Estudi bioinformàtic de la funcionalitat i conservació de l'*splicing* alternatiu

Jordi Morata Chirivella

ESTUDI BIOINFORMÀTIC DE LA FUNCIONALITAT I CONSERVACIÓ DE L'SPLICING ALTERNATIU

DEPARTAMENT DE BIOQUÍMICA I BIOLOGIA MOLECULAR
UNIVERSITAT DE BARCELONA

PROGRAMA DE DOCTORAT D'AQÜICULTURA
BIENNI 2004-2006

Memòria de la tesi realitzada per en Jordi Morata Chirivella a l'Institut de Biologia Molecular de Barcelona (IBMB-CSIC), sota la direcció del Dr. Francisco Javier de la Cruz Montserrat i la tutoria del Dr. Josep Lluís Gelpí Buchaca, per optar al títol de doctor per la Universitat de Barcelona

Jordi Morata Xavier de la Cruz Josep Lluís Gelpí

Barcelona, 2012

"*It is remarkable how the remarkable has become unremarked*"
Richard Fortey

"Possunt quia posse videntur"
(Poden els que creuen que poden)
Virgili

Agraïments

Arribats a l'hora d'agrair, intentaré seguir dos de les màximes que m'ha recordat més d'un cop mon pare: "*Es de bien nacido ser agradecido*" i "*Lo bueno, si breve, dos veces bueno*", aquesta última, però, difícilment la compliré. Les circumstàncies m'obliguen a estendre'm.

Primer de tot, tota la meva gratitud al meu director de tesi, en Xavier de la Cruz, per escollir-me per realitzar aquesta tesi després de rebre molts cops de porta a la cara, però sobretot per tots aquests anys i aquesta manera de fer ciència, que creia ja desapareguda, on la discussió, la crítica, la modèstia, la confiança i l'honestedat són essencials a l'hora de treballar.

Al grup mai no hem sigut molts però sempre ben escollits i molt ben avinguts. Agraït estic: als veterans, al Sergi Lois, el David Piedra i a en David Talavera, per tota l'ajuda en les primeres capbussades en el món de Perl i Linux, el bon humor i els amuntegaments dels *journals club*s; a en Iago Vázquez i les nostres converses acalorades sobre ocells, mosquits dolents i política; al cop de mà de la Gemma Mas de Xaxars i la Meritxell Domeño, dos personatges dos; al duo de físics reciclats, en Santi Béjar i la Montse Barbany, crítics a més no poder, curiosos, sempre de ben humor, pares i massa pacients (sobretot amb els meus conflictes inacabables amb el món Apple); al tió de la Nora i la Sílvia; i a l'última incorporació, a la Casandra Riera, una glopada d'aire fresc des d'una *terra in salo*.

Mitja tesi l'he realitzat a les instal·lacions del MMB, agrair doncs a en Modesto Orozco i a en Francisco Javier Luque per permetre'm l'ús i a en Josep Lluís Gelpí, la Isabel Navarro, en Joaquim Gutiérrez i en Josep Planas per les facilitats administratives. Tota la meva gratitud a la gentada del MMB: a en Jose Alcántara i l'Adam Hospital per l'ajuda tècnica, molt més enllà dels que els hi pertocava, en els meus mil i un problemes informàtics; al bon ambient que vaig haver de deixar enrere, gràcies a gent com el Nacho Faustino, l'Anna Lisa Arcella, l'Agustí Emperador (bou!), l'Antonella, l'Oliver, la Laura Orellana i tots els que em deixo. Sou masses, així que un agraïment massiu i indiscriminat a cadascun de vosaltres. I disculpes pel meu volum de veu massa sovint massa alt. No m'oblido dels companys de patiments immobiliaris, en Guillem Portella, el curiós infinit, i en Carles Fenollosa, l'optimista irremeiable.

Abans de passar-me a l'enemic, havia sigut un *experimental*. Com no vaig tenir l'ocasió de ser agraït en el seu moment, ara aprofito l'avinentesa: A la Núria Montserrat (i al David!), per liar-me sempre de mala manera i seguir creient en mi després de totes les meves desbarrades; a la Bàrbara Castellana, per veure aquell cartellet, per totes les batalles perdudes i per seguir-nos aguantant; a l'hora del safareig de cada divendres amb la Mònica Díaz, el Uoang Castillo i l'Encarni Capilla; a la Concepció Soler i a tots els *immunos*, bojos i genials. I a tots els fisiòlegs que em deixo (Lamia, Pablo, Kelly!, etc etc etc) i que no acabessin farts de mi, moltes gràcies per aquells anys.

Al pis de dalt rondaven els microbiòlegs, on treballava en *Charli* Maluquer de Motes, company d'infinites *tanganes*, de discussions científiques envolades i viatges caòtics, a més a més d'inventor del terme "*moratada*". Agraït a l'Enrique Amenedo, pel rescat personal en un moment més que crític. Agraït al *trio maravillas*: Ayalkibet, Jesús Rodman i Byron Tomas (sin mote), exòtics tots ells a la seva manera, per la camaraderia i tot el que he pogut arribar a conèixer en les nostres converses interminables. També hi ha massa *micros* però, així que a tots vosaltres (Lejla, Anna, Sarah, Tariq, etc etc) moltes gràcies per ser-hi… i en especial a una altra microbiòloga que deixaré pel final.

Estic obligat a recordar dos companys de carrera, en Gorka Navarrete, sempre lluny però apareixent sempre en moments clau per donar-me un nou impuls, i a en Tadeusz, doctor, músic i figura renaixentista, per aquell dia al Soho, *respect for all*! I d'aquí enllaço amb els meus dos filòsofs no *heideggerians* preferits, en Sisco, baixista i polemista, i Érika, sempre allà, sempre removent-me el magí. Gràcies també al Feliu (x2) i l'Eva i la resta de la tribu musical badalonina (Peter, April, Erran, Nigga, etc.) pels milions d'experiències, dalt i baix de l'escenari.

Una breu menció per als meus dos professos de biologia de l'Institut Fort Pius, en Luis Ugedo i la Teresa Cuartielles, per inocular-me la passió per la biologia i després aconsellar-me que no se m'ocorregués per res del món fer biologia. Contradiccions típiques de biòlegs!

Durant mesos i anys, tothom al Parc creia que era de Valls, tot per una confusió administrativa. No, no sóc de Valls però allà hi tinc grans amics, un vàlvula d'escapament de la ciutat i la tesi. Agraït, *de_tot_cor*, estic a la parella de pixapins Jordi & Meritxell, pel suport inalterable i la confiança cega; a les caminades de cabra sense mapa amb el clan Domènech (Manel, Jordi i Joan); al Ramon i la família perdiu, a l'exiliat aragonès, als habitants del Pla de Santa Maria (Peke, David, Simon) i a totes les respectives. Gràcies també a dos grans amics del *p'tit* barri, l'Edu i l'Àlex, retrobats just a temps.

Per acabar, agrair, parafrasejant Pangloss, a la millor família de tots els mons, on vaig créixer en un ambient on no res no deixava de ser prou important per no ser conegut i après. A l'Òscar, per la perseverança per fer-me atractiva la ciència (i la música i mil coses més), tot i que no vaig acabar fent física; al Virgili i la Glòria i els nebodets, per tota la informació inútil (que no ho és pas), per amansir-me el caràcter i assabentar-se ni fa un any que feia la tesi; i al Rafel i la Bàrbara, per omplir-me el cap de pardalets, cançons, còmics i projectes estranys i fer-me parar boig en general. A ma mare, la Immaculada, i a mon pare, Virgilio, per obligar-me a preguntar menys (fenomen conegut com a "globus sonda") i fer anar més els diccionaris i les enciclopèdies. I a tots plegats per aguantar durant tots aquests anys els meus renecs, estirabots passats de voltes i negacions diverses de la realitat!

I com a epíleg i prefaci, tota l'estima del món per a la meva arbequina preferida, l'Elisenda, per haver superat tantes llunes plenes sense no gaires rascades del licantrop.

Barcelona/Valls/Dublín, 2012

CONTINGUT

Taula general

ABREVIATURES

AI	Isoforma alternativa
AS	Splicing alternatiu
CAGE	*Cap Analysis of Gene Expression*
cDNA	DNA complementari
DcR	Divergència de la seqüència *cis*-reguladora
DNA	Àcid Desoxiribonucleic
DP	Divergència de la seqüència proteica
EST	*Expressed Sequence Tag*
GO	*Gene Ontology*
HITS-CLIP	*High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation*
MI	Isoforma principal
miRNA	micro RNA
mRNA	RNA missatger
NMD	*Nonsense Mediated Decay*
NGS	*High-throughput Next Generation Sequencing*
PDB	*Protein Data Bank*
RAS	Splicing alternatiu regulador
RNA	Àcid Ribonucleic
SNP	*Single nucleotide polymorphism*
snRNP	*Small nuclear ribonucleoprotein*
TD	Diversitat transcripcional
TSS	Lloc d'inici de la transcripció

1. INTRODUCCIÓ

L'splicing és el mecanisme que té lloc durant la maduració del pre-mRNA, en el qual els introns són exclosos i els exons es lliguen consecutivament per formar el mRNA madur. Entenem com a splicing alternatiu la combinació de diferents patrons d'exclusió/inclusió d'exons i introns (Graveley, 2001; Kim et al., 2008), el qual permet la generació de diversos transcrits alternatius que, un cop traduïts, resulten en diverses isoformes proteiques provinents totes d'un mateix gen.

1.1 Descripció i importància de l'splicing alternatiu

Tot i que fa més de trenta anys que es va descriure aquest fenomen (Chow et al., 1977; Early et al., 1980; Gilbert, 1978), no ha estat fins aquesta última dècada en la que s'ha pres consciència de la importància biològica de l'splicing alternatiu. Avui dia ja es dona per fet que no només és un del principals generadors de la diversitat de transcrits i proteïnes (Graveley, 2001; Nilsen and Graveley, 2010), sino que també té un paper regulador de la expressió gènica i, per tant, l'splicing alternatiu pot ser un contribuïdor destacat en la determinació del fenotip de cèl·lules, teixits i organismes (Hartmann and Valcarcel, 2009; Wang et al., 2008).

1.1.1 Freqüència de l'splicing alternatiu

Les estimacions que s'han anat obtenint sobre la freqüència de l'splicing alternatiu han anat augmentant al llarg dels anys, de manera parella a l'avenç en la sensibilitat i profunditat de cobertura de les tècniques de detecció genòmica (Taula 1). Gràcies a l'aplicació de les noves tècniques de seqüenciació massiva de DNA i RNA (*High Thoughput next generation DNA sequencing*) s'han pogut obtenir unes estimacions de presència d'splicing alternatiu superiors al 90% en gens humans, el 86% dels quals tenen una freqüència de la isoforma alternativa superior al 15% (Pan et al., 2008; Wang et al., 2008). No obstant això, encara hi ha cert debat sobre els biaixos i errors a l'hora d'analitzar aquests resultats (Schwartz et al., 2011) i

sobre la funcionalitat de tots aquests transcrits. Per esvair els dubtes biològics que suscita seria necessari un aprofundiment en l'estudi dels transcrits sotmesos a NMD (de Lima Morais and Harrison, 2010), la quantitat d'mRNAs que són traduïts (Tanner et al., 2007) i quina proporció és considerada al·lèlica (Graveley, 2008).

Pel que fa a altres espècies, la proporció de gens amb splicing alternatiu s'ha relacionat amb el grau de complexitat de l'organisme (Keren et al., 2010; Nilsen and Graveley, 2010). Les estimacions són del 40-75% a ratolí (Bingham et al., 2008; Kim et al., 2007), del 30% a porc (Nygard et al., 2010), entre el 23 i el 42% a vaca (Chacko and Ranganathan, 2009b; Kim et al., 2007), d'un 21-43% a pollastre (Chacko and Ranganathan, 2009a; Kim et al., 2007), del 60% a *Drosophila* (Graveley et al., 2011) i sensiblement inferior a *C. elegans* (Kim et al., 2007; Stolc et al., 2004).

Any	AS estimat	Tècnica	Referència
2000	38%	ESTs	(Brett et al., 2000)
2001	42%	ESTs	(Modrek et al., 2001)
2003	74%	Exon junction arrays	(Johnson et al., 2003)
2004	40%	cDNAs	(Imanishi et al., 2004)
2004	45%	ESTs	(Gupta et al., 2004)
2007	73%	Exon array	(Clark et al., 2007)
2008	92-94%	mRNA-seq	(Wang et al., 2008)
2008	95%	mRNA-seq	(Pan et al., 2008)

Taula 1: Freqüència de gens amb splicing alternatiu a humà. Recull d'alguns treballs en els que s'ha caracteritzat de manera massiva (*genome-wide*) el percentatge de gens amb splicing alternatiu a humà.

1.1.2 Tècniques de detecció transcriptòmica

El desenvolupament de les tecnologies NGS ha revolucionat el món de la transcriptòmica, com ja van fer en el seu moment els estudis basats en ESTs i microarrays, i avui dia és possible la seqüenciació massiva i directa de cDNA mitjançant la tècnica de l'RNA-seq. A banda del descobriment de nous transcrits, de nous esdeveniments d'splicing alternatiu específics de teixit i nous gens

(Denoeud et al., 2008; Wang et al., 2008), s'han aconseguit altres fites que afecten la descripció d'aquests transcrits com és l'estudi massiu de les dianes a l'RNA dels factors d'splicing mitjançant la tècnica HITS-CLIP (Xiao and Lee, 2010), una definició més clara del inicis de transcripció mitjançant tècniques derivades de CAGE (Valen et al., 2009) i la identificació de transcrits *sense*, *antisense* i de fusió gènica (Maher et al., 2009; Ozsolak and Milos, 2011).

Les tecnologies NGS han permès superar certs problemes relacionats amb els microarrays mitjançant la reducció del soroll de fons, l'increment de la sensibilitat i precisió, i, al cap i a la fi, una estimació final més acurada dels nivells totals de transcrits (Fu et al., 2009), tot plegat sense la necessitat d'hibridacions i de coneixements previs de les estructures gèniques (Wang et al., 2008; Xiao and Lee, 2010). L'anàlisi de totes aquestes dades, però, suposa tot un repte computacional, sobretot a l'hora d'associar de manera no ambigua les lectures (*reads*) a isoformes (Trapnell et al., 2010).

1.1.3 Desenvolupament i diferenciació tissular

Diferents conjunts de gens poden ser regulats a nivell de transcripció i splicing alternatiu a l'hora de definir els perfils d'expressió gènica específics de teixits i cèl·lules i, per tant, també dels diferents estadis de desenvolupament cel·lular d'un organisme (Blencowe, 2006). Tot i això, és interessant destacar que, en un teixit o programa cel·lular concret, no s'ha trobat cap solapament significatiu entre gens coregulats a nivell de transcripció per una banda i entre gens coregulats a nivell d'splicing per l'altra (Castle et al., 2008; Pan et al., 2004), el que suggereix que l'especificitat tissular s'assoleix mitjançant mecanismes paral·lels sobre conjunts de gens amb expressió específica de teixit.

Fins fa ben poc, la confirmació experimental d'aquestes diferències en els patrons d'splicing entre teixits ha patit de grans dificultats tècniques (Wang et al., 2008) però, gràcies a l'adveniment de les tècniques d'NGS, en els darrers temps s'han publicat multitud de treballs en els quals s'han descrit la gran diversitat de transcrits en teixits humans i murins (Castle et al., 2008; Kwan et al., 2008; Pan et al., 2008; Sultan et al., 2008), entre teixits cancerígens i normals (Venables et al., 2009) i en el desenvolupament de teixits i òrgans concrets, tant de ratolí com humà

(Bland et al., 2010; Johnson et al., 2009; Kalsotra et al., 2008; Tang et al., 2009) , entre d'altres.

A banda del conegut increment en la diversitat de transcrits, aquests estudis han mostrat una clara especificitat tissular dels esdeveniments d'splicing, els quals són específics en el 60% dels casos (Hallegger et al., 2010; Wang et al., 2008). Entre els teixits amb més isoformes específiques trobem el sistema nerviós central, el múscul esquelètic i els testicles (Castle et al., 2008; de la Grange et al., 2010) i, tot i que els transcrits alternatius rarament són únics per només un teixit concret (Xie et al., 2002), podem trobar proporcions extremes entre teixits en els esdeveniments anomenats "*switch-like*" (Wang et al., 2008) relacionats amb la regulació de funcions molt concretes de teixit.

S'ha postulat que l'splicing específic de teixit pot ser una conseqüència de les variacions entre teixits de l'expressió relativa (o en l'activitat) dels factors reguladors d'splicing que trobem majoritàriament de manera ubiqua (hnRNP, SR) (Grosso et al., 2008; Kalsotra et al., 2008). També s'ha descrit factors reguladors d'splicing específics de teixits, com els NOVA i PTB a les neurones (Boutz et al., 2007; Ule et al., 2005) o FOX1 i 2 a neurones i múscul, entre d'altres (Zhang et al., 2008). Els miRNAs també activen programes d'splicing alternatiu específics en la diferenciació neuronal o en el desenvolupament post-natal del cor (Kalsotra et al., 2010; Makeyev et al., 2007) o influeixen sobre els factors reguladors d'splicing (Boutz et al., 2007). Tenint en compte que el 40% dels llocs d'unió de l'miRNA es troba en la regió 3'UTR (Majoros and Ohler, 2007), l'splicing alternatiu en aquesta zona pot reduir els llocs d'unió d'miRNAs, eliminant així la seva capacitat reguladora en cèl·lules proliferants (Sandberg et al., 2008).

1.1.4 Incidència de la variació interindividual en l'splicing alternatiu

La precisió de les noves tècniques de seqüenciació ha permès un pas més en l'aprofundiment del coneixement de les diferències al·lèliques entre individus. Una preocupació legítima és determinar fins a quin punt aquestes variacions intraespecífiques poden emmascarar les estimacions d'splicing alternatiu global. La variació interindividual dels patrons d'splicing alternatiu entre mostres de còrtex de cerebel de sis individus va ser d'entre el 10 i el 30% (Wang et al., 2008),

resultats que concorden amb el ~21% estimat anteriorment degut a polimorfismes associats a splicing que alteren les proporcions relatives de les isoformes alternatives (Nembaware et al., 2004). Altres estudis (Hull et al., 2007; Kwan et al., 2007) també han identificat esdeveniments d'splicing al·lèlics relacionats amb SNPs, però disten encara de ser estudis exhaustius com per a poder proposar una conclusió global (Graveley, 2008). D'aquesta manera, la influència de la variació intraespecífica en l'splicing alternatiu encara queda per determinar.

1.1.5 Implicació de l'splicing alternatiu en malalties

L'splicing alternatiu afecta la majoria dels nostres gens i, per això mateix, no és rar que estigui implicat en multitud de malalties més o menys greus. Les patologies són una alteració en el fenotip dels individus d'una mateixa espècie i això ens pot donar una idea de quin és el grau d'implicació de l'splicing alternatiu en la diferenciació tissular o del desenvolupament. Tot simplificant, les patologies degudes a splicing es poden separar entre aquelles produïdes per la generació d'isoformes aberrants, les que són degudes a l'expressió d'isoformes "correctes" en el moment o lloc inadequat o, finalment, pot ser degut al balanç anormal entre isoformes.

Les isoformes aberrants poden ser produïdes per mutacions en seqüències bàsiques per el desenvolupament normal de l'splicing. El 15% de les mutacions presents a la Human Gene Mutation Database (HGMD) afecten llocs d'splicing, però pot arribar a ser fins el 50% si afegim les mutacions que generen llocs críptics d'splicing i les que afecten seqüències inhibidores/estimuladores de l'splicing a introns i exons (Blencowe, 2006; Matlin et al., 2005). Aquests percentatges farien de l'splicing alternatiu un dels principals responsables de les malalties hereditàries (Lopez-Bigas et al., 2005). Alguns exemples són els de la atrofia muscular espinal, originada per una mutació que produeix l'exclusió d'un exó a *SMN2* (Tazi et al., 2009), o l'activació d'un lloc críptic d'splicing a la *LMNA* que genera la progèria de Hutchinson-Gilford (De Sandre-Giovannoli and Levy, 2006). Les isoformes aberrants també estan àmpliament esteses en els diferents processos cancerígens (Venables, 2004; Venables, 2006), les quals han estat relacionades amb la resistència contra els tractaments anti-càncer (Hartmann and Valcarcel, 2009),

però que per una altra banda també es poden utilitzar com a biomarcadors de diferents càncers (Brinkman, 2004; Yi and Tang, 2011).

És en el càncer on trobem també múltiples exemples d'expressió d'isoformes no patològiques (o el balanç entre elles) en estadis inadequats, les quals poden afectar molts altres gens, com és el cas de la sobreexpressió de la proteïna SF2/ASF. Aquesta proteïna desencadena la transformació de la cèl·lula i el desenvolupament del tumor degut a l'alteració que provoca en el balanç entre isoformes de diversos reguladors clau del creixement cel·lular (Ben-Dov et al., 2008). L'expressió d'una isoforma inapropiada pot ser bàsic en el desenvolupament de càncer, com succeeix amb l'expressió de la isoforma embrionària de la piruvat quinasa (*PKM2*) en l'estadi adult, tot permetent a la cèl·lula cancerígena mantenir alts nivells de captació de glucosa i producció de lactosa sota condicions aeròbiques, el que es coneix com a efecte Warburg (Christofk et al., 2008; Hartmann and Valcarcel, 2009). Per una altra banda, la proporció alterada entre les dues isoformes de *WT1*, les quals només difereixen en tres aminoàcids, deguda a una mutació en un lloc d'splicing és la que provoca l'aparició del síndrome de Fraser i altres malalties derivades (Faustino and Cooper, 2003).

Altres malalties estan relacionades amb el número de repeticions de seqüències curtes que poden alterar els patrons d'splicing ja sigui perquè segresten proteïnes d'unió a l'RNA i no permeten realitzar el procés d'splicing correctament en altres pre-mRNA's o perquè són *hotspots* de generació de nous llocs d'splicing, com és el cas de les repeticions *Alu*, molt abundants i amb molts llocs d'splicing potencials (Tazi et al., 2009). Les malalties associades a l'envelliment també han estat relacionades amb l'aparició d'splicing aberrant (Meshorer and Soreq, 2002), en especial en les malalties neurodegeneratives com l'Alzheimer i la desregulació de l'splicing controlat per NOVA (Tollervey et al., 2011).

1.2 Efecte a nivell de nucleòtid

La lògica de l'splicing alternatiu rau en la mateixa estructura gènica, en la qual trobem exons i introns intercalats. Durant el processat del pre-mRNA, s'ha d'enretirar els introns i mantenir els exons escollits que codificaran per a una proteïna. El complex de l'spliceosoma, format per cinc snRNPs i més de 200

proteïnes (Hartmann and Valcarcel, 2009), és l'encarregat d'aquesta funció mitjançant el reconeixement d'una sèrie de senyals que l'informen sobre on tallar i empalmar. Aquestes seqüències senyal són el llocs d'splicing als extrems 5' i 3' de l'intró que indiquen el punt de tall (5' *splice site* i 3' *splice site*), el nus (*branch point*) i el tram de pirimidines (Kim et al., 2008) . A més a més, tenim una sèrie de seqüències curtes i poc conservades incrustades dins d'exons i introns que tenen una funció estimuladora o inhibidora de la inclusió d'un determinat exó (Matlin et al., 2005).

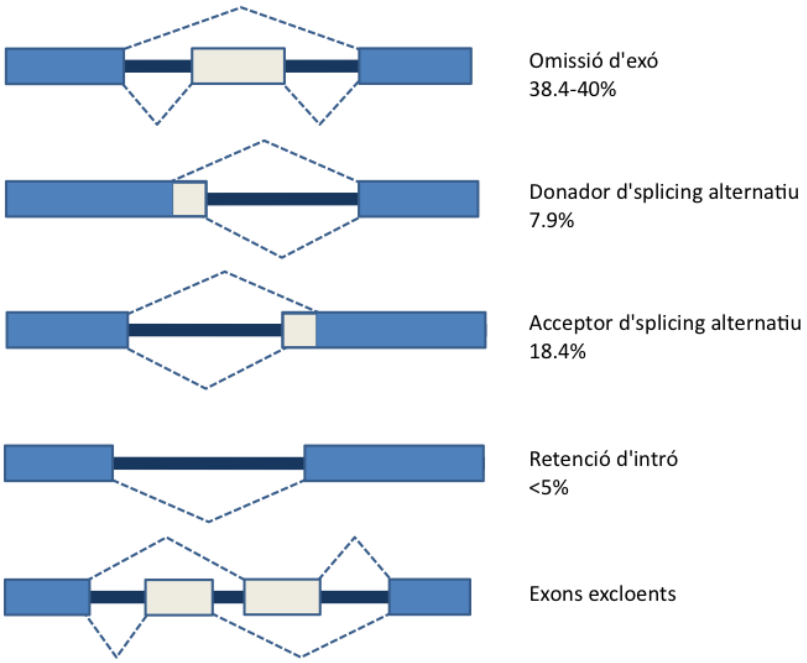

Figura 1: Patrons d'splicing més freqüents. Les capses blaves representen els exons constitutius; les barres fosques, els introns; i les capses blanques els exons alternatius. Les línies discontínues assenyalen les diferents combinacions d'splicing. Els percentatges indiquen la freqüència estimada de cada esdeveniment a metazous més complexos (Keren et al., 2010; Kim et al., 2008).

De la interacció dels diferents factors reguladors d'splicing amb unes seqüències o unes altres (que poden estar competint (Yu et al., 2008)), combinat amb la taxa de transcripció (Kornblihtt, 2007), l'aparició d'estructures secundàries a l'RNA (Shepard and Hertel, 2008) i l'estat de la cromatina (Kolasinska-Zwierz et al., 2009), entre molts altres factors, resulta en un patró d'splicing o un altre. Els principals patrons es resumeixen en omissió d'exó, exons excloents, acceptors/donadors d'splicing alternatius i retenció d'intró (Figura 1) (Kim et al., 2008). Combinant tots aquests patrons amb d'altres menys freqüents, obtenim un

panorama força complicat: s'han descrit més de 150 tipus diferents d'esdeveniments d'splicing (Nagasaki et al., 2006).

Aquesta és una visió per força simplificadora però ja que l'objecte d'aquesta tesi no ha estat centrada en els mecanismes moleculars d'splicing, sino en les conseqüències a nivell de proteïna, m'he limitat a fer una breu pinzellada sobre aquest tema.

1.3 Efectes a nivell de proteïna

Una altra qüestió pendent és determinar quantes isoformes provinents d'splicing alternatiu s'arriben a traduir i quantes d'elles són degradades per mitjà del mecanisme del NMD. S'han fet esforços considerables en aquesta direcció en diverses espècies (Power et al., 2009; Tanner et al., 2007; Tress et al., 2008) mitjançant dades provinents d'espectrometria de masses, encara amb resultats limitats. La importància dels processos d'NMD és discutida i encara no tenim uns valors de referència de la seva incidència a nivell de proteïna. Les estimacions de la quantitat d'isoformes candidates a ser regulades mitjançant NMD estan al voltant d'un 25%-33% dels exons alternatius a humà (Lareau et al., 2007; Stamm et al., 2005). Per una altra banda, s'ha estudiat la conservació de les senyals de NMD, que es mantenen en un 9% dels gens ortòlegs a humà i ratolí (Mudge et al., 2011) o entre un 27-35% dels ortòlegs d'humà, ratolí, vaca i rata (de Lima Morais and Harrison, 2010) i poden afectar gens d'importància cabdal per l'organisme. Aquests i altres autors han postulat que els processos de NMD acoblats a l'splicing alternatiu no són ni molt menys soroll, sinó que formen part d'un mecanisme més ampli, conegut com a RUST (*Regulated Unproductive Splicing and Translation*), que permet una altra via de regulació de l'expressió gènica en moments o localitzacions concrets (Lareau et al., 2007; Lewis et al., 2003).

Tot i la gran variabilitat de processos que tenen lloc a nivell de nucleòtid, els efectes que produeix l'splicing alternatiu a nivell de seqüència de proteïna es limiten a insercions/delecions i/o substitucions de fragments de seqüència (Kondrashov and Koonin, 2001; Kondrashov and Koonin, 2003), la severitat dels quals abasta des de substitucions subtils fins grans delecions de dominis sencers (Figura 2) (Ben-Dov et al., 2008). De totes maneres, només amb la informació dels

canvis a nivell de seqüència aminoacídica és complicat predir quin impacte tindrà l'splicing en l'estructura i funció de la proteïna.

Figura 2: Efecte de l'splicing a nivell de seqüència proteica. La seqüència en color vermell correspon a la part variable entre les isoformes d'un mateix gen. La inserció/deleció representa el canvi que trobem a IL4_HUMAN, isoformes 1 i 2, i la substitució, a GAPD1_HUMAN, isoformes 1 i 6 (nomenclatura de gen a Uniprot/Swissprot)

La informació experimental dels efectes estructurals de l'splicing alternatiu sobre la proteïna és reduïda. Fins ara només 14 proteïnes humanes tenen estructures al PDB per, com a mínim, dos isoformes alternatives, de les quals la majoria són degudes a insercions d'entre 2 i 30 aminoàcids (Hegyi et al., 2011). Aquest canvis tendeixen a evitar els dominis globulars (o a afectar-los marginalment), solen situar-se en regions *coiled* o *loop* localitzades a la superfície i estan enriquits en zones desordenades de la proteïna (Hegyi et al., 2011; Romero et al., 2006). Per una altra banda, diversos estudis han proposat que l'splicing alternatiu pot modificar la localització i funció de la isoforma alternativa si altera o elimina els dominis transmembrana de les proteïnes de membrana, convertint-les així en proteïnes solubles, i/o eliminant el pèptid senyal (Davis et al., 2006; Stamm et al., 2005; Xing et al., 2003).

A continuació menciono una sèrie de casos que han estat estudiats experimentalment, amb o sense estructura cristal·litzada, que ens informen dels efectes funcionals i estructurals de l'splicing a nivell de proteïna .

Entre les parelles resoltes tenim dos isoformes de l'ectodisplasina (EDA), un factor de necrosi tumoral implicat en el desenvolupament ectodèrmic. L'splicing

alternatiu provoca una inserció de només dos residus (Glu308 i Val309) que afecta la forma del lloc d'unió a receptor de la isoforma EDA-A1, a més a més de generar alteracions conformacionals i de càrrega superficial. El resultat és un control de tipus interruptor de l'especificitat del receptor (Hymowitz et al., 2003).

De l'ús alternatiu del primer exó de la sulfotransferasa (SULT)2B1 s'obté una substitució a l'extrem N-terminal i una alteració en l'especificitat de substrat: la isoforma 2B1b actua sobre el colesterol i la 2B1a sobre l'hormona esteroide pregnenolona (Fuda et al., 2002; Hegyi et al., 2011). En el cas de la proteïna Piccolo, proteïna *scaffolding* de la pre-sinapsi i important en l'alliberació de neurotransmissors, l'afinitat pel Ca^{2+} per cada una de les dues isoformes és modulada per un reajustament a l'estructura secundària deguda a una inserció de només 9 residus (Garcia et al., 2004). S'han descrit dues isoformes del receptor de la insulina (IR-A, IR-B) que solament difereixen en 12 aminoàcids degut a l'splicing alternatiu de l'exó 11, però que provoquen canvis considerables en les afinitats d'unió a IGF-I, IGF-II i, en menor mesura, a insulina però també genera diferències en la internalització i reciclatge del receptor i, en definitiva, en la senyalització intracel·lular (Belfiore et al., 2009; Denley et al., 2003).

