology.

BAC contig is 1.3 Mb and covers most of the Acrg region, with two presumably small gaps flanking marker W37.

Gene Identification within Acrg

One way to identify genes from cloned DNA is to map CpG-rich islands that are often associated with the 5′ end of genes and are underrepresented in noncoding DNA [17]. On the basis of the low CpG content of syntenic human chromosome 13 [18], the Acrg minimal region was predicted to be gene-poor. To gain information about the location of genes on the Acrg contig, mouse clones were mapped using restriction enzymes that cleave within CpG-rich DNA (data not shown). Consistent with our prediction, much of the contig was nearly devoid of CpG-containing restriction sites for *Not*I, *Eag*I, and *Bss*HII. However, the center of the contig displayed a relatively higher density of CpG residues as evidenced by a single *Not*I site and numerous *Eag*I and *Bss*HII sites. Therefore, we predicted that the gene or genes underlying the embryonic lethality would map to this region.

To identify candidate genes, we analyzed the highly refined map of human chromosome 13q22 that is syntenic to the Acrg minimal region. Before draft sequence availability, the most powerful resource for mapped genes and ESTs was Genemap 97–99. This database provides the radiation hybrid map position of over 30,000 expressed sequence tags (ESTs) distributed throughout the human genome. Because the Acrg minimal region maps 1–2 cM proximal to *Ednrb* in the mouse (Fig. 1), we hybridized human ESTs that mapped within a similar distance proximal to *EDNRB* to the mouse Acrg contig (Table 1). From a total of 17 human ESTs analyzed, 7 clones cross-hybridized to the mouse contig. Four of these mapped within the CpG-rich core of the Acrg minimal region. Three corresponded to the *AK000009, UCHL3,* and *LMO7* genes. R74041 was represented by multiple single-exon ESTs, but subsequent hybridization of a corresponding mouse cDNA to multitissue poly(A) northern blots did not yield evidence of expression. The three remaining human ESTs mapped to YACs outside the Acrg minimal region.

Once we had obtained draft sequence from the mouse RPCI-23 (C57BL/6J) BAC contig (Fig. 1), we were able to carry out a comparative analysis for conserved sequences between the mouse Acrg minimal region and the corresponding syntenic human locus. We used the PipMaker program [19], which computes alignments of similar regions in two DNA sequences and thus identifies exons as well as potential regulatory elements. Using a criterion of ≥ 70% identity over a stretch of at least 100 bp, we identified the three genes found by cross-species hybridization, EST R74041, and one additional gene, *KIAA0603*, that also mapped to the CpG-rich core (Fig. 2). PipMaker identified nearly every exon of all four genes. Exons that were not recognized by PipMaker were usually small or coded for untranslated exons. We also observed 136 regions of significant conservation that were not previously assigned to any of the four gene sequences (Fig. 2). Subsequent BLAST searches of these conserved regions against dbEST identified one additional 3′ exon of *KIAA0603* and a novel coding exon of *UCHL3* represented rarely in dbEST. An additional four conserved regions corresponded to rare, single-exon ESTs that overlapped with known *LMO7* exons that may represent novel spliced products or intermediates. The remaining 130 conserved regions identified by PipMaker were not represented in the nonredundant or EST databases. These were distributed throughout the region, including the gene desert that is distal to *LMO7* (Fig. 2). These nonexonic conserved regions did not overlap with CpG islands, and their base composition was indistinguishable from neighboring DNA. They may represent regulatory elements or possess unknown but conserved functions.

To verify these findings, we referenced the University of California at Santa Cruz (UCSC) and National Center for Biotechnology Information (NCBI) human genome maps. According to UCSC, the syntenic human locus also possesses a CpG-rich core containing the four genes we identified, but no additional genes were predicted in the regions 1.1 Mb proximal or 1.2 Mb distal to this region. There also were no CpG islands either 0.7 Mb proximal or 1.1 Mb distal to the region. According to the NCBI human map, no genes were

