CMLS, Cell. Mol. Life Sci. 61 (2004) 488–496
1420-682X/04/040488-09
DOI 10.1007/s00018-003-3383-2

CMLS Cellular and Molecular Life Sciences

Research Article

Recent evolutionary origin within the primate lineage of two pseudogenes with similarity to members of the transforming growth factor-β superfamily

S. Eketjäll, H. Jörnvall, P. Lönnerberg[‡], S. Kobayashi[#] and C. F. Ibáñez*

Division of Molecular Neurobiology, Department of Neuroscience, Karolinska Institute, 17177 Stockholm (Sweden), e-mail: carlos.ibanez@neuro.ki.se

Received 9 October 2003; received after revision 13 November 2003; accepted 2 December 2003

Abstract. Using a search engine called Motifer, we searched the public database of the human genome for genes matching a consensus pattern of cysteine residues derived from members of the transforming growth factor-beta (TGF-β) superfamily. We identified two genes (named *MDF451* and *MDF628*) that display sequence similarity to members of the TGF-β superfamily in the arrangement of six conserved cysteine residues. Phylogenetic analyses revealed that *MDF451* and *MDF628* constitute a distinct subgroup within the TGF-β superfamily, distantly related to the GDNF subfamily of ligands. Both genes could be identified in several primate species in addition to human, including chimpanzee, gorilla, guereza, and green and gray monkey, but not in rodents or other non-primate mammals, and appear not to be present in the genomes of mouse, rat or zebrafish. RNAs for *MDF451* and *MDF628* were expressed at low levels within distinct regions of the human central nervous system, including adult cerebellum, adult spinal cord and fetal brain. Despite expression at the RNA level, both genes presented a transcribed upstream stop codon that would prevent translation of the TGF-β-like reading frame. The coding potential of alternative reading frames was not immediately apparent. The two genes may represent TGF-β-like pseudogenes that have recently appeared in evolution in a common ancestor of the primate lineage by duplication from a GDNF/TGF-β-like ancestral gene.

Key words. Human genome; mammalian evolution; nervous system; growth factor.

The transforming growth factor-β (TGF-β) superfamily constitutes the largest known group of polypeptide growth factors, and includes the TGF-βs, activins, bone morphogenetic proteins (BMPs), growth and differentiation factors (GDFs) and neurotrophic factors of the GDNF ligand subfamily. Members of the TGF-β superfamily exhibit a large range of biological activities, including the regulation of cell proliferation, lineage determination, differentiation, migration, adhesion and apoptosis during development, homeostasis and repair in practically all tissues, from flies to humans.

TGF-β superfamily members are characterized by a common three-dimensional fold containing a cysteine-knot [1]. The pattern of cysteine residues is highly conserved in the primary sequence of different TGF-β superfamily members. So powerful is the structural constraint imposed by the conserved spacing of cysteine residues that even very distant members of the superfamily, such as TGF-β2 and GDNF, which are completely divergent in the sequence segments between the cys-

* Corresponding author.
Present addresses: ‡ Global Genomics, Karolinska Science Park, Stockholm (Sweden)
Department of Biochemistry, College of Pharmacy, Nihon University, Chiba (Japan)

teines [2], have an almost indistinguishable three-dimensional structure [3, 4].

Most TGF-β ligands signal through a complex of two transmembrane receptor serine-threonine kinases belonging to two distinct subfamilies: the type I receptors of approximately 55 kDa, and the type II receptors of approximately 70 kDa. Both receptors cooperate in ligand binding: type II receptors phosphorylate type I receptors, and the latter activate a member of the Smad family of signal transducers, which then translocates to the nucleus where it participates in various DNA-binding complexes [5–7]. An exception to the scheme presented above is the GDNF ligand subfamily, which utilizes a completely unrelated receptor system, composed of a GPI-anchor ligand-binding subunit, the GFRα receptor, and a signaling subunit, the receptor tyrosine kinase c-Ret [8].

We have taken advantage of the high conservation of the cysteine pattern of TGF-β superfamily members to identify novel members of this family using bioinformatic techniques. We report here the discovery of two pseudogenes displaying sequence similarity to members of the TGF-β gene superfamily specifically expressed in the human nervous system.

Materials and methods

Bioinformatics and motifer

Relevant sequences were obtained from GenBank and EMBL databases [9, 10]. Sequence alignments and further processing were done using programs from the Genetics computer Group of the University of Wisconsin (GCG-package) and CLUSTALX [11, 12]. Motifer was used as previously described [13]. BLAST [14] was run using default parameters.

DNA preparation and PCR analysis

Genomic DNA was purified from whole blood using QIAamp (Qiagen), according to the manufacturer's instructions. Blood from several species was obtained from local zoos. Specific primers derived from the human sequence of *MDF451* and *MDF628* were designed as follows: MDF451 upper 5′-ACC ATT GGG AGT TGG GGC TT-3′ and *MDF451* lower 5′-CCA TCA AAT GCA TCG GCA AG-3′, amplifying a 246-bp fragment, and *MDF628* upper 5′-AAG GGC CTG GGG GAT GTT GTC-3′ and *MDF628* lower 5′-TCA CCA CCA GGC CAT GCA GG-3′, amplifying a 237-bp fragment. Fragments were amplified using Taq polymerase (Promega) under the following PCR conditions: 30 s at 90 °C, 1 min at 50 °C, 1 min at 72 °C for 40 cycles, and subsequently subcloned into the pCRII-TOPO vector (Invitrogen). Degenerated PCR was performed as described above, using a different primer combination within the 246-bp and 237-bp fragments.

RNase protection assay and RT-PCR

The subcloned PCR fragments were linearized with either *Bam*HI (*MDF451*) or with *Xho*I (*MDF628*) and transcribed with T7 RNA polymerase or Sp6 RNA polymerase, respectively, to generate RNA probes. Ten micrograms of total human RNA (Clontech) was hybridized to [α-^{32}P]CTP-labeled probes according to the manufacturer's instructions (Ambion). Protected bands were visualized using a STORM Phosphorimager and ImageQuant software (Molecular Dynamics). Molecular weight markers were obtained by radiolabeling a commercial DNA ladder. For analysis of gene expression by RT-PCR, cDNA was synthesized from total human RNA using ProSTAR from Stratagene. PCR reactions were performed using the following primers: *MDF451*, forward 5′-CTA AGT CCA TTC TTG CTA TCC-3′ and reverse 5′-CCA TCA AAT GCA TCG GCA AG-3′; *MDF628*, forward 5′-CCC TGG GTG AGT TGC AAC TTG-3′ and reverse 5′-GCT CCC TGC CTG GGA-3′.

