PART II
Introducció

En el marc del projecte EuroImage i per l'interès existent al departament en la regió q24-q26 del cromosoma 15 humà aquest segon apartat del treball està centrat en l'anàlisi transcripcional d'aquesta regió, en l'identificació de nous gens mapats entre q24 i q26, i en una caracterització preliminar dels mateixos a nivell de patró d'expressió, seqüència, estructura genòmica i exploració del seu paper potencial en funció de la seqüència aminoacídica predita. A més a més, l'anàlisi de gens a 15q24-q26 ha comportat l'observació de l'existència de paralogia amb la regió p13.3-p12 del cromosoma 19 humà. S'inclou l'anàlisi d'aquestes dues regions cromosòmiques amb l'objectiu d'obtenir una caracterització més profunda de les mateixes i de la seva relació des del punt de vista de l'evolució del genoma.

I. Cromosoma 15 humà. Reordenaments cromosòmics. Inestabilitat genòmica

Una de les característiques evidenciades els darrers anys és l'existència d'una freqüència significativament elevada de reordenaments i alteracions citogenètiques a nivell del braç llarg del cromosoma 15 humà. Alteracions d'aquest tipus poden donar lloc al que s'anomenen patologies o transtorns d'origen genòmic (Lupski, 1998b; 2003; Stankiewicz & Lupski, 2002) (Taula 5). És a dir, malalties causades per la pèrdua, guany o disrupció de l'integritat genòmica d'un gen o nombre de gens. Aquest tipus de desordres d'origen genòmic, ja sigui per deleció, duplicació, translocació o inversió, es distingeixen de les tradicionals patologies mendelianes en que el seu origen no són mutacions puntuals sinó que s'originen per reordenaments genòmics que afecten fragments relativament grans de DNA. Generalment, aquests reordenaments es produeixen per mecanismes de recombinació, en contraposició amb les mutacions puntuals, les quals usualment procedeixen d'errors de replicació o reparació. En el cas dels transtorns d'origen genòmic, s'han associat determinades estructures i seqüències genòmiques amb punts concrets de trencament i reordenament, suggerint l'existència d'una predisposició a reordenaments per la presència de patrons redundants de seqüència genòmica. Aquestes regions d'homologia creuada actuarien com a responsables de crear inestabilitat en el genoma i afavorir l'aparició de reorganitzacions genòmiques (Shaw & Lupski, 2004).

Taula 5. *Alguns exemples de transtorns genòmics humans i el reordenament cromosòmic al qual han estat associats (adaptat de (Emanuel & Shaikh, 2001)).*

Transtorn genòmic	***Reordenament***	***Localització***	***Tamany reordenament (Mb)***	***Referències***
Charcot-Marie-Tooth tipus 1A (CMT1A)	Duplicació intersticial	17p12	1.5	(Chance *et al.*, 1994; Lupski, 1998a)
Neuropatia hereditària (HNPP)	Deleció	17p12	1.5	(Chance *et al.*, 1994; Lupski, 1998a)
Síndrome de Smith-Magenis	Deleció	17p11.2	5	(Chen *et al.*, 1997)
Duplicació 17p11.2	Duplicació intersticial	17p11.2	5	(Potocki *et al.*, 2000)
Neurofibromatosi tipus I (NF1)	Deleció	17q11.2	1.5	(Dorschner *et al.*, 2000)
Síndrome Prader-Willi (PWS)	Deleció	15q11-15q13	4	(Amos-Landgraf *et al.*, 1999; Christian *et al.*, 1999)
Síndrome d'Angelman (AS)	Deleció	15q11-15q13	4	(Amos-Landgraf *et al.*, 1999; Christian *et al.*, 1999)
Duplicació invertida 15 (Huang *et al.*)	Cromosoma marcador extranumerari	15q11-15q14	4	(Huang *et al.*, 1997)
Síndrome de Williams-Beuren (WBS)	Deleció	7q11.23	1.6	(Perez Jurado *et al.*, 1998)
Síndrome de DiGeorge i velocardiofacial (DGS/VCFS)	Deleció	22q11.2	3	(Edelmann *et al.*, 1999)
Síndrome ull de gat (CES)	Cromosoma marcador extranumerari	22q11.2	3	(McTaggart *et al.*, 1998)
Ictiosi lligada al cromosoma X	Deleció	Xp22	1.9	(Ballabio & Andria, 1992)
Hemofília A	Inversió	Xq28	0.5	(Naylor *et al.*, 1996)

En el cas específic del cromosoma 15 humà existeixen diversos exemples de malalties o síndromes que es troben associades a alteracions a nivell genòmic. Són destacables les delecions a nivell de la regió 15q11-q13 presents en individus amb les síndromes de Prader-Willi i Angelman (PWS/AS) (Amos-Landgraf *et al.*, 1999; Christian *et al.*, 1999; Khan & Wood, 1999). De forma semblant, els cromosomes 15 dicèntrics identificats en certs casos de PWS constitueixen un exemple de la capacitat de reordenament d'aquest cromosoma (Webb *et al.*, 1995). La duplicació invertida [inv dup(15)] és el segon reordenament més comú que afecta el cromosoma 15 i dóna lloc a un cromosoma 15 extranumerari (Blennow *et al.*, 1995; Huang *et al.*, 1997). S'han identificat duplicacions proximals de 15q en casos d'autisme i individus amb graus variables de retard mental (Cook *et al.*, 1997). A la regió més proximal de 15q, s'han observat triplicacions intersticials en fenotips caracteritzats per alteracions mentals i motores (Schinzel *et al.*, 1994). S'han detectat també triplicacions, duplicacions i translocacions entre el cromosoma 15 i el cromosoma 7 en pacients afectats de dismorfologia lleu i atacs de tipus epilèptic (Bettelheim *et al.*, 1998; Jewett *et al.*, 1998). A 15q també s'han descrit tetrasomies distals associades a retard mental, hipotonia i alteracions morfològiques lleus (Blennow *et al.*, 1994; Rowe *et al.*, 2000). Delecions i duplicacions intersticials a nivell distal de 15q també han estat publicades (Browne *et al.*, 2000; Han *et al.*, 1999; Verma *et al.*, 1996). Finalment, un exemple clar i concret d'alteració a 15q associada a patologia és la translocació entre 15q25 i 12p13 present en pacients amb fibrosarcoma congènit (CFS) (Knezevich *et al.*, 1998).

Durant les darreres dècades el desenvolupament tecnològic ha possibilitat optimitzar la resolució de les anàlisis de la seqüència genòmica humana. L'ús d'eines citogenètiques com el bandejat cromosòmic per a l'identificació de reordenaments genòmics s'ha vist substituïda per tècniques més específiques i de més resolució com l'hibridació *in situ* fluorescent (FISH) o el pintat cromosòmic. Més recentment la tecnologia d'arrays-CGH (hibridació genòmica comparada sobre microarrays) s'ha començat a implementar amb èxit per identificar delecions i duplicacions genòmiques amb resolucions encara majors (Bruder *et al.*, 2001; Shaw *et al.*, 2004). Aquests estudis han evidenciat que la majoria de reordenaments genòmics no tenen lloc a l'atzar, sino que es tracta d'errors inherents als processos de manteniment d'un genoma tan complex com l'humà.

II. Origen i significació dels fenòmens de paralogia

La presència de gens paràlegs o de regions de paralogia en una mateixa espècie i en un moment de temps és un reflex de la història evolutiva del genoma de l'organisme. La paralogia entre seqüències acostuma a originar-se per duplicació o amplificació seguida d'un procés de divergència successiva més o menys dràstica en funció de la pressió selectiva que va sent exercida sobre aquelles seqüències. En canvi, es parla d'ortologia quan es fa referència a l'existència d'homologia entre seqüències després d'un procés d'especiació, per tant homologia entre seqüències corresponents al mateix gen en espècies diferents.

Els mecanismes principals per generar fenòmens de paralogia en vertebrats són les duplicacions regionals i els processos de tetraploidització (a partir de dues duplicacions del genoma sencer) (Ohno *et al.*, 1968). Els parells de gens paràlegs derivats d'un gen ancestral comú, estan sotmesos a diferents pressions selectives que determinaran la progressiva divergència entre ells. Els grups de gens paràlegs acaben formant algunes de les nombroses famílies i subfamílies de gens que s'han anat descobrint amb la disponibilitat de la seqüència genòmica dels organismes. El grau de conservació i similitud entre gens d'un mateix grup paràleg és variable, i la seva identificació i classificació pot basar-se en les identitats de seqüència, o fins i tot en alguns casos, en dades funcionals. Les dificultats principals per a identificar grups paràlegs

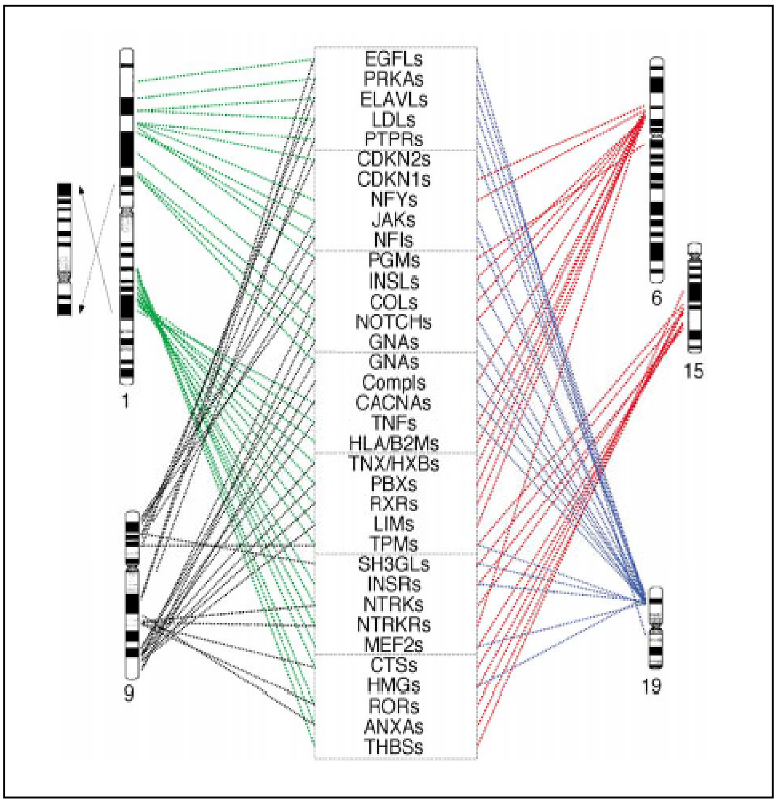


Figura 5. *Paralogia entre els cromosomes humans 1, 6, 9, 15 i 19. Adaptat de (Lundin et al., 2003)*

recauen en els fenòmens de silenciament diferencial de gens per donar lloc a pseudogens (gens no actius que poden retenir durant cert temps la seva seqüència i estructura original, i patir divergència com a conseqüència d'una pressió selectiva diferenciada) i en els reordenaments genòmics (Lundin, 1979; 1993).

Els estudis portats a terme fins ara que han analitzat regions paràlogues suggereixen que hi ha hagut un mínim de dues duplicacions genòmiques en les etapes inicials de l'evolució dels vertebrats. Els grans canvis morfològics a nivell d'òrgans, com en el cas del sistema nerviós, per exemple, coincideixen amb increments en bloc del nombre de gens al genoma. La presència de regions de paralogia al genoma humà és un reflexe d'aquest augment (Lundin *et al.*, 2003). Un exemple descrit de grup paràleg està constituït per regions dels cromosomes 1, 6p, 9, 15 i 19p humans (Figura 5) (Lundin *et al.*, 2003). En aquest cas s'hi impliquen nombrosos gens pertanyents al complexe major d'histocompatibilitat (MHC) (Katsanis *et al.*, 1996). Un altre exemple de paralogia al genoma humà inclou els cromosomes 2q, 7, 12q13 i 17q, on hi destaquen els grups de gens Hox entre d'altres (Lundin, 1993). L'existència de paralogia ha estat descrita pels segments dels cromosomes humans 11, 15 i 19. S'ha proposat que han derivat d'un cromosoma ancestral comú conjuntament amb regions sintèniques en els cromosomes 7 i 9 de ratolí (Seldin *et al.*, 1991).

En qualsevol cas, l'estudi de regions paràlogues conjuntament amb les corresponents regions ortòlogues en espècies allunyades evolutivament en major o menor grau, pot permetre conèixer la història evolutiva del genoma i dels cromosomes dels vertebrats.

III. Duplicacions segmentàries

L'observació relativament recent de que patologies específiques com algunes mencionades en els apartats anteriors són causades per reordenaments cromosòmics recurrents, ha indicat la possibilitat de que l'inestabilitat genòmica i predisposició a reorganitzacions del DNA d'aquestes regions estigui directament relacionada amb l'estructura i seqüència de les mateixes (Emanuel & Shaikh, 2001; Ji *et al.*, 2000; Lupski, 1998b). Els reordenaments que impliquen regions grans del genoma poden dividir-se en funció de la complexitat i tamany de les regions flanquejants. En qualsevol dels casos, els reordenaments poden comportar que els nivells d'expressió de nombrosos gens estiguin afectats i produeixin determinats fenotips.

Taula 6. *Presència de duplicacions segmentàries al genoma humà. Dades basades en la seqüència genòmica del Juny del 2002. S'analitzen duplicacions segmentàries de més de 5 kb i amb identitats de seqüència majors al 90%. Adaptat de (Cheung et al., 2003).*

Cromosoma	*Tamany (pb)*	*Duplicacions intracromosòmiques (pb)*	*Duplicacions intercromosòmiques (pb)*	*Duplicacions totals (pb)*	*Errors*	*% cromosomes*
1	246,874,334	5,278,549	2,854,898	7,056,274	4,369,406	1.8
2	240,681,600	4,917,160	3,298,723	6,892,585	2,311,522	1.0
3	194,908,136	2,128,493	1,654,201	3,146,570	3,979,610	2.0
4	192,019,378	2,599,650	2,164,382	4,061,432	2,482,740	1.3
5	180,966,400	3,519,480	1,464,945	4,530,406	2,297,998	1.3
6	170,309,517	2,358,252	743,875	2,877,392	569,918	0.3
7	157,432,793	8,636,434	2,614,326	10,139,669	205,130	0.1
8	143,874,322	2,318,984	1,125,241	2,612,280	3,956,756	2.8
9	132,438,756	7,248,232	4,801,871	8,341,767	1,589,734	1.2
10	134,416,750	5,279,301	1,375,341	6,334,458	1,250,157	0.9
11	137,442,545	3,622,080	1,670,412	4,363,619	2,028,875	1.5
12	131,300,572	1,894,547	971,490	2,816,187	3,383,730	2.6
13	113,446,104	918,255	1,202,102	1,855,806	146,198	0.1
14	104,324,908	531,219	820,880	1,335,177	13,814	0.0
15	99,217,355	4,593,233	2,344,618	5,634,201	1,739,894	1.8
16	81,671,585	4,917,218	2,228,116	6,012,178	2,113,843	2.6
17	80,052,782	4,775,137	646,968	5,274,195	2,145,614	2.7
18	77,516,809	525,636	700,654	1,226,290	1,443,775	1.9
19	60,013,307	2,700,984	704,757	3,156,687	335,190	0.6
20	62,842,997	592,441	873,152	1,052,248	147,940	0.2
21	44,626,493	481,879	1,303,776	1,504,333	0	0.0
22	47,748,585	1,741,766	1,374,363	2,770,386	0	0.0
X	14,924,9818	2,625,206	2,927,714	5,518,712	2,185,046	1.5
Y	58,368,225	5,959,836	3,524,276	8,461,355	56,204	0.1
Sense mapar	1,391,854	179,709	378,110	407,013	116,923	8.4
Total	3,043,135,925	80,343,681	43,769,191	107,381,220	4,369,406	1.8

La complexitat dels fenotips observats associats a anomalies genòmiques relativament grans suggereix un paper directe pels gens inclosos en la regió problema, però alhora un efecte genòmic global degut a alteracions en la regulació d'altres gens relacionats. Observacions més recents han detectat la presència de seqüències repetides complexes d'1 a 500 kb a nivell de punts de trencament de reordenaments genòmics i també immerses en les pròpies seqüències que pateixen el reordenament. Les duplicacions segmentàries poden resultar de la duplicació de regions del genoma representant gens, pseudogens, grups de gens contigus o altres fragments cromosòmics. La freqüència amb la que es troben seqüències d'aquestes característiques al genoma humà s'estima entre el 3'5 i 5% (Cheung *et al.*, 2001; Samonte & Eichler, 2002). Es distribueixen de forma no uniforme, apareixent significativament concentrades a nivell de regions concretes del genoma, especialment a les regions pericentromèriques i subtelomèriques (Shaw & Lupski, 2004) (Taula 6). Les identitats entre aquestes seqüències superen el 95% i poden arribar a ser del 99%, fet que constitueix una de les principals dificultats per a aconseguir caracteritzar-les i determinar amb exactitud la seva seqüència i localització (Chen *et al.*, 1997; Lupski, 1998b; Shaw & Lupski, 2004).

Els estudis sobre la presència de reordenaments cromosòmics a nivell del braç llarg del cromosoma 15 han identificat seqüències repetides de tamanys de fins a 60 kb anomenades LCR15. Han estat localitzades a 15q11-q13, 15q22, 15q24, 15q25 i 15q26, i presenten identitats significatives entre elles i amb seqüències presents a altres cromosomes (6q, 7p i 12p) (Gratacos *et al.*, 2001; Pujana *et al.*, 2001) (Figura 6). S'ha postulat que existeixen com a mínim 30 copies d'aquestes seqüències a 15q. En el cas concret d'aquestes seqüències al cromosoma 15 es parla de LCR15 (low copy repeats 15).

Els duplicons LCR15 presenten una mida variable entre 13 i 60 kb, amb unes identitats de seqüència superiors al 90% i contenen seqüències amb similaritat significativa a tres gens humans ja descrits: gens golgina-like (GLP), a un gen codificant per una proteïna amb dominis SH3 (SH3P18) i al gen de la dinamina 1 (DNM1) (Gratacos *et al.*, 2001; Pujana *et al.*, 2001). La presència de gens o pseudogens ha estat descrita en moltes de les duplicacions segmentàries estudiades fins al moment, són agents potenciadors de fenòmens de recombinació. L'eucromatina que constitueix el DNA que s'expressa habitualment (gens, pseudogens) es troba menys condensada, més oberta i per tant, presenta una major predisposició a patir reorganitzacions. Aquest fet afegit a una pressió selectiva determinada i a fenòmens de conversió gènica afavorint la conservació de seqüència dels gens que s'hi troben inclosos comporta que aquestes regions genòmiques siguin considerades punts calents de recombinació genòmica (Chen *et al.*, 1997; Lupski, 1998b). La localització d'aquestes seqüències a 15q suggereix que poden tenir un paper en la generació de reordenaments cromosòmics associats a anomalies genòmiques del cromosoma 15 humà com, per exemple, l'autisme entre d'altres (Han *et al.*, 1999; Silva *et al.*, 2002).

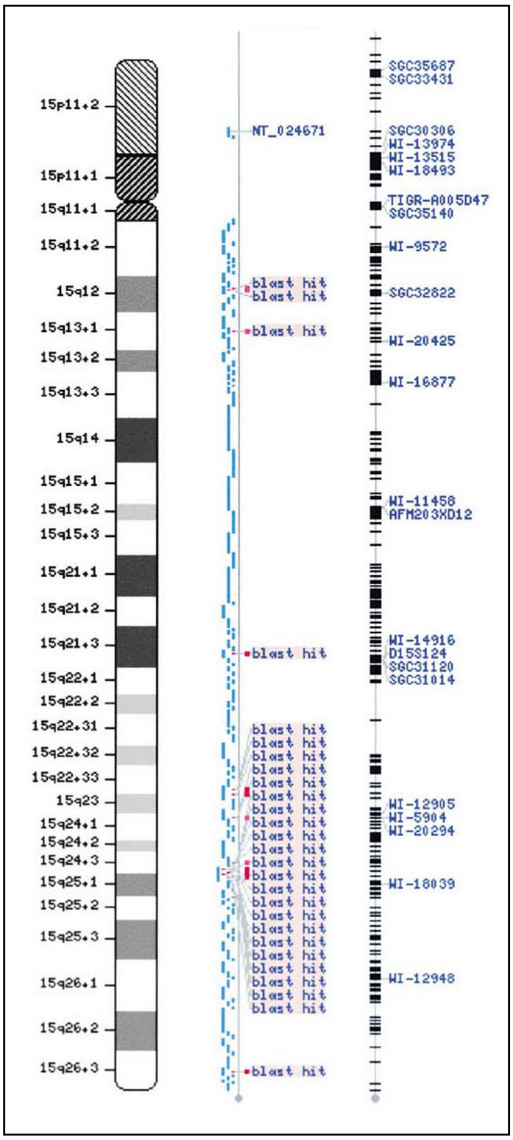


Figura 6. *Distribució de duplicons identificats al braç llarg del cromosoma 15 humà. Adaptat de (Gratacos et al., 2001)*

IV. Transtorn d'ansietat associat a 15q24-q26

L'ansietat en humans és una resposta de protecció enfront l'adversitat. Es caracteritza per una sèrie de respostes específiques del sistema nerviós autònom i per comportaments d'autodefensa. Una reacció ansiosa excessiva pot esdevenir un desavantatge i comprometre greument la qualitat de vida de l'individu. Es distingeixen sis formes de transtorn d'ansietat: el transtorn de pànic, l'ansietat generalitzada, la fòbia social, la fòbia específica, el transtorn per estrès post-traumàtic i els transtorns obsessivo-compulsius (American Psychiatric Association. & American Psychiatric Association. Task Force on DSM-IV., 2000).

Nombrosos estudis han demostrat l'existència de factors familiars en la manifestació dels transtorns d'ansietat (Fyer *et al.*, 1995; Noyes *et al.*, 1987). L'existència d'un cert grau d'heretabilitat s'ha evidenciat mitjançant comparacions entre bessons. S'ha aconseguit detectar una major concordància de fenotip en parelles de bessons monozigòtics en comparació amb els bessons dizigòtics (Kendler *et al.*, 1995; Perna *et al.*, 1997). En referència al risc de patir un transtorn d'ansietat, aquests estudis han establert que els factors genètics expliquen entre el 30% i 50% de la variabilitat que existeix entre individus (Hettema *et al.*, 2001; Kendler, 2001). La variabilitat restant ha de ser explicada per factors ambientals específics de cada individu. A més a més, els estudis de risc han confirmat l'existència d'heretabilitat en aquest tipus de transtorn atribuïnt valors de risc a patir transtorns d'ansietat 4 a 10 vegades majors en individus amb familiars afectats respecte a individus sense antecedents familiars de la malaltia (Crowe *et al.*, 1983; Hettema *et al.*, 2001; Starcevic *et al.*, 1993; Weissman *et al.*, 1997).

Els estudis de lligament global del genoma per tal d'identificar gens o regions genòmiques implicades en l'evolució d'aquest tipus de transtorn no han obtingut resultats concrets o concloents (Crowe *et al.*, 2001; Knowles *et al.*, 1998; Weissman, 1993). Tot i això, alguns estudis han identificat regions de susceptibilitat al cromosoma 7p (Crowe *et al.*, 2001; Logue *et al.*, 2003) i al cromosoma 9q (Thorgeirsson *et al.*, 2003). S'ha detectat cosegregació del transtorn de panic amb síndromes que inclouen cistitis, desordres tiroideus o migranya crònica (Weissman *et al.*, 2004). S'han obtingut dades d'associació significativa entre el transtorn de pànic i marcadors del cromosoma 13 (Weissman *et al.*, 2004), així com amb el receptor d'adenosina 2A del cromosoma 22 (Hamilton *et al.*, 2004).

Als anys 90, l'observació d'una major incidència dels transtorns d'ansietat en un estudi de famílies afectades per laxitud articular va suggerir que podia existir un mateix locus o regió cromosòmica responsable per a les dues afectacions (Bulbena *et al.*, 1993; Martin-Santos *et al.*, 1998). L'anàlisi de set genealogies en les quals es detectava cosegregació de transtorns d'ansietat i laxitud articular va donar lloc l'any 2001 a l'identificació d'associació entre els transtorns d'ansietat i la presència d'una duplicació de la regió q24-q26 del cromosoma 15 humà (Gratacos *et al.*, 2001). Aquest polimorfisme genòmic va ser identificat mitjançant tècniques citogènetiques (FISH), i va permetre distingir diferents tipus de duplicacions (centromèriques, telomèriques, invertides, directes). L'ocurrència de mutacions *de novo* i la presència de mosaicisme en graus variables van fer proposar un model no-mendelià com a patró d'herència d'aquest factor de susceptibilitat. La duplicació inicialment reportada es va anomenar DUP25 i també es va detectar en el 7% de la població general (Gratacos *et al.*, 2001). S'ha estimat que aquesta regió té una mida d'aproximadament 17 Mb, que conté de 50 a 60 gens i que es caracteritza per un elevat grau d'inestabilitat genòmica degut a la presència de duplicacions segmentàries flanquejants i al llarg de tot 15q, tal com s'ha mencionat anteriorment (Pujana *et al.*, 2001).

L'observació d'associació entre 15q24-q26 i els transtorns d'ansietat fou de gran interès ja que es tractava de la primera associació identificada entre un polimorfisme genòmic d'aquest estil i una malaltia psiquiàtrica amb una incidència poblacional tan significativa. A més a més, en aquest cas s'aconseguia identificar un nou tipus de mutació genòmica associada a malaltia que no estaba lligada a cap dels loci contigus. L'importància i rellevància dels resultats obtinguts per Gratacòs i col.laboradors va impulsar diversos estudis d'associació entre 15q24-q26 i transtorns mentals en poblacions diferents. S'han aplicat tècniques de FISH, PCR quantitativa i hibridació (MAPH) per tal de dur a terme els mateixos estudis en altres poblacions i famílies afectades de transtorn d'ansietat (Hollox & Armour, 2003; Schumacher *et al.*, 2003; Tabiner *et al.*, 2003; Weiland *et al.*, 2003; Zhu *et al.*, 2004). Fins aquest moment, cap dels estudis publicats ha estat capaç de replicar en altres poblacions els resultats obtinguts a la població catalana usada en el primer estudi per Gratacòs *et al*. El per què de la no replicació de resultats pot ser indicatiu de la dificultat en la detecció de la duplicació mitjançant FISH, de l'existència de diversos graus de mosaicisme entre individus i línies cel.lulars, dels patrons d'herència no mendelians, de les observacions d'inestabilitat genòmica en la regió i de la possibilitat de que no es tracti d'una duplicació *per se* (veure Discussió).

De tota manera, ja sigui per la potencial associació a transtorns d'ansietat, per la naturalesa inestable de la regió o com a contribució imprescindible a l'obtenció de la seqüència completa del genoma humà i els gens que hi són continguts, la regió q24-q26 ha esdevingut una zona d'alt interès per a ser caracteritzada a nivell transcripcional i de contingut gènic. És per això que, com a part del Consorci EuroImage i per la nostra implicació en l'identificació de gens nous humans (veure part I), part dels esforços del nostre grup es van adreçar cap a l'identificació de gens mapats prèviament a la regió 15q24-q26 segons les bases de dades públiques.

