The identification and characterization of human Sister-of-Mammalian Grainyhead (SOM) expands the *grainyhead*-like family of developmental transcription factors

Stephen B. TING*, Tomasz WILANOWSKI*, Loretta CERRUTI*, Lin-Lin ZHAO*, John M. CUNNINGHAM† and Stephen M. JANE*[1]

*Rotary Bone Marrow Research Laboratories, Royal Melbourne Hospital Research Foundation, c/o Royal Melbourne Hospital Post Office, Grattan Street, Parkville, Victoria 3050, Australia, and †Division of Experimental Hematology, St Jude Children's Research Hospital, Memphis, TN 38101, U.S.A.

The *Drosophila* gene *grainyhead* is the founding member of a large family of genes encoding developmental transcription factors that are highly conserved from fly to human. The family consists of two main branches, with *grainyhead* as the ancestral gene for one branch and the recently cloned *Drosophila CP2* as the ancestral gene for the other. We now extend this family with the identification of another novel mammalian member, Sister-of-Mammalian Grainyhead (SOM), which is phylogenetically aligned with *grainyhead*. SOM is closely related to the other mammalian homologues of *grainyhead*, including Mammalian Grainyhead (MGR) and Brother-of-MGR, sharing a high degree of sequence identity with these factors in the functional DNA-binding, protein dimerization and activation domains. Protein interaction studies demonstrate that SOM can heterodimerize with MGR and Brother-of-MGR, but not with the more distant members of the family. Like *grainyhead*, the SOM gene too produces several distinct isoforms with differing functional properties through alternative splicing. The tissue distributions of these isoforms differ and all display highly restricted expression patterns. These findings indicate that SOM, like its family members, may play important roles in mammalian development.

Key words: *Drosophila*, *grainyhead*, homologue, mammalian development, transcription factor.

INTRODUCTION

The *grainyhead*-like genes encode a rapidly expanding family of developmental transcription factors. The founding member, *Drosophila grainyhead* (also known as NTF-1 or Elf-1), was originally identified as a transcriptional activator of the *Dopa decarboxylase* gene [1–4]. Subsequent studies revealed important roles in early fly development through repression of dorsal/ventral and terminal patterning genes *decapentaplegic*, *tailless* and *zerknüllt* and in later stages through transcriptional regulation of key developmental genes such as *engrailed*, *fushi tarazu* and *Ultrabithorax* [1,2,5–11]. Flies carrying a null mutation of the *grainyhead* gene die during embryogenesis with abnormalities of the cuticle, head skeleton and trachea [4]. The lack of patterning defects in these mutant embryos is due to the maternal provision of Grainyhead (GRH) during early embryogenesis. This is evidenced by *in situ* hybridization studies, which show *grainyhead* mRNA synthesis during oogenesis with deposition in the developing oocyte, and by the expansion of the *tailless* expression domain observed in embryos derived from females carrying germ-line clones lacking *grh* [10,11].

The diversity of the developmental roles played by *grainyhead* is mediated through the formation of tissue-specific protein isoforms and also through the formation of target gene-specific heteromeric protein complexes [8]. A neuroblast-specific isoform of the protein that arises from alternative splicing is critical for normal development of the central nervous system. Flies carrying a mutation that abolishes the production of this isoform exhibit pupal and adult lethality and demonstrate non-co-ordinate movements [12].

Five mammalian members of the *grh*-like family have been identified previously [13–21]. These genes are divided into two distinct phylogenetic groups on the basis of sequence alignments [21]. The first group consists of CP2, LBP-1a and LBP-9, which are homologues of the recently identified *Drosophila* gene *dCP2* [21]. These genes are, in general, widely expressed and play roles in diverse cellular and developmental events, which include T-cell proliferation, globin gene expression and steroid biosynthesis [18,22–24]. The second group consists of *Drosophila grainyhead* and its two mammalian homologues, namely, Mammalian Grainyhead (MGR) and Brother-of-MGR (BOM). These genes have highly restricted patterns of expression and show high levels of sequence conservation with *grh*, particularly in the DNA-binding and protein dimerization domains [21]. The sequence differences between the two arms of the phylogenetic tree have functional consequences, with the *grh*-like factors MGR, BOM and GRH capable of interacting with each other, but not with members of the CP2-like group [9,21].

We now report the identification and characterization of a third mammalian member of the *grh*-like family, Sister-of-MGR

Abbreviations used: BOM, Brother-of-MGR; CAT, chloramphenicol acetyltransferase; EST, expressed sequence tag; GRH, *Drosophila* Grainyhead; GST, glutathione S-transferase; MGR, Mammalian Grainyhead; RACE, rapid amplification of cDNA ends; RT, reverse transcriptase; SCL, stem cell leukaemia; SOM, Sister-of-MGR.

[1] To whom correspondence should be addressed (e-mail jane@wehi.edu.au).

The nucleotide sequence data for human SOM have been submitted to the DDBJ, EMBL, GenBank® and GSDB Nucleotide Sequence Databases. The accession number is pending.

(SOM). Similar to its homologues, SOM is restricted in its expression patterns and its use of protein partners, emphasizing the functional diversity that has evolved in the two branches of this multi-gene family.

EXPERIMENTAL

Cloning of SOM

The genomic sequence of SOM was initially identified in the High Throughput Genomic Sequence database by searching with the highly conserved protein dimerization domain of the previously cloned mammalian *grh*-like factors, MGR and BOM. Gene-specific primers were derived from this sequence and used in reverse transcriptase (RT)–PCR analysis of a panel of human tissues. A fragment of the predicted size was amplified from tonsil and verified as a novel family member by sequencing. The subsequent deposition in the databases of a human expressed sequence tag (EST) that contained significant identity with, but was clearly distinct from, the activation domain of MGR and BOM facilitated the identification of a larger SOM fragment. Gene-specific primers derived from this activation domain and our previously cloned dimerization domain were used to amplify a fragment from tonsil cDNA. The sequence of this fragment confirmed the identity with MGR and BOM in the three functional domains but was divergent outside these regions. No initiating methionine residue was identified in these fragments, so 5′-rapid amplification of cDNA ends (5′-RACE; Marathon RACE; Clontech) was employed to clone the N-terminus from testis cDNA. This resulted in the isolation of the first coding exon of SOM1. Ultimately, the entire contiguous cDNA of SOM1 was amplified in one PCR from testis cDNA using specific oligonucleotide primers from the 5′- and 3′-untranslated regions respectively. The PCR product was cloned according to the manufacturer's instructions (Invitrogen Topo TA cloning kit). Topoclone number 5 was used for sequencing of the full-length cDNA. At this stage, topoclone number 2 was noted to be slightly smaller on an *Eco*RI excision digest, and upon sequencing, exon 2 was noted to be absent. This clone was subsequently labelled SOM3.

SOM2 was identified as an EST from human renal epithelial cells (accession no. AK074386), and was identical with SOM1 except for the first exon. This difference was supported by the alignment of SOM1 and SOM2 with the two human SOM genomic clones (accession nos. AL138902 and AL031431), which showed exon 1a of SOM2 to be upstream of exon 1b of SOM1 and that both of these isoforms continued in-frame with the same exon 2 up to the stop codon. Using RT–PCR with a panel of human cDNA tissues and a specific oligonucleotide primer for exon 1a, and a downstream 3′ oligonucleotide primer, the existence of SOM2 in specific tissues was confirmed. Primer sequences are listed below.

PCR

For RT–PCR, first strand cDNA was prepared from 2 μg of mRNA from primary tissues using random hexamers. For the expression studies, each cDNA sample was appropriately diluted to give similar amplification of S14 RNA under the same PCR conditions. The primer sequences are detailed below. The PCR conditions were 94 °C for 2 min followed by 35 cycles at 94 °C for 30 s, 60 °C for 30 s and 72 °C for 1 min with a final extension at 72 °C for 5 min. All PCR products were electrophoresed on 1.5 % agarose gels, transferred to nitrocellulose and analysed by Southern blotting using ^{32}P-radiolabelled internal oligonucleotides as probes. Membranes were then autoradiographed for 2 h at −70 °C.

