UNIVERSIDAD AUTÓNOMA DE MADRID
FACULTAD DE CIENCIAS
DEPARTAMENTO DE BIOLOGÍA MOLECULAR

Estudio comparativo del splicing alternativo del gen NCR3 en diferentes especies de mamíferos y sus posibles implicaciones funcionales.

Memoria presentada por:

Alberto Rastrojo Lastras

para optar al grado de Doctor en Ciencias por la Universidad Autónoma de Madrid.

Trabajo dirigido por **Begoña Aguado Orea** y realizado en el Centro de Biología Molecular "Severo Ochoa" (UAM-CSIC).

Noviembre, 2013

Summary

Alternative splicing (AS) is a major source of transcriptome and proteome diversity in higher eukaryotes allowing the generation of several structurally and functionally distinct mRNAs and protein isoforms from a single gene. Differential AS variants expression has been implicated in tissue and cellular differentiation, which in an evolutionary context could account for the phenotypic divergence of mammals that share a common repertoire of genes. On the other hand, the miss-regulation of AS is one of the mayor sources of human disease. However, there are not many detailed comparative analyses showing the alternative splice forms generated from particular genes among different organisms. The work presented here is a deep study of the splice variants expressed by the "*Natural Cytotoxicity Receptor 3"* (*NCR3*) gene and its comparative analysis among 13 mammals.

NCR3 is a member of the NCR family, which represents the major NK cells triggering receptors. It has been involved in the recognition and killing of tumoral and infected cells, and in the maturation of dendritic cells. By using a combination of nested RT-PCR and RNA-seq analysis, it was possible to detect the expression of several *NCR3* transcripts in all the analysed mammalian species, except in *Mus musculus*, where *NCR3* is a pseudogen. It was observed an increase in the number of coding variants in primates, mainly by the internal splicing of exon 2 and the presence of exons 4II and 4III, which are only expressed in higher primates (*Hominoidea*). In contrast, the diversity of non-coding variants is similar among all analysed species, although there are some conserved ones, which could have a potential regulatory role. The nine human splice variants, six coding and three non-coding, were only detected at high expression levels in immune-related tissues, being the coding *A*, *B* y *C* the most abundant ones.

The predicted human protein isoforms are potential transmembrane type I receptors, which extracellular domains are predicted to be Immunoglobulin type V (IgV) or type C (IgC). The interaction of these potential NCR3 extracellular domains with the ligand B7-H6 was analysed using flow citometry assays. Firstly, they were over-expressed in insect cells finding that defective glycosylation avoids binding. In a second approach, it was used a mammalian system detecting only specific binding of IgV domain to B7H6, indicating that B7H6 is not a ligand for IgC containing isoforms.

In conclusion, all the presented data suggest a potential role of AS in the NCR3 functions and immune system regulation and evolution.

Índice

Abreviaturas

AANE	Aminoácidos no esenciales
ATP	Adenosina-triphosphato
BSA	Albúmina sérica bovina
cDNA	DNA complementario
Ct	Ciclo umbral
CTD	Dominio C-terminal de la RNA polimerasa II
DAPI	*4',6-diamidino-2-phenylindole*
DMEM	*Dulbecco's Modified Eagle Medium*
DMSO	*Dimethyl sulfoxide*
DNA	Ácido desoxirribonucleico
dNTP	Desoxirribonucleótidos
ds cDNA	cDNA de doble cadena
DTT	*Dithiothreitol*
ECL	Quimioluminescencia mejora
EDTA	Ácido etilendinitrilotetracético
EGFP	Proteína verde fluorescente potenciada
FACs	Citometría de flujo
FBS	Suero fetal bovino
g	Gramo
GFP	Proteína verde fluorescente
HEPES	*2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid*
hIgG1	Inmunoglobulina humana tipo IgG1
hIgG1-Fc	Dominio FC de las hIgG1
hnRNP	*Heterogeneous nuclear ribonucleoprotein*
HRP	Peroxidasa del rábano
IF	Inmunofluorescencia
IPTG	*Isopropyl β-D-1-thiogalactopyranoside*
Kb	Kilobase (1000 pb)
kDa	Kilodalton
Ma	Millones de años
Mb	Megabase (1000 Kb)
MCS	Sitio de clonaje múltiple
MHC	Complejo Mayor de Histocompatibilidad
NK	Natural Killer
O/N	Toda la noche
ORF	Fase abierta de lectura
pb	Pares de bases
PBMC	Célula mononuclear de sangre periférica
PBS	Tampón fosfaro salino
PCa	Método del fosfato cálico
PCR	Reacción en cadena de la polimerasa
PFA	Paraformaldehido
qPCR	PCR cuantitativa
RNA	Ácido ribonucleico
RNA Pol II	RNA polimerasa II
RPKMs	*Reads Per Kilobase per Million mapped reads*
rpm	Revoluciones por minuto
RPMI	*Roswell Park Memorial Institute medium*
RT	Temperatura ambiente
SDS	*Sodium dodecyl sulfate*
SF1	Factor de splicing 1
snRNP	*Small nuclear ribonucleoproteins*
SR	*Serine-arginine rich protein family*
TEMED	*N,N,N′,N′-Tetramethylethane-1,2-diamine*
U	Unidades de actividad enzimática
UTR	Untranslated region
WB	Western Blot
X-gal	*5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside*

1. Introducción

Introducción

El descubrimiento de la compleja arquitectura de los genes de los organismos eucariotas fue un gran avance en la comprensión de los mecanismos de regulación de la expresión génica. A diferencia de los organismos procariotas cuyos RNA mensajeros son co-lineales con la correspondiente secuencia genómica, en los organismos eucariotas la información contenida en los genes está interrumpida por fragmentos de secuencia no codificante denominados intrones (Figura 1-1). Durante la transcripción los intrones deben ser eliminados de los mensajeros inmaduros o pre-mensajeros, uniéndose los fragmentos de secuencia (exones) que portan la información necesaria para la síntesis de proteínas, en un proceso denominado maduración o "*splicing*" (Figura 1-1). La diferencia de longitud de los mensajeros observada entre el núcleo y el citoplasma permitía intuir la existencia de este proceso de maduración en organismos eucariotas (Sharp, 2005). Sin embargo, el origen de tal diferencia no se resolvió hasta que en la década de los 70s, a través de híbridos DNA-RNA, se detectó que diferentes fragmentos de un mismo mensajero maduro de adenovirus mapeaban en distintas regiones alejadas entre sí del genoma del virus, evidenciándose la presencia de un fragmento de secuencia en el DNA genómico que no estaba presente en el mensajero maduro, así se acababa de identificar el primer intrón (Berget et al., 1977; Blanchard et al., 1978; Chow et al., 1977). Este sorprendente descubrimiento rápidamente se reconoció como un proceso necesario para la expresión de la mayoría de los genes eucariotas (Calame et al., 1980; Early et al., 1980; Kole and Weissman, 1982; Leff et al., 1986; Rogers et al., 1980).

Desde los primeros estudios del proceso de splicing en organismos eucariotas se observó que esta maduración de los pre-mensajeros no era un proceso sistemático de eliminación de intrones, siguiendo un patrón único para cada gen, sino que era un proceso dinámico, posibilitando la generación de varios mensajeros maduros

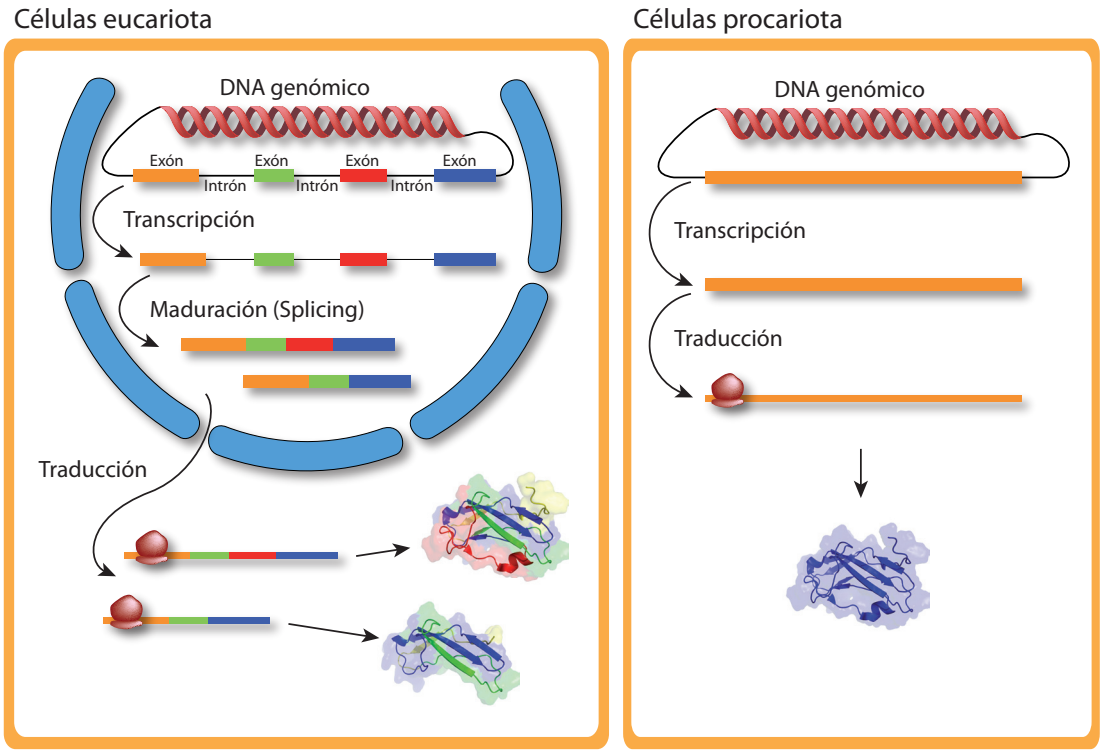


Figura 1-1. Representación esquemática del proceso de expresión génica en organismos eucariotas (izquierda) y procariotas (derecha).

estructuralmente diferentes a partir de un único gen por la eliminación diferencial de fragmentos del pre-mensajero (Leff et al., 1986). Por lo tanto, mediante este procesamiento alternativo o "*splicing alternativo*" (AS, *"Alternative splicing"*) se podían obtener varias isoformas proteicas diferentes a partir de la secuencia de un único gen (Figura 1-1) (Caceres and Kornblihtt, 2002; Kalsotra and Cooper, 2011; Kelemen et al., 2013; Stamm et al., 2005), lo que ponía de manifiesto que la capacidad codificante de los genomas de los organismos eucariotas era mucho mayor de la esperada.

Este descubrimiento puso el foco científico en el papel que el RNA podía desempeñar en la regulación de la expresión génica, más allá de ser un simple intermediario entre genes y proteínas, especialmente en la expansión de la capacidad codificante de los genomas de organismos eucariotas, cuya dimensión no se reveló en su magnitud hasta la secuenciación del genoma humano (Lander et al., 2001; Venter et al., 2001). Basándose en el número de secuencias de mensajeros y proteínas depositadas en las bases de datos se había estimado que el genoma humano debía estar compuesto por al menos 150.000 genes, sin embargo, en los proyectos de secuenciación sólo se detectaron aproximadamente 23.000 genes (Lander et al., 2001; Venter

Especie	Tamaño del genoma (pb)	Nº de genes codificantes	Nº de mensajeros	Densidad génica (Nº genes/Mb)
Homo sapiens	3.324.592.091	20.769	195.565	6,2
Pan troglodytes	2.995.917.117	18.759	29.160	6,3
Gorilla gorilla	2.828.888.833	20.962	35.727	7,4
Pongo abelli	3.109.347.532	20.424	29.447	6,6
Macaca mulata	3.093.871.206	21.905	44.725	7,1
Felis catus	2.365.745.914	19.493	22.656	8,2
Canis lupus	2.392.715.236	19.856	29.884	8,3
Oryctolagus cuniculus	2.604.023.284	19.293	24.964	7,4
Rattus norvegicus	2.573.362.844	22.941	29.189	8,9
Mus musculus	3.480.528.190	23.139	93.480	6,6
Bos taurus	2.649.685.036	19.994	26.740	7,5
Gallus gallus	1.072.544.086	15.508	17.954	14,5
Danio rerio	1.505.581.940	26.247	54.869	17,4
Takifugu rubripes	393.312.790	18.523	48.706	47,1
Drosophila melanogaster	168.736.537	13.937	29.173	82,6
Arabidopsis thaliana	135.670.229	27.249	41.671	200,8
Caenorhabditis elegans	103.022.290	20.532	57.844	199,3
Saccharomyces cerevisiae	12.157.105	6.692	7.126	550,5
Escherichia coli	4.907.865	5.086	5.193	1036,3
Bacillus subtilis	4.215.606	4.185	4.371	992,7
Haemophilus influenzae	1.830.138	1.709	1.745	933,8

Tabla 1-1. Comparativa del tamaño del genoma, número de genes codificantes y número de mensajeros descritos. La información se obtuvo de Ensembl.

et al., 2001), lo que apuntaba al AS como potencial responsable de esta aparente discrepancia. Actualmente se estima que el 95% de los genes humanos producen dos o más mensajeros maduros diferentes a través de AS (Pan et al., 2008), siendo éste considerado uno de los principales factores en la generación de diversidad transcriptómica y proteómica (Barbosa-Morais et al., 2012; Blencowe, 2006; Chen et al., 2012; Merkin et al., 2012; Wang et al., 2008).

En un contexto evolutivo, la expansión de la capacidad codificante del genoma a través de AS se ha convertido en los últimos años en uno de los elementos fundamentales en el estudio de la complejidad y la diversidad entre especies (Barbosa-Morais et al., 2012; Merkin et al., 2012), ya que permite explicar la reducida diferencia detectada en el número de genes de especies tan alejadas como *Drosophila melanogaster* o *Caenorhabditis elegans* y *Homo sapiens*, entre otros (Tabla 1-1). En especial, en el grupo de los mamíferos, que comparten un mismo conjunto de genes básicos (Barbosa-Morais et al., 2012; Boue et al., 2003; Merkin et al., 2012), el AS se postula como un elemento crucial para explicar su diversidad (Tabla 1-1).

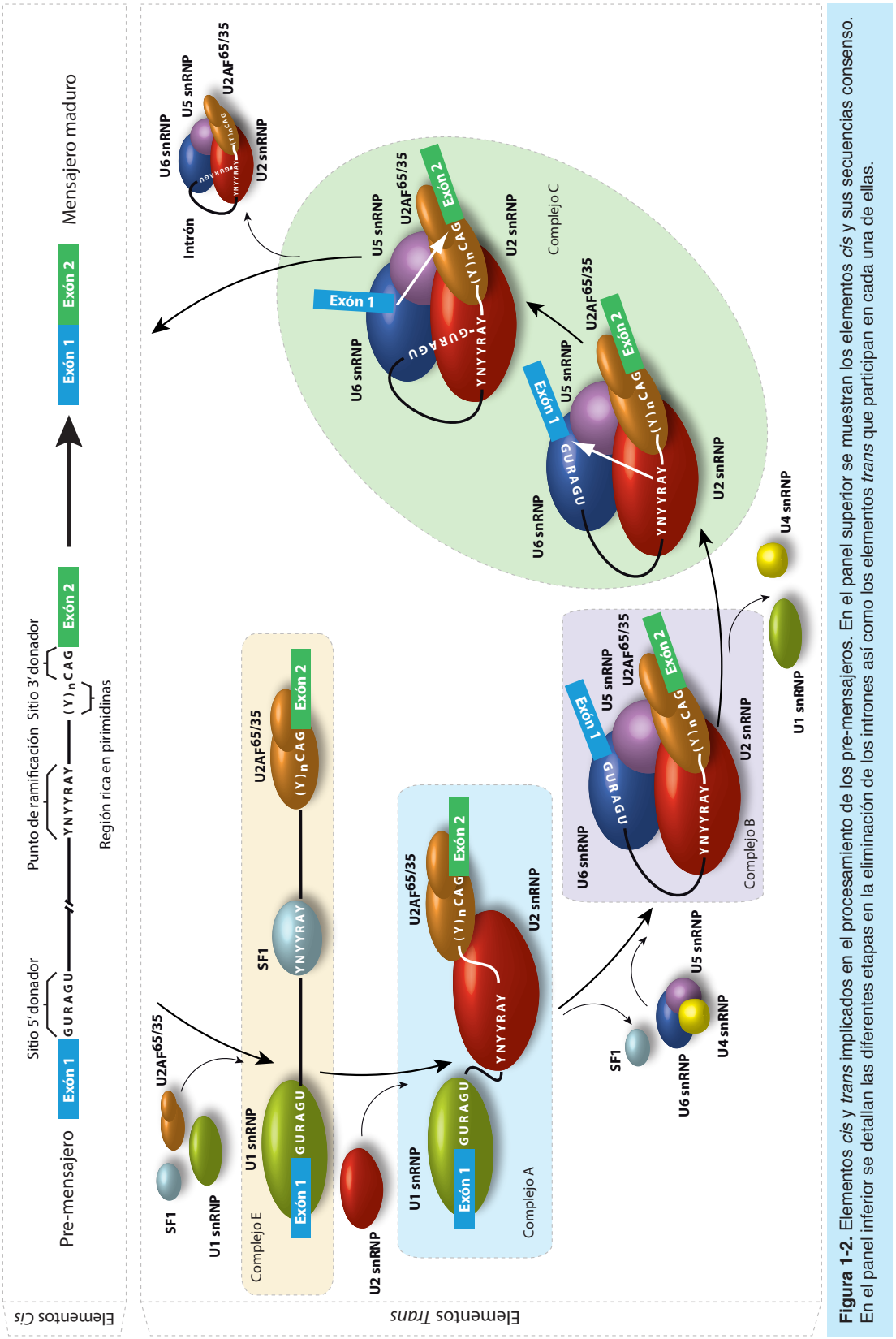


Figura 1-2. Elementos *cis* y *trans* implicados en el procesamiento de los pre-mensajeros. En el panel superior se muestran los elementos *cis* y sus secuencias consenso. En el panel inferior se detallan las diferentes etapas en la eliminación de los intrones así como los elementos *trans* que participan en cada una de ellas.

1.1. Mecanismo y regulación del splicing

El proceso de splicing comienza con la identificación de los exones y los intrones en los pre-mensajeros a través de una serie de secuencias consenso, conocidas como elementos *cis*, que definen las fronteras exón/intrón (Figura 1-2). Estos elementos básicos son reconocidos por complejos ribonucleoproteicos o snRNP (*"small nuclear ribonucleoproteins"*) y diversas proteínas, denominados en conjunto elementos *trans*, que tras la unión con sus elementos *cis* correspondientes dan lugar a la formación de un complejo multiribonucleoproteico llamado *spliceosoma* que, tras varias etapas de restructuración dinámica, cataliza la escisión de los intrones (De Conti et al., 2013; Villate et al., 2008; Wahl et al., 2009).

1.1.1 Mecanismo básico del splicing

El extremo 5' de los intrones de los pre-mensajeros está definido por la presencia de una secuencia consenso denominada sitio 5' donador, que es reconocida por el U1 snRNP gracias a la complementariedad de bases (Figura 1-2). Por otro lado, el extremo 3' queda definido por la presencia del sitio de ramificación ("*Branch point"*) reconocido inicialmente por la proteína SF1 (*"Splicing Factor 1"*), seguido por una región rica en pirimidinas ("*polypyrimidin track*") y el sitio 3' aceptor siendo ambos reconocidos por el heterodímero $U2AF^{65/35}$. La unión de estos elementos iniciales forma el primer estadio del spliceosoma o complejo "E" (Figura 1-2). La proteína SF1 es entonces desplazada por la unión del U2 snRNP que interacciona con la secuencia del sitio de ramificación, siendo su unión estabilizada por la interacción con la proteína $U2AF^{65/35}$, formándose el complejo "A" (Figura 1-2). El complejo pre-catalítico "B", se forma por la incorporación del trímero de snRNP (U4/U5-U6) al sitio 5' donador, que acaba por desestabilizar la interacción del U1 snRNP provocando su salida del complejo, así como la del U4 snRNP. Se cree que el centro catalítico se forma en la superficie de contacto del U6 snRNP con el U2 snRNP, mientras que el U5 snRNP realiza funciones de estabilización del complejo (Wahl et al., 2009). Finalmente, una serie de alteraciones estructurales, así como la interacción de diferentes proteínas conducen a la formación del complejo catalítico "C", encargado de la eliminación de las secuencias intrónicas (Figura 1-2).

Bioquímicamente, el proceso de escisión de los intrones se produce en dos etapas, a través de dos trans-esterificaciones (Wahl et al., 2009). En la primera, el grupo

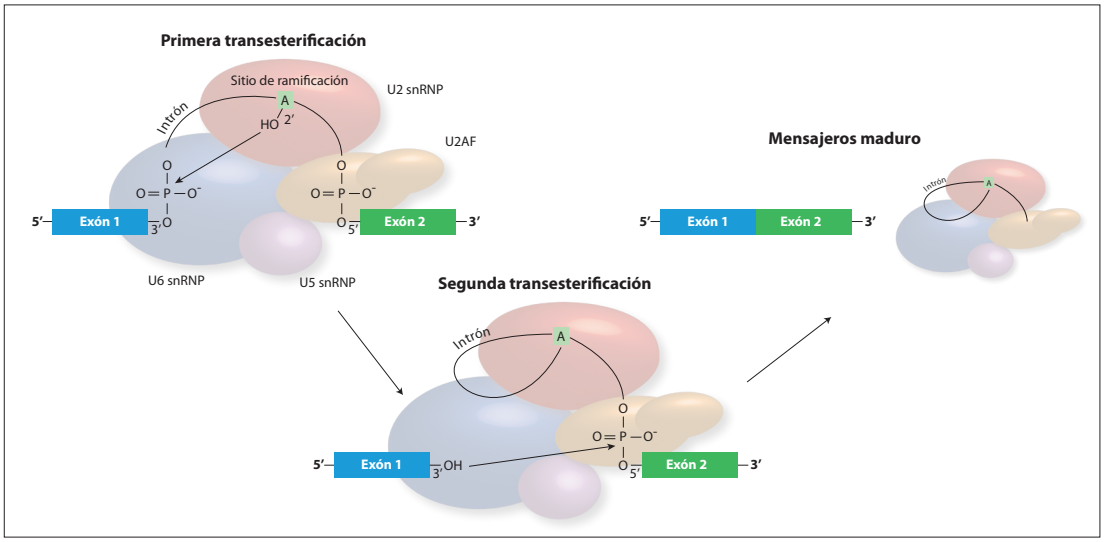


Figura 1-3. Detalle de las trans-esterificaciones necesarias para la escisión de un intrón.

hidroxilo del carbono 2' de la adenosina conservada en el sitio de ramificación, ataca el fosfato del primer nucleótido del intrón (Figura 1-3). Esto lleva a la escisión del intrón por su extremo 5', quedando éste ligado al sitio de ramificación (Figura 1-3). En la segunda etapa, el grupo hidroxilo del carbono 3' del último nucleótido del primer exón ataca al grupo fosfato del primer nucleótido del siguiente exón, ligándose ambos exones y liberando el intrón en forma de lazo, que será degradado (Figura 1-3).

1.1.2 Regulación del proceso de splicing

Este mecanismo, conceptualmente simple, no puede explicar por sí solo la gran diversidad estructural detectada experimentalmente en los mensajeros maduros. Se han descrito eventos de inclusión/exclusión de exones concretos, exones que aparecen de forma mutuamente excluyentes en los mensajeros maduros, retención de intrones, uso de sitios 5' donadores y 3' aceptores alternativo que reducen o incrementan el tamaño de los exones, y combinaciones de todos ellos para formar los mensajeros maduros (Figura 1-4).

La razón de esta variabilidad se haya en la flexibilidad que confiere al proceso de splicing la laxitud de las secuencias que definen los exones e intrones (Izquierdo and Valcarcel, 2006). En *Saccharomyces cerevisiae*, con tan sólo 250 intrones y en la que apenas se han descrito eventos de splicing alternativo (Izquierdo and Valcarcel, 2006; Spingola et al., 1999), las secuencias consenso necesarias para el proceso de splicing están muy bien conservadas (GUAUGU en el sitio 5', UACUAAC en el

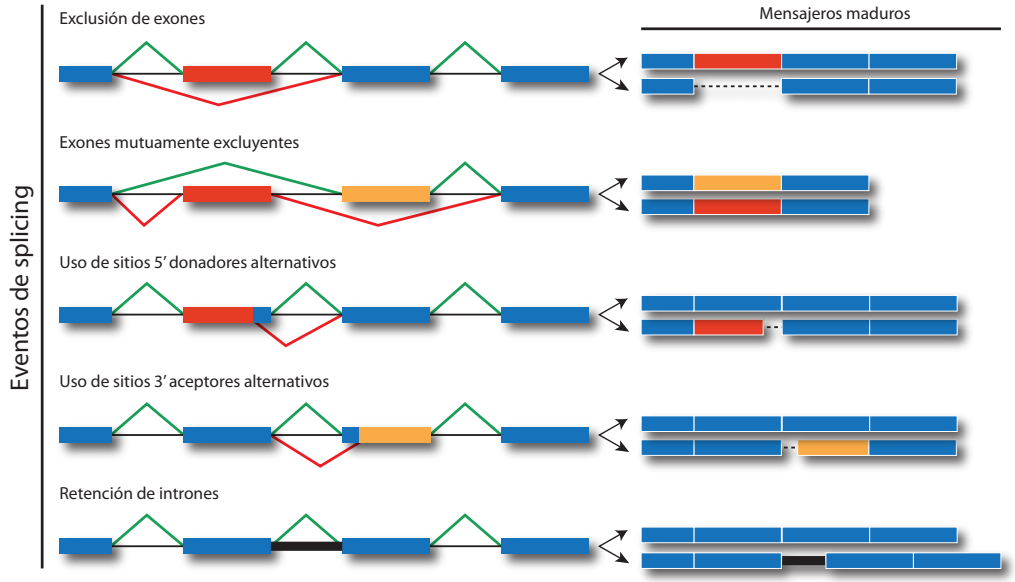


Figura 1-4. Clasificación de los diferentes eventos de splicing alternativo descritos. Las cajas azules represetan exones y las líneas negras intrones. Las cajas rojas o naranjas representan los exones sometidos a AS, indicado con líneas rojas y verdes.

sitio de ramificación y AG en el sitio 3'), por lo que la complementariedad de bases entre éstas y los factores *trans* es suficiente para la eliminación de los intrones. Sin embargo, en eucariotas superiores las secuencias presentan un alto grado de degeneración (GURUGU en el sitio 5', YNYYRAY en el sitio de ramificación y AG en el sitio 3'), quedando la mayoría de los intrones definidos por apenas dos dinucleótidos, GU y AG, en los extremos (De Conti et al., 2013; Izquierdo and Valcarcel, 2006). Esta degeneración de las secuencias supone un alto grado de libertad en la identificación de las fronteras exón/intrón lo que podría ser la base para explicar la enorme variabilidad de eventos de splicing detectada (Figura 1-4). Sólo en el caso de los exones constitutivos, los que mayoritariamente forman parte de los mensajeros maduros, el grado de conservación de las secuencias es ligeramente superior, aunque no suficiente como para producir interacciones estables con los distintos factores *trans* a través de la complementariedad de bases (De Conti et al., 2013; Izquierdo and Valcarcel, 2006; Shen and Green, 2006). Además, una pequeña fracción (~1%) de los intrones están definidos por los dinucleótidos AT y AC que son procesados por un spliceosoma alternativo en el que los snRNP U1 y U2 son reemplazados por sus homólogos U11 y U12, cuya expresión diferencial regula el procesamiento de estos intrones y en consecuencia la inclusión o exclusión de los exones rodeados por éstos (Kreivi and Lamond, 1996; Tarn and Steitz, 1996; Wahl et al., 2009; Wu and Krainer, 1999).

Uno de los factores que puede modular el reconocimiento de los sitios de splicing por los factores *trans* es la propia interacción entre éstos. La unión del U1 snRNP al sitio 5' donador puede estabilizar la unión de la proteína $U2AF^{65/35}$ al sitio 3' aceptor (Figura 1-2 y 1-5). Sin embargo, en función de la estructura génica se han propuesto dos modelos sobre cómo podría ocurrir esta pequeña estabilización. En eucariotas superiores, en los que el tamaño medio de los exones internos es de 140 pb y el tamaño de los intrones unas diez veces mayor (Faustino and Cooper, 2003; Villate et al., 2008), se cree que la interacción entre estos factores ocurre a través de los exones (De Conti et al., 2013), es decir, que el splicing ocurre a través de la definición de los exones (Figura 1-5). Esta hipótesis se sustenta en experimentos en los que la reducción artificial del tamaño de un exón (<50 pb) provoca su exclusión del mensajero final (De Conti et al., 2013; Dominski and Kole, 1991), probablemente por la interferencia estérica entre el heterodímero $U2AF^{65/35}$ ligado al sitio 3' aceptor en el intrón anterior y el U1 snRNP unido al sitio 5' donador del intrón posterior. Inversamente, la expansión artificial de exones pequeños fomenta su inclusión o el uso de sitios crípticos en el interior de los exones cuando la expansión supera las 300 pb (Berget, 1995). Por el contrario, en levaduras y en otros eucariotas inferiores, con exones mayores e intrones pequeños (<100 pb), se piensa que el splicing depende de la definición de los intrones (Figura 1-5). En este sentido, la expansión artificial de intrones en *Saccharomyces* y *Drosophila* conduce a su retención. Por lo tanto, el tamaño de exones e intrones es un factor que puede determinar la fortaleza del reconocimiento de los sitios de splicing, y en consecuencia, influir en la inclusión o exclusión de algunos exones en el mensajero final (Figura 1-5).

Adicionalmente, en eucariotas superiores se ha desarrollado un complejo sistema auxiliar que permite la regulación del AS a través de la modulación del reconocimiento de los distintos sitios de splicing. A lo largo de exones e intrones existen secuencias potenciadoras, conocidos como *"Exon Sequence Enhancers"* (ESE) e "*Intron Sequence Enhancers"* (ISE), que son reconocidos por proteínas de la familia *"Serine-arginine rich"* (SR). Estas proteínas se caracterizan por la presencia de dominios RS (ricos en arginina y serina) con la capacidad de interaccionar con RNA de doble cadena, permitiéndoles estabilizar la interacción de los snRNPs con sus sitios de splicing correspondiente (Figura 1-5) y fomentando así la inclusión de los exones en el mensajero final. Por otro lado, también existen secuencias inhibitorias, "*Exon Sequence Silencers"* (ESS) e "*Intron Sequence Silencers"* (ISS),

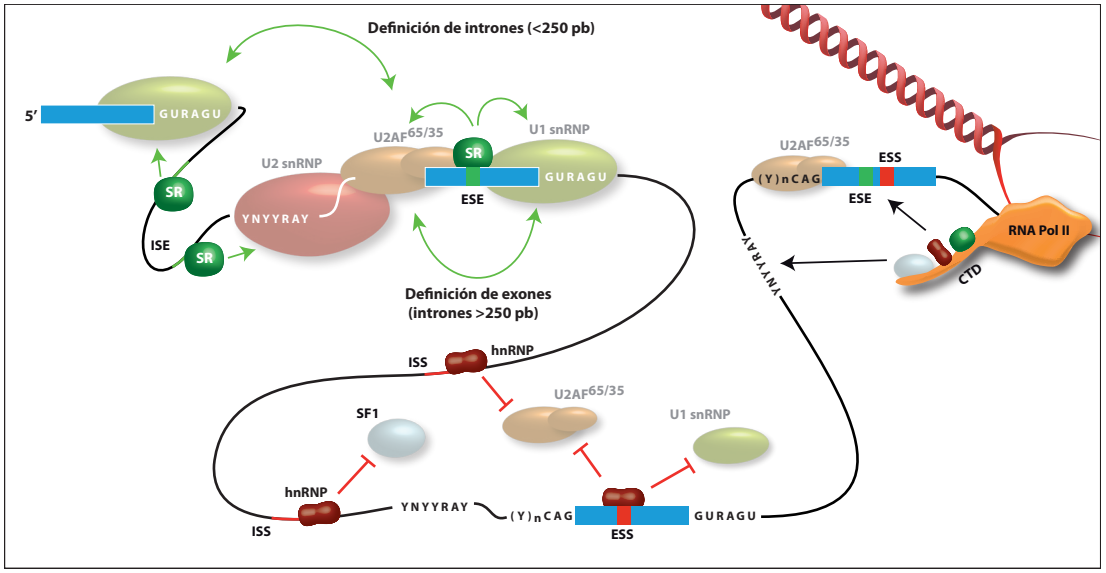


Figura 1-5. Regulación del proceso de splicing. El reconocimiento de los elementos *cis* está influido por el tamaño de los exones e intrones (definición de exones o intrones), por la presencia de secuencias potenciadoras (ESE e ISE) e inhibidoras (ESS e ISS) o por la disponibilidad de los factores de splicing, regulada por la unión de éstos al dominio CTD de las RNA Pol II, entre otros. Las cajas azules representan los exones, mientras que la líneas negras representan los intrones. Las fechas verdes indican influencias positivas y las barras rojas negativas.

a los que se unen diferentes ribonucleoproteínas de la familia *"Heterogeneous nuclear ribonucleoproteins"* (hnRNP) (Heinrich et al., 2009). Los hnRNP promueven la exclusión de los exones dificultando la interacción de los diferentes factores de splicing con sus correspondientes secuencias (Figura 1-5).

Aunque de manera general, las proteínas SR fomentan la inclusión de los exones y los hnRNP su exclusión, se han descrito situaciones concretas en las que el comportamiento de estos factores es el inverso. Además, los factores auxiliares se expresan de manera diferencial, lo que permite controlar la arquitectura de los mensajeros y regular las variantes de splicing que se expresan en cada tejido, tipo celular o en respuesta a unas condiciones fisiológicas concretas. Esta compleja red regulatoria, denominada el "*código del splicing*", es uno de los principales factores implicados en la diferenciación tisular y celular (Blencowe, 2006).

1.1.3 Splicing alternativo y transcripción

Una de las grandes cuestiones sobre la maduración de los pre-mensajeros fue determinar cuándo ocurría la escisión de los intrones. Se han descrito casos de procesamiento post-transcripción en los que los mensajeros inmaduros, principalmente por la retención de algún intrón, se acumulan en el núcleo a modo de reservorios, que más tarde, tras una estimulación adecuada, terminan su maduración para producir

una rápida respuesta a dichos estímulos (Hernández-Torres et al., 2013; Vargas et al., 2011). Sin embargo, los últimos estudios indican que al menos el 80% del splicing ocurre de manera co-transcripcional, indicando que el AS es un proceso íntimamente conectado con la transcripción (Dujardin et al., 2013).

El dominio C-terminal de la RNA polimerasa II (RNA Pol II) o CTD ("*C-terminal domain*") tiene un papel decisivo en la regulación de la transcripción de los genes codificantes de proteína (Dujardin et al., 2013). El estado de fosforilación de este dominio CTD modula la actividad de la RNA polimerasa II por la unión de diversos factores de transcripción en función de diferentes estímulos, regulando tanto la actividad transcripcional general como la adición de la caperuza (CAP) y la cola de poliadenilas (PolyA) a los mensajeros. Asimismo, se han descrito interacciones de diversos factores implicados en el proceso de splicing (U1 snRNP, proteínas SR, hnRNPs, etc.) con el dominio CTD, siendo éstas dependientes de su estado de fosforilación (Figura 1-5). Se ha propuesto que esta interacción regula la disponibilidad de los factores de splicing (tanto básicos como auxiliares) e influye, por lo tanto, en el procesamiento de los pre-mensajeros, provocando la inclusión/exclusión de exones determinados en respuesta a diversos estímulos (Dujardin et al., 2013). Por otro lado, el correcto procesamiento de los exones iniciales y finales está influido por la interacción de la maquinaria de splicing con las proteínas del complejo implicado en la adición del CAP, el *"CAP-Binding protein complex"*) (CBPC), y con los factores relacionados con la adición de la cola de PolyA, el *"cleavege/polyadenylation specificity factor"* (CPSF), el *"Cleavege stimulation factor"* (CstF),, el *"Clevage factor"* (CF), y la *"poly(a) polymerase"* (PAP,) (Dujardin et al., 2013).

La velocidad de elongación de la RNA Pol II es otro factor que influye sobre el procesamiento de los pre-mensajeros (Dujardin et al., 2013). Se ha propuesto un modelo competitivo en el que la velocidad de elongación influye sobre la unión de los factores de splicing con los elementos *cis*. Así, una velocidad de elongación lenta permitiría el reconocimiento de sitios de splicing débiles (secuencias muy degeneradas), mientras que en velocidades de elongación más rápidas, un potencial sitio de splicing posterior más potente (secuencia más conservada) podría competir con el sitio más débil en la unión de los distintos factores de splicing (tanto básicos como auxiliares), de modo que en el primer caso el exón sería incluido mientras que en segundo éste sería excluido. Otra hipótesis que podría explicar esta influencia de

la velocidad de elongación sobre el procesamiento de los mensajeros apunta a la estructura secundaria del RNA naciente, influida por la velocidad de elongación, que podría impedir o facilitar el reconocimiento de las distintas secuencias de los sitios de splicing (Dujardin et al., 2013).

La velocidad de elongación de la RNA Pol II depende de múltiples factores. El estado energético de la célula es uno de ellos, siendo éste dependiente en gran medida del contenido de mitocondrias de la célula (das Neves et al., 2010). Dado que la segregación de las mitocondrias ocurre de manera estocástica durante la división celular, esta variación podría influir sobre el procesamiento diferencial de los mensajeros en las dos células hijas (das Neves et al., 2010; Rastrojo et al., 2013). Otro de los factores moduladores de la velocidad de transcripción es el estado de condensación de la cromatina. En este sentido, en estudios a gran escala se ha detectado una mayor presencia de nucleosomas sobre exones con sitios de splicing débiles que sobre exones constitutivos (De Conti et al., 2013; Luco et al., 2011). Se ha sugerido que la presencia de los nucleosomas sobre estos exones débiles podría reducir la velocidad de elongación de manera que se facilite el reconocimiento de estos sitios por la maquinaria de splicing. Este fenómeno establece una conexión entre los mecanismos epigenéticos y el AS, permitiendo explicar la expresión diferencial de algunas variantes de splicing en función del estadio de desarrollo e incluso entre tejidos y tipos celulares, como un factor complementario a la expresión diferencial de los distintos potenciadores e inhibidores del splicing (Luco et al., 2011).

1.1.4 Implicaciones funcionales del splicing alternativo

El descubrimiento del AS y el creciente conocimiento de los procesos implicados en su regulación han puesto de manifiesto que la capacidad codificante del genoma de los organismos eucariotas es mucho mayor de la que se puede esperar de un análisis lineal de la secuencia genómica. Mediante AS es posible utilizar diferentes combinaciones de exones de un mismo gen para la producción de diversas proteínas con funciones potencialmente diferentes. Sin embargo, este proceso combinatorio no es aleatorio sino que está fuertemente regulado, lo que demuestra que el genoma, y su expresión, tienen una enorme plasticidad, permitiendo a los organismos adaptarse a las circunstancias ambientales concretas y al contexto fisiológico a través de la expresión diferencial de variantes de splicing, e isoformas proteicas, que les permitan responder a los estímulos de manera específica. Por lo tanto, el AS se

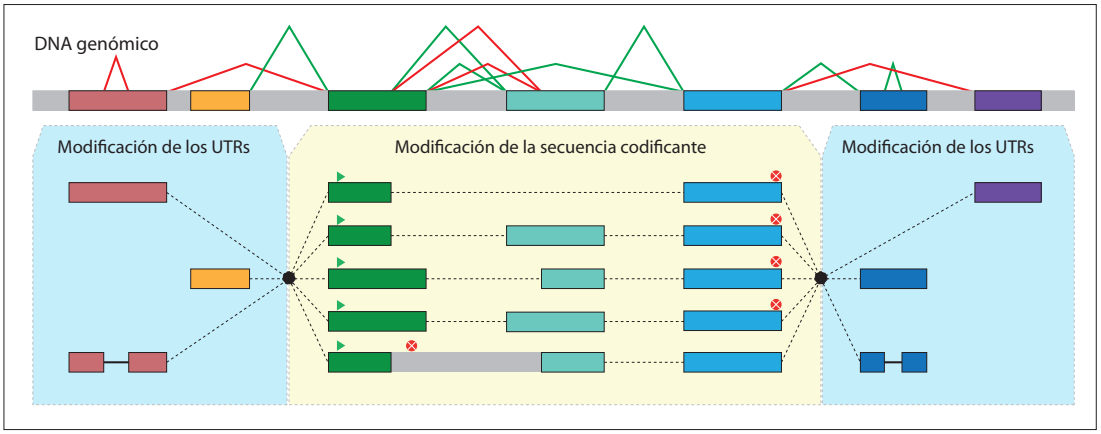


Figura 1-6. Clasificación de los diferentes eventos de splicing en función de su influencia en la modificación de la secuencia codificante o los extremos UTR.

puede considerar como un mecanismo de regulación de la expresión génica. En general, estos mecanismos de regulación de la expresión génica actúan estimulando o reprimiendo globalmente la expresión, sin embargo, a través del AS, además de ejercer un control cuantitativo, también es posible un control cualitativo.

Conceptualmente, los distintos tipos de eventos de splicing alternativo, descritos en la Figura 1-4, se pueden agrupar en dos categorías: los que afectan a las regiones 5' y 3' no traducidas de los mensajeros (5' UTR y 3'UTR) y los que afectan a la secuencia codificante (Figura 1-6). En la primera categoría, la modificación de los extremos UTR puede afectar a la localización subcelular de los mensajeros, a su estabilidad, a la eficiencia de su traducción o la sensibilidad a la degradación mediada por microRNAs, controlando de este modo la expresión cuantitativa del gen (Figura 1-6) (Holt and Bullock, 2009; Martin and Ephrussi, 2009). Un ejemplo de este fenómeno es el gen *SC35*, cuya expresión produce varios mensajeros maduros, que codifican la misma proteína, pero que presentan diferencias en sus extremos 3' UTR lo que provoca cambios en la vida media de éstos, controlando los niveles de expresión de la proteína (Sureau and Perbal, 1994). Curiosamente, la proteína SC35 pertenece a la familia de proteínas reguladoras del splicing SR de modo que la regulación de su expresión a través de AS afecta recíprocamente el AS de otros mensajeros (Sureau and Perbal, 1994). La expresión del co-receptor CD3ζ, implicado en la señalización intracelular mediada por antígeno, también está regulada de manera similar. La región 3'UTR del mensajero de CD3ζ contiene un intrón con varios *"AU-rich elements"* (AREs), regiones ricas en adeninas y uradinas, que aumenta la estabilidad del mensajero y la eficiencia de su traducción (Chowdhury et al., 2006; Martinez and Lynch, 2013). Sin

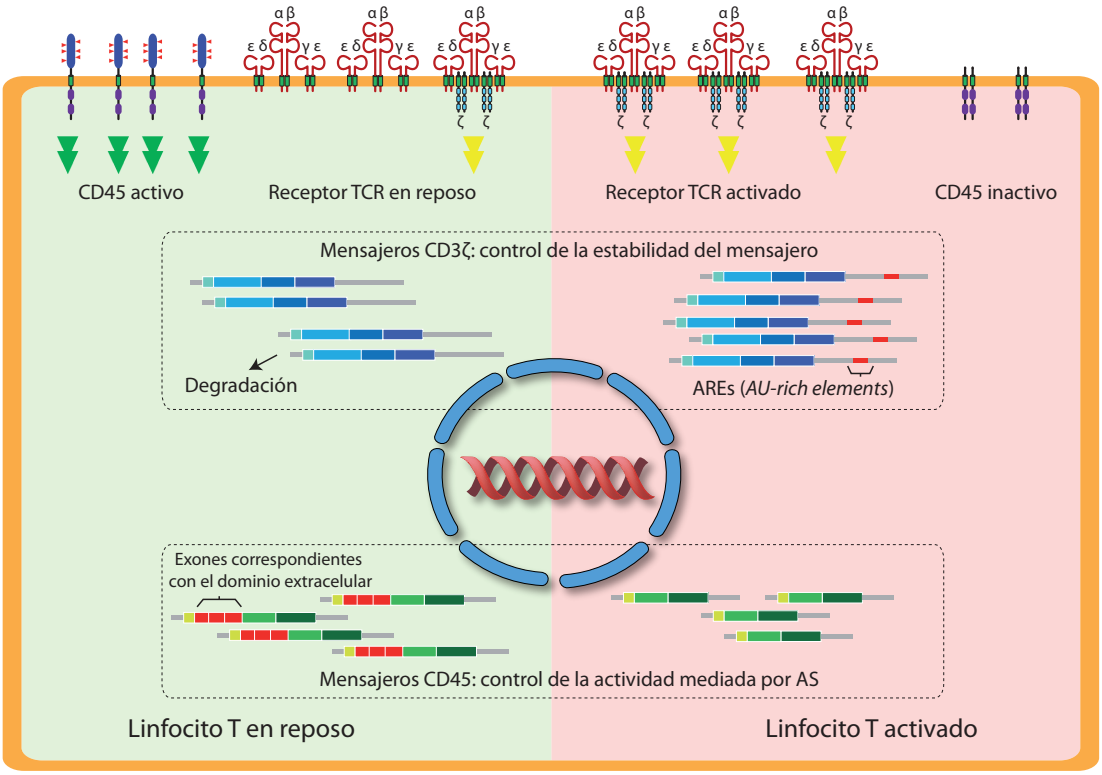


Figura 1-7. Ejemplos de la influencia del AS sobre la expresión génica, cualitativamente (CD45) y cuantitativamente (CD3ζ), en función de la activación de los linfocitos T. Los rectángulos de colores representan exones de la secuencia codificante, mientras que los grises representan las regiones UTRs.

embargo, esta región es eliminada mediante AS en células en reposo, de manera que la expresión del receptor en estas células es menor que tras la activación mediada por antígeno, cuando este intrón es retenido y las secuencias allí presente ejercen su influencia (Figura 1-7).

Por otro lado, los eventos de splicing que afectan a la secuencia codificante permiten ejercer un control cualitativo de la expresión génica (Figura 1-6) que ha sido asociado con diversos procesos biológicos (Kelemen et al., 2013; Stamm et al., 2005). Mediante AS es posible obtener isoformas proteicas con una diferente composición de dominios estructurales, lo que puede afectar a la función que desempeñan. Entre otros, se han descrito eventos de AS que conducen a la producción de isoformas constitutivamente activas por la ausencia de dominios proteicos reguladores, de dominantes negativos de la isoforma canónica por la pérdida/ganancia de alguno de estos dominios, de isoformas con diferente localización subcelular, de isoformas con diferente capacidad de unión a otras proteínas o ligandos afectando incluso a la especificidad de sustrato de algunos enzimas, etc. (Kelemen et al., 2013; Stamm et al., 2005). El receptor de

insulina es un claro ejemplo de este tipo de regulación. Durante el desarrollo fetal se detecta la expresión de la isoforma A, producida por la exclusión del exón 11, que además de presentar una alta afinidad por insulina es capaz de reconocer factores de crecimiento de la familia *"Insulin-like growth factor"* (IGF), mientras que en adulto la isoforma B (que incluye el exón 11) es la más expresada, no siendo capaz de interaccionar con dichos factores de crecimiento (Frasca et al., 1999; Pandini et al., 2002). Otro ejemplo, quizás el más estudiado, es el caso del receptor de membrana CD45, fundamental para la activación de los linfocitos T. Tras la estimulación de estos linfocitos, los exones correspondientes con el dominio extracelular del receptor CD45 son excluidos del mensajero, produciéndose receptores truncados que homodimerizan, lo que antes no era posible debido a la presencia de glicosilaciones en el receptor completo, provocando la inactivación de su actividad fosfatasa y contribuyendo a la regulación de la activación de los linfocitos T (Figura 1-7) (Lynch, 2004; Martinez and Lynch, 2013; Martinez et al., 2012).

Dentro del grupo de eventos de splicing que afectan a la secuencia codificante de los mensajeros destacan aquellos que provocan la aparición de codones prematuros de terminación de la traducción (PTC - *"Premature stop codons"*), considerados por lo tanto no codificantes. Generalmente, los mensajeros que presentan un PTC son eliminados por el sistema de degradación de mensajeros sin sentido o *"Nonsense-mediated decay"* (NMD) (Chang et al., 2007). Este mecanismo de degradación selectiva supone un vínculo entre la maquinaria de splicing y la traducción de los mensajeros. En las uniones de exones, tras la eliminación de los intrones, permanecen asociados un complejo proteico denominado *"Exon junction complex"* (EJC), que participa en el transporte de los mensajeros al citoplasma y son eliminados por el avance del ribosoma durante la traducción (Chang et al., 2007). Sin embargo, si el codón de terminación se localiza más allá de 50-55 pb hacia el extremo 5' de la última unión de exones, el complejo EJC de esta última unión no es eliminado por el ribosoma, lo que desencadena una cascada de señalización que acaba en la degradación del mensajero. Inicialmente se consideró este proceso como un mecanismo de control para evitar la producción de isoformas proteicas aberrantes, que eventualmente podrían ser perjudiciales para el organismo. Sin embargo, estudios recientes apuntan que este mecanismo podría ser utilizado para controlar los niveles globales de expresión. Así, el control de la producción de este tipo de variantes no codificantes podría influir sobre los niveles de expresión de las variantes codificantes, y por lo

tanto, de las proteínas correspondientes, sin modificar los niveles de expresión del gen a nivel de RNA (Hwang and Kim, 2013).

Además de la influencia sobre rutas y procesos biológicos concretos, el AS ejerce una influencia global, que ha sido relacionada con la diferenciación celular y tisular. A través de la expresión diferencial de factores reguladores de splicing, como MBNL1 y CUGBP-1 en tejido muscular (Bland et al., 2010; Kalsotra et al., 2008) o RBFOX1 en el sistema nervioso (Fogel et al., 2012) entre otros, junto a mecanismos epigenéticos, que modifican tanto la accesibilidad de los sitios de splicing como la unión de los factores regulados, se estable un programa concreto de expresión génica, tanto cuantitativo como cualitativo, que conduce al desarrollo tisular y celular específico. En este sentido, se está investigando el papel que el AS puede jugar en el proceso de diferenciación de las células madre, lo que podría tener importantes aplicaciones terapéuticas (Fu et al., 2009). Esta influencia global, trasladada al contexto evolutivo, ha sido propuesta como uno de los factores fundamentales en la diversificación de las especies (Blencowe, 2006; Merkin et al., 2012; Mudge et al., 2011; Pan et al., 2008). Recientemente, mediante estudios a gran escala en vertebrados, se ha observado un alto grado de diversidad en cuando a la influencia del AS en distintas especies, encontrado el mayor grado de AS en el grupo de los primates (Barbosa-Morais et al., 2012). Además, se muestra que el perfil de AS tiene una gran divergencia entre especies, siendo el patrón de expresión de los tejidos de una especie más próximo al de otros tejidos de esta misma, que al tejido equivalente en otra especie (Barbosa-Morais et al., 2012), aunque se han encontrado algunos elementos comunes que podrían ser la base para definir un tejido o tipo celular (Merkin et al., 2012). Por ello, se ha propuesto al AS como uno de los motores de la evolución. En este sentido, cambios en el patrón de splicing de un determinado gen, podría generar nuevas variantes de splicing que eventualmente podrían ser seleccionadas, permitiendo la evolución de las especies sin alterar la función original del gen (Boue et al., 2003).

1.1.5 Splicing alternativo y enfermedades

Pese a la fuerte regulación del AS y a los mecanismos de eliminación de mensajeros aberrantes por NMD, la alteración del correcto procesamiento de los mensajeros es uno de los principales factores en el desarrollo de enfermedades. En humano, el 60% de las mutaciones causantes de enfermedad están relacionadas con el splicing (Villate et al., 2008; Wang and Cooper, 2007). Éstas pueden afectar a los

sitios de splicing (5' donador, punto de ramificación o 3' aceptor) o a las secuencias activadoras e inhibidoras, alterando el correcto reconocimiento de exones e intrones. También es posible que mutaciones silenciosas en los exones o mutaciones en los intrones permitan la aparición de nuevos sitios de splicing o secuencias reguladoras, provocando cambios en el patrón normal de splicing. Estas mutaciones pueden afectar a la expresión de un gen concreto, como en la ß-talasemia o la fibrosis quística atípica (Villate et al., 2008), pero también pueden afectar a elementos de la maquinaria del splicing o a factores auxiliares, lo que provoca una alteración global del splicing alternativo, como ocurre en la Retinitis pigmentosa o en la Atrofia muscular espinal (Villate et al., 2008). Un caso especial de este tipo de alteraciones es aquel en el que la alteración no reside en el factor de splicing afectado, como es el caso de la Distrofia miotónica tipo 1, en la que la expansión de tripletes CUG en el extremo 3' UTR del gen *DMPK* conduce al secuestro del factor de splicing MBNL1, lo que provoca la sobre-expresión del factor CUGBP-1, que en combinación altera globalmente el procesamiento de varios mensajeros produciendo esta compleja enfermedad (Kanadia et al., 2003; Miller et al., 2000). El splicing aberrante también ha sido asociado al desarrollo de procesos cancerígenos, su metástasis y la resistencia de éstos a la quimioterapia (Pal et al., 2012). Uno de los casos mejor estudiados es el del gen *CD44*, cuya expresión produce varios mensajeros codificantes en varios tipos celulares y por lo tanto, varios receptores de membrana implicados en la adhesión celular. Sin embargo, se ha observado una sobre-expresión del gen en células metastáticas y un cambio en el patrón de AS, expresando preferentemente la variante CD44v6, contra la que se están desarrollando nuevas estrategias terapéuticas (Bennett et al., 1995a; Bennett et al., 1995b; Heider et al., 2004).

1.1.6 Splicing alternativo y nuevas tecnologias

Aunque el avance en el conocimiento del AS y sus implicaciones en diversos procesos biológicos, tanto normales como en enfermedades, ha sido inmenso, la comunidad científica sigue considerándolo como algo anecdótico, quedando su investigación relegada al ámbito de lo patológico, ignorando por tanto, su potencial influencia en el resto de procesos biológicos. Sólo en los últimos años, con el avance de la tecnología, y en especial en las técnicas de secuenciación masiva de RNA (RNA-seq), se han logrado grandes avances en la compresión del alcance del AS en diferentes procesos biológicos, tanto normales como patológicos. Sin embargo,

estas técnicas de secuenciación masiva, aunque nos proporcionan una visión global, presentan algunas limitaciones que suponen un freno a dicho avance. Entre otros, cabe destacar la deficiente anotación de los genomas, usados como referencia para los estudios de RNA-seq, por el creciente uso de predicciones bioinformáticas, que, aunque bienintencionadas, suponen la generalización de los parámetros biológicos, que en ocasiones conduce a graves errores, sobre todo en el uso de secuencias de una especie para la predicción de la anotación en otras especies. Otra de las grandes limitaciones de estas técnicas de secuenciación masiva reside en el análisis informático de la ingente información obtenida, y en especial, en el ensamblaje de mensajeros, que dada la naturaleza fragmentaria de las secuencias obtenidas, es un proceso complejo lleno de potenciales incertidumbres en la asignación de las lecturas a mensajeros concretos. Por ello, la mayoría de los estudios a gran escala se limitan a la anotación existente en las bases de datos o a anotaciones arbitrarias generadas por la combinación de los exones descritos, ignorando los potenciales eventos de splicing aún no descritos.

Por lo tanto, el análisis del AS, mediante técnicas convencionales, sigue siendo fundamental para obtener una visión real del alcance de este fenómeno en cada uno de los genes. En especial, resulta crucial el análisis del AS en condiciones normales, para una mejor compresión de los cambios observados en procesos patológicos y el desarrollo de terapias más eficaces y específicas carentes de efectos secundarios indeseados (Villate et al., 2008). Asimismo, el estudio comparativo del AS en diferentes especies puede ayudarnos a entender su contribución a la evolución de las especies, a través de la plasticidad que este mecanismo proporciona al genoma y su expresión. El análisis del AS en otras especies, además, puede ayudar a comprender la función de variantes de splicing conservadas en humano, e incluso aquellas no conservadas pueden mejorar nuestra compresión sobre los mecanismos de regulación de la expresión génica y su contribución a los procesos biológicos en los que éstas están implicadas, proporcionando vías alternativas para el desarrollo de nuevas terapias.

1.2. La región del Complejo Mayor de Histocompatibilidad de clase III

Dado el papel central que el AS tiene sobre la regulación de la expresión génica, confiriendo al genoma de una plasticidad antes inimaginable, el objetivo principal del laboratorio es el estudio comparativo de este fenómeno en diferentes especies con especial interés en el papel diversificador del AS y su implicación en el desarrollo de

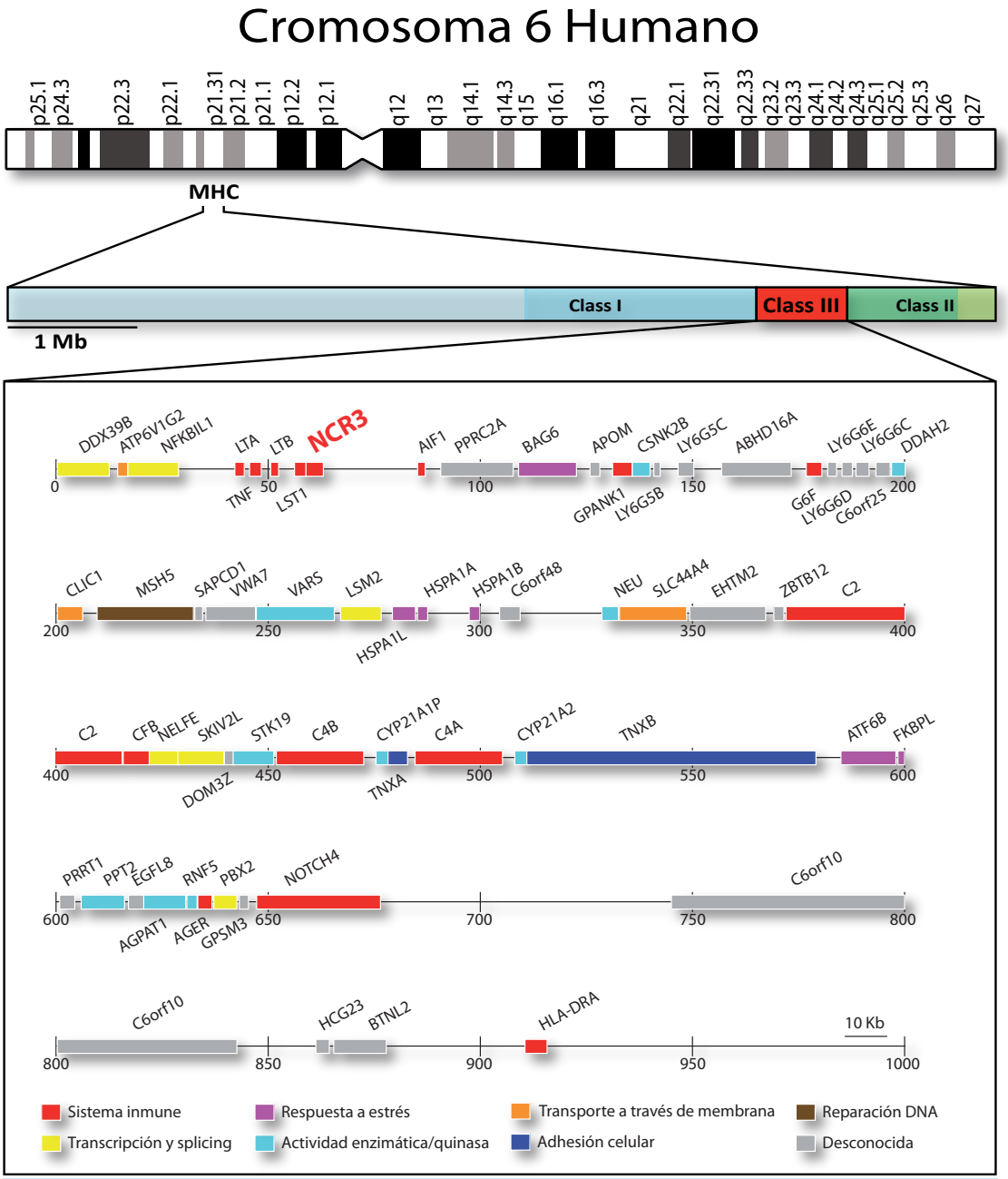


Figura 1-8. Localización de la región del MHC humano en el cromosoma 6. En la parte inferior se detallan los genes descritos en la región de clase III, representados a escala, lo que permite observar la concentración génica. Adicionalmente, se indica con un código de colores la implicación de cada uno de los genes en diversas funciones biológicas.

enfermedades. Para ello, se eligió la región del Complejo Mayor de Histocompatibilidad de clase III, *"Major Histocompatibility Complex"* (MHC), situada en el cromosoma 6p21.3 humano entre las regiones de clase I y II (Figura 1-8), en las que están codificados las moléculas de histocompatibilidad implicadas en la presentación de antígenos a los linfocitos T y B (Horton et al., 2004; Stephens et al., 1999).

La región central de clase III es una de las regiones con mayor densidad génica del genoma humano, con 71 genes codificantes, 5 pseudogenes y 20 genes no codificantes (rRNA, miRNA, ncRNAs, snRNA y snoRNA) en tan sólo 0.9 Mb (Xie et al., 2003). Entre los genes codificantes se encuentran algunos con funciones relacionadas con el sistema inmune como los componentes del complemento *C4A*, *C4B* y Factor B (*BF*), citoquinas pro-inflamatorias como Factor de Necrosis Tumoral (*"Tumor Necrosis Factor", TNF-α*) o las linfotoxinas α y β (*LTα y LTβ*) (Xie et al., 2003), y receptores activadores de células "*Natural Killer"* ("*Natural Citotoxicity Receptor 3"*, *NCR3*) (Neville and Campbell, 1999) (Figura 1-8). Sin embargo, también encontramos genes con un amplio espectro de funciones, incluyendo enzimas como la *ácido lisofosfatídico acil transferasa* ("*1-acylglycerol-3-phosphate-O-acyltransferase*", *AGPAT1*) (Aguado and Campbell, 1998), la *palmitoil proteína tioesterasa* (*PPT2*) (Aguado and Campbell, 1999) o la *sialidasa Neu 1* (Milner et al., 1997), una *valyl tRNA sintetasa* (*VARS2*) (Hsieh and Campbell, 1991), el *receptor de productos de glicosilación avanzada* ("*advanced glycosylation end product-specific receptor"*, *AGER*) (Neeper et al., 1992), un miembro de la familia IkappaB (*NFKBIL1*) (Albertella and Campbell, 1994) y genes implicados en la regulación transcripcional y el procesamiento de RNA mensajeros como el *"pre-B-cell leukemia homeobox 2"* (*PBX2*) (Aguado and Campbell, 1995) y *"DEAD box protein 39B"* (*DDX39b*) (Luo et al., 2001) (Figura 1-8).

Las regiones de clase I y II presentan un alto grado de polimorfismo, lo que permite a las distintas moléculas de MHC de tipo I y II reconocer los diversos antígenos a los que deben enfrentarse, sin embargo, este alto grado de variación también está implicado en la susceptibilidad a diversas enfermedades (Vandiedonck and Knight, 2009). La región de clase III no es una excepción en este sentido, habiéndose descrito numerosas enfermedades con una fuerte vinculación con esta región como la inmunodeficiencia variable común ("*common variable immunodeficiency*", CVID) (Cucca et al., 1998; Schroeder et al., 1998), la artritis reumatoide (Bali et al., 1999;

Chiba et al., 2011; Hu et al., 2011; Jenkins et al., 2000; Kilding and Wilson, 2005), la diabetes tipo I (Kumar et al., 2012) o el lupus eritematoso sistémico ("*systemic lupus erythematosus*", SLE) (Boteva et al., 2012), así como un mayor o menor grado de afección en procesos infecciosos tales como la malaria (Delahaye et al., 2007), la hepatitis C (Shichi et al., 2005) o el "Sindrome de inmunodeficiencia adquirida" (SIDA) (Guergnon et al., 2012), lo que la convierte en una región modelo para estudios genéticos (Vandiedonck and Knight, 2009).

Aunque muchos de los genes localizados en la región de clase III no muestran una asociación directa con el funcionamiento del sistema inmune, se puede observar un alto grado de conservación de éstos y su organización (sintenia), junto a las clases I y II, en la mayoría de los mamíferos (Deakin et al., 2006; Hurt et al., 2004; Peelman et al., 1996; Qin et al., 2008; Renard et al., 2006; Walter et al., 2002). Sin embargo, se cree que el origen de esta organización génica se remonta incluso hasta los anfibios, hace unos 350 millones de años (Ma) (Deakin et al., 2006; Flajnik et al., 2012). Por ello, esta región resulta idónea para el estudio comparativo del splicing alternativo, minimizando el efecto de otros procesos sobre éste. En este sentido, estudios previos en el laboratorio, han demostrado que los genes de la región de MHC de clase III muestran un alto grado AS, encontrando una gran diversidad entre especies (Calvanese et al., 2008; Hernández-Torres et al., 2013; López-Díez et al., 2013; Mallya et al., 2002, 2006).

1.3 Splicing alternativo del gen *NCR3*

Este trabajo se centra en el estudio del AS del gen *NCR3*, localizado en la región del MHC de clase III (Figura 1-8), en su extremo telomérico, entre los genes *LST1* y *AIF1* (Neville and Campbell, 1999). El gen *NCR3* se describió en 1996 (entonces llamado *1C7*) (Nalabolu et al., 1996) mediante el rastreo de una librería de cDNA por hibridación contra un *"Yeast artificial chromosome"* (YAC) que contenía una parte importante de la región MHC de clase III humano, cuya secuencia era aún desconocida. Se detectaron tres cDNAs de diferentes tamaños que hibridaban en la misma región del YAC, que presentaban diferencias en sus extremos 3', lo que sugería un procesamiento alternativo (Nalabolu et al., 1996).

Neville et al. (1999) (Neville and Campbell, 1999) estudiaron los potenciales genes que se habían descrito en la región MHC de clase III de la que se acababa de

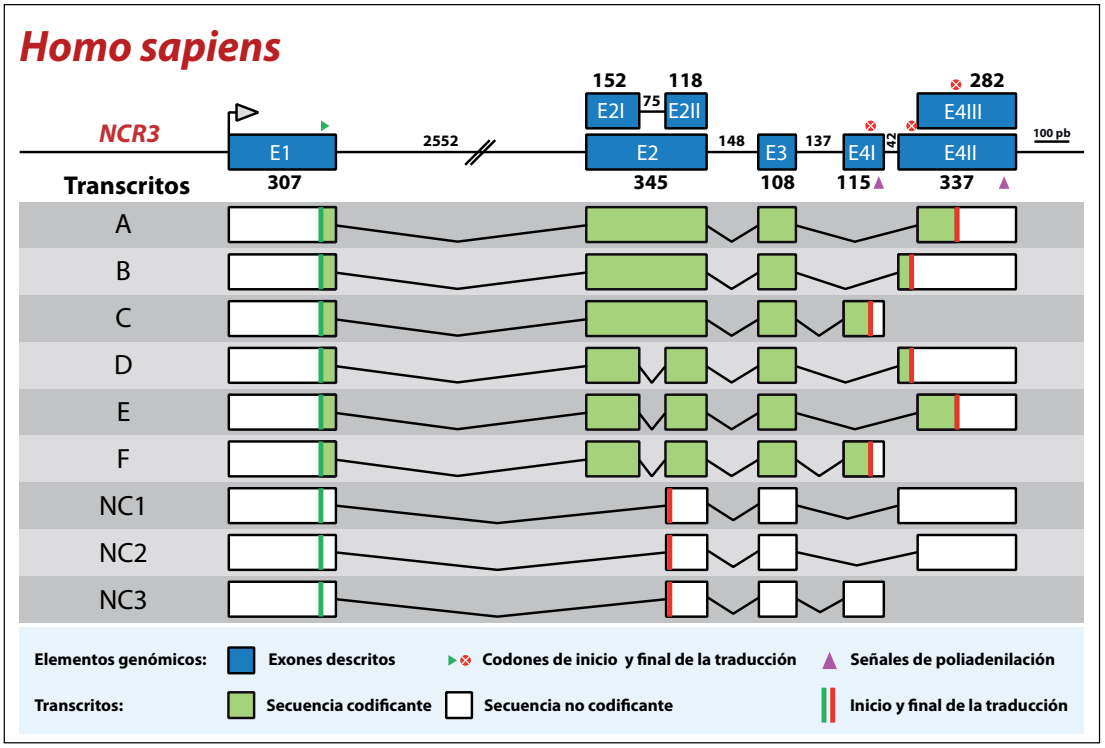


Figura 1-9. Splicing alternativo del gen *NCR3*. En la parte superior se representa la estructura exónica del gen *NCR3*, mientras que en la parte inferior se detalla las variantes de splicing descritas por *Neville et al.* (1999). Los números sobre la estructura exónica se refieren al tamaño en pares de bases (pb) .

obtener la secuencia completa (The MHC sequencing consortium, 1999).Utilizando las secuencias previamente descritas descubrieron que los cDNAs del gen *NCR3* estaban formados por 4 exones, siendo los tres primeros comunes y el último diferente para cada uno de los cDNAs (exones 4I, 4II y 4III) (Figura 1-9), lo que confirmaba su generación por AS. Por tanto, los tres transcritos descritos podían codificar tres proteínas diferentes en su extremo C-terminal y fueron nombrados *1C7a*, *1C7b* y *1C7c* (referidos aquí como *A*, *B* y *C*, Figura 1-9). Basándose en los exones que habían conseguido posicionar diseñaron cebadores para hacer un análisis más profundo de la expresión del gen *NCR3* usando RT-PCR en varias líneas celulares. Además de estos mensajeros encontraron seis nuevos. Tres de ellos, nombrados *1C7e*, *1C7d*, y *1C7f* y referidos aquí como *D*, *E* y *F*, presentaban el exón 2 dividido en dos partes por la ausencia de 75 pb, causado por un evento de splicing alternativo, en combinación con los exones 4 descritos (Figura 1-9). Los tres restantes sólo conservaban la segunda mitad del exón 2, combinados con los exones finales alternativos. Estos mensajeros, nombrados aquí como *NC1*, *NC2* y *NC3*, presentan un cambio de fase que desencadena la aparición de un codón prematuro de parada, y por lo tanto, se trata de mensajeros no codificantes (Figura 1-9).

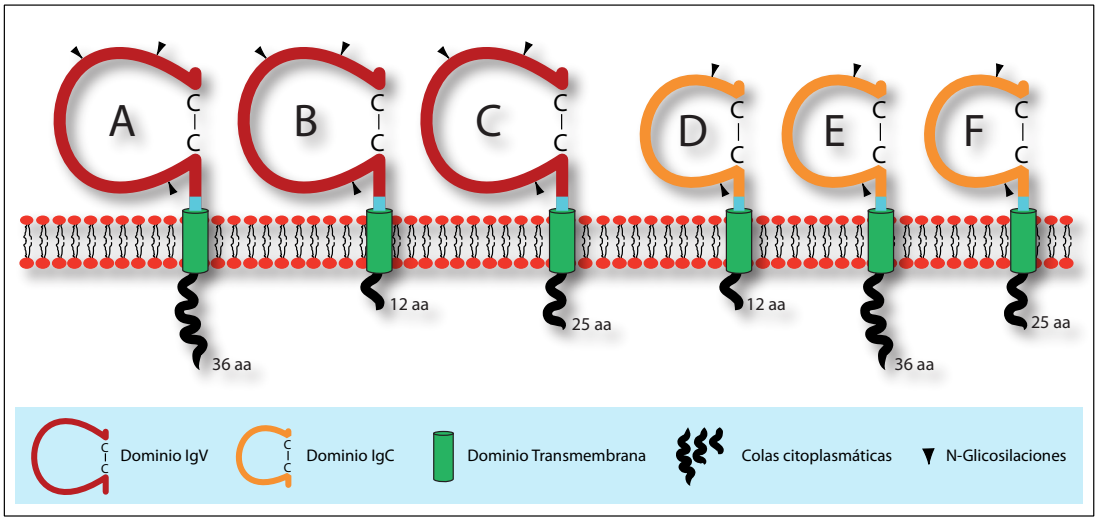


Figura 1-10. Representación esquemática de las potenciales isoformas proteicas generadas por el splicing alternativo del gen *NCR3*.

La estructura genómica del gen *NCR3* consta, por lo tanto, de 5 potenciales exones (Figura 1-9). El exón 1, de 307 pb desde el inicio de transcripción, que contiene el codón de iniciación de la traducción en la posición +264, precede al intrón 1 de 2552 pb. El exón 2 comprende 345 pb, pero puede subdividirse en dos exones independientes de 152 pb (2I) y 118 pb (2II), dejando un pequeño intrón intermedio de 75 pb. El intrón 2, de 148 pb, precede al exón 3 que comprende 108 pb. Después de un pequeño intrón de 137 pb encontramos el primer exón 4 (4I) de 115 pb cuyo codón de terminación se sitúa en la posición +77 con respecto al inicio del exón. Tras este primer exón final existe una pequeña región intrónica de 42 pb a la que le sigue el exón 4II de 337 pb, solapante con el exón 4III (282 pb), que comenzaría 55 pb después del primero. Pese a que las secuencias de los exones 4II y 4III solapan en una parte importante, la secuencia teóricamente codificante de ambos exones no lo hace, ya que en los 55 pb no compartidos existe un codón de terminación (+38 respecto del exón 4II) en fase con el inicio de la traducción descrito, y por lo tanto, la secuencia que el exón 4II comparte con el exón 4III se comportaría como UTR para los mensajeros portadores de este exón. Existe en el exón 4III otro codón de terminación en fase con el codón de inicio de la traducción en la posición +110 con respecto al propio exón (Figura 1-9).

En base a los resultado obtenidos por *Neville et al* (1999) (Neville and Campbell, 1999) se deduce que el gen *NCR3* puede generar seis potenciales isoformas proteicas diferentes (A-F en Figura 1-10). Estas hipotéticas proteínas tendrían un

péptido señal compuesto por los 18 primeros aminoácidos, codificados en el exón 1. Las isoformas A, B y C contendrían un potencial dominio inmunoglobulina (Ig) de tipo V (IgV, aminoácidos 19-128) codificado por el exón 2 completo, estando las cisteínas de las posiciones 39 y 108 implicadas en la formación del puente disulfuro característico de este tipo de dominio (Figura 1-10). En las variantes *D*, *E* y *F* el exón 2 se encuentra dividido en 2 partes, sin embargo, la fase de lectura desde el codón de inicio de la traducción se mantiene, lo que provoca una deleción de 25 aminoácidos que genera un cambio en el tipo de dominio inmunoglobulina codificado, de tipo V a tipo C2 (IgC, Figura 1-10) (Neville and Campbell, 1999). Todas las isoformas contienen una región transmembrana codificada por el exón 3, que estaría formada por los aminoácidos 139 a 161 en las isoformas A, B y C, y por los aminoácidos 114 a 136 en D, E y F (Figura 1-10). El resto de las proteínas se comportaría como dominio citoplasmático, codificados por los tres exones 4 alternativos (4I, 4II y 4III), de 36 aa en el caso de las isoformas A y E, 25 aa en C y F y tan sólo 12 en las variantes B y D (Figura 1-10).

Tan sólo unos meses después de la descripción del AS del gen *NCR3*, se identificó la función de este gen. *Pende et al.* (1999) (Pende et al., 1999) identificaron un nuevo receptor presente en células NK, al que denominaron NKp30, implicado en la activación de estas células frente a diferentes células tumorales. Tras el rastreo de una librería de cDNA, identificaron la secuencia del mensajero correspondiente con este nuevo receptor, que resultó ser idéntica a la variante *1C7c*, recientemente descrita. Sin embargo, resulta llamativo, que pese a mencionar la existencia de otras variantes de splicing, éstas y sus posibles implicaciones funcionales fueran ignoradas, siendo ésta la tendencia seguida en la investigación de este receptor.

1.4 Las células *Natural Killer* y el receptor NCR3

Pese a que la expresión del gen *NCR3* puede producir al menos seis potenciales isoformas proteicas diferentes, en este apartado se habla del receptor NCR3 de manera genérica sin especificar una isoforma concreta, dado que la mayoría de los trabajos referidos tampoco hacen esta especificación.

El receptor NCR3, perteneciente a la familia *Natural Citotoxicity Receptor*, es uno de los tres principales receptores activadores de células *Natural killer* (NK), que se identificaron inicialmente por su capacidad de eliminar células tumorales en ausencia

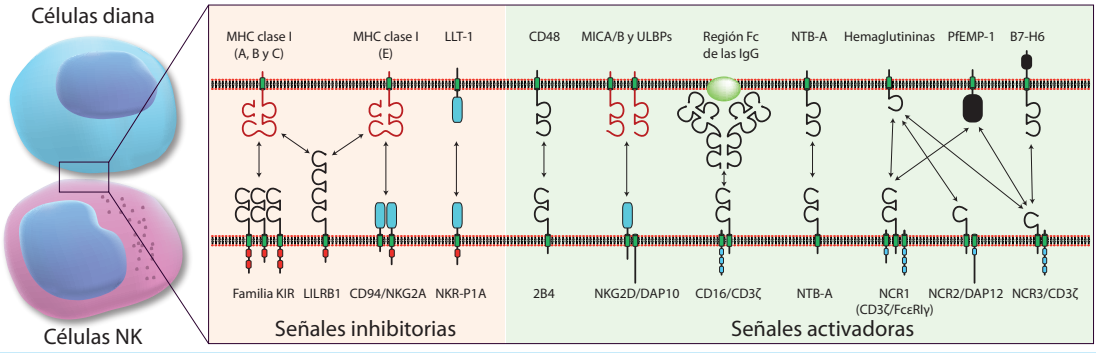


Figura 1-11. Principales receptores inhibidores y activadores, y sus ligandos, implicados en la regulación de la actividad de las células NK.

de estimulación previa, y por no presentar en su superficie receptores generados por recombinación homóloga como en los linfocitos T o B (receptor TCR y BCR, respectivamente) (Bryceson et al., 2006; Moretta et al., 2002a). También son responsables de la eliminación de células infectadas por virus y otros microorganismos infecciosos como bacterias, hongos y parásitos (Bogdan, 2012; Jost and Altfeld, 2013; Mavoungou et al., 2007; Schmidt et al., 2013), proporcionando una respuesta rápida, antes de que se desencadene la respuesta del sistema inmune adaptativo. Por ello, se considera que las células NK son las principales efectoras del sistema inmune innato, ofreciendo una respuesta primaria frente a procesos infecciosos y el desarrollo de tumores (Bryceson et al., 2006). Las células NK son linfocitos derivados de la médula ósea, que completan su maduración en órganos linfoides secundarios (*"Secondary lymphoid tissue"*, SLT) (Caligiuri, 2008), como la amígdala o los nodos linfáticos, aunque se han detectados precursores inmaduros comunes con los linfocitos T en el timo (Moretta et al., 2002c; Sanchez et al., 1994). Comparten con sus parientes, los linfocitos T citotóxicos ($CD8^+$), los mecanismos para la eliminación de las células diana, a través de la formación de una sinapsis inmunológica con ésta, la secreción de perforina y granzima, y la inducción de apoptosis a través de ligandos como FasL y TRAIL (Miletic et al., 2013).

A diferencia de los linfocitos T y B, cuya activación depende de los péptidos presentados en las moléculas MHC de clase I y II respectivamente, la activación de las células NK no depende de la presentación antigénica, aunque si requiere de la presencia de moléculas de MHC de clase I para el control de su citotoxicidad. La activación de las células NK está regulada por un fino equilibrio de señales que les permiten discriminar las células sanas de aquellas que han sido infectadas o están sufriendo un proceso de transformación. Por un lado, las células NK poseen

diversos receptores inhibidores que reconocen las moléculas de MHC de clase I como indicador del estado de salud de la célula diana (Figura 1-11). Las moléculas clásicas del MHC de clase I (tipo A, B y C) son reconocidas, principalmente, por miembros de la familia de receptores *"Killer immunoglobin (Ig)-like receptors"* (KIR) en humano, mientras que las moléculas MHC de clase I no clásicas (tipo E) son reconocidas por el heterodímero CD94/NKG2A, perteneciente a la familia de las lectinas (Figura 1-11) (Bryceson et al., 2006; Caligiuri, 2008; Jost and Altfeld, 2013; Long et al., 2013; Moretta et al., 2002a). Ambos tipos de receptores poseen en sus respectivos dominios intracelulares motivos *"Immunoreceptor tyrosine-based inhibitory motif"* (ITIM), que son fosforilados tras la interacción con las moléculas de MHC de clase I, permitiendo la asociación de las fosfatasas *"Src homology 2 domain-containing phosphatase"* (SHP-1 o SHP-2) cuya actividad conduce a inhibición de las células NK (Bryceson et al., 2006; Caligiuri, 2008). Esta señal inhibitoria, mediada por la presencia de moléculas de MHC de clase I en la célula diana, es dominante sobre las señales activadoras, evitándose de este modo la eliminación de células sanas. Por lo tanto, la ausencia o disminución de la expresión de las moléculas de MHC de clase I, lo que ocurre en procesos infecciosos o en el desarrollo tumoral para evitar la respuesta citotóxica de los linfocitos T $CD8^+$ (Lanier, 2008), es el primer paso para la activación de las células NK, aunque no es suficiente para desencadenar la respuesta citotóxica de estas células, sino que es necesaria la estimulación de éstas a través de los diversos receptores activadores.

Aunque ha quedado demostrado que la activación de las células NK, en ausencia de señales inhibitorias, depende de la estimulación coordinada de varios receptores y co-receptores, cabe destacar el papel principal que tienen el receptor NKG2D y los miembros de la familia NCR en la activación de estas células frente a diversos tipos de células tumorales y células infectadas (Figura 1-11) (Moretta et al., 2000). El receptor NKG2D, perteneciente a la familia de las lectinas y expresado de manera constitutiva en las células NK, y también en células T, reconoce a los ligandos *"MHC class I chain-related proteins A and B"* (MICA/B) y a los miembros de la familia *"UL16 binding protein"* (ULBP), que se inducen en las células diana en situaciones de estrés genotóxico y durante procesos de transformación, permitiendo su reconocimiento y eliminación (Figura 1-11) (Chitadze et al., 2013; Jost and Altfeld, 2013; Sutherland et al., 2006). Por otro lado, los receptores de la familia NCR (NCR1, NCR2 y NCR3), pertenecientes a la super-familia de las inmunoglobulinas, colaboran, entre sí y con

otros co-receptores, en la activación de las células NK frente a células tumorales. Sin embargo, sólo en el caso del receptor NCR3 se ha identificado un potencial ligando en estas células, B7-H6 que parece expresarse de manera exclusiva en células tumorales (Figura 1-11) (Brandt et al., 2009; Joyce et al., 2011; Kaifu et al., 2011; Li et al., 2011). También se ha sugerido, con cierta controversia, que el reconocimiento de las células tumorales por parte de los tres miembros de la familia NCR podría estar mediado por la presencia de diferentes glicosaminoglicanos (GAGs) en la membrana celular (Bloushtain et al., 2004; Hecht et al., 2009; Hershkovitz et al., 2008; Ito et al., 2012; Porgador, 2005; Warren et al., 2005). Estos receptores han sido también implicados en la identificación y eliminación de células infectadas por diferentes virus. Así NCR1 y NCR2, reconocen la hemaglutinina del virus de la gripe y del virus Sendai (Arnon et al., 2001; Jarahian et al., 2009) y la hemaglutinina-neuraminidasa del virus de la enfermedad de Newcastle (Jarahian et al., 2009), NCR1 y NCR3 reconocen las hemaglutininas de Vaccinia y Ectromelia (Jarahian et al., 2011) y NCR3 también es capaz de interaccionar con la proteína pp65 del citomegalovirus humano, sin embargo, en este caso la señalización es inhibida por dicha interacción, siendo este un mecanismos utilizado por el virus para evadir la respuesta inmune (Arnon et al., 2005). Además, tanto NCR1 como NCR3, han sido implicados en la respuesta inmune frente a la malaria, a través del reconocimiento del ligando *"P. falciparum erythrocyte membrane protein–1"* (PfEMP-1) en la superficie de los eritrocitos infectados por el parásito *Plasmodium falciparum* (Jost and Altfeld, 2013; Mavoungou et al., 2007). Tanto NCR1 como NCR3 se expresan de manera constitutiva en las células NK, mientras que la expresión del receptor NCR2 se induce tras la estimulación con IL-2 (de Rham et al., 2007), lo que propone un papel de apoyo para este receptor.

Una característica común a estos receptores activadores es la ausencia de motivos que permitan la transmisión de la señal al interior celular, por lo que en la mayoría de los casos éstos se asocian con proteínas adaptadoras que poseen un dominio "*Immunoreceptor tyrosine-based activating motif*" (ITAM) u otros motivos susceptibles de ser fosforilados, que generalmente interaccionan a través de puentes salinos creados con aminoácidos cargados positivamente localizados en la región transmembrana de los receptores (Biassoni et al., 2003). NKG2D se asocia con la proteína adaptadora DAP10, que tras su fosforilación, permite la señalización a través de la ruta *"Phosphatidylinositide 3-kinase"* (PI3K). NCR1 se asocia con CD3ζ y FcεRIγ, ambos con motivos ITAM, NCR3 se asocia con CD3ζ y NCR2 interacciona

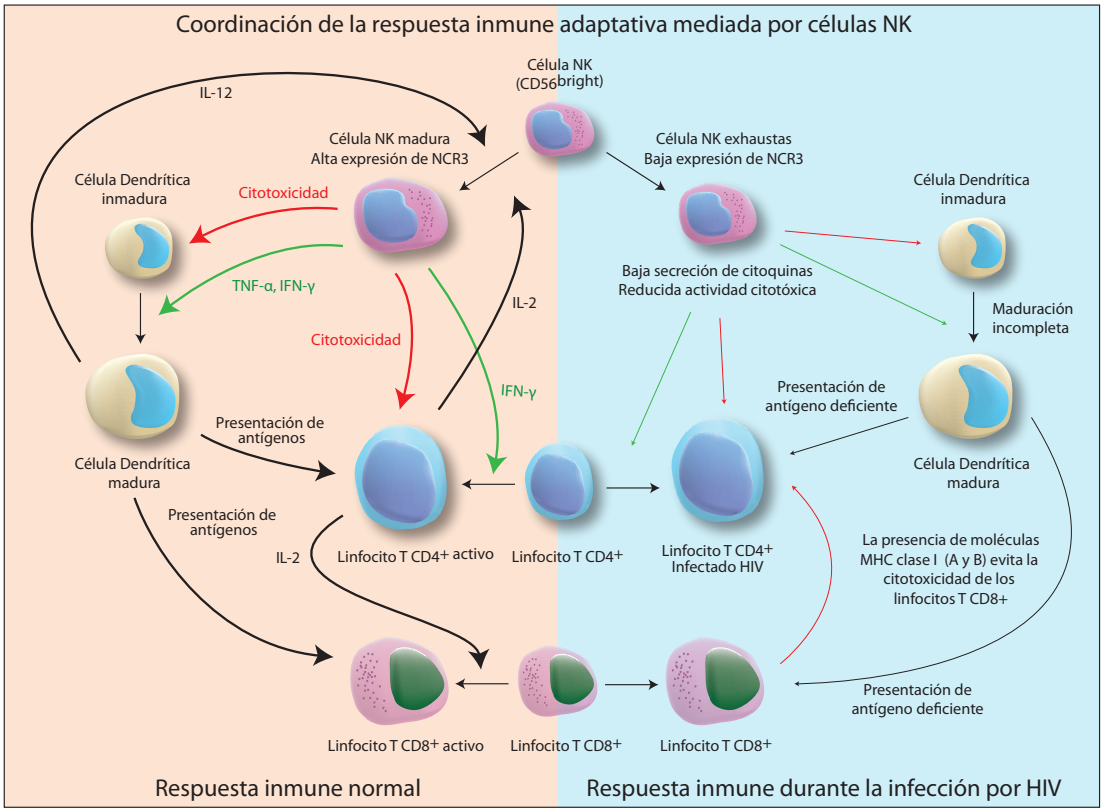


Figura 1-12. Mecanismo de control de las células NK sobre la respuesta inmune adaptativa. En el panel de la izquierda se muestra este proceso durante la respuesta inmune normal. En el panel de la derecha se muestra la descoordinación de este proceso durante la infección por HIV. El grosor de las fechas indica la intensidad de las conexiones.

con DAP12, que también posee un motivo ITAM (Biassoni et al., 2003; Moretta et al., 2002b). La fosforilación de estás proteínas adaptadoras tras la adecuada estimulación de los receptores, permite la interacción de las quinasas Syk y ZAP-70, que en última instancia desencadenan la señalización necesaria para la activación de la respuesta citotóxica, incluyendo la reorganización del citoesqueleto para formar la sinapsis inmunológica con la célula diana, la exocitosis de las vesículas de gramzima y perforina, y la activación de la rutas de señalización "*Nuclear factor kappa-light-chain-enhancer of activated B cells*" (NF-κB) y "*Nuclear factor of activated T-cells*" (NFAT) (Lanier, 2008).

Las células NK, aunque no presentan la diversidad clonal de los linfocitos T o B, también muestran un alto grado de diversidad, en especial en la expresión de los distintos receptores inhibidores de la familia KIR, expresando cada célula NK individual un sub-conjunto de estos receptor (Caligiuri, 2008), lo que les permite reconocer pequeñas variaciones en la expresión de las moléculas de MHC de clase I. Cabría

esperar que la ausencia de receptores inhibidores, o la deficiencia global de expresión de moléculas de MHC de clase I en el organismo, provocase una hiperactividad de las células NK. Sin embargo, los resultados experimentales han permitido comprobar que las células NK presentan en realidad una reducida capacidad citotóxica en estas situaciones (Bryceson et al., 2006; Cruz-Munoz and Veillette, 2010; Kim et al., 2005; Long et al., 2013). Se ha postulado que las señales inhibitorias, a través de un mecanismo aún desconocido denominado "*educación*", están implicadas en el establecimiento de la auto-tolerancia de las células NK. Se cree que la ausencia global de señales inhibitorias evita el desarrollo de la citotoxicidad de estas células a través de la estimulación de los receptores activadores, dado que sin éstas no es posible la discriminación de las células sanas de aquellas que están sufriendo un proceso infeccioso o tumoral. Por lo tanto, la educación de las células NK es un mecanismo conservador que evita la auto-reactividad de las células NK, y sólo en el caso de que estas células reciban esta educación a través de las señales inhibitorias podrán desarrollar completamente su citotoxicidad.

Una excepción a este proceso de educación es la citotoxicidad mediada por anticuerpos (*"Antibody-dependent cellular cytotoxicity"*, ADCC) que las células NK pueden desarrollar a través del receptor de inmunoglobulinas de baja afinidad CD16 (FcγRIIIA), que se asocia con el adaptador CD3ζ para la transmisión de la señalización (Figura 1-11). La expresión de este receptor permite que las células NK eliminen células marcadas con anticuerpos (tipo IgG). Esta capacidad no está limitada por el proceso de educación debido a que el marcaje con anticuerpos procede de una respuesta del sistema inmune adaptativo y por lo tanto, ya ha superado todos los controles necesarios para evitar el desarrollo de auto-reactividad. Esta comunicación entre las células NK y el sistema inmune adaptativo no es la única conexión existente entre ellos, y aunque las células NK se clasifican dentro del sistema inmune innato, desarrollan un papel activo en la regulación del sistema inmune adaptativo, a través de la secreción de citoquinas, principalmente IFN-γ y TNF-α, y contribuyen a la maduración de las células dendríticas (DC) en los órganos linfoides secundarios (Caligiuri, 2008).