Library screening

Human phage cDNA libraries were purchased from Clontech and Stratagene. Hybridization screening of library plaques was performed on Hybond-N filters (Amersham Pharmacia Biotech) at 65 °C using [α-^{32}P]dCTP probe labeled by PCR, and detected either by Phosphorimager or BiomaxMS X-ray film (Kodak).

Results

Identification of novel genes displaying sequence similarity to members of the TGF-β gene superfamily in the human genome

To identify new members of the TGF-β superfamily, we utilized a search engine called Motifer [13], a program that allows the unbiased search of sequence databases based on the pattern of any residue or combinations of residues and the distance between them from a prototypic query sequence. The pattern used for database searches conformed to a formula derived from the primary sequences of members of the TGF-β superfamily as follows: Cys-(18–24)-Cys-(3)-Cys-(18–42)-Cys-Cys-(20–35)-Cys-(1)-Cys, where the numbers in parentheses indicate the distances separating the corresponding Cys residues.

We searched the publicly available database of the human genome using Motifer and the amino acid pattern shown above. For this search, we did not exclude sequences of known members of the TGF-β superfamily. The first 300 positions of the output file of this search corresponded to previously described genes of the TGF-β superfamily. Immediately below in the list, two new sequences were identified displaying conservation of all but the first Cys residue of the query. The two sequences retrieved from

the human genome database were named Motifer-derived factors (MDF) 451 and 628, respectively, and their sequences are shown in figure 1. Neither *MDF451* nor *MDF628* could be identified applying other algorithms, including BLAST, to the same database. *MDF451* is present on human chromosome 9, while *MDF628* is on chromosome 2.

The absence of the first Cys of the query in the two retrieved sequences was intriguing, but could have been explained by the presence of an intron somewhere upstream of the first Cys present in the new sequences (but see below). It should be noted in this respect that the position of introns is not perfectly conserved among different members of the TGF-β superfamily. In the case of GDNF and BMP-2, for example, an intron is located upstream of the pro-hormone proteolytic cleavage site, so that the complete mature sequence is encoded in a single exon [15]. However, in other TGF-β superfamily members such as BMP-6 and TGF-β1, the mature sequences are interrupted by three and two introns, respectively [16].

No stop codons could be found within the predicted TGF-β-like coding sequences of either *MDF451* or *MDF628*. However, visual inspection of genomic sequences revealed the presence of in-frame stop codons upstream of the first Cys residue in both *MDF451* and *MDF628*. These did not appear to represent errors in the database since their presence could be confirmed independently in PCR fragments amplified from human DNA (data not shown). In addition, both genes contained a stop codon 1 or 4 residues downstream of the last Cys, respectively. Importantly, this is a feature that is highly conserved in all members of the TGF-β superfamily but that was not present in the query. No other Cys in addition to those conforming to the query pattern could be found in either sequence.

A multiple alignment was constructed using conceptual translations of the TGF-β-like frames of *MDF451* and *MDF628* and a set of amino acid sequences from representative members of different TGF-β subfamilies using CLUSTALX. A portion of this alignment is shown in figure 2. A phylogenetic tree was derived from the multiple alignment and is shown in figure 3. Bootstrap values are indicated in the nodes of the tree, and give an estimate of the probability that the corresponding node represents a true and distinct branch in the tree. Nodes with a low bootstrap value are not well supported by the data and suggest that alternative relationships may also be possible. In this case, both *MDF451* and *MDF628* appeared in a distinct branch of the tree leading to the GDNF ligand subfamily. This branch had a relatively high bootstrap value of 993, indicating that it was well supported by the data, suggesting that *MDF451* and *MDF628* may represent distant relatives of GDNF subfamily ligands.

MDF451

```
CAGATGATAGACAGAGGTGCCACTTTTGGTAAGTTCACACAACTATTTTCAACAACTAAG
Q  M  I  D  R  G  A  T  F  G  K  F  T  Q  L  F  S  T  T  K

TCCATTCTTGCTATCCAATATGCTCCACCGTCCTCAAATTTCTTATAAGCCACCATTGGG
S  I  L  A  I  Q  Y  A  P  P  S  S  N  F  L  *  A  T  I  G

AGTTGGGGCTTCATATCAGAAAATGAGTGGCCATGTCTACGGTTCTGCTTTGGGAAGACC
S  W  G  F  I  S  E  N  E  W  P  C  L  R  F  C  F  G  K  T

ACTCTGGGCAGTGTGGAGGGGGAGCTTGGGAAAGGAGAGGAACATACAGAGGCCAGGAAG
T  L  G  S  V  E  G  E  L  G  K  G  E  E  H  T  E  A  R  K

CTGGAGGGATGCTGTCTCAATAATCCAGGTGGGATATCAATTAAGGTGGAGAACAGGGAG
L  E  G  C  C  L  N  N  P  G  G  I  S  I  K  V  E  N  R  E

AGGACTCATGACTATTCCCAAGTTCAACTCAATGAGACTTGCCGATGCATTTGATGGGAG
R  T  H  D  Y  S  Q  V  Q  L  N  E  T  C  R  C  I  *  W  E
```