Primers

The following primers were used to amplify probes for cloning of SOM isoforms and for RT–PCR. For amplification of full-length human SOM1: SOM1 5′-primer, 5′-GGAGATGTGCCAAACTGT-3′; SOM1 3′-primer, 5′-TGTGGAGAGGTTGTGTGT-3′. For amplification of SOM2-specific N-terminal fragment: SOM2 5′-primer, 5′-AGTCGAATGAACTTGATTTCAG-3′; SOM2 3′-primer, 5′-TCCAGACACGTTCTCTGT-3′. For amplification of SOM1- and SOM3-specific N-terminal fragments: SOM1/3 5′-primer, 5′-AGCAGAAGAATGTGGATG-3′; SOM1/3 3′-primer, 5′-TTTGTTGAGGTAGGCCATGGGTGACTC-3′ (SOM1 fragment, 790 bp; SOM3 fragment, 603 bp). Hypoxanthine–guanine phosphoribosyltransferase: 5′-primer, 5′-ATGGACAGGACTGAACGTCT-3′; 3′-primer, 5′-CTTGCGACCTTGACCATCTT-3′; 5′-RACE primers. First round PCR: SOM 3′-primer, 5′-GCAACACTGTCATCATCTCC-3′; AP1 5′-primer, 5′-CCATCCTAATACGACTCACTATAGGGC-3′. Nested PCR: SOM 3′-primer, 5′-ACTCTCATCATGGCCTTTGTGG-3′; AP2 5′-primer, 5′-ACTCACTATAGGGCTCGAGCGGC-3′.

DNA construction and transactivation analysis

Protein domains involved in transcriptional activation were defined using the Mammalian Matchmaker two-hybrid assay kit (Clontech). We engineered *Eco*RI restriction sites by PCR to the 5′-ends of cDNA fragments encoding the conserved activation domain in SOM1 and SOM2 (amino acid residues 30–95) and the N-terminal part of the SOM3 protein (amino acid residues 1–32). The PCR products were subcloned into the *Eco*RI site of the pM vector containing the GAL4-DNA-binding domain and sequenced in their entirety. The 293T cell line was transfected in triplicate with these plasmids using the calcium phosphate transfection system (Gibco BRL, Gaithersburg, MD, U.S.A.). The pM vector alone and the pM vector containing the VP16 activation domain served as the negative and positive controls respectively. A firefly luciferase reporter construct driven by the thymidine kinase promoter was used for controlling the transfection efficiency. After 48 h, the amount of chloramphenicol acetyltransferase (CAT) produced in each transfected cell line was quantified with the CAT ELISA colorimetric immunoassay kit (Roche Diagnostics GmbH, Mannheim, Germany). This value was corrected for transfection efficiency based on the luciferase activity of the sample (measured in a Monolight 2001 luminometer with the Promega luciferase assay kit) and also adjusted with respect to the protein content of the lysate.

Yeast two-hybrid analysis of protein interactions

For analysis of mammalian protein interactions, indicated fragments of the cDNAs encoding the protein dimerization domains of CP2, LBP-1a, MGR, BOM and SOM and full-length SCL (stem cell leukaemia) were derived by PCR or restriction digest and cloned into pGAD424 (a GAL4 transactivation domain vector; GAL4AD; Clontech). The PCR fragments were sequenced in their entirety and the restriction fragments were sequenced across the fusion junction with GAL4AD. The resulting plasmids were co-transformed into yeast with pGB-SOM, a yeast expression vector containing the protein dimerization domain of SOM fused to the GAL4 DNA-binding domain (GAL4DBD) as described previously [25]. Positive interactions met the two criteria of growing on selection media plates and testing β-galactosidase-positive. Expression of the various plasmids was confirmed in yeast by Western-blot analysis using antibodies that recognize GAL4DBD or GAL4AD (Clontech; results not shown).

Expression of glutathione S-transferase (GST) fusion proteins and affinity chromatography

Human MGR, BOM and SOM and GRH cDNAs were cloned in frame with the GST coding sequence in the pGEX vectors (Pharmacia, Piscataway, NJ, U.S.A.). The GST fusion proteins were expressed in the *Escherichia coli* strain BL21. Fusion proteins were purified on glutathione–Sepharose (Pharmacia), and their integrity confirmed with Coomassie Blue staining after SDS/PAGE. For *in vitro* protein–protein interaction assays, 1 μg of GST or GST fusion protein was incubated for 1 h at 4 °C with 10 μl of glutathione–Sepharose beads, which had been preblocked with 0.5 % milk. After extensive washing, the beads were resuspended in 200 μl of binding buffer [10 mM Tris/HCl (pH 7.9)/500 mM KCl/0.1 mM EDTA/150 μg/ml BSA/0.1 % Nonidet P40/10 % glycerol] and incubated for 1 h at room temperature (22 °C) with [^{35}S]methionine-labelled SOM. After extensive washing, retained proteins were eluted by boiling in SDS loading buffer and analysed by SDS/PAGE and autoradiography [26].

RESULTS

Cloning of SOM, a novel mammalian member of the *grainyhead*-like family of transcription factors

We have recently reported [21] the identification and characterization of two novel mammalian members of the *grh*-like family of developmental transcription factors, MGR and BOM, which rewrite the phylogeny of this family. As a continuation of the above-mentioned work, we located additional sequences in the High Throughput Genomic Sequence database and a human EST that shared identity with human MGR and BOM in the highly conserved protein dimerization and activation domains. Gene-specific primers derived from these domains were used to amplify a fragment from tonsil cDNA. The sequence of this fragment confirmed the identity with MGR and BOM in the functional domains, but was divergent outside these regions. No initiating methionine was identified in this fragment and so 5′-RACE (Marathon RACE) was employed to clone the N-terminus from testis cDNA. This resulted in the isolation of the first coding exon of SOM1. Ultimately, the entire contiguous cDNA of SOM1 was amplified from testis cDNA and sequenced (see cloning of SOM in the Experimental section). Alignment of the predicted amino acid sequence from the full-length cDNA (which we have named SOM) with MGR and BOM revealed > 60 % overall similarity at amino acid level (Figure 1A). This was significantly higher in the DNA-binding, protein dimerization and transactivation domains (Table 1). We subsequently identified the full-length mouse SOM cDNA from a brain cDNA library (results not shown). The similarity between the predicted human and murine proteins is very high, with 90 % identity.