Aquestes modificacions estructurals solen ser molt subtils i, per tant, difícils de detectar. No és el cas dels canals d'ions, sovint afectats per l'splicing alternatiu i els efectes dels quals són mesurables amb tècniques electrofisiològiques (Stamm et al., 2005). Aquests efectes cobreixen tot el ventall de possibilitats: des de la inactivació total del canal fins lleugeres modulacions de les eficiències d'unió al lligand i temps de resposta. Un cas paradigmàtic és el dels canals BK (*calcium-activated potassium channel*) amb presència d'splicing tant a invertebrats com a vertebrats: tot i que els exons alternatius no estan conservats entre espècies, vindrien a ser *hot-spots* d'splicing que permetrien un ajustament funcional del canal per les diverses necessitats de les espècies (Fodor and Aldrich, 2009). La comprovació per *patch-clamp* de la funcionalitat de les diferents isoformes ha permès descartar isoformes aberrants mitjançant tècniques no computacionals (Ha et al., 2000).

1.3.1 L'splicing alternatiu en les vies cel·lulars

La funció de les proteïnes ve determinada per la xarxa d'interaccions amb altres proteïnes i molècules (DNA, RNA, metabòlits, etc.) en un estat cel·lular concret, formant part de macrocomplexos i organitzant-se en vies (de senyalització, metabòliques, de regulació de l'expressió gènica, etc.) (Eisenberg et al., 2000) Dins d'aquestes xarxes, trobem proteïnes amb un alt nombre d'interaccions (*hubs*) que tenen una importància capital en el bon funcionament de la cèl·lula. El debat encara resta obert sobre si només aquests nodes són essencials o ho són tots (Przytycka et al., 2010), però no hi ha cap dubte que afecten de manera considerable la viabilitat cel·lular. L'splicing alternatiu també actua sobre gens implicats en aquests *hubs*, en especial a les interaccions proteïna-proteïna i d'unió al DNA (Floris et al., 2008), tot aportant així un nou nivell de complexitat al sistema d'interacció proteica (que pot ser entre isoformes d'un mateix gen) i, per tant, al fenotip final de la cèl·lula. A continuació n'esmentaré uns exemples il·lustratius.

La p53 és un factor de transcripció ubic clau en el manteniment de l'estabilitat genètica. A l'hora de respondre a senyals d'estrès, la p53 regula l'expressió d'una bateria de gens implicats en la reparació del DNA, apoptosi i cicle cel·lular per donar la resposta adequada i evitar així el desenvolupament de càncer. S'ha determinat l'existència de 10 isoformes generades per splicing alternatiu i altres mecanismes (Khoury and Bourdon, 2010), les quals només tenen en comú el domini d'unió a DNA. S'ha suggerit que aquestes isoformes poden tenir funcions autònomes depenent del cicle cel·lular, modular l'activitat d'unió a DNA de la isoforma "clàssica" (tant sigui pel seu segrest al citosol o per la formació d'hetero-oligomers) i, fins i tot, regular l'expressió de la p53 clàssica de manera equivalent a la regulació de promotor (Beno et al., 2011; Marcel and Hainaut, 2009).

Hi ha casos molt més generals on l'splicing afecta molts nodes i interaccions al llarg de la via. És el cas del complex de l' NF-κB, un factor de transcripció pleiotròpic que és responsable de les variacions en els nivells d'expressió de múltiples gens involucrats en la resposta immune i inflamatòria, en el desenvolupament, en la supervivència cel·lular i l'apoptosi. S'han descrit esdeveniments d'splicing alternatiu pràcticament a tots els nivells de la via de senyalització de l'NF-κB

(Leeman and Gilmore, 2008), incloent el complex de l' NF-κB, i modulant així la seva pròpia activitat.

Un cas similar de regulació general per splicing és el que trobem en la biosíntesi i captació del colesterol, on l'alteració de les proporcions entre la isoforma completa i les alternatives sembla ser el mecanisme per regular l'expressió i activitat dels gens implicats. Dos gens clau, la reductasa HMGCR i el receptor de l'LDL, entre d'altres, tenen l'splicing alternatiu regulat en resposta als nivells d'esterol cel·lular (Medina et al., 2011)

En altres casos, l'splicing pot afectar famílies dins d'una via, tal com passa amb les proteïnes SMAD. Aquestes família inclou les mitjanceres intracel·lulars de la senyal de la via del TGF-β, factor determinant en el desenvolupament tissular, entre d'altres funcions. S'ha descrit isoformes provinents d'splicing en diverses SMADs, algunes d'elles amb expressió diferencial tissular (SMAD6) o temporal (SMAD2), amb funció de dominant negatiu (SMAD8) i, en general, força regulades en la seva transcripció per tal de mantenir les proporcions entre les diferents isoformes, les quals poden servir per realitzar petits ajustaments en el senyal de la via (Tao and Sampath, 2010).

Una altra família provinent de duplicacions gèniques ancestrals i essencial en el desenvolupament embrionari en metazous són els gens Pax. Aquests factors de transcripció s'organitzen en quatre subfamílies, tots els membres de les quals tenen splicing (Short and Holland, 2008). Les isoformes resultants per cada gen tenen funcions prou diferenciades i crítiques en diferents processos, com en el desenvolupament del múscul esquelètic per part de Pax3/7 (Buckingham and Relaix, 2007), amb fortes dependències en la dosi de cada isoforma de Pax6 durant el desenvolupament de l'iris i el cos ciliar de l'ull (Davis et al., 2009) o vitals durant la diferenciació dels limfòcits B (Robichaud et al., 2008).

Les proteïnes estructurals també estan afectades per splicing alternatiu i, de la mateixa manera, les seves interaccions. La proteïna gegant Titin (o connectina) és essencial en l'assemblatge, integritat i elasticitat del sarcòmer (Trinick and Tskhovrebova, 1999). El gen de Titin codifica per múltiples isoformes amb expressió específica depenent de la fase de desenvolupament o del teixit (a múscul

esquelètic o cardíac), i la proporció incorrecta entre aquestes isoformes pot dur a malalties del cor (Guo et al., 2010).

1.4 Altres fonts de diversitat transcripcional i proteòmica

A banda de l'splicing alternatiu com a font de diversitat transcripcional trobem altres fenòmens que, gràcies a l'avenç en les tècniques de seqüenciació o derivades, s'han pogut valorar en la seva justa mesura. L'inici i terminació alternatius de la transcripció en són els més destacats. La seva freqüència i efecte és similar als de l'splicing alternatiu amb el qual se solapen als extrem 5' i 3' i, per tant, amb el que estan fortament relacionats (Shabalina et al., 2010). La poliadenilació alternativa és també un procés més freqüent del que s'estimà inicialment, fins més del 50% dels gens a mamífers (Lutz, 2008), i pot dur també a la formació de noves isoformes (Yan and Marr, 2005), les quals sovint són específiques de teixit (Wang et al., 2008). La poliadenilació alternativa i l'splicing estan altament correlacionats entre teixits, fet que suggereix que són dos processos coordinats (Marguerat and Bahler, 2010).

Per acabar, el *trans-splicing* i l'RNA editing són dos processos que es creia restringits a eucariotes més simples però que treballs recents, i prou polèmics, han afirmat que són presents en eucariotes més complexos (Li et al., 2009a; Li et al., 2011). Pel que fa al *trans-splicing*, se suposa que és un fenomen no gaire freqüent encara que estudis bioinformàtics han afirmat tot el contrari (Herai and Yamagishi, 2010). La proteïna traduïda prové d'un mRNA quimèric format per dos fragments de pre-mRNA's de gens diferents. S'ha pogut demostrar que s'expressa en teixits sans i no només en processos cancerígens (Li et al., 2009a). Per la seva banda, l'RNA editing consisteix en substitucions post-transcripcionals de bases simples a l'mRNA, el més freqüent del quals és la desaminació de l'adenosina i la substitució C-to-U. Tot plegat pot alterar l'aminoàcid que es traduirà a la seqüència peptídica, tal com ha estat provat, però falta aclarir si és tan estès com s'ha considerat (tot descartant anotacions incorrectes de SNP's o polimorfismes al·lèlics) i la seva relació amb alguna patologia (Li et al., 2009b; Li et al., 2011; Marguerat and Bahler, 2010).

Tot plegat ens mostra que el panorama transcripcional i proteòmic és molt més complex i ric del que es pensava d'antuvi, sense oblidar el rol cada cop més important que tenen els RNA's no codificants.

1.5 Estudis comparatius entre espècies

La conservació al llarg de l'evolució d'un determinat patró d'splicing pot ser un argument ben sòlid a l'hora de defensar la viabilitat i funcionalitat biològica d'una isoforma (Keren et al., 2010). En el cas contrari, la no conservació es pot considerar tant com un indici de l'aparició d'isoformes específiques d'espècie com una evidència de soroll transcripcional (Takeda et al., 2008), però tot i així amb algun possible rol regulador (Artamonova and Gelfand, 2007; Keren et al., 2010). En definitiva, els estudis comparatius d'splicing entre espècies ens poden ajudar a esclarir com l'splicing alternatiu amplia el repertori funcional d'un llinatge i inferir la importància de l'splicing alternatiu en els processos d'especiació i adaptació (Xing and Lee, 2006).

S'han fet una gran quantitat d'estudis comparatius a gran escala entre humà i ratolí en els quals s'ha estimat la conservació d'splicing alternatiu sobretot a nivell d'exó i/o esdeveniment d'splicing (Cusack and Wolfe, 2005; Irimia et al., 2009; Mudge et al., 2011; Nurtdinov et al., 2007), d'altres més centrats a nivell de cDNA i transcrit (Mollet et al., 2010; Takeda et al., 2008), en bases de dades de transcrits i EST's (Kim et al., 2007; Zambelli et al., 2010) o dades genòmiques directament (Koralewski and Krutovsky, 2011), entre molts d'altres. Els valors obtinguts no ofereixen una explicació unívoca però sembla haver-hi una conservació baixa dels exons alternatius (Keren et al., 2010; Modrek and Lee, 2003), sobretot en aquelles isoformes amb expressió menor (Nurtdinov et al., 2007), encara que s'ha trobat nivells alts de conservació d'aquests exons a gens expressats al cervell i en gens implicats en processos de regulació de la transcripció, desenvolupament i processat d'RNA (Yeo et al., 2005).

A nivells més elevats de l'organització biològica, més enllà del nivell de seqüència i motius de nucleòtids, els estudis comparatius de l'splicing alternatiu entre espècies són minsos. És per això que ens hem proposat avançar un pas més en el coneixement d'aquest camp poc conegut, estudiar l'impacte d'splicing alternatiu a

nivell de proteïna i desxifrar quines conseqüències poden tenir en el fenotip i, per tant, en les diferències entre espècies.

OBJECTIUS

2. OBJECTIUS

El present treball s'inclou dins de l'estudi bioinformàtic de la funcionalitat i la conservació entre humà i ratolí de l'splicing alternatiu. Per aprofundir en aquesta àrea de coneixement, ens vam marcar els següent objectius:

- **Estudiar la contribució de l'splicing alternatiu a la diversitat fenotípica**. Mitjançant un estudi comparatiu entre humà i ratolí, ens proposem establir les relacions entre l'splicing alternatiu i altres fonts de diversitat fenotípica com són les divergències de la seqüència proteica, de la regió *cis*-reguladora gènica i de l'expressió gènica.

- **Esclarir la conservació de l'splicing alternatiu regulador**. Mitjançant un estudi comparatiu entre humà i ratolí, ens proposem establir el nivell de l'alteració de l'arquitectura de dominis, els canvis de seqüència que el provoquen, i el possible impacte funcional i regulador de l'expressió gènica.

- **Esbrinar la relació de l'splicing alternatiu, i de l'splicing alternatiu regulador, amb el fenomen de la dominància gènica**. L'splicing alternatiu és un ferm candidat a influir en la dominància, donada la seva capacitat d'introduir variabilitat i la relació inversa entre l'splicing alternatiu i la paralogia, sent aquest últim associat amb la haploinsuficiència. Volem comprovar fins a quin punt aquesta hipòtesi és certa.

RESULTATS I DISCUSSIÓ

3. Estudi de la relació entre l'splicing alternatiu i les divergències de seqüència proteica i de la regió *cis*-reguladora: enllaçant les fonts de diversitat fenotípica

3.1 Introducció

Les diferències qualitatives (com pot ésser la presència/absència d'una proteïna donada) i les diferències quantitatives (per exemple, les variacions en la quantitat de producte proteic) en la composició del proteoma són just a la base de la diversitat fenotípica entre organismes. Avui dia, gràcies a un gran conjunt d'estudis experimentals, computacionals i teòrics, estem començant a conèixer els mecanismes moleculars en el que es basen aquestes diferències com són la duplicació gènica (Lynch and Conery, 2000; Ohno, 1970), la divergència en la seqüència proteica (Hoekstra and Coyne, 2007), divergència en la regió *cis*-reguladora (Carroll, 2000; Wray, 2007), diferències en l'splicing del pre-mRNA (Blencowe et al., 2009; Nilsen and Graveley, 2010), així com canvis en altres mecanismes reguladors de l'expressió gènica (com variacions en l'exportació del mRNA nuclear, la degradació d'mRNA, etc.) (Alonso and Wilkins, 2005).

Entre aquests mecanismes, la capacitat dels gens per codificar diversos transcrits és una font destacada de les diferències a nivell de proteoma entre eucariotes (Blencowe et al., 2009; Nilsen and Graveley, 2010; Tress et al., 2007). Aquesta diversitat transcripcional (TD, per les sigles en anglès) pot ser deguda a l'splicing alternatiu del pre-mRNA o per inicis i/o terminacions alternatives de la traducció, encara que s'accepta de manera general que l'splicing alternatiu és el màxim contribuïdor a aquesta diversitat (Nilsen and Graveley, 2010). És més: el potencial de l'splicing alternatiu per mostrejar l'espai funcional proteic és tan ampli (Kim et al., 2008; Nilsen and Graveley, 2010; Talavera et al., 2007b) que s'ha postulat com

un contribuïdor a les diferències de complexitat entre organismes (Lander et al., 2001; Marden, 2008). La idea de que les variacions en els patrons d'splicing alternatiu poden tenir aquest rol està sent recolzada per un nombre creixent d'estudis que relacionen l'splicing i diverses malalties (Garcia-Blanco et al., 2004; Tazi et al., 2009). També, des d'un punt de vista més general, aquesta idea ha estat explorada i avaluada per molts investigadors que han examinat la quantitat de diferències en l'splicing alternatiu entre organismes (Kim et al., 2007), la conservació dels esdeveniments d'splicing entre espècies (Artamonova and Gelfand, 2007; Calarco et al., 2009; Calarco et al., 2007; Modrek and Lee, 2003; Nurtdinov et al., 2003; Ohler et al., 2005; Pan et al., 2005; Sorek et al., 2004; Takeda et al., 2008; Thanaraj et al., 2003; Yeo et al., 2005), l'impacte de l'splicing a nivell funcional i estructural (Valenzuela et al., 2004), el paper dels codons de terminació prematurs (Wetterbom et al., 2009), la regulació diferencial (Blekhman et al., 2010), etc. Els resultats de tots aquests estudis assenyalen cap a una relació clara entre les diferències interespecífiques i la distribució de les propietats de l'splicing alternatiu.

Uns altres processos diferents des del punt de vista mecanístic, però relacionats amb l'splicing alternatiu, són l'inici/terminació alternatiu de la transcripció, els quals són dos fonts addicionals de la TD d'origen gènic (Landry et al., 2003; Shabalina et al., 2010). Poden generar proteïnes amb extrems N-ter o C-ter diferents, o d'altres modificacions a la seqüència més substancials (Landry et al., 2003; Shabalina et al., 2010), les quals poden comptar amb propietats funcionals similars a les que tenen les isoformes provinents d'splicing alternatiu, com són els dominants negatius o l'aparició de noves localitzacions cel·lulars (Landry et al., 2003). Dades recents indiquen que aquests mecanismes poden ser comparables als obtinguts per l'splicing alternatiu (Latchman, 2008; Shabalina et al., 2010).

A mesura que avancem en el coneixement de les diferents fonts de variabilitat del proteoma, és cada cop més evident que és necessari entendre les relacions entre aquestes fonts si pretenem obtenir una explicació completa de les diferències entre organismes (Alonso and Wilkins, 2005; Copley, 2008). En el cas concret de la TD, no coneixem de quina manera la seva contribució a les diferències fenotípiques està relacionada amb la divergència de les zones codificants i del promotor

(Castillo-Davis et al., 2004; Chiba et al., 2008). La TD pot jugar un paper secundari: per exemple, la TD podria modular l'efecte de les mutacions a la regió *cis*-reguladora per mitjà de la producció d'una quantitat determinada d'isoformes inactives o de dominats negatius, amb una distribució que depengués del teixit. Però també podria ser un actor principal, ja que si ens fixem en la seva naturalesa reguladora (Barberan-Soler et al., 2009; Lois et al., 2007), les diferències interespecífiques podrien ser equivalents a les mutacions a la regió reguladora del gen. Això podria ocórrer especialment en els casos en els que la identitat de seqüència aminoacídica està per sota del llindar de la divergència funcional.

En principi, esperem certa relació entre la TD i les divergències de seqüència proteica i de la regió *cis*-reguladora, donat el solapament entre les seves seqüències codificants. Per exemple, les seqüències exòniques a 5' i 3' implicades en l'exclusió d'intró se solapen amb la regió codificant (Fichant, 1992; Whamond and Thornton, 2006); els *enhancers* i *silencers* d'splicing poden trobar-se dins dels exons (Goren et al., 2006; Schaal and Maniatis, 1999; Smith and Valcarcel, 2000). Per tant, els canvis interespecífics en les regions codificadores estan probablement relacionats amb les variacions en els patrons de TD. Per una altra banda, la quantitat creixent d'estudis que connecten la transcripció amb l'splicing alternatiu (Chern et al., 2008; Kornblihtt, 2006; Kornblihtt et al., 2004; Pandit et al., 2008), junt amb l'existència d'una lleu correlació entre la divergència de les regions codificants i promotores (Chiba et al., 2008), ens suggereix l'existència d'una relació entre la zona *cis*-reguladora i la divergència a nivell de TD. Malgrat això, i tot i la importància del tema per a poder entendre les diferències fenotípiques, no hi ha cap estudi que connecti la TD i qualsevol d'aquestes dues variables implicades en el fenotip. Per tal d'esclarir aquest assumpte, hem estudiat la relació entre els patrons de TD (presència/absència de la TD i equivalència d'isoformes) i les divergències de les seqüències proteiques (DP, per les sigles en anglès) i de la regió *cis*-reguladora (DcR, per les sigles en anglès). També hem indagat quina relació existeix entre els patrons de TD i la divergència de l'expressió gènica, ja que és un fenotip de baix nivell que és resultat de diverses contribucions a la diversitat fenotípica (regulació del promotor, estat de la cromatina, etc.).

Com es veurà en el que segueix, els nostres resultats, obtinguts mitjançant un estudi a nivell genòmic de la TD d'humà i ratolí, proporcionen una caracterització del lligam entre la TD per una banda i de la DP, DcR i la divergència de l'expressió gènica per una altra. Primer, mostrem que la coincidència de TD (la fracció d'ortòlegs humà-ratolí amb més d'un transcrit en tots dos gens, Figura 1A) està relacionada amb la DP, la DcR i, menys significativament, amb la divergència de l'expressió gènica. Segon, obtenim uns resultats oposats pel que fa a l'equivalència d'isoformes (la fracció d'isoformes alternatives que són equivalents a totes dues espècies, Figura 1B), la qual no està relacionada amb la DP i la DcR. Aquests resultats situen la contribució de la TD com una font de diferències fenotípiques respecte la DP i la DcR, tot indicant que la TD pot tenir un impacte en les diferències fenotípiques abans que la DP i la DcR puguin fer-ho.

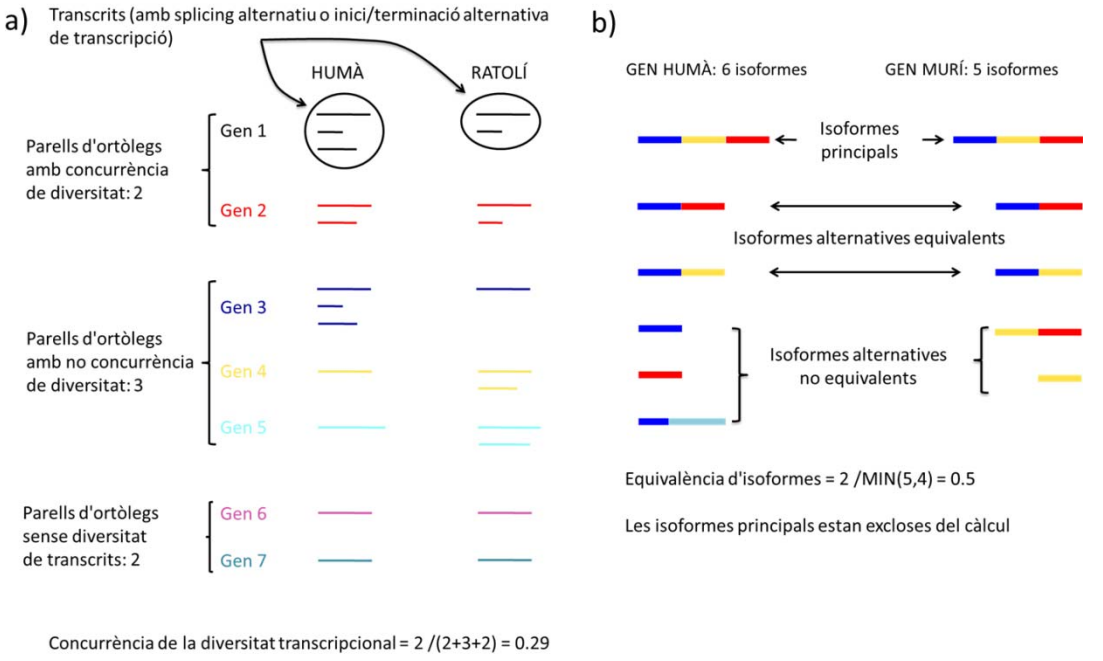

Figura 1: Descripció de les variables utilitzades en aquest treball per a caracteritzar la diversitat transcripcional. (A) Concurrència de la diversitat transcripcional i (B) equivalència d'isoformes. La conservació de la diversitat transcripcional és la fracció d'ortòlegs en els quals tots dos gens codifiquen per més d'un transcrit. L'equivalència d'isoformes és la fracció mitjana d'isoformes alternatives conservades entre ortòlegs (veure *Materials i Mètodes*). En aquesta figura només mostrem com es va calcular la fracció d'isoformes alternatives conservades; tot seguit, es va fer la mitjana d'aquest valor sobre el total de parelles d'ortòlegs en els que es va comparar isoformes alternatives

3.2 Material i mètodes

3.2.1 El conjunt de dades TD

En aquest treball hem comparat la diversitat transcripcional (TD) entre ortòlegs d'humà i ratolí. Per tant, el primer requisit era obtenir una llista de gens d'humà i de ratolí amb els seus transcrits corresponents. Per aconseguir aquesta llista, vam combinar dos bases de dades manualment revisades: UniProt/SwissProt (Boeckmann et al., 2003; The_UniProt_Consortium, 2010) i RefSeq (Pruitt et al., 2007). Vam utilitzar el següent protocol tant per humà com per ratolí. Primer de tot, vam recuperar tots els gens d'humà per UniProt/SwissProt i RefSeq. Segon, vam agrupar els mateixos gens de les dues fonts (UniProt/SwissProt i RefSeq) mitjançant l'identificador GeneId. Tercer, vam eliminar els gens que tinguessin més conflictes a la seqüència del que esperaríem a l'atzar. Per això vam modelitzar la distribució de conflictes amb una distribució Poisson en la que es va calcular la probabilitat d'observar un conflicte en una seqüència a partir de la informació de seqüència d'UniProt/SwissProt. Aleshores vam excloure tots aquells gens en els que el nombre de conflictes observats tinguessin una probabilitat inferior al 0.05 (la informació de RefSeq per aquests gens també va ser descartada). Quart, vam descartar les isoformes provinents de RefSeq que comencessin per XP_ o ZP_. Cinquè, vam recuperar les seqüències de transcrits per cada gen que estiguessin disponibles per cada base de dades. Sisè, vam unificar totes les seqüències mitjançant el programa d'agrupació de seqüències CD-Hit (Li et al., 2001) ja que, en molts casos, versions lleugerament diferents del mateix transcrit estan emmagatzemades tant a SwissProt com RefSeq. Vam aplicar paràmetres molt restrictius a l'hora d'executar CD-Hit: només vam permetre un residu de diferència entre les mides de les proteïnes, i el nombre de desaparellaments (*mismatches)* entre seqüències equivalents havia d'estar entre els límits esperats pels conflictes de seqüència calculats prèviament. Setè, per assegurar-nos que no estàvem agrupant isoformes no equivalents, vam imposar la màxima distància entre desaparellaments (és a dir, el nombre de residus entre els dos desaparellaments més allunyats a l'alineament), k, que havia de seguir la següent distribució:

$$(N-k)\cdot\frac{\binom{k-1}{n-2}}{\binom{N}{n}} \qquad (1)$$

on n és el nombre de desaparellaments a l'alineament, i N és la longitud de la seqüència.

Llavors vam utilitzar el programa InParanoid (Ostlund et al., 2010), a fi i efecte d'obtenir les relacions d'ortologia entre els gens d'humà i ratolí. Al final del procés, vam obtenir 13970 parelles d'ortòlegs humà-ratolí, després d'excloure aquells casos en els que no es va trobar cap ortòleg o en les situacions que no s'hagués pogut establir cap parella única humà-ratolí. Aquest conjunt de dades constituí el punt de partida per tots els anàlisis posteriors. S'ha de precisar que degut a que no sempre es va poder calcular les variables desitjades (per exemple, problemes a l'hora de definir les regions promotores o absència de dades sobre l'expressió gènica) el nombre final de gens utilitzats poden ser diferents per cada cas particular.

3.2.2 La concurrència de la TD

La concurrència de TD, explicada a la Figura 1A, va ser definida com la fracció de parelles d'ortòleg per les quals tots dos gens codifiquessin per més d'un transcrit. Vam representar aquesta fracció en funció de les quatre variables escollides per aquest estudi: identitat de seqüència proteica, identitat de seqüència de la regió promotora i distància entre motius, i correlació de l'expressió gènica. Una definició més formal de la concurrència de la TD és la següent:

$$\text{Prob}\{[(TD_{Hs}=YES)\cap(TD_{Mm}=YES)]/var=X\} \qquad (2)$$

on TD_{Hs} és una variable indicadora que té dos valors: YES, quan l'ortòleg humà té més d'un transcrit, i NO, quan només en tenen un. TD_{Mm} és l'equivalent per ratolí de TD_{Hs}. "var" és una de les quatre variables mencionades anteriorment (identitat de seqüència de proteïna, etc.), i X representa els valors que pren (per exemple, per la identitat de seqüència proteica, X pot tenir valor d'entre 0 i 100%).

3.2.3 Equivalència d'isoformes

L'equivalència d'isoformes, explicada ja a la Figura 1B, va ser definida com la fracció mitjana de isoformes alternatives conservades entre ortòlegs. Una vegada més vam representar aquesta fracció en funció de les quatre variables escollides per aquest estudi: identitat de seqüència proteica, identitat de la seqüència de la regió promotora i distància entre motius, i correlació de l'expressió gènica. Una definició més formal de l'equivalència d'isoformes és la següent:

$$\frac{1}{N_X}\sum_{i=1}^{N_X}\frac{c_i}{\min(n_{Hs,i},n_{Mm,i})} \qquad (3)$$

on N_X és el nombre d'ortòlegs, per un valor donat de la variable seleccionada (per exemple, al 95% d'identitat de seqüència), per la qual les isoformes van ser comparades (pel procés de comparació, veure el protocol més endavant); el subíndex "i" recorre totes les parelles d'ortòlegs, c_i és el nombre d'isoformes homòlogues per parella "i", i $\min(n_{Hs,i},n_{Mm,i})$ és el nombre mínim de transcrits per humà i ratolí per parella, $n_{Hs,i}$ i $n_{Mm,i}$, respectivament. El mínim va ser escollit ja que és el valor màxim que pot ser adoptat per c_i.

3.2.4 Isoformes homòlogues

A l'hora de calcular l'equivalència d'isoformes aplicant la Equació (3), és necessari conèixer si les isoformes per un gen humà concret tenen algun equivalent o homòleg entre les isoformes de l'ortòleg corresponent a ratolí. Per això vam utilitzar un procediment desenvolupat al grup, i basat en xarxes neuronals (Talavera et al., 2007a), el qual utilitza seqüències proteiques completes. Aquesta aproximació té l'avantatge, si ho comparem amb les aproximacions basades en la presència/absència d'exons, que ens permet el tractament d'esdeveniments multi-exònics, els quals representen un percentatge gens menyspreable dels fenòmens d'splicing alternatiu i d'iniciació de la transcripció (Nagasaki et al., 2006).

La informació necessària per aplicar el procediment de comparació d'isoformes és: (i) les isoformes principals dels dos ortòlegs; i (ii) per un dels ortòlegs, els canvis de seqüència entre la isoforma principal i l'alternativa, tal com estan definits a UniProt/SwissProt (el nostre procediment requereix el coneixement dels canvis de seqüència entre isoformes per només un dels gens). Aquesta comparació no es va fer quan no teníem informació disponible d'UniProt/SwissProt. La sortida del programa és senzilla: una resposta YES/NO, indicant si les isoformes alternatives són homòlogues, i un índex de fiabilitat per a aquesta resposta, que abasta des de 0 (fiabilitat mínima) fins a 9 (màxima fiabilitat). Només vam analitzar les comparacions que tinguessin una fiabilitat igual o superior a 5. A més a més, vam decidir descartar tots els casos en els que la identitat de seqüència entre ortòlegs fos inferior al 50%, ja que el mètode va ser entrenat amb seqüències la majoria de les quals tenien una identitat superior al 50%. També vam cenyir les nostres comparacions a aquelles parelles on la diferència de mida fos inferior del 20% de la longitud de la isoforma principal, llindar establert en base a una inspecció manual de les prediccions. En realitat, vam explorar tres escenaris alternatius: diferències de mida igual a zero, diferències per sota del 10% i diferències per sota del 20%.

3.2.5 Controls pels biaixos de mostratge i de base de dades. Efectes específics al·lèlics

En el nostre estudi sobre la connexió entre la concurrència de la TD i la resta de variables (DP, DcR i divergència de l'expressió gènica), els biaixos a qualsevol de les bases de dades utilitzades (UniProt/SwissProt i RefSeq) o en el nostre procediment de mostreig podien fer aparèixer relacions espúries. Amb l'objectiu de tenir en compte aquests efectes a l'hora d'interpretar els resultats, vam dibuixar en tots els casos el valor esperat de la concurrència de TD sota l'assumpció d'independència. És a dir, vam representar:

$$\text{Prob}\{(TD_{Hs}=YES)/var=X\} \cdot \text{Prob}\{(TD_{Mm}=YES)]/var=X\} \qquad (4)$$

on el significat de les diferents variables és el mateix que a (2). Les corbes corresponents a aquest control apareixen en gris a les figures 2-5 i 7, juntament amb el seu embolcall (+/- 1 desviació estàndard).