MDF628

```
CTTTCCTACAGGGCTGGCCTGGAGATGCGGGGAAGGAGGTGGTCGAGGCAGTCTTCCTTC
L  S  Y  R  A  G  L  E  M  R  G  R  R  W  S  R  Q  S  S  F

TGCAGTGACTCAGGAGCCAGCTCCTAAAAGGGCCTGGGGGATGTTGTGGGGAAGGTCCTG
C  S  D  S  G  A  S  S  *  K  G  L  G  D  V  V  G  K  V  L

AAGGAGGAATGGGGAGGGCCAGGATCCTGCAGGGGCAAGTGCAGTCTGGTTGGATCTCTC
K  E  E  W  G  G  P  G  S  C  R  G  K  C  S  L  V  G  S  L

CAGGCAGGGAGCCTCTTAAAATCCCAGAATTTCCAGGTGGGCATCTGCTGCCTTTCGGCA
Q  A  G  S  L  L  K  S  Q  N  F  Q  V  G  I  C  C  L  S  A

TGGGAGCAGCCATGGCCATCAATGAGGGTAATTTCCTTGTATCATGGGTCCCAGGGACAG
W  E  Q  P  W  P  S  M  R  V  I  S  L  Y  H  G  S  Q  G  Q

TGTCCCTGCATGGCCTGGTGGTGAGGGGGCAGGGCCTTCAAAGTCAGATGACTATATTTG
C  P  C  M  A  W  W  *  G  G  R  A  F  K  V  R  *  L  Y  L
```

Figure 1. Human *MDF451* and *MDF628* genomic DNA sequence and conceptual translation. Sequences identified with Motifer corresponding to the motifs C-3-C-21-C-C-23-C-1-C (for MDF628) and C-3-C-27-C-C-28-C-1-C (for *MDF451*) are underlined. Cysteine residues are indicated in bold and in-frame stop codons by asterisks. Accession numbers of genomic fragments in GenBank containing the *MDF451* and *MDF628* sequences are AC006952 and AC007563, respectively.

```
hMDF451   CLR--FCFG--KTTLG-SVEGELGKGEEHTEARKLEG------CCLNNPGGISIKVENRERTHD---YSQVQ----LNETCRCI---
hMDF628   CRG--KCSLVGSLQAG-----SLLKSQNFQV-----G-----ICCL--SAWEQP-WPSMRVIS------LYHG---SQGQCPCMAWW
hGDNF     CSG--SCDA-AETTYD-KILKNLSRNRRLVSDKV--G----QACCR--PIAFDDDLSFLDDNLV---YHILRK--HSAKRCGCI---
xVg1      CYG--ECPYPLTEILNGS-NHAIL--QTLVHSIE--PEDIPLPCCV--PTKMSPISMLFYDNNDNVVLRHYEN--MAVDECGCR---
mBMP4     CHGDFTCPFPLADHLN-STNHAIV--QTLVNSV---NSSIPKACCV--PTELSAISMLYLDEYDKVVLKNYQEFTMVVEGCGS----
Xnr1      CEG--ACPIPLNETFK-PTNHAYM--KSVVKLYQ--PERVECPLCV--PVKMSPLSMLYYEGDE-VVLRHHQE--MIVEECGCS---
mNodal    CEG--ECPNPVGEEFH-PTNHAYI--QSLLKRYQ--PHRVPSTCCA--PVKTKPLSMLYVDNGR-VLLEHHKD--MIVEECGCL---
mTGFb2    CAG--ACPY-LWSSDT---QHTKVL--SLYNTIN--PEASASPCCVS--QDLEPLTILYYIGNTPKI-EQLSN--MIVKSCKCS---
mGDF9     CKGD--CPRAVRHRYG-SPVHTMVQN-IIYE--KLDP-SVPRPSCV--PGKYSPLSVLTIEPDGSIAYKEYED--MIATRCTCR---
```

Figure 2. Sequence alignment of conceptual translations of human *MDF628* and *MDF451* with homologous segments of human GDNF, *Xenopus* Veg-1 (xVg1), mouse BMP4, *Xenopus* Nodal related-1 (Xnr1), mouse Nodal, mouse TGF-β2 (mTGFb2) and mouse GDF9. Cysteine residues are highlighted. The alignment was made with CLUSTALX.

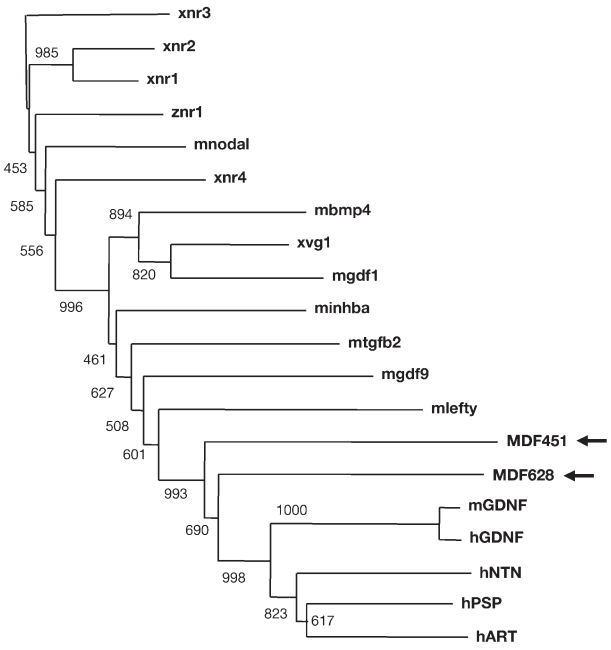

Figure 3. Phylogenetic tree based on amino acid sequences of representative members of the TGF-β superfamily and conceptual translations of human *MDF451* and *MDF628*. Bootstrap values are indicated at each node. The tree was based on an alignment made with CLUSTALX and was calculated using NJPLOT.

Both *MDF451* and *MDF628* are novel genes, not previously described in the GenBank database. Interestingly, no expressed sequence tag (EST) from any species could be found matching either of the two MDF sequences, indicating that both genes may be expressed at very low levels or restricted to a specific developmental stage or tissue.