Phylogenetic analysis of the extended *grh*-like family

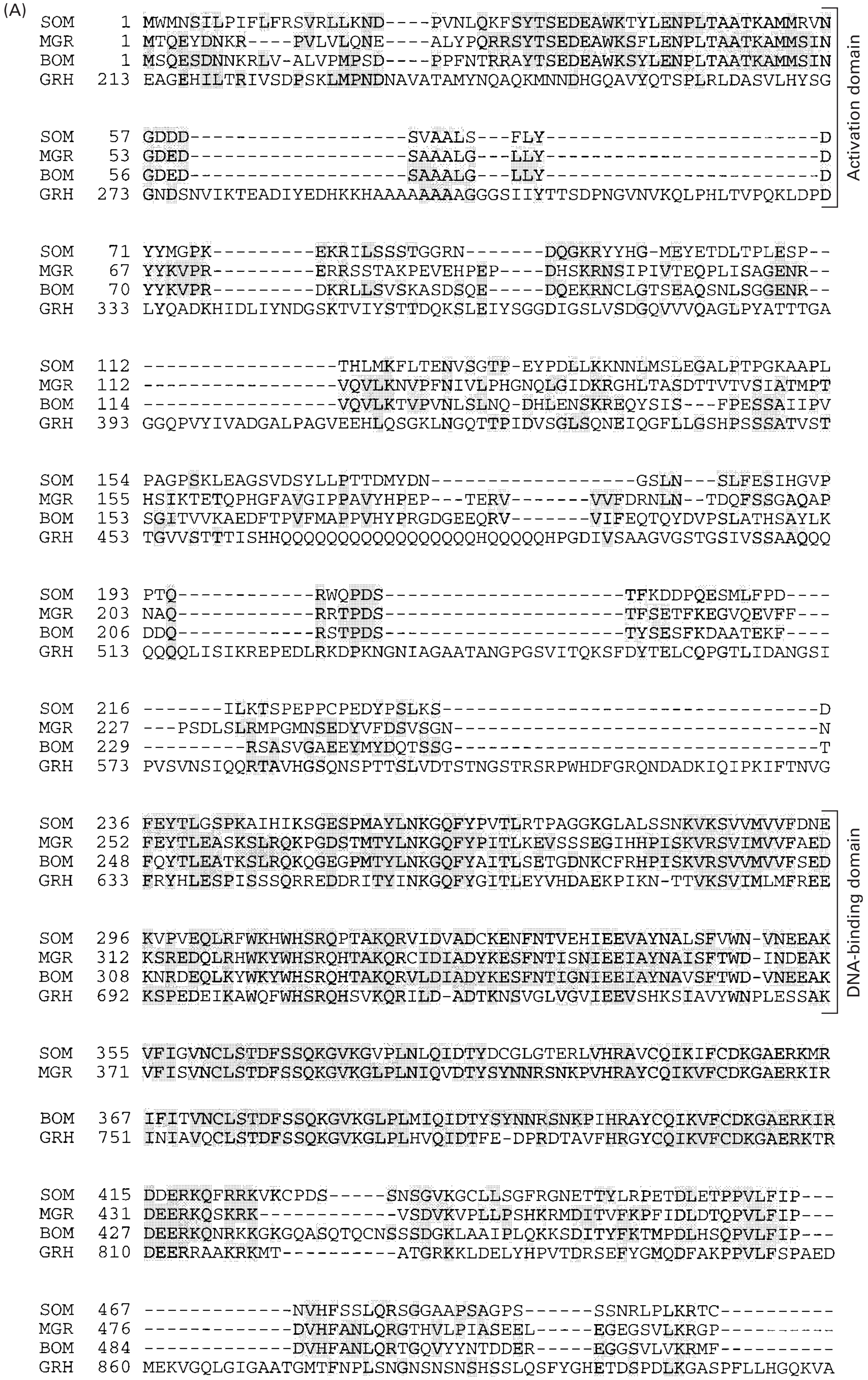
The sequence conservation in this family from fly to mouse to man suggested that related genes would be present in more diverse organisms. We have identified previously homologues of MGR and BOM in *Xenopus laevis* (S. M. Jane and J. M. Cunningham, unpublished work). Further trawling of the databases revealed that homologues of SOM also exist in *Xenopus*. Phylogenetic analysis of these sequences confirmed our previous grouping of this family into two distinct arms, one descended from *Drosophila CP2* and the other from *grainyhead* (Figure 2). Homologues of these genes are also present in *Danio rerio*, *Xenopus tropicalis*, *Oryzias latipes* and *Takifugu rubripes* (results not shown).

The phylogenetic grouping of the *grh*-like family has functional consequences

The grainyhead-like factors achieve functional diversity through the formation of homo- and heteromeric complexes. We have shown previously that protein interactions between members of this family were confined to factors that segregate on the distinct arms of the phylogenetic tree. To determine whether SOM adhered to this, we examined if protein complexes could be formed between SOM and the *grh*-like factors and between SOM and the CP2-like proteins. We utilized the yeast two-hybrid assay system for this purpose, as we have shown that results in this system gave an accurate reflection of interactions, which were subsequently confirmed using GST chromatography and co-immunoprecipitation. The dimerization domains of the three mammalian *grh*-like factors were cloned into a yeast expression vector in-frame with the GAL4 DNA-binding domain and a second yeast vector in-frame with the GAL4 activation domain. The constructs were co-transfected into yeast in various combinations, and interactions were assessed by the activation of a histidine reporter gene and growth on selective media plates. As shown in Figure 3(A), SOM is capable of homodimerization, as well as heterodimerization with MGR and BOM. No interactions were observed between SOM and CP2 or LBP-1a. No colonies were observed with an unrelated transcription factor (SCL). Significant activation of the histidine reporter in these experiments was accompanied by concomitant activation of the second reporter, LacZ, with all colonies staining blue with X-gal (results not shown). To confirm the findings for yeast, we performed GST pull-down experiments utilizing a GST–SOM fusion protein coupled with glutathione–Sepharose and ^{35}S-radiolabelled *in vitro* transcribed/translated SOM, BOM and MGR. As shown in Figure 3(B), ^{35}S-radiolabelled SOM was strongly retained on GST–SOM (panel 1). GST–SOM also retained ^{35}S-radiolabelled MGR (panel 2) and BOM (panel 3), but not an unrelated control, breast-cancer susceptibility gene 1 (panel 4). To determine whether the ability of SOM to interact with the factors in this arm of the phylogenetic tree extended back to GRH, ^{35}S-radiolabelled *in vitro* transcribed/translated GRH was added to the GST–SOM matrix. As shown in Figure 3(C), GRH was specifically retained on this matrix, but not on the control GST column. It has been shown previously that GRH does not interact with CP2. These findings again emphasize the functional importance of the phylogenetic division of this family.

Identification of isoforms of SOM

In addition to a range of protein partners, the *grh*-like genes achieve functional diversity through the presence of tissue-specific isoforms. This is particularly evident in *Drosophila*, where disruption of the neural isoform of GRH causes pupal/adult lethality. MGR and BOM also exist as different isoforms, some of which lack key functional domains. To facilitate the search for isoforms of SOM, we used a combination of genomic mapping of the SOM locus, database trawling and RT–PCR of samples from



```
SOM    1 MWMNSILPIFLFRSVRLLKND----PVNLQKFSYTSEDEAWKTYLENPLTAATKAMMRVN
MGR    1 MTQEYDNKR----PVLVLQNE----ALYPQRRSYTSEDEAWKSFLENPLTAATKAMMSIN
BOM    1 MSQESDNNKRLV-ALVPMPSD----PPFNTRRAYTSEDEAWKSYLENPLTAATKAMMSIN
GRH  213 EAGEHILTRIVSDPSKLMPNDNAVATAMYNQAQKMNNDHGQAVYQTSPLRLDASVLHYSG

SOM   57 GDDD-------------------SVAALS---FLY------------------------D
MGR   53 GDED-------------------SAAALG---LLY------------------------D
BOM   56 GDED-------------------SAAALG---LLY------------------------D
GRH  273 GNDSNVIKTEADIYEDHKKHAAAAAAAAGGGSIIYTTSDPNGVNVKQLPHLTVPQKLDPD

SOM   71 YYMGPK---------EKRILSSSTGGRN-------DQGKRYYHG-MEYETDLTPLESP--
MGR   67 YYKVPR---------ERRSSTAKPEVEHPEP----DHSKRNSIPIVTEQPLISAGENR--
BOM   70 YYKVPR---------DKRLLSVSKASDSQE-----DQEKRNCLGTSEAQSNLSGGENR--
GRH  333 LYQADKHIDLIYNDGSKTVIYSTTDQKSLEIYSGGDIGSLVSDGQVVVQAGLPYATTTGA

SOM  112 -----------------THLMKFLTENVSGTP-EYPDLLKKNNLMSLEGALPTPGKAAPL
MGR  112 -----------------VQVLKNVPFNIVLPHGNQLGIDKRGHLTASDTTVTVSIATMPT
BOM  114 -----------------VQVLKTVPVNLSLNQ-DHLENSKREQYSIS---FPESSAIIPV
GRH  393 GGQPVYIVADGALPAGVEEHLQSGKLNGQTTPIDVSGLSQNEIQGFLLGSHPSSSATVST

SOM  154 PAGPSKLEAGSVDSYLLPTTDMYDN------------------GSLN---SLFESIHGVP
MGR  155 HSIKTETQPHGFAVGIPPAVYHPEP---TERV-------VVFDRNLN--TDQFSSGAQAP
BOM  153 SGITVVKAEDFTPVFMAPPVHYPRGDGEEQRV-------VIFEQTQYDVPSLATHSAYLK
GRH  453 TGVVSTTTISHHQQQQQQQQQQQQQQQQQHQQQQQHPGDIVSAAGVGSTGSIVSSAAQQQ

SOM  193 PTQ------------RWQPDS---------------------TFKDDPQESMLFPD----
MGR  203 NAQ------------RRTPDS---------------------TFSETFKEGVQEVFF---
BOM  206 DDQ------------RSTPDS---------------------TYSESFKDAATEKF----
GRH  513 QQQQLISIKREPEDLRKDPKNGNIAGAATANGPGSVITQKSFDYTELCQPGTLIDANGSI

SOM  216 -------ILKTSPEPPCPEDYPSLKS---------------------------------D
MGR  227 ---PSDLSLRMPGMNSEDYVFDSVSGN--------------------------------N
BOM  229 ---------RSASVGAEEYMYDQTSSG--------------------------------T
GRH  573 PVSVNSIQQRTAVHGSQNSPTTSLVDTSTNGSTRSRPWHDFGRQNDADKIQIPKIFTNVG

SOM  236 FEYTLGSPKAIHIKSGESPMAYLNKGQFYPVTLRTPAGGKGLALSSNKVKSVVMVVFDNE
MGR  252 FEYTLEASKSLRQKPGDSTMTYLNKGQFYPITLKEVSSSEGIHHPISKVRSVIMVVFAED
BOM  248 FQYTLEATKSLRQKQGEGPMTYLNKGQFYAITLSETGDNKCFRHPISKVRSVVMVVFSED
GRH  633 FRYHLESPISSSQRREDDRITYINKGQFYGITLEYVHDAEKPIKN-TTVKSVIMLMFREE

SOM  296 KVPVEQLRFWKHWHSRQPTAKQRVIDVADCKENFNTVEHIEEVAYNALSFVWN-VNEEAK
MGR  312 KSREDQLRHWKYWHSRQHTAKQRCIDIADYKESFNTISNIEEIAYNAISFTWD-INDEAK
BOM  308 KNRDEQLKYWKYWHSRQHTAKQRVLDIADYKESFNTIGNIEEIAYNAVSFTWD-VNEEAK
GRH  692 KSPEDEIKAWQFWHSRQHSVKQRILD-ADTKNSVGLVGVIEEVSHKSIAVYWNPLESSAK

SOM  355 VFIGVNCLSTDFSSQKGVKGVPLNLQIDTYDCGLGTERLVHRAVCQIKIFCDKGAERKMR
MGR  371 VFISVNCLSTDFSSQKGVKGLPLNIQVDTYSYNNRSNKPVHRAYCQIKVFCDKGAERKIR

BOM  367 IFITVNCLSTDFSSQKGVKGLPLMIQIDTYSYNNRSNKPIHRAYCQIKVFCDKGAERKIR
GRH  751 INIAVQCLSTDFSSQKGVKGLPLHVQIDTFE-DPRDTAVFHRGYCQIKVFCDKGAERKTR

SOM  415 DDERKQFRRKVKCPDS-----SNSGVKGCLLSGFRGNETTYLRPETDLETPPVLFIP---
MGR  431 DEERKQSKRK------------VSDVKVPLLPSHKRMDITVFKPFIDLDTQPVLFIP---
BOM  427 DEERKQNRKKGKGQASQTQCNSSSDGKLAAIPLQKKSDITYFKTMPDLHSQPVLFIP---
GRH  810 DEERRAAKRKMT----------ATGRKKLDELYHPVTDRSEFYGMQDFAKPPVLFSPAED

SOM  467 -------------NVHFSSLQRSGGAAPSAGPS------SSNRLPLKRTC----------
MGR  476 -------------DVHFANLQRGTHVLPIASEEL-----EGEGSVLKRGP----------
BOM  484 -------------DVHFANLQRTGQVYYNTDDER-----EGGSVLVKRMF----------
GRH  860 MEKVGQLGIGAATGMTFNPLSNGNSNSNSHSSLQSFYGHETDSPDLKGASPFLLHGQKVA
```