Aunque las células NK muestran una gran diversidad, como se ha mencionado, se pueden clasificar en dos poblaciones principales: las que presentan un fenotipo $CD56^{dim}$, que suponen el 90% del total de las células NK en circulación y presentan

altos niveles de citotoxicidad, y las CD56^{bright}, mayoritariamente localizadas en los órganos linfoides secundarios, con una menor actividad citotóxica y una mayor capacidad de respuesta a citoquinas (Deguine and Bousso, 2013).

Esta segunda población (CD56^{brigh}) ha sido relacionada con la maduración de las células dendríticas inmaduras. Las células dendríticas se localizan en los órganos periféricos, y tras la detección de agentes patógenos, y por lo tanto, tras la captación de antígenos, migran hacia los órganos linfoides secundarios para actuar como células presentados de antígeno ("*Antigen presenting cell*", APC) para la estimulación de la respuesta del sistema inmune adaptativo. En este contexto de inflación, las células dendríticas, a través de la secreción de IL-12 inducen la proliferación y activación de las células NK CD56^{bright} en los órganos linfoides secundarios que adquieren un fenotipo similar al de las células NK circulantes (CD56^{dim}CD94/NKG2D^{+}KIR^{-}NCR^{+}). Esta activación de las células NK induce la secreción de IFN-γ y TNF-α, que promueve la maduración de las células dendríticas recíprocamente, y confiere a las células NK de la capacidad de eliminar células dendríticas inmaduras, lo que no afecta a las células dendríticas maduras por la mayor expresión en éstas de moléculas de MHC de clase I (especialmente de tipo E), junto a la expresión de los ligandos CD80 (B7.1) y CD83 (B7.2), necesarios para la co-activación de los linfocitos T (Figura 1-12). Se ha propuesta al receptor NCR3 como un regulador clave para este proceso, siendo necesario para la activación y proliferación de las células NK, así como para la estimulación de la secreción de citoquinas, y considerándose el principal receptor activador necesario para la eliminación de células dendríticas inmaduras (Caligiuri, 2008; Della Chiesa et al., 2003; Ferlazzo et al., 2004; Ferlazzo et al., 2002; Vitale et al., 2005; Wai et al., 2010). Aunque el mecanismo exacto es aún desconocido, se cree que la secreción de exosomas con la proteína BAT-3 (BAG6) por parte de las células dendríticas, principalmente por las inmaduras, podría ser el ligando reconocido por NCR3 en este proceso (Pogge von Strandmann et al., 2007; Simhadri et al., 2008).

La fuerte auto-regulación de la activación y la maduración células NK y las células dendríticas podría tener un papel crucial en el desarrollo de SIDA en pacientes infectados con el "*Human Immunodeficiency Virus"* (HIV). Estos pacientes, pese a desarrollar inicialmente una fuerte respuesta citotóxica de los linfocitos T CD8^{+}, presentan una reducción en su capacidad de respuesta inmune generalizada. Entre otros factores, cabe destacar la aparición de células NK exhaustas, que aunque

muestran marcadores de activación (CD69 y HLA-DR), se caracterizan por una reducida expresión de receptores activadores, entre los que destaca NCR3, y una baja secreción de citoquinas. Esto conduce a un desequilibrio de la maduración de las células dendríticas y por ende, a una baja respuesta de los linfocitos T frente al virus (Figura 1-12) (De Maria and Moretta, 2008; Rutjens et al., 2007). Adicionalmente, las células T $CD4^+$ infectadas por el virus HIV, muestran una reducida expresión de moléculas de MHC de clase I tipo A y B, lo que evita la activación de los linfocitos T $CD8^+$, aunque mantienen una elevada expresión de moléculas tipo E, lo que evita la activación de las células NK. Por lo tanto, el virus HIV podría escapar al control del sistema inmune a través del desequilibrio generado en el proceso de maduración de las células dendríticas mediado por las células NK y en especial por el receptor NCR3.

Esta hipótesis, se sustenta en parte en los datos obtenidos del análisis de chimpancés infectados con HIV, que son capaces de soportar una infección crónica sin desarrollar SIDA. En chimpancé, las células NK no expresan el receptor NCR3 de manera constitutiva, sino que se induce tras la activación de éstas, lo que sugiere que la maduración conjunta de las células NK y las células dendríticas ocurre a través de otro mecanismo, lo que permite a esta especie resistir la infección, no mostrando desequilibrios en la función de sus células NK, que responden de manera normal (De Maria and Moretta, 2008; Jost and Altfeld, 2013; Rutjens et al., 2007; Rutjens et al., 2010).

1.5 Implicaciones funcionales del AS del gen *NCR3*

En los últimos años se ha conseguido un gran avance en el estudio de las funciones del receptor NCR3, habiendo sido implicado en la activación de la citotoxicidad de las células NK frente a células tumorales, células infectadas por diversos virus y parásitos, y resulta un factor crucial en la maduración de las células dendríticas. Asimismo, se han identificado diversos ligandos para la estimulación del receptor NCR3 en estas diversas situaciones (B7-H6, hemaglutinina, PfEMP-1, BAT3 y GAGs). Sin embargo, apenas se sabe nada de las potenciales implicaciones de las distintas variantes de splicing del gen *NCR3*. Este fenómeno se debe a que la investigación del receptor NCR3 ha estado restringida mayoritariamente al área de la inmunología, siendo el uso de anticuerpos una de sus principales herramientas, lo que imposibilita en la mayoría de los casos la detección de isoformas proteicas.

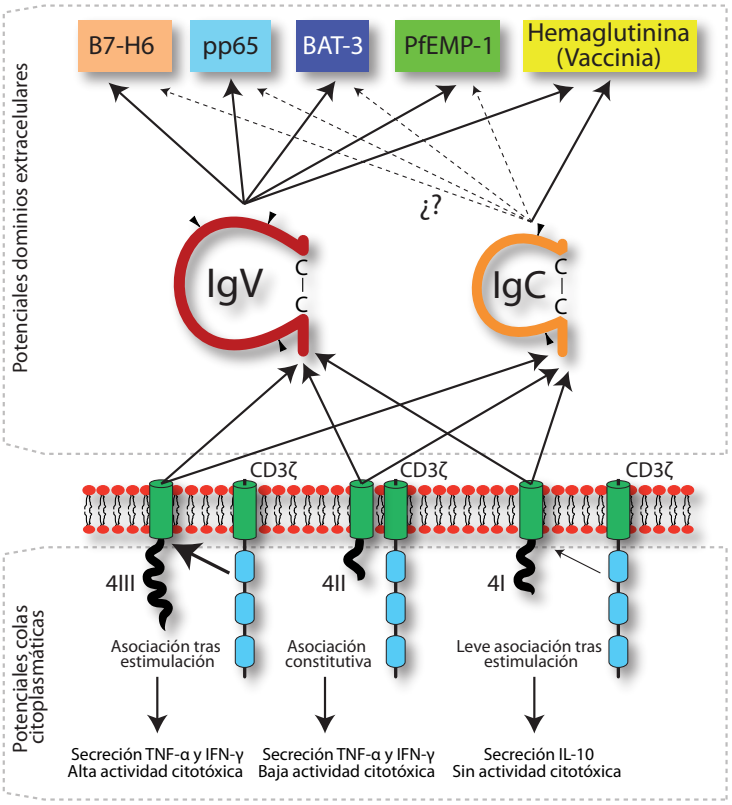


Figura 1-13. Potenciales implicaciones funcionales del AS del gen *NCR3* en humano. La division conceptual de las isoformas descritas del receptor NCR3 permite dividir la influencia del AS en dos términos. Por un lado, la influencia sobre la composicion del dominio extracelulas (arriba) que podría influir sobre el reconocimiento de los distintos ligandos descritos. Por otro lado, la transmisión de la señal intracelulas (abajo) se ve afectada por el AS.

Las seis potenciales isoformas proteicas descritas del gen *NCR3* comparten el péptido señal y la región transmembrana, sin embargo, presentan diferencias, generadas a través de AS, en los dominios extracelulares y en las colas citoplasmáticas. Extracelularmente podemos dividir las isoformas en dos grupos: aquellas que presentan un dominio IgV o un dominio IgC. Cabría esperar que estos dos grupos de isoformas presentaran diferencias en cuanto a la especificidad de ligando o a la afinidad por éstos, lo que podría explicar la enorme diversidad de ligandos descritos para el receptor NCR3 y la diversidad de funciones atribuidas a éste (Figura 1-13). En este sentido, en la bibliografía sólo existe un trabajo que se hace eco, aunque de manera anecdótica, de esta potencial influencia, y analiza la interacción de los dos potenciales dominios extracelulares con la hemaglutinina del virus vaccinia en células infectadas, no observando diferencias significativas (Jarahian et al., 2011).

Por otro lado, también se podrían clasificar las isoformas de NCR3 en función de las tres posibles colas citoplasmáticas, codificadas en los tres posibles exones finales (4I, 4II o 4III). Aunque ninguna de las colas citoplasmáticas parece presentar dominios conservados que nos permitan inferir diferencias en la señalización intracelular, cabría preguntarse si existen alguna diferencia en este señalización. En este sentido,

Delahaye et al. (2011) (Delahaye et al., 2011), han descrito una influencia del perfil de expresión de las variantes *A*, *B* y *C* en la supervivencia de pacientes con sarcomas gastrointestinales, mostrando que los pacientes con mayor expresión de las variantes *A* y *B* presentan un mejor pronóstico que aquellos en los que la variante mayoritaria es la *C*. Además, en estudios funcionales *in vitro* usando la línea celular NKL transfectadas con estas tres variantes, observan que la activación a través de las isoformas A y B conduce a la secreción de IFN-γ y TNF-α, aunque sólo en el caso de la isoforma A éstas mostraron capacidad citotóxica. Sin embargo, las células transfectadas con la variante *C* no secretaron éstas citoquinas tras su estimulación ni mostraron capacidad citotóxica, aunque sí detectaron un ligero incremento en la secreción de la citoquina inmunosupresora IL-10. En consonancia, observaron que la isoforma A se asocia con CD3ζ tras la estimulación, mientras que la isoforma B lo hace de manera constitutiva y la isoforma C sólo incremente ligeramente su asociación, lo que provoca cambios en la activación de la ruta NF-κB, que en última instancia es responsable de la estimulación de la secreción de citoquinas (Delahaye et al., 2011). El valor predictivo del perfil de las variantes *A*, *B y C* en el pronóstico de pacientes de sarcomas gastrointestinales, se ha intentado trasladar a pacientes con HIV, en los hay una represión de las expresión de los receptores activadores de la familia NCR, en especial NCR3. Sin embargo, parece que en pacientes de HIV no hay diferencias en el perfil de expresión de estas variantes (Prada et al., 2013). Por lo tanto, y aunque el estudio funcional de las distintas isoformas del gen *NCR3* está lejos de completarse, los primeros indicios apuntan que el AS del gen podrían tener importantes consecuencias en el funcionamiento del sistema inmune, que hasta la fecha han sido ignoradas (Figura 1-13).

1.6 *NCR3* en otras especies

En las principales bases de datos podemos encontrar anotaciones del gen *NCR3* (completas o parciales) en un total de 44 especies, siendo en su mayoría mamíferos (42), y de los que 14 son primates. Además, también encontramos secuencias de rana (*Xenopues tropicalis*), que presenta al menos siete copias del gen *NCR3* (Flajnik et al., 2012), y de una especie de tortuga (*Pelodiscus sinensis*). Sin embargo, en la mayoría de las especies en las que se ha descrito el gen *NCR3*, las anotaciones se basan en predicciones automáticas (36/44, ver Tabla 1-2), principalmente utilizando las secuencias descritas en humano como referencia, asumiendo en algunas casos

Especie	Nombre común	NCBI	Ensembl
Ailuropoda melanoleuca	Panda	-	Predicción
Bos taurus	Vaca	Secuencia depositada	Secuencia depositada
Callithrix jacchus	Tití	-	Predicción
Canis lupus familiaris	Perro	Predicción	Predicción
Cebus apella	Mono capuchino	-	-
Ceratotherium simum simum	Rinoceronte	Predicción	-
Colobus guereza	Colobo	-	-
Condylura cristata	Topo de nariz estrellada	Predicción	-
Echinops telfairi	Erizo de madagascar	-	Predicción
Equus caballus	Caballo	Predicción	Predicción
Erinaceus europaeus	Erizo	-	Predicción
Felis catus	Gato	Predicción	Predicción
Gorilla gorilla	Gorilla	Predicción	Predicción
Homo sapiens	Humano	Secuencia depositada*	Secuencia depositada
Ictidomys tridecemlineatus	Ardilla de tierra	Predicción	-
Loxodonta africana	Elefante	-	Predicción
Macaca fascicularis	Mono cangrejero	-	-
Macaca mullata	Mono rhesus	Secuencia depositada*	Secuencia depositada
Macropus eugenii	Wallaby	-	Predicción
Mesocricetus auratus	Hámster dorado	Predicción	-
Microcebus murinus	Lemur ratón	-	Predicción
Microtus ochrogaster	Topillo	Predicción	-
Mus musculus	Ratón	Predicción	Predicción
Mustela putorius furo	Hurón	-	Predicción
Myotis lucifugus	Murciélago	-	Secuencia depositada
Nomascus leucogenys	Gibón	Predicción	Predicción
Ochotona princeps	Pica americano	Predicción	Predicción
Octodon degus	Degú	Predicción	-
Odobenus rosmarus divergens	Morsa	Predicción	-
Orcinus orca	Orca	Predicción	-
Oryctolagus cuniculus	Conejo	-	Secuencia depositada
Otolemur garnettii	Gálago	Predicción	Predicción
Ovis aries	Oveja	Predicción	-
Pan paniscus	Bonobo	Predicción	-
Pan troglodytes	Chimpacé	Secuencia depositada	Secuencia depositada
Papio anubis	Papión	Predicción	-
Papio cynocephalus	Papión	-	-
Pelodiscus sinensis	Tortuga china	-	Predicción
Pongo abelli	Orangután de Sumatra	Predicción	Predicción
Pongo pygmaeus	Orangután de Borneo	-	-
Procavia capensis	Damán	-	Predicción
Rattus norvegicus	Rata	Secuencia depositada	Secuencia depositada
Saimiri boliviensis	Mono ardilla	Predicción	-
Sorex araneus	Musaraña	Predicción	Predicción
Sus scrofa	Cerdo	Secuencia depositada*	Secuencia depositada*
Tarsius syrichta	Tarsero filipino	-	Predicción
Trichechus manatus latirostris	Manatí	Predicción	-
Tursiops truncatus	Delfín	Predicción	Predicción
Xenopus tropicalis	Rana	-	Predicción

Tabla 1-2. Especies en las que se ha descrito el gen *NCR3*, ya sea por predicción o gracias a secuencias depositadas de la especie correspondiente. Las especies sombreadas son las incluidas porteriormente en este trabajo.
*Parcialmente basado en predicciones

la posible existencia de las mismas variantes de splicing.

Desde el punto de vista funcional, el estudio del gen *NCR3* en otras especies se ha visto restringido por tratarse de un pseudogen en ratón (*Mus músculus*) (Hollyoake et al., 2005), sin embargo, tanto en rata (*Rattus norvegicus*) como en chimpancé (*Pan troglodytes*) se ha estudiado su implicación en la activación de las células NK (Hsieh et al., 2006; Rutjens et al., 2007). Además, en rata, a diferencia de humano, se ha observado que la maduración de las células dendríticas mediada por las células NK no depende del receptor NCR3 (Wai et al., 2010), revelando una regulación diferencial del sistema inmune. Esto resulta lógico si tenemos en cuenta que la única isoforma descrita en rata es homóloga de la isoforma C humana, cuya estimulación no promueve la secreción de las citoquinas necesarias para la maduración de las células dendríticas (Delahaye et al., 2011).

Aunque los indicios de la influencia del AS en el sistema inmune han aumentado en los últimos años, su implicación sigue siendo infravalorada (Lynch, 2004; Martinez and Lynch, 2013). El sistema inmune resulta crucial para la supervivencia de los organismos, y por lo tanto, se encuentra en continua evolución. Un ejemplo de esto es la expansión de las moléculas de MHC de clase I y II, y su alto grado de polimorfismo, que permiten al sistema inmune adaptarse de manera eficiente a los distintos patógenos a los que debe enfrentarse. De manera análoga, la generación de diversos receptores mediante AS podría responder a una estrategia similar, aumentando la capacidad de respuesta del sistema inmune. En este sentido, se han descrito diversos receptores, además de NCR3, que podrían presentar varias isoformas entre los que cabe destacar NCR1, NCR2 y varios miembros de la familia NKG2 (Hollyoake et al., 2005; LaBonte et al., 2000).

Por lo tanto, el estudio del AS del gen *NCR3* resulta crucial para comprender mejor el funcionamiento de los mecanismos biológicos en los que está implicado. Del mismo modo, el análisis comparativo en diferentes especies puede mejorar nuestra visión de estos procesos y ayudarnos a entender cómo el AS ha podido influir en el desarrollo del sistema inmune así como en la diversificación de las especies.

2. Objetivos

Objetivos

En función de los antecedentes descritos en el apartado anterior se plantearon los siguientes objetivos para este trabajo:

1. Anotar el gen *NCR3* en todas las especies incluidas en este estudio y analizar la conservación de su entorno genómico: para el desarrollo de este objetivo se recurrió al análisis comparativo de la información disponible en las distintas bases de datos. En los casos en los que no hubiera información se recurrió a la secuenciación del DNA genómico.

2. Analizar las variantes de splicing del gen *NCR3* diferentes especies de mamíferos: se utilizaron muestras de diferentes tejidos de cada una de las especies incluidas en este apartado para el análisis de las variantes de splicing mediante PCR anidada, que se complementó con el análisis de los datos disponibles procedentes de experimentos de secuenciación masiva (RNA-seq).

3. Analizar de las variantes de splicing del gen *NCR3* en sangre de diferentes primates: en este apartado se analizaron muestras de sangre de diferentes primates para la detección de variantes de splicing mediante PCR anidada. Adicionalmente, se analizaron secuencias de RNA-seq disponibles para la búsqueda de variantes de splicing en las especies de primates.

4. Caracterizar las variantes de splicing del gen *NCR3* en humano: se cuantificó la expresión en diferentes tejidos de las variantes de splicing mayoritarias detectadas en humano. Se analizaron las variantes codificantes mediante transfección en células de mamíferos para determinar su localización subcelular, la formación de dímeros y la presencia de glicosilaciones. Adicionalmente, se estudió la interacción de las isoformas detectadas con el ligando celular B7-H6.

3. Materiales y métodos

Materiales y métodos

3.1 Análisis informático

3.1.1 Bases de Datos

La búsqueda de secuencias relativas al gen *NCR3* (DNA genónico, RNA mensajeros y proteínas) de las distintas especies incluidas en este trabajo y la anotación de éstas se obtuvieron de las siguiente bases de datos: *NCBI* (*Nacional Center for Biotechnology Information*, http://www.ncbi.nlm.nih.gov) (Benson et al., 2000), *Ensembl* (http://www.ensembl.org) (Hubbard et al., 2002), *MGI* (*Mouse Genome Informatics*, http://www.informatics.jax.org/) (Eppig et al., 2000), *UCSC* (*University of California Santa Cruz*, http://genome.ucsc.edu/) (Kent et al., 2002), Uniprot (http://www.uniprot.org/) (O'Donovan et al., 2002), *KEGG* (*Kyoto Encyclopedia of Genes and Genomes*, http://www.genome.jp/kegg/) (Kanehisa and Goto, 2000; Kanehisa et al., 2012), *RGD* (*Rat Genome Database*, http://rgd.mcw.edu/wg/home) (Laulederkind et al., 2013), GeneCards (http://www.genecards.org/) (Rebhan et al., 1997) y *HGNC* (*HUGO Gene Nomenclature Committee*, http://www.genenames.org/) (Gray et al., 2013). La divergencia evolutiva entre especies se obtuvo de *TimeTree* (http://www.timetree.org/) (Hedges et al., 2006).

3.1.2 Análisis y alineamiento de secuencias

Los análisis de las secuencias obtenidas en el desarrollo de este trabajo mediante secuenciación Sanger se realizaron con el programa *4Peaks*. La búsqueda de las posibles fases de lectura se determinó con *ORF Finder* (http://www.ncbi.nlm.nih.gov/projects/gorf/), *Translator* (http://www.fr33.net/translator.php) y *Transeq* (http://www.ebi.ac.uk/Tools/st/emboss_transeq/) (Rice et al., 2000).

El análisis y comparación de las secuencias disponibles en las bases de datos, así como las obtenidas en el desarrollo de este trabajo, se realizó con las siguientes herramientas de alineamiento: *Blast* (http://blast.ncbi.nlm.nih.gov/) (Camacho et al., 2009), *Multalin* (http://multalin.toulouse.inra.fr/multalin/) (Corpet, 1988), *Align* (http://www.ebi.ac.uk/Tools/psa/emboss_needle/) (Rice et al., 2000), *ClustalW* (http://www.ebi.ac.uk/Tools/msa/clustalw2/) (Goujon et al., 2010; Larkin et al., 2007), *WebLogo* (http://weblogo.berkeley.edu/) (Crooks et al., 2004) y Spidey (http://www.ncbi.nlm.nih.gov/spidey/) (Wheelan et al., 2001).

El análisis de la presencia de secuencias diana de enzimas de restricción se realizó con la aplicación *Webcutter* (http://rna.lundberg.gu.se/cutter2/).

3.1.3 Predicción de sitios de splicing y secuencias de poliadenilación

La predicción de potenciales sitios donadores y aceptor de splicing se determinó mediante el uso de las siguientes aplicaciones: *FSPLICE* (http://linux1.softberry.com/berry.phtml), *NetGene2* (http://www.cbs.dtu.dk/services/NetGene2/) (Brunak et al., 1991), *ASSP* (http://wangcomputing.com/assp/index.html) (Wang and Marin, 2006) y *GENSCAN* (http://genes.mit.edu/GENSCAN.html) (Burge and Karlin, 1997).

Para la predicción de secuencias de poliadenilación se utilizó *POLYAH* (http://linux1.softberry.com/berry.phtml), *PolyAPred* (http://www.imtech.res.in/raghava/polyapred/index.html)(Ahmed et al., 2009) y *Poly(A) Signal Miner* (http://dnafsminer.bic.nus.edu.sg/PolyA.html).

3.1.4 Análisis de dominios estructurales y predicción de modificaciones post-traduccionales

La predicción general de dominios conservados en las potenciales isoformas detectadas en este trabajo se determinó con *SMART* (http://smart.embl-heidelberg.de/) (Letunic et al., 2012; Schultz et al., 1998), *PFAM* (http://pfam.sanger.ac.uk/) (Sonnhammer et al., 1997), *InterProScan* (http://www.ebi.ac.uk/InterProScan/) (Zdobnov and Apweiler, 2001) y *Prosite* (http://prosite.expasy.org/) (Sigrist et al., 2002).

La búsqueda de potenciales péptidos señal se realizó con *SignalP* (http://www.cbs.dtu.dk/services/SignalP/) (Petersen et al., 2011). La predicción de dominios transmembrana se realizó con *TMHMM* (http://www.cbs.dtu.dk/services/TMHMM/)

(Krogh et al., 2001), *TMprep* (http://www.ch.embnet.org/software/TMPRED_form.html) y *SOSUI* (http://bp.nuap.nagoya-u.ac.jp/sosui/) (Hirokawa et al., 1998). El peso molecular teórico de las isoformas detectadas se calculó con el programa *Protein Molecular Weight* (http://www.bioinformatics.org/SMS/prot_mw.html).

La presencia de potenciales modificaciones post-traduccionales se determinó con las siguientes herramientas: *NetNGlyc* (http://www.cbs.dtu.dk/services/NetNGlyc/) (Gupta, 2002), *NetOGlyc* (http://www.cbs.dtu.dk/services/NetOGlyc/) (Steentoft et al., 2013), *NetPhos* (http://www.cbs.dtu.dk/services/NetPhos/) (Blom et al., 1999) y *SH3-Hunter* (http://cbm.bio.uniroma2.it/SH3-Hunter/) (Ferraro et al., 2007).

3.1.5 Diseño y análisis de cebadores

Los cebadores utilizados se diseñaron utilizando el programa *Primer3plus* (www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) (Rozen and Skaletsky, 2000; Untergasser et al., 2007). La especificidad de éstos se analizó con *Primer-blast* (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) y su calidad se evaluó con el programa *IDT OligoAnalyzer* (http://www.idtdna.com/analyzer/Applications/OligoAnalyzer/). La síntesis de los oligonucleótidos se envió a ser realizada por Sigma.

3.2 Cultivos bacterianos

3.2.1 Cepas y medios de cultivo

Para el crecimiento de plásmidos se utilizaron bacterias *E.coli* competentes de las cepas DH5α y XL1-Blue preparadas por el método del cloruro de rubidio en el servicio de fermentación del CBMSO.

Para la producción de bácmidos recombinantes (sistema de baculovirus) se utilizó la cepa DH10bac™ (Invitrogen). Estas bacterias portan el bácmido vacío (bMON14272) como un mega-plásmido, confiriéndoles resistencia a kanamicina. Además, portan un plásmido (pMON7124) que les confiere resistencia a tetraciclina y permite la expresión de la transposasa necesaria para la recombinación entre el bácmido y el plásmido de interés para la producción de bácmidos recombinantes.

Para el crecimiento de bacterias se empleó medio Luria Broth (LB; 10 g/l de bactotriptona, 5 g/l de extracto de levadura y 10 g/l de NaCl), líquido o sólido (basado en el anterior al que se le añade 15 g/l de agar bacteriológico). Ambos medios fueron

preparados en el Servicio de Cultivos del CBMSO. Los clones bacterianos de interés se preservaron en LB líquido con glicerol al 10% a -70ºC.

3.2.2 Transformación

Para purificar los plásmidos de interés, se transformaron bacterias competentes (DH5α o XL1-Blue) por el método del choque térmico. Para ello se incubaron en hielo 50 µl de bacterias competentes con 10 µl de una mezcla de ligación (detallado en el apartado 3.4.5) o 20 ng de plásmido durante 30 min. Posteriormente se sometieron a un choque térmico de 40 segundos a 42ºC, tras lo que se incubaron 30 minutos en hielo. Después se añadió 0,5 ml de medio LB atemperado y se incubaron durante 1 hora a 37ºC para permitir la expresión de los genes de resistencia presentes en los plásmidos de interés. Tras la expresión fenotípica se centrifugaron las bacterias para reducir el volumen de medio y se sembraron en placas Petri con medio LB sólido suplementado con ampicilina (50 µg/ml) y se incubaron 16-20 horas a 37ºC. En el caso de plásmidos derivados de pGEM®-T Easy (Promega) se suplementó el medio LB sólido con X-gal (100 µg/ml) e IPTG (40 µg/ml) para habilitar la selección por color.

La transformación de bacterias DH10bac™ siguió básicamente el mismo protocolo, salvo porque se utilizó 100 µl de bacterias competentes y 3 ng de plásmido. La expresión fenotípica se extendió por 5 horas y se sembró 1/10 de las bacterias transformadas en placas Petri con medio LB sólido suplementado con kanamicina (50 µg/ml), tetraciclina (10 µg/ml), gentamicina (7 µg/ml), X-gal (100 µg/ml) e IPTG (40 µg/ml). Las placas se incubaron a 37ºC durante 48-72 horas.

3.2.3 Análisis de las colonias positivas

El análisis de las colonias positivas se realizó mediante *Polimerase Chain Reaction* (PCR). Se preparó una mezcla de reacción de PCR con los componentes indicados en el apartado 3.4.3, excepto porque no se añadió molde a la reacción. Se seleccionaron colonias utilizando puntas estériles y se re-sembraron en una nueva placa Petri con medio LB sólido y los suplementos y antibióticos adecuados, que se incubó a 37ºC hasta la comprobación del resultado de la amplificación. Cada punta utilizada para la selección de colonias se sumergió en la mezcla de PCR, siendo el plásmido (o bácmido) liberado de las bacterias en la primera fase de la amplificación, el molde de la reacción. El programa de PCR utilizado fue el estándar (ver apartado 3.4.3), excepto en el análisis de las colonias procedentes de la transformación de bacterias

DH10Bac™ donde la fase de extensión a 72ºC fue de 3 minutos.

El resultado de la amplificación se analizó cargando 5 µl de cada producto amplificado en geles de agarosa (1-2%, ver apartado 3.4.6). Las colonias positivas se crecieron en 2 ml de LB líquido con los suplementos y antibióticos adecuados y se incubaron 16 horas a 37ºC con agitación (180 rpm). Para la obtención de grandes cantidades de plásmido, se sembraron 100 µl de este cultivo de 2 ml crecido 16 horas en 200-500 ml de medio LB líquido con los suplementados adecuados durante 16-20 horas.

3.3 Cultivos celulares

3.3.1 Líneas celulares y mantenimiento

En este trabajo se han utilizado las siguientes líneas celulares: HeLa (células epitelioides obtenidas de un adenocarcinoma cervical), HEK293T (células embrionarias de riñón transformadas con el virus SV40), K562 (granulocitos indiferenciados derivados de una leucemia mieloide crónica), Jurkat (linfocitos T obtenidas de una leucemia aguda), Raji (linfocitos B obtenidos de un linfoma de Burkitt), U937 (monocitos derivados de un linfoma), YT (células NKs derivadas de un linfoma), COS-7 (fibroblastos de riñón de mono verde africano transformados con el virus SV40), CHO-K1 (células epitelioides de ovario de hámster chino) y CHO-745 (células epitelioides de ovario de hámster chino deficientes en xylosyltransferasa I) y High-Five™ (células de ovario de la polilla *Trichoplusia ni*, Invitrogen).

El medio de cultivo de las células adherentes HeLa, HEK293T y COS-7 fue *Dulbecco's Modified Eagle Medium* (DMEM) suplementado con antibióticos (75 U/ml de estreptomicina y 75 µg/ml de penicilina G), 2 mM de L-glutamina y 10% (v/v) de suero fetal de ternera (FBS) descomplementado a 56ºC durante 30 minutos. Las líneas CHO-K1 y CHO-745 se crecieron en medio DMEM suplementando con un 50% (v/v) de medio Ham's F12, antibióticos (75 U/ml de estreptomicina y 75 µg/ml de penicilina G), 2 mM de L-glutamina y 10% (v/v) de FBS. Las células en suspensión K562, Jurkat, Raji, U937 e YT se crecieron en medio *Roswell Park Memorial Institute* (RMPI) suplementado con antibióticos (75 U/ml de estreptomicina y 75 µg/ml de penicilina G), 2 mM de L-glutamina y 10% (v/v) de FBS. Todas las líneas celulares se crecieron a 37ºC en una atmósfera de 5% de CO_2 y 95% de humedad.

Las células High-Five™ pueden crecer en forma adherente y en suspensión. Para la

expansión y mantenimiento de las células se crecieron en forma adherente, mientras que para la producción de proteínas recombinantes se crecieron en suspensión. El medio utilizado para el crecimiento adherente fue TC100 suplementado con antibióticos (75 U/ml de estreptomicina, 75 μg/ml de penicilina G, 10 μg/ml de tetraciclina y 7 μg/ml de gentamicina), 2 mM de L-glutamina, 45 μM de aminoácidos no esenciales (AANE) y 10% (v/v) de FBS. Para el crecimiento de las células en suspensión se empleo el medio ExpressFive® STM (Invitrogen) suplementado con antibióticos (75 U/ml de estreptomicina, 75 μg/ml de penicilina G, 10 μg/ml de tetraciclina y 7 μg/ml de gentamicina), 2 mM de L-glutamina y 45 μM de AANE. Las células High-Five™ se crecieron a 28ºC y agitación (180 rpm) en el caso de crecimiento en suspensión.

Todas las células adherentes se pasaron antes de llegar a confluencia (~90%) para su mantenimiento. Para ello, se eliminó el medio de cultivo, se lavaron las células con una solución de EDTA (0,02%, p/v), tras lo que se añadió 1 ml de una solución de Tripsina (0,05%, p/v) y EDTA (0,02%, p/v). Se incubaron las células a 37ºC hasta que se despegaron del sustrato y se detuvo la tripsinización añadiendo 10 ml de medio completo. Las células se centrifugaron 5 minutos a 500 xg y tras eliminar el sobrenadante, se resuspendieron en 10 ml de medio completo fresco. Finalmente, se sembraron las células en medio completo fresco a la densidad necesaria. Las células en suspensión se mantuvieron entre $2{,}5x10^5$-$5x10^5$ células/ml para lo que se diluyeron 1:3 o 1:4 en medio completo fresco cada 2-3 días.

La criopreservación de las líneas celulares se realizó en FBS suplementado con 10% de dimetilsulfóxido (DMSO) estéril en criotubos (Nalgene). Para asegurar una congelación lenta se utilizó un *Nalgene® Cryo 1ºC Freezing Container* en el que se guardaron los criotubos a -80ºC durante 2-3 días, tras lo que se guardaron definitivamente en nitrógeno líquido (-196ºC). La descongelación se realizó a 37ºC en agitación, tras lo que se lavaron las células con medio completo fresco para eliminar el DMSO antes de sembrarlas a la densidad requerida en medio completo fresco.

3.3.2 Transfección transitoria

La transfección de células COS-7 se realizó con el reactivo TransIT®-LT1 (Mirus) siguiendo las indicaciones del fabricante. El método del fosfato cálico (PCa) se empleó para la transfección de células HEK293T, para ello se prepararon las soluciones A (250 mM de $CaCl_2$) y B (50 mM HEPES, 1,5 mM Na_2HPO_4 y 140 mM NaCl, ajustado a pH

Formato placa	Volúmen de A/B	ml de medio	µg de plásmido
P100	500 µl	5	12,5
M6 (pocillo)	200 µl	2	2
M24 (pocillo)	50 µl	0,5	1

Tabla 3-1. Cantidades de reactivos utilizados para la transfección transitoria por el método PCa en función del formato de placa de cultivo utilizado.

7,05), se esterilizaron por filtración (0,22 μm) y se conservaron a 4ºC. Se sembraron las células el día anterior para alcanzar una confluencia del 60-70% en el momento de la transfección, excepto para la producción de proteínas recombinantes donde la confluencia fue del 30%. Se cambió el medio 4 horas antes de la transfección por medio fresco o medio con 1% FBS para la producción de proteínas. En el momento de la transfección se mezcló el plásmido de interés a transfectar con la solución A, tras lo que se añadió esta mezcla sobre la solución B. Se incubó un minuto en agitación y se añadió a las células gota a gota. Las cantidades de plásmido y los volúmenes de las soluciones A y B se detallan en la Tabla 3-1. Tras la transfección se incubaron las células 24-48 horas a 37ºC, excepto para la producción de proteínas recombinantes que se incubaron 72 horas.

Para la transfección de células Hive-Five™ con bácmidos recombinantes (ver apartado 3.13.1) se sembraron células el día anterior en placas de seis pocillos (M6) a una densidad de $4x10^5$ células/pocillo en 2 ml de medio completo. Se añadió 2 μg de bácmido en 100 μl de medio completo sin FBS que se mezcló con 6 μl de Cellfectin® (Invitrogen) preparados en 100 μl de medio completo sin FBS. Esta mezcla se incubó 45 minutos a temperatura ambiente (RT). Durante la incubación se lavaron las células con medio completo sin FBS dejándolas con 800 μl de medio completo sin FBS. Tras la incubación se añadió la mezcla a las células gota a gota y se incubaron 5 horas a 28ºC. Finalmente se retiró la mezcla de transfección y se les añadió 2 ml de medio completo. Tras 72 horas de incubación a 28ºC se recogió el sobrenadante de las células donde se encuentra el baculovirus recombinante generado.

3.4 Técnicas básicas de biología molecular

3.4.1 Extracción de DNA plasmídico

En función del volumen de cultivo bacteriano se utilizaron tres kits comerciales diferentes. Para volúmenes pequeños (2-10 ml) se utilizo el kit *Wizard® Plus SV Minipreps DNA Purification System* (Promega), para volúmenes entre 50 y 200 ml se utilizó el kit *Wizard® Plus SV Midipreps DNA Purification System* (Promega) y

para volúmenes mayores de 200 ml se empleó el kit *QIAGEN Plasmid Plus Maxi kit* (QIAGEN). En todos los casos se siguieron las indicaciones de los fabricantes.

3.4.2 Extracción de bácmidos

La extracción de los bácmidos recombinantes se realizó a partir de cultivos de bacterias DH10bac™ de 2 ml crecidas durante 16-20 horas a 37ºC y siguiendo el protocolo recomendado por los fabricantes del sistema *Bac-to-Bac System® Baculovirus Expression* (Invitrogen).

3.4.3 Reacción en cadena de la polimerasa (PCR)

Las amplificaciones convencionales se realizaron con la polimerasa *Gotaq® Green Master Mix* (Promega), siendo la mezcla de reacción la detallada en la Tabla 3-2A. Las amplificaciones destinadas al clonaje se realizaron con la polimerasa de alta fidelidad *Phusion®* (NEB) siendo la mezcla de reacción tipo la descrita en la Tabla 3-2B. La cantidad de molde en las amplificaciones dependió de su naturaleza: 25-50 ng de plásmido, 50-200 ng de DNA genómico o 1 *µ*l de cDNA preparado como se detalla en el apartado 3.5.3.

Todas las amplificaciones se realizaron en los termocicladores *MJ mini* (BIO-RAD) y *MyCycler* (BIO-RAD) utilizando el programa descrito en la Tabla 3-2C, salvo que se indique lo contrario.

3.4.4 Digestión con enzimas de restricción

Para la digestión con enzimas de restricción (Promega) de plásmidos o fragmentos de DNA se utilizó la siguiente mezcla de reacción: 1-2 *µ*g de DNA, 10 U de cada uno de los enzimas necesarios, 2 *µ*l del tampón adecuado suministrado por el fabricante y agua destillada estéril hasta un volumen final de 20 *µ*l. Se incubó la mezcla de digestión durante 1-2 horas a 37ºC tras lo que se procedió a su purificación o a su análisis electroforético en geles de agarosa.

3.4.5 Ligación

En todas las ligaciones se utilizó una relación molar vector:inserto de 1:3, salvo que se indique lo contrario. La mezcla de ligación utilizada constó de los siguientes componentes: 50-100 ng de plásmido digerido (o 50 ng de pGEM®-t Easy), X ng de inserto (donde X es tres veces la cantidad de moleculas de vector, lo que depende del

A) PCR estándar

	[Initial]	[Final]	Mezcla (µL)
Gotaq® Green Master Mix	2X	1X	10
Cebador Directo	10 µM	0,5 µM	1
Cebador Reverso	10 µM	0,5 µM	1
Molde	-	-	x
H_2O (NFW)	-	-	Hasta 20 µl

B) PCR para clonajes

	[Initial]	[Final]	Mezcla (µL)
Polimerasa Phusion®	2U/µl	1 U	0,5
Tampón GC	5X	1X	4
dNTPs	10 mM	0,4 mM	0,8
Cebador Directo	10 µM	0,5 µM	1
Cebador Reverso	10 µM	0,5 µM	1
Molde	-	-	x
H_2O (NFW)	-	-	Hasta 20 µl

C) Programa estándar

Tiempo	Temperatura (ºC)	
5′	95	
30″	95	35 ciclos
30″	60	
30′′	72	
10′	72	
∞	4	

Tabla 3-2. Mezclas de reacciones para las amplificaciones por PCR. A) mezcla de reacción de las amplificaciones por PCR estándar. B) mezcla de reacción de las amplificaciones por PCR destinadas al clonaje. C) Programa estándar utilizado para todas las amplificaciones. NFW: *Nuclease Free Water*, agua libre de nucleasas.

tamaño del inserto), 2 µl de tampón de reacción 10X suministrado por el fabricante de la ligasa (300 mM Tris-HCl pH 7,8, 100 mM $MgCl_2$, 100m M DTT y 10 mM ATP), 3 U de DNA ligasa del fago T4 (Promega) y agua destilada estéril hasta un volumen final de 20 µl. La reacción se incubó 1 hora a 37ºC.

3.4.6 Electroforesis en geles de agarosa

Para el desarrollo de las electroforesis de DNA se prepararon geles de agarosa del porcentaje adecuado (e indicado en cada caso) en tampón TAE (40 mM Tris, 20 mM Ácido acético glacial y 1 mM EDTA) suplementado con Bromuro de etidio (50 µg/ml). La electroforesis se desarrollaron en cubetas de BIO-RAD, usando como tampón de electroforesis TAE, generalmente durante 45 minutos a 100 voltios.

Para cargar las muestras en los geles de agarosa se utilizó tampón de carga 10X (30% v/v glicerol, 5 mM EDTA y 0,1% p/v de azul de bromofenol), excepto en las muestras de las amplificaciones desarrolladas con la polimesara *Gotaq® Green Master Mix* que se cargaron directamente en el gel. Además, se cargaron marcadores de tamaños conocidos en función de las necesidades (1Kb o 100 pb, NEB). El resultado de la electroforesis se reveló en un transiluminador de luz UV (Uvidoc).

3.4.7 Purificación de fragmentos y productos de PCR

Los productos de interés, procedentes de amplificaciones o de digestiones, se purificaron tras la escisión de la banda correspondiente en el gel un agarosa con el kit *Wizard® SV Gel and PCR Clean-Up System* (Promega). Este kit también se utilizó para la purificación directa de productos amplificados por PCR. Los fragmentos purificados se cuantificaron espectrofotométricamente en un *Nanodrop ND-1000* (Thermo) y se utilizaron para ser clonados o para su secuenciación.

3.4.8 Secuenciación

La secuenciación de plásmidos y productos de PCR purificados se llevó a cabo empleando el secuenciador *ABI Prism 377* (PE Biosystems) en el servicio de Genómica del IIB (Instituto de Investigaciones Biomédicas "Alberto Sols"). Para la secuenciación de plásmidos se utilizaron los cebadores universales T7 (5´-TAATACGACTCACTATAGGG-3´), Sp6 (5'-GATTTAGGTGACACTATAG-3') y BGH (5'- TAGAAGGCACAGTCGAGG-3'), disponibles en el Servicio, o cebadores específicos del producto a secuenciar, que en el caso de productos de PCR fueron los mismos que los empleados en la amplificación.

3.5 Obtención de muestras, extracción de RNA y DNA genómico y síntesis de cDNA

3.5.1 Obtención de muestras

Las muestras de diferentes tejidos de *Homo sapiens*, *Macaca mulatta*, *Rattus norvegicus*, *Mus musculus*, *Bos taurus* y *Sus scrofa* se obtuvieron de Biochain a través de su distribuidor europeo AMS (Tabla 3-3 y Tabla-S1).

Las muestras de sangre de diferentes primates (*Gorilla gorilla*, *Colobus guereza*, *Papio cynocephalus*, *Macaca fascicularis* y *Cebus apella*, ver Tabla S1*)* se obtuvieron gracias a la colaboración del Zoo Aquarium de Madrid a través de su director Jesús Fernández Morán y la veterinaria Eva Martínez Nevado. Las muestras de sangre de donantes sanos fueron suministradas por la Dra. Aurora Viejo del Centro de donantes del Hospital Universitario "La Paz" de Madrid (Tabla S1). Todas las muestras de sangre, tanto de primates como humanas, fueron suministradas en tubos cerrados con EDTA como anticoagulante y conservadas a 4ºC durante 24 horas hasta recibir los resultados negativos de los análisis de patógenos infecciosos. Las muestras de

Tejidos	*Homo sapiens*	*Macaca mulatta*	*Rattus norvegicus*	*Mus musculus*	*Bos taurus*	*Sus scrofa*
Sangre	A	-	-	-	-	-
Cerebro	A,F	A	A	A	A	A
Corazón	F	A	A	A	A	A
Riñón	A,F,T	A	A	A	A	A
Hígado	A,F,T	A	A	A	A	A
Pulmón	A,F,T	A	A	A	A	A
Páncreas	A	A	A	A	A	A
Bazo	A,F	-	-	-	-	-
Timo	A,F	-	-	-	-	-
Embrión	-	-	19 días	11, 15 y 17 días	-	-

Tabla 3-3. Resumen de las muestras de tejido disponibles en cada una de las especies indicadas (A, tejido adulto; F, tejido fetal; T, tejido tumoral y -, no disponible). Para una descripción detallada consultar Tabla S1.

primates no humanos no eran sometidas a estos análisis pero se trataron de la misma forma para su mejor comparación.

3.5.2 Extracción de RNA y DNA genómico

Para la obtención de RNA de líneas celulares cultivadas en el laboratorio se recolectaron $2x10^6$ células y se utilizó el kit *SV Total RNA Isolation System* (Promega) siguiendo las indicaciones del fabricante. En el caso de muestras de sangre, se utilizaron tubos *Leucosep™* (Greiner Bio One) para la separación de células mononucleares (PBMC, *Peripheral Blood Mononuclear Cell*). Para ello, los tubos *Leucosep™* se rellenaron con 3 ml de Ficoll-Paque (Sigma) y se centrifugaron durante 1 minuto a 1000 xg. Posteriormente, se añadieron entre 3 y 5 ml de sangre y se centrifugaron durante 15 minutos a 1500 xg, quitando el freno de la centrífuga. La fracción celular enriquecida en células mononucleares, se recogió con una pipeta y se pasó a tubos de 1.5 ml. Se centrifugaron las células durante 5 minutos a 500 xg, tras lo que se lavaron dos veces con PBS (*Phosphate Buffered Saline*). A partir del pellet de células mononucleares (PBMCs) se procedió a la extracción de RNA total utilizando el kit *SV Total RNA Isolation System* o DNA genómico con el kit *FlexiGene® DNA* (QIAGEN).

La concentración de RNA total obtenida de las muestras de sangre (PBMCs), medida espectrofotométricamente (*Nanodrop ND-1000*), suele ser insuficiente para la síntesis directa de cDNA, por lo que se procedió a su concentración mediante precipitación. Para ello, se añadió a las muestras acetato sódico ajustado a pH 5 (300 mM final o 1/10 del volumen total) y 150 μg/ml de glicógeno para mejorar la eficiencia de la precipitación. Posteriormente se añadieron dos volúmenes de etanol absoluto frío y se incubó 30 minutos a -70ºC o toda la noche a -20ºC. Tras la incubación se centrifugó a 16000 xg durante 15 minutos a 4ºC, se eliminó el sobrenadante y se lavo el pellet

con etanol al 70%. Finalmente se resuspendió el pellet seco en el volumen de agua libre de nucleasas (NFW) deseado para conseguir la concentración adecuada.

Todos las muestras de RNA, tanto comerciales como obtenidos en el laboratorio fueron testadas para determinar su integridad con un Bioanalizador 2100 (Agilent) en el servicio de Genómica del CBMSO.

3.5.3 Síntesis de cDNA

La síntesis de cDNA se realizó a partir de 1 μg de RNA total utilizando el kit *ImProm-II™ ReverseTranscription System* (Promega) siguiendo las instrucciones del fabricante. En la Tabla 3-4 se resume los componentes de la reacción. La mezcla 1, que contiene el RNA total y el cebador Oligo $(dt)_{15}$, se desnaturalizó 5 minutos a 70ºC, tras lo que se pasó rápidamente a 4ºC. Entonces se añadió la mezcla 2 sobre la mezcla 1 y se incubó 5 minutos a 25ºC para permitir el anillamiento del RNA con el cebador. Luego se incubó 1,5 horas a 42ºC para que se llevara a cabo la síntesis, tras lo que se desnaturalizó la muestra a 70ºC durante 5 minutos para inactivar el enzima. En todas las reacciones de síntesis de cDNA se añadió un control negativo (RT-) con los mismos componentes que la reacción normal (RT+) salvo que no se añadió retrotranscriptasa, lo que nos permitió verificar la ausencia de DNA genómico en las muestras de RNA.

La verificación de la síntesis de cDNA, y la posible contaminación de DNA genómico, se comprobó mediante PCR, tanto en RT+ como en RT-, amplificando el cDNA del mensajero de actina B con los cebadores ACTB_F (5'-TGGTGGTGAAGCTGTAGCC-3') y ACTB_R (5'-CTTCGCGGGCGACGATGC-3') comunes para todas las especies. El tamaño del fragmento amplificado del mensajero de actina B es de 540 pb

Mezcla 1	[Inicial]	[Final]	Volumen (µl)
Oligo $(dt)_{15}$	0,5 µg/ul	0,5 µg	1
RNA total	x	1 µg	x
H_2O (NFW)	-	-	Hasta 5 µl
Mezcla 2	**[Inicial]**	**[Final]**	**Volumen (µl)**
Tampón de reacción	5X	1X	4
dNTPs	10 mM	0,5 mM	1
$MgCl_2$	25 mM	3 mM	2,4
Retrotranscriptasa	1 U/µl	1 U	1
H_2O (NFW)	-	-	Hasta 15 µl

Tabla 3-4. Mezclas de reacción necesarias para la síntesis de cDNA. NFW: *Nuclease Free Water*, agua libre de nucleasas.

aproximadamente (dependiendo de la especie), mientras que la amplificación de su DNA genómico resultaría en una banda de 1200 pb aproximadamente, por la presencia de un intrón entre las posiciones de los cebadores utilizados. La amplificación se realizó siguiendo el programa estándar de PCR (ver apartado 3.4.3). El resultado de la amplificación se analizó cargando 5 μl de cada producto amplificado en geles de agarosa del 1% (ver apartado 3.4.6).

3.6 Análisis del splicing alternativo mediante PCR anidada

La PCR anidada consiste en utilizar el producto de una amplificación previa como molde para realizar una segunda amplificación con otros cebadores que se ubican dentro de la primera secuencia amplificada (Figura 3-1.1). Esta metodología se usó para analizar el splicing alternativo del gen *NCR3* en las distintas especies para lo que se diseñaron parejas de cebadores directos en el exón 1 en el entorno del ATG iniciador y parejas de cebadores reversos en el exón 4 (o 4I en primates), próximos al codón de terminación de la traducción descrito para cada especie. En primates se diseñaron parejas de cebadores reversos adicionales, comunes para los exones 4II y 4III, próximos al codón de terminación presente en el exón 4III, que se combinaron con los cebadores directos diseñados en el exón 1 (Tabla 3-5). La primera amplificación se realizó utilizando 1 μl de cDNA de cada muestra (ver apartado 3.4.3) y la combinación de los cebadores directo y reverso más externos (Figura 3-1.1). Para la segunda amplificación se utilizó 1 μl de la primera reacción de PCR como molde, y la pareja de cebadores más internos (Figura 3-1.1). Ambas amplificaciones se realizaron con la mezcla de reacción y el programa estándar descritos en el apartado 3.4.3.

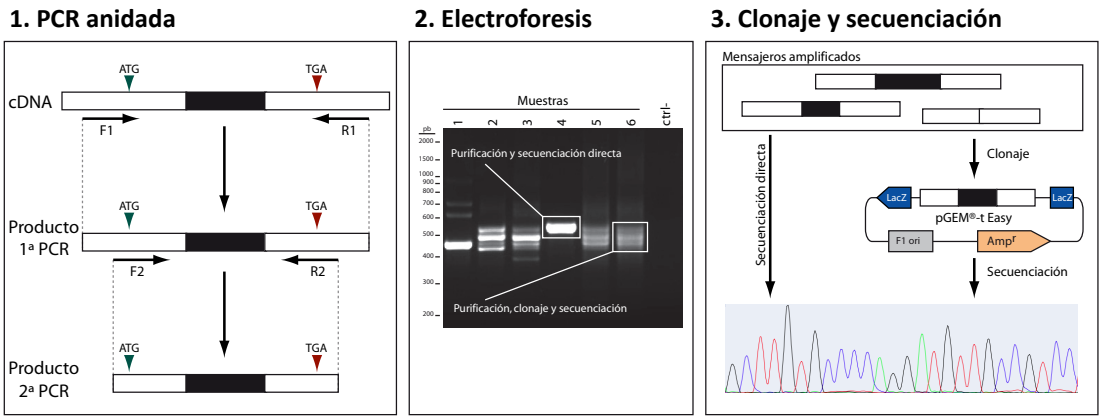


Figura 3-1. Representación esquemática de las etapas realizadas en el proceso de identificación de las distintas variantes de splicing mediante PCR anidada.

El resultado de la amplificación en los diferentes tejidos y líneas celulares se resolvió mediante electroforesis en geles de agarosa del 2% (ver apartado 3.4.6) (Figura 3-1.2). En función del resultado de la amplificación y de la separación electroforética se siguieron dos estrategias para identificar las variantes de splicing. La primera consistió en la purificación de cada variante amplificada directamente tras la escisión de la banda correspondiente en el gel de agarosa, que se envió a secuenciar bidireccionalmente utilizando los mismos cebadores empleados en la segunda amplificación (Figura 3-1.2 y 3-1.3). En los casos en los que el aislamiento de los diferentes productos amplificados no era posible, por dar lugar a bandas cercanas, éstos se clonaron en el vector pGEM®-t Easy (Figura 3-1.2 y 3-1.3). Para ello se purificó el producto de PCR directamente (o un grupo de bandas aisladas del gel) y se utilizó 10 μl de esta mezcla como inserto para una reacción de ligación con pGEM®-t Easy (ver apartado 3.4.5). Con el producto ligado se transformaron bacterias competentes, que tras crecer fueron seleccionadas por PCR utilizando los

Especie	Nombre oligonucleótido	Exón/Intrón	Sentido	Secuencia (5'-3')
Homo sapiens y Macaca mulatta	1C7289F	E1	D	CCACCTGGGACATCTTCCG
	RT26891HF	E1	D	ATGGCCTGGATGCTGTTGC
	1C7C3688R	E4I	R	GAGGACTAGGGACATCTGGG
	Hs1C7E4IR1	E4I	R	ATTGGGGTCTTTTGAAGAGGAC
	1C7AB3906R	E4II/E4III	R	AGCAGATGTGCTGAGCTCC
	Hs1C7E4IIR2	E4II/E4III	R	CAGGAAAGGGCAGTTGCC
Mus musculus	RT68682MF	E1	D	ATGGCCAAGGTGCTCCTGG
	RT68702McarF	E1	D	ATCTTCATCATGGTTTATCC
	RT72941McarR	E4	R	AGAATCACTTCTCAGAGGC
	RT72989McarR	E4	R	GTAAGGACTTATTGTTGGC
Rattus norvegicus	E1_OutRF	E1	D	CGCTTCGCCAATGGGAGG
	E1_InRF	E1	D	ATGGCCAAGGTGCTGCTGG
	E4_InRR	E4	R	GGGATTTAGAATCGCTCCTC
	E4_OutRR	E4	R	GGCCTCTTTGAGTGCTGGAT
Bos taurus y Sus scrofa	E1_InBSF	E1	D	CATGGCCCAGATGCTGTTG
	E1_OutBSF	E1	D	CCTCATCAACAGGAACACCC
	E4I_InBSR	E4	R	CACTCATCAGAGGCCATCC
	E4I_OutBSR	E4	R	TCACTGGGGTCTGGAATCAC
Multi-especie (secuenciación DNA genómico)	PrimE1F1	E1	D	TTCCTCCTCCACCCAGACC
	PrimI1R1	I1	R	GCATCTAGTCCAGCCTCCTG
	PrimE3F1	E3	D	CTGTCAGCTTTCTCTCTGTGG
	PrimE4II-IIIR1	E4II/E4III	R	AGTTGCCAAGGAGGAGTCAT
Homo sapiens, Gorilla gorilla, Pongo pigmaeus, Colobus guereza, Papio cynocephalus, Macaca fascicularis y Cebus apella	PanPrimE1F1	E1	D	TCACTGCTCAGATCCCCTTC
	PanPrimE1F2	E1	D	AACTGGGACATCTTCCGACA
	PanPrimE4IR1	E4I	R	AGGAGTGGCAGTGTGTTCC
	PanPrimE4IR2	E4I	R	GGCTCTGGAATCACTCCTC
	PanPrimE4II-IIIR1	E4II/E4III	R	CTTGGACCTTTCCAGGTCAG
	PanPrimE4II-IIIR2	E4II/E4III	R	GCTGAGCTCCCACATGGT

Tabla 3-5. Relación de los oligonucleótidos utilizados como cebadores en el análisis de las variantes de splicing del gen *NCR3* en cada una de las especies mediante PCR anidada. También se incluyen los cebadores multi-especie utilizados para la amplificación por PCR del DNA genómico del gen *NCR3* de algunos primates.

mismos cebadores internos que en la amplificación original. El análisis electroforético de la amplificación procedente de cada colonia permitió la selección de las distintas variantes de splicing y tras la purificación de los plásmidos correspondientes éstos fueron secuenciados bidireccionalmente utilizando los oligonucleótidos universales T7 y Sp6. El análisis descrito se realizó por triplicado en cada una de las muestras para la identificación de todas las variantes de splicing.

3.7 Análisis del splicing alternativo mediante RNA seq

3.7.1 Obtención de secuencias de RNA-seq

Para analizar el splicing alternativo mediante RNAseq se utilizaron secuencias disponibles públicamente en las bases de datos procedentes de experimentos de RNA-seq de muestras de las distintas especies incluidas en este trabajo (Tabla S2). La búsqueda de éstas fue bibliográfica a través de *PubMed* (http://www.ncbi.nlm.nih.gov/pubmed), mediante el rastreo directo de la base de datos *GEO* (di de que viene GEO) (http://www.ncbi.nlm.nih.gov/geo/) o procedentes del proyecto *Encyclopedia of DNA Elements* (*ENCODE*) (http://genome.ucsc.edu/ENCODE/). Los datos se descargaron del repositorio *Sequence Read Archive* (*SRA*) (http://www.ncbi.nlm.nih.gov/sra), en su formato propio (formato "sra"). Para la conversión de este formato al convencional "fastq" se utilizo el comando "fastq-dump" del paquete *SRA-toolkit 2.0*, disponible en el repositorio *SRA*.

Además de los datos disponibles en las bases de datos, se utilizaron las secuencias obtenidas en el laboratorio a partir de la secuenciación de dos muestras procedentes de una mezcla de RNAs obtenidos de sangre (PBMCs) de dos donantes sanos (Tabla S1). La extracción de RNA total de ambas muestras de sangre siguió el mismo protocolo que el descrito en el apartado 3.5.2. Se preparó una mezcla equilibrada de ambas muestras que se empleó para la síntesis de cDNA de doble cadena (ds cDNA) utilizando el kit *MINT* (Evrogen). Tras la síntesis de ds cDNA se dividió la muestras en dos partes iguales. Ambas muestras fueron pre-amplificadas utilizando el kit *TRIMMER* (Evrogen) siguiendo las instrucciones del fabricante. Una de las dos muestras fue normalizada con el kit *TRIMMER* (Evrogen), para reducir la contribución de los genes de alta expresión e incrementar la detección de genes de baja expresión, y por lo tanto, de variantes de splicing de baja expresión. La secuenciación se realizó en un *Genome Analyzer II* (Illumina) en "*The Wellcome Trust Sanger Institute*", gracias

a una colaboración con el Dr Harold Swerdlow, donde se realizó la fragmentación y la preparación de las librerías de cada muestra siguiendo las indicaciones de Illumina. Cada muestra fue secuenciada (108 ciclos por cada extremo, 2x108) en una línea independiente y el filtrado de calidad se realizó de acuerdo a los estándares de Illumina. Se obtuvieron 23.123.872 lecturas (2x108) en la muestra sin normalizar y 20.277.787 de lecturas (2x108) de la muestra normalizada. La calidad de las lecturas se analizó con el programa FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

3.7.2 Análisis de las secuencias de RNAseq

Para el análisis de las secuencias de RNA-seq se procedió al alineamiento de éstas contra el genoma de referencia de cada especie obtenido de Ensembl o de NCBI (Tabla 3-6). Para el alineamiento se utilizaron los programas Bowtie (versión 0.12.7) (Langmead et al., 2009) y Tophat (versión 2.0.2) (Trapnell et al., 2009) y los parámetros utilizados fueron los determinados por defecto, suministrando a Tophat la anotación del genoma disponible (Tabla 3-6) y permitiendo la detección de nuevas uniones de exones no descritas en dicha anotación. La conversión de formatos de lecturas alineadas se realizo con el programa Samtools (versión 0.0.18) (Li et al., 2009) y la visualización de los alineamientos se realizó con el navegador IGV (Robinson et al., 2011).

Las uniones de exones procedentes del gen *NCR3* en cada especie se obtuvieron filtrando el fichero de resultados "*Junctions.bed*" (Tophat), mediante el uso de scripts de python desarrollados en esta Tesis. Sólo se consideraron válidas las uniones de exones presentes en más de una muestra o que estuvieran soportadas por más de una lectura (además de tener más de ocho nucleótidos en cada extremo de la unión).

Especie	Versión	Base de datos	Dirección
Homo sapiens	GRCh37	Ensembl	ftp://ftp.ensembl.org/pub/release-71/fasta/homo_sapiens/
Pan troglodytes	CHIMP2.1.4	Ensembl	ftp://ftp.ensembl.org/pub/release-70/fasta/pan_troglodytes/
Gorilla gorilla	Gor3.1	Ensembl	ftp://ftp.ensembl.org/pub/release-70/fasta/gorilla_gorilla/
Pongo abelli	PPYG2	Ensembl	ftp://ftp.ensembl.org/pub/release-71/fasta/pongo_abelii/
Macaca mullata	MMUL 1	Ensembl	ftp://ftp.ensembl.org/pub/release-70/fasta/macaca_mulatta/
Macaca fascicularis	MacFas_Jun2011	NCBI	http://www.ncbi.nlm.nih.gov/bioproject/PRJEA48347/
Papio anubis	Panu_2.0	NCBI	http://www.ncbi.nlm.nih.gov/bioproject/PRJNA169345/
Rattus norvegicus	Rnor 5.0	Ensembl	ftp://ftp.ensembl.org/pub/release-70/gtf/rattus_norvegicus/
Mus musculus	GRCm38	Ensembl	ftp://ftp.ensembl.org/pub/release-70/gtf/mus_musculus/
Bos taurus	UMD3.1	Ensembl	ftp://ftp.ensembl.org/pub/release-70/gtf/bos_taurus/
Sus scrofa	Sscrofa10.2	Ensembl	ftp://ftp.ensembl.org/pub/release-71/gtf/sus_scrofa/

Tabla 3-6. Referencia de las versiones de los genomas utilizadas de las distintas especies.

El número de lecturas en las uniones de exones se normalizó por el número total de lecturas alineadas en cada muestra (en millones). Las uniones de exones detectadas en *Homo sapiens* (la especie de referencia) se nombraron como H1-H20. Las uniones de exones detectadas en otras especies que se localicen en las mismas posiciones relativas que en humano recibieron el mismo nombre. Las nuevas uniones de exones detectadas en cada especies sin homólogos en humano se nombraron con la primera letra del nombre de la especie y un número, excepto en el grupo de primates (*Gorilla gorilla*, *Colobus guereza*, *Papio cynocephalus*, *Macaca mulatta, Macaca fascicularis* y *Cebus apella*) en el que se utilizó la letra "P" (de Primates) seguido de un número para nombrar a las nuevas uniones de exones.

Para el cálculo de la expresión del gen *NCR3* en *Reads Per Kilobase per Million mapped reads* (RPKMs) se contabilizó el número de lecturas alineadas en los exones del gen *NCR3* y este valor fue normalizado por la suma de la longitud de los exones (en Kb) y por el número total de lecturas alineadas (en millones). Para ello se utilizó Samtools y scripts de python desarrollados en esta Tesis.

3.8 Anotación y estudio de la conservación del gen *NCR3* y su entorno

El análisis de la conservación del gen *NCR3* se realizó con el programa mVista (http://genome.lbl.gov/vista/mvista/submit.shtml), suministrándole la secuencia genómica del gen de cada una de las especies y la anotación de éste corregida según lo descrito en el apartado 4.1 de esta Tesis.

Para el análisis de la sintenia de la región de MHC de clase III se obtuvieron las secuencias de las variantes canónicas de todos los genes de la región humana (Xie et al., 2003) usando BioMart (Smedley et al., 2009). Estas secuencias se utilizaron para la búsqueda de homólogos en las demás especies mediante el uso de tBlastn (Camacho et al., 2009). Los resultados del alineamiento se filtraron con scripts de python desarrollados en esta Tesis.

El estudio de la densidad génica se realizó también con scripts de python desarrollados en esta Tesis. El genoma de cada especie se recorrió en ventanas de 1 Mb, en las que se calculó el porcentaje del contenido en G y C, se obtuvo el número de genes de la anotación disponible y se extrajo el número de uniones de exones diferentes registradas en dicha anotación (Tabla 3-6). De los resultados del alineamiento de las secuencias de RNA-seq de cada especie se extrajo la información correspondiente a

uniones de exones detectadas (fichero "*junctions.bed*" de Tophat). Se agruparon las uniones de exones detectadas en todas las muestras analizadas y se consideraron válidas aquellas soportadas por al menos cinco lecturas. La representaciones gráficas de estos datos se realizó con el uso del módulo matplotlib de python. Las uniones de exones detectadas se compararon con las descritas mediante el uso de scripts de python desarrollados en esta Tesis. Para la generación de los diagramas de Venn se utilizó el módulo Venn/Euler de R.

3.9 PCR cuantitativa (qPCR)

La PCR cuantitativa (qPCR) permite conocer de manera precisa los niveles de expresión de uno o varios mensajeros y su comparación entre muestras. Existen dos aproximaciones principales en la qPCR: relativa, que permite comparar los niveles de expresión de un mensajero en diferentes muestras, y absoluta, indicada para comparar los niveles de expresión de varios mensajeros en una misma muestra.

En la primera aproximación se necesita la utilización de uno o varios mensajeros normalizadores (*GAPDH*, *POLR2A*, *Actina*, etc.) para corregir las potenciales diferencias existentes en las cantidades de RNA de cada una de las muestras que se utilizó para la síntesis de cDNA, las diferencias en la degradación de cada una de las muestras e incluso las diferencias de carga de las distintas muestras en el propio proceso de amplificación. En estas circunstancias, dado que sólo se evalúa la expresión un mensajero problema y los mensajeros de los normalizadores, cabría esperar que la eficiencia de amplificación en las distintas muestras fuese constante, de modo que podemos utilizar el método $2^{\Delta\Delta Ct}$ o de Livak (Livak and Schmittgen, 2001) para el análisis de los resultados. Sin embargo, la eficiencia de las distintas parejas de cebadores se puede medir a través del análisis de curvas de dilución de cDNA o del mismo producto de PCR amplificado (amplicón) purificado o clonado en un vector. Si calculamos de este modo la eficiencia podremos aplicar el método de Pffafl (Pfaffl, 2001) para el análisis que es más preciso que el primero.

En la aproximación absoluta de la qPCR no es necesario el uso de normalizadores, puesto que se trata de cuantificar varios mensajeros de una única muestra. Sin embargo, es necesario el uso de curvas de dilución de estándares de concentración conocida (generalmente amplicones clonados en un vector) para poder interpolar los resultados en la curva y conocer la concentración de los distintos mensajeros

en la muestra, que ya estarán corregidos por la eficiencia de la pareja de cebadores definida por la curva de dilución de estándares.

Sin embargo, en este trabajo se han comparado los niveles de expresión de diferentes mensajeros entre ellos en la misma muestra y entre muestras, de manera que se ha seguido un procedimiento híbrido entre estos dos. Se ha utilizado una cuantificación absoluta que permite la comparación intra-muestra, pero además los resultados fueron normalizados para habilitar la comparación entre muestras.

3.9.1 Diseño de los cebadores

Dado el alto porcentaje de secuencia compartida por más de una variante de splicing del gen *NCR3* se diseñaron manualmente cebadores específicos de eventos de splicing concretos. Éstos son capaces de hibridar con más de una variante, sin embargo la combinación de dos de estos cebadores, diseñados en dos eventos de splicing, permite la amplificación específica de una variante concreta (Figura 3-2 y Tabla 3-7). Además se diseñaron cebadores comunes (HsNCR3F y HsNCR3R) que permitieran la cuantificación global de todas las variantes de splicing (Figura 3-2 y Tabla 3-7).

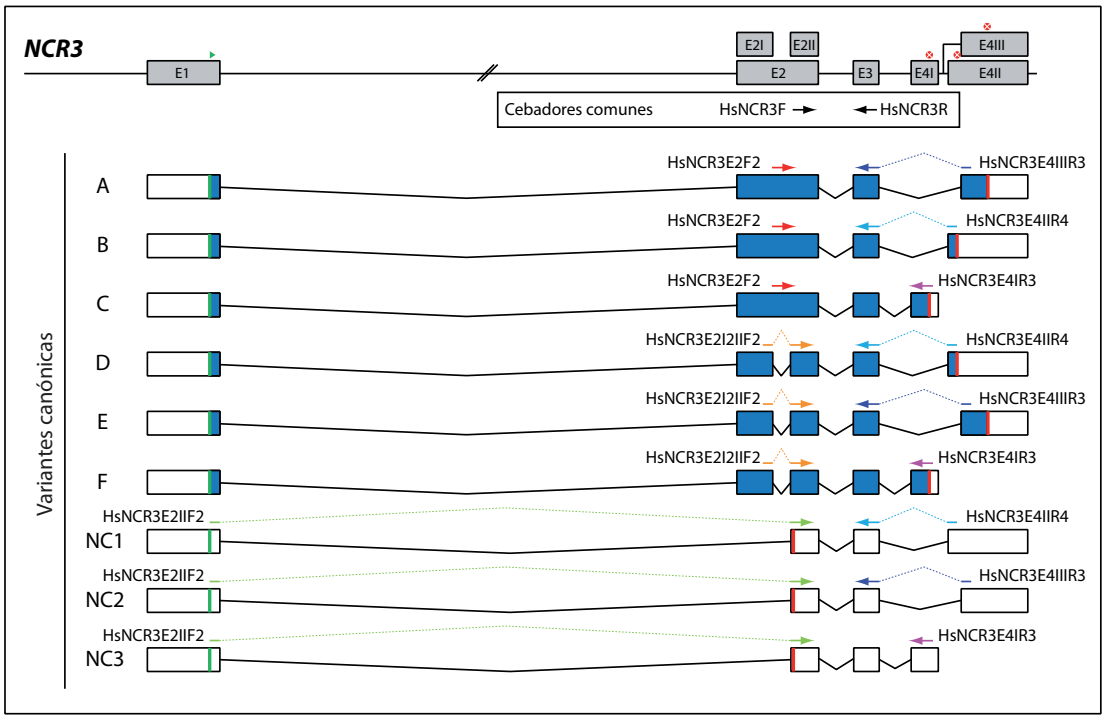


Figura 3-2. Representación gráfica del diseño de cebadores utilizado para la cuantificación de las nueve variantes canónicas del gen *NCR3*.

Para el diseño de los cebadores destinados a la amplificación de los mensajeros normalizadores se utilizó el programa Primer3plus, de manera que el amplicón estuviese interrumpido por al menos un intrón en la secuencia genómica, evitándose así la amplificación inespecífica de DNA genómico contaminante. Los normalizadores utilizados fueron el mensajero del enzima glucolítico Gliceraldehído-3-fosfato deshidrogenasa (*GADPH*, NM_002046.4) y el mensajero de la subunidad grande de la RNA polimerasa II (*POLR2A*, NM_000937.4).

Todos los cebadores se analizaron informáticamente para asegurar su especificidad y se testaron experimentalmente mediante PCR estándar (ver aparatado 3.4.3) antes de proceder a la cuantificación definitiva. La especificidad de los cebadores de cada variante del gen *NCR3* se corroboró además mediante el análisis de la amplificación cruzada. Para ello se enfrentó cada pareja de cebadores con todas las variantes canónicas clonadas en el vector pGEM®-t Easy, obtenidas durante el proceso de análisis del splicing del gen *NCR3* en humano (ver apartado 3.6), como molde en reacciones en qPCR (ver más abajo).

3.9.2 Preparación de las curvas de dilución de estándares

Para la preparación de las curvas de estándares se utilizaron los vectores derivados de pGEM®-t Easy en los que habían sido clonadas las nueve variantes canónicas del gen *NCR3*. También se clonaron en pGEM®-t Easy los productos de la amplificación de los mensajeros de los genes normalizadores *POLR2A* y *GAPDH*, usando los mismos cebadores que se usaron para la qPCR (Tabla 3-7) y cDNA procedente de células HeLa como molde siguiendo el protocolo estándar descrito en lo apartado 3.4. Todos estos vectores fueron linearizados mediante digestión con el enzima de restricción *SpeI*, tras lo que fueron purificados y cuantificadas espectrofotométricamente (*Nanodrop ND-1000*) por triplicado. Entonces se calculó el número de moléculas por *µ*l, considerando que el peso molecular medio de 1 pb es 0,66 ng/pmol y la siguiente fórmula:

$$\frac{N^{o}Moléculas}{\mu l} = \frac{[ng/\mu l]}{PM\ (ng/pmol)} x \frac{10^{12} moles}{pmoles} \quad xNA$$

Donde *PM* es el peso molecular del vector calculado como 0,66 ng/pmol * N, siendo N la longitud del vector en pb. *NA* es el Número de Avogadro (6,023x10^{23} moléculas/mol). Se prepararon diluciones seriadas desde 10 a 1x10^{5} moléculas/*µ*l de cada

Nombre oligonucleótido	Exón/Intrón	Sentido	Secuencia (5'-3')
HsNCR3E2F2	E2	D	CACTTGCTTCTTCCCGTTTC
HsNCR3E2I2IIF2	E2I-E2II	D	AGGGAAGGAGGCTGAGCT
HsNCR3E2IIF2	E1-E2II	D	ATCATGGTCCATCCAGGCTG
HsNCR3E4IR3	E4I	R	GCAGTGTGTTCCCATGTGAC
HsNCR3E4IIR4	E3-E4II	R	GGGGAATCCGGAGAGAGTAG
HsNCR3E4IIIR3	E3-E4III	R	CTTTCCAGGTCAGACATTTGC
HsNCR3F	E2II	D	TACGTGTGCAGAGTGGAGGT
HsNCR3R	E3	R	GGCCACAGAGAGAAAGCTGA
HsPOLR2AF	E10	D	GCAAATTCACCAAGAGAGACG
HsPOLR2AR	E11	R	CACGTCGACAGGAACATCAG
HsGAPDHE2-3F1	E2-E3	D	GAGTCAACGGATTTGGTCGT
HsGAPDHE5R1	E5	R	TTGATTTTGGAGGGATCTCG

Tabla 3-7. Oligonucleótidos utilizados como cebadores en la cuantificación por qPCR de las variantes de splicing del gen *NCR3*. También se indican los cebadores utilizados para la amplificación de los mensajeros normalizadores.

plásmido linearizado que se utilizaron como curva de estándares.

3.9.3 Amplificación

La amplificación se realizó en el termociclador *ABI Prism 7900HT* (Applied Biosystems) utilizando la *Taq* polimerasa del kit *Power Sybr® Green Master Mix* (Applied Biosystems). Se utilizó la mezcla de reacción que se detalla en la Tabla 3-8 y el programa descrito en la Figura 3-3. Como molde se utilizó 1 μl de cDNA procedente de cada una de las muestras humanas preparado como se describe en el apartado 3.5.3. En el caso de las curvas de estándares como molde de las reacciones se utilizó 1 μl de los plásmido linearizado descritos en el apartado anterior a la concentración requerida. La detección de fluorescencia se realizó a 75ºC para evitar la cuantificación de los dímeros de oligonucleótidos, siendo la temperatura de fusión o melting de todos los amplicones superior a dicha temperatura (determinada experimentalmente). Todas las muestras se analizaron por triplicado siendo necesario el desarrollo de varios experimentos de qPCR en placas de 384 pocillos. Para reducir las posibles diferencias entre experimentos se utilizaron calibradores inter-experimento. Estos calibradores fueron los plásmidos utilizados para las curvas de estándares a una concentración de $1x10^5$ moléculas/μl, que permitieron ajustar

	[Inicial]	[Final]	Volumen (µl)
Power Sybr® Green Master Mix	2X	1X	5
Mezcla de cebadores	2,5 µM	0,25 µM	1
Molde	-	-	1
H_2O (NFW)	-	-	Hasta 10 µl

Tabla 3-8. Mezcla de reacción utilizada para las reacciones de qPCR. NFW: *Nuclease Free Water*, agua libre de nucleasas.

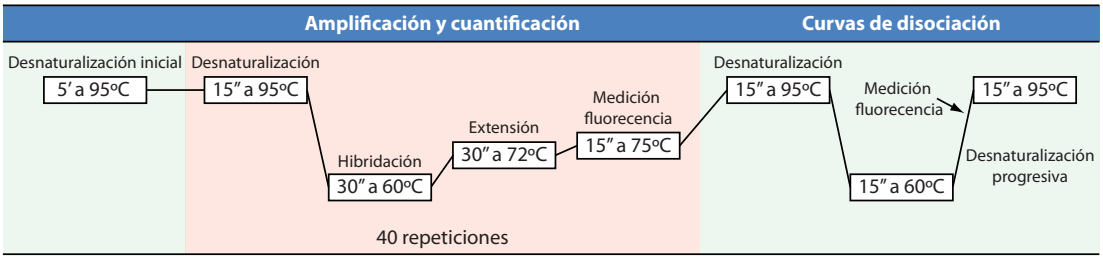


Figura 3-3. Programa utilizado en la cuantificación por qPCR de las variantes de splicing del gen *NCR3*.

las curvas de estándares para adaptarlas a las circunstancias concretas de cada experimento. Tras el proceso de amplificación se procedió al análisis de las curvas de disociación para comprobar la presencia de un único producto amplificado en cada pocillo (Figura 3-3). Para ello, al final del proceso de amplificación, se añadió una etapa de desnaturalización controlada, aumentando la temperatura medio grado cada 15 segundos y midiendo la fluorescencia en cada temperatura. La pérdida de fluorescencia debida a la desnaturalización, puesto que el *Sybr® Green* sólo se intercala en los productos de cadena doble, permite determinar de manera precisa la temperatura de fusión o *melting* de cada producto amplificado.

3.9.4 Análisis de los resultados

Tras el desarrollo de los distintos experimentos se procedió al análisis de los resultados. Se utilizó el programa SDS (Applied Biosystems) para obtener los valores de la cuantificación, permitiendo al programa establecer la línea de base (ruido de fondo) de manera automática. En qPCR se utiliza el Ciclo Umbral o *Ct* (*threshold cycle*) como parámetro fundamental, que es el ciclo, durante la fase lineal de la amplificación, en el que la señal de fluorescencia es superior a la fluorescencia basal o ruido de fondo. Por lo tanto, el *Ct* es inversamente proporcional a la cantidad inicial de molde. Con los valores de la cuantificación de las curvas de estándares se calculó la eficiencia de cada pareja de cebadores. Para ello se representó el *Ct* obtenido frente al logaritmo de la concentración (Figura 3-4) y se utilizaron los parámetros de la recta definida para calcular la eficiencia con la siguiente fórmula:

$$E = 10^{\frac{-1}{m}} - 1 \quad \text{(Pfaffl, 2001)}$$

Donde *m* es la pendiente de la recta (Figura 3-4).

Para la cuantificación de las distintas variantes de splicing del gen *NCR3* se interpolaron los *Ct* obtenidos en las rectas de referencia, resultado de la cuantificación de los

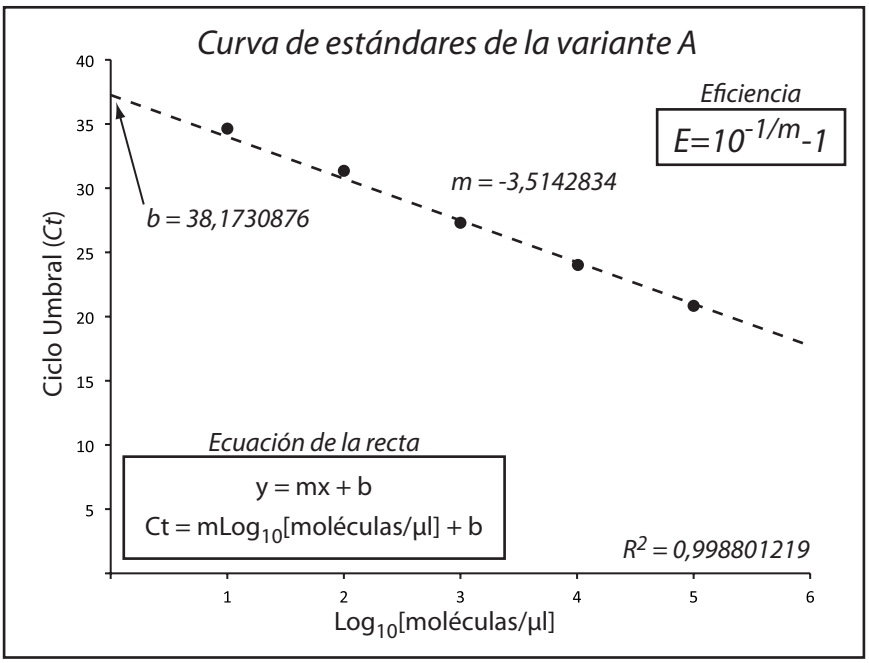


Figura 3-4. Ejemplo del análisis de las curvas de estándares realizado. m, pendiente de la recta; b, intersección de la recta con el eje de abscisas, R^2, coeficiente de correlación.

estándares de cada variante, para obtener la concentración (10^{Ct} = moléculas/μl). Los valores de expresión fueron normalizados por la media geométrica de los valores obtenidos de la amplificación de los mensajeros normalizadores en cada muestra. Los valores obtenidos de la amplificación utilizando los cebadores comunes para todas las variantes de *NCR3* en bazo adulto tras la normalización se consideró como el 100% de expresión, escalando el resto de los valores con respecto a este para su comparación.

En cada experimento independiente se ajustaron las rectas de referencia (curvas de estándares) mediante el uso de los calibradores inter-experimento. Dado que los calibradores son muestras de plásmido linearizado de concentración conocida, los resultados de su cuantificación (*Ct*) permiten recalcular la intersección de la recta de referencia con el eje de abcisas (*b*) (Figura 3-4), permitiendo ajustar la recta a las circunstancias concretas de cada experimento, asumiendo que la eficiencia de la amplificación de cada pareja de cebadores es constante y por lo tanto, la pendiente de la curva es siempre la misma.

3.10 Plásmidos y construcciones

Las amplificaciones desarrolladas para la construcción de los distintos vectores y plásmidos (vectores con inserto) se realizaron con la polimerasa de alta fidelidad *Phusion®* (NEB) como se describe en el apartado 3.4.3. Todas las construcciones fueron confirmadas mediante secuenciación bidireccional (ver apartado 3.4.8).

3.10.1 Vectores construidos

Además de los vectores comerciales pcDNA3.1™/Zeo(-) (Invitrogen) y pcDNA3.1™/V5-His TOPO® (Invitrogen) se construyeron los siguientes vectores para la realización de esta Tesis como esqueleto para la obtención de los plásmidos destinados a la expresión de proteínas recombinantes en células de mamíferos e insecto:

pZC: derivado del vector pcDNA3.1™/Zeo(-) (Invitrogen). Se clonó la secuencia codificante del péptido señal de la proteína CD33 humana (número de acceso P20138). Para ello se usaron los oligonucleótidos CD33F y CD33R (Tabla 3-9) que contenían toda la secuencia del péptido señal además de los sitios de reconocimiento por los enzimas *NheI* y *XbaI* que se desnaturalizaron a 95ºC y luego se redujo la temperatura medio grado cada minuto hasta los 25ºC para permitir la hibridación. Tras la hibridación se digirió con los enzimas *NheI* y *XbaI* (Figura 3-5A) y se clonó en pcDNA3.1™/Zeo(-) digerido con los mismos enzimas.

pZCV: para la construcción de este vector se utilizo pZC como esqueleto. Se usó la misma estrategia que en el caso anterior para clonar la secuencia codificante del epítopo V5 a través de la hibridación de los oligonucleótidos V5F y V5R (Tabla 3-9), que tras su digestión con los enzimas *XbaI* y *XhoI* se clonaron en el vector pZC digerido del mismo modo (Figura 3-5B).

pZC-Fc: para la construcción de este vector se amplificó la secuencia codificante de la porción Fc de la inmunoglobulina G1 humana (hIgG1-Fc) usando como molde el plásmido pIgPlus (R&D Systems) y los cebadores FC_EcoRI_XhoI_F y FC_HindIII_R (Tabla 3-9). El producto amplificado se digirió con los enzimas *EcoRI* y *XhoI* y se clonó en el vector pZC (Figura 3-5C).

pZ-EGFP: se obtuvo la secuencia codificante de la proteína EGFP mediante amplificación con los cebadores EcoRV_EGFP_F y EGFP_HindIII_R (Tabla 3-9) usando el vector pEGFP-N1 (NEB) como molde. El producto amplificado, sin el ATG

Plásmido	Vector	Inserto	Nombre oligonucleótido	Sentido	Secuencia (5'-3')
pZC	pcDNA3.1™/Zeo(-) (Invitrogen)	Secuencia del péptido señal de CD33	CD33F	D	CCC**GCTAGC**ATGCCGCTGCTGCTACTGCTGCCCCTGC
					TGTGGGCAGGGGCCCTGGCTATG**TCTAGA**CCC
			CD33R	R	GGG**TCTAGA**CATAGCCAGGGCCCCTGCCCACAGCAG
					GGGCAGCAGTAGCAGCAGCGGCAT**GCTAGC**GGG
pZCV	pZC	Secuencia del epítopo V5	V5F	D	CCC**TCTAGA**GGTAAGCCTATCCCTAACCCTCTCCTCGG
					TCTCGATTCTACGCCC**CTCGAG**CCC
			V5R	R	GGG**CTCGAG**GGGCGTAGAATCGAGACCGAGGAGAG
					GGTTAGGGATAGGCTTACC**TCTAGA**GGG
pZC-FC	pZC	hIgG1-Fc (pIgPlus)	FC_EcoRI_XhoI_F	D	CCC**GAATTCTCGAG**CCCAAATCTTGTGAC
			FC_HindIII_R	R	CCC**AAGCTT**TCATTTACCCGGAGAC
pZ-EGFP	pcDNA3.1™/Zeo(-) (Invitrogen)	EGFP	EcoRV_EGFP_F	D	CCC**GATATC**GTGAGCAAGGGCGAG
			EGFP_HindIII_R	R	CCC**AAGCTT**TTACTTGTACAGCTC
pV5-NCR3A	pZCV	*NCR3A*	HsNCR3E2V5F	D	CCC**CTCGAG**TGGGTGTCCCAGCCCCCT
			HSNCR3E4IIIV5R	R	GGG**GAATTC**TCAGCCTCCTGGGACTGG
pV5-NCR3B	pZCV	*NCR3B*	HsNCR3E2V5F	D	CCC**CTCGAG**TGGGTGTCCCAGCCCCCT
			HSNCR3E4IIV5R	R	GGG**GAATTC**TCAGAGTTGGGGGAATCC
pV5-NCR3C	pZCV	*NCR3C*	HsNCR3E2V5F	D	CCC**CTCGAG**TGGGTGTCCCAGCCCCCT
			HSNCR3E4IV5R	R	GGG**GAATTC**CTAGGGACATCTGGGCTC
pV5-NCR3D	pZCV	*NCR3D*	HsNCR3E2V5F	D	CCC**CTCGAG**TGGGTGTCCCAGCCCCCT
			HSNCR3E4IIV5R	R	GGG**GAATTC**TCAGAGTTGGGGGAATCC
pV5-NCR3E	pZCV	*NCR3E*	HsNCR3E2V5F	D	CCC**CTCGAG**TGGGTGTCCCAGCCCCCT
			HSNCR3E4IIIV5R	R	GGG**GAATTC**TCAGCCTCCTGGGACTGG
pV5-NCR3F	pZCV	*NCR3F*	HsNCR3E2V5F	D	CCC**CTCGAG**TGGGTGTCCCAGCCCCCT
			HSNCR3E4IV5R	R	GGG**GAATTC**CTAGGGACATCTGGGCTC
pV5-B7H6	pZCV	*B7H6*	HsB7H6_PZCV_F	D	GGG**GAATTC**GTAGAGATGATGGCAGGG
			HsB7H6_PZCV_R	R	CCC**GGATCC**TTACTGTAGGGGTAACAG
pV5-B7.1	pZCV	*B7.1*	pZCV-B7.1_XhoI_F	D	CCC**CTCGAG**GTTATCCACGTGACCAAG
			pZCV-B7.1_EcoRI_R	R	CCC**GAATTC**TTATACAGGGCGTACAC
pNCR3C-V5His	pcDNA3.1™/V5-His TOPO® (Invitrogen)	*NCR3C*	HsNCR3_F2	D	CCC**AAGCTT**ATGGCCTGGATGCTGTTG
			pNCR3c_R3XbaI	R	GGG**TCTAGA**CCGGGACATCTGGGCTCTGG
pB7H6-His	pcDNA3.1™/V5-His TOPO® (Invitrogen)	*B7H6*	hsB7H6HisF	D	GGG**AAGCTT**ACCATGACGTGGAGGGCT
			hsB7H6HisR	R	GGG**ACCGGT**CTGTAGGGGTAACAG
pAL7-IgV	pAL7	*NCR3 ectodominio IgV*	pAL7NCR3E1-2F1_5EcoRI	D	CCC**GAATTC**TTTGTGCTCTCTGGGTGTCCCA
			pAL7NCR3RXbaI	R	GGG**TCTAGA**GGGACTGTACCAGCCCC
pAL7-IgC	pAL7	*NCR3 ectodominio IgC*	pAL7NCR3E1-2F1_5EcoRI	D	CCC**GAATTC**TTTGTGCTCTCTGGGTGTCCCA
			pAL7NCR3RXbaI	R	GGG**TCTAGA**GGGACTGTACCAGCCCC
pZC-NCR3A-Fc	pZC-FC	*NCR3 ectodominio IgV*	pZC-NCR3_XbaI_F	D	CCC**TCTAGA**CTCTGGGTGTCCCAGCC
			pZC-NCR3_xhoI_R	R	CCC**CTCGAG**CCGAAGGAGGAGGAC
pZC-NCR3E-Fc	pZC-FC	*NCR3 ectodominio IgC*	pZC-NCR3_XbaI_F	D	CCC**TCTAGA**CTCTGGGTGTCCCAGCC
			pZC-NCR3_xhoI_R	R	CCC**CTCGAG**CCGAAGGAGGAGGAC
pZC-CTLA4-Fc	pZC-FC	*CTLA4 (ectodominio)*	pZC-CTLA-4_XbaI_F	D	CCC**TCTAGA**AAAGCAATGCACGTGG
			pZC-CTLA-4_xhoI_R	R	CCC**CTCGAG**GTCAGAATCTGGGCAC
pZC-B7H6-Fc	pZC-FC	*B7H6 (ectodominio)*	pZC-B7H6_XbaI_F	D	CCC**TCTAGA**GATCTGAAAGTAGAG
			pZC-B7H6_xhoI_R	R	CCC**CTCGAG**GGAAAAATTATCTGTC
pZC-B7.1-Fc	pZC-FC	*B7.1 (ectodominio)*	pZC-B7.1_XbaI_F	D	CCC**TCTAGA**GTTATCCACGTGACCAAG
			pZC-B7.1_xhoI_R	R	CCC**CTCGAG**GTTATCAGGAAAATGC
pNCR3A-GFP	pZ-EGFP	*NCR3 (ectodominio IgV y región transmembrana)*	accATG_NCR3_F	D	CCC**GCTAGC**ACCATGGCCTGGATGCTG
			NCR3_TM_EcoRV	R	CCC**GAATTC**TCAGATATCTTTGCCCTGGTA
pNCR3E-GFP	pZ-EGFP	*NCR3 (ectodominio IgC y región transmembrana)*	accATG_NCR3_F	D	CCC**GCTAGC**ACCATGGCCTGGATGCTG
			NCR3_TM_EcoRV	R	CCC**GAATTC**TCAGATATCTTTGCCCTGGTA
pCTLA4-GFP	pZ-EGFP	*CTLA4 (ectodominio y región transmembrana)*	CT4GFP_F	D	CCC**GCTAGC**ACCATGGCTTGCCTTGG
			CT4GFP_R	R	CCC**GATATC**GCTTCTTTTCTTTAGC
pB7H6-GFP	pZ-EGFP	*B7H6 (ectodominio y región transmembrana)*	NheI_B7H6_F	D	GGG**GCTAGC**ACCATGACGTGGAGGGCT
			B7H6_TM_EcoRV	R	CCC**GATATC**TGAAGATGATTTG

Tabla 3-9. Oligonucleótidos utilizados como cebadores en la cuantificación por qPCR de las variantes de splicing del gen *NCR3*. También se indican los cebadores utilizados para la amplificación de los mensajeros normalizadores.

iniciador para evitar su expresión, se digirió con los enzimas *EcoRV* y *HindIII* para ser clonado en el vector pcDNA3.1™/Zeo(-) (Invitrogen) digerido con los mismos enzimas (Figura 3-5D).