PART II: Objectius

En el marc del consorci EuroImage, identificació de nous gens humans continguts en la regió q24-q26 del cromosoma 15 humà i conseqüent caracterització a nivell de seqüència nucleotídica, patró d'expressió i homologia en altres espècies

Identificació i anàlisi de paralogia entre les regions q24-q26 del cromosoma 15 i p13.3-p12 del cromosoma 19 humans

Identificació i caracterització de nous gens humans localitzats a la regió p13.3-p12 del cromosoma 19

PART II
Resultats

Resultats

Com a membres del consorci EuroImage i com a resultat de l'enfoc i concentració en una regió genòmica específica, la regió q24-q26 del cromosoma 15 humà, es va avançar significativament en l'identificació i caracterització de diversos gens, la majoria dels quals es presenten a les publicacions que constitueixen els apartats següents d'aquest treball. L'anàlisi transcripcional de la regió ha permès identificar l'existència de paralogia entre 15q24-q26 i la regió p13.3-p12 del cromosoma 19, així com aprofundir en la seva naturalesa i la seva significació a nivell evolutiu. Aquest últim punt queda reflectit en els articles sobre gens del cromosoma 15 amb paràlegs al 19 i els treballs identificant gens nous a la mencionada regió del cromosoma 19.

I. Identificació, expressió i mapatge del gen humà C15orf4

En aquest cas es presenta l'identificació i caracterització d'un nou gen humà amb similitud de seqüència a la proteïna YmL30 dels ribosomes mitocondrials de llevat. Les dades obtingudes de mapatge del gen C15orf4 confirmen la seva presència dins de la regió 15q24-q26. S'identifiquen els gens ortòlegs en altres espècies confirmant l'existència real d'aquest gen amb una funció biològica predita conservada al llarg de l'evolució. L'expressió de C15orf4 presenta uns nivells basals ubicus i un enriquiment a testicle, suggerint una possible funció específica a nivell de teixit.

Ref. *DNA Seq. 2001;12(2):91-6*

Cloning, mapping and expression analysis of C15orf4, a novel human gene homologous to the yeast mitochondrial ribosomal protein YmL30 gene

Laura Carim, Lauro Sumoy, Nuria Andreu, Xavier Estivill and Mònica Escarceller

Medical and Molecular Genetics Center,
Institut de Recerca Oncològica,
Hospital Duran i Reynals, Autovia de Castelldefels km 2,7
L'Hospitalet de Llobregat, 08907 Barcelona, Spain

Correspondence: Mònica Escarceller
Phone: 34-93-260-7775
Fax: 34-93-260-7776
e-mail: mescarceller@iro.es

Running title:

C15orf4 homologous to yeast YmL30

Abstract

We have identified C15orf4, a novel human gene showing homology to the yeast mitochondrial ribosomal protein YmL30. C15orf4 encodes a transcript of 1,006 nt with an ORF of 279 amino acids and a predicted protein size of 31.7 kDa. Expression of C15orf4 is enriched in testis. C15orf4 was positioned to chromosome 15q24 by radiation hybrid mapping. We have identified the C15orf4 mouse orthologue as well as homologues in other species.

Keywords: C15orf4, EUROIMAGE, chromosome 15q24, ribosomal YmL30.

Introduction

Within the EUROIMAGE full-length cDNA sequencing project underway in our laboratory (Adams et al 1991; Deloukas et al. 1998; Lennon et al. 1996; Schuler et al. 1998) we sequenced cDNA clones corresponding to EST clusters with the aim of identifying new genes. Clusters were selected on the basis of clone size, chromosomal localization and tissue distribution of transcripts. The EST contigs were built and analyzed *in silico* and representative clones were chosen for sequencing.

Using this approach, we have characterized a new gene, C15orf4, which shows a significant similarity to a putative ORF in *Drosophila melanogaster,* to the "decoy" gene in *Arabidopsis thaliana* and to the mitochondrial ribosomal protein (MRP) YNL252c gene in *Saccharomyces cerevisiae.*

The yeast ORF YNL252c gene is the synonym of YmL30 mitochondrial ribosomal protein of the large subunit (Graack and Wittmann-Liebold, 1998; Kitakawa et al, 1997; Sen-Gupta et al, 1997). Mitochondrial ribosomal proteins (MRPs) are the counterparts in that organelle of the cytoplasmic ribosomal proteins in the host. MRPs fulfill similar functions in protein biosynthesis but they are distinct in number, features and primary structures from the cytoplasmic riboproteins. To date, most of the information about MRPs has been obtained from the yeast *S. cerevisiae* and 50 different MRPs have been determined, although biochemical data and mutational analysis propose a total number substantially higher. Ribosomes of all species contain a core of structurally and functionally conserved riboproteins. Additional non-conserved proteins are considered to maintain functions specific for the respective species. The function of the MRPs may go beyond the mere biosynthesis of the small number of proteins encoded by the mitochondrial DNA. The characterization of the complete set of human MRPs and the elucidation of their role is a process at its beginning (Graack and Wittmann-Liebold, 1998).

We present here the cloning, mapping and expression analysis of C15orf4, a novel human gene homologous to the yeast MRP YmL30 gene.

Material and Methods.

cDNA isolation and sequence analysis

ESTs from UniGene cluster Hs.14018 (http://www.NCBI.nlm.nih.gov/UniGene) were assembled using the EST CAP assembly program (http://gcg.tigem.it/cgi-bin/uniestass.pl) and Sequencher (GeneCodes) sequence assembly software. Clones were obtained from the EUROIMAGE distribution centers. Sequence was determined by primer walking using the Perkin-Elmer BigDye reagents on an ABI PRISM-377 fluorescent automated sequencer and custom synthesized sequencing primers (LifeTech).

Full-length cDNA sequence was obtained using the rapid amplification of cDNA ends (RACE) method on SMART™ RACE cDNA from adult human placenta (Clontech), according to the manufacturer's kit instructions. The following primers were used: G1 (5' TGTTCAGGGATGTTCGGTCA 3'), G2 (5' CTTTCTTCTTTGCCAGTCGC 3') and G3 (5' GCTGTTAGGGGTGGCGG 3') for 5' C15orf4 extension. PCR extended products were subcloned into the pGEM-T easy vector (Promega) and sequenced as above. We generated nine independently generated extended clones to determine the cDNA ends.

Sequence comparisons were performed using ClustalW 1.7 (http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html).

Boxed multiple sequence alignments were obtained with the BOXSHADE 3.21 program (http://www.ch.embnet.org/software/BOX_form.html). To search for known motifs or functional domains, protein pattern and domain databases consulted were Prosite, SMART and Pfam (http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-domain.html).

The human C15orf4 nucleotide and protein sequences are available in GenBank under Acc. No. AF210056 and the mouse C15orf4 othologue ones under Acc. No. AF217090. The C15orf4 name has been approved by the Human Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/).

Northern blot analysis

A multiple-tissue northern blot (MTN II blot, Clontech) was hybridized to a 1-kb *Eco*RI-*Not*I restriction product corresponding to the cDNA insert from IMAGE clone 259218; and to a 2-kb ß-actin cDNA supplied commercially (Clontech) as control for quantification. Probes were labeled using a random primer DNA labeling kit (Amersham Pharmacia). Blots were hybridized overnight at 65ºC in ExpressHyb solution (Clontech) and washed at 68ºC in 0.2XSSC/0.5%SDS.

C15orf4 radiation hybrid mapping

To precisely localize the C15orf4 gene we used the Stanford TNG4 whole genome radiation hybrid panel (Stewart et al. 1997). Two point linkage analysis was performed using the RHMAP-2.0 on the RH Server at the Stanford Human Genome Center (http://www-shgc.stanford.edu/RH/index.html). We used primers corresponding to STS marker SHGC-15202 (D15S1261): F (5’

TCTAATCCCAGACTTGTCTGAGC 3’) and R (5’ TGTGGGTCACTAAGGATGAGC 3’). The PCR conditions were 1 cycle at 94ºC for 1 min 30 s; 30 cycles at 94ºC for 15 s, 62ºC for 23 s and 72ºC for 30 s; and 1 cycle at 72ºC for 5 min.

BAC assignment was obtained through BLAST searching against the BAC ends database at TIGR (http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_search.html) and contiguous BACs were determined from the ctc.ace clone tracking database (http://genome.wustl.edu/gsc/cgi-bin/ace/ctc_choices/ctc.ace).

Results and Discussion

Cloning of C15orf4

In our effort to identify new genes, we constructed and analyzed *in silico* unique gene EST clusters on the basis of clone size, chromosomal localization and tissue expression. A partial human cDNA sequence with a single open reading frame (ORF) named C15orf4 was identified during the analysis of EST clusters within the physical region in 15q24.

The UniGene cluster representative of the clones was Hs.14018. Human IMAGE cDNA clones whose ESTs extended most 5’ and 3’ in the cDNA were chosen for sequencing: 1963245 (GenBank Acc.No. AI355098), 1723436 (GenBank Acc.No. AI188252) and 259218 (GenBank Acc.No. N29438).

Since the clones did not cover the entire transcript, the full-length cDNA sequence was obtained by 5' RACE extension (see Methods). The assembly of both IMAGE and RACE clones gave as a result a total transcript length of 1,006 bp (including the polyA tail), with an ORF (from nt 12 to 849) encoding a 279 amino acid product with a calculated mass of 31.7 kDa and a pI=6.6. A polyadenylation signal (AATAAA) was observed at nt 967 and a polyA tail at the end (987 nt).

Analysis with protein domain identification software did not reveal the presence of any important feature with the exception of a coiled coil domain from amino acids 66th to 88th.

At the protein level, the most significative hit obtained after BLAST homology searching against “non redundant” databases (TBLASTN program at NCBI (http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/); Altschul et al. 1997) was a Drosophila translated genomic fragment in chromosome 3L, region 62A10-62B5 (GenBank Acc. No. AC005557). Further search in the Drosophila databases to find out if the genomic region could correspond to a characterized gene gave no positive result (Berkeley Drosophila Genome Project, http://www.fruitfly.org/;http://www.sanger.ac.uk/Projects/D_melano gaster/). We propose that the level of homology between both proteins (39% identity and 57% similarity within the 230 amino acid alignment) is suggestive of the existence of a distant C15orf4 gene homologue in the fruit fly.

The next significative BLAST homology hits were the *A. thaliana* “decoy” mRNA (GenBank Acc.No. U87586) and the *S. cerevisiae* ORF YNL252c gene (GenBank Acc.No. Z71528). Within the regions aligning with the human C15orf4 gene, the Arabidopsis and yeast ORFs showed 42% and 31% identity; 55% and 45% similarity, respectively.

The *Arabidopsis* “decoy” gene was described by Zhang and Somerville when analyzing the phenotypical defects in early embryogenesis caused by the twn2-1 mutation (Zhang and Somerville, 1997). This new gene encoded a protein of 19.1 kDa with low sequence homology to the yeast mitochondrial ribosomal protein, YmL30. The authors pursued no further research on the gene and at present the function of the protein remains unknown.

The yeast ORF YNL252c gene is the synonym of YmL30 ribosomal protein of the large subunit (Munich Information Center for Protein Sequences; http://www.mips.biochem.mpg.de/proj/yeast/). Its complete ORF had an estimated molecular mass of 30 kDa and the mature form is calculated to be 16 kDa by electrophoresis. It localizes in the mitochondria and null mutants show slow growth and slightly increased thiabendazole sensitivity (Graack and Wittmann-Liebold, 1998; Kitakawa et al, 1997; Sen-Gupta et al, 1997). It has not been determined whether its function is essential or not for mitochondrial function (Graack and Wittmann-Liebold, 1998).

BLAST searches against “mouse” and “other” EST databases identified possible C15orf4 homologues in other species. We selected two murine IMAGE clones which once sequenced, completed the entire cDNA transcript in mouse: 1891319 (GenBank Acc.No. AI266903) and 2123770 (GenBank Acc.No. AI930359). The mouse protein showed a remarkably high level of homology with its human counterpart: 80% identity and 87% similarity (Fig. 1).

We were also able to identify the partial sequence of close C15orf4 homologues in other species: *Rattus sp.* (83% identity, 90% similarity in the aligned regions); the zebrafish *Danio rerio* (58% identity, 68% similarity); the trematode *Schistosoma mansoni* (33% identity, 52% similarity); and a single EST in Drosophila corresponding to the genomic clone above described (GenBank Acc.No. AI387313) (Fig.1).

The identification of the mouse homologue, as well as the possibility of the protein being represented in a broad range of species seems to be in agreement with it holding an essential role in eukaryotic cells. One attractive possibility is that C15orf4 could be a novel MRP member of the mitochondrial ribosome.

Expression of C15orf4

Expression studies of C15orf4 with northern blots of human tissues (MTN II blot, Clontech), were carried out by hybridizing with a specific probe (see Methods). In adult tissues, basal expression was largely ubiquitous (Fig. 2), showing a 1.1-kb mRNA species. Remarkably, C15orf4 transcript signal was highly enriched in testis. The localized expression pattern suggests a tissue specific role in humans for C15orf4.

Mapping of C15orf4

Chromosomal localization of the human C15orf4 gene was determined by radiation hybrid mapping using the Stanford TNG4 panel. The gene was linked to STS SHGC-101328 with a lod score of 8.2 at an approximate distance of 120 kb. This STS is contained in RPCI-11 BAC clones 97O12 and 1127D9, in agreement with the fact that D15S1261, the marker contained in UniGene cluster Hs.14018, is contained in BAC 173D20 which overlaps with BACs 97O12 and 1127D9, in a physical contig mapped in 15q24 (D15S151-D15S202).

Acknowledgments

We are grateful to A. Puig and D. Otero for technical support with DNA sequencing. We wish to thank the HGMP Resource Center in Hinxton, UK, and the RZPD in Berlin, Germany, for supplying us with IMAGE cDNA clones. This work has been supported by EU Biomed Project No. BMH4-CT97-2284 to X. E and by CICYT-IN95-0347. M.E. is funded by the Spanish Ministry of Education (CIDYT contract FPI-070-97) and L.S. by the Catalan autonomous government (CIRIT-RED contract 1998-64).

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A., Olde B, Moreno RF, et al.: Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651-6 (1991).

Altschul SF, Maden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman, DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search program. Nucleic Acids Res 25, 3389-402 (1997).

Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L, Matise TC, McKusick KB, Beckmann JS, Bentolila S, Bihoreau M, Birren BB, Browne J, Butler A, Castle AB, Chiannilkulchai N, Clee C, Day PJ, Dehejia A, Dibling T, Drouot N, Duprat S, Fizames C, Bentley DR, et al.: A physical map of 30,000 human genes. Science 282: 744-746 (1998).

Graack HR and Wittmann-Liebold B. Mitochondrial ribosomal proteins (MRPs) of yeast. Biochem J 329: 433-48 (1998). Rewiew.

Kitakawa M, Graack HR, Grohmann L, Goldschmidt-Reisin S, Herfurth E, Wittmann-Liebold B, Nishimura T, Isono KSen-Gupta. Identification and characterization of the genes for mitochondrial ribosomal proteins of Saccharomyces cerevisiae. Eur J Biochem 245:449-56 (1997).

Lennon G, Auffray C, Polymeropoulos M, Soares MB: The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. Genomics 33:151-2 (1996).

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birre BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Hudson TJ, et al.: A gene map of the human genome. Science 274: 540-6 (1996).

Sen-Gupta M, Guldener U, Beinhauer J, Fiedler T, Hegemann JH. Sequence analysis of the 33 kb long region between ORC5 and SUI1 from the left arm of chromosome XIV from Saccharomyces cerevisiae. Yeast 13(9):849-60 (1997).

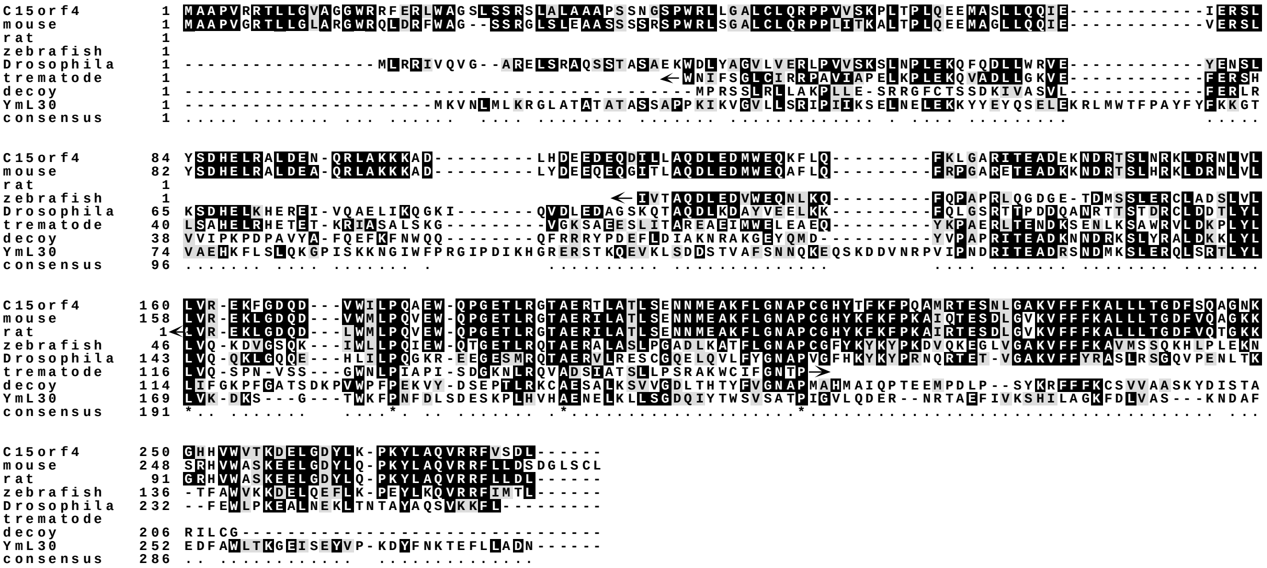
Stewart, EA, McKusick, KB, Aggarwal, A, Bajorek, E, Brady, S, Chu, A, Fang, N, Hadley, D, Harris, M, Hussain, S, Lee, R, Maratukulam, A, O'Connor, K, Perkins, S, Piercy, M, Qin, F, Reif, T, Sanders, C, She, X, Sun, WL, Tabar, P, Voyticky, S, Cowles, S, Fan, JB, Cox, DR, and et al.: An STS-based radiation hybrid map of the human genome. Genome Res 7:422-33 (1997).

Zhang JM and Somerville CR. Suspensor-derived polyembryony caused by altered expression of valyl-tRNA synthetase in the twn2 mutant of Arabidopsis. Proc Natl Acad Sci U S A 94:7349-55 (1997).

Figure legends.

Figure 1. Multiple sequence alignments of human C15orf4 protein and its homologues in mouse, rat (partial ORF), zebrafish (partial), Drosophila, the trematode *S. mansoni* (partial), the *A. thaliana* "decoy" protein and the yeast YmL30. Partial sequences are indicated by arrows. Identical residues are printed in reverse type and similar residues are shaded. Consensus sequence is shown at the bottom, identical amino acids are marked with asterisks and similar amino acids, in at least three species, with dots.

Figure 2. Multiple-tissue northern blot analysis of C15orf4. The 1-kb *Eco*RI-*Not*I restriction product was used as a probe revealing a testis enriched 1.1-kb mRNA species. C15orf4 and β–actin control transcripts are labeled.



```
C15orf4      1 MAAPVRRTLLGVAGGWRRFERLWAGSLSSRSLALAAAPSSNGSPWRLLGALCLQRPPVVSKPLTPLQEEMASLLQQIE------------IERSL
mouse        1 MAAPVGRTLLGLARGWRQLDRFWAG--SSRGLSLEAASSSSRSPWRLSGALCLQRPPLITKALTPLQEEMAGLLQQIE------------VERSL
rat          1
zebrafish    1
Drosophila   1 -----------------MLRRIVQVG--ARELSRAQSSTASAEKWDLYAGVLVERLPVVSKSLNPLEKQFQDLLWRVE------------YENSL
trematode    1                                            ←WNIFSGLCIRRPAVIAPELKPLEKQVADLLGKVE------------FERSH
decoy        1 ---------------------------------------------MPRSSLRLLAKPLLE-SRRGFCTSSDKIVASVL------------FERLR
YmL30        1 ----------------------MKVNLMLKRGLATATATASSAPPKIKVGVLLSRIPIIKSELNELEKKYYEYQSELEKRLMWTFPAYFYFKKGT
consensus    1 ..... ....... ... ......  .... ........ ....... ............. . .... .........            .....

C15orf4     84 YSDHELRALDEN-QRLAKKKAD---------LHDEEDEQDILLAQDLEDMWEQKFLQ---------FKLGARITEADEKNDRTSLNRKLDRNLVL
mouse       82 YSDHELRALDEA-QRLAKKKAD---------LYDEEQEQGITLAQDLEDMWEQAFLQ---------FRPGARETEADKKNDRTSLHRKLDRNLVL
rat          1
zebrafish    1                                        ←IVTAQDLEDVWEQNLKQ---------FQPAPRLQGDGE-TDMSSLERCLADSLVL
Drosophila  65 KSDHELKHEREI-VQAELIKQGKI-------QVDLEDAGSKQTAQDLKDAYVEELKK---------FQLGSRTTPDDQANRTTSTDRCLDDTLYL
trematode   40 LSAHELRHETET-KRIASALSKG---------VGKSAEESLITAREAEIMWELEAEQ---------YKPAERLTENDKSENLKSAWRVLDKPLYL
decoy       38 VVIPKPDPAVYA-FQEFKFNWQQ--------QFRRRYPDEFLDIAKNRAKGEYQMD----------YVPAPRITEADKNNDRKSLYRALDKKLYL
YmL30       74 VAEHKFLSLQKGPISKKNGIWFPRGIPDIKHGRERSTKQEVKLSDDSTVAFSNNQKEQSKDDVNRPVIPNDRITEADRSNDMKSLERQLSRTLYL
consensus   96 ....... .... ....... .          .......... ..............         .... ....... ........ .......

C15orf4    160 LVR-EKFGDQD---VWILPQAEW-QPGETLRGTAERTLATLSENNMEAKFLGNAPCGHYTFKFPQAMRTESNLGAKVFFFKALLLTGDFSQAGNK
mouse      158 LVR-EKLGDQD---VWMLPQVEW-QPGETLRGTAERILATLSENNMEAKFLGNAPCGHYKFKFPKAIQTESDLGVKVFFFKALLLTGDFVQAGKK
rat          1←LVR-EKLGDQD---LWMLPQVEW-QPGETLRGTAERILATLSENNMEAKFLGNAPCGHYKFKFPKAIRTESDLGVKVFFFKALLLTGDFVQTGKK
zebrafish   46 LVQ-KDVGSQK---IWLLPQIEW-QTGETLRQTAERALASLPGADLKATFLGNAPCGFYKYKYPKDVQKEGLVGAKVFFFKAVMSSQKHLPLEKN
Drosophila 143 LVQ-QKLGQQE---HLILPQGKR-EEGESMRQTAERVLRESCGQELQVLFYGNAPVGFHKYKYPRNQRTET-VGAKVFFYRASLRSGQVPENLTK
trematode  116 LVQ-SPN-VSS---GWNLPIAPI-SDGKNLRQVADSIATSLLPSRAKWCIFGNTP→
decoy      114 LIFGKPFGATSDKPVWPFPEKVY-DSEPTLRKCAESALKSVVGDLTHTYFVGNAPMAHMAIQPTEEMPDLP--SYKRFFFKCSVVAASKYDISTA
YmL30      169 LVK-DKS---G---TWKFPNFDLSDESKPLHVHAENELKLLSGDQIYTWSVSATPIGVLQDER--NRTAEFIVKSHILAGKFDLVAS---KNDAF
consensus  191 *.. .......   ....*. .. ....... .*....................*.................................... ...

C15orf4    250 GHHVWVTKDELGDYLK-PKYLAQVRRFVSDL------
mouse      248 SRHVWASKEELGDYLQ-PKYLAQVRRFLLDSDGLSCL
rat         91 GRHVWASKEELGDYLQ-PKYLAQVRRFLLDL------
zebrafish  136 -TFAWVKKDELQEFLK-PEYLKQVRRFIMTL------
Drosophila 232 --FEWLPKEALNEKLTNTAYAQSVKKFL---------
trematode
decoy      206 RILCG--------------------------------
YmL30      252 EDFAWLTKGEISEYVP-KDYFNKTEFLLADN------
consensus  286 .. ............  ..............
```

Figura 1 (Carim-Todd *et al.*, 2001)

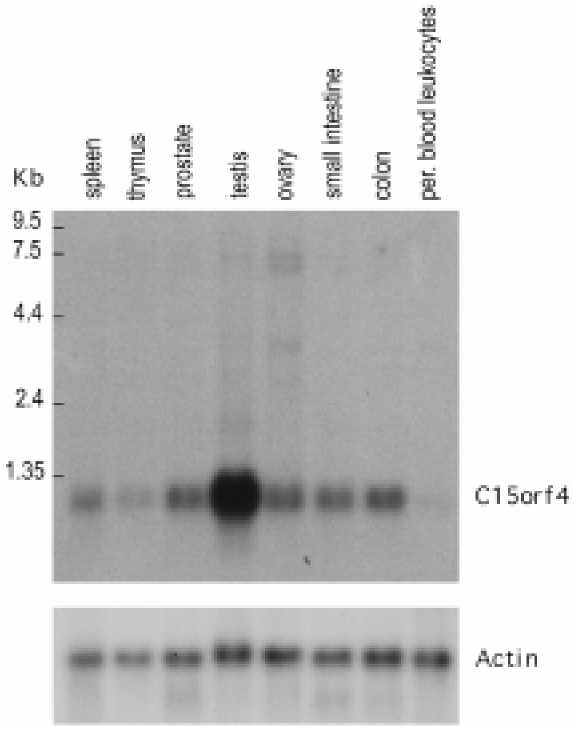


Figura 2 (Carim-Todd *et al.*, 2001)

II. Identificació i caracterització dels gens HMG20A i HMG20B

La publicació següent demostra l'existència d'una nova classe de proteïnes amb dominis HMG (high mobility group) amb una potencial funció en la regulació de la transcripció i conformació de la cromatina. L'aïllament del gen HMG20A mapat a la regió q24 del cromosoma 15 humà porta a l'identificació d'un gen paràleg, HMG20B, al cromosoma 19. Tots dos transcrits presenten una distribució generalitzada a nivell de teixit. La disponibilitat en aquell moment de seqüència genòmica provisional a les bases de dades públiques va permetre establir l'estructura exó-intró de tots dos gens i analitzar la seva conservació a aquest nivell.