MDF451 and *MDF628* are present in primates but absent in rodents

Our efforts to identify homologous DNA sequences of *MDF451* and *MDF628* in non-primate mammalian species were unsuccessful. PCR reactions using specific or degenerate primers and genomic DNA as template failed to give any specific product. The non-primate species included in this survey were rat, mouse, dog, pig, mink whale, pygmy sperm whale, bottlenose dolphin, horse, elephant, moose, wallaro, bear, seal, lynx and wolverine. Recently, the mouse genome has been completed by researchers in the public domain [10] and at Celera. An estimated 95% of all mouse genes are represented in this database. Using the human sequences of *MDF451* and *MDF628* as queries, we searched the mouse genome database with the BLAST algorithm using both DNA and protein queries. No sequences displaying significant similarity to either gene could be obtained from these searches. However, matches were obtained to other, presumably more distant, members of the TGF-β superfamily such as activin A. We also searched other partial or complete genomes available in public databases, including those of rat, *Fugu*, zebrafish, *Drosophila* and *Caenorhabditis elegans*, but no MDF-like sequences could be detected. Using the same conditions, sequences similar to human GDNF were found in the genomes of mouse, rat, zebrafish and *Fugu*, but not in *Drosophila* or *C. elegans*, as expected. These results strongly suggest that neither of the two genes is present in rodents or non-primate mammalian genomes, although at this time we cannot rule out that they may be found among portions of these genomes not yet represented in public databases.

Although we have been unable so far to identify sequences similar to *MDF451* and *MDF628* in rodents, PCR amplification from genomic DNA of several primates allowed us to recover these genes in a number of other species including gorilla, chimpanzee, gray monkey, green monkey and guereza. The new sequences identified displayed high similarity to human *MDF451* and *MDF628*, although they were clearly distinct (fig. 4A, B). Multiple sequence alignment and phylogenetic analysis with CLUSTALX showed a similar evolutionary relationship in both genes from different primate species (fig. 4C). In agreement with evolutionary relationships established by other genetic markers [17, 18], human, chimpanzee and gorilla formed a group separated from green and gray monkey and guereza which were more related to each other (fig. 4C).

MDF451 and *MDF628* are predominantly expressed in human nervous tissue

Using specific PCR primers, we amplified DNA fragments corresponding to *MDF451* and *MDF628* from human genomic DNA. Automated sequencing of the PCR products yielded sequences with 100% match to the corresponding genes identified from the genome database. Expression of RNA for *MDF451* and *MDF628* was analyzed in human tissues by RNase protection assay using riboprobes derived from the cloned PCR fragments. From the samples included in this survey, *MDF451* and *MDF628* appeared to be predominantly expressed in human nervous tissue, including fetal and adult brain (fig. 5A). Strong expresion of *MDF628* was detceted in human adult cerebellum (fig. 5A). Weak expression of *MDF451* could also be seen in skeletal muscle, testis and spinal cord, and of *MDF628* in testis and prostate (fig. 5A). These data indicated that both *MDF451* and *MDF628* are expressed genes with putative functions in the human nervous system.

Each riboprobe used in these experiments extended from the first Cys residue to the C-terminal stop codon in each gene and was fully protected, indicating that these sequences correspond to exons in the respective genes. The protected bands seen in RNase protection experiments

A

MDF451

```
Gray monkey   TGTCTACGGTTCTGCTTTGGGAAGATCACTCTGGGCAGTGTGGAGGGGGAGCTTGAGAAAGGAGTGGAACATACAGAGGCCAGGAAGCTGGAGGGATGC
Chimpanzee    TGTCTACGGTTCTGCTTTGGgAAGACCACTCTGGGCAGTGTGgAGGGGGAGCCTGGGAAAGGAGAGGAACATACAGAGGCCAGGAAGCTGGAGGGATGC
Green monkey  TGTCTACGGTTCTGCTTTGGGAAGATCACTCTGGGCAGTGTGGAGGGGGAGCTTGAGAAAGGAGTGGAACATACAGAGGCCAGGAAGCTGGAGGGATGC
Gorilla       TGTCTACGGTTCTGCTTTGGGAAGACCACTCTGGGCAGTGTGgAGGGGGACCTTGGGAAAGGAGAGGAACATACAGAGGCCAGGAAGCTGGAGGGATGC
Guereza       TGTCTACAGTTCTGCTTTGGGAAGATCACTCTGGGCAGTGTGGAGGGGGAGCTTGGGAAAGGAGAGGAACATACAGAGGCTAGGAAGCTGGAGGGATGC
Human         TGTCTACGGTTCTGCTTTGGGAAGACCACTCTGGGCAGTGTGGAGGGGGAGCTTGGGAAAGGAGAGGAACATACAGAGGCCAGGAAGCTGGAGGGATGC

Gray monkey   TGTCTCAATGATCCAGGTGGGATATCAATTAAGGTGGAGAAGAGGGAGAGGACCCATGACTATTCCCAAGTTCAACTCAATGAGACTTGCCGATGCATT
Chimpanzee    TGTCTCAATAATCCAGGTGGGATATCAATTAAGGTGGAGAACAGGGAGAGGACTCATGACTATTCCCAAGTTCAACTCAATGAGACTTGCCGATGCATT
Green monkey  TGTCTCAATGATCCAGGTGGGATATCAATTCAGGTGGAGAAGAGGGAGAGGACCCATGACTATTCCCAAGTTCAACTCAATGAGACTTGCCGATGCATT
Gorilla       TGTCTCAATAATCCAGGTGGGATATCAATTAAGGTGGAGAACAGGGAGAGGACTCATGACTATTCCCAAGTTCAACTCAATGAGACTTGCCGATGCATT
Gueretza      TGTCTCAATGATCCAGGTGGGATATCAATTAAGGTGGAGAAGAGGGAGAGGACCCATGACTATTCCCAAGTTCAACTCAATGAGACTTGCCGATGCATT
Human         TGTCTCAATAATCCAGGTGGGATATCAATTAAGGTGGAGAACAGGGAGAGGACTCATGACTATTCCCAAGTTCAACTCAATGAGACTTGCCGATGCATT
```