Figure 1 For legend see facing page

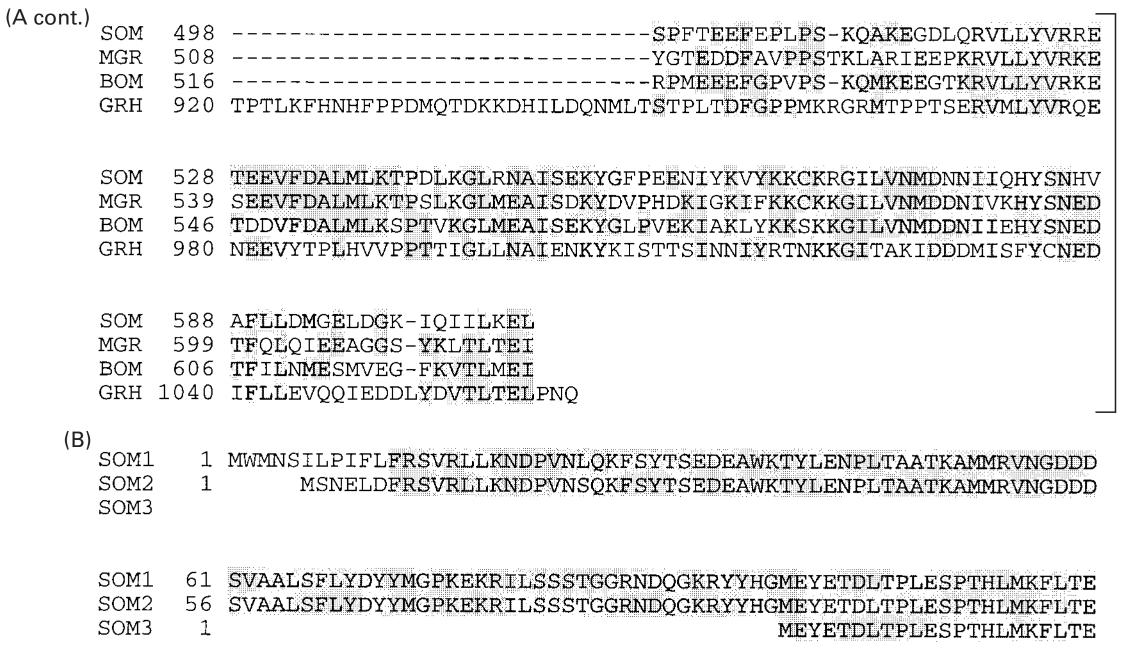

(A cont.)
SOM 498 ------------------------------SPFTEEFEPLPS-KQAKEGDLQRVLLYVRRE
MGR 508 ------------------------------YGTEDDFAVPPSTKLARIEEPKRVLLYVRKE
BOM 516 ------------------------------RPMEEEFGPVPS-KQMKEEGTKRVLLYVRKE
GRH 920 TPTLKFHNHFPPDMQTDKKDHILDQNMLTSTPLTDFGPPMKRGRMTPPTSERVMLYVRQE
SOM 528 TEEVFDALMLKTPDLKGLRNAISEKYGFPEENIYKVYKKCKRGILVNMDNNIIQHYSNHV
MGR 539 SEEVFDALMLKTPSLKGLMEAISDKYDVPHDKIGKIFKKCKKGILVNMDDNIVKHYSNED
BOM 546 TDDVFDALMLKSPTVKGLMEAISEKYGLPVEKIAKLYKKSKKGILVNMDDNIIEHYSNED
GRH 980 NEEVYTPLHVVPPTTIGLLNAIENKYKISTTSINNIYRTNKKGITAKIDDDMISFYCNED
SOM 588 AFLLDMGELDGK-IQIILKEL
MGR 599 TFQLQIEEAGGS-YKLTLTEI
BOM 606 TFILNMESMVEG-FKVTLMEI
GRH 1040 IFLLEVQQIEDDLYDVTLTELPNQ
Dimerization domain
(B)
SOM1 1 MWMNSILPIFLFRSVRLLKNDPVNLQKFSYTSEDEAWKTYLENPLTAATKAMMRVNGDDD
SOM2 1 MSNELDFRSVRLLKNDPVNSQKFSYTSEDEAWKTYLENPLTAATKAMMRVNGDDD
SOM3
SOM1 61 SVAALSFLYDYYMGPKEKRILSSSTGGRNDQGKRYYHGMEYETDLTPLESPTHLMKFLTE
SOM2 56 SVAALSFLYDYYMGPKEKRILSSSTGGRNDQGKRYYHGMEYETDLTPLESPTHLMKFLTE
SOM3 1 MEYETDLTPLESPTHLMKFLTE

Figure 1 Identification of a novel mammalian gene showing identity with MGR, BOM and GRH

(**A**) Alignment of the predicted amino acid sequences of SOM, MGR, BOM and GRH. Amino acid identity between two or more factors is denoted by shading. The DNA-binding domain and the protein dimerization domains conserved in all factors are labelled, as is the activation domain, conserved in the mammalian factors only. Spaces are inserted to maintain the alignment. (**B**) Alignment of the predicted amino acid sequences of the N-termini of the three isoforms of SOM. The shaded areas indicate residues identical in two or more of the isoforms. Downstream of the area shown, all the isoforms are identical.