Cytogenet Cell Genet 88:62–67 (2000)

Cytogenetics and
Cell Genetics

HMG20A and HMG20B map to human chromosomes 15q24 and 19p13.3 and constitute a distinct class of HMG-box genes with ubiquitous expression

L. Sumoy, L. Carim, M. Escarceller, M. Nadal, M. Gratacòs, M.A. Pujana, X. Estivill, and B. Peral

Medical and Molecular Genetics Center, Institut de Recerca Oncològica, Hospital Duran i Reynals, Barcelona (Spain)

Abstract. The HMG box encodes a conserved DNA binding domain found in many proteins and is involved in the regulation of transcription and chromatin conformation. We describe HMG20A and HMG20B, two novel human HMG box-containing genes, discovered within the EURO-IMAGE Consortium full-length cDNA sequencing initiative. The predicted proteins encoded by these two genes are 48.4% identical (73.9% within the HMG domain). The HMG domain of both HMG20 proteins is most similar to that of yeast NHP6A (38% to 42%). Outside of this domain, HMG20 proteins lack any significant homology to other known proteins. We determined the genomic structure and expression pattern of HMG20A and HMG20B. Both genes have several alternative transcripts, expressed almost ubiquitously. HMG20A maps to chromosome 15q24 (near D15S1227) and HMG20B to 19p13.3 (between D19S209 and D19S216). The HMG20 genes define a distinct class of mammalian HMG box genes.

Completion to full length of the sequences of unique cDNA clones represented in dbEST is a key step toward the characterization of all human genes (Auffray et al., 1995; Lennon et al., 1996; Collins et al., 1998). With this aim and working within the Euro-IMAGE Consortium, we have discovered a novel family of genes that have an HMG domain as their most outstanding feature.

The basic signature of the HMG domain, originally defined in **h**igh-**m**obility-**g**roup protein 1, HMG-1 (Jantzen et al., 1990), comprises approximately 70 amino acid residues forming three α-helices which mediate the interaction of HMG proteins with the minor groove of DNA. HMG domains recognize irregular DNA structures, such as non-B DNA, cruciforms, and cisplatin adducts, and are capable of bending DNA (Landsman and Bustin, 1993; Laudet et al., 1993; Baxevanis and Landsman, 1995).

There are two major types of HMG box genes based on amino acid sequence and DNA binding specificity. Members of the HMG1/2 class of proteins have low sequence target specificity, are expressed in many tissues, and usually have more than one HMG domain; they regulate nucleosome assembly. The TCF/SOX class includes sequence-specific DNA-binding proteins, with a single HMG domain, which regulate tissue-specific transcription (Grosschedl et al., 1994; Soullier et al., 1999).

We describe here the sequence, genomic structure, expression pattern, and chromosome location of two novel human HMG box genes, which we have named HMG20A and HMG20B. Based on sequence conservation criteria, HMG20 genes constitute a distinct class of mammalian HMG genes.

Materials and methods

cDNA isolation and sequencing

Expressed sequence tags (ESTs) from Unigene clusters Hs.69594 and Hs.32317 (http://www.NCBI.nlm.nih.gov/UniGene), corresponding to HMG20A and HMG20B respectively, were assembled using the EST CAP

Supported by EU Biomed Project BMH4-CT97-2284 to X.E. (which also funded L.C.) and by grants for research equipment and infrastructure (CICYT-IN95-0347 and CIRIT-SGR00444). M.E. and B.P. are funded by the Ministerio de Educación, Spain (MEC-CIDYT CT97-2284 and CT96-1364, respectively). L.S. is supported by CIRIT, Generalitat de Catalunya (CIRIT-RED Contract CT97-2284). B.P.'s present address is Neurology Unit, Fundación Jiménez Díaz, Avda. Reyes Católicos, 2, 28040 Madrid (Spain).

Received 22 June 1999; revision accepted 20 October 1999.

Request reprints from Dr. Lauro Sumoy, Medical and Molecular Genetics Center,
Institut de Recerca Oncològica, Hospital Duran i Reynals,
Autovia de Castelldefels Km 2,7, L'Hospitalet de Llobregat,
08907 Barcelona (Spain); telephone: 34-93-260-7775;
fax: 34-93-260-7776; e-mail: lsumoy@iro.es.

KARGER
Fax + 41 61 306 12 34
E-mail karger@karger.ch
www.karger.com

© 2000 S. Karger AG, Basel
0301–0171/00/0882–0062$17.50/0

Accessible online at:
www.karger.com/journals/ccg

assembly program (http://gcg.tigem.it/cgi-bin/uniestass.pl) and Sequencher (GeneCodes) sequence assembly software. Additional ESTs were found by BLAST searching dbEST. IMAGE cDNA clones chosen for sequencing were: 548013, 270903, 140081, and 147037 for human HMG20A; 532078, 587808, 308007, 267627, 179729, and 1695848 for human HMG20B; and 522081 and 834855 for mouse *Hmg20B.* None of the IMAGE cDNA clones corresponding to mouse *Hmg20A* was available for sequencing. Sequence was determined by primer walking using the Perkin-Elmer BigDye reagents on an ABI PRISM-377 fluorescent automated sequencer and custom-synthesized oligonucleotides (LifeTech).

Full-length cDNA sequence was obtained using the rapid amplification of cDNA ends (RACE) method on Marathon-Ready cDNA from adult human heart (Clontech). Primers were: G4 (5′-CAG TAG TGG CGT GGA TTG TTG GT-3′) and G5 (5′-GCC TCT GTT CAT TGC CTT CTG CT-3′) for 5′ HMG20A extension; and C2F (5′-ATG AAG TTA CAG GCT AGC AC-3′), G1 (5′-CAG TGA GGA GGC AGT AAA TGA AG-3′), and G2 (5′-AAG TTG TTG CCT ATT CAG TGT TAC-3′) for HMG20A 3′ extension. PCR-extended products were subcloned into the pGEM-T-easy vector (Promega) and sequenced as above. Nucleotide sequences for the cDNAs of HMG20A, HMG20B, and *Hmg20B* are available from GenBank under Accession Nos. AF146222, AF146223, and AF146224, respectively. Official gene symbols are ISGN approved.

Protein sequence analysis

Protein sequences were aligned using the PILEUP and GAP programs (GCG). Boxed multiple sequence alignments shown in the figures were obtained with the BOXSHADE program (http://www.isrec.isb-sib.ch:8080/software/BOX_form.html). To detect conserved protein domains, we used Pfam, PROSITESCAN, PRINTS, PRODOM, BLOCKS, PROFILESCAN, TOPITS, SMART, and PREDICTPROTEIN, available at http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-domain.html, http://coot.embl-heidelberg.de/SMART, and http://www.embl-heidelberg.de/predictprotein.

Genomic cloning of HMG20A

Genomic sequence of the HMG20A gene was obtained from clones generated in the process of establishing a physical map of the 15q24 chromosomal region. A high-density human bacterial artificial chromosome (BAC) library (CITB, Research Genetics) was screened by nucleic acid hybridization using STS WI-5695 (D15S739). Four positive BAC clones were obtained (80J7, 121A10, 194E21, and 204M2) that contained inserts of approximately 130 kb. After restriction analysis and Southern blotting, we confirmed that the four BAC clones were almost identical and contained the entire HMG20A gene. Exon/intron structure was determined by direct sequencing of BAC DNA or PCR products obtained using exonic primers (Big Dye, Perkin-Elmer/Applied Biosystems). Thermocycling parameters for BAC sequencing were 5 min at 95 °C, 35 cycles of 30 s each at 95 °C, and 4 min at 60 °C.

We determined the exon/intron structure of the HMG20B gene by direct comparison between its cDNA and genomic sequences using the GAP program (GCG).

Northern blot hybridization

Commercial human Multiple Tissue Northern blots (Clontech) were hybridized to random primed cDNA probes. Probes for HMG20A were inserts from IMAGE clones 140081 (*Eco*RI-*Hin*dIII) and 548013 (*Bam*HI-*Hin*dIII); for HMG20B the insert from clone 532078 (*Eco*RI-*Bam*HI); and for β-actin a 2-kb cDNA supplied commercially (Clontech). Blots were hybridized overnight at 65 °C in ExpressHyb hybridization solution (Clontech) and washed at 68 °C in 0.2 × SSC, 0.5 % SDS.

HMG20A radiation hybrid mapping

To precisely localize the HMG20A gene, we used the Stanford G3 and TNG whole genome panels (Stewart et al., 1997). Two-point linkage analysis was performed using the RHMAP-2.0 on the RH Server at the Stanford Human Genome Center (http://www-shgc.stanford.edu/RH/index.html). We used primers C2F (used also for RACE) and C2R (5′-AGT TCC AAA CAC ATG TAC AC-3′). The PCR conditions were 1 cycle at 94 °C for 2 min, 35 cycles of 30 s each at 94 °C, 58 °C for 30 s, and 74 °C for 30 s, and 1 cycle at 74 °C for 7 min.

Fluorescence in situ hybridization (FISH) analysis of HMG20A

BAC clone 80J7 containing the HMG20A gene was used as a probe for FISH on metaphase chromosome spreads as described (Nadal et al., 1997). The only modification was in the last three post-hybridization washes, which were in 2 × SSC at 42 °C instead of 0.1 × SSC at 60 °C.

Results

Isolation of the HMG20A gene

In the process of characterizing the q24 region in chromosome 15, we screened a human genomic BAC library using STS WI-5695. The end sequence of a 12-kb *Bam*HI single-copy subclone, contained in all four BACs obtained from the screen, was found to have sequence similarity to two dbEST entries (Accession Nos. AA113373 and AA919704, from human and mouse origin, respectively). Human EST AA113373 was found to belong to a Unigene cluster; cDNA clones corresponding to ESTs from the same cluster were sequenced within the EURO-IMAGE full-length cDNA sequencing project underway in our laboratory. The longest cDNA, IMAGE clone 548013, contained a complete open reading frame. The full-length cDNA sequence was obtained by 5′ and 3′ RACE extension. Assembly of the different clones gave as a result a total transcript length of 3,773 bp.

Analysis of the protein sequence encoded by this cDNA showed the presence of an HMG box, which was the only amino acid region of significant conservation. Because of this, we named this gene HMG20A, conforming with international human gene nomenclature rules (Shows et al., 1979). The human HMG20A gene encodes a predicted 347 amino acid protein (Fig. 1) with an expected molecular weight of 40.1 kDa. The amino acid sequence of the mouse *Hmg20A* gene was partially deduced from EST sequences (Accession Nos. AI574467, AI574369, AA823250, AI119468, AI119138, and AA144479 for the *N*-terminal portion; AA919704 for the internal *C*-terminal portion; and AA249948 for the *C*-terminal portion; Fig. 1) and found to be very well conserved (100% within the HMG domain).

The entire transcript for HMG20A spans 10 exons (Fig. 2). All the junctions between exons and introns are in accordance with the rule that introns begin with a GT dinucleotide and end with AG (Table 1). The first exon encodes a 5′ untranslated region (UTR), and the last one encodes a very long 3′ UTR with five non-canonical polyadenylation signals within 200 nucleotides upstream of the polyA tail in the longest RACE-extended cDNA clone.

Finding of a second HMG20 gene

A second mammalian gene very closely resembling HMG20A was found by EST sequence database searching. Human ESTs similar, but nonidentical, to HMG20A were found, identical to a partial cDNA (Accession No. AF072836, unpublished data). The longest clone, IMAGE clone 532078, contained an open reading frame encoding a 317 amino acid protein (Fig. 1), 35.8 kDa in predicted size and smaller than HMG20A. The sequence of four other clones was also determined and found to be mostly identical. Protein sequence comparison, with a 70.7% overall similarity and 48.4% overall

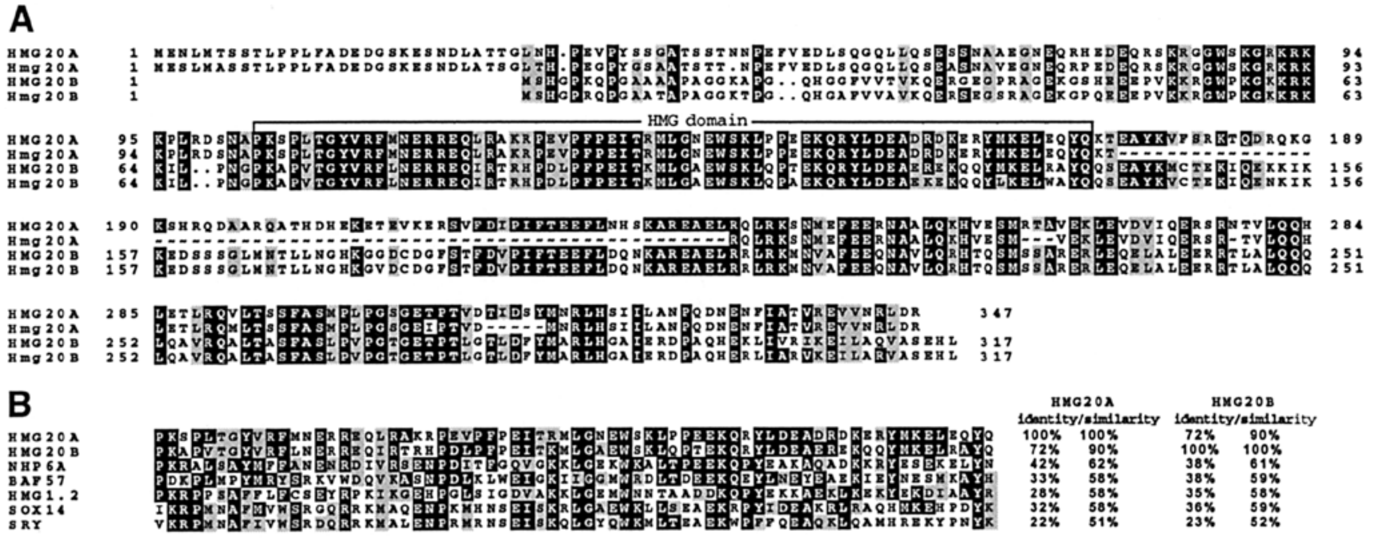

A
HMG20A 1 MENLMTSSTLPPLFADEDGSKESNDLATTGLNH.PEVPYSSGATSSTNNPEFVEDLSQGQLLQSESSNAAEGNEQRHEDEQRSKRGGWSKGRKRK 94
Hmg20A 1 MESLMASSTLPPLFADEDGSKESNDLATSGLTH.PEGPYGSAATSTT.NPEFVEDLSQGQLLQSEASNAVEGNEQRPEDEQRSKRGGWSKGRKRK 93
HMG20B 1 MSHGPKQPGAAAAPAGGKAPG..QHGGFVVTVKQERGEGPRAGEKGSHEEEPVKKRGWPKGKKRK 63
Hmg20B 1 MSHGPRQPGAATAPAGGKTPG..QHGAFVVAVKQERSEGSRAGEKGPQEEEPVKKRGWPKGKKRK 63
HMG domain
HMG20A 95 KPLRDSNAPKSPLTGYVRFMNERREQLRAKRPEVPFPEITRMLGNEWSKLPPEEKQRYLDEADRDKERYNKELEQYQKTEAYKVFSRKTQDRQKG 189
Hmg20A 94 KPLRDSNAPKSPLTGYVRFMNERREQLRAKRPEVPFPEITRMLGNEWSKLPPEEKQRYLDEADRDKERYNKELEQYQKT---------------
HMG20B 64 KIL..PNGPKAPVTGYVRFLNERREQIRTRHPDLPFPEITKMLGAEWSKLQPTEKQRYLDEAEREKQQYMKELRAYQQSEAYKMCTEKIQEKKIK 156
Hmg20B 64 KIL..PNGPKAPVTGYVRFLNERREQIRTRHPDLPFPEITKMLGAEWSKLQPAEKQRYLDEAEKEKQQYLKELWAYQQSEAYKVCTEKIQENKIK 156
HMG20A 190 KSHRQDAARQATHDHEKETEVKERSVFDIPIFTEEFLNHSKAREAELRQLRKSNMEFEERNAALQKHVESMRTAVEKLEVDVIQERSRNTVLQQH 284
Hmg20A --RQLRKSNMEFEERNAALQKHVESM---VEKLEVDVIQERSR-TVLQQH
HMG20B 157 KEDSSSGLMNTLLNGHKGGDCDGFSTFDVPIFTEEFLDQNKAREAELRRLRKMNVAFEEQNAVLQRHTQSMSSARERLEQELALEERRTLALQQQ 251
Hmg20B 157 KEDSSSGLMNTLLNGHKGVDCDGFSTFDVPIFTEEFLDQNKAREAELRRLRKMNVAFEEQNAVLQRHTQSMSSARERLEQELALEERRTLALQQQ 251
HMG20A 285 LETLRQVLTSSFASMPLPGSGETPTVDTIDSYMNRLHSIILANPQDNENFIATVREVVNRLDR 347
Hmg20A LETLRQMLTSSFASMPLPGSGEIPTVD-----MNRLHSIILANPQDNENFIATVREVVNRLDR
HMG20B 252 LQAVRQALTASFASLPVPGTGETPTLGTLDFYMARLHGAIERDPAQHEKLIVRIKEILAQVASEHL 317
Hmg20B 252 LQAVRQALTASFASLPVPGTGETPTLGTLDFYMARLHGAIERDPAQHERLIARVKEILARVASEHL 317
B
HMG20A identity/similarity
HMG20B identity/similarity
HMG20A PKSPLTGYVRFMNERREQLRAKRPEVPFPEITRMLGNEWSKLPPEEKQRYLDEADRDKERYNKELEQYQ 100% 100% 72% 90%
HMG20B PKAPVTGYVRFLNERREQIRTRHPDLPFPEITKMLGAEWSKLQPTEKQRYLDEAEREKQQYMKELRAYQ 72% 90% 100% 100%
NHP6A PKRALSAYMFFANENRDIVRSENPDITFGQVGKKLGEKWKALTPEEKQPYEAKAQADKKRYESEKELYN 42% 62% 38% 61%
BAF57 PDKPLMPYMRYSRKVWDQVKASNPDLKLWEIGKIIGGMWRDLTDEEKQEYLNEYEAEKIEYNESMKAYH 33% 58% 38% 59%
HMG1.2 PKRPPSAFFLFCSEYRPKIKGEHPGLSIGDVAKKLGEMWNNTAADDKQPYEKKAEKLKEKYEKDIAAYR 28% 58% 35% 58%
SOX14 IKRPMNAFMVWSRGQRRKMAQENPKMHNSEISKRLGAEWKLLSEAEKRPYIDEAKRLRAQHMKEHPDYK 32% 58% 36% 59%
SRY VKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYK 22% 51% 23% 52%

Fig. 1. (**A**) Amino acid sequence comparison between human HMG20A and HMG20B and murine *Hmg20A* and *Hmg20B.* Amino acids that are identical in all genes are boxed in black, while similar amino acids are boxed in gray. The HMG domain is bracketed. Dots indicate gaps introduced in the sequences for optimal alignment; dashes in the mouse *Hmg20A* putative sequence are unknown residues (due to lack of ESTs covering the region or to poor EST sequence). (**B**) Multiple sequence alignment of the HMG domains of HMG20 proteins compared to the prototype HMG classes SRY and HMG-1 and to the best matches obtained by BLAST comparison, yeast NHP6A, and mammalian BAF57 and SOX14. Black and gray boxes indicate identities and similarities, respectively, between at least three of the seven sequences. Numbers on the right indicate percent identity and similarity to HMG20A and HMG20B.

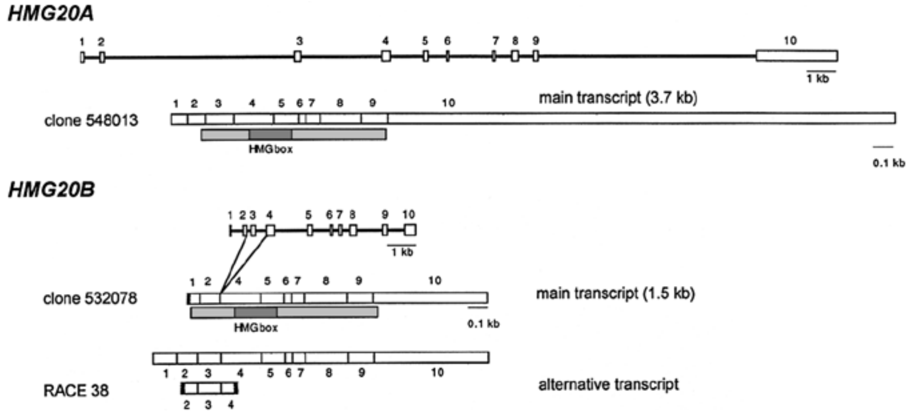


Fig. 2. Genomic structures of the HMG20A gene, based on the sequencing and Southern blot mapping data, and of HMG20B gene, based on cDNA-genomic sequence alignment comparison (see Table 1). The respective open reading frames are shown as boxes immediately beneath the cDNA sequences; the HMG domain is shaded. Exons are depicted as numbered boxes in the genomic DNA and cDNA. An alternative splicing isoform of HMG20B includes exon 3 (RACE38).

identity (with 91.3% similarity and 73.9% identity within the HMG domain) suggests that this second gene is a close homolog of HMG20A, and thus we have named it HMG20B (Fig. 1).

The 1,524-bp-long 532078 clone probably represents the predominant HMG20B gene product. A few rare EST sequences and a single PCR product obtained in one of the RACE experiments indicated that the HMG20B gene may have alternatively spliced transcripts expressed at very low levels or in very specific cell types (not shown).

Two partial mouse cDNA clones derived from ESTs matching human HMG20B were sequenced, which overlapped to construct a 1,632-bp transcript cDNA. The open reading frame is the same size and shares 93.7% identity with the human HMG20B protein (94.2% identity within the HMG domain).

The genomic sequences of the human HMG20B and mouse *Hmg20B* genes have been determined by others (GenBank Accession Nos. AC005786 and AF067430, respectively; unpublished data). The human HMG20B genomic structure was obtained directly by comparison between the respective cDNA and genomic sequences (Fig. 2) and has at least 9 exons. The exact number of exons could not be determined, since there appear to be minor splice variants.

HMG20A gene expression analysis

HMG20A Northern blot hybridization showed two major transcripts of approximately 4 and 9 kb and a barely detectable 1.5-kb mRNA (Fig. 3). The 3.7-kb full-length cDNA sequence may appear as 4 kb on Northern blots due to extensive polyade-

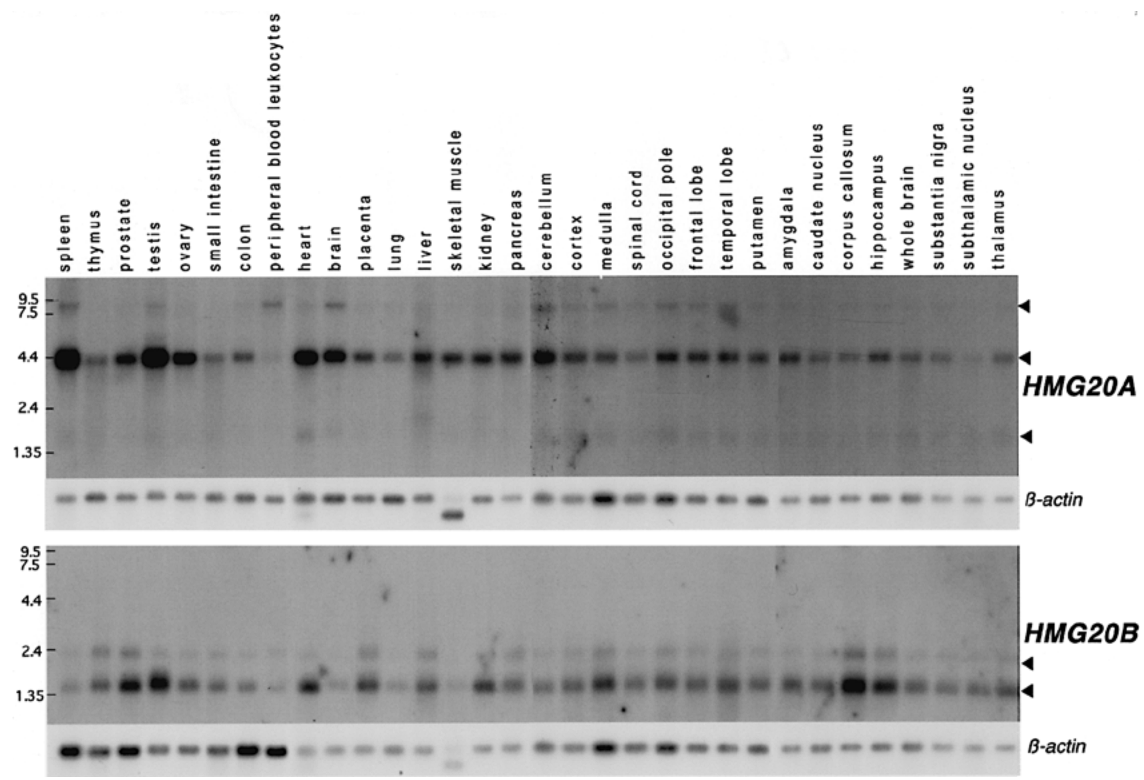


Fig. 3. Northern blot hybridization analysis of HMG20A and HMG20B. cDNA probes corresponding to clone 548013 (for HMG20A), 532078 (for HMG20B), and a commercially supplied clone for β-actin, used as a loading control, were hybridized to polyA RNA from multiple human tissues blotted onto a nylon membrane. Three size mRNA transcripts can be detected for HMG20A of approximately 9, 4, and 1.5 kb in size. Two bands are detected with the HMG20B probe at 2.6 and 1.5 kb. The positions of the detected bands are marked by the triangles on the right.