MDF628

```
Gray monkey   TGTAGGGGGAAGTGCAGCCTGGTTGGATCTCTCCAGGCAGGAAGCCTCTTAAAATCCCAGAATTTCCAGGTGAGCACCTGCTGCCTT
Chimpanzee    TGCAGGGGCAAGTGCAGTCTGGTTGGATCTCTCCAGGCAGGGAGCCTCTTAAAATCCCAGAATTTCCAGGTGGGCACCTGCTGCCTT
Green monkey  TGCAGGGGAAAGTGCAGCCTGGTTGGATCTCTCCAGGCAGGAAGCCTCTTAAAATCCCAGAATTTCCAGGTgAGCACCTGcTGCCTT
Gorilla       TGCAGGGGCAAGTGCAGTCTGGTTGGATCTCTCCAGGCAGGGAGCCTCTTAAAATCCCAGAATTTCCAGGTGGGCACCTGCTGCCTT
Human         TGCAGGGGCAAGTGCAGTCTGGTTGGATCTCTCCAGGCAGGGAGCCTCTTAAAATCCCAGAATTTCCAGGTGGGCATCTGCTGCCTT

Gray monkey   TCAGCATGGGTGCAGCCATGGCCATCAATGAGGGTAATTTCCTTGTATCGTGGGTCCCAGGGACAGTGTCCCTGTATGGCCTGGTGG
Chimpanzee    TCGGCATGGGAGCAGCCATGGCCATCAATGAGGGTAATTTCCTTGTATCATGGGTCCCAGGGACAGTGTCCCTGCATGGCCTGGTGG
Green monkey  TCAGCATGGgAGCAGCCATGGCCATCAATGAGGGTAATTTCCTTGTATTGTGGGTCCCAGGGACAGTGCCCCTGCATGGCCTGGTGG
Gorilla       TCGGCATGGGAGCAGCCATGGCCATCAATGAGGGTAATTTCCTTGTATCATGGGTCCCAGGGACAGTGTCCCTGCATGGCCTGGTGG
Human         TCGGCATGGGAGCAGCCATGGCCATCAATGAGGGTAATTTCCTTGTATCATGGGTCCCAGGGACAGTGTCCCTGCATGGCCTGGTGG
```

B

MDF451

```
Gray monkey   CLRFCFGKITLGSVEGELEKGVEHTEARKLEGCCLNDPGGISIKVEKRERTHDYSQVQLNETCRCI*
Chimpanzee    CLRFCFGKTTLGSVEGEPGKGEEHTEARKLEGCCLNNPGGISIKVENRERTHDYSQVQLNETCRCI*
Green monkey  CLRFCFGKITLGSVEGELEKGVEHTEARKLEGCCLNDPGGISIQVEKRERTHDYSQVQLNETCRCI*
Gorilla       CLRFCFGKTTLGSVEGDLGKGEEHTEARKLEGCCLNNPGGISIKVENRERTHDYSQVQLNETCRCI*
Guereza       CLQFCFGKITLGSVEGELGKGEEHTEARKLEGCCLNDPGGISIKVEKRERTHDYSQVQLNETCRCI*
Human         CLRFCFGKTTLGSVEGELGKGEEHTEARKLEGCCLNNPGGISIKVENRERTHDYSQVQLNETCRCI*
```

MDF628

```
Gray monkey   CRGKCSLVGSLQAGSLLKSQNFQVSTCCLSAWVQPWPSMRVISLYRGSQGQCPCMAWW*
Chimpanzee    CRGKCSLVGSLQAGSLLKSQNFQVGTCCLSAWEQPWPSMRVISLYHGSQGQCPCMAWW*
Green monkey  CRGKCSLVGSLQAGSLLKSQNFQVSTCCLSAWEQPWPSMRVISLYCGSQGQCPCMAWW*
Gorilla       CRGKCSLVGSLQAGSLLKSQNFQVGTCCLSAWEQPWPSMRVISLYHGSQGQCPCMAWW*
Human         CRGKCSLVGSLQAGSLLKSQNFQVGICCLSAWEQPWPSMRVISLYHGSQGQCPCMAWW*
```

C

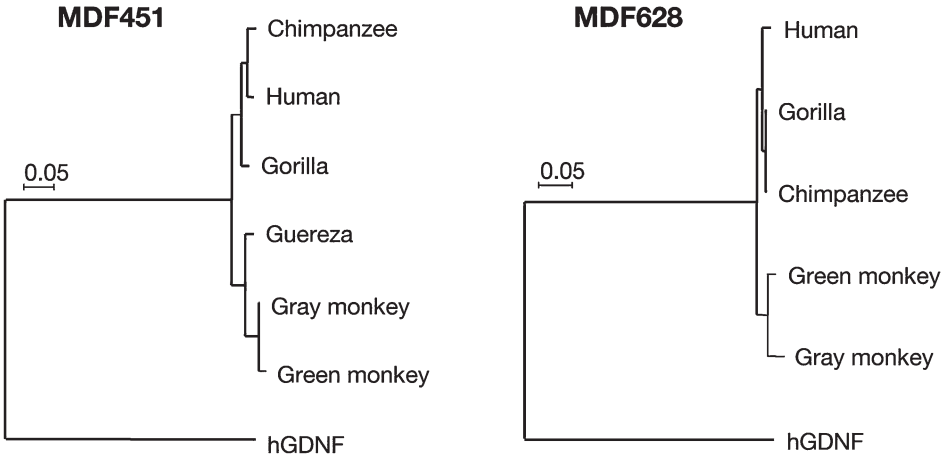

Figure 4. (*A*) DNA sequences of *MDF451* and *MDF628* from different primate species. (*B*) Conceptual amino acid sequences of *MDF451* and *MDF628* from different primate species. (*C*) Phylogenetic relationships of *MDF451* and *MDF628* genes in primate species. Trees were made with CLUSTALX and NJPLOT based on amino acid sequences derived by conceptual translation. Human GDNF (hGDNF) was used as an outgroup in both cases.

were unlikely to represent contaminating genomic DNA because (i) they could still be detected after prolonged treatment of RNA samples with RNase-free DNase and (ii) they were not detected if human genomic DNA was used instead of RNA. Nevertheless, the expression levels of both *MDF451* and *MDF628* appeared to be low, a notion supported by their absence from current EST databases.

We also investigated *MDF451* and *MDF628* by RT-PCR, a less quantitative but more sensitive technique than RNase protection assay. Expression of *MDF451* was detected in fetal brain and in testis (fig. 5B). Using RT-PCR, expression of *MDF628* was detected more broadly and included heart and mammary gland, in addition to cerebellum (fig. 5B).