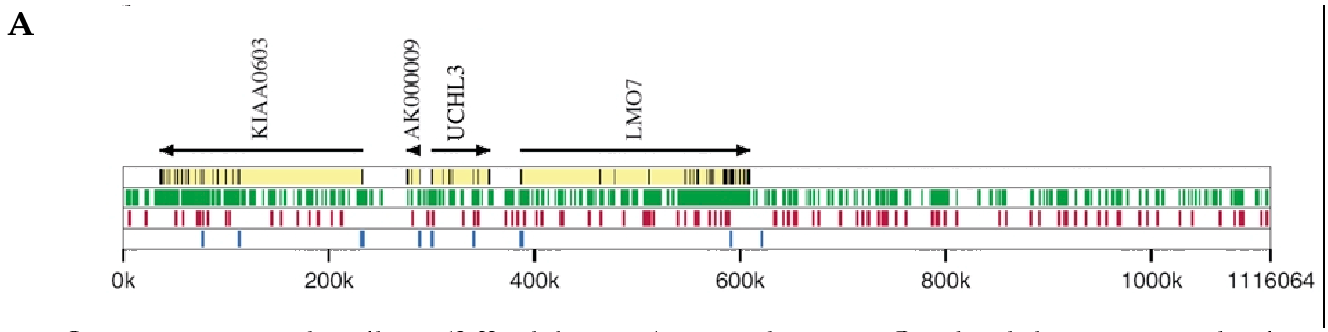


FIG. 2. Comparative sequence analysis of human 13q22 with the mouse Acrg minimal region using PipMaker. The human sequence is the reference sequence, and black horizontal lines correspond to regions of sequence conservation. (A) The percent identity plot (Pip) overview. The first line depicts exons (black) and introns (yellow); the second line depicts moderately conserved elements (green); the third line depicts strongly conserved elements (red); and CpG islands (blue) are shown on the fourth line. (B) (see supplementary data) The PIP. Exons (blue), introns (yellow), strongly conserved elements (red), and CpG islands (green) are depicted along the 1.1-Mb aligned sequence. Two colors are used where features overlap.

predicted either 1.7 Mb proximal or 2.4 Mb distal to the cluster of four genes that we identified. Both maps indicate that the Acrg gene cluster is flanked by "gene deserts," regions over 500 kb that are devoid of genes. Such regions are predicted to compose up to 20% of the human genome [1]. Therefore, although we do not know the precise location of the proximal and distal breakpoints of the Acrg minimal region relative to the human map, the genes that we identified are the only candidates within over 2.8 Mb on the human map, which should encompass the entire Acrg minimal region.

Although UCSC and NCBI cited the use of gene prediction analysis in the region, we further analyzed the human sequence using GeneMachine [20]. GeneMachine automatically runs the gene prediction programs GenScan [21] and HMMGene [22] and the exon prediction program MZEF [23], as well as performing BLAST [24] searches against nucleotide, EST, and protein databases.

The gene and exon prediction programs identified many putative exons in the region that we claim is part of a gene desert. In the final 500 kb of our sequence (positions above 620 kb), GenScan predicted 16 genes with a total of 55 exons, when interspersed repeats in the human sequence were not masked, and MZEF predicted 86 exons in the same (unmasked) region. However, there were no cases in which a single segment simultaneously was predicted to be an exon by at least one exon/gene prediction program, had a reliable BLAST hit to at least one database, and showed strong interspecies conservation. Indeed in only six cases did two of the three sources of evidence point to the same segment, and in each case the evidence was weak. Five had only one BLAST hit and a low-probability exon prediction; one showed interspecies conservation and a low-probability exon prediction. The six segments were scattered throughout the gene desert.

The most compelling case for an additional gene was provided by the EST AI825938, which had a spliced (four-exon) alignment to the human sequence just downstream of LM07, and where one exon of the EST (positions 629,335–629,414) was predicted to be an exon by HMMGene and MZEF (but not GenScan). However, there are no additional database matches to that segment, and the absence of human-mouse alignments at that position appears not to be caused by missing data (the human BAC is finished and the mouse sequence is in the interior of a long contig). There also is no predicted open reading frame. Thus, we found the evidence unconvincing. In any case, this potential gene occurs at the extreme end of the region that we believe to be free of protein-coding genes, leaving a 490-kb region where evidence for the presence of a gene is even weaker.

Analysis of the Acrg Gene Cluster

The structure of the CpG-rich gene cluster is highly conserved between human and mouse. In both species, the relative order of the genes and their orientation of transcription is the same; *KIAA0603* and *AK000009* are transcribed in the opposite direction to *UCHL3* and *LMO7* (Fig. 2). Analysis of base composition within the cluster indicates conserved CpG islands at the 5′ ends of all four genes. The coding sequences of these genes are highly conserved, reaching on average 87% nucleotide identity.