Table 1 Amino-acid sequence comparison of SOM with the GRH-like genes

n.s., not significant.

	Amino acid identity/similarity to SOM (%)			
	Overall	Activation domain	DNA-binding domain	Dimerization domain
MGR	46/61	76/87	59/76	56/73
BOM	45/61	75/87	57/73	61/77
CP2	33/48	n.s.	42/60	38/57

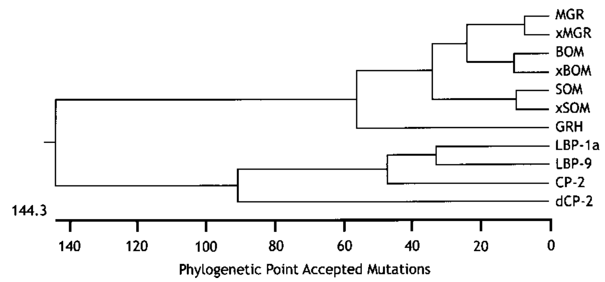


Figure 2 Phylogenetic analysis of the *grainyhead*-like family

Deduced protein sequences from the cDNAs of human SOM, MGR and BOM, *Xenopus laevis* SOM, MGR and BOM (denoted by the prefix x) and *Drosophila CP2* (dCP-2) identified in our laboratory, and the sequences of CP2, LBP-1a, LBP-9 and GRH downloaded from the Genbank® databases were aligned using the Clustal method.

various tissues. As shown in Figure 4, three distinct RNA isoforms have been identified. SOM1 is the original cDNA we isolated. SOM2 is identical with SOM1 except that it utilizes an alternative first coding exon. In contrast, SOM3 originates from the same first exon as SOM1, but lacks exon 2 that encodes a significant component of the core transactivation domain.

The isoforms of SOM are differentially expressed

Our initial expression studies utilized a cDNA probe to a region common to all the identified isoforms of SOM. A human multitissue Northern blot was screened and it demonstrated a single band in placenta and kidney, migrating at approx. 3 kb, consistent with the size predicted for all three SOM transcripts, which do not differ significantly in size (Figure 5A). The lack of transcripts in other tissues on the Northern blot suggested that the expression levels of the SOM isoforms outside the kidney and placenta were very low. To circumvent this and to determine whether differential expression patterns existed, we performed RT–PCR on a range of human tissues using primer pairs that would allow discrimination of the different isoforms. The primers for SOM2 were specific, as one primer was contained in exon 1a, and the primers for SOM1 and SOM3 were shared but spanned exon 2 and thus yielded different-sized products for each isoform. Samples of interest shown in Figure 5(B) and the complete experiment summarized in Figure 5(C) reveal different patterns of expression for each isoform. SOM1 was expressed in a range of tissues, including brain, pancreas, testis, placenta, prostate, colon and kidney. The expression pattern of SOM2 was far more restricted, with some overlap with SOM1 in brain, pancreas,

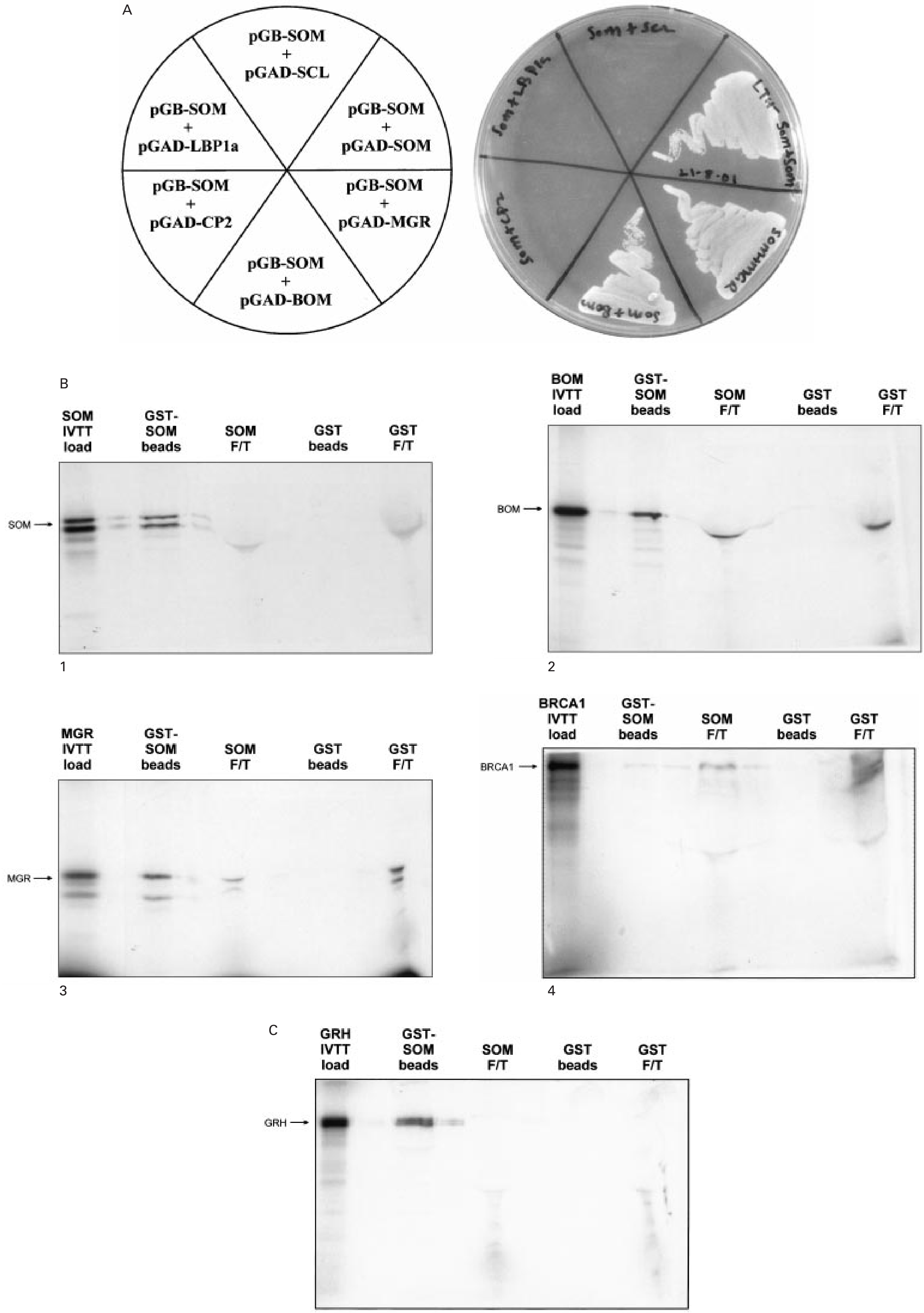


Figure 3 For legend, see facing page.