Table 1. Intron-exon boundary sequences of HMG20A and HMG20B

Gene	3′ Splice acceptor[a]	Exon	Size(bp)	5′ Splice donor[a]	Intron	Size (kb)
HMG20A	RACE +1 TGAGAGGGGCTG	1	87	TGTCGTCAGCAGgtcagctgcatg	1	> 0,600
	ttttcctttcagAGAGATGGAAAA	2	93	GGCTACCACTGGgtaagcagctgc	2	˜6,000
	cctattattcagGTTAAATCACCC	3	148	CATGAAGATGAGgtaagctgaagt	3	˜2,800
	tttgttttgtagCAACGAAGTAAA	4	213	GAGGAAAAACAGgtaattgttcct	4	> 1,300
	tttctttcctagCGCTACCTTGAT	5	133	CTCATAGGCAAGgtatcaaaacca	5	0,740
	ttatattwttagATGCAGCCCGGC	6	32	CATGATCATGAGgtaattagccat	6	> 1,600
	cttgattcacagAAAGAAACAGAG	7	76	ACCATAGCAAAGgtgattactgaa	7	0,663
	gtatgtcctcagCTCGGGAAGCAG	8	216	TGCCCTTGCCTGgtaagtcctccc	8	0,670
	atctcacttcagGAAGTGGAGAGA	9	143	CGTTAGGGAATGgtgagtgctcac	9	˜6,000
	ctgattctacagGTCTTAGAACTC	10	2,632	TTTTCTACGTCGaaaaaaaaaaaa	polyA	
HMG20B	cDNA end CGGAGCGGCCATGT	1	48	CGCGGCCGCCGCgtgagtgcactg	1	0,344
	ccttggctccagGCCGGCGGGCGG	2	109	CACGAGGAGGAGgtgagagtccct	2	0,102
	ctccgcccacagCCCATTTTCCTC	3	121	GCCCACTCTAAGgtcccgcccact	3	0,359
	ccggttctgcagCCGGTGAAGAAA	4	204	ACGGAAAAGCAGgtgggcggggcg	4	0,953
	gctgccccccagCGGTACCTGGAT	5	121	AGATCAAGAAAGgtgggaggggtc	5	0,600
	ctcccccgccagAAGACTCGAGCT	6	47	AATGGACACAAGgtaagcgacctt	6	0,245
	ccttcgtcttagGGTGGGGACTGC	7	73	ACCAAAACAAAGgtgagcggtaac	7	0,266
	cccccggcgcagCGCGTGAGGCGG	8	216	TGCCGGTGCCGGgtgcgggccacg	8	0,873
	cctgtcccccagGCACGGGCGAAA	9	133	CCAGGTCGCCAGgtgtgtgccggg	9	0,395
	ctctcgtttcagCGAGCACCTGTG	10	573	AATTTGTGTTTTaaaaaaaaaaaa	polyA	

[a] Intron sequence is shown in lowercase and exon sequence in uppercase; methionine and stop codons are underscored.

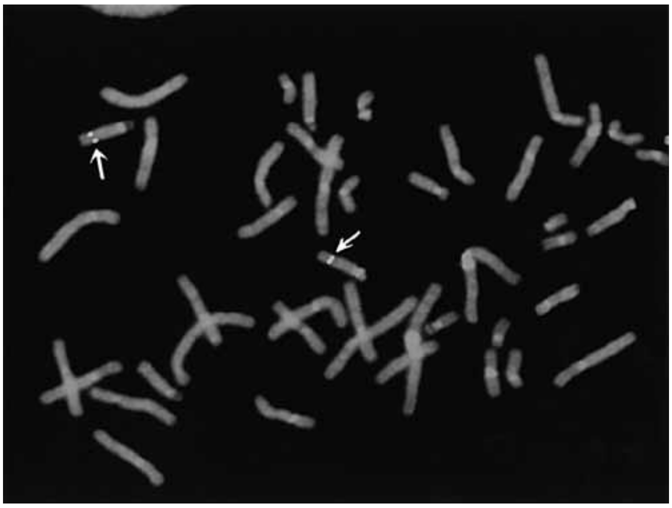


Fig. 4. Mapping of the human HMG20A gene by FISH. The 80J7 BAC clone containing the human HMG20A gene was labeled with biotin and hybridized to human metaphase chromosomes. Hybridization of 80J7 localizing HMG20A to chromosome region 15q24 is indicated by arrows.

nylation or a sizing imprecision. The 4-kb form is the most abundant HMG20A mRNA and is expressed almost ubiquitously–spleen, testis, and heart having the highest levels and peripheral blood leukocytes lower levels. Brain expression is mostly uniform, with higher levels being found in the cerebellum and lower levels in the spinal cord and the subthalamic nucleus. The 9-kb HMG20A transcript is differentially expressed in spleen, testis, leukocytes, and brain and is barely detectable in the remaining tissues.

Expression of HMG20B

HMG20B is also expressed among a wide variety of tissues, with Northern blot hybridization analysis showing two major mRNA forms of approximately 2.6 and 1.5 kb in size. The 2.6-kb form is expressed at lower levels and can be detected in thymus, prostate, placenta, liver, kidney, and pancreas. The 1.5-kb mRNA corresponds with the size of the full-length HMG20B cDNA clones we have sequenced. The highest expression levels of the 1.5-kb form are seen in the prostate, testis, heart, and kidney, while brain, spleen, lung, skeletal muscle, and leukocytes show lower levels. Within the brain, the 1.5-kb form of HMG20B is expressed almost uniformly, with increased levels in the corpus callosum and hippocampus.

Chromosome location of HMG20A and HMG20B

FISH analysis localized human HMG20A to 15q24 (Fig. 4). The Unigene EST cluster had been mapped between D15S114 and D15S989 (Deloukas et al., 1998). Our own radiation hybrid mapping, using the G3 panel, determined that HMG20A was linked to STS SHGC-15284 (close to D15S1227), with a lod score of 9.46. Using the TNG panel, HMG20A was found at an estimated distance of 180 kb from SHGC-20921 (D15S984) with a lod score of 5.79.

The chromosome map position of HMG20B was inferred from genomic clone sequence annotation to be in 19p13.3, between D19S209 and D19S216 (Deloukas et al., 1998).

Discussion

We have found two new genes, HMG20A and HMG20B, which add to a list of over 250 different genes coding for proteins with an HMG domain, a protein motif capable of binding to DNA (Baxevanis and Landsman, 1995; Soullier et al., 1999). Protein sequence comparison indicates that the HMG20A and HMG20B proteins have a much higher resemblance to each other than to any other HMG-box gene. The conservation in HMG20 gene structure in exons with the strongest amino acid similarity is consistent with the two genes deriving from a common ancestor.

HMG20 protein sequence comparisons show conservation with other known proteins only within the HMG domain. The closest match to both HMG20 proteins is the yeast NHP6A nonhistone chromatin binding protein (Kolodrubetz and Burgum, 1990) (Fig. 1B). NHP6A has a single HMG domain with non–sequence-specific DNA binding properties and is involved in potentiating the transcriptional activation (Paull et al., 1993). Among mammalian HMG-box genes most similar to HMG20 (Fig. 1B) is BAF57, encoding a subunit of the SWI/SNF complex (Wang et al., 1998). However, the level of similarity is not high enough to establish homology to HMG20A and HMG20B.

Parsimony, maximum likelihood, and protein distance phylogenetic comparisons excluded the HMG20 genes from the TCF/SOX subfamily (Laudet et al., 1993) and placed them along with the remaining subgroups (including NHP6 and BAF57) (Baxevanis and Landsman, 1995). This information would suggest that HMG20 genes belong to the HMG1/2 type of non–sequence-specific HMG proteins. However, the lack of conservation outside of the HMG domain suggests that the HMG20 genes constitute a distinct class of HMG genes.

Both HMG20A and HMG20B are transcribed with a wide tissue distribution (Fig. 3). At first, the wide expression pattern of HMG20A and HMG20B suggests that they could be performing a housekeeping role as nonhistone components of chromatin, like HMG-1. The other possibility is that, although they have a generalized pattern of expression, they could act locally through interaction with tissue-specific transcription factors.

The HMG20A gene was confirmed to map to 15q24 by FISH and radiation hybrid mapping techniques. Hereditary conditions known to map in the 15q24 region include nocturnal frontal lobe epilepsy (Phillips et al., 1998) and severe mental retardation with spasticity and pigmentary tapetoretinal degeneration (Mitchell et al., 1998). The corresponding syntenic region in mouse chromosome 9 lacks mutations for which the gene remains unknown (Blake et al., 1999).

HMG20B is located in chromosome 19p13.3, between the markers D19S209 and D19S216. No known human diseases have been mapped to this region (McKusick, 1998). In mice, *Hmg20B* appears to map to the 43-cM region of mouse chromosome 10, which is syntenic to human 19p13.3 (by neighbor gene reference). Mutations in this region include *jittery (jt),* a recessive sublethal mutation affecting neuromotor coordination and male fertility (Kapfhamer et al., 1996), and *grizzled (gr),* a recessive mutation causing hair pigmentation and tail defects with reduced viability (Bloom and Falconer, 1966). It will be very interesting to determine whether *Hmg20B* is associated with either of these two mutations.

Acknowledgements

We wish to thank A. Puig and D. Otero for technical support, M. Riutort for advice on phylogenetic analysis, and S. de la Luna for helpful criticisms to this manuscript. The HGMP Resource Center in Hinxton (UK) and the RZPD in Berlin (Germany) supplied us with IMAGE cDNA clones.

References

Auffray C, Behar G, Bois F, Bouchier C, Da Silva C, Devignes MD, Duprat S, Houlgatte R, Jumeau MN, Lamy B, et al.: IMAGE: molecular integration of the analysis of the human genome and its expression. CR Acad Sci III 318:263–272 (1995).

Baxevanis AD, Landsman D: The HMG-1 box protein family: classification and functional relationships. Nucl Acids Res 23:1604–1613 (1995).

Blake JA, Richardson JE, Davisson MT, Eppig JT: The Mouse Genome Database (MGD): genetic and genomic information about the laboratory mouse. The Mouse Genome Database Group. Nucl Acids Res 27:95–98 (1999).

Bloom JL, Falconer DS: *"Grizzled",* a mutant in linkage group X of the mouse. Genet Res 7:159–167 (1966).

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

Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L, Matise TC, McKusick KB, Beckmann JS, Bentolila S, Bihoreau M, Birren BB, Browne J, Butler A, Castle AB, Chiannilkulchai N, Clee C, Day PJ, Dehejia A, Dibling T, Drouot N, Duprat S, Fizames C, Bentley DR, et al: A physical map of 30,000 human genes. Science 282:744–746 (1998).

Grosschedl R, Giese K, Pagel J: HMG domain proteins: architectural elements in the assembly of nucleoprotein structures. Trends Genet 10:94–100 (1994).

Jantzen HM Admon A, Bell SP, Tjian R: Nucleolar transcription factor hUBF contains a DNA-binding motif with homology to HMG proteins. Nature 344:830–836 (1990).

Kapfhamer D, Sweet HO, Sufalko D, Warren S, Johnson KR, Burmeister M: The neurological mouse mutations *jittery* and *hesitant* are allelic and map to the region of mouse chromosome 10 homologous to 19p13.3. Genomics 35:533–538 (1996).

Kolodrubetz D, Burgum A: Duplicated NHP6 genes of *Saccharomyces cerevisiae* encode proteins homologous to bovine high mobility group protein 1. J Biol Chem 265:3234–3239 (1990).

Landsman D, Bustin M: A signature for the HMG-1 box DNA-binding proteins. Bioessays 15:539–546 (1993).

Laudet V, Stehelin D, Clevers H: Ancestry and diversity of the HMG box superfamily. Nucl Acids Res 21:2493–2501 (1993).

Lennon G, Auffray C, Polymeropoulos M, Soares MB: The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. Genomics 33:151–152 (1996).

McKusick VA: Mendelian Inheritance in Man: Catalogs of Human Genes and Genetic Disorders. (Johns Hopkins University Press, Baltimore 1998).

Mitchell SJ, McHale DP, Campbell DA, Lench NJ, Mueller RF, Bundey SE, Markham AF: A syndrome of severe mental retardation, spasticity, and tapetoretinal degeneration linked to chromosome 15q24. Am J hum Genet 62:1070–1076 (1998).

Nadal M, Moreno S, Pritchard M, Preciado MA, Estivill X, Ramos-Arroyo MA: Down syndrome: characterisation of a case with partial trisomy of chromosome 21 owing to a paternal balanced translocation (15;21)(q26;q22.1) by FISH. J med Genet 34:50–54 (1997).

Paull TT, Haykinson MJ, Johnson RC: Yeast HMG proteins NHP6A/B potentiate promoter-specific transcriptional activation in vivo and assembly of preinitiation complexes in vitro. Genes Dev 7:1521–1534 (1993).

Phillips HA, Scheffer IE, Crossland KM, Bhatia KP, Fish DR, Marsden CD, Howell SJ, Stephenson JB, Tolmie J, Plazzi G, Eeg-Olofsson O, Singh R, Lopes-Cendes I, Andermann E, Andermann F, Berkovic SF, Mulley JC: Autosomal dominant nocturnal frontal-lobe epilepsy: genetic heterogeneity and evidence for a second locus at 15q24. Am J hum Genet 63:1108–1116 (1998).

Shows TB, Alper CA, Bootsma D, Dorf M, Douglas T, Huisman T, Kit S, Klinger HP, Kozak C, Lalley PA, Lindsley D, McAlpine PJ, McDougall JK, Meera Khan P, Meisler M, Morton NE, Opitz JM, Partridge CW, Payne R, Roderick TH, Rubinstein P, Ruddle FH, Shaw M, Spranger JW, Weiss K: International system for human gene nomenclature (1979): ISGN (1979). Cytogenet Cell Genet 25:96–116 (1979).

Soullier S, Jay P, Poulat F, Vanacker JM, Berta P, Laudet V: Diversification pattern of the HMG and SOX family members during evolution. J molec Evol 48:517–527 (1999).

Stewart EA, McKusick KB, Aggarwal A, Bajorek E, Brady S, Chu A, Fang N, Hadley D, Harris M, Hussain S, Lee R, Maratukulam A, O'Connor K, Perkins S, Piercy M, Qin F, Reif T, Sanders C, She X, Sun WL, Tabar P, Voyticky S, Cowles S, Fan JB, Cox DR, et al: An STS-based radiation hybrid map of the human genome. Genome Res 7:422–33 (1997).

Wang W, Chi T, Xue Y, Zhou S, Kuo A, Crabtree GR: Architectural DNA binding by a high-mobility-group/kinesin-like subunit in mammalian SWI/SNF-related complexes. Proc natl Acad Sci, USA 95:492–498 (1998).

III. Identificació i caracterització de TM6SF1 i paralogia entre cromosoma 15 i 19

En aquest cas l'identificació i caracterització d'un nou gen a la regió 15q24-q26, TM6SF1 (transmembrane 6 superfamily 1), permet descobrir l'existència de paralogia entre 11 gens del cromosoma 15 i gens localitzats a 19p13.3-p12. TM6SF1 codifica una proteïna de membrana d'expressió significativa a melsa, testicle i leucòcits. La similaritat de seqüència entre TM6SF1 i el seu paràleg, TM6SF2, és del 68%. L'anàlisi d'aquestes dues regions cromosòmiques i anàlisis comparatives amb el genoma de ratolí revelen l'existència de regions de sintènia conservada que donen suport als estudis que proposen una història evolutiva comú per aquests dos cromosomes.

Cytogenet Cell Genet 90:255–260 (2000)

Cytogenetics and Cell Genetics

Cloning of the novel gene TM6SF1 reveals conservation of clusters of paralogous genes between human chromosomes 15q24→q26 and 19p13.3→p12

L. Carim-Todd, M. Escarceller, X. Estivill and L. Sumoy

Medical and Molecular Genetics Center, Institut de Recerca Oncològica, Hospital Duran i Reynals, L'Hospitalet de Llobregat, Barcelona (Spain)

Abstract. As the result of the EUROIMAGE Consortium sequencing project, we have isolated and characterized a novel gene on chromosome 15, TM6SF1. It encodes a 370 amino acid product with enhanced expression in spleen, testis and peripheral blood leukocytes. We have identified another gene, paralogous to TM6SF1 on chromosome 19p12, TM6SF2, with an overall similarity of 68% and 52% identity at the protein level. This conservation has led us to uncover a series of eleven genes in 19p13.3→p12 with close homology to genes in 15q24→q26. The percentage of sequence similarity between each paralogous pair of genes at the protein level ranges between 43 and 89%. A partial conservation of synteny with mouse chromosomes 7, 8 and 9 is also observed. The corresponding orthologous genes in mouse of human TM6SF1 and TM6SF2 show a high degree of amino acid sequence conservation.

The EUROIMAGE Consortium was established with the aim of completing to full-length sequences of unique cDNA clones represented in dbEST as a key step towards the characterization of all human genes (Adams et al., 1991; Lennon et al., 1996; Schuler et al., 1996; Schuler, 1997; Deloukas et al., 1998). Working within this Consortium, we have undertaken the characterization of transcripts mapping in the 15q24→q26 region and have come across the TM6SF1 (transmembrane 6 superfamily 1) gene, a novel gene lacking homology to any other known gene. A homologous putative gene, TM6SF2, was found on 19p12 by database searching. The conservation between chromosome 15q24→q26 and chromosome 19 can be extended to ten more genes with matching sequences on chromosome 19p13.3→p12. While some of these genes have already been identified and characterized by others, a few have been uncovered by our cDNA sequencing effort and some still remain as incomplete cDNA sequences in the databases. This observation of sequence conservation between human chromosomes 15q24→q26 and 19p13.3→p12 supports the hypothesis of a common origin for human chromosomes 11, 15 and 19 (Lundin, 1993).

It has been proposed that the vertebrate genome has evolved through a series of large regional duplications. The remnants of these events are still apparent in the form of clusters of sequence conservation between different chromosomes that could be derived from a common ancestral chromosome. Genes in such clusters are therefore paralogous and can have similar functions, redundant or complementary in nature depending on the repertoire of functional targets and tissues in which they are expressed. Numerous examples of linked or syntenic genes with shared homology have been described

Supported by EU Biomed Project No. BMH4-CT97-2284 to X.E and by CICYT-IN95-0347. M.E. is funded by the Spanish Ministry of Education (CIDYT contract FPI-070-97) and L.S. by the Catalan autonomous government (CIRIT-RED contract 1998-64).

Received 19 April 2000; revision accepted 28 July 2000.

Request reprints from Lauro Sumoy, Medical and Molecular Genetics Center,
Institut de Recerca Oncològica, Hospital Duran i Reynals,
Autovia de Castelldefels Km 2.7, L'Hospitalet de Llobregat,
E–08907 Barcelona (Spain); telephone: 34-93-260-7775;
fax: 34-93-260-7776; e-mail: lsumoy@iro.es

KARGER

Fax + 41 61 306 12 34
E-mail karger@karger.ch
www.karger.com

© 2000 S. Karger AG, Basel
0301–0171/00/0904–0255$17.50/0

Accessible online at:
www.karger.com/journals/ccg

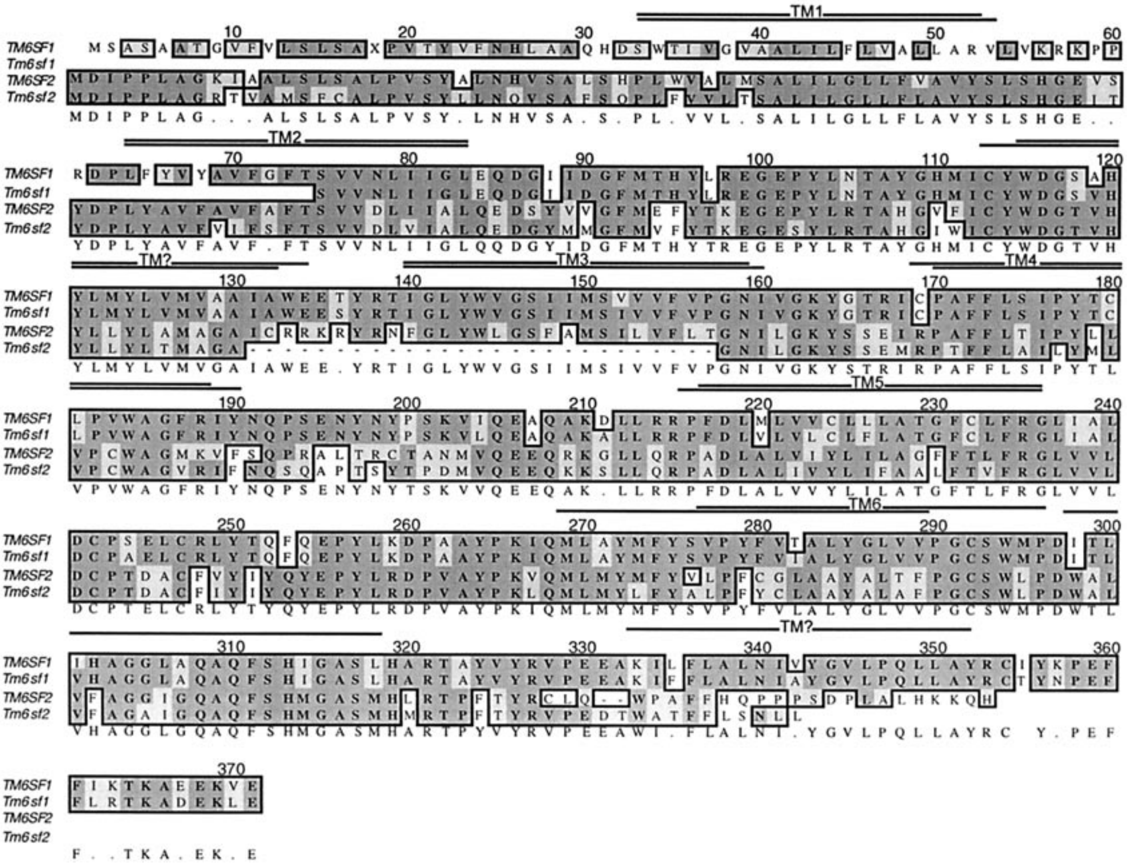


Fig. 1. Multiple sequence alignment including human TM6SF1 and TM6SF2 along with their corresponding mouse orthologs, Tm6sf1 and Tm6sf2 respectively. Residues conserved in the four proteins are highlighted using dark boxes, similar residues are boxed using lighter background. Residue X in TM6SF1 amino acid sequence can be I or T. The position of the predicted transmembrane domains (TM) is represented. TM? Indicates a possible transmembrane domain.

forming part of clusters of variable size on mouse and human chromosomes (Lundin, 1993). Examples of this phenomenon are human chromosomes 4 and 5, which have been found to contain 13 groups of paralogous genes. Human chromosomes 2, 7, 12, 14 and 17 constitute another example, sharing the presence of extensive clusters that include the Hox and collagen genes along with other genes (Hart et al., 1987). Finally, extensive orthologies have been found between mouse chromosomes 7 and 9, and human chromosomes 11, 15 and 19. It has been proposed that they all derive from a single ancestral chromosome and have appeared after a series of regional duplications that occurred during vertebrate evolution (Lalley et al., 1991; Lyon and Kirby, 1992).

We report here the identification and characterization of TM6SF1, the sequence of which further reveals the existence of a paralogous gene on chromosome 19p12 along with another ten pairs of paralogous genes on chromosomes 15q24→q26 and 19p13.3→p12. A partial conservation of synteny with mouse chromosomes 7, 8 and 9 is also observed for some of the genes reported here.

Materials and methods

cDNA isolation and sequencing

ESTs from Unigene cluster Hs.226144 (http://www.NCBI.nlm.nih.gov/UniGene) were assembled using the EST CAP assembly program (http://gcg.tigem.it/cgi-bin/uniestass.pl) and Sequencher (GeneCodes) sequence assembly software. Additional ESTs were found by searching the dbEST database. We chose IMAGE cDNA clone 22284 for sequencing. Sequence was determined by primer walking using the PerkinElmer BigDye reagents on an ABI PRISM-377 fluorescent automated sequencer and custom synthesized sequencing primers (LifeTech).

Full-length cDNA sequence was obtained using the rapid amplification of cDNA ends (RACE) method on fetal brain SMART cDNA (Clontech). Primers were: G1A (5′ TGAAACCAGCCCAGACAGGAAGAC 3′), G2A (5′ GAGCAGAGCCATCCCAGTAGCAGA 3′) and G3A (5′ GCTGCCCTCATCCTGTTCCTGGTA 3′) for 5′ extension. PCR extended products were subcloned into the pGEM-T easy vector (Promega) and sequenced as above. Nucleotide sequences for the TM6SF1 and TM6SF2 cDNAs are available from GenBank under Accession Numbers AF255922 and AF255923 respectively. The gene symbols are ISGN approved.

Protein sequence analysis

Protein sequences were aligned using ClustalW and boxed multiple sequence alignments shown in the figures were obtained with MacVector 6.5.3. To detect conserved protein domains we used Pfam, PROSITESCAN, PRINTS, PRODOM, BLOCKS, PROFILESCAN, TOPITS, SMART and PREDICTPROTEIN available at http://www.hgmp.mrc.ac.uk/Genome-

Table 1. The eleven pairs of paralogous genes studied: Unigene clusters, IMAGE clones and references are noted, the number of residues in the ORF, the score and percentage of similarity and the number of amino acids aligned are also shown

	Hs.[a] gene		Hs. Unigene	IMAGE clones		Hs. gene chr.15/chr.19 alignment		ORF length (aa)		Sequence aligned (aa)	References
	chr. 15	chr. 19		chr. 15	chr. 19	e-value	% similarity	chr. 15	chr. 19		
1	TM6SF1	TM6SF2	Hs.226144 Hs.175441	22284 40627	1520283	6e-98	68	369	351	319	Present article
2	EXLD1	EXLD2p	Hs.29283	156915		3e-94	66	360	438	320	Sumoy et al., in preparation
3	CSPG1	CSPG3	Hs.2159	1113913		2e-91	43	2316	1321	1053	Doege et al., 1991 Prange et al., 1998
4	MEF2A	MEF2B	Hs.182280			8e-44	62	507	365	210	Yu et al., 1992
5	EEF2-like 15 partial	EEF2	Hs.19348	415504 203133 295310		5e-65	43	764	858	597	Rapp et al., 1989 Present article
6	BTBD1	BTBD1-19 partial	Hs.21332	549016 28577 2325042		e-143	89	482	297	297	Carim et al., submitted
7	IRO403-15 partial	IRO403-19p partial	Hs.24835	172219 38269 938081 726301		e-173	69	637	560	573	Carim et al., in preparation
8	HMG20A	HMG20B	Hs.69594	548013	532078	5e-80	71	347	317	281	Sumoy et al., 2000
9	SH3GL3	SH3GL1	Hs.80315	278699		e-139	78	347	368	368	Giachino et al., 1997
10	SIN3A	SIN3B	Hs.172444 Hs.22583	1626863 626487 950137		0	67	1272	1130	1158	Sumoy et al., in preparation Halleck et al., 1995
11	DRIL2	DRIL1	Hs.10431			e-100	52	560	593	598	Kortschak et al., 1998

[a] Hs: *Homo sapiens*

Web/prot-domain.html, http://coot.embl-heidelberg.de/SMART and http://www.embl-heidelberg.de/predictprotein. Exons were predicted on genomic sequence by Fex, Genemark and Genscan in the NIX program, available from the HGMP Resource Centre at http://www.hgmp.mrc.ac.uk.

Northern blot analysis

For detection of TM6SF1, a human multiple-tissue Northern blot (MTN II blot, Clontech) was hybridized with a 1.4-kb *Hin*dIII-*Not*I probe belonging to IMAGE human cDNA clone 22284 and a 2-kb β-actin probe supplied commercially (Clontech) was used as control for quantification. Probes were labeled using a random primer DNA labeling kit (Amersham Pharmacia). The blot was hybridized overnight at 65 °C in ExpressHyb hybridization solution (Clontech) and washed at 68 °C in 0.2 × SSC, 0.5 % SDS.