En el primer bloc del nostre treball, vam obtenir (Figures 2-4) una relació monotònica entre la concurrència de la TD i la DP, i per les dos mesures de DcR. Podria haver passat que, degut a que estàvem treballant amb una mostra del total de la població, aquesta relació monotònica fos resultat d'un efecte de mostreig. A fi i efecte de descartar aquesta possibilitat, vam idear un control format per 100000 versions del nostre conjunt de dades, en el qual la TD fos assignada a l'atzar a cadascun del gens d'humà i ratolí, tot preservant les freqüències de TD totals per cada espècie. D'aquesta manera vam simular el procés de mostreig a partir d'una població sense relació entre la concurrència de la TD i la resta de variables i vam obtenir la relació mitjana entre elles. Tot seguit vam comparar aquest escenari amb els nostres resultats. Les corbes corresponents a aquest control es poden veure a la Figura 6, juntament amb el seu embolcall (+/- 1 s.d.).

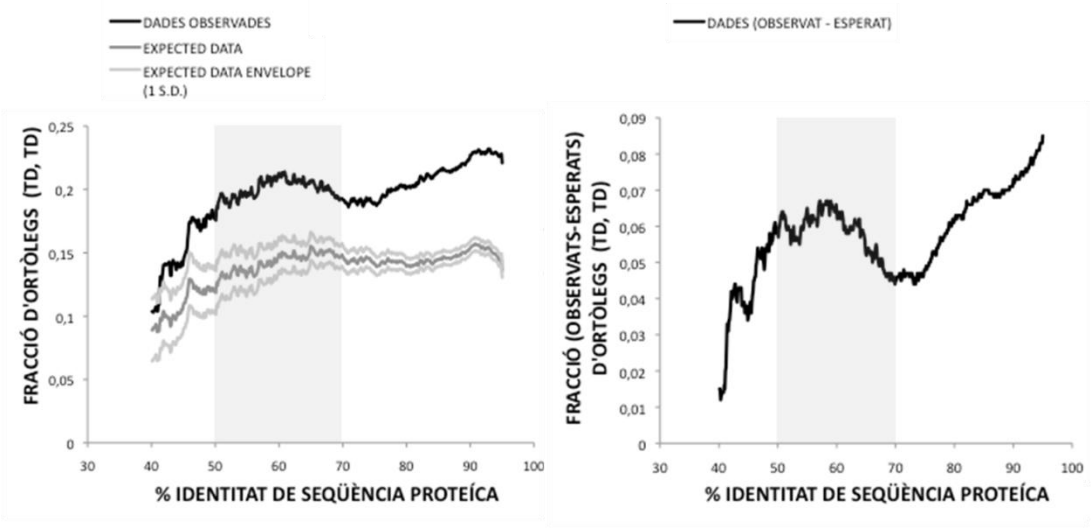

Figura 2. Relació entre la concurrència de la TD i la DP. A l'esquerra, mostrem els valors observats i esperats per a la concurrència de la TD. Els valors esperats varen ser obtinguts sota l'assumpció d'independència pel que fa a la TD entre els ortòlegs d'humà i ratolí. S'hi adjunta un embolcall de desviació estàndard d'1. Podem veure que les corbes són diferents i que només es troben en valors baixos d'identitat de seqüència. A la dreta, hem representat la diferència entre els valors observats i esperats. Podem veure que la concurrència de la TD tendeix a disminuir a mesura que les proteïnes ortòlogues divergeixen. L'àrea enfosquida correspon als límits d'identitat de seqüència per sobre dels quals la funció proteica és, en general, conservada.

Finalment, un tercer aspecte que vam tenir en compte fou la possibilitat que part de la TD observada fos deguda a variabilitat d'origen al·lèlic, o individual, la qual no té rellevància per les diferències fenotípiques interespecífiques. Pel que fa a l'splicing alternatiu, s'ha postulat que entre el 5% i el 15% (Nembaware et al., 2004; Wang et al., 2008) correspon a variabilitat interindividual, la qual probablement no influirà en la diversitat fenotípica interespecífica. A priori, però, no podem excloure la possibilitat que part dels nostres resultats corresponguin a variacions individuals o al·lèliques, encara que esperem que aquest efecte sigui poc freqüent ja que moltes isoformes de les bases de dades revisades manualment són el resultat d'estudis específics amb la intenció de caracteritzar un gen donat i la seva funció.

3.2.6 Càlcul de la DP

Les seqüències proteiques, tan les isoformes principals com les alternatives, van ser obtingudes a partir de les bases de dades d'UniProt/SwissProt i RefSeq. Vam utilitzar el percentatge d'identitat de seqüència aminoacídica entre els ortòlegs d'humà i ratolí com a referència de la DP. Aquest percentatge va ser calculat, després d'alinear les seqüències (fent servir l'algoritme de programació dinàmica de Needleman i Wunsch (Needleman and Wunsch, 1970)), com el nombre total de parells de residus idèntics dividit per la mida mitjana de les dues proteïnes alineades.

3.2.7 Càlcul de la DcR

Les regions *cis*-reguladores pels gens d'humà i ratolí van ser definides com els 1000 parells de bases situats a 5' del lloc d'inici de transcripció (TSS). Aquestes seqüències van ser obtingudes des del UCSC genome server (Fujita et al., 2011). Ja que no estava del tot clar què és el constitueix la regió *cis*-reguladora d'un gen, vam provar diverses definicions (2000 i 500 pb a 5' del TSS i de -700 a +300 pb en relació al TSS (Zhang, 2007)), amb les quals vam obtenir resultats similars.

Per una parella d'ortòlegs donada, les regions reguladores d'humà i ratolí van ser comparades fent servir dos mesures diferents: el percentatge d'identitat de seqüència i la divergència dels motius compartits (*shared motif divergence*) (Castillo-Davis et al., 2004). La identitat de seqüència va ser calculada com el

percentatge de parells de bases idèntics després d'alinear-les amb ClustalW (Thompson et al., 1994). Per una altra banda, la divergència de motius compartits entre dos seqüències és (Castillo-Davis et al., 2004) "la fracció d'ambdues seqüències que no conté un regió de similitud local significativa". La divergència de motius compartits varia entre 0 (identitat de motius completa) i 1 (absència de motius compartits); va ser calculada mitjançant el programa distribuït per Castillo-Davis i col·laboradors (Castillo-Davis et al., 2004).

3.2.8 Càlcul de la correlació de l'expressió gènica

Les dades d'expressió gènica es van descarregar des del servidor de biogps (http://biogps.gnf.org/): les dades per humà i ratolí (Su et al., 2004) foren U133A/GNF1H i GNF1M, respectivament. En elles, cada gen està caracteritzat pel seu nivell d'expressió a 26 teixits. Per a la divergència de l'expressió gènica entre els ortòlegs d'humà-ratolí, vam seguir en Liao i Zhang (Liao and Zhang, 2006) i vam calcular el coeficient de correlació de Pearson entre els perfils d'expressió tissular respectius per cada parell d'ortòlegs:

$$\frac{1}{N_{Tissue}} \cdot \frac{\sum_{i=1}^{N_{Tissue}} (HsExp_i . MmExp_i - < HsExp > . < MmExp >)}{\sigma_{HsExp} . \sigma_{MmExp}} \qquad (5)$$

on N_{Tissue} és igual a 26; $HsExp_i$ i $MmExp_i$ són els nivells d'expressió gènica pel teixit "i" a humà i ratolí, respectivament; i <HsExp>, <MmExp> són les mitjanes corresponents; σ_{HsExp} i σ_{MmExp}, les desviacions estàndard. Recordar que el càlcul només es va dur a terme en aquells ortòlegs amb informació disponible de l'expressió gènica.

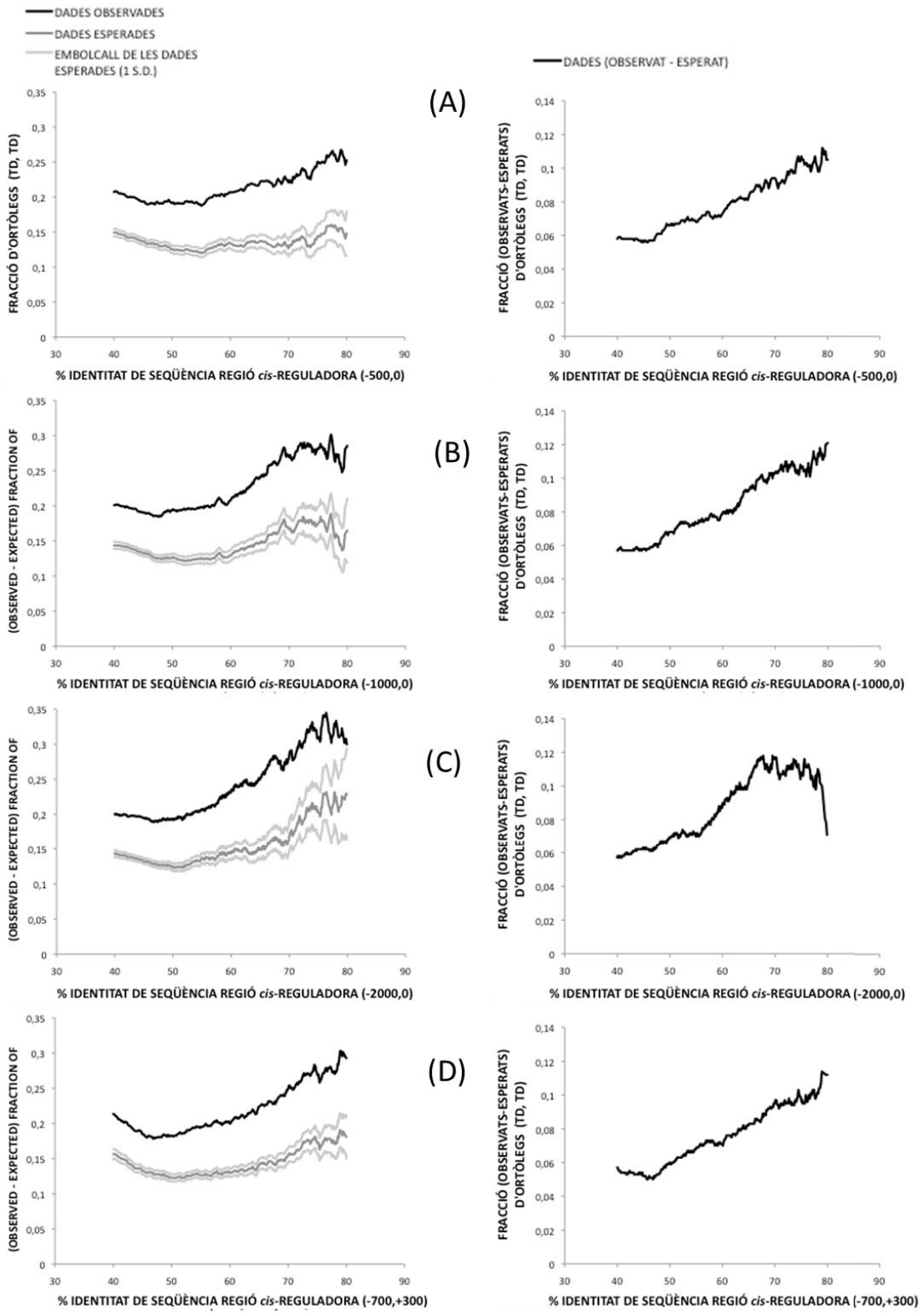

Figura 3 Relació entre la concurrència de la TD i DcR (representat per la identitat de seqüència de la regió *cis*-reguladora). Vam utilitzar quatre definicions de la regió *cis*-reguladora: 500 (A), 1000 (B) i 2000 (C) parells de bases a 5' del TSS, i de -700 a +300 parells de bases en relació al TSS (D). Per tots els casos, mostrem dos gràfics: a l'esquerra, les corbes observades i esperades; a la dreta, la diferència entre elles. En tots els casos, observem un relació evident entre la concurrència de la TD i la divergència a la regió *cis*-reguladora.

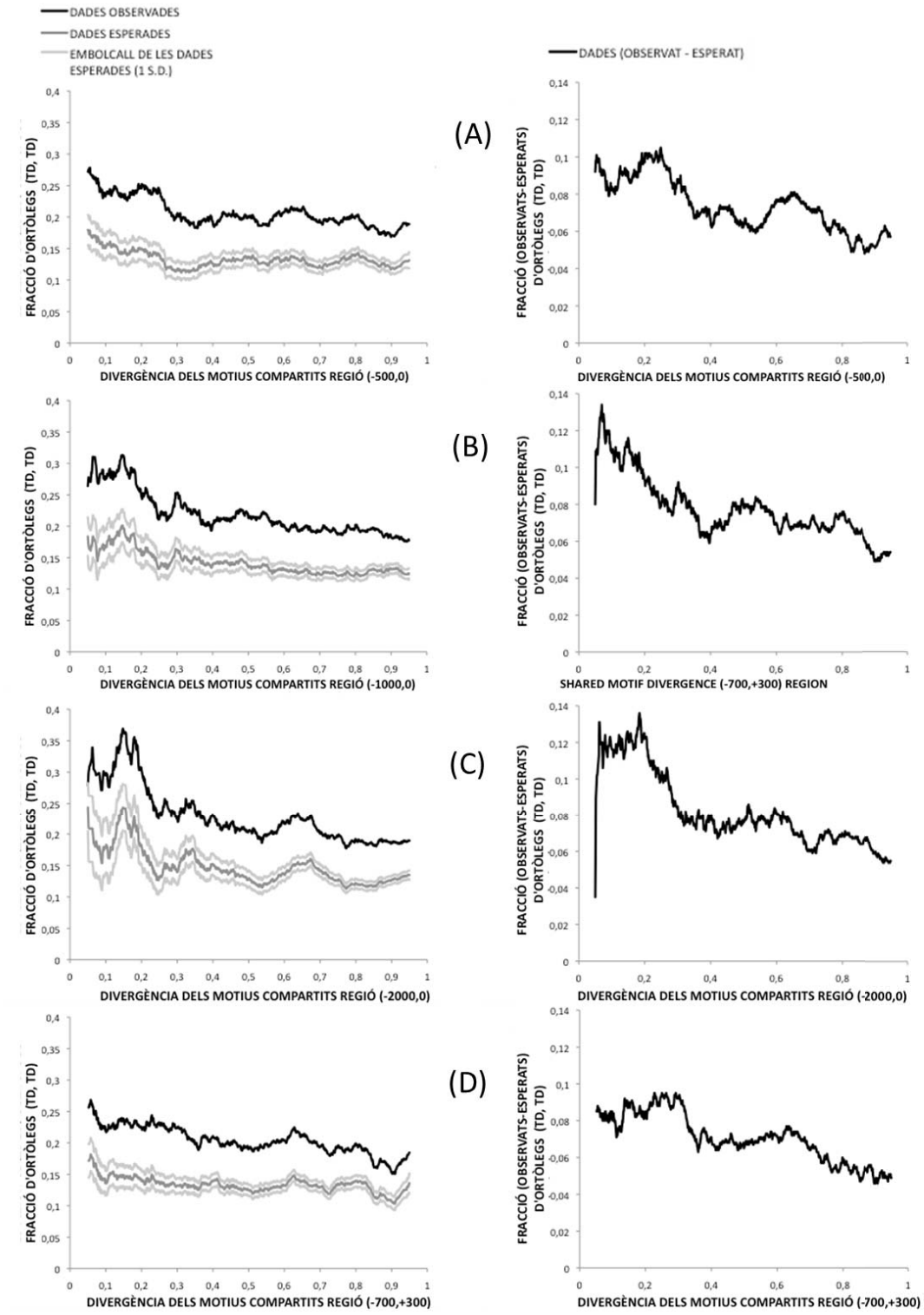

Figura 4. Relació entre la concurrència de la TD i DcR (representat per la divergència dels motius compartits (Castillo-Davis et al., 2004)). Com en el cas de la identitat de seqüència a la regió *cis*-reguladora, vam utilitzar quatre definicions de la regió *cis*-reguladora: 500 (A), 1000 (B) i 2000 (C) parells de bases a 5' del TSS, i de -700 a +300 parells de bases en relació al TSS (D). Per tots els casos, mostrem dos gràfics: a l'esquerra, les corbes observades i esperades; a la dreta, la diferència entre elles. En tots els casos, observem un relació evident entre la concurrència de la TD i la divergència dels motius conservats.

3.3 Resultats

Tal com hem explicat, vam utilitzar una sèrie de descriptors per caracteritzar la TD: (i) Coincidència de TD (és a dir, la fracció de parells d'ortòlegs en els quals tots dos gens codifiquen per més d'una isoforma, Figura 1A) i (ii) equivalència d'isoformes (la fracció d'isoformes alternatives equivalents entre ortòlegs, Figura 1B). Per cada un dels descriptors vam estudiar la seva relació amb la DP, la DcR i la divergència de l'expressió gènica entre els ortòlegs humà-ratolí (veure *Materials i Mètodes*).

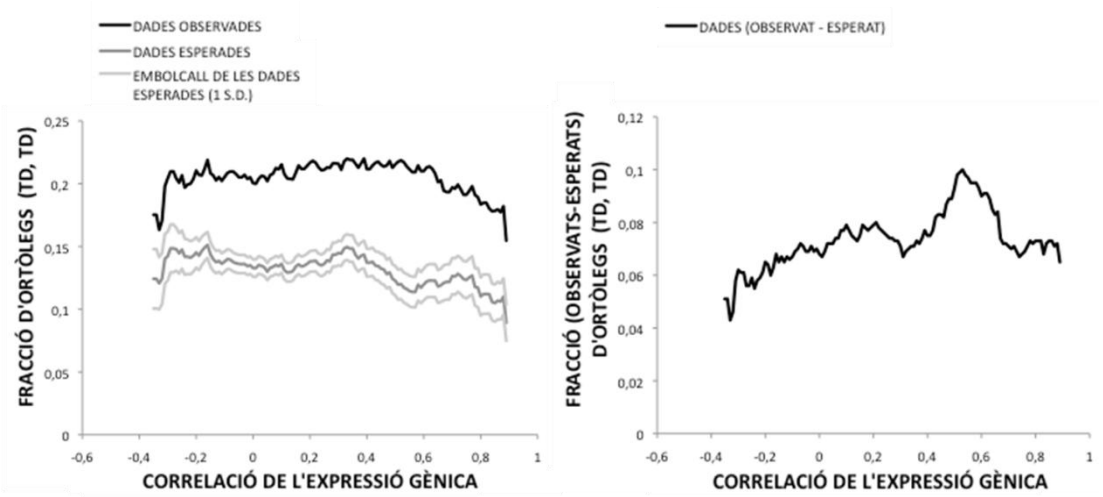

Figura 5. Relació entre la concurrència de la TD i la divergència de l'expressió gènica. A l'esquerra del gràfic, es mostren els valors observats i esperats per a la concurrència de la TD. Els valors esperats s'obtingueren sota l'assumpció d'independència per a la TD entre els ortòlegs d'humà i ratolí. S'hi adjunta un embolcall de desviació estàndard d'1. No podem apreciar una diferència clara entre les corbes, a banda de la distància entre elles. A la dreta, mostrem la diferència entre els valors observats i esperats. Podem veure, encara que no tan clarament com per la DP i la DcR, que la concurrència de la disminueix a mesura que l'expressió gènica divergeix.

3.3.1 Resum del conjunt de dades TD

En primer lloc, vam obtenir un recull de bona qualitat dels gens i les seves isoformes per humà i ratolí, amb les seves seqüències corresponents. Per això vam fusionar les bases de dades d'UniProt/SwissProt (Boeckmann et al., 2003; The_UniProt_Consortium, 2010) i RefSeq (Pruitt et al., 2007), tot aplicant una sèrie de filtres de qualitat segons el seguit de criteris de seqüència explicats (veure *Materials i Mètodes*); a partir d'aquest conjunt de gens vàrem generar una llista d'ortòlegs humà-ratolí mitjançant InParanoid (Ostlund et al., 2010). En el nostre

conjunt de dades final, format per 13970 parells d'ortòlegs humà-ratolí, 46% i 30% dels gens d'humà i de ratolí, respectivament, tenien més d'una isoforma i foren considerats com a gens amb TD. S'ha de puntualitzar que aquest conjunt de dades va incloure tant gens multiexònics com gens amb un únic exó, degut a que la TD pot ser deguda als fenòmens d'inici/terminació alternatiu de la transcripció.

Degut al processament manual de les dades que han aplicat els responsables de generar les bases de dades d'UniProt/SwissProt (Boeckmann et al., 2003; The_UniProt_Consortium, 2010) i de RefSeq (Pruitt et al., 2007), i al posterior filtratge que hem realitzat per seleccionar les seqüències d'aquest treball, els valors de la TD obtinguts són inferiors als proposats en altres estudis, com és el cas de l'splicing alternatiu (Pan et al., 2008; Tress et al., 2007; Wang et al., 2008). De totes maneres, el nostre percentatge de gens amb TD (46% i 30% per humà i ratolí, respectivament) representen una fracció destacada de les estimacions més altes per al percentatge de gens amb diversitat transcripcional a humà i ratolí, al voltant del 90% (Wang et al., 2008) i del 50% (Bingham et al., 2008), respectivament (tot afegint el fet que aquestes estimacions són per gens multiexònics i , per tant, hauríem de reduir-les de l'ordre d'un 5%-10% per fer-les comparables amb les nostres dades). No obstant això, per descartar la possibilitat que els nostres resultats siguin deguts a un efecte de mostratge, a un biaix a la base de dades o producte d'un efecte de la variabilitat individual, vam concebre els tres controls que es mostren a les Figures 2-7 (veure també *Materials i Mètodes*).

A banda de la qualitat de les dades, un avantatge destacat a l'hora d'utilitzar les bases de dades ja mencionades és que ens proveeixen de seqüències d'isoformes alternatives que han estat observades experimentalment. És a dir, aquestes seqüències, que formen part del nucli de la segona part d'aquest treball (*Resultats* secció "3. Equivalència d'isoformes vs DP, DcR i divergència de l'expressió gènica "), no són inferides. Aquest és un aspecte important, ja que els productes de l'splicing alternatiu poden ser, en un nombre gens menyspreable de casos, resultat de canvis de seqüència complicats (Nagasaki et al., 2006) que presenten moltes dificultats a l'hora de ser deduïdes a partir de certes fonts de dades.

3.3.2 Concurrència de la TD respecte DP, DcR i divergència de l'expressió gènica

La concurrència de la TD va ser representada com a funció de la DP (simbolitzada pel percentatge d'identitat de seqüència proteica) (Figura 2, en negre, gràfic a l'esquerra), DcR (definida per dos variables: percentatge d'identitat de seqüència i distància entre motius) (Figures 3 i 4, en negre, gràfics a l'esquerra) i divergència en l'expressió gènica (representada pel coeficient de correlació de Pearson entre els patrons d'expressió gènica d'humà i ratolí) (Figura 5, en negre, gràfic a l'esquerra).

Com a referència, vam dibuixar a la mateixa figura (Figures 2-5, en gris, gràfics a l'esquerra) la concurrència de la TD esperada (Equació 4), que vam obtenir sota l'assumpció d'independència entre la TD d'humà i ratolí (veure *Materials i Mètodes*). Les fluctuacions observades per aquest terme són d'origen mostral, però també poden incloure efectes tècnics (degut als límits de disponibilitat de dades a les bases de dades, qualitat de les dades, etc.). La corba de la concurrència de la TD esperada ens va permetre discernir si les tendències presents a les dades observades eren degudes a una relació real entre la TD a humà i ratolí o tenien un origen tècnic (quan s'observés una coincidència entre les dades observades com les esperades). S'ha d'emfatitzar que en la nostra anàlisi, quan parlem de tendències només ens referim a aquelles de naturalesa monotònica, per exemple, corbes creixents o decreixents. Aquestes són les que amb més probabilitat es poden interpretar com una conseqüència del solapament entre el motius de seqüència que codifiquen per la TD i aquells que codifiquen per la resta de propietats (veure *Introducció*). Hem de puntualitzar que vam obtenir resultats molt similars per les diferents definicions de la regió *cis*-reguladora (veure Figures 3 i 4) i, per aquesta raó, només mostrarem els resultats per una de les regions *cis*-reguladores, en concret, la regió que cobreix els 1000 parells de bases situats a 5' del lloc d'inici de transcripció (TSS).

Els nostres resultats (Figures 2-5, en negre, gràfics a l'esquerra) mostren una corba pràcticament monotònica per a la concurrència de la TD vs. DP i per les dos mesures de la DcR. També vam apreciar que les corbes per les dades observades i esperades mostraven un comportament monotònic similar. Això ens indica que

part de les tendències observades podrien ser degudes a biaixos a les bases de dades. A fi i efecte de refinar aquesta visió, vam calcular i visualitzar la diferència entre la concurrència de TD observada i esperada. En tots quatre casos, vam veure que hi havia una tendència significativa en la direcció esperada (Figures 2-5, gràfics a la dreta): vam obtenir uns coeficients de correlació d'Spearman de 0.62, 0.99, -0.81 i 0.55, per TD vs. DP, vs. les dos mesures de DcR i vs. la correlació de l'expressió gènica, respectivament, amb els p-valors propers a zero en tots els casos. És a dir, la variació entre les corbes observades i esperades no era constant: quan anàvem des d'identitat de seqüència elevades cap a baixes, la corba observada tendia gradualment cap a la corba esperada; hem de mencionar que totes dues corbes només es creuaven en el cas de la DP. En resum, aquests resultats ens indiquen que també hi ha una relació entre la concurrència de la TD i la resta de variables, a banda de possibles biaixos d'origen tècnic representats per la corba de valors esperats.

Com ja hem dit abans (*Resultats* secció "1.Resum del conjunt de dades TD"), degut als filtres tan estrictes que vàrem aplicar a les dades, els percentatges observats de gens amb TD, 46% i 30% per humà i ratolí, respectivament, eren inferiors respecte certes estimacions recents (Pan et al., 2008; Tress et al., 2007; Wang et al., 2008). Degut a aquesta diferència, podria ser que les tendències que hem discutit prèviament (Figures 2-4) fossin degudes a un efecte de mostreig més que no pas a l'existència d'una tendència real. Amb l'objectiu de descartar aquesta possibilitat, vam simular el procés de mostreig a partir d'una població sense relació entre la TD i la resta de variables: vam generar a l'atzar 100000 mostres amb les característiques del nostre conjunt de dades (13970 parells d'ortòlegs, i 46% i 30% d'humà i ratolí, respectivament, amb TD), però assumint independència entre la TD als ortòlegs d'humà i ratolí. Els resultats s'exposen a la Figura 6, on podem veure que la corba mitjana de la simulació de mostreig és bàsicament plana (com esperàvem, segons l'assumpció d'independència) i que hi ha diferències significatives entre les relacions observades i les corbes mitjanes de la simulació de mostreig (Coeficient de correlació d'Spearman per a la diferència entre observat i les corbes mitjanes de la simulació de mostreig a (A), (B) i (C) de 0.77, 0.89, -0.93, respectivament, amb p-valors ~0 en tots tres casos). Aquest resultat mostra que és

bastant poc probable que les tendències que s'han discutit prèviament siguin degudes a un efecte de mostreig.

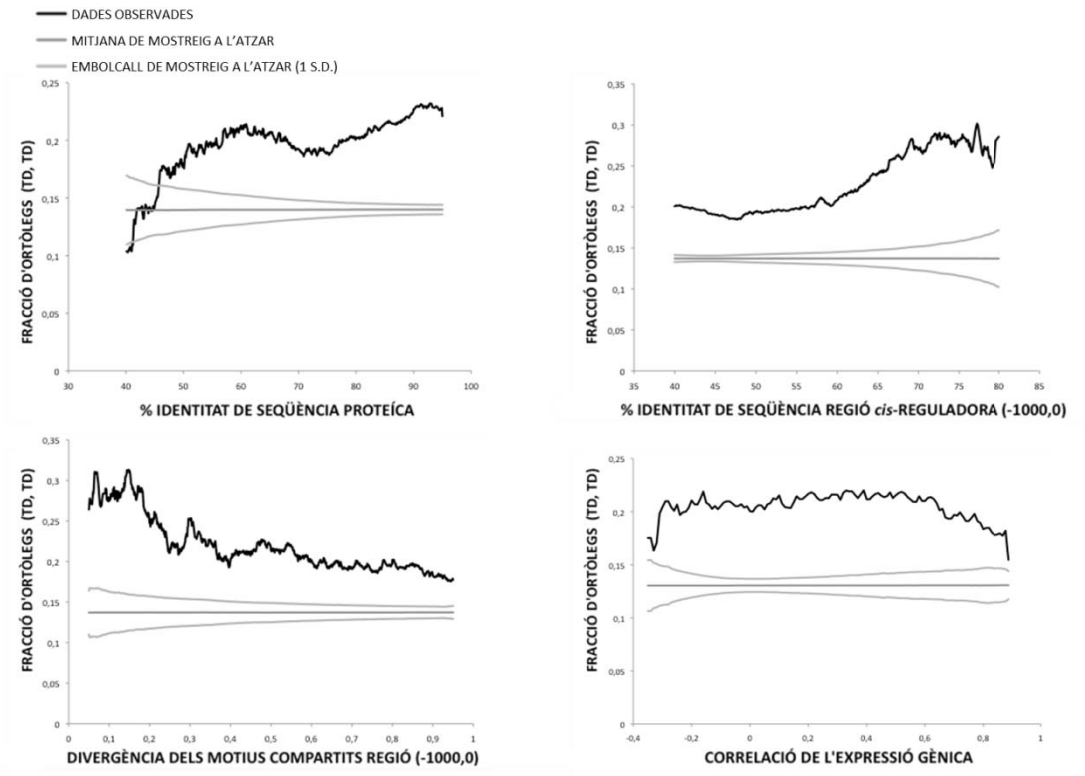

Figura 6. Exclusió d'un possible efecte de mostreig en els resultats sobre la concurrència de la TD. A fi i efecte de descartar la possibilitat que les tendències observades a les Figures 2-4 fossin degudes a un efecte de mostreig i que, en realitat, no n'hi hagués, de tendències, vam reproduir el resultat d'aquestes figures (a més a mes de la Figura 5) fent servir 100000 versions a l'atzar del nostre conjunt de dades original (veure *Material i mètodes*). En negre, mostrem els resultats observats (igualment mostrats en negre a les Figures 2-4); en gris fosc, tracem els resultats de la mitjana d'aquestes 100000 versions a l'atzar i, en gris clar, l'embolcall d'1 de la desviació estàndard de la mitjana. Els quatre dibuixos corresponen a les quatre variables treballades en aquest estudi: (A) la identitat de seqüència proteica; (B) i (C) la identitat de seqüència i divergència dels motius compartits de la regió *cis*-reguladora (concretament, 1000 parells de bases a 5' del TSS); i (D) la correlació de l'expressió gènica. Les diferències observades a (A), (B) i (C) entre les dades originals i la seva versió a l'atzar ens indiquen que les tendències de les Figures 2-4 no són, amb molta probabilitat, resultat de fluctuacions del mostreig.