Three of the four genes within the Acrg cluster contain conserved domains in their coding regions that may predict their corresponding functions. Human and mouse *KIAA0603* contain a TBC domain that has been found in the human *USP6* (formerly *Tre2*) oncogene and the yeast regulators of cell-cycle *BUB2* and *CDC16* [25]. *USP6* encodes a de-ubiquitinating enzyme, whereas the activities of *BUB2* and *CDC16* remain unknown [26]. Human *UCHL3* encodes a de-ubiquitinating enzyme [27]; because mouse *Uchl3* is 97% identical with no amino acid changes in catalytic residues, it is likely to perform the same function. Human and mouse *LMO7* contain three separate protein–protein interaction domains. At its C terminus, *LMO7* contains a LIM domain (Lin-11, Isl-1, Mec-3) that is found in various proteins including homeodomain transcription factors and kinases, many of which regulate cell morphology and growth [28]. A subset of *LMO7* splice forms contains an F-box that has been shown to confer substrate specificity to a class of ubiquitin ligase complexes [29]. The

doi:10.1006/geno.2002.6692, available online at http://www.idealibrary.com on IDEAL

FBX20 cDNA is shown on the NCBI and UCSC maps as a separate gene upstream of *LMO7*, but from northern analysis and EST walking, we have confirmed that *FBX20* is part of *LMO7* (data not shown). *LMO7* also contains a PDZ (PSD-95, Dlg, Z0-1) domain that is found in various unrelated proteins, many of which associate with the cytoskeleton [30]. Finally, the *AK000009* sequence is novel.

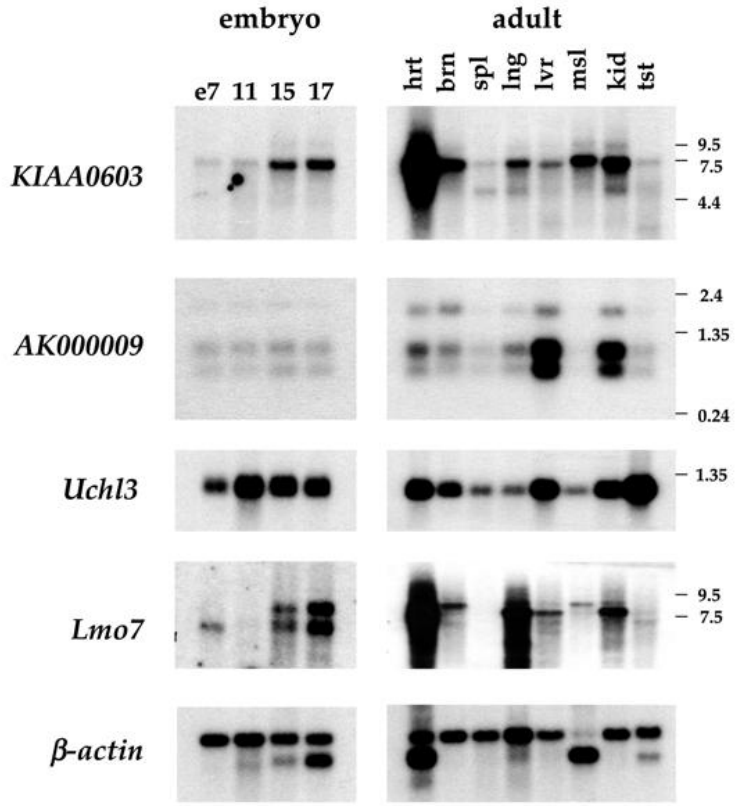
TABLE 2: Comparison of intronic versus gene desert sequence

	GC *Alu*	MIR	L1	L2	LTR	DNA	cons	vcons
Intron	37.8	9.6	1.8	10.9	4.2	4.2	3.1	54.7 2.0
Desert	36.4	3.8	1.7	25.2	4.4	8.7	2.4	45.1 3.6

Percentages of G or C nucleotides and several classes of interspersed repeats in the intronic and putative gene-desert portions of the human sequence plus percentages of nonrepetitive sequence that was conserved (aligned by PipMaker to the mouse sequence) or very conserved (included in a gap-free segment of at least 100 bp that aligns with at least 70% nucleotide identity). L1 and L2, LINE1 and LINE2 elements; cons, conserved; vcons, very conserved.