Results and discussion

Isolation and characterization of TM6SF1 and TM6SF2

A comprehensive effort to identify all transcripts mapping to 15q24→q26 in silico by using existing EST cluster information available from the Unigene database has resulted in the identification of over 100 new putative genes in our laboratory (Carim et al., 1999; Sumoy et al., 2000; Carim et al., 2000a, 2000b; Auffray et al., unpublished). Full length cDNA sequencing within the EUROIMAGE Consortium allowed us to determine the sequence of a maximal length clone representative of Unigene cluster Hs.226144 (IMAGE clone 22284 EST GenBank Acc. No. T87220). The final assembly of the different clones corresponding to this cluster resulted in a 1.4-kb mRNA, which contained an ORF spanning 1.1-kb of the transcript. The encoded protein was 370 amino acids long with a calculated mass of 41.6 kDa and an estimated pI of 7.55 (Fig. 1). The 5′ untranslated region (UTR) contained an in-frame stop codon at nt position 31. The gene was designated TM6SF1 following the Human Gene Nomenclature Committee instructions (http://www.gene.ucl.ac.uk/nomenclature/).

Database searches with BLAST programs using TM6SF1 sequence gave a significant hit with finished genomic sequence from chromosome 19 (Acc. No. AC003967). Database searches to identify ESTs corresponding to TM6SF1′s counterpart on chromosome 19 identified IMAGE clone 1520283 (EST GenBank Acc. No. AA907902). We sequenced this clone and detected a partial ORF of 350 residues. Using exon prediction programs on the available genomic sequence we identified a 5′ methionine that gave as a result an ORF of 351 amino acids with 68% similarity and 52% identity to TM6SF1. Named TM6SF2, this gene encodes a 39.5-kDa protein with an estimated pI of 8.29. We also identified ESTs that would correspond to the mouse orthologs of TM6SF1 and of TM6SF2, indicating that these genes appear to be intact and transcribed in mice (Fig. 1). The analysis of the amino acid sequence of the human genes with protein pattern and domain prediction software revealed the presence of six significative transmembrane domains plus two other identified with lower significance in both TM6SF1 and TM6SF2 protein products (Fig. 1). Otherwise no other protein features were detected which could help to elucidate the function of this pair of genes.

We used the available human genomic sequence in the public databases to establish the exon-intron structure of both genes as shown in Table 2. We detected remarkable conservation in the length and sequence of the main coding exons (2, 3, 4, 5, 6, 7, 8 and 9) and significant divergence in the size of the

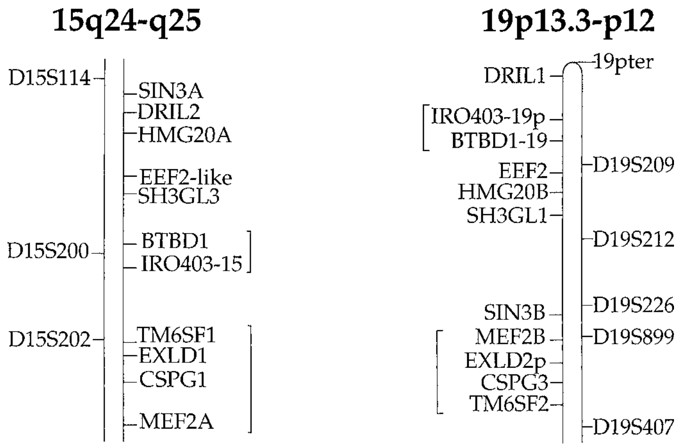


Fig. 2. Schematic illustrating the relative positioning of the eleven pairs of paralogous genes mentioned in the text on human chromosomes 15 and 19. Genes have been positioned using public sequence contig and gene mapping information from radiation hybrid mapping panels G3, GB4 and TNG4 (Schuler et al., 1996; http://www.ncbi.nlm.nih.gov/genemap99).

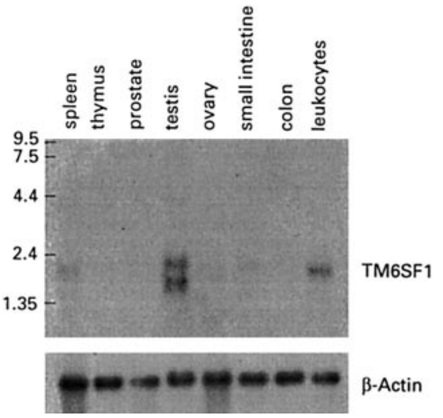


Fig. 3. Northern blot analysis of TM6SF1 mRNA expression in adult human tissues. Both TM6SF1 and β-actin as quantification control are indicated.

corresponding introns, a finding that is in accordance with selective pressure acting strongly on coding sequence. This exon-intron structure conservation results in conservation of the distribution of the transmembrane domains between the coding exons.

Expression analysis of TM6SF1 was carried out by hybridization with a specific probe using a human tissue Northern blot (MTN II Human blot, Clontech) (see Methods). In adult tissues, TM6SF1 expression was detected in spleen, testis and peripheral blood leukocytes, while it appeared absent from the rest of tissues analyzed. The size of the majority mRNA form is in accordance with clone 22284 and corresponds to a 1.4-kb transcript (Fig. 3).

Paralogous gene clusters on chromosomes 15q24→q26 and 19p13.3→p12

After the identification of TM6SF1's paralogous gene on chromosome 19p12 and having reviewed previous literature describing paralogy between these two chromosomes, we began the search for other examples of paralogy using already described genes and partial and full-length transcripts mapping to 15q24→q26 isolated in our lab within the EUROIMAGE cDNA sequencing project. By database searches we have identified three other genes which mapped contiguously to TM6SF2 on 19p12, within BAC sequence contig NT_000936, which also shared homology to counterparts on chromosome 15q24→q26 (Fig. 2 and Table 1, examples 1 to 4). The close relative positions of the genes on chromosome 19 are not conserved between their chromosome 15 counterparts. Contrary to the well studied example of the Hox clusters which have been kept together because this provides an evolutionary advantage, it is possible that, although the genes described here initially traveled together, there was no selective pressure to keep them close to one another.

At the protein level the percentage of similarity between each of the three paralogous pairs of genes is found to range between 43 and 68% (Table 1). These include EXLD1 and EXLD2p (extracellular link domain containing protein) (Sumoy et al., unpublished data), two new hyaluronic acid binding protein genes with homology to cartilage matrix link protein CRTL1 (Osborne-Lawrence et al, 1990); aggrecan (CSPG1, Acc. No. AAA62824) and neurocan (CSPG3, Acc. No. AAC80576), two large aggregating proteoglycan core proteins with hyaluronic acid binding activity; and MEF2A (Acc. No. NP–005578) and MEF2B (Acc. No. NP_005910) genes, which encode myocyte enhancer factors that are responsible for tissue-specific transcription in skeletal and cardiac muscle (Yu et al., 1992).

When comparing the genomic structure of genes EXLD1/EXLD2p, MEF2A/MEF2B and CSPG1/CSPG3 we have found a variable degree of conservation in their exon-intron organization. Conservation between paralogous genes is especially significant along the coding exons, as would be expected from the action of selective pressure on functional genes. Domain composition relative to the exon-intron structure is also conserved within paralog pairs (data not shown). We have found great variation in the size and sequence of introns.

Partial conservation of synteny with mouse chromosomes 7 and 8 can be detected in the case of CSPG1/CSPG3 and MEF2A/MEF2B paralog genes. The mouse homologue genes for CSPG1 and MEF2A, which in humans are on 15q24→q26, map to mouse chromosome 7. On the other hand, the rodent homologues of CSPG3 and MEF2B, mapped in humans to 19p12, have been localized to mouse chromosome 8. Paralogous gene pairs such as IGF1R/INSR (insulin growth factor 1 receptor, AAB59399, and insulin receptor, NP_000199) and MANA/MANB (cytoplasmic alpha-mannosidase A, NP_006706, and lysosomal alpha-mannosidase B,

Table 2. Exon-intron structure of TM6SF1 and TM6SF2 genes

TM6SF1

3'Splice acceptor	Exon	Size (bp)	5' Splice donor	Intron	Size (bp)
5'UTR GGGGGGGGCGGG	1	146	GGCCCAGCATGA**gt**----------	1	-
tggtcgttgc**ag**TTCCTGGACTAT	2	104	CACTGTTCTATG**gt**acgtctccac	2	-
tctttgctct**ag**TGTATGCAGTTT	3	98	TACTTGAGAGAG**gt**atgggatcac	3	3635
tgctttcttt**ag**GGTGAACCGTAT	4	104	CATAGCATGGGA**gt**aagtcagttc	4	2239
ttgttgttac**ag**GGAAACTTATAG	5	83	GAAACATTGTAG**gt**aagaaacttt	5	753
tcttttaaat**ag**GGAAGTATGGAA	6	122	TACCCCTCAAAG**gt**gattttatta	6	-
actctatttt**ag**GTTATTCAAGAA	7	105	TTCAGAGGTTTG**gt**aagcataaca	7	1978
ttttgtcctt**ag**ATTGCTTTGGAT	8	93	CCTAAAATTCAG**gt**caagtagtta	8	510
tgttcaatac**ag**ATGCTGGCATAT	9	120	GGTCTGGCTCAG**gt**actaagaata	9	9006
taaatgcaac**ag**GCTCAGTTTTCT	10	442	ATCATCCATCTCaaaaaaaaaaaa		

Detail of the exon-intron structure of TM6SF1; the entire transcript consists of 10 exons. All exon-intron boundaries are in accordance with the rule that introns begin with dinucleotide GT and end in AG (in bold). The length of introns 1, 2 and 6 is not shown because the corresponding genomic DNA is in the public databases as a provisional contig of unordered sequence fragments (Acc. No. AC069400, AC018910, AC069400).

TM6SF2

3'Splice acceptor	Exon	Size (bp)	5' Splice donor	Intron	Size (bp)
5'UTR TGAGGGAAACTG	1	337	GGCGCTCTCGCA**gt**gcgtgcggag	1	1995
ccctctccac**ag**CCCCCTGTGGGT	2	104	CACTCTATGCTG**gt**gagtcagtgg	2	579
cactcactgc**ag**TCTTCGCTGTCT	3	98	TACACCAAGGAG**gt**acttggggga	3	68
ctgcccccac**ag**GGAGAGCCATAC	4	104	CATCTGCAGAAG**gt**gcaggaggca	4	403
tcccctgctc**ag**GAAGAGATACCG	5	83	GAAACATTCTTG**gt**aaggacaagg	5	932
gcaccatgga**ag**GCAAATACAGCT	6	125	ACCGCCAACATG**gt**gagtgcctta	6	542
cccctgggcc**ag**GTGCAAGAGGAA	7	102	TTCCGGGGCCTG**gt**gagtctgcgc	7	272
tgtctgctcc**ag**GTGGTGCTTGAT	8	93	CCTAAGGTGCAG**gt**gagaggggag	8	1011
ccgcccctgc**ag**ATGCTGATGTAC	9	120	GGCATCGGCCAG**gt**gaggtggcgg	9	1616
tctggctacc**ag**GCACAGTTCTCG	10	57	TTCACCTACCGT**gt**gcctgaggac	10	78
ctggcctaccgtTGCCTTCAGTGG	11	206	ATGGTAGTTTCA**gt**CCAGTGGGTG	11	59
gagcctcccacaGCTGTCACCATG	12	104	CTTTCAATTTCCaaaaaaaaaaaa		

Detail of the exon-intron structure of TM6SF2 using available finished genomic sequence from chromosome 19; the entire transcript consists of 12 exons. All exon-intron boundaries, except for exons 11 and 12, are in accordance with the rule that introns begin with dinucleotide GT and end in AG (in bold).

AAC51362) constitute further examples of synteny conservation between human chromosome 15 and mouse chromosome 7, and between human chromosome 19 and mouse chromosome 8 (Lalley et al., 1991; Lyon and Kirby, 1992) (http://WWW.informatics.jax.org).

Further illustrating the existence of sequence conservation between chromosome 15q24→q26 and 19p13.3→p12, we have found seven other new paralogous pairs of sequences between these two regions (Table 1, examples 5 to 11). These include novel complete genes (BTBD1, HMG20A/HMG20B), partial sequence from unidentified transcripts (EEF2-like-15, BTBD1-19, IRO403-15/403-19p) and already described and characterized genes (EEF2, SH3GL3/SH3GL1, SIN3A/SIN3B and DRIL2/DRIL1). SH3GL3 gene maps on chromosome 15 and its mouse ortholog has been localized to chromosome 7, while the mouse homologue of SH3GL1, the human paralogous gene on chromosome 19, is not yet mapped. Human SIN3B maps to chromosome 19 and its mouse homologue does so on chromosome 8, and SIN3A, on human chromosome 15, has its mouse counterpart on chromosome 9. The latter observation follows the pattern of split conservation of synteny between human chromosome 15 and mouse chromosomes 7 and 9, already illustrated by several known examples: MPI/GPI (mannose phosphate isomerase AAF37697, glucose phosphate isomerase NP_000166); cytochromes P450 CYP11A/CYP2B3 and CYP19/CYP2A6 (NM_000781, M29873, NM_000103 and AF182275 respectively); LIPC/LIPE (hepatic lipase NM_000236 and hormone sensitive lipase L11706 respectively) (Lundin, 1989). These mapping data also agree with the possibility that a process of tetraploidization took place on an ancestral common chromosome for human chromosomes 11, 15 and 19 and mouse chromosomes 7, 8 and 9 before both species diverged (Lundin, 1989, Seldin et al., 1991).

In contrast, other genes on human chromosome 19 have their orthologous gene on mouse chromosome 10, as is the case for HMG20B, suggesting additional correspondences of human chromosome 19 besides mouse chromosome 8. Therefore, additional chromosomal rearrangements must have taken place after the divergence between primates and rodents, which would explain the differences in gene order between both species (Lundin, 1989, Seldin et al., 1991).

Although not conclusive, the limited number of paralogous genes identified up to date is in agreement with the possibility of an ancestral relationship between chromosomes 15 and 19 proposed by Lundin (1989; 1993). If this hypothesis is true, the two chromosomes must have gone through extensive reshuffling, reflected by the differences in gene order between chromosomes and the existence of other interspersed genes without known paralogs. Alternatively, a more simple explanation of the differences in order and in the degree of conservation between each pair of paralogous genes would be that they originated in independent events, at different times and through different mechanisms. The finishing of sequencing the human and other mammalian genomes will provide the means to definitely establish the extent of conservation between the two regions. The identification of new genes on these chromosomes will be of great importance to better understand the evolutionary history that led to the current gene ordering and composition of human chromosomes.

Acknowledgements

We are grateful to A. Puig and D. Otero for technical support with DNA sequencing. We wish to thank the HGMP Resource Center in Hinxton (UK) and the RZPD in Berlin (Germany) for supplying us with IMAGE cDNA clones.

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, et al: Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656 (1991).

Altschul SF, Maden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman, DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search program. Nucl Acids Res 25:3389–3402 (1997).

Carim L, Sumoy L, Nadal M, Estivill X, Escarceller M: Cloning, expression and mapping of PDCD9, the human homologue of *Gallus gallus* pro-apoptotic protein p52. Cytogenet Cell Genet 87:85–88 (1999).

Carim L, Sumoy L. Andreu N, Estivill X, Escarceller M: Identification and expression analysis of C15orf3, a novel gene on chromosome 15q21.1 → q21.2. Cytogenet Cell Genet 88:330–332 (2000a).

Carim L, Sumoy L, Andreu N, Estivill X, Escarceller M: Cloning, expression and mapping of VPS33B, the human orthologue of rat vps33b. Cytogenet Cell Genet 89:92–95 (2000b).

Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L, Matise TC, McKusick KB, Beckmann JS, Bentolila S, Bihoreau M, Birren BB, Browne J, Butler A, Castle AB, Chiannilkulchai N, Clee C, Day PJ, Dehejia A, Dibling T, Drouot N, Duprat S, Fizames C, Bentley DR, et al: A physical map of 30,000 human genes. Science 282:744–746 (1998).

Doege KJ, Sasaki M, Kimura T, Yamada Y: Complete coding sequence and deduced primary structure of the human cartilage large aggregating proteoglycan, aggrecan. Human-specific repeats, and additional alternatively spliced forms. J biol Chem 266:894–902 (1991).

Giachino C, Lantelme E, Lanzetti L, Saccone S, Bella Valle G, Migone N: A novel SH3-containing human gene family preferentially expressed in the central nervous system. Genomics 41:427–434 (1997).

Halleck MS, Pownall S, Harder KW, Duncan AM, Jirik FR, Schlegel RA: A widely distributed putative mammalian transcriptional regulator containing multiple paired amphipathic helices, with similarity to yeast SIN3. Genomics 26:403–406 (1995).

Hart CP, Fainsod A, Ruddle FH: Sequence analysis of the murine Hox-2.2, -2.3 and -2.4 homeoboxes: Evolutionary and structural comparisons. Genomics 1:182–195 (1987).

Kortschak RD, Reimann H, Zimmer M, Eyre HJ, Saint R, Jenne DE: The human dead ringer/bright homolog, DRIL1: cDNA cloning, gene structure, and mapping to D19S886 marker on 19p13.3 that is strictly linked to Peutz-Jeghers syndrome. Genomics 51:288–292 (1998).

Lalley PA, Peters J, Doolittle DP, Hillyard AL, Searle AG: Report of the comparative committee for human, mouse and other rodents. Cytogenet Cell Genet 58:1152–1189 (1991).

Lennon G, Auffray C, Polymeropoulos M, Soares MB: The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. Genomics 33:151–152 (1996).

Lyon MF, Kirby MC: Mouse chromosome atlas. Mouse Genome 90:22–44 (1992).

Lundin LG: Gene homologies with emphasis on paralogous genes and chromosomal regions. Life Sci Adv (Genet) 8:89–104 (1989).

Lundin LG: Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics 16:1–19 (1993).

Osborne-Lawrence SL, Sinclair AK, Hicks RC, Lacey SW, Eddy RL Jr, Byers MG, Shows TB, Duby AD: Complete amino acid sequence of human cartilage link protein (CRTL1) deduced from cDNA clones and chromosomal assignment of the gene. Genomics 8:562–567 (1990).

Prange CK, Pennacchio LA, Lieuallen K, Fan W, Lennon GG: Characterization of the human neurocan gene, CSPG3. Gene 221:199–205 (1998).

Rapp G, Klaudiny J, Hagendorff G, Luck MR, Scheit KH: Complete sequence of the coding regions of human elongation factor 2 (EF2) by enzymatic amplification of cDNA from human ovarian granulosa cells. Biol Chem Hoppe-Seyler 370:1071–1075 (1989).

Schuler GD: Pieces of the puzzle: expressed sequence tags and the catalog of human genes. J molec Med 75:694–669 (1997).

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birre BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Hudson TJ, et al: A gene map of the human genome. Science 274:540–546 (1996).

Seldin MF, Saunders AM, Rochelle JM, Howard TA: A proximal mouse chromosome 9 linkage map that further defines linkage groups homologous with segments of human chromosomes 11, 15 and 19. Genomics 9:678–685 (1991).

Sumoy L, Carim L, Escarceller M, Nadal M, Gratacòs M, Pujana MA, Estivill X, Peral B: HMG20A and HMG20B map to human chromosomes 15q24 and 19p13.3 and constitute a distinct class of HMG-box genes with ubiquitous expression. Cytogenet Cell Genet 88:62–67 (2000).

Yu YT, Breitbart RE, Smoot LB, Lee Y, Mahdavi V, Nadal-Ginard B: Human myocyte-specific enhancer factor 2 comprises a group of tissue-restricted MADS box transcription factors. Genes Dev 6:1783–1798 (1992).

IV. Identificació i caracterització del gen humà BTBD1

L'identificació i caracterització del gen BTBD1 a 15q24 i la descripció d'un gen paràleg al cromosoma 19, BTBD2, són els resultats principals presentats en la publicació següent. Les dues proteïnes contenen dominis BTB que han estat implicats en interaccions entre proteïnes. L'estudi de la distribució tissular de BTBD1 a nivell transcripcional mostra nivells destacats a cor i múscul esquelètic. L'identificació i anàlisi de les seqüències aminoacídiques dels gens ortòlegs murins i bovins indiquen que els gens BTBD1 i 2 constitueixen una nova família de proteïnes. L'existència de BTBD1 i BTBD2 dóna suport a la hipòtesi d'una història evolutiva comuna per als cromosomes 15 i 19 humans.

ELSEVIER

Gene 262 (2001) 275–281

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Identification and characterization of *BTBD1*, a novel BTB domain containing gene on human chromosome 15q24

Laura Carim-Todd, Lauro Sumoy*, Nuria Andreu, Xavier Estivill, Mònica Escarceller

Medical and Molecular Genetics Center, Institut de Recerca Oncològica, Hospital Duran i Reynals, Av. Gran Via s/n km 2,7L'Hospitalet de Llobregat, 08907 Barcelona, Spain

Received 16 May 2000; received in revised form 7 August 2000; accepted 24 October 2000
Received by M. D'Urso

Abstract

Working within the EUROIMAGE full-length cDNA sequencing project we have isolated *BTBD1*, a novel human gene with a BTB/POZ domain. This motif is found in developmentally regulated zinc finger proteins and in the kelch family of actin-associated proteins, and is thought to mediate protein-protein interactions. The *BTBD1* gene encodes a transcript of 3188 nt with an ORF of 482 amino acids and a predicted protein product size of 52.7 kDa. Northern blot analysis revealed an enhanced *BTBD1* expression in heart and skeletal muscle. We have identified a paralogous *BTBD1* counterpart gene on chromosome 19, *BTBD2*. *BTBD1* was mapped to chromosome 15q24. Conservation of multiple pairs of genes between 15q24 and 19p13.3-p12 suggests their possible common chromosomal origin. We show the existence of the murine *BTBD1* and *BTBD2* orthologous genes, as well as the partial rat and bovine homologs. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: BTB (POZ) domain containing 1; EUROIMAGE; 15q24; 19p

1. Introduction

The major goals of the Human Genome Project are the identification of all human genes and the construction of a genome-wide transcript map. The EUROIMAGE Consortium was established in 1997 with the aim of completing the sequence of unique cDNA clones represented in dbEST, obtaining for each one the precise map location and the expression pattern data (Adams et al., 1991; Lennon et al., 1996; Schuler et al., 1996; Schuler, 1997; Deloukas et al., 1998). Our laboratory, as a member of this Consortium, is engaged in the isolation, precise mapping and characterization of novel human genes. We report here the cloning, tissue distribution and chromosome location of a novel BTB domain containing gene, *BTBD1*.

The BTB domain (*B*road-Complex, *T*ramtrack and *B*ric à brac) (Zollman et al., 1994) is also known as POZ domain (*Po*x virus and *Z*inc finger) (Bardwell and Treisman, 1994). The BTB domain is a 120-amino-acid sequence first identified in a set of *Drosophila* and poxvirus genes and it is widely represented in eukaryotic genomes. This evolutionarily conserved protein-protein interaction motif is often found at the N-terminus of developmentally regulated zinc-finger transcription factors, as well as in some actin-associated proteins bearing the kelch motif. Approximately two-thirds of the full-length human BTB genes also encode C_2H_2 zinc finger modules, whereas one half of the remaining entries contain the kelch motif. It is known that the domain mediates homomeric dimerization and in some instances heterodimeric associations with other BTB domains (Bardwell and Treisman, 1994). The crystal structure of the dimerized PLZF (human promyelocytic leukemia zinc finger protein) BTB domain has been solved and consists of a tightly intertwined homodimer with an extensive hydrophobic interface (Ahmad et al., 1998).

2. Materials and methods

2.1. cDNA isolation and sequence analysis

EST clusters were assembled using the EST CAP assembly program (http://www.tigem.it) and the Sequencher software for Macintosh (GeneCodes Corporation). Clones were

Abbreviations: BTB domain, Broad-Complex, Tramtrack and Bric á brac; Poz domain, Pox virus and Zinc finger; BTBD1, BTB/POZ domain protein 1; BTBD2, BTB/POZ domain protein 2, cDNA, complementary DNA; EST, expressed sequence tag; nt, nucleotide; bp, base pair; RACE, rapid amplification of cDNA ends; ORF, open reading frame; UTR, untranslated region; htgs, high-throughput genomic sequence

* Corresponding author. Tel.: 34-93-260-7775; fax: 34-93-260-7776.
E-mail address: lsumoy@iro.es (L. Sumoy).

0378-1119/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0378-1119(00)00513-8

obtained from the EUROIMAGE distribution centers (HGMP, RZPD). Sequences were determined by primer walking using custom synthesized primers (LifeTech) with the Perkin–Elmer BigDye reagents on an ABI-377 fluorescent automated sequencer. Each clone was sequenced on both strands with at least three independent reads per base.

BTBD1 full-length 5′ cDNA sequence was obtained using the rapid amplification of cDNA ends (RACE) method on RACE MARATHON™ cDNA from adult heart tissue (Clontech), according to the manufacturers kit instructions. The following primers were used: G1 (5′ ATCCATTGTGCTTTTGTCTATTGTATC 3′), G2 (5′ ATTATCTGCCCTAAGATGTTTGGTGAG 3′) and G3 (5′ GTTATGACCACTCTTTATACTGCCAAG 3′). We sought four independently generated fully extended clones to determine the cDNA 5′ end. PCR extended products were subcloned into the pGEM-T easy vector (Promega) and sequenced as above. *BTBD2* was extended by RACE on human fetal brain SMART cDNA (Clontech) using primers: G3 (5′ CGCTTCGCCTTCCTCTTCAACAAC 3′), G4 (5′ ATCCATACAGCCCAAATCCCACC 3′), G2 (5′ TTGTCGGCTCGCAGGTTCTTC 3′). PCR products were subcloned and sequenced as above.

Sequence comparisons were performed using ClustalW and boxed multiple sequence alignments were obtained with the BOXSHADE 3.21 program (http://www.isrec.isb-sib.ch/software/BOX_form.html). Protein pattern and domain databases consulted for prediction of known motifs or functional domains were Prosite, SMART and Pfam (http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-domain.html).

2.2. Northern blot analysis

For detection of *BTBD1*, a human multiple-tissue northern blot (MTN blot, Clontech) was hybridized with a 0.8 kb *Hind*III-*EcoR*I probe belonging to the 5′ end of human IMAGE cDNA clone 28577 and a 2 kb β-actin probe supplied commercially (Clontech) was used as control for quantification. Probes were labeled using a random primer DNA labeling kit (Amersham Pharmacia). Blots were hybridized overnight at 65°C in ExpressHyb hybridization solution (Clontech) and washed at 68°C in 0.2 × SSC/0.5%SDS.

2.3. Mapping of BTBD1

To precisely localize the *BTBD1* gene we used the Stanford TNG4 whole genome radiation hybrid panel (Stewart et al., 1997). Two point linkage analysis was performed using the RHMAP-2.0 on the RH Server at the Stanford Human Genome Center (http://www-shgc.stanford.edu/RH/index.html). We used primers corresponding to STS marker SHGC-15202 (D15S1261): F (5′ CAGTTTAGTGACAGGGAAT 3′) and R (5′ TCTTGTTGTTAGCATTTGTA 3′). The PCR conditions were 1 cycle at 94°C for 3 min; 35 cycles at 94°C for 30 s, 54°C for 30 s and 72°C for 1 min; and 1 cycle at 72°C for 5 min.