Además se utilizó el vector pAL7, disponible en el laboratorio, derivado del vector pFast-Bac™1 (Invotrogen), en el que se clonó la secuencia codificante del péptido señal de la melitina (*Apis mellifera*) tras el promotor de la polihedrina y las secuencias codificantes de los epítopos V5 y 6xHis, en fase, en el extremo 3' del sitio de clonaje múltiple (MCS).

3.10.2 Plásmidos construidos

A partir de los vectores anteriores se obtuvieron las siguientes construcciones (plásmidos):

pV5-NCR3 (variante *A-F*), pV5-B7H6 y pV5-B7.1: el vector pZCV se usó como esqueleto para el clonaje de las distintas secuencias codificantes de las variantes de splicing del gen *NCR3* y de los mensajeros de *B7H6* (NM_001202439.1) y de *B7.1* (NM_005191.3). Se amplificaron mediante PCR las secuencias codificantes completas mencionadas, excepto la secuencia codificante del péptido señal, con los cebadores indicados en la Tabla 3-9. En el caso de las variantes del gen *NCR3* se utilizó como molde los plásmidos derivados de pGEM®-t Easy obtenidos previamente (ver apartado 3.6). Para la amplificación de los mensajeros de los genes *B7H6 y B7.1* se utilizó 1 μl de cDNA procedente de sangre de un donante sano. Los productos amplificados fueron digeridos con *XhoI* y *EcoRI*, excepto *B7H6* que fue digerido con *EcoRI* y *BamHI*, y clonados en el vector pZCV digerido del mismo modo que los insertos correspondientes manteniendo la fase de lectura con la secuencia codificante del péptido señal de CD33 y del epítopo V5 presentes en el vector (Figura 3-5E-F).

pNCR3C-V5His: se amplificó la secuencia codificante completa de la variante *C* del gen *NCR3*, excepto el codón de terminación para fusionar esta secuencia a las de V5 y His, con los cebadores HsNCR3_F2 y pNCR3c_R3XbaI (Tabla 3-9) y usando como molde el plásmido derivado de pGEM®-t Easy obtenido previamente. Tras la digestión del producto amplificado con *HindIII* y *XbaI* se clonó en el vector pcDNA3.1™/V5-His TOPO® (Invitrogen) digerido del mismo modo, quedando el inserto en fase con la secuencia codificante de los epítopos V5 y 6xHis (Figura 3-5G).

pB7H6-His: como en el caso anterior el vector utilizado como esqueleto fue pcDNA3.1™/V5-His TOPO® (Invitrogen). Se amplificó la secuencia codificante completa del mensajero de *B7H6*, excepto el codón de terminación para fusionar estas secuencias a His, usando como molde 1 μl de cDNA procedente de células HeLa y los cebadores hsB7H6HisF y hsB7H6HisR (Tabla 3-9). Se digirió el inserto con *HindIII* y *AgeI* y se clonó en el vector indicado digerido con los mismos enzimas de modo que se eliminó la secuencia codificante del epítopo V5 (Figura 3-5H).

pEGFP: para la obtención de este plásmido se digirió el plásmido pEGFP-N1 (NEB) con los enzimas *NheI* y *NotI*. El fragmento portador de la secuencia codificante

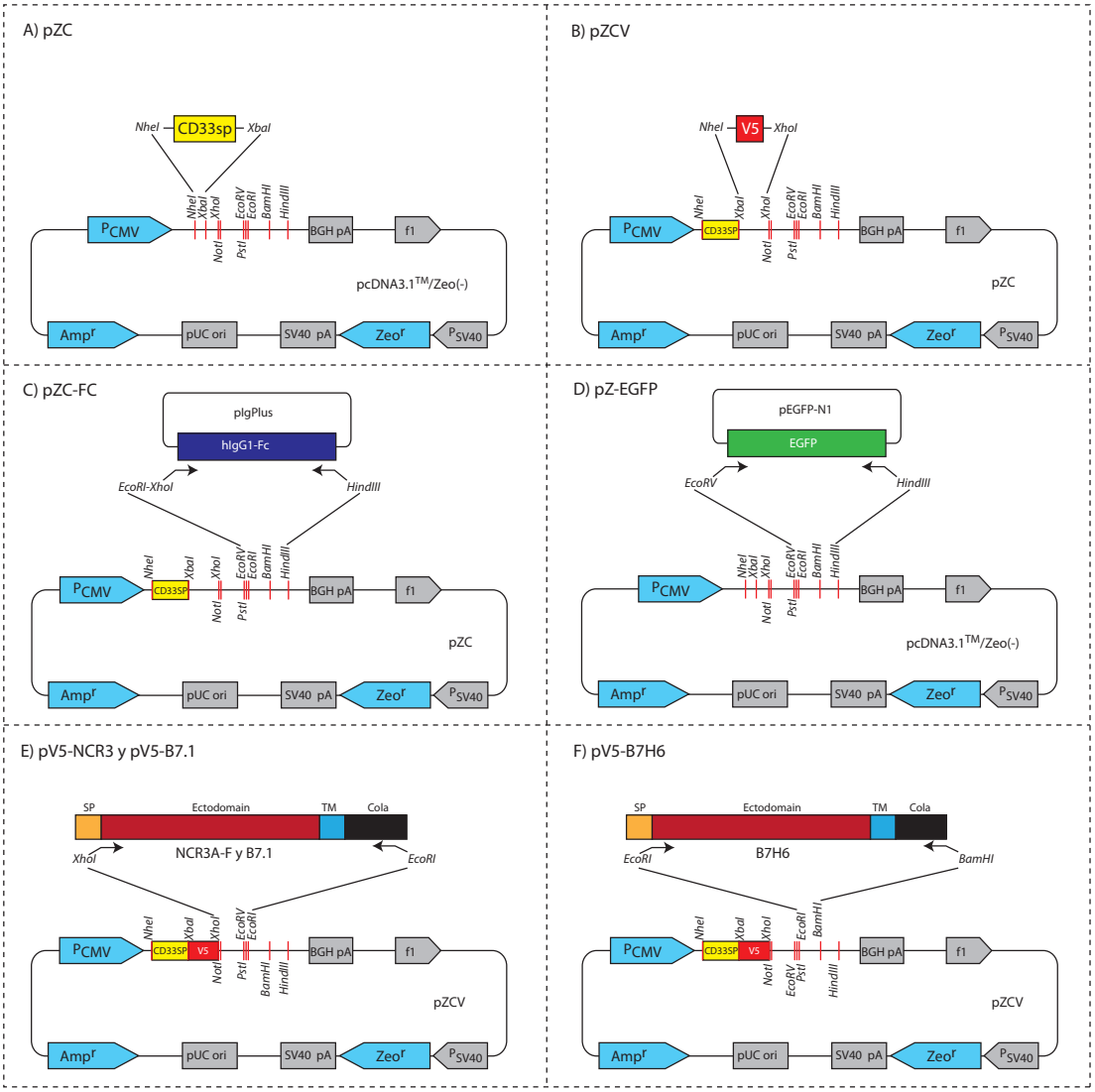


Figura 3-5. Representación esquemática del proceso realizo para la construcciones de vectores y plásmidos.

de la proteína EGFP se clonó en el vector pcDNA3.1^{TM}/Zeo(-) (Invitrogen) digerido con los mismos enzimas de restricción (Figura 3-5I).

pAL7-IgV y pAL7-IgC: la secuencia codificante correspondiente con los aminoácidos 17-139 de las isoformas A, B y C, y 17-114 de las isoformas D, E y F se clonaron en el vector pAL7, que posteriormente se utilizaron para la producción de las proteínas recombinantes IgV e IgC (ver apartado 3.13). Se amplificaron tales secuencias codificantes con los cebadores pAL7NCR3E1-2F1_5EcoRI y pAL7NCR3RXbaI (Tabla 3-9) utilizando como molde los plásmidos de pGEM®-t Easy de las variantes *A* y *E* del gen *NCR3*, previamente clonadas. Se digirieron los

productos amplificados con *EcoRI* y *XbaI*, y se clonaron el vector pAL7 digerido del mismo modo, de manera que los insertos estuvieran en fase con la secuencia codificante del peptido señal de la melitina en el extremo 5' y con los epitopos V5 y 6xHis en el extremo 3' (Figura 3-5J).

pZC-NCR3A-Fc, pZC-NCR3E-Fc, pZC-CTLA4-Fc, pZC-B7H6-Fc y pZC-B7.1-Fc: estos plásmidos se utilizaron para la producción de proteínas recombinantes en células de mamífero (ver apartado 3.14). Se clonaron las secuencias correspondientes con los dos potenciales dominios extracelulares o ectodominios de las variantes de splicing del gen *NCR3* (aminoácidos 19-143 de las isoformas A, B y C y 19-118 de las isoformas D, E y F), la secuencia codificante del ectodominio de la isoforma canónica del gen *CTLA4* (P16410, animoácidos 36-160) y la de los ectodominios de los ligandos B7H6 y B7.1 (Q68D85, aminoácidos 25-262 y P33681, 35-242, respectivamente). Se utilizaron los cebadores indicados en la Tabla 3-9 en cada caso para la amplificación de los insertos por PCR. Como molde se utilizaron los plásmidos derivados de pGEM®-t Easy de las variantes *A* y *E* de *NCR3* para la amplificación de las secuencias de éste, cDNA de HeLa para la amplificación de los ligandos B7H6 y B7.1 y cDNA procedente de sangre de un donante sano para la amplificación de la secuencia codificante del ectodominio de CTLA4. Todos los insertos fueron digeridos con los enzimas *XbaI* y *XhoI* y fueron clonados en el vector pZC-FC digerido con los mismos enzimas, quedando en fase con la secuencia codificante de la porción Fc de la hIgG1, necesaria para la posterior purificación de las proteínas recombinantes (Figura 3-5K).

pNCR3A-GFP, pNCR3E-GFP, pCTLA4-GFP y pB7H6-GFP: se utilizó el vector pZ-EGFP construido previamente para clonar la secuencia codificante de los dos posibles dominios extracelular y la región transmembrana de las isoformas del gen *NCR3* (aminoácidos 1-165 de las isoformas A, B y C y aminoácidos 1-140 de las isoformas D, E y F), del ectodominio y de la región transmembrana del ligando B7H6 (Q68D85, aminoácidos 1-293) y del ectodominio y de la región transmembrana del receptor CTLA4 (P16410, animoácidos 1-194). Se usaron los cebadores indicados en la Tabla 3-9 y se utilizaron los plásmidos derivados de pGEM®-t Easy de las variantes *A* y *E* de *NCR3* para la amplificación de las secuencias de éste, cDNA de HeLa para la amplificación del ligando B7H6 y cDNA procedente de sangre de un donante sano para la amplificación de CTLA4. Todos los insertos fueron digeridos con los enzimas

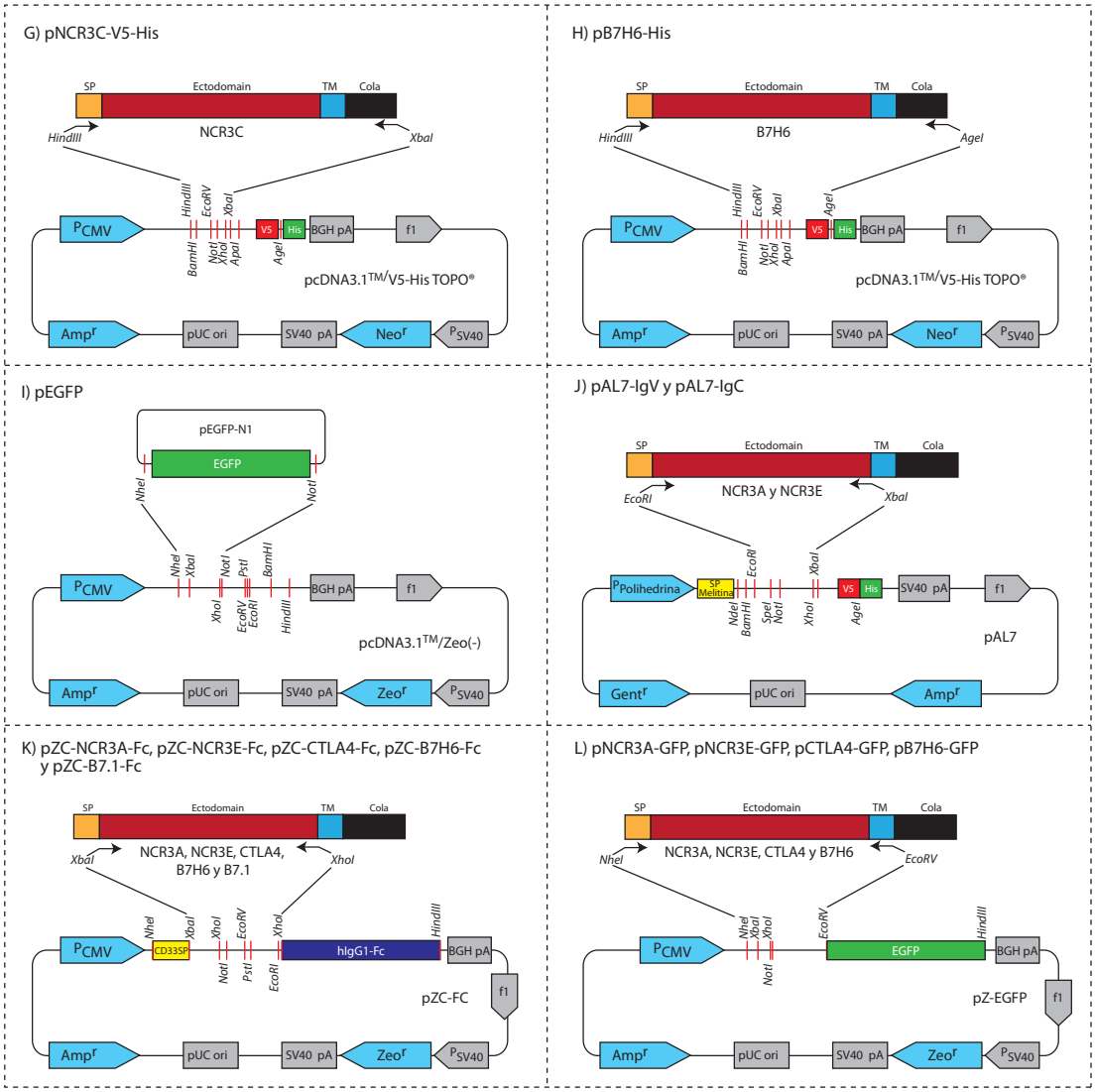


Figura 3-5 continuación. Representación esquemática del proceso realizo para la construcciones de vectores y plásmidos.

NheI y *EcoRV* y fueron clonados en el vector pZ-EGFP digerido con los mismos enzimas (Figura 3-5L).

3.11 Electroforesis en geles de poliacrilamida y Western Blot

3.11.1 Preparación de muestras

Para la preparación de muestras de proteína se utilizó tampón de carga 2X (120 mM Tris ajustado a pH 6,8 con HCl, 4% SDS, 20% glicerol, 10% ß-mercaptoetanol y 0,02% azul de bromofenol). Para las electroforesis en condiciones no reductoras se utilizó el mismo tampón al que no se le añadió ß-mercaptoetanol. La preparación de extractos

celulares se realizó con tampón de carga 1X (50 μl/cm^2, con o sin ß-mercaptoetanol) tras haber lavado las células dos veces con PBS. Se hirvieron las muestras 5 minutos a 95ºC antes de cargarlas en los geles de acrilamida.

3.11.2 Deglicosilaciones

Para el análisis de la glicosilación se utilizó los siguientes enzimas: PNGase F (NEB), O-glycosidase (NEB) y Neuraminidase (NEB), siguiendo las indicaciones del fabricante en cada caso. Se lisaron pocillos M6 de células transfectadas con las diferentes construcciones portadoras de las variantes de splicing de *NCR3* en 150 μl de tampón de desnaturalización (0,5% SDS y 40 mM de DTT) y usaron 14,5 μl de éstos para cada ensayo de deglicosilación. En el caso de las proteínas recombinantes purificadas se utilizaron 2 μg de proteína para cada ensayo.

3.11.3 Preparación de geles de poliacrimalida y desarrollo de las electroforesis

Las muestras de proteína preparadas se cargaron en minigeles de poliacrilamida (SDS-PAGE) preparados como se describe en la Tabla 3-10 y montados en cubetas *Mini-protean®* (BIO-RAD). El porcentaje del gel separador fue del 10% o del 12% en función de los tamaños que fueran necesarios resolver. Las electroforesis se realizaron a 100-200 voltios durante 45-60 minutos utilizando tampón Tris-Glicina (25 mM Tris, 250 mM Glicina y 0,1% SDS). En todos las electroforesis se utilizó el marcador de peso molecular *Precision Plus Protein™ Dual Color Standars* (BIO-RAD).

Gel separador	[Stock]	[Final]	Geles 10% (10 ml)	Geles 12% (10 ml)
Acrilamida/Bisacrilamida	40%	-	2,5 ml	3 ml
Tris-HCl pH 8,8	1 M	375 mM	3,75 ml	3,75 ml
SDS	10%	0,1%	100 µl	100 µl
H_2O destilada	-	-	3,55 ml	3,05 ml
APS	10%	0,1%	100 µl	100 µl
TEMED	-	-	4 µl	4 µl

Gel concentrador	[Stock]	[Final]	Gel 5% (4 ml)
Acrilamida/Bisacrilamida	40%	5%	0,5 ml
Tris-HCl, pH 6,8	1 M	125 mM	0,5 ml
SDS	10%	0,1%	40 µl
Bromophenol blue	1%	0,001%	4 µl
H_2O destilada	-	-	2,92 ml
APS	10%	0,1%	40 µl
TEMED	-	-	4 µl

Tabla 3-10. Componentes utilizados en la preparación de minigeles de poliacrilamida discontinuos.

3.11.4 Tinción con azul de Coomassie

Tras la electroforesis los geles fueron fijados con una solución de metanol (50%) y ácido acético glacial (10%) durante 30 minutos. La tinción se realizó sumergiendo los geles fijados en un solución de Azul de Coomassie G250 (0,001%), metanol (20%) y ácido acétido glacial (10%) durante 1 hora. La destinción se realizó con una solución de metanol (25%) y ácido acético glacial (7,5%) durante toda la noche.

3.11.5 Análisis por Western Blot

La transferencia de los geles a membranas de nitrocelulosa (Whatman), tras la electroforesis, se realizó con el sistema de transferencia húmeda *Trans-Blot® Cell* (BIO-RAD). El tampón de transferencia utilizado fue el mismo que el usado para la electroforesis suplementado con 20% de metanol. La transferencia se realizó a 100 voltios durante 1 hora. Se comprobó la transferencia mediante la tinción de las membranas con una solución de Rojo Ponceau S (0,1%) y ácido acético glacial (1%). Se bloquearon las membranas con PBS suplementado con Tween 20 (0,05%) y 5% de leche en polvo desnatada (solución de bloqueo) durante una hora en agitación. Tras el bloqueo se incubaron la membranas con los anticuerpos primarios necesarios (ver Tabla 3-11) en solución de bloqueo durante 1 hora. Entonces se lavaron las membranas con un exceso de PBS suplementado con Tween 20 (0,05%) durante 30 minutos. Tras ello se incubaron las membranas con el anticuerpo secundario correspondiente (ver Tabla 3-11), preparado en PBS suplementado con Tween 20 (0,05%), durante 30 minutos. Finalmente se lavaron las membranas con un exceso de PBS antes de tratarlas con la solución de revelado *ECL PLUS™ Westren Blotting Detection Reagent* (Amershan). Finalmente se colocaron las membranas en carcasas opacas (FisherBiotech) para la exposición de películas autoradigráficas (Curix RP2 Plus, Agfa).

Anticuerpos Primarios	Epítopo	Tipo	Casa comercial	Referencia	Diluciones
Anti-V5	Péptido V5	Monoclonal de ratón	Sigma	V8012	WB:1/5000, IF:1/500, FACs: 1/1000
Anti-Penta His	Péptido 6xHis	Monoclonal de ratón	QIAGEN	34660	WB:1/5000, IF:1/500, FACs: 1/1000
Anticuerpos Secundarios	**Epítopo**	**Tipo**	**Casa comercial**	**Referencia**	**Diluciones**
Anti-mIgG unido a HRP	Porción Fc de la IgG murina	Policlonal de oveja	Sigma	A5906	WB:1/5000
Anti-mIgG unido a Alexa 488®	Porción Fc de la IgG murina	Policlonal de burro	Invitrogen	A-21202	FACs: 1/500
Anti-hIgG1-Fc unido a HRP	Porción Fc de la IgG humana	Policlonal de cabra	Abcam	Ab98567	WB:1/5000
Anti-hIgG1-Fc unido a Alexa 488®	Porción Fc de la IgG humana	Policlonal de cabra	Invitrogen	A11013	FACs:1/500
Anti-hIgG1-Fc unido a Alexa 647®	Porción Fc de la IgG humana	Policlonal de cabra	Invitrogen	A-21445	FACs:1/500

Tabla 3-11. Listado de los anticuerpos comerciales utilizados. WB, Western blot; IF, Inmunofluorescencia; FACs, citometría de flujo; HRP, *Horseradish peroxidase.*

3.12 Inmunofluorescencia

Para realizar inmunofluorescencias se colocaron cubreobjetos circulares de vidrio en el fondo de la placa de cultivo (M24) antes de sembrar las células (COS-7). Al día siguiente las células fueron transfectadas como se describe en el apartado 3.3.2 y 24 horas tras la transfección se lavaron dos veces con PBS antes de fijarlas con PFA (paraformaldehido) al 4% durante 20 minutos a 4ºC, seguido de dos lavados con PBS. Posteriormente se trataron las células con una solución 50 mM de cloruro de amonio durante 30 minutos a 4ºC para reducir la autofluorescencie del PFA. Se lavaron las células dos veces con PBS tras lo que se permeabilizaron con PBS suplementado con Triton X-100 (0,1%) durante 30 minutos a 4ºC. Entonces se bloquearon con PBS suplementado con BSA (5%, "*Bovine serum albumin*") y Triton X-100 (0,1%) durante 30 minutos a 4ºC (solución de bloqueo). Tras el bloqueo se incubaron las células con el anticuerpo primario (ver Tabla 3-11) preparado en solución de bloqueo durante 1 hora a 4ºC. Se lavaron las células dos veces con PBS suplementado con Triton X-100 (0,1%) y se incubaron con el anticuerpo secundario correspondiente (ver Tabla 3-11) preparado en solución de bloqueo durante 1 hora a 4ºC. Se lavaron las células dos veces con PBS suplementado con Triton X-100 (0,1%) y se incubaron 5 minutos con una dilución 1/5000 de DAPI (4',6-diamidino-2-phenylindole) preparada en PBS para la tinción de los núcleos, tras lo que se lavaron de nuevo las células dos veces con PBS. Los cubreobjetos de montaron en portaobjetos utilizando *ProLong® Gold antifade reagents* (Invitrogen). Las inmunofluorescencias en condiciones no permeabilizantes siguieron el mismo protocolo salvo porque no se añadió Triton X-100 a las soluciones. Para el análisis de las preparaciones se utilizó un microscopio vertical *Axioskop 2 plus* (Zeiss) acoplado a una cámara ccd color. Para la toma de fotografías se utilizó el microscopio de barrido láser confocal y multifotón *LSM710* acoplado a un microscopio invertido *AxioObserver* (Zeiss).

3.13 Purificación de proteínas recombinantes en células de insecto

3.13.1 Obtención de baculovirus recombinantes

Para la obtención de baculovirus para la producción de las proteínas recombinantes correspondientes con los dos potenciales ectodominios de las isoformas del gen *NCR3* (IgV e IgC) se transfectaron bacterias competentes DH10bac™, como se indica en el apartado 3.2.2, con los plásmidos pAL7-IgV y pAL7-IgC (ver apartado

3.10.2). Tras la recombinación de estos plásmidos con el genoma del bácmido en las bacterias DH10Bac™ se extrajo el bácmido recombinante y se transfectó en células Hive-Five™ como se indica en el apartado 3.3.2. y 72 horas tras la transfección se recogió el sobrenadante de las células que contenía los baculovirus recombinantes. A este lote de virus se le denominó Pase 0. Para la expansión de los baculovirus se infectaron células Hive-Five™ sembradas al 50% del confluencia en placas p100 con el virus de Pase 0. Para ello se lavaron las células con medio completo sin suero y se añadió 3 ml de medio sin suero y 30 μl del virus Pase 0. Se dejó incubar 2 horas a 28ºC, tras lo que se eliminó el sobrenadante y se añadió 8 ml de medio completo y se incubaron las células 72 horas a 28ºC. Tras la incubación se recogió el sobrenadante que se denominó Pase 1. Con este lote de virus se infectaron células Hive-Five™ sembradas al 50% de confluencia en placas p150. Para ello, se lavaron las células con medio sin suero y se añadieron 10 ml de medio sin suero y 100 μl de virus de Pase 1. Se dejó incubar 2 horas a 28ºC tras lo que se eliminó le medio y se añadieron 20 ml de medio completo con suero. Se incubaron las células 72 horas a 28ºC, tras lo que se recogió el lote de virus del Pase 2.

3.13.2 Producción de las proteínas recombinantes

Para la producción de las proteínas recombinantes se infectaron 200 millones de células Hive-Five™ crecidas en suspención (ver apartado 3.3.1) a una densidad de 1.5 millones/ml con 20 ml de los baculovirus recombinantes de Pase 2. Para la infección simplemente se añadió la cantidad de virus indicada sobre las células y se incubaron 72 horas a 28ºC y agiración (180 rpm). Se siguió el mismo procedimiento para la producción de la proteína recombinante KHV (ectodominio de la proteína KHV ORF4 del herpesvirus de carpa, YP_001096043) cuyo virus estaba disponible en el laboratorio del Dr Antonio Alcamí y fue amablemente cedido por la Dra Soledad Blanco.

3.13.3 Purificación de las proteínas recombinantes

Tras la infección de las células Hive-Five™ en suspensión se centrifugaron las células a 6000 xg y se recogió el sobrenadante (aproximadamente 150 ml) conteniendo las proteínas recombinantes. Se concentró el sobrenadante en una *Amicon stirred cells* (Millipore) hasta un volumen de 2.5-3 ml usando una membrana cuyo tamaño de poro fue 3 veces inferior al tamaño de las proteínas a purificar (5 kDa en el caso de

Tampón	[Inicial]	[Final]	TPI10 10 mM Imidazol	TPI20 20 mM Imidazol	TPI40 40 mM Imidazol	TPI60 60 mM Imidazol	TPI250 250 mM Imidazol
Volumen	-	-	100 ml	100 ml	10 ml	10 ml	10 ml
Tampón fosfato*	100 mM	50 mM	50 ml	50 ml	5 ml	5 ml	5 ml
NaCl	3 M	0,3 M	10 ml	10 ml	1 ml	1 ml	1 ml
Imidazol	4 M	-	0,25 ml	0,5 ml	0,1 ml	0,15 ml	0,625 ml
H_2O Destilada	-	-	39,75 ml	39,5 ml	3,9 ml	3,85 ml	3,375 ml

Tabla 3-12. Tampones utilizados para la purificación de las proteínas recombinantes producidas con el sistema de baculovirus.
*100 mM Tampón fosfato ph 7,45: 405 ml de 0,2 M de Na_2HPO_4 (MERCK) + 95 ml de 0,2 M NaH_2PO_4 + 500 ml de H_2O destilada.

las proteínas IgV e IgC y de 10 kDa en el caso de la proteína KHV). Después de la concentración se intercambió el medio por tampón TPI10 (ver Tabla 3-12) aplicando el sobrenadante concentrado en una columna *PD-10* (GE Healthcare) pre-equilibrada con 5 volúmenes de tampón TPI10. Se eluyeron las proteínas de la columna *PD-10* con 4 ml de tampón TPI10. El eluido se incubó 1 hora a 4ºC con 600 µl de resina *Ni-NTA Agarose* (QIAGEN) lavada con 30 ml de tampón TPI10. Tras la unión de las proteínas se recogió el eluido (fracción de proteínas no unidas) y se lavó la resina con 40 ml de tampón TPI20 (ver Tabla 3-12), 1 ml de tampón TPI40 (ver Tabla 3-12) y 3 ml de tampón TPI60 (ver Tabla 3-12). Finalmente se eluyeron las proteínas unidas a la resina con 2.4 ml de tampón TPI250 (ver Tabla 3-12) que se recogieron en fracción de 600 µl en tubos independientes.

3.13.4 Análisis de la purificación y diálisis

Durante el proceso de purificación se tomaron muestras de cada una de las etapas y se cargaron en minigeles de poliacrilamida para su evaluación mediante electroforesis desnaturalizante y posterior tinción con azul de Coomassie (ver apartado 3.11). En todos los casos las proteínas recombinantes eluyeron mayoritariamente en las fracciones 2-3.

Se juntaron las fracciones con mayor concentración de proteína recombinante purificada y se dializaron usando columnas *Vivaspin 500* (Sartorios) siguiendo las indicaciones del fabricante. Se utilizaron columnas de 5 kDa de tamaño de poro para las proteínas IgV e IgC y de 10 kDa para la proteína KHV. El tampón de diálisis fue en siguiente: 20 mM HEPES ajustado a pH 7,4 con NaOH, 150 mM de NaCl y 10% glicerol. Las proteínas dializadas se conservaron en alícuotas a -20ºC.

3.13.5 Cuantificación

La cuantificación inicial de las proteínas dializadas se realizó espectrofotométricamente (*Nanodrop ND-1000*). Tras conocer esta concentración preliminar, se cargaron varias cantidades de las proteínas purificadas y dializadas en minigeles de poliacrilamida en los que se incluyeron cantidades conocidas de BSA (Promega). Tras la electroforesis y la tinción con azul de Coomassie de los minigeles, se cuantificaron las proteínas densitométricamente utilizando el programa ImageJ.

3.14 Purificación de las proteínas recombinantes en células de mamífero

3.14.1 Producción y purificación

Para la producción de proteínas recombinantes en células de mamífero se clonaron los diferentes ectodominios en fase con la secuencia codificante del dominio Fc de la inmunoglobulina G1 humana (hIgG1-Fc) clonada previamente en el vector pZC-FC (ver apartado 3.10). Se transfectaron 20-40 placas p100 de células HEK293T al 30% de confluencia (200-400 millones de células totales) en 5 ml de medio completo con 1% de FBS utilizando el método del PCa (ver apartado 3.3.2). A las 72 horas tras la transfección se recogió el sobrenadante, que se centrifugó 15 minutos a 15000 xg para eliminar posibles restos celulares, tras lo que se pasó por filtros de 0,45 μm. El sobranadante se aplicó a una velocidad de 1 ml/minuto sobre una columna *Hitrap Protein A HP* (Ge Healthcare) lavada previamente con 10 ml de tampón TP (Tabla 3-13) seguido por 10 ml de tampón TC (Tabla 3-13) y 10 ml más de tampón TP. Tras aplicar el sobrenadante se lavó la columna con 10 ml de tampón TP y se eluyeron las proteínas unidas a la resina con 5 ml de tampón TC, recogiendo fracciones de 500 μl que se neutralizaron con 90 μl de una solución de Tris-HCl 1M ajustada a pH 8,8.

Tampón TP, pH 7	[Inicial]	[Final]	Volumen (ml)
Na_2HPO_4	0,2 M	20 mM	2,89
NaH_2PO_4	0,2 M	20 mM	2,11
NaCl	3 M	150 mM	2,5
H_2O destilada	-	-	Hasta 50 ml
Tampón TC, pH 3	**[Inicial]**	**[Final]**	**Volumen (ml)**
Citrato sódico	0,2 M	100 mM	11,44
Ácido cítrico	0,2 M	100 mM	13,56
H_2O destilada	-	-	Hasta 50 ml

Tabla 3-13. Tampones utilizados para la purificación de las proteínas recombinantes producidas en células de mamíferos.

3.14.2 Análisis de la purificación, diálisis y cuantificación

Se tomaron alícuotas de todas las etapas de purificación que se evaluaron mediante electroforesis desnaturalizante en minigeles de poliacrilamida y posterior tinción con azul de Coomassie (ver aparatado 3.11). Las fracciones con mayor cantidad de proteína pura se juntaron y se dializaron siguiendo el mismo procedimiendo que en las proteínas recombinantes producidas en células de insecto (ver apartado 3.13), salvo porque el tamaño de poro de las columnas *Vivaspin 500* fue de 30 kDa para todas las proteínas producidas en células de mamíferos. La cuanfiticación, tras la diálisis, se realizó del mismo modo que en el caso de las proteínas recombinantes producidas en células de insecto (ver apartado 3.13).

3.15 Ensayos de interacción mediante citometría de flujo

Las proteínas recombinantes purificadas se utilizaron para el estudio de la interacción de éstas con el ligando celular B7H6. Para ello se lavaron las células, transfectadas transitoriamente o sin transfectar, con una solución de EDTA (0,02%) y se despegaron de la placa de cultivo mediante pipeteo. Se pasaron las células a tubos de 15 ml y se centrifugaron 5 minutos a 500 xg. Se eliminó el sobrenadante y se lavaron las células con un exceso de PBS. Se alicuotaron 100.000 células para cada ensayo en tubos de citómetro. Se centrifugaron las células y se bloquearon con 200 μl de PBS suplementado con 2% de BSA (PBSB) durante 30 minutos a RT. Tras el bloqueo se lavaron las células dos veces con 2 ml de PBSB. Se incubaron las células con 10 μg/ml de las proteínas recombinantes correspondientes en cada caso preparadas en PBSB o sólo con PBSB 30 minutos a RT, tras lo que se lavaron dos veces con 2 ml de PBSB. En el caso de las proteínas recombinantes producidas en células de insecto, se incubaron las células 30 minutos a RT con 100 μl de los anticuerpos primarios anti-V5 o anti-Penta His, según las necesidades, a la dilución adecuada (ver Tabla 3-11) y preparados en PBSB. Tras la incubación con los anticuerpos primarios las células se lavaron dos veces con PBSB y se resuspendieron en 100 μl de los anticuerpos secundarios adecuados (ver Tabla 3-11) diluidos en PBSB. Se incubaron las células 30 minutos con los anticuerpos secundarios tras lo que se lavaron dos veces con PBSB. Finalmente se resuspendieron las células con 500 μl de PBS y se midieron en un citómetro *FACSCalibur* (Becton Dickinson). El análisis de los resultados se realizó con el software *FlowJo*.

Reactivo	Casa Comecial	Reactivo	Casa Comercial
Acetato sódico	MERCK	Glicina	Sigma
Ácido acétido glacial	Panreac	Glicógeno	GE Healthcare
Ácido clorhídrico (HCl)	MERCK	HEPES	MERCK
Acrilamida/bisacrilamida (40%, 37.5:1)	BIORAD	Hidróxido sódico (NaOH)	MERCK
Agar bacteriológico	Pronadisa	IPTG	Sigma
Agarosa	Pronadisa	Kanamicina	Sigma
Albúmina sérica bovina (BSA)	Sigma	L-glutamina	Sigma
Aminoácidos no esenciales (AANE)	Gibco	Medio DMEM	Gibco
Ampicilina	Sigma	Medio Ham's F12	Invitrogen
Azul de bromofenol	Sigma	Medio RMPI	Gibco
Azul de Coomassie G250	Sigma	Medio TC100	Invitrogen
Bactotriptona	Pronadisa	Metanol	MERCK
Bromuro de etidio	Sigma	Paraformaldehido	Sigma
Cloruro de amonio (NH_4Cl)	MERCK	Penicilina G	Sigma
Cloruro de calcio ($CaCl_2$)	Sigma	Persulfato amónico (APS)	BIORAD
Cloruro de sodio (NaCl)	MERCK	Rojo Ponceau S	Sigma
DAPI	ROCHE	SDS	MERCK
Dimetil sulfóxido (DMSO)	Sigma	ß-mercaptoethanol	Sigma
EDTA	Sigma	Suero fetal de ternera (FBS)	Sigma
Estreptomicina	Sigma	TEMED	BIORAD
Etanol Absoluto	MERCK	Tetraciclina	Sigma
Extracto de levadura	Pronadisa	Tripsina	Sigma
Fosfato de disodio (Na_2HPO_4)	MERCK	Tritón X-100	Panreac
Fosfato monosódico (NaH_2PO_4)	MERCK	Trizma base (Tris)	Sigma
Gentamicina	Sigma	Tween® 20	Sigma
Glicerol	MERCK	X-Gal	Sigma

Tabla 3-14. Reactivos utilizados en el desarrollo de este trabajo.

3.16 Reactivos

La relación de reactivos y las empresas suministradores se detalla en la Tabla 3-14.

4. Resultados

Resultados

4.1 Anotación del gen *NCR3* y análisis de la conservación de su entorno genómico

El estudio del splicing alternativo de un determinado gen mediante PCR anidada depende del diseño de cebadores específicos, por lo que el conocimiento de la secuencia del gen bajo estudio (completa o parcial) y su anotación precisa son fundamentales. Por otro lado, el análisis de las secuencias procedentes de experimentos de RNA-seq requiere de la disponibilidad del genoma completo de la especie de las que proceden y una buena anotación es crítica para el análisis de las uniones de exones. Por lo tanto, el análisis de la información génica disponible sobre *NCR3* en las distintas bases de datos es el punto de partida de este estudio.

El gen *NCR3* está descrito en ocho de las 13 especies analizadas en este trabajo, sin embargo, la anotación en éstas presenta algunas deficiencias (Tabla 4-1). Para completar y corregir esta información se procedió a la anotación manual a través del análisis de la homología con especies emparentadas y el uso de herramientas bioinformáticas.

En las otras cinco especies restantes, todas del grupo de los primates (*Pongo pygmaneus*, *Colobus guereza*, *Papio cynocephalus*, *Macaca fascicularis* y *Cebus apella*), no se dispone de un genoma de referencia, por lo que se recurrió a la secuenciación del DNA genómico del gen *NCR3* en ellas. Para ello se elaboró un mapa de regiones consenso con secuencias disponibles de primates, lo que permitió el diseño de cebadores multi-especie para la amplificación de dos regiones del DNA

			NCBI			Ensembl			Uniprot
Especie	**Variante**	**Tipo**	**Gen**	**Mensajeros**	**Proteína**	**Gen**	**Mensajeros**	**Proteína**	**Proteína**
Homo sapiens	A	Codificante	259197	NM_147130.2	NP_667341.1	ENSG00000204475	ENST00000340027	ENSP00000342156	O14931-1
Homo sapiens	B	Codificante	259197	NM_001145466.1	NP_001138938.1	ENSG00000204475	ENST00000376073	ENSP00000365241	O14931-3
Homo sapiens	C	Codificante	259197	NM_001145467.1	NP_001138939.1	ENSG00000204475	ENST00000376072	ENSP00000365240	O14931-2
Homo sapiens	D	Codificante	-	-	-	-	-	-	O14931-6
Homo sapiens	E	Codificante	-	-	-	-	-	-	O14931-4
Homo sapiens	F	Codificante	-	-	-	ENSG00000204475	ENST00000376071	ENSP00000365239	O14931-5
Homo sapiens	NC1	No codificante	-	-	-	-	-	-	-
Homo sapiens	NC2	No codificante	-	-	-	-	-	-	-
Homo sapiens	NC3	No codificante	-	-	-	-	-	-	-
Homo sapiens	NC4	No codificante	-	-	-	ENSG00000204475	ENST00000491161	No codificante	-
Homo sapiens	NC5	No codificante				ENSG00000204475	ENST00000495600	No codificante	-
Homo sapiens	NC6	No codificante	-	-	-	ENSG00000204475	ENST00000478506	ENSP00000436198	-
Pan troglodytes	Ptro_A	Codificante	449613	NM_001009016.1	NP_001009016.1	ENSPTRG00000017968	ENSPTRT00000033186	ENSPTRP00000030663	P61484
Pan paniscus	Ppan-A	Codificante	100976229	XM_003831584.1	XP_003831632.1	-	-	-	-
Pan paniscus	Ppan-B	Codificante	100976229	XM_003831585.1	XP_003831633.1	-	-	-	-
Gorilla gorilla	Ggor-A	Codificante	-	-	-	ENSGGOG00000023848	ENSGGOT00000026196	ENSGGOP00000025342	G3SB68
Gorilla gorilla	Ggor-B	Codificante	101144701	XM_004043720.1	XP_004043768.1	-	-	-	-
Gorilla gorilla	Ggor-C	Codificante	-	-	-	ENSGGOG00000023848	ENSGGOT00000025954	ENSGGOP00000020547	G3RXL1
Pongo abelli	Pabe-A	Codificante	100448302	XM_002816723.1	XP_002816769.1	-	-	-	-
Pongo abelli	Pabe-B	Codificante	100448302	XM_002816722.1	XP_002816768.1	ENSPPYG00000016441	ENSPPYT00000019119	ENSPPYP00000018384	H2PII6
Pongo pygmaeus	-	-	-	-	-	-	-	-	-
Colobus guereza	-	-	-	-	-	-	-	-	-
Papio anubis	Panu-A	Codificante	101013418	XM_003897341.1	XP_003897390.1	-	-	-	-
Papio anubis	Panu-B	Codificante	101013418	XM_003897342.1	XP_003897391.1	-	-	-	-
Papio cynocephalus	-	-	-	-	-	-	-	-	-
Macaca mulatta	Mmul-A	Codificante	-	-	-	-	-	-	Q8MJ02-5
Macaca mulatta	Mmul-B	Codificante	-	-	-	ENSMMUG00000008854	ENSMMUT00000012386	ENSMMUP00000011616	-
Macaca mulatta	Mmul-C	Codificante	715574	NM_001042640.1	NP_001036105.1	ENSMMUG00000008854	ENSMMUT00000012387	ENSMMUP00000011617	Q8MJ02-1
Macaca mulatta	Mmul-E	Codificante	-	-	-	ENSMMUG00000008854	ENSMMUT00000012383	ENSMMUP00000011615	-
Macaca mulatta	Mmul-F	Codificante	715574	AY035216.1	AAK63118.1	-	-	-	Q8MJ02-2
Macaca mulatta	Mmul-NC6	No codificante	715574	AY035217.1	AAK63119.1	-	-	-	Q8MJ02-3
Macaca fascicularis	Mfas-C	Codificante	-	AJ278389.1	-	-	-	-	P61483
Cebus apella	-	-	-	-	-	-	-	-	-
Mus musculus	-	pseudogen	667612	-	-	ENSMUSG00000086076	-	-	-
Rattus norvegicus	Rnor-C	Codificante	294251	NM_181822.2	NP_861543.2	ENSRNOG00000000854	ENSRNOT00000001139	ENSRNOP00000001139	Q8CFD9
Bos taurus	Btau-C	Codificante	294251	NM_001040524.2	NP_001035614.1	ENSBTAG00000018343	ENSBTAT00000024407	ENSBTAP00000024407	Q32LF2
Sus scrofa	C	No codificante	100270822	XM_003128312.1	XP_003128360.1	ENSSSCG00000001407	ENSSSCT00000035932	-	-
Sus scrofa	SS7	Codificante	100270822	EU282355	ABX82821.1	ENSSSCG00000001407	ENSSSCT00000001538	-	B9TSR4

Los identificadores coloreados en gris se corresponden con predicciones.

Tabla 4-1. Información disponible de los mensajeros del gen *NCR3* en las bases de datos de cada una de las especies estudiadas en esta Tesis. Además, se ha incluido en esta tabla a *Pan paniscus*, *Pongo abelli* y *Papio anubis* ya que se han utilizado como referencia para la anotación de sus parientes más próximos.

genómico del gen *NCR3* en estas especies. La primera región (o región 5'), de unos 180 pb, contiene parte del exón 1, incluyendo el ATG iniciador, y una porción del intrón 1. La segunda región (o región 3'), de 570 pb, contiene los exones 3, 4I, 4II y 4III, así como los intrones entre dichos exones, incluyendo los codones de terminación de la traducción (Figura 4-1 y Tabla 3-5 en materiales y métodos). La información obtenida de la secuenciación del DNA genómico del gen *NCR3* en estas especies unido al estudio de la homología con otras especies emparentadas nos ha permitido la obtención de la anotación necesaria para el estudio del splicing alternativo. Con esta información se pudieron diseñar cebadores específicos que amplificaran los transcritos del gen *NCR3* en los distintos primates desde el ATG iniciador hasta los posibles codones de terminación.

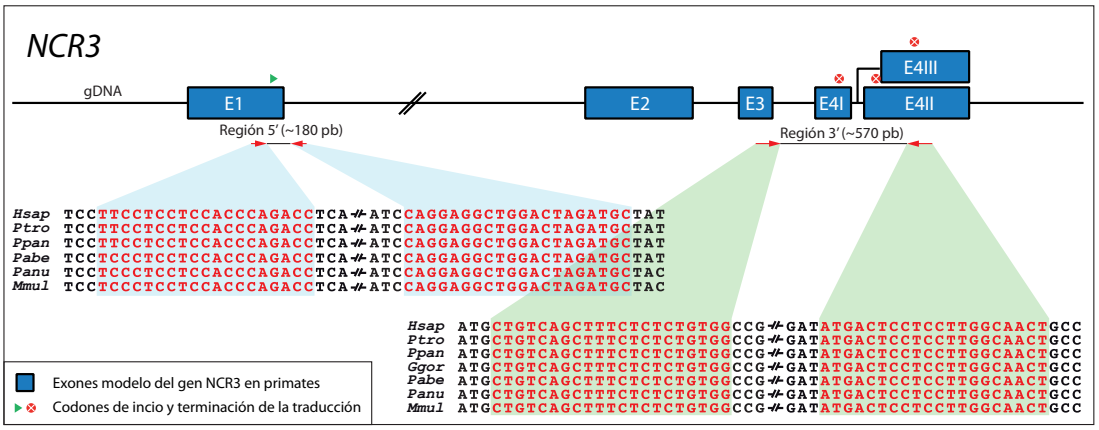


Figura 4-1. Alineamiento de la secuencia genómica disponible de disversos primates para el diseño de cebadores multi-especie, detallados en la materiales y métodos (Tabla 3-5). *Hsap, Homo sapiens; Ptro, Pan troglodytes; Ppan, Pan paniscus; Ggor, Gorilla gorilla; Pabe, Pongo abelli; Panu, Papio anubis y Mmul, Macaca mulatta.*

4.1.1 Homo sapiens (*NCR3*)

En humano, pese a la descripción de nueve transcritos diferentes del gen *NCR3* (ver Figura 1-9 en introducción) (Neville and Campbell, 1999), la información disponibles en las bases de datos sobre las variantes de splicing se encuentra dispersa e incompleta (Tabla 4-1 y Figura 4-2). Los mensajeros de las variantes *A*, *B* y *C* están presentes en todas las bases de datos principales, mientras que las variantes *D*, *E* y *F*, teóricamente codificantes como las primeras, sólo están presentes en Uniprot, a excepción del transcrito de la variante *F* que también se encuentra registrada en Ensembl. Las variantes no codificantes *NC1*, *NC2* y *NC3* han quedado fuera de las bases de datos, así como los mensajeros de las variantes codificantes *D* y *E* (Tabla 4-1 y Figura 4-2). Por otro lado, en Ensembl, están anotados tres nuevos mensajeros no codificantes (nombradas aquí como *NC4*, *NC5* y *NC6*). Las variantes *NC4* y *NC5*, basados en secuencias depositadas (BG341330.1 y BQ053092.1 respectivamente), presentan los mismos exones que la variante B a excepción del exón 1, que en ambos casos se reduce por el uso de sitios 5' donadores alternativos anteriores al ATG iniciador, sin que exista ningún otro inicio anterior (Tabla 4-1 y Figura 4-2). La variante *NC6* está basada en una predicción por homología con una secuencia de mensajero descrita en *Macaca mulatta* (AY035217.1), que está compuesta por los mismos exones que la variante C excepto el exón 2, que presente una reducción en su extremo 3' debido al uso de un sitio 5' donador alternativo (Figura 4-2). El uso de este sitio donador alternativo provoca un cambio de la fase de lectura que conduce a la aparición de un codón prematuro de terminación de la traducción en el

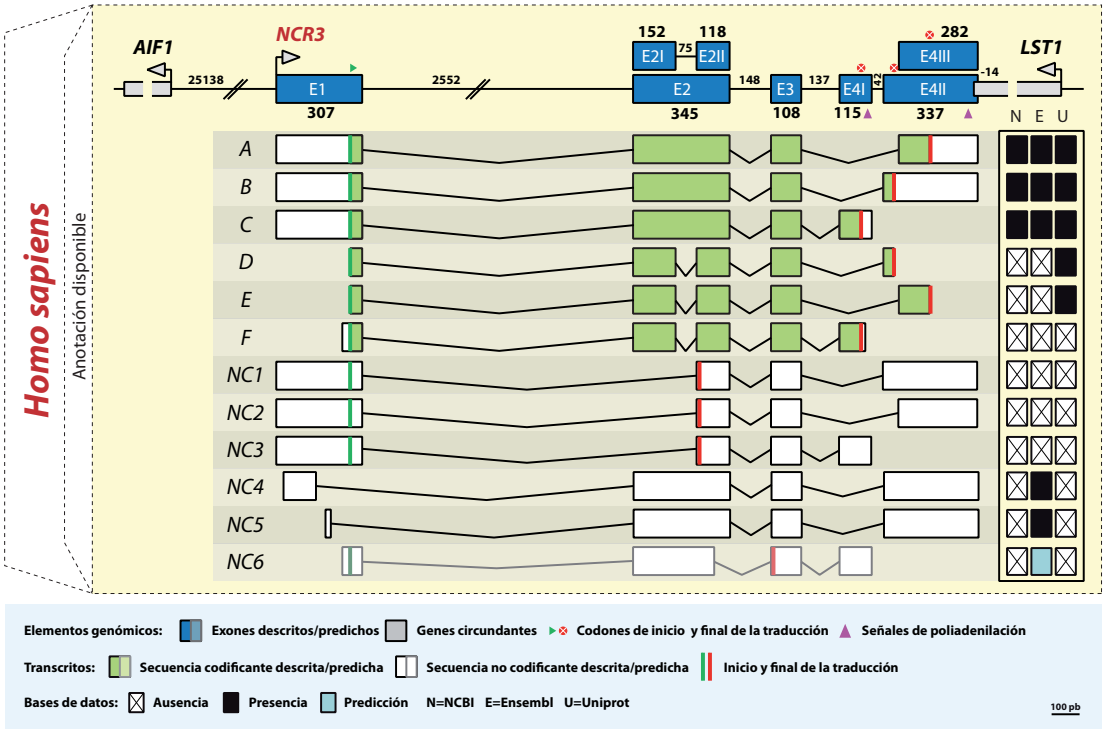


Figura 4-2. Anotación disponible del gen *NCR3* en *Homo sapiens*. Se han incluido las variantes *NC1*, *NC2* y *NC3* con fines comparativos, aunque no haya registro de éstas en las bases de datos.

exón 3, considerándose por ello un mensajero no codificante. Aunque la información disponible en las bases de datos es incompleta se puede considerar que los exones que formar el gen *NCR3* en humano son los mismos descritos inicialmente (Neville and Campbell, 1999), con la posible presencia de sitios alternativos de splicing en los exones 1 y 2, definidos por las nuevas variantes no codificantes anotadas en las bases de datos (Figura 4-2).

4.1.2 Pan troglodytes (*Ptro-ncr3*)

En las distintas bases de datos sólo se puede encontrar un mensajero anotado en esta especies (*Ptro-a*) (Tabla 4-1), equivalente a la variante *A* humana, lo que permite identificar los exones 1, 2, 3 y 4III según la nomenclatura utilizada en humano (Figura 4-3, anotación disponible). En Bonobo (*Pan paniscus*), otra especie de chimpancé, existen dos mensajeros anotados por predicción, equivalentes a las variantes *A* y *B* humanas, lo que permite la identificación del exón 4II del gen *Ptro-ncr3* por homología de secuencia (Figura 4-3, *Pan paniscus*). Por otro lado, la alta conservación de secuencia entre humano y chimpancé (~98%) (Sequencing and Consortium, 2005),

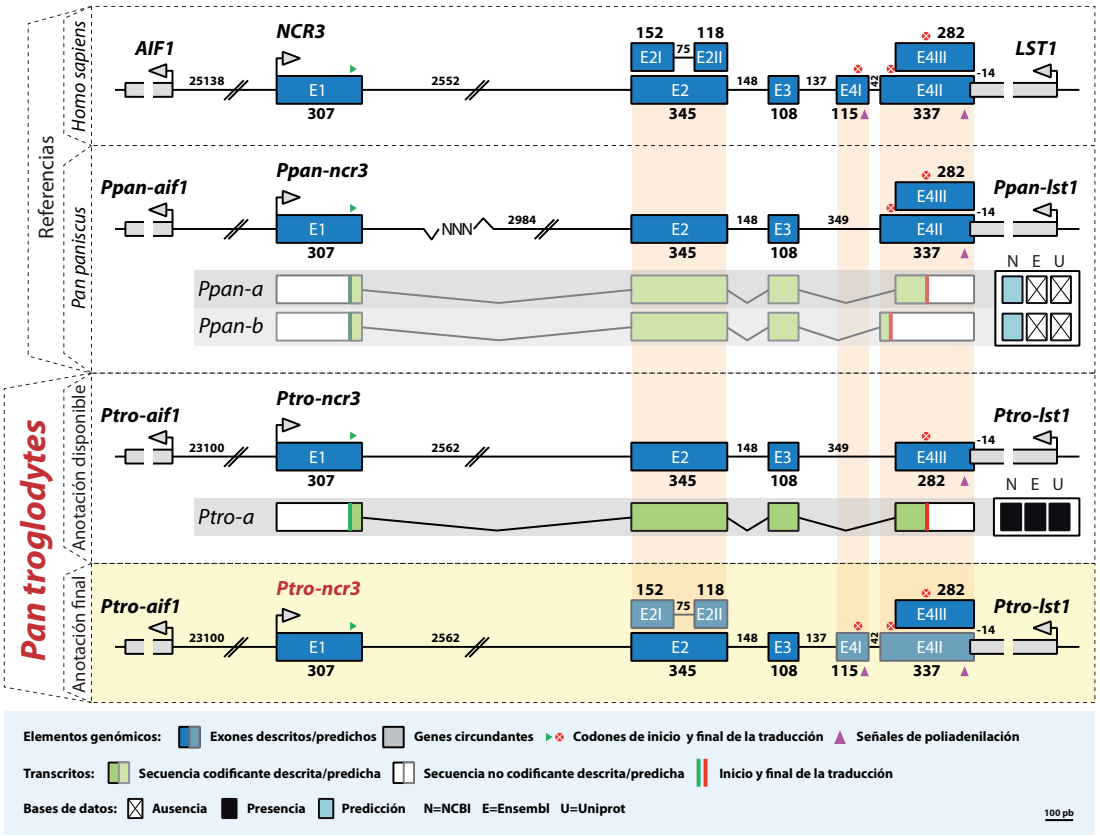


Figura 4-3. Anotación del gen *Ptro-ncr3*. En la parte superior se muestran las especies utilizadas como referencia para la corrección de la anotación disponible, que se muestra debajo junto a la anotación final. Las proyecciones, sombreados naranjas, indican el uso de los exones de referencia para la anotación final de *Pan troglodytes*.

permitió predecir la localización del exón 4I (Figura 4-3, anotación final). También, se puede observar la presencia de sitios de splicing (donador y aceptor) en el interior del exón 2 de chimpancé en las mismas posiciones relativas que en humano, lo que podrían dar origen a la división del exón 2 en los exones 2I y 2II, y por lo tanto, se podría generar en esta especie variantes de splicing similares a las humanas *D*, *E* y *F*, que codificarían un dominio inmunoglobulina tipo C. Asimismo, podrían producirse variantes de splicing similares a las humanas *NC1*, *NC2* y *NC3*, en las que las que el exón 2I es excluido del mensajero final. La anotación completa permite observar que la estructura exónica del gen *Ptro-ncr3* es similar a la de su homólogo humano, incluyendo la posición del ATG iniciador en el exón 1, los distintos codones de parada en los exones 4 alternativos y los dos sitios de poliadenilación (Figura 4-3, anotación final).

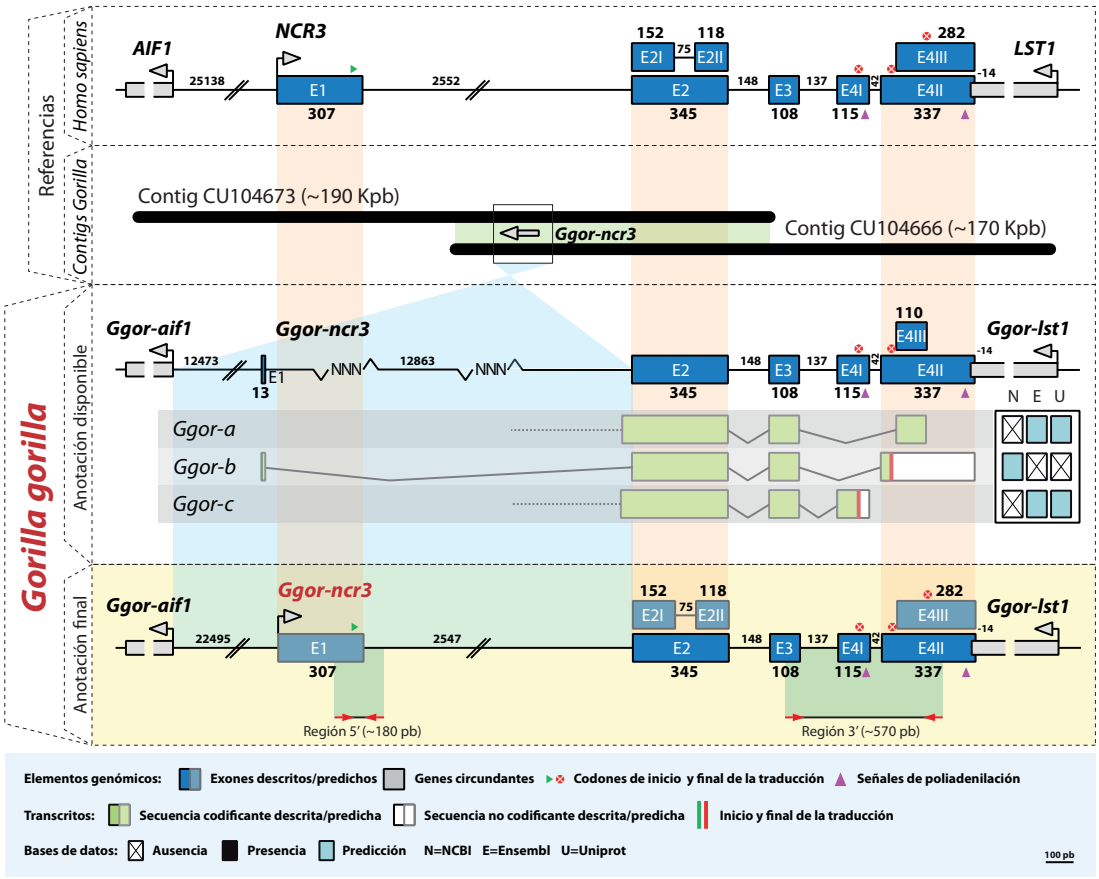


Figura 4-4. Anotación del gen *Ggor-ncr3*. En la parte superior se muestran las especies utilizadas como referencia para la corrección de la anotación disponible, que se muestra debajo junto a la anotación final. Las proyecciones, sombreados naranjas, azules y verdes, indican el uso de los exones y *contigs* de referencia y las secuencias obtenidas durante la secuenciación del DNA genómico, respectivamente, para la anotación final de *Gorilla gorilla.*

4.1.3 Gorilla gorilla (*Ggor-ncr3*)

El genoma disponible de gorila contiene muchos fragmentos sin secuenciar o *gaps*, y esa es la causa de que el gen *Ggor-ncr3* se encuentre anotado deficientemente (Tabla 4-1 y Figura 4-4). La anotación disponible en Ensembl muestra dos mensajeros similares a las variantes *A* y *C* humanas (*Ggor-a* y *Ggor-c* respectivamente), cuya traducción parcial se puede encontrar también en Uniprot (Tabla 4-1). Ambos mensajeros carecen del primer exón por encontrarse éste en una región aún sin secuenciar y presentan un exón 2 ligeramente mayor (Figura 4-4, anotación disponible). En NCBI un mensajero similar a la variante *B* humana (*Ggor-b*) es la única secuencia disponible, cuyo exón 1, predicho de manera automática, no presenta ninguna homología con humano ni con ningún otro primate (Figura 4-4, anotación disponible). La búsqueda de secuencias mediante Blast permitió conocer la existencia de dos

grandes fragmentos o *contigs* del cromosoma 6 de gorila (CU104666 y CU104673) que contienen la secuencia completa del DNA genómico del gen *Ggor-ncr3* (Figura 4-4, contigs gorilla). El ensamblaje "*in silico*" de estos dos fragmentos y la secuencia genómica de referencia permitió la identificación del exón 1, así como la obtención de la secuencia completa del intrón 1 (Figura 4-4, anotación final). Adicionalmente, se utilizaron los cebadores multi-especie para amplificar y secuenciar las regiones 5' y 3' del DNA genómico del gen *Ggor-ncr3* (Figura 4-1) utilizando como molde DNA genómico extraído de muestras de sangre disponibles de esta especie (Tabla S1), lo que permitió confirmar la fiabilidad de la secuencia genómica ensamblada "*in silico*" (Figura 4-4, anotación final). El análisis comparativo de la estructura exónica obtenida revela una gran similitud con humano y chimpancé, incluyendo las posiciones del ATG iniciador, los codones de terminación y los sitios de poliadenilación. Además, se puede detectar la presencia de sitios de splicing en el interior del exón 2 de *Ggro-ncr3*, que podrían originar variantes similares a las humanas *D*, *E* y *F*, así como variantes no codificantes (Figura 4-4, anotación final).

4.1.4 Pongo pygmaeus (*Ppyg-ncr3*)

El genoma del Orangután de Borneo (*Pongo pygmaeus*) no ha sido secuenciado, y en las principales bases de datos no existen referencias al gen *Ppyg-ncr3*. Por ello, se procedió a amplificar y secuenciar las regiones 5' y 3' del DNA genómico del gen *Ppyg_ncr3* (Figura 4-1) a partir de DNA genómico extraído de muestras de sangre disponibles (Tabla S1). Las secuencias obtenidas, junto a la de los exones 2 y 3, procedente del análisis de las variantes de splicing (ver apartado 4.3), se compararon con las secuencias disponibles de su pariente más próximo, el Orangután de Sumatra (*Pongo abelli*) (Figura 4-5, *Pongo abelli*), revelando una total identidad de secuencia entre ambas especies. En *Pongo abelli* se han descrito dos mensajeros por predicción equivalentes a las variantes humanas *A* y *B*, *Pabe-a* y *Pabe-b* respectivamente (Tabla 4-1 y Figura 4-5), lo que permitió definir los exones 1, 2, 3, 4II y 4III de *Ppyg-ncr3* por homología de secuencia. La predicción del exón 4I y la posible presencia de sitios de splicing internos en el exón 2 se determinaron por homología de secuencia con humano (Figura 4-5, anotación final). De manera general se puede observar que la estructura exónica inferida del gen *Ppyg-ncr3* es muy similar a la de su homólogo humano. Sin embargo, el codón de terminación presente en el exón 4I está desplazado 30 pb hacia el 5' del exón (posición +47), como ocurre en *Pongo*

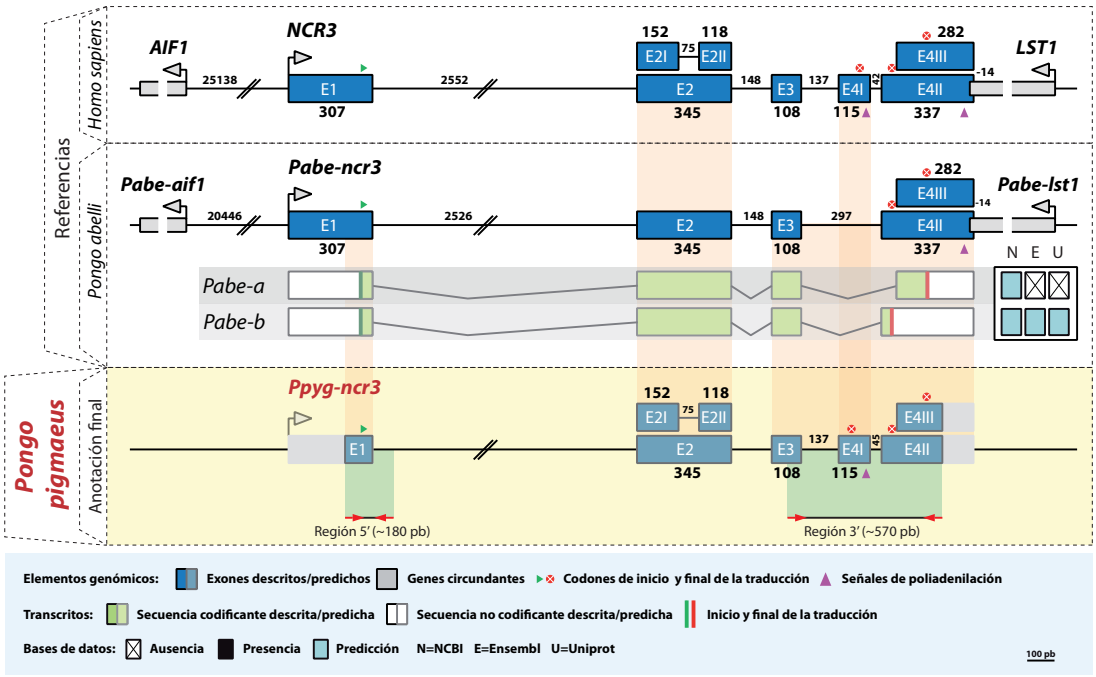


Figura 4-5. Anotación del gen *Ppyg-ncr3*. En la parte superior se muestran las especies utilizadas como referencia para la corrección de la anotación disponible, que se muestra debajo junto a la anotación final. Las proyecciones, sombreados naranjas y verdes, indican el uso de los exones de referencia y las secuencias obtenidas durante la secuenciación del DNA genómico, respectivamente, para la anotación final de *Pongo pygmaeus*.

abelli (resultado no mostrado), lo que provocaría una reducción en la longitud de la cola citoplasmática codificada en este exón. Por otro lado, aunque no se ha podido secuenciar el sitio de poliadenilación común para los exones 4II y 4III, en *Pongo abelli* se encuentra en la misma posición relativa que en humano (Figura 4-5, *Pongo abelli*).

4.1.5 Colobus guereza (*Cgue-ncr3*)

En esta especie no ha sido descrito el gen *Cgue-ncr3* y no existe ningún proyecto de secuenciación del genoma de *Colobus guereza* que permita la búsqueda por homología. Por lo tanto, se procedió a la amplificado de las regiones 5' y 3' del DNA genómico del gen *Cgue-ncr3*, usando los cebadores multi-especie (Figura 4-1) y el DNA genómico extraído de las muestras de sangre disponible (Tabla S1). El resultado de la secuenciación de estos fragmentos unido a los datos obtenidos del análisis de las variantes de splicing (ver apartado 4.3) y a la predicción por homología de secuencia con humano permitieron elaborar una versión preliminar de la estructura exónica del gen *Cgue-ncr3*, que muestra una gran similitud con el resto de primates analizados (Figura 4-6, anotación final). Las diferencias más importantes se concentran en el exón 4I, que presenta una longitud mayor que en humano por la inserción de

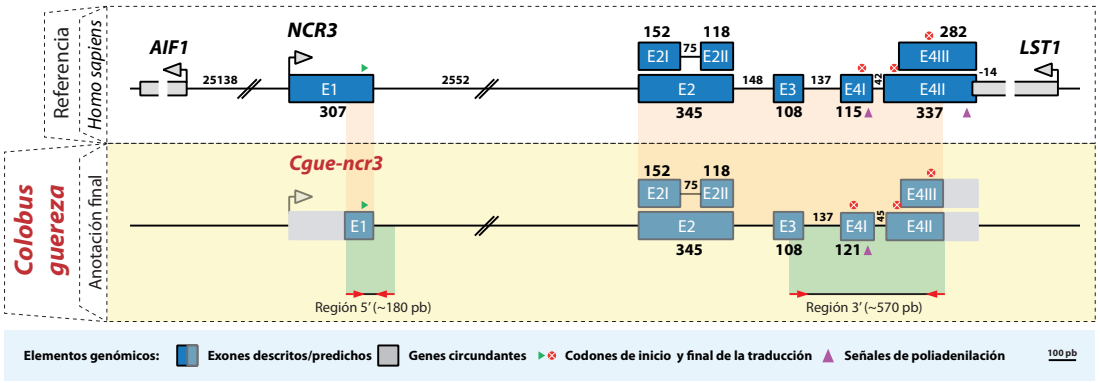


Figura 4-6. Anotación del gen *Cgue-ncr3*. En la parte superior se muestra la estrucutura exónica de *Homo sapiens*, utilizada como referencia para el análisis de las secuencias obtenidas del DNA genómico del gen *Cgue-ncr3* de *Colobus guereza* (sombreados verdes). Además, se incluyen las secuencias de los exones 2 y 3 obtenidos durante el análisis de las variantes de splicing expresadas en esta especie (ver apartado 4.3). Los sombrados naranjas indican el uso de la secuencia humana de referencia para definir los exones en *Colobus guereza*.

6 nucleótidos tras el sitio de poliadenilación, situado en la misma posición relativa que en humano (Figura 4-6). Además, el codón de terminación presente en este exón se encuentra en la posición +47, como en las especies del género *Pongo* aquí analizadas, lo que reduciría la longitud de la cola citoplasmática codificada por este exón con respecto a humano. El análisis de la secuencia obtenida de los exones 4II y 4III no revela grandes diferencias con respecto a humano o el resto de primates analizadas, presentando los codones de parada de la traducción en las mismas posiciones relativas. Por otro lado, la homología del exón 2 de *Cgue-ncr3* con el exón 2 humano ha permitido detectar sitios consenso de splicing en el interior de este exón que podrían dar lugar a un splicing interno similar al descrito en humano (Figura 4-6, anotación final).

4.1.6 Papio cynocephalus (*Pcyn-ncr3*)

El gen *Pcyn-ncr3* no ha sido descrito en *Papio cynocephalus*, ni existe en las bases de datos ninguna secuencia de mensajero procedente de este gen. Además, el genoma de esta especie no ha sido secuenciado, por lo que se procedió a la amplificación del DNA genómico del gen *Pcyn-ncr3* y su posterior secuenciación (Figura 4-1 y Tabla S1). Para el análisis de los resultados de la secuenciación se utilizó como referencia la secuencia genómica de una especie próxima, *Papio anubis*, en la que además existe anotación por predicción de dos mensajeros del gen *Panu-ncr3*, similares a las variantes *A* y *B* humanas, *Panu-a* y *Panu-b* respectivamente (Tabla 4-1 y Figura 4-7). La secuenciación de la región 5' del DNA genómico del gen *Pcyn-ncr3* no mostró ningún cambio con respecto a la de su homólogo en *Papio anubis*, permitiendo

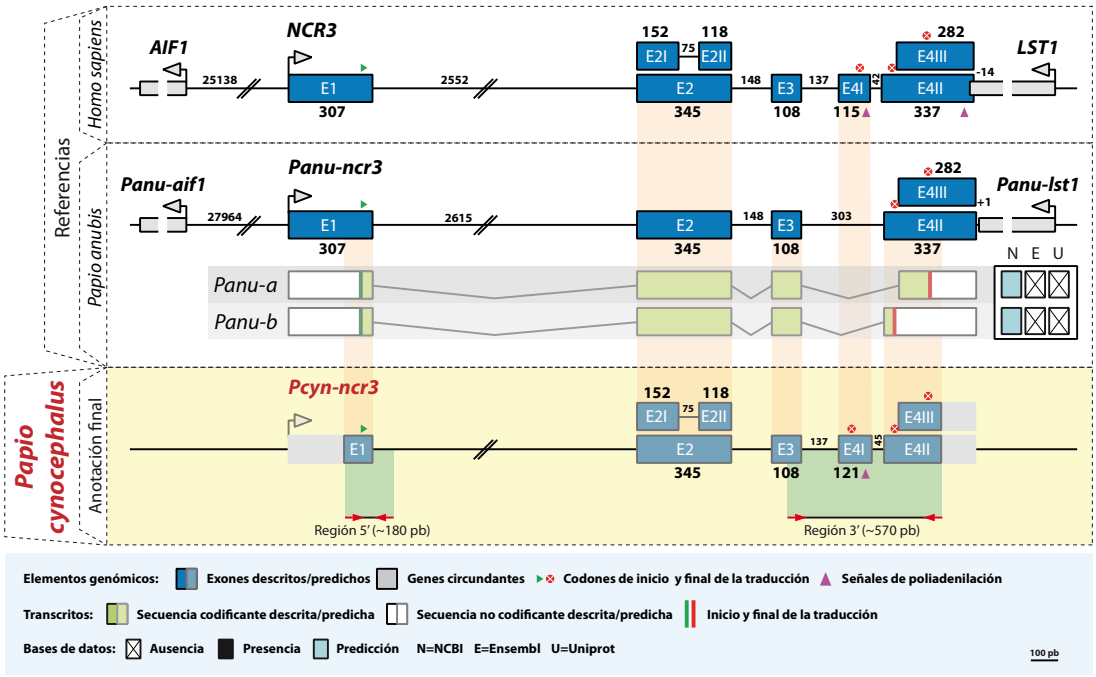


Figura 4-7. Anotación del gen *Pcyn-ncr3*. En la parte superior se muestran las especies utilizadas como referencia para la corrección de la anotación disponible, que se muestra debajo junto a la anotación final. Además, se incluyen las secuencias de los exones 2 y 3 obtenidos durante el análisis de las variantes de splicing expresadas en esta especie (ver apartado 4.3). Las proyecciones, sombreados naranjas y verdes, indican el uso de los exones de referencia y las secuencias obtenidas durante la secuenciación del DNA genómico para la anotación final de *Papio cynocephalus*.

determinar la posición del ATG iniciador, localizado en la misma posición relativa que en humano (Figura 4-7, anotación final). En la secuencia obtenida de la región 3' se pudo localizar los exones 3, 4II y 4III mediante la comparación con *Papio anubis*, mientras que el exón 4I se identificó en ambas especies de papiones mediante el análisis comparativo con humano (Figura 4-7, anotación final). El exón 4I, en ambos papiones, presenta la misma inserción de 6 pb observada en *Colobus guereza* tras el sitio de poliadenilación, estando éste localizado en la misma posición relativa que en humano. El codón de terminación presente en este exón (4I) en *Papio anubis* se localiza en la posición +35 (no mostrado), 52 pb antes que en humano, mientras que en *Papio cynocephalus* existe un polimorfisco en la posición +34 (G/C) que elimina ese codón de parada, encontrando el siguiente en la posición +47, similar a lo observado en *Colobus guereza* y en el género *Pongo*. La posición de los codones de parada en los exones 4II y 4III se sitúan en las mismas posiciones relativas que en humano (Figura 4-7, anotación final). Aunque la secuencia obtenida de la región 3' de *Pcyn-ncr3* no permite determinar la presencia de la señal de poliadenilación compartida por los exones 4II y 4III, podemos observar que la señal de poliadenilación

presente en *Papio anubis*, localizada en la misma posición relativa que en humano, tiene un polimorfismo que evitaría su reconocimiento, no encontrando ninguna otra señal de poliadenilación (Figura 4-7, *Papio anubis*). Finalmente, la secuencia del exón 2 de *Pcyn-ncr3*, obtenida en el análisis de las variantes de splicing (ver apartado 4.3), presenta sitios de splicing internos en las mismas posiciones relativas que en humano, que también observamos en *Papio anubis* (no mostrado).

4.1.7 Macaca mulatta (*Mmul-ncr3*)

Macaca mulatta es una especie ampliamente utilizada en investigación, por lo que la anotación disponible en esta especies es extensa. En NCBI se pueden encontrar tres mensajeros anotados para el gen *Mmul-ncr3*, aunque sólo la variante *Mmul-c*, equivalente a la variante *C* humana, aparece como secuencia de referencia. Adicionalmente, se pueden encontrar en NCBI las secuencias depositadas de las variantes *Mmul-f* y *Mmul-nc6*, equivalentes a las humanas *F* y *NC6* (Tabla 4-1 y Figura 4-8, anotación disponible). En Ensembl existe registro de las variantes *Mmul-b*, *Mmul-c* y *Mmul-e*, aunque sorprende comprobar que se trata de predicciones basadas en las secuencias de humano, incluso la variante *Mmul-c* de la que existe una secuencia depositada en NCBI. En Uniprot encontramos registro de las proteínas correspondientes con las variantes *Mmul-a*, de la que no existe registro a nivel de

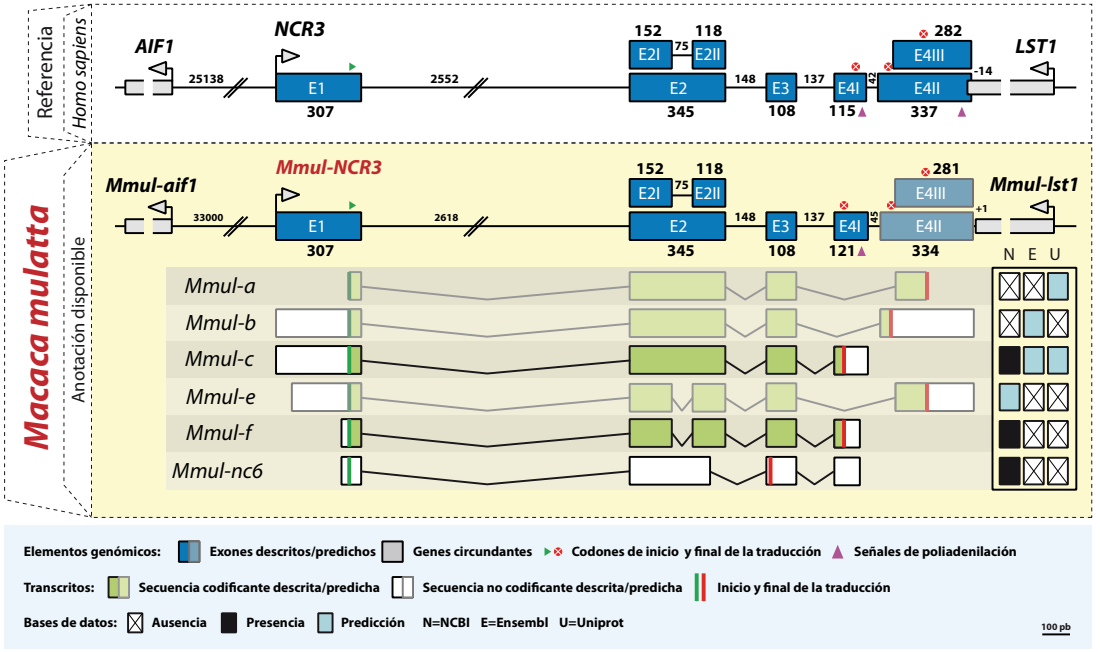


Figura 4-8. Anotación del gen *Mmul-ncr3*. En la parte superior se muestra la estrucutura exónica de *Homo sapiens*, utilizada como referencia. En la parte inferior se muestra la anotaciones disponibles para *Macaca mulatta*.

RNA mensajero, *Mmul-c* y *Mmul-e*, todas ellas predicciones. Además, existe otra secuencia de proteína registrada en Uniprot procedente de la traducción del mensajero *Mmul-nc6*, teóricamente no codificante por la presencia de un codón prematuro de terminación de la traducción (Figura 1-8). El análisis comparativo de la secuencia del gen *Mmul-ncr3* con su homólogo humano revela que la estructura exónica es similar en ambas especies (Figura 1-8). Como en el colobo (*Colobus guereza*) y en las especies de papiones (*Papio anubis y Papio* cynocephalus) aquí analizadas, se puede observar una inserción de 6 pb tras el sitio de poliadenilación presente en el exón 4I, incrementando ligeramente su tamaño con respecto a humano. La secuencia genómica del gen *Mmul-ncr3* disponible en NCBI y Ensembl son idénticas, sin embargo, existe cierta discrepancia en las secuencias disponibles de la variante *Mmul-c* entre las distintas bases de datos. En la secuencia genómica el codón de parada presente en el exón 4I se localiza en la posición +35 desde el inicio del exón, similar a lo observado en *Papio anubis,* siendo éste el codón de parada descrito en Ensembl. Sin embargo, en NCBI el codón de parada descrito para la variante *Mmul-c* se sitúa en la posición +47, aunque en la secuencia genómica estaría en la posición +35, lo que se debe a la presencia de tres polimorfismos en la secuencia del mensajero depositada, afectando uno de ellos al codón de parada presente en la secuencia genómica en la posición +35. Además de la secuencia de referencia para la variante *Mmul-c*, en NCBI, podemos encontrar otras tres secuencias depositadas que se corresponden con la misma variante (AJ554301.1, AY035214.1, AY035215.1), estando el polimorfismo de la posición +35 en las dos primeras, mientras que en la tercera la secuencia del exón 4I es idéntica a la secuencia genómica. Asimismo, en los mensajeros de las variantes *Mmul-f* y *Mmul-nc6* (AY035217.1 y AY035216.1, respectivamente) tampoco existe el polimorfismo en la posición +35 del exón 4I, de modo que se ha considerado el codón de parada en la posición +35 como el de referencia para esta especie (Figura 1-8, anotación final). Por otro lado, los codones de parada de la traducción presente en los exones 4II y 4III están en las mismas posiciones relativas que en humano. El sitio de poliadenilación común a estos exones presenta un polimorfismo que evita el reconocido como tal en los programas de predicción, no existiendo otro sitio alternativo en estos exones ni en la secuencia posterior, similar a lo detectado en *Papio anubis*.

4.1.8 Macaca fascicularis (*Mfas-ncr3*)

Aunque el gen *Mfas-ncr3* no ha sido descrito en esta especie, la búsqueda mediante Blast reveló la existencia de una secuencia de cDNA depositada en las bases de datos (AJ278389.1, Tabla 4-1), que contiene los exones equivalentes a los que componen la variante *C* humana (*Mfas-c*, Figura 1-9). Por otro lado, tampoco existe un genoma de referencia para esta especie, aunque si existe un borrador del que se extrajo la secuencia genómica completa del gen *Mfas-ncr3*. El cDNA de la variante *Mfas-c,* aunque parcial en su extremo 5', permitió definir los exones 1 (parcial), 2, 3 y 4I (Figura 1-9, anotación final). Para determinar la localización de los exones 4II y 4III se recurrió al análisis de la homología con la especie más próxima, *Macaca mulatta* (Figura 1-9, *Macaca mulatta*). Asimismo, el análisis comparativo con *Macaca mulatta* permitió definir el exón 1 completo, así como la potencial división del exón 2 en dos fragmentos (exones 2I y 2II), por la presencia de sitios consenso de splicing en las mismas posiciones relativas. La estructura exónica inferida no difiere de la descrita en *Macaca mulatta*. El exón 4I, 6 pb mayor que el equivalente humano por una inserción tras el sitio de poliadenilación, presenta el codón de parada en la posición +35, como lo observado en *Papio anubis* y en *Macaca mulatta* (ver apartado anterior). La secuencia de los exones 4II y 4III en esta especie no muestra ningún cambio con

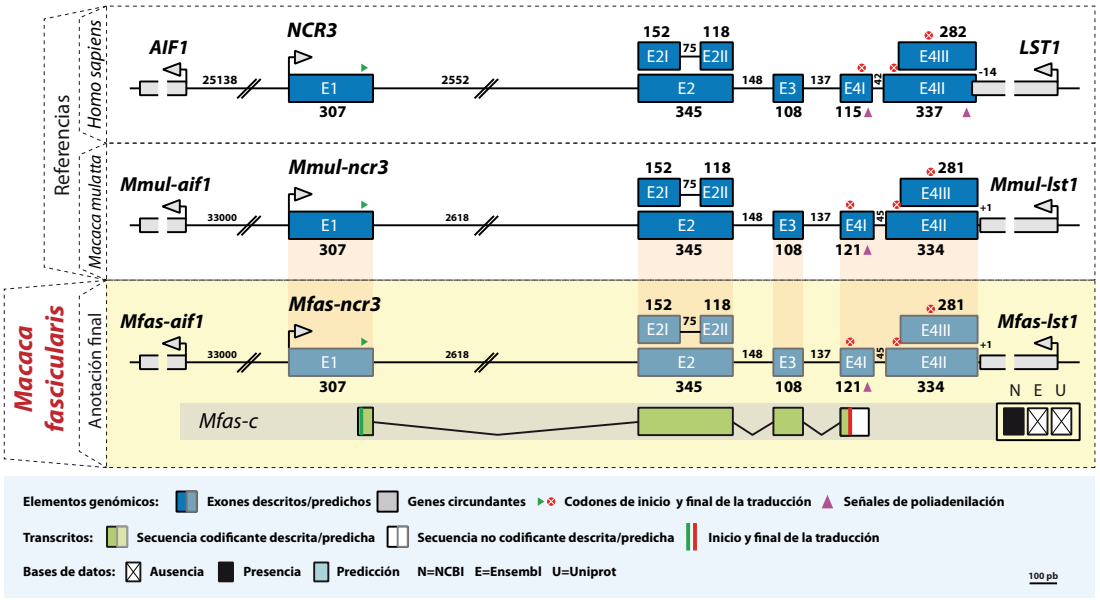


Figura 4-9. Anotación del gen *Mfas-ncr3*. En la parte superior se muestran las especies utilizadas como referencia para establecer la anotación final que se muestra debajo. Las proyecciones, sombreados naranjas, indican el uso de los exones de referencia para la anotación final de *Macaca fascicularis*.

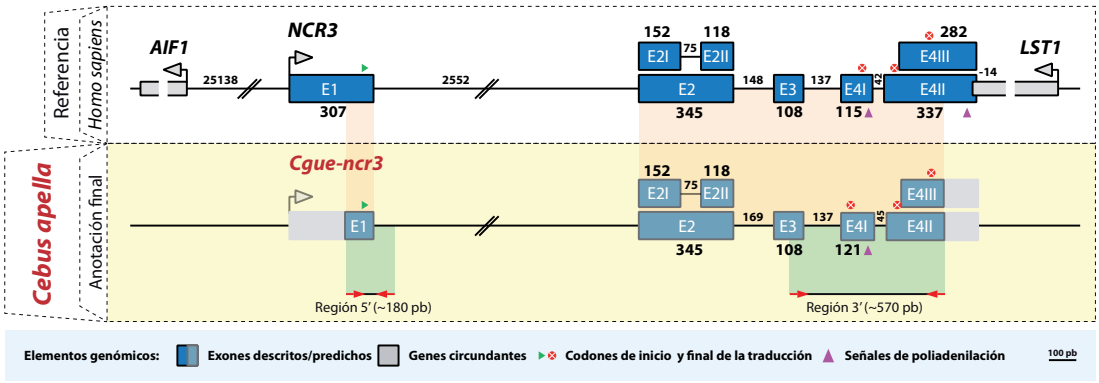


Figura 4-10. Anotación del gen *Cape-ncr3*. En la parte superior se muestra la estrucutura exónica de *Homo sapiens*, utilizada como referencia para el análisis de las secuencias obtenidas del DNA genómico del gen *Cgue-ncr3* de *Cebus apella* (sombreados verdes). Además, se incluyen las secuencias de los exones 2 y 3 obtenidos durante el análisis de las variantes de splicing expresadas en esta especie (ver apartado 4.3). Los sombrados naranjas indican el uso de la secuencia humana de referencia para definir los exones en *Cebus apella.*

respecto a *Macaca mulatta*, y por lo tanto, presenta los codones de parada en las mismas posiciones relativas que ésta. El sitio de poliadenilación presente en humano en estos exones muestra el mismo polimorfismo en la secuencia del gen *Mfas-ncr3* que el detectado en *Macaca mulatta* (Figura 4-9, anotación final).

4.1.9 Cebus apella (*Cape-ncr3*)

La información sobre esta especie en las bases de datos es escasa, principalmente por la ausencia de un proyecto de secuenciación de su genoma, no existiendo ninguna secuencia relacionada con el gen *Cape-ncr3*. Por lo tanto, se recurrió a la secuenciación de las regiones 5' y 3' del gen utilizando el DNA extraído de las muestras de sangre disponibles (Figura 4-1 y Tabla S1). Con las secuencias obtenidas, unidas a las de los exones 2 y 3 obtenidas en al análisis de las variantes de splicing (ver apartado 4.3), se pudo establecer la estructura exónica preliminar del gen *Cape-ncr3*. El análisis comparativo con humano revela una gran similitud en la estructura exónica entre ambas especies. El ATG iniciador del gen *Cape-ncr3* se localiza en la misma posición relativa que en humano, y podemos detectar los sitios de splicing interno en el exón 2 presentes en humano (Figura 4-10). El exón 4I presenta la misma inserción de 6 pb tras el sitio de poliadenilación conservado que detectamos en las especies del género *Macaca* y *Papio*, así como en *Colobus guereza*. En esta especie el codón de parada presente en el exón 4I se sitúa en el posición +35 como en el género *Macaca* o en *Papio anubis*. Como en el resto de los primates analizados las posiciones relativas de los codones de parada presentes en los exones 4II y 4III son similares a *Homo sapiens*, sin embargo, en esta especie no

se pudo determinar la presencia de la secuencia de poliadenilación común para estos exones puesto que el fragmento secuenciado no contenía tal sitio (Figura 4-10).