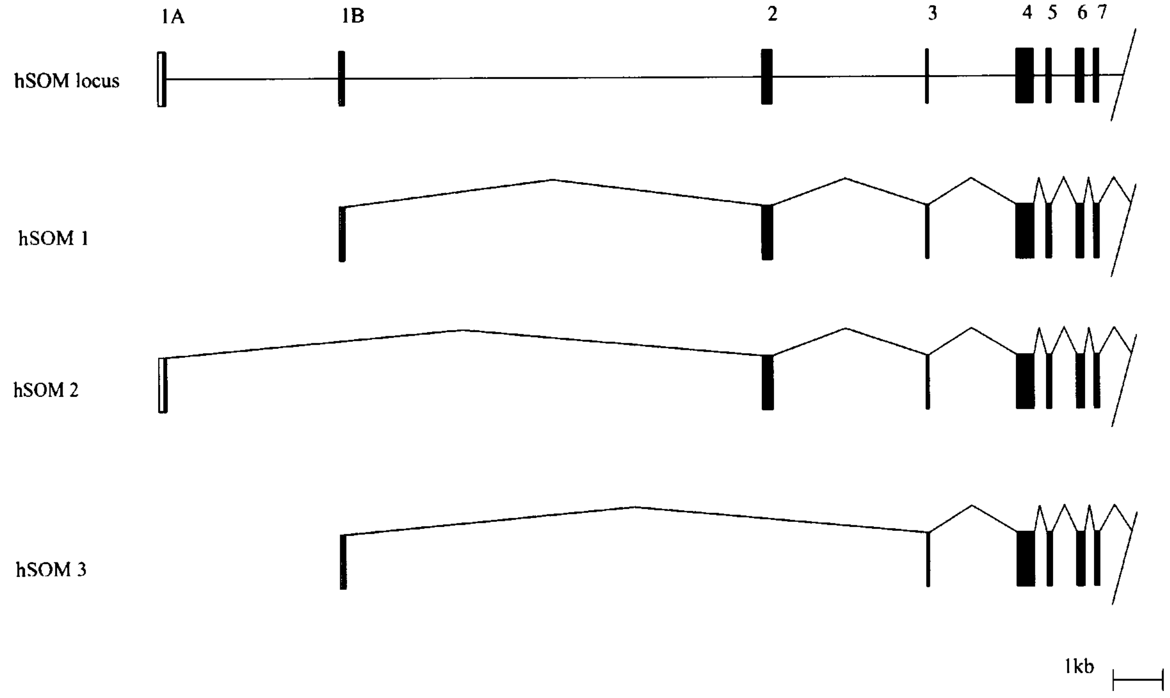


Figure 4 The SOM genomic locus encodes three distinct isoforms

The structure of the human SOM genomic locus is shown. It was derived from an alignment of a human genomic clone (accession no. AL138902) with the cDNA sequences using the BLAST algorithm. The three SOM isoforms were isolated and sequenced using RT–PCR of samples from various tissue sources. Scale bar = 1 kb.

placenta and kidney, but unique expression in tonsil and thymus. SOM3 was less widely expressed when compared with SOM1, but the three tissues, brain, pancreas and testis, in which SOM3 expression was identified also co-expressed SOM1 and SOM2. Although SOM protein is detectable by Western-blot analysis using monoclonal antisera raised against the protein, the small molecular-mass differences between isoforms SOM1 and SOM2 (< 1 kDa) would make their individual detection at the protein level impossible. The predicted smaller (approx. 7 kDa) SOM3 protein size was not seen (results not shown). The detection of SOM3 at the RT–PCR level (above), but its absence in Western-blot analysis, may reflect the low levels of expression of this isoform.

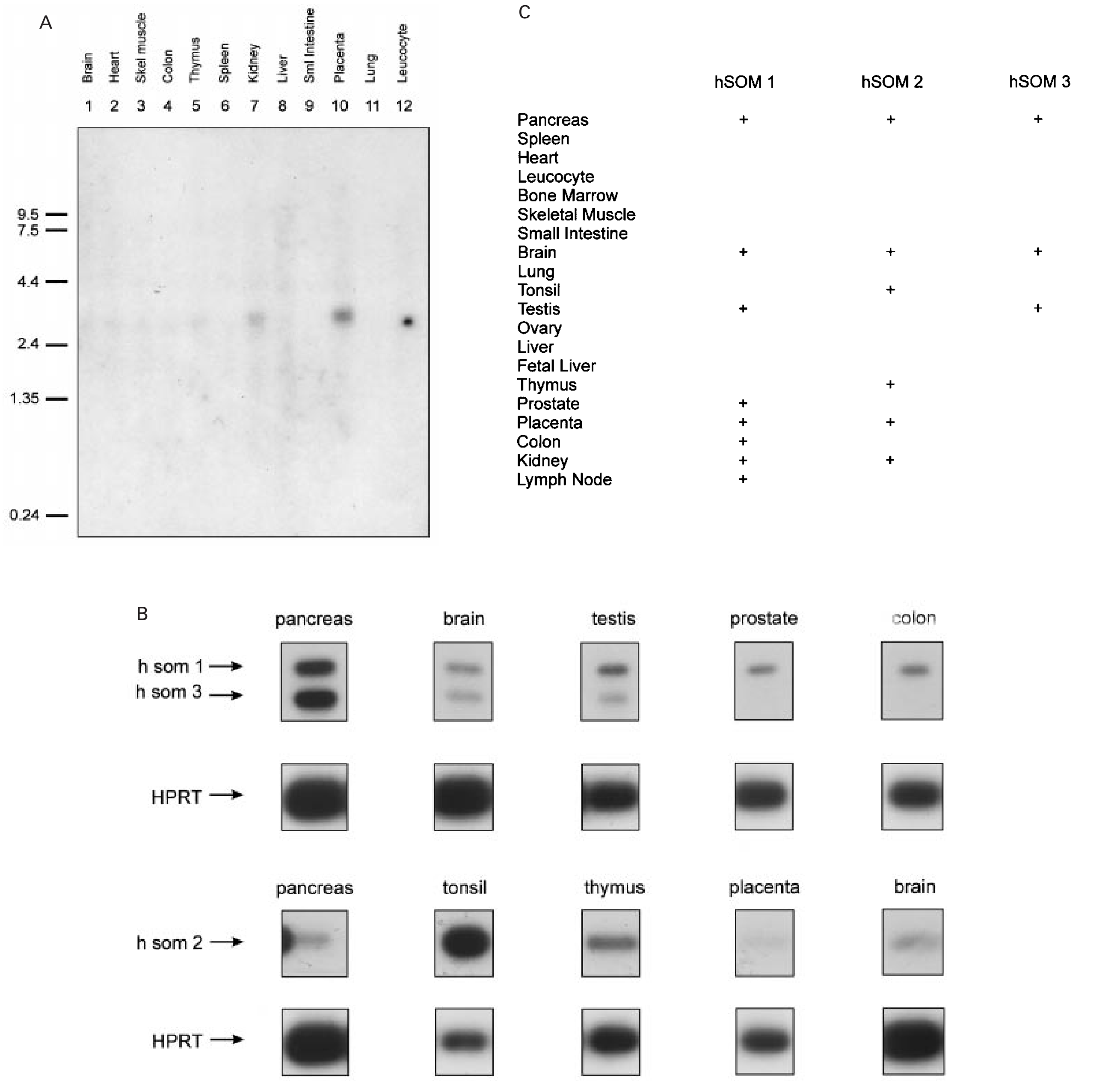
Functional diversity of the different SOM isoforms

Our sequence alignments coupled with our earlier functional studies of MGR and BOM suggested that a conserved trans-activation domain was present in the N-terminal region of SOM (Table 1). The sequence identity between the MGR, BOM and SOM in this region extended from amino acids 30–95 of SOM. This domain is contained in the SOM1 and SOM2 isoforms, but is disrupted by the loss of exon 2 in SOM3. To examine the potential of the SOM isoforms to function as transcriptional activators, we generated mammalian expression vectors carrying fusion proteins between the GAL4 DNA-binding domain and the conserved core-activation domain common to SOM1 and SOM2 or the N-terminal region of SOM3 (Figure 6). The vectors containing the GAL4 DNA-binding domain alone (GAL4) or this domain fused to the activation domain of VP16 (VP16) served as the negative and positive controls respectively.

The constructs were co-transfected into the human 293T cell line with a reporter plasmid containing five concatamerized GAL4 DNA-binding sites upstream of the CAT gene. Transcriptional activation of the CAT gene was observed with VP16 and with the construct containing the core activation domain. In contrast, no activation was observed with the GAL4 vector alone or with the N-terminal region of SOM3. These findings confirm

Figure 3 Protein–protein interactions of the *grainyhead*-like family

(**A**) SOM interacts with itself, MGR and BOM, but fails to interact with CP2 or LBP-1a. The *Saccharomyces cerevisiae* reporter strain HF7C was transformed with the indicated plasmids (left panel). pGB-SOM contains the conserved dimerization domain fused to GAL4DBD. pGAD-SOM, pGAD-MGR, pGAD-BOM, pGAD-CP2 and pGAD-LBP-1a contain the conserved dimerization domains of their respective factors fused to GAL4AD. pGAD-SCL served as an unrelated control. Transformants were streaked on to synthetic medium plates lacking tryptophan, leucine and histidine (LTH$^-$) and incubated at 30 °C for 3 days (right panel). (**B**) GST chromatography with GST–SOM. A fusion protein between GST and SOM (GST–SOM) and GST alone were expressed in *E. coli*, and 1 μg of protein was bound to glutathione–Sepharose beads. The fusion protein and GST coupled with the beads were both incubated with 2 μl of ^{35}S-labelled *in vitro* translated (IVTT) SOM, BOM or MGR or the unrelated control breast-cancer susceptibility (BCRA1) gene in binding buffer for 1 h at room temperature (see the Experimental section). The beads were spun and the non-binding supernatant (labelled as flow-through – F/T) was collected. After extensive washing, the GST fusion protein or GST-coupled beads were mixed with SDS loading buffer and subjected to SDS/PAGE (labelled GST beads). The unbound flow-throughs were also electrophoresed (labelled flow through – F/T). The migration of the various radiolabelled loads is indicated (SOM, BOM, MGR and BCRA1, panels 1–4 respectively). (**C**) GST chromatography with radiolabelled GRH. The experimental details are identical to those in (**B**), except that ^{35}S-labelled *in vitro* translated GRH was used.