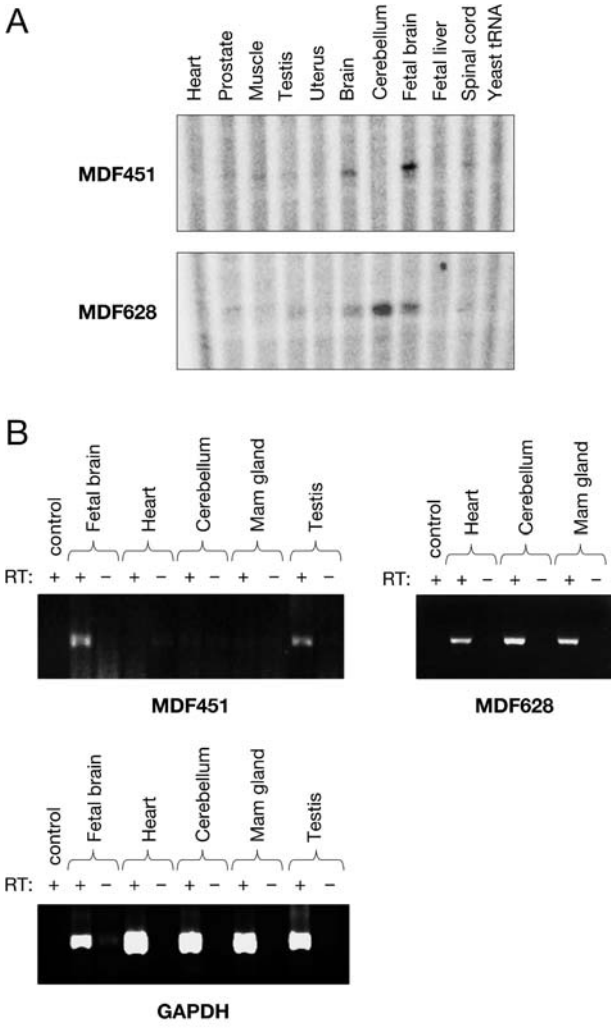

Figure 5. Expression of *MDF451* and *MDF628* in human tissues. (*A*) Total RNA (Clontech) from the indicated tissues and brain regions (adult, unless otherwise indicated) was analyzed for expression by RNase protection analysis using riboprobes extending from the first Cys residue to the stop codon of the TGF-β-like reading frame in both *MDF451* and *MDF628*. (*B*) Total RNA from the indicated tissues was analyzed for expression by RT-PCR using primers specific for *MDF451* and *MDF628*. Lack of signal in parallel control reactions run in the absence of reverse transcriptase (RT) indicated that the products detected were derived from RNA and not from contaminating genomic DNA. Amplification of glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) was used as positive control and indicated comparable amounts of RNA among the different samples.

Molecular analysis of the human *MDF628* locus

Extensive screening of both human cerebellum and fetal brain phage cDNA libraries failed to yield any cDNA corresponding to either of the two genes, again indicating that the expression levels of these genes may be very low. However, during the screening of the human cerebellum library, under stringent conditions, a transcript with similarity to *MDF628* was recovered (fig. 6). This cerebellar transcript (hereby referred to as *628R*, for *MDF628*-related) was ~3.8-kb-long and consisted of two exons of ~60-bp and ~3.7-kb in length, respectively, separated by a ~17.6-kb long intron. The reading frame with similarity to TGF-β superfamily proteins in the second exon of the *628R* gene has multiple stop codons and also lacks several consensus Cys residues (fig. 6). The positions of the 628R exons relative to *MDF628* in human chromosome 2 are shown in figure 6C. The absence of these Cys residues and the presence of in-frame stop codons may explain why this sequence was not detected in the Motifer search. The *628R* gene is located in the vicinity of *MDF628* on human chromosome 2. Closer inspection of this locus revealed a number of ≈200-bp-long repetitive segments in this region, unrelated to the reading frame with similarity to TGF-β superfamily proteins, upstream and downstream of the *628R/MDF628* locus over a region of approximately 500 kb. PCR experiments using genomic DNA from different species and degenerate primers based on the *628R* sequence revealed that, similar to *MDF628*, the *628R* gene also appears to be restricted to primates and absent in rodents (data not shown). Database searches on the mouse genome database using BLAST failed to retrieve any sequences with significant similarity to *628R*, suggesting that this entire locus on human chromosome 2 may be specific to primates.

Coding capacity of *MDF451* and *MDF628* genes

Finally, to shed light on the potential coding capacity of the *MDF451* and *MDF628* genes, we investigated whether the upstream stop codons found in the TGF-β-like reading frames of these sequences formed part of exonic or intronic sequences. To this end, we performed RNase protection assays using probes extending over this region in each of the genes. For *MDF451*, the riboprobe used extended from about 100 nucleotides upstream of the first Cys codon to the end of the TGF-β-like open reading frame. This probe was fully protected by human fetal brain RNA (fig. 7A). For *MDF628*, the riboprobe used extended from about 320 nucleotides upstream of