Una quantitat indeterminada d'isoformes en el nostre conjunt de dades podria correspondre a TD d'origen genètic, deguda a diferències al·lèliques entre individus (Nembaware et al., 2004), o a soroll estocàstic (Melamud and Moult, 2009) o una combinació de tots dos. De nou, per excloure la possibilitat que aquesta variabilitat pogués introduir tendències no desitjades en els nostres resultats, vam realitzar 10000 simulacions, on a cadascuna de les quals un percentatge donat de TD era eliminat a l'atzar pel 15% dels gens de cada espècie (de manera independent a humà i ratolí, veure *Materials i Mètodes*); aquest valor

va ser escollit perquè coincidia amb el llindar superior de la variabilitat d'origen genètic proposat per Nembaware i col·laboradors (Nembaware et al., 2004). A continuació, vam recalcular els gràfics de les Figures 2-5. Els resultats obtinguts (Figura 7) foren similars a les dades originals (Figures 2-5) i, per tant, van revelar que les tendències originals no corresponien a un biaix degut a la TD intraespecífica, sinó més bé a diferències interespecífiques.

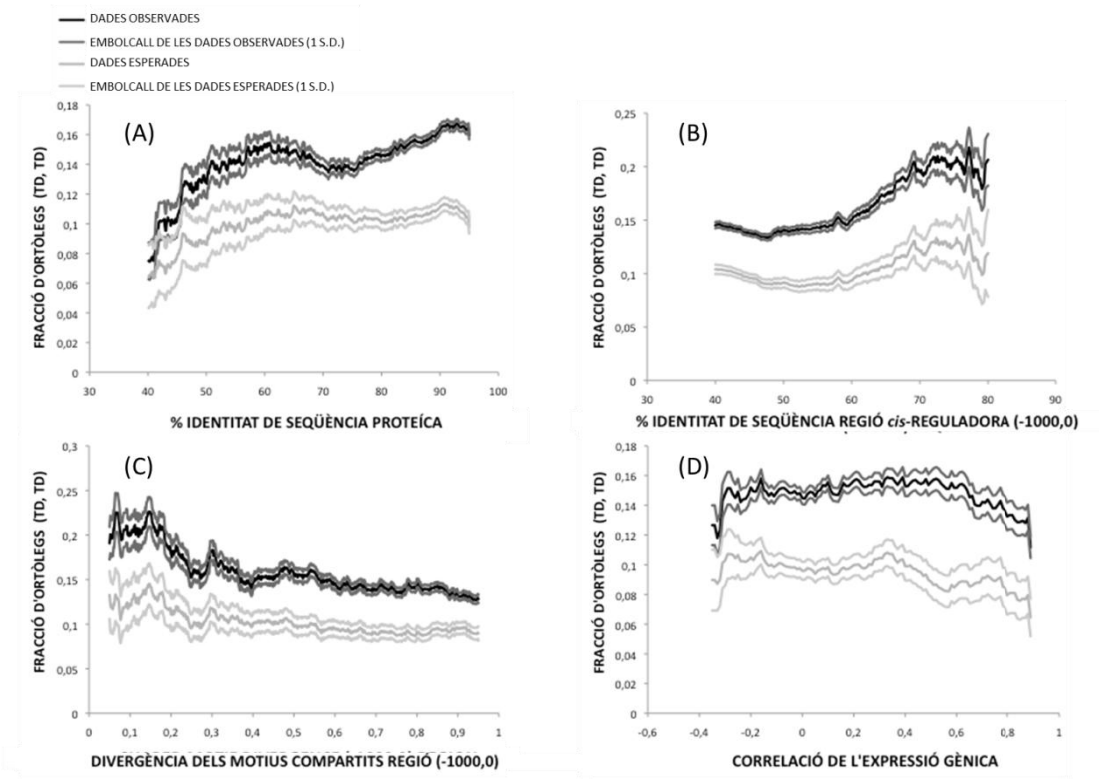

Figura 7. Inalterabilitat dels resultats sobre la concurrència de la TD respecte la variabilitat individual o estocàstica. Després d'eliminar a l'atzar TD del nostre conjunt de dades original, vam reproduir els resultats per a les gràfiques de la concurrència de la TD respecte les quatre variables estudiades en aquest treball: (A) identitat de seqüència proteica; (B) i (C) identitat de seqüència i divergència dels motius compartits de la regió *cis*-reguladora (en aquest cas, 1000 parells de bases a 5' del TSS); i (D) la correlació de l'expressió gènica. En aquest cas, també vam obtenir un embolcall d'1 de la desviació estàndard per als valors observats. Les tendències resultants són comparables a les que hem vist a les dades observades originals (Figures 2-5) i mostren que és poc probable que aquestes tendències originals fossin degudes a la variabilitat individual o estocàstica a nivell molecular.

3.3.3 Equivalència d'isoformes respecte DP, DcR i la divergència de l'expressió gènica

Per aquelles parelles d'ortòlegs en les quals tots dos gens tinguessin TD, vam comparar les isoformes alternatives per intentar esbrinar si l'equivalència

d'isoformes tenia algun tipus de relació amb la DP, la DcR i la divergència d'expressió gènica. La comparació d'isoformes es va dur a terme utilitzant el mètode de comparació basat en xarxes neuronals (Talavera et al., 2007a) tal com expliquem a la Secció "*3.2.4 Isoformes homòlogues*", a *Material i mètodes*.

No vam trobar cap tendència en els nostres resultats (Figura 8), és a dir, no hi havia cap relació significativa entre l'equivalència d'isoformes i la resta de variables (DP, DcR o correlació de l'expressió gènica). La solidesa d'aquest resultat va ser confirmada mitjançant l'aplicació del nostre protocol de comparació amb uns llindars de diferència de longitud del 0% i el 20%, sense trobar diferències €substancials (veure Figures addicionals 1 i 2).

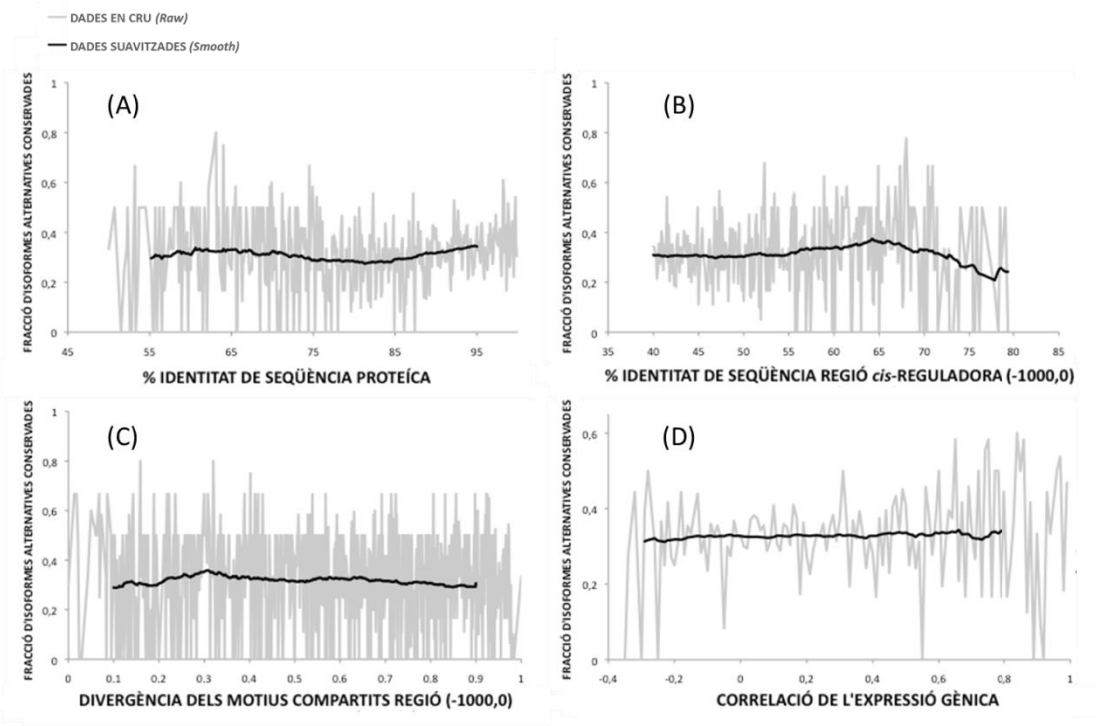

Figura 8. Relació entre l'equivalència d'isoformes i les divergències de proteïna, de la regió *cis*-reguladora i de l'expressió gènica. (A) DP (representada per la identitat de seqüència proteica), (B) i (C) DcR (representada per la identitat de seqüència i la divergència dels motius compartits) i (D) la divergència de l'expressió gènica (representada per la correlació de l'expressió gènica). Les dades en cru estan traçades en gris, i les dades suavitzades per finestra flotant estan representades en negre. No vam poder veure cap relació clara per cap dels quatre casos, ja que aquesta era molt lleu o inexistent.

3.4 Discussió

En els darrers anys, s'han aconseguit avanços significatius en la identificació de les fonts moleculars de la diversitat fenotípica (Castillo-Davis et al., 2004). A més a més d'aprofundir en el coneixement d'aquestes fonts, és necessari establir quines relacions hi ha entre elles (Copley, 2008). Per exemple, si actuen plegades o de manera independent, i fins a quin punt, en el procés de diferenciació de fenotips (Copley, 2008; Lemos et al., 2005); si hi ha algun grau d'equivalència entre elles, de manera que diferents fenòmens moleculars poden tenir un paper comparable o similar en diferents espècies (com ha estat suggerit justament per l'splicing alternatiu i la duplicació gènica (Kopelman et al., 2005; Su et al., 2006)); etc. Un grapat d'estudis recents han començat a clarificar aquest assumpte pel que fa a la DP, la DcR i la divergència en l'expressió gènica (Castillo-Davis et al., 2004; Chiba et al., 2008; Tirosh et al., 2008; Wall et al., 2005; Wang and Rekaya, 2009). Malgrat això, no hi havia cap estudi fins ara relacionant aquests fenòmens amb la TD. En aquest treball hem posat fil a l'agulla en aquest tema, tot explorant l'existència d'una relació entre, per una banda, la TD (descrita utilitzant dos mesures: la concurrència de TD i l'equivalència d'isoformes, Figura 1) i, per una altra banda, la DP, la DcR i l' expressió gènica.

Hem trobat que la concurrència de la TD està relacionada amb les altres fonts de diferències fenotípiques (Figures 2-4). És especialment interessant la relació entre la concurrència de TD i la DP (Figura 2, gràfic a l'esquerra) ja que, quan ho complementem amb informació addicional sobre la seqüència i funció proteica, ens permet distingir dos regions rellevants per a la diversitat fenotípica i el seu origen molecular. És ben conegut que la funció de la proteïna i la divergència de seqüència estan lligades, i que molts aspectes de la funcionalitat de la proteïna estan conservats per sobre d'un llindar d'identitat de seqüència, els quals són: 60-70% per la funció enzimàtica (Devos and Valencia, 2000; Rost, 2002; Tian and Skolnick, 2003); 30-40% per la geometria global de les interaccions proteiques (Aloy et al., 2003); 50% per l'estructura quaternària (Levy et al., 2008); per sobre del 65% per a la conservació de la parella proteica (Yu et al., 2004); i entre el 60 i 80% per una sèrie de propietats associades a funció (Devos and Valencia, 2001). Basant-nos en aquestes dades, podem definir un llindar d'identitat de seqüència d'entre el 50% i

el 70% per sota i per sobre del qual la funció proteica és diferent o conservada, respectivament. Això ens suggereix que, en general, la TD pot contribuir a les diferències interespecífiques, tot i que la DP no ho faci: podria ser el cas de gens situats en la regió de conservació de funció (identitat de seqüència proteica > 50-70%). Dades recents (Blekhman et al., 2010; Calarco et al., 2007) concorden amb aquesta idea, al menys pel que fa a espècies molt properes. En un treball experimental i comparatiu de l'splicing alternatiu a humà i ximpanzé, Calarco i col·laboradors (Calarco et al., 2007) trobaren exemples clars on els patrons d'splicing alternatiu poden variar entre ambdues espècies, i conclouen que probablement tinguin un impacte en les diferències fenotípiques. Per exemple, el gen de TAF6 (*TATA-box-binding protein-associated factor 6*) codifica per una isoforma a humà que no es troba a ximpanzé (Calarco et al., 2007). Per una altra banda, Blekhman i col·laboradors (Blekhman et al., 2010) arribaren a una conclusió similar quan van comparar els patrons d'expressió de transcrits a humà respecte els de ximpanzé i macaco rhesus (*Macaca mulatta*). En concret, varen observar que les isoformes curtes (les quals són generalment de naturalesa reguladora (Lois et al., 2007)) exclusives d'humà estan associades amb gens que participen en l'estructura anatòmica i en processos de morfogènesi (Blekhman et al., 2010). Degut a que les espècies estudiades tenen una similitud mitjana aminoacídica superior al 90% (Magness et al., 2005; Watanabe et al., 2004), clarament per sobre del llindar que ens indica divergència funcional (un valor proper al 50-70%, com ja hem explicat), tots aquests resultats indiquen que l'splicing alternatiu pot contribuir a les diferències fenotípiques entre humà i la resta de primats quan la identitat de seqüència codificant és massa elevada per introduir cap contribució real. Una situació semblant també ha estat descrita en un estudi comparatiu de l'splicing alternatiu en subespècies de *Mus musculus* (Harr and Turner, 2010), on la identitat de seqüència entre aquestes subespècies és, lògicament, molt alta.

També vam observar una relació entre la concurrència de TD i la DcR (Figures 3-4) i entre la concurrència de TD i la correlació de l'expressió gènica (Figura 5). Aquesta última és molt més feble que la primera, un fet que no ens hauria de sorprendre ja que és un fenotip regulat per a diversos factors (com pot ésser la presència/absència d'*enhancers*, l'estat de la cromatina, etc.) i, a més a més, se sap

que la qualitat de les dades sobre expressió, normalment baixa, afecta els estudis comparatius entre espècies (Liao and Zhang, 2006). Pel que fa a la DcR, la situació és lleugerament diferent (Figures 3-4): es pot observar una tendència molt més evident en la direcció esperada (creixent amb el percentatge d'identitat de seqüència i decreixent amb la divergència dels motius compartits). En aquest cas, però, i contràriament al que hem vist pel cas de la DP, la interpretació de les dades és limitada degut a la falta de llindars generals relacionats amb la divergència funcional. No obstant això, com en el cas de la DP, donat el fet que la concurrència de TD és baixa per conservacions altes de la regió *cis*-reguladora, podem especular que la TD pot contribuir a les diferències fenotípiques entre organismes quan la divergència a la regió *cis*-reguladora és encara massa baixa com per a poder tenir un impacte funcional.

A més a més de la concurrència de la TD, vam estudiar la relació entre fracció d'isoformes alternatives equivalents (Talavera et al., 2007a) i DP, DcR i correlació de l'expressió gènica. Els resultats obtinguts foren els mateixos per totes les comparacions (Figura 8): no hi ha cap relació entre l'equivalència d'isoformes i la resta de variables. Aquest resultat va d'alguna manera contra la intuïció que teníem, ja que esperàvem al menys una tendència entre l'equivalència d'isoformes i la DP, degut al solapament entre les regions codificadores i les senyals de l'splicing alternatiu (Fichant, 1992; Goren et al., 2006; Schaal and Maniatis, 1999; Smith and Valcarcel, 2000; Whamond and Thornton, 2006). Hi ha diversos factors que poden explicar aquesta contradicció. A continuació, els discutirem breument considerant aquells factors en els que estigui implicat l'splicing alternatiu, ja que és el contribuïdor a la TD més conegut. Primer de tot, les insercions/delecions i les substitucions, les quals són les principals propietats que determinen la similitud d'isoformes i, per tant, l'equivalència, depenen de l'estructura gènica, la qual està sovint ben conservada entre els ortòlegs d'humà i ratolí (Meyer and Durbin, 2004; Pavesi et al., 2008; Waterston et al., 2002). Amb tota seguretat, estarà conservada dins del rang de valors que hem escollit per aquest estudi. Per tant, no esperem cap variació destacada en l'equivalència d'isoformes en relació a aquestes variables. I en segon lloc, no podem descartar la importància que pot tenir una petita fracció d'errors en el nostre protocol de comparació d'isoformes. De totes maneres, és poc probable que aquestes alteressin substancialment la tendència real, ja que la regió

més propensa a cometre errors del nostre protocol (identitats de seqüència que voregin el 50%) és on menys ortòlegs humà-ratolí tenim. En resum, basant-nos en els nostres resultats (Figura 8) pensem que no hi ha cap tendència, o és lleu, entre l'equivalència d'isoformes i la DP, la DcR i la divergència en l'expressió gènica.

Donat que les isoformes poden realitzar multitud de papers reguladors, aquest resultat ens indica que les diferències en la naturalesa de les isoformes poden jugar algun rol en determinar diferències fenotípiques quan les altres variables mostren un nivell de divergència baix. Cal destacar el cas de la DP, per la qual només vam trobar equivalència d'isoformes parcial per sobre del llindar de divergència funcional (per identitats de seqüència proteiques superiors al 50%). Aquest resultat està en la línia, de nou, de les conclusions obtingudes en els estudis comparatius entre humà i primats (Blekhman et al., 2010; Calarco et al., 2007) i entre les subespècies de ratolí (Harr and Turner, 2010).

3.5 Conclusions

Aquest treball subministra, per primera volta, un model quantitatiu per a la relació entre la TD (centrant-nos en dos de les seves característiques: la concurrència de la TD i l'equivalència d'isoformes) i la DP i la DcR, i mostra com aquest model pot ser utilitzat per entendre les contribucions relatives d'aquests factors a la diversitat fenotípica entre espècies. En el primer bloc, hem exposat com la concurrència de TD varia d'acord amb la DP i la DcR; en concret, hem descrit com fins i tot a nivells de divergència no funcionals de la seqüència proteica i de la regió *cis*-reguladora, la concurrència de la TD introdueix una variació funcional potencial, és a dir, la TD pot estar ja contribuint a la diversitat fenotípica. Aquesta conclusió ve reforçada pel segon bloc, en el qual mostrem que l'equivalència d'isoformes no està relacionada, o de manera molt lleu, a la DP i la DcR. Això ens fa pensar que, abans que la divergència per proteïna i per la regió *cis*-reguladora amb implicació funcional pugui produir-se, trobem isoformes específiques a humà i ratolí que poden contribuir a les diferències entre organismes

3.6 Figures addicionals

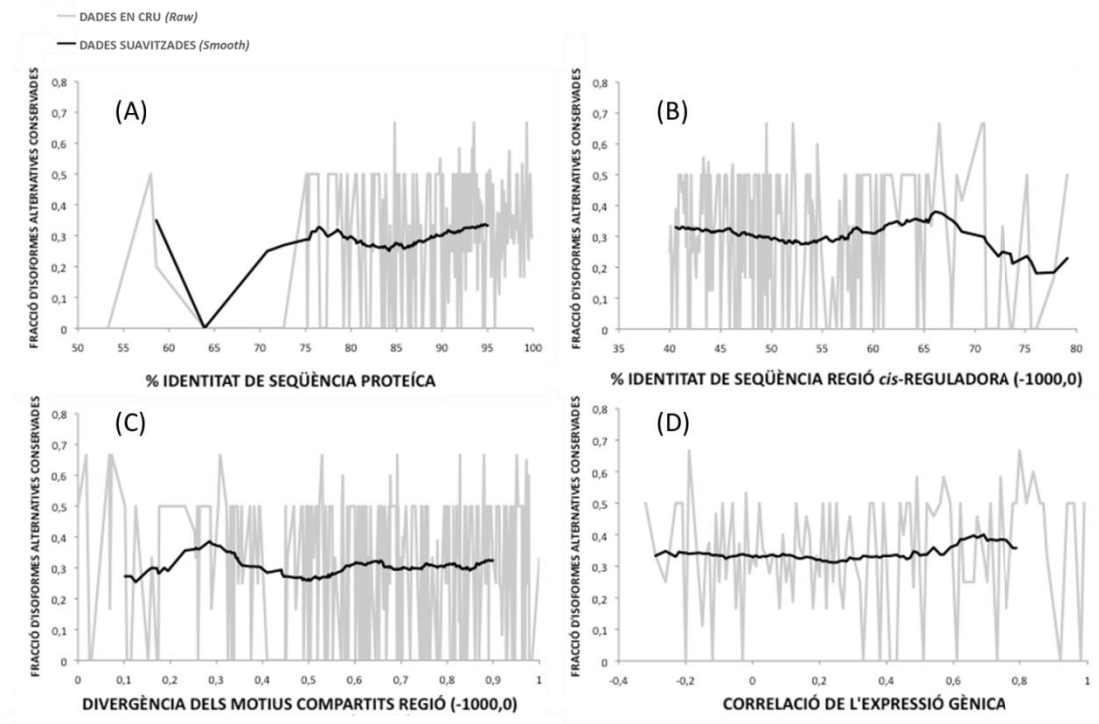

Figura Addicional 1. Relació entre l'equivalència d'isoformes i les divergències de la seqüència proteica, de la regió cis-reguladora i de l'expressió gènica. Aquesta figura és equivalent a la Figura 8 però en aquest cas les dades van ser obtingudes després d'imposar que no hi hagués diferències de mida entre les isoformes principals d'humà i ratolí. (A) DP (descrita per la identitat de seqüència proteica), (B) i (C) DcR (representada per la identitat de seqüència i la divergència dels motius compartits) i (D) la divergència de l'expressió gènica (representada per la correlació de l'expressió gènica). Les dades en cru estan representades en gris i les dades suavitzades per finestra flotant estan traçades en negre. No vam poder veure cap relació clara per cap dels quatre casos, ja que aquesta era molt lleu o inexistent.

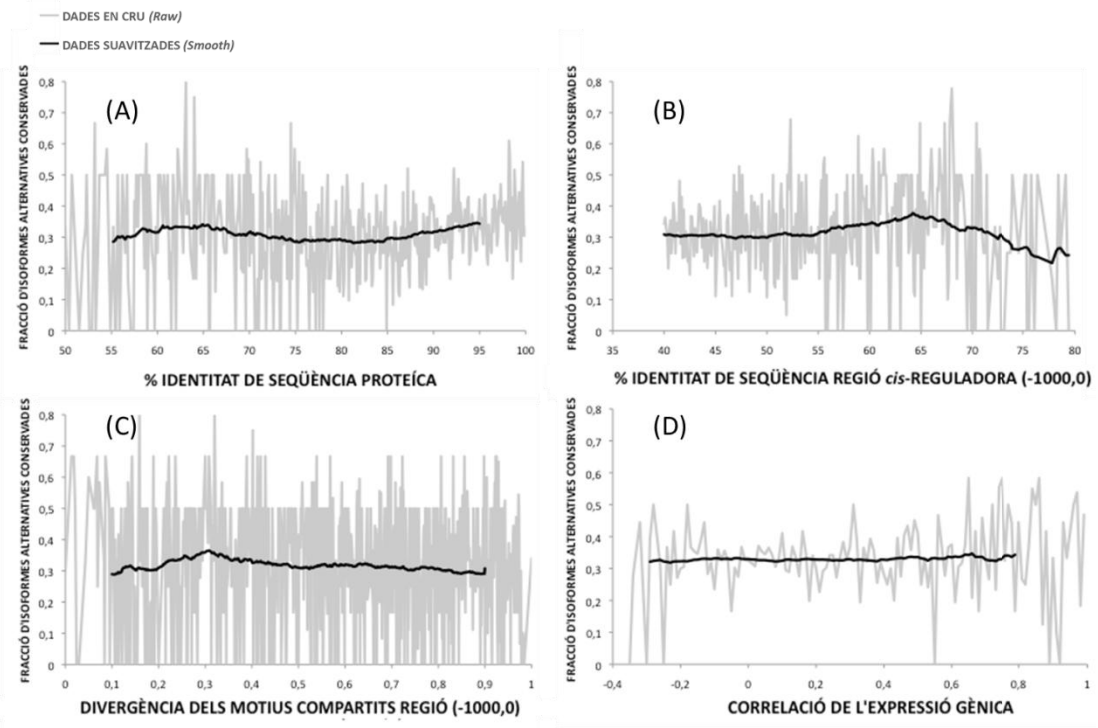

Figura Addicional 2. Relació entre l'equivalència d'isoformes i la divergència de proteïna, de la regió *cis*-reguladora i de l'expressió gènica. Aquesta figura és equivalent a la Figura 8 però, en aquest cas, les dades van ser obtingudes després d'imposar que la diferència entre la longitud de les isoformes principals d'humà i ratolí fossin inferiors al 20% de la mida mitjana. (A) DP (definida per la identitat de seqüència proteica), (B) i (C) DcR (representada per la identitat de seqüència i la divergència dels motius compartits) i (D) la divergència de l'expressió gènica (descrita per la correlació de l'expressió gènica). Les dades en cru estan representades en negre i les dades suavitzades per finestra flotant estan dibuixades en negre. No vam poder veure cap relació clara per cap dels quatre casos, ja que aquesta era molt lleu o inexistent.

4. L'impacte de l'splicing alternatiu en l'arquitectura de dominis a humà i ratolí: implicacions per a la conservació de la regulació de la funció proteica

4.1 Introducció

Tal com hem descrit a la introducció d'aquesta tesi, l'splicing alternatiu és una de les fonts més importants de diversitat proteica (Nilsen and Graveley, 2010). Té la capacitat de modificar la seqüència proteica de dos maneres diferents: mitjançant substitucions (on part de la seqüència proteica és substituïda per un fragment de seqüència de mida variable) i/o insercions/delecions (Indels) a localitzacions específiques de la seqüència (Ben-Dov et al., 2008; Kondrashov and Koonin, 2003; Valenzuela et al., 2004). Aquestes modificacions de la seqüència estan associades a un àmplia i variada sèrie d'efectes funcionals, que poden abastar des de modulacions subtils (com els canvis en l'especificitat de substrat (Hymowitz et al., 2003)) fins a mecanismes més dràstics de *switch on/off*, com la generació de dominants negatius (Barberan-Soler et al., 2009; Stamm et al., 2005).

L'splicing alternatiu aprofita la manera com està distribuïda la funció al llarg de la seqüència proteica. Les proteïnes tenen una estructura modular on els diferents dominis presents són els responsables de les activitats més diverses, com és l'activitat enzimàtica, les interaccions proteïna-proteïna o la unió al DNA, entre d'altres (Creighton, 1993)(Figura 1). La quantitat i la disposició d'aquests dominis al llarg de la seqüència proteica defineix l'arquitectura de dominis de la proteïna, la qual està tan íntimament relacionada amb la seva funció (Hegyi and Gerstein, 2001) que alguns mètodes de predicció de la funció proteica es basen completament en aquesta arquitectura i la seva conservació (Forslund and

Sonnhammer, 2008). L'splicing alternatiu, doncs, es beneficia de la modularitat dels dominis per generar diversitat funcional.

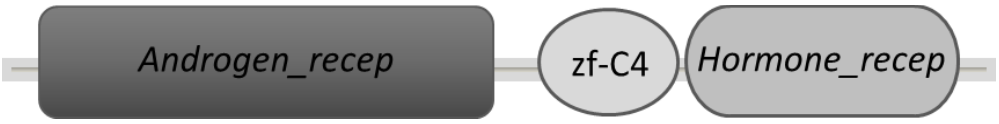

Figura 1: L'estructura modular de les proteïnes. En aquest exemple, veiem com s'organitzen els dominis del receptor d'andrògen (ANDR_HUMAN). Està format per un domini modulador a l'extrem N-ter, un domini d'unió al DNA (*zf-C4*) i un domini d'unió a lligand al C-ter. Una isoforma generada per AS s'expressa al cor i al múscul esquelètic i no conté el domini *"Androgen-recep"*. L'anotació de dominis ha estat obtinguda mitjançant Pfam.

En els darrers anys, una sèrie d'estudis ens han mostrat les diverses formes per les quals l'splicing alternatiu pot alterar l'arquitectura proteica (Garcia et al., 2004; Stetefeld and Ruegg, 2005; Xing et al., 2003). En una revisió sobre els factors de transcripció eucariotes, Latchman (Latchman, 1990) va descriure com els reguladors de la isoforma principal poden originar-se per la pèrdua del domini d'unió a lligand. En un treball posterior, López (Lopez, 1995) va descriure, de manera més extensa, com l'splicing alternatiu en factors de transcripció pot donar lloc a isoformes amb l'arquitectura de dominis modificada en relació a la isoforma principal. També va descriure la relació entre aquest canvi de seqüència i el rol funcional de la isoforma, tot recalcant el fet que algunes isoformes podrien actuar com a reguladors de la isoforma principal (com són, per exemple, els dominants negatius). Una generalització d'aquests resultats estesos cap a d'altres famílies proteiques es pot trobar en el treball de Kriventseva i col·laboradors (Kriventseva et al., 2003), en el qual van mostrar que l'splicing alternatiu tendeix tant a eliminar com a afectar greument els dominis de la proteïna, provocant així l'aparició de possibles isoformes reguladores. Liu i Altman (Liu and Altman, 2003) van observar que no tots els dominis proteics tenen la mateixa probabilitat de patir splicing: alguns dominis que suporten splicing d'una manera desproporcionada tenen tendència a formar part de proteïnes associades a processos com la diferenciació, el desenvolupament, etc. En un estudi sobre la composició proteica d'ENCODE, Tress i col·laboradors (Tress et al., 2007) van trobar molts exemples on l'splicing alternatiu introdueix grans canvis de seqüència, incloent alguns casos interessants

on els esdeveniments d'splicing exclouen dominis sencers. Encara que aquests autors anticipen la possibilitat que part d'aquest splicing alternatiu no tingui cap rol específic i que gaudeixin de la tolerància de la cèl·lula, també consideren que és possible que tinguin un paper regulador. Recentment, fent servir un conjunt de dades diferent als anteriors, Floris i col·laboradors (Floris et al., 2008) han observat que en molts casos les isoformes mostren arquitectures de dominis alterades, amb dominis que es troben absents de manera parcial o total. Aquests autors ens indiquen que aquestes isoformes estan relacionades amb un paper regulador de la isoforma principal com pot ser que, per exemple, les isoformes alternatives siguin antagonistes de l'efecte funcional de la isoforma principal.