3. Results

In our effort to identify new genes, we construct and analyze in silico unique gene EST clusters. Using this approach, a human cDNA sequence with a partial single open reading frame (ORF) was identified during the analysis of EST clusters spanning chromosome 15q. The cDNA clones associated to the ESTs belonged to Unigene cluster Hs. 21332; among all, IMAGE clones 549016 (EST GenBank Acc. No. AA083385), 2325042 (EST GenBank Acc. No. AI677979) and 28577 (EST GenBank Acc. No. R40907) were chosen for sequencing (Fig. 1A). Clone 549016 was chimeric and clone 28577 did not cover the complete mRNA sequence. Further completion of the *BTBD1* transcript was performed using the RACE approach on human heart cDNA. The final assembly of all the different clones gave a total transcript 3188 bp long containing an ORF (from nt 83–1528) encoding a 482 amino acid product with a calculated mass of 52.7 kDa and an estimated pI of 5.7. The 5′ untranslated region (UTR) contained an in-frame stop codon at nt position 56. A polyadenylation signal (AATAAA) was observed at nt 3,146 and a polyA tail at the end (3169 nt).

The gene was designated *BTBD1* following the Human Gene Nomenclature Committee instructions (http://www.gene.ucl.ac.uk/nomenclature/). *BTBD1* nucleotide and protein sequences are available in GenBank under Acc. No. AF257241.

Analysis of the amino acid sequence of BTBD1 with protein domain identification software revealed the presence of one single type BTB domain spanning residues 69 to 175 (Fig. 1A, B).

BLAST homology searching against non-redundant databases (NCBI) (http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/; Altschul et al., 1997) revealed that BTBD1 was remarkably similar to a hypothetical partial ORF on chromosome 19p13.3 (GenBank Acc. No. AAC26984). With lower significance, hits following were the F38H4.7 'similar to BTB protein' in *C. elegans* (GenBank Acc. No. CAB01179) and the human gene products KIAA0952 protein (GenBank Acc. No. BAA76796) and a second unnamed hypothetical protein (GenBank Acc. No. CAB70908).

Launching BLAST searches against the unfinished High Throughput Genomic Sequences (htgs) database we obtained a perfect match to BAC clone RP11-17G13 (GenBank Acc. No. AC018910). The availability of the genomic sequence allowed us to determine the exon-intron structure of *BTBD1*. The entire transcript consists of eight exons. Details of the exon-intron structure are shown in Table 1. All the junctions between exons and introns are in accordance with the rule that introns begin with a GT dinucleotide and end with AG. The first exon encodes a 5′ untranslated region (UTR), and the last one encodes a very long 3′ UTR with five non-canonical polyadenylation signals.

Based on the remarkable homology at the protein level between BTBD1 and the partial putative ORF on chromosome 19 (82% identity, 89% similarity in the aligned region), we tried to identify this other complete putative gene,

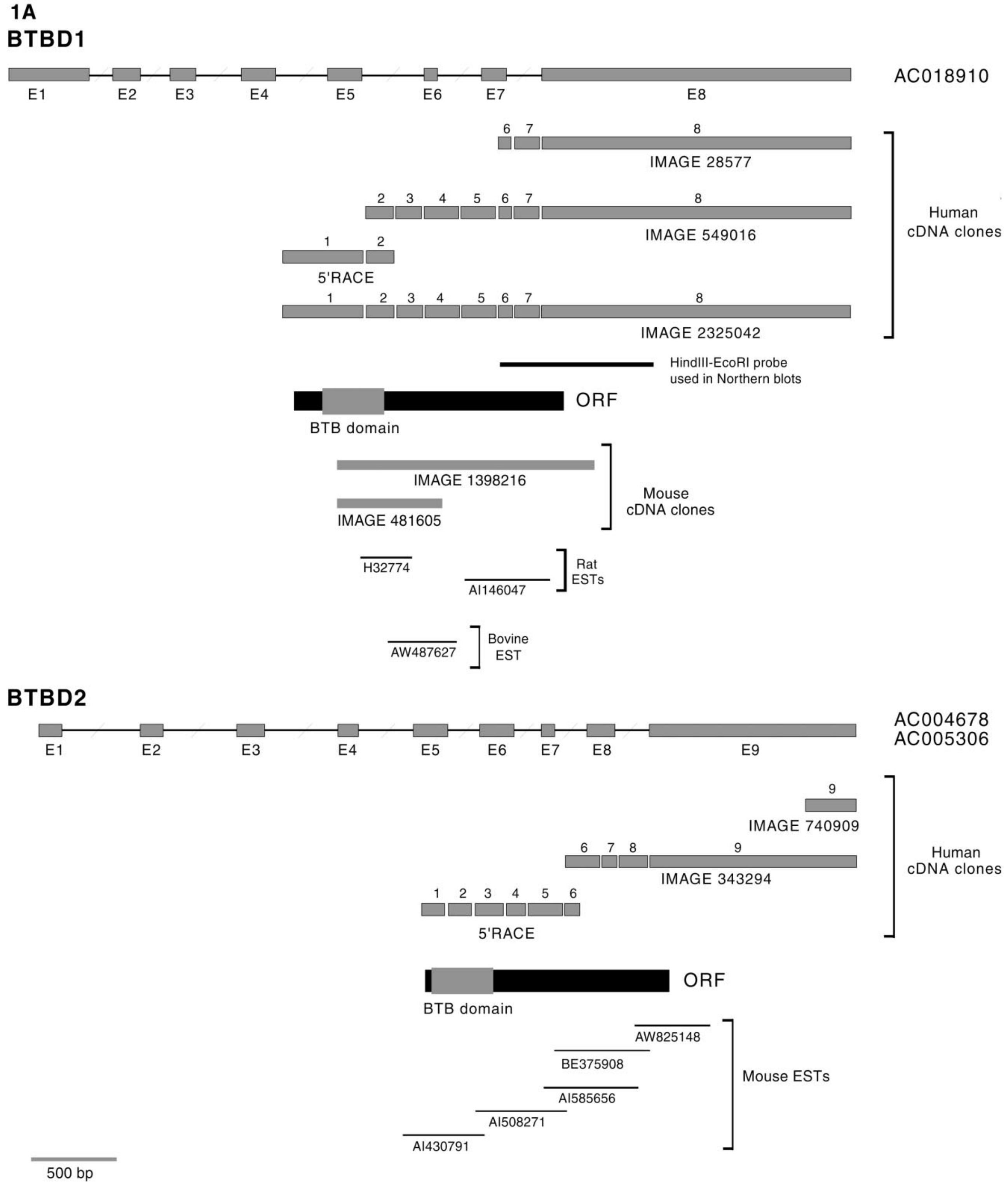


Fig. 1. (A) Diagram illustrating the assembly procedure of the human, mouse, rat and bovine *BTBD* genes. Introns are not proportional in size (see Table 1 for information about intron sizes when available).

(*Continued overleaf*)

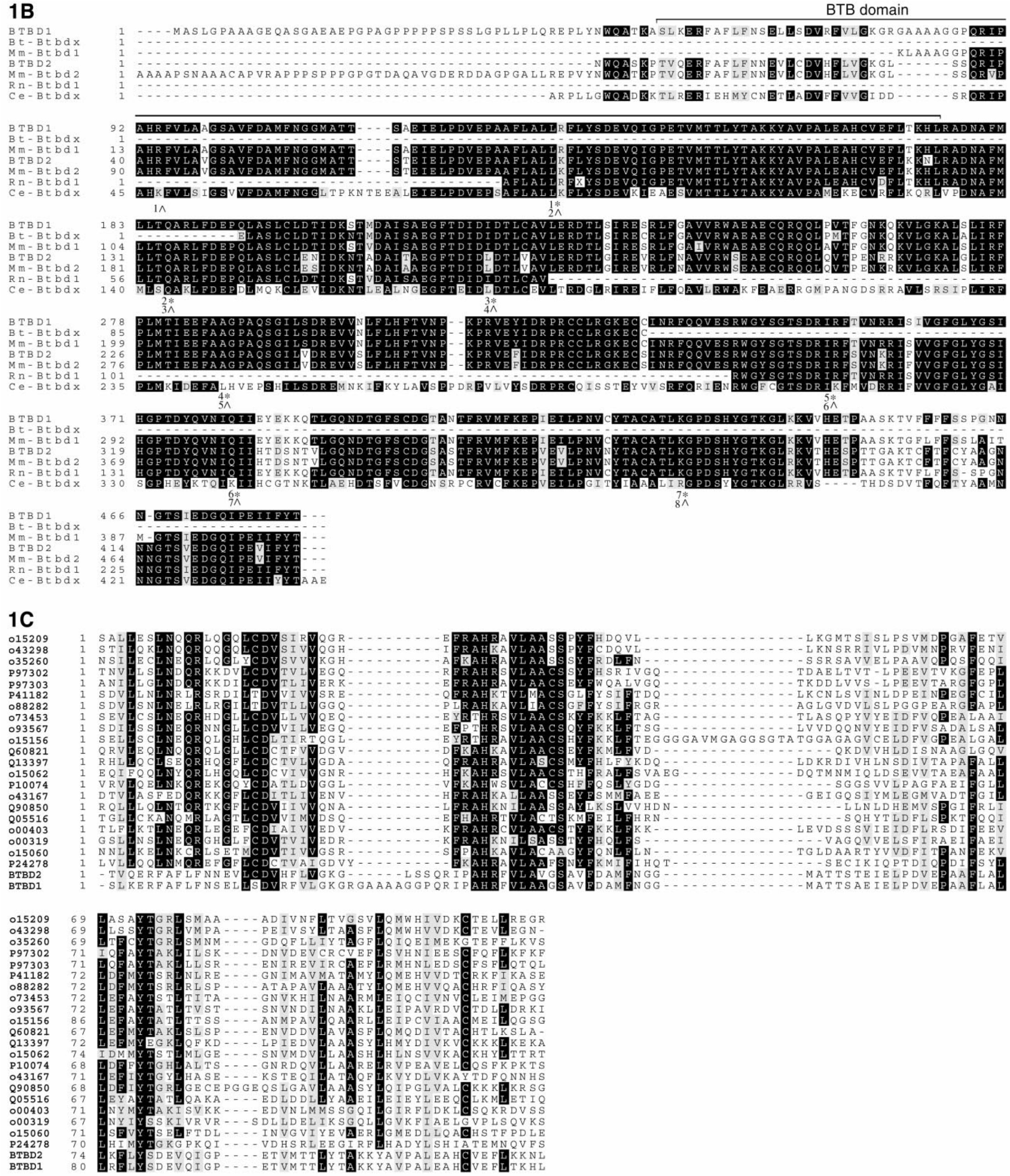


Fig. 1. (B) Amino acid multiple sequence alignment of human BTBD1 and BTBD2 along with the corresponding murine orthologs and the rat and bovine homolog partial sequences. The BTB domain is highlighted. The position of introns is indicated with a * for BTBD1 and with a ^ for BTBD2. (C) Multiple sequence alignment of BTB domains including human BTBD1 and BTBD2. Proteins are named according to the accession number assigned by the Swiss-Prot database (http://srs.ebi.ac.uk/swissprot). Bt, *Bos taurus*; Mm, *Mus musculus*; Rn, *Rattus norvegicus*; Ce, *Caenorhabditis elegans*.

BTBD2. We selected three human cDNA clones that matched perfectly with the genomic sequence: IMAGE clone 2154200 (EST GenBank Acc. No. AI445279), IMAGE clone 740909 (EST GenBank Acc. No. AA478274) and IMAGE clone 343294 (EST GenBank Acc. No. W68046). The last two clones belonging to cluster Unigene Hs.25817, associated to cosmid R27216 on chromosome 19 (GenBank Acc. No. AAC26984), were fully sequenced and the assembly of them gave only partial coding sequence. We further extended the 5′ end of the *BTBD2* cDNA by 1 kb using the RACE technique on human fetal brain RNA but this did not cover the entire coding region of the gene (Fig. 1A). Repeated attempts on other tissue sources known to express *BTBD2* were unsuccessful.

To find putative homologues of human *BTBD1* and *BTBD2* in other species we performed BLAST homology searches against 'mouse' and 'other' dbEST (NCBI). We obtained significant scores with mouse ESTs GenBank Acc. Nos. AA929256, AA060112 that corresponded to the *BTBD1* mouse orthologous gene. The following IMAGE clones were fully sequenced and assembled: 1398216 and 481605 (Fig. 1A). We also identified mouse ESTs belonging to the *BTBD2* mouse ortholog, GenBank Acc. Nos. AI430791, AI508271, AI585656, BE375908 and AW825148 and we were able to predict partial protein sequence. A high degree of amino acid sequence conservation between each other was detected (98% similarity between BTBD1 and its mouse ortholog and 95% similarity for BTBD2 and its own mouse counterpart). Other putative *BTBD1* orthologs represented by partial sequences from rat (GenBank Acc. Nos. H32774 and AI146047) and bovine (GenBank Acc. Nos. AW487627) were found to share 97 and 98% similarity respectively in the aligned regions (Fig. 1B).

3.1. Expression pattern of BTBD1

Expression analysis of *BTBD1* was carried out by hybridization with a specific probe using a human tissue northern blot (MTN Human blot, Clontech) (see Section 2). In adult tissues expression showed a majoritary 3.2 kb mRNA species (Fig. 2). Ubiquitous basal expression of the gene was detected in all tissues, with higher levels in heart and skeletal muscle, lower in brain, kidney and pancreas. A faint signal was detected in liver, placenta and lung. A 2.4 kb secondary transcript was barely detectable in heart and skeletal muscle.

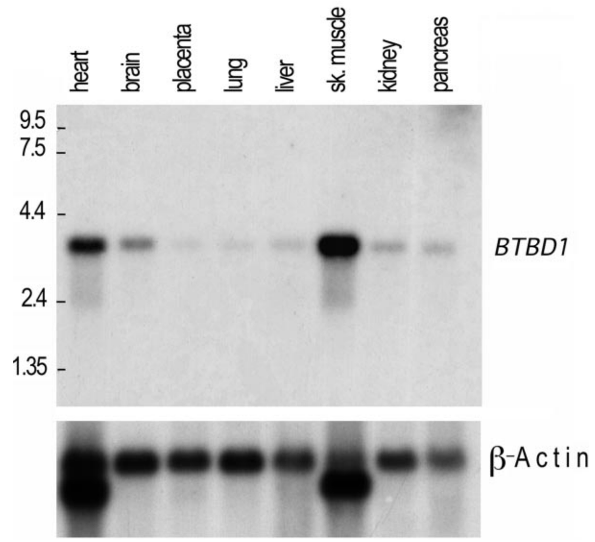


Fig. 2. Northern blot analysis of *BTBD1* mRNA expression in adult human tissues. The *BTBD1* cDNA probe reveals an approximately 3.2 kb size mRNA species. *BTBD1* and β-actin transcripts are indicated.

Table 1
Exon-intron structure of the *BTBD*genes[a]

3′ Splice acceptor	Exon	Size (bp)	5′ Splice donor	Intron	Size (bp)
***BTBD1* Chr15**					
5′UTR GAGGCAGCGCCG	1	483	GGCGCTGCTGAG**gt**gagcggcagc	1	–
attgttttcc**ag**ATTTCTATATTC	2	157	TTACTTACTCAG**gt**aagtaaatgt	2	–
ttatatgttt**ag**GCTCGATTATTT	3	148	ATATTGATATAG**gt**aagttcgcta	3	–
ttttgcctat**ag**ATACACTCTGTG	4	198	AATTTGCAGCAG**gt**aagggtataa	4	–
ttatttccac**ag**GTCCTGCTCAAT	5	193	TGATCGAATCAG**gt**atctgttata	5	–
tctttctctc**ag**ATTCACAGTTAA	6	88	GTGAATATACAG**gt**acagtttcct	6	–
taaaaattac**ag**ATCATTGAATAT	7	147	GCAACACTCAAA**gt**aagagtcatg	7	–
gttctcattt**ag**GGTCCAGATTCC	8	1796	ATATATATATATaaaaaaaaaaaa	polyA	
***BTBD2* Chr19**					
AACTGGCAGGCC	1	125	CCCCGCGCACAG**gt**gggcgccccg	1	17833
ctcccgtacc**ag**GTTCGTGCTGGC	2	120	CGCACTGCTCAA**gt**aatgcttccg	2	4167
ttgcttctgc**ag**GTTTCTCTACTC	3	157	CTGCTCACGCAG**gt**gggcggggcc	3	–
gtctcttgcc**ag**GCGCGACTCTTC	4	106	ACATTGACCTGG**gt**aagggcccag	4	515
tccaccccac**ag**ACACGCTGGTGG	5	198	AGTTCGCTGCAG**gt**aacagagctc	5	2311
cctgtgctgc**ag**GTCCCGCACAGT	6	193	TGACCGCATCAG**gt**ggggcttggg	6	246
ctgccacggc**ag**GTTCTCAGTCAA	7	88	GTGAACATCCAG**gt**accagcccca	7	189
acctccccgc**ag**ATTATTCACACC	8	147	GCCACGCTCAAG**gt**gcgccgccgg	8	180
tattccctac**ag**GGCCCAGACTCC	9	1196	AATGCACCTGCCaaaaaaaaaaaa	polyA	

[a] Detail of the exon/intron structure of the *BTBD1* and *BTBD2* genes; the entire *BTBD1* transcript consists of eight exons. For BTBD2 nine exons have been identified. All the junctions between exons and introns are in accordance with the rule that introns begin with a GT dinucleotide and end with AG.

3.2. Mapping of *BTBD1*

Chromosomal localization of the human *BTBD1* gene was determined by radiation hybrid mapping using the Stanford TNG4 panel (Stewart et al., 1997). The gene was linked to STS SHGC-2198 with a LOD score of 7.28 at an approximate distance of 116 kb. STS SHGC-2198 is the microsatellite marker D15S200.

Consistent with this result, by BLAST searching against 'htgs' database at NCBI, we have found that the *BTBD1* gene is present in the genomic sequence of RPCI-11 BAC clone 17G13 (AC018910) which contains SHGC-35969 (http://genome.wustl.edu/gsc/cgi-bin/ace/ctc_choices/ ctc.ace). Both WI-13449 (belonging to Unigene cluster Hs. 21332) and SHGC-35969 had been mapped previously on chromosome 15q24 between D15S115-D15S152, where D15S200 also maps.

4. Discussion

Working within the EUROIMAGE full-length cDNA sequencing project, we have identified and characterized a novel human gene, *BTBD1,* and shown the existence of a close paralog gene, *BTBD2*. We have deduced the partial putative amino acid sequence of the *BTBD1* and *BTBD2* orthologous genes in mouse, as well as identified the partial rat and bovine homologues, suggesting the existence of a new group of mammalian proteins. This fact together with the identification of a closely related gene in such a distant species as *C. elegans* may be an indication of a conserved functional role of this gene in basic cellular processes that has been maintained throughout the evolution of these species.

The most relevant feature of BTBD1 is the presence of a BTB/POZ domain in its amino acid sequence (Fig. 1). The BTB domain defines a large family whose members function in a variety of biological processes (Zollman et al., 1994; Bardwell and Treisman, 1994). The majority of BTB-containing genes also encode zinc-finger or kelch motifs. In Drosophila, proteins containing both BTB and zinc finger domains have been associated with a variety of processes including nucleosome/chromatin disruption, pattern formation, metamorphosis, oogenesis, and eye and limb development (Bardwell and Treisman, 1994 and references therein). Also in Drosophila, the 'kelch (BTB containing) protein' is localized in specialized structures of the intercellular bridges that connect the nurse cells of the developing oocyte, probably through the oligomerization activity of the BTB domain (Xue and Cooley, 1993). In humans, two BTB-zinc finger genes, PLZF and Bcl6/LAZ3, are associated with chromosomal translocation breakpoints and their BTB domain has been shown to mediate transcriptional repression through the local control of chromatin conformation, interacting with components of the histone deacetylase co-repressor complexes SMRT and N-CoR (Deweindt et al., 1995; Dhordain et al., 1995; Wong and Privalsky, 1998). In the ZID (zinc finger protein with interaction domain) the BTB domain inhibits DNA binding and localizes the protein to the nucleus (Bardwell and Treisman, 1994; Deweindt et al., 1995).

Since BTBD1 contains neither zinc finger nor kelch motifs, perhaps it represents a further class of BTB domain proteins. The ability of the BTB domain to form homo and heterodimeric associations with other BTB domains suggests the possibility that the BTB motif could be allowing combinatorial diversity of complexes of BTBD1 with other proteins. These complexes might be tissue-specific, and as the *BTBD1* uneven expression pattern indicates, BTBD1 could perform distinct roles, particularly in heart and skeletal muscle in which enhanced expression is detected.

We have finely mapped *BTBD1* on chromosome 15q24 and found a remarkably homologous gene on chromosome 19, *BTBD2*. Both genes share a high degree of similarity, reflected in the amino acid sequence conservation (89% similarity) and in the exon-intron structure of both genes (Fig. 1 and Table 1). The identification of a chromosome 19 *BTBD1* paralogous gene has confirmed and extended a recent observation by our group: we have identified at least nine genes in chromosome 15q24-q26 that share close homology with counterparts in 19p13.3-12 (Carim-Todd et al. in press). In reviewing the literature and the current transcript map of the genome (http://www.ncbi.nlm.nih.gov/genemap98/) looking for other possible cases, we have found a series of loci with paralogs in 15q and 19p. Previously, it had been proposed that three chromosomal regions on 15q, 5q and 19p correspond to large clusters of genes possibly derived by tetraploidization of an ancestral genome (Lundin, 1993 and references therein; Hallbook, 1999). The existence of paralogous gene clusters in 15q and 19p is consistent with this and the hypothesis that the vertebrate genome has evolved through a series of chromosomal duplications. *BTBD1* and *BTBD2* would be an example of genes that originated through events such as these. They may be performing similar, redundant or complementary functions. Experiments will be needed to determine the functional overlap between the two BTBD protein products.

Acknowledgements

We are grateful to A. Puig and D. Otero for technical support with DNA sequencing. We wish to thank the HGMP Resource Center in Hinxton, UK, and the RZPD in Berlin, Germany, for supplying us with IMAGE cDNA clones. This work has been supported by EU Biomed Project No. BMH4-CT97-2284 to X. E and by CICYT-IN95-0347. M.E. is funded by the Spanish Ministry of Education (CIDYT contract FPI-070-97) and L.S. by the Catalan autonomous government (CIRIT-RED contract 1998-64).

References

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al., 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252, 1651–1656.

Ahmad, K.F., Engel, C.K., Prive, G.G., 1998. Crystal structure of the BTB domain from PLZF. Proc. Natl. Acad. Sci. USA 95, 12123–12128.

Altschul, S.F., Maden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search program. Nucleic Acids Res. 25, 3389–3402.

Bardwell, V.J., Treisman, R., 1994. The POZ domain: a conserved protein-protein interaction motif. Genes Dev. 8, 1664–1677.

Deloukas, P., Schuler, G.D., Gyapay, G., Beasley, E.M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T.C., McKusick, K.B., Beckmann, J.S., Bentolila, S., Bihoreau, M., Birren, B.B., Browne, J., Butler, A., Castle, A.B., Chiannilkulchai, N., Clee, C., Day, P.J., Dehejia, A., Dibling, T., Drouot, N., Duprat, S., Fizames, C., Bentley, D.R., et al., 1998. A physical map of 30,000 human genes. Science 282, 744–746.

Deweindt, C., Albagli, O., Bernardin, F., Dhordain, P., Quief, S., Lantoine, D., Kerckaert, J.P., Leprince, D., 1995. The LAZ3/BCL6 oncogene encodes a sequence-specific transcriptional inhibitor: a novel function for the BTB/POZ domain as an autonomous repressing domain. Cell Growth Differ. 6, 1495–1503.

Dhordain, P., Albagli, O., Ansieau, S., Koken, M.H., Deweindt, C., Quief, S., Lantoine, D., Leutz, A., Kerckaert, J.P., Leprince, D., 1995. The BTB/POZ domain targets the LAZ3/BCL6 oncoprotein to nuclear dots and mediates homomerisation in vivo. Oncogene 11, 2689–2697.

Hallbook, F., 1999. Evolution of the vertebrate neurotrophin and trk receptor gene families. Curr. Opin. Neurobiol. 9, 616–621.

Lennon, G., Auffray, C., Polymeropoulos, M., Soares, M.B., 1996. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. Genomics 33, 151–152.

Lundin, L.G., 1993. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house man. Genomics 16, 1–19.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birre, B.B., Butler, A., Castle, A.B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P.J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., East, C., Hudson, T.J., et al., 1996. A gene map of the human genome. Science 274, 540–546.

Schuler, G.D., 1997. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. J. Mol. Med. 75, 694–769.

Stewart, E.A., McKusick, K.B., Aggarwal, A., Bajorek, E., Brady, S., Chu, A., Fang, N., Hadley, D., Harris, M., Hussain, S., Lee, R., Maratukulam, A., O'Connor, K., Perkins, S., Piercy, M., Qin, F., Reif, T., Sanders, C., She, X., Sun, W.L., Tabar, P., Voyticky, S., Cowles, S., Fan, J.B., Cox, D.R., et al., 1997. An STS-based radiation hybrid map of the human genome. Genome Res. 7, 422–433.

Wong, C.W., Privalsky, M.L., 1998. Components of the SMRT corepressor complex exhibit distinctive interactions with the POZ domain oncoproteins PLZF PLZF-RARalpha, and BCL-6. J. Biol. Chem. 273, 27695–27702.

Xue, F., Cooley, L., 1993. Kelch encodes a component of intercellular bridges in Drosophila egg chambers. Cell 72, 681–693.

Zollman, S., Godt, D., Prive, G.G., Couderc, J.L., Laski, F.A., 1994. The BTB domain, found primarily in zinc finger proteins, defines an evolutionarily conserved family that includes several developmentally regulated genes in Drosophila. Proc. Natl. Acad. Sci. USA 91, 10717–10721.

V. Identificació i caracterització del gen UBXD1

El següent article és resultat de l'observació de paralogia entre 15q24-q26 i 19p13.3-p12 que va impulsar l'anàlisi *in silico* d'aquesta última regió cromosòmica. L'identificació de gens humans no coneguts al cromosoma 19 és conseqüència directa d'aquesta aproximació. En aquest cas es descriu l'identificació del gen UBXD1. L'anàlisi de la seva seqüència permet definir una nova classe de proteïnes amb dominis UBX, conservades al llarg de l'evolució. La seva caracterització a nivell de patró d'expressió mostra una distribució ubícua amb nivells més elevats a testicle. Finalment, la detecció dels corresponents ortòlegs a rata i ratolí són resultats que permeten inferir o suggerir funcions potencials per aquests tipus de gens alhora que defineixen una nova classe de proteïnes amb dominis UBX conservades al llarg de l'evolució dels vertebrats.