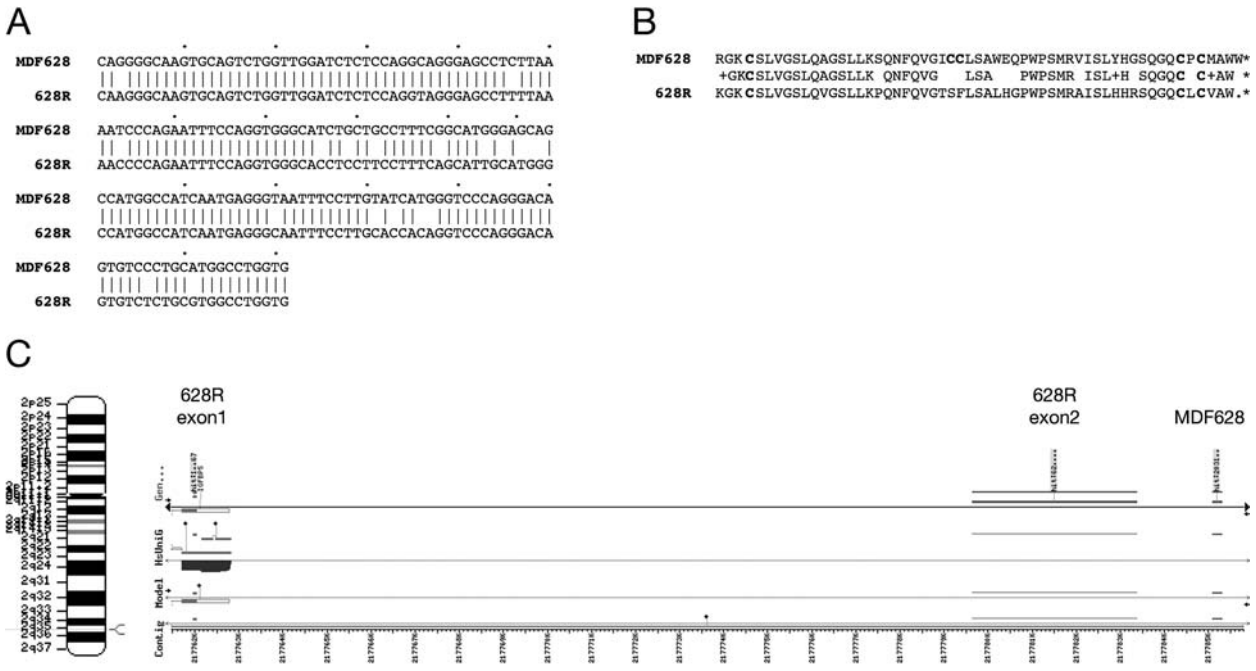

Figure 6. Alignment of *MDF628* and *628R*. Nucleotide sequences (*A*) and conceptual amino acid translations of the TGF-β-like reading frames (B) are shown. Note that *628R* is missing two of the Cys residues in the region with most similarity to TGF-β-like factors. The positions of MDF628 and 628R transcripts on human chromosome 2 are shown (*C*). The diagram was produced with the GeneViewer feature of the NCBI human genome website.

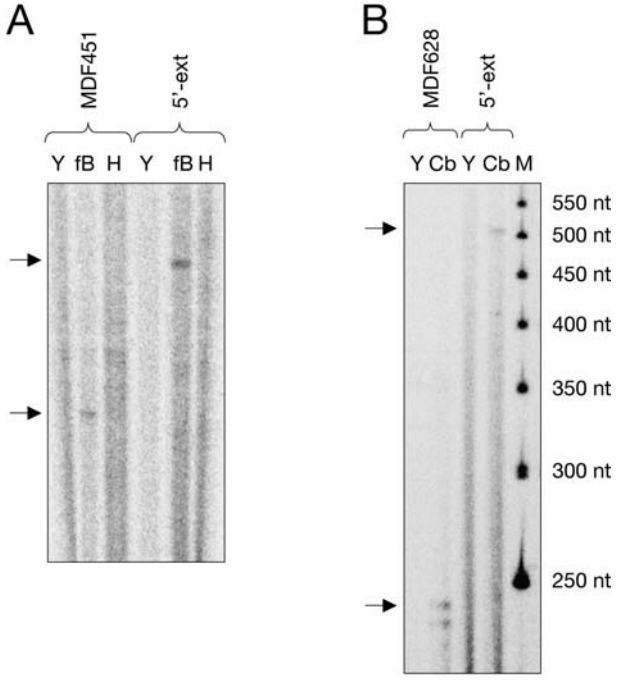

Figure 7. Investigation of the transcription of upstream stop codons in *MDF451* and *MDF628* by RNase protection analysis with extended riboprobes. Riboprobes encompassing the TGF-β-like reading frames of *MDF451* (*A*) and *MDF628* (*B*) as well as extended riboprobes (5′-ext) extending over the upstream stop codons in each gene gave fully protected bands. Y, yeast tRNA control; Cb, human cerebellum; H, human heart; fB, human fetal brain; M, molecular-weight marker.

the first Cys codon to the end of the TGF-β-like open reading frame. The probe was fully protected by human cerebellar RNA (fig. 7B). These results indicated that in both cases, the upstream stop codon present in the TGF-β-like reading frames of *MDF451* and *MDF628* corresponded to the exonic portions of the genes. The absence of other protected species with the extended probes ruled out partial or incomplete splicing events from immature transcripts. Together, these data suggested that the *MDF451* and *MDF628* genes are unlikely to encode members of the TGF-β superfamily. On the other hand, the coding potential of the alternative reading frames was not immediately apparent from manual inspection of the corresponding genomic sequences.

Discussion

Here we have described the discovery of two human genes with similarity to members of the TGF-β superfamily. The two genes were expressed at low levels, and preliminary analyses indicated that their expression appeared to be enriched in distinct areas of the human central nervous system, including adult cerebellum, spinal cord and fetal brain.

Several lines of evidence suggest that both sequences correspond to genes recently evolved within the primate lineage. Their apparent absence in rodents is intriguing,

given that rodents and primates are relatively closely related groups within the evolutionary tree of placental mammals [17, 18]. Further studies using DNA from more distant mammalian species should clarify whether these genes were present in the last common ancestor of primates and rodents but subsequently lost in the latter, or whether they originated de novo within the primate lineage after this separated from the rodent subgroup. Recent estimates indicate that about 5% of all human genes will not be found in the mouse, so the two genes may belong to this small group of primate-specific genes. Human-specific genes are expected to show low levels and restricted patterns of expression.

The presence of a conserved translation stop at the predicted distance from the last cysteine codon in both MDF genes is interesting, given that a translation stop sequence at this position was not part of the query used in our Motifer database searches. Of note, a sizeable fraction (6 out of 11) of the nucleotide differences among different primate sequences of the *MDF628* gene were found in the third codon position of the TGF-β-like reading frame. Together, these features suggest that *MDF451* and *MDF628* have their evolutionary origin within the TGF-β gene superfamily, and may have originated by gene duplication of a TGF-β-like gene in an ancestral primate species.

Given that we have so far been unable to isolate full-length cDNAs or proteins for either *MDF451* or *MDF628*, what, if any, kinds of proteins these two genes may encode remains uncertain. Stop codons in the upstream regions of both *MDF451* and *MDF628* were found to form part of their corresponding mature transcripts, indicating that neither gene is likely to have the capacity to encode a TGF-β-like protein. Intriguingly, a highly related gene, termed *628R*, found in the same genomic region as *MDF628*, was also expressed in human cerebellum and appeared to be restricted to primate genomes, but absent in rodents. The *628R* transcript did not appear to contain any sizeable open reading frame, leaving the identity of its protein product undefined.