En tots aquests estudis, encara que s'emfatitza en el coneixement de la naturalesa del canvis de seqüència i el seu efecte en les arquitectures de dominis, hi ha lloc per postular sobre el paper funcional de l'splicing alternatiu. És més, en molts d'aquests treballs es proposa un rol regulador de la funció gènica, en el sentit de control sobre les quantitats d'isoforma principal, per part de les isoformes alternatives. No obstant això, aquesta proposta rarament ha anat acompanyada d'una descripció del mecanisme que expliqui aquest paper regulador. Tal com va mostrar de manera sistemàtica López (Lopez, 1995; Lopez, 1998) i, més recentment, Talavera et al. (Talavera et al., 2009) en el cas dels factors de transcripció, per relacionar aquest paper regulador amb els canvis de seqüència introduïts per l'splicing és necessari tenir en compte les diferències en l'arquitectura de dominis entre les isoformes principal i alternativa, i la seva co-expressió. Pel que fa a l'extensa família dels enzims epigenètics, Lois i col·laboradors (Lois et al., 2007) van descriure diferents mecanismes reguladors relacionats amb l'splicing alternatiu: dominants negatius, efectes similars a la regulació del promotor, etc. Del treball de Stamm i col·laboradors també obtenim una visió de com els dominants negatius de la isoforma principal poden ser generats per la pèrdua de dominis (Stamm et al., 2005). Brenner i col·laboradors (Lareau et al., 2007) han defensat un paper regulador de l'splicing en combinació amb el NMD, en concret del mecanisme anomenat RUST (*Regulated Unproductive Splicing and Translation*). En resum, hi ha una evidència àmplia que ens mostra que l'splicing alternatiu pot tenir un rol regulador mitjançant l'alteració de l'arquitectura de dominis de la proteïna, però també assenyalant que els

mecanismes reguladors resultants poden ser diversos. Donades aquestes conclusions, combinades amb la rellevància de la regulació a l'hora d'establir les diferències de complexitat entre organismes (Carroll, 2005), hem decidit estudiar la conservació de l'splicing alternatiu entre humà i ratolí, tot centrant-nos en les propietats reguladores de les isoformes alternatives. Ens hem proposat estudiar tres punts principals: (i) la relació entre l'splicing alternatiu i l'expressió gènica; (ii) la conservació dels mecanismes reguladors basats en l'splicing alternatiu entre els ortòlegs d'humà i ratolí; (iii) la conservació d'esdeveniments específics d'splicing alternatiu regulador.

4.2 Material i mètodes

4.2.1 Obtenció del conjunt de dades

El conjunt de dades que es va fer servir en aquest treball prové del que s'ha descrit al Capítol 3. En aquest conjunt de dades, tenim 13970 parelles d'ortòlegs d'humà i ratolí, amb informació de tot tipus sobre identitat de seqüència proteica, presència d'AS i nombre d'isoformes, expressió gènica, etc. Per a més detalls de com es van obtenir aquestes dades, consulteu la secció de *Material i mètodes* del Capítol 3.

4.2.2 Definició i quantificació dels diferents tipus d'splicing alternatiu

A continuació, es descriu la classificació de les quatre classes d'splicing alternatiu que s'han definit per aquest treball (veure també la Figura 2) i es donen les dades completes corresponents a les Taula 1 i 2.

Classe 1: Sense dominis ni a la isoforma principal (MI) ni a la isoforma alternativa (AI). En aquesta classe, agrupem tots aquells casos en el quals no vam obtenir informació sobre dominis ni a la isoforma principal (MI, per les sigles en anglès) ni per la isoforma alternativa (AI, per les sigles en anglès). En aquest cas, no vam podem establir ni interpretar el paper regulador de les isoformes alternatives, degut a la falta d'informació sobre l'arquitectura de dominis de la proteïna i l'efecte conseqüent que introduiria l'AS sobre ella.

		Visió gènica		Visió d'isoforma	
Classe	*Sub-classe*	*Casos*	*%*	*Casos*	*%*
Classe 1		1263	16,31	2181	14,60
Classe 2	0-49%	2605	33,63	4502	30,14
	50-79%	1455	18,78	2208	14,78
	80-100%	556	7,18	764	5,11
Classe 3		1159	14,96	1619	10,84
Classe 4	1 domini perdut	1882	24,30	2601	17,41
	>1 domini perdut	812	10,48	1063	7,12
		-	-	14938	100

Taula 1: Resum de la classificació dels esdeveniments d'AS per humà. Es mostren els resultats organitzats segons dues visions: a la *visió d'isoforma* hem comparat les arquitectures de dominis entre totes les isoformes alternatives (AI) i la isoforma principal (MI) pel gen corresponent i les hem classificat segons els criteris esmentats. El nombre total i el percentatge es refereixen al nombre d'AI incloses en el conjunt de dades. Per la seva banda, la *visió gènica* resumeix la visió d'isoforma per gen: compta la presència de cada una de les classes per gen. El percentatge és respecte el nombre total de gens amb AS però la suma de casos no és igual al nombre de gens amb AS, ja que un mateix gen pot tenir parelles MI-AI de diferents classes. (Per exemple, en un gen amb 5 AI, si tres isoformes són de la Classe 2 i dues isoformes de la Classe 4, per la visió d'isoforma tindrem un 3 a la casella de la Classe 2 i un 2 a la casella de la Classe 4. Des de la visió gènica, tindrem un 1 a la casella Classe 2 i un 1 a la casella Classe 4, ja que només es compta la presència en aquell gen d'alguna de les classes, no importa quants cops).

		Visió gènica		Visió d'isoforma	
Classe	*Sub-classe*	*Casos*	*%*	*Casos*	*%*
Classe 1		666	13,75	1048	13,02
Classe 2	0-49%	1798	37,13	2783	34,57
	50-79%	898	18,54	1238	15,38
	80-100%	375	7,74	479	5,95
Classe 3		604	12,47	751	9,33
Classe 4	1 domini perdut	929	19,18	1217	15,12
	>1 domini perdut	429	8,86	535	6,65
		-	-	8051	100

Taula 2: Resum de la classificació dels esdeveniments d'AS per ratolí. Aquesta taula és l'equivalent de la taula 1 però amb dades de ratolí. Veure Taula 1 per a més detalls.

Classe 2: Variants de la mateixa funció. En aquest cas, tant la isoforma principal com l'alternativa presenten la mateixa estructura de dominis. En aquesta categoria la naturalesa reguladora de l'AS és substancialment diferent que a la resta de classes ja que l'AS probablement introdueixi canvis que modulin subtilment la

funció, mitjançant la modificació de les cues dels dominis o de les zones interdominis. És a dir, tant la isoforma principal com l'alternativa exploren regions properes de l'espai funcional proteic. Per assignar un esdeveniment d'AS a aquesta classe, vam emprar el següent protocol: (i) calcular l'arquitectura de dominis de les isoformes principal i alternativa; (ii) comprovar que totes dues isoformes tinguessin exactament el mateix nombre d'elements funcionals (dominis, hèlixs transmembrana, pèptids senyal i dominis desordenats) i disposats en el mateix ordre. Degut a que en aquesta classe la interpretació funcional depèn de la cobertura de dominis funcionals al llarg de la seqüència proteica, vam dividir aquesta classe en tres subclasses d'acord a la fracció de residus que pertanyen a algun domini funcional: entre el 0% i el 50%, entre el 50% i el 80% i, finalment, entre el 80% i el 100%. A la primera subclasse, la fracció de dominis és tan baixa que impedeix un anàlisi funcional fiable de l'AS, basant-nos únicament en l'arquitectura de dominis. La segona la situació és millor i, finalment, a la tercera subclasse l'anàlisi funcional esdevé molt fiable, degut a que pràcticament tota la proteïna pertany a un domini funcional.

Classe 3: Aparició de nous dominis. Aquesta classe inclou aquells casos on l'AS ha modificat l'arquitectura de dominis de la isoforma principal mitjançant la introducció de nous dominis, la reordenació de dominis ja existents o degut a la introducció de petits canvis de seqüència però amb rellevància funcional (hèlixs transmembrana, pèptids senyal o regions desordenades). A l'hora de col·locar un esdeveniment d'splicing en aquesta classe, vam utilitzar el protocol següent: (i) calcular l'arquitectura de dominis de les isoformes principal i alternativa; (ii) comprovar que la isoforma alternativa presentés algun element funcional que no aparegués a la isoforma principal. Aquest element podia ser un nou domini, una hèlix transmembrana, una pèptid senyal diferent o una regió desordenada

Classe 4: Pèrdua de dominis coneguts. Aquesta classe inclou aquells casos on l'AS ha modificat l'arquitectura de dominis de la isoforma principal mitjançant l'eliminació de domini(s) específic(s). Aquesta classe és especialment interessat degut a que inclou el mecanisme àmpliament conegut dels reguladors dominants negatius. Hem dividit aquesta classes en dues sub-classes: (a) quan la pèrdua és de

només un domini i (b) quan diversos dominis s'eliminen. A l'hora de destinar un esdeveniment d'AS a aquesta classe, vam emprar el següent protocol: (i) calcular l'arquitectura de dominis de les isoformes principal i alternativa; (ii) Comprovar si la isoforma alternativa ha perdut un o més dominis en referència a la isoforma principal; (iii) Verificar que cap element funcional nou ha aparegut a la isoforma alternativa, com hèlixs transmembranes, dominis desordenats i/o pèptids senyals, i els que restin, mantinguin el mateix ordre que a la isoforma principal. Si aquest fos el cas, aquest cas concret seria recol·locat a la classe 3.

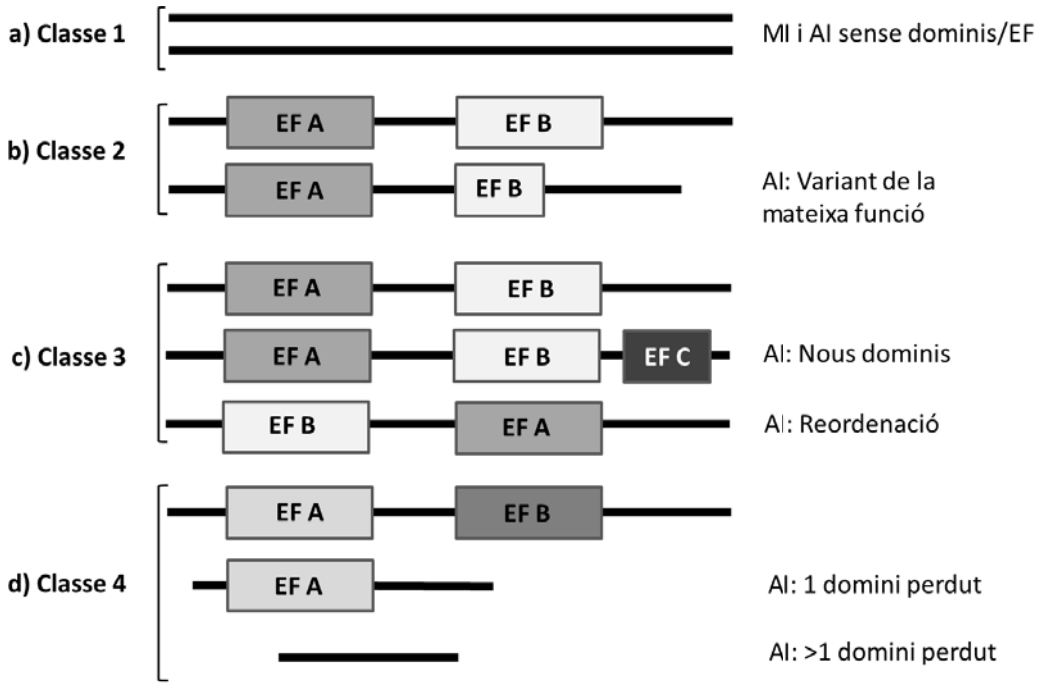

Figura 2: Classificació dels esdeveniments d'splicing alternatiu. S'esquematitzen les quatre classes que s'han descrit al text. Per claredat, no s'han mostrat les tres subclasses de la Classe 2. Per a que un domini es consideri absent, ha de tenir en aquell domini menys del 50% de residus que tenia el domini corresponent a la MI. EF: element funcional (dominis, hèlix transmembrana, regió desordenada, pèptid senyal); MI: isoforma principal; AI: isoforma alternativa.

4.2.3 Identificació d'elements funcionals

4.2.3.1 Assignació de dominis

A fi d'establir l'arquitectura de dominis de totes les isoformes, vam utilitzar el programa Cd-Search (CDD versió 2.22, 27-05-2010) (Marchler-Bauer and Bryant, 2004), a partir del qual vam recuperar la informació de dominis que provinguessin de la base de dades Pfam (Finn et al., 2010). Del dominis predits resultants, vam resoldre els solapaments de dominis i vam considerar absents aquells dominis que,

a la isoforma alternativa, han perdut un 50% o més dels residus que tenen a la isoforma principal. A la fi del procés, vam obtenir l'anotació de dominis Pfam i el seu ordre respectiu dins de totes les isoformes.

4.2.3.2 Hèlixs transmembrana

Vam emprar el programa Memsat3 (Jones, 2007) que ens permet predir quins residus estan implicats en hèlixs transmembrana i, a més a més, ens dona la topologia general més probable. Només vam tenir en compte les hèlixs transmembrana predites que tinguessin una fiabilitat superior a 2, que garanteix una elevada probabilitat d'encert, que fou del 80% en el test de referència emprat pels seus programadors (Jones, 2007). En els casos que el nombre d'hèlixs transmembrana entre la isoforma principal i alternativa fossin diferents, vam analitzar manualment si l'splicing alternatiu estava implicat en la introducció d'aquesta diferència.

4.2.3.3 Pèptids senyal

Per identificar el residus implicats en algun pèptid senyal, vam utilitzar el programa TargetP (Emanuelsson et al., 2007), el qual ens indicà la localització subcel·lular de les proteïnes. Aquest programa proporciona un valor de fiabilitat que vam considerar només en el cas que fos 1, és a dir, quan la predicció és més restrictiva i fiable.

4.2.3.4 Regions desordenades

La identificació de residus desordenats es va realitzar mitjançant el programa Disopred2 (Ward et al., 2004). En aquest cas, només vam considerar les prediccions amb una fiabilitat igual o superior a 5. Vam considerar només els casos on la isoforma alternativa presentés més d'un 80% de residus desordenats, sempre i quan aquest valor fos més elevat que el percentatge de residus desordenats a la isoforma principal. En els casos que així va succeir, vam analitzar manualment si l'splicing alternatiu n'estava implicat.

4.2.4 Comparació dels esdeveniments d'splicing alternatiu

Tal com s'ha descrit al Capítol 3, vam comparar els esdeveniments d'splicing mitjançant el nostre software Splash (Talavera et al., 2007a), basat en xarxes neuronals. Aquest procediment ens indica si un parell d'esdeveniments d'splicing

són equivalents (o homòlegs) i ens dona un índex de fiabilitat d'aquest resultat. La comparació no sempre es pot dur a terme, per causes tècniques d'origen de les dades (per exemple, quan totes dues isoformes provenien de l'anotació VARSPLIC de SwissProt) o per diferències en la mida de les isoformes principal i alternativa. En el cas dels esdeveniments d'splicing que generen fenòmens de la Classe 4, vam afegir a aquest protocol un procediment semiautomàtic per recuperar el màxim de parelles d'isoformes principal i alternativa per comparar amb Splash. En aquest cas no ens vam cenyir a les anotacions de SwissProt sobre splicing, sinó que vam generar nosaltres mateixos, sempre que vam poder, aquestes anotacions mitjançant alineaments amb ClustalW (Thompson et al., 1994) seguides d'una inspecció manual. Per a més detalls, consultar la secció "*3.2.4 Isoformes homòlogues*" de la secció de *Material i mètodes* del Capítol 3.

4.2.5 Dades d'expressió gènica

Aquest protocol i les dades obtingudes ja han estat descrites al Capítol 3 (secció "*3.2.8 Càlcul de la correlació de l'expressió gènica*", dins de *Material i mètodes*). A continuació, tornem a resumir el protocol breument:

Les dades d'expressió gènica es van obtenir del servidor de biogps (http://biogps.gnf.org/): les dades per humà i ratolí van ser U133A/GNF1H i GNF1M, respectivament. En cadascun d'ells, cada gen està caracteritzat per la seva expressió a 26 teixits. Pel que fa a la divergència de l'expressió gènica entre ortòlegs d'humà i ratolí, vam seguir Liao i Zhang (Liao and Zhang, 2006), tot calculant el coeficient de correlació de Pearson entre els perfils d'expressió tissular respectius per a cada parell d'ortòleg. Recordar que el càlcul només es va dur a terme en aquells ortòlegs amb informació disponible de l'expressió gènica.

4.3 RESULTATS

4.3.1. Desacoblament entre l'AS i els nivells d'expressió gènica

Donat el fet que l'AS és un ferm candidat a tenir un paper regulador en les quantitats de producte gènic (Floris et al., 2008; Talavera et al., 2009; Tress et al., 2007), vam explorar si hi havia cap relació entre AS i els nivells d'expressió gènica. Per assolir aquest objectiu, vam seguir la mateixa aproximació que ja van emprar

Talavera i col·laboradors (Talavera et al., 2009), en la que dividiren la població de parelles d'ortòlegs en quatre subgrups, depenent si els gens humans i/o els gens murins presentaven AS: (AS,AS), (AS,noAS), (noAS,AS) i (noAS,noAS). Quan vam representar l'histograma de freqüències de la correlació de l'expressió gènica per cadascun d'aquests quatre grups, no vam trobar cap diferència significativa entre ells (Figura 3, X^2 = 0,12; p-valor > 0,05). Així, aquesta similitud entre les distribucions sembla indicar-nos un desacoblament entre l'AS i l'expressió gènica.

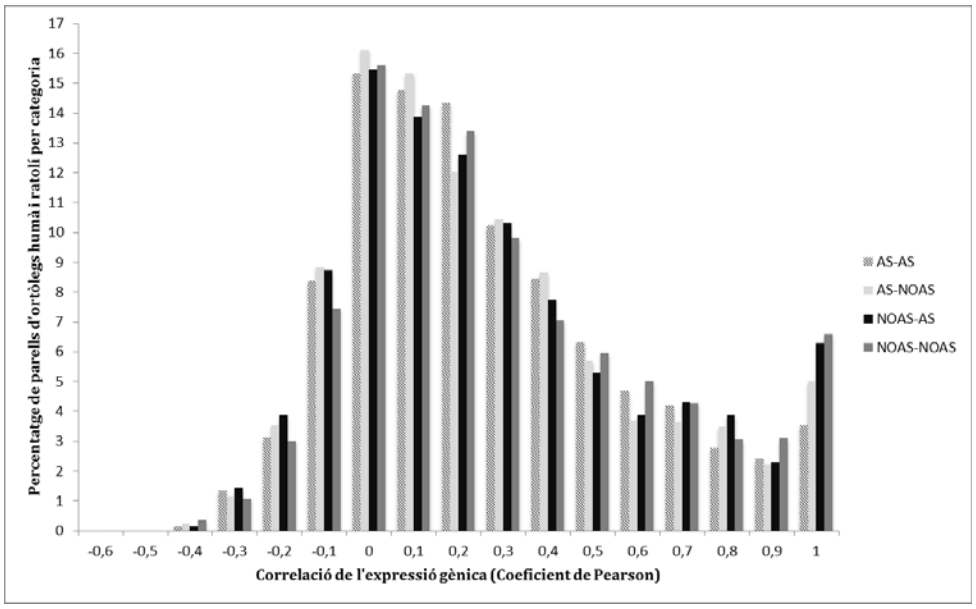

Figura 3: Comparació de l'expressió gènica entre ortòlegs d'humà i ratolí. Es compara el coeficient de correlació de Pearson de l'expressió gènica pels quatre grups definits: (AS,AS), (AS,noAS), (noAS,AS), (noAS, noAS), per humà i ratolí, respectivament.

4.3.2 Conservació global dels mecanismes reguladors d'AS entre humà i ratolí

A la Taula 3 mostrem els resultats de l'experiment de comparació dels mecanismes reguladors basats en AS entre 2914 parelles d'ortòlegs d'humà i ratolí. L'objectiu d'aquest experiment és establir el grau de conservació en els mecanismes de regulació entre humà i ratolí. Per a cadascuna de les parelles d'ortòlegs disponibles, vam comparar si totes dues espècies presentaven els mateixos o diferents mecanismes reguladors. Per exemple, comprovàrem si tots dos ortòlegs expressaven isoformes alternatives que es poguessin comportar com dominants

negatius (o el que és el mateix, si totes dues isoformes per humà i ratolí estaven incloses dins de la Classe 4), etc.

Els nostres resultats mostren que hi ha uns nivells de conservació desiguals entre els diferents mecanismes reguladors basats en AS. Per les tres subclasses de la Classe 2, *Variants de la mateixa funció*, la conservació es situa per sobre del 60% tant en humà com en ratolí (65,7% i 61,7% respectivament); això és particularment valuós en el cas de la Classe 2.3 (on s'assoleix un grau de conservació del 60,9% i 64,5%, en humà i ratolí, respectivament), degut a que els canvis es poden interpretar funcionalment amb més facilitat. Pel que fa a la Classe 4, *Pèrdua de dominis coneguts*, el percentatge de conservació no arriba al 50% (44,8% i 48,6% per humà i ratolí, respectivament). Finalment, la Classe 3, *Nous dominis*, és la menys conservada de totes, amb percentatges del 29,1% i 27,4% per humà i ratolí, respectivament. Vam excloure la Classe 1, *Sense dominis a MI/AI*, d'aquest anàlisi ja que, degut a la falta d'anotacions funcionals, ens va ser impossible proposar cap explicació sobre el mecanisme regulador en qüestió.

		Ratolí							
		Classe 1	*Classe 2.1*	*Classe 2.2*	*Classe 2.3*	*Classe 3*	*Classe 4.1*	*Classe 4.2*	
Humà	*Classe 1*	385	20	3	1	9	12	3	433
	Classe 2.1 (<50%)	19	913	30	2	124	264	116	1468
	Classe 2.2 (50-79%)	0	31	366	21	67	89	49	623
	Classe 2.3 (80-100%)	0	3	11	131	27	24	7	203
	Classe 3	14	145	78	18	150	95	47	547
	Classe 4.1 (1 domini perdut)	13	332	105	37	104	304	95	990
	Classe 4.2 (>1 domini perdut)	1	147	45	5	34	115	153	500
		432	1591	638	215	515	903	470	-

Taula 3: Resum de les dades obtingudes en la comparació entre ortòlegs dels mecanismes reguladors basats en AS. Les zones enfosquides corresponen als esdeveniments conservats entre humà i ratolí

4.3.3. Conservació dels esdeveniments de la Classe 4 (*Pèrdua de dominis coneguts*)

Un aspecte interessant d'aquest treball fou cercar els esdeveniments equivalents d'AS (és a dir, processos d'AS homòlegs entre espècies) entre ortòlegs d'humà i ratolí. Vam concentrar el nostre anàlisi en la Classe 4, *Pèrdua de dominis*, ja que se li ha assignat un paper regulador.

Vam poder analitzar amb Splash un elevat nombre de les comparacions d'esdeveniments d'AS amb patró de dominis de la Classe 4 tant per humà (3093 de 3482 possibles, 90%) com per ratolí (3067 de 3461 possibles, 88,6%) , gràcies a un procediment semiautomàtic, dissenyat a fi i efecte de poder recuperar el màxim nombre de casos (veure secció "*4.2.4 Comparació dels esdeveniments d'splicing alternatiu*" a *Material i mètodes*). Pel que fa a cada espècie, a la Taula 4 es resumeixen el percentatge d'esdeveniments de la Classe 4 que van ser equivalents amb algun esdeveniment d'splicing (de totes les Classes estudiades) de l'altra espècie. La fracció d'isoformes de la Classe 4 amb esdeveniments d'AS equivalents és lleugerament superior per ratolí (32,3%), així com la proporció d'esdeveniments que incloïen fenòmens de la Classe 4 amb només la pèrdua d'un domini (26,7% i 34,2% per humà i ratolí, respectivament). Tot i això, la conservació de l'splicing és, en general, baixa per tots els casos de la Classe 4.

	% Equivalents 1 domini perdut	*% Equivalents >1 domini perdut*	*% Equivalents Total*
Humà Classe 4	26,7%	20,1%	24,3%
(Ratolí qualsevol Classe)	(525/1966)	(227/1127)	(752/3093)
Ratolí Classe 4	34,2%	28,7%	32,3%
(Humà qualsevol Classe)	(685/2005)	(305/1062)	(990/3067)

Taula 4: Fracció d'esdeveniments d'AS equivalents de la Classe 4 per espècie. Anàlisi de la conservació de l'AS on un dels ortòlegs sigui de la Classe 4. Entre parèntesi, el nombre d'esdeveniments d'AS equivalents respecte el nombre total de casos que es van poder analitzar per cadascuna de les espècies que van ser, en total, 3093 i 3067 per humà i ratolí respectivament.

A continuació, vam reduir aquest mateix anàlisi als casos comuns de Classe 4, és a dir, als casos en que totes dues espècies havien de presentar un fenomen de la Classe 4. D'aquesta manera, vam poder estudiar l'equivalència/no equivalència entre espècies del mecanisme d'AS que donava lloc a la pèrdua de dominis. A més a més, vam limitar l'estudi a aquells casos on aquestes isoformes alternatives de la Classe 4 per humà i ratolí haguessin perdut exactament els mateixos dominis, independentment de la composició de dominis que tingués a cada espècie la isoforma principal (en unes altres paraules, que les isoformes principals

respectives presentessin diferent arquitectura de dominis no era important: el que sí que ho era fou que les isoformes alternatives manquessin dels mateixos dominis). Amb aquest conjunt de casos, vam calcular l'equivalència/no equivalència dels esdeveniments d'AS.

Vam obtenir 735 parelles d'esdeveniments d'AS (MI-AI)(MI-AI) amb patrons de la Classe 4 comuns, que correspongueren a un total de 333 parelles d'ortòlegs. Des del punt de vista d'isoforma, 350 esdeveniments d'AS (47,6%) van resultar ser equivalents pel que fa al mecanisme d'AS. Per a millorar la qualitat d'aquestes prediccions, vam considerar només les casos amb fiabilitat igual o superior a 5. Pel cas dels esdeveniments d'AS equivalents, 182 casos (52%) complien aquesta condició. Per una altra banda, 300 esdeveniments (40,8%) van ser considerats no equivalents pel que fa al mecanisme d'AS, tot i que el resultat a nivell de l'arquitectura de dominis fos el mateix per ambdues espècies. A la Taula 5 es resumeixen aquestes dades.

	Esdeveniments d'AS	*%*	*Fiabilitat >5*	*% Fiabilitat >5*
Equivalents	350	47,6%	182	52%
No equivalents	300	40,8%	285	95%
Sense resultat	85	11,6%	-	-
	735	100%	467	71,8%

Taula 5: Resum de les comparacions d'esdeveniments d'AS amb Classe 4 comuna des del punt de vista d'isoforma. Les dades estan referides al total de les comparacions possibles, 735, que corresponien a 333 grups d'ortòlegs.

A la Taula 6 es mostren aquests mateixos resultats però des del punt de vista de gen o ortòleg. Per a tal fet, vam cercar per a cada un dels 333 ortòlegs la presència d'equivalències/no equivalències entre esdeveniments de la Classe 4 i una fiabilitat màxima en la comparació SPLASH. El percentatge de ortòlegs amb algun esdeveniment d'AS equivalent s'enfila fins el 53% però, tot i això, no podem dir que els processos d'AS amb els quals s'obtenen fenòmens de Classe 4 estiguin especialment conservats entre humà i ratolí. A més a més, s'ha de tenir en compte

que no deixen de ser uns resultats modestos tenint en compte els valors de partida (13970 ortòlegs).

	Ortòlegs	*%*	*Fiabilitat >5*	*% Fiabilitat >5*
Equivalents	178	53,4%	100	56,2%
No equivalents	124	37,2%	119	96%
Sense resultat	31	9,3%	-	-
	333	100%	219	72,5%

Taula 6: Resum de les comparacions d'esdeveniments d'AS amb Classe 4 comuna per ortòlegs (visió de gen). En aquests cas, per cada parella d'ortòlegs es quantifica la presència d'algun esdeveniment d'AS equivalent i de màxima fiabilitat i, si no hi hagués cap, la presència de l'esdeveniment d'AS no equivalent amb fiabilitat més alta. Per diversos motius, no vam obtenir resultat d'equivalència d'AS a la categoria "*Sense resultat*" (veure *Material i mètodes*, secció 4.2.4 "*Conservació dels esdeveniments d'splicing alternatiu"*)

4.4 DISCUSSIÓ

En el present capítol ens hem centrat en l'estudi de la conservació de l'splicing alternatiu, entre humà i ratolí, des del punt de vista de les propietats reguladores de les isoformes alternatives. Els tres blocs que l'han constituït han estat: (i) la (in)dependència entre l'AS i l'expressió gènica; (ii) la conservació dels mecanismes reguladors basats en AS entre els ortòlegs d'humà i ratolí; i (iii) la conservació d'esdeveniments específics d'splicing alternatiu regulador (Classe 4).

Pel que fa al primer bloc, els nostres resultats ens indiquen que l'AS i l'expressió gènica actuen de manera independent (Figura 3). Així, l'AS tindria una funció complementària a la regulació al promotor (regió *cis*-reguladora) de l'expressió gènica, una font ben coneguda de diferències fenotípiques (Carroll, 2005), l'efecte del qual seria anàleg però en un nivell diferent. Al Capítol 3 (Figures 5 i 8) ja hem pogut observar la baixa relació obtinguda entre la concurrència de la diversitat transcripcional (i la fracció d'isoformes homòlogues) i la divergència de l'expressió gènica entre humà i ratolí. Resultats previs (Talavera et al., 2009) ja van assenyalar cap a la independència d'aquests dos fenòmens (AS i expressió gènica) pel cas de 559 factors de transcripció. Anteriorment, s'havia postulat aquesta independència (Pan et al., 2004) en els programes d'expressió que dirigeixen la determinació i diferenciació tissular. Així doncs, aquestes dades afegides als nostres resultats, ens

confirmen que l'AS i la regulació clàssica del promotor operen a diferent nivell i de manera independent i, per tant, l'AS pot introduir diferències fenotípiques pròpies entre teixits, individus o espècies. El següent pas lògic d'aquest raonament va ser preguntar-se, donat el fet que l'AS sembla tenir un rol regulador, quina proporció d'aquests mecanismes reguladors basats en AS estan conservats entre humà i ratolí.

Per respondre aquesta qüestió, en primer lloc vam fixar-nos en els efectes que té l'AS a nivell de la arquitectura de dominis de les isoformes i fins a quin punt aquestes alteracions estan conservades entre espècies. Després de definir els diferents tipus o classes de fenòmens que es poden produir (veure secció "*4.2.2 Definició i quantificació dels diferent tipus d'splicing alternatiu*"), vam observar que no tots els mecanismes reguladors estan conservats de la mateixa manera, fet que permetria a l'AS introduir diferències de complexitat entre humà i ratolí (Taula 3). La Classe 2 (*Variants de la mateixa funció*) va resultar ser la més conservada (valors al voltant del 60%). És en aquests casos que l'AS modularia l'expressió mitjançant petites modificacions dels extrems dels dominis i les regions interdominis, sense alterar dràsticament la funcionalitat de la proteïna. La Classe 3 (*Nous dominis*) va ser la menys conservada (valors inferiors al 30%) però aquí és necessària una puntualització. Aquesta categoria pot incloure artefactes provinents d'una anotació incorrecta de dominis Pfam en casos on aquesta predicció hagués de distingir entre diverses arquitectures de dominis amb fiabilitats molt properes. Pel que fa a la darrera classe, la Classe 4 (*Pèrdua de dominis*), presentà una conservació inferior al 50% entre els ortòlegs d'humà i ratolí però, tot i ser la categoria més restrictiva en les seves condicions, vam poder obtenir un nombre destacat de casos. Cal assenyalar que tot aquest anàlisi ha estat limitat per la cobertura de casos d'splicing, és a dir, d'isoformes definides per gen, i segurament una millor cobertura podria modular la descripció que hem proporcionat.