ELSEVIER

Biochimica et Biophysica Acta 91489 (2000) 1–4

BIOCHIMICA ET BIOPHYSICA ACTA
BBA

www.elsevier.com/locate/bba

Short sequence-paper

Identification and characterization of UBXD1, a novel UBX domain-containing gene on human chromosome 19p13, and its mouse ortholog

Laura Carim-Todd, Mònica Escarceller, Xavier Estivill, Lauro Sumoy *

Medical and Molecular Genetics Center, Institut de Recerca Oncològica, Hospital Duran i Reynals, Av. Gran Via s/n, km 2,7, L'Hospitalet de Llobregat, 08907 Barcelona, Spain

Received 19 July 2000; accepted 3 October 2000

Abstract

We have identified a novel human gene, UBXD1, on chromosome 19p13, which encodes a putative protein containing a UBX domain. Expression analysis showed an enhanced presence in testis. We identified the corresponding orthologous genes in mouse and rat. The characterization of UBXD1 has allowed us to define a new class of UBX domain-containing proteins conserved during evolution. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: UBXD1; UBX domain; EUROIMAGE; Ubiquitin; cDNA sequencing

The identification of all human genes and the construction of a genome-wide transcript map are one of the major goals of the Human Genome Project. In order to achieve this, the EUROIMAGE Consortium was set up to characterize unique cDNAs represented in dbEST [1,2,3,4,5]. Within this Consortium we have analyzed in silico over 400 EST clusters (EST CAP assembly program, http://www.tigem.it; Sequencher, GeneCodes Corporation). A putative open reading frame (ORF) was identified in cluster Hs.11081 and the following IMAGE clones were fully sequenced by primer walking (custom synthesized primers, LifeTech, Perkin-Elmer BigDye reagents on an ABI-377): 345397 (EST GenBank Acc. No. W72616), 1649435 (EST GenBank Acc. No. AI027145) and 2190872 (EST GenBank Acc. No. AI682078). Clone 345397, representing a 1820 kb transcript, contained an in frame stop codon at nucleotide (nt) position 87 preceding an ORF of 388 amino acids with a calculated mass of 43.80 kDa and an estimated p*I* of 5.54. Clone 1649435 did not show a 5′ stop codon, and the resulting ORF was a 441 amino acid product with a predicted mass of 49.75 kDa and an estimated p*I* of 6.46 (Fig. 1A).

Analysis with protein domain identification software revealed the presence of a unique UBX domain spanning amino acids 329–410 within the C-terminal region of the protein (Fig. 1B). The gene was therefore designated UBXD1 following the Human Gene Nomenclature Committee rules (http://www.gene.ucl.ac.uk/nomenclature/).

BLAST searches (http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/) [6] with UBXD1 against the unfinished high throughput genomic sequences (htgs) database gave a perfect match to the working draft sequence from clone CTB_50L17 on chromosome 19, GenBank Acc. No. AC011498. Using the non-redundant database the most significant hit was a partially characterized genomic sequence from chromosome 19p13, GenBank Acc. No. AF190465. Using this information we established the exon–intron structure of the gene as shown in Table 1. Analysis of the genomic sequence upstream from the 5′ end of our longest cDNA detected an in frame stop codon. Assuming this corresponds to a transcribed sequence the resulting transcript would be at least 1659 nt in length, which is in accordance with the sizes of mRNAs observed by Northern blot (Fig. 2).

We identified murine ESTs (Unigene cluster Mm.28000) corresponding to the orthologous gene in this species. We sequenced clone 1195567 (EST GenBank Acc. No. AA711688) obtaining a 442 amino acid ORF with a predicted mass of 49.79 kDa and a theoretical p*I* of 8.69 (Fig. 1A). Both the human and mouse genes coincide in the position of the most 5′ methionine and significantly diverge upstream from this position indicating this must act as an initiation codon. Further database searches identified ESTs corresponding to the rat orthologous gene

* Corresponding author. Fax: +34-93-260-7776; E-mail: lsumoy@iro.es

0167-4781/00/$ – see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0167-4781(00)00248-7

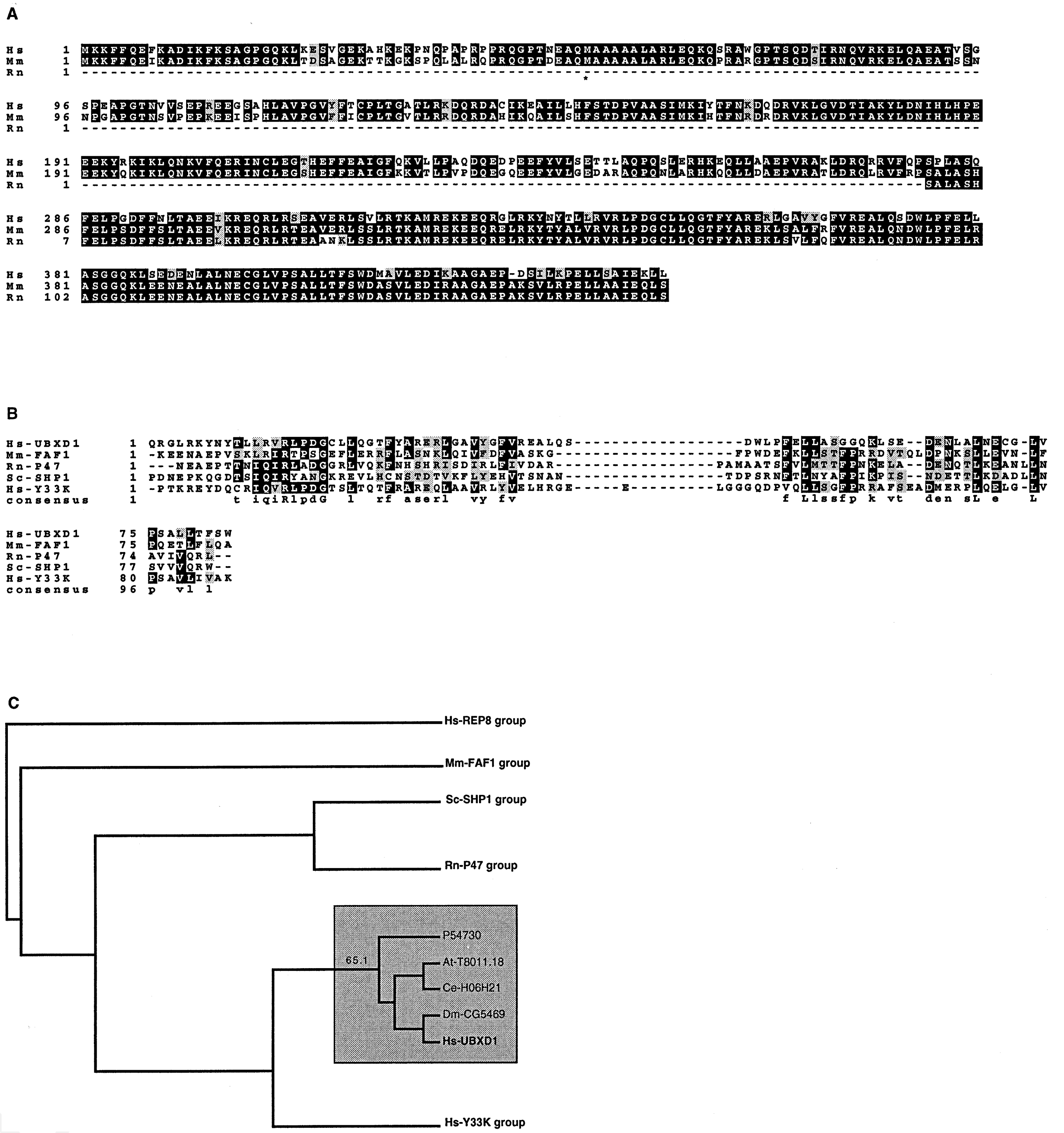

A
Hs 1 MKKFFQEFKADIKFKSAGPGQKLKESVGEKAHKEKPNQPAPRPPRQGPTNEAQMAAAAALARLEQKQSRAWGPTSQDTIRNQVRKELQAEATVSG
Mm 1 MKKFFQEIKADIKFKSAGPGQKLTDSAGEKTTKGKSPQLALRQPRQGPTDEAQMAAAAALARLEQKQPRARGPTSQDSIRNQVRKELQAEATSSN
Rn 1 ---
*
Hs 96 SPEAPGTNVVSEPREEGSAHLAVPGVYFTCPLTGATLRKDQRDACIKEAILLHFSTDPVAASIMKIYTFNKDQDRVKLGVDTIAKYLDNIHLHPE
Mm 96 NPGAPGTNSVPEPKEEISPHLAVPGVFFICPLTGVTLRRDQRDAHIKQAILSHFSTDPVAASIMKIHTFNRDRDRVKLGVDTIAKYLDNIHLHPE
Rn 1 ---
Hs 191 EEKYRKIKLQNKVFQERINCLEGTHEFFEAIGFQKVLLPAQDQEDPEEFYVLSETTLAQPQSLERHKEQLLAAEPVRAKLDRQRRVFQPSPLASQ
Mm 191 EEKYQKIKLQNKVFQERINCLEGSHEFFEAIGFKKVTLPVPDQEGQEEFYVLGEDARAQPQNLARHKQQLLDAEPVRATLDRQLRVFRPSALASH
Rn 1 ---SALASH
Hs 286 FELPGDFFNLTAEEIKREQRLRSEAVERLSVLRTKAMREKEEQRGLRKYNYTLLRVRLPDGCLLQGTFYARERLGAVYGFVREALQSDWLPFELL
Mm 286 FELPSDFFSLTAEEVKREQRLRTEAVERLSSLRTKAMREKEEQRELRKYTYALVRVRLPDGCLLQGTFYAREKLSALFRFVREALQNDWLPFELR
Rn 7 FELPSDFFSLTAEELKREQRLRTEAANKLSSLRTKAMREKEEQRELRKYTYALVRVRLPDGCLLQGTFYAREKLSVLFQFVREALQNDWLPFELR
Hs 381 ASGGQKLSEDENLALNECGLVPSALLTFSWDMAVLEDIKAAGAEP-DSILKPELLSAIEKLL
Mm 381 ASGGQKLEENEALALNECGLVPSALLTFSWDASVLEDIRAAGAEPAKSVLRPELLAAIEQLS
Rn 102 ASGGQKLEENEALALNECGLVPSALLTFSWDASVLEDIRAAGAEPAKSVLRPELLAAIEQLS
B
Hs-UBXD1 1 QRGLRKYNYTLLRVRLPDGCLLQGTFYARERLGAVYGFVREALQS------------------DWLPFELLASGGQKLSE--DENLALNECG-LV
Mm-FAF1 1 -KEENAEPVSKLRIRTPSGEFLERRFLASNKLQIVFDFVASKG-------------------FPWDEFKLLSTFPRRDVTQLDPNKSLLEVN-LF
Rn-P47 1 ---NEAEPTTNIQIRLADGGRLVQKFNHSHRISDIRLFIVDAR-----------------PAMAATSFVLMTTFPNKELA--DENQTLKEANLLN
Sc-SHP1 1 PDNEPKQGDTSIQIRYANGKREVLHCNSTDTVKFLYEHVTSNAN-----------------TDPSRNFTLNYAFPIKPIS--NDETTLKDADLLN
Hs-Y33K 1 -PTKREYDQCRIQVRLPDGTSLTQTFRAREQLAAVRLYVELHRGE-----E---------LGGGQDPVQLLSGFPRRAFSEADMERPLQELG-LV
consensus 1 t iqiRlpdG l rf aserl vy fv f Llssfp k vt den sL e L
Hs-UBXD1 75 PSALLTFSW
Mm-FAF1 75 PQETLFLQA
Rn-P47 74 AVIVQRL--
Sc-SHP1 77 SVVVQRW--
Hs-Y33K 80 PSAVLIVAK
consensus 96 p vl l
C
Hs-REP8 group
Mm-FAF1 group
Sc-SHP1 group
Rn-P47 group
65.1
P54730
At-T8011.18
Ce-H06H21
Dm-CG5469
Hs-UBXD1
Hs-Y33K group

Fig. 1. (A) ClustalW multiple sequence alignment of human UBXD1 protein and its homolog in mouse, Ubxd1, along with the predicted homolog protein in rat. Identical residues are printed in reverse type and similar residues are shaded. The asterisk marks the methionine in the human alternative transcript (Genbank Acc. No.: AF272893 and AF272894 for human UBXD1; AF272895 for mouse Ubxd1). (B) Alignment with published UBX-containing proteins and the resulting consensus sequence. Identical residues in at least half of the sequences are printed in reverse type and similar residues are shaded; Hs, *Homo sapiens*; Mm, *Mus musculus*; Rn, *Rattus norvegicus*; Dm, *Drosophila melanogaster*; At, *Arabidopsis thaliana*; Ce, *Caenorhabditis elegans*. (C) Unrooted consensus tree of a set of unique UBX domains and UBXD1-related. Only forks with significant values above 50% are shown. Domains labeled by gene name (when referenced in this article) or by the Swiss-Prot or TREMBL accession number.

(Unigene cluster Rn.7230) and resulted in an 800 bp contig corresponding to the 3′ end of the gene.

The degree of conservation at the amino acid level between the three mammalian genes is remarkable: 80% identity and 87% similarity between human and mouse, and 79% identity and 88% similarity between human and rat in the aligned region. Mouse and rat sequences are highly similar (96% identity and 97% similarity) as would

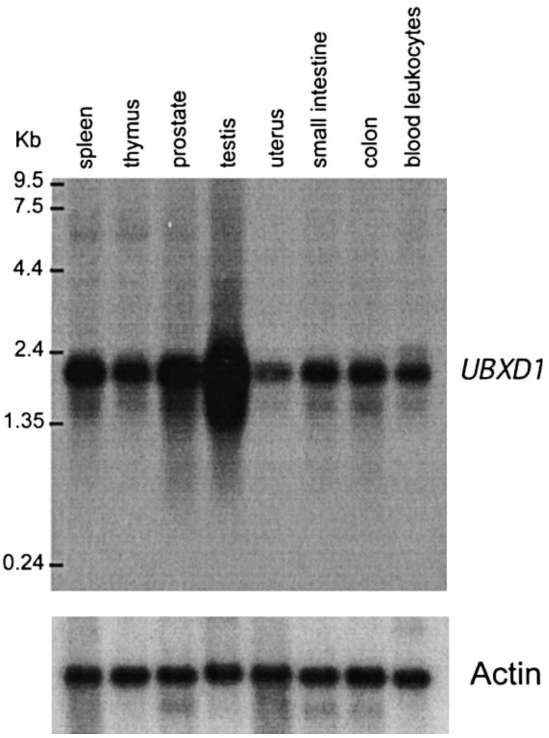


Fig. 2. Northern blot analysis of UBXD1 expression pattern in adult human tissues. Three alternative transcripts, sizes 1.6, 1.8 and 5.5 kb, can be observed in all tissues and enhanced expression is evident in testis. UBXD1 and β-actin transcripts are indicated.

be expected from more closely related species (Fig. 1A). Significant similarity hits with other species such as *Drosophila melanogaster* (CG5469, Acc. No. AE003798, 53% similar), *Arabidopsis thaliana* (T8011.18, Acc. No. AC006069, 41% similar) and *Caenorhabditis elegans* (H06H12.6, Acc. No. AF099920, 55% similar) were also obtained, indicating conservation of this gene during evolution.

The UBX domain found in UBXD proteins is constituted by scattered invariant conserved amino acids separated by variable sequence (consensus shown in Fig. 1C). The UBX domain is localized on the C-terminal end of 80% of the proteins that contain it. Structural analysis predicts two α helices but no apparent secondary structure conservation is observed between different UBX-containing proteins. A rough classification of the different UBX domains based on amino acid similarity shows that UBXD1 and the closest matching proteins fall within a separate group (Fig. 1C). UBX domains were compared with Protpars (PHYLIP package) [7]; 100 replicates were generated by bootstrapping with Seqboot; the consensus tree was generated with Consense (PHYLIP) and printed using TreeView [8].

UBX is described in Pfam as a ‘domain present in ubiquitin regulatory proteins’, illustrated by Y33K, a human putative ORF with homology to ubiquitin conjugating enzymes [9]. In Prosite, the equivalent UX domain (PROSITE entry PS50033, www.isrec.isb-sib.ch) is defined as a ‘UBA-associated domain, present in possibly distant ubiquitin homologs’. Since the UBX domain is not present in ubiquitination proteins themselves and none of the UBX-containing proteins has been directly implicated in the ubiquitination pathway [10], these definitions are misleading. The function of other UBX-containing proteins can give insights into possible roles of the UBX domain. FAF-1 (FAS associated factor 1) has been shown to potentiate Fas-induced apoptosis [10]. Functional studies with mutant FAF-1 indicate that the UBX-containing C-terminal region acts to repress the FAS-responsive apoptosis inducing function of FAF-1 [12]. Expression pattern results for UBXD1 could also help to predict possible roles of this

Table 1
Exon–intron structure of *UBXD1* gene

3′ Splice acceptor	Exon	Size (bp)	5′ Splice donor	Intron	Size (bp)	Sequence identity (mouse/human) (%)
5′UTR TAATTTTCTTCC	1	183	AGAGTCCGTGGG**gt**gcgtgaggaa	1	?	78
agaggagtat**ag**GGCTTGAACTGA	1a	69	TTGACACAGCCC**gt**aagttcatgg	2a	1128	–
gtctgttccc**ag**GGAAAAGGCCCA	2	164	TCCGAAACCAGG**gt**gagatatggc	2	407	81
gctgccttcc**ag**TGAGAAAGGAAC	3	65	GGGACCAACGTG**gt**aagaacagcc	3	965	75
tcaccgcccc**ag**GTATCTGAGCCC	4	129	GCCATTCTCTTG**gt**gagtggcacc	4	3948	82
ccctcctccc**ag**CACTTCTCCACC	5	98	CACCATTGCCAA**gt**gagcgtgtgg	5	692	85
cacctacccc**ag**GTACCTGGACAA	6	76	AAGGTGTTTCAG**gt**gggcgtgcct	6	629	85
ctccaccccc**ag**GAGCGCATTAAC	7	85	CCCAGGATCAGG**gt**aagtggacag	7	116	78
gccacgctgt**ag**AGGACCCCGAGG	8	220	GCAGAGGCTCAG**gt**gggcctgacg	8	86	77
ccgcgtgtgc**ag**GTCCGAGGCGGT	9	131	GCCTCCTGCAGG**gt**gggcaccaac	9	85	82
cccactcccc**ag**GCACTTTCTACG	10	149	GAGTGCGGGCTG**gt**gagtgtcggc	10	425	77
cccctaccgc**ag**GTGCCCTCTGCC	11	612	ACAGGTTGTTTGaaaaaaaaaaaa			71

Detail of the exon–intron structure of UBXD1 gene; the entire transcript consists of 11 exons plus an alternatively spliced exon 1a, which contains an in frame stop codon that results in a shorter protein at the N-terminal end. All exon–intron boundaries are in accordance with the rule that introns begin with dinucleotide GT and end in AG (in bold). The length of intron 1 remains unknown because predictions were made using sequence from unordered genomic fragments available from the public databases. The percentage of identity between the human and mouse *UBXD1* genes is indicated for each exon except for the alternatively spliced exon 1a which has not been identified in mouse yet. For exon 1 the percentage of identity is referred to the coding part only.

protein (Fig. 2). Reproduction-8 (Rep-8) is a mammalian UBX domain-containing gene proposed to have a role in reproduction and with strong expression in testis [11]. The highly renewing nature of this tissue would be in accordance with enhanced expression of genes involved in apoptosis, cell cycling or targeting of proteins for degradation. All three functions would be in accordance with UBXD1 being highly expressed in testis and less in other tissues with lower renewal rates. Experiments will be needed to test all these hypotheses.

We thank A. Puig and D. Otero, the HGMP Resource Center in Hinxton, UK and the RZPD in Berlin, Germany. This work has been supported by EU Biomed Project No. BMH4-CT97-2284 to X.E. and by CICYT-IN95-0347. M.E. was funded by CIDYT contract FPI-070-97 and L.S. by CIRIT-RED contract 1998-64.

References

[1] M.D. Adams, J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno et al., Science 252 (1991) 1651–1656.

[2] G. Lennon, C. Auffray, M. Polymeropoulos, M.B. Soares, Genomics 33 (1996) 151–152.

[3] G.D. Schuler, M.S. Boguski, E.A. Stewart, L.D. Stein, G. Gyapay, K. Rice, R.E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek, S. Bentolila, B.B. Birre, A. Butler, A.B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P.J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, T.J. Hudson et al., Science 274 (1996) 540–546.

[4] G.D. Schuler, J. Mol. Med. 75 (1997) 694.

[5] P. Deloukas, G.D. Schuler, G. Gyapay, E.M. Beasley, C. Soderlund, P. Rodriguez-Tome, L. Hui, T.C. Matise, K.B. McKusick, J.S. Beckmann, S. Bentolila, M. Bihoreau, B.B. Birren, J. Browne, A. Butler, A.B. Castle, N. Chiannilkulchai, C. Clee, P.J. Day, A. Dehejia, T. Dibling, N. Drouot, S. Duprat, C. Fizames, D.R. Bentley et al., Science 282 (1998) 744–746.

[6] S.F. Altschul, T.L. Maden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Nucleic Acids Res. 25 (1997) 3389–3402.

[7] J. Felsenstein, Cladistics 5 (1989) 164–166.

[8] R. Page, Comp. Appl. Biosc. 12 (1996) 357–358.

[9] K. Hofmann, P. Bucher, Trends Biochem. Sci. 21 (5) (1996) 172–173.

[10] K. Chu, X. Niu, L.T. Williams, Proc. Natl. Acad. Sci. USA 92 (1995) 11894–11898.

[11] T. Fröhlich, W. Risau, I. Flamme, J. Cell Sci. 111 (1998) 2353–2363.

[12] Y. Yamabe, A. Yokoi, O. Imamura, M. Matsui, A. Matsunaga, M. Taketo, M. Sugawara, Y. Furuichi, Gene 227 (1) (1999) 39–47.

VI. Identificació i caracterització del gen FSD1 al cromosoma 19

L'objecte de la publicació següent és un nou gen humà, FSD1, d'expressió específica al sistema nerviós central. La seva identificació és resultat de l'anàlisi del transcripcional de la regió p13.3-p12 del cromosoma 19. Es va demostrar l'existència de gens ortòlegs murins i bovins significativament conservats. Una potencial seqüència paràloga de FSD1 va ser localitzada al cromosoma 9 humà, recolzant la relació evolutiva amb el cromosoma 19 descrita a la literatura. Els experiments de transferència de Northern indiquen que FSD1 pateix fenòmens de transcripció alternativa i l'anàlisi de la seva seqüència genòmica permet distingir que es tracta d'un gen format per 13 exons.

ELSEVIER

Biochimica et Biophysica Acta 1518 (2001) 200–203

BIOCHIMICA ET BIOPHYSICA ACTA
BBA

www.bba-direct.com

Short sequence-paper

Characterization of human *FSD1*, a novel brain specific gene on chromosome 19 with paralogy to 9q31

Laura Carim-Todd, Mònica Escarceller, Xavier Estivill, Lauro Sumoy *

Medical and Molecular Genetics Center, Institut de Recerca Oncològica, Departament de Genètica Molecular, Hospital Duran i Reynals, Av. Gran Via s/n, km 2.7, L'Hospitalet de Llobregat, 08907 Barcelona, Spain

Received 22 November 2000; received in revised form 17 January 2001; accepted 18 January 2001

Abstract

We have characterized a novel human gene, *FSD1*, on chromosome 19. FSD1 has a BBC, FN3 and SPRY domain, it is distantly related to the midline 1 gene and is expressed only in the brain. We have established its exon–intron structure and we have identified the corresponding orthologous genes in other species. In addition, the identification of *FSD1* has led us to identify a homologous counterpart sequence on chromosome 9. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: FSD1 domain; BBC domain; FN3 domain; SPRY domain; FSD1L; EUROIMAGE; cDNA sequencing; Paralogy

The EUROIMAGE Consortium was established in 1997 with the aim of completing the sequence and identification of unique cDNA clones represented in dbEST, toward the final goal of characterizing all the transcripts in the human genome [1–5]. Working within this consortium we have studied in silico over 400 EST clusters (EST CAP assembly program, http://www.tigem.it; Sequencher, GeneCodes Corporation). This effort led to the identification of a putative open reading frame (ORF) in Unigene cluster Hs.28144. The IMAGE clones selected for further sequencing (custom synthesized primers, LifeTech, Perkin-Elmer BigDye reagents on an ABI-377) and analysis were: 179664 (EST GenBank accession number H51132), 192295 (EST GenBank accession number H39024) and 180766 (EST GenBank accession number R87684). The full coding cDNA was obtained using the rapid amplification of cDNA ends (RACE) approach on fetal brain SMART cDNA (Clontech) with the following primers: G1 (5′-TCTGGTGGGATGTGGACGCA-3′), G2A (5′-TGGTCAATCTTGCTGTCCTCATCC-3′) and G3A (5′-CTCAAGTTCCTGCCTGTGCCC-3′). The extended transcript was 1729 bp and contained a 5′ methionine at position 45 leading to an ORF of 496 amino acids with a calculated mass of 55.8 kDa and an estimated p*I* of 6.54.

We analyzed the resulting protein sequence with pattern and domain identification software which revealed the presence of three domains (http://smart.embl-heidelberg.de/smart): BBC (coiled-coil region found downstream to some B-box domains) spanning residues 4–130, FN3 (fibronectin type 3 internal repeat present both in intracellular and extracellular proteins) from residue 165 to 258 and SPRY between residue 353 and 474 (a domain of unknown function present in SP1a and in the ryanodine receptor) (Fig. 1A). The gene was designated *FSD1*, fibronectin type 3 and SPRY (sp1a, ryanodine) domain containing (with coiled-coil motif) 1, following the Human Gene Nomenclature Committee rules (http://www.gene.ucl.ac.uk/nomenclature/).