4.1.10 Rattus norvegicus (*Rnor-ncr3*)

En gen *Rnor-ncr3*, en rata, sólo cuenta con un mensajero anotado, similar a la variante humana *C*, *Rnor-c* (Tabla 4-1 y Figura 4-11), presente en todas las bases de datos, basaba en secuencias depositadas. Por lo tanto, los exones anotados en el gen *Rnor-ncr3* se corresponden con los exones 1, 2, 3 y 4I humano (Figura 4-11). La búsqueda por homología de los exones 4II y 4III y el uso de herramientas de predicción de exones, no dieron ningún resultado positivo, por lo que se puede decir que estos exones no existen en esta especie. El análisis comparativo de los exones anotados muestra algunas diferencias entre rata y humano. En rata, la región no codificante del exón 1 es ligeramente mayor que en humano, sin embargo, el codón de inicio de la traducción se sitúa en la misma posición relativa (Figura 4-11). El exón 2, muestra la misma longitud que en humano, y presenta sitios de splicing internos similares a los descritos en humano, sin embargo, el sitio 5' donador se sitúa 1 nucleótido antes que en humano debido a un polimorfismo. El uso de este sitio de splicing generaría un cambio en la fase de lectura, lo que conduciría a la aparición de codones prematuros de terminación de la traducción en la segunda mitad del exón 2. Por otro lado, el exón 3 es 24 pb más largo en *Rattus norvegicus*, con respecto a humano, lo que amplía su secuencia codificante. Además, el exón 4 (4I en humano) es 1 pb más largo en humano, sin embargo, la posición del codón de terminación de la traducción en rata se sitúa 17 pb antes que en humano, en la posición +59, lo que

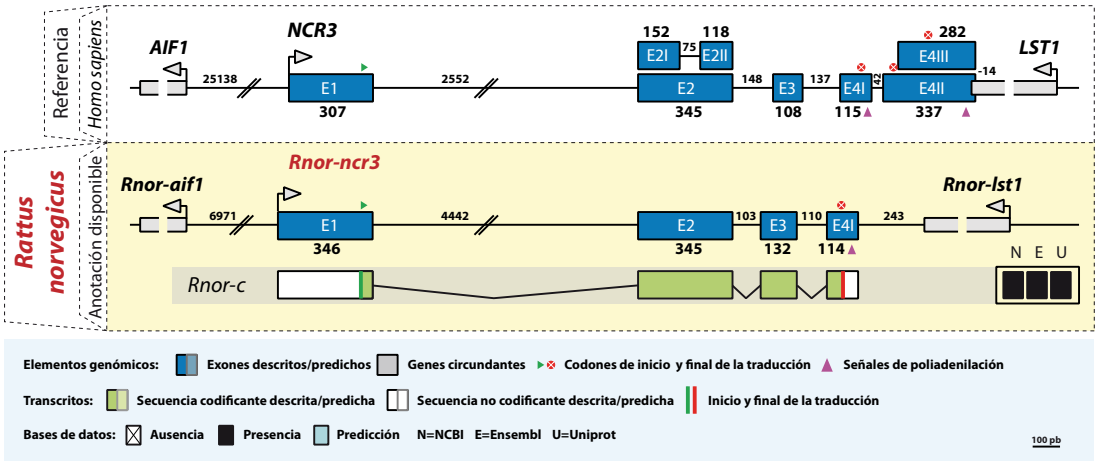


Figura 4-11. Anotación del gen *Rnor-ncr3*. En la parte superior se muestra la estrucutura exónica de *Homo sapiens*, utilizada como referencia. En la parte inferior se muestra la anotaciones disponibles para *Rattus norvegicus.*

reduce la longitud de la cola citoplasmática codificada en este exón con respecto a humano. Finalmente, se puede observar la presencia en la secuencia del exón 4 de un sitio de poliadenilación en una posición similar a lo observado en humano (Figura 4-11).

4.1.11 Mus musculus (*Mmus-ncr3*)

Estudios previos en esta especie demostraron que el gen *Mmus-ncr3* es un pseudogen en 12 cepas de ratón de laboratorio por la presencia de dos codones prematuros de terminación de la traducción en el extremo 5' del exón 2 (Hollyoake et al., 2005). Por el contrario, en *Mus caroli*, una especie de ratón de campo, la presencia de varios polimorfismos revierten estos codones de terminación, habiéndose detectado dos mensajeros potencialmente codificantes del gen *Mcaro-ncr3* (Figura 4-12, *Mus caroli*) (Hollyoake et al., 2005). Sin embargo, en la bibliografía existen algunas citas sobre la detección de mensajeros procedentes de este gen en *Mus musculus* mediante Northern Blot (Sivakamasundari et al., 2000), por lo que se decidió incluir esta especie en el presente estudio para un nuevo análisis. El gen *Mmus-ncr3* está presente en las principales bases de datos, sin embargo, la anotación sólo cubre los tres primeros exones, equivalentes a los exones 1, 2 y 3 en humano (Figura 4.12). El estudio comparativo con las secuencias disponibles de *Rattus norvegicus* permitió prolongar el exón 1, que sólo estaba descrito desde el ATG iniciador teórico (Figura 4.12). El exón 2 anotado del gen *Mmus-ncr3* es tres nucleótidos más corto que en el resto de las especies analizadas, sin embargo, las secuencia del sitio donador de splicing de este exón está tres nucleótidos más abajo en la secuencia genómico, por lo que se prolongó la anotación hasta este punto, siendo consistente con lo descrito en *Mus caroli*. El exón 3 anotado, de 139 pb, no presenta las secuencias consenso para lo sitios aceptor y donador de splicing. La comparación con el exón 3 de *Mcaro-ncr3* permitió corregir este error, prolongando la anotación 2 nucleótidos en el extremo 5' y reduciendo 15 nucleótidos en el extremo 3' (Figura 4.12). Como resultado el exón 3 ha quedado reducido a 126 pb, mayor que en humano, pero consistente con *Mus caroli* y similar a los descrito en *Rattus norvegicus* (Figura 4.12). El exón 4, no anotado en esta especie, se predijo por homología con la secuencia del exón 4 descrito en *Mus caroli*. Como en éste, se puede determinar la posición teórica del codón de parada de la traducción en la posición +53 de este exón, similar a lo descrito en rata (Figura 4.12). La búsqueda de secuencias de poliadenilación en este

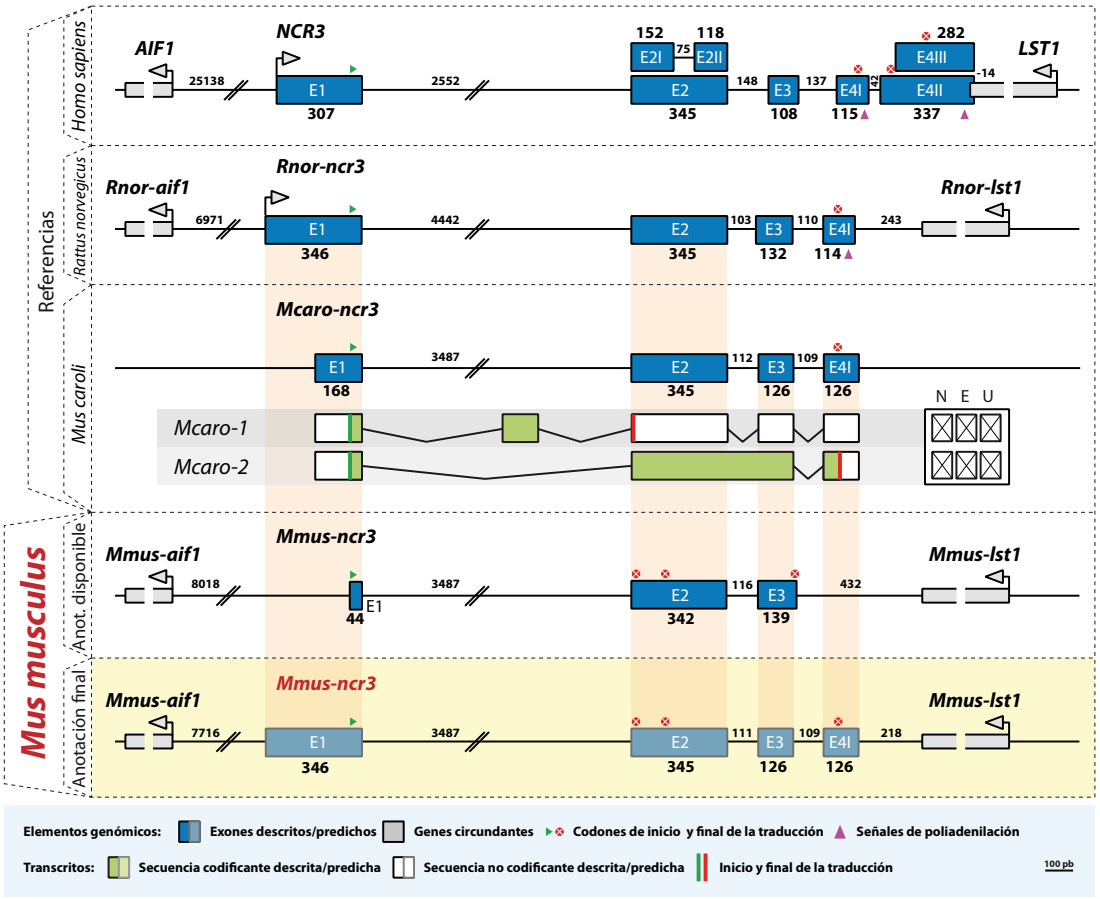


Figura 4-12. Anotación del gen *Mmus-ncr3*. En la parte superior se muestran las especies utilizadas como referencia para establecer la anotación final que se muestra debajo. Las proyecciones, sombreados naranjas, indican el uso de los exones de referencia para la anotación final de *Mus musculus*.

exón no dieron resultado, sin embargo, en la misma posición relativa que en rata se localiza el sitio de poliadenilación, se puede observar en *Mus musculus* un vestigio de dicha secuencia.

4.1.12 Bos taurus (*Btau-ncr3*)

En las principales bases de datos se pueden encontrar dos mensajeros anotados del gen *Btau-ncr3* (*Btau-c1* y *Btau-c2*), basados en secuencias depositadas, cuya traducción teórica daría como resultado una isoforma equivalente a la isoforma *C* humana, aunque estos mensajeros difieren en sus exones iniciales (Figura 4-13). El mensajero anotado en Ensembl (*Btau-c2*) muestra el exón 1 divido en dos fragmentos o exones, ambos localizados en la región equivalente al exón 1 humano (Figura 4-13). El primer fragmento es considerado como UTR, mientras que en el segundo, considerado como el exón 1, se localiza el ATG iniciador en una posición relativa

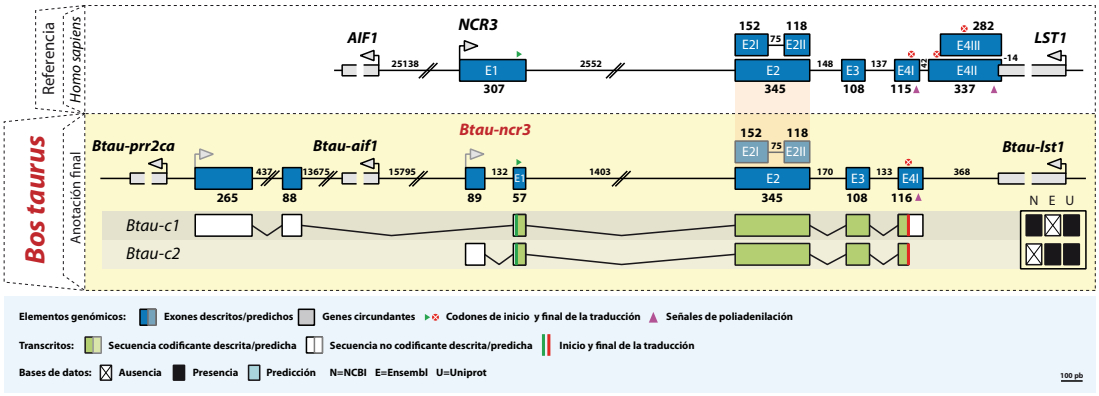


Figura 4-13. Anotación del gen *Btau-ncr3*. Anotación del gen *Mmul-ncr3*. En la parte superior se muestra la estructura exónica de *Homo sapiens*, utilizada como referencia. En la parte inferior se muestra la anotación final. Las proyecciones, sombreados naranjas, indican el uso de los exones de referencia para la anotación final de *Bos taurus*.

equivalente a la detectada en el resto de las especies aquí analizadas. El transcrito descrito en NCBI (*Btau-c1*) presenta este segundo exón descrito en Ensembl, precedido de 2 exones adicionales situados ~30 kpb por encima de éste, próximos al gen Btau-prr2ca y cerca de una isla CpG, quedando el gen *Btau-aif1* en el intrón entre estos exones y el exón 1 (Figura 4-13). El resto de exones descritos en ambas bases de datos son similares a los presentes en humano y otros primates, no existiendo evidencia alguna de la presencia de los exones 4II y 4III en esta especie. El análisis de la secuencia del exón 2 muestras la existencia de potenciales sitios de splicing internos en posiciones equivalentes a las encontradas en humano (Figura 4.13, anotación final). Los exones 3 y 4, presentan una longitud similar a los homólogos humanos y de otros primates, estando el codón de terminación de la traducción del exón 4 localizado en la posición +47, similar a lo detectado en el género *Pongo*, en *Papio cynocephalus* y en *Colobus guereza*. Finalmente, cabe resaltar que el sitio de poliadenilación se encuentra en una posición equivalente al encontrado en humano para el exón 4I (Figura 4.13).

4.1.13 Sus scrofa (*Sscro-ncr3*)

Al inicio de este trabajo el gen *Sscro-ncr3* no había sido descrito, ni se disponía de la secuencia del genoma de cerdo. Por ello se procedió a la búsqueda de secuencias mediante Blast, usando la secuencia codificante de la variante *Btau-c* (*Btau-c1* o *Btau-c2*) como referencia por su proximidad, lo que permitió identificar un cósmido procedente de *Sus scrofa* (BX548169) que contenía la secuencia genómica completa del gen *Sscro-ncr3* (Figura 4-14, cósmido Sus). La homología entre las secuencias

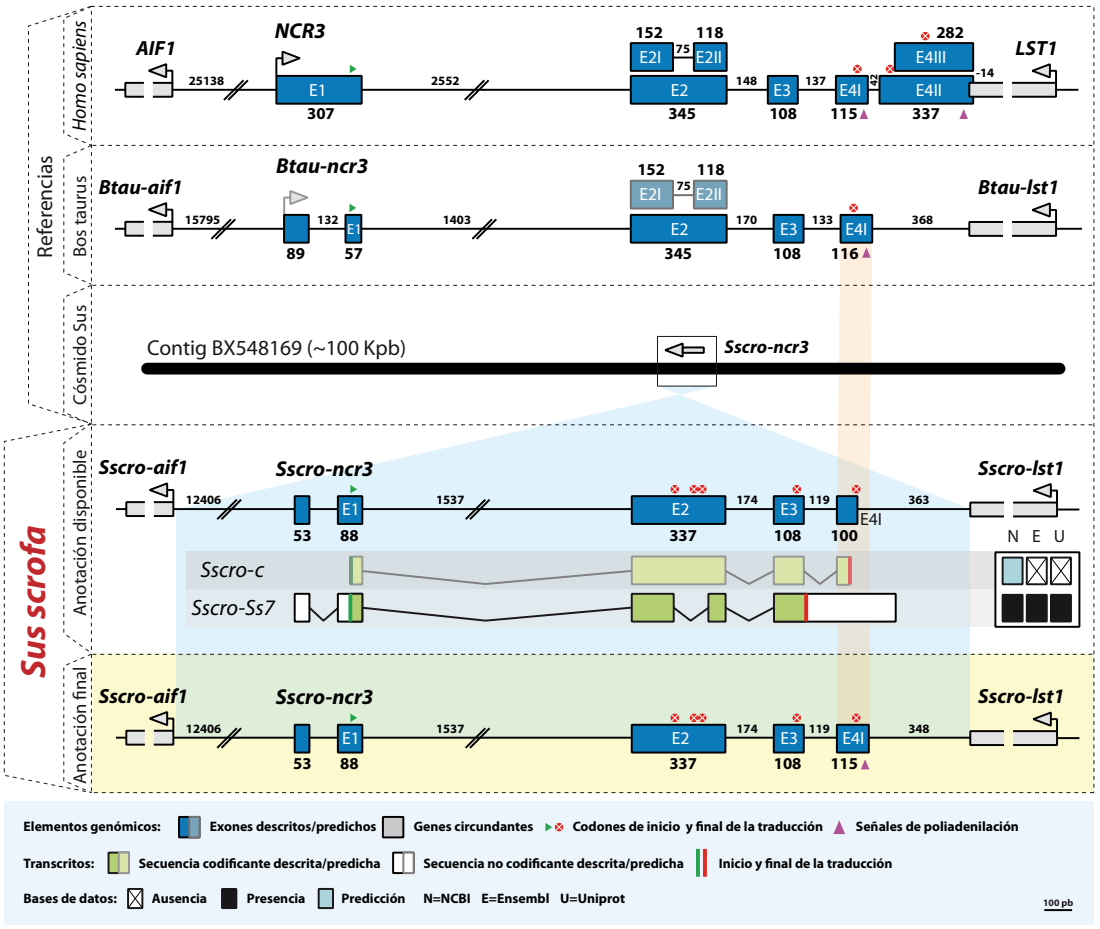


Figura 4-14. Anotación del gen *Sscro-ncr3*. En la parte superior se muestran las especies utilizadas como referencia para establecer la anotación final que se muestra debajo junto a la anotación disponible para *Sus scrofa*. Las proyecciones, sombreados naranjas y azules, indican el uso de los exones y secuencias de referencia respectivamente, para la anotación final de *Sus scrofa*.

de *Sus scrofa* y *Bos taurus* permitió la obtención de una anotación preliminar, posteriormente confirmada por la anotación disponible en la actualidad en las bases de datos (Tabla 4-1 y Figura 4-14, anotación disponible). El ATG iniciador se localiza en el exón 1, en la misma posición relativa que en el resto de especies analizadas. En el exón 2, de tan sólo 337 pb, la fase de lectura análoga a la detectada en otras especies se ve alterada por la presencia de dos deleciones. La primera, de tan sólo un nucleótido en la posición +68 (con respecto al inicio del exón), produce un cambio en la pauta de lectura que provoca la aparición de un codón de terminación de la traducción en la posición +155 del propio exón, alterando toda la secuencia codificante teórica. La segunda, de siete nucleótidos, localizada en la posición +196, introduce un nuevo cambio en la fase de lectura, que no restaura la fase observada en otras especies, y produce la aparición de dos nuevos codones de terminación

en el exón 2, en las posiciones +224 y +254. Sendos cambios en la fase de lectura afectan también a los exones 3, en el que aparece otro codón de terminación, y 4 que presentaría el codón de terminación en la posición +70 (Figura 4-14). En conjunto, la presencia de todos estos codones prematuros de terminación de la traducción sugieren que el gen *Sscro-ncr3* podría ser un pseudogen, como su homólogo en *Mus musculus*. Sin embargo, podemos detectar la secuencia consenso del sitio de poliadenilación en el exón 4, que en el caso de ratón parece degenerada (Figura 4-14, anotación final). En NCBI se puede encontrar un mensajero anotado por predicción que contiene los 4 exones descritos en este trabajo (*Sscro-c)* (Figura 4-14, anotación disponible). Sorprendentemente, la secuencia del mensajero predicha en NCBI no presenta estos codones de parada, sino que muestra algunas correcciones como la inserción de dos nucleótidos indeterminados ('N') en el exón 2 en las posiciones +68 y +196 para restaurar la fase de lectura. Al hacer esto, aparecen dos nuevos codones de terminación en la segunda mitad del exón 2, que son corregidos en la traducción teórica, de modo que el único codón de terminación de la traducción se localice en la posición +47 del exón 4, similar a lo descrito en *Bos taurus, Colobus guereza, Papio cynocephalus* y en el género *Pongo*. Por otro lado, en Ensembl, encontramos anotados dos mensajeros, siendo uno de ellos el mismo que el presente en NCBI, sin las correcciones introducidas en ésta, de manera que el mensajero es considerado como no codificante (Tabla 4-1). Además, en Ensembl, encontramos otro transcrito anotado (denominado aquí *Sscro-ss7*), basado en dos secuencias depositas, una parcial (HQ658981.1) y otra del cDNA completo (EU282355.1). El exón 1 de este mensajero está divido en dos fragmentos, de manera análoga a lo observado en el mensajero *Btau-c2*, siendo el primer fragmento considerado UTR, y estando localizado el ATG iniciador en el segundo fragmento (Figura 4-14, anotación disponible). El exón 2, también divido en dos fragmentos por un evento de splicing interno, diferente del detectado en humano y otras especies, carece de la porción del exón 2 donde se detectan los tres codones de parada (Figura 4-14). Además, este evento de splicing, provoca la restauración de la pauta de lectura, de modo que la secuencia teóricamente codificada por la segunda parte del exón 2 y el exón 3 es similar a la observada en otras especies. Los exones 3 y 4 aparecen como un único exón en el mensajero *Sscro-ss7*, debido a la retención del intrón entre ambos, encontrándose un codón de terminación de la traducción en fase con el ATG iniciador al principio de este intrón (Figura 4-14, anotación disponible).

4.1.14 Análisis de la conservación y el entorno genómico del gen NCR3

El análisis de la información disponible en las distintas bases de datos de las especies bajo estudio, así como la secuenciación del DNA genómico de algunas de éstas ha permitido llevar a cabo el estudio de la conservación de estas secuencias, así como estudiar la estructura exónica del gen *NCR3* a lo largo de la evolución (Figura 4-15). Globalmente se puede observar que la mayor conservación de secuencia se detecta en el grupo de los primates, encontrado niveles elevados de conservación incluso en las secuencias intrónicas (Figura 4-15). A medida que aumenta la distancia evolutiva con respecto a humano se detecta una concentración de la conservación

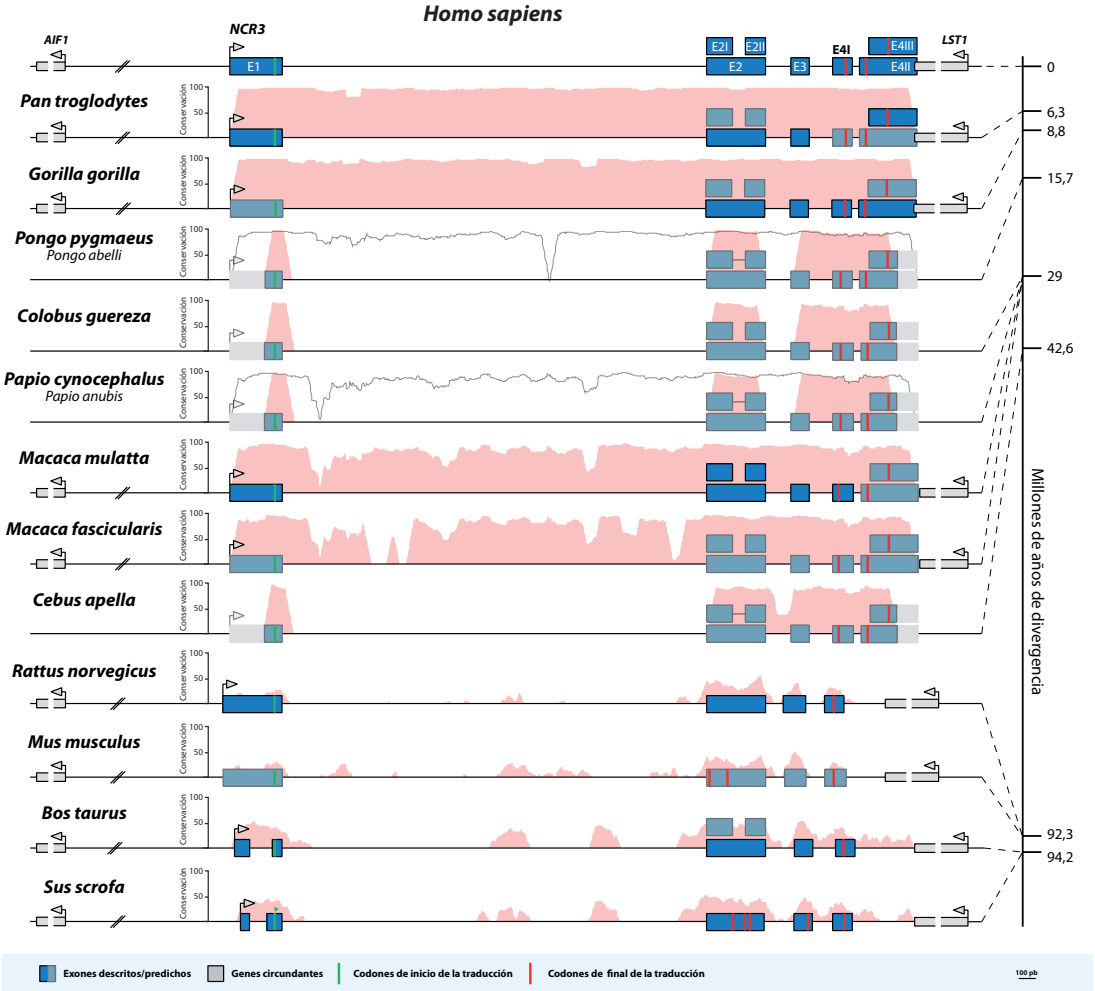


Figura 4-15. Análisis de la conservación de la secuencia genómica del gen *NCR3* en las especies incluidas en este trabajo. El histograma rojo sobre la estructura exónica de cada especie representa la conservación (%) con respecto a la secuencia de *Homo sapiens*. En *Pongo pygmaeus*, *Colobus guereza*, *Papio cynocephalus* y *Cebus apella* el análisis está restringido a las regiones de secuencia conocida. En *P. pygmaeus* y *P. cynocephalus* se han incluido los histogramas (línea negra) correspondiente con el análisis de la conservación de la secuencia en *Pongo abelli* y *Papio anubis* respectivamente, como referencia por su proximidad.

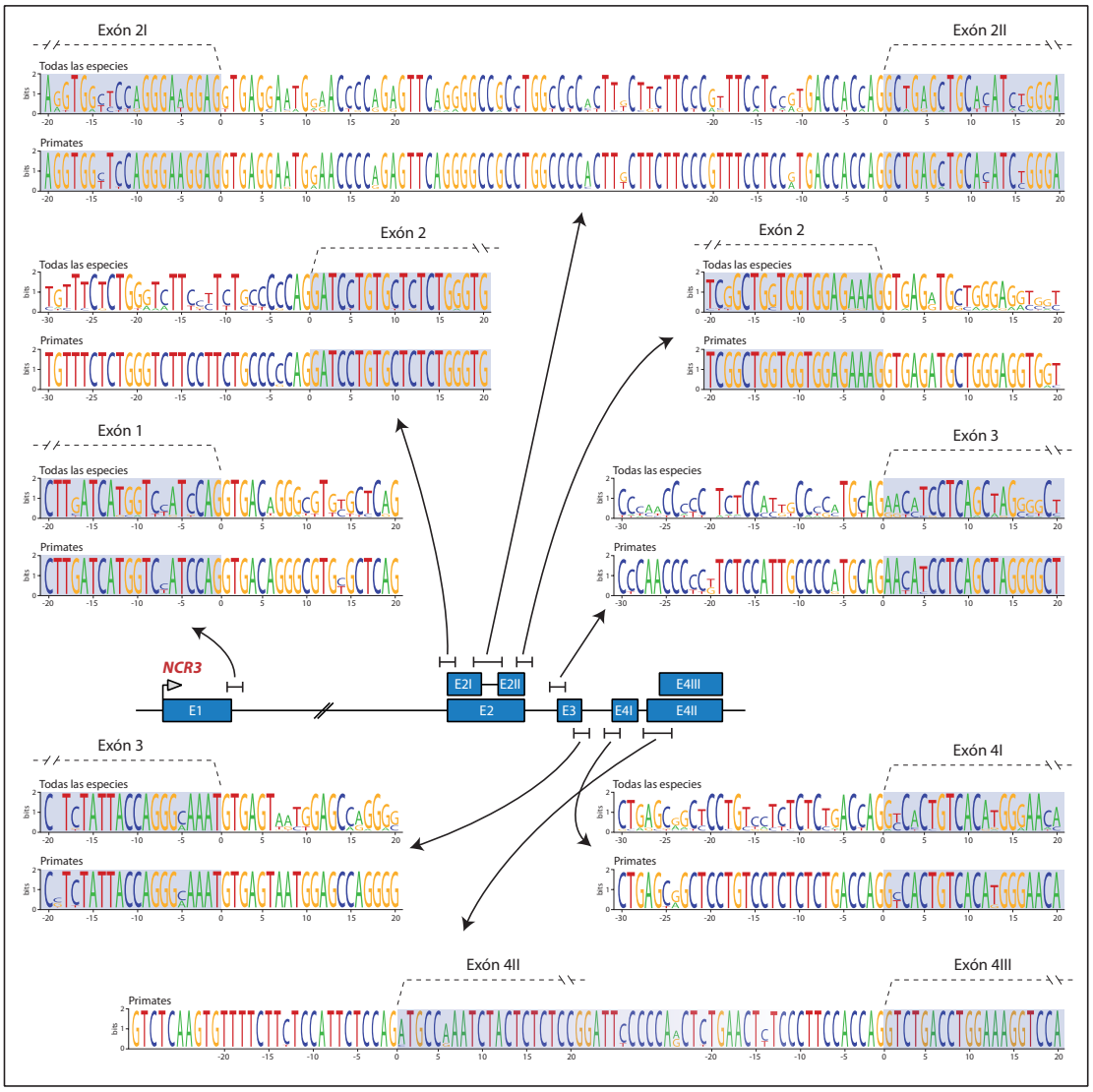


Figura 4-16. Conservación de los sitios de splicing. En el centro se muestra la estructura exónica genérica del gen *NCR3* en la que se indican con barras las fronteras exón intrón. En cada una de éstas se alinearon con ClustalW las secuencias de todas las especies incluidas en este trabajo, tras lo que se obtuvo la secuencia consenso con la aplicación Weblogo (secuencias consensos superiores). Además, se realizó el mismo análisis incluyendo sólo las secuencias del grupo de los primates (secuencias consenso inferiores).

en las secuencias exónicas y la frontera exón-intrón (Figura 4-15). En *Bos taurus*, *Sus scrofa*, *Mus musculus* y *Rattus norvegicus* no existen evidencias que permitieran definir los exones 4II y 4III, sin embargo, resulta llamativa la presencia de cierta conservación de secuencia en la región intergénica entre el gen *NCR3* y *LST1* en ungulados (*Bos taurus* y *Sus scrofa*) donde se localizan dichos exones en primates (Figura 4-15).

Además de la secuencia, la estructura exónica también ha sido conservada a lo largo de la evolución (Figura 4-15). El exón 1, y sobre todo la región codificante de éste, ha permanecido inalterado, salvo en el grupo formado por *Sus scrofa* y *Bos taurus*, en los que parece haber desarrollado un splicing interno. Además en *Bos taurus* existen evidencias de la inclusión de exones anteriores a éste, aunque muy alejados en secuencias, más de 30 kpb, que pese a no modificar la secuencia codificante podría tener consecuencias sobre la estabilidad y vida media del mensajero. El exón 2, que codifica el dominio inmunoglobulina característico de este gen, ha sido bien conservado, manteniendo un tamaño constante, excepto en *Sus scrofa* por la presencia de dos deleciones que producen cambios en la fase de lectura. En la mayoría de las especies analizadas se puede detectar la presencia de sitios potenciales de splicing interno en el exón 2 que podrían generar un cambio en el dominio inmunoglobulina codificado como el descrito en humano (Figura 4-15). El exón 3, que codifica principalmente el dominio transmembrana, también se ha conservado estructuralmente, sin embargo, éste muestra un incrementado de tamaño en el linaje de los roedores. Pese a que su tamaño ha fluctuado sólo ligeramente, el exón 4, muestra una gran conservación, aunque ligeramente inferior al resto de exones. Lo más interesante de este exón es la modificación progresiva de la posición del codón de terminación de la traducción, en especial en el grupo de los primates. En el género *Macaca*, en *Cebus apella* y en *Papio anubis*, se localiza a tan sólo 35 pb del inicio del exón, mientras que en *Colobus guereza*, *Papio cynocephalus* y en el género *Pongo* está desplazado hacia el 3', en la posición +47 con respecto al inicio del exón, lo que incrementa el tamaño de la cola citoplasmática de las proteínas teóricas. En *Gorilla gorilla*, *Pan troglodytes* y *en Homo sapiens* el desplazamiento del códon de terminaición hacia el 3' del exón es aún mayor, encontrándose en la posición +77. Finalmente, pese a ser exclusivo de primates, los exónes 4II y 4III muestra una diferencia importante en cuanto a la presencia de un sitio de poliadenilación. Éste sólo está presente en primates superiores, género *Pongo*, *Gorilla gorilla*, *Pan troglodytes* y *Homo sapiens*, mientras que en el resto de primates analizados la presencia de polimorfismos altera esta secuencia consenso, lo que podría tener consecuencias en la expresión de variantes que portasen dichos exones. En análisis detallado de las fronteras exón-intrón permite observar una alta conservación de las secuencias de los sitios 5' donador y 3' aceptor (Figura 4-16), con la única excepción del sitio 5' donador del exón 2I, que en roedores presenta un polimorfismo (G/T).

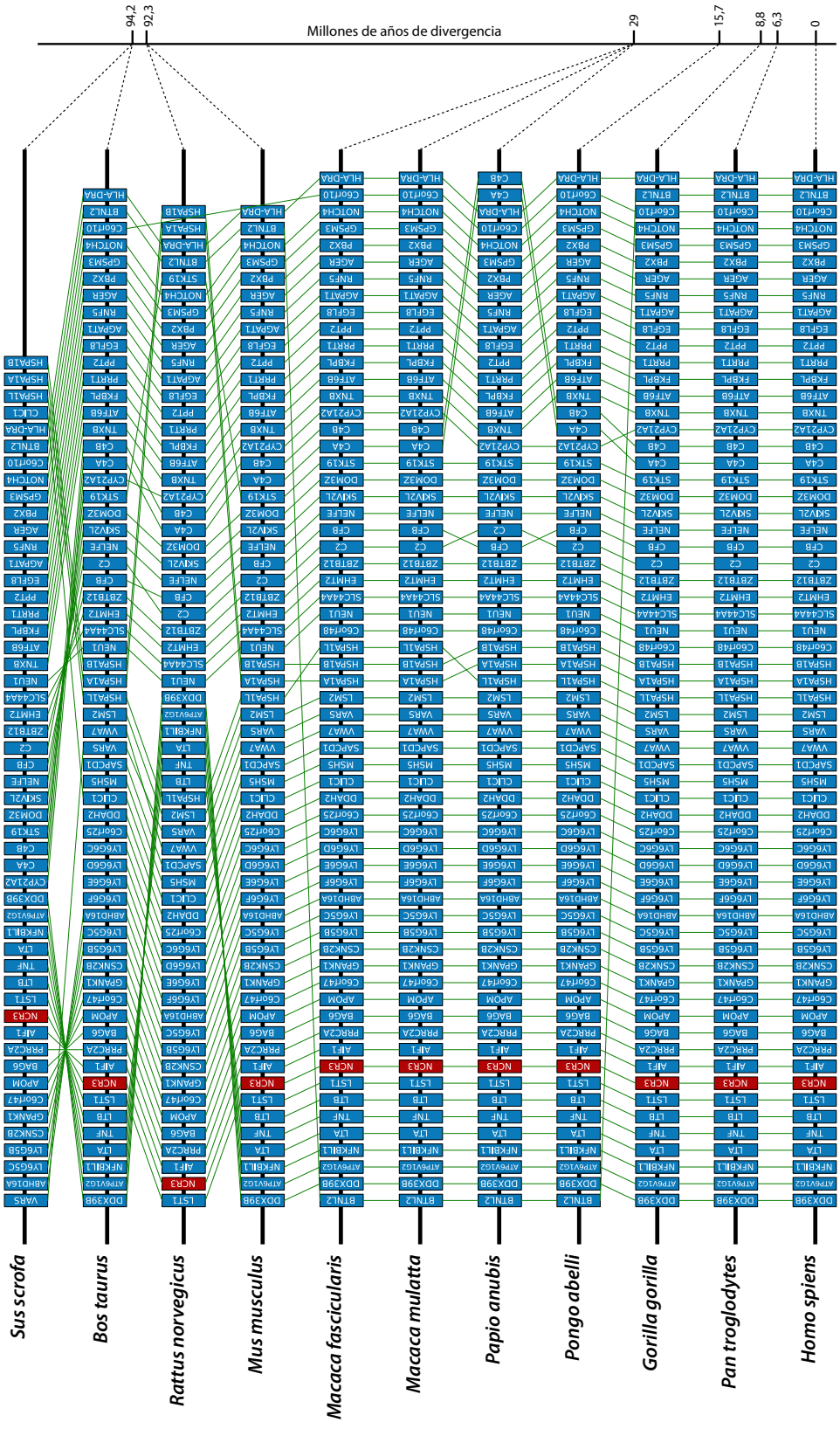


Figura 4-17. Análisis de la sintenia de la región de MHC de clase III en las especies incluidas en este estudio. Cada una de las cajas azules representan un gen, y las líneas verdes indican la conexión de los genes entre especies. El gen *NCR3* se ha destacada en rojo.

El entorno genómico del gen *NCR3* también muestra una alta conservación. En todas las especies analizadas de las que existe un genoma de referencia se ha podido comprobar la presencia de los genes *AIF1* y *LST1*, a ambos lados del gen NCR3, y en orientación opuesta a éste. La distancia entre *AIF1* y *NCR3* es muy similar en todos los primates (~25-30 kpb), mientras que en roedores y en ungulados esta distancia es considerablemente menor (~7-8 kpb en roedores y ~12-15 kpb en ungulados). Por otro lado, la distancia entre *NCR3* y *LST1* es muy pequeña en primates, llegando a solapar parcialmente con los exones 4II y 4III, mientras que en roedores y ungulados la distancia es superior (~200-370 pb), debido a la ausencia de estos exones en estas especies (Figura 4-15).

En humano, el gen *NCR3* se localiza en la región del MHC de clase III, que muestra una alta conservación entre humano, ratón, rata y cerdo (Hurt et al., 2004; Peelman et al., 1996; Renard et al., 2006; Walter et al., 2002; Xie et al., 2003). Por ello se decidió extender el estudio del entorno genómico a la región del MHC de clase III utilizando las secuencias de genes incluidos en dicha región en humano como referencia para la búsqueda de los genes homólogos en el resto de las especies de las que se dispone de la secuencia genómica completa. Gracias a este procedimiento se ha podido comprobar que la mayoría de los genes presentes en esta región en humano se encuentran conservados en regiones homólogas en las demás especies (Figura 4-17), detectándose en todas ellas una gran conservación global de la sintenia de la región, exceptuando *Rattus norvegicus* y *Sus scrofa*, en las que existe algunos reordenamientos importantes, aunque la mayoría de los genes siguen estando agrupados en la misma región.

Adicionalmente, se decidió estudiar la densidad génica de la región del MHC de clase III , ya que es una de las peculiaridades que muestra esta región en humano (The MHC sequencing consortium, 1999). Para esta tarea se han utilizado las anotaciones disponibles y se ha analizado en cada especie la densidad génica global, así como la densidad génica del cromosoma en el que se localiza la región del MHC de clase III de cada una de las especies. Globalmente se puede considerar que la región del MHC de clase III es una de las regiones con mayor densidad génica en todas las especies analizadas (Figura 4-18). Sólo en ungulados se observa una densidad menor de la región, aunque superior a la media del genoma en estas especies. Este fenómeno es aún más claro cuando se limita el análisis al cromosoma en el que se localiza esta

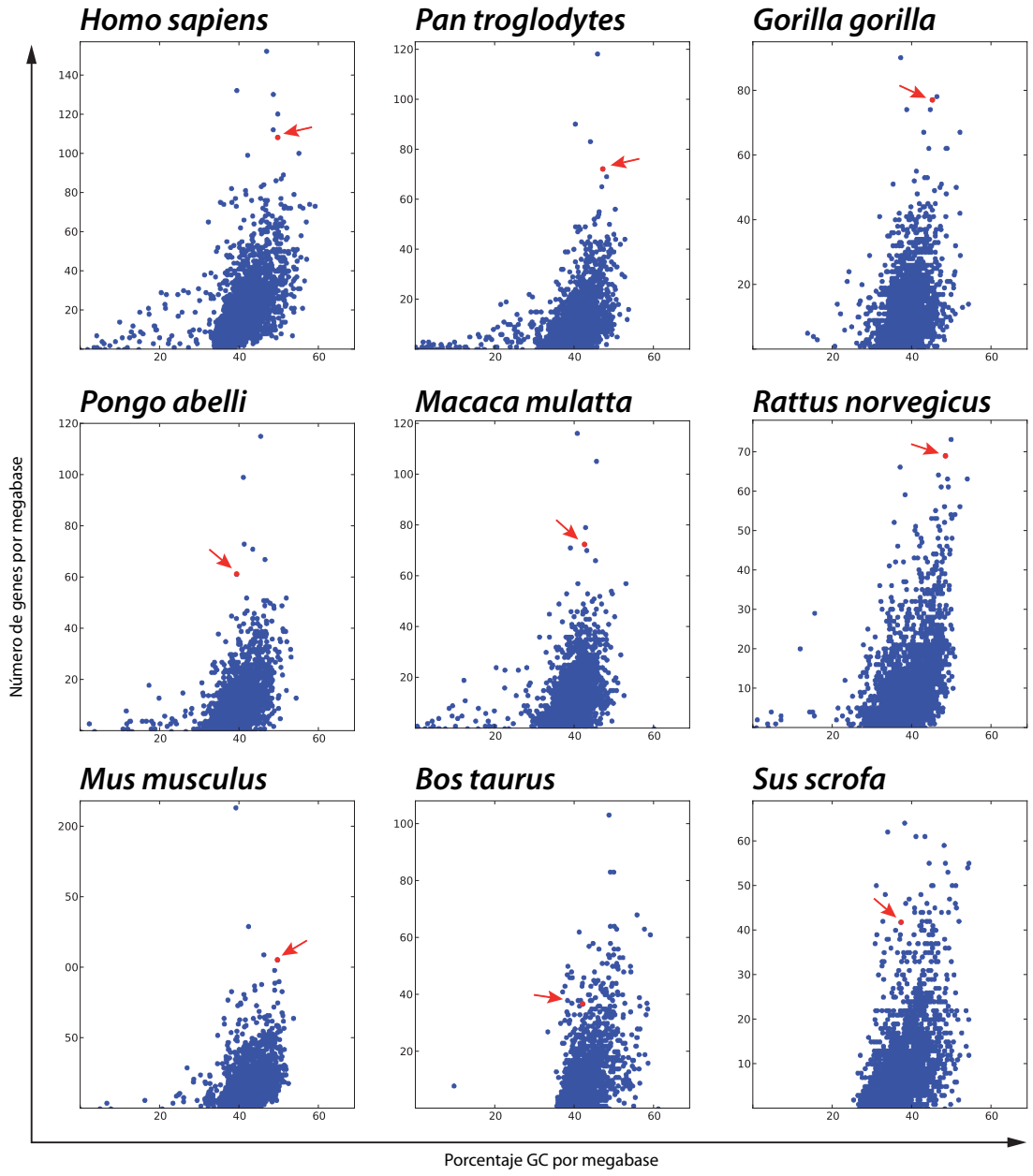


Figura 4-18. Análisis de la densidad génica de la región de MHC de clase III. En las especies en las que se dispuso de la secuencia completa de su genoma, así como de la anotación de éste, se calculó la densidad génica en ventanas de una megabase, así como el contenido en GC.

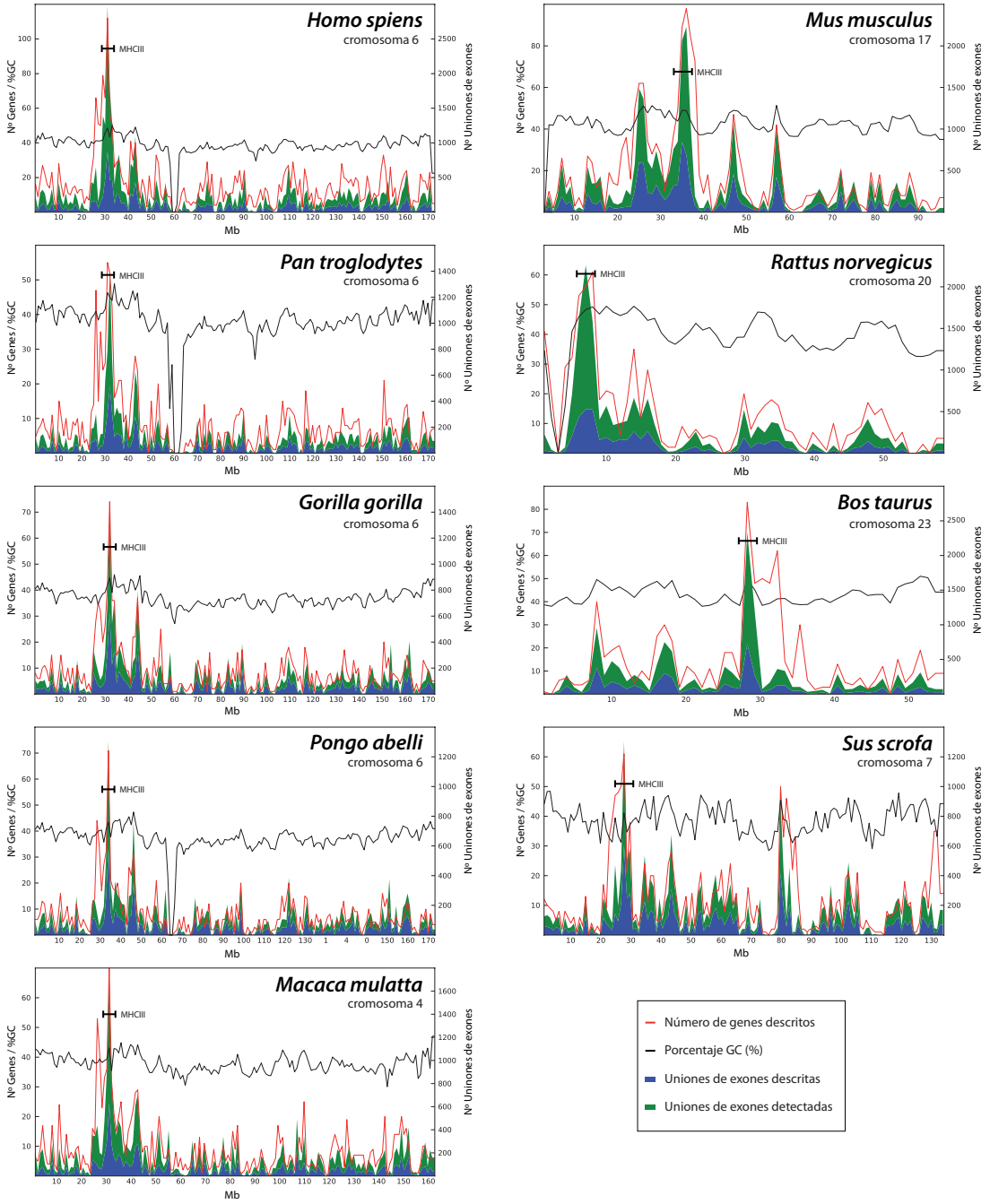


Figura 4-19. Análisis de la densidad génica local. En cada especie se estudió la densidad génica (línea roja) del cromosoma en el que se localizada la región de MHC de clase III (indicada con una barra negra). Además se muestra el contenido en GC (línea negra), las uniones de exones descritas (histograma azul) y las uniones de exones detectadas mediante RNA-seq (histograma verde).

región en cada una de las especies (Figura 4-19). En el perfil de densidad génica de los cromosomas de cada especie la región del MHC de clase III se hace evidente a simple vista, superando en varias veces la densidad media del cromosoma en todos los casos (Figura 4-19).

Finalmente, se decidió comparar la anotación disponible en las bases de datos con la información global obtenida del análisis de secuencias de RNA-seq realizados en este trabajo. Para ello se extrajo la información de las uniones de exones descritas de las anotaciones disponibles para cada especie y se comparó con los datos de uniones de exones detectadas en el análisis de los datos de RNA-seq disponibles (Figura 4-19 y Tabla S2). De manera general se puede afirmar que el conocimiento actual sobre el splicing alternativo aún está lejos de ser completo, puesto que en todas las especies analizadas mediante RNA-seq se detectan de decenas a centenas de miles de uniones de exones nuevas. La disponibilidad de datos de RNA-seq no es equitativa entre las especies analizadas en este trabajo, existiendo mucho más datos de humano y ratón que en el resto de las especies, sin embargo, el número de uniones de exones detectadas incrementa el número total de uniones de exones totales entre 1,23 y 1,74 veces en todas las especies (Figura 4-20). Si se analiza en detalle esta información, se puede observar que el número de uniones de exones detectadas en el cromosoma en el que encontramos la región del MHC de clase III en cada especie es superior al número de uniones de exones descritas en toda su longitud (Figura 4-19). Del mismo modo, la región del MHC de clase III muestra un déficit en el número de uniones de exones descritas con respecto a las que se pueden detectarse mediante RNA-seq (Figura 4-19). Conviene resaltar que las uniones de exones detectadas mediante RNA-seq han sido filtradas, sólo considerándose válidas aquellas soportadas por al menos cinco lecturas, lo que ha reducido el número de uniones de exones detectadas considerablemente. Por otro lado, la limitación en el número de datos de RNA-seq y el limitado número de tipo celulares o tejidos de los que preceden dichos datos permite especular que el número de uniones de exones y de variantes de splicing será aún mucho mayor del aquí expuesto, poniendo de manifiesto el gran déficit existente en el conocimiento del número real de variantes de splicing.

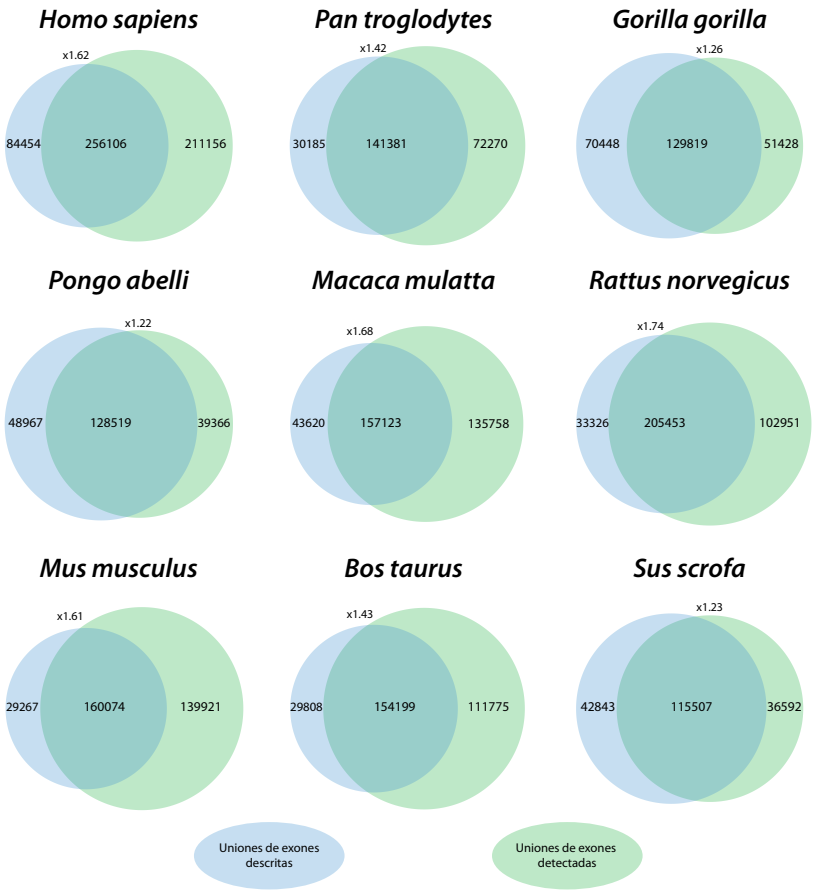


Figura 4-20. Comparación del número de uniones de exones descritas en las anotaciones disponibles (azul) y las detectadas mediante RNA-seq (verde). Los números sobre los diagramas de Venn indican el incremento en el número total de uniones de exones diferentes descritas y detectadas con respecto a las descritas.

4.2 Análisis de las variantes de splicing del gen *NCR3* en diferentes especies de mamíferos

Una vez conocida las secuencia del gen *NCR3* y su anotación se procedió al análisis de las variantes de splicing expresadas a partir de éste en cada especies. Para ello se han utilizado dos métodos complementarios. Por una lado, se ha realizado un análisis detallado de las variantes de splicing expresadas en los diferentes tejidos de cada especie mediante PCR anidada, y adicionalmente hemos estudiado las uniones de exones que podemos detectar mediante el análisis de secuencias de RNA-seq disponibles. La unión de los resultados obtenidos con ambos métodos permite determinar en mayor detalle el grado de splicing alternativo que muestra el gen *NCR3* en cada una de las especies y su análisis comparativo.

La división de este apartado y el análisis de las variantes de splicing desarrollado en el grupo de los primates es meramente práctica, dado que la metodología utilizada es similar. Sin embargo, la disponibilidad de muestras de diferentes tejidos en el grupo de especies que hemos agrupado bajo el epígrafe mamíferos es muy diversa (Tabla 3-3, materiales y métodos), mientras que en el grupo de los primates sólo se dispuso de muestras de sangre. Por ello, se consideró conveniente la separación en dos apartados para una mejor compresión de los resultados.

4.2.1 *Homo sapiens*

Los resultados obtenidos mediante PCR anidada en los distintos tejidos y líneas celulares humanas permitieron confirmar la existencia de los nueve mensajeros descritos previamente por Neville *et al.* (Neville and Campbell, 1999) (Figura 4-21), no encontrándose transcritos adicionales. Se detectó expresión del gen *NCR3* en todas las muestras analizadas, sin embargo, la expresión de las nueve variantes muestran un patrón diferencial. Así, mientras que en las líneas celulares Jurkat, Raji e YT se expresan todas las variantes (Figura 4-21), cabe destacar la falta de expresión de las variantes *C*, *F* y *NC3* en las líneas celulares U937 y HeLa (Figura 4-21A), la ausencia de la variante *NC3* en células K562 (Figura 4-21A) y la expresión exclusiva de las isoformas *B* y *D* en la línea celular U937 y la variante *B* en células HeLa (Figura 4-21B).

Los resultados obtenidos en los distintos tejidos humanos adultos, fetales y tumorales también evidencian la expresión de estos nueve mensajeros, existiendo igualmente

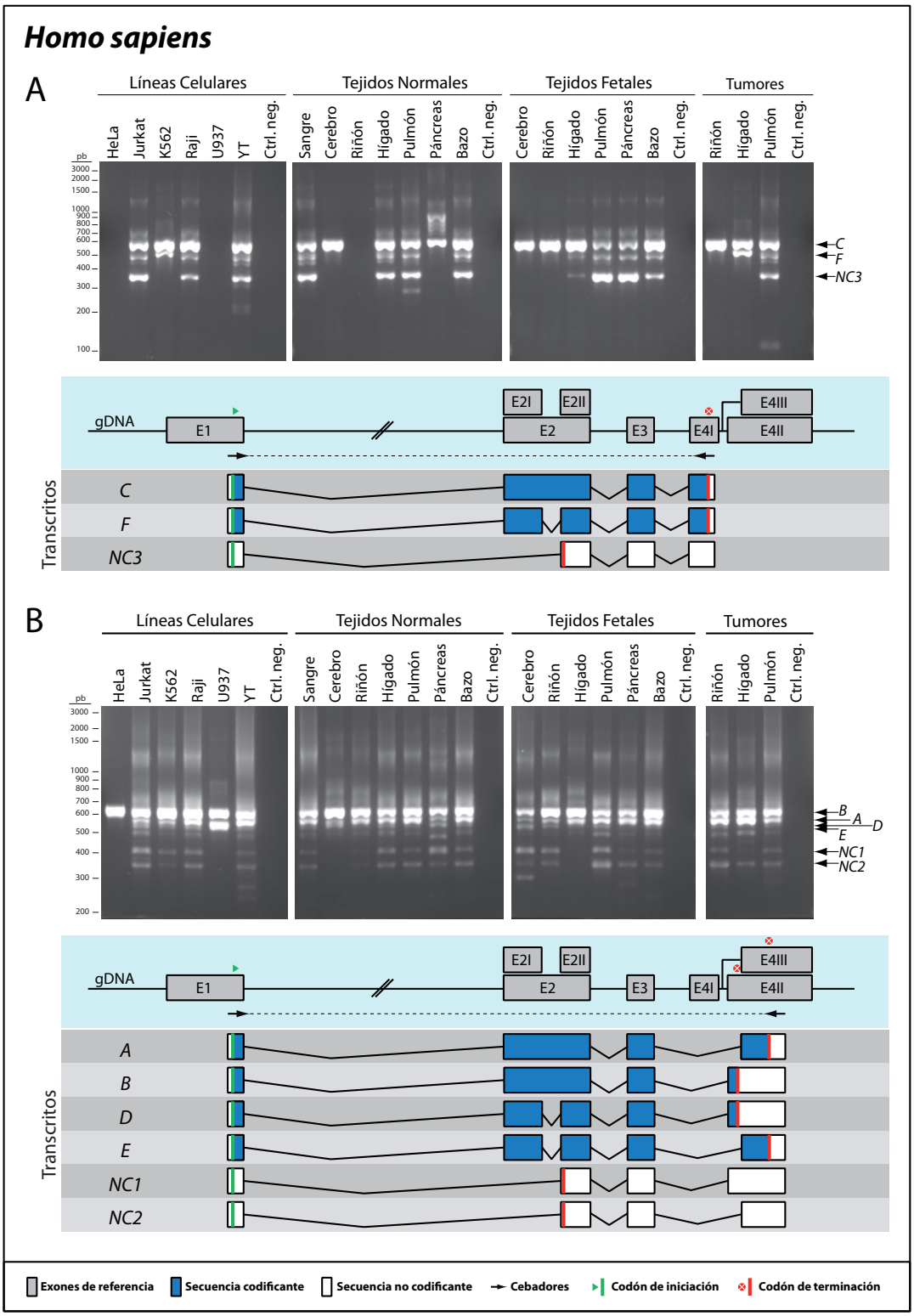


Figura 4-21. Análisis del AS del gen *NCR3* mediante PCR anidada. A) Amplificación de los mensajeros portadores del exón 4I. B) Amplificación de los mensajeros portadores de los exones 4II o 4III.

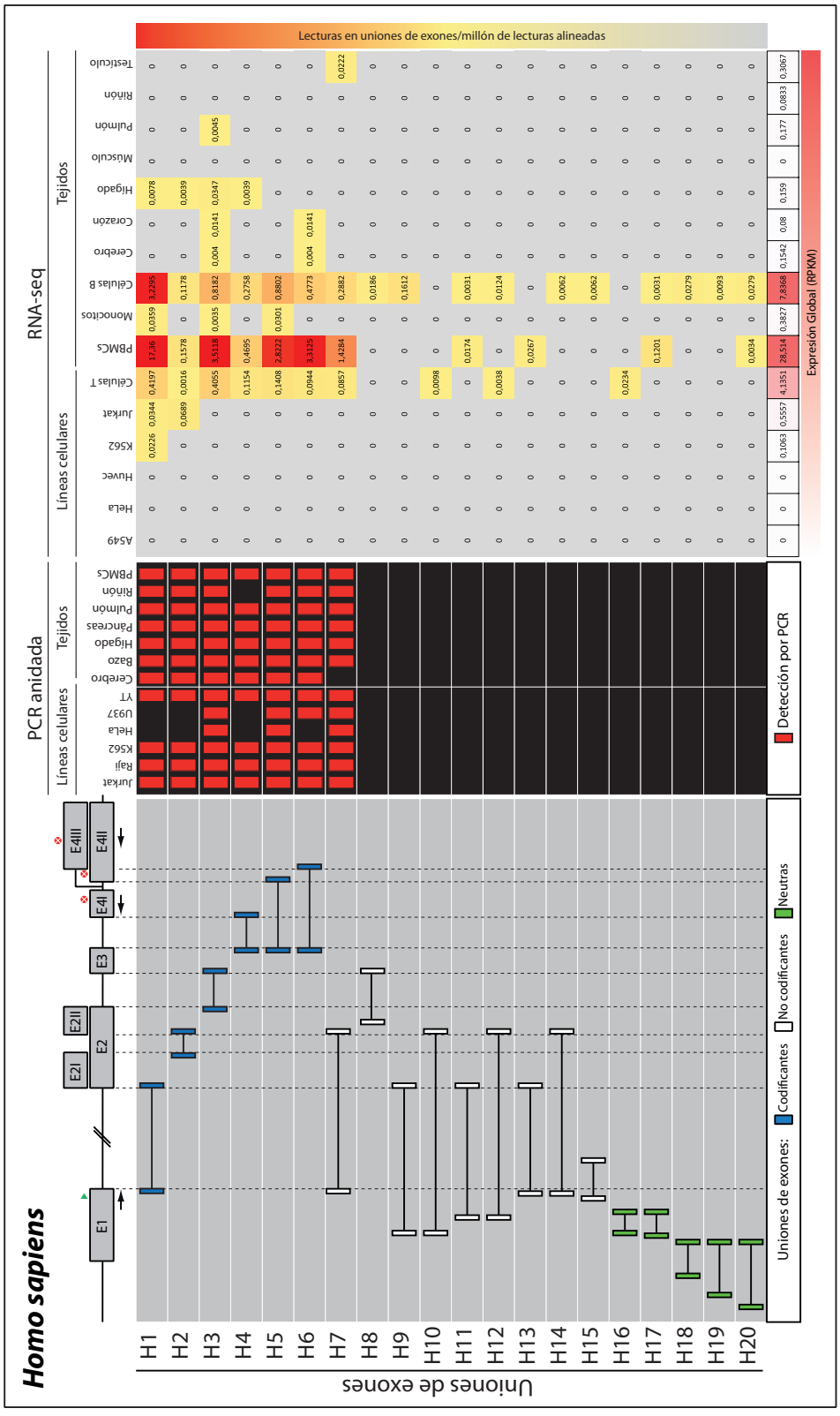


Figura 4-22. Análisis del AS del gen *NCR3* mediante RNA-seq. A la izquierda se representan esquemáticamente las uniones de exones detectadas mediante por PCR anidada y RNA-seq. En el panel central se muestra un resumen de la información obtenida mediante PCR anidada. En el panel derecho se muestra las uniones de exones detectadas en las secuencias de RNA-seq disponibles (Tabla S2) medidas como lecturas alineadas en cada unión de exones normalizadas por el número tal de lecturas alineadas.

un patrón de expresión diferencial. En PBMCs procedentes de sangre se puede observar la expresión de los nueve transcritos (Figura 4-21). En tejido de riñón adulto sólo se expresan las variantes *A*, *B*, *E* y ligeramente *NC1* y *NC2*, mientras que en el estadio fetal podemos observar la expresión de las variantes *A*, *B*, *C*, *D*, *NC1* y *NC2*. Además, en tejido de riñón tumoral, cabe destacar, un patrón de expresión más próximo al estadio fetal que el mostrado por el tejido adulto como demuestra la expresión de los mensajeros *A*, *B*, *C*, *E*, *NC1* y *NC2* (Figura 4-21). Asimismo, en los tejidos cerebral, páncreatico y hepático también pueden observarse diferencias llamativas en la expresión de algunas variantes entre el estadio fetal y adulto. En cerebro adulto se detecta expresión de las variantes *A*, *B* y *C*, mientras que en tejido fetal se observa además expresión de las variantes *D* y *NC1*. En hígado adulto se expresan todas las variantes, sin embardo, en tejido fetal destaca la ausencia de expresión de las variantes *D*, *E*, *NC1* y *NC2* (Figura 4-21). En páncreas adulto se expresan todas las variantes a excepción de las variantes *F* y *NC3*, que si se expresan en estadio fetal, mientras que no lo hacen las variantes *D* y *E*. Por el contrario, el tejido pulmonar se expresan todas las variantes, tanto en tejido adulto como en fetal y tumoral (Figura 4-21).

Adicionalmente al uso de la PCR anidada en la búsqueda de variantes de splicing hemos analizado las variantes splicing procedentes del gen *NCR3* mediante secuenciación masiva de RNA (RNA-seq) en una muestra de RNA procedente de sangre periférica (PBMCs) de un donante sano, así como en los datos disponibles en las bases de datos de RNA-seq (Tabla S2). De esta forma, pudimos detectar uniones de exones que combinadas conforman los nueve mensajeros amplificados por PCR anidada (uniones de exones H1 a H7, Figura 4-22), consideradas aquí como uniones de exones canónicas. Además, se detectaron nuevas uniones de exones, que proceden de variantes de splicing no identificadas hasta la fecha. Estas nuevas uniones de exones se clasificaron en 2 categorías: uniones de exones no codificantes (uniones H8 a H15, Figura 4-22) y uniones de exones neutras (uniones H16 a H20, Figura 4-22).

Las primeras procederían de mensajeros no codificantes por la aparición de codones prematuros de terminación de la traducción (uniones H8, H13 y H15) o por la aparente ausencia de codones de iniciación de la traducción (uniones H9 a H11) (Figura 4-22 y Tabla S3). También hemos incluido en este grupo la unión de exones H14, que

pese a mantener la fase de lectura, presenta la deleción de gran parte del exón 2 imposibilitando así el plegamiento del dominio inmunoglobulina codificado en éste y por tanto, generando un proteína potencialmente defectiva.

Por otro lado, las uniones de exones que hemos llamado neutras (uniones H16 a H20, Figura 4-22) son eventos de splicing en la región 5´ no traducida del exón 1 o entre en exón 1 y un potencial exón o exones anteriores no descritos (Figura 4-22). El análisis de las secuencias revela que no existe ningún otro ATG iniciador anterior al ya descrito en ninguno de los mensajeros teóricos que portasen estas uniones de exones ni combinaciones de éstas. Por lo tanto, la presencia de estas uniones de exones no altera la secuencia codificante de los mensajeros. Sin embargo, la presencia de tales uniones de exones podría tener implicaciones en la estabilidad de los mensajeros.

Analizando todas la uniones en las distintas muestras de RNA-seq disponibles, podemos ver un alto grado de detección de estas uniones de exones en células T, B y linfocitos procedentes de sangre periférica (PBMCs), donde la expresión global del gen es mayor, siendo las uniones de exones canónicas (H1 a H7) las más abundantes en todas las muestras (Figura 4-22). Destaca el alto número de uniones de exones que proceden de mensajeros no codificantes (uniones de exones H8 a H15), sin embargo, el número de lecturas por muestra y el número de muestras en las que detectamos tales uniones de exones es bajo (Figura 4-22 y Tabla S3) y por lo tanto, se trata de variantes poco abundantes. Del mismo modo, las uniones de exones neutras son poco frecuentes, de modo que sólo una pequeña proporción de los mensajeros maduros portará dichas uniones de exones (Figura 4-22). Si comparamos la expresión de las distintas uniones de exones se puede observar que las variantes *A*, *B* y *C* son mayoritarias (uniones H1 y H3), mientras que las variantes *D*, *E* y *F* son minoritarias dado la baja expresión de la unión de exones H2 característica de estos mensajeros, incluso menor que la unión de exones H7, característica de los mensajeros no codificantes *NC1*, *NC2* y *NC3* (Figura 4-22). En cuanto a las distintas colas citoplasmáticas existe cierta especificidad en función del linaje celular, así en células T las variantes *B* y *D* con el exón 4II (unión H5) son mayoritarias, seguidas de las variantes *C* y *F* con el exón 4I (unión H4) o *A* y *E* 4III (unión H6), en células B la variantes mayoritarias portan el exón 4II (*B* y *D,* unión H5), seguidas de las que presentan los exones 4III (*A* y *E* unión H6) o 4I (*C* y *F* unión H4)

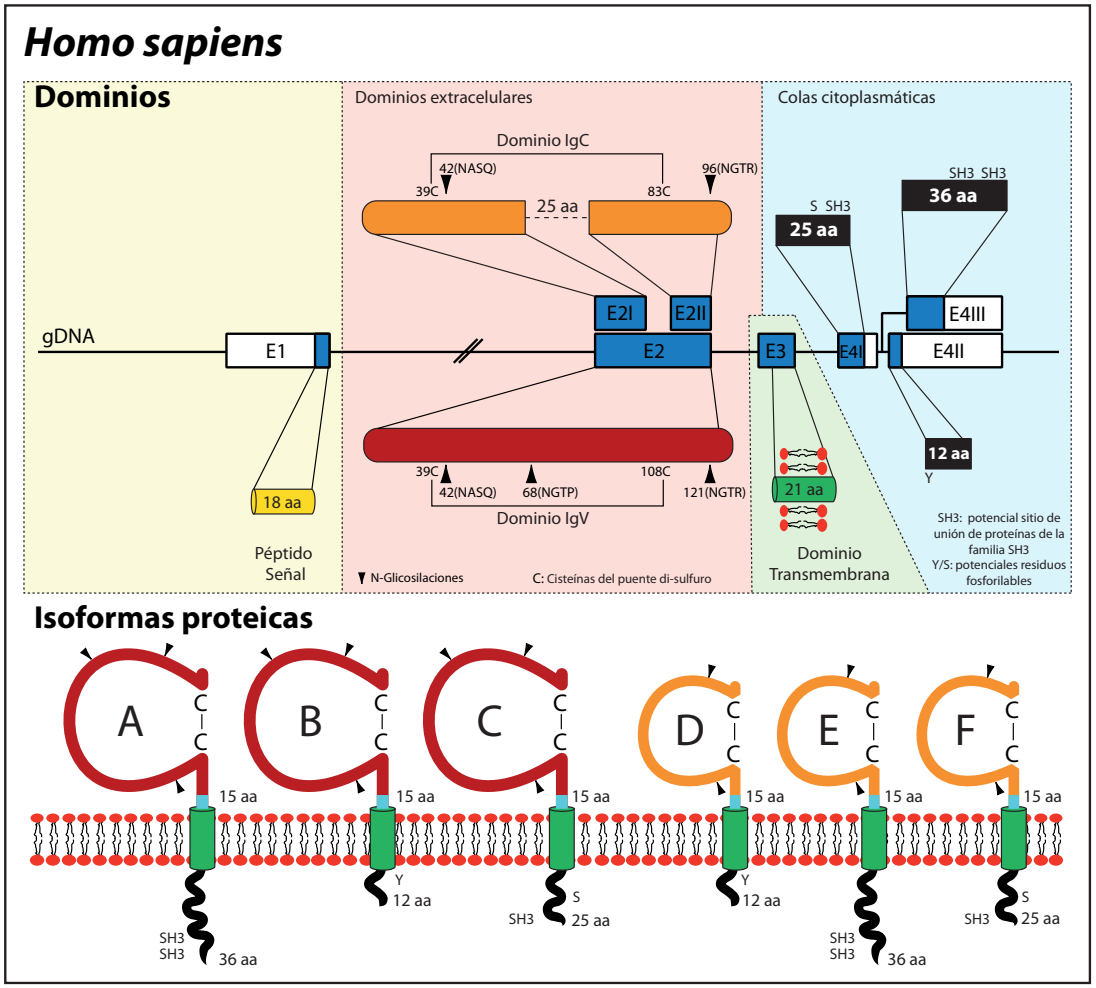


Figura 4-23. Potenciales isoformas codificantes del gen *NCR3*. En la parte superior se indican los posibles dominios codificados en los diferentes exones. En la parte inferior se representa de manera esquemática la configuración estructural de las posibles isoformas proteínas generadas mediante AS.

y en PBMCs las variantes *A* y *E* con el exón 4III (unión H6) se expresan más que las que porta el exón 4II (*B* y *D*) o el exón 4I (*C* y *F*). En monocitos, aunque la expresión global es considerablemente menor, destaca la expresión exclusiva de la variante *C*, definida por las uniones de exones H1, H3 y H4 (exón 4I) (Figura 4-22). En el resto de los tejidos sólo se detectan uniones de exones aisladas dado la baja expresión del gen *NCR3* en éstos, excepto en hígado donde se detectan las uniones de exones que definen las variantes *C* (H1, H3 y H4) y *F* (H1, H2, H3 y H4) (Figura 4-22).

Teniendo en cuenta los resultados de PCR anidada y RNA-seq conjuntamente podemos concluir que las potenciales proteínas codificadas en las múltiples variantes de splicing detectadas son las codificadas por los mensajeros *A*, *B*, *C*, *D*, *E* y *F*, aunque éstos podrían además presentar eventos de splicing como los

detectados en las uniones de exones neutras que modificarían su extremo 5´. Todas las isoformas proteicas presentarían el péptido señal codificado en el exón 1 y la región transmembrana codificada en el exón 3, que además codifica el cuello de 15 aa que separa el dominio inmunoglobulina del dominio transmembrana e implicado en la unión de ligandos (Figura 4-23) (Hartmann et al., 2012). Tres de ellas estarían compuestas por un dominio inmunoglobulina tipo V codificado en el exón 2 completo (isoformas A, B, C), que presenta tres potenciales N-glicosilaciones, en combinación con cada una de las tres colas citoplasmáticas codificadas en cada uno de los exones 4 alternativos. Las otras tres variantes codificantes (isoformas D, E, F), debido al splicing interno del exón 2, estarían compuesta por un dominio inmunoglobulina tipo C, con dos potenciales N-glicosilaciones, combinado con las tres colas citoplasmáticas (Figura 4-23). Las colas citoplasmáticas no muestran ningún dominio conservado, sin embargo, la cola codificada en el exón 4I (variantes *C* y *F*) muestra una serina que podría ser fosforilada y un potencial sitio de unión de proteínas de la familia SH3, la codificada en el exón 4II (variantes *B* y *D*) presenta una tiroxina candidata a fosforilación y la codificada por el exón 4III (variantes *A* y *E*) presentan dos sitios potenciales de unión de proteínas de la familia SH3, según predicciones bioinformáticas (Figura 4-23).

4.2.2 *Macaca Mulatta*

La conservación de secuencia entre *Macaca mulatta* y *Homo sapiens* permitió utilizar la misma estrategia y los mismos cebadores (Tabla 3-5, materiales y métodos) que en el caso de humano. La amplificación con los cebadores correspondientes a los mensajeros portadores del exón 4I resultó en la identificación de cuatro variantes de splicing (Figura 4-24). El análisis de la composición de exones reveló que tres de estas variantes (*Mmul*-c y *Mmul*-f, *Mmul-nc3)* presentaban la misma composición de exones que las variantes *C*, *F* y *NC3* descritas en humano. Además, se detecto una variante nueva, nombrada *Mmul-1*, compuesta por la suma de los exones 1, 2I, 3 y 4I (Figura 4-24). Las variantes *Mmul-ncr3* y *Mmul-1* son transcritos no codificantes por la aparición de codones prematuros de terminación de la traducción ocasionados por cambios en la fase de lectura generados por las uniones de exones que definen estos mensajeros (unión de los exón 1 y 2I y unión de los exones 1 y 2II respectivamente). Cabe destacar que no se detectó expresión del mensajero *Mmul-NC6* descrito en las bases de datos (Figura 4-8, en apartado anterior) en ninguno de los tejidos

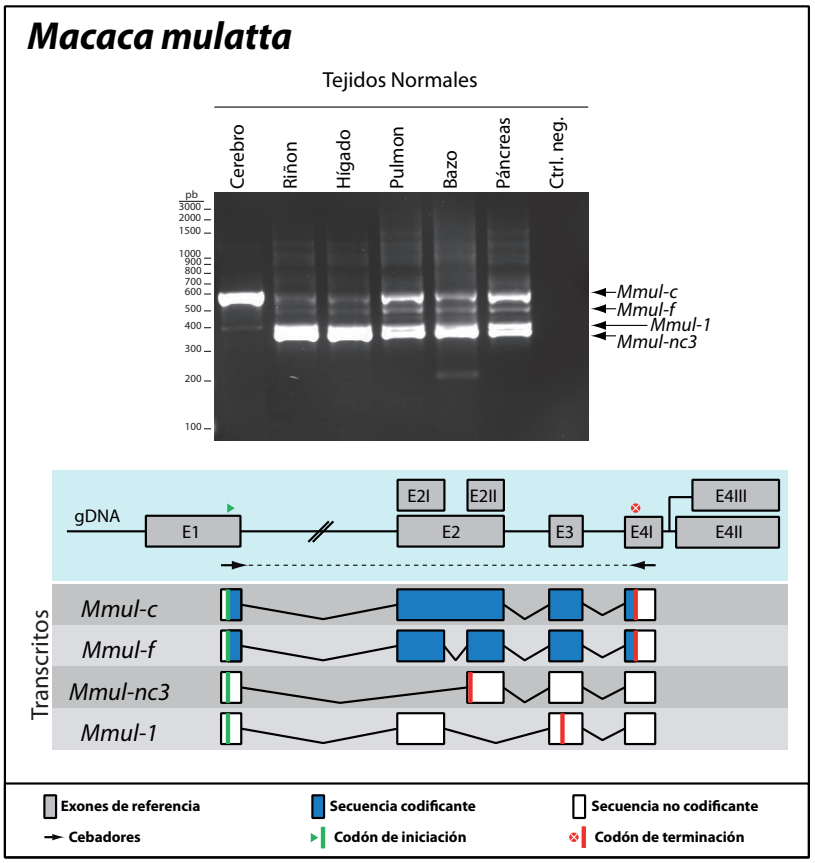


Figura 4-24. Análisis del AS del gen *Mmul-ncr3* mediante PCR anidada. En la parte superior se muestra un gel representativo de la amplificación en los tejido disponibles. En la parte inferior se representan la estructura de los mensajeros identificados.

analizados. Por otro lado, al realizar amplificaciones con los cebadores diseñados en los exones 4II y 4III, no se detectó ningún transcrito que portase estos exones en ninguno de los tejidos analizados. Modificando las condiciones de la PCR anidada (aumentando el número de ciclos y la cantidad de molde inicial) se amplificaron dos mensajeros inmaduros (datos no mostrados). Éstos presentaban los mismo exones que las variantes *Mmul-c* y *Mmul-nc3* (incluído el exón 4I), seguidos de los exones 4II y 4III y los intrones entre los exones 3 y 4I y entre éste y los exones 4II y 4III.

La expresión de las variantes detectadas muestra un patrón diferencial en los tejidos analizados (Figura 4-24). En tejido cerebral destaca la detección exclusiva de la variante *Mmul-c*. En el resto de tejidos estudiados detectamos la expresión de las variantes *Mmul-nc3* y *Mmul-1*, sin embargo, las variantes *Mmul-c* y *Mmul-f* presentan una mayor expresión en pulmón, bazo y páncreas (Figura 4-24).

Mediante el análisis de secuencias procedentes de experimentos de RNA-seq se detectaron las uniones de exones que definen los mensajeros *Mmul-c*, *Mmul-f* y *Mmul-nc3* (uniones de exones H1 a H4 y H7 en Figura 4-25). No se detectó expresión de la unión de exones P1 (Figura 4-24), característica de la nueva variante detectada por PCR anidada en este estudio (*Mmul-1*), ni tampoco se detectaron uniones de exones que incluyesen a los exones 4II y 4III, confirmado el resultado obtenido mediante PCR anidada (Figura 4-24). Adicionalmente se detectó una unión de exones de tipo neutra (unión H16 en Figura 4-24), que provoca un evento de splicing en la región no codificante del exón 1, en las mismas posiciones relativas que la unión de exones detectada en *Homo sapiens* (Figura 4-22). Igual que en humano, esta nueva unión de exones no modifica la secuencia codificante de los potenciales mensajeros. Los mayores niveles de expresión detectados mediante RNA-seq se encuentran en bazo, pulmón y células linfoblásticas (Figura 4-24).

Conjuntamente, los resultados del análisis mediante PCR anidada y mediante RNA-seq, revelan que en esta especie podrían producirse sólo dos isoformas proteícas, Mmul-c y Mmul-f (Figura 4-25). Los dominios estructurales que presentan estas isoformas son los mismos que las isoformas equivalentes humanas (C y F). La única diferencia reseñable es la longitud de la cola citoplasmática, de 25 aa en humano, y que en esta especie sería de tan sólo 11 aa (Figura 4-25). Esta reducción se debe a la presencia de un codón de terminación de la traducción en la posición +35, confirmado en la secuenciación de las variantes detectadas, mientras que en humano se localiza

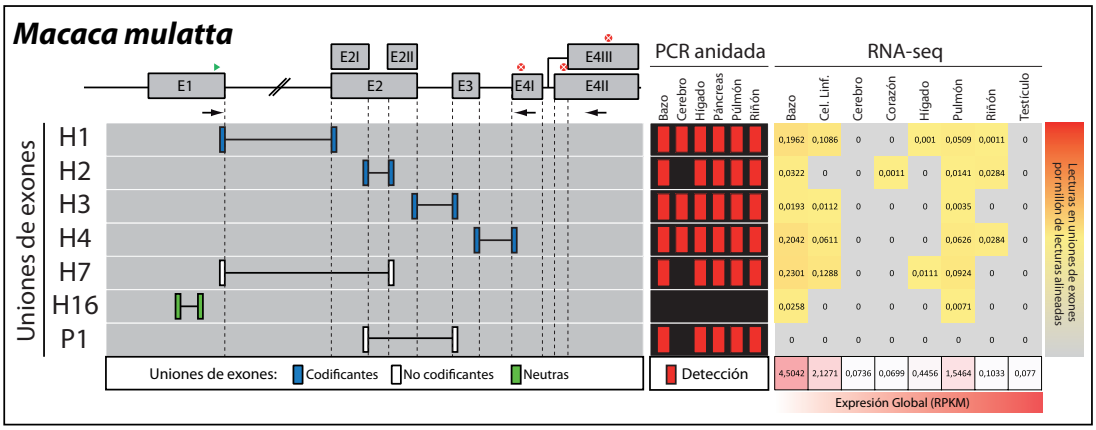


Figura 4-25 Análisis del AS del gen *Mmul-ncr3* mediante RNA-seq. En el panel de la izquierda se representan esquemáticamente las uniones de exones detectadas mediante por PCR anidada y RNA-seq. En el panel central se muestra un resumen de la información obtenida mediante PCR anidada y en el panel de la derecha se muestran las uniones de exones detectadas en las secuencias de RNA-seq disponibles (Tabla S2) medidas como lecturas alineadas en cada unión de exones normalizadas por el número tal de lecturas alineadas.

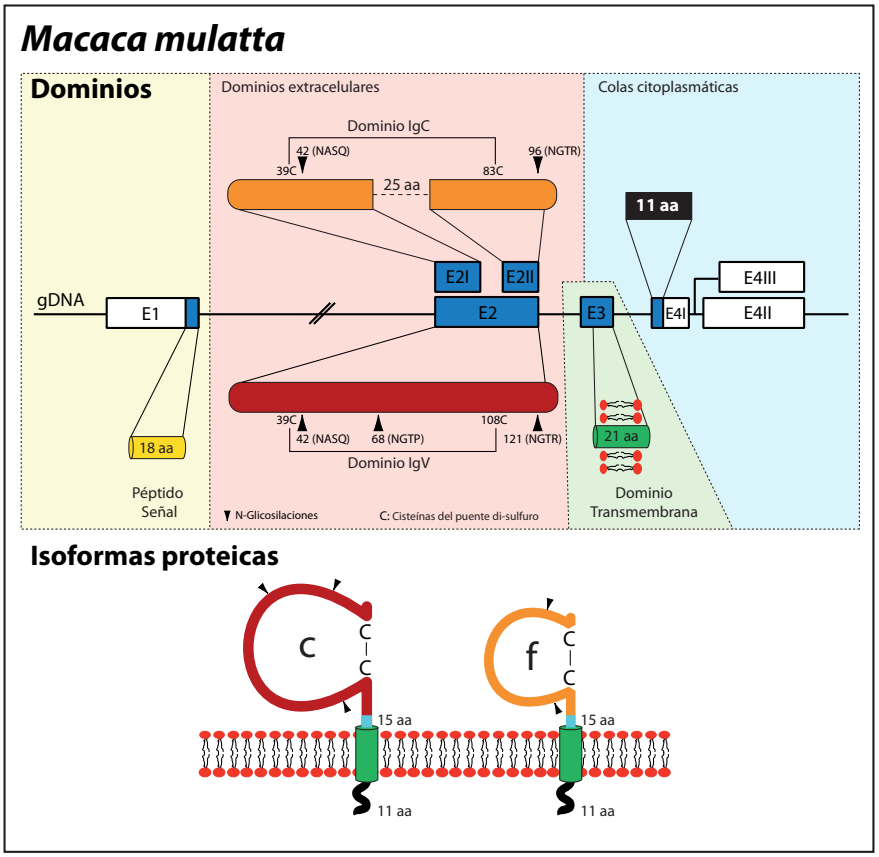


Figura 4-26. Potenciales isoformas codificantes del gen *Mmul-ncr3*. En la parte superior se indican los posibles dominios codificados en los diferentes exones. En la parte inferior se representa de manera esquemática la configuración estructural de las posibles isoformas proteínas generadas mediante AS.

en la posición +77. La serina potencialmente fosforilable detectada en la isoforma C de humano está presente también en la cola citoplasmática de macaco, sin embargo, al tratarse del último aminoácido de la proteína, el entorno bioquímico en este caso no es el adecuado para su reconocimiento. La reducción del tamaño de la cola citoplasmática en esta especie elimina el potencial sitio de unión de proteínas de la familia SH3 detectado en humano. No se detectaron mensajeros que portasen los exones 4II y 4III, ni uniones de exones que permitieran inferir su existencia, de modo que no se producirán isoformas proteicas como las humanas A, B, D y E (Figura 4-24), aunque si se haya predicho su existencia por homología de secuencia (Figura 4-8, en apartado anterior).

4.2.3 *Rattus norvegicus*

Mediante PCR anidada en los distintos tejidos disponibles de *Rattus norvegicus* se pudo identificar 5 mensajeros (Figura 4-27). En todos los tejidos analizados se detectó la variante descrita *Rnor-c*. Adicionalmente, se detectaron cuatro nuevos transcritos no descritos (Figura 4-27). El mensajero *Rnor-1* está formado por los mismos exones que la variante *Rnor-c*, portando además un nuevo exón localizado en el intrón 1 (Figura 4-27). La incorporación de este nuevo exón hace que el mensajero sea no codificante por la presencia en este exón de un codón prematuro de terminación de la traducción (Figura 4-27). Los mensajeros *Rnor-2* y *Rnor-4* también se han considerado no codificantes ya que ambos presentan deleciones en

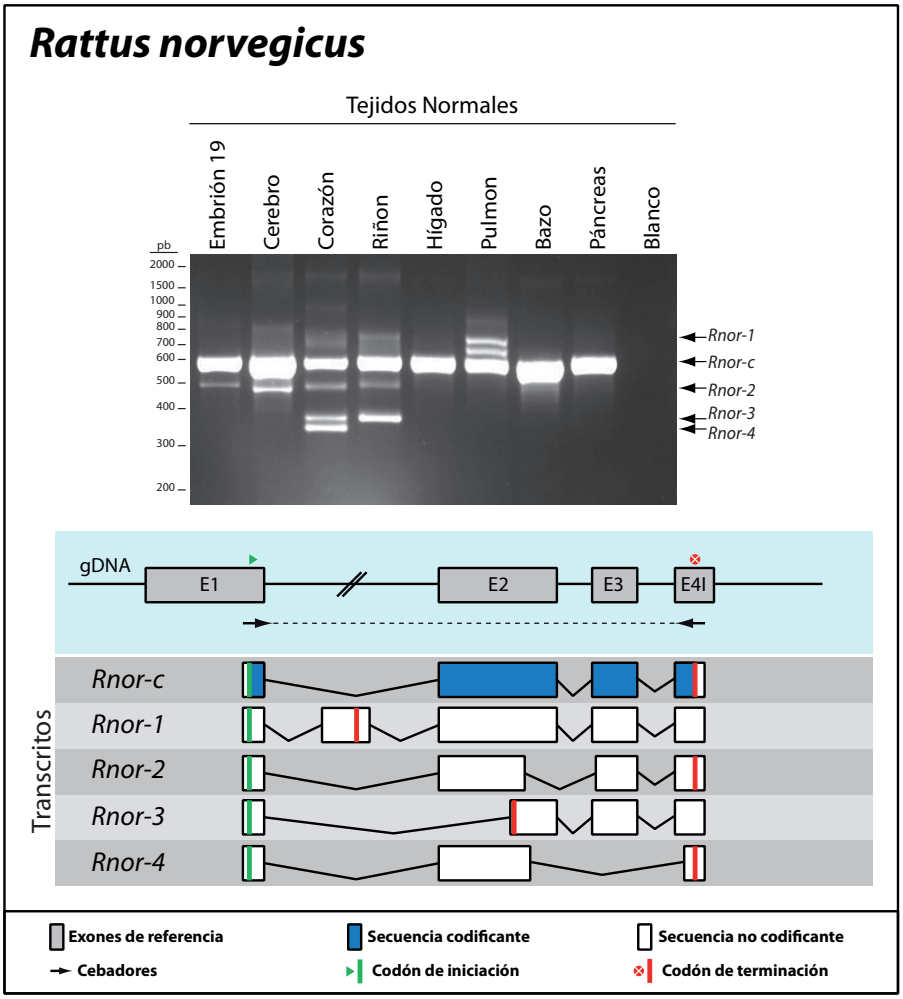


Figura 4-27. Análisis del AS del gen *Rnor-ncr3* mediante PCR anidada. En la parte superior se muestra un gel representativo de la amplificación en los tejido disponibles. En la parte inferior se representan la estructura de los mensajeros identificados.

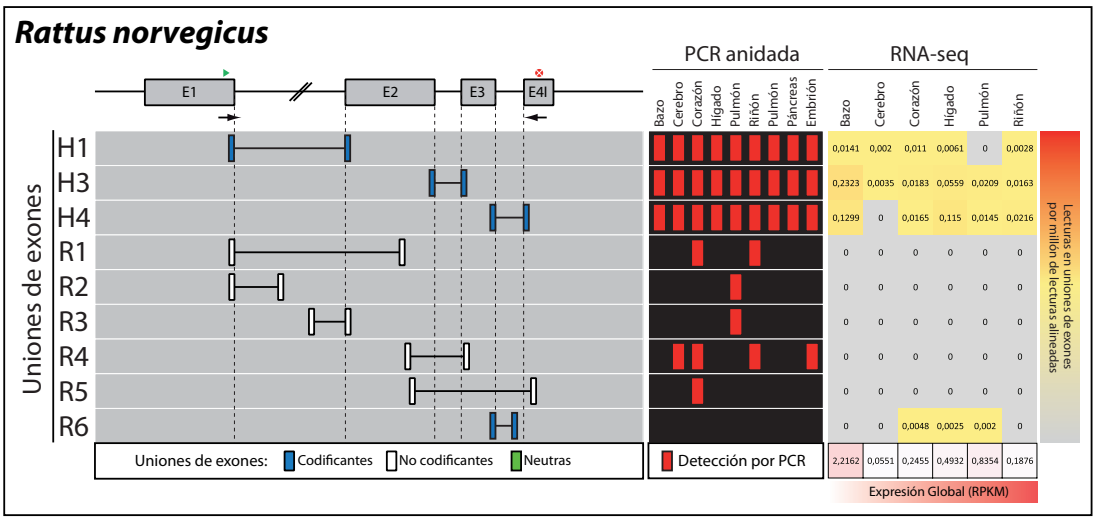


Figura 4-28 Análisis del AS del gen *Rnor-ncr3* mediante RNA-seq. En el panel de la izquierda se representan esquemáticamente las uniones de exones detectadas mediante por PCR anidada y RNA-seq. En el panel central se muestra un resumen de la información obtenida mediante PCR anidada y en el panel de la derecha se muestran las uniones de exones detectadas en las secuencias de RNA-seq disponibles (Tabla S2) medidas como lecturas alineadas en cada unión de exones normalizadas por el número tal de lecturas alineadas.

el extremo 3' del exón 2 en la región en la que está codificada la cisteína necesaria para formar el puente disulfuro característico de los dominios inmunoglobulina y por lo tanto, comprometiendo su plegamiento. En ambos mensajeros la fase de lectura se mantiene como en la variante *Rnor-C*, careciendo el transcrito *Rnor-4* además del exón 3 (Figura 4-27). La variante *Rnor-3*, similar a la humana *NC3*, presenta sólo la mitad 3' del exón 2, aunque el sitio 3' aceptor no guarda relación con el detectado en humano. Como en la variante *NC3*, aparece un codón de terminación de la traducción al inicio del fragmento conservado del exón 2, y por lo tanto, se trata de un mensajero no codificante (Figura 4-27). La expresión de estas nuevas variantes muestra cierta especificidad tisular, así la variante *Rnor-1* sólo se expresa en pulmón, el mensajero *Rnor-4* se expresa exclusivamente en corazón, la variante *Rnor-3* la detectamos en corazón y riñón, y la variante *Rnor-2* se observa en embrión, cerebro, corazón y riñón. Sin embargo, la variante *Rnor-c* se expresa en todos los tejidos, mayoritariamente en comparación con las otras variantes (Figura 4-27).