	hSOM 1	hSOM 2	hSOM 3
Pancreas	+	+	+
Spleen			
Heart			
Leucocyte			
Bone Marrow			
Skeletal Muscle			
Small Intestine			
Brain	+	+	+
Lung			
Tonsil		+	
Testis	+		+
Ovary			
Liver			
Fetal Liver			
Thymus		+	
Prostate	+		
Placenta	+	+	
Colon	+		
Kidney	+	+	
Lymph Node	+		

Figure 5 Expression of SOM in primary human tissues

(**A**) Northern blot of multiple human tissues. A 480 bp cDNA fragment common to all three SOM isoforms but not homologous with BOM or MGR was used to probe a human multi-tissue Northern blot (Clontech). The size standards are indicated. (**B**) Tissue-specific expression of the different SOM isoforms. First strand cDNA transcribed from polyadenylated [poly(A)$^{+}$] RNA from multiple primary tissues was used as a template to PCR-amplify products using primers specific for the different SOM isoforms. Samples were amplified with primers specific for hypoxanthine–guanine phosphoribosyltransferase (HPRT) RNA to confirm the integrity of the template. A panel of selected tissues is shown. These primer pairs all span an intron and thus discriminate between mRNA and genomic DNA-derived signal. Thirty-five cycles of amplification were used. All PCR products were electrophoresed on 1% agarose, transferred to nitrocellulose and probed with an internal radiolabelled oligonucleotide specific for the predicted product. (**C**) Summary of expression patterns of the SOM isoforms.

the presence of a highly conserved activation domain in SOM1 and SOM2, which is lacking in SOM3.

DISCUSSION

These studies detail the identification and characterization of a novel gene *SOM*, which expands the highly conserved *grainyhead*-like family of developmental transcription factors. SOM, like the other recently described mammalian members of this family MGR and BOM, is most closely aligned from both a sequence and functional viewpoint with the founding member of the family GRH [21]. Phylogenetic analysis confirms that SOM, MGR, BOM and GRH all exist in a distinct division of this multi-gene family, separate from the CP2-like members. We have identified homologues of each of these genes in *Xenopus laevis* and additional homologues in zebrafish (*Danio rerio*), western

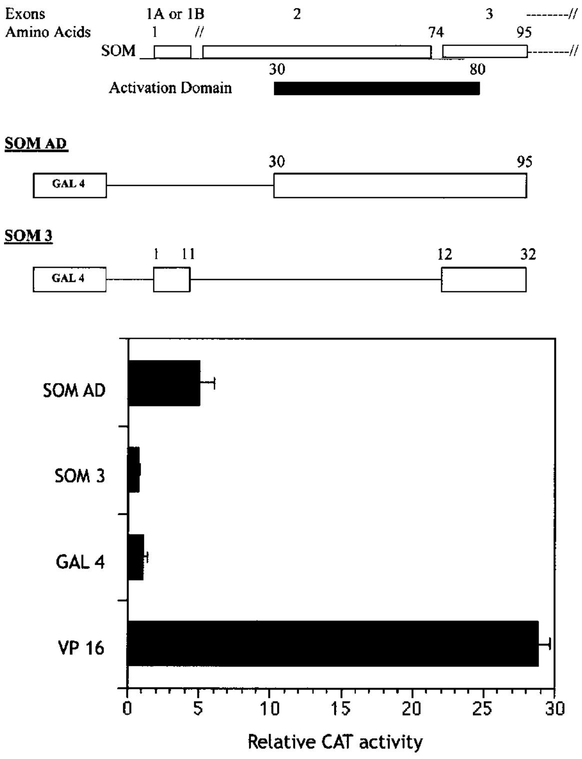


Figure 6 SOM1 and SOM2 contain a conserved activation domain that is absent in SOM3

Mammalian expression plasmids containing the GAL4 DNA-binding domain fused in-frame to the conserved activation domain of SOM1 and SOM2 (SOM AD) or the N-terminal region of SOM3 and linked to the CAT reporter gene were transfected into 293T cells. The empty expression vector (pM) or a vector containing the VP16 activation domain (VP16) served as negative and positive controls respectively (top panel). Cell lysates were prepared after 48 h and analysed for CAT activity (bottom panel).

clawed frog (*Xenopus tropicalis*), Japanese medaka (*Oryzias latipes*), puffer fish (*Takifugu rubripes*) and nematode (*Caenorhabditis elegans*), suggesting that the members of this family play critical developmental roles in a wide range of organisms. The phylogenetic segregation of the *grh*-like genes may have important functional consequences. SOM is only capable of forming multi-protein complexes with the other members of the *grh*-like arm of the family. This finding is in keeping with previous studies showing that MGR and BOM also fail to interact with the CP2-like proteins [21] and that GRH and CP2 could not heterodimerize [9]. It is also consistent with the sequence comparisons in Table 1 and the results of Wilanowski et al. [21], which demonstrate that MGR, BOM, SOM and GRH have no significant identity with dCP2 in the protein dimerization domain.

An important aspect of the functional diversity of this family is the existence of tissue-specific isoforms that arise from alternative splicing. This is particularly evident for the *grh*-like branch of the phylogenetic tree, in which isoforms lacking critical functional domains have been identified and, in the case of *grainyhead*, linked to abnormal phenotypes in the setting of perturbed expression [12,21]. Three isoforms of SOM have been identified thus far. One of these, SOM3, lacks the transactivation domain which is present in the other two SOM isoforms and which is also highly conserved in BOM and MGR. The potential functional importance of SOM3 is not difficult to envisage, particularly in tissues where it is co-expressed with SOM1. Conceivably, SOM1 could function as an activator and the shorter species as a repressor in this context. This regulatory mechanism is a feature of many other transcription factors, including NF-E4 ([24]; S. M. Jane and J. M. Cunningham, unpublished work), AML1 (the acute myeloid leukaemia 1 gene product) [27], cAMP-response-element-binding protein [28], Egr3 [29] and octamer-binding protein 2 [30]. As the shorter SOM species retains its protein dimerization and DNA-binding properties, it could also play a dominant negative role. A precedent for this is observed with a *grainyhead* mutant lacking the N-terminal activation domain that inhibits full-length *grainyhead* site-dependent activation. This mutant preferentially binds to full-length *grainyhead*, inhibiting its ability to homodimerize, leading to a failure of gene activation despite the presence of one intact activation domain. Ectopic expression of this dominant negative protein during fly development leads to embryonic lethality with cuticular defects [31]. The ability of the short form of SOM to homodimerize or heterodimerize with other members of the *grainyhead* family may have similar functional consequences, potentially playing a key role in the regulation of BOM and MGR target genes. The co-expression of the three isoforms of SOM, MGR and BOM in brain and kidney identifies the tissues in which this regulatory mechanism could operate.

The lack of co-expression of SOM2 and SOM3 in tonsil and thymus suggests that the ability of SOM2 to activate gene expression in these two tissues may be obligate. Interestingly, neither of these tissues expresses SOM1. At this stage, it is not clear what functional differences exist between the SOM1 and SOM2 isoforms, but their distinct expression patterns suggest that they may play divergent roles.

At this stage, the importance of SOM in mammalian development is unknown. The elucidation of its functional properties will depend on the identification of target genes and the study of loss of function mutants in mouse models. A range of human homologues of *grainyhead* target genes have been identified, many of which play important roles during embryogenesis. We are proceeding with protein–DNA interaction studies with recombinant SOM and candidate regulatory elements to identify these targets.