Analysis of the Acrg Noncoding Sequence

Several authors have observed that comparisons between human and mouse genomic sequences are effective at identifying protein-coding regions [31,32]. We sought to determine if patterns of sequence conservation differ between introns and gene deserts to help gene-modeling efforts in low-GC regions. The human genomic region analyzed in this article has a relatively low G+C level (37.78%). Human/mouse comparative analyses have been performed in regions of comparable G+C content, namely SNCA (α-synuclein) at 36.33% [33], CFTR (cystic fibrosis transmembrane conductance regulator) at 37.35% [34], and FHIT (fragile histidine triad) at 38.39% [35]. However, this is our first opportunity to observe patterns of sequence conservation in what may be part of a gene desert.



We sought to determine the ways in which sequence in gene deserts differs from intronic sequence of low G+C content. We began by extracting the putative intron sequences from the four human genes, omitting 100 bp on either side of each exon in hopes of excluding most elements that modulate splicing. This provided 462,536 bp of intronic sequence. The final 462,536 bp of the human sequence was used as our sample of putative desert data.

Intraspecies properties that we determined included GC level and the density of various classes of interspersed repeats. The intronic regions contain a two- to threefold higher density of *Alu* elements and a two- to threefold lower density of LINE1 and LTR elements (Table 2). An earlier study [36] observed that in a larger sample of AT-rich regions, *Alu* elements were 1.5 times more common in introns than in nontranscribed regions.

To compare interspecies conservation, we analyzed alignments generated by PipMaker, in which interspersed repeats were excluded. A higher percentage of the intronic sequence than of the desert sequence (54.7% versus 45.1%) was aligned by PipMaker to mouse contigs, suggesting that overall nucleotide divergence is slightly higher in deserts. On the other hand, for very strongly aligning segments, defined somewhat arbitrarily as segments of at least 100 bp that aligned without a gap and with at least 70% nucleotide identity, the density in the desert sequence was about 1.8-fold higher than in the introns. Specifically, in the introns there were 43 such segments totaling 5982 bp (2.0% of the 299,725 nonrepetitive base pairs in introns), whereas in the desert there were 61 such segments totaling 8683 bp (3.6% of 241,492 bp). There was no difference when we compared the nucleotide and CpG content of the conserved elements from the introns and the desert. Further analysis of these conserved elements is required to determine whether they are significant.

FIG. 3. Expression of candidate genes in mouse embryos and adult tissues. mRNAs from the tissues indicated were hybridized to *KIAA0603, AK000009, Uchl3, Lmo7,* and β-actin probes. The order of the lanes from left to right is embryos from 7, 11, 15, and 17 dpc, followed by adult heart, brain, spleen, lung, liver, skeletal muscle, kidney, and testis. Molecular size standards are shown to the right of each panel.

Expression Analysis

To be a candidate underlying the deletion phenotype, a gene must be expressed at or before the time when the *Ednrb*$^{s\text{-}1Acrg}$ developmental defects are manifested. Because *Ednrb*$^{s\text{-}1Acrg}$ homozygotes arrest at embryonic day 8.5, we analyzed expression of each gene from the Acrg minimal region at embryonic days 7, 11, 15, and 17, as well as in various adult mouse tissues (Fig. 3).

KIAA0603 was expressed at embryonic days 7, 11, 15, and 17 as a single transcript of 7.5 kb. This transcript was also found in all adult tissues analyzed, with highest levels found in heart. *AK000009* mRNA was expressed during all stages of development as transcripts of approximately 0.6, 1.0, and 2.0 kb. These forms of *AK000009* were also present in all adult tissues analyzed except skeletal muscle, with highest levels of expression in liver and kidney. *Uchl3* mRNA was abundant throughout development as a single transcript of ~ 0.8 kb in length. *Uchl3* expression persisted in adult mice without any tissue specificity, and the highest levels of mRNA were found in testis. However, *Uchl3* knockout mice are viable, thus ruling it out as a candidate for the *Ednrb*$^{s\text{-}1Acrg}$ phenotype [37]. *Lmo7* was expressed at embryonic day 7 as a single transcript of ~ 5 kb. No transcript was detectable at day 11, but expression was detected at days 15 and 17 as two transcripts of 5 and 6 kb. In adult mouse, *Lmo7* mRNA was present in multiple forms in a tissue-specific pattern with transcripts ranging in length from 4.4 to 7.5 kb. It was expressed highly in heart and not expressed in spleen. Therefore, three of the four genes that we identified are candidates for the *Ednrb*$^{s\text{-}1Acrg}$ phenotype.