El en análisis de las secuencias de RNA-seq disponibles de esta especie se pueden detectar las uniones de exones que forman la variante canónica (*Rnor-c*) en la mayoría de los tejidos (uniones de exones H1, H3 y H4 en Figura 4-28). Las uniones de exones características de los nuevos transcritos detectados mediante PCR (uniones de exones R1 a R5 en Figura 4-28) no se detectan mediante RNA-seq, lo que podría

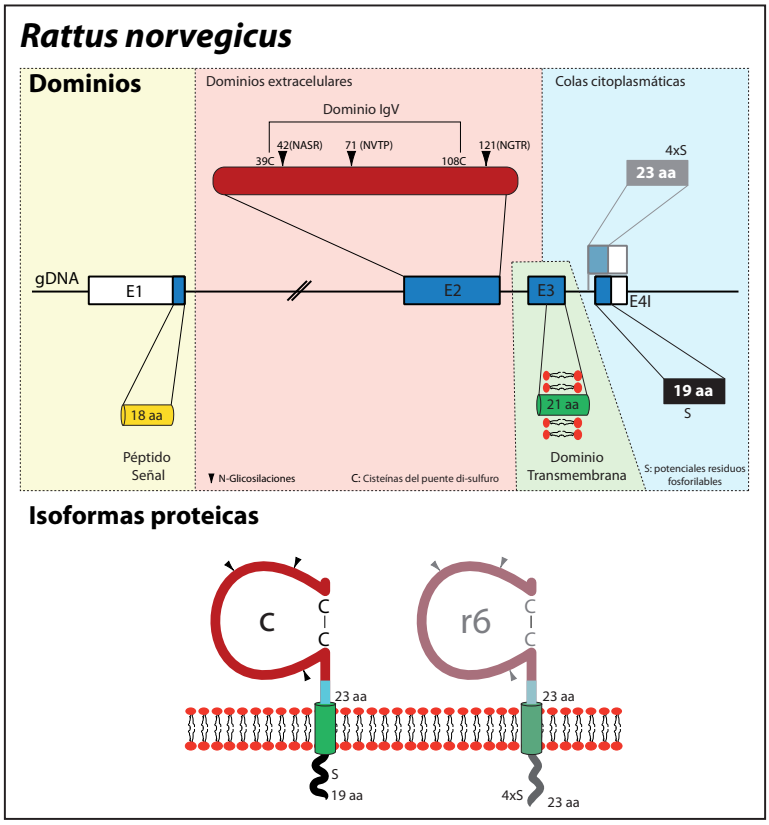


Figura 4-29. Potenciales isoformas codificantes del gen *Rnor-ncr3*. En la parte superior se indican los posibles dominios codificados en los diferentes exones. En la parte inferior se representa de manera esquemática la configuración estructural de las posibles isoformas proteínas generadas mediante AS.

deberse a la baja expresión de dichas variantes. Además de las uniones de exones mencionadas, detectamos una nueva unión no detectada por PCR anidada (unión de exones R6). Aunque el número de lecturas detectadas correspondientes con esta unión de exones es muy bajo, la combinación de ésta con las uniones de exones H1 y H3 podría producir mensajeros codificantes. La unión de exones R6 supone la retención parcial del intrón entre los exones 3 y 4, de manera que la fase de lectura del exón 4 se ve alterada, y por lo tanto, modificando la posible cola citoplasmática de la proteína final (Figura 4-28). La expresión global del gen *Rnor-ncr3*, medida mediante RNA-seq, indica que los mayores niveles de expresión se encuentran en bazo seguido de pulmón, con cierta expresión en hígado (Figura 4-28).

En análisis del splicing alternativo en esta especie permite inferir la existencia de dos potenciales proteínas, la isoforma Rnor-c ya descrita, y una nueva potencial isoforma, denominada aquí Rnor-r6 (por la unión de exones que la define), que difiere de la primera en la secuencia de la cola citoplasmática. Aunque las herramientas

de predicción de dominios no detectan la presencia de ningún dominio conservado en ninguna de las dos potenciales colas citoplasmáticas, sí presentan serinas potencialmente fosforilables, una en el caso de la variante canónica y cuatro en el caso de Rnor-R6 (Figura 4-29). Como en humano, los dominios característicos de ambas variantes son el péptido señal codificado en el exón 1, el dominio inmunoglobulina tipo V codificado en el exón 2 (con tres potenciales sitios de N-glicosilación) y un dominio transmembrana codificado en el exón 3, además de las colas citoplasmáticas ya mencionadas. En exón 3, 24 pb mayor que en humano, codifica además el cuello entre el dominio inmunoglobulina y el dominio transmembrana, que en esta especie es de 23 aa, ocho más que en humano.

4.2.4 *Mus musculus*

Pese a la descripción del gen *Mmus-ncr3* como un pseudogen en diferentes cepas de ratón (Hollyoake et al., 2005), en la bibliografía se pueden encontrar referencias sobre la detección de transcritos procedentes de este gen (Sivakamasundari et al., 2000). Por ello, se decidió re-analizar la posible expresión de este gen mediante PCR anidada y RNA-seq. No se detectó ningún mensajero procedente de este gen mediante PCR anidada en los tejidos disponibles (Tabla 3-3 en materiales y métodos), lo que confirma su condición de pseudogen. Sin embargo, en el resultado del análisis de los datos disponibles de RNA-seq en esta especie (Tabla S2) se puede detectar el alineamiento de lecturas en la región en la que se localiza en gen, lo que podría deberse a niveles de transcripción residuales. Además, se pueden detectar dos uniones de exones, con un bajo número de lecturas (en bazo y pulmón), que

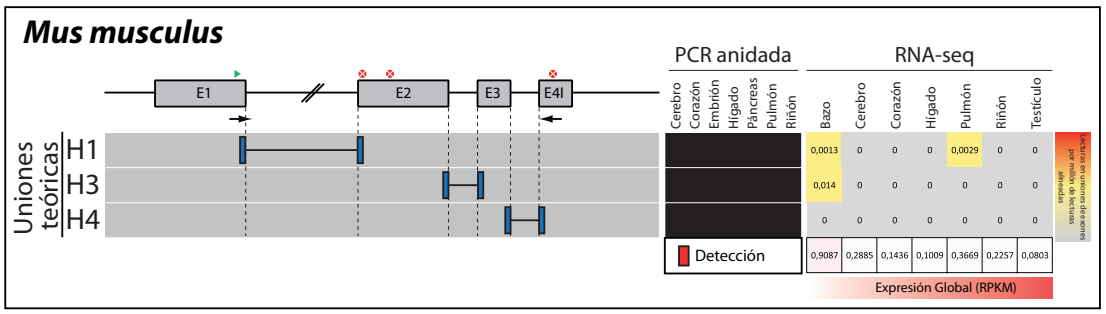


Figura 4-30 Análisis del AS del gen *Mmus-ncr3* mediante RNA-seq. En el panel de la izquierda se representan esquemáticamente las uniones de exones detectadas mediante por PCR anidada y RNA-seq. En el panel central se muestra un resumen de la información obtenida mediante PCR anidada y en el panel de la derecha se muestran las uniones de exones detectadas en las secuencias de RNA-seq disponibles (Tabla S2) medidas como lecturas alineadas en cada unión de exones normalizadas por el número tal de lecturas alineadas.

se corresponden con la unión de los exones 1-2, y 2-3, cuyo origen podría ser dicha transcripción residual (Figura 4-30 y Tabla S3). Por lo tanto, se puede concluir que el gen *Mmus-ncr3* es un pseudogen que sin embargo, presenta ciertos niveles de transcripción residual.

4.2.5 *Bos taurus*

El análisis de las variantes de splicing mediante PCR anidada en vaca permitió identificar siete mensajeros diferentes (Figura 4-31). El mensajero *Btau-c*, equivalente a la variante *C* humana, se expresa en todos los tejidos analizados. También se puede observar expresión en todos los tejidos excepto en pulmón de la variante no codificante *Btau-nc3*, no descrita en esta especie y similar a la humana *NC3* (Figura 4-31). Además de estas variantes conservadas con humano, se han detectado cinco nuevas variantes (Figura 4-31). La variante *Btau-1*, detectada en todos los tejidos, y formada por los exones 1, 2 y 4, mantiene la fase de lectura hasta el codón de terminación descrito en el exón 4 (posición +47). Los otros cuatro mensajeros (*Btau-2*, *Btau-3*, *Btau-4* y *Btau-5*), de tipo no codificante, presentan eventos de splicing alternativo que provocan la deleción total (*Btau-5*) o parcial (*Btau-2*, *Btau-3* y *Btau-4*) del exón 2 y por lo tanto, las potenciales proteínas codificadas carecerían del dominio inmunoglobulina característico de este receptor (Figura 4-31). En las variantes *Btau-2*, *Btau-3* y *Baut-5* la fase de lectura se mantiene, sin embargo, se han considerado no codificantes por las grandes deleciones que presentan. En la variante *Btau-4*, sin embargo, el splicing interno sufrido en el exón 2 provoca un cambio de fase que genera la aparición de un codón de terminación de la traducción prematuro al principio del exón 3 (Figura 4-31). La expresión de estas variantes presenta especificidad de tejido, observándose expresión de *Btau-2* sólo en corazón, *Btau-3* y *Btau-4* en riñón e hígado y *Btau-5* en riñón y pulmón (Figura 4-31).

Mediante el análisis de las secuencias disponibles procedentes de experimentos de RNA-seq en esta especie (Table S2) se detecta mayoritariamente las uniones de exones que caracterizan la variante *Btau-c* (uniones de exones H1, H3 y H4 en Figura 4-32). No se detectaron las uniones de exones que definen las nuevas variantes observadas mediante PCR anidada (uniones de exones H7 y B1-B5), ni siquiera, la unión de exones H7, característica de la variante no codificante *Btau-nc3* conservada en otras especies (Figura 4-32). Por otro lado, detectamos varias uniones de exones de tipo neutra localizadas en la región anterior al exón 1. En el apartado anterior, se

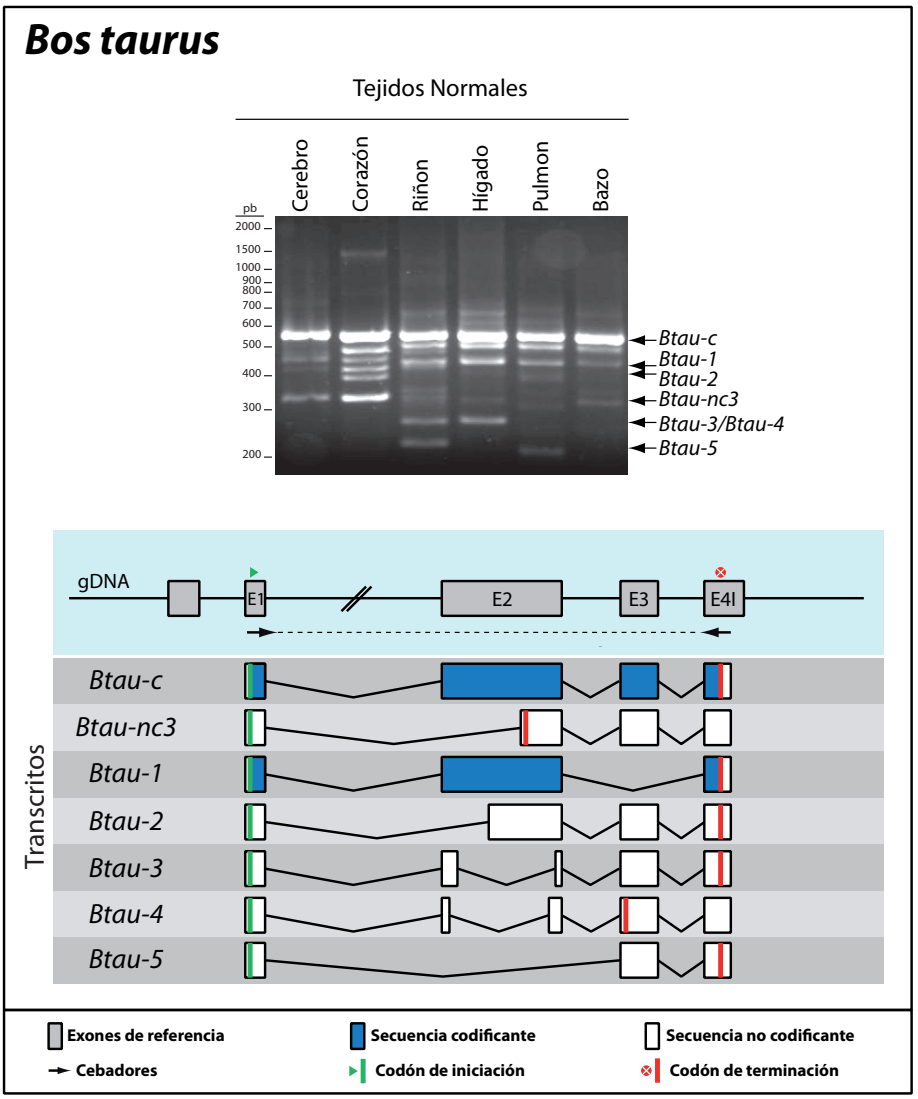


Figura 4-31. Análisis del AS del gen *Btau-ncr3* mediante PCR anidada. En la parte superior se muestra un gel representativo de la amplificación en los tejido disponibles. En la parte inferior se representan la estructura de los mensajeros identificados.

detalló la existencia en las bases de datos de dos mensajeros (*Btau-c1* y *Btau-c2*), cuya traducción produciría la isoforma Btau-c, aunque presentan diferencias en los exones iniciales (Figura 4-13, en apartado anterior). Mediante el análisis de las secuencias de RNA-seq se detectan las uniones de exones características de ambos mensajeros: la unión de exones B6 en el caso de *Btau-c2* y las uniones de exones B10 y B12 en el caso de *Btau-c1* (Figura 4-32). Los niveles de expresión detectados de la unión de exones B6, similares a los detectados de las uniones de exones H1, H3 y H4, sugiere que ésta es una unión de exones mayoritaria, es decir, que la variante *Btau-c2* puede considerarse como la variante canónica en esta especie (Figura 4-32).

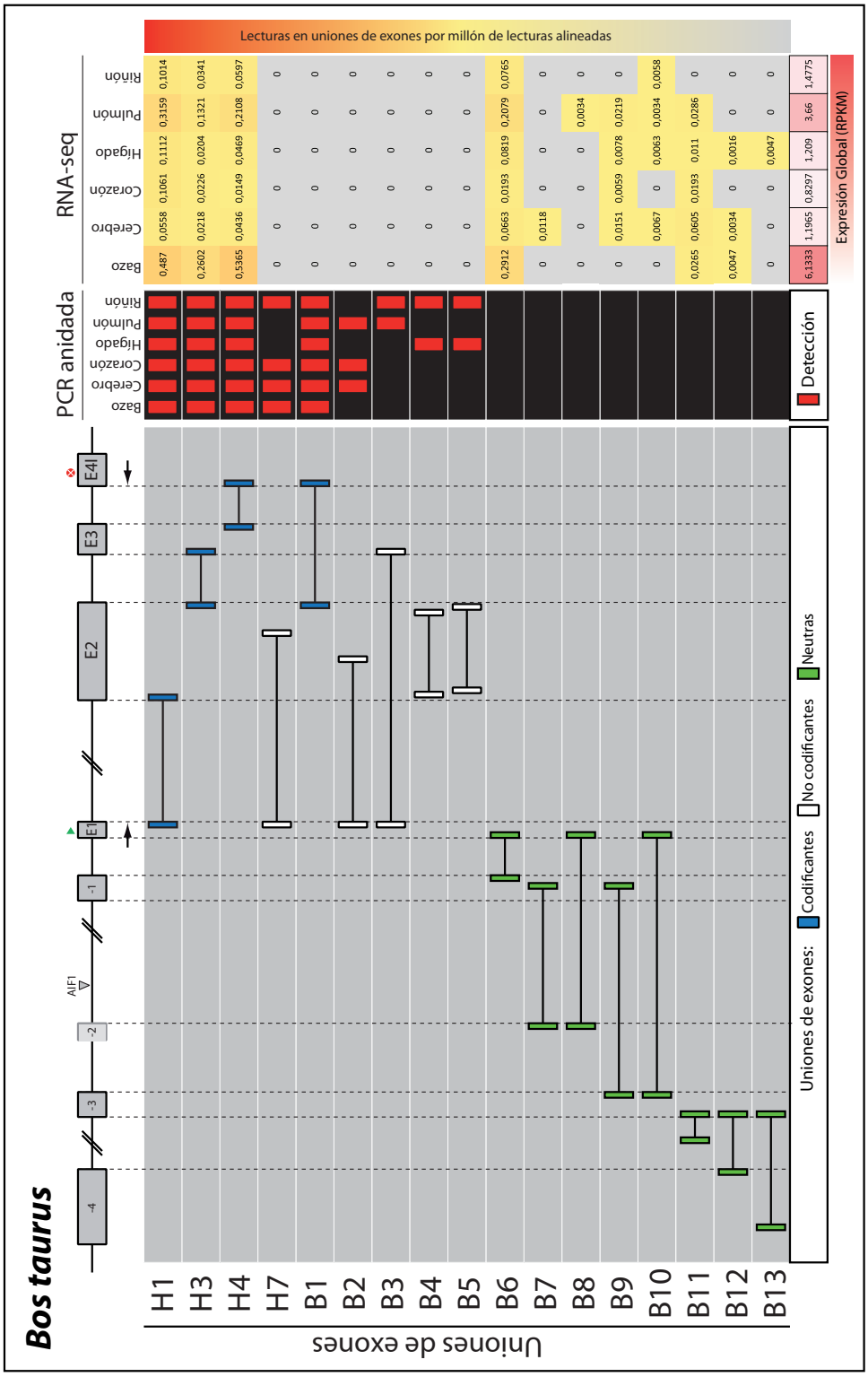


Figura 4-32 Análisis del AS del gen *Btau-ncr3* mediante RNA-seq. En el panel de la izquierda se representan esquemáticamente las uniones de exones detectadas mediante por PCR anidada y RNA-seq. En el panel central se muestra un resumen de la información obtenida mediante PCR anidada y en el panel de la derecha se muestran las uniones de exones detectadas en las secuencias de RNA-seq disponibles (Tabla S2) medidas como lecturas alineadas en cada unión de exones normalizadas por el número tal de lecturas alineadas.

Además de estas uniones de exones ya descritas, se pueden observar la presencia de cinco nuevas uniones de exones neutras (Figura 4-32). Las uniones de exones B7 y B8 permiten definir un nuevo exón, denominado aquí exón -2, enlazando la primera con el exón denominado -1 (la primera mitad del exón 1, en referencia a humano), aunque sólo incorporaría una parte de éste. La segunda, unión de exones B8, une este nuevo exón -2 con el exón 1 (Figura 4-32). Las uniones de exones B11 y B13, podrían definir también dos nuevos exones, unidos ambos con el exón -3 (Figura 4-32), estando el primero en el intrón entre los exones -4 y -3, y el segundo siendo una versión alternativa del exón -4, utilizando un sitio donador alternativo anterior al descrito (Figura 4-32). Ninguna de estas uniones de exones neutras (B6-B13) modifican la secuencia codificante ni ninguna combinación de éstas. Además los bajos niveles de expresión detectados para las uniones de exones B7-B13 indica que los mensajeros que porten dichas uniones serán de expresión minoritaria, no siendo así en el caso de la unión de exones B6, que parece ser constitutiva dado sus niveles de expresión, similares a los de las uniones de exones canónicas (H1, H3 y

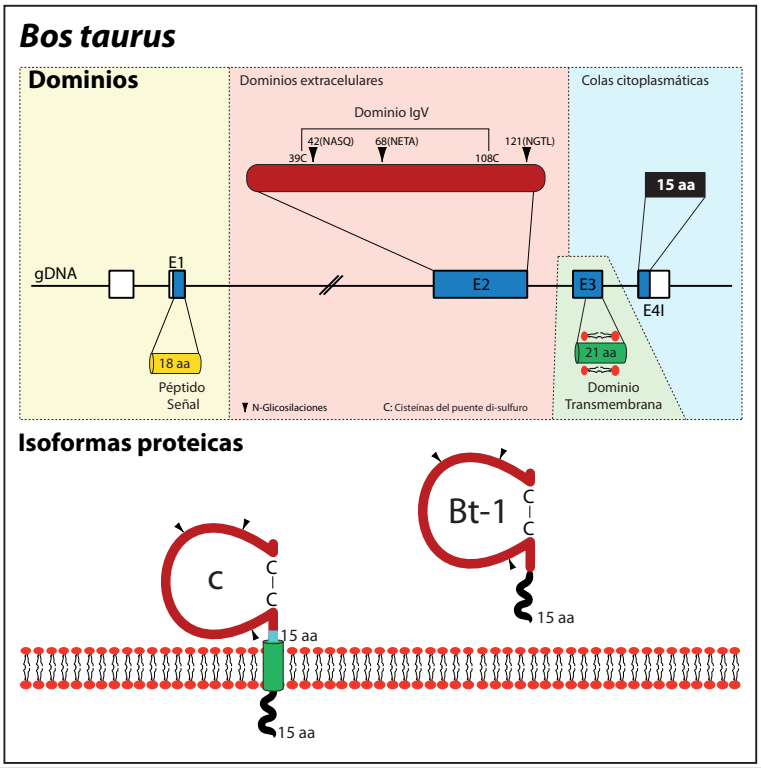


Figura 4-33. Potenciales isoformas codificantes del gen *Btau-ncr3*. En la parte superior se indican los posibles dominios codificados en los diferentes exones. En la parte inferior se representa de manera esquemática la configuración estructural de las posibles isoformas proteínas generadas mediante AS.

H4). De manera global se pueden apreciar que los mayores niveles de expresión del gen *Btau-ncr3* se detectan en bazo y pulmón, sin embargo, los niveles de expresión en el resto de tejidos es superior al detectado en otras especies, por lo que se deduce que la expresión del gen es más ubicua en esta especie (Figura 4-32).

Las isoformas proteicas potenciales, dado los resultados obtenidos, son dos: la isoforma Btau-c y una nueva isoforma denominada aquí Btau-1 (Figura 4-33). Ambas comparten el péptido señal codificado en el exón 1, así como el dominio inmunoglobulina tipo V, que presente tres sitios de N-glicosilación conservados en las mismas posiciones relativas que en humano, y la cola citoplasmática codificada en el exón 4, de tan sólo 15 aa, que conserva la serina potencialmente fosforilable presente en humano, aunque no es reconocida como tal por los programas de predicción. La isoforma Btau-c, como en humano, además presenta un dominio transmembrana codificado en el exón 3, junto con un cuello de 15 aa que separa el dominio inmunoglobulina del domino transmembrana (Figura 4-33). Por lo tanto, la función potencial de esta isoforma, al igual que en humano, se desarrollará en la membrana plasmática. Por otro lado, la isoforma Btau-1 carece de dominio transmembrana, por la exclusión del exón 3, y por lo tanto, podría tratarse de una isoforma soluble del receptor (Figura 4-33).

4.2.6 *Sus scrofa*

En el apartado anterior se manifestó la posibilidad de que el gen *Sscro-ncr3* fuese un pseudogen por la presencia de tres codones prematuros de terminación de la traducción en el exón 2 (Figura 4-14). Pese a esto y desconociendo la existencia del mensajero *Sscro-7*, anotado recientemente, se decidió analizar mediante PCR anidada la transcripción potencial del gen *Sscro-ncr3* en diferentes tejidos. Sorprendentemente, se pudieron amplificar y secuenciar diez transcritos diferentes (Figura 4-34). Lo más interesante al analizar la secuencia de estos transcritos detectados es que la mayoría evita la región del exón 2 donde se localizan los codones prematuros de terminación gracias a eventos de splicing alternativo, excepto el mensajero *Sscro-8* (Figura 4-34). Además, todos presentan la retención del intrón 3, entre el exón 3 y el exón 4 (Figura 4-34). Los mensajeros *Sscro-1*, *Sscro-2* y *Sscro-9* presentan el mismo evento de splicing en el interior del exón 2, dividiendo éste en dos fragmentos de 120 pb y 62 pb, sin que la unión de éstos restaure la fase de lectura homóloga a la detectada en otras especies. Estos tres mensajeros presentan además diferentes eventos de

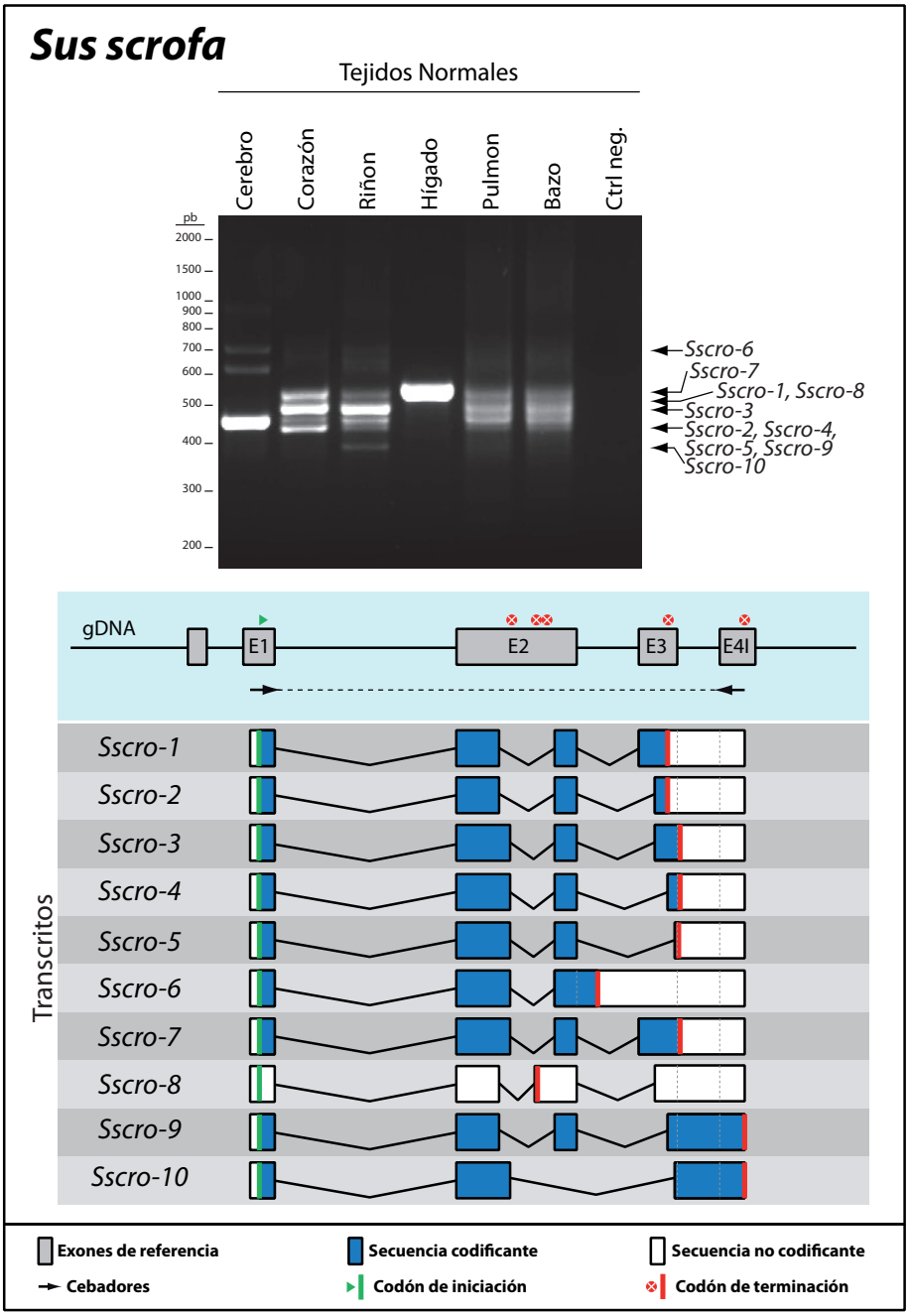


Figura 4-34. Análisis del AS del gen *Sscro-ncr3* mediante PCR anidada. En la parte superior se muestra un gel representativo de la amplificación en los tejido disponibles. En la parte inferior se representan la estructura de los mensajeros identificados.

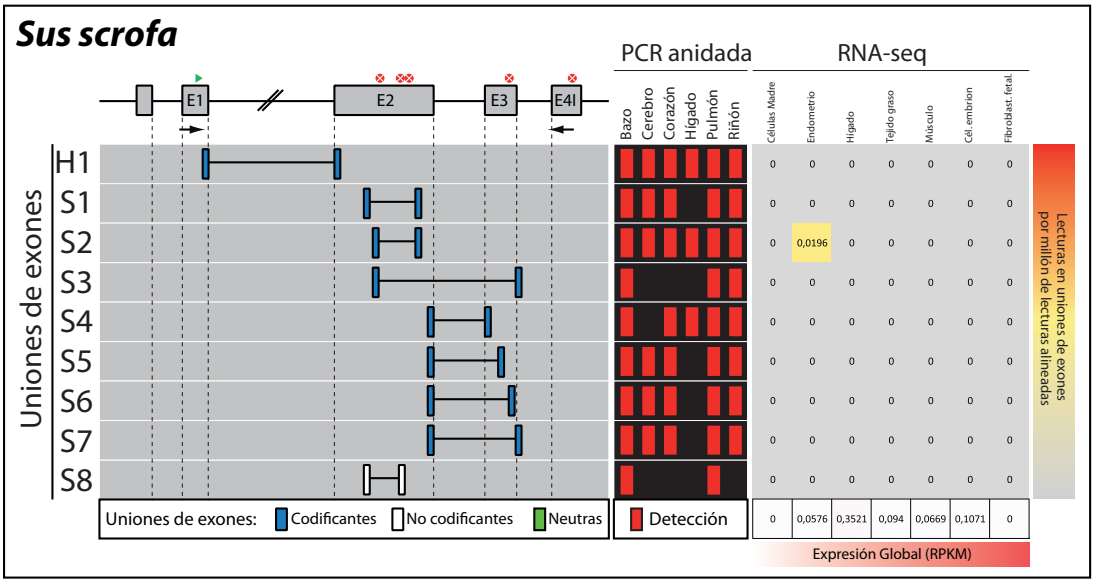


Figura 4-35 Análisis del AS del gen *Sscro-ncr3* mediante RNA-seq. En el panel de la izquierda se representan esquemáticamente las uniones de exones detectadas mediante por PCR anidada y RNA-seq. En el panel central se muestra un resumen de la información obtenida mediante PCR anidada y en el panel de la derecha se muestran las uniones de exones detectadas en las secuencias de RNA-seq disponibles (Tabla S2) medidas como lecturas alineadas en cada unión de exones normalizadas por el número tal de lecturas alineadas.

splicing entre el exón 2 y el exón 3, que aparece unido al exón 4 por la retención de intrón 3 (Figura 4-34). El mensajero *Sscro-1* presenta el exón 3 completo, mientras que los mensajeros *Sscro-2* y *Sscro-9* sólo conservan fragmentos de éste de 63 pb y 27 pb, respectivamente. En los transcritos *Sscro-1* y *Sscro-2* se observan codones de terminación en exón 3 en la misma posición para ambos, y en la variante *Sscro-9* la secuencia codificante se prolonga hasta un codón de terminación presente en la posición +70 del exón 4 (Figura 4-34). Los mensajeros *Sscro-3*, *Sscro-4*, *Sscro-5*, *Sscro-6* y *Sscro-7* presentan un evento de splicing interno del exón 2 común entre ellos y diferente del anterior. En éstos el exón 2 queda divido en dos fragmentos de 151 pb y 62 pb, siendo el sitio 5' donador el detectado en otras especies, desplazado un nucleótido por la deleción en la posición +68 de este exón, y el sitio aceptor el mismo utilizado en los mensajeros *Sscro-1*, *Sscro-2* y *Sscro-9* (Figura 4-34). La unión de estos dos fragmentos restaura la fase de lectura en el segundo fragmento del exón 2 y en el exón 3, recuperando la homología con la secuencia codificante de otras especies. Como los primeros mensajeros, éstos también presentan diversos eventos de splicing entre el exón 2 y 3 portando fragmentos de diferentes tamaños de este último o el exón 3 completo en el caso del mensajero *Sscro-7*. Las variantes *Sscro-3*, *Sscro-4* y *Sscro-5* presentan fragmentos de 63, 27 y 7 pb del exón 3,

respectivamente, y en todos los casos la potencial traducción de los mensajeros acabaría en un codón de terminación al inicio del intrón 3 retenido (Figura 4-34). El mensajero *Sscro-6* presenta además la retención del intrón 2, entre los exones 2 y 3, terminando la traducción hipotética en dicho intrón retenido. El mensajero *Sscro-7*, descrito en las bases de datos, presenta el exón 3 completo, cuya fase de lectura ha sido restaurada por el splicing interno del exón 2, que también afectaría al exón 4 de no ser por la retención del intrón 3, donde observamos la presencia de un codón de parada de la traducción en las misma posición que en las variantes *Sscro-3*, *Sscro-4* y *Sscro-5* (Figura 4-34). Los transcritos *Sscro-8* y *Sscro-10* presentan otros eventos de splicing en el exón 2. En el primero (*Sscro-8*) el exón 2 queda también divido en dos fragmentos, de 120 pb como en los primeros mensajeros y de 118 pb generado por el uso de un sitio 3' aceptor localizado en la misma posición relativa que el observado en otras especies. Este evento de splicing del exón 2 no elimina dos de los tres codones de terminación detectados en la secuencia genómica por lo que la traducción potencial acabaría en el primero de ellos (Figura 4-34). El mensajero *Sscro-8* presenta 63 pb del exón 3, así como el exón 4 y el intrón entre éstos. Finalmente, el mensajero *Sscro-10*, presenta sólo una mitad del exón 2, de 151 pb como en los mensajeros *Sscro-3*, *Sscro-4*, *Sscro-5*, *Sscro-6* y *Sscro-7*, además porta sólo 7 pb del exón 3, junto al exón 4 y el intrón 3 retenido (Figura 4-34). La fase de lectura de este mensajero se extiende hasta el codón de terminación detectado en la posición +70 del exón 4, similar a lo observado en el transcrito *Sscro-9* (Figura 4-34).

La disponibilidad de secuencias procedentes de experimentos de RNA-seq en esta especie son limitados encontrando muestras de tejidos menos habituales (Tabla S2). El análisis de estos datos sólo permitió la identificación de una unión de exones, la unión de exones S2 característica de las variantes *Sscro-3*, *Sscro-4*, *Sscro-5*, *Sscro-6* y *Sscro-7* (Figura 4-35). Sólo detectamos esta unión de exones en tejido endometrial, donde pudimos observar otra unión de exones en el alineamiento (no mostrado) que unía los exones 1 y 2, es decir, la unión de exones H1, pero fué descartada por los programas de alineamiento por presentar sólo cuatro nucleótidos en el exón 1, siendo el mínimo permitido de ocho nucleótidos.

Las potenciales proteínas codificadas por las variantes de splicing detectadas son muy diferentes a las observadas en otras especies, y no codifican *NCR3* ni ninguna proteína parecida. Todas, excepto el mensajero *Sscro-8*, son potencialmente

codificantes, sin embargo el cambio de fase ocasionado por la deleción de una citosina en la posición +68 del exón 2 modifica por completo la secuencia codificada desde este punto, no detectándose conservación de la secuencia con ninguna especie. Sólo los mensajeros que portan la unión de exones S2, es decir, los mensajeros *Sscro-3*, *Sscro-4*, *Sscro-5*, *Sscro-6* y *Sscro-7*, presentan la restauración de la fase de lectura en la segunda parte del exón 2 y el exón 3 o parte de éste. Esta restauración de la traducción afectaría también al exón 4, encontrando un codón de terminación en fase en la posición +47, similar a lo descrito en vaca, de no ser por la retención del intrón 3 que provoca la aparición de un codón de terminación en éste. En el exón 1, como en el resto de las especies analizadas, está codificado en péptido señal. Sin embargo, pese a la restauración parcial de la fase de lectura del exón 2, y la correcta traducción de las primeras 67 pb iniciales del exón 2, las dos cisteínas fundamentales para el plegamiento de un dominio inmunoglobulina se pierden, la primera por el cambio de fase en la posición +68 y el segundo debido a la unión de exones S2, que excluye la secuencia donde se codificaría la cisteína, y por lo tanto, los programas de predicción no detectan homología alguna que permita inferir dominios estructurales en el exón 2. Los mensajeros *Sscro-3* y *Sscro-7*, que presentan 63 pb del exón 3 y el exón completo respectivamente, podrían presentar un dominio transmembrana completo, gracias a la restauración de la fase de lectura (Figura 4-34).

4.3 Análisis de las variantes de splicing del gen *NCR3* en sangre de diferentes primates

Gracias a una colaboración con el Zoo-Aquarium de Madrid se dispuso de muestras de sangre de diferentes primates (Tabla S1). Asimismo, la colaboración del Banco de Donantes del Hospital Universitario la Paz de Madrid permitió disponer de muestras de sangre de donantes sanos para realizar comparativas con humanos. Tanto las muestras humanas como las de los diferentes primates fueron procesadas siguiendo en mismo procedimiento para la obtención de RNA y DNA genómico (ver materiales y métodos). La metodología general utilizada para el análisis de las variantes de splicing del gen *NCR3* en las muestras de sangre de primates es básicamente la misma que en el apartado anterior; PCR anidada y análisis de secuencias de RNA-seq disponibles. Sin embargo, el diseño de los cebadores necesarios para las PCR anidadas no fue específico para cada especie, sino que se diseñaron cebadores comunes para todas las especies incluidas en este apartado (Figura 4-36). Para el diseño de estos cebadores se alinearon las secuencias de los exones 1, 4I, 4II y 4III de los primates disponibles en las bases de datos, así como las secuencias obtenidas en este trabajo (ver apartado 4.1). El uso de los cebadores comunes para todas las especies reduce el riesgo de introducir sesgos en las amplificaciones de modo que la comparación sea más fiel (Figura 4-36).

4.3.1 Análisis de las variantes de splicing mediante PCR anidada

En *Homo sapiens* ya se analizó la expresión del gen *NCR3* en muestras de sangre (PBMCs) comercial donde se detectaron las nueve variantes canónicas (ver apartado 4.2). Sin embargo, para realizar una comparación adecuada repetimos este análisis en muestras de sangre procesadas en el laboratorio siguiendo el mismo procedimiento que en el caso de las muestras de primates. De nuevo se amplificaron y secuenciaron los nueve transcritos canónicos y como en el análisis anterior no se detectaron nuevas variantes (Figura 4.37).

En *Gorilla gorilla* se detectó expresión de ocho transcritos diferentes (Figura 4-37), siendo siete equivalentes a las variantes humanas *A*, *C*, *E*, *F*, *NC1*, *NC2* y *NC3* (*Ggor-a*, *Ggor-c*, *Ggor-e*, *Ggor-f*, *Ggor-nc1*, *Ggor-nc2* y *Ggor-nc3* respectivamente). Además, se amplificó una nueva variante, denominada aquí *Ggor-1*, de tipo no codificante, que está compuesta por los exones 1, 3 y 4III, además de un pequeño

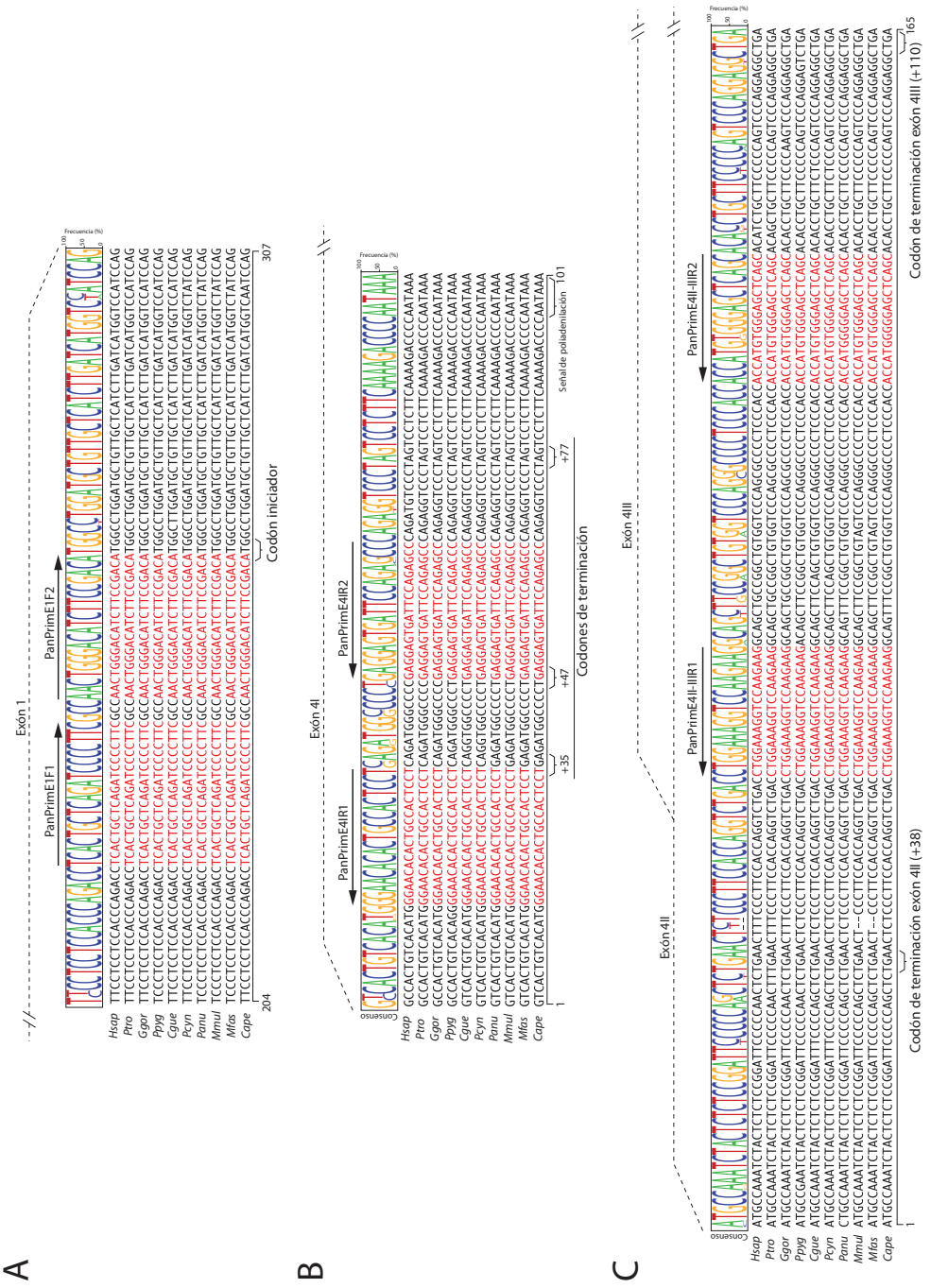


Figura 4-36 Alineamiento de las secuencias de los exones 1 (A), 4I (B) y 4II/4III (C) del gen *NCR3* en las distintas especies de primates para el diseño de los cebadores comunes.

fragmento del exón 2 (77 pb) debido al uso de un sitio 5' donador alternativo (Figura 4-37). El uso de este sitio alternativo, diferente del descrito en humano, provoca un cambio en la fase de lectura, y aunque no aparece un codón de terminación hasta el principio del exón 4III, la exclusión de una gran parte del exón 2 evitaría el correcto plegamiento del dominio inmunoglobulina característico de este gen, por lo que se ha considerado como un mensajero no codificante. No se detectó expresión de variantes potencialmente codificantes que portasen el exón 4II, es decir, las variantes equivalentes a las humanas *B* y *D* (Figura 4.37).

En las muestras de *Pongo pygmaeus* si se detectó expresión de las variantes homólogas a las humanas *B* y *D* (*Ppyg-b* y *Ppyg-d* respectivamente, Figura 4-37), además de las variantes *Ppyg-c* y *Ppyg-nc1*, similares a las variantes *C* y *NC1* descritas en humano, siendo esta última de tipo no codificante (Figura 4-37). Además, se amplificaron dos nuevas variantes consideradas no codificantes; *Ppyg-1* y *Ppyg-2* (Figura 4-37). La primera muestra un exón 2 reducido en su extremo 3' por el uso de un sitio donador alternativo en la misma posición relativa que el detscrito en humano característico de la variante no codificante *NC6*, descrita por homología en las bases de datos, y siendo detectada la unión de exones en este estudio mediante el análisis de secuencias de RNA-seq. Al igual que en la variante *NC6*, el uso de este sitio donador provoca la aparición de un codón prematuro de terminación de la traducción al inicio del exón 3 (Figura 4-37). Sin embargo, a diferencia de la variante humana, en esta especie el exón final incluido es el 4II en lugar del 4I (Figura 4-37). El transcrito *Ppyg-2* sólo presenta un pequeño fragmento de 88 pb del extremo 5' del exón 2, y tan sólo 47 pb del extremo 3' del exón 3, además del exón 4III completo (Figura 4-37). Este evento de splicing mantiene la fase de lectura canónica, de modo que la traducción hipotética de este mensajero utilizaría el codón de terminación descrito en el exón 4III, sin embargo, la deleción de una parte muy considerable del exón 2 no permitiría el correcto plegamiento de la proteína y por lo tanto se ha considerado no codificante (Figura 4-37). Curiosamente en esta especie, al contrario que en *Gorilla gorilla*, no se expresan variantes codificantes que porten el exón 4III (Figura 4-37).

La amplificación en las muestras de *Colobus guereza* permitió identificar cuatro mensajeros diferentes. Dos de ellos, *Cgue-c* y *Cgue-nc3*, presentan la misma composición exónica que la variante codificante *C* y la no codificante *NC3* detectadas en humano respectivamente (Figura 4-37). Además, se detectaron dos nuevas

variantes de tipo no codificante; *Cgue-1* y *Cgue-2*. La primera de ellas muestra la misma composición exónica que la variante *Mmul-1* detectada en el apartado anterior en *Macaca mulatta*. La segunda, a diferencia de la primera que porta el exón 4I, presenta el exón 4III, siendo ésta la única variante detectada en la que se incluya este exón (Figura 4-37). Ambas variantes no codificantes, *Cgue-1* y *Cgue-2*, comparte con el mensajero *Mmul-1* el mismo evento de splicing que utiliza el sitio donador interno del exón 2 en la posición +152, conservado en humano y otros primates, y el sitio canónico aceptor del exón 3 (Figura 4-37). lll

En *Papio cynocephalus* sólo detectamos variantes que portan el exón 4I (Figura 4-37). Tres de los cuatro mensajeros amplificados son homólogos a las variantes *C*, *F* y *NC3* humanas (*Pcyn-c*, *Pcyn-f* y *Pcyn-nc3* respectivamente). También detectamos expresión de una nueva variante no codificante, *Pcyn-1*, similar a los mensajeros *Mmul-1* y *Cgue-1* detectados en *Macaca mulatta* y *Colobus guereza* (Figura 4-37).

El resultado en *Macaca fascicularis* permitió detectar la expresión de siete mensajeros diferentes (Figura 4-37). Tres de ellos, portadores del exón 4I, siendo dos estructuralmente equivalentes a las variantes humanas *C* y *NC3* (*Mfas-c Mfas-nc3* respectivamente). El tercero, *Mfas-1*, presenta la misma estructura exónica que las variantes *Mmul-1*, *Cgue-1* y *Pcyn-1*, y como éstos se trata de un mensajero no codificante (Figura 4-37). Utilizando los cebadores diseñados para amplificar las variantes que portasen los exones 4II o 4III se detectó la expresión de cuatro mensajeros, *Mfas-2*, *Mfas-3*, *Mfas-4* y *Mfas-5* (Figura 4-37). Todos ellos presentan los exones 3, 4I y 4II (incluyendo al exón 4III), además de la retención del intrón entre el exón 4I y 4II (Figura 4-37), lo que sugiere que se trata de pre-mensajeros inmaduros. En mensajero *Mfas-2* presenta además un nuevo exón localizado en el interior del intrón 1, similar a lo observado en la variante *Rnor-1* detectada en *Rattus norvegicus* (ver Figura 4.27 en apartado 4.2), aunque no existe homología en la secuencia entre estos nuevos exones en las dos especies. La presencia de este nuevo exón provoca la aparición de un codón de terminación en éste de manera que el transcrito es considerado no codificante o inmaduro. El mensajero *Mfas-3* podría tratarse de un versión alternativa o inmadura de la variante *Mfas-c*, puesto que la secuencia codificante es la misma que en ésta, pese a presentar el exón 4II y la retención del intrón mencionada arriba (Figura 4-37). Los mensajero *Mfas-4* y *Mfas-5*, ambos no codificantes, sólo muestran la mitad 3' del exón 2, similar a lo observado en

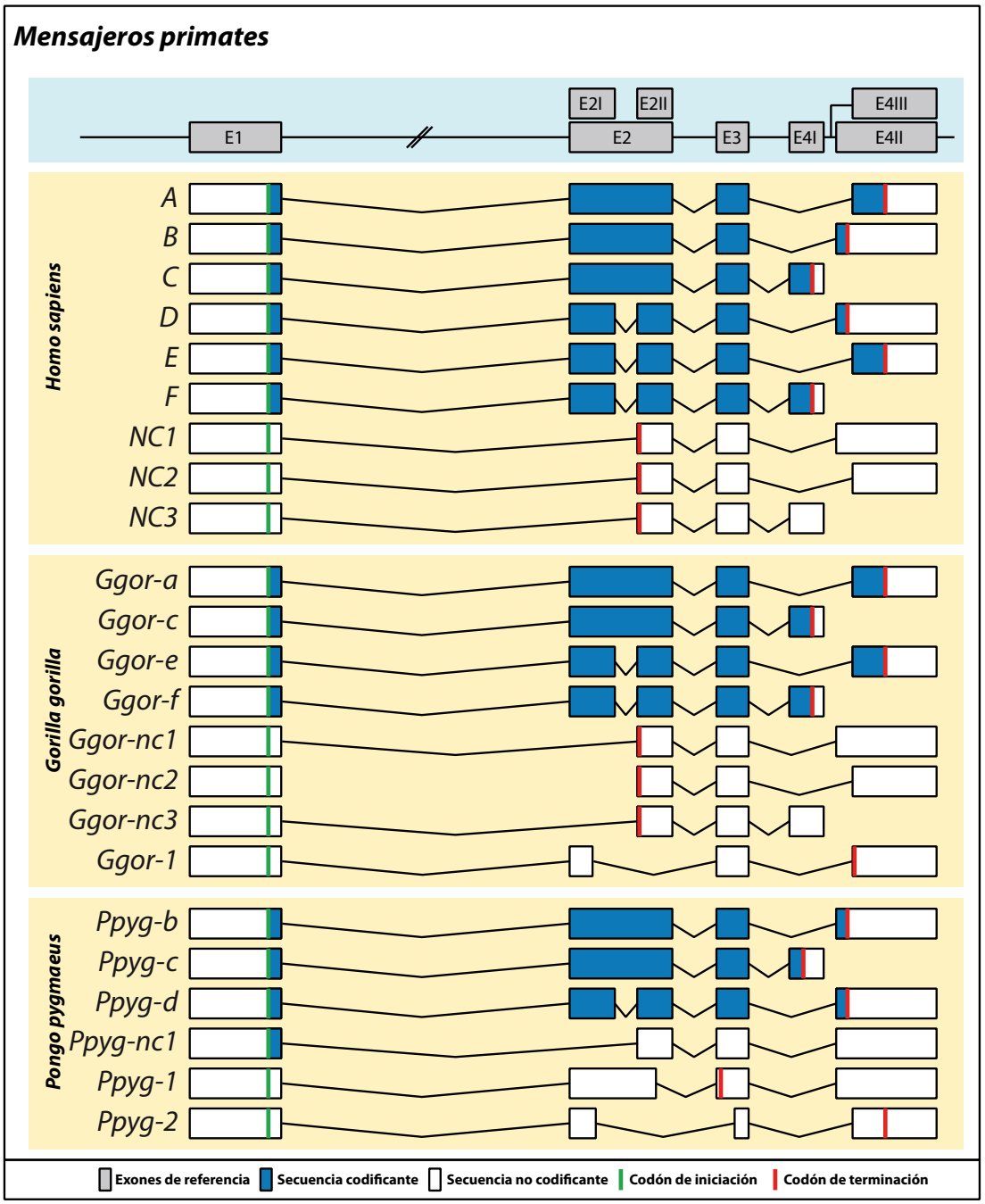


Figura 4-37. Representación de la estructura de los mensajeros identificados mediante PCR anidada en las muestras de sangre disponibles de las especies de primates (Tabla S1).

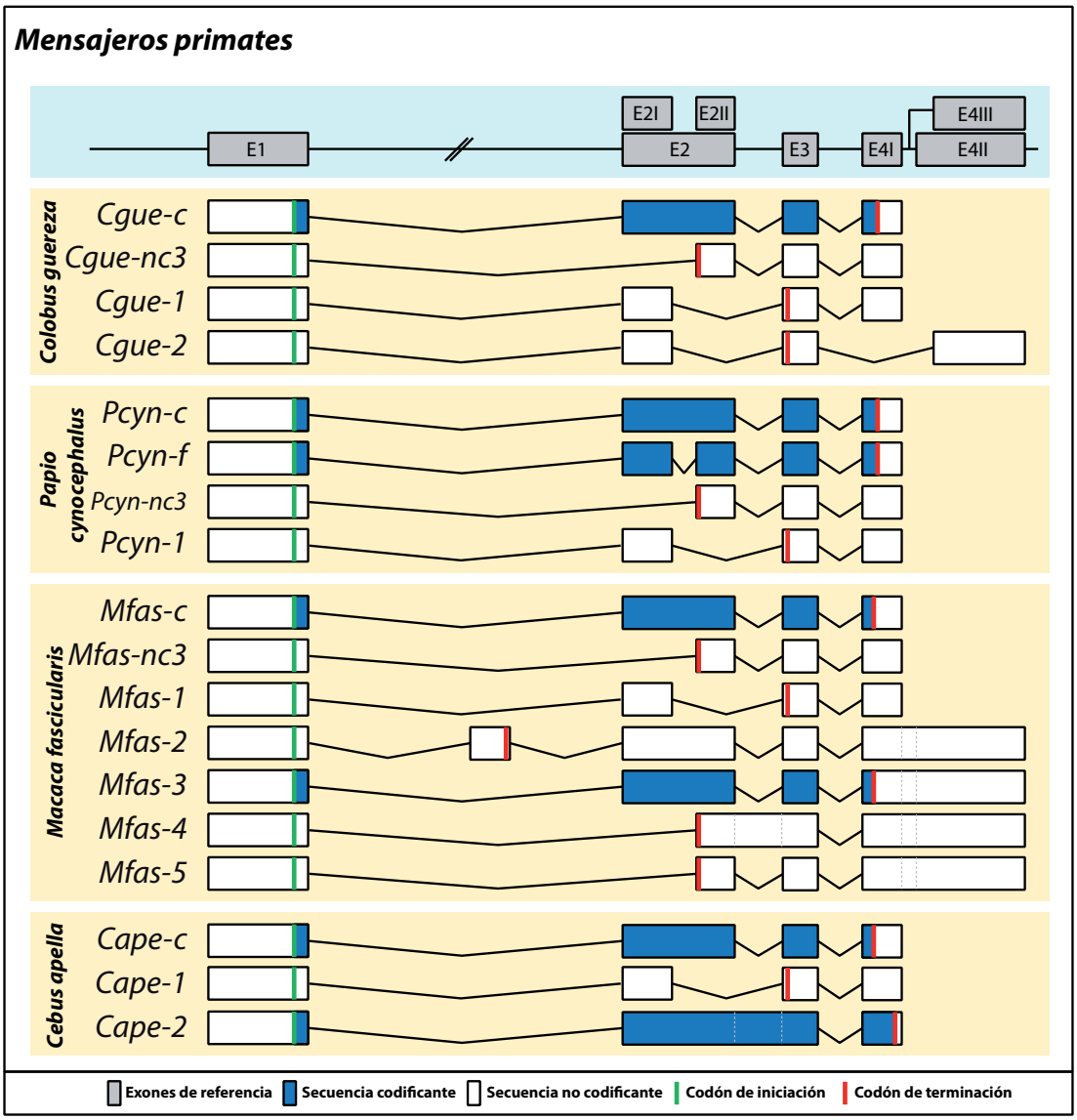


Figura 4-37 continuación. Representación de la estructura de los mensajeros identificados mediante PCR anidada en las muestras de sangre disponibles de las especies de primates (Tabla S1).

el mensajero *Mfas-nc3* y en los homólogos de otras especies. Además, el mensajero *Mfas-4* presenta la retención del intrón 2, lo que sugiere que podrían tratarse de mensajeros inmaduros (Figura 4-37).

Finalmente, en *Cebus apella*, se detectaron tres mensajeros formados con el exón 4I y ninguno que portase los exones 4II o 4III (Figura 4-37). Se detectó la variante canónica, *Cape-c*, y la variante no codificante *Cape-1* estructuralmente similar a las variantes *Mmul-1*, *Cgue-1*, *Ppyg-1* y *Mfas-1*. Además, se detectó una nueva variante potencialmente codificante, *Cape-2*, que presente los mismos exones que la variante

Cape-c además de la retención del intrón 2 (Figura 4-37). La retención del intrón genera un cambio en la fase de lectura, que no provoca la aparición de codones prematuros de terminación, sino que prolonga la secuencia codificante más allá de la posición del codón de parada canónico detectado en el exón 4l. Aunque podría tratarse de un pre-mensajero, la traducción hipotética de este mensajero produciría una versión soluble del receptor, debido al cambio de fase generado (Figura 4-37).

4.3.2 Análisis de las variantes de splicing mediante RNA-seq

Las secuencias procedente de experimentos de RNA-seq disponibles en las bases de datos de los diferentes primates incluidos en este trabajo son muy limitados (Tabla S2). Sólo se han realizado experimentos de RNA-seq en muestras procedentes de *Gorilla gorilla* y *Pongo pygmaeus*, aunque de este último no existe un genoma de referencia, por lo que se utilizó el genoma de su pariente *Pongo abelli* para el alineamiento (ver materiales y métodos). La disponibilidad de datos de RNA-seq en *Pan troglodytes* es ligeramente superior a la de estas dos especies, y se decidió su inclusión en este estudio pese a no disponer de muestras de sangre en las que analizar la expresión del gen *NCR3* mediante PCR anidada. Además, en *Papio anubis*, especie emparentada estrechamente con *Papio cynocephalus*, se encontraron algunas secuencias de experimentos de RNA-seq, por lo que también se incluyó en este apartado. El análisis de las uniones de exones detectadas en humano, descrita en el apartado anterior, se muestra en la Figura 4.38 como referencia para la comparación con los resultados obtenidos en el análisis de los distintos primates.

El análisis de las secuencias de RNA-seq procedentes de *Pan troglodytes* permitió detectar la expresión de las uniones de exones H1, H3, H5 y H6 (Figura 4-38 y Tabla S3), características de las variantes *Ptro-a* y *Ptro-b*, homólogas de las humanas *A* y *B* respectivamente, lo que permite confirmar las predicciones de las bases de datos realizadas en esta especie. Resulta llamativo no observar la expresión de la unión H4, característica de la variante canónica *C* detectada en todas las especies aquí analizadas. Asimismo, tampoco se detectó expresión de la unión de exones H2 en esta especie, que en humano permite definir las variantes *D*, *E* y *F* (Figura 4-38).

En *Gorilla gorilla* se detectaron las uniones de exones H1, H3 y H6 (Figura 4-38 y Tabla S3), que permiten definir la variante *Ggor-a*, detectada mediante PCR en este estudio. No se detectó, en los datos disponibles, expresión de las uniones de exones

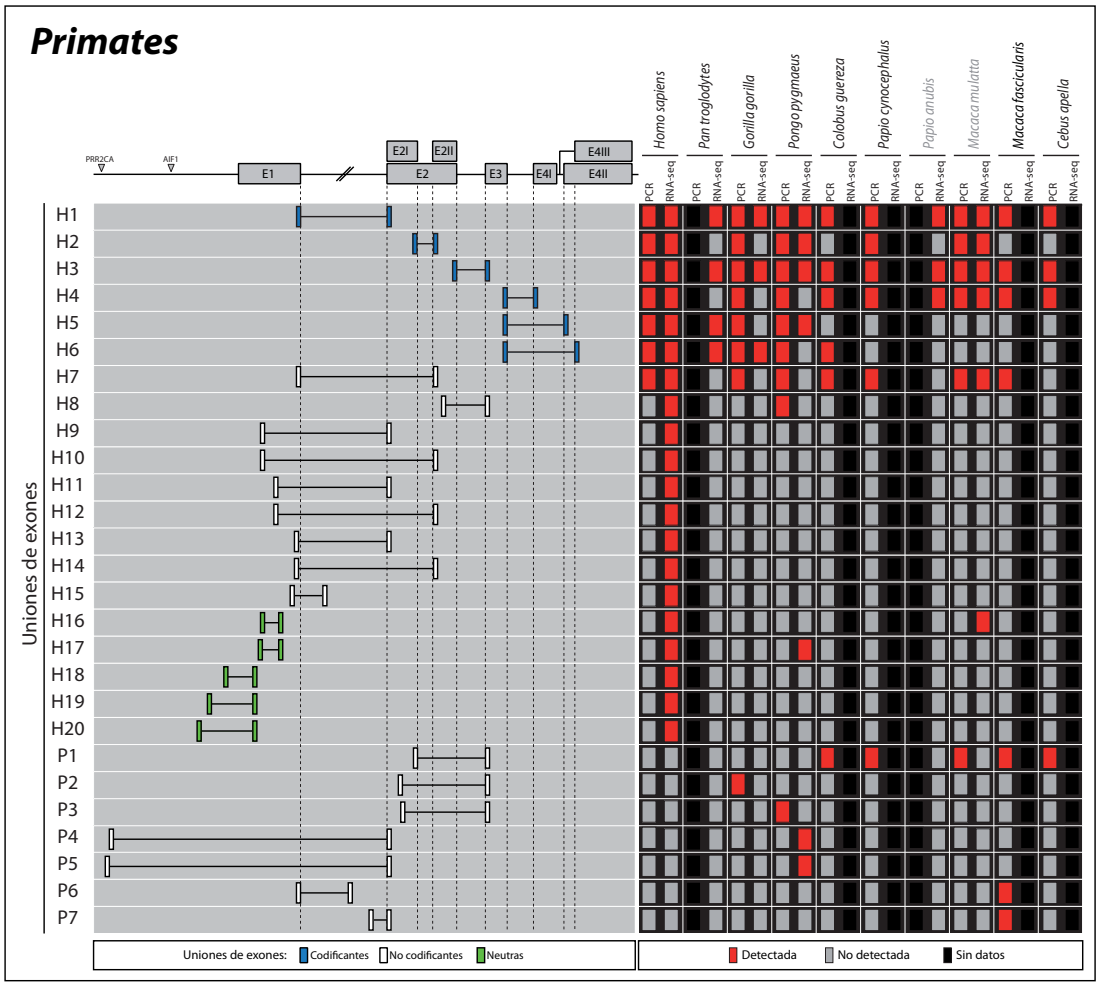


Figura 4-38 Análisis del AS del gen *NCR3* mediante RNA-seq en primates. En el panel de la izquierda se representan esquemáticamente las uniones de exones detectadas mediante por PCR anidada y RNA-seq. La detección de cada una de ellas se muestra en el panel de la derecha, indicando si se detectó por PCR o mediante análisis de secuencias de RNA-seq (Tabla S2). Se ha incluído la información de *Macaca mulatta* para poder facilitar la comparación.

H2, H4, H5 ni de la nueva unión de exones P2 característica del mensajero *Ggor-1*, todas ellas detectadas mediante PCR (Figuras 4-37 y 4-38).

El análisis de las secuencias disponibles de *Pongo pygmaeus* permitió confirmar la expresión de las uniones de exones H1, H2, H3 y H5 (Figura 4-38 y Tabla S3), confirmando la presencia de las variantes *Ppyg-b* y *Ppyg-d* detectadas mediante PCR. Se detectaron además tres uniones de exones, una equivalente a la unión neutra H17 detectada en las muestras de humano, y dos nuevas uniones denominadas P4 y P5 (Figura 4-38). Ambas nuevas uniones de exón muestran cierta similitud a la unión de exones B12 detectada en *Bos taurus*, que define el exón -4, próximo al gen *Btau-prr2ca* y adyacente a una isla CpG (ver Figura 4-32 en apartado 4.2). Los extremos

5’ de las uniones de exones P4 y P5 se localizan en una posición equivalente a la descrita en vaca, cerca del gen *Ppyg-prr2ca* y una isla CpG, sin embargo, los extremos 3’ se localiza en el sitio aceptor canónico del exón 2 a diferencia de los observado en vaca (Figura 4-38). Estas uniones de exones, P4 y P5, excluyen al exón 1, sin que en el nuevo exón definido por éstas exista ningún ATG iniciador en fase, por lo que ambas uniones de exones son consideradas no codificantes.

El resultado del análisis de las secuencias de *Papio anubis*, donde detectamos las uniones de exones H1, H3 y H4 (Figura 4-38 y Tabla S3), nos permite inferir la existencia en esta especie de la variante canónica *Panu-c*. Como en *Papio cynocephalus,* mediante PCR anidada, no detectamos mediante RNA-seq en *Papio anubis* ninguna unión de exones que confirme la expresión de los exones 4II ni 4III (Figura 4-38). Tampoco detectamos expresión en las secuencias de *Papio anubis* de las uniones de exones H7 ni P1, características de las variantes *Pcyn-nc3* y *Pcyn-1* respectivamente, detectadas en *Papio cynocephalus* mediante PCR, y en otras especies de primates (Figura 4-37 y 4-38).

4.3.3 Análisis de las potenciales isoformas proteicas

En humano, tanto por PCR anidada como a través del análisis de secuencias de RNA-seq, ha quedado sobradamente probado que existen seis potenciales variantes codificantes, descritas en detalle en el apartado anterior (Figura 4-23 y 4-39). En *Pan troglodytes* podemos deducir la existencia de dos variantes codificantes, Ptro-a y Ptro-b, que muestran las mismas características estructurales que las homólogas humanas A y B, respectivamente. Ambas isoformas muestran un cambio puntual en el segundo sitio de N-glicosilación (E/G respecto a humano) que no parece afectar a la predicción de dicha modificación post-traduccional (Figura 4-39 y 4.40). Además la isoformas Ptro-a muestra otro cambio puntual (H/Q) en la cola citoplasmática con respecto a humano, que tampoco afecta a la predicción del sitio SH3 detectado en esa región (Figura 4-39 y 4.40).

Los resultados obtenidos en *Gorilla gorilla* permiten definir cuatro potenciales isoformas proteicas; Ggor-a, Ggor-c, Ggor-e y Ggor-f (Figura 4-39). Pese a que existen dos cambios de aminoácido en la región transmembrana, una en la cola citoplasmática de las isoformas Ggor-a y Ggor-e, y otro en el dominio intracelular de las isoformas Ggor-c y Ggor-f (Figura 4-40) los dominios y motivos estructurales son

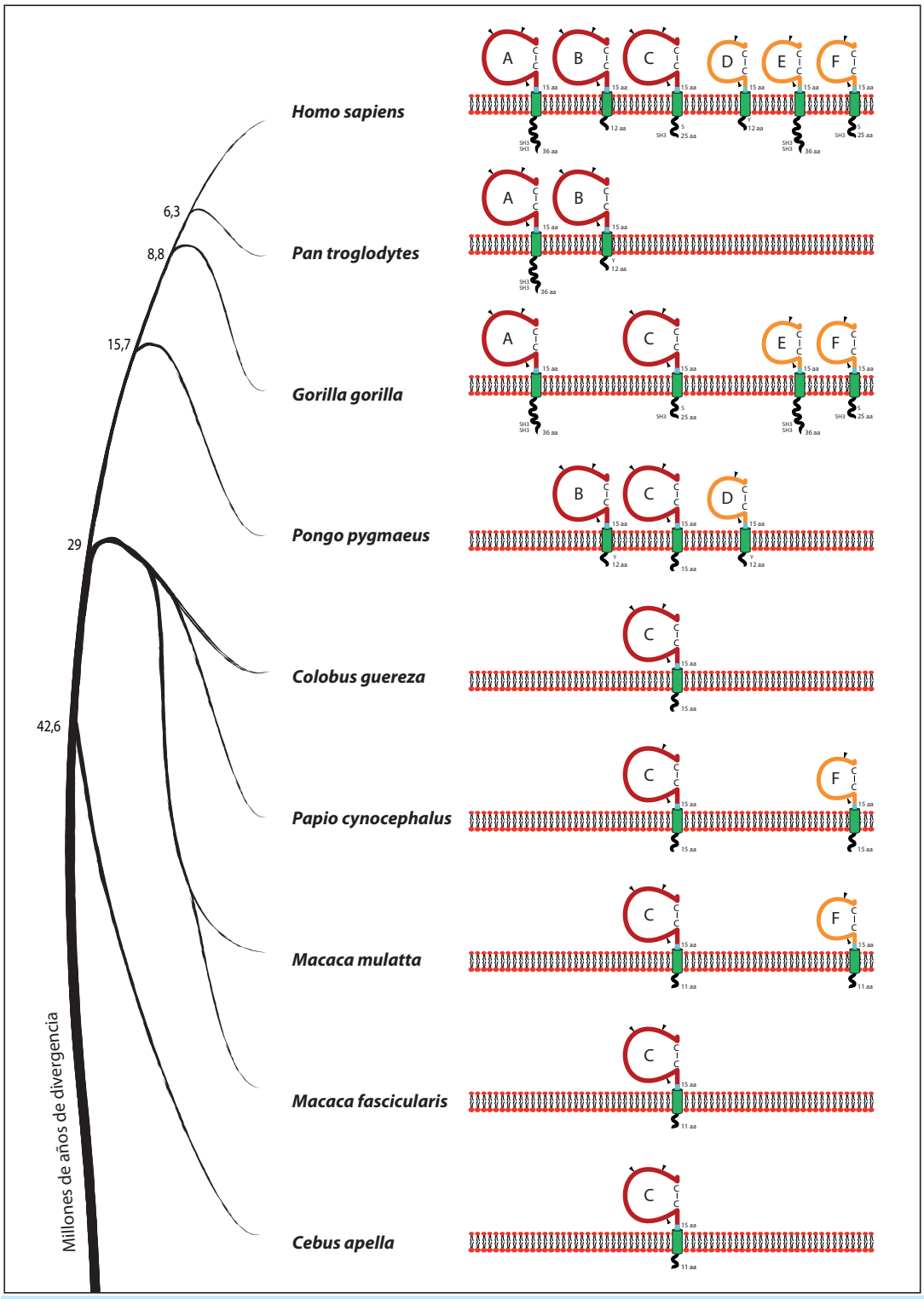


Figura 4-39 Potenciales isoformas proteicas expresadas en cada primate. A la izquierda se muestra la divergencia evolutica de cada una de las especies con respecto a *Homo sapiens*. En la parte derecha se han representado esquemáticamente las isoformas potenciales expresadas en cada especie de primate. Se ha incluido a *Macaca mulatta* para facilitar la comparación.

los mismos que los descritos para las isoformas homólogas humanas.

En *Pongo pygmaeus* detectamos tres potenciales isoformas proteicas; Ppyg-b, Ppyg-c y Ppyg-d (Figura 4-39). En esta especie sólo detectamos dos cambios puntuales con respecto a humano, en las colas citoplasmáticas (Figura 4-40). Sin embargo, detectamos una diferencia importante en cuanto a la longitud de la cola citoplasmática de la isoforma Ppyg-c, de tan sólo 15 aa, diez menos que en humano. Esto se debe al desplazamiento del codón de terminación hacia el 5' del exón comentado en el apartado 4.1. Esta reducción elimina la secuencia adyacente a la serina potencialmente fosforilable detectada en humano, que pese a estar conservada no es reconocida por los programas de predicción (Figura 4-40). Asimismo, el sitio de unión de proteínas de la familia SH3 detectada en la cola citoplasmática homóloga humano no esta presente en esta especie por la reducción de secuencia de ésta (Figura 4-39 y 4-40).

Sólo se puede inferir la existencia de una isoforma en *Colobus guereza*, la isoforma canonica Cgue-c (Figura 4-39). Aunque presenta algunos cambios puntuales, las características estructurales no difieren de la isoforma C humana, salvo por la cola citoplasmática, reducida en esta especie del mismo modo que en *Pongo pygmaeus*, de modo que no presenta el sitio de unión SH3 y la serina no parece ser fosforilable (Figura 4-40).

En *Papio cynocephalus*, además de la isoforma canónica Pcyn-c, que muestra los mismos cambios estructurales que la isoforma Cgue-c y Ppyg-C, con la reducción de la cola citoplasmática, también podemos inferir la existencia de la isoforma Pcyn-f. Como la isoforma F humana, la predicción de dominios revela que la deleción de 25 aa en el interior del dominio inmunoglobulina provoca en la configuración del dominio, pasando a ser de tipo C, en lugar de tipo V como en la variante canónica (Figura 4-39). La variante canónica en este especie, Pcyn-c, presenta un cambio puntual con respecto a humano (N/S) perdiendo la potencial N-glicosilación en esa posición (Figura 4-40).

En *Macaca fascicularis* y *Cebus apella* solo se expresa la isoforma canónica, Mfas-c y Cape-c respectivamente (Figura 4-39). En ambas especies la reducción del tamaño de la cola citoplasmática es aún mayor que en *Pongo pygmaeus*, *Colobus guereza* o *Papio cynocephalus*. Respecto a humano el dominio intracelular en estas especies

es 14 aa menor, perdiendo el sitio potencial de unión de proteínas de la familia SH3 y la serina fosforilable, que aunque está presente, es el último aminoácido de ambas isoformas (Figura 4-39 y 4-40).

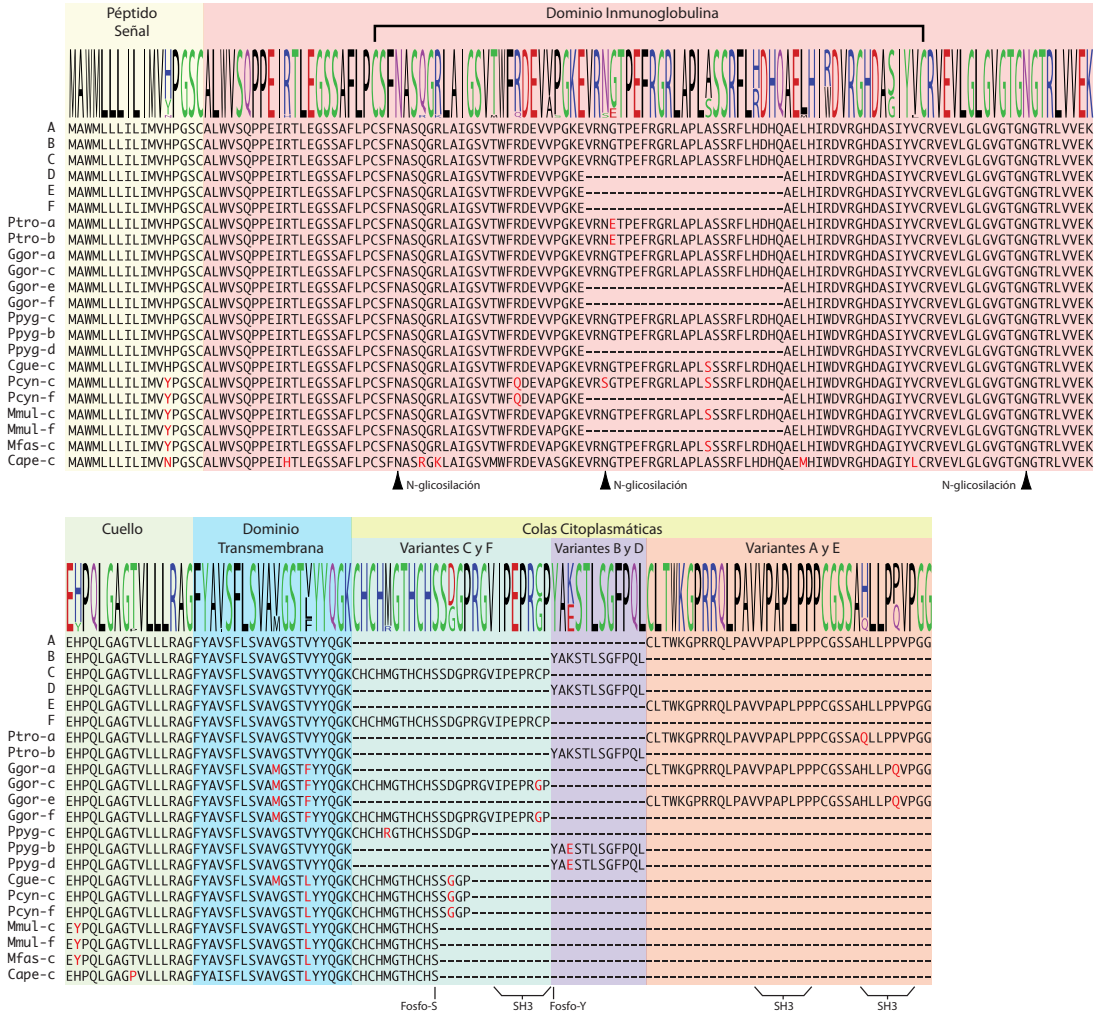


Figura 4-40 Alineamiento de las secuencias de proteínas de las potenciales isoformas generadas mediante AS del gen *NCR3* en las distintas especies de primates analizadas. En la parte superior del alineamiento se muestra la secuencia consenso en la que el tamaño de las letras (aa) representan la frecuencia de aparación en cada posición. Se indica los potenciales dominios estructurales en la parte superior, y los posibles motivos y modificación post-traduccionales en la parte inferior. Los aminoácidos resaltados en rojo representan cambios con respecto a las secuencias de las isoformas detectadas en humano. Se ha incluido a *Macaca mulatta* para facilitar la comparación.

4.4 Caracterización de las variantes de splicing del gen *NCR3* en humano

El análisis detallado del AS del gen *NCR3* demuestra la expresión de varios mensajeros en todas las especies analizadas, contribuyendo a la variabilidad transcriptómica, salvo en *Mus musculus* donde es un pseudogen. Entre las variantes detectadas destaca el alto número de mensajeros no codificantes expresados en todas las especies estudiadas. Sin embargo, la expresión de varios mensajeros potencialmente codificantes, principalmente en primates y en especial en *Homo sapiens*, que expresa seis potenciales variantes codificantes, revela que el AS ejerce un papel crucial en la generación de variabilidad proteómica. Este hecho pone de manifiesto que el AS podría tener importantes consecuencias en las funciones del gen *NCR3* y por lo tanto, se hace necesaria la caracterización a nivel de proteína de dichas variantes. En este apartado se ha desarrollado una serie de experimentos para la caracterización de las principales variantes detectadas en humano que pueden ayudar a comprender las implicaciones funcionales que el AS puede tener sobre el gen *NCR3*.

4.4.1 Cuantificación de las variantes de splicing canónicas mediante PCR cuantitativa (qPCR)

El análisis de las variantes de splicing del *NCR3* en diferentes tejidos humanos demuestra la expresión de nueves mensajeros diferentes, seis codificantes y tres no codificantes. Los resultados obtenidos del análisis de secuencias procedentes de experimentos de RNA-seq sugiere la posible existencia de otras variantes no codificantes minoritarias cuya estructura exónica completa es desconocida. Por ello, se decidió cuantificar, mediante qPCR, las variantes mayoritarias o canónicas en diversos tejidos adultos, fetales y tumorales, así como en líneas celulares. Se diseñaron cebadores específicos para cada variante, además de una pareja de cebadores que amplificase todas las variantes (Figura 3-2 y Tabla 3-7 en materiales y métodos).

Los resultados confirman la información obtenida en el análisis de las secuencias de RNA-seq, siendo las variantes *A*, *B* y *C* mayoritarias en todos los tejidos donde se detecta expresión del gen *NCR3* (Figura 4-41A). La mayor expresión del gen *NCR3* se detecta en bazo adulto y en células NK (línea YT). También se detectaron niveles altos de expresión en tejido pulmonar procedente de un tumor, lo que llevó a establecer una colaboración con el Hospital Universitario Ramón y Cajal (Madrid) para

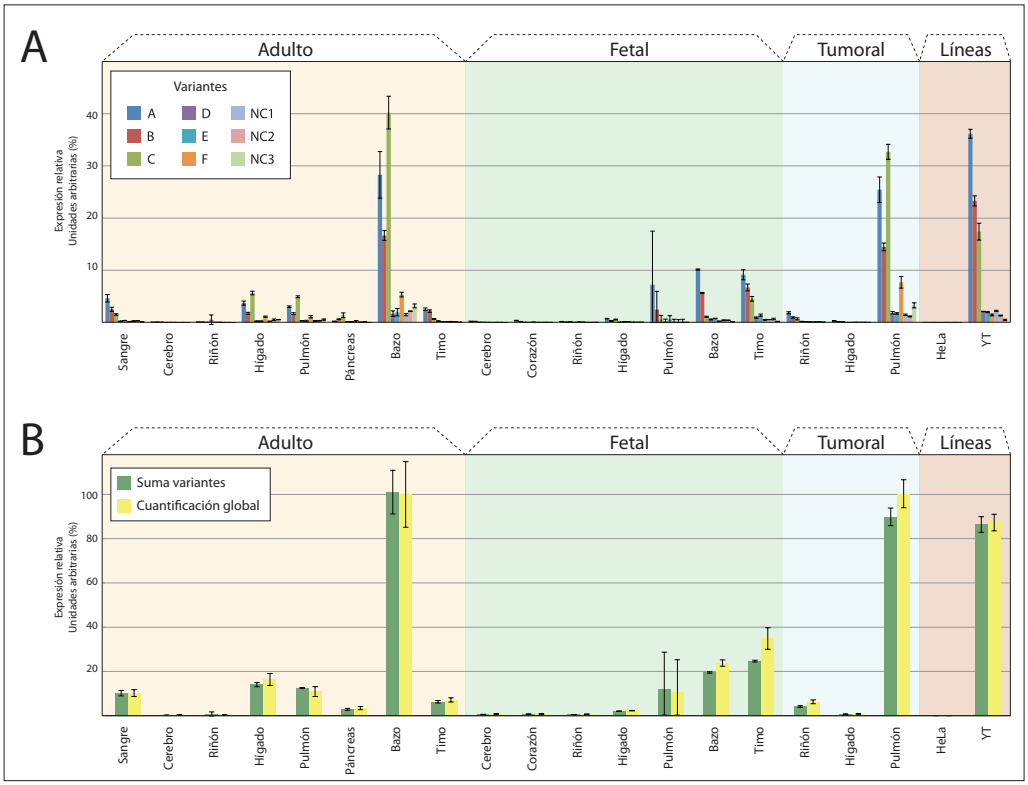


Figura 4-41 Cuantificación de las nueve variantes canónicas del gen *NCR3* humano. A) Cuantificación mediante qPCR de cada una de las variantes en los distintos tejidos y líneas celulas indicadas. B) Comparación de la cuantificación global de la expresión del gen *NCR3* (cebadores comunes para todas las variantes) y el sumatorio de la cuantificación de las nueve variantes canónicas.

la exploración de esta potencial sobre-expresión. El resultado del análisis de varias muestras de tumores pulmonares no resultó concluyente, por lo que se atribuyó la aparente sobre-expresión a la mala calidad de la muestra original, que presentaba signos de degradación (no mostrado). Además de los tejidos señalados, se puede apreciar expresión del gen *NCR3* en varias tejidos adultos, como sangre (PBMCs), hígado, pulmón y timo, y fetales como bazo y timo, aunque los niveles de expresión son relativamente bajos, aproximadamente cuatro veces inferior a los detectados en bazo adulto y en células NKs (Figura 4-41A).

La variante *C* es mayoritaria en tejidos adultos como bazo, hígado y pulmón, seguida por las variante *A* y *B* (Figura 4-41A). Por otro lado, en bazo y timo fetales, en sangre adulta y en células NKs la variante mayoritaria es la *A*, seguida por los mensajeros *B* y *C*. El resto de variantes son minoritarias respecto de éstas, sin embargo, se detectan niveles relativamente altos de expresión de la variante *F* en tejidos donde la expresión de la variante *C* es elevada, es decir, en bazo, hígado y pulmón adultos (Figura 4-41A).

El uso de qPCR absoluta (ver materiales y métodos) permite estimar la expresión global del gen *NCR3* a través de la suma de la expresión de todos los mensajeros cuantificados. Esta estimación se corresponde fielmente con la cuantificación desarrollada mediante el uso de los cebadores comunes para todas las variantes (Figura 4-41B) lo que indica que las potenciales variantes, detectadas mediante RNA-seq, que porten las uniones de exones H9-H15 son muy minoritarias (ver Figura 4-22).

4.4.2 Sobre-expresión, dímeros, glicosilaciones y localización subcelular

Para la caracterización de las potenciales isoformas proteicas humanas se clonaron las secuencias codificantes de éstas en vectores de expresión, incluyendo en el extremo 5' la secuencia codificante del péptido señal de la proteína CD33 seguida de la secuencia codificante del epítopo V5, de modo que éste quedara en el extremo N-terminal de las potenciales (construcciones pV5-NCR3 en materiales y métodos). El análisis, mediante Western Blot, de extractos de células HEK293T transfectadas transitoriamente con las construcciones antes mencionadas permitió la detección de la expresión de las seis potenciales isoformas (Figura 4-42A). El uso de condiciones no reductoras en el desarrollo del ensayo de Western Blot indicó que todas las isoformas presentan capacidad de dimerización, dado que el tamaño aparente en estas condiciones es aproximadamente el doble del detectado en condiciones reductoras (Figura 4-42A).

Los pesos moleculares teóricos de las seis isoformas proteicas humanas, incluyendo el epítopo V5, son: A, 21.49 kDa; B, 19.13 kDa; C, 20.53 kDa; D, 16.29 kDa; E, 18.64 kDa; F, 17.69 kDa. Sin embargo, los pesos moleculares observados mediante Western Blot en condiciones reductoras exceden considerablemente estos tamaños teóricos en todos los casos, sugiriendo la presencia de modificaciones post-traduccionales (Figura 4-42A). Por ello se procedió al análisis del estado de glicosilación de las distintas isoformas proteícas. Para ello se trataron extractos de células HEK293T, transfectadas con las distintas construcciones, con diferentes glicosidasas tras lo que fueron analizadas mediante Western Blot. El resultado demuestra que todas las isoformas sufren una fuerte N-glicosilación, que es responsable de la diferencia de peso molecular observada (Figura 4.-42B). Las isoformas A, B y C, presentan tres potenciales sitios de N-glicosilación, mientras que las isoformas D, E y F, carecen del segundo sitio que presentan las primeras. Sin embargo, la disminución del peso

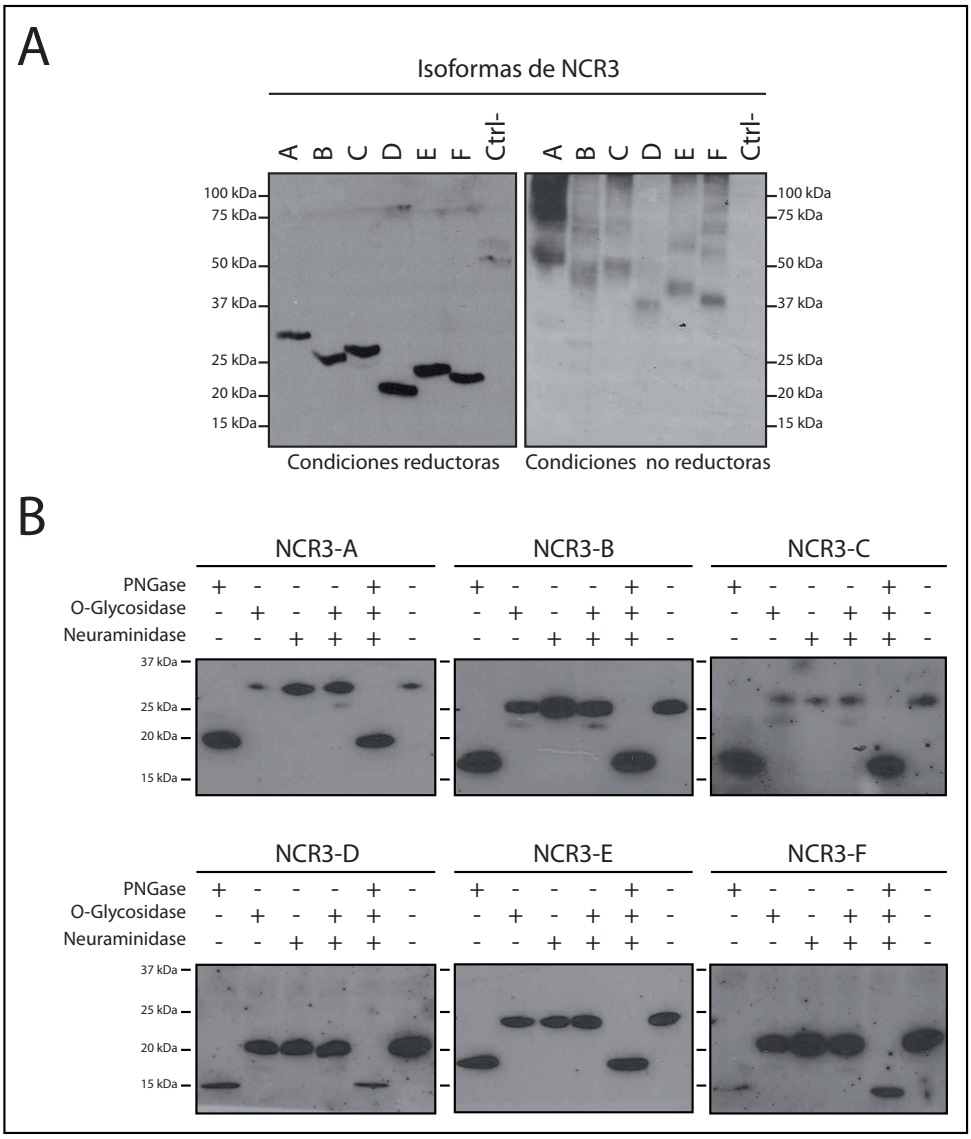


Figura 4-42 Análisis de la expresión de las seis isoformas del gen NCR3. A) Análisis por Western Blot de la expresión de las isoformas humanas en células HEK293T, en condiciones reductoras (izquierda) o no reductoras (derecha). B) Estudio de la glicosilación de las seis isoformas sobre-expresadas en células HEK293T.

aparente observada en todas las isoformas tras el tratamiento con PNGase es similar, lo que podría indicar que el segundo sitio potencial de glicosilación presente en las isoformas A, B y C no está glicosilado o que la potencial glicosilación no es muy ramificada (Figura 4-42B).