Extant pseudogenes are known as transcriptionally silent genes sharing a common ancestry with known active genes. Mutations accumulate quickly in pseudogenes, and extant pseudogenes are characterized by the presence of multiple in-frame stop codons that prevent translation of a protein product. The sequence of mutation events that lead to the inactivation of an active gene and the appearance of the corresponding pseudogene are unknown. However, during the early evolution of a pseudogene, an active gene may first be inactivated by one or more mutations that prevent protein translation without affecting transcription. With time, lack of selection pressure ultimately results in the accumulation of further mutations in promoter and regulatory sequences that, in addition, abrogate transcription. Thus, although pseudogenes are usually not known to be transcribed, many extant pseudogenes may have given rise to RNA early during their evolution. In keeping with this notion, there are now several examples of transcribed pseudogenes [see for example ref. 19], although their functional relevance remains unclear [for a review, see ref. 20]. Interestingly, however, a recent study reported that an expressed pseudogene of the *Makorin-1* gene regulates the mRNA stability of its homologous coding gene in trans [21]. Disruption of this pseudogene produced polycystic kidneys and bone deformity in transgenic mice [21], thus demonstrating a specific regulatory role of an expressed pseudogene and pointing to the functional significance of non-coding RNAs. Given their apparently recent evolutionary origin, *MDF451* and *MDF628* may be transcribed pseudogenes originating from the duplication of an ancestral TGF-β-like gene in the primate lineage. However, our data do not rule out the possibility that the MDFs may have existed prior to the divergence of primates but were then eliminated in other mammalian species, although we think this is a less likely scenario.

The completion of the sequence of the human genome has opened a new era in genomic research. This study illustrates several of the complexities lying ahead as we undertake the task of annotating all human genes, elucidating their evolutionary histories, and characterizing their products.

Acknowledgements. We thank E. Sonnehammer for help with early access to private sequence databases, B. O. Röken and T. Möller at Kolmårdens Zoo, K. Bernodt at Skansen, A. Holmström and B. Jonsson at the Skansen Aquarium for kindly providing blood samples, J. Smeds for human genomic DNA, P. Kotokorpi for technical assistance and X. Li for secretarial help. This work was initially funded by a grant from the Pharmacia Corporation and later by grants from the Swedish Medical Research Council (K99-33X-10908-06C), the Göran Gustafssons Stiftelse and Karolinska Institutet.

1 McDonald N. Q. and Hendrickson W. A. (1993) A structural superfamily of growth factors containing a cystine knot motif. Cell **73:** 421–424
2 Lin L.-F. H., Doherty D., Lile J., Bektesh S. and Collins F. (1993) GDNF: a glial cell line-derived neurotrophic factor for midbrain dopaminergic neruons. Science **260:** 1130–1132
3 Schlunegger M. and Grüter M. (1992) An unusual feature revealed by the crystal struture at 2.2 Å resolution of human transforming growth factor-b2. Nat. **358:** 430–434
4 Eigenbrot C. and Gerber N. (1997) X-ray structure of glial cell-derived neurotrophic factor at 1.9 angstrom resolution and implications for receptor binding. Nat. Struct. Biol. **4:** 435–438
5 Massagué J. and Chen Y.-G. (2000) Controlling TGF-b signaling. Genes Dev **14:** 627–644
6 Dijke P. ten, Miyazono K. and Heldin C.-H. (2000) Signaling inputs converge on nuclear effecters in TGF-beta signaling. Trends Biochem. Sci. **25:** 64–70
7 Wrana J. (2000) Crossing smads, science's signal transduction knowledge environment. http//stke.sciencemag.org/content/full/sigtrans; 2000/23/re1
8 Airaksinen M. S., Titievsky A. and Saarma M. (1999) GDNF family neurotrophic factor signaling: four masters, one servant? Mol. Cell. Neurosci. **13:** 313–325

9 Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J. et al. (2001) Initial sequencing and analysis of the human genome. Nature **409:** 860–921
10 Waterston R. H., Lindblad-Toh K., Birney E., Rogers J., Abril J. F., Agarwal P. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature **420:** 520–562
11 Higgins D. G. and Sharp P. M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene **73:** 237–244
12 Jeanmougin F., Thompson J. D., Gouy M., Higgins D. G. and Gibson T. J. (1998) Multiple sequence alignment with Clustal X. Trends Biochem. Sci. **23:** 403–405
13 Jörnvall H. (1999) Motifer, a search tool for finding amino acid sequence patterns from nucleotide sequence databases. FEBS Lett. **456:** 85–88
14 Altschul S. F., Gish W., Miller W., Myers E. W. and Lipman D. J. (1990) Basic local alignment search tool. J. Mol. Biol. **215:** 403–410
15 Woodbury D., Schaar D. G., Ramakrishnan L. and Black I. B. (1998) Novel structure of the human GDNF gene. Brain Res. **803:** 95–104
16 Gitelman S. E., Kobrin M., Lee A., Fet V., Lyons K., Hogan B. L. et al. (1997) Structure and sequence of the mouse Bmp6 gene. Mamm. Genome **8:** 212–214
17 Murphy W. J., Eizirik E., Johnson W. E., Zhang Y. P., Ryder O. A. and O'Brien S. J. (2001) Molecular phylogenetics and the origins of placental mammals. Nature **409:** 614–618
18 Madsen O., Scally M., Douady C. J., Kao D. J., DeBry R. W., Adkins R. et al. (2001) Parallel adaptive radiations in two major clades of placental mammals. Nature **409:** 610–614
19 Olsen M. A. and Schechter L. E. (1999) Cloning, mRNA localization and evolutionary conservation of a human 5-HT7 receptor pseudogene. Gene **227:** 63–69
20 Mighell A. J., Smith N. R., Robinson P. A. and Markham A. F. (2000) Vertebrate pseudogenes. FEBS Lett. **468:** 109–114
21 Hirotsune S., Yoshida N., Chen A., Garrett L., Sugiyama F., Takahashi S. et al. (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature **423:** 91–96