Finalment, vam centrar-nos en la Classe 4 (*Pèrdua de dominis coneguts*) per desenvolupar un estudi més exhaustiu de la conservació dels esdeveniments d'splicing entre espècies, no només de l'alteració del patró de dominis, sinó també de la conservació. Vam escollir aquesta categoria ja que és un mecanisme

regulador que ha estat ben documentat i establert experimentalment (Gamba, 2001; Janicke et al., 2009; Stamm et al., 2005; van der Vaart and Schaaf, 2009). Els resultats (Taules 4-6) ens indiquen que els esdeveniments d'splicing que duen a la pèrdua de dominis no estan gaire conservats, tot i que els esdeveniments equivalents siguin lleugerament superiors. D'aquest resultat podem extreure diverses explicacions. Un primer raonament seria de tipus tècnic: aquestes diferències només respondrien a una mostra molt petita del total de casos possibles (333 ortòlegs d'un total de 13970) i seria necessari una cobertura més extensa d'AS i de fenòmens de pèrdua de dominis per poder extreure alguna conclusió més sòlida. Un segon enfocament seria que aquestes diferències de seqüència provoquen alteracions en el patró de dominis i, per tant, impliquen divergència funcional. Així doncs, la baixa conservació dels esdeveniments d'AS significarien una introducció de diferències fenotípiques i/o complexitat entre espècies.

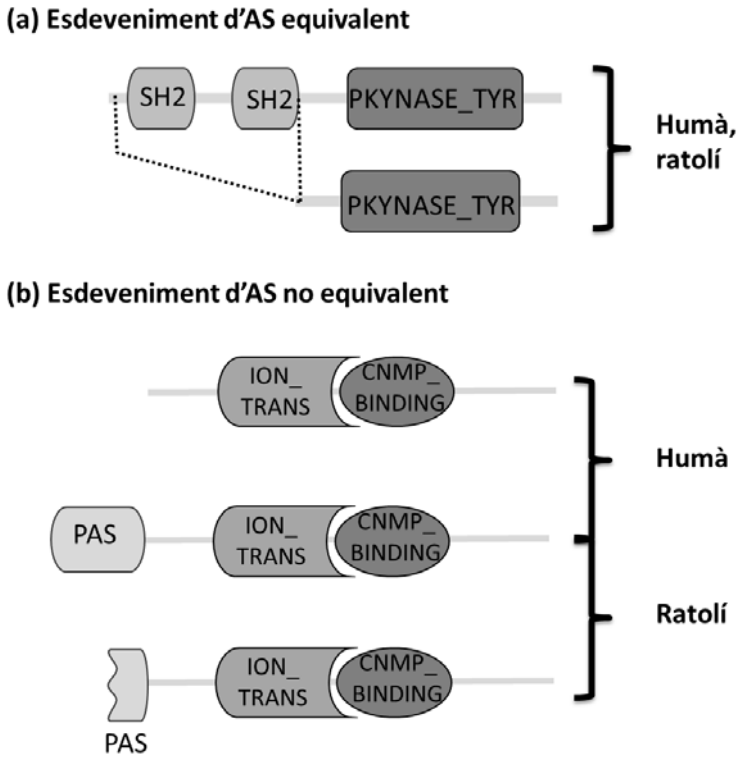

Figura 4: AS i el seu impacte en l'arquitectura de dominis. (a) L'ortòleg d'humà i ratolí per la quinasa ZAP-70 presenta un esdeveniment d'AS equivalent. El canvi de seqüència (deleció de dos dominis SH2) i l'alteració de l'arquitectura de dominis és la mateixa, i per tant, esperem el mateix impacte funcional en totes dues espècies. (b) L'esdeveniment d'AS pels ortòlegs de la *potassium voltage-gated channel subfamily H member 2* és no equivalent: l'alteració de seqüència és diferent. No obstant, l'impacte a nivell de dominis és comparable, ja que tant a humà com a ratolí el domini PAS es perd. Aquest pot ser un exemple de convergència funcional, on la funcionalitat de les dues isoformes d'humà i ratolí és similar, tot i les diferències en el canvi de seqüència.

Finalment, un tercer argument defensaria que una part de les diferències que hem trobat no implicarien diferències funcionals, tot i ser divergents a nivell de seqüència i esdeveniment d'splicing. És a dir, tots dos esdeveniments que duen a la pèrdua de dominis tindrien el mateix impacte funcional a partir de diferents processos d'AS, el que ve a ser un cas de convergència funcional (Figura 4). Aquesta visió ja ha estat suggerida en diversos treballs (Fodor and Aldrich, 2009; Lois et al., 2007), el primer d'ells en el cas dels canals de potassi dependents de calci, i, en el segon treball, en un conjunt de reguladors epigenètics. Així doncs, tot i que la funcionalitat de les isoformes, en general, estigui conservada, els mecanismes a nivell de seqüència, com l'AS, que l'han generada poden deure's a vies evolutives divergents. Aquesta opció és especialment interessant en el cas de les alteracions del patró de dominis de la Classe 4, pel seu paper establert de reguladors (per exemple, dominants negatius), però seria molt interessant també generalitzar aquest anàlisi a tota la resta d'esdeveniments d'AS per les diferents categories funcionals del proteoma.

4.5 CONCLUSIONS

La regulació de l'expressió gènica és un fenomen cabdal en l'establiment de les diferències fenotípiques entre organismes. És per això que en aquest capítol hem estudiat quines són les propietats i la conservació dels mecanismes reguladors basats en AS. Per una banda, hem observat que l'AS i la regulació de l'expressió gènica convencional són independents i que actuarien a diferents nivells. A continuació, hem descrit una baixa conservació dels mecanismes reguladors per AS pel que fa a l'alteració de l'arquitectura de dominis de les isoformes. En concret, per aquells esdeveniments d'AS que comporten pèrdua de dominis, la conservació del mecanisme d'AS és baixa, encara que la funcionalitat s'hagi conservat, el que ens fa pensar en un cas de convergència funcional. Tot plegat ens fa concloure que l'AS disposa d'un espai dins la regulació gènica, independent dels altres mecanismes de regulació, on introduir variabilitat i influir en la determinació de les diferències fenotípiques entre humà i ratolí.

5. La implicació de l'splicing alternatiu en el fenomen de la dominància gènica

5.1 Introducció

El concepte de dominància genètica deriva del fet que la contribució dels diferents al·lels d'un gen autosòmic donat, d'un organisme com a mínim diploide, és sovint no additiva (Kondrashov and Koonin, 2004). Entendre les bases moleculars d'aquest fenomen és basic per incrementar el coneixement de les vies metabòliques i el seu impacte en el fenotip, així com en el desenvolupament dels processos evolutius (Kondrashov and Koonin, 2004). Des d'un punt de vista més aplicat, és necessari per conèixer amb més claredat els processos que duen a l'aparició de malalties (Furney et al., 2006; Lopez-Bigas et al., 2006; Osada et al., 2009).

	Punt de vista EVOLUTIU (Al·lel salvatge)	Punt de vista GENÈTICA HUMANA (Al·lel mutant)	Relació amb la dosi gènica
Gen Haplosuficient	Dominant	Recessiu	Insensible
Gen Haploinsuficient	Recessiu	Dominant	Sensible

Taula 1. Resum terminològic sobre la dominància. La part ombrejada correspon a la que s'utilitzà en aquest treball

El debat sobre què és exactament la dominància s'ha estès durant més d'un segle, des que Gregor Mendel fes les primeres quantificacions d'aquest fenomen (Furney et al., 2006). A partir de pocs caràcters qualitatius, Mendel va concloure que una sèrie de "factors" (al·lels) dominants amagaven els efectes d'uns altres de recessius (Veitia and Birchler, 2010). Posteriorment, durant el primer terç del segle XX, les investigacions es van enfocar cap a l'explicació dels mecanismes moleculars de la

dominància. Dues teories es van enfrontar: per una banda la teoria fisiològica de Sewall Wright i, per l'altra, la teoria evolutiva de R.A. Fisher (Bagheri, 2006; Kondrashov and Koonin, 2004). Segons Fisher, els al·lels mutants serien recessius respecte els al·lels salvatges i aquesta dominància hauria aparegut gràcies a la selecció natural, que protegiria els heterozigots dels efectes adversos de mutacions recurrents deletèries (Fisher, 1928). Per la seva banda, Wright va criticar que, segons la teoria de Fisher, era necessari una pressió selectiva enorme per mantenir i fixar aquesta dominància, pressió que considerava més bé feble i no exempta de ser afectada per la deriva gènica. En contraposició, Wright postulà que la dominància no era resultat de l'adaptació, sinó una conseqüència de la fisiologia de l'organisme i que, per tant, els al·lels salvatges de gens sensibles a dosi haurien de ser recessius, tot just el contrari del que defensà Fisher (Wright, 1934).

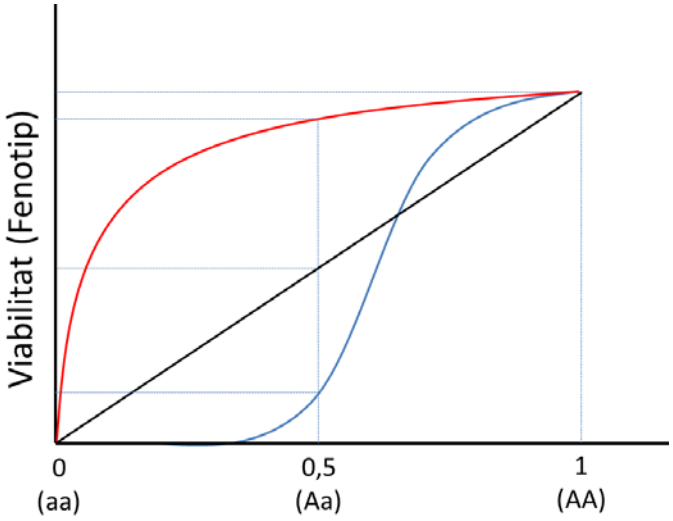

Figura 1: Relació entre dosi gènica i viabilitat (fenotip) depenent del tipus de dominància. La línia vermella correspon a gens haplosuficients i la línia negra a gens haploinsuficients. La línia blava correspon a gens haploinsuficients altament sensibles a dosi, com podrien ser els que formen part de subunitats de complexos multiproteics. Adaptat de la Figura 1 de (Kondrashov and Koonin, 2004)

Cap al darrer quart del segle XX, la teoria de Wright va rebre un fort suport gràcies al treball de Kacser i Burns (Kacser and Burns, 1981), els quals van modelar l'efecte de mutacions en gens que codifiquen per enzims en vies metabòliques. Van veure que la dominància dels al·lels salvatges dels enzims era una conseqüència de les propietats del flux metabòlic, el qual podia resistir fins a un 50% de disminució

de la concentració de l'enzim sense alterar de manera notable la via o cicle metabòlic. Aquest resultat concorda amb la predicció de Wright ja que en el cas d'un gen que formi part d'una via amb diversos passos, el fenotip associat amb una mutació d'aquest gen ha de ser insensible a la dosi gènica (Kondrashov and Koonin, 2004). Per una altra banda, estudis sobre l'aneuploïdia van poder simular un efecte extrem de la dosi gènica i, per tant, donar pistes sobre la dominància. Van identificar que els efectes en l'expressió gènica per la pèrdua/guany de còpia no afectaven només el cromosoma aneuploide (monosòmic, trisòmic, etc.) en qüestió, sinó que s'estenien a altres regions del genoma. Aquest efecte era produït per una sèrie de gens individuals presents als cromosomes aneuploides que eren, de manera predominant, factors de transcripció, proteïnes d'unió a la cromatina i membres de cascades de transducció de senyal, categories funcionals que després van ser relacionades amb els gens haploinsuficients (recessius des del punt de vista d'al·lel salvatge) (Birchler et al., 2005). Així doncs, el principals efectes de l'aneuploïdia (dosi gènica) eren deguts a l'alteració de l'estequiometria dels components reguladors de l'expressió gènica i no només a la variació del nombre de còpies dels gens reguladors. En definitiva, l'aneuploïdia va resultar ser una eina per identificar una sèrie de gens sensibles a dosi i haploinsuficients que confirmaren les derivacions de la teoria de Wright, tal com des d'aleshores han fet diversos treballs (Kondrashov and Koonin, 2004; Qian and Zhang, 2008; Veitia, 2002). En aquest punt és important assenyalar quina ha estat la terminologia utilitzada en els treballs que han tractat la dominància aquest darrers anys (Birchler et al., 2005; Kondrashov and Koonin, 2004), donat que els termes de dominant i recessiu poden dur a moltes confusions. S'entén com a gens haplosuficients aquells casos en els que l'al·lel salvatge és dominant sobre el mutant i és insensible de la dosi gènica, i gens haploinsuficients seran aquells en els que l'al·lel mutant afecti significativament el fenotip en estat heterozigot (Figura 1). Encara així aquests termes no estan exempts de confusions: en genètica humana el significat d'haplosuficient i haploinsuficient és just el contrari, ja que s'observa des del punt de vista dels al·lels anormals (Furney et al., 2006). Tot i això, nosaltres seguirem amb el sentit que s'empra en biologia evolutiva, que és la que hem plantejat en primer lloc, és a dir, des del punt de vista dels al·lels salvatges (Kondrashov and Koonin, 2004). Així doncs, per una banda, dominant i

haplosuficient i, per l'altra, recessiu i haploinsuficient seran sinònims per aquest treball. Cal recordar que en les bases de dades que utilitzarem el sentit és l'oposat i el que s'ha definit com a dominant a OMIM o MGI és recessiu des del punt de vista evolutiu i, per tant, haploinsuficient per aquest treball. El mateix raonament fem pels gens recessius per OMIM o MGI. A la taula 1 es resumeix l'equivalència entre les diferents terminologies.

Tal com hem esmentat, entre els gens insensibles a dosi gènica (haplosuficients o dominants) trobem els gens que codifiquen per proteïnes amb funció enzimàtica. Per contra, entre els haploinsuficients, sensibles a dosi gènica o recessius, podem distingir els gens que codifiquen per proteïnes amb funcions reguladores o estructurals, factors de transcripció, de senyalització o d'unió a diversos substrats, entre d'altres (Birchler et al., 2005; Kondrashov and Koonin, 2004). Evidentment, no és una relació unívoca ja que pot succeir que enzims amb una baixa taxa de renovació poden tenir un comportament haploinsuficient o, en el cas contrari, proteïnes estructurals o reguladores que es trobin a concentracions molt baixes poden ser considerades haplosuficients (Kondrashov and Koonin, 2004). És en aquest camp, en el de les proteïnes no enzimàtiques, que s'ha obert un ventall de possibilitats en el paper de la dominància. Mitjançant un enfocament mecanístic, s'ha trobat relacions entre la dosi gènica, la funció proteica i l'estructura de complexos proteics (Birchler et al., 2005; Teichmann and Veitia, 2004). Un aspecte particularment important ha estat la integració dins la teoria de la dominància de l'efecte de les duplicacions gèniques, les quals són una font destacada de variabilitat en la dosi gènica i el producte proteic (Kondrashov and Koonin, 2004; Veitia, 2005). En aquest treballs, han trobat que el nombre de paràlegs per gens haploinsuficients són significativament més alts que pels gens haplosuficients, constatació que es deriva de la teoria de Wright on la fixació per selecció és més probable en gens dependents de dosi que no pas en els haplosuficients (Kondrashov and Koonin, 2004). En un context molecular, aquesta relació ens fa pensar que l'splicing alternatiu, en concret l'splicing alternatiu regulador (RAS), pugui tenir també un paper destacat dins de la dominància genètica per dos motius: en primer lloc, diversos estudis han trobat una relació inversa entre AS i paralogia (Kopelman et al., 2005; Talavera et al., 2007b); en segon lloc, RAS és també una font important de variabilitat d'isoformes, amb propietats reguladores

susceptibles de tenir algun rol en la dominància. No obstant això, fins allà on sabem, no hi ha hagut fins ara cap estudi amb l'objectiu de comprendre la contribució de l'AS a la dominància genètica. L'objectiu d'aquest treball ha estat estudiar aquesta contribució i integrar l'splicing alternatiu dins del marc de la teoria fisiològica de la dominància.

5.2 Material i mètodes

5.2.1 Obtenció de les dades sobre dominància i splicing a humà

El protocol per humà es basa en el que trobem descrit en el treball de Kondrashov (Kondrashov and Koonin, 2004), tot integrant les modificacions suggerides en treballs posteriors (Qian and Zhang, 2008) i una sèrie de filtres de qualitat propis. Aquest protocol es basa en l'ús de dades OMIM (Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/omim) per a obtenir un model sobre l'efecte que té la dominància sobre el fenotip. Aquesta base de dades està constituïda per un recull de gens amb al·lels anormals que tenen descrita la seva herència mendeliana i l'efecte patològic que provoquen, és a dir, el fenotip resultant. Descrivim a continuació els passos d'aquest protocol:

Pas 1: Obtenció dels conjunts inicials de gens dominants i recessius. Es basa en una cerca local exhaustiva de la base de dades OMIM, per tal de trobar els gens humans amb al·lels anormals responsables de malalties, tant "dominants" com "recessius". La cerca va incloure aquestes paraules combinades amb "haploinsuficient" i "haplosuficient", però només utilitzant l'arrel dels termes (per exemple: recessi*, haploin*, etc) per a poder recuperar el màxim de casos. Depenent en quin camp de la fitxa d'OMIM aparegués aquesta paraula, se li donava una puntuació o una altra, el que ens donava un índex de qualitat (per exemple, que aparegués la paraula "dominant" en la descripció de la malaltia tenia molt més valor que no pas en la "Sinopsi clínica", en la que sovint es pot trobar informació poc clara o contradictòria). Només vam considerar aquells gens on el tipus d'herència hagués sigut degudament demostrat per OMIM.

Pas 2: Filtratge de les dades. Els gens contradictoris, és a dir, aquells que haguéssim trobat que eren alhora dominants i recessius, van ser sotmesos a una inspecció

manual. Si la contradicció es confirmava, eren descartats. Tot seguit es va identificar si hi havia "dominants negatius" i es va comprovar manualment que fossin descrits com a dominants, tal com ens han posat en avís altres treballs (Qian and Zhang, 2008) (en efecte, l'expressió "dominant negatiu" pot donar lloc a errors en la cerca textual).

5.2.2 Obtenció de les dades sobre dominància i splicing a ratolí

Per obtenir les dades de ratolí vam fer servir la mateixa idea del procés emprat per humà, però fent servir l'equivalent d'OMIM a ratolí, que és la base de dades del MGI (Mouse Genome Informatics, http://www.informatics.jax.org/) (Blake et al., 2011). El protocol també consistí de dues passes.

Pas 1: Obtenció dels conjunts inicials de gens dominants i recessius. Les dades de ratolí van ser obtingudes a partir d'un conjunt de dades, generat expressament pel personal del MGI a petició nostra, que incloïa informació de la herència de tots els al·lels i l'efecte patològic en l'animal. En aquest cas, la informació es trobava només en termes de dominant, recessiu i les variants respectives.

Pas 2: Filtratge de les dades. Vam descartar els al·lels amb herència "codominant" i els "semidominants" es van considerar dominants. De la mateixa manera que havíem fet amb OMIM, els termes "dominants" va ser considerats haploinsuficients i "recessiu", haplosuficient. Tots els gens que tinguessin contradiccions van ser descartats, ja que no teníem cap altra dada de qualitat, com a OMIM, que ens fes decidir entre dominant o recessiu.

5.2.3 Generació del conjunt de dades sobre dominància i splicing alternatiu

Tots aquests gens amb informació sobre herència per humà i ratolí van ser creuats amb la nostra base de dades (veure Capítol 3), per obtenir finalment una taula amb informació sobre el gen, tipus d'herència, índex de qualitat (en el cas d'humà), pertinència a algun parell d'ortòlegs humà-ratolí propi, presència d'AS, nombre d'isoformes, presència d'alguna isoforma RAS (veure Capítol 4) i tota la informació obtinguda fins el moment (veure Capítol 3 i 4).

A banda de tot aquest procés, vam generar un conjunt de dades amb la funció de control extern de les taules mencionades anteriorment. A tal fi, vam creuar la

informació de dominància obtinguda a OMIM/MGI amb dades d'Ensembl (Flicek et al., 2012). Mitjançant BioMart, vam recuperar gens unívocs amb el nombre d'isoformes conegudes que es transcriuen per cadascun dels gens (Taula 4), obtenint així una taula amb informació sobre gen, tipus de dominància i presència d'AS

5.2.4 Dades addicionals de caracterització de la dominància

A banda de la informació bàsica sobre dominància i splicing, vam voler caracteritzar les dades, tal com s'ha fet en treballs precedents, amb informació addicional de categories funcionals, pertinença a algun complex proteic i, com a novetat, freqüències d'SNP's a la seqüència *cis*-reguladora (l'objectiu fonamental d'aquest pas és enriquir la visió biològica que es deriva de les nostres dades). A continuació es descriu com es van obtenir aquestes dades.

Gene ontology (GO). Vam cercar possibles enriquiments diferencials entre els gens haplosuficients i haploinsuficients, en categories de funció molecular de GO al nivell 3 (Gene Ontology, http://www.geneontology.org/) (Carbon et al., 2009), fent servir la pàgina de FatiGO (http://babelomics.bioinfo.cipf.es) (Medina et al., 2010). A banda, també vam analitzar les categories de Gene Ontology, però sense normalitzar al nivell 3, a les utilitats de GOrilla (Gene Ontology enRIchment anaLysis and visuaLizAtion tool, http://cbl-gorilla.cs.technion.ac.il/) (Eden et al., 2007; Eden et al., 2009)

Complexos proteics. Només en humà vam afegir la informació sobre la pertinença d'aquests gens a algun complex proteic a partir de la web de CORUM (Comprehensive Resource of Mammalian protein complexes) (Ruepp et al., 2010) i les anotacions trobades a UniProt/SwissProt. La pertinença a complexos proteics és rellevant des del punt de vista de la teoria de la dominància, ja que està relacionada amb els mecanismes moleculars que l'expliquen (Teichmann and Veitia, 2004; Veitia, 2002)

SNPs. Pel que fa a la distribució d'SNP's al promotor, vam obtenir les seqüències de 1000 parells de bases a 5' del TSS a partir del servidor d'UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables) (Karolchik et al., 2004). La versió per humà fou GRCh37/h19 (Feb. 2009) i NCBI37/mm9 (Jul.2007) per ratolí. Les

anotacions d'SNPs es van recuperar a partir de la plataforma Galaxy (http://main.g2.bx.psu.edu/) (Blankenberg et al., 2010; Giardine et al., 2005; Goecks et al., 2010).

5.3 Resultats

5.3.1 Descripció de les dades

Pel que fa a humà, vam obtenir 1370 gens amb informació inequívoca i de qualitat sobre l'herència. Aquest valor és molt superior al que s'ha obtingut en treballs anteriors (Kondrashov and Koonin, 2004; Qian and Zhang, 2008). D'aquests, 574 gens (41,9%) van resultar ser haplosuficients i 796 gens (58,1%) presentaven un comportament haploinsuficient (Taula 2). A ratolí, vam recuperar 692 gens amb informació sobre herència, del quals 560 (81%) eren haplosuficients i 132 gens (19%) haploinsuficients (Taula 3). Cal dir que un 66% dels al·lels i un 63% dels gens de les dades originals de MGI eren haplosuficients. Per enriquir i contrastar aquestes dades, paral·lelament, vam recuperar 1583 gens humans amb informació no contradictòria a OMIM i Ensembl, dels quals 948 (59,9%) foren haploinsuficients i 635 gens (40,1%) haplosuficients. Pel que fa a ratolí i amb dades d'MGI i Ensembl, vam obtenir 746 gens, dels quals 149 (20%) foren haploinsuficients (Taula 4). En general, tot i obtenir més gens, els percentatges són similars.

	Gens sense AS (%)	Gens amb AS (%)	Gens amb RAS (%)	Totals
Gens haplosuficients	320 (55,8%)	254 (44,2%)	70 (12,2%)	574 (41,9%)
Gens haploinsuficients	439(55,2%)	357 (44,8%)	118 (14,8%)	796 (58,1%)
Total gens	760 (55,5%)	610 (44,5%)	188 (13,7%)	1370

Taula 2 . Resum de les dades obtingudes sobre dominància, presència d'AS i RAS a Humà

	Gens sense AS (%)	Gens amb AS (%)	Gens amb RAS (%)	Totals
Gens haplosuficients	353 (63%)	207 (37%)	66 (31,9%)	560 (81%)
Gens haploinsuficients	90 (68,2%)	42 (31,8%)	10 (23,8%)	132 (19%)
Total gens	443 (64%)	249 (36%)	76 (11%)	692

Taula 3 . Resum de les dades obtingudes sobre dominància, presència d'AS i RAS a Ratolí

		Gens sense AS (%)	Gens amb AS (%)	Totals
Humà	Gens haplosuficients	150 (23,6%)	485 (76,4%)	635 (59,9%)
	Gens haploinsuficients	206 (21,7%)	742 (78,3%)	948 (40,1%)
	Total	356 (22,5%)	1227 (77,5%)	1583
Ratolí	Gens haplosuficients	211 (35,4%)	386 (64,6%)	597 (80%)
	Gens haploinsuficients	62 (41,6%)	87 (58,4%)	149 (20%)
	Total	273 (36,6%)	473 (63,4%)	746

Taula 4 . Resum de les dades obtingudes sobre dominància i presència d'AS per humà i ratolí amb dades provinents d'Ensembl

El següent pas en la caracterització de les dades va ser comprovar si aquests gens formaven part de complexes proteics o no. Amb dades només per humà, 643 gens (47%) tenien aquesta informació, dels quals 407 (63,3%) eren haploinsuficients i significativament més enriquits del que esperàvem (X^2 = 2x10^{-4}, p-valor<0,05) respecte els gens haplosuficients (236 gens), que resultaren estar esbiaixats cap gens no pertanyents a complexes proteics (Taula 5).

	Pertinença a complex proteic (%)	No pertinença a complex proteic (%)	Total
Gens Haplosuficients	236 (36,7%)	338 (46,5%)	574
Gens Haploinsuficients	407 (63,3%)	389 (53,5%)	796
Total	643 (46,9%)	727 (53,1%)	1370

Taula 5 . Resum de les dades obtingudes sobre dominància i pertinença a complex proteic per humà.

5.3.2 Anàlisi Gene Ontology

Una dada interessant és veure quina distribució de categories de Gene Ontology presenten les nostres dades, ja que ens permet relacionar-les amb l'impacte que té la dominància en els processos biològics i és rellevant a l'hora d'interpretar-les en termes mecanístics. Per a cadascuna de les espècies, vam comparar les llistes de gens haplosuficients i haploinsuficients amb FatiGO i vam estudiar quins termes de nivell 3 estaven enriquits en cada una d'elles. Pel que fa a humà, entre els gens haploinsuficients (dominants a OMIM) vam trobar que les categories de funció molecular de *protein binding* (GO:0005515), *nucleic acid binding* (GO:0003676), *transcriptor activator activity* (GO:0016563) i *RNA polymerase II transcription factor activity* (GO:0003702) eren significativament superiors (p valor >0,05) als gens haplosuficients, els quals només presentaven la categoria d'activitat hidrolasa (GO:0016787) com a significativament diferent. En el cas de ratolí, els gens haploinsuficients presentaven les categories de funció molecular *nucleic acid binding*, *chromatin binding* (GO:0003682), *structural constituent of ribosome* (GO:0003735) i *structural constituent of eye lens* (GO:0005212) com a significativament representades. Excepte per a la darrera categoria, les dades són raonablement consistents amb les dades obtingudes d'humà ja que, en ambdues espècies, les categories significatives estan relacionades amb el control de l'expressió gènica. En l'anàlisi de GOrilla, vam obtenir resultats similars però sense normalitzar al nivell 3 de les categories GO de funció molecular. Els gens haploinsuficients, tant per humà com ratolí, estaven enriquits en categories d'unió a DNA (en especial a regions reguladores), unió a cromatina, unió a proteïna (a enzims, proteïnes quinases, etc.) i relacionades amb l'activitat de factors transcripcionals.

5.3.3 Dominància i AS

A continuació, vam comprovar la hipòtesi principal d'aquest treball: si hi havia diferències en la distribució de l'splicing alternatiu, tal com hem definit al Capítol 3, entre els gens haplosuficients i haploinsuficients. Aplicant el test de la chi-quadrat tant als resultats de la taula 2 (per humà) com als de la taula 3 (per ratolí) no vam trobar diferències entre els dos grups (X^2 = 0,77 i 0,28, respectivament, p-valor≥0,05). Pel que fa al control amb dades d'Ensembl (Taula 4), tampoc vam

trobar diferències entre els gens haplosuficients i haploinsuficients per humà i ratolí (X^2 = 0,37 i 0,15, respectivament, p-valor≥0,05). Aquest resultat ens indica que no hi ha cap diferència significativa pel que fa a la presència d'AS entre els gens haplosuficients i haploinsuficients o, el que és el mateix, no hem trobat cap relació entre dominància i AS. Tampoc vam trobar diferències significatives en la distribució del nombre d'isoformes entre els gens haplosuficients i haploinsuficients (X^2 = 0,09 i 0,16 per humà i ratolí respectivament, p-valor≥0,05) (Figura 2a), el que ens fa deduir que no hi ha cap relació entre dominància i nombre d'isoformes.

En el cas de RAS (Capítol 4), la dificultat principal fou aconseguir un nombre acceptable de casos. Tot i això, vam obtenir 188 gens amb alguna isoforma RAS a humà i 76 a ratolí. En tot cas, no vam trobar una distribució diferencial entre els gens haplosuficients i haploinsuficients (X^2 = 0,16 tant per humà com ratolí, p-valor≥0,05), confirmant d'aquesta manera que per RAS tampoc hi ha cap relació significativa entre dominància i AS.

Pel que fa a la conservació de l'splicing alternatiu entre els ortòlegs d'humà i ratolí, no vam trobar diferències entre els gens que tenien conservada la dominància i els que no la tenien. En tots dos casos, un 70% de les parelles presentaren el mateix patró de presència/absència d'splicing.

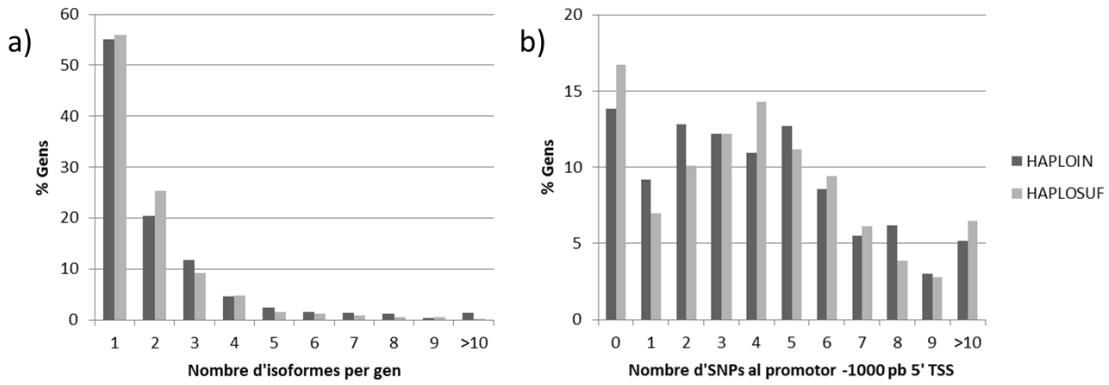

Figura 2: Nombre d'isoformes per gen i nombre d'SNP's al promotor per humà

5.3.4 Relacions d'ortologia i dominància

Pel que fa a l'ortologia, vam trobar que al voltant del 88% dels gens tant a humà com ratolí estaven inclosos en alguna parella d'ortòlegs (veure Capítol 3). El nombre d'ortòlegs pels quals vam recuperar informació sobre la dominància en totes dues espècies va ser de 203 (Taula 6). D'aquestes, un 75,9 % (154 parelles) presentaren conservació en la dominància, de les quals una proporció destacada foren haplosuficients (105 ortòlegs). Pel conjunt de les dades, la conservació de la dominància, és a dir, parelles d'ortòlegs que fossin tots dos haplosuficients o haploinsuficients, era significativament superior del que esperàvem a l'atzar (X^2 = 4,5x10^{-14}, p-valor<0,05).