DISCUSSION

Genomic Organization of the Acrg Minimal Region

We initiated positional cloning of the Acrg minimal region to identify genes required for embryonic development. We found a single cluster of four genes within the 1.4- to 2-Mb Acrg contig that is flanked by a gene desert. In addition, PipMaker identified 130 regions of conservation that were nonexonic. According to studies of orthologous pairs of human and mouse genes, the average coding sequence identity is 85% and noncoding identity is 68% [38]. Because our cross-species hybridization and PipMaker parameters were sensitive to identities within this range, we believe that the search was exhaustive within the mouse sequence. Our use of the GeneMachine program failed to identify additional genes. These results were further supported by the NCBI and UCSC maps, in which the only human genes found within the corresponding Acrg minimal region were the four genes we described.

It is striking that such a large number of highly conserved nonexonic sequences were identified by PipMaker and that none were associated with CpG islands and only one was associated with a low-probability exon prediction. Moreover, there was a 1.8-fold enrichment of these elements in the gene desert relative to the gene cluster. This is in contrast to results obtained from a similar analysis of the imprinted gene cluster at 11p15. Onyango *et al.* [39] found 82 nonexonic regions conserved with mouse, where all but one were clustered around genes and 42% overlapped with CpG islands. Because none of our conserved regions overlap with CpG islands and are not solely associated with genes, they may not define classic regulatory elements. Deletion of these sequences may determine whether they are significant.

Our results at 13q22 are consistent with other reports referring to the gene-poor qualities of this chromosome. On the basis of the number of mapped ESTs versus cytogenetic length, human chromosome 19 was predicted to have a significant excess of genes whereas human chromosome 13 was predicted to have a significant deficit of genes [40]. Only human autosomes 21, 18, and 13 are tolerated as trisomies, a result that suggests low gene density and/or small chromosome size. Additional evidence for gene scarcity on chromosome 13 comes from Gardiner [18], who showed that chromosome 13 is devoid of CpG-rich isochores that are associated with gene-dense regions. This finding predicted that chromosome 13 had the lowest gene density in the genome.

To obtain additional information about gene density on chromosome 13, we calculated the number of known human genes using the UCSC human genome map. We found 141 known genes with an average density of 1 gene per 700 kb, predicting only 3 known genes in our 2-Mb region. In contrast, there are 526 known genes on chromosome 19, which is 30% smaller than chromosome 13 with an average density of one known gene per 130 kb. More recently, Venter *et al.* confirmed that chromosome 13 is the least gene-rich, with only 5 genes per megabase versus 23 genes per megabase on chromosome 19 [1]. Chromosome 13 also contained the largest gene desert of over 3 Mb.

The distal end of mouse chromosome 14 is also predicted to be gene-poor on the basis of functional studies. Analysis of data from the SLT suggested that the piebald locus is gene-poor. Out of seven loci mutagenized, nearly half of all the mutations mapped to the *piebald* locus [41]. This suggested that the region displayed no haploinsufficiency. Consistent with this, mice that are haploid for a deletion that removes 30% of chromosome 14 are viable with no obvious haploinsufficient phenotypes [42].

Candidates for Genes Required for Embryonic Development

Based on their map position within the Acrg minimal region and their temporal expression during mouse development, *KIAA0603, AK000009*, and *Lmo7* are candidates for the *Ednrb*$^{s\text{-}1Acrg}$ embryonic lethality. *KIAA0603* contains a TBC domain that has been found in the human *USP6* oncogene, whose product has de-ubiquitinating enzymatic activity [25]. Mutations in several ubiquitin pathway enzymes have disrupted embryonic development [43–45]. Because *AK000009* is a novel gene, further analysis is required to understand its role in development.

Lmo7 is a strong candidate for the Acrg deletion phenotype. It is expressed throughout embryogenesis and in multiple adult tissues. LMO7 contains a PDZ and LIM domain, both of which mediate protein–protein interactions [28,30]. There are at least six proteins in addition to LMO7 that contain an N-terminal PDZ and a C-terminal LIM domain. The LIM domains of several of these proteins were shown to interact with various kinases [46,47], whereas PDZ domains often associate with the cytoskeleton [48,49]. LIM/PDZ-containing proteins are likely to have important roles in signal transduction, cell shape changes, motility, and cell adhesion, all of which are essential for normal embryogenesis. As an example, mice homozygous for a mutation in the PDZ domain-containing protein Afidin showed developmental defects during and after gastrulation, including impaired migration of mesoderm similar to Acrg-deletion embryos [50]. It is interesting to note that some splice forms of LMO7 also contain an N-terminal F-box, which is yet another protein interaction domain that was shown to recruit phosphorylated substrates to the SCF ubiquitin–ligase complexes [51]. Through the F-box, LMO7 may recruit LIM and PDZ domain-binding proteins for degradation.