La localización subcelular de todas las isoformas se determinó mediante inmunofluorescencia de células COS-7 transfectadas transitoriamente con las construcciones pV5-NCR3. En todos los casos, las células transfectadas con cada

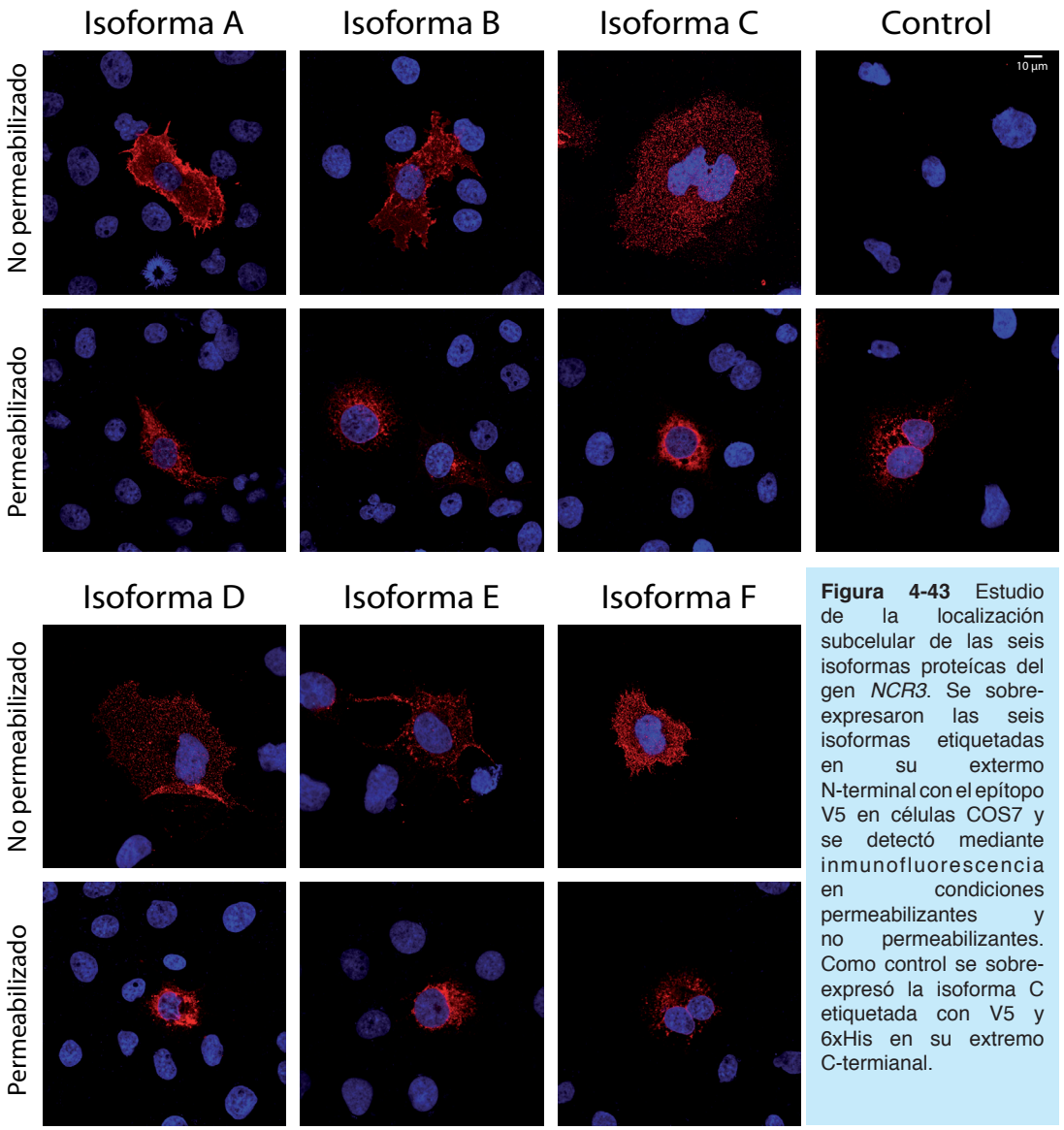


Figura 4-43 Estudio de la localización subcelular de las seis isoformas proteícas del gen *NCR3*. Se sobre-expresaron las seis isoformas etiquetadas en su extermo N-terminal con el epítopo V5 en células COS7 y se detectó mediante inmunofluorescencia en condiciones permeabilizantes y no permeabilizantes. Como control se sobre-expresó la isoforma C etiquetada con V5 y 6xHis en su extremo C-termianal.

una de las variantes codificantes del gen *NCR3* mostraron reactividad al marcaje con el anticuerpo anti-V5, tanto en condiciones permeabilizantes como en condiciones no permeabilizantes (Figura 4-43). Dado que el epítopo V5 se introdujo en el extremo N-terminal, podemos concluir que todas las isoformas humanas se localizan en la membrana plasmática con su extremo N-terminal hacia el exterior, y por tanto, se trata de receptores de membrana tipo I. Como control se utilizó un vector de expresión en el que se clonó la secuencia codificante de la variante *C*, etiquetada en el extremo C-terminal con los epítopos V5 y 6xHis (pNCR3C-V5His). Las células transfectadas con esta construcción no presentaron reactividad mediante inmunofluorescencia cuando se utilizaron condiciones no permeabilizantes. Sin embargo, en condiciones

permeabilizantes se detectó reactividad intracelular en células transfectadas con la construcción control así como con las construcciones pV5-NCR3, mostrando un patrón reticular en todos los casos (Figura 4-43). Este patrón reticular es característico de proteínas de membrana dado que su traducción tiene lugar en el retículo endoplasmático, lugar donde además se produce la glicosilación que hemos podido comprobar que presentan todas las isoformas.

4.4.3 Interacción de las diferentes isoformas proteicas con el ligando celular B7H6

Recientemente, durante el desarrollo de esta Tesis, se ha descrito la existencia de un ligando celular del receptor NKp30 (Brandt et al., 2009), refiriéndose con este nombre a las isoformas A, B o C del gen *NCR3*, es decir, aquellas que presenta un dominio IgV en su porción extracelular. La importancia de este ligando (B7H6), perteneciente a la familia de receptores B7, reside es su especificidad de expresión, ya que ésta se restringe a células tumorales. Esto permite su reconocimiento por las células NKs a través de NKp30 y por lo tanto, posibilitando la eliminación de estas células tumorales. Por ello hemos querido analizar la interacción entre las diferentes isoformas del gen *NCR3* y el ligando celular B7H6. Puesto que las seis isoformas proteicas de *NCR3* presentan sólo dos ectodominios diferentes, IgV en las isoformas A, B y C o IgC en las isoformas D, E y F, los ensayos se centraron en la interacción de éstos dos ectodominios con el ligando B7H6.

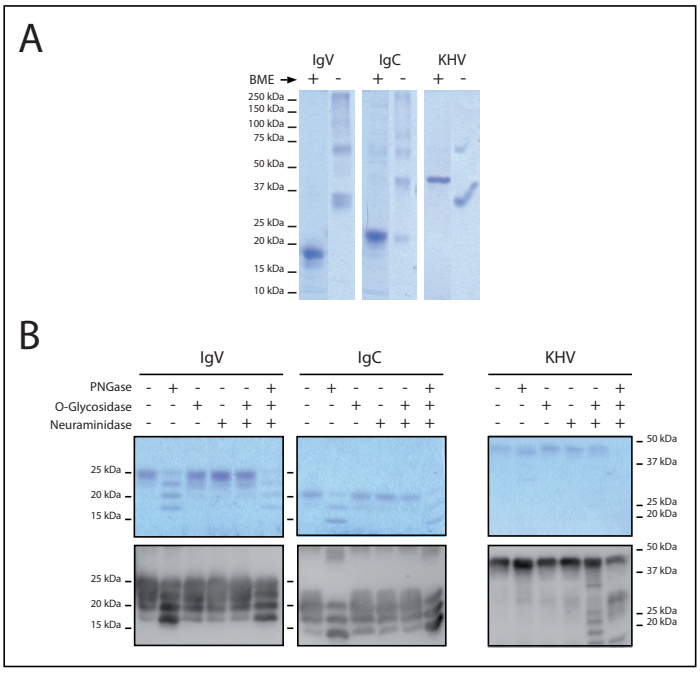


Figura 4-44 Análisis de las proteínas recombinantes producidas en células de insecto. A) Análisis de la capacidad de dimerización mediante electroforesis en condiciones reductoras (BME+) y no reductoras (BME-). B) Estudio del estado de glicosilación de las proteínas recombinantes mediante electroforesis en condiciones reductoras tras el tratamiento con glicosidasas (Detección mediante tinción con azul de Coomassie y mediante western blot con anti-V5)

Debido a la disponibilidad y sencillez de la metodología se escogió un sistema de producción de proteínas recombinantes basado en baculovirus. Se utilizó un vector derivado de pFastBac, en el que previamente se había clonado la secuencia codificante del péptido señal de la melitina y las secuencias codificantes de los epítopos V5 y 6xHis. En este vector se clonaron las secuencias codificantes correspondientes con los ectodominios IgV (isoformas A, B y C) e IgC (isoformas D, E y F) entre la secuencia del péptido señal de la melitina y la secuencia de los epítopos. Las proteínas recombinantes obtenidas se denominaron IgV e IgC respectivamente, refiriéndose al tipo de dominio extracelular. Como control se utilizó la proteína recombinante correspondiente al dominio extracelular de la proteína KHV_ORF_4 (ver materiales y métodos), procedente del herpesvirus de carpa, disponible en el laboratorio, y producido del mismo modo que las anteriores. Por conveniencia se denominó KHV a esta proteína control.

El análisis de las proteínas recombinantes producidas demostró que IgV e IgC formaban dímeros, mostrando el mismo comportamiento observado en células HEK293T (Figura 4-44A). También se comprobó el estado de glicosilación de éstas, observando que ambas presentan N-glicosilaciones, sin embargo, se pueden apreciar bandas de menor peso molecular lo que podría ser indicativo de una glicosilación parcial (Figura 4-44B). En lo que se refiere a la proteína control KHV, no forma dímeros y podría estar ligeramente glicosilada (Figura 4-44).

Dado que se ha descrito que el ligando B7H6 se expresa en líneas celulares tumorales (Brandt et al., 2009), se analizó la unión de las proteínas recombinantes con varias líneas celulares mediante citometría de flujo (Figura 4-45). Las proteínas IgV e IgC se unieron a la superficie de las células HeLa, HEK293T y K562, mientras que la proteína control KHV no mostró ninguna interacción con éstas (Figura 4-45). Esta interacción podría responder a la expresión endógena del ligando B7H6 en estas líneas celulares.

Por ello se procedió a la sobre-expresión del ligando B7H6, transfectando células HEK293T con un vector en el que se clonó la secuencia codificante del gen *B7H6*, incluyendo en su extremo 5' la secuencia codificante del péptido señal de CD33 y del epítopo V5, de modo que la proteína producida quedara etiquetada en el extremo N-terminal (pV5-B7H6). Se comprobó la eficiencia de la transfección con este vector marcando las células transfectadas con el anticuerpo anti-V5 (Figura 4-46 A y C).

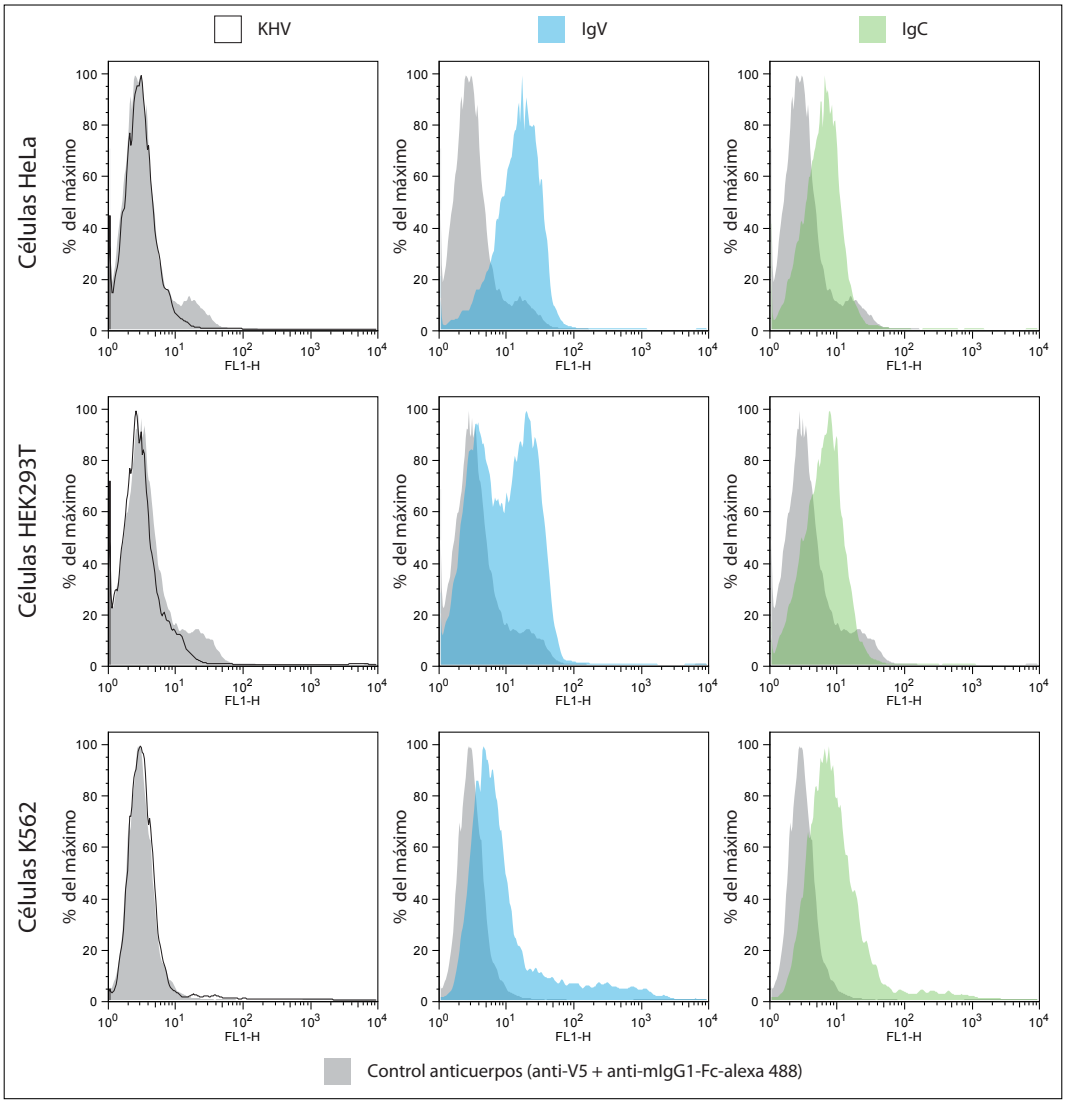


Figura 4-45 Análisis de la interacción de las proteínas recombinantes IgV e IgC con diferentes líneas celulas mediante citometría de flujo. En el panel de la izquierda se muestra la interacción de la proteína no relacionada KHV con cada una de las líneas celulas, histogramas vacíos (líneas negras). En el panel central se muestra la interacción de la proteína IgV con cada línea celular (histogramas azules) y en el de la derecha se muestra la interacción de la proteína IgC (histogramas verdes). Los histogramas grises representan la fluorescencia basal de las células marcadas con los anticuerpos indicados en la figura. Esta figura es representativa de al menos tres réplicas.

Pese a la buena sobre-expresión, la interacción de las proteínas recombinantes IgV e IgC mostraron la misma interacción que la presentada con células transfectadas con un vector vacío (Figura 4-46B). Esto sugería una unión inespecífica de NCR3 a las líneas celulares o una falta de unión entre NCR3 y su ligando B7H6.

Al mismo tiempo que se desarrollaba este trabajo se publicó la estructura cristalográfica del dominio IgV de las isoformas A, B o C del gen *NCR3* (Joyce et al., 2011) y el

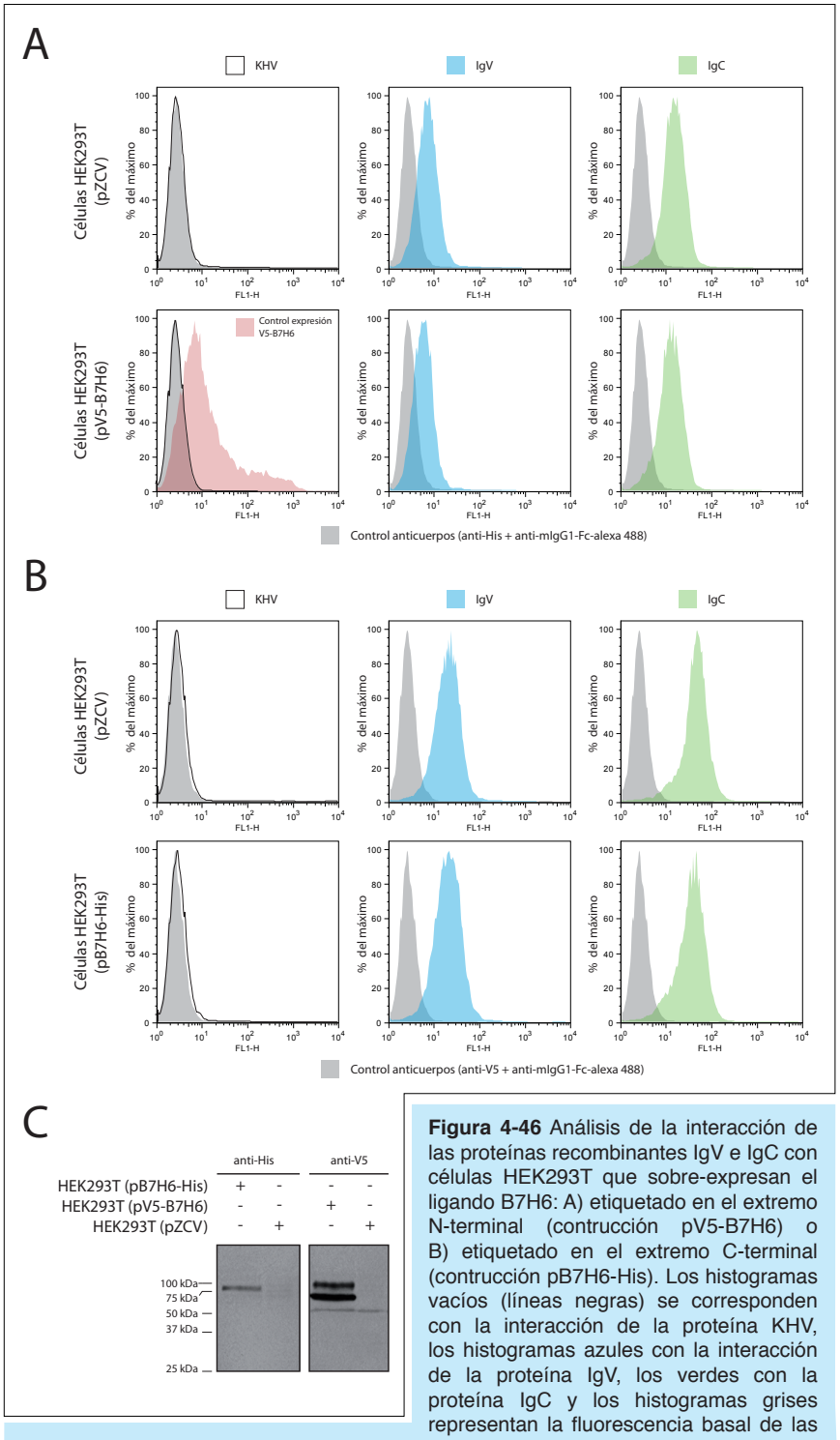


Figura 4-46 Análisis de la interacción de las proteínas recombinantes IgV e IgC con células HEK293T que sobre-expresan el ligando B7H6: A) etiquetado en el extremo N-terminal (contrucción pV5-B7H6) o B) etiquetado en el extremo C-terminal (contrucción pB7H6-His). Los histogramas vacíos (líneas negras) se corresponden con la interacción de la proteína KHV, los histogramas azules con la interacción de la proteína IgV, los verdes con la proteína IgC y los histogramas grises representan la fluorescencia basal de las células transfectadas marcadas con los anticuerpos indicados en la figura. En el panel A se muestra además el marcaje con anti-V5 de las células transfectadas con la construcción pV5-B7H6 como control de la sobre-expresión C) Análisis por western blot de la sobre-expresión del ligando B7H6. Esta figura es representativa de al menos tres réplicas.

modelo de la interacción de éste con el dominio extracelular del ligando B7H6 (Li et al., 2011). En ambas publicaciones se sugiere un modelo de interacción del extremo N-terminal del ligando B7H6 y el extremo N-terminal del dominio IgV. La sobre-expresión del ligando B7H6 etiquetado en su extremo N-terminal con el epítopo V5 podría estar evitando la interacción con los dominio IgV e IgC.

Para resolver la sospecha de la interferencia del epítopo V5 en la interacción de las isoformas de NCR3 y el ligando B7H6, se clonó este último en un vector de expresión incorporando la secuencia codificante del epítopo 6xHis en el extremo 3', quedando éste en el extremo C-terminal de la proteína codificada (pB7H6-His). Sin embargo, y pese a la correcta sobre-expresión de la construcción pB7H6-His en las células HEK293T (Figura 4-46C), el resultado de la interacción de las proteínas IgV e IgC con estas células no mostró un incremento con respecto a la interacción con células

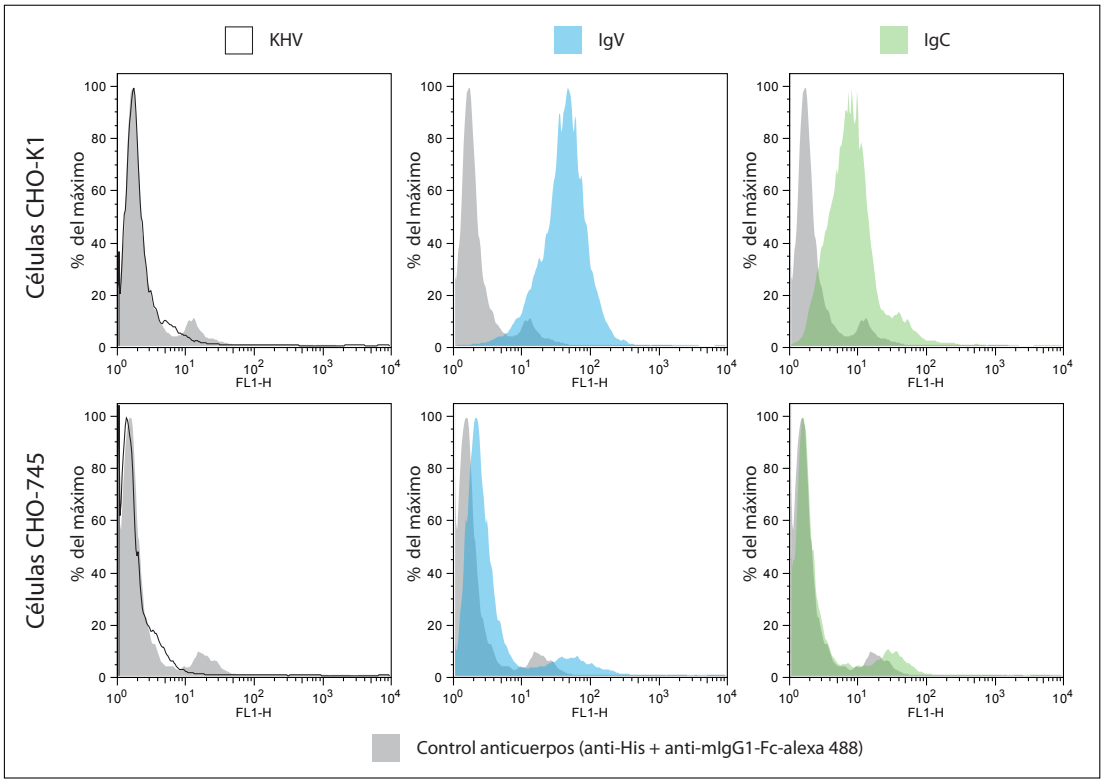


Figura 4-47 Análisis de la interacción de las proteínas recombinantes IgV e IgC con células CHO-K1 y CHO-745 Los histogramas vacíos (líneas negras) se corresponden con la interacción de la proteína KHV, los histogramas azules con la interacción de la proteína IgV, los verdes con la proteína IgC y los histogramas grises representan la fluorescencia basal de las células transfectadas marcadas con los anticuerpos indicados en la figura. En el panel A se muestra además el marcaje con anti-V5 de las células transfectadas con la construcción pV5-B7H6 como control de la sobre-expresión C) Análisis por western blot de la sobre-expresión del ligando B7H6. Esta figura es representativa de al menos tres réplicas.

transfectadas con el vector vacío pZCV (Figura 4-46B). Esto indicaba que el epítopo en el N terminal no era la causa de la falta de unión-especificidad.

La búsqueda de ligandos para el receptor NCR3, o concretamente al ectodominio IgV, durante años se ha centrado en los glucosaminoglicanos (GAGs), en especial en el heparan sulfato (HS), lo que produjo una intensa discusión en las publicaciones científicas (Bloushtain et al., 2004; Hecht et al., 2009; Hershkovitz et al., 2008; Porgador, 2005; Warren et al., 2005). Lejos de resolverse este problema, los resultados publicados indican que la presencia o ausencia de GAGs en la superficie celular tiene un papel importante en la interacción del receptor NCR3 (IgV) con la superficie de las células diana (Bloushtain et al., 2004; Hecht et al., 2009; Hershkovitz et al., 2008; Porgador, 2005; Warren et al., 2005). Por este motivo se decidió analizar la interacción de los dos posibles ectodominios del receptor NCR3 con la superficie de células CHO-K1 y CHO-745, siendo estas últimas deficientes en la producción de GAGs (Esko et al., 1985). Ambas proteínas recombinantes, IgV e IgC, interaccionaron con la superficie de las células CHO-K1 (Figura 4-47), sin embargo, no hubo interacción con la superficie de las células mutantes CHO-745 (Figura 4-47), lo que podría indicar que la interacción observada en células HeLa, HEK293T y K562 se debía a la presencia de GAGs y no a la interacción con el ligando B7H6 expresado de manera endógena en estas células (Brandt et al., 2009).

El estado de glicosilación del receptor NCR3 (IgV) ha sido también materia de debate en el estudio de la interacción de éste con los GAGs de la superficie celular (Hershkovitz et al., 2008), deduciéndose que la correcta glicosilación es esencial para esta interacción y por lo tanto, el sistema elegido para la producción de los receptores recombinantes es crucial para su estudio. Recientemente, se ha publicado un trabajo en el que se estudia en detalle la interacción del receptor NCR3 (IgV), y mutantes de éste deficientes en los tres potenciales sitios de N-glicosilación, con el ligando celular B7H6 (Hartmann et al., 2012). La conclusión de este trabajo asegura que la glicosilación del receptor, esencialmente en el primer sitio de N-glicosilación, es fundamental para la interacción con el ligando B7H6. Por ello, se decidió cambiar el sistema de producción de las proteínas recombinantes a un sistema más próximo a la naturaleza del receptor, de células humanas concretamente células HEK293T, ya que la glicosilación en células de insecto se mostró deficiente. Para ello, se clonaron los dos potenciales ectodominios del receptor NCR3, IgV e IgC, en un

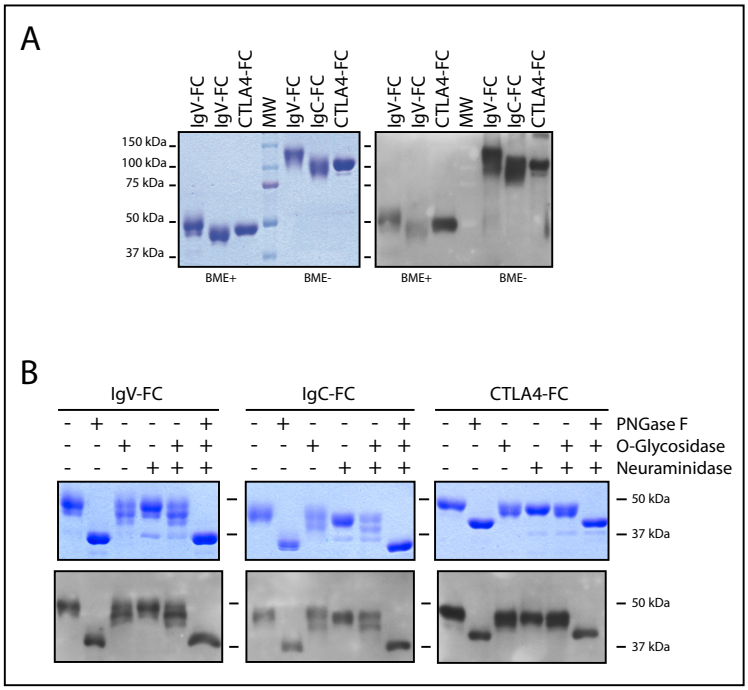


Figura 4-48 Análisis de las proteínas recombinantes producidas en células HEK293T. A) Análisis de la capacidad de dimerización mediante electroforesis en condiciones reductoras (BME+) y no reductoras (BME-). B) Estudio del estado de glicosilación de las proteínas recombinantes mediante electroforesis en condiciones reductoras tras el tratamiento con glicosidasas. Se muestra la tinción de los geles con azul de Coomassie así como la detección con anti-hIgG1-Fc mediante western blot.

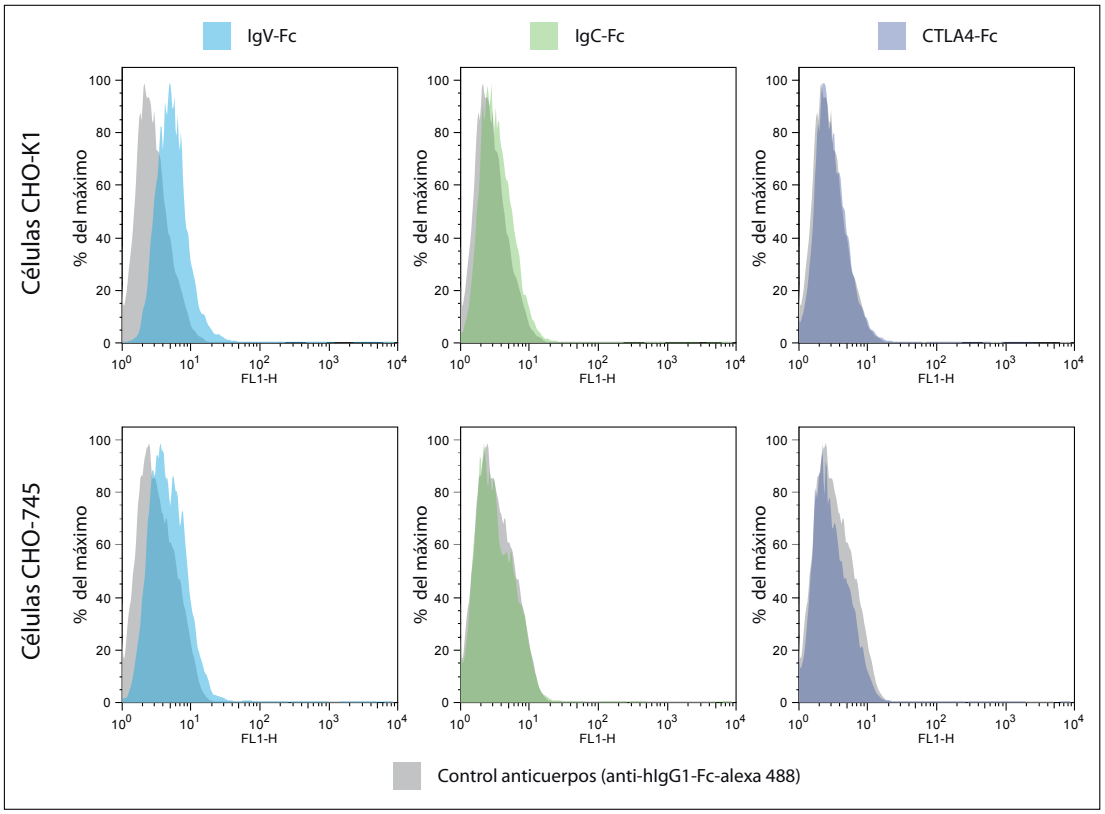


Figura 4-49 Análisis de la interacción de las proteínas recombinantes IgV-Fc, IgC-Fc y CTLA4-Fc con células CHO-K1 y CHO-745. Los histogramas azules con la interacción de la proteína IgV-Fc, los verdes con la proteína IgC-Fc, los morados con CTLA4-Fc y los histogramas grises representan la fluorescencia basal de las células transfectadas marcadas con los anticuerpos indicados en la figura. Esta figura es representativa de al menos tres réplicas.

vector de expresión donde previamente se había clonado la secuencia codificante del péptido señal de la proteína CD33 y la secuencia codificante del dominio constante de la inmunoglobulina humana (hIgG1-Fc). Estas construcciones se transfectaron transitoriamente en células HEK293T y las proteínas recombinantes fusionadas a hIgG1-Fc (IgV-Fc e IgC-Fc) se purificaron del sobrenadante utilizando columnas de proteína A (ver materiales y métodos). Como control se clonó y purificó del mismo modo la proteína recombinante correspondiente con el ectodominio del receptor CTLA4 fusionado a hIgG1-Fc (CTLA4-Fc). Las proteínas recombinantes producidas en células HEK293T mostraron capacidad de dimerización (Figura 4-48A) y el análisis de las glicosilación permitió comprobar la correcta glicosilación de las proteínas IgV-Fc, IgC-Fc y CTLA4-Fc (Figura 4-48B), no observándose bandas adicionales correspondientes con glicosilaciones parciales como las observadas en el caso de las proteínas recombinantes producidas en células de insecto (Figura 4-44B). El análisis de la interacción de las proteínas sobre-expresadas en células HEK293T con las células CHO-K1 y CHO-745 no mostró ninguna interacción significativa (Figura 4-49). Sólo IgV-Fc presentó una ligera interacción con las células CHO-K1 con respecto a las células CHO-745 (Figura 4-49), demostrando que la correcta glicosilación de las proteínas recombinantes evita la interacción inespecífica de éstas con los GAGs de la superficie celular.

Para el estudio de la interacción de estas nuevas proteínas recombinantes se transfectaron células HEK293T con la construcción pV5-B7H6. Como control del ensayo se utilizó la construcción pV5-B7.1, generada del mismo que la anterior pero que corresponde a la secuencia codificante del ligando de CTLA4. La proteína IgV-Fc interaccionó con las células control con la misma intensidad que con las células transfectadas con las construcciones pV5-B7H6 o pV5-B7.1 (Figura 4-50). Sin embargo, la proteína IgC-Fc no mostró ninguna interacción, mientras que la proteína control sólo interaccionó con las células transfectadas con pV5-B7.1 (Figura 4-50).

Del mismo modo que en el ensayo de interacción con las proteínas producidas en células de insecto, este resultado podría deberse a la presencia del epítopo V5 en el extremo N-terminal del ligando B7H6, lo que no parece afectar a la interacción de CTLA4-Fc y V5-B7.1. Por ello, se transfectaron células HEK293T con la construcción pB7H6-His o con una construcción quimera del ligando B7H6 fusionado a GFP en su dominio intracelular, sin epítopos en el extremo N-terminal (pB7H6-GFP). La unión

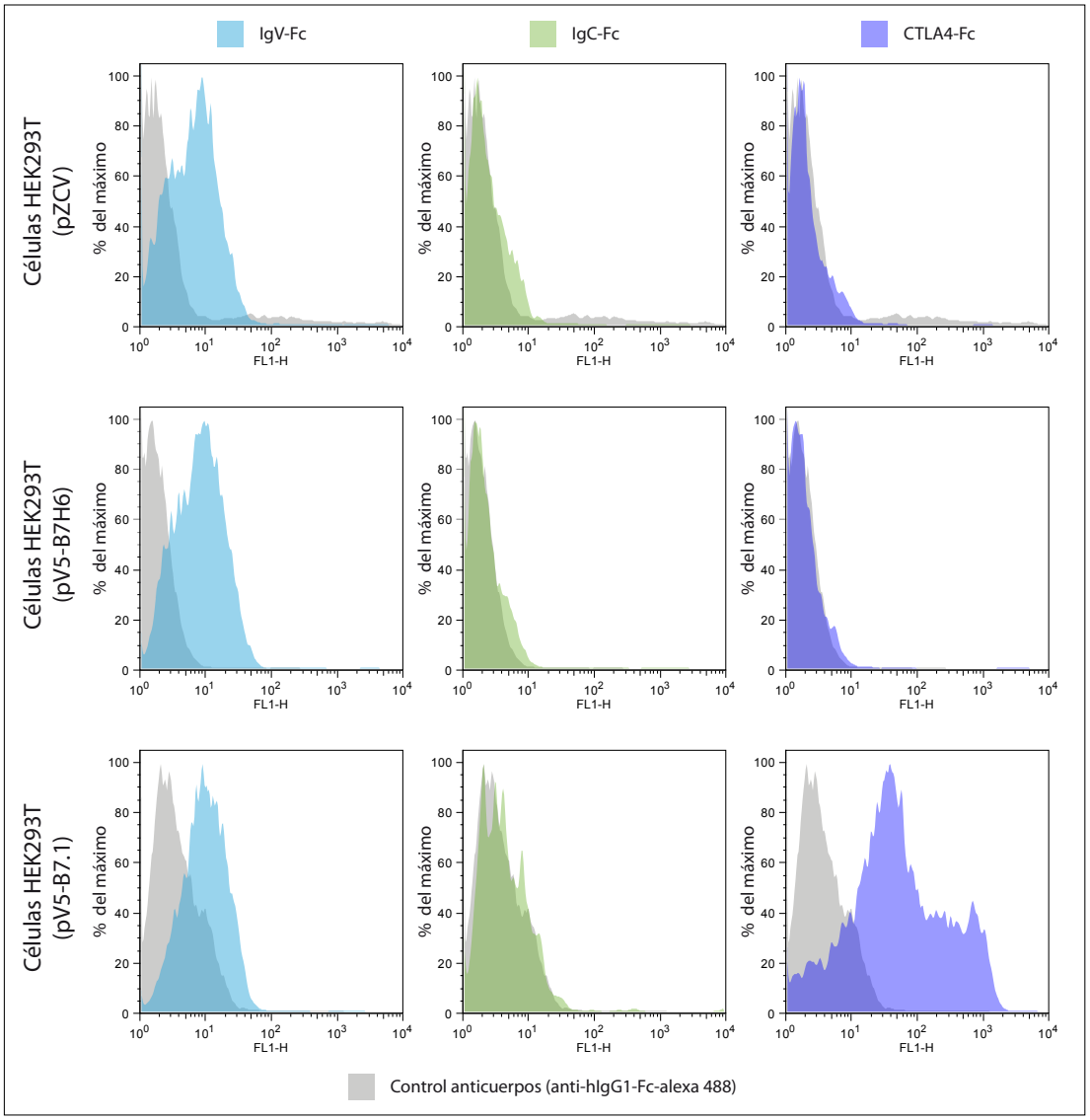
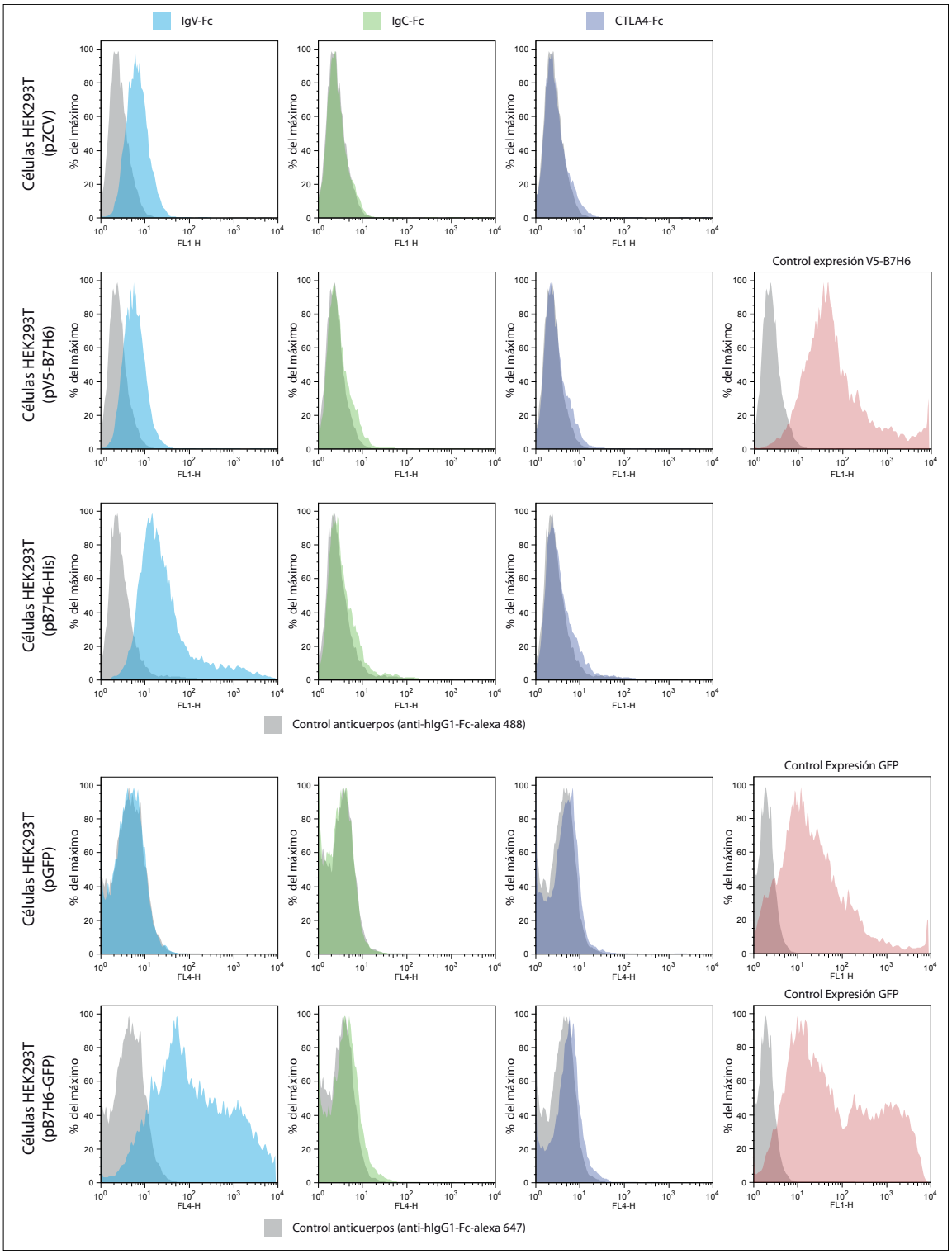


Figura 4-50 Análisis de la interacción de las proteínas recombinantes IgV-Fc, IgC-Fc y CTLA4-FC con células HEK293T que sobre-expresan el ligando B7H6 o el ligando B7.1, ambos etiquetados en el extremo N-terminal. Los histogramas azules con la interacción de la proteína IgV-Fc, los verdes con la proteína IgC-Fc, los morados con CTLA4-Fc y los histogramas grises representan la fluorescencia basal de las células transfectadas marcadas con los anticuerpos indicados en la figura. Esta figura es representativa de al menos tres réplicas.

Figura 4-51 (Página siguiente) Análisis de la interacción de las proteínas recombinantes IgV-Fc, IgC-Fc y CTLA4-FC con células HEK293T que sobre-expresan el ligando B7H6, etiquetado en el extremo N-terminal (pV5-B7H6) o en el C-terminal con el tag 6xHis (pB7H6-His) o fusionado con GFP en el extremo C-terminal (pB7H6-GFP). Los histogramas azules muestran la interacción de la proteína IgV-Fc, los verdes con la proteína IgC-Fc y los morados con CTLA4-Fc. Los histogramas grises representan la fluorescencia basal de las células transfectadas marcadas con los anticuerpos indicados en la figura. Los histogramas rojos (derecha) muestran el marcaje de las células transfectadas con la contrucción pV5-B7H6 con el anticuerpo anti-V5 o la expresión de GFP y B7H6-GFP como control de sobre-expresión. Esta figura es representativa de al menos tres réplicas.


IgV-Fc
IgC-Fc
CTLA4-Fc
Células HEK293T (pZCV)
Células HEK293T (pV5-B7H6)
Control expresión V5-B7H6
Células HEK293T (pB7H6-His)
% del máximo
FL1-H
Control anticuerpos (anti-hIgG1-Fc-alexa 488)
Células HEK293T (pGFP)
Control Expresión GFP
Células HEK293T (pB7H6-GFP)
FL4-H
Control anticuerpos (anti-hIgG1-Fc-alexa 647)

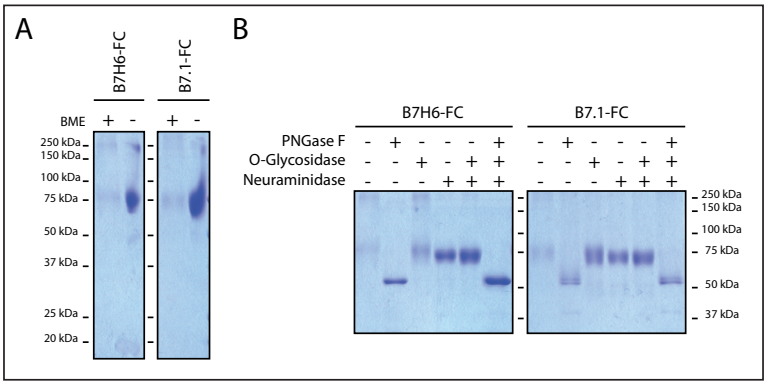


Figura 4-52 Análisis de las proteínas recombinantes producidas en células HEK293T. A) Análisis de la capacidad de dimerización mediante electroforesis en condiciones reductoras (BME+) y no reductoras (BME-). B) Estudio del estado de glicosilación de las proteínas recombinantes mediante electroforesis en condiciones reductoras tras el tratamiento con glicosidasas.

de la proteína IgV-Fc se vio incrementada cuando las células sobre-expresaron la construcción B7H6-His o B7H6-GFP (Figura 4-51). Sin embargo, como en el experimento anterior, la proteína IgC-Fc no mostró interacción con las células transfectadas con pV5-B7H6, pB7H6-His o pB7H6-GFP, lo que sugiere que el ligando de las isoformas D, E y F no es B7H6.

Para confirmar que las isoformas D, E y F, que comparten el ectodominio tipo IgC, no interacciona con el ligando B7H6, se sobre-expresaron en células HEK293T construcciones quimera de las isoformas A (IgV) y E (IgC) del receptor NCR3, así como de CTLA4, fusionados a GFP en su dominio intracelular, sin epítopos en el extremo N-terminal (pNCR3A-GFP, pNCR3E-GFP y pCTLA4-GFP respectivamente). Además, se utilizaron las proteínas purificadas recombinantes B7H6-Fc y B7.1-Fc, correspondientes a los ectodominios de los ligandos B7H6 y B7.1 respectivamente, fusionados a hIgG1-Fc, producidas en células HEK293T (Figura 4-52). La proteína recombiante B7.1-Fc mostró una fuerte interacción con las células que sobre-expresaron el receptor CTLA4-GFP, no detectándose interacción con otras células. Finalmente, la proteína B7H6-Fc mostró una ligera interacción con las células que sobre-expresaron el receptor NCR3A-GFP, no detectándose unión a las células transfectadas con NCR3E-GFP (Figura 4-53). El ligero incremento de la interacción de B7H6-Fc con las células que sobre-expresan NCR3A-GFP podría deberse a la baja expresión de este receptor, sin embargo la expresión del receptor NCR3E-GFP es mayor y no se observó interacción de B7H6-Fc con estas células, confirmando que B7H6 no es el ligando de las isoformas D, E y F de NCR3 (Figura 4-53).

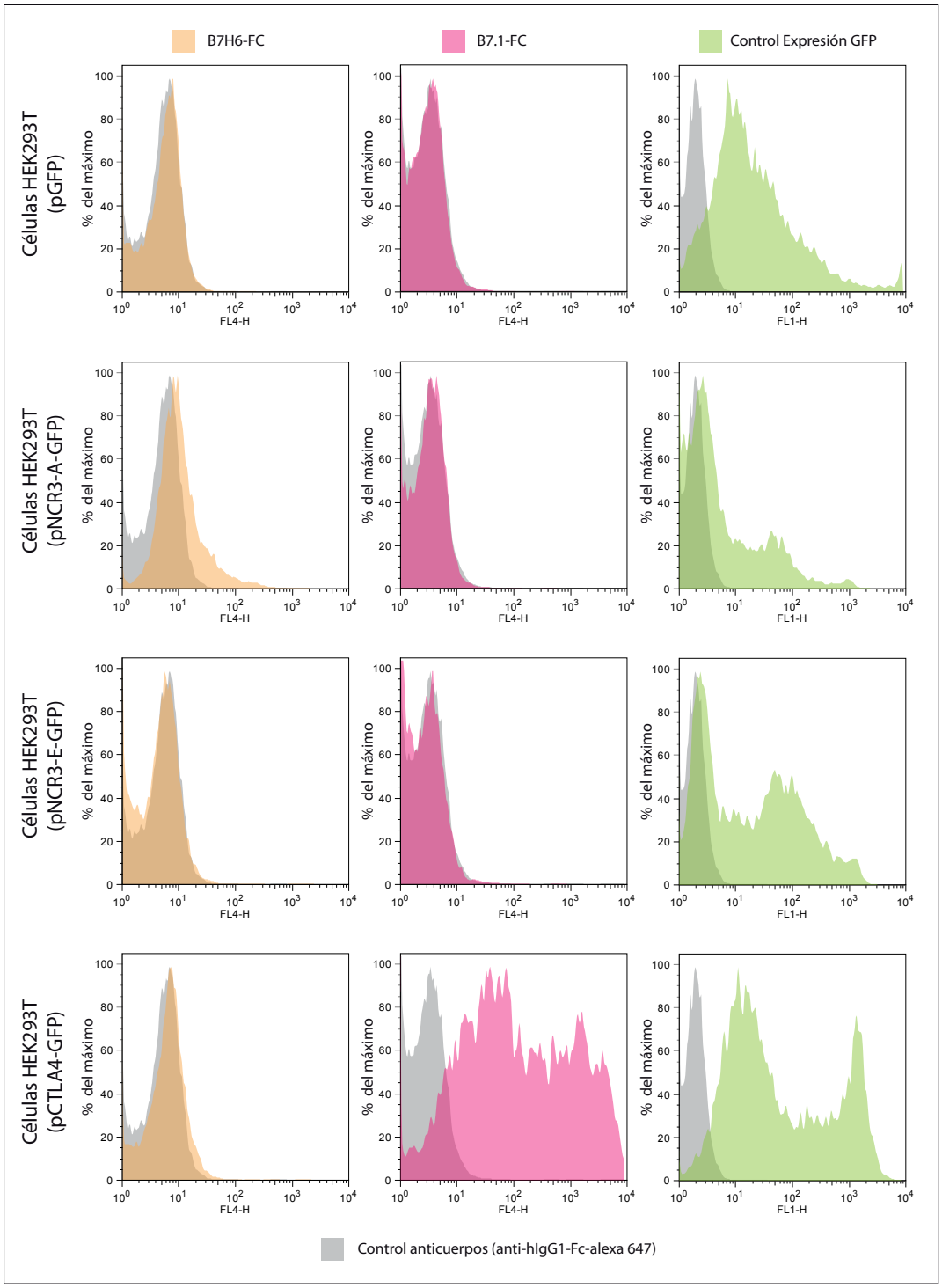


Figura 4-53. Análisis de la interacción de las proteínas recombinantes B7H6-Fc y B7.1-FC con células HEK293T que sobre-expresan los receptores NCR3-A, NCR3-E y CTLA4 fusionados con GFP en su extremo C-terminal. Los histogramas naranjas representan la interacción de la proteína B7H6-Fc y los rosas con la proteína B7.1-Fc. Los histogramas grises representan la fluorescencia basal de las células transfectadas marcadas con los anticuerpos indicados en la figura. Los histogramas verdes (derecha) muestran la expresión de GFP como control de sobre-expresión. Esta figura es representativa de al menos tres réplicas.

5. Discusión

Discusión

5.1 Anotación de genomas

La secuenciación del genoma humano (Lander et al., 2001; Venter et al., 2001) fue una revolución científica y social, que proporcionó una inmensa cantidad de información que sentó las bases para majorar la comprensión de los mecanismos moleculares subyacentes a todos los procesos biológicos. Este avance supuso un impulso para la comunidad científica que se volcó en la obtención de la secuencia completa del genoma de otros organismos. En los últimos años, gracias al desarrollo de la tecnología, y especialmente al uso de la secuenciación masiva, el número de genomas disponibles ha crecido de manera exponencial (Liolios et al., 2008). Actualmente, existen 7.382 genomas disponibles, de los cuales 7.071 son de organismos procariotas (6.837 de bacteria y 234 de arqueas) y 311 de organismos eucariotas, incluyendo 21 mamíferos (12 genomas completos y 9 borradores) (Tabla 5-1) (Pagani et al., 2012). Además, existen 24.506 proyectos de secuenciación en curso entre los que cabe destacar 138 genomas de diferentes mamíferos (Tabla 5-1) (Pagani et al., 2012). La disponibilidad de toda esta información y su comparación entre diferentes especies ha mejorado nuestra compresión de la bases moleculares que determinan la complejidad y diversidad de las especies, permitiendo desterrar viejos paradigmas. En especial, en lo que se refiere a mamíferos, se ha podido comprobar que el número de genes no es un parámetro fundamental en la generación de la diversidad existente entre estas especies, siendo muy similar en todas ellas, y encontrando además un alto grado de conservación entre éstos (Tabla 1-1, introducción) (Barbosa-Morais et al., 2012; Blencowe, 2006; Merkin et al., 2012). Este hallazgo ha sugerido la existencia de otros mecanismos moleculares que pudieran explicar la diversidad entre las especies de mamíferos, siendo el AS uno de los más relevantes, habiéndose descrito su influencia

	Disponibles		Incompletos
	Completos	Borradores	
Arqueas	160	74	411
Bacterias	2337	4500	18656
Eucariotas (mamíferos)	151 (12)	160 (9)	5439 (138)
Total	7382		24506

Tabla 5-1. Número de genomas secuenciados y proyectos de secuenciación en curso (incompletos). La información se obtuvo de las base de datos *Genomes OnLine Databases* (GOLD, http://www.genomesonline.org/). No se incluyen los multiples genomas secuenciados de una misma especie, como por ejemplo lo distintos genomas humanos secuenciados.

en diversos trabajos (Barbosa-Morais et al., 2012; Blencowe, 2006; Merkin et al., 2012).

Por otro lado, el avance de la tecnología de secuenciación masiva ha supuesto el abandono de metodologías tradicionales como la generación y secuenciación de librerías de *Expressed Sequence Tag* (EST) o la secuenciación de mensajeros completos, por lo que el crecimiento de estas secuencias en las bases de datos ha sufrido un retroceso a favor de las procedentes de esta novedosa tecnología (Benson et al., 2013). Además, hay que tener en cuenta la existencia previa de un gran sesgo en el número de secuencias de mensajeros y EST disponibles en cada especie, encontrando un mayor número de éstas en humano, ratón y otras especies usadas frecuentemente como modelos en investigación (Tabla 1-1, Introducción) lo que dificultaba realizar verdaderos estudios comparativos-cuantitativos. Por lo tanto, para la anotación de los genomas, necesaria para la interpretación de la información contenida en éstos, se recurre al uso de métodos predictivos como el análisis del contenido en GC, la búsqueda de sitios consenso de splicing o al análisis de la homología con las secuencias de mensajeros descritas en otras especies.

El uso de este tipo de herramientas bioinformáticas son de gran utilidad para localizar y anotar los genes contenidos en los genomas, sin embargo, como hemos observado en el caso de NCR3, el uso de estas herramientas puede producir errores que dificulten la interpretación de estas anotaciones. En este sentido, se detectaron errores en la anotación del gen *NCR3* en *Gorilla gorilla* y en *Sus scrofa*. En el primero, la presencia de un hueco sin secuenciar en el genoma disponible (Scally et al., 2012), hizo que se predijera en el NCBI la existencia de un exón 1 aberrante y en Ensembl la expansión artificial del exón 2 para obtener una secuencia codificante completa. Estos errores se han subsanado secuenciando parte del exón 1 y utilizando la

secuencia de dos grandes fragmentos del cromosoma 6 depositados en las bases de datos, que han sido recientemente utilizados para obtener la secuencia completa de la región del MHC de clase III en esta especie (Wilming et al., 2013). En *Sus scrofa*, pese a la disponibilidad de la secuencia completa del DNA genómico del gen *Sscro-ncr3*, en el NCBI se predice la existencia de un mensajero potencialmente codificante, compuesto por los cuatro exones detectados en esta especie, y al que se le introducen diversas modificaciones para obtener una secuencia de proteína, que a tenor de los resultados obtenidos en este trabajo es inviable. También se ha observado un deterioro de la anotación de los pseudogenes, que en caso del gen *Mmus-ncr3* (*Mus musculus*) ha llevado a la desaparición del exón 4 y la alteración de los límites de otros exones en las bases de datos.

Por otro lado, la gran homología de secuencia observada en el gen *NCR3*, así como la de su entorno genómico (MHC de clase III), ha facilitado la predicción, en las bases de datos, de las potenciales variantes de splicing de cada especie basadas en las variantes descritas en otras especies.. En *Macaca mulatta* se había predicho la existencia de variantes equivalentes a las humanas *A*, *B* y *E* que portan los exones 4II y 4III, sin embargo, el análisis de la expresión del gen *Mmul-ncr3* en diversos tejidos, mediante PCR anidada y mediante secuenciación masiva, revela que estas variantes codificantes no se expresan en esta especie, probablemente por la ausencia de un sitio de poliadenilación consenso para los exones 4II y 4III, lo que parece confirmarse también con lo observado en *Macaca fascicularis*. De manera similar se ha predicho la existencia de las variantes *Panu-a* y *Panu-b* en *Papio anubis* en el que tampoco existe una secuencia de polyadenilación consenso. La extrapolación de esta ausencia de secuencia a su pariente *Papio cynocephalus*, en el que no se ha detectado experimentalmente expresión de dichos exones, podría indicar que las predicciones de las variantes *Panu-a* y *Panu-b* realizadas en *Papio anubis* son incorrectas. En este sentido se propone un modelo de evolución de los exones finales alternativos en el que primero aparecen los sitios de splicing y por lo tanto, los exones, pero éstos no son funcionales hasta que los sitios de poliadenilación se desarrollan por completo. En *Colobus guereza* y *Cebus apella* se podría especular, en función del modelo propuesto, sobre la ausencia de sitios consenso de poliadenilación, dado que en estas especies no se ha detectado expresión de variantes de splicing codificantes portadoras de los exones 4II y 4III. Por lo tanto, la anotación de los sitios consenso de poliadenilación podrían ser una herramienta útil para mejorar la predicción de las

variantes de splicing, que debería tomarse en consideración.

Además de la ausencia de sitios de poliadenilación, las variantes de splicing que se expresan en cada una de las especies podría estar regulada por la presencia de secuencias potenciadoras o inhibidoras, o mediante la expresión diferencial de los distintos factores regulados que se unen a estas secuencias. Aunque la conservación de secuencia del gen *NCR3* entre las especies incluidas en este trabajo es muy alta (ver Figura 4-15, Resultados), incluyendo la alta homología en los sitios de splicing (ver Figura 4-16, Resultados), pequeños cambios podrían modular la fortaleza del reconociendo de los sitios activadores e inhibidores, que en definitiva podría modular la expresión de las distintas variantes de splicing. En este sentido, el uso de las variantes de splicing detectadas en una especie para la predicción de las variantes que se expresan en otras especies podrían ser erróneas. Por ejemplo, en *Gorilla gorilla* se ha predicho la expresión de la variante *Ggor-b*, basada en la variante humana B, aunque no se ha detectado su expresión experimentalmente, ni tampoco se detectó expresión en *Pongo pygmaeus* de la hipotética variante *Ppyg-a*, cuya homóloga *Pabe-a* si ha sido predicha en *Pongo abelli.* Por el contrario, la predicción de la variante humana no codificante *NC6*, basada en un mensajero descrito en *Macaca mulatta*, podría existir dado que se ha detectado, mediante RNA-seq, la unión de exones característica de esta variante. Por lo tanto, y pese a que el uso de herramientas bioinformáticas resultan imprescindibles en la actualidad, sigue siendo necesario realizar estudios experimentales detallados de las variantes de splicing expresadas por cada gen, como el presente, que nos permitan obtener una visión más real del alcance del AS y sus potenciales implicaciones funcionales.

5.2 Secuenciación masiva y PCR anidada

El avance de la tecnología ha puesto a la secuenciación masiva y en particular a la aplicación de RNA-seq a la cabeza de los estudios transcriptómicos, relevando a tecnologías obsoletas como los microarrays (Pan et al., 2008). Mediante RNA-seq es posible la secuenciación de fragmentos de RNA (~200-400 pb) procedentes del transcriptoma completo de un organismo, obteniendo una visión global y cuantitativa de éste, permitiendo observar cambios en diferentes situaciones fisiológicas y/o patológicas. Sin embargo, pese a las virtudes de esta tecnología, también posee limitaciones, sobre todo aquellas derivadas de la preparación de las muestras y del análisis computacional de la información obtenida. En primer lugar, el RNA que va

a ser secuenciado es fragmentado de manera aleatoria, tras lo que se seleccionan los fragmentos adecuados para dicho proceso de secuenciación (~200-400 pb). Asumiendo que dicha fragmentanción es completamente aleatoria, se puede deducir que los extremos 5' y 3' de los mensajeros van a estar infra-representados en las secuencias obtenidas, ya que los fragmentos procedentes de estos extremos serán de menor tamaño y habrán sido descartados durante el proceso de selección por tamaño. Tras esta selección, el RNA es sometido a una retrotranscripción con hexámeros al azar, tras lo que se une a cada fragmento unos adaptadores que contienen secuencias que posteriormente serán utilizados para pre-amplificar toda la muestras. Como es evidente, la calidad de las muestra original puede condicionar la eficiencia de cada una de estas etapas de la preparación influyendo en el resultado y su interpretación.

El alineamiento de estas secuencias de pequeño tamaño (32-250 pb) frente al genoma de referencia puede conllevar algunas incertidumbres, sobre todo si estas secuencias proceden de genes con parálogos o de secuencias muy repetidas, dificultando la obtención de alineamientos inequívocos. De manera similar, la asignación de éstas secuencias a diferentes mensajeros generados por AS y que comparten una gran propoción de secuencia como el caso del gen *NCR3*, es un proceso complejo que en ocasiones no pueden resolverse con total certeza. Además, las secuencias procedentes de uniones de exones, que no alinean en toda su longitud contra el genoma de referencia, sino que deben ser divididas para su alineamiento, pueden estar también infra-valoradas, en esta ocasión debido al análisis computacional. De manera general, para que una union de exones sea considerada válida debe tener al menos ocho nucleótidos en cada exón de procedencia, por lo que todas aquellas secuencias, que originalmente procedan de uniones de exones, pero que no cumplan este criterio serán descartadas, provocando una representación inferior a la que cabría esperar si se compara con la cobertura observada en los exones adyacentes.

Por otro lado, el uso de PCR anidada, cuyo foco se limita a unos cuantos genes, destaca por su especificidad y su capacidad de magnificación. Mediante PCR anidada podemos detectar mensajeros cuya expresión podría ser de tan sólo unas pocas copias (1-10) en la muestra. Sin embargo, esta metodología también presenta limitaciones. Por un lado, su potencia podría mostrar como relevante algunos mensajeros detectados cuya expresión real es muy minoritaria, que incluso podrían

ser mensajeros inmaduros o pre-mensajeros. El diseño de los cebadores también resulta crucial para conseguir una buena eficiencia y especificidad, sin embargo, su uso conlleva una limitación en las variantes que se pueden detectar mediante esta técnica, puesto que todas aquellas en las que los cebadores diseñados no estén presentes serán invisibles mediante PCR, lo que no ocurre mediante RNA-seq dado su carácter global.

Dadas las limitaciones de ambas técnicas, se decidió diseñar los cebadores para detectar aquellas variantes de splicing que afectasen a la secuencia codificante mediante PCR anidada y completar el análisis del AS utilizando secuencias de RNA-seq disponibles en las bases de datos para obtener un visión global y detallada al mismo tiempo, habiendo quedado demostrado que la combinación de ambas tecnologías puede ayudarnos a reducir las limitaciones particulares de cada una y mejorando nuestra visión del AS del gen *NCR3*. Sirva como ejemplo el análisis realizado en humano, que mediante PCR anidada ha permitido detectar la expresión de nueve mensajeros, mientras que mediante RNA-seq, además de validar las uniones de exones de éstos, se han detectado nuevas uniones de exones en el 5´UTR que en su mayoría no podrían haber sido detectadas por PCR anidada debido a la localización de los cebadores. Además, el carácter cuantitativo del RNA-seq permite inferir que las variantes mayoritariamente expresadas en todos las muestras analizadas son la *A, B* y *C*, lo que ha sido confirmado mediante PCR cuantitativa.

5.3 Splicing Alternativo del gen *NCR3* en diferentes especies

Aunque el splicing alternativo del gen *NCR3* humano se describió antes incluso de conocer las funciones de este gen (Nalabolu et al., 1996; Neville and Campbell, 1999), estas variantes han sido ampliamente ignoradas en los estudios funcionales realizados. Pese a ello, se han atribuido diversas funciones al receptor NCR3 de manera genérica, desde la activación de las células NK frente a células tumorales o infectadas por diferentes patógenos (virus, parásitos y hongos), hasta la regulación de la respuesta inmune adaptativa a través de la inducción de la maduración de las células dendríticas y mediante la eliminación de éstas cuando no presentan la maduración adecuada. Sin embargo, la mayoría de los estudios funcionales realizados se han basado en el uso de anticuerpos frente al extremo N-terminal del receptor, lo que imposibilita la discriminación de isoformas proteicas por tratarse de una región conservada. En este contexto, cabría preguntarse si la diversidad

funcional descrita podría deberse al AS que presenta este gen, y por lo tanto, que cada potencial isoforma proteica descrita fuese sólo responsable parcialmente de las funciones atribuidas al receptor NCR3.

Por otro lado, los estudios del gen *NCR3* en otras especies, tanto a nivel de AS como funcionales, se han visto limitados por tratarse de un pseudogen en ratón (Hollyoake et al., 2005), uno de los principales animales modelo de investigación, sólo existiendo algunos estudios funcionales en *Rattus norvegicus* y en *Pan troglodytes* (Backman-Petersson et al., 2003; Hsieh et al., 2006; Rutjens et al., 2007; Rutjens et al., 2010), en los que el AS también ha sido ignorado. De manera similar a lo observado en humano, el AS podría tener una influencia crucial en el desarrollo funcional del gen *NCR3* en estas especies, de modo que su estudio podría mejorar nuestra compresión de esta influencia, y trasladar los resultados a humano, además de contribuir a profundizar en el conocimiento de los mecanismos subyacentes a la diversidad de los mamíferos.

En este sentido, se realizó un análisis del AS del gen *NCR3* mediante PCR anidada en diversas muestras de tejidos en las especies clasificadas como mamíferos, incluyendo distintas líneas celulares en el caso de humano, y en muestras de sangre de diferentes primates, incluyendo también humanas. De manera general, se puede decir que la expresión del gen *NCR3* produce diversas variantes de splicing en todas las especies analizadas, a excepción de *Mus musculus*, donde se ha podido verificar una ausencia de expresión mediante PCR anidada, lo que confirma definitivamente que se trata de un pseudogen. Considerando únicamente la variantes codificantes detectadas mediante PCR anidada en todas las especies, podemos observar que todas éstas presentan variantes equivalentes expresadas en humano (Figura 5-1A, conservadas), excepto las variantes descritas en *Sus scrofa* y las variantes *Cape-2 y Btau-1* dectadas en *Cebus apella* y *Bos taurus* respectivamente (Figura 5-1A, no conservadas).

La potencial isoforma proteica derivada de la variante no conservada *Cape-2* (Figura 5-1A) no presentaría un dominio transmembrana como consecuencia de la retención del intrón entre los exones 2 y 3 que introduce un cambio en la fase de lectura y prolonga la secuencia codificante sobrepasando el codón de parada canónico, lo que podría indicar que se trata de una isoforma soluble. Sin embargo, la nula homología del extremo C-terminal de esta potencial isoforma y la ausencia de dominios funcionales podría indicar que esta variante es simplemente un mensajero inmaduro. Por otro

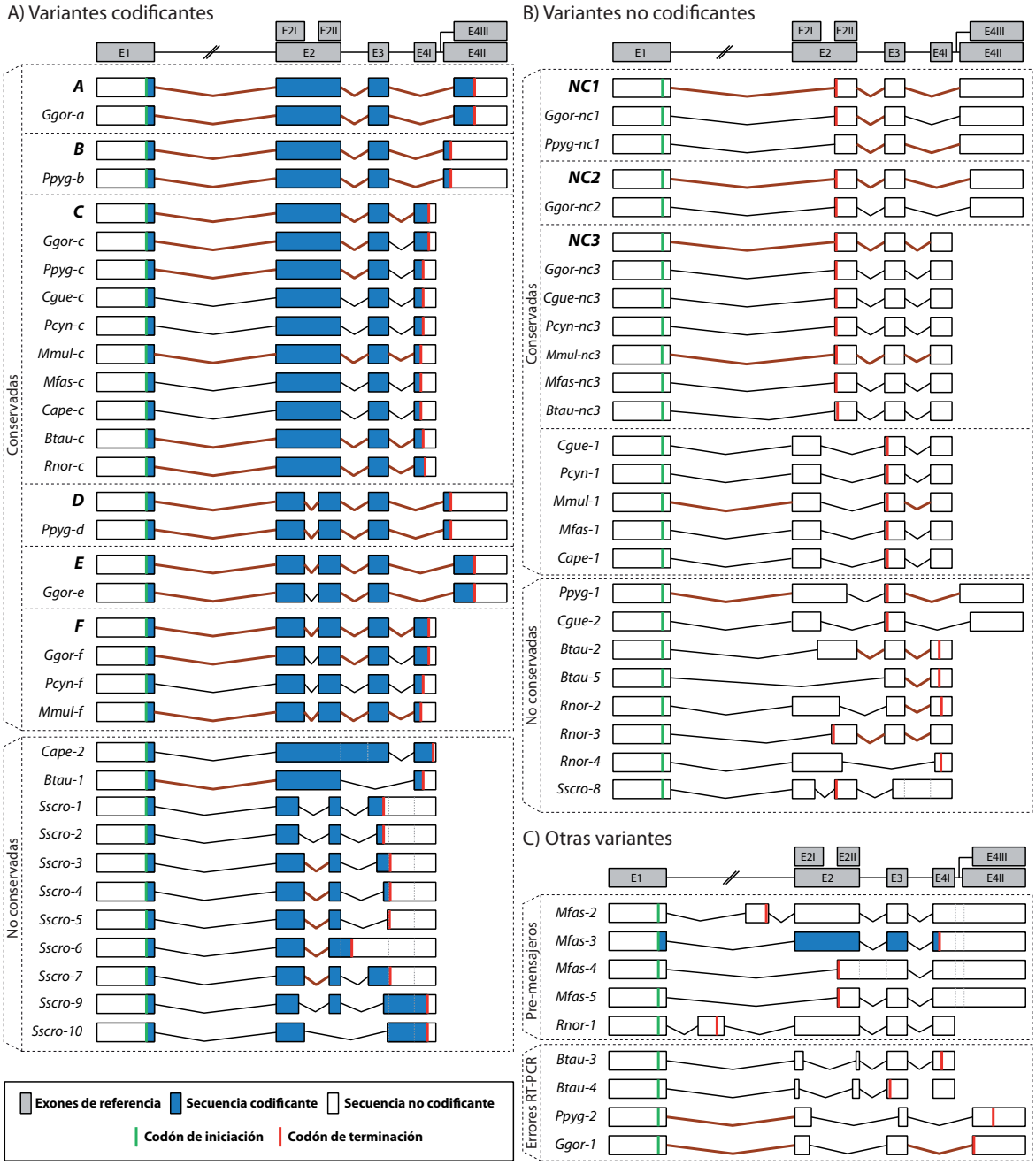


Figura 5-1. Resumen de las variantes de splicing detectadas mediante PCR anidada. Se han agrupado las distintas variantes detectadas con criterios funcionales y por el grado de conservación en las distintas especies analizadas. Las uniones de exones que además fueron detectadas mediante análisis de secuencias de RNA-seq se han resaltado en rojo.

lado, en la variante *Btau-1* (Figura 5-1A), la exclusión del exón 3 completo, que no provoca ningún cambio en la fase de lectura, también conduciría a la ausencia del dominio transmembrana indicando que la isoforma codificada podría ser también de tipo soluble. La expresión de isoformas solubles generadas por AS ha sido descrita en otros receptores como CTLA-4, lo que contribuye a la regulación de la actividad de este receptor a través de la disminución de su presencia en la membrana de los linfocitos T, y estando implicadas en la inhibición de éstas células (Magistrelli et al., 1999; Ward et al., 2013). Por lo tanto, la posible existencia de isoformas solubles podría ser reflejo de mecanismos de regulación similares, especialmente en *Bos taurus* dada la detección ubicua de la variante *Btau-1* en todos los tejidos analizados. Además, en un estudio recientemente publicado se ha observado que la interacción de una proteína recombinante soluble de NCR3, provoca la inhibición de la señalización del receptor NCR3 en ensayos *in vitro* (Binici et al., 2013), disminuyendo la activación de las células NK y reduciendo su secreción de citoquinas. En este sentido, el descubrimiento de isoformas potencialmente solubles, junto al efecto inhibitorio descrito, podría permitir el desarrollo de nuevas estrategias terapeúticas, en especial en el tratamiento de trastornos autoinmunes.

Por otro lado, la expresión del gen *Sscro-ncr3 (Sus scrofa*), que inicialmente podría haberse considerado como un pseudogen debido a las delecciones detectadas en el exón 2, produce nueve variantes de splicing potencialmente codificantes, que no presentan homología con humano ni ninguna otra especie. Este fenómeno se debe, principalmente, a los cambios de fase que introducen las delecciones detectadas en el exón, sin embargo, la expresión de diferentes variantes de splicing es un ejemplo de la plasticidad de la expresión génica mediada por AS. En este sentido, aunque el gen *Sscro-ncr3* esté en el camino de convertirse en un potencial pseudogen, también podría estar evolucionando, gracias al AS, para producir un nuevo gen, con funciones diferentes.

En relación a las variantes codificantes conservadas (Figura 5-1A), se puede observar que la mayor diversidad se encuentra en primates, lo que se debe fundamentalmente a la existencia de los exones 4II y 4III, y por la expresión de variantes que portan el exón 2 divido en dos fragmentos (exones 2I y 2II). En *Rattus norvegicus* y en *Bos taurus* sólo se detectó expresión de las variantes homólogas a la *C* humana, pese a la detección en vaca de los sitios consenso de splicing en el interior del exón 2 que

podrían originar una variante homóloga a la *F* humana (Figura 5-1A, conservadas).

Aunque la presencia de los exones 4II y 4III es evidente en todas las especies de primates, sólo en *Gorilla gorilla, Pongo pygmaeus* y *Homo sapiens* se detectó expresión de variantes codificantes que portasen estos exones, lo que podría reflejar una evolución reciente de estos exones, que se encontrarían aún en un estadio primitivo en primates inferiores como *Colobus* guereza, *Papio cynocephalus,* Macaca *mulatta, Macaca fascicularis* o *Cebus apella.* Para corroborrar esta hipótesis se analizaron las secuencias genómicas de todas las especies en las que existe una secuencia disponible completa del gen *NCR3*, sólo encontrando los exones 4II y 4III en especies de primates simiiformes (Figura 5-2 y Tabla S4), lo que posiciona el origen de estos exones hace 44.2 MA. Por otro lado, la falta de expresión de estos exones en algunas de éstas especies y su relación con la presencia de un sitio consenso de poliadenilación parece tener también un claro origen evolutivo, sólo detectandose la presencia de esta secuencia en primates superiores (superfamilia Hominoidea), que se separaron de los otros primates hace 18.8 MA, lo que se correlaciona directamente con los datos de expresión obtenidos en este trabajo (Figura 5-2). Por lo tanto, se puede considerar que las variantes codificantes portadoras de los exones 4II y 4III detectadas en primates superiores son de reciente evolución, y en consecuencia podrían desarrollar nuevas funciones en el sistema inmune.

En contraposión con estos nuevos exones, el exón 4I es evolutivamente más primitivo, dada su presencia transversal en todas las especies de mamíferos (Tabla S4), y por lo tanto, las variantes portadoras de este exón presentarán la función primordial del gen *NCR3*. En consonancia con esto, se puede observar expresión de este tipo de variantes de manera mayoritaria (Figura 5-1), tanto codificantes como no codificantes. Sin embargo, también se observan algunas diferencias en este exón, en cuanto a la posición del codón de terminación, que podrían reponder a una dinámica evolutiva (Figura 5-1). En la mayoría de las especies analizadas el codón de terminación se sitúa entre las posiciones +47 y +59 con respecto al inicio del exón (Tabla S4), mientras que en primates como *Macaca mulatta, Macaca fascicularis, Papio anubis* y *Cebus apella*, y también en *Echinops telfairi* (erizo), éste se localiza en la posición +35. En el otro extremo, los primates superiores de la superfamilia Hominoidea, junto con *Loxodonta africana* (elefante), presentan el codón de terminación en la posición +77 (Tabla S4). Como consecuencia la longitud de la cola citoplasmática

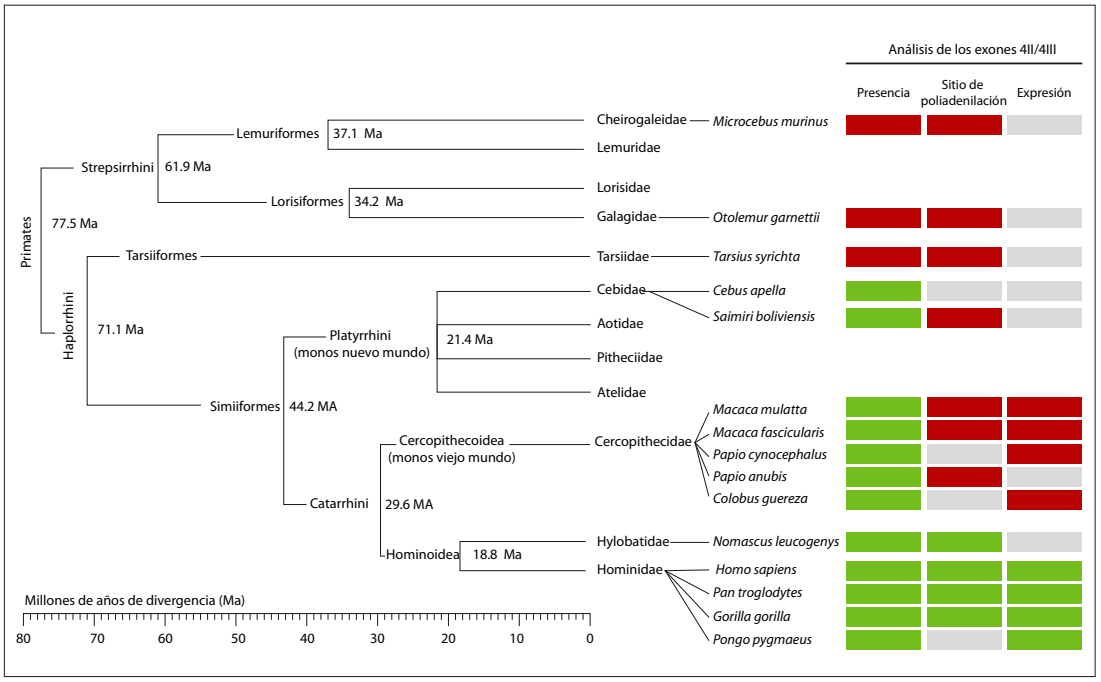


Figura 5-2. Análisis del origen evolutivo de los exones 4II y 4III. Se representa esquemáticamente la filogenia de los primates, incluyendo en cada grupo el resultado del análisis de la presencia de los exones 4II/4III y del sitio consenso de poliadenilación. En las especies incluidas en este trabajo se incluye además la información de la expresión de variantes codificantes que porten estos exones. Los cuadros verdes representa presencia o expresión, los rojos denotan ausencia o falta de expresión, y los grises indican ausencia de información. En la Tabla S4 se amplía esta información al resto de especies en las que se ha descrito el gen *NCR3* en las que se dispone de una secuencia completa.

codificada en este exón ha sufrido varios cambios que podrían responder a las diversas necesidades evolutivas, pudiendo influir en la interacción del receptor NCR3 con CD3ζ. De manera similar se ha descrito en humano una influencia de las distintas colas citoplasmáticas, codificadas en los exones 4I, 4II o 4III, en la modulación de la señalización intracelular a través de la asociación diferencial con CD3ζ (Delahaye et al., 2011), y dado que éstas no presentan dominios funcionales conservados, se podría especular con que las diferencias funcionales observadas podrían responder a las diferentes longitudes.

Además de diversas variantes codificantes, se han detectado diversas variantes de tipo no codificantes (Figura 5-1B). Aunque generalmente son consideradas como errores de la maquinaria de splicing o ruido transcripcional, sorprende observar la conservación de algunas de estas variantes no codificantes como las similaras a las detectadas en humano (*NC1*, *NC2* y *NC3*), además de las variantes similares a *Mmul-1* detectadas en diversos primates, y que de manera similiar a las primeras sólo presentan una de las dos partes del exón 2 (Figura 5-1B). De la misma forma que en

las variantes codificantes, son mayoría aquellas formadas con el exón 4I, es decir, las similires a *NC3* y *Mmul-1* (Figura 5-1B). La conservación de estas variantes no codificantes permite suponer que éstas presentan algún tipo de función conservada, que bien podría ser el control de la expresión de las variantes codificantes. Así una mayor expresión de este tipo de variantes, supondría un descenso de la expresión de las variantes codificantes sin que se altere la expresión global del gen (Hwang and Kim, 2013). También se han detectado diversas variantes no codificantes no conservadas (Figura 5-1B), que también podrían estar implicadas en la regulación de la expresión de las variantes codificantes, aunque su especificidad de especie podría indicar que su origen es el ruido transcripcional o errores de la maquinaria de splicing. En este sentido, se ha propuesto que este tipo de variantes, más que simple basura transcripcional, podrían ser un indicio de procesos evolutivos mediados por AS. La plasticidad en el reconocimiento de los sitios de splicing, permite que la maquinaria de splicing utilice sitios crípticos o utilice combinaciones de sitios diferentes, de modo que a través del AS se pueden producir nuevas variantes, sin alterar sustancialmente las variantes funcionales, que son sometidas a un estricto control a través de NMD, y que sólo eventualmente serán seleccionadas para desarrollar una nueva función (Pickrell et al., 2010; Zhang et al., 2009).

Finalmente, mediante PCR anidada detectamos otras variantes, mayoritariamente no codificantes, que hemos considerado pre-mensajeros o errores de RT-PCR (Figura 5-1C). En el primer grupo se han incluído todas aquellas variantes no codificantes que presentan retención de intrones o que presentan nuevos exones como las variantes *Mfas-2* y *Rnor-1*. Evidentemente, estas variantes, expresadas de manera específica en distintas especies, podrían ser también consideradas como ruido transcripcional entendido como errores de la maquinaria o ensayos de nuevas variantes, en especial aquellas que incorporan exones nuevos. Sin embargo, se ha propuesto que este tipo de exones, dentro de intrones (Figura 5-1C), podría deberse a la eliminación por etapas de intrones grandes, apareciendo pseudo-exones intermediarios que serán eliminados posteriormente para producir los mensajeros maduros (Suzuki et al., 2013). Por otro lado, se han identificado varios mensajeros que podrían deberse a errores durante la retrotranscripción de las muestras de RNA o durante la amplificación mediante PCR denominados *Template-switching* (cambio de molde) (Odelberg et al., 1995) (Figura 5-1C). Estas variantes presentan regiones repetidas en sus respectivas uniones de exones características, además de una ausencia

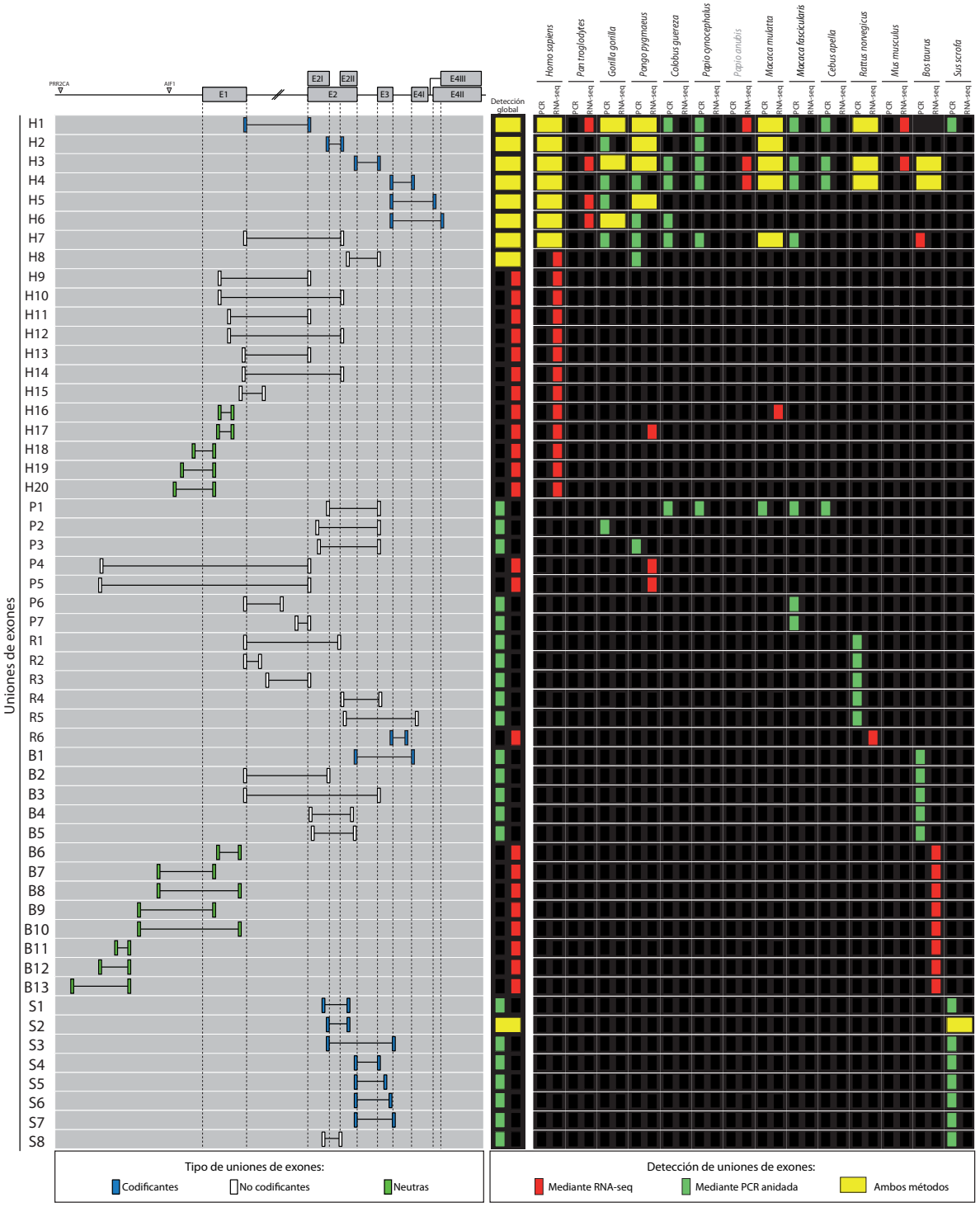


Figura 5-3. Resumen de las uniones de exones detectadas en todas las especies. A la izquierda se representan esquemáticamente las uniones de exones detectadas mediante PCR anidada y RNA-seq. En el panel derecho se muestra las uniones de exones detectadas en los datos de RNA-seq disponibles (Tabla S2) y las uniones detectadas mediante PCR anidada en todas las muestras (Tabla S1).

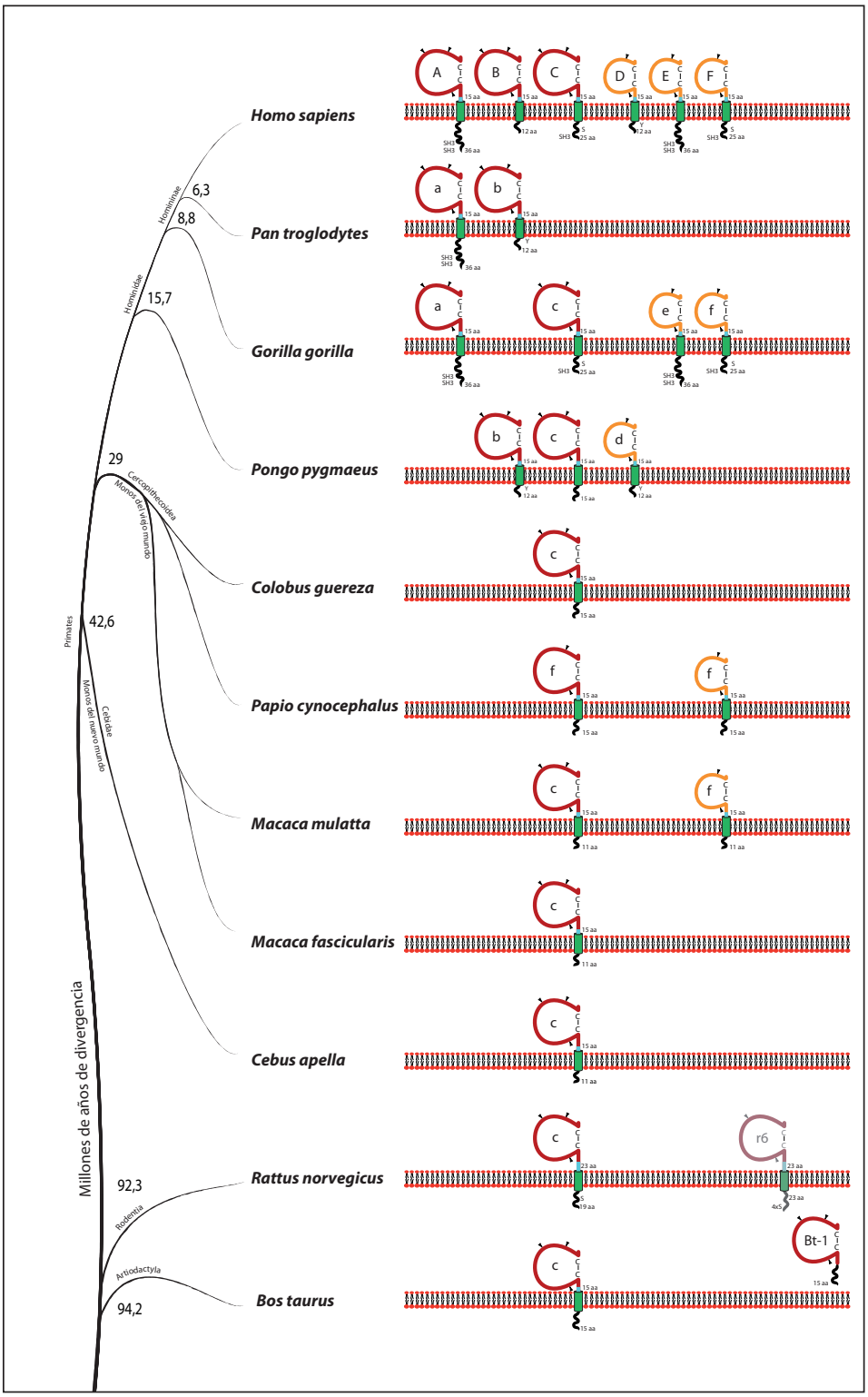


Figura 5-4. Resumen de las posibles isoformas proteicas expresadas en cada una de las especies analizadas. Se ha excluido a *Mus musculus* por no expresar ningún mensajero y a *Sus scrofa* porque las posibles isoformas codificantes no muestras homología con *NCR3*.

de sitios consenso de splicing, lo que es indicativo de este tipo de errores que se producen por la hibridación parcial de una molécula parcialmente sintetizada con una región homóloga en el molde.