		Ratolí		
		Haploinsuficient	Haplosuficient	Total
Humà	Haploinsuficient	49 (24,1%)	42 (20,7%)	91
	Haplosuficient	7 (3,5%)	105 (51,7%)	112
	Total	56	147	203

Taula 6. Resum de les dades obtingudes sobre la conservació de la dominància entre parelles d'ortòlegs d'humà i ratolí.

5.3.5. SNPs en regió promotora

Per finalitzar, vam voler explorar quina distribució hi havia d'SNPs al promotor dels gens (Figura 2b). Esperàvem obtenir un nombre major d'SNPs en els promotors dels gens haplosuficients, ja que aquests tenen la capacitat d'acumular més variabilitat sense alterar de manera greu la funció o el fenotip resultant de l'expressió d'aquest gen. Tot i això, no hi ha cap relació significativa en la distribució de SNPs al promotor entre els dos tipus de gens (X^2 = 0,19 i 0,30 a humà i ratolí, respectivament, p-valor≥0,05).

5.4 Discussió

L'objectiu d'aquest treball és el d'establir la relació entre dos fenòmens associats a l'impacte fenotípic i a la regulació funcional dels gens, dominància i splicing alternatiu, respectivament. El nostre punt de partida van ser, per una banda, els

resultats d'una sèrie d'articles (Kondrashov and Koonin, 2004; Veitia, 2005) en els quals es va trobar una relació entre duplicació gènica i dominància i, per l'altra banda, el fet que es coneixia la relació inversa entre paralogia i la presència d'splicing alternatiu en humà i altres espècies (Hughes and Friedman, 2008; Irimia et al., 2008; Irimia et al., 2009; Jin et al., 2008; Kopelman et al., 2005; Talavera et al., 2007b). Aquestes dues observacions ens van fer pensar de forma natural en la possibilitat que l'splicing alternatiu i la dominància podien tenir una relació inversa a la trobada prèviament entre duplicació gènica i dominància.

En contra del que esperàvem, els nostres resultats (Taules 2-4) no mostren l'existència de cap relació significativa que ens permeti recolzar aquesta hipòtesi, a cap nivell: ni amb les dades de presència d'splicing alternatiu, ni amb el nombre d'isoformes ni amb una distribució diferencial de les isoformes RAS entre els gens haplosuficients i haploinsuficients. Aquests resultats aparentment contradictoris amb el nostre punt de partida, son però consistents amb el treball recentment publicat per Roux i Robinson-Rechavi (Roux and Robinson-Rechavi, 2011). Aquests autors fan servir un conjunt més gran de dades de paralogia i AS, així com un nou tipus d'anàlisi, i troben que no hi ha una relació entre splicing i paralogia. Això aclariria la manca de relació que trobem entre AS i dominància.

En l'aspecte comparatiu, vam poder recuperar una quantitat considerable de parelles d'ortòlegs (203), la majoria de les quals (75,9%) presentaren conservació de la dominància (Taula 6). Observacions prèvies (Veitia, 2005) apuntaven a una baixa conservació dels gens humans haploinsuficients en ratolins, en part degut a la incapacitat per a detectar correctament els efectes fenotípics per a gens haploinsuficients murins en heterozigosi (Veitia, 2002). En el present treball, no hem pogut demostrar aquesta baixa conservació. Ans al contrari: la conservació de l'haploinsuficiència era significativament superior a l'esperat. Així doncs, encara que partíem d'un conjunt de dades esbiaixades cap a gens i al·lels haplosuficients per ratolí, els resultats ens indiquen que la conservació de la dominància, tant per gens recessius com dominants, és significativament superior a l'esperat.

La caracterització dels dos tipus de gens segons les categories de Gene Ontology van coincidir prou bé amb resultats anteriors (Birchler et al., 2005; Kondrashov and Koonin, 2004), donant suport al nostre protocol de recollida de dades, encara

que pràcticament no vam poder trobar cap categoria enriquida entre els gens haplosuficients tant per humà com per ratolí. Entre els gens haploinsuficients, les funcions moleculars d'unió (a DNA, a proteïna, etc.) i les tasques relacionades amb la transcripció van ser les majoritàries. Les funcions estructurals estaven representades però en menor mesura. De totes maneres, tal com s'ha assenyalat a la introducció, aquesta categorització de Gene Ontology s'ha de prendre amb certa precaució, ja que els gens en qüestió poden tenir un comportament contrari al que s'esperava per condicions especials de concentració, renovació de la proteïna, etc. A més a més, certs gens, tot i comportar un desordre dominant (des del punt de vista d'OMIM), no tenen perquè ser haploinsuficients, en especial en els casos on s'expressi alguna isoforma que actuï com a dominant negatiu (Figura 3) (Kondrashov and Koonin, 2004; Qian and Zhang, 2008; Veitia, 2002).

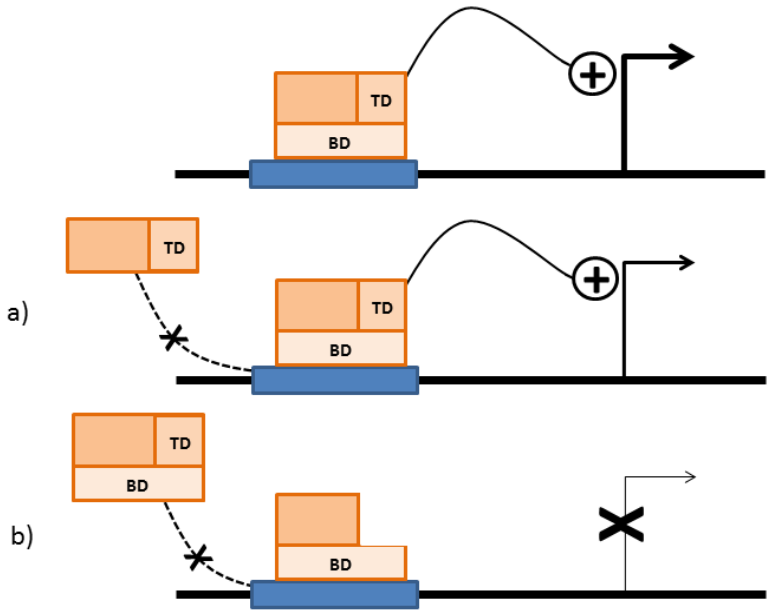

Figura 3: Haploinsuficiència i dominant negatiu. Donat un factor de transcripció compost per un domini d'unió a una regió específica del promotor (DNA-BD) i un domini transactivador de la transcripció (TD) podem observar diferents comportaments depenent de l'afectació de l'al·lel mutat. a) Exemple típic de fenotip haploinsuficient depenent de dosi gènica on l'al·lel mutat perd la capacitat d'unió al DNA, però manté el domini TD. b) Exemple de dominant negatiu: l'al·lel mutat perd la capacitat d'activar la transcripció però manté el domini d'unió al DNA. Aquesta isoforma competeix amb l'al·lel salvatge i l'expressió del gen es pot veure greument afectada en heterozigosi (Veitia, 2002).

Per una altra banda, els gens haploinsuficients són majoria (63'3 % de 643 gens humans amb informació sobre complexos proteics) i significativament diferents pel que fa a la pertinença a algun complex proteic (Taula 5). Els gens haplosuficients, per contra, estaven enriquits en gens que no han estat identificats participant en cap complex proteic. Aquesta informació està en concordança amb els resultats obtinguts en treballs previs (Birchler et al., 2005; Birchler and Veitia, 2010; Papp et al., 2003; Veitia, 2002; Veitia, 2005) en els quals afirmen que, degut a que la dosi de cada una de les subunitats dels complexos multiproteics estan sovint fortament regulada, l'increment o decreixement d'aquesta dosi pot produir efectes notables en el fenotip. Això els fa pensar que aquestes subunitats proteiques tendiran a ser codificades per gens haploinsuficients. Tot plegat és un resultat totalment coincident amb el desenvolupament de la teoria fisiològica de la dominància .

Finalment, i relacionat amb aquest darrer punt, esperàvem que els gens haploinsuficients tinguessin menys quantitat d'SNPs al promotor (-1000 pb 5' TSS) pel mateix raonament desenvolupat a l'hora de tractar els complexos proteics. S'ha vist que els SNPs al promotor poden afectar el nivell d'expressió general del gen (Chorley et al., 2008; Shastry, 2009) i, des del punt de vista d'aquest treball, podria alterar la dosi gènica d'un al·lel concret. En principi doncs, els gens haplosuficients haurien de poder absorbir més SNPs al promotor sense que hi hagués una alteració destacada del fenotip. Però no hem obtingut cap resultat significatiu que ens pugui permetre defensar aquesta postura. Això pot haver estat degut a causes diverses de tipus tècnic, com podrien ser una mala cobertura d'SNPs al promotor o potser una mala definició del què és el promotor encara que, tal com hem tractat al capítol 3, amb la mesura adoptada de -1000 pb a 5' a partir del TSS recuperem la màxima informació sobre el promotor de la majoria de gens. Per una altra banda, la independència d'aquests dos fenòmens de manera global no contradiu el fet que, en gens concrets, els SNPs sí que generen haploinsuficiència (Masotti et al., 2005; Prokunina and Alarcon-Riquelme, 2004). Finalment, també pot ocórrer que molts dels SNPs recollits no afectin la funcionalitat del gen, és a dir, que siguin canvis neutres que no alteren de cap manera la regulació del gen; en aquest cas, esperem que tinguin una distribució similar tant en gens haplosuficients com haploinsuficients. Per obtenir una comparació més acurada entre dominància i

SNPs al promotor, seria necessari obtenir un conjunt de dades d'SNPs en els qual s'hagi comprovat la seva implicació, encara que sigui lleu, en l'alteració de la maquinària transcripcional.

5.5 Conclusions

L'objectiu del present treball era la caracterització de la relació entre l'splicing alternatiu i la dominància, una relació previsible en base als estudis previs sobre paralogia i dominància. Per tal d'assolir-ho, hem caracteritzat i actualitzat les dades sobre dominància obtingudes a partir d'OMIM, i hem integrat en aquesta visió dades sobre ratolí amb les quals hem fet un primer estudi comparatiu amb humà. En els punts ja estudiats anteriorment (categories GO, complexos proteics, etc.), hem obtingut resultats comparables als d'altres autors. Però, de manera interessant, pel que fa a les hipòtesis que ens vàrem plantejar a l'inici, no hem aconseguit trobar cap relació entre splicing alternatiu, inclòs RAS, i dominància. Això ens fa dubtar de l'existència d'una relació inversa establerta entre paralogia i splicing alternatiu, en consonància amb resultats recentment obtinguts (Roux and Robinson-Rechavi, 2011). Finalment, tampoc hem vist cap relació entre la variabilitat nucleotídica al promotor i el tipus de dominància, el que ens suggereix independència dels dos fenòmens, encara que seria necessari una classificació més acurada dels SNPs per obtenir resultats més clarificadors.

RESUM I CONCLUSIONS

6. Resum

L'estudi de les diferències fenotípiques entre espècies, i entre individus, ha estat una de les grans qüestions fonamentals en els camps de la biologia evolutiva i la genètica. Ben aviat, es va fer palès que la regulació de l'expressió gènica tindria un paper clau en establir aquestes diferències de complexitat. L'adveniment de les tècniques massives de seqüenciació no van sinó confirmar aquesta visió primerenca. Avui dia coneixem un grapat de mecanismes que determinen aquestes diferències entre organismes, com són la divergència de seqüència proteica, la duplicació gènica o la divergència de la regió *cis*-reguladora, entre d'altres.

En la darrera dècada, l'splicing alternatiu ha anat afermant-se com a mecanisme post-transcripcional freqüent i ha anat prenent protagonisme com a font de variabilitat de transcrits i isoformes proteiques, a més a més de jugar un paper regulador de l'expressió gènica. Per tant, l'splicing alternatiu és un ferm candidat a introduir diferències substancials al proteoma que expliquin la diversitat fenotípica entre organismes.

Així doncs, aquest treball es va marcar com a objectiu aclarir fins a quin punt la variabilitat que introduïa l'splicing alternatiu tenia implicacions en el fenotip, quina era la seva conservació i si actuava de manera coordinada o independent amb d'altres mecanismes. En primer lloc, vam estudiar la relació que hi havia entre l'splicing alternatiu i les altres fonts moleculars de diversitat fenotípica i si era possible que l'splicing alternatiu pogués introduir variabilitat amb implicacions fenotípiques per si sola . A continuació, ens vam centrar en els mecanismes reguladors de l'expressió gènica basats en splicing alternatiu, analitzant les seves propietats i la seva conservació entre espècies. Finalment, vam examinar la implicació de l'splicing alternatiu en el fenomen de la dominància gènica, ja que és un procés conegut que determina diferències fenotípiques intraespecífiques.

El primer pas fou, doncs, comparar l'splicing alternatiu amb d'altres fonts moleculars de diferències fenotípiques: les divergències de la seqüència proteica, de la regió *cis*-reguladora del gen i de l'expressió gènica entre humà i ratolí. En un estudi massiu de les propietats de tots aquests fenòmens entre 13970 parelles d'ortòlegs, vam observar que l'splicing alternatiu podia introduir diferències abans que les altres variables poguessin fer-ho. Quan les identitats de seqüència proteica o de la regió *cis*-reguladora eren massa elevades com per introduir diferències, l'splicing alternatiu ja presentava patrons prou diferents en la concurrència d'splicing entre humà i ratolí. A més a més, la relació entre l'equivalència d'isoformes amb aquestes divergències també va resultar ser molt lleu, fet que ens va fer pensar que l'splicing alternatiu pot introduir isoformes específiques que contribueixin a les diferències entre espècies abans que les altres divergències puguin fer-ho.

Pel que fa al segon bloc, vam investigar la conservació i propietats dels mecanismes reguladors de l'expressió gènica basats en AS. Primer de tot, vam confirmar la independència entre les divergències d'expressió gènica i l'splicing alternatiu, fet que ens indica que actuen a diferents nivells. A continuació, vam definir i classificar aquests mecanismes reguladors depenent com l'splicing alternatiu alterava l'arquitectura de dominis de les isoformes. La conservació d'aquests efectes, dels mecanismes reguladors basats en AS, va resultar ser baixa per tots els casos. Pel que fa als esdeveniments on es perdien un o més dominis a les isoformes alternatives, a més a més de ser baixa la conservació del mecanisme, també ho va ser l'equivalència dels esdeveniments d'splicing alternatiu. Així, tot i tenir efectes a nivell de seqüència no homòlegs, la funció es conservava, fet que ens porta a suggerir que aquests esdeveniments d'AS són un exemple de convergència funcional.

Per últim, ens vam fixar en el procés de la dominància, abastament conegut, que introdueix diferències fenotípiques clares entre individus de la mateixa espècie, sobretot en el cas de malalties. Donat el fet que es coneixia una relació inversa entre paralogia i haploinsuficiència, per una banda, i paralogia i splicing per l'altra, sumat a la capacitat d'introduir variabilitat per part de l'splicing alternatiu, vam endegar aquest estudi amb la idea de descriure la relació entre dominància i

splicing. El resultat final ens va mostrar una independència dels dos processos, fet que ens va fer qüestionar la relació entre paralogia i splicing alternatiu. Per la resta de variables estudiades, la caracterització de la dominància va concordar amb els resultats de treballs anteriors.

7. Conclusions

A continuació, enumero les principals conclusions a les que s'ha arribat en aquest treball de tesi:

- Hem generat, per primer cop, un model quantitatiu que relaciona l'splicing alternatiu amb les divergències de seqüència proteica, de la regió *cis*-reguladora i de l'expressió gènica, que pot ser emprat per entendre les contribucions relatives d'aquests factors a la diversitat fenotípica entre espècies.
- L'splicing alternatiu pot introduir diferències entre espècies abans que les divergències de seqüència proteica i de la regió *cis*-reguladora puguin fer-ho i, per tant, contribuir a les diferències fenotípiques entre espècies.
- La relació entre l'splicing alternatiu i la divergència de l'expressió gènica és lleu, el que ens fa pensar que constitueixen dos nivells diferents de regulació.
- Els mecanismes reguladors de l'expressió gènica basats en splicing tenen una conservació molt diversa, sent la més alta entre la categoria *Variants de la mateixa funció*.
- Els esdeveniments d'splicing de la Classe 4, *Pèrdua de dominis*, presenten una baixa conservació però, tot i això, la funció es conserva, el que seria doncs un exemple de convergència funcional.
- No hem observat cap relació entre el fenomen de la dominància gènica i l'splicing alternatiu, en cap de les seves mesures.
- Hem realitzat un primer estudi comparatiu de la dominància entre humà i ratolí, en el qual hem vist que l' haploinsuficiència està més conservada.
- Per la resta de variables, el nostre treball concorda amb estudis anteriors a l'hora de relacionar els tipus de dominància amb les categories funcionals de Gene Ontology i pertinença a algun complex proteic, però no hem trobat cap relació entre SNPs al promotor i tipus de dominància.

BIBLIOGRAFIA

8. BIBLIOGRAFIA

Alonso, C. R., Wilkins, A. S., 2005. The molecular elements that underlie developmental evolution. Nat Rev Genet. 6, 709-15.

Aloy, P., Ceulemans, H., Stark, A., Russell, R. B., 2003. The relationship between sequence and interaction divergence in proteins. J Mol Biol. 332, 989-98.

Artamonova, II, Gelfand, M. S., 2007. Comparative genomics and evolution of alternative splicing: the pessimists' science. Chem Rev. 107, 3407-30.

Bagheri, H. C., 2006. Unresolved boundaries of evolutionary theory and the question of how inheritance systems evolve: 75 years of debate on the evolution of dominance. J Exp Zool B Mol Dev Evol. 306, 329-59.

Barberan-Soler, S., Lambert, N. J., Zahler, A. M., 2009. Global analysis of alternative splicing uncovers developmental regulation of nonsense-mediated decay in C. elegans. RNA. 15, 1652-60.

Belfiore, A., Frasca, F., Pandini, G., Sciacca, L., Vigneri, R., 2009. Insulin receptor isoforms and insulin receptor/insulin-like growth factor receptor hybrids in physiology and disease. Endocr Rev. 30, 586-623.

Ben-Dov, C., Hartmann, B., Lundgren, J., Valcarcel, J., 2008. Genome-wide analysis of alternative pre-mRNA splicing. J Biol Chem. 283, 1229-33.

Beno, I., Rosenthal, K., Levitine, M., Shaulov, L., Haran, T. E., 2011. Sequence-dependent cooperative binding of p53 to DNA targets and its relationship to the structural properties of the DNA targets. Nucleic Acids Res. 39, 1919-32.

Bingham, J. L., Carrigan, P. E., Miller, L. J., Srinivasan, S., 2008. Extent and diversity of human alternative splicing established by complementary database annotation and microarray analysis. OMICS. 12, 83-92.

Birchler, J. A., Riddle, N. C., Auger, D. L., Veitia, R. A., 2005. Dosage balance in gene regulation: biological implications. Trends Genet. 21, 219-26.

Birchler, J. A., Veitia, R. A., 2010. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. New Phytol. 186, 54-62.

Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E., Eppig, J. T., 2011. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. Nucleic Acids Res. 39, D842-8.

Bland, C. S., Wang, E. T., Vu, A., David, M. P., Castle, J. C., Johnson, J. M., Burge, C. B., Cooper, T. A., 2010. Global regulation of alternative splicing during myogenic differentiation. Nucleic Acids Res. 38, 7651-64.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., Taylor, J., 2010. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. Chapter 19, Unit 19 10 1-21.

Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M., Gilad, Y., 2010. Sex-specific and lineage-specific alternative splicing in primates. Genome Res. 20, 180-9.

Blencowe, B. J., 2006. Alternative splicing: new insights from global analyses. Cell. 126, 37-47.

Blencowe, B. J., Ahmad, S., Lee, L. J., 2009. Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. Genes Dev. 23, 1379-86.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365-70.

Boutz, P. L., Stoilov, P., Li, Q., Lin, C. H., Chawla, G., Ostrow, K., Shiue, L., Ares, M., Jr., Black, D. L., 2007. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. Genes Dev. 21, 1636-52.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J., Bork, P., 2000. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett. 474, 83-6.

Brinkman, B. M., 2004. Splice variants as cancer biomarkers. Clin Biochem. 37, 584-94.

Buckingham, M., Relaix, F., 2007. The role of Pax genes in the development of tissues and organs: Pax3 and Pax7 regulate muscle progenitor cell functions. Annu Rev Cell Dev Biol. 23, 645-73.

Calarco, J. A., Superina, S., O'Hanlon, D., Gabut, M., Raj, B., Pan, Q., Skalska, U., Clarke, L., Gelinas, D., van der Kooy, D., Zhen, M., Ciruna, B., Blencowe, B. J., 2009. Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. Cell. 138, 898-910.

Calarco, J. A., Xing, Y., Caceres, M., Calarco, J. P., Xiao, X., Pan, Q., Lee, C., Preuss, T. M., Blencowe, B. J., 2007. Global analysis of alternative splicing differences between humans and chimpanzees. Genes Dev. 21, 2963-75.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., 2009. AmiGO: online access to ontology and annotation data. Bioinformatics. 25, 288-9.

Carroll, S. B., 2000. Endless forms: the evolution of gene regulation and morphological diversity. Cell. 101, 577-80.

Carroll, S. B., 2005. Evolution at Two Levels: On Genes and Form. PLoS Biol. 3, e245.

Castillo-Davis, C. I., Hartl, D. L., Achaz, G., 2004. cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res. 14, 1530-6.

Castle, J. C., Zhang, C., Shah, J. K., Kulkarni, A. V., Kalsotra, A., Cooper, T. A., Johnson, J. M., 2008. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet. 40, 1416-25.

Clark, T. A., Schweitzer, A. C., Chen, T. X., Staples, M. K., Lu, G., Wang, H., Williams, A., Blume, J. E., 2007. Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biol. 8, R64.

Copley, R. R., 2008. The animal in the genome: comparative genomics and evolution. Philos Trans R Soc Lond B Biol Sci. 363, 1453-61.

Creighton, T. E., 1993. Proteins : structures and molecular properties. W.H. Freeman, New York.

Cusack, B. P., Wolfe, K. H., 2005. Changes in alternative splicing of human and mouse genes are accompanied by faster evolution of constitutive exons. Mol Biol Evol. 22, 2198-208.

Chacko, E., Ranganathan, S., 2009a. Comprehensive splicing graph analysis of alternative splicing patterns in chicken, compared to human and mouse. BMC Genomics. 10 Suppl 1, S5.

Chacko, E., Ranganathan, S., 2009b. Genome-wide analysis of alternative splicing in cow: implications in bovine as a model for human diseases. BMC Genomics. 10 Suppl 3, S11.

Chern, T. M., Paul, N., van Nimwegen, E., Zavolan, M., 2008. Computational analysis of full-length cDNAs reveals frequent coupling between transcriptional and splicing programs. DNA Res. 15, 63-72.

Chiba, H., Yamashita, R., Kinoshita, K., Nakai, K., 2008. Weak correlation between sequence conservation in promoter regions and in protein-coding regions of human-mouse orthologous gene pairs. BMC Genomics. 9, 152.

Chorley, B. N., Wang, X., Campbell, M. R., Pittman, G. S., Noureddine, M. A., Bell, D. A., 2008. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat Res. 659, 147-57.

Chow, L. T., Gelinas, R. E., Broker, T. R., Roberts, R. J., 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell. 12, 1-8.

Christofk, H. R., Vander Heiden, M. G., Harris, M. H., Ramanathan, A., Gerszten, R. E., Wei, R., Fleming, M. D., Schreiber, S. L., Cantley, L. C., 2008. The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. Nature. 452, 230-3.

Davis, M. J., Hanson, K. A., Clark, F., Fink, J. L., Zhang, F., Kasukawa, T., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Teasdale, R. D., 2006. Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units. PLoS Genet. 2, e46.

Davis, N., Yoffe, C., Raviv, S., Antes, R., Berger, J., Holzmann, S., Stoykova, A., Overbeek, P. A., Tamm, E. R., Ashery-Padan, R., 2009. Pax6 dosage requirements in iris and ciliary body differentiation. Dev Biol. 333, 132-42.

de la Grange, P., Gratadou, L., Delord, M., Dutertre, M., Auboeuf, D., 2010. Splicing factor and exon profiling across human tissues. Nucleic Acids Res. 38, 2825-38.

de Lima Morais, D. A., Harrison, P. M., 2010. Large-scale evidence for conservation of NMD candidature across mammals. PLoS One. 5, e11695.

De Sandre-Giovannoli, A., Levy, N., 2006. Altered splicing in prelamin A-associated premature aging phenotypes. Prog Mol Subcell Biol. 44, 199-232.

Denley, A., Wallace, J. C., Cosgrove, L. J., Forbes, B. E., 2003. The insulin receptor isoform exon 11- (IR-A) in cancer and other diseases: a review. Horm Metab Res. 35, 778-85.

Denoeud, F., Aury, J. M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C., Jaillon, O., Artiguenave, F., 2008. Annotating genomes with massive-scale RNA sequencing. Genome Biol. 9, R175.

Devos, D., Valencia, A., 2000. Practical limits of function prediction. Proteins. 41, 98-107.

Devos, D., Valencia, A., 2001. Intrinsic errors in genome annotation. Trends Genet. 17, 429-31.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., Hood, L., 1980. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. Cell. 20, 313-9.

Eden, E., Lipson, D., Yogev, S., Yakhini, Z., 2007. Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol. 3, e39.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., Yakhini, Z., 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 10, 48.

Eisenberg, D., Marcotte, E. M., Xenarios, I., Yeates, T. O., 2000. Protein function in the post-genomic era. Nature. 405, 823-6.

Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H., 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. 2, 953-71.

Faustino, N. A., Cooper, T. A., 2003. Pre-mRNA splicing and human disease. Genes Dev. 17, 419-37.

Fichant, G. A., 1992. Constraints acting on the exon positions of the splice site sequences and local amino acid composition of the protein. Hum Mol Genet. 1, 259-67.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., Bateman, A., 2010. The Pfam protein families database. Nucleic Acids Res. 38, D211-22.

Fisher, R. A., 1928. The Possible Modification of the Response of the Wild Type to Recurrent Mutations. The American Naturalist. 62, 115-126.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A. K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H. S., Ritchie, G. R., Ruffier, M., Schuster, M., Sobral, D., Tang, Y. A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X. M., Harrow, J., Herrero, J., Hubbard, T. J., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., Searle, S. M., 2012. Ensembl 2012. Nucleic Acids Res. 40**,** D84-90.

Floris, M., Orsini, M., Thanaraj, T. A., 2008. Splice-mediated Variants of Proteins (SpliVaP) - data and characterization of changes in signatures among protein isoforms due to alternative splicing. BMC Genomics. 9**,** 453.

Fodor, A. A., Aldrich, R. W., 2009. Convergent evolution of alternative splices at domain boundaries of the BK channel. Annu Rev Physiol. 71**,** 19-36.

Forslund, K., Sonnhammer, E. L., 2008. Predicting protein function from domain content. Bioinformatics. 24**,** 1681-7.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y., Zeng, R., Khaitovich, P., 2009. Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics. 10**,** 161.

Fuda, H., Lee, Y. C., Shimizu, C., Javitt, N. B., Strott, C. A., 2002. Mutational analysis of human hydroxysteroid sulfotransferase SULT2B1 isoforms reveals that exon 1B of the SULT2B1 gene produces cholesterol sulfotransferase, whereas exon 1A yields pregnenolone sulfotransferase. J Biol Chem. 277**,** 36161-6.

Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D., Kent, W. J., 2011. The UCSC Genome Browser database: update 2011. Nucleic Acids Res. 39**,** D876-82.

Furney, S. J., Alba, M. M., Lopez-Bigas, N., 2006. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics. 7**,** 165.

Gamba, G., 2001. Alternative splicing and diversity of renal transporters. Am J Physiol Renal Physiol. 281**,** F781-94.

Garcia-Blanco, M. A., Baraniak, A. P., Lasda, E. L., 2004. Alternative splicing in disease and therapy. Nat Biotechnol. 22**,** 535-46.

Garcia, J., Gerber, S. H., Sugita, S., Sudhof, T. C., Rizo, J., 2004. A conformational switch in the Piccolo C2A domain regulated by alternative splicing. Nat Struct Mol Biol. 11**,** 45-53.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., Nekrutenko, A., 2005. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 15**,** 1451-5.

Gilbert, W., 1978. Why genes in pieces? Nature. 271**,** 501.

Goecks, J., Nekrutenko, A., Taylor, J., 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11**,** R86.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., Ast, G., 2006. Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. Mol Cell. 22**,** 769-81.

Graveley, B. R., 2001. Alternative splicing: increasing diversity in the proteomic world. Trends in Genetics. 17**,** 100-107.

Graveley, B. R., 2008. The haplo-spliceo-transcriptome: common variations in alternative splicing in the human population. Trends Genet. 24**,** 5-7.

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., Brown, J. B., Cherbas, L., Davis, C. A., Dobin, A., Li, R., Lin, W., Malone, J. H., Mattiuzzo, N. R., Miller, D., Sturgill, D., Tuch, B. B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R. E., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J. E., Wan, K. H., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P. J., Brenner, S. E., Brent, M. R., Cherbas, P., Gingeras, T. R., Hoskins, R. A., Kaufman, T. C., Oliver, B., Celniker, S. E., 2011. The developmental transcriptome of Drosophila melanogaster. Nature. 471, 473-9.

Grosso, A. R., Gomes, A. Q., Barbosa-Morais, N. L., Caldeira, S., Thorne, N. P., Grech, G., von Lindern, M., Carmo-Fonseca, M., 2008. Tissue-specific splicing factor gene expression signatures. Nucleic Acids Res. 36, 4823-32.

Guo, W., Bharmal, S. J., Esbona, K., Greaser, M. L., 2010. Titin diversity--alternative splicing gone wild. J Biomed Biotechnol. 2010, 753675.

Gupta, S., Zink, D., Korn, B., Vingron, M., Haas, S. A., 2004. Genome wide identification and classification of alternative splicing based on EST data. Bioinformatics. 20, 2579-85.

Ha, T. S., Jeong, S. Y., Cho, S. W., Jeon, H., Roh, G. S., Choi, W. S., Park, C. S., 2000. Functional characteristics of two BKCa channel variants differentially expressed in rat brain tissues. Eur J Biochem. 267, 910-8.

Hallegger, M., Llorian, M., Smith, C. W., 2010. Alternative splicing: global insights. FEBS J. 277, 856-66.

Harr, B., Turner, L. M., 2010. Genome-wide analysis of alternative splicing evolution among Mus subspecies. Mol Ecol. 19 Suppl 1, 228-39.

Hartmann, B., Valcarcel, J., 2009. Decrypting the genome's alternative messages. Curr Opin Cell Biol. 21, 377-86.