It is striking that three of the genes within the cluster possess conserved motifs linked to the ubiquitin pathway. It will be interesting to determine whether these genes are coregulated, and whether they affect the same processes.

Supplementary data for this article are available on IDEAL (http://www.idealibrary.com).

Materials and Methods

Construction of a contig of the Ednrb$^{s\text{-}1Acrg}$ minimal region. *D14Mit* markers *176, 105, 8, 145, 164,* and *38,* which map to the Acrg minimal region, were used to screen a 129/SvJ mouse P1 library (Incyte Genomics). Using the P1 vector SP6 and T7 promoters, RNA probes to P1 insert ends were synthesized and hybridized to the contig to determine orientation and overlap among P1 clones. RNA end probes from P1s 176, 8, 164, and 145 defined gaps in the contig and were used to screen a 129/SvJ mouse BAC library (Research Genetics). BAC insert end probes were generated to determine orientation and overlap to continue chromosome walking. BAC insert end probes consisted of PCR products derived from sequence of "mini-BACs," subclones of the genomic insert ends using *Nsi*I digestion and re-ligation. Following six chromosome walks from BACs 1, 15, 23, 42, 50, and 51, two gaps remained that were spanned by YAC clones mapped to the region by the Whitehead Institute/MIT Center for Genome Research and by Metallinos [52]. P1, BAC, and YAC clone sizes were determined by pulse-field gel electrophoresis, and overlap among clones to estimate contig size was determined by Southern blot hybridization. To obtain the sequence of the Acrg minimal region, the RPCI-23 (C57BL/6J) BAC library (BACPAC Resources) was screened using 129/SvJ BAC insert end probes. Draft sequence with 5.6× coverage was subsequently generated by W. R. McCombie's group at Cold Spring Harbor Laboratory and deposited into the HTGS database of GenBank. Accession numbers of the BAC sequences from proximal to distal are AC074357, AC074207, AC074304, AC079638, AC083913, AC074211, and AC080021.

Southern and northern blot analysis. BAC DNA was extracted according to supplier's instructions. Digested BAC DNA was separated in Tris-borate-EDTA (TBE) gels and transferred to Hybond N+ membranes. Southern blots were hybridized in Church buffer [53] at 65°C and washed in 0.1× SSC/0.1% SDS at 23°C and 65°C. Human cDNA clones corresponding to mapped ESTs were obtained from Research Genetics. For cross-species hybridization, radiolabeled human cDNA probes (Table 1) were hybridized to mouse BAC blots in Church buffer at 65°C and washed at low stringency in 1× SSC/0.1% SDS at 23°C and 50°C. Northern blots (Clontech) were hybridized according to manufacturer's instructions. Radiolabeled probes were synthesized from corresponding mouse cDNA clones: *KIAA0603* (AW988132), *AK000009* (5′ fragment of AA915052), *Uchl3* (kocDNA), *Lmo7*(G5), and β-actin (Clontech).

Sequence analysis. Genemap 97–99 and BLAST (ncbi.nlm.nih.gov) were used for identification of ESTs corresponding to the Acrg minimal region. PipMaker (bio.cse.psu.edu) was used to identify homology between syntenic mouse and human genomic sequences. The UCSC human genome site (genome.ucsc.edu) was used for identification of syntenic human ESTs as well as for base composition analysis. MacVector was used for base composition analysis of mouse genomic sequence. Gene prediction analysis was performed using GeneMachine [20].

Acknowledgments

We thank W. Richard McCombie's group at Cold Spring Harbor Laboratory for generating draft sequence of our RPCI-23 mouse BAC contig; Izabela Makalowska at Pennsylvania State University for performing the GeneMachine analysis; and Goga Vukmirovic at Princeton University for assistance with CpG island mapping. L.J.K. was supported by an NRSA award from the National Institutes of Health. E.S. is supported by an National Institutes of Health training grant. W.M. was supported by grant HG02238 from the National Human Genome Research Institute. S.M.T. is an Investigator of the Howard Hughes Medical Institute.

RECEIVED FOR PUBLICATION AUGUST 8;
ACCEPTED DECEMBER 3, 2001.