Adicionalmente, se han analizado las secuencias procedentes de experimentos de RNA-seq disponibles en las bases de datos para el estudiar el AS del gen *NCR3*. Pese a la desigual disponibinilidad de estas secuencias para las distintas especies, su análisis ha permitido refrendar la mayoría de las uniones de exones detectadas mediante PCR, especialmente las de tipo codificante (Figura 5-1 y 5-3). En lo que se refiere a las uniones de exones codificantes canónicas (H1-H6), cabe destacar que su análisis permite confirmar la hipótesis de un origen reciente de los exones 4II y 4III, no detectado expresión de las uniones de exones H5 y H6 en *Macaca mulatta*, aunque si observamos expresión en *Pongo pygmaeus, Gorilla gorilla*, *Pan troglodytes* y *Homo sapiens* (Figura 5-2 y 5-3). Además de estas uniones de exones, se han detectado mediante esta metodología numerosas uniones de exones nuevas, que en su mayoría no podrían haberse detectado mediante PCR por exceder los límetes definidos por los cebadores diseñados. Sin embargo, la mayoría se puede clasificar como no codificantes o neutras, que además muestran un expresión muy baja, sólo encontrándose en muestras en las que la profundidad de la secuenciación es muy elevada, como en las muestras humanas procedentes de PBMCs, monocitos o células B, o en las muestras de vaca procedentes de la réplica 1 (Tabla S3), lo que revela la limitación de este técnica en la detección de variantes de splicing de baja expresión. Por otro lado, mediante el análisis de estas secuencias no se han detectado la mayoría de las uniones de exones de tipo no codificante detectadas por PCR anidada, lo que podría deberse a esta limitación mencionada, siendo la capacidada de detección de la PCR anidada mucha más profunda. Sin embargo, cabe destacar la unión de exones no codificante H7, característica de las variantes similares a las humanas *NC1*, *NC2* y *NC3*, que además de su detección mediante PCR anidada en diversas especies (Figura 5-1), también se ha detectado mediante RNA-seq en *Homo sapiens* y en *Macaca mulatta*, dos de las especies en las que existen un mayor número de secuencias disponibles (Tabla S1 y S2). Dentro de las uniones de exones neutras, cabe destacar la unión de exones B6, detectada en *Bos taurus*, presentando unos niveles de expresión similares al de las uniones de exones canónicas en esta especie (H1, H3 y H4) (ver Figura4-32, Resultados), lo que permite deducir su presencia en todos los mensajeros, lo que podría explicar la expresión

ubicua del gen *Btau-ncr3* en todos los tejidos analizados mediante RNA-seq a través de la regulación de la estabilidad de los mensajeros portadores. De manera similar, se han detectado uniones de exones en *Homo sapiens* (H16 y H17), *Macaca mulatta* (H16) y *Pongo pygmaeus* (H17), que pese a mostrar niveles de expresión inferiores podrían influir también sobre la estabilidad del mensajero o incorporarse a éstos en respuesta a diferentes estímulos para así controlar la expresión de las isoformas proteicas, de manera análoga a lo descrito para otros genes como SC35 o CD3ζ (Chowdhury et al., 2006; Sureau and Perbal, 1994).

El carácter cualitativo del RNA-seq permite además inferir los niveles de expresión globales del gen *NCR3* en las distintas especies observándose que de manera mayoritaria éste se expresa en tejido linfoides, como bazo y sangre, aunque también se detectan niveles relativamente altos en tejidos muy irrigados como pulmón y corazón (Tabla S3). Una excepción a este patrón de expresión lo encontramos en *Bos taurus*, donde detectamos una expresión elevada en todos los tejidos y en *Pongo pygmaeus*, en el se observan niveles elevados de expresión en cerebro, lo que podría indicar un posible papel de este gen en la inmunidad del tejido cerebral.

5.4 Implicaciones funcionales del splicing alternativo del gen *NCR3*

Tomando los resultados de PCR anidada y RNA-seq en conjunto, pese a la limitación de las muestras y las secuencias de RNA-seq disponibles, se puede observar un claro incremento de la potencial diversidad proteica en primates, que es aún mayor en los superiores por la expresión de variantes portadoras de los exones 4II y 4III (Figura 5-4), y en especial en humano. Mediante qPCR se detectaron en esta especie los mayores niveles de expresión del gen *NCR3* en tejidos linfoides como bazo y timo, y en células NK (ver Figura 4-41, resultados). Las variantes mayoritarias detectadas en todos los tejidos fueron la *A*, *B* y C, que mostraron un patrón diferencial en función del tejido. Así, en bazo, la mayor expresión se correspondió con la variante inmunosupresora *C*, mientras que en células NK y en sangre se observaron mayores niveles de expresión de la variante *A*. Esta diferencia podría deberse a las diferencias funcionales observadas en las distintas poblaciones de células NK, que en los tejidos linfoides secundarios presenta un fenotipo regulador ($CD56^{dim}$), mientras que en circulación muestran un fenotipo más citotóxico ($CD56^{brigth}$). Por lo tanto, se podría considerar que la variante *C* está implicada en procesos de regulación, mientras que la variante *A* está asociada con la activación de la citotoxicidad de las células NK, en

consonancia con los descrito en pacientes con tumores gastrointestinales (Delahaye et al., 2011). Por otro lado, la expresión de las variantes codificantes *D*, *E* y *F*, muy inferior a la de las primeras, podría reflejar una funcionalidad secundaria para estas isoformas, que además, han sido detectadas en otras especies de primates como *Pongo pygmaeus*, *Gorilla gorilla, Papio cynocephalus* y *Macaca mulatta*, encontrando en este último niveles de expresión comparables con los de la variante *C* (ver Tabla S3). Del mismo modo, las variantes no codificantes canónicas muestran también niveles de expresión bajos, como cabría esperar. Finalmente, el uso de cebadores comunes para todas estas variantes, ha permitido confirmar que la contribución de otras potenciales variantes, como las detectadas mediante RNA-seq, es menor.

Pese a los bajos niveles de expresión de detectados de las variantes *D*, *E* y *F*, se decidió caracterizar las potenciales isoformas proteicas, para analizar su viabilidad, junto a las isoformas A, B y C. En primer lugar, se observó sobre-expresión de las seis isoformas codificantes en células HEK293T, así como su capacidad de formar dímeros. Adicionalmente, se pudo comprobar la presencia de N-glicosilaciones, además de determinar que todas se localizan en la membrana plasmática con el extremo N-terminal en el exterior. No se observaron diferencias significativas en la expresión de las distintas isoformas, ni agregados en las inmunofluorescencias que pudieran sugerir un procesamiento aberrante de ninguna de ellas, lo que en conjunto, y pese a tratarse de sobre-expresiones *in vitro*, permite concluir que las isoformas D, E y F podrían expresarse *in vivo*, del mismo modo que las isoformas A, B y C.

Desde un punto de vista estructural, el dominio IgV, característico de las isoformas A, B y C, está formado por dos hojas beta, compuestas por cuatro (ABED) y tres (GFC) láminas beta respectivamente, unidas por un puente disulfuro (Li et al., 2011). A diferencia de otros dominios inmunoglobulina de tipo IgV, la conexión entre las láminas no se realiza a través de dos pequeñas láminas beta (C' y C''), sino que existen dos hélices alfa que realizan esta función (α1 y α2) . La obtención de la estructura tridimensional de este posible dominio extracelular del receptor NCR3 unido al ligando B7H6 permitió determinar que el bucle entre las láminas F y G, y la hélice α2 son las principales regiones implicadas en la interacción con este ligando (Figura 5-5) (Joyce et al., 2011; Kaifu et al., 2011; Li et al., 2011). Sin embargo, estas hélices alfa no están presentes en el potencial dominio IgC, característico de las isoformas D, E y F, debido a la ausencia de 25 aa con respecto al primero como consecuencia

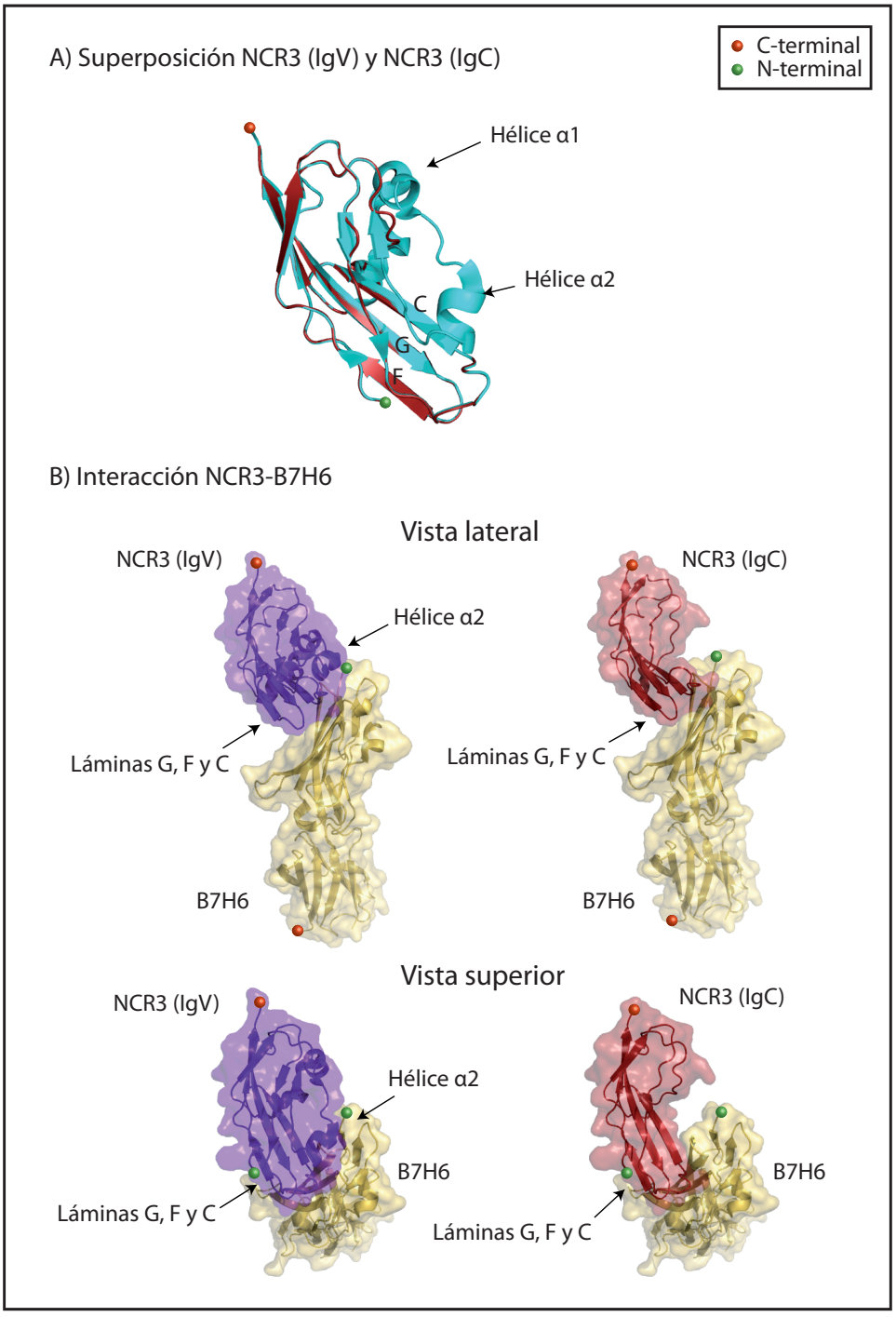


Figura 5-5. Análisis estructural de los dos posibles dominios extracelulares del receptor NCR3. El modelo estructural del dominio IgC se obtuve mediante predicción a través del servicio Swiss-model (http://swissmodel.expasy.org/) utilizando la estructura descrita del dominio IgV como molde (PDB:3pv6). A) Se muestras la superposición del modelo obtenido para el dominio IgC (rojo) con IgV (azul), resaltando en amarillo la región ausente en IgC. B) Comparación de la interacción conB7H6. El panel izquierdo muestra la estructura descrita (3pv6), y en el panel derecho se ha sustituido el dominio IgV por el modelo de IgC.

del splicing interno del exón 2 (Figura 5-5). Por lo tanto, basándose únicamente en esta ausencia se puede deducir que este segundo dominio no interaccionaría con el ligando B7H6.

El análisis de la interacción de cada uno de estos dominios extracelulares, producidos de manera recombinante en el sistema de baculovirus, con el ligando celular B7H6 no dio resultados positivos. Dado que las células de insecto producen una glicosilación diferente a la de mamíferos se concluyó que la correcta glicosilación de éstos dominios es fundamental para la interacción con sus ligandos.

Además, se pudo comprobar que esta deficiente glicosilación inducía la interacción de éstas proteínas recombinantes con los GAGs presentes en la membrana plasmática de las células diana. Esto ya había sido observado previamente con la variante IgV proponiendo a los GAGs como potenciales ligandos de NCR3 (Bloushtain et al., 2004; Hecht et al., 2009; Hershkovitz et al., 2008; Ito et al., 2012; Porgador, 2005; Warren et al., 2005).

Ante los resultados negativos, y la publicación de un estudio en el que se mostraba que la glicosilación del dominio extracelular tipo V es fundamental para la interacción con el ligando B7H6, indujo a realizar un cambio en el sistema de purificación de las proteínas recombinantes. La producción de los posibles dominios extracelulas (IgV e IgC) en células humanas HEK293T permitió obtener una correcta glicosilación de estas proteínas recombinantes, lo que además se pudo verificar por la nula interacción con los GAGs presentes en la membrana celular, en comparación con células que no presentan estas moléculas en la membrana. Sin embargo, tampoco se obtuvieron resultados positivos cuando se analizó la interacción de ambos dominios con células que sobre-expresaban el ligando B7H6 etiquetado en su extremo N-terminal. Por el contrario, la sobre-expresión del ligando B7H6 etiquetado en su extremo C-terminal, permitió observar la interacción del dominio extracelular tipo IgV, mientras que el dominio IgC no mostró ninguna interacción. Este mismo resultado se obtuvo cuando se sobre-expresó B7H6 fusionado a GFP en su extremo C-terminal. Por lo tanto, la incorporación de un epítopo en el extremo N-terminal del ligando impide la interacción, lo que está en consonancia con la estructura tridimensional publicada, estando el extremo N-terminal del ligando B7H6 muy próxima a la superficie de contacto con NCR3 (IgV) (Figura 5-5). Adicionalmente, se utilizó una aproximación inversa, es decir, sobre-expresando en las células las isoformas A y E de NCR3, que presentan

los dominios IgV e IgC respectivamente, fusionadas en su extremo C-terminal con la proteína GFP, y analizando la interacción del dominio extracelular de B7H6 producido de manera recombinante. Esta nueva aproximación confirmó los resultados obtenidos anteriormente, no observando interacción de ligando recombinante B7H6 y la isoforma E del receptor, aunque si se detectó interacción con la isoforma A. En conclusión, se puede afirmar que B7H6 es un ligando de las isoformas A, B y C, todas ellas portadoras de un dominio extracelular tipo IgV, pero no es un ligando para las isoformas D, E y F, cuyo dominio extracelular es de tipo IgC.

Este descubrimiento pone de manifiesto que el AS alternativo del gen *NCR3* tiene una influencia directa sobre la funcionalidad de este gen, indicando que mediante la expresión de diferentes isoformas se puede modular el reconocimiento del ligando B7H6, y por lo tanto, influir sobre la respuesta inmune. De manera análoga, se puede especular con una interacción diferencial de estas isoformas con otros ligandos descritos para el receptor NCR3, como BAT3, hemaglutinina o PfEMP-1, además de sugerir la posibilidad de que existan otros ligandos desconocidos. Aún en el caso de considerar a estas isoformas alternativas (D, E y F) como artefactuales o anecdóticas por su bajos niveles de expresión a nivel de RNA, este hallazgo debe tenerse en consideración por su potencial influencia en el sistema inmune. De manera similar a lo descrito para el receptor CD45, cuya actividad esta controlada mediante AS (Lynch, 2004; Martinez and Lynch, 2013; Martinez et al., 2012) (ver Figura 1-7, Introducción), estas variantes minoritarias del gen *NCR3* podrían expresarse en respuesta a diferentes estímulos, aún por determinar, de manera que sustituyan a las isoformas mayoritarias, como versiones inactivas de éstas, al menos en lo que se refiere al ligando B7H6, controlando la respuesta de las células NK frente a este ligando. Por otro lado, suponiendo que estas isoformas minoritarias fuesen simplemente versiones inactivas, su expresión podría ser un parámetro a considerar en el desarrollo de algunos trastornos inmunes, de manera similar al valor predictivo atribuido a la expresión de las variantes *A*, *B* y *C* en pacientes con tumores gastrointestinales (Delahaye et al., 2011).

En conjunto, los resultado aquí expuestos no son más que un estímulo para continuar estudiando las implicaciones funcionales del splicing alternativo del gen *NCR3* y su influencia en la actividad de las células NK. En este sentido, sería interesante estudiar la interacción de todos los ligandos atruidos al receptor NCR3 con las isoformas

D, E y F, así como la búsqueda de nuevos ligandos. Asimismo, el análisis de los niveles de expresión de las distintas variantes en células NK sometidas a diferentes estímulos podría ayudarnos a comprender el papel que cada una de ellas tiene en la funcionalidad del este gen.

6. Conclusiones

Conclusiones

1. Las anotaciones disponibles en las bases de datos basadas en la homología de secuencia con otras especies o en predicciones bioinformáticas deben ser validadas experimentalmente para obtener una visión real de la influencia del splicing alternativo.

2. El gen *NCR3* presenta una alta homología de secuencia en todas las especies de mamíferos analizadas, tanto en su secuencia como en su estructura exónica.

3. En todas las especies analizadas, mediante PCR anidada y RNA-seq, el gen *NCR3* se expresa produciendo varios mensajeros maduros diferentes, excepto en *Mus musculus* .En *Sus scrofa*, pese a la detección de expresión, se podría considerar un pseudogen.

4. Las variantes portadoras del exón 4I se expresan en todas las especies analizadas indicando que la función primigenia del gen *NCR3* es desarrollado por éstas. Sin embargo, la longitud de la cola citoplasmática codificada en este exón ha cambiado a lo largo de la evolución indicando una posible modulación de su función a través de la interacción con su co-receptor intracelular CD3ζ.

5. La diversidad de las variantes codificantes aumenta siguiendo una tendencia evolutiva, encontrando la mayor diversidad en los primates. Esto se debe fundamentalmente al splicing alternativo del exón 2, sólo detectado en primates y que permite la expresión de dominios extracelulares de tipo IgV o IgC, y a la presencia exclusiva de los exones 4II y 4III que sólo se expresan en primates superiores (*Hominoidea*) por la presencia de un sitio consenso de poliadenilación.

6. El número de variantes no codificantes detectadas no muestra una tendencia evolutiva, encontrando algunas específicas de especie, aunque se ha observado la conservación generalizada de las variantes *NC1*, *NC2, NCR3* y *Mmul-1*, lo que sugiere un papel regulador para estas variantes.

7. Los transcritos correspondientes con las isoformas humanas A, B y C son las más abundantes en todos lo tejidos analizados. Sin embargo, la expresión y análisis de las seis potenciales isoformas proteicas detectadas revela que todas ellas muestran características similares indicando su potencial viabilidad.

8. La correcta glicosilación del receptor NCR3 es fundamental para la interacción con el ligando B7H6, siendo ésta además inhibida por la presencia de epítopos en el extremo N-terminal del ligando.

9. El ligando B7H6 interacciona de manera específica con las isoformas de NCR3 que presentan un dominio extracelular tipo IgV. Sin embargo, las isoformas que muestran un dominio tipo IgC, generado a través de splicing alternativo, no interaccionan con este ligando, lo que podría ser uno de los factores responsables de la diversidad funcional atribuida a este receptor.

7. Bibliografía

Bibliografía

A

Aguado, B., Campbell, R.D., 1995. The novel gene G17, located in the human major histocompatibility complex, encodes PBX2, a homeodomain-containing protein. Genomics 25, 650-659.

Aguado, B., Campbell, R.D., 1998. Characterization of a human lysophosphatidic acid acyltransferase that is encoded by a gene located in the class III region of the human major histocompatibility complex. J Biol Chem 273, 4096-4105.

Aguado, B., Campbell, R.D., 1999. Characterization of a human MHC class III region gene product with S-thioesterase activity. The Biochemical journal 341 (Pt 3), 679-689.

Ahmed, F., Kumar, M., Raghava, G.P., 2009. Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies. In Silico Biol 9, 135-148.

Albertella, M.R., Campbell, R.D., 1994. Characterization of a novel gene in the human major histocompatibility complex that encodes a potential new member of the I kappa B family of proteins. Hum Mol Genet 3, 793-799.

Arnon, T.I., Achdout, H., Levi, O., Markel, G., Saleh, N., Katz, G., Gazit, R., Gonen-Gross, T., Hanna, J., Nahari, E., Porgador, A., Honigman, A., Plachter, B., Mevorach, D., Wolf, D.G., Mandelboim, O., 2005. Inhibition of the NKp30 activating receptor by pp65 of human cytomegalovirus. Nat Immunol 6, 515-523.

Arnon, T.I., Lev, M., Katz, G., Chernobrov, Y., Porgador, A., Mandelboim, O., 2001. Recognition of viral hemagglutinins by NKp44 but not by NKp30. European journal of immunology 31, 2680-2689.

B

Backman-Petersson, E., Miller, J.R., Hollyoake, M., Aguado, B., Butcher, G.W., 2003. Molecular characterization of the novel rat NK receptor 1C7. European journal of immunology 33, 342-351.

Bali, D., Gourley, S., Kostyu, D.D., Goel, N., Bruce, I., Bell, A., Walker, D.J., Tran, K., Zhu, D.K., Costello, T.J., Amos, C.I., Seldin, M.F., 1999. Genetic analysis of multiplex rheumatoid arthritis families. Genes and immunity 1, 28-36.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C.M., Wilson, M.D., Kim, P.M., Odom, D.T., Frey, B.J., Blencowe, B.J., 2012. The evolutionary landscape of alternative splicing in vertebrate species. Science 338, 1587-1593.

Bennett, K.L., Jackson, D.G., Simon, J.C., Tanczos, E., Peach, R., Modrell, B., Stamenkovic, I., Plowman, G., Aruffo, A., 1995a. CD44 isoforms containing exon V3 are responsible for the presentation of heparin-binding growth factor. J Cell Biol 128, 687-698.

Bennett, K.L., Modrell, B., Greenfield, B., Bartolazzi, A., Stamenkovic, I., Peach, R., Jackson, D.G., Spring, F., Aruffo, A., 1995b. Regulation of CD44 binding to hyaluronan by glycosylation of variably spliced exons. J Cell Biol 131, 1623-1633.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2013. GenBank. Nucleic acids research 41, D36-42.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L., 2000. GenBank. Nucleic Acids Res 28, 15-18.

Berget, S.M., 1995. Exon recognition in vertebrate splicing. The Journal of biological chemistry 270, 2411-2414.

Berget, S.M., Moore, C., Sharp, P.A., 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proceedings of the National Academy of Sciences of the United States of America 74, 3171-3175.

Beyer, M., Mallmann, M.R., Xue, J., Staratschek-Jox, A., Vorholt, D., Krebs, W., Sommer, D., Sander, J., Mertens, C., Nino-Castro, A., Schmidt, S.V., Schultze, J.L., 2012. High-resolution transcriptome of human macrophages. PloS one 7, e45466.

Biassoni, R., Cantoni, C., Marras, D., Giron-Michel, J., Falco, M., Moretta, L., Dimasi, N., 2003. Human Natural Killer cell receptors: insights into their molecular function and structure. Journal of Cellular and Molecular Medicine 7, 376-387.

Binici, J., Hartmann, J., Herrmann, J., Schreiber, C., Beyer, S., G√ºler, G.n., Vogel, V., Tumulka, F., Abele, R., M√§ntele, W., Koch, J., 2013. A soluble fragment of the tumor antigen BCL2-associated athanogene 6 (BAG-6) is essential and sufficient for inhibition of NKp30-dependent cytotoxicity of natural killer cells. Journal of Biological Chemistry.

Blanchard, J.M., Weber, J., Jelinek, W., Darnell, J.E., 1978. In vitro RNA-RNA splicing in adenovirus 2 mRNA formation. Proceedings of the National Academy of Sciences of the United States of America 75, 5344-5348.

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 research 38, 7651-7664.

Blekhman, R., Marioni, J.C., Zumbo, P., Stephens, M., Gilad, Y., 2010. Sex-specific and lineage-specific alternative splicing in primates. Genome Research 20, 180-189.

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

Blom, N., Gammeltoft, S., Brunak, S., 1999. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294, 1351-1362.

Bloushtain, N., Qimron, U., Bar-Ilan, A., Hershkovitz, O., Gazit, R., Fima, E., Korc, M., Vlodavsky, I., Bovin, N.V., Porgador, A., 2004. Membrane-associated heparan sulfate proteoglycans are involved in the recognition of cellular targets by NKp30 and NKp46. Journal of immunology 173, 2392-2401.

Bogdan, C., 2012. Natural killer cells in experimental and human leishmaniasis. Front Cell Infect Microbiol 2, 69.

Boteva, L., Morris, D.L., Cortes-Hernandez, J., Martin, J., Vyse, T.J., Fernando, M.M., 2012. Genetically determined partial complement C4 deficiency states are not independent risk factors for SLE in UK and Spanish populations. Am J Hum Genet 90, 445-456.

Boue, S., Letunic, I., Bork, P., 2003. Alternative splicing and evolution. Bioessays 25, 1031-1034.

Brandt, C.S., Baratin, M., Yi, E.C., Kennedy, J., Gao, Z., Fox, B., Haldeman, B., Ostrander, C.D., Kaifu, T., Chabannon, C., Moretta, A., West, R., Xu, W., Vivier, E., Levin, S.D., 2009. The B7 family member B7-H6 is a tumor cell ligand for the activating natural killer cell receptor NKp30 in humans. J Exp Med.

Brunak, S., Engelbrecht, J., Knudsen, S., 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. J Mol Biol 220, 49-65.

Bryceson, Y.T., March, M.E., Ljunggren, H.G., Long, E.O., 2006. Activation, coactivation, and costimulation of resting human natural killer cells. Immunological reviews 214, 73-91.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J Mol Biol 268, 78-94.

C

Caceres, J.F., Kornblihtt, A.R., 2002. Alternative splicing: multiple control mechanisms and involvement in human disease. Trends in genetics : TIG 18, 186-193.

Calame, K., Rogers, J., Early, P., Davis, M., Livant, D., Wall, R., Hood, L., 1980. Mouse Cmu heavy chain immunoglobulin gene segment contains three intervening sequences separating domains. Nature 284, 452-455.

Caligiuri, M.A., 2008. Human natural killer cells. Blood 112, 461-469.

Calvanese, V., Mallya, M., Campbell, R.D., Aguado, B., 2008. Regulation of expression of two LY-6 family genes by intron retention and transcription induced chimerism. BMC Mol Biol 9, 81.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC bioinformatics 10, 421.

Chang, S.T., Sova, P., Peng, X., 2011. Next-Generation Sequencing Reveals HIV-1-Mediated Suppression of RNA Expression in a CD4 T Cell Line.

Chang, Y.F., Imam, J.S., Wilkinson, M.F., 2007. The nonsense-mediated decay RNA surveillance pathway. Annual review of biochemistry 76, 51-74.

Chen, L., Tovar-Corona, J.M., Urrutia, A.O., 2012. Alternative splicing: a potential source of functional innovation in the eukaryotic genome. Int J Evol Biol 2012, 596274.

Chiba, T., Matsuzaka, Y., Warita, T., Sugoh, T., Miyashita, K., Tajima, A., Nakamura, M., Inoko, H., Sato, T., Kimura, M., 2011. NFKBIL1 confers resistance to experimental autoimmune arthritis through the regulation of dendritic cell functions. Scandinavian journal of immunology 73, 478-485.

Chitadze, G., Bhat, J., Lettau, M., Janssen, O., Kabelitz, D., 2013. Generation of Soluble NKG2D Ligands: Proteolytic Cleavage, Exosome Secretion and Functional Implications. Scandinavian journal of immunology 78, 120-129.

Chow, L.T., Roberts, J.M., Lewis, J.B., Broker, T.R., 1977. A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. Cell 11, 819-836.

Chowdhury, B., Krishnan, S., Tsokos, C.G., Robertson, J.W., Fisher, C.U., Nambiar, M.P., Tsokos, G.C., 2006. Stability and Translation of TCR Œð mRNA Are Regulated by the Adenosine-Uridine-Rich Elements in Splice-Deleted 3,Ä≤ Untranslated Region of Œð-Chain. The Journal of Immunology 177, 8248-8257.

Corpet, F., 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16, 10881-10890.

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res 14, 1188-1190.

Cruz-Munoz, M.E., Veillette, A., 2010. Do NK cells always need a license to kill? Nature immunology 11, 279-280.

Cucca, F., Zhu, Z.B., Khanna, A., Cossu, F., Congia, M., Badiali, M., Lampis, R., Frau, F., De Virgiliis, S., Cao, A., Arnone, M., Piras, P., Campbell, R.D., Cooper, M.D., Volanakis, J.E., Powis, S.H., 1998. Evaluation of IgA deficiency in Sardinians indicates a susceptibility gene is encoded within the HLA class III region. Clin Exp Immunol 111, 76-80.

D

das Neves, R.P., Jones, N.S., Andreu, L., Gupta, R., Enver, T., Iborra, F.J., 2010. Connecting variability in global transcription rate to mitochondrial variability. PLoS Biol 8, e1000560.

De Conti, L., Baralle, M., Buratti, E., 2013. Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip Rev RNA 4, 49-60.

De Maria, A., Moretta, L., 2008. NK cell function in HIV-1 infection. Curr HIV Res 6, 433-440.

de Rham, C., Ferrari-Lacraz, S., Jendly, S., Schneiter, G., Dayer, J.M., Villard, J., 2007. The proinflammatory cytokines IL-2, IL-15 and IL-21 modulate the repertoire of mature human natural killer cell receptors. Arthritis Res Ther 9, R125.

Deakin, J.E., Papenfuss, A.T., Belov, K., Cross, J.G., Coggill, P., Palmer, S., Sims, S., Speed, T.P., Beck, S., Graves, J.A., 2006. Evolution and comparative analysis of the MHC Class III inflammatory region. BMC Genomics 7, 281.

Deguine, J., Bousso, P., 2013. Dynamics of NK cell interactions in vivo. Immunological reviews 251, 154-159.

Delahaye, N.F., Barbier, M., Fumoux, F., Rihet, P., 2007. Association analyses of NCR3 polymorphisms with P. falciparum mild malaria. Microbes Infect 9, 160-166.

Delahaye, N.F., Rusakiewicz, S., Martins, I., Ménard, C., Roux, S., Lyonnet, L., Paul, P., Sarabi, M., Chaput, N., Semeraro, M., Minard-Colin, V., Poirier-Colame, V., Chaba, K., Flament, C., Baud, V., Authier, H., Kerdine-Römer, S., Pallardy, M., Cremer, I., Peaudecerf, L., Rocha, B., Valteau-Couanet, D., Gutierrez, J.C., Nunès, J.a., Commo, F., Bonvalot, S., Ibrahim, N., Terrier, P., Opolon, P., Bottino, C., Moretta, A., Tavernier, J., Rihet, P., Coindre, J.-M., Blay, J.-Y., Isambert, N., Emile, J.-F., Vivier, E., Lecesne, A., Kroemer, G., Zitvogel, L., 2011. Alternatively spliced NKp30 isoforms affect the prognosis of gastrointestinal stromal tumors. Nature medicine 17, 700-707.

Della Chiesa, M., Vitale, M., Carlomagno, S., Ferlazzo, G., Moretta, L., Moretta, A., 2003. The natural killer cell-mediated killing of autologous dendritic cells is confined to a cell subset expressing CD94/NKG2A, but lacking inhibitory killer Ig-like receptors. European journal of immunology 33, 1657-1666.

Dominski, Z., Kole, R., 1991. Selection of splice sites in pre-mRNAs with short internal exons. Mol Cell Biol 11, 6075-6083.

Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Gomez Acuna, L.I., Fiszbein, A., Godoy Herz, M.A., Nieto Moreno, N., Munoz, M.J., Allo, M., Schor, I.E., Kornblihtt, A.R., 2013. Transcriptional elongation and alternative splicing. Biochimica et biophysica acta 1829, 134-140.

E

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, 313-319.

Eppig, J.T., Blake, J.A., Davisson, M.T., Kadin, J.A., Richardson, J.E., 2000. The mouse genome database: a resource for today and tomorrow. Lab Anim (NY) 29, 39-43.

Esko, J.D., Stewart, T.E., Taylor, W.H., 1985. Animal cell mutants defective in glycosaminoglycan biosynthesis. Proceedings of the National Academy of Sciences of the United States of America 82, 3197-3201.

F

Faustino, N.A., Cooper, T.A., 2003. Pre-mRNA splicing and human disease. Genes Dev 17, 419-437.

Ferlazzo, G., Pack, M., Thomas, D., Paludan, C., Schmid, D., Strowig, T., Bougras, G., Muller, W.A., Moretta, L., Munz, C., 2004. Distinct roles of IL-12 and IL-15 in human natural killer cell activation by dendritic cells from secondary lymphoid organs. Proceedings of the National Academy of Sciences of the United States of America 101, 16606-16611.

Ferlazzo, G., Tsang, M.L., Moretta, L., Melioli, G., Steinman, R.M., Munz, C., 2002. Human dendritic cells activate resting natural killer (NK) cells and are recognized via the NKp30 receptor by activated NK cells. The Journal of experimental medicine 195, 343-351.

Ferraro, E., Peluso, D., Via, A., Ausiello, G., Helmer-Citterich, M., 2007. SH3-Hunter: discovery of SH3 domain interaction sites in proteins. Nucleic Acids Res 35, W451-454.

Flajnik, M.F., Tlapakova, T., Criscitiello, M.F., Krylov, V., Ohta, Y., 2012. Evolution of the B7 family: co-evolution of B7H6 and NKp30, identification of a new B7 family member, B7H7, and of B7's historical relationship with the MHC. Immunogenetics.

Fogel, B.L., Wexler, E., Wahnich, A., Friedrich, T., Vijayendran, C., Gao, F., Parikshak, N., Konopka, G., Geschwind, D.H., 2012. RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. Human molecular genetics 21, 4171-4186.

Frasca, F., Pandini, G., Scalia, P., Sciacca, L., Mineo, R., Costantino, A., Goldfine, I.D., Belfiore, A., Vigneri, R., 1999. Insulin receptor isoform A, a newly recognized, high-affinity insulin-like growth factor II receptor in fetal and cancer cells. Mol Cell Biol 19, 3278-3288.

Fu, R.H., Liu, S.P., Ou, C.W., Yu, H.H., Li, K.W., Tsai, C.H., Shyu, W.C., Lin, S.Z., 2009. Alternative splicing modulates stem cell differentiation. Cell Transplant 18, 1029-1038.

G

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., Lopez, R., 2010. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res 38, W695-699.

Gray, K.A., Daugherty, L.C., Gordon, S.M., Seal, R.L., Wright, M.W., Bruford, E.A., 2013. Genenames.org: the HGNC resources in 2013. Nucleic Acids Res 41, D545-552.

Guergnon, J., Dalmasso, C., Broet, P., Meyer, L., Westrop, S.J., Imami, N., Vicenzi, E., Morsica, G., Tinelli, M., Zanone Poma, B., Goujard, C., Potard, V., Gotch, F.M., Casoli, C., Cossarizza, A., Macciardi, F., Debre, P., Delfraissy, J.F., Galli, M., Autran, B., Costagliola, D., Poli, G., Theodorou, I., Riva, A., 2012. Single-nucleotide polymorphism-defined class I and class III major histocompatibility complex genetic subregions contribute to natural long-term nonprogression in HIV infection. J Infect Dis 205, 718-724.

Gupta, R., 2002. Prediction of glycosylation across the human proteome and the correlation to protein function 1 Introduction 2 Methods. 322, 310-322.

H

Hackett, N.R., Butler, M.W., Shaykhiev, R., Salit, J., Omberg, L., Rodriguez-Flores, J.L., Mezey, J.G., Strulovici-Barel, Y., Wang, G., Didon, L., Crystal, R.G., 2012. RNA-Seq quantification of the human small airway epithelium transcriptome. BMC Genomics 13, 82.

Hartmann, J., Tran, T.-V., Kaudeer, J., Oberle, K., Herrmann, J., Quagliano, I., Abel, T., Cohnen, A., Gatterdam, V., Jacobs, A., Wollscheid, B., Tampé, R., Watzl, C., Diefenbach, A., Koch, J., 2012. The stalk domain and the glycosylation status of the activating natural killer cell receptor NKp30 are important for ligand binding. The Journal of biological chemistry 287, 31527-31539.

Hecht, M.L., Rosental, B., Horlacher, T., Hershkovitz, O., De Paz, J.L., Noti, C., Schauer, S., Porgador, A., Seeberger, P.H., 2009. Natural cytotoxicity receptors NKp30, NKp44 and NKp46 bind to different heparan sulfate/heparin sequences. Journal of proteome research 8, 712-720.

Hedges, S.B., Dudley, J., Kumar, S., 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22, 2971-2972.

Heider, K.H., Kuthan, H., Stehle, G., Munzert, G., 2004. CD44v6: a target for antibody-based cancer therapy. Cancer Immunol Immunother 53, 567-579.

Heinrich, B., Zhang, Z., Raitskin, O., Hiller, M., Benderska, N., Hartmann, A.M., Bracco, L., Elliott, D., Ben-Ari, S., Soreq, H., Sperling, J., Sperling, R., Stamm, S., 2009. Heterogeneous nuclear ribonucleoprotein G regulates splice site selection by binding to CC(A/C)-rich regions in pre-mRNA. The Journal of biological chemistry 284, 14303-14315.

Hernández-Torres, F., Rastrojo, A., Aguado, B., 2013. Intron retention and transcript chimerism conserved across mammals: Ly6g5b and Csnk2b-Ly6g5b as examples. BMC Genomics 14, 199.

Hershkovitz, O., Jarahian, M., Zilka, A., Bar-Ilan, A., Landau, G., Jivov, S., Tekoah, Y., Glicklis, R., Gallagher, J.T., Hoffmann, S.C., Zer, H., Mandelboim, O., Watzl, C., Momburg, F., Porgador, A., 2008. Altered glycosylation of recombinant NKp30 hampers binding to heparan sulfate: a lesson for the use of recombinant immunoreceptors as an immunological tool. Glycobiology 18, 28-41.

Hirokawa, T., Boon-Chieng, S., Mitaku, S., 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 14, 378-379.

Hollyoake, M., Campbell, R.D., Aguado, B., 2005. NKp30 (NCR3) is a pseudogene in 12 inbred and wild mouse strains, but an expressed gene in Mus caroli. Molecular biology and evolution 22, 1661-1672.

Holt, C.E., Bullock, S.L., 2009. Subcellular mRNA localization in animal cells and why it matters. Science 326, 1212-1216.

Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., Lush, M.J., Povey, S., Talbot, C.C., Jr., Wright, M.W., Wain, H.M., Trowsdale, J., Ziegler, A., Beck, S., 2004. Gene map of the extended human MHC. Nature reviews. Genetics 5, 889-899.

Hsieh, C.L., Nagasaki, K., Martinez, O.M., Krams, S.M., 2006. NKp30 is a functional activation receptor on a subset of rat natural killer cells. European journal of immunology 36, 2170-2180.

Hsieh, S.L., Campbell, R.D., 1991. Evidence that gene G7a in the human major histocompatibility complex encodes valyl-tRNA synthetase. The Biochemical journal 278 (Pt 3), 809-816.

Hu, H.J., Jin, E.H., Yim, S.H., Yang, S.Y., Jung, S.H., Shin, S.H., Kim, W.U., Shim, S.C., Kim, T.G., Chung, Y.J., 2011. Common variants at the promoter region of the APOM confer a risk of rheumatoid arthritis. Exp Mol Med 43, 613-621.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., Clamp, M., 2002. The Ensembl genome database project. Nucleic Acids Res 30, 38-41.

Hurt, P., Walter, L., Sudbrak, R., Klages, S., Müller, I., Shiina, T., Inoko, H., Lehrach, H., Günther, E., Reinhardt, R., Himmelbauer, H., 2004. The Genomic Sequence and Comparative Analysis of the Rat Major Histocompatibility Complex. 1, 631-639.

Hwang, J., Kim, Y.K., 2013. When a ribosome encounters a premature termination codon. BMB Rep 46, 9-16.

I

Iskow, R.C., Gokcumen, O., Abyzov, A., Malukiewicz, J., Zhu, Q., Sukumar, A.T., Pai, A.a., Mills, R.E., Habegger, L., Cusanovich, D.a., Rubel, M.a., Perry, G.H., Gerstein, M., Stone, A.C., Gilad, Y., Lee, C., 2012. Regulatory element copy number differences shape primate expression profiles. Proceedings of the National Academy of Sciences of the United States of America 109, 12656-12661.

Ito, K., Higai, K., Shinoda, C., Sakurai, M., Yanai, K., Azuma, Y., Matsumoto, K., 2012. Unlike natural killer (NK) p30, natural cytotoxicity receptor NKp44 binds to multimeric alpha2,3-NeuNAc-containing N-glycans. Biological & pharmaceutical bulletin 35, 594-600.

Izquierdo, J.M., Valcarcel, J., 2006. A simple principle to explain the evolution of pre-mRNA splicing. Genes Dev 20, 1679-1684.

J

Jarahian, M., Fiedler, M., Cohnen, A., Djandji, D., Hammerling, G.J., Gati, C., Cerwenka, A., Turner, P.C., Moyer, R.W., Watzl, C., Hengel, H., Momburg, F., 2011. Modulation of NKp30- and NKp46-mediated natural killer cell responses by poxviral hemagglutinin. PLoS pathogens 7, e1002195.

Jarahian, M., Watzl, C., Fournier, P., Arnold, A., Djandji, D., Zahedi, S., Cerwenka, A., Paschen, A., Schirrmacher, V., Momburg, F., 2009. Activation of natural killer cells by newcastle disease virus hemagglutinin-neuraminidase. Journal of virology 83, 8108-8121.

Jenkins, S.C., March, R.E., Campbell, R.D., Milner, C.M., 2000. A novel variant of the MHC-linked hsp70, hsp70-hom, is associated with rheumatoid arthritis. Tissue Antigens 56, 38-44.

Jost, S., Altfeld, M., 2013. Control of human viral infections by natural killer cells. Annual review of immunology 31, 163-194.

Joyce, M.G., Tran, P., Zhuravleva, M.A., Jaw, J., Colonna, M., Sun, P.D., 2011. Crystal structure of human natural cytotoxicity receptor NKp30 and identification of its ligand binding site. Proc Natl Acad Sci U S A.

K

Kaifu, T., Escaliere, B., Gastinel, L.N., Vivier, E., Baratin, M., 2011. B7-H6/NKp30 interaction: a mechanism of alerting NK cells against tumors. Cellular and molecular life sciences : CMLS 68, 3531-3539.

Kalsotra, A., Cooper, T.A., 2011. Functional consequences of developmentally regulated alternative splicing. Nature reviews. Genetics 12, 715-729.

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. Proceedings of the National Academy of Sciences of the United States of America 105, 20333-20338.

Kanadia, R.N., Johnstone, K.A., Mankodi, A., Lungu, C., Thornton, C.A., Esson, D., Timmers, A.M., Hauswirth, W.W., Swanson, M.S., 2003. A muscleblind knockout model for myotonic dystrophy. Science 302, 1978-1980.

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, 27-30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M., 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40, D109-114.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., Stamm, S., 2013. Function of alternative splicing. Gene 514, 1-30.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D., 2002. The human genome browser at UCSC. Genome Res 12, 996-1006.

Kilding, R., Wilson, A.G., 2005. Mapping of a novel susceptibility gene for rheumatoid arthritis in the telomeric MHC region. Cytokine 32, 71-75.

Kim, S., Poursine-Laurent, J., Truscott, S.M., Lybarger, L., Song, Y.-J., Yang, L., French, A.R., Sunwoo, J.B., Lemieux, S., Hansen, T.H., Yokoyama, W.M., 2005. Licensing of natural killer cells by host major histocompatibility complex class I molecules. Nature 436, 709-713.

Kole, R., Weissman, S.M., 1982. Accurate in vitro splicing of human beta-globin RNA. Nucleic acids research 10, 5429-5445.

Kreivi, J.P., Lamond, a.I., 1996. RNA splicing: unexpected spliceosome diversity. Current biology : CB 6, 802-805.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305, 567-580.

Kumar, N., Kaur, G., Tandon, N., Mehra, N., 2012. Tumor necrosis factor-associated susceptibility to type 1 diabetes is caused by linkage disequilibrium with HLA-DR3 haplotypes. Human immunology 73, 566-573.

L

LaBonte, M.L., Levy, D.B., Letvin, N.L., 2000. Characterization of rhesus monkey CD94/NKG2 family members and identification of novel transmembrane-deleted forms of NKG2-A, B, C, and D. Immunogenetics 51, 496-499.

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.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology 10, R25.

Lanier, L.L., 2008. Up on the tightrope: natural killer cell activation and inhibition. Nat Immunol 9, 495-502.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947-2948.

Laulederkind, S.J., Liu, W., Smith, J.R., Hayman, G.T., Wang, S.J., Nigam, R., Petri, V., Lowry, T.F., de Pons, J., Dwinell, M.R., Shimoyama, M., 2013. PhenoMiner: quantitative phenotype curation at the rat genome database. Database (Oxford) 2013, bat015.

Leff, S.E., Rosenfeld, M.G., Evans, R.M., 1986. Complex transcriptional units: diversity in gene expression by alternative RNA processing. Annual review of biochemistry 55, 1091-1117.

Letunic, I., Doerks, T., Bork, P., 2012. SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res 40, D302-305.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 25, 2078-2079.

Li, Y., Wang, Q., Mariuzza, R.A., 2011. Structure of the human activating natural cytotoxicity receptor NKp30 bound to its tumor cell ligand B7-H6. J Exp Med 208, 703-714.

Liolios, K., Mavromatis, K., Tavernarakis, N., Kyrpides, N.C., 2008. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic acids research 36, D475-D479.

Livak, K.J., Schmittgen, T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25, 402-408.

Long, E.O., Kim, H.S., Liu, D., Peterson, M.E., Rajagopalan, S., 2013. Controlling natural killer cell responses: integration of signals for activation and inhibition. Annual review of immunology 31, 227-258.

López-Díez, R., Rastrojo, A., Villate, O., Aguado, B., 2013. Complex tissue-specific patterns and distribution of multiple RAGE splice variants in different mammals. En revisión.

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., Misteli, T., 2011. Epigenetics in alternative pre-mRNA splicing. Cell 144, 16-26.

Luo, M.L., Zhou, Z., Magni, K., Christoforides, C., Rappsilber, J., Mann, M., Reed, R., 2001. Pre-mRNA splicing and mRNA export linked by direct interactions between UAP56 and Aly. Nature 413, 644-647.

Lynch, K.W., 2004. Consequences of regulated pre-mRNA splicing in the immune system. Nature reviews. Immunology 4, 931-940.

M

Magistrelli, G., Jeannin, P., Herbault, N., Benoit De Coignac, A., Gauchat, J.F., Bonnefoy, J.Y., Delneste, Y., 1999. A soluble form of CTLA-4 generated by alternative splicing is expressed by nonstimulated human T cells. European journal of immunology 29, 3596-3602.

Mallya, M., Campbell, R.D., Aguado, B., 2002. Transcriptional analysis of a novel cluster of LY-6 family members in the human and mouse major histocompatibility complex: five genes with many splice forms. Genomics 80, 113-123.

Mallya, M., Campbell, R.D., Aguado, B., 2006. Characterization of the five novel Ly-6 superfamily members encoded in the MHC, and detection of cells expressing their potential ligands. Protein Sci 15, 2244-2256.

Martin, K.C., Ephrussi, A., 2009. mRNA localization: gene expression in the spatial dimension. Cell 136, 719-730.

Martinez, N.M., Lynch, K.W., 2013. Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn. Immunological reviews 253, 216-236.

Martinez, N.M., Pan, Q., Cole, B.S., Yarosh, C.a., Babcock, G.a., Heyd, F., Zhu, W., Ajith, S., Blencowe, B.J., Lynch, K.W., 2012. Alternative splicing networks regulated by signaling in human T cells. RNA (New York, N.Y.) 18, 1029-1040.

Mavoungou, E., Held, J., Mewono, L., Kremsner, P.G., 2007. A Duffy binding-like domain is involved in the NKp30-mediated recognition of Plasmodium falciparum-parasitized erythrocytes by natural killer cells. J Infect Dis 195, 1521-1531.

Merkin, J., Russell, C., Chen, P., Burge, C.B., 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science 338, 1593-1599.

Miletic, A., Krmpotic, A., Jonjic, S., 2013. The evolutionary arms race between NK cells and viruses: who gets the short end of the stick? European journal of immunology 43, 867-877.

Miller, J.W., Urbinati, C.R., Teng-Umnuay, P., Stenberg, M.G., Byrne, B.J., Thornton, C.A., Swanson, M.S., 2000. Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. Embo J 19, 4439-4448.

Milner, C.M., Smith, S.V., Carrillo, M.B., Taylor, G.L., Hollinshead, M., Campbell, R.D., 1997. Identification of a sialidase encoded in the human major histocompatibility complex. J Biol Chem 272, 4549-4558.

Moretta, A., Biassoni, R., Bottino, C., Mingari, M.C., Moretta, L., 2000. Natural cytotoxicity receptors that trigger human NK-cell-mediated cytolysis. Immunol Today 21, 228-234.

Moretta, A., Bottino, C., Mingari, M.C., Biassoni, R., Moretta, L., 2002a. What is a natural killer cell? Nature immunology 3, 6-8.

Moretta, L., Biassoni, R., Bottino, C., Mingari, M.C., Moretta, A., 2002b. Natural killer cells: a mystery no more. Scand J Immunol 55, 229-232.

Moretta, L., Bottino, C., Pende, D., Mingari, M.C., Biassoni, R., Moretta, A., 2002c. Human natural killer cells: their origin, receptors and function. European journal of immunology 32, 1205-1211.

Mudge, J.M., Frankish, A., Fernandez-Banet, J., Alioto, T., Derrien, T., Howald, C.d., Reymond, A., Guig√≥, R., Hubbard, T., Harrow, J., 2011. The Origins, Evolution, and Functional Potential of Alternative Splicing in Vertebrates. Molecular biology and evolution 28, 2949-2959.

N

Nalabolu, S.R., Shukla, H., Nallur, G., Parimoo, S., Weissman, S.M., 1996. Genes in a 220-kb region spanning the TNF cluster in human MHC. Genomics 31, 215-222.

Neeper, M., Schmidt, A.M., Brett, J., Yan, S.D., Wang, F., Pan, Y.C., Elliston, K., Stern, D., Shaw, A., 1992. Cloning and expression of a cell surface receptor for advanced glycosylation end products of proteins. J Biol Chem 267, 14998-15004.

Neville, M.J., Campbell, R.D., 1999. A new member of the Ig superfamily and a V-ATPase G subunit are among the predicted products of novel genes close to the TNF locus in the human MHC. J Immunol 162, 4745-4754.

O

O'Donovan, C., Martin, M.J., Gattiker, A., Gasteiger, E., Bairoch, A., Apweiler, R., 2002. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinform 3, 275-284.

Odelberg, S.J., Weiss, R.B., Hata, A., White, R., 1995. Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. Nucleic acids research 23, 2049-2057.

P

Pagani, I., Liolios, K., Jansson, J., Chen, I.M., Smirnova, T., Nosrat, B., Markowitz, V.M., Kyrpides, N.C., 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic acids research 40, D571-579.

Pal, S., Gupta, R., Davuluri, R.V., 2012. Alternative transcription and alternative splicing in cancer. Pharmacology & Therapeutics 136, 283-294.

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. Nature genetics 40, 1413-1415.

Pandini, G., Frasca, F., Mineo, R., Sciacca, L., Vigneri, R., Belfiore, A., 2002. Insulin/insulin-like growth factor I hybrid receptors have different biological characteristics depending on the insulin receptor isoform involved. The Journal of biological chemistry 277, 39684-39695.

Peelman, L.J., Chardon, P., Vaiman, M., Mattheeuws, M., Van Zeveren, A., Van de Weghe, A., Bouquet, Y., Campbell, R.D., 1996. A detailed physical map of the porcine major histocompatibility complex (MHC) class III region: comparison with human and mouse MHC class III regions. Mammalian genome : official journal of the International Mammalian Genome Society 7, 363-367.

Pende, D., Parolini, S., Pessino, A., Sivori, S., Augugliaro, R., Morelli, L., Marcenaro, E., Accame, L., Malaspina, A., Biassoni, R., Bottino, C., Moretta, L., Moretta, A., 1999. Identification and molecular characterization of NKp30, a novel triggering receptor involved in natural cytotoxicity mediated by human natural killer cells. J Exp Med 190, 1505-1516.

Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8, 785-786.

Petkov, S.G., Marks, H., Klein, T., Garcia, R.S., Gao, Y., Stunnenberg, H., Hyttel, P., 2011. In vitro culture and characterization of putative porcine embryonic germ cells derived from domestic breeds and Yucatan mini pig embryos at Days 20-24 of gestation. Stem Cell Res 6, 226-237.

Pfaffl, M.W., 2001. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res 29, e45.

Pickrell, J.K., Pai, A.A., Gilad, Y., Pritchard, J.K., 2010. Noisy splicing drives mRNA isoform diversity in human cells. PLoS genetics 6, e1001236.

Pipes, L., Li, S., Bozinoski, M., Palermo, R., Peng, X., Blood, P., Kelly, S., Weiss, J.M., Thierry-Mieg, J., Thierry-Mieg, D., Zumbo, P., Chen, R., Schroth, G.P., Mason, C.E., Katze, M.G., 2013. The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. Nucleic acids research 41, D906-914.

Pogge von Strandmann, E., Simhadri, V.R., von Tresckow, B., Sasse, S., Reiners, K.S., Hansen, H.P., Rothe, A., B√∂ll, B., Simhadri, V.L., Borchmann, P., McKinnon, P.J., Hallek, M., Engert, A., 2007. Human Leukocyte Antigen-B-Associated Transcript 3 Is Released from Tumor Cells and Engages the NKp30 Receptor on Natural Killer Cells. Immunity 27, 965-974.

Porgador, A., 2005. Natural cytotoxicity receptors: pattern recognition and involvement of carbohydrates. ScientificWorldJournal 5, 151-154.

Prada, N., Antoni, G., Commo, F., Rusakiewicz, S., Semeraro, M., Boufassa, F., Lambotte, O., Meyer, L., Gougeon, M.-L., Zitvogel, L., 2013. Analysis of NKp30/NCR3 isoforms in untreated HIV-1-infected patients from the ANRS SEROCO cohort. Oncoimmunology 2, e23472.

Q

Qin, J., Mamotte, C., Cockett, N.E., Wetherall, J.D., Groth, D.M., 2008. A map of the class III region of the sheep major histocompatibilty complex. BMC Genomics 9, 409.

R

Rastrojo, A., Neves, R., Aguado, B., Guantes, R., Iborra, F., 2013. Global variability in gene expression and alternative splicing is modulated by mitochondrial content. En revisión.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D., 1997. GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13, 163.

Renard, C., Hart, E., Sehra, H., Beasley, H., Coggill, P., Howe, K., Harrow, J., Gilbert, J., Sims, S., Rogers, J., Ando, a., Shigenari, a., Shiina, T., Inoko, H., Chardon, P., Beck, S., 2006. The genomic sequence and analysis of the swine major histocompatibility complex. Genomics 88, 96-110.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16, 276-277.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nature biotechnology 29, 24-26.

Rogers, J., Early, P., Carter, C., Calame, K., Bond, M., Hood, L., Wall, R., 1980. Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. Cell 20, 303-312.

Rozen, S., Skaletsky, H., 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132, 365-386.

Rutjens, E., Mazza, S., Biassoni, R., Koopman, G., Radic, L., Fogli, M., Costa, P., Mingari, M.C., Moretta, L., Heeney, J., De Maria, A., 2007. Differential NKp30 inducibility in chimpanzee NK cells and conserved NK cell phenotype and function in long-term HIV-1-infected animals. Journal of immunology 178, 1702-1712.

Rutjens, E., Mazza, S., Biassoni, R., Koopman, G., Ugolotti, E., Fogli, M., Dubbes, R., Costa, P., Mingari, M.C., Greenwood, E.J., Moretta, L., De Maria, A., Heeney, J.L., 2010. CD8+ NK cells are predominant in chimpanzees, characterized by high NCR expression and cytokine production, and preserved in chronic HIV-1 infection. European journal of immunology 40, 1440-1450.

S

Samborski, A., Graf, A., Krebs, S., Kessler, B., Bauersachs, S., 2013. Deep sequencing of the porcine endometrial transcriptome on day 14 of pregnancy. Biol Reprod 88, 84.

Sanchez, M.J., Muench, M.O., Roncarolo, M.G., Lanier, L.L., Phillips, J.H., 1994. Identification of a common T/natural killer cell progenitor in human fetal thymus. The Journal of experimental medicine 180, 569-576.

Scally, A., Dutheil, J.Y., Hillier, L.W., Jordan, G.E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S.H., Schwalie, P.C., Tang, Y.A., Ward, M.C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L.N., Ayub, Q., Ball, E.V., Beal, K., Bradley, B.J., Chen, Y., Clee, C.M., Fitzgerald, S., Graves, T.A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G.K., Lunter, G., Meader, S., Mort, M., Mullikin, J.C., Munch, K., O'Connor, T.D., Phillips, A.D., Prado-Martinez, J., Rogers, A.S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J.T., Stenson, P.D., Turner, D.J., Vigilant, L., Vilella, A.J., Whitener, W., Zhu, B., Cooper, D.N., de Jong, P., Dermitzakis, E.T., Eichler, E.E., Flicek, P., Goldman, N., Mundy, N.I., Ning, Z., Odom, D.T., Ponting, C.P., Quail, M.A., Ryder, O.A., Searle, S.M., Warren, W.C., Wilson, R.K., Schierup, M.H., Rogers, J., Tyler-Smith, C., Durbin, R., 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483, 169-175.

Schmidt, S., Zimmermann, S.-Y., Tramsen, L., Koehl, U., Lehrnbecher, T., 2013. Natural Killer Cells and Antifungal Host Response. Clinical and Vaccine Immunology 20, 452-458.

Schroeder, H.W., Jr., Zhu, Z.B., March, R.E., Campbell, R.D., Berney, S.M., Nedospasov, S.A., Turetskaya, R.L., Atkinson, T.P., Go, R.C., Cooper, M.D., Volanakis, J.E., 1998. Susceptibility locus for IgA deficiency and common variable immunodeficiency in the HLA-DR3, -B8, -A1 haplotypes. Mol Med 4, 72-86.

Schultz, J., Milpetz, F., Bork, P., Ponting, C.P., 1998. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A 95, 5857-5864.

Sequencing, T.C., Consortium, A., 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69-87.

Sharp, P.A., 2005. The discovery of split genes and RNA splicing. Trends Biochem Sci 30, 279-281.

Shen, H., Green, M.R., 2006. RS domains contact splicing signalsand promote splicing by a commonmechanism in yeast through humans. Genes Dev 20, 1755-1765.

Shichi, D., Kikkawa, E.F., Ota, M., Katsuyama, Y., Kimura, A., Matsumori, A., Kulski, J.K., Naruse, T.K., Inoko, H., 2005. The haplotype block, NFKBIL1-ATP6V1G2-BAT1-MICB-MICA, within the class III-class I boundary region of the human major histocompatibility complex may control susceptibility to hepatitis C virus-associated dilated cardiomyopathy. Tissue Antigens 66, 200-208.

Sigrist, C.J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., Bucher, P., 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 3, 265-274.

Simhadri, V.R., Reiners, K.S., Hansen, H.P., Topolar, D., Simhadri, V.L., Nohroudi, K., Kufer, T.A., Engert, A., Pogge von Strandmann, E., 2008. Dendritic cells release HLA-B-associated transcript-3 positive exosomes to regulate natural killer function. PLoS ONE 3, e3377.

Sivakamasundari, R., Raghunathan, A., Chowdhury, C.Z.R.-r., Weissman, S.M., 2000. Expression and cellular localization of the protein encoded by the 1C7 gene : a recently described component of the MHC. 7, 723-732.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., Kasprzyk, A., 2009. BioMart--biological queries made easy. BMC Genomics 10, 22.

Sonnhammer, E.L., Eddy, S.R., Durbin, R., 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28, 405-420.

Spingola, M., Grate, L., Haussler, D., Ares, M., 1999. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. Rna 5, 221-234.

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.

Steentoft, C., Vakhrushev, S.Y., Joshi, H.J., Kong, Y., Vester-Christensen, M.B., Schjoldager, K.T., Lavrsen, K., Dabelsteen, S., Pedersen, N.B., Marcos-Silva, L., Gupta, R., Bennett, E.P., Mandel, U., Brunak, S., Wandall, H.H., Levery, S.B., Clausen, H., 2013. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. EMBO J 32, 1478-1488.

Stephens, R., Horton, R., Humphray, S., Rowen, L., Trowsdale, J., Beck, S., 1999. Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. Journal of Molecular Biology 291, 789-799.

Sureau, A., Perbal, B., 1994. Several mRNAs with variable 3' untranslated regions and different stability encode the human PR264/SC35 splicing factor. Proceedings of the National Academy of Sciences of the United States of America 91, 932-936.

Sutherland, C.L., Rabinovich, B., Chalupny, N.J., Brawand, P., Miller, R., Cosman, D., 2006. ULBPs, human ligands of the NKG2D receptor, stimulate tumor immunity with enhancement by IL-15. Blood 108, 1313-1319.

Suzuki, H., Kameyama, T., Ohe, K., Tsukahara, T., Mayeda, A., 2013. Nested introns in an intron: evidence of multi-step splicing in a large intron of the human dystrophin pre-mRNA. FEBS letters 587, 555-561.

T

Tarn, W.Y., Steitz, J.a., 1996. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. Cell 84, 801-811.

The MHC sequencing consortium, 1999. Complete sequence and gene map of a human major histocompatibility complex. Nature 401, 921-923.

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics (Oxford, England) 25, 1105-1111.

U

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., Leunissen, J.A., 2007. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res 35, W71-74.

V

Vandiedonck, C., Knight, J.C., 2009. The human Major Histocompatibility Complex as a paradigm in genomics research. Brief Funct Genomic Proteomic 8, 379-394.

Vargas, D.Y., Shah, K., Batish, M., Levandoski, M., Sinha, S., Marras, S.A., Schedl, P., Tyagi, S., 2011. Single-molecule imaging of transcriptionally coupled and uncoupled splicing. Cell 147, 1054-1065.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell,

T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., Zhu, X., 2001. The sequence of the human genome. Science 291, 1304-1351.

Villate, O., Rastrojo, a., López-Díez, R., Hernández-Torres, F., Aguado, B., 2008. Differential splicing, disease and drug targets. Infectious disorders drug targets 8, 241-251.

Vitale, M., Della Chiesa, M., Carlomagno, S., Pende, D., Arico, M., Moretta, L., Moretta, A., 2005. NK-dependent DC maturation is mediated by TNFalpha and IFNgamma released upon engagement of the NKp30 triggering receptor. Blood 106, 566-571.

W

Wahl, M.C., Will, C.L., Luhrmann, R., 2009. The spliceosome: design principles of a dynamic RNP machine. Cell 136, 701-718.

Wai, L.E., Garcia, J.A., Martinez, O.M., Krams, S.M., 2010. Distinct Roles for the NK Cell-Activating Receptors in Mediating Interactions with Dendritic Cells and Tumor Cells. J Immunol.

Walter, L., Hurt, P., Himmelbauer, H., Sudbrak, R., Günther, E., 2002. Physical mapping of the major histocompatibility complex class II and class III regions of the rat. Immunogenetics 54, 268-275.

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, 470-476.

Wang, G.S., Cooper, T.A., 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. Nature reviews. Genetics 8, 749-761.

Wang, M., Marin, A., 2006. Characterization and prediction of alternative splice sites. Gene 366, 219-227.

Ward, F.J., Dahal, L.N., Wijesekera, S.K., Abdul-Jawad, S.K., Kaewarpai, T., Xu, H., Vickers, M.A., Barker, R.N., 2013. The soluble isoform of CTLA-4 as a regulator of T-cell responses. European journal of immunology 43, 1274-1285.

Warren, H.S., Jones, A.L., Freeman, C., Bettadapura, J., Parish, C.R., 2005. Evidence that the cellular ligand for the human NK cell activation receptor NKp30 is not a heparan sulfate glycosaminoglycan. Journal of immunology 175, 207-212.

Wheelan, S.J., Church, D.M., Ostell, J.M., 2001. Spidey: a tool for mRNA-to-genomic alignments. Genome Res 11, 1952-1957.

Wilming, L.G., Hart, E.A., Coggill, P.C., Horton, R., Gilbert, J.G., Clee, C., Jones, M., Lloyd, C., Palmer, S., Sims, S., Whitehead, S., Wiley, D., Beck, S., Harrow, J.L., 2013. Sequencing and comparative analysis of the gorilla MHC genomic sequence. Database : the journal of biological databases and curation 2013, bat011.

Wu, Q., Krainer, A.R., 1999. AT-AC Pre-mRNA Splicing Mechanisms and Conservation of Minor Introns in Voltage-Gated Ion Channel Genes MINIREVIEW AT-AC Pre-mRNA Splicing Mechanisms and Conservation of Minor Introns in Voltage-Gated Ion Channel Genes. 19.

X

Xiao, S., Xie, D., Cao, X., Yu, P., Xing, X., Chen, C.C., Musselman, M., Xie, M., West, F.D., Lewin, H.A., Wang, T., Zhong, S., 2012. Comparative epigenomic annotation of regulatory DNA. Cell 149, 1381-1392.

Xie, T., Rowen, L., Aguado, B., Ahearn, M.E., Madan, A., Qin, S., Campbell, R.D., Hood, L., 2003. Analysis of the gene-dense major histocompatibility complex class III region and its comparison to mouse. Genome Res 13, 2621-2636.

Z

Zdobnov, E.M., Apweiler, R., 2001. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847-848.

Zhang, Z., Xin, D., Wang, P., Zhou, L., Hu, L., Kong, X., Hurst, L., 2009. Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. BMC Biol 7, 1-13.

8. Anexo

Tablas suplementarias

Especie	Nombre científico	Tejido/Línea celular	Sexo	Edad	Estadio del desarrollo	Procedencia de la muestras	Descripción
Humano	*Homo sapiens*	Sangre	Hombre	50 años	Adulto	Comercial	PBMCs
Humano	*Homo sapiens*	Sangre	Mujer	60 años	Adulto	Hospital Universitario "La Paz" de Madrid	PBMCs
Humano	*Homo sapiens*	Sangre	Hombre	38 años	Adulto	Hospital Universitario "La Paz" de Madrid	PBMCs
Humano	*Homo sapiens*	Sangre	Mujer	31 años	Adulto	Hospital Universitario "La Paz" de Madrid	PBMCs
Humano	*Homo sapiens*	Cerebro	Hombre	30 años	Adulto	Comercial	Cerebro adulto
Humano	*Homo sapiens*	Riñón	Hombre	65 años	Adulto	Comercial	Riñón adulto
Humano	*Homo sapiens*	Hígado	Hombre	64 años	Adulto	Comercial	Hígado adulto
Humano	*Homo sapiens*	Pulmón	Hombre	24 años	Adulto	Comercial	Pulmón adulto
Humano	*Homo sapiens*	Páncreas	Hombre	44 años	Adulto	Comercial	Páncreas adulto
Humano	*Homo sapiens*	Bazo	Hombre	29 años	Adulto	Comercial	Bazo adulto
Humano	*Homo sapiens*	Timo	Hombre	21 años	Adulto	Comercial	Timo adulto
Humano	*Homo sapiens*	Cerebro	Mujer	28 semanas	Fetal	Comercial	Cerebro fetal
Humano	*Homo sapiens*	Corazón	Mujer	33 semanas	Fetal	Comercial	Corazón fetal
Humano	*Homo sapiens*	Riñón	Mujer	30 semanas	Fetal	Comercial	Riñón fetal
Humano	*Homo sapiens*	Hígado	Hombre	38 semanas	Fetal	Comercial	Hígado fetal
Humano	*Homo sapiens*	Pulmón	Mujer	37 semanas	Fetal	Comercial	Pulmón fetal
Humano	*Homo sapiens*	Bazo	Hombre	40 semanas	Fetal	Comercial	Bazo fetal
Humano	*Homo sapiens*	Timo	Hombre	24 semanas	Fetal	Comercial	Timo fetal
Humano	*Homo sapiens*	Riñón	Mujer	28 años	Tumoral	Comercial	Riñón tumoral
Humano	*Homo sapiens*	Hígado	Hombre	65 años	Tumoral	Comercial	Hígado tumoral
Humano	*Homo sapiens*	Pulmón	Mujer	39 años	Tumoral	Comercial	Pulmón tumoral
Humano	*Homo sapiens*	HeLa	Mujer	-	Líneas celulares establecidas	Obtenida en el laboratorio	Células epiteliales derivadas de un adenocarcinoma de cuello de útero
Humano	*Homo sapiens*	Jurkat	Hombre	-	Líneas celulares establecidas	Obtenida en el laboratorio	Células T derivadas de una leucemia
Humano	*Homo sapiens*	K562	Mujer	-	Líneas celulares establecidas	Obtenida en el laboratorio	Granulocitos indiferenciados derivados de una leucemia
Humano	*Homo sapiens*	Raji	Hombre	-	Líneas celulares establecidas	Obtenida en el laboratorio	Células B derivadas de un limfoma
Humano	*Homo sapiens*	U937	Hombre	-	Líneas celulares establecidas	Obtenida en el laboratorio	Monocitos derivadas de un limfoma
Humano	*Homo sapiens*	YT	-	-	Líneas celulares establecidas	Obtenida en el laboratorio	Células NKs derivadas de un limfoma
Gorilla	*Gorilla gorilla*	Sangre	Hembra	14 años	Adulto	Zoo-Acuarium de Madrid	PBMCs
Gorilla	*Gorilla gorilla*	Sangre	Hembra	11 años	Adulto	Zoo-Acuarium de Madrid	PBMCs
Orangután de Borneo	*Pongo pygmaeus*	Sangre	Macho	20 años	Adulto	Zoo-Acuarium de Madrid	PBMCs
Orangután de Borneo	*Pongo pygmaeus*	Sangre	Macho	18 años	Adulto	Zoo-Acuarium de Madrid	PBMCs
Colobo	*Colobus guereza*	Sangre	Macho	12 años	Adulto	Zoo-Acuarium de Madrid	PBMCs
Babuíno amarillo	*Papion cynocephalus*	Sangre	Macho	-	Adulto	Zoo-Acuarium de Madrid	PBMCs
Babuíno amarillo	*Papion cynocephalus*	Sangre	Macho	-	Adulto	Zoo-Acuarium de Madrid	PBMCs
Babuíno amarillo	*Papion cynocephalus*	Sangre	Hembra	-	Adulto	Zoo-Acuarium de Madrid	PBMCs
Macaco Rhesus	*Macaca mulatta*	Cerebro	-	-	Adulto	Comercial	Cerebro adulto
Macaco Rhesus	*Macaca mulatta*	Corazón	Hembra	4.5 años	Adulto	Comercial	Corazón adulto
Macaco Rhesus	*Macaca mulatta*	Riñón	Hembra	4.5 años	Adulto	Comercial	Riñón adulto
Macaco Rhesus	*Macaca mulatta*	Hígado	Macho	4.5 años	Adulto	Comercial	Hígado adulto
Macaco Rhesus	*Macaca mulatta*	Pulmón	Hembra	3.5 años	Adulto	Comercial	Pulmón adulto
Macaco Rhesus	*Macaca mulatta*	Páncreas	Macho	2.5 años	Adulto	Comercial	Páncreas adulto
Macaco cangrejero	*Macaca fascicularis*	Sangre	Macho	14 años	Adulto	Zoo-Acuarium de Madrid	PBMCs
Macaco cangrejero	*Macaca fascicularis*	Sangre	Macho	12 años	Adulto	Zoo-Acuarium de Madrid	PBMCs
Mono capuchino	*Cebus apella*	Sangre	Hembra	14 años	Adulto	Zoo-Acuarium de Madrid	PBMCs

Tabla S1. Detalle de las muestras de RNA utilizadas en el análisis del splicing alternativo del gen *NCR3*.

Especie	Nombre científico	Tejido/Línea celular	Sexo	Edad	Estadio del desarrollo	Procedencia de la muestras	Descripción
Rata	*Rattus norvegicus*	Cerebro	Macho	10 semanas	Adulto	Comercial	Cerebro adulto (4 individuos)
Rata	*Rattus norvegicus*	Corazón	Hembra	8-11 semanas	Adulto	Comercial	Corazón adulto (7 individuos)
Rata	*Rattus norvegicus*	Riñón	Macho y Hembra	8-12 semanas	Adulto	Comercial	Riñón adulto (3 mahos y 3 hembras)
Rata	*Rattus norvegicus*	Hígado	Hembra	13 semanas	Adulto	Comercial	Hígado (1 individuo)
Rata	*Rattus norvegicus*	Pulmón	Macho	24 semanas	Adulto	Comercial	Pulmón adulto (3 individuos)
Rata	*Rattus norvegicus*	Páncreas	Hembra	9 semanas	Adulto	Comercial	Páncreas adulto (1 individuo)
Rata	*Rattus norvegicus*	Bazo	Hembra	13 semanas	Adulto	Comercial	Bazo adulto (10 individuos)
Rata	*Rattus norvegicus*	Embrión 19 días	-	19 días	Fetal	Comercial	Embrión (2 individuos)
Ratón	*Mus musculus*	Cerebro	Macho	12 semanas	Adulto	Comercial	Cerebro adulto (1 individuo)
Ratón	*Mus musculus*	Corazón	-	-	Adulto	Comercial	Corazón adulto (15 individuos)
Ratón	*Mus musculus*	Riñón	Hembra	4 semanas	Adulto	Comercial	Riñón adulto (20 individuos)
Ratón	*Mus musculus*	Hígado	Hembra	8 semanas	Adulto	Comercial	Hígado adulto (1 individuo)
Ratón	*Mus musculus*	Pulmón	Hembra	4 semanas	Adulto	Comercial	Pulmón adulto (1 individuo)
Ratón	*Mus musculus*	Páncreas	Hembra	8.5 semanas	Adulto	Comercial	Páncreas adulto (1 individuo)
Ratón	*Mus musculus*	Embrión 11 días	-	11 días	Fetal	Comercial	Embrión (5 individuos)
Ratón	*Mus musculus*	Embrión 15 días	-	15 días	Fetal	Comercial	Embrión (2 individuos)
Ratón	*Mus musculus*	Embrión 17 días	-	17 días	Fetal	Comercial	Embrión (2 individuos)
Vaca	*Bos taurus*	Cerebro	Hembra	2 años	Adulto	Comercial	Cerebro adulto
Vaca	*Bos taurus*	Corazón	Hembra	2 años	Adulto	Comercial	Corazón adulto
Vaca	*Bos taurus*	Riñón	Hembra	2 años	Adulto	Comercial	Riñón adulto
Vaca	*Bos taurus*	Hígado	Hembra	2 años	Adulto	Comercial	Hígado adulto
Vaca	*Bos taurus*	Pulmón	Hembra	2 años	Adulto	Comercial	Pulmón adulto
Vaca	*Bos taurus*	Bazo	Hembra	2 años	Adulto	Comercial	Bazo adulto
Cerdo	*Sus scrofa*	Cerebro	Hembra	7.5 meses	Adulto	Comercial	Cerebro adulto
Cerdo	*Sus scrofa*	Corazón	Hembra	7.5 meses	Adulto	Comercial	Corazón adulto
Cerdo	*Sus scrofa*	Riñón	Hembra	7.5 meses	Adulto	Comercial	Riñón adulto
Cerdo	*Sus scrofa*	Hígado	Hembra	7.5 meses	Adulto	Comercial	Hígado adulto
Cerdo	*Sus scrofa*	Pulmón	Hembra	7.5 meses	Adulto	Comercial	Pulmón adulto
Cerdo	*Sus scrofa*	Bazo	Hembra	7.5 meses	Adulto	Comercial	Bazo adulto

Tabla S1 continuación. Detalle de las muestras de RNA utilizadas en el análisis del splicing alternativo del gen *NCR3*.

Anexo

Homo sapiens

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
A549	Línea celular	A549	2x75	70.457.234	133.257.238	Proyecto Encode	SRR307907 SRR307908	Células epiteliales derivadas de un carcinoma de pulmón
HeLa	Línea celular	HeLa	2x50	54.497.457	98.872.112	No publicado	-	Células epiteliales derivadas de un adenocarcinoma de cuello de útero
HUVEC	Línea celular	Huvec	2x75	72.436.137	17.085.217	Proyecto Encode	SRR065509	Células derivadas del endotelio de la vena umbilical
K562	Línea celular	K562	1x75	62.732.735	44.264.804	Proyecto Encode	SRR521468 SRR534174 SRR534175	Granulocitos indiferenciados derivados de una leucemia
Jurkat	Línea celular	Jurkat	1x36	45.756.579	29.041.789	Proyecto Encode	SRR492030 SRR492031	Células T derivadas de un limfoma
Células T+PMA	Línea celular	Células T	1x50	46.251.043	49.354.842	Martinez et al., 2012	SRR408754 SRR408755 SRR408756	Células JSL-1 (linfocitos T) activados con PMA
Células T 1	Línea celular	Células T	1x50	47.881.096	53.412.582	Martinez et al., 2012	SRR408751 SRR408752 SRR408753	Células JSL-1 (linfocitos T) en reposo
Células T 2	Línea celular	Células T	1x75	30.932.008	24.870.321	Chang et al., 2011	SRR497705	Células T (SUP-T1) derivadas de un limfoma
Células T 3	Línea celular	Células T	1x75	32.330.868	26.819.184	Chang et al., 2011	SRR497706	Células T (SUP-T1) derivadas de un limfoma
Células T 4	Línea celular	Células T	1x75	30.121.789	20.078.000	Chang et al., 2011	SRR497707	Células T (SUP-T1) derivadas de un limfoma
Células T 5	Línea celular	Células T	1x75	32.383.846	30.268.430	Chang et al., 2011	SRR497708	Células T (SUP-T1) derivadas de un limfoma
Células T 6	Línea celular	Células T	1x75	25.428.142	19.737.092	Chang et al., 2011	SRR497709	Células T (SUP-T1) derivadas de un limfoma
Células T+HIV 1	Línea celular	Células T	1x75	31.362.049	20.160.072	Chang et al., 2011	SRR497699	Células T (SUP-T1) derivadas de un limfoma e infectadas con HIV
Células T+HIV 2	Línea celular	Células T	1x75	35.603.888	23.993.255	Chang et al., 2011	SRR497700	Células T (SUP-T1) derivadas de un limfoma e infectadas con HIV
Células T+HIV 3	Línea celular	Células T	1x75	29.770.516	12.974.856	Chang et al., 2011	SRR497701	Células T (SUP-T1) derivadas de un limfoma e infectadas con HIV
Células T+HIV 4	Línea celular	Células T	1x75	23.936.312	13.200.351	Chang et al., 2011	SRR497702	Células T (SUP-T1) derivadas de un limfoma e infectadas con HIV
Células T+HIV 5	Línea celular	Células T	1x75	24.419.121	12.731.631	Chang et al., 2011	SRR497703	Células T (SUP-T1) derivadas de un limfoma e infectadas con HIV
Células T+HIV 6	Línea celular	Células T	1x75	23.372.815	13.079.882	Chang et al., 2011	SRR497704	Células T (SUP-T1) derivadas de un limfoma e infectadas con HIV
PBMCs_1	Tejido	PBMCs	2x108	23.123.872	18.732.316	No publicado	-	PBMCs obtenidos de un donante sano
PBMCs_2	Tejido	PBMCs	2x108	20.277.787	14.371.738	No publicado	-	PBMCs obtenidos de un donante sano
PBMCs_3	Tejido	PBMCs	2x101	236.532.985	367.740.205	Proyecto Encode	SRR545689 SRR545690	PBMCs obtenidos de un donante sano
PBMCs+LPS	Tejido	PBMCs	2x101	83.210.901	79.723.569	No publicado	SRR629558	PBMCs tratados con LPS

Tabla S2. Secuencias procedentes de experimentos de RNA-seq utilizados en el análisis del splicing alternativo del gen *NCR3*.

Homo sapiens

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Monocitos 1	Tejido	Monocitos	2x77	202.993.398	396.413.943	Proyecto Encode	SRR545695 SRR545696 SRR545697 SRR545698 SRR545699 SRR545700	Monocitos (CD14+)
Monocitos 2	Tejido	Monocitos	2x101	36.443.544	18.826.146	Beyer et al., 2012	SRR452328	Monocitos M1
Monocitos 3	Tejido	Monocitos	2x101	30.952.877	27.895.713	Beyer et al., 2012	SRR452329	Monocitos M1
Monocitos 4	Tejido	Monocitos	2x101	22.072.388	13.649.664	Beyer et al., 2012	SRR452330	Monocitos M1
Monocitos 5	Tejido	Monocitos	2x101	51.417.553	51.452.790	Beyer et al., 2012	SRR452331	Monocitos M2
Monocitos 6	Tejido	Monocitos	2x101	15.585.860	12.525.415	Beyer et al., 2012	SRR452332	Monocitos M2
Monocitos 7	Tejido	Monocitos	2x101	49.258.388	38.807.954	Beyer et al., 2012	SRR452333	Monocitos M2
Células B	Tejido	Células B	2x77	169.661.447	322.655.255	Proyecto Encode	SRR534319 SRR534320 SRR534321 SRR534322 SRR534323 SRR534324	Células B (CD20+)
Cerebro 1	Tejido	Cerebro	1x32	17.246.957	17.932.079	Pan et al., 2008	SRR036966	Cerebro
Cerebro 2	Tejido	Cerebro	1x32	31.940.303	32.349.055	Pan et al., 2008	SRR014262	Córtex Cerebral
Cerebro 3	Tejido	Cerebro	1x75	24.513.415	18.202.098	Barbosa-Morais et al., 2012	SRR306838	Cerebro
Cerebro 5	Tejido	Cerebro	1x75	22.576.705	2.336.081	Barbosa-Morais et al., 2012	SRR306840	Cerebro
Cerebro 6	Tejido	Cerebro	1x75	24.325.223	15.500.956	Barbosa-Morais et al., 2012	SRR306841	Cerebro
Cerebro 4	Tejido	Cerebro	1x75	18.850.030	10.792.955	Barbosa-Morais et al., 2012	SRR306839	Cerebro
Cerebro 7	Tejido	Cerebro	1x75	17.422.994	3.519.874	Barbosa-Morais et al., 2012	SRR306842	Cerebro
Cerebro 8	Tejido	Cerebro	1x75	7.913.181	2.491.510	Barbosa-Morais et al., 2012	SRR306843	Cerebro
Cerebro 9	Tejido	Cerebro	1x75	32,698,558	22.803.117	Barbosa-Morais et al., 2012	SRR306844	Cerebelo
Cerebro 10	Tejido	Cerebro	1x75	46.755.221	27.686.265	Barbosa-Morais et al., 2012	SRR306845 SRR306846	Cerebelo
Corazón 1	Tejido	Corazón	1x32	20.169.301	20.169.301	Pan et al., 2008	SRR014263	Corazón
Corazón 2	Tejido	Corazón	1x75	24.128.204	17.703.319	Barbosa-Morais et al., 2012	SRR306847	Corazón
Corazón 3	Tejido	Corazón	1x75	30.896.351	18.212.717	Barbosa-Morais et al., 2012	SRR306848 SRR306849	Corazón
Corazón 4	Tejido	Corazón	1x75	25.197.713	14.621.873	Barbosa-Morais et al., 2012	SRR306850	Corazón
Hígado 1	Tejido	Hígado	1x75	43.147.061	28.457.392	Barbosa-Morais et al., 2012	SRR306854 SRR306855	Hígado
Hígado 2	Tejido	Hígado	1x75	23.866.499	17.612.733	Barbosa-Morais et al., 2012	SRR306856	Hígado
Hígado 3	Tejido	Hígado	1x32	18.517.121	20.019.668	Pan et al., 2008	SRR014264	Hígado
Hígado 4	Tejido	Hígado	1x32	5.225.421	5.822.277	Blekhman et al., 2010	SRR032116	Hígado
Hígado 5	Tejido	Hígado	1x32	5.378.161	6.017.610	Blekhman et al., 2010	SRR032117	Hígado
Hígado 6	Tejido	Hígado	1x32	4.983.423	5.328.583	Blekhman et al., 2010	SRR032118	Hígado
Hígado 7	Tejido	Hígado	1x32	5.477.129	6.064.704	Blekhman et al., 2010	SRR032119	Hígado
Hígado 8	Tejido	Hígado	1x32	6.592.168	7.442.403	Blekhman et al., 2010	SRR032120	Hígado
Hígado 9	Tejido	Hígado	1x32	6.820.205	7.755.211	Blekhman et al., 2010	SRR032121	Hígado
Músculo	Tejido	Músculo	1x32	22.640.454	24.302.639	Pan et al., 2008	SRR014266	Músculo esquelético
Pulmón 1	Tejido	Pulmón	1x32	25.862.057	27.155.165	Pan et al., 2008	SRR014265	Pulmón

Tabla S2 continuación. Secuencias procedentes de experimentos de RNA-seq utilizados en el análisis del splicing alternativo del gen *NCR3*.

Homo sapiens

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Pulmón 2	Tejido	Pulmón	1x43	13.025.469	14.322.935	Hackett et al., 2012	SRR094912	Pulmón fumador
Pulmón 3	Tejido	Pulmón	1x43	19.179.727	19.062.028	Hackett et al., 2012	SRR094913	Pulmón fumador
Pulmón 4	Tejido	Pulmón	1x43	16.615.218	16.981.614	Hackett et al., 2012	SRR094911	Pulmón fumador
Pulmón 5	Tejido	Pulmón	1x43	14.204.752	13.128.969	Hackett et al., 2012	SRR094910	Pulmón fumador
Pulmón 6	Tejido	Pulmón	1x43	15.201.510	16.238.322	Hackett et al., 2012	SRR094909	Pulmón fumador
Pulmón 7	Tejido	Pulmón	1x43	15.837.368	16.447.042	Hackett et al., 2012	SRR094908	Pulmón fumador
Pulmón 8	Tejido	Pulmón	1x43	17.670.811	19.571.068	Hackett et al., 2012	SRR094907	Pulmón no fumador
Pulmón 9	Tejido	Pulmón	1x43	19.788.339	20.255.816	Hackett et al., 2012	SRR094906	Pulmón no fumador
Pulmón 10	Tejido	Pulmón	1x43	15.886.268	17.166.022	Hackett et al., 2012	SRR094864	Pulmón no fumador
Pulmón 11	Tejido	Pulmón	1x43	17.193.664	18.376.019	Hackett et al., 2012	SRR094862	Pulmón no fumador
Pulmón 12	Tejido	Pulmón	1x43	17.356.854	18.381.134	Hackett et al., 2012	SRR094863	Pulmón no fumador
Riñón 1	Tejido	Riñón	1x75	22.493.518	16.399.162	Barbosa-Morais et al., 2012	SRR306851	Riñón
Riñón 2	Tejido	Riñón	1x75	20.684.752	14.615.495	Barbosa-Morais et al., 2012	SRR306852	Riñón
Riñón 3	Tejido	Riñón	1x75	31.386.619	19.274.749	Barbosa-Morais et al., 2012	SRR306853	Riñón
Testículo 1	Tejido	Testículo	1x75	9.224.054	4.279.539	Barbosa-Morais et al., 2012	SRR306857	Testículo
Testículo 2	Tejido	Testículo	1x75	32.444.809	22.483.468	Barbosa-Morais et al., 2012	SRR306858	Testículo

Macaca mulatta

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Bazo 1	Tejido	Bazo	2x40	40.625.316	85.438.428	Merkin et al., 2012	SRR594453	Bazo
Bazo 2	Tejido	Bazo	2x80	97.713.278	155.136.020	Merkin et al., 2012	SRR594462	Bazo
Bazo 3	Tejido	Bazo	2x40	39.812.460	82.044.659	Merkin et al., 2012	SRR594471	Bazo
Cerebro 1	Tejido	Cerebro	1x75	19.068.947	10.087.465	Barbosa-Morais et al., 2012	SRR306777	Cerebro
Cerebro 2	Tejido	Cerebro	1x75	22.554.234	10.939.123	Barbosa-Morais et al., 2012	SRR306778	Cerebro
Cerebro 3	Tejido	Cerebro	1x75	21.461.283	11.510.153	Barbosa-Morais et al., 2012	SRR306779	Cerebro
Cerebro 4	Tejido	Cerebro	1x75	25.528.147	12.268.379	Barbosa-Morais et al., 2012	SRR306780	Cerebro
Cerebro 5	Tejido	Cerebro	1x75	21.141.815	11.112.775	Barbosa-Morais et al., 2012	SRR306781	Cerebro
Cerebro 6	Tejido	Cerebro	2x40	35.066.763	68.959.635	Merkin et al., 2012	SRR594446	Cerebro
Cerebro 7	Tejido	Cerebro	2x80	107.669.551	181.384.566	Merkin et al., 2012	SRR594455	Cerebro
Cerebro 8	Tejido	Cerebro	2x40	26.487.487	53.237.336	Merkin et al., 2012	SRR594464	Cerebro
Corazón 1	Tejido	Corazón	1x75	28.636.572	11.692.168	Barbosa-Morais et al., 2012	SRR306782	Corazón
Corazón 2	Tejido	Corazón	1x75	20.815.484	9.782.071	Barbosa-Morais et al., 2012	SRR306783	Corazón
Corazón 3	Tejido	Corazón	2x40	35.248.042	66.738.600	Merkin et al., 2012	SRR594448	Corazón
Corazón 4	Tejido	Corazón	2x80	109.193.093	179.013.629	Merkin et al., 2012	SRR594457	Corazón
Corazón 5	Tejido	Corazón	2x40	36.101.619	69.365.864	Merkin et al., 2012	SRR594466	Corazón
Hígado 1	Tejido	Hígado	1x75	21.711.196	11.654.151	Barbosa-Morais et al., 2012	SRR306786	Hígado
Hígado 2	Tejido	Hígado	1x75	32.224.651	19.864.182	Barbosa-Morais et al., 2012	SRR306787 SRR306788	Hígado
Hígado 3	Tejido	Hígado	2x40	28.555.788	55.504.501	Merkin et al., 2012	SRR594450	Hígado
Hígado 4	Tejido	Hígado	2x80	113.094.939	190.674.223	Merkin et al., 2012	SRR594459	Hígado
Hígado 5	Tejido	Hígado	2x40	26.700.682	52.435.141	Merkin et al., 2012	SRR594468	Hígado
Pulmón 1	Tejido	Pulmón	2x40	38.295.330	77.866.802	Merkin et al., 2012	SRR594451	Pulmón
Pulmón 2	Tejido	Pulmón	2x80	112.732.867	189.031.791	Merkin et al., 2012	SRR594460	Pulmón
Pulmón 3	Tejido	Pulmón	2x40	32.399.751	65.433.067	Merkin et al., 2012	SRR594469	Pulmón
Riñón 1	Tejido	Riñón	1x75	17.581.272	7.033.256	Barbosa-Morais et al., 2012	SRR306784	Riñón
Riñón 2	Tejido	Riñón	1x75	24.115.366	7.919.122	Barbosa-Morais et al., 2012	SRR306785	Riñón
Riñón 3	Tejido	Riñón	2x40	31.891.747	61.746.518	Merkin et al., 2012	SRR594449	Riñón

Tabla S2 continuación. Secuencias procedentes de experimentos de RNA-seq utilizados en el análisis del splicing alternativo del gen *NCR3*.