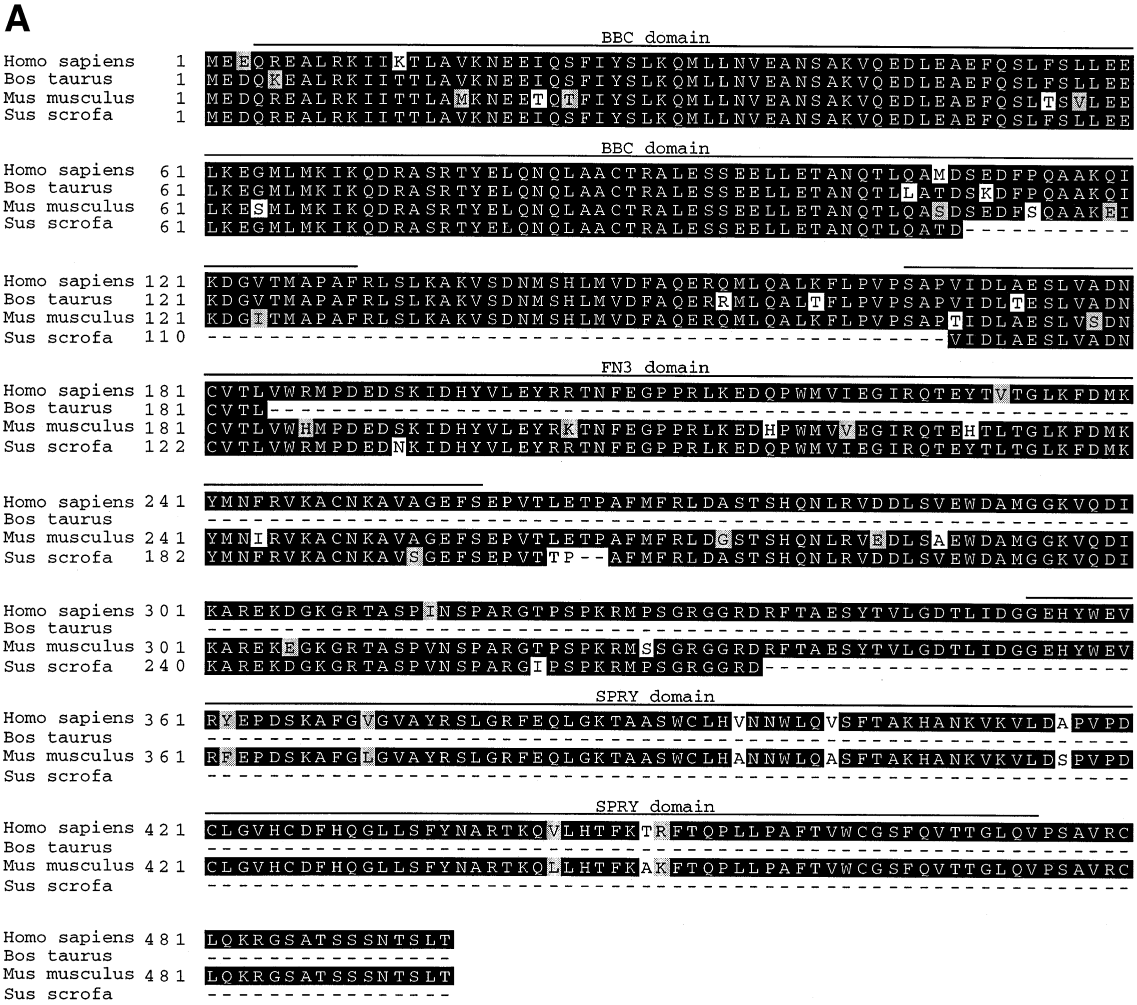
Proteins with similar domain content include midline development proteins such as MID1, a gene causing Opitz syndrome when mutated and that has been shown to associate to microtubules throughout the cell cycle [6,7]. However, FSD1 lacks the N-terminal ring finger and B-box domains that are present in the midline proteins and that characterize the B-box ring finger family. Nevertheless, an uncharacterized protein from *Drosophila* (GenBank accession number AAF52977) has been identified showing the same domain content as FSD1. This fact together with the existence of *FSD1* orthologous genes in other species indicates that these could constitute a new family of proteins with the above mentioned motifs: BBC domain, FN3 domain and C-terminal SPRY domain.

BLAST searches (http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/) [8] with *FSD1* against the non-redundant data-

* Corresponding author. Fax: +34-93-260-7776; E-mail: lsumoy@iro.es

0167-4781/01/$ – see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0167-4781(01)00178-6

base gave a match to finished human genomic sequence from chromosome 19, clone CTB-144D21 (GenBank accession number AC008616). Using the information available from this sequence we were able to establish the exon–intron structure of *FSD1* as shown in Table 1. We analyzed the 5′ upstream genomic sequence with the NIX package (http://menu.hgmp.mrc.ac.uk/menu-bin/Nix/Nix.pl) in order to predict further 5′ exons. No *FSD1* exons were predicted with significant probability upstream from the region corresponding to the longest cDNA end. We did not detect any in frame stop codons upstream of the first methionine, which meant we could not unequivocally establish whether we had a complete ORF. We took a cross species comparison approach to address this issue. BLAST searches against dbEST with *FSD1* identified the corresponding orthologous gene in *Mus musculus* (GenBank accession numbers AU035250 and AU035795). Using mouse unfinished High Throughput Genomic Sequence (GenBank accession number AC073737) we predicted the complete ORF for the mouse *Fsd1* gene. It results in a 496 amino acid protein with an expected mass of 55.5 kDa and a theoretical p*I* of 6.22. The degree of conservation between the human and murine genes is very high: 86% identity and similarity at the DNA level and 92% identity and 95% similarity at the protein level (Fig. 1A). The exon–intron structure is also conserved (Table 1). We also identified the orthologous genes in *Bos taurus* (EST GenBank accession numbers AW653084 and AW445529) and *Sus scrofa* (EST GenBank accession numbers AW359643, AW785163 and AW436727). The partial sequences of these genes show a remarkable degree of homology, especially at the amino acid level (Fig. 1A), an indication of the conservation of *FSD1* during vertebrate evolution. The position of the 5′ initiation codon is shared by the four orthologous genes, and significant divergence in amino acid and DNA sequence is detected upstream from this position. This fact and the presence of upstream stop codons in the mouse and pig transcripts



```
A
                                         BBC domain
Homo sapiens    1  MEEQREALRKIIKTLAVKNEEIQSFIYSLKQMLLNVEANSAKVQEDLEAEFQSLFSLLEE
Bos taurus      1  MEDQKEALRKIITTLAVKNEEIQSFIYSLKQMLLNVEANSAKVQEDLEAEFQSLFSLLEE
Mus musculus    1  MEDQREALRKIITTLAMKNEETQTFIYSLKQMLLNVEANSAKVQEDLEAEFQSLTSVLEE
Sus scrofa      1  MEDQREALRKIITTLAVKNEEIQSFIYSLKQMLLNVEANSAKVQEDLEAEFQSLFSLLEE

                                         BBC domain
Homo sapiens   61  LKEGMLMKIKQDRASRTYELQNQLAACTRALESSEELLETANQTLQAMDSEDFPQAAKQI
Bos taurus     61  LKEGMLMKIKQDRASRTYELQNQLAACTRALESSEELLETANQTLLATDSKDFPQAAKQI
Mus musculus   61  LKESMLMKIKQDRASRTYELQNQLAACTRALESSEELLETANQTLQASDSEDFSQAAKEI
Sus scrofa     61  LKEGMLMKIKQDRASRTYELQNQLAACTRALESSEELLETANQTLQATD-----------

Homo sapiens  121  KDGVTMAPAFRLSLKAKVSDNMSHLMVDFAQERQMLQALKFLPVPSAPVIDLAESLVADN
Bos taurus    121  KDGVTMAPAFRLSLKAKVSDNMSHLMVDFAQERRMLQALTFLPVPSAPVIDLTESLVADN
Mus musculus  121  KDGITMAPAFRLSLKAKVSDNMSHLMVDFAQERQMLQALKFLPVPSAPTIDLAESLVSDN
Sus scrofa    110  ------------------------------------------------VIDLAESLVADN

                                         FN3 domain
Homo sapiens  181  CVTLVWRMPDEDSKIDHYVLEYRRTNFEGPPRLKEDQPWMVIEGIRQTEYTVTGLKFDMK
Bos taurus    181  CVTL--------------------------------------------------------
Mus musculus  181  CVTLVWHMPDEDSKIDHYVLEYRKTNFEGPPRLKEDHPWMVVEGIRQTEHTLTGLKFDMK
Sus scrofa    122  CVTLVWRMPDEDNKIDHYVLEYRRTNFEGPPRLKEDQPWMVIEGIRQTEYTLTGLKFDMK

Homo sapiens  241  YMNFRVKACNKAVAGEFSEPVTLETPAFMFRLDASTSHQNLRVDDLSVEWDAMGGKVQDI
Bos taurus         ------------------------------------------------------------
Mus musculus  241  YMNIRVKACNKAVAGEFSEPVTLETPAFMFRLDGSTSHQNLRVEDLSAEWDAMGGKVQDI
Sus scrofa    182  YMNFRVKACNKAVSGEFSEPVTTP--AFMFRLDASTSHQNLRVDDLSVEWDAMGGKVQDI

Homo sapiens  301  KAREKDGKGRTASPINSPARGTPSPKRMPSGRGGRDRFTAESYTVLGDTLIDGGEHYWEV
Bos taurus         ------------------------------------------------------------
Mus musculus  301  KAREKEGKGRTASPVNSPARGTPSPKRMSSGRGGRDRFTAESYTVLGDTLIDGGEHYWEV
Sus scrofa    240  KAREKDGKGRTASPVNSPARGIPSPKRMPSGRGGRD------------------------

                                         SPRY domain
Homo sapiens  361  RYEPDSKAFGVGVAYRSLGRFEQLGKTAASWCLHVNNWLQVSFTAKHANKVKVLDAPVPD
Bos taurus         ------------------------------------------------------------
Mus musculus  361  RFEPDSKAFGLGVAYRSLGRFEQLGKTAASWCLHANNWLQASFTAKHANKVKVLDSPVPD
Sus scrofa         ------------------------------------------------------------

                                         SPRY domain
Homo sapiens  421  CLGVHCDFHQGLLSFYNARTKQVLHTFKTRFTQPLLPAFTVWCGSFQVTTGLQVPSAVRC
Bos taurus         ------------------------------------------------------------
Mus musculus  421  CLGVHCDFHQGLLSFYNARTKQLLHTFKAKFTQPLLPAFTVWCGSFQVTTGLQVPSAVRC
Sus scrofa         ------------------------------------------------------------

Homo sapiens  481  LQKRGSATSSSNTSLT
Bos taurus         ----------------
Mus musculus  481  LQKRGSATSSSNTSLT
Sus scrofa         ----------------
```

Fig. 1. A: ClustalW multiple sequence alignment of human FSD1 and the mouse predicted orthologous protein, along with the pig and bovine partial sequence. Identical residues are printed in reverse type and similar residues are shaded. The position of the predicted domains is also indicated. B: Alignment between human chromosome 19 FSD1 (GenBank accession number AF316829), chromosome 9 FSD1L (GenBank accession number AF316830) predicted from ESTs in Unigene cluster Hs.55846 and the paralogous sequence on chromosome 9q31.1–31.3 obtained from finished genomic sequence. An asterisk indicates a stop codon.

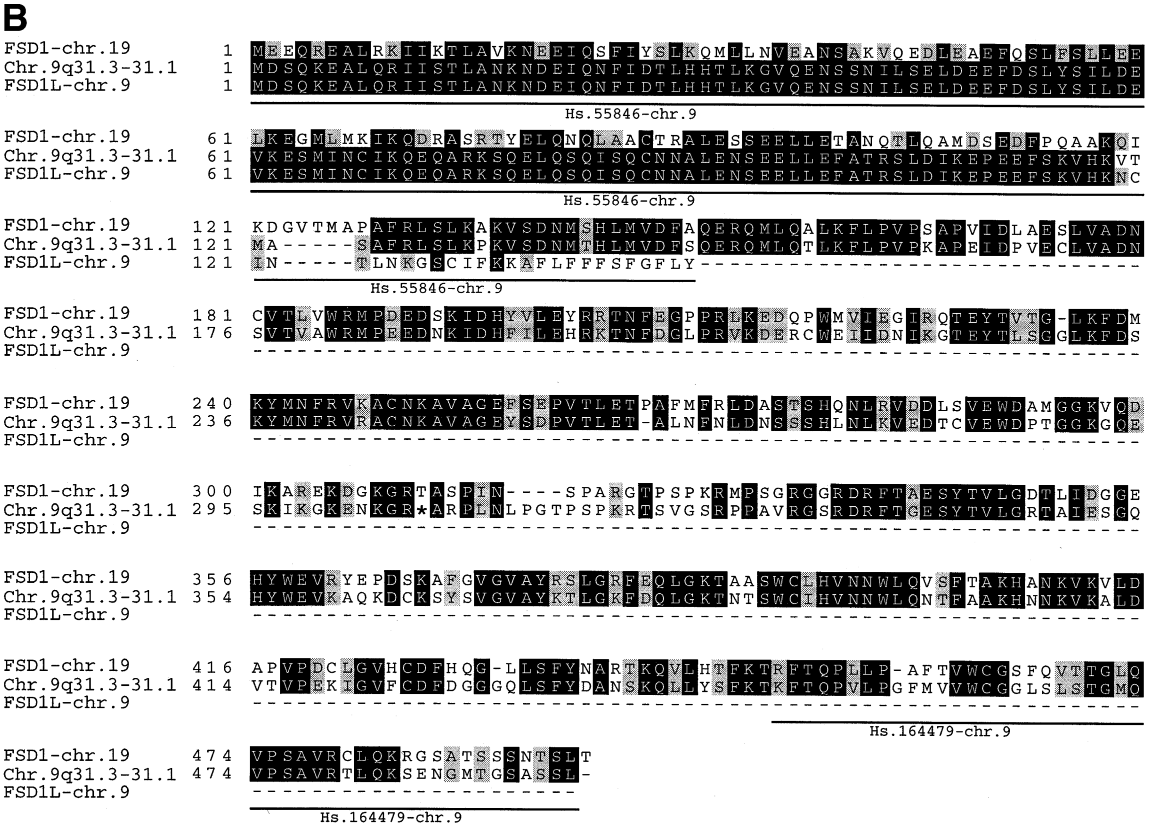

B
FSD1-chr.19 1 MEEQREALRKIIKTLAVKNEEIQSFIYSLKQMLLNVEANSAKVQEDLEAEFQSLFSLLEE
Chr.9q31.3-31.1 1 MDSQKEALQRIISTLANKNDEIQNFIDTLHHTLKGVQENSSNILSELDEEFDSLYSILDE
FSD1L-chr.9 1 MDSQKEALQRIISTLANKNDEIQNFIDTLHHTLKGVQENSSNILSELDEEFDSLYSILDE
Hs.55846-chr.9
FSD1-chr.19 61 LKEGMLMKIKQDRASRTYELQNQLAACTRALESSEELLETANQTLQAMDSEDFPQAAKQI
Chr.9q31.3-31.1 61 VKESMINCIKQEQARKSQELQSQISQCNNALENSEELLEFATRSLDIKEPEEFSKVHKVT
FSD1L-chr.9 61 VKESMINCIKQEQARKSQELQSQISQCNNALENSEELLEFATRSLDIKEPEEFSKVHKNC
Hs.55846-chr.9
FSD1-chr.19 121 KDGVTMAPAFRLSLKAKVSDNMSHLMVDFAQERQMLQALKFLPVPSAPVIDLAESLVADN
Chr.9q31.3-31.1 121 MA------SAFRLSLKPKVSDNMTHLMVDFSQERQMLQTLKFLPVPKAPEIDPVECLVADN
FSD1L-chr.9 121 IN-----TLNKGSCIFKKAFLFFFSFGFLY------------------------------
Hs.55846-chr.9
FSD1-chr.19 181 CVTLVWRMPDEDSKIDHYVLEYRRTNFEGPPRLKEDQPWMVIEGIRQTEYTVTG-LKFDM
Chr.9q31.3-31.1 176 SVTVAWRMPEEDNKIDHFILEHRKTNFDGLPRVKDERCWEIIDNIKGTEYTLSGGLKFDS
FSD1L-chr.9 --
FSD1-chr.19 240 KYMNFRVKACNKAVAGEFSEPVTLETPAFMFRLDASTSHQNLRVDDLSVEWDAMGGKVQD
Chr.9q31.3-31.1 236 KYMNFRVRACNKAVAGEYSDPVTLET-ALNFNLDNSSSHLNLKVEDTCVEWDPTGGKGQE
FSD1L-chr.9 --
FSD1-chr.19 300 IKAREKDGKGRTASPIN----SPARGTPSPKRMPSGRGGRDRFTAESYTVLGDTLIDGGE
Chr.9q31.3-31.1 295 SKIKGKENKGR*ARPLNLPGTPSPKRTSVGSRPPAVRGSRDRFTGESYTVLGRTAIESGQ
FSD1L-chr.9 --
FSD1-chr.19 356 HYWEVRYEPDSKAFGVGVAYRSLGRFEQLGKTAASWCLHVNNWLQVSFTAKHANKVKVLD
Chr.9q31.3-31.1 354 HYWEVKAQKDCKSYSVGVAYKTLGKFDQLGKTNTSWCIHVNNWLQNTFAAKHNNKVKALD
FSD1L-chr.9 --
FSD1-chr.19 416 APVPDCLGVHCDFHQG-LLSFYNARTKQVLHTFKTRFTQPLLP-AFTVWCGSFQVTTGLQ
Chr.9q31.3-31.1 414 VTVPEKIGVFCDFDGGGQLSFYDANSKQLLYSFKTKFTQPVLPGFMVVWCGGLSLSTGMQ
FSD1L-chr.9 --
Hs.164479-chr.9
FSD1-chr.19 474 VPSAVRCLQKRGSATSSSNTSLT
Chr.9q31.3-31.1 474 VPSAVRTLQKSENGMTGSASSL-
FSD1L-chr.9 ----------------------
Hs.164479-chr.9

Fig. 1 (*continued*).

would support that the identified conserved methionine is indeed the initiation codon in the human, mouse, pig and bovine *FSD1* genes.

Further database searches detected homology between *FSD1* and finished genomic sequence from chromosome 9q31.1–31.3 (GenBank accession numbers AL158070 and AL161627). A more detailed analysis of this hit revealed the existence of a 145 amino acid ORF with 47% identity and 72% similarity to the 5′ region of *FSD1*. It corresponded to an uncharacterized Unigene cluster, Hs.55846, and contained four ESTs indicating that this gene is expressed (Fig. 1B). It only presents the BBC domain and has been named *FSD1L* (*FSD1-like*), in agreement with the Human Gene Nomenclature Committee. The homology between *FSD1* and this region of chromosome 9 extends further with an identity ranging from 55 to 58% (70–80% similarity) and localizes on chromosome 9q31 immediately upstream from the fukutin gene (Gen-

Table 1
Exon–intron structure of *FSD1*

3′ Splice acceptor	Exon	Size (bp)	5′ Splice donor	Intron	Size (bp)	Sequence identity (%) (mouse/human)
5′-CGCGGGGCCCGC	1	> 58	GAAGAACAGAGG**gt**aggacggggt	1	1184	–
ctggtggtgc**ag**GAGGCCCTGAGG	2	96	CTGAACGTGGAG**gt**gaaggcggtg	2	156	80
gttccgaccc**ag**GCGAACTCGGCG	3	132	TACGAGCTGCAG**gt**gagggctgag	3	1552	84
tggtatccac**ag**AACCAGCTGGCT	4	102	GACTTTCCTCAG**gt**gggtgcctct	4	2289	89
tcctctctct**ag**GCTGCCAAGCAA	5	23	AATCAAAGATGG**gt**aagacactgg	5	179	95
tccttcctgc**ag**AGTGACCATGGC	6	122	AGTTCCTGCCTG**gt**gagaggggca	6	1246	87
cgtcttggtc**ag**TGCCCAGCGCAC	7	210	ACACCGTGACAG**gt**aagggcagtg	7	5130	88
cttgcctacc**ag**GTCTCAAGTTTG	8	99	TGGAGACACCAG**gt**gactggattc	8	1065	86
gcccggcccc**ag**CGTTCATGTTCC	9	160	CTCCCCAGCCAG**gt**agcctgcccc	9	366	86
tgcccaccac**ag**AGGTACTCCATC	10	80	ACACAGTTCTGG**gt**aaggaagggg	10	4034	90
cctgcgccac**ag**GGGACACGCTGA	11	252	ACTTCCACCAAG**gt**gaccccaagc	11	110	88
ccccgacccc**ag**GCCTCCTGTCCT	12	89	CCTGCTTTCACG**gt**gagctgccct	12	96	84
gctgtccctc**ag**GTATGGTGTGGC	13	304	TCAGACACTGGCaaaaaaaaaaaa	–	–	90

Detail of the exon–intron structure of *FSD1* gene; the entire transcript consists of 13 exons. All exon–intron boundaries are in accordance with the rule that introns begin with dinucleotide GT and end in AG (in bold). The percentage of identity between the human *FSD1* and the mouse predicted orthologous gene is indicated for each exon. For exon 13 the percentage of identity refers to the coding part only.

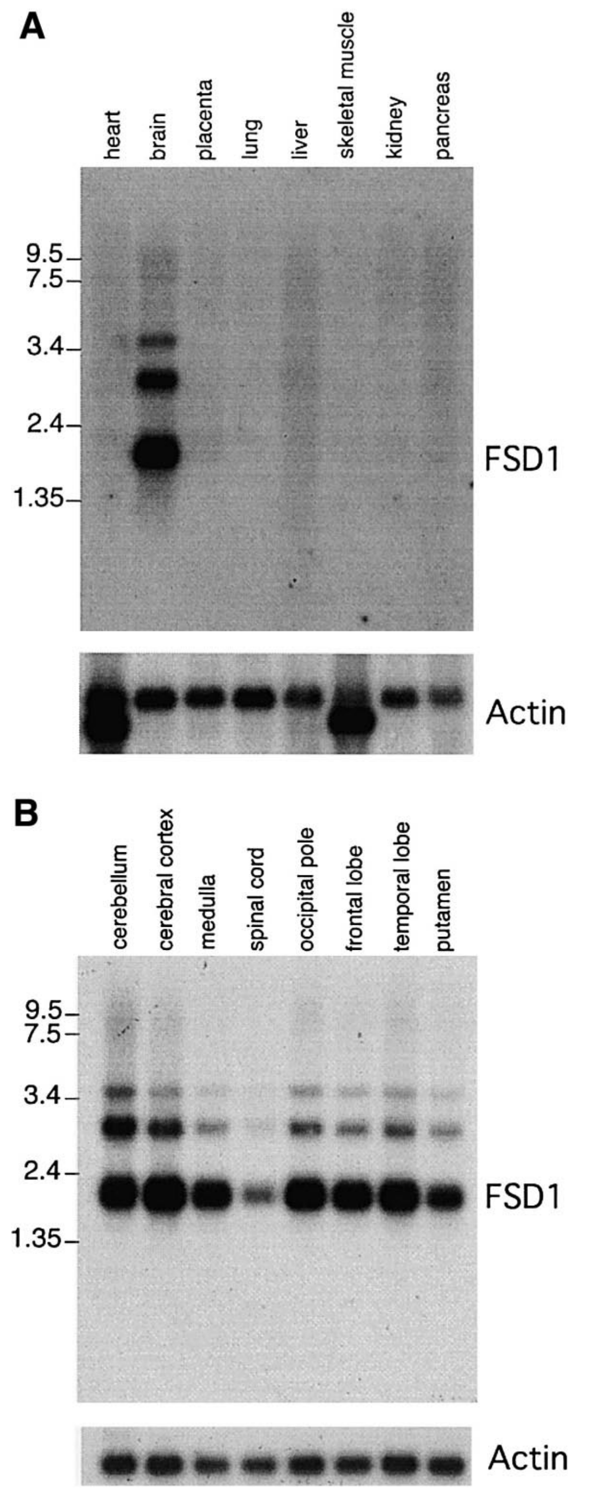


Fig. 2. Northern blot analysis of *FSD1*. A: Human multiple tissue blot (Clontech) was hybridized with a 0.8 kb *Pst*I fragment from clone 179664. B: A 0.6 kb *Eco*RI probe derived from one of the RACE clones was hybridized to a human brain tissue Northern blot (Clontech). Three alternative transcripts are evident in brain tissue, corresponding to sizes 1.8 kb, 3 kb and 3.5 kb. Significantly lower expression is detected in spinal cord, whilst it remains high in the rest of cerebral tissues analyzed. *FSD1* and β-actin transcripts are indicated.

Bank accession number AB038490) [9]. A stop codon was identified in the genomic chromosome 9 sequence aligned with *FSD1* (clone RP11-287A8, GenBank accession number AL161627), indicating that this could correspond to a pseudogene (Fig. 1B). However, a cluster, Hs.164479, containing three ESTs matches perfectly downstream from this position but no methionine that could act as an initiation codon has been identified.

The characterization of *FSD1*'s expression pattern can give insights into the function of this protein. Northern blot studies show specific expression of three mRNA species in the human central nervous system (reduced in spinal cord), indicating a specific role of *FSD1* in neural tissues (Fig. 2A,B). The most abundant mRNA is about 1.8–2.0 kb long, in accordance with the size obtained by 5′ RACE extension. The two alternative 3 and 3.5 kb transcripts most likely correspond to splice variants without representation in dbEST in the region flanking the overlap with the 1.8 kb mRNA, and could not be extended by RACE. Functional experiments will be needed to determine the specific role of *FSD1* in the human brain and the processes in which it is implicated.

We are grateful to A. Puig and D. Otero, the HGMP Resource Center in Hinxton, UK, and the RZPD in Berlin, Germany. Supported by EU Biomed Project No. BMH4-CT97-2284 to X.E. and by CICYT-IN95-0347. M.E. is funded by the Spanish Ministry of Education (CIDYT contract FPI-070-97) and L.S. by the Catalan autonomous government (CIRIT-RED contract 1998-64).

References

[1] M.D. Adams, J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno et al., Science 252 (1991) 1651–1656.

[2] G. Lennon, C. Auffray, M. Polymeropoulos, M.B. Soares, Genomics 33 (1996) 151–152.

[3] G.D. Schuler, M.S. Boguski, E.A. Stewart, L.D. Stein, G. Gyapay, K. Rice, R.E. White, P. Rodriguez-Tome, A. Aggarwal, E. Bajorek, S. Bentolila, B.B. Birre, A. Butler, A.B. Castle, N. Chiannilkulchai, A. Chu, C. Clee, S. Cowles, P.J. Day, T. Dibling, N. Drouot, I. Dunham, S. Duprat, C. East, T.J. Hudson et al., Science 274 (1996) 540–546.

[4] G.D. Schuler, J. Mol. Med. 75 (1997) 694.

[5] P. Deloukas, G.D. Schuler, G. Gyapay, E.M. Beasley, C. Soderlund, P. Rodriguez-Tome, L. Hui, T.C. Matise, K.B. McKusick, J.S. Beckmann, S. Bentolila, M. Bihoreau, B.B. Birren, J. Browne, A. Butler, A.B. Castle, N. Chiannilkulchai, C. Clee, P.J. Day, A. Dehejia, T. Dibling, N. Drouot, S. Duprat, C. Fizames, D.R. Bentley et al., Science 282 (1998) 744–746.

[6] S. Schweiger, J. Foerster, T. Lehmann, V. Suckow, Y.A. Muller, G. Walter, T. Davies, H. Porter, H. van Bokhoven, P.W. Lunt, P. Traub, H.H. Ropers, Proc. Natl. Acad. Sci. USA 96 (6) (1999) 2794–2799.

[7] S. Cainarca, S. Messali, A. Ballabio, G. Meroni, Hum. Mol. Genet. 8 (8) (1999) 1387–1396.

[8] S.F. Altschul, T.L. Maden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Nucleic Acids Res. 25 (1997) 3389–3402.

[9] Y. Saito, M. Mizuguchi, A. Oka, S. Takashima, Ann. Neurol. 47 (6) (2000) 756–764.