References

1. Venter, J. C., *et al.* (2001). The sequence of the human genome. *Science* **291:** 1304–1351.
2. Lander, E. S., *et al.* (2001). Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. *Nature* **409:** 860–921.
3. Rinchick, E. M., and Russell, L. B. (1990). Germ-line deletion mutations in the mouse: tools for intensive functional and physical mapping of regions of the mammalian genome. In *Genome Analysis,* Vol. 1: *Genetic and Physical Mapping* (K. Davies and S. M. Tilghman, Eds.), pp. 121–158. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
4. Schimenti, J., and Bucan, M. (1998). Functional genomics in the mouse: phenotype-based mutagenesis screens. *Genome Res.* **8:** 698–710.
5. Ramirez-Solis, R., Liu, P., and Bradley, A. (1995). Chromosome engineering in mice. *Nature* **378:** 720–724.
6. Thomas, J. W., LaMantia, C., and Magnuson, T. (1998). X-ray-induced mutations in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **95:** 1114–1119.
7. You, Y., Browning, V. L., and Schimenti, J. C. (1997). Generation of radiation-induced deletion complexes in the mouse genome using embryonic stem cells. *Methods* **13:** 409–421.
8. Russell, W. L. (1951). X-ray induced mutations in mice. *Cold Spring Harbor Symp. Quant. Biol.* **16:** 327–336.
9. Hosoda, K., *et al.* (1994). Targeted and natural (piebald-lethal) mutations of endothelin-B receptor gene produce megacolon associated with spotted coat color in mice. *Cell* **79:** 1267–1276.
10. Shin, M. K., Levorse, J. M., Ingram, R. S., and Tilghman, S. M. (1999). The temporal requirement for endothelin receptor-B signalling during neural crest development. *Nature* **402:** 496–501.
11. Lane, P. W. (1966). Association of megacolon with two recessive spotting genes in the mouse. *J. Hered.* **57:** 29–31.
12. Mayer, T. C. (1965). The development of *piebald* spotting in mice. *Dev. Biol.* **11:** 319–334.
13. Metallinos, D. L., *et al.* (1994). Fine structure mapping and deletion analysis of the murine *piebald* locus. *Genetics* **136:** 217–223.
14. O'Brien, T. P., Metallinos, D. L., Chen, H., Shin, M. K., and Tilghman, S. M. (1996). Complementation mapping of skeletal and central nervous system abnormalities in mice of the *piebald* deletion complex. *Genetics* **143:** 447–461.
15. Welsh, I. C., and O'Brien, T. P. (2000). Loss of late primitive streak mesoderm and interruption of left-right morphogenesis in the Ednrb(s-1Acrg) mutant mouse. *Dev. Biol.* **225:** 151–168.
16. Copp, A. J. (1995). Death before birth: clues from gene knockouts and mutations. *Trends Genet.* **11:** 87–93.
17. Cross, S. H., and Bird, A. P. (1995). CpG islands and genes. *Curr. Opin. Genet. Dev.* **5:** 309–314.
18. Gardiner, K. (1996). Base composition and gene distribution: critical patterns in mam-