Macaca mulatta

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Riñón 4	Tejido	Riñón	2x80	108.637.672	177.583.266	Merkin et al., 2012	SRR594458	Riñón
Riñón 5	Tejido	Riñón	2x40	40.389.069	80.892.790	Merkin et al., 2012	SRR594467	Riñón
Testículo 1	Tejido	Testículo	1x75	23.550.934	11.390.840	Barbosa-Morais et al., 2012	SRR306789	Testículo
Testículo 2	Tejido	Testículo	1x75	32.751.616	20.290.775	Barbosa-Morais et al., 2012	SRR306790	Testículo
Células linfoblásticas 1	Tejido-Línea	Células linfoblásticas	1x36	6.935.762	7.722.459	Iskow et al., 2012	SRR504803	Células linfoblásticas
Células linfoblásticas 2	Tejido-Línea	Células linfoblásticas	1x36	8.722.182	9.882.573	Iskow et al., 2012	SRR504804	Células linfoblásticas
Células linfoblásticas 3	Tejido-Línea	Células linfoblásticas	1x36	8.848.655	10.490.643	Iskow et al., 2012	SRR504805	Células linfoblásticas
Células linfoblásticas 4	Tejido-Línea	Células linfoblásticas	1x36	13.335.185	14.446.788	Iskow et al., 2012	SRR504806	Células linfoblásticas
Células linfoblásticas 5	Tejido-Línea	Células linfoblásticas	1x36	15.212.293	17.789.985	Iskow et al., 2012	SRR504807	Células linfoblásticas
Células linfoblásticas 6	Tejido-Línea	Células linfoblásticas	1x36	17.484.757	19.582.478	Iskow et al., 2012	SRR504808	Células linfoblásticas

Rattus norvegicus

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Bazo 1	Tejido	Bazo	2x50	108.073.128	255.471.881	Merkin et al., 2012	SRR594426	Bazo
Bazo 2	Tejido	Bazo	2x80	112.500.969	235.904.870	Merkin et al., 2012	SRR594435	Bazo
Bazo 3	Tejido	Bazo	2x40	29.882.026	75.462.266	Merkin et al., 2012	SRR594444	Bazo
Cerebro 1	Tejido	Cerebro	2x80	119.038.900	221.611.967	Merkin et al., 2012	SRR594419	Cerebro
Cerebro 2	Tejido	Cerebro	2x80	96.368.839	164.999.044	Merkin et al., 2012	SRR594428	Cerebro
Cerebro 3	Tejido	Cerebro	2x40	32.262.802	68.132.919	Merkin et al., 2012	SRR594437	Cerebro
Corazón 1	Tejido	Corazón	2x40	35.802.852	87.856.189	Merkin et al., 2012	SRR594421	Corazón
Corazón 2	Tejido	Corazón	2x80	67.008.998	137.464.524	Merkin et al., 2012	SRR594430	Corazón
Corazón 3	Tejido	Corazón	2x40	25.869.367	60.476.295	Merkin et al., 2012	SRR594439	Corazón
Hígado 1	Tejido	Hígado	2x36	26.181.362	62.924.829	Merkin et al., 2012	SRR594423	Hígado
Hígado 2	Tejido	Hígado	2x80	131.658.529	271.766.412	Merkin et al., 2012	SRR594432	Hígado
Hígado 3	Tejido	Hígado	2x40	42.043.440	98.994.403	Merkin et al., 2012	SRR594441	Hígado
Pulmón 1	Tejido	Pulmón	2x50	117,269,672	261.218.964	Merkin et al., 2012	SRR594424	Pulmón
Pulmón 2	Tejido	Pulmón	2x80	84.087.214	163.703.525	Merkin et al., 2012	SRR594433	Pulmón
Pulmón 3	Tejido	Pulmón	2x40	20.835.407	46.211.300	Merkin et al., 2012	SRR594442	Pulmón
Riñón 1	Tejido	Riñón	2x50	114.089.612	258.393.051	Merkin et al., 2012	SRR594422	Riñón
Riñón 2	Tejido	Riñón	2x80	116.656.722	241.346.373	Merkin et al., 2012	SRR594431	Riñón
Riñón 3	Tejido	Riñón	2x40	40.114.787	90.431.863	Merkin et al., 2012	SRR594440	Riñón

Mus musculus

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Bazo_1	Tejido	Bazo	2x50	114.072.257	255.432.575	Merkin et al., 2012	SRR594400	Bazo
Bazo_2	Tejido	Bazo	2x80	114.814.142	214.692.757	Merkin et al., 2012	SRR594408	Bazo
Bazo_3	Tejido	Bazo	2x50	113.321.013	250.239.575	Merkin et al., 2012	SRR594417	Bazo
Cerebro_1	Tejido	Cerebro	1x75	76.170.802	53.064.101	Barbosa-Morais et al., 2012	SRR306758 SRR306759 SRR306760 SRR306761	Cerebro
Cerebro_2	Tejido	Cerebro	1x75	19.726.026	9.550.600	Barbosa-Morais et al., 2012	SRR306762	Cerebro
Cerebro_3	Tejido	Cerebro	1x75	41.340.785	24.268.078	Barbosa-Morais et al., 2012	SRR306763	Cerebro
Cerebro_4	Tejido	Cerebro	2x50	87.264.604	177.016.239	Merkin et al., 2012	SRR594393	Cerebro
Cerebro_5	Tejido	Cerebro	1x75	20.947.739	10.693.287	Barbosa-Morais et al., 2012	SRR306764	Cerebelo
Cerebro_6	Tejido	Cerebro	1x75	28.710.250	17.900.601	Barbosa-Morais et al., 2012	SRR306765	Cerebelo
Cerebro_7	Tejido	Cerebro	2x80	118.824.353	200.364.582	Merkin et al., 2012	SRR594402	Cerebro
Cerebro_8	Tejido	Cerebro	2x80	32.511.234	66.218.941	Merkin et al., 2012	SRR594410	Cerebro

Tabla S2 continuación. Secuencias procedentes de experimentos de RNA-seq utilizados en el análisis del splicing alternativo del gen *NCR3*.

Mus musculus

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Cerebro_9	Tejido	Cerebro	1x75	25.094.445	19.201.472	Barbosa-Morais et al., 2012	SRR306757	Cerebro
Corazón_1	Tejido	Corazón	1x75	44.668.984	25.008.946	Barbosa-Morais et al., 2012	SRR306766	Corazón
Corazón_2	Tejido	Corazón	1x75	24.493.681	16.469.930	Barbosa-Morais et al., 2012	SRR306767	Corazón
Corazón_3	Tejido	Corazón	1x75	25.903.961	18.154.201	Barbosa-Morais et al., 2012	SRR306768	Corazón
Corazón_4	Tejido	Corazón	2x36	35.175.982	78.095.213	Merkin et al., 2012	SRR594395	Corazón
Corazón_5	Tejido	Corazón	2x40	15.968.605	33.571.300	Merkin et al., 2012	SRR594412	Corazón
Hígado_1	Tejido	Hígado	1x75	48.306.727	22.309.280	Barbosa-Morais et al., 2012	SRR306772	Hígado
Hígado_2	Tejido	Hígado	2x50	116.292.478	299.833.542	Merkin et al., 2012	SRR594397	Hígado
Hígado_3	Tejido	Hígado	1x75	18.444.416	17.069.549	Barbosa-Morais et al., 2012	SRR306773	Hígado
Hígado_4	Tejido	Hígado	2x80	134.045.721	208.430.656	Merkin et al., 2012	SRR594405	Hígado
Hígado_5	Tejido	Hígado	1x75	34.010.208	29.932.099	Barbosa-Morais et al., 2012	SRR306774	Hígado
Hígado_6	Tejido	Hígado	2x40	34.824.609	104.527.412	Merkin et al., 2012	SRR594414	Hígado
Pulmón_1	Tejido	Pulmón	2x36	34.050.626	74.914.430	Merkin et al., 2012	SRR594398	Pulmón
Pulmón_2	Tejido	Pulmón	2x80	62.362.901	100.665.712	Merkin et al., 2012	SRR594406	Pulmón
Pulmón_3	Tejido	Pulmón	2x50	111.879.091	232.907.600	Merkin et al., 2012	SRR594415	Pulmón
Riñón_1	Tejido	Riñón	1x75	45.613.332	15.921.231	Barbosa-Morais et al., 2012	SRR306769	Riñón
Riñón_2	Tejido	Riñón	1x75	23.639.764	15.974.541	Barbosa-Morais et al., 2012	SRR306770	Riñón
Riñón_3	Tejido	Riñón	1x75	29.158.234	18.200.720	Barbosa-Morais et al., 2012	SRR306771	Riñón
Riñón_4	Tejido	Riñón	2x50	119.274.786	249.873.082	Merkin et al., 2012	SRR594396	Riñón
Riñón_5	Tejido	Riñón	2x75	118.885.190	231.393.521	Merkin et al., 2012	SRR594404	Riñón
Riñón_6	Tejido	Riñón	2x36	29.821.800	79.881.989	Merkin et al., 2012	SRR594413	Riñón
Testículo_1	Tejido	Testículo	1x75	29.762.970	19.245.597	Barbosa-Morais et al., 2012	SRR306776	Testículo

Bos taurus

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Bazo 1	Tejido	Bazo	2x80	116.884.523	213.812.023	Merkin et al., 2012	SRR594480	Bazo
Bazo 2	Tejido	Bazo	2x36	24.283.167	56.041.684	Merkin et al., 2012	SRR594489	Bazo
Bazo 3	Tejido	Bazo	2x36	31.925.721	74.034.095	Merkin et al., 2012	SRR594498	Bazo
Cerebro 1	Tejido	Cerebro	2x80	104.353.205	198.374.899	Merkin et al., 2012	SRR594473	Cerebro
Cerebro 2	Tejido	Cerebro	2x36	28.445.194	59.970.662	Merkin et al., 2012	SRR594482	Cerebro
Cerebro 3	Tejido	Cerebro	2x40	33.628.113	67.381.914	Merkin et al., 2012	SRR594491	Cerebro
Corazón 1	Tejido	Corazón	2x80	117.554.231	224.113.913	Merkin et al., 2012	SRR594475	Corazón
Corazón 2	Tejido	Corazón	2x36	21.185.451	44.087.142	Merkin et al., 2012	SRR594484	Corazón
Corazón 3	Tejido	Corazón	2x40	37.897.106	79.910.394	Merkin et al., 2012	SRR594493	Corazón
Hígado 1	Tejido	Hígado	2x80	103.019.718	212.781.708	Merkin et al., 2012	SRR594477	Hígado
Hígado 2	Tejido	Hígado	2x36	26.021.524	58.088.762	Merkin et al., 2012	SRR594486	Hígado
Hígado 3	Tejido	Hígado	2x36	29.192.793	64.209.790	Merkin et al., 2012	SRR594495	Hígado
Pulmón 1	Tejido	Pulmón	2x80	103.196.444	198.008.329	Merkin et al., 2012	SRR594478	Pulmón
Pulmón 2	Tejido	Pulmón	2x36	22.498.703	51.969.728	Merkin et al., 2012	SRR594487	Pulmón
Pulmón 3	Tejido	Pulmón	2x36	35.817.736	81.251.842	Merkin et al., 2012	SRR594496	Pulmón
Riñón 1	Tejido	Riñón	2x80	115.720.336	229.521.360	Merkin et al., 2012	SRR594476	Riñón
Riñón 2	Tejido	Riñón	2x36	27.567.792	61.990.962	Merkin et al., 2012	SRR594485	Riñón
Riñón 3	Tejido	Riñón	2x36	22.535.945	52.169.075	Merkin et al., 2012	SRR594494	Riñón

Tabla S2 continuación. Secuencias procedentes de experimentos de RNA-seq utilizados en el análisis del splicing alternativo del gen *NCR3*.

Sus scrofa

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Células Madre	Línea	Célula madre	1x100	69.988.504	30.640.738	Xiao et al., 2012	SRR414980	Células Madre
Endometrio	Tejido	Endometrio	1x75	27.690.231	51.003.684	Samborski et al., 2013	SRR651712	Endometrio
Hígado 1	Tejido	Hígado	2x91	19.404.478	33.835.103	No publicado	SRR167669	Hígado
Hígado 2	Tejido	Hígado	2x91	20.066.681	17.777.135	No publicado	SRR167672	Hígado
Tejido graso abdominal 1	Tejido	Hígado	2x91	20.000.000	31.247.038	No publicado	SRR167670	Tejido graso abdominal
Tejido graso abdominal 2	Tejido	Hígado	2x91	20.000.000	31.783.448	No publicado	SRR167673	Tejido graso abdominal
Músculo 1	Tejido	Hígado	2x91	19.510.475	30.943.454	No publicado	SRR167671	Músculo (longissimus dorsi)
Músculo 2	Tejido	Hígado	2x91	19.582.399	32.910.501	No publicado	SRR167674	Músculo (longissimus dorsi)
Células embrionarias germinales	Líneas celular	Embrión	1x35	28.699.514	27.422.964	Petkov et al., 2011	SRR066365	Células embrionarias germinales
Fibroblastos fetales	Líneas celular	Embrión	1x35	26.530.289	26.680.992	Petkov et al., 2011	SRR066366	Fibroblastos fetales

Pan troglodytes

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Cerebro_1	Tejido	Cerebro	1x75	20.083.064	10.216.781	Barbosa-Morais et al., 2012	SRR306818	Cerebro
Cerebro_2	Tejido	Cerebro	2x100	13.947.644	16.976.155	Barbosa-Morais et al., 2012	SRR306817	Cerebro
Cerebro_3	Tejido	Cerebro	2x100	20.408.261	17.448.268	Barbosa-Morais et al., 2012	SRR306816	Cerebro
Cerebro_4	Tejido	Cerebro	2x100	17.394.854	11.870.615	Barbosa-Morais et al., 2012	SRR306815	Cerebro
Cerebro_5	Tejido	Cerebro	2x100	22.234.086	16.577.832	Barbosa-Morais et al., 2012	SRR306814	Cerebro
Cerebro_6	Tejido	Cerebro	2x100	23.317.655	6.728.233	Barbosa-Morais et al., 2012	SRR306813	Cerebro
Cerebro_7	Tejido	Cerebro	1x75	32.043.112	20.342.086	Barbosa-Morais et al., 2012	SRR306812	Cerebro
Cerebro_8	Tejido	Cerebro	1x75	19.384.434	11.393.185	Barbosa-Morais et al., 2012	SRR306811	Cerebro
Corazón_1	Tejido	Corazón	1x75	31.468.011	21.246.561	Barbosa-Morais et al., 2012	SRR306825	Corazón
Corazón_2	Tejido	Corazón	1x75	43.064.259	22.863.564	Barbosa-Morais et al., 2012	SRR306822	Corazón
Hígado_1	Tejido	Hígado	1x75	29.737.439	21.452.678	Barbosa-Morais et al., 2012	SRR306819	Hígado
Hígado_2	Tejido	Hígado	1x75	17.876.248	9.090.165	Barbosa-Morais et al., 2012	SRR306824	Hígado
Células linfoblásticas_1	Tejido-Línea	Células linfoblásticas	1x36	4.356.426	4.435.528	Iskow et al., 2012	SRR504802	Células linfoblásticas
Células linfoblásticas_2	Tejido-Línea	Células linfoblásticas	1x36	8.127.615	8.410.821	Iskow et al., 2012	SRR504801	Células linfoblásticas
Células linfoblásticas_3	Tejido-Línea	Células linfoblásticas	1x36	8.216.282	8.625.163	Iskow et al., 2012	SRR504800	Células linfoblásticas
Células linfoblásticas_4	Tejido-Línea	Células linfoblásticas	1x36	8.076.830	8.511.882	Iskow et al., 2012	SRR504799	Células linfoblásticas
Células linfoblásticas_5	Tejido-Línea	Células linfoblásticas	1x36	8.473.201	8.704.029	Iskow et al., 2012	SRR504798	Células linfoblásticas
Células linfoblásticas_6	Tejido-Línea	Células linfoblásticas	1x36	17.332.962	18.067.450	Iskow et al., 2012	SRR504797	Células linfoblásticas
Riñón_1	Tejido	Riñón	1x75	25.454.775	18.542.015	Barbosa-Morais et al., 2012	SRR306821	Riñón
Riñón_2	Tejido	Riñón	1x75	34.169.060	25.709.032	Barbosa-Morais et al., 2012	SRR306820	Riñón
Testículo_1	Tejido	Testículo	1x75	26.745.671	13.849.948	Barbosa-Morais et al., 2012	SRR306823	Testículo

Gorilla gorilla

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Cerebro_1	Tejido	Cerebro	1x75	35.257.547	23.571.763	Barbosa-Morais et al., 2012	SRR306800	Cerebro
Cerebro_2	Tejido	Cerebro	2x100	16.254.814	24.116.487	Barbosa-Morais et al., 2012	SRR306801	Cerebro
Cerebro_3	Tejido	Cerebro	1x75	28.305.051	18.835.684	Barbosa-Morais et al., 2012	SRR306802	Cerebro
Cerebro_4	Tejido	Cerebro	1x75	20.661.901	13.653.780	Barbosa-Morais et al., 2012	SRR306803	Cerebro
Corazón_1	Tejido	Corazón	1x75	28.286.878	21.774.036	Barbosa-Morais et al., 2012	SRR306804	Corazón
Corazón_2	Tejido	Corazón	1x75	30.588.563	22.474.735	Barbosa-Morais et al., 2012	SRR306805	Corazón
Riñón_1	Tejido	Riñón	1x75	19.804.877	14.877.049	Barbosa-Morais et al., 2012	SRR306806	Riñón
Riñón_2	Tejido	Riñón	1x75	29.684.063	24.571.449	Barbosa-Morais et al., 2012	SRR306807	Riñón
Hígado	Tejido	Hígado	1x75	32.830.718	25.065.167	Barbosa-Morais et al., 2012	SRR306808	Hígado
Mezcla	Tejido	Mezcla de tejidos	2x100	206.526.535	173.134.209	Pipes et al., 2013	SRR832926	Mezcla de tejidos

Tabla S2 continuación. Secuencias procedentes de experimentos de RNA-seq utilizados en el análisis del splicing alternativo del gen *NCR3*.

Pongo pygmaeus

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Cerebro_1	Tejido	Cerebro	1x75	36.457.958	19.283.411	Barbosa-Morais et al., 2012	SRR306791	Cerebro
Cerebro_2	Tejido	Cerebro	2x100	17.675.725	18.693.129	Barbosa-Morais et al., 2012	SRR306792	Cerebro
Cerebro_3	Tejido	Cerebro	1x75	20.807.820	12.057.640	Barbosa-Morais et al., 2012	SRR306793	Cerebro
Corazón_1	Tejido	Corazón	1x75	36.798.263	20.377.300	Barbosa-Morais et al., 2012	SRR306794	Corazón
Corazón_2	Tejido	Corazón	1x75	31.482.282	14.677.549	Barbosa-Morais et al., 2012	SRR306795	Corazón
Riñón_1	Tejido	Riñón	1x75	30.547.227	14.987.922	Barbosa-Morais et al., 2012	SRR306796	Riñón
Riñón_2	Tejido	Riñón	1x75	30.043.284	16.470.773	Barbosa-Morais et al., 2012	SRR306797	Riñón
Hígado_1	Tejido	Hígado	1x75	21.355.541	15.106.269	Barbosa-Morais et al., 2012	SRR306798	Hígado
Hígado_2	Tejido	Hígado	1x75	35.683.453	22.891.629	Barbosa-Morais et al., 2012	SRR306799	Hígado

Papio anubis

Muestra	Supergrupo	Grupo	Tipo Lecturas	Lecturas	Lecturas alineadas	Referencia	SRA_Id	Descripción
Mezcla 1	Tejido	Mezcla de tejidos	1x100	151.524.634	101.784.864	Pipes et al., 2013	SRR832912	Mezcla de tejidos
Mezcla 2	Tejido	Mezcla de tejidos	1x100	67.763.503	60.500.250	Pipes et al., 2013	SRR832907	Mezcla de tejidos
Mezcla 3	Tejido	Mezcla de tejidos	1x100	71.477.607	57.630.056	Pipes et al., 2013	SRR832906	Mezcla de tejidos

Tabla S2 continuación. Secuencias procedentes de experimentos de RNA-seq utilizados en el análisis del splicing alternativo del gen *NCR3*.

	Muestras	Lecturas alineadas totales	Gen NCR3		Uniones de Exones (número de lecturas)																										
			Lecturas	RPKM	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17	H18	H19	H20	P1	P2	P3	P4	P5	P6	P7
Homo sapiens	Hígado 1	28.457.392	27	0,64	2	1	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado 2	17.612.733	4	0,15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado 3	20.019.668	4	0,13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado 4	5.822.277	1	0,12	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado 5	6.017.610	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado 6	5.328.583	1	0,13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado 7	6.064.704	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado 8	7.442.403	1	0,09	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado 9	7.755.211	2	0,17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Músculo	24.302.639	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pulmón	27.155.165	34	0,84	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pulmón fumador 1	14.322.935	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pulmón fumador 2	19.062.028	5	0,18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pulmón fumador 3	16.981.614	4	0,16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pulmón fumador 4	13.128.969	2	0,10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pulmón fumador 5	16.238.322	2	0,08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pulmón fumador 6	16.447.042	4	0,16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pulmón no fumador 1	19.571.068	4	0,14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pulmón no fumador 2	20.255.816	5	0,17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pulmón no fumador 3	17.166.022	1	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pulmón no fumador 4	18.376.019	4	0,15	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Pulmón no fumador 5	18.381.134	3	0,11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Riñón 1	16.399.162	3	0,12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Riñón 2	14.615.495	2	0,09	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Riñón 3	19.274.749	1	0,03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Testículo 1	4.279.539	2	0,31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Testículo 2	22.483.468	10	0,30	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pan troglodytes	Cerebro 1	10.216.781	1	0,081	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 2	16.976.155	4	0,195	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 3	17.448.268	2	0,095	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 4	11.870.615	1	0,070	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 5	16.577.832	3	0,150	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 6	6.728.233	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 7	20.342.086	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 8	11.393.185	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Corazón 1	21.246.561	2	0,078	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Corazón 2	22.863.564	1	0,036	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado 1	21.452.678	10	0,386	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado 2	9.090.165	16	1,458	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Células linfoblásticas 1	4.435.528	1	0,187	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Células linfoblásticas 2	8.410.821	22	2,167	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Células linfoblásticas 3	8.625.163	4	0,384	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Células linfoblásticas 4	8.511.882	25	2,433	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Células linfoblásticas 5	8.704.029	5	0,476	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabla S3. Uniones de exones detectadas mediante RNA-seq. Se indica el número de lecturas que alinean con cada unión de exones correspondiente en cada una de las muestras de RNA-seq analizadas.

	Muestras	Lecturas alineadas totales	Gen NCR3		Uniones de Exones (número de lecturas)																										
			Lecturas	RPKM	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17	H18	H19	H20	P1	P2	P3	P4	P5	P6	P7
Pan t.	Células linfoblásticas 6	18.067.450	35	1,605	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Riñón 1	18.542.015	3	0,134	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Riñón 2	25.709.032	2	0,064	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Testículo 1	13.849.948	6	0,359	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gorilla gorilla	Cerebro 1	23.571.763	1	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 2	24.116.487	3	0,10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 3	18.835.684	1	0,04	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 4	13.653.780	1	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Corazón 1	21.774.036	3	0,11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Corazón 2	22.474.735	1	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado	25.065.167	6	0,20	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Mezcla	173.134.209	3	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Riñón 1	14.877.049	4	0,22	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Riñón 2	24.571.449	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pongo pygmaeus	Cerebro_1	19.283.411	110	4,73	2	0	10	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	12	0	0
	Cerebro_2	18.693.129	48	2,13	0	2	3	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	7	0	0
	Cerebro_3	12.057.640	51	3,50	3	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
	Corazón_1	20.377.300	7	0,28	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Corazón_2	14.677.549	1	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Riñón_1	14.987.922	6	0,33	0	0	3	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Riñón_2	16.470.773	3	0,15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado_1	15.106.269	30	1,65	0	0	5	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado_2	22.891.629	39	1,41	0	0	5	0	6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Papio anubis	Mezcla 1	101.784.864	41	0,33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Mezcla 2	60.500.250	72	0,98	3	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Mezcla 3	57.630.056	48	0,69	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Macaca mulatta	Bazo 1	85.438.428	412	3,99	13	0	0	15	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bazo 2	155.136.020	635	3,38	45	15	9	30	0	0	72	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0
	Bazo 3	82.044.659	610	6,14	12	0	0	20	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 1	10.087.465	0	0,00	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 2	10.939.123	3	0,23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 3	11.510.153	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 4	12.268.379	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 5	11.112.775	1	0,07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 6	68.959.635	5	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 7	181.384.566	33	0,15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cerebro 8	53.237.336	5	0,08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Corazón 1	11.692.168	1	0,07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Corazón 2	9.782.071	1	0,08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Corazón 3	66.738.600	3	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Corazón 4	179.013.629	16	0,07	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Corazón 5	69.365.864	7	0,08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado 1	11.654.151	13	0,92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Hígado 2	19.864.182	14	0,58	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabla S3 continuación. Uniones de exones detectadas mediante RNA-seq. Se indica el número de lecturas que alinean con cada unión de exones correspondiente en cada una de las muestras de RNA-seq analizadas.

Macaca mulatta

Muestras	Lecturas alineadas totales	Gen NCR3		Uniones de Exones (número de lecturas)																									
		Lecturas	RPKM	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17	H18	H19	H20	P1	P2	P3	P4	P5	P6	P7
Hígado 3	55.504.501	24	0,36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hígado 4	190.674.223	30	0,13	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hígado 5	52.435.141	15	0,24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pulmón 1	77.866.802	179	1,90	2	0	0	3	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pulmón 2	189.031.791	251	1,10	24	8	2	8	0	0	34	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0
Pulmón 3	65.433.067	130	1,64	0	0	0	7	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 1	7.033.256	3	0,35	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 2	7.919.122	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 3	61.746.518	6	0,08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 4	177.583.266	7	0,03	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 5	80.892.790	5	0,05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Testículo 1	11.390.840	1	0,07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Testículo 2	20.290.775	2	0,08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Células linfoblásticas 1	7.722.459	7	0,75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Células linfoblásticas 2	9.882.573	46	3,85	4	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Células linfoblásticas 3	10.490.643	23	1,81	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Células linfoblásticas 4	14.446.788	35	2,00	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Células linfoblásticas 5	17.789.985	41	1,90	0	0	1	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Células linfoblásticas 6	19.582.478	58	2,45	0	0	0	4	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Rattus norvegicus

Muestras	Lecturas alineadas	Gen NCR3		Uniones de Exones (número de lecturas)																									
		Lecturas	RPKM	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17	H18	H19	H20	R1	R2	R3	R4	R5	R6
Bazo 1	255.471.881	418	1,75	0	0	16	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bazo 2	235.904.870	721	3,28	10	0	134	82	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bazo 3	75.462.266	114	1,62	0	0	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cerebro 1	221.611.967	11	0,05	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cerebro 2	164.999.044	10	0,06	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cerebro 3	68.132.919	3	0,05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corazón 1	87.856.189	18	0,22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corazón 2	137.464.524	14	0,11	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Corazón 3	60.476.295	23	0,41	2	0	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hígado 1	62.924.829	34	0,58	0	0	5	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hígado 2	271.766.412	113	0,45	5	0	24	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Hígado 3	98.994.403	42	0,45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pulmón 1	261.218.964	163	0,67	0	0	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pulmón 2	163.703.525	146	0,96	0	0	9	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Pulmón 3	46.211.300	38	0,88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 1	258.393.051	15	0,06	0	0	3	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 2	241.346.373	62	0,28	2	0	9	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 3	90.431.863	19	0,23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabla S3 continuación. Uniones de exones detectadas mediante RNA-seq. Se indica el número de lecturas que alinean con cada unión de exones correspondiente en cada una de las muestras de RNA-seq analizadas.

Mus musculus

Muestras	Lecturas alineadas	Gen NCR3		Uniones de Exones (número de lecturas)																			
		Lecturas	RPKM	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17	H18	H19	H20
Bazo 1	255.432.575	64	0,39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bazo 2	214.692.757	161	1,16	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bazo 3	250.239.575	191	1,18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cerebro 1	53.064.101	10	0,29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cerebro 2	9.550.600	4	0,65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cerebro 3	24.268.078	5	0,32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cerebro 4	177.016.239	17	0,15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cerebro 5	10.693.287	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cerebro 6	17.900.601	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cerebro 7	200.364.582	37	0,29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cerebro 8	66.218.941	25	0,58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cerebro 9	19.201.472	4	0,32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corazón 1	25.008.946	1	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corazón 2	16.469.930	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corazón 3	18.154.201	5	0,43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corazón 4	78.095.213	7	0,14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corazón 5	33.571.300	2	0,09	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hígado 1	22.309.280	6	0,42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hígado 2	299.833.542	1	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hígado 3	17.069.549	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hígado 4	208.430.656	5	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hígado 5	29.932.099	2	0,10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hígado 6	104.527.412	3	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pulmón 1	74.914.430	11	0,23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pulmón 2	100.665.712	18	0,28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pulmón 3	232.907.600	90	0,60	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 1	15.921.231	1	0,10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 2	15.974.541	1	0,10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 3	18.200.720	11	0,93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 4	249.873.082	13	0,08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 5	231.393.521	16	0,11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Riñón 6	79.881.989	2	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Testículo 1	19.245.597	1	0,08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabla S3 continuación. Uniones de exones detectadas mediante RNA-seq. Se indica el número de lecturas que alinean con cada unión de exones correspondiente en cada una de las muestras de RNA-seq analizadas.

Bos taurus

Muestras	Lecturas alineadas	Gen NCR3		Uniones de Exones (número de lecturas)																																
		Lecturas	RPKM	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17	H18	H19	H20	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13
Bazo 1	213.812.023	1064	7,03	155	0	97	181	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	114	0	0	0	0	17	3	0
Bazo 2	56.041.684	215	5,42	14	0	10	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
Bazo 3	74.034.095	312	5,95	36	0	11	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0
Cerebro 1	198.374.899	293	2,09	24	0	13	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	7	0	9	4	36	2	0
Cerebro 2	59.970.662	22	0,52	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0
Cerebro 3	67.381.914	47	0,99	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Corazón 1	224.113.913	74	0,47	20	0	4	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	4	0	13	0	0
Corazón 2	44.087.142	35	1,12	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Corazón 3	79.910.394	51	0,90	2	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hígado 1	212.781.708	226	1,50	42	0	13	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	5	4	7	1	3
Hígado 2	58.088.762	54	1,31	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0
Hígado 3	64.209.790	37	0,81	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
Pulmón 1	198.008.329	441	3,15	88	0	34	87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	77	0	2	13	2	17	0	0
Pulmón 2	51.969.728	113	3,07	14	0	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
Pulmón 3	81.251.842	274	4,76	19	0	12	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0
Riñón 1	229.521.360	206	1,27	30	0	11	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0	4	0	0	0
Riñón 2	61.990.962	89	2,03	6	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
Riñón 3	52.169.075	42	1,14	4	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0

Sus scrofa

Muestras	Lecturas alineadas	Gen NCR3		Uniones de Exones (número de lecturas)																										
		Lecturas	RPKM	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17	H18	H19	H20	S1	S2	S3	S4	S5	S6	S7	S8
Células Madre	30.640.738	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Endometrio	51.003.684	2	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Hígado 1	33.835.103	1	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hígado 2	17.777.135	8	0,66	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tejido graso abdominal 1	31.247.038	4	0,19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tejido graso abdominal 2	31.783.448	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Músculo 1	30.943.454	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Músculo 2	32.910.501	3	0,13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Células embrionarias germinales	27.422.964	2	0,11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fibroblastos fetales	26.680.992	0	0,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabla S3 continuación. Uniones de exones detectadas mediante RNA-seq. Se indica el número de lecturas que alinean con cada unión de exones correspondiente en cada una de las muestras de RNA-seq analizadas.

Especie	Nombre común	Bases de datos		Exón 4I	Exones 4II y 4III		
		NCBI	Ensembl	Codón de parada	Presencia	Secuencia de poliadenilación	Expresión
Ailuropoda melanoleuca	Panda	-	P	53	-	-	ND
Bos taurus	Vaca	S	S	47	-	-	ND
Canis lupus familiaris	Perro	P	P	47	-	-	ND
Cebus apella	Mono capuchino	-	-	35	+	ND	-
Ceratotherium simum simum	Rinoceronte	P	-	47	-	-	ND
Colobus guereza	Colobo	-	-	47	+	ND	-
Echinops telfairi	Erizo de madagascar	-	P	35	-	-	ND
Equus caballus	Caballo	P	P	47	-	-	ND
Erinaceus europaeus	Erizo	-	P	47	-	-	ND
Felis catus	Gato	P	P	47	-	-	ND
Gorilla gorilla	Gorilla	P	P	77	+	+	+
Homo sapiens	Humano	S	S	77	+	+	+
Loxodonta africana	Elefante	-	P	77	-	-	ND
Macaca fascicularis	Mono cangrejero	-	-	35	+	-	-
Macaca mullata	Mono rhesus	S	S	35	+	-	-
Mesocricetus auratus	Hámster dorado	P	-	53	-	-	ND
Microcebus murinus	Lemur ratón	-	P	47	-	-	ND
Microtus ochrogaster	Topillo	P	-	53	-	-	ND
Mus musculus	Ratón	P	P	53	-	ND	ND
Mustela putorius furo	Hurón	-	P	53	-	-	ND
Nomascus leucogenys	Gibón	P	P	77	+	+	ND
Ochotona princeps	Pica americano	P	P	53	-	-	ND
Octodon degus	Degú	P	-	80	-	-	ND
Odobenus rosmarus divergens	Morsa	P	-	53	-	-	ND
Orcinus orca	Orca	P	-	47	-	-	ND
Oryctolagus cuniculus	Conejo	-	S	53	-	-	ND
Otolemur garnettii	Gálago	P	P	47	-	-	ND
Ovis aries	Oveja	P	-	47	-	-	ND
Pan paniscus	Bonobo	P	-	77	+	+	ND
Pan troglodytes	Chimpacé	S	S	77	+	+	+
Papio anubis	Papión	P	-	35	+	-	ND
Papio cynocephalus	Papión	-	-	47	+	ND	-
Pongo abelli	Orangután de Sumatra	P	P	47	+	+	ND
Pongo pygmaeus	Orangután de Borneo	-	-	47	+	ND	+
Procavia capensis	Damán	-	P	71	-	-	ND
Rattus norvegicus	Rata	S	S	59	-	ND	ND
Saimiri boliviensis	Mono ardilla	P	-	47	+	-	ND
Sus scrofa	Cerdo	P	P	47	-	-	ND
Tarsius syrichta	Tarsero filipino	-	P	41	-	-	ND
Trichechus manatus latirostris	Manatí	P	-	54	-	-	ND
Tursiops truncatus	Delfín	P	P	47	-	-	ND

Tabla S4. Análisis del origen evolutivo de los exones finales del gen NCR3. En cada de una de las especies se indica su disponibilidad en las bases de datos, especificando si la anotación es una predicción (P) o está basada en secuencias de mensajero depositadas (S). Se incluye la posición del codón de parada del exón 4I en todas estas especies, referido al inicio del exón. Además, se muestra el resultado del análisis de la presencia de los exones 4II/4III y del sitio consenso de poliadenilación común para éstos. En las especies incluidas en este trabajo se incluye además la información de la expresión de variantes codificantes que porten estos exones. Los símbolos '+' representa presencia o expresión, los símbolos '-' denotan ausencia o falta de expresión y con 'ND' se indica ausencia de información.

Artículos publicados

Rastrojo *et al. BMC Genomics* 2013, **14**:223
http://www.biomedcentral.com/1471-2164/14/223

BMC Genomics

RESEARCH ARTICLE Open Access

The transcriptome of *Leishmania major* in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq

Alberto Rastrojo[1], Fernando Carrasco-Ramiro[1], Diana Martín[1], Antonio Crespillo[1], Rosa M Reguera[2], Begoña Aguado[1*] and Jose M Requena[1*]

Abstract

Background: Although the genome sequence of the protozoan parasite *Leishmania major* was determined several years ago, the knowledge of its transcriptome was incomplete, both regarding the real number of genes and their primary structure.

Results: Here, we describe the first comprehensive transcriptome analysis of a parasite from the genus *Leishmania*. Using high-throughput RNA sequencing (RNA-seq), a total of 10285 transcripts were identified, of which 1884 were considered novel, as they did not match previously annotated genes. In addition, our data indicate that current annotations should be modified for many of the genes. The detailed analysis of the transcript processing sites revealed extensive heterogeneity in the spliced leader (SL) and polyadenylation addition sites. As a result, around 50% of the genes presented multiple transcripts differing in the length of the UTRs, sometimes in the order of hundreds of nucleotides. This transcript heterogeneity could provide an additional source for regulation as the different sizes of UTRs could modify RNA stability and/or influence the efficiency of RNA translation. In addition, for the first time for the *Leishmania major* promastigote stage, we are providing relative expression transcript levels.

Conclusions: This study provides a concise view of the global transcriptome of the *L. major* promastigote stage, providing the basis for future comparative analysis with other development stages or other *Leishmania* species.

Keywords: Gene Expression, RNA-seq, Transcript annotation, mRNAs, *Leishmania*, Trypanosomatids

Background

Species of the genus *Leishmania* are protozoan parasites and aetiological agents of a spectrum of clinical diseases, known as leishmaniases, ranging from disfiguring skin lesions to life-threatening visceral infection. The World Health Organization (WHO) estimates that 350 million people worldwide are at risk of infection, and this disease is considered a major public health problem. Two million new cases of leishmaniasis (1.5 million for cutaneous forms and 500 000 for visceral leishmaniasis) occur annually [1]. The genus *Leishmania* belongs to the order Trypanosomatida [2], which also includes, among others, *Trypanosoma brucei* and *Trypanosoma cruzi*, causative agents of two other important human infectious diseases: sleeping sickness and Chagas disease, respectively. The evolutionary origin of these organisms is found in the deepest roots of the eukaryotic tree [3], and are characterized by markedly original molecular features.

In 1999, the complete sequence of chromosome 1 of *Leishmania major* was published and showed a remarkable feature of the gene organization in *Leishmania*, i.e. genes are grouped in large clusters sharing the same transcriptional direction. Thus, from the left end of chromosome 1, the first 29 genes are all located on the same DNA strand, whereas the remaining 50 genes are located on the other strand [4]. When transcriptional activity was examined by nuclear run-on analyses using single-stranded DNA probes, the protein-coding strand was found to be more strongly transcribed than the non-coding strand in the majority of the chromosome 1

* Correspondence: baguado@cbm.uam.es; jmrequena@cbm.uam.es
[1]Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, 28049 Madrid, Spain
Full list of author information is available at the end of the article

genes [5]. Furthermore, it was found that the RNA polymerase initiates transcription within the strand-switch region of chromosome 1. Similarly, in chromosome 3, which contains two convergent clusters of 67 and 30 genes, nuclear run-on analyses indicated that transcription initiates upstream of the most-5' gene of the two long polycistronic clusters [6]. After whole genome sequences for *Leishmania* and other trypanosomatids (i.e. *T. brucei* and *T. cruzi*) were completed, it was confirmed that in these organisms most genes are organized into large clusters on the same DNA strand.

Another remarkable molecular feature found in trypanosomatids is that transcription initiation by RNA polymerase II (RNAP II) is not regulated on a per gene basis; instead, most genes are transcribed polycistronically. Genome-wide chromatin immunoprecipitation analysis of *L. major* promastigotes showed acetylated histone H3 peaks at the 5' ends of all polycistronic protein-coding gene clusters, indicating that global regulation of transcription initiation may be achieved by epigenetic regulation of H3 acetylation at the origins of polycistronic transcription units [7]. In a recent publication, the J base (a modification of thymine, which is introduced with some frequency in the DNA of trypanosomatids) was shown to define the RNAP II transcription termination sites in *L. major* and *L. tarentolae* [8].

In contrast to operons in bacteria, polycistronic units in trypanosomatids require processing before translation, and the mature mRNAs are processed from primary transcripts by coupled *trans*-splicing and polyadenylation [9]. During *trans*-splicing, a conserved spliced leader RNA (SL RNA or mini-exon) is added to the 5' end of all mRNAs, providing the cap structure for translation. The differential expression of mature mRNAs from a single polycistronic unit is thought to be achieved by post-transcriptional control, i.e. mRNA levels are regulated by RNA stability and/or differential translation [10-12].

In 2005, the sequence for the 36 chromosomes of the *L. major* genome (32.8 Mb) was published, and provided a framework for future comparative genomic studies [13]. Using bioinformatic analyses, 911 RNA genes, 39 pseudogenes, and 8272 protein-coding genes were predicted. Within the latter group, only 36% can be assigned to a putative function based on sequence conservation with protein characterized in other eukaryotic organisms. Most *L. major* genes have orthologs in the *T. brucei* and *T. cruzi* genomes [14]. However, more than 60% of the predicted genes remain annotated as hypothetical. A major challenge lies ahead to discover whether or not these genes are expressed at any moment in the life cycle and, therefore, may be catalogued as functional genes. On the other hand, both known and putative genes lack annotated 5' and 3' untranslated regions (UTRs), and for only a few genes these regions have been experimentally determined [15]. In *Leishmania*, and related trypanosomatids, these flanking regions (largely the 3'-UTRs) have been involved in regulating the steady-state level and translational status of specific mRNAs along the cell cycle and in the different life cycle stages [10-12].

Recent advances in sequencing technologies, known as deep sequencing or next-generation sequencing (NGS), are becoming invaluable tools, among others, for reconstructing of the entire transcriptome of a given organism [16,17]. In this study, we employed the power of NGS on RNA analysis (RNA-seq) to provide a comprehensive characterization of the poly-A transcriptome for the promastigote stage of *Leishmania major*. A total of 10285 transcripts were identified, of which 1884 did not match with previously annotated genes and therefore were categorized as novel genes. In addition, the RNA-seq analysis generated valuable information on both the relative abundance of the RNAs and the structures of their corresponding genes (i.e. ORFs, and 5'- and 3'-UTRs).

Methods

Leishmania culture and RNA isolation

Promastigotes of *L. major* Friedlin strain (MHOM/IL/80/Friedlin; clone V1) were cultured at 26°C in RPMI medium supplemented with 10% fetal bovine serum, 100 U/ml penicillin G and 0.1 mg/mL streptomycin sulphate. Promastigotes were grown to mid log phase by seeding cultures at 1×10^6 cells/mL, and collected for RNA isolation when the culture density reached 6.1×10^6 cells/mL (mid-logarithmic phase of growth). Total RNA was isolated using the Aurum™ Total RNA Mini Kit (Biorad), and treated with RNAse-free DNAse I. RNA samples were quantified by absorbance at 260 nm using the Nanodrop ND-1000 (Thermo Scientific), all samples showed an A_{260}/A_{280} ratio higher than 2.0. In addition, RNA integrity was checked in a bioanalyzer (Agilent 2100).

RNA-seq and data processing

RNA-seq was performed at the Massive Sequencing Platform of Cantoblanco (CSIC-PCM, Madrid, Spain). Standard libraries for massive sequencing were generated using the TruSeq RNA Sample Prep Kit (Illumina). Briefly, poly-A^+ RNA was selected by oligo-dT chromatography, and RNA fragmentation was achieved using divalent cations under elevated temperature. Afterwards, these fragments were used to generate a cDNA library, and cDNA fragments corresponding in size to about 300-400 bp were selected on an agarose gel. Two cDNA libraries were constructed: first strand synthesis of one of them was initiated with only random hexadeoxynucleotide primers (Illumina standard protocol); however, for the first strand synthesis of the second library, we introduced as an additional component the 5'-T_{15}VN-3' oligonucleotide together with the random hexamer primers present in the

Rastrojo et al. BMC Genomics 2013, 14:223
http://www.biomedcentral.com/1471-2164/14/223

kit. Afterwards, the second strand of the cDNA was synthesized. The cDNA ends were repaired and adenylated, subsequently adapters were added at both ends. Finally, the library was enriched in ligated fragments by limited PCR amplification. Sequencing was carried out in a GAIIx Illumina system. Each library was sequenced in two separated lines. Single reads of 75 nucleotides were obtained, and raw reads were subject to quality-filtered using the standard Illumina process and analyzed using FASTQC tool [18]. Reads were mapped to the last assembled version of *L. major* genome, obtained from the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/pathogens/Leishmania/major/V6_211210/), using Bowtie [19]. In the alignment of reads, a maximal of three mismatches was allowed within the whole read (aligner V mode). Nevertheless, in order to select the best alignment in terms of number of mismatches, the option "—best" was used. Also, the option "-k1" was elected, i.e. if in the course of the search Bowtie found 2 (or more) possible alignments for a given read, the program selected one of the alignments at random. We analyzed different alignment conditions in terms of multi-hits in order to obtain the best and accurate results from our data. Allowing up to 10 multi-hits for a single read, the main differences with the transcripts assembled with no multi-hits restriction were found at gene-tandem repeat regions. In those regions the assembled transcripts were reduced to the UTRs, losing the coding regions. Therefore, no restriction in the number of multi-hits was introduced, except for SL-containing reads, in which reads mapping to more than 10 sites were excluded for further analysis. Finally, mapped reads were assembled into transcripts using Cufflinks [20].

Identification of *trans*-splicing and polyadenylation sites

Among the non-aligned reads, a search for reads containing 8 (or more) nucleotides identical to the 3'-end of the SL sequence (AACTAACGCT ATATAAGTAT CAGTTTCTGT ACTTTATTG) was performed using a custom Perl script. No mismatches were allowed. Afterwards, the SL-matching nucleotides were stripped from the reads and the remaining sequence was used to map the position of the *trans*-splicing site in the reference genome. Similarly, reads spanning potential polyadenylation sites were extracted from the non-aligned sequences by an in house Perl script, which finds reads with A-runs (higher than 5 nucleotides in length) located at an end of the read sequence. These reads were mapped back to the reference genome.

Additional sequencing analysis tools

Samtools software [21] was used to interconvert alignment formats, and to assign the annotated genes to transcripts generated from Seqdata, a local version of Blastx program [22]. The IGV browser was used [23] for visualization of mapped reads and assembling of transcripts to its genome context. Consensus sequences were analyzed using a local version of WebLogo tool [24]. BLAST searches for sequence homologies were performed in the following databases: GeneDB [25], TritrypDB [26] and GenBank at the NCBI [27].

Results and discussion

RNA-seq data and delineation of transcripts

RNA isolated from an axenic culture of *L. major* promastigotes (Friedlin strain, clone V1) was sequenced, after poly(A) + selection, on an Illumina GAII platform generating a total of 14 656 121 sequence reads (75-nt long). RNA-seq data from this study have been submitted to the EBI-ENA Sequence Read Archive (SRA) under accession number ERP002077. Allowing up to three mismatches, 14 027 356 reads (95.71%) were aligned to the reference *L. major* Friedlin genome [13]. After initial assembling, it was possible to define a total of 6937 transcripts; a number lower than the 8272 protein-coding genes previously predicted to exist in the *L. major* genome [13]. However, as shown in Figure 1, this difference was not derived from a low coverage of RNA-seq data. Instead, the transcript assembly indicated that most of the *Leishmania* genome seems to be transcribed, and many assembled transcripts contain two or more annotated coding-genes (Figure 1). In fact, the genome coverage of the RNA-seq reads generated in this study was around 90.75%, even though reads for tRNAs, SL-RNAs and other small RNAs were not obtained. Several possibilities may be envisioned to accommodate this observation. First, existence of stable polycistronic transcripts; however, to date there are not descriptions of mature polycistronic transcripts in *Leishmania*. Nevertheless, the existence of a functional bicistronic transcript has been demonstrated in *T. cruzi* [28]. A second possibility is that some RNA processing intermediates with larger half-life may be represented in the RNA-seq reads. This hypothesis is very plausible as there are many reports describing processing intermediates that are clearly detected by Northern blot analysis. For example, at least 10 stable cytoplasmic poly(A) + RNAs, ranging in size from 1.7 to 13 kb and related to the 3.2-kb DHFR-TS mRNA have been observed in antifolate-resistant *Leishmania* promastigotes [29]. In other studies, polycistronic intermediates were demonstrated using a combination of genic and intergenic probes [30]. Third, antisense transcription might be contributing to create polycistronic transcript, since RNA-Seq data were derived from non-oriented, unidirectional sequencing of RNA molecules. There are several reports describing the existence of antisense transcription in *Leishmania*. For example Monnerat and co-workers [31], analyzing the transcriptional activity of a 30-Kb region from *L. major* chromosome 27, found that while the non-coding strand

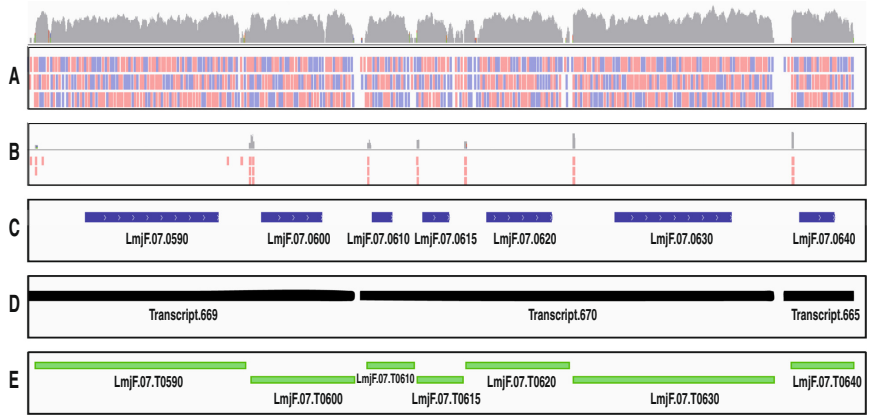


Figure 1 Transcript assembling and annotation from RNASeq data. The figure shows a region of the chromosome 7. Panel **A**: reads aligned in this region; a small window of the total mapped reads is shown (bottom panel); the relative coverage (logarithmic scale) is depicted as a sky-line on the panel. Each vertical dash represents a read. Reads aligned with the plus strand of the chromosome are shown in pink and those aligned with the minus strand in violet (note that the direction of the reads was assigned arbitrarily, as sequencing was not oriented). Panel **B**: mapping of SL-containing reads. Panel **C**: previously annotated *L. major* genome (GeneDB database). Panel **D**: crude transcripts as assembled by Cufflinks. Panel **E**: new transcript annotation after mapping of both SL addition sites and the 3' ends generated by polyadenylation. The images were generated after loading the RNA-seq data in the Integrative Genomics Viewer (IGV 2.1) [23].

generally appears to be transcribed at levels close to background, several regions appeared to be transcribed at significant levels, albeit still substantially lower than the coding strand. A fourth possibility, a background derived from sequencing of contaminating DNA, may be discarded, since there are many intergenic regions from which there were no reads (see gaps without reads on Figure 1A). However, if DNA contamination were present in the RNA samples, reads should be mapped to all chromosomal locations.

In order to further delineate *Leishmania* transcripts, we took advantage of the expected addition of the 39-nucleotide long mini-exon sequence at the 5'-end of all *Leishmania* mRNAs [32,33]. Thus, we searched among the non-aligned reads (628 765; 4.29% of total reads) for sequences containing at the 5'-end eight (or more) nucleotides identical to the 3'-end of the mini-exon sequence. A total of 188 398 sequence reads met these criteria. After trimming the mini-exon sequences, these reads were aligned to the *L. major* genome (Figure 1B) and, as a result, 22 592 different mini-exon addition sites were defined.

Interestingly, only 44, of the 188 398 reads containing SL sequences, were mapped in antisense orientation (related to the coding strand), suggesting that *trans*-splicing occurs almost exclusively in sense transcripts and that antisense transcripts (if produced at meaningful levels) should not be processed by the addition of mini-exon sequences. In a recent published work [8], the authors describe the role that base J plays in termination of RNAP II transcription in *L. tarentolae*, mentioning that the vast majority of SL-containing reads were restricted to the coding strand. Proper transcription termination and avoidance of readthrough of transcriptional stops seemed to be vital for *Leishmania* [8].

As illustrated in Figure 1, most of the SL-containing reads mapped at expected locations, i.e. upstream of annotated genes and a significant number of reads were found for each putative splicing acceptor site (considering both main and alternative sites). However, exceptions for this rule were also found. Thus, from time to time, single reads containing SL sequences were mapped at unexpected positions, such as coding sequences or 3'-UTRs. Furthermore, the position of those reads was not accompanied by a breakdown in the reads density as occurs for the rest of SL addition sites. A plausible interpretation for these findings is that the *trans*-splicing machinery generates a low, but detectable number of events in which the mini-exon is misplaced. Keeping in mind this idea, we excluded in the transcript defining process those mini-exon addition sites that were defined by a sole read and located at unexpected positions.

Finally, using as criterion for defining the 5' end of a transcript the location of a SL addition site, most of the

polycistronic transcripts obtained after the initial assembling could be split up, giving a total number of 10 285 transcripts (Table 1). Only 73 of these transcripts remained as polycistronic, with 72 bicistronics and one tetracistronic (*LmjF.30.T1460-1470-1480-1490*). It would be interesting to analyze whether these bicistronic transcripts really exist or they are only evidencing current annotation deficiencies in the *L. major* database. A detailed list of the *L. major* transcriptome is provided as an Additional file 1. Transcripts were named using the systematic identifiers for the annotated genes [13], and were interdigitated numbers to name the new transcripts. In order to distinguish between transcript and genes, a T preceding the transcript number was included. By way of example, Figure 1 (panel C) shows the previously annotated genes existing in a region of chromosome 7, and in panel E are shown the new transcripts (and their names) mapped at that chromosomal region.

Table 1 Transcriptome of *Leishmania major* promastigotes

Chromosome	Number of transcripts	Bicistronic transcripts	Non-annotated genes	Mis-annotated genes (*)
1	92		7	1
2	93	2	19	9 (1)
3	110		13	9
4	140	1	11	14 (1)
5	146	1	22	4 (2)
6	154		19	7 (3)
7	158		28	6 (7)
8	171	1	36	3 (1)
9	189	1	21	6 (3)
10	175	4	32	5 (4)
11	171	1	34	1 (2)
12	183		41	12 (1)
13	207	1	38	7 (5)
14	194		35	5 (2)
15	196	2	28	11 (2)
16	213	1	37	8 (3)
17	211	1	52	5
18	242		70	5 (2)
19	216		38	4 (1)
20	215	5	40	9
21	264		37	11 (2)
22	214		45	7 (1)
23	253	3	53	8 (3)
24	286	1	48	14 (1)
25	302	4	50	22 (1)
26	330		51	10 (2)
27	340	2	59	10 (1)
28	403	1	84	13 (2)
29	372	3	77	10
30	465	2[a]	69	11 (3)
31	463	5	126	51 (12)
32	509	6	85	25 (4)
33	492	5	115	13 (4)
34	570	6	88	25 (1)
35	673	5	133	9 (10)
36	873	9	143	40 (7)
Genome	**10285**	**73**	**1884**	**410 (94)**

*In brackets is indicated the number of genes that might be truncated by addition of SL in secondary trans-splicing sites.
[a]Transcript LmjF.30.T1460-1470-1480-1490 is tetracistronic.

With the sole exception of genes *LmjF.02.0400*, *LmjF.09.0690*, *LmjF27.0280*, *LmjF33.1760*, *LmjF35.2600* and *LmjF35.2610*, transcripts were found for all the currently annotated genes at GeneDB database [25]. These six genes code for hypotethical proteins, but, at least, gene *LmjF35.2610* seems to be encoding a protein since the predicted amino acid sequence contains a region with similarity to ubiquitin and also an AT hook, DNA-binding motif; furthermore, the gene is present in other *Leishmania* species [26]. Thus, the lack of expression of these genes, and in particular of *LmjF35.2610*, in *L. major* promastigotes is a finding that would merit further studies.

Interestingly, 1884 new transcripts were found spanning genomic regions lacking annotated genes; hence, they were categorized as non-annotated genes (Table 1). These findings suggest that the gene content of *L. major* would be approximately 20% higher than previously believed [13]. Similar results have been reported after determining the *T. brucei* transcriptome by RNA-seq [34]. Nevertheless, it is likely that many of these new transcripts may have roles other than protein-coding function; some may even be merely processing products resulting from the unusual polycistronic gene organization and processing of the *Leishmania* genome. In this regard, non-coding transcripts, derived from intercoding regions of *T. brucei* VSG genes, were found to be *trans*-spliced, polyadenylated and present in polyribosomes [35]. Therefore, the new transcripts described in this work might be considered non-coding (nc) RNAs until shown to be otherwise.

Concerning the 5'-end mapping, we have shown that 410 annotated genes are mis-predicted, regarding the translation start codons, as splice acceptor sites were found exclusively downstream of the previously assigned ATG. An example for a clear mis-annotation is shown in Figure 2. Thus, three SL addition sites were found to exist in the middle of the ORF currently annotated as *LmjF.04.0860*; however, no SL addition sites were found at the 5' end of the annotated gene and no reads were mapped at the region coding for the N-terminal moiety of LmjF.04.0860. Translation from the nearest ATG codon found after the main SL addition site gives a protein corresponding to the last 240 amino acids of the annotated

LmjF.04.0860 protein (Figure 2B). Interestingly, the new protein is similar in size and sequence to that encoded by the gene *Tb927.9.8290*, which has an authentic annotation in the *T. brucei* databases (Figure 2C). In the GeneDB database this hypothetical protein is categorized as conserved, since it is also encoded in the genomes of *T. cruzi* and other *Leishmania* species. As a structural feature, the protein contains the domain SSF55129, which is typical of the ribosomal protein L30p/L7e superfamily. In summary, these data support the conclusion that mature mRNAs containing the LmjF.04.0860 ORF, as annotated in the GeneDB database, do not exist.

On the other hand, for 94 annotated genes, alternative splice addition sites were mapped into the ORF, suggesting that different proteins might be generated from a single gene. In this regard, there is a documented case of alternative *trans*-splicing in the *T. cruzi* LYT1 gene, in which the different maturation of the mRNA leads to the expression of protein isoforms showing different compartmental and functional properties [36]. Overall, our transcriptomic study has uncovered that the current annotation of the *L. major* genome had clear limitations that are corrected by the data reported in this work.

Determination of RNA levels from RNA-seq data

RNA-seq is an accurate method for quantifying transcript levels. The strength of this method is that it produces digital counts of transcript abundance, in contrast to the analog-style signals obtained from fluorescent dye–based microarrays. This technique has been validated by several studies and found to be highly reproducible, with very little technical variability and can measure mRNA levels over several orders of magnitude [37,38]. A useful parameter is FPKM (fragments – or reads- per kilobase of transcript per million mapped reads), which reflects the abundance of a transcript in the sample by normalizing for RNA length and for the total read number in the measurement [39]. Thus, the presence and abundance of a given RNA can be calculated and subsequently compared with the amount in any other sequenced sample, now or in the future.

Table 2 contains a list with the 50 most abundant transcripts detected in promastigotes of *L. major*. Two of the top three on the list are transcripts corresponding to the heat shock protein 70 (HSP70); this finding is not unexpected taking into account that this protein make up 2.1% of the total protein in unstressed *Leishmania* promastigotes [40]. Additionally, the fact that 20 out of 50 correspond to transcripts encoding ribosomal proteins might indicate that a direct correlation between transcript levels and protein abundance would be a general rule in *Leishmania*. Another conspicuous observation is that 18 out of the 50 most abundant transcripts derive from genes located on chromosome 35. However, at first glance, these genes do not seem to be concentrated at specific regions of the chromosome; rather they seem to be randomly distributed. Other abundant transcripts are those encoding nucleoside transporters, histone H4, peptidases, cyclophilin, LACK and tubulins (Table 2). Two abundant transcripts, *LmjF.35.T2220* and *LmjF.35.T2210*, encode KMP-11, a protein found tightly associated with lipophosphoglycan, the major cell surface glycoconjugate of *Leishmania* promastigotes [41]. In addition, there are other transcripts encoding for hypothetical proteins that are expected to be abundant ones. Thus, transcripts *LmjF.31.T0900* (the fifth on the list) would be coding for a hypothetical, small protein (79 amino acids), annotated as LmjF31.0900, which is also present in the genomes of related trypanosomatids. Interestingly, this transcript was identified in previous studies to be both abundant and differentially expressed in promastigotes by using oligonucleotide microarrays [42], and one of the most abundant transcripts in metacyclic *L. major* grown in culture by using SAGE methodologies [43]. Another abundant transcript, *LmjF.31.T0966*, located at a region lacking annotated genes contains high sequence identity with gene *LmjF31.0900*. The structural relationship between both transcripts (*LmjF.31.T0900* and *LmjF.31.T0966*) and the functional role of the encoded protein (LmjF31.0900) are two aspects that merit further studies. Transcript *LmjF.36.T3620*, containing the annotated *LmjF36.3620* gene and coding for a hypothetical protein, was also found among the most abundant transcripts in *L. major* promastigotes [42] and metacyclics forms [43].

On the other hand, five transcripts (*LmjF.19.T0983*, *LmjF.20.T1285*, *LmjF.31.T0964*, *LmjF.31.T0895*, and *LmjF.35.T4191*), among the most abundant in *L. major* promastigotes (Table 2), do not contain previously annotated genes. The sequence of transcript *LmjF.19.T0983* was found to be conserved in the genomes of different *Leishmania* species (*L. braziliensis, L. donovani, L. mexicana* and *L. infantum*) but conserved sequences were not detected in the genomes of related trypanosomatids (i.e. *T. brucei* and *T. cruzi*). Interestingly, a cDNA (named DRS-2) derived from this transcript was previously described in *L. major as* an mRNA whose expression increases during metacyclogenesis [15]. Similarly, sequences homologous to transcript *LmjF.31.T0964* were found in the genomes of all *Leishmania* species sequenced to date, but absent in the genus *Trypanosoma*. The sequences of transcripts *LmjF.20.T1285*, *LmjF.31.T0895* and *LmjF.35.T4191* were found to be well conserved in the genomes of *L. donovani*, *L. mexicana* and *L. infantum*, but seemed to be absent from the *L. braziliensis* genome. It is clear that a challenge for the future will be to understand the nature (coding or not) of these transcripts, and certainly for the additional 1879 new transcripts that have been described in this work (Table 1).

Tandemly repeated, multi-copy gene loci are frequent in the *Leishmania* genome [11]. In fact, the list shown in

Rastrojo *et al. BMC Genomics* 2013, **14**:223
http://www.biomedcentral.com/1471-2164/14/223

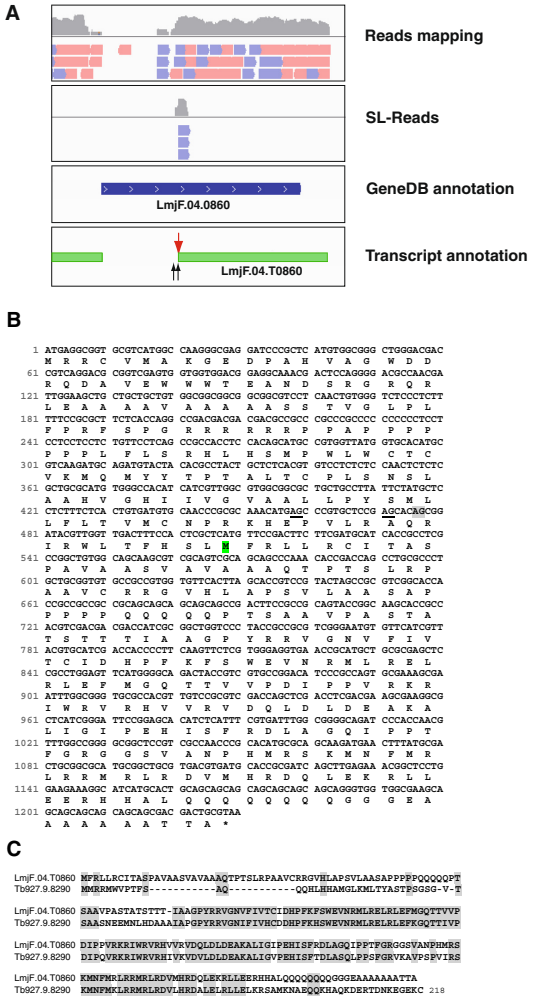


Figure 2 Mis-annotation of *LmjF04.0860* gene. (**A**) Upper panels show the mapping of RNA-seq reads (either total or SL-containing reads) in the genomic region containing the annotated *LmjF04.0860* gene; the bottom panel contains the transcripts delimited in this region. Arrows indicate SL addition sites, the red arrow points at the main SL addition site. Reads aligned with the plus strand of the chromosome are shown in violet and those aligned with the minus strand in pink (note that the direction of the reads was a consequence of sequencing process, as sequencing was not oriented). (**B**) Nucleotide sequence (and predicted amino acid sequence) of the *LmjF04.0860* gene as annotated in the GeneDB database [25]. Shaded in gray is shown the position of the main AG dinucleotide used for *trans*-splicing in the *LmjF.04.T0860* transcripts, and underlined are those AG dinucleotide representing alternative SL addition sites. Shaded in green it is shown the first ATG found downstream the SL addition sites. (**C**) Alignment between the protein predicted in the *LmjF.04.T0860* transcript and the Tb927.9.8290 protein annotated in the *T. brucei* GeneDB database [25]. Identical amino acids are shaded in gray.

Rastrojo *et al. BMC Genomics* 2013, **14**:223
http://www.biomedcentral.com/1471-2164/14/223

Table 2 The 50 most abundant transcripts in *L. major* promastigotes

Transcript	Gene[a]	FPKM[b] (±SD)	Remarks[c]
LmjF.28.T2770	LmjF28.2770	1357.39 ± 5.12	heat-shock protein (HSP70; gene *HSP70-II*)
LmjF.35.T0240	LmjF35.0240	1034.68 ±12.39	ribosomal protein L30
LmjF.28.T2780	LmjF28.2780	987.24 ± 4.45	heat-shock protein hsp70 (HSP70; gene *HSP70-I*)
LmjF.36.T1940	LmjF36.1940	952.68 ± 4.37	inosine-guanosine transporter (NT2)
LmjF.31.T0900	LmjF31.0900	809.07 ± 7.12	hypothetical protein, conserved
LmjF.28.T2205	LmjF28.2205	792.33 ± 8.16	ribosomal protein S29
LmjF.35.T2220	LmjF35.2220	780.12 ± 5.99	kinetoplastid membrane protein-11 (KMP11)
LmjF.19.T0983	Non-annotated	674.85 ± 5.26	-
LmjF.35.T0600	LmjF35.0600	672.00 ± 5.81	ribosomal protein L18a
LmjF.06.T0010	LmjF06.0010	666.85 ± 8.31	histone H4
LmjF.35.T3800	LmjF35.3800	617.33 ± 7.36	ribosomal protein L23
LmjF.36.T3620	LmjF36.3620	616.48 ± 5.07	hypothetical protein, conserved
LmjF.35.T2210	LmjF35.2210	603.26 ± 4.69	kinetoplastid membrane protein-11 (KMP11)
LmjF.28.T2460	LmjF28.2460	596.61 ± 6.27	ribosomal protein S29
LmjF.20.T1285	Non-annotated	560.05 ± 5.48	-
LmjF.31.T0964	Non-annotated	539.44 ± 5.07	-
LmjF.31.T0895	Non-annotated	524.19 ± 10.51	-
LmjF.35.T3290	LmjF35.3290	514.13 ± 5.67	ribosomal protein L31
LmjF.13.T0570	LmjF13.0570	496.01 ± 5.02	ribosomal protein S12
LmjF.35.T3790	LmjF35.3790	493.33 ± 8.18	ribosomal protein L23
LmjF.35.T4191	Non-annotated	490.94 ± 5.15	-
LmjF.35.T3760	LmjF35.3760	483.03 ± 7.04	ribosomal protein L27A/L29
LmjF.30.T3340	LmjF30.3340	482.87 ± 5.89	ribosomal protein L9
LmjF.35.T2050	LmjF35.2050	464.76 ± 5.1	ribosomal protein L32
LmjF.08.T0640	LmjF08.0640	452.92 ± 3.18	hypothetical protein
LmjF.14.T0850	LmjF14.0850	451.18 ± 3.65	calpain-like cysteine peptidase
LmjF.35.T1910	LmjF35.1910	448.56 ± 5.89	ribosomal protein L15
LmjF.35.T0420	LmjF35.0420	446.58 ± 5.12	ribosomal protein S3A
LmjF.35.T1920	LmjF35.1920	446.57 ± 9.14	ribosomal protein L36
LmjF.25.T0910	LmjF25.0910	436.47 ± 3.61	cyclophilin a
LmjF.35.T3780	LmjF35.3780	427.63 ± 4.95	ribosomal protein L27A/L29
LmjF.28.T2740	LmjF28.2740	426.82 ± 4.82	activated protein kinase c receptor
LmjF.13.T0450	LmjF13.0450	425.12 ± 4.7	hypothetical protein, conserved
LmjF.20.T1280	LmjF20.1280	424.16 ± 3.22	small myristoylated protein 4
LmjF.31.T0966	Non-annotated[d]	419.44 ± 8.52	hypothetical protein, conserved
LmjF.28.T2750	LmjF28.2750	414.41 ± 4.15	activated protein kinase c receptor
LmjF.31.T1170	LmjF31.1170	414.01 ± 3.8	hypothetical protein
LmjF.35.T0410	LmjF35.0410	411.95 ± 4.52	ribosomal protein S3A
LmjF.15.T1240	LmjF15.1240	410.71 ± 3.38	nucleoside transporter 1
LmjF.24.T2230	LmjF24.2230	409.67 ± 3.11	hypothetical predicted multi-pass transmembrane protein
LmjF.35.T3280	LmjF35.3280	406.24 ± 5.44	ribosomal protein L31
LmjF.35.T0400	LmjF35.0400	403.61 ± 4.43	ribosomal protein S3A
LmjF.24.T1280	LmjF24.1280	403.46 ± 3.63	amastin-like surface protein
LmjF.13.T0370	LmjF13.0370	403.05 ± 3.84	alpha tubulin
LmjF.35.T1670	LmjF35.1670	403.05 ± 6.46	ribosomal protein L26
LmjF.13.T0360	LmjF13.0360	403.04 ± 3.84	alpha tubulin
LmjF.13.T0350	LmjF13.0350	399.2 ± 3.82	alpha tubulin

Table 2 The 50 most abundant transcripts in *L. major* promastigotes *(Continued)*

LmjF.13.T0380	LmjF13.0380	399.06 ± 3.82	alpha tubulin
LmjF.33.T3230	LmjF33.3230	396.93 ± 7.5	ribosomal protein L44
LmjF.13.T0330	LmjF13.0330	396.86 ± 3.81	alpha tubulin

[a] GeneDB identification code.
[b] FPKM, fragments (reads) per kilobase per million mapped reads.
[c] Hypothetical: predicted by informatics tools; conserved: present in other trypanosomatids (i.e. *T. brucei* and *T. cruzi* genome).
[d] This transcript has 97% of sequence identity with gene LmjF31.0900.

Table 2 contains several examples: genes *LmjF.28.2770* and *LmjF.28.2780*, coding for HSP70; *LmjF.35.3790* and *LmjF.35.3800*, coding for ribosomal protein L23; *LmjF.35.2220* and *LmjF.35.2210*, coding for KMP-11; *LmjF35.0420*, *LmjF35.0410* and *LmjF35.0400*, coding for ribosomal protein S3A; *LmjF28.2740* and *LmjF28.2750*, coding for activated protein kinase c receptor (also known as LACK in *Leishmania* [44]); and the five *LmjF13.0330-0370* genes, coding for alpha tubulin. More frequently, the tandemly arranged genes have identical or highly conserved sequences in their protein coding regions. When RNA-seq reads are mapped at two or more places in the reference genome, due to sequence identity, the assembling algorithms, as that used in this study, make an equal distribution of the reads among the putative transcripts. Obviously, this fact may lead to miscalculation of the expression levels when two transcripts share conserved regions but also contain divergent ones. This can be illustrated analyzing the expression levels of transcripts derived from the *HSP70* locus, i.e. transcripts *LmjF.28.T2770* and *LmjF.28.T2780* (Table 2). Two types of genes, *HSP70-I* and *HSP70-II*, are present in different *Leishmania* species [45]. Both types of genes have identical 5'-UTR and coding sequences, but divergent 3'-UTRs; in addition, analysis of the steady-state mRNA levels in *L. infantum* promastigotes indicated that transcripts derived from the *HSP70-II* gene are one order of magnitude more abundant than *HSP70-I* transcripts [46]. Indeed, according to the FPKM values shown in Table 2, the level of *HSP70-II* transcripts (i.e. *LmjF.28.T2770* transcript) is higher than the level of *HSP70-I* transcripts (i.e. *LmjF.28.T2780*); however, the difference is lower than expected. Fortunately, the counter nature of the RNA-seq data allows defining with precision the transcription level of a particular region of a given gene, and Figure 3A shows the results of this analysis for the *HSP70* locus. While, as expected, the reads mapped to the 5'-UTR + CDS of both genes are equivalent, the number of reads containing sequences belonging to the 3'UTR of *HSP70-II* gene (*LmjF.28.2770*) was 6,24 fold higher than the number of reads corresponding to the 3'UTR of *HSP70-I* gene (*LmjF.28.2780*), even though the 3'-UTR lengths are very similar (1096 and 1084 nucleotides, respectively). These results indicate that the steady-state level for *LmjF.28.T2770* (*HSP70-II*) transcripts is clearly higher than that for *LmjF.28.T2780* (*HSP70-I*) transcripts, giving similar results to those determined by classical methods of mRNA expression levels [46]. Thus, this analysis demonstrated the usefulness of RNA-seq for studies of transcript abundance, and the necessity, however, of knowing and taking into account the differences in the UTR in order to determine transcript levels in a more accurate manner. To further illustrate the usefulness of RNA-seq for determining transcript abundance, we searched for another repeated genes among the more abundant transcripts listed in Table 2. We selected those coding for ribosomal protein L23 (Figure 3B). In the *L. major* genome database [25], there are two tandemly linked *L23* genes (*LmjF.35.3790 and LmjF.35.3800*). A sequence comparison showed that both genes have identical coding regions, but marked differences both in length and sequence in the 3'-UTRs. According to the number of reads obtained for each region of the genes, it was evident that transcript *LmjF.35.T3800* would be more abundant than transcript *LmjF.35.T3790* in the promastigote stage. Thus, after correcting by the length of the 3'-UTRs (93 nucleotides for gene *LmjF.35.3790* and 299 nucleotides for gene *LmjF.35.3800*), the relative steady-state level of transcript *LmjF.35.T3790* was estimated to be 3-fold lower than that for transcript *LmjF.35.T3800*.

On the other hand, our analysis evidenced a negligible level of single nucleotide polymorphisms (SNPs) in the assembled transcripts regarding the reference genome; this is a surprising discovery taking into account that *Leishmania* is an aneuploid organism, in which disomic and trisomic chromosomes are more frequently observed than monosomic ones [47,48]). Similarly, this very low rate of heterozygosity was noted when sequencing the *L. major* genome [13] and, more recently, when Rogers and co-workers re-sequenced the *L. major* genome using the Illumina methodology [48].

Heterogeneity of *trans*-splicing and polyadenylation sites

The addition of a 39-nt mini-exon (or spliced leader, SL) to the 5' end of all mRNAs in *Leishmania* and related trypanosomatids provides the 5' cap structure for mRNA translation [32]. As noted above, we have obtained a large number of mini-exon-containing reads and this facilitated the 5' end-mapping for most transcripts. Furthermore, we mapped two or more SL addition sites for around 50% of the genes, suggesting the existence of a remarkable

Rastrojo *et al. BMC Genomics* 2013, **14**:223
http://www.biomedcentral.com/1471-2164/14/223

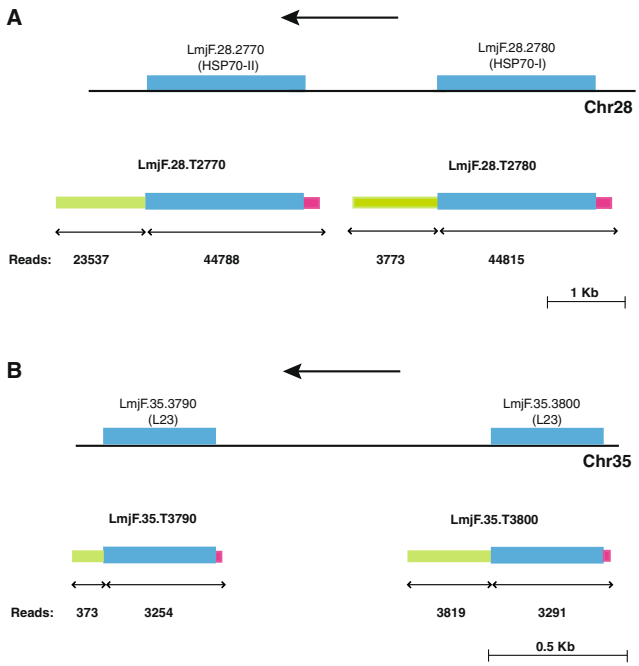


Figure 3 Relative expression levels of transcripts derived from loci coding for HSP70 and ribosomal protein L23. Panel **A**: *upper*, current annotation of the two types of genes found in the *L. major HSP70* locus, *LmjF28.2770* (also known as *HSP70-II*) and *LmjF28.2780* (*HSP70-I*); *bottom*, transcript annotation, as defined in this study, for the *L. major HSP70* locus. Panel **B**: *upper*, current annotation of the two genes coding for the ribosomal protein L23, *LmjF35.3790* and *LmjF35.3800*; *bottom*, transcripts annotated in this study. Each locus is composed by two types of genes with identical ORFs (blue boxes). Transcript mapping has allowed the identification of 5'-UTRs (purple boxes) and 3'-UTRs (green boxes). The number of reads mapped by Cufflinks within the different regions is shown at bottom.

heterogeneity in the selection of the SL addition site. A similar observation has been reported in RNA-seq studies carried out in a related trypanosmatid, *T. brucei* [34,49,50]. For bioinformatics analyses, when the distance between two consecutive SL addition sites was lower than 500 nucleotides, they were considered alternative addition sites for a given gene. Furthermore, taking into account the numbers of reads mapping at each site, they were categorized as either main addition sites or alternative splicing sites. Thus, SL addition sites were separated into two categories: i) main SL addition sites, including unique SL addition sites or the most frequent SL addition sites when two or more sites were mapped in the same transcript; ii) alternative SL addition sites, i.e. the rest of SL addition sites in transcripts containing two of more SL addition sites. In order to avoid possible bias, SL-containing reads mapping to ten or more different genomic positions were excluded from this analysis. Most of these "multi-hit" reads mapped to the 5'-UTR of gene families containing ten or more members (Crespillo et al, manuscript in preparation). Finally, a total of 9530 SL addition sites were classified as main sites and 4531 as alternative ones. Looking for sequence signatures associated with the SL addition sites, a compositional analysis in the ± 40 nucleotide region surrounding the SL addition site was carried out (Figure 4). Addition of SL occurs after the well-known AG dinucleotide, even though a slight difference was observed between main and alternative sites. Thus, whereas for the main SL addition sites, the A and G frequencies were 99.82% and 99.87%, respectively, for the alternative sites the A and G frequencies were 97.26 and 99.23%. Additionally, in agreement with previous analysis [51], a preference for a C

Rastrojo *et al. BMC Genomics* 2013, **14**:223
http://www.biomedcentral.com/1471-2164/14/223

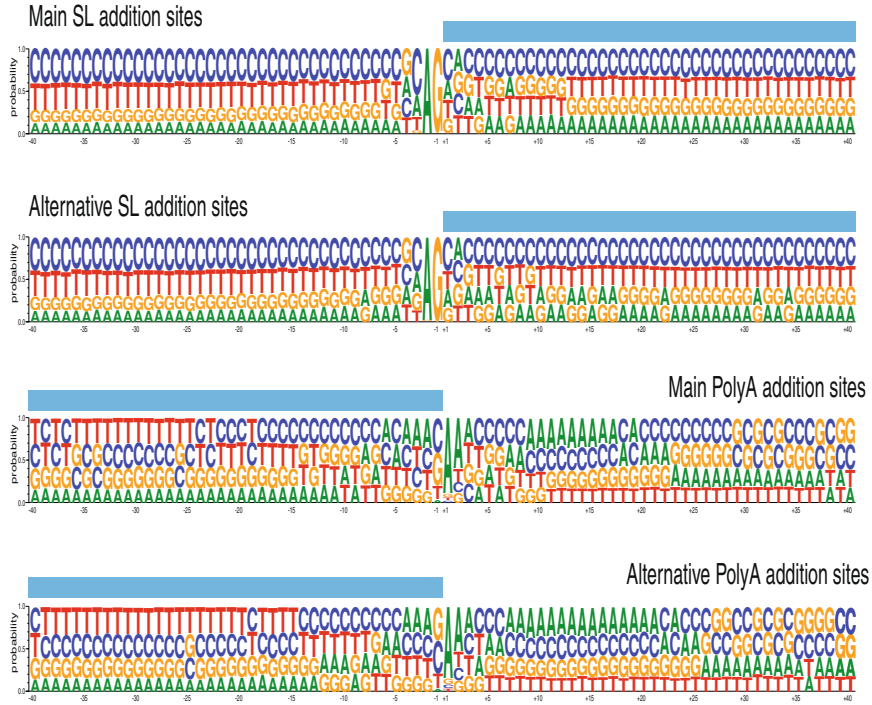


Figure 4 Nucleotide frequencies for sequences surrounding SL and polyadenylation addition sites. Panels show the compositional profiles of sequences around the main SL addition sites (n = 9530), alternative SL addition sites (n = 4531), main polyadenylation sites (n = 3178) and alternative polyadenylation sites (n = 1238).

before the AG dinucleotide was observed. Again, this preference was more marked in the main addition sites (55.26%) than in the alternative addition sites (40.87%). Another noticeable feature of the sequence surrounding the AG addition site is a clear richness in pyrimidine nucleotides (Figure 4). This T + C richness is more pronounced in the upstream region: in the -40 to -21 positions the T + C frequencies are higher than 70%. Likewise, the T + C content is higher in the regions upstream the main addition sites (71.28%) than the alternative ones (68.68%).

Heterogeneity in the polyadenylation sites in the *Leishmania* transcripts was also observed; however, the number of reads found denoting polyadenylation events was lower (7894 reads) than those mapped at the 5'end (see above), in spite that we prepared a second library in which an oligo-dT for priming was included in the cDNA reaction (see Methods section). Difficulties in the identification of polyadenylation sites were also experienced by other authors [34]. Recently, P.J. Myler and coworkers have deposited in the TriTrypDB database [26] a large number of SL- and polyadenylation sites for *L. major*; these new data further illustrate the complexity of *trans*-splicing and polyadenylation site selection in *Leishmania*. A comparative study between our data and those from Myler's laboratory is underway. Nevertheless, some conclusions may be drawn from the analysis of those reads mapping at the polyadenylation sites derived from our data. Polyadenylation sites were categorized as main (3178 different sites) and alternative (1238 sites). A compositional analysis of the regions surrounding the polyadenylation sites for both categories is shown in Figure 4. Searching for possible consensus sequence, we followed the consensus criteria defined by Cavener [52]: a consensus status is assigned to a single base when the frequency of a nucleotide at a certain position is greater than

50% and greater than twice the relative frequency of the second most frequent nucleotide; a pair of bases were assigned co-consensus status if the sum of the relative frequencies of the two nucleotides exceeded 75%. The application of this rule leads to a very short consensus for the polyadenylation addition site, which may be defined as (C/G) AA; the noteworthy differences between main and alternative sites were: i) C is more frequent in the main sites (40.62%) than in the alternative ones (38.77%); ii) the frequencies for A residues at position 2 and 3 are higher in the main sites (88.92 and 59.1%, respectively) than those found in the alternative sites (83.84 and 51.93%, respectively). An unresolved question related to the polyadenylation consensus sequence is whether the polyadenylation occurs either before or after the AA dinucleotide. Although our data cannot elucidated this question, it is likely that the adenosines of the consensus sequences are encoded residues as it is well known that poly(A) polymerases prefer an initial adenosine residue for attachment of the poly(A) tail, and therefore the selection of the polyadenylation site would be strengthened by the presence of adenosine residues [53].

Conclusions

Sequencing and annotation of the genomes for some *Leishmania* species [13,54] have constituted an important milestone for the study of many biological aspects of this group of parasites. The availability of these genome sequences [25,26] now enables database mining and identification of different protein sets in *Leishmania*. This information provided new approaches to study the pattern of gene expression during differentiation and development by the use of DNA microarrays [55]. In current genome databases, the *Leishmania* genes lacks the definition of 5' and 3' UTRs; however, it should be noticed that recently P.J. Myler and coworkers have incorporated SL and polyA sites for most genes of *L. major* in the TriTrypDB datase [26]. The RNA-seq study described here represents the first annotation of the *L. major* transcriptome, in which the genes have been delimited in their translated and untranslated regions. As a result, we have uncovered many cases of mis-annotated genes, and more importantly we have found 1884 new genes (previously non-annotated) in the promastigote stage. In addition, we have determined relative expression levels for each one of the 10285 transcripts detected in *L. major* promastigotes. In summary, the data generated by this study constitute a framework for future analysis aimed to determine differential gene expression either along the life cycle or among different *Leishmania* species.

Additional file

Additional file 1: An Excel file containing the complete transcriptome information for each one of the 36 *L. major* chromosomes. The coordinates for each transcript are provided (Locus column), in addition to location of main (SL) and alternative SL (Alt_SL) addition sites, location of main (polyA) and alternative (Alt_polyA) polyadenylation sites. It is also indicated (GeneDB_ID) whether the transcript corresponds to a previously annotated gene (in this case the corresponding annotated gene is provided) or represents a new gene (denoted as unknown). The relative levels, expressed as FPKM, for each transcript are also provided. Finally, transcripts evidencing mis-annotated translation start codons are remarked as truncated.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
FC, BA and JMR were responsible for design and coordination of this study. AR carried out most of the bioinformatics analyses; DM, AC, RMR and JMR contributed to data analysis. JMR and BA wrote the manuscript. All the authors read and approved the final manuscript.

Acknowledgements
We are extremely grateful to Dr Julie Sheldon for English style corrections and critical reading of the manuscript. This work was funded by Ministerio de Ciencia y Tecnología [BFU2009-08986 to J.M.R., BFU 2008-03126 to B.A.], Comunidad Autonoma de Madrid (S2010/BMD-2361 to J.M.R.), and the Fondo de Investigaciones Sanitarias [ISCIII-RETIC RD06/0021/0008-FEDER to J. M.R. and R.M.R]. A.R. holded a postgraduate fellowship (FPU) from the Ministerio de Educación y Ciencia. Also, an institutional grant from Fundación Ramón Areces is acknowledged.

Author details
[1]Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, 28049 Madrid, Spain. [2]Departamento de Ciencias Biomédicas, Universidad de León, 24071 León, Spain.

Received: 4 September 2012 Accepted: 25 February 2013
Published: 4 April 2013

References

1. Desjeux P: **Leishmaniasis: current situation and new perspectives.** *Comp Immunol Microbiol Infect Dis* 2004, **27:**305–318.
2. Moreira D, Lopez-Garcia P, Vickerman K: **An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea.** *Int J Syst Evol Microbiol* 2004, **54:**1861–1875.
3. Baldauf SL: **The deep roots of eukaryotes.** *Science* 2003, **300:**1703–1706.
4. Myler PJ, Audleman L, DeVos T, Hixson G, Kiser P, Lemley C, Magness C, Rickel E, Sisk E, Sunkin S, *et al*: ***Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes.** *Proc Natl Acad Sci U S A* 1999, **96:**2902–2906.
5. Martinez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler PJ: **Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region.** *Mol Cell* 2003, **11:**1291–1299.
6. Martinez-Calvillo S, Nguyen D, Stuart K, Myler PJ: **Transcription initiation and termination on *Leishmania major* chromosome 3.** *Eukaryot Cell* 2004, **3:**506–517.
7. Thomas S, Green A, Sturm NR, Campbell DA, Myler PJ: **Histone acetylations mark origins of polycistronic transcription in *Leishmania major*.** *BMC Genomics* 2009, **10:**152.
8. van Luenen HGAM, Farris C, Jan S, Genest PA, Tripathi P, Velds A, Kerkhoven RM, Nieuwland M, Haydock A, Ramasamy G, *et al*: **Glucosylated hydroxymethyluracil, DNA base j, prevents transcriptional readthrough in *Leishmania*.** *Cell* 2012, **150:**909–921.
9. LeBowitz JH, Smith HQ, Rusche L, Beverley SM: **Coupling of poly(A) site selection and trans-splicing in *Leishmania*.** *Genes Dev* 1993, **7:**996–1007.
10. Fernandez-Moya SM, Estevez AM: **Posttranscriptional control and the role of RNA-binding proteins in gene regulation in trypanosomatid protozoan parasites.** *Wiley Interdiscip Rev RNA* 2010, **1:**34–46.
11. Requena JM: **Lights and shadows on gene organization and regulation of gene expression in *Leishmania*.** *Front Biosci* 2011, **17:**2069–2085.
12. Kramer S: **Developmental regulation of gene expression in the absence of transcriptional control: the case of kinetoplastids.** *Mol Biochem Parasitol* 2012, **181:**61–72.

13. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, *et al*: **The Genome of the Kinetoplastid Parasite, *Leishmania major*.** *Science* 2005, **309:**436–442.
14. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, *et al*: **Comparative genomics of trypanosomatid parasitic protozoa.** *Science* 2005, **309:**404–409.
15. Coulson RM, Connor V, Ajioka JW: **Using 3' untranslated sequences to identify differentially expressed genes in *Leishmania*.** *Gene* 1997, **196:**159–164.
16. Martin JA, Wang Z: **Next-generation transcriptome assembly.** *Nat Rev Genet* 2011, **12:**671–682.
17. Siegel TN, Gunasekera K, Cross GAM, Ochsenreiter T: **Gene expression in *Trypanosoma brucei*: lessons from high-throughput RNA sequencing.** *Trends Parasitol* 2011, **27:**434–441.
18. FASTQC: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/.
19. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10:**R25.
20. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28:**511–515.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078–2079.
22. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC Bioinformatics* 2009, **10:**421.
23. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer.** *Nat Biotechnol* 2011, **29:**24–26.
24. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome Res* 2004, **14:**1188–1190.
25. GeneDB: www.genedb.org.
26. TritrypDB: www.tritrypdb.org.
27. NCBI: www.ncbi.nlm.nih.gov.
28. Jager AV, De Gaudenzi JG, Cassola A, D'Orso I, Frasch AC: **mRNA maturation by two-step trans-splicing/polyadenylation processing in trypanosomes.** *Proc Natl Acad Sci U S A* 2007, **104:**2035–2042.
29. Kapler GM, Beverley SM: **Transcriptional mapping of the amplified region encoding the dihydrofolate reductase-thymidylate synthase of *Leishmania major* reveals a high density of transcripts, including overlapping and antisense RNAs.** *Mol Cell Biol* 1989, **9:**3959–3972.
30. Soto M, Requena JM, Alonso C: **Isolation, characterization and analysis of the expression of the *Leishmania* ribosomal PO protein genes.** *Mol Biochem Parasitol* 1993, **61:**265–274.
31. Monnerat S, Martinez-Calvillo S, Worthey E, Myler PJ, Stuart KD, Fasel N: **Genomic organization and gene expression in a chromosomal region of *Leishmania major*.** *Mol Biochem Parasitol* 2004, **134:**233–243.
32. Agabian N: **Trans splicing of nuclear pre-mRNAs.** *Cell* 1990, **61:**1157–1160.
33. Agami R, Shapira M: **Nucleotide sequence of the spliced leader RNA gene from *Leishmania mexicana amazonensis*.** *Nucleic Acids Res* 1804, **1992:**20.
34. Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S, Tschudi C: **The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution.** *PLoS Pathog* 2010, **6:**e1001090.
35. Aline RF Jr, Scholler JK, Stuart K: **Transcripts from the co-transposed segment of variant surface glycoprotein genes are in *Trypanosoma brucei* polyribosomes.** *Mol Biochem Parasitol* 1989, **32:**169–178.
36. Benabdellah K, Gonzalez-Rey E, Gonzalez A: **Alternative trans-splicing of the *Trypanosoma cruzi* LYT1 gene transcript results in compartmental and functional switch for the encoded protein.** *Mol Microbiol* 2007, **65:**1559–1567.
37. Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LW, Sasidharan R, Reinke V, Waterston RH