This work was supported by grants from the NHMRC (Australia), The Wellcome Trust (to S. M. J.), The Cancer Council of Victoria (to S. B. T.), National Institutes of Health (no. P01 HL53749-06), Cancer Center Support CORE (no. P30 CA 21765), the American Lebanese Syrian Associated Charities (ALSAC) and the Assisi Foundation of Memphis.

REFERENCES

1 Bray, S. J., Burke, B., Brown, N. H. and Hirsh, J. (1989) Embryonic expression of a family of *Drosophila* proteins that interact with a central nervous system regulatory element. Genes Dev. **3**, 1130–1145

2 Dynlacht, B. D., Attardi, L. D., Admon, A., Freeman, M. and Tjian, R. (1989) Functional analysis of NTF-1, a developmentally regulated *Drosophila* transcription factor that binds neuronal *cis* elements. Genes Dev. **3**, 1677–1688

3 Johnson, W. A., McCormick, C. A., Bray, S. J. and Hirsh, J. (1989) A neuron-specific enhancer of the *Drosophila dopa decarboxylase* gene. Genes Dev. **3**, 676–686

4 Bray, S. J. and Kafatos, F. C. (1991) Developmental function of Elf-1: an essential transcription factor during embryogenesis in *Drosophila*. Genes Dev. **5**, 1672–1683

5 Biggin, M. D. and Tjian, R. (1988) Transcription factors that activate the *Ultrabithorax* promoter in developmentally staged extracts. Cell (Cambridge, Mass.) **53**, 699–711

6 Soeller, W. C., Poole, S. J. and Kornberg, T. (1988) *In vitro* transcription of the *Drosophila engrailed* gene. Genes Dev. **2**, 68–81

7 Dynlacht, B. D., Hoey, T. and Tjian, R. (1991) Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. Cell (Cambridge, Mass.) **66**, 563–576

8 Attardi, L. D. and Tjian, R. (1993) *Drosophila* tissue-specific transcription factor NTF-1 contains a novel isoleucine-rich activation motif. Genes Dev. **7**, 1341–1353

9 Uv, A. E., Thompson, C. R. L. and Bray, S. J. (1994) The *Drosophila* tissue-specific factor *grainyhead* contains novel DNA-binding and dimerization domains that are conserved in the human protein CP2. Mol. Cell. Biol. **14**, 4020–4031

10 Huang, J.-D., Dubnicoff, T., Liaw, G.-J., Bai, Y., Valentine, S. A., Shirokawa, J. M., Lengyel, J. A. and Courey, A. J. (1995) Binding sites for transcription factor NTF-1/Elf-1 contribute to the ventral repression of *decapentaplegic*. Genes Dev. **9**, 3177–3189

11 Liaw, G.-J., Rudolph, K. M., Huang, J.-D., Dubnicoff, T., Courey, A. J. and Lengyel, J. A. (1995) The torso response element binds GAGA and NTF-1/Elf-1, and regulates tailless by relief of repression. Genes Dev. **9**, 3163–3176

12 Uv, A. E., Harrison, E. J. and Bray, S. J. (1997) Tissue-specific splicing and functions of the *Drosophila* transcription factor *Grainyhead*. Mol. Cell. Biol. **17**, 6727–6735

13 Jones, K. A., Luciw, P. A. and Duchange, N. (1988) Structural arrangements of transcription control domains within the 5′-untranslated leader regions of the HIV-1 and HIV-2 promoters. Genes Dev. **2**, 1101–1114

14 Wu, F. K., Garcia, J. A., Harrich, D. and Gaynor, R. B. (1988) Purification of the human immunodeficiency virus type 1 enhancer and TAR binding proteins EBP-1 and UBP-1. EMBO J. **7**, 2117–2129

15 Kato, H., Horikoshi, M. and Roeder, R. G. (1991) Repression of HIV transcription by a cellular protein. Science **251**, 1476–1478

16 Lim, L. C., Swendeman, S. L. and Sheffery, M. (1992) Molecular cloning of the α-globin transcription factor CP2. Mol. Cell. Biol. **12**, 828–835

17 Yoon, J.-B., Li, G. and Roeder, R. G. (1994) Characterization of a family of related cellular transcription factors which can modulate human immunodeficiency virus type I transcription *in vitro*. Mol. Cell. Biol. **14**, 1776–1785

18 Sueyoshi, T., Kobayasi, R., Nishio, K., Aida, K., Moore, R., Wada, T., Handa, H. and Negishi, M. (1995) A nuclear factor (NF2d9) that binds to the male-specific p450 (Cyp 2d-9) gene in mouse liver. Mol. Cell. Biol. **15**, 4158–4166

19 Huang, N. and Miller, W. L. (2000) Cloning of factors related to HIV-inducible LBP proteins that regulate steroidogenic factor-1-independent human placental transcription of the cholesterol side-chain cleavage enzyme, p450scc. J. Biol. Chem. **275**, 2852–2858

20 Rodda, S., Sharma, S., Scherer, M., Chapman, G. and Rathjen, P. (2001) CRTR-1, a developmentally regulated transcriptional repressor related to the CP2 family of transcription factors. J. Biol. Chem. **275**, 2852–2858

21 Wilanowski, T., Tuckfield, A., Cerruti, L., O'Connell, S., Saint, R., Parekh, V., Tao, J., Cunningham, J. M. and Jane, S. M. (2002) A highly conserved family of mammalian developmental transcription factors related to *Drosophila grainyhead*. Mech. Dev. **114**, 37–50

22 Jane, S. M., Nienhuis, A. W. and Cunningham, J. M. (1995) Hemoglobin switching in man and chicken is mediated by a heteromeric complex between the ubiquitous transcription factor CP2 and a developmentally specific protein. EMBO J. **14**, 97–105

23 Volker, J. L., Rameh, L. E., Zhu, Q., DeCaprio, J. and Hansen, U. (1997) Mitogenic stimulation of resting T cells causes rapid phosphorylation of the transcription factor LSF and increased DNA-binding activity. Genes Dev. **11**, 1435–1446

24 Zhou, W.-L., Clouston, D. R., Wang, X., Cerruti, L., Cunningham, J. M. and Jane, S. M. (2000) Induction of human fetal globin gene expression by a novel erythroid factor, NF-E4. Mol. Cell. Biol. **20**, 7662–7672

25 Fields, S. and Song, O. (1989) A novel genetic system to detect protein–protein interactions. Nature (London) **340**, 245–246

26 Amrolia, P. J., Ramamurthy, L., Saluja, D., Tanese, N., Jane, S. M. and Cunningham, J. M. (1997) The activation domain of the enhancer binding protein p45NF-E2 interacts with TAF$_{II}$130 and mediates long-range activation of the α and β-globin gene loci in an erythroid cell line. Proc. Natl. Acad. Sci. U.S.A. **94**, 10051–10056

27 Zhang, Y. W., Bae, S. C., Huang, G., Fu, Y. X., Lu, X., Ahn, M. Y., Kanno, Y., Kanno, T. and Ito, Y. (1997) A novel transcript encoding an N-terminally truncated AML1/PEBP2 αB protein interferes with transactivation and blocks granulocytic differentiation of 32Dcl3 myeloid cells. Mol. Cell. Biol. **17**, 4133–4145

28 Descombes, P. and Schibler, U. (1991) A liver-enriched transcriptional activator protein, LAP, and a transcriptional inhibitory protein, LIP, are translated from the same mRNA. Cell (Cambridge, Mass.) **67**, 569–579

29 O'Donovan, K. J. and Baraban, J. M. (1999) Major Egr3 isoforms are generated via alternate translation start sites and differ in their abilities to activate transcription. Mol. Cell. Biol. **19**, 4711–4718

30 Tanaka, M., Lai, J. S. and Herr, W. (1992) Promoter-selective activation domains in Oct-1 and Oct-2 direct differential activation of an snRNA and mRNA promoter. Cell (Cambridge, Mass.) **68**, 755–767

31 Attardi, L. D., Von Seggern, D. and Tjian, R. (1993) Ectopic expression of wild-type or a dominant-negative mutant of transcription factor NTF-1 disrupts normal *Drosophila* development. Proc. Natl. Acad. Sci. U.S.A. **90**, 10563–10567

Received 23 September 2002/19 November 2002; accepted 9 December 2002
Published as BJ Immediate Publication 28 January 2003, DOI 10.1042/BJ20021476