malian genome organization. *Trends Genet*. **12:** 519–524.
19. Schwartz, S., *et al*. (2000). PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res*. **10:** 577–586.
20. Makalowska, I., Ryan, J. F., and Baxevanis, A. D. (2001). GeneMachine: gene prediction and sequence annotation. *Bioinformatics* **17:** 843–844.
21. Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*. **268:** 78–94.
22. Krogh, A. (2000). Using database matches with for HMMGene for automated gene detection in *Drosophila*. *Genome Res*. **10:** 523–528.
23. Zhang, M. Q. (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci. USA* **94:** 565–568.
24. Altschul, S. F., *et al*. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25:** 3389–3402.
25. Richardson, P. M., and Zon, L. I. (1995). Molecular cloning of a cDNA with a novel domain present in the tre-2 oncogene and the yeast cell cycle regulators BUB2 and cdc16. *Oncogene* **11:** 1139–1148.
26. Papa, F. R., and Hochstrasser, M. (1993). The yeast DOA4 gene encodes a deubiquitinating enzyme related to a product of the human tre-2 oncogene. *Nature* **366:** 313–319.
27. Wilkinson, K. D., *et al*. (1989). The neuron-specific protein PGP 9.5 is a ubiquitin carboxyl-terminal hydrolase. *Science* **246:** 670–673.
28. Bach, I. (2000). The LIM domain: regulation by association. *Mech. Dev*. **91:** 5–17.
29. Patton, E. E., Willems, A. R., and Tyers, M. (1998). Combinatorial control in ubiquitin-dependent proteolysis: don't Skp the F-box hypothesis. *Trends Genet*. **14:** 236–243.
30. Fanning, A. S., and Anderson, J. M. (1999). PDZ domains: fundamental building blocks in the organization of protein complexes at the plasma membrane. *J. Clin. Invest*. **103:** 767–772.
31. Jang, W., *et al*. (1999). Comparative sequence of human and mouse BAC clones from the mnd2 region of chromosome 2p13. *Genome Res*. **9:** 53–61.
32. Ansari-Lari, M. A., *et al*. (1998). Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. *Genome Res*. **8:** 29–40.
33. Touchman, J. W., *et al*. (2001). Human and mouse α-synuclein genes: comparative genomic sequence analysis and identification of a novel gene regulatory element. *Genome Res*. **11:** 78–86.
34. Ellsworth, R. E., *et al*. (2000). Comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane conductance regulator genes. *Proc. Natl. Acad. Sci. USA* **97:** 1172–1177.
35. Shiraishi, T., *et al*. (2001). Sequence conservation at human and mouse orthologous common fragile regions, FRA3B/FHIT and Fra14A2/Fhit. *Proc. Natl. Acad. Sci. USA* **98:** 5722–5727.
36. Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev*. **9:** 657–663.
37. Kurihara, L. J., Semenova, E., Levorse, J. M., and Tilghman, S. M. (2000). Expression and functional analysis of Uch-L3 during mouse development. *Mol. Cell Biol*. **20:** 2498–2504.
38. Makalowski, W., Zhang, J., and Boguski, M. S. (1996). Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res*. **6:** 846–857.
39. Onyango, P., *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.
40. Schuler, G. D., *et al*. (1996). A gene map of the human genome. *Science* **274:** 540–546.
41. Searle, A. G. (1974). *Mutation Induction in Mice*. Academic Press, New York.
42. Cattanach, B. M., Burtenshaw, M. D., Rasberry, C., and Evans, E. P. (1993). Large deletions and other gross forms of chromosome imbalance compatible with viability and fertility in the mouse. *Nat. Genet*. **3:** 56–61.
43. Harbers, K., *et al*. (1996). Provirus integration into a gene encoding a ubiquitin-conjugating enzyme results in a placental defect and embryonic lethality. *Proc. Natl. Acad. Sci. USA* **93:** 12412–12417.
44. Huang, Y., Baker, R. T., and Fischer-Vize, J. A. (1995). Control of cell fate by a deubiquitinating enzyme encoded by the *fat facets* gene. *Science* **270:** 1828–1831.
45. Zhen, M., Schein, J. E., Baillie, D. L., and Candido, E. P. (1996). An essential ubiquitin-conjugating enzyme with tissue and developmental specificity in the nematode *Caenorhabditis elegans*. *EMBO J*. **15:** 3229–3237.
46. Durick, K., Wu, R. Y., Gill, G. N., and Taylor, S. S. (1996). Mitogenic signaling by Ret/ptc2 requires association with enigma via a LIM domain. *J. Biol. Chem*. **271:** 12691–12694.
47. Kuroda, S., *et al*. (1996). Protein–protein interaction of zinc finger LIM domains with protein kinase C. *J. Biol. Chem*. **271:** 31029–31032.
48. Guy, P. M., Kenny, D. A., and Gill, G. N. (1999). The PDZ domain of the LIM protein enigma binds to β-tropomyosin. *Mol. Biol. Cell* **10:** 1973–1984.
49. Vallenius, T., Luukko, K., and Makela, T. P. (2000). CLP-36 PDZ-LIM protein associates with nonmuscle α-actinin-1 and α-actinin-4. *J. Biol. Chem*. **275:** 11100–11105.
50. Ikeda, W., *et al*. (1999). Afadin: A key molecule essential for structural organization of cell–cell junctions of polarized epithelia during embryogenesis. *J. Cell Biol*. **146:** 1117–1132.
51. Skowyra, D., Craig, K. L., Tyers, M., Elledge, S. J., and Harper, J. W. (1997). F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin–ligase complex . *Cell* **91:** 209–219.
52. Metallinos, D. L. (1994). *Molecular and genetic characterization of the piebald locus in mice*. p. 151. Thesis, Princeton Univ., Princeton, NJ.
53. Church, G. M., and Gilbert, W. (1984). Genomic sequencing. *Proc. Natl. Acad. Sci. USA* **81:** 1991–1995.