Hegyi, H., Gerstein, M., 2001. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. Genome Res. 11, 1632-40.

Hegyi, H., Kalmar, L., Horvath, T., Tompa, P., 2011. Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder. Nucleic Acids Res. 39, 1208-19.

Herai, R. H., Yamagishi, M. E., 2010. Detection of human interchromosomal trans-splicing in sequence databanks. Brief Bioinform. 11, 198-209.

Hoekstra, H. E., Coyne, J. A., 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution. 61, 995-1016.

Hughes, A. L., Friedman, R., 2008. Alternative splicing, gene duplication and connectivity in the genetic interaction network of the nematode worm Caenorhabditis elegans. Genetica. 134, 181-6.

Hull, J., Campino, S., Rowlands, K., Chan, M. S., Copley, R. R., Taylor, M. S., Rockett, K., Elvidge, G., Keating, B., Knight, J., Kwiatkowski, D., 2007. Identification of common genetic variation that modulates alternative splicing. PLoS Genet. 3, e99.

Hymowitz, S. G., Compaan, D. M., Yan, M., Wallweber, H. J., Dixit, V. M., Starovasnik, M. A., de Vos, A. M., 2003. The crystal structures of EDA-A1 and EDA-A2: splice variants with distinct receptor specificity. Structure. 11, 1513-20.

Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., Barrero, R. A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M., Yura, K., Miyazaki, S., Ikeo, K., Homma, K., Kasprzyk, A., Nishikawa, T., Hirakawa, M., Thierry-Mieg, J., Thierry-Mieg, D., Ashurst, J., Jia, L., Nakao, M., Thomas, M. A., Mulder, N., Karavidopoulou, Y., Jin, L., Kim, S., Yasuda, T., Lenhard, B., Eveno, E., Yamasaki, C., Takeda, J., Gough, C., Hilton, P., Fujii, Y., Sakai, H., Tanaka, S., Amid, C., Bellgard, M., Bonaldo Mde, F., Bono, H., Bromberg, S. K., Brookes, A. J., Bruford, E., Carninci, P., Chelala, C., Couillault, C., de Souza, S. J., Debily, M. A., Devignes, M. D., Dubchak, I., Endo, T., Estreicher, A., Eyras, E., Fukami-Kobayashi, K., Gopinath, G. R., Graudens, E., Hahn, Y., Han, M., Han, Z. G., Hanada, K., Hanaoka, H., Harada, E., Hashimoto, K., Hinz, U.,

Hirai, M., Hishiki, T., Hopkinson, I., Imbeaud, S., Inoko, H., Kanapin, A., Kaneko, Y., Kasukawa, T., Kelso, J., Kersey, P., Kikuno, R., Kimura, K., Korn, B., Kuryshev, V., Makalowska, I., Makino, T., Mano, S., Mariage-Samson, R., Mashima, J., Matsuda, H., Mewes, H. W., Minoshima, S., Nagai, K., Nagasaki, H., Nagata, N., Nigam, R., Ogasawara, O., Ohara, O., Ohtsubo, M., Okada, N., Okido, T., Oota, S., Ota, M., Ota, T., Otsuki, T., Piatier-Tonneau, D., Poustka, A., Ren, S. X., Saitou, N., Sakai, K., Sakamoto, S., Sakate, R., Schupp, I., Servant, F., Sherry, S., Shiba, R., Shimizu, N., Shimoyama, M., Simpson, A. J., Soares, B., Steward, C., Suwa, M., Suzuki, M., Takahashi, A., Tamiya, G., Tanaka, H., Taylor, T., Terwilliger, J. D., Unneberg, P., Veeramachaneni, V., Watanabe, S., Wilming, L., Yasuda, N., Yoo, H. S., Stodolsky, M., Makalowski, W., Go, M., Nakai, K., Takagi, T., Kanehisa, M., Sakaki, Y., Quackenbush, J., Okazaki, Y., Hayashizaki, Y., Hide, W., Chakraborty, R., Nishikawa, K., Sugawara, H., Tateno, Y., Chen, Z., Oishi, M., Tonellato, P., Apweiler, R., Okubo, K., Wagner, L., Wiemann, S., Strausberg, R. L., Isogai, T., Auffray, C., Nomura, N., Gojobori, T., Sugano, S., 2004. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol. 2, e162.

Irimia, M., Rukov, J. L., Penny, D., Garcia-Fernandez, J., Vinther, J., Roy, S. W., 2008. Widespread evolutionary conservation of alternatively spliced exons in Caenorhabditis. Mol Biol Evol. 25, 375-82.

Irimia, M., Rukov, J. L., Roy, S. W., Vinther, J., Garcia-Fernandez, J., 2009. Quantitative regulation of alternative splicing in evolution and development. Bioessays. 31, 40-50.

Janicke, R. U., Graupner, V., Budach, W., Essmann, F., 2009. The do's and don'ts of p53 isoforms. Biol Chem. 390, 951-63.

Jin, L., Kryukov, K., Clemente, J. C., Komiyama, T., Suzuki, Y., Imanishi, T., Ikeo, K., Gojobori, T., 2008. The evolutionary relationship between gene duplication and alternative splicing. Gene. 427, 19-31.

Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., Shoemaker, D. D., 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science. 302, 2141-4.

Johnson, M. B., Kawasawa, Y. I., Mason, C. E., Krsnik, Z., Coppola, G., Bogdanovic, D., Geschwind, D. H., Mane, S. M., State, M. W., Sestan, N., 2009. Functional and evolutionary insights into human brain development through global transcriptome analysis. Neuron. 62, 494-509.

Jones, D. T., 2007. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics. 23, 538-44.

Kacser, H., Burns, J. A., 1981. The molecular basis of dominance. Genetics. 97, 639-66.

Kalsotra, A., Wang, K., Li, P. F., Cooper, T. A., 2010. MicroRNAs coordinate an alternative splicing network during mouse postnatal heart development. Genes Dev. 24, 653-8.

Kalsotra, A., Xiao, X., Ward, A. J., Castle, J. C., Johnson, J. M., Burge, C. B., Cooper, T. A., 2008. A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. Proc Natl Acad Sci U S A. 105, 20333-8.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., Kent, W. J., 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32, D493-6.

Keren, H., Lev-Maor, G., Ast, G., 2010. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet. 11, 345-55.

Khoury, M. P., Bourdon, J. C., 2010. The isoforms of the p53 protein. Cold Spring Harb Perspect Biol. 2, a000927.

Kim, E., Goren, A., Ast, G., 2008. Alternative splicing: current perspectives. Bioessays. 30, 38-47.

Kim, E., Magen, A., Ast, G., 2007. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 35, 125-31.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X. S., Ahringer, J., 2009. Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet. 41, 376-81.
Kondrashov, F. A., Koonin, E. V., 2001. Origin of alternative splicing by tandem exon duplication. Hum Mol Genet. 10, 2661-9.
Kondrashov, F. A., Koonin, E. V., 2003. Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. Trends Genet. 19, 115-9.
Kondrashov, F. A., Koonin, E. V., 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. Trends Genet. 20, 287-90.
Kopelman, N. M., Lancet, D., Yanai, I., 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat Genet. 37, 588-9.
Koralewski, T. E., Krutovsky, K. V., 2011. Evolution of exon-intron structure and alternative splicing. PLoS One. 6, e18055.
Kornblihtt, A. R., 2006. Chromatin, transcript elongation and alternative splicing. Nat Struct Mol Biol. 13, 5-7.
Kornblihtt, A. R., 2007. Coupling transcription and alternative splicing. Adv Exp Med Biol. 623, 175-89.
Kornblihtt, A. R., de la Mata, M., Fededa, J. P., Munoz, M. J., Nogues, G., 2004. Multiple links between transcription and splicing. RNA. 10, 1489-98.
Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S., Sunyaev, S., 2003. Increase of functional diversity by alternative splicing. Trends Genet. 19, 124-8.
Kwan, T., Benovoy, D., Dias, C., Gurd, S., Provencher, C., Beaulieu, P., Hudson, T. J., Sladek, R., Majewski, J., 2008. Genome-wide analysis of transcript isoform variation in humans. Nat Genet. 40, 225-31.
Kwan, T., Benovoy, D., Dias, C., Gurd, S., Serre, D., Zuzan, H., Clark, T. A., Schweitzer, A., Staples, M. K., Wang, H., Blume, J. E., Hudson, T. J., Sladek, R., Majewski, J., 2007. Heritability of alternative splicing in the human genome. Genome Res. 17, 1210-8.
Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson,

M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., 2001. Initial sequencing and analysis of the human genome. Nature. 409**,** 860-921.

Landry, J. R., Mager, D. L., Wilhelm, B. T., 2003. Complex controls: the role of alternative promoters in mammalian genomes. Trends Genet. 19**,** 640-8.

Lareau, L. F., Brooks, A. N., Soergel, D. A., Meng, Q., Brenner, S. E., 2007. The coupling of alternative splicing and nonsense-mediated mRNA decay. Adv Exp Med Biol. 623**,** 190-211.

Latchman, D. (Ed.) 2008. Gene Regulation: A Eukaryotic Perspective. Springer.

Latchman, D. S., 1990. Eukaryotic transcription factors. Biochem J. 270**,** 281-9.

Leeman, J. R., Gilmore, T. D., 2008. Alternative splicing in the NF-kappaB signaling pathway. Gene. 423**,** 97-107.

Lemos, B., Bettencourt, B. R., Meiklejohn, C. D., Hartl, D. L., 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. 22**,** 1345-54.

Levy, E. D., Boeri Erba, E., Robinson, C. V., Teichmann, S. A., 2008. Assembly reflects evolution of protein complexes. Nature. 453**,** 1262-5.

Lewis, B. P., Green, R. E., Brenner, S. E., 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A. 100**,** 189-92.

Li, H., Wang, J., Ma, X., Sklar, J., 2009a. Gene fusions and RNA trans-splicing in normal and neoplastic human cells. Cell Cycle. 8**,** 218-22.

Li, J. B., Levanon, E. Y., Yoon, J. K., Aach, J., Xie, B., Leproust, E., Zhang, K., Gao, Y., Church, G. M., 2009b. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science. 324**,** 1210-3.

Li, M., Wang, I. X., Li, Y., Bruzel, A., Richards, A. L., Toung, J. M., Cheung, V. G., 2011. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 333**,** 53-8.

Li, W., Jaroszewski, L., Godzik, A., 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics. 17**,** 282-3.

Liao, B. Y., Zhang, J., 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol. 23**,** 530-40.

Liu, S., Altman, R. B., 2003. Large scale study of protein domain distribution in the context of alternative splicing. Nucleic Acids Res. 31**,** 4828-35.

Lois, S., Blanco, N., Martinez-Balbas, M., de la Cruz, X., 2007. The functional modulation of epigenetic regulators by alternative splicing. BMC Genomics. 8**,** 252.

Lopez-Bigas, N., Audit, B., Ouzounis, C., Parra, G., Guigo, R., 2005. Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett. 579**,** 1900-3.

Lopez-Bigas, N., Blencowe, B. J., Ouzounis, C. A., 2006. Highly consistent patterns for inherited human diseases at the molecular level. Bioinformatics. 22**,** 269-77.

Lopez, A. J., 1995. Developmental role of transcription factor isoforms generated by alternative splicing. Dev Biol. 172, 396-411.
Lopez, A. J., 1998. Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. Annu Rev Genet. 32, 279-305.
Lutz, C. S., 2008. Alternative polyadenylation: a twist on mRNA 3' end formation. ACS Chem Biol. 3, 609-17.
Lynch, M., Conery, J. S., 2000. The evolutionary fate and consequences of duplicate genes. Science. 290, 1151-5.
Magness, C. L., Fellin, P. C., Thomas, M. J., Korth, M. J., Agy, M. B., Proll, S. C., Fitzgibbon, M., Scherer, C. A., Miner, D. G., Katze, M. G., Iadonato, S. P., 2005. Analysis of the Macaca mulatta transcriptome and the sequence divergence between Macaca and human. Genome Biol. 6, R60.
Maher, C. A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N., Chinnaiyan, A. M., 2009. Transcriptome sequencing to detect gene fusions in cancer. Nature. 458, 97-101.
Majoros, W. H., Ohler, U., 2007. Spatial preferences of microRNA targets in 3' untranslated regions. BMC Genomics. 8, 152.
Makeyev, E. V., Zhang, J., Carrasco, M. A., Maniatis, T., 2007. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. Mol Cell. 27, 435-48.
Marcel, V., Hainaut, P., 2009. p53 isoforms - a conspiracy to kidnap p53 tumor suppressor activity? Cell Mol Life Sci. 66, 391-406.
Marchler-Bauer, A., Bryant, S. H., 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 32, W327-31.
Marden, J. H., 2008. Quantitative and evolutionary biology of alternative splicing: how changing the mix of alternative transcripts affects phenotypic plasticity and reaction norms. Heredity. 100, 111-20.
Marguerat, S., Bahler, J., 2010. RNA-seq: from technology to biology. Cell Mol Life Sci. 67, 569-79.
Masotti, C., Armelin-Correa, L. M., Splendore, A., Lin, C. J., Barbosa, A., Sogayar, M. C., Passos-Bueno, M. R., 2005. A functional SNP in the promoter region of TCOF1 is associated with reduced gene expression and YY1 DNA-protein interaction. Gene. 359, 44-52.
Matlin, A. J., Clark, F., Smith, C. W., 2005. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 6, 386-98.
Medina, I., Carbonell, J., Pulido, L., Madeira, S. C., Goetz, S., Conesa, A., Tarraga, J., Pascual-Montano, A., Nogales-Cadenas, R., Santoyo, J., Garcia, F., Marba, M., Montaner, D., Dopazo, J., 2010. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res. 38, W210-3.
Medina, M. W., Gao, F., Naidoo, D., Rudel, L. L., Temel, R. E., McDaniel, A. L., Marshall, S. M., Krauss, R. M., 2011. Coordinately regulated alternative splicing of genes involved in cholesterol biosynthesis and uptake. PLoS One. 6, e19420.
Melamud, E., Moult, J., 2009. Stochastic noise in splicing machinery. Nucleic Acids Res. 37, 4873-86.
Meshorer, E., Soreq, H., 2002. Pre-mRNA splicing modulations in senescence. Aging Cell. 1, 10-6.
Meyer, I. M., Durbin, R., 2004. Gene structure conservation aids similarity based gene prediction. Nucleic Acids Res. 32, 776-83.
Modrek, B., Lee, C. J., 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet. 34, 177-80.
Modrek, B., Resch, A., Grasso, C., Lee, C., 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res. 29, 2850-9.

Mollet, I. G., Ben-Dov, C., Felicio-Silva, D., Grosso, A. R., Eleuterio, P., Alves, R., Staller, R., Silva, T. S., Carmo-Fonseca, M., 2010. Unconstrained mining of transcript data reveals increased alternative splicing complexity in the human transcriptome. Nucleic Acids Res. 38, 4740-54.
Mudge, J. M., Frankish, A., Fernandez-Banet, J., Alioto, T., Derrien, T., Howald, C., Reymond, A., Guigo, R., Hubbard, T., Harrow, J., 2011. The origins, evolution and functional potential of alternative splicing in vertebrates. Mol Biol Evol.
Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M., Gotoh, O., 2006. Automated classification of alternative splicing and transcriptional initiation and construction of visual database of classified patterns. Bioinformatics. 22, 1211-6.
Needleman, S. B., Wunsch, C. D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 48, 443-53.
Nembaware, V., Wolfe, K. H., Bettoni, F., Kelso, J., Seoighe, C., 2004. Allele-specific transcript isoforms in human. FEBS Lett. 577, 233-8.
Nilsen, T. W., Graveley, B. R., 2010. Expansion of the eukaryotic proteome by alternative splicing. Nature. 463, 457-463.
Nurtdinov, R. N., Artamonova, II, Mironov, A. A., Gelfand, M. S., 2003. Low conservation of alternative splicing patterns in the human and mouse genomes. Hum Mol Genet. 12, 1313-20.
Nurtdinov, R. N., Neverov, A. D., Favorov, A. V., Mironov, A. A., Gelfand, M. S., 2007. Conserved and species-specific alternative splicing in mammalian genomes. BMC Evol Biol. 7, 249.
Nygard, A. B., Cirera, S., Gilchrist, M. J., Gorodkin, J., Jorgensen, C. B., Fredholm, M., 2010. A study of alternative splicing in the pig. BMC Res Notes. 3, 123.
Ohler, U., Shomron, N., Burge, C. B., 2005. Recognition of unknown conserved alternatively spliced exons. PLoS Comput Biol. 1, 113-22.
Ohno, S., 1970. Evolution by gene duplication. Springer-Verlag, Berlin, New York,.
Osada, N., Mano, S., Gojobori, J., 2009. Quantifying dominance and deleterious effect on human disease genes. Proc Natl Acad Sci U S A. 106, 841-6.
Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D. N., Roopra, S., Frings, O., Sonnhammer, E. L., 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38, D196-203.
Ozsolak, F., Milos, P. M., 2011. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 12, 87-98.
Pan, Q., Bakowski, M. A., Morris, Q., Zhang, W., Frey, B. J., Hughes, T. R., Blencowe, B. J., 2005. Alternative splicing of conserved exons is frequently species-specific in human and mouse. Trends Genet. 21, 73-7.
Pan, Q., Shai, O., Lee, L. J., Frey, B. J., Blencowe, B. J., 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 40, 1413-5.
Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A. L., Mohammad, N., Babak, T., Siu, H., Hughes, T. R., Morris, Q. D., Frey, B. J., Blencowe, B. J., 2004. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell. 16, 929-41.
Pandit, S., Wang, D., Fu, X. D., 2008. Functional integration of transcriptional and RNA processing machineries. Curr Opin Cell Biol. 20, 260-5.
Papp, B., Pal, C., Hurst, L. D., 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature. 424, 194-7.
Pavesi, G., Zambelli, F., Caggese, C., Pesole, G., 2008. Exalign: a new method for comparative analysis of exon-intron gene structures. Nucleic Acids Res. 36, e47.
Power, K. A., McRedmond, J. P., de Stefani, A., Gallagher, W. M., Gaora, P. O., 2009. High-throughput proteomics detection of novel splice isoforms in human platelets. PLoS One. 4, e5001.

Prokunina, L., Alarcon-Riquelme, M. E., 2004. Regulatory SNPs in complex diseases: their identification and functional validation. Expert Rev Mol Med. 6, 1-15.
Pruitt, K. D., Tatusova, T., Maglott, D. R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35, D61-5.
Przytycka, T. M., Singh, M., Slonim, D. K., 2010. Toward the dynamic interactome: it's about time. Brief Bioinform. 11, 15-29.
Qian, W., Zhang, J., 2008. Gene dosage and gene duplicability. Genetics. 179, 2319-24.
Robichaud, G. A., Perreault, J. P., Ouellette, R. J., 2008. Development of an isoform-specific gene suppression system: the study of the human Pax-5B transcriptional element. Nucleic Acids Res. 36, 4609-20.
Romero, P. R., Zaidi, S., Fang, Y. Y., Uversky, V. N., Radivojac, P., Oldfield, C. J., Cortese, M. S., Sickmeier, M., LeGall, T., Obradovic, Z., Dunker, A. K., 2006. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc Natl Acad Sci U S A. 103, 8390-5.
Rost, B., 2002. Enzyme function less conserved than anticipated. J Mol Biol. 318, 595-608.
Roux, J., Robinson-Rechavi, M., 2011. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. Genome Res. 21, 357-63.
Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., Mewes, H. W., 2010. CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res. 38, D497-501.
Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., Burge, C. B., 2008. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science. 320, 1643-7.
Schaal, T. D., Maniatis, T., 1999. Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. Mol Cell Biol. 19, 261-73.
Schwartz, S., Oren, R., Ast, G., 2011. Detection and removal of biases in the analysis of next-generation sequencing reads. PLoS One. 6, e16685.
Shabalina, S. A., Spiridonov, A. N., Spiridonov, N. A., Koonin, E. V., 2010. Connections between alternative transcription and alternative splicing in mammals. Genome Biol Evol. 2, 791-9.
Shastry, B. S., 2009. SNPs: impact on gene function and phenotype. Methods Mol Biol. 578, 3-22.
Shepard, P. J., Hertel, K. J., 2008. Conserved RNA secondary structures promote alternative splicing. RNA. 14, 1463-9.
Short, S., Holland, L. Z., 2008. The evolution of alternative splicing in the Pax family: the view from the Basal chordate amphioxus. J Mol Evol. 66, 605-20.
Smith, C. W., Valcarcel, J., 2000. Alternative pre-mRNA splicing: the logic of combinatorial control. Trends Biochem Sci. 25, 381-8.
Sorek, R., Shamir, R., Ast, G., 2004. How prevalent is functional alternative splicing in the human genome? Trends Genet. 20, 68-71.
Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A., Soreq, H., 2005. Function of alternative splicing. Gene. 344, 1-20.
Stetefeld, J., Ruegg, M. A., 2005. Structural and functional diversity generated by alternative mRNA splicing. Trends Biochem Sci. 30, 515-21.
Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M. F., Rifkin, S. A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P. E., Bussemaker, H. J., White, K. P., 2004. A gene expression map for the euchromatic genome of Drosophila melanogaster. Science. 306, 655-60.
Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., Hogenesch, J. B., 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 101, 6062-7.

Su, Z., Wang, J., Yu, J., Huang, X., Gu, X., 2006. Evolution of alternative splicing after gene duplication. Genome Res. 16, 182-9.
Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., Yaspo, M. L., 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 321, 956-60.
Takeda, J., Suzuki, Y., Sakate, R., Sato, Y., Seki, M., Irie, T., Takeuchi, N., Ueda, T., Nakao, M., Sugano, S., Gojobori, T., Imanishi, T., 2008. Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. Nucleic Acids Res. 36, 6386-95.
Talavera, D., Hospital, A., Orozco, M., de la Cruz, X., 2007a. A procedure for identifying homologous alternative splicing events. BMC Bioinformatics. 8, 260.
Talavera, D., Orozco, M., de la Cruz, X., 2009. Alternative splicing of transcription factors' genes: beyond the increase of proteome diversity. Comp Funct Genomics. 905894.
Talavera, D., Vogel, C., Orozco, M., Teichmann, S. A., de la Cruz, X., 2007b. The (in)dependence of alternative splicing and gene duplication. PLoS Comput Biol. 3, e33.
Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., Surani, M. A., 2009. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 6, 377-82.
Tanner, S., Shen, Z., Ng, J., Florea, L., Guigo, R., Briggs, S. P., Bafna, V., 2007. Improving gene annotation using peptide mass spectrometry. Genome Res. 17, 231-9.
Tao, S., Sampath, K., 2010. Alternative splicing of SMADs in differentiation and tissue homeostasis. Dev Growth Differ. 52, 335-42.
Tazi, J., Bakkour, N., Stamm, S., 2009. Alternative splicing and disease. Biochim Biophys Acta. 1792, 14-26.
Teichmann, S. A., Veitia, R. A., 2004. Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. Genetics. 167, 2121-5.
Thanaraj, T. A., Clark, F., Muilu, J., 2003. Conservation of human alternative splice events in mouse. Nucleic Acids Res. 31, 2544-52.
The UniProt Consortium, 2010. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. 38, D142-8.
Thompson, J. D., Higgins, D. G., Gibson, T. J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673-80.
Tian, W., Skolnick, J., 2003. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. 333, 863-82.
Tirosh, I., Weinberger, A., Bezalel, D., Kaganovich, M., Barkai, N., 2008. On the relation between promoter divergence and gene expression evolution. Mol Syst Biol. 4, 159.
Tollervey, J. R., Wang, Z., Hortobagyi, T., Witten, J. T., Zarnack, K., Kayikci, M., Clark, T. A., Schweitzer, A. C., Rot, G., Curk, T., Zupan, B., Rogelj, B., Shaw, C. E., Ule, J., 2011. Analysis of alternative splicing associated with aging and neurodegeneration in the human brain. Genome Res.
Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28, 511-5.
Tress, M. L., Bodenmiller, B., Aebersold, R., Valencia, A., 2008. Proteomics studies confirm the presence of alternative protein isoforms on a large scale. Genome Biol. 9, R162.
Tress, M. L., Martelli, P. L., Frankish, A., Reeves, G. A., Wesselink, J. J., Yeats, C., Olason, P. I., Albrecht, M., Hegyi, H., Giorgetti, A., Raimondo, D., Lagarde, J., Laskowski, R. A.,

Lopez, G., Sadowski, M. I., Watson, J. D., Fariselli, P., Rossi, I., Nagy, A., Kai, W., Storling, Z., Orsini, M., Assenov, Y., Blankenburg, H., Huthmacher, C., Ramirez, F., Schlicker, A., Denoeud, F., Jones, P., Kerrien, S., Orchard, S., Antonarakis, S. E., Reymond, A., Birney, E., Brunak, S., Casadio, R., Guigo, R., Harrow, J., Hermjakob, H., Jones, D. T., Lengauer, T., Orengo, C. A., Patthy, L., Thornton, J. M., Tramontano, A., Valencia, A., 2007. The implications of alternative splicing in the ENCODE protein complement. Proc Natl Acad Sci U S A. 104, 5495-500.

Trinick, J., Tskhovrebova, L., 1999. Titin: a molecular control freak. Trends Cell Biol. 9, 377-80.

Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J. S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., Zeeberg, B. R., Kane, D., Weinstein, J. N., Blume, J., Darnell, R. B., 2005. Nova regulates brain-specific splicing to shape the synapse. Nat Genet. 37, 844-52.

Valen, E., Pascarella, G., Chalk, A., Maeda, N., Kojima, M., Kawazu, C., Murata, M., Nishiyori, H., Lazarevic, D., Motti, D., Marstrand, T. T., Tang, M. H., Zhao, X., Krogh, A., Winther, O., Arakawa, T., Kawai, J., Wells, C., Daub, C., Harbers, M., Hayashizaki, Y., Gustincich, S., Sandelin, A., Carninci, P., 2009. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res. 19, 255-65.

Valenzuela, A., Talavera, D., Orozco, M., de la Cruz, X., 2004. Alternative splicing mechanisms for the modulation of protein function: conservation between human and other species. J Mol Biol. 335, 495-502.

van der Vaart, M., Schaaf, M. J., 2009. Naturally occurring C-terminal splice variants of nuclear receptors. Nucl Recept Signal. 7, e007.

Veitia, R. A., 2002. Exploring the etiology of haploinsufficiency. Bioessays. 24, 175-84.

Veitia, R. A., 2005. Gene dosage balance: deletions, duplications and dominance. Trends Genet. 21, 33-5.

Veitia, R. A., Birchler, J. A., 2010. Dominance and gene dosage balance in health and disease: why levels matter! J Pathol. 220, 174-85.

Venables, J. P., 2004. Aberrant and alternative splicing in cancer. Cancer Res. 64, 7647-54.

Venables, J. P., 2006. Unbalanced alternative splicing and its significance in cancer. Bioessays. 28, 378-86.

Venables, J. P., Klinck, R., Koh, C., Gervais-Bird, J., Bramard, A., Inkel, L., Durand, M., Couture, S., Froehlich, U., Lapointe, E., Lucier, J. F., Thibault, P., Rancourt, C., Tremblay, K., Prinos, P., Chabot, B., Elela, S. A., 2009. Cancer-associated regulation of alternative splicing. Nat Struct Mol Biol. 16, 670-6.

Wall, D. P., Hirsh, A. E., Fraser, H. B., Kumm, J., Giaever, G., Eisen, M. B., Feldman, M. W., 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A. 102, 5483-8.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., Burge, C. B., 2008. Alternative isoform regulation in human tissue transcriptomes. Nature. 456, 470-6.

Wang, Y., Rekaya, R., 2009. A comprehensive analysis of gene expression evolution between humans and mice. Evol Bioinform Online. 5, 81-90.

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., Jones, D. T., 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol. 337, 635-45.

Watanabe, H., Fujiyama, A., Hattori, M., Taylor, T. D., Toyoda, A., Kuroki, Y., Noguchi, H., BenKahla, A., Lehrach, H., Sudbrak, R., Kube, M., Taenzer, S., Galgoczy, P., Platzer, M., Scharfe, M., Nordsiek, G., Blocker, H., Hellmann, I., Khaitovich, P., Paabo, S., Reinhardt, R., Zheng, H. J., Zhang, X. L., Zhu, G. F., Wang, B. F., Fu, G., Ren, S. X., Zhao, G. P., Chen, Z., Lee, Y. S., Cheong, J. E., Choi, S. H., Wu, K. M., Liu, T. T., Hsiao, K. J., Tsai, S. F., Kim, C. G., S, O. O., Kitano, T., Kohara, Y., Saitou, N., Park, H. S., Wang, S. Y., Yaspo, M. L., Sakaki, Y., 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature. 429, 382-8.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Von Niederhausern, A. C., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S. P., Zdobnov, E. M., Zody, M. C., Lander, E. S., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420**,** 520-62.

Wetterbom, A., Gyllensten, U., Cavelier, L., Bergstrom, T. F., 2009. Genome-wide analysis of chimpanzee genes with premature termination codons. BMC Genomics. 10**,** 56.

Whamond, G. S., Thornton, J. M., 2006. An analysis of intron positions in relation to nucleotides, amino acids, and protein secondary structure. J Mol Biol. 359**,** 238-47.

Wray, G. A., 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 8**,** 206-16.

Wright, S., 1934. Physiological and Evolutionary Theories of Dominance. The American Naturalist. 68**,** 24-53.

Xiao, X., Lee, J. H., 2010. Systems analysis of alternative splicing and its regulation. Wiley Interdiscip Rev Syst Biol Med. 2**,** 550-65.

Xie, H., Zhu, W. Y., Wasserman, A., Grebinskiy, V., Olson, A., Mintz, L., 2002. Computational analysis of alternative splicing using EST tissue information. Genomics. 80**,** 326-30.

Xing, Y., Lee, C., 2006. Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes. Nat Rev Genet. 7**,** 499-509.

Xing, Y., Xu, Q., Lee, C., 2003. Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. FEBS Lett. 555**,** 572-8.

Yan, J., Marr, T. G., 2005. Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. Genome Res. 15**,** 369-75.

Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T., Burge, C. B., 2005. Identification and analysis of alternative splicing events conserved in human and mouse. Proc Natl Acad Sci U S A. 102**,** 2850-5.
Yi, Q., Tang, L., 2011. Alternative Spliced Variants as Biomarkers of Colorectal Cancer. Curr Drug Metab.
Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J. D., Bertin, N., Chung, S., Vidal, M., Gerstein, M., 2004. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res. 14**,** 1107-18.
Yu, Y., Maroney, P. A., Denker, J. A., Zhang, X. H., Dybkov, O., Luhrmann, R., Jankowsky, E., Chasin, L. A., Nilsen, T. W., 2008. Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. Cell. 135**,** 1224-36.
Zambelli, F., Pavesi, G., Gissi, C., Horner, D. S., Pesole, G., 2010. Assessment of orthologous splicing isoforms in human and mouse orthologous genes. BMC Genomics. 11**,** 534.
Zhang, C., Zhang, Z., Castle, J., Sun, S., Johnson, J., Krainer, A. R., Zhang, M. Q., 2008. Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. Genes Dev. 22**,** 2550-63.
Zhang, M. Q., 2007. Computational analyses of eukaryotic promoters. BMC Bioinformatics. 8 Suppl 6**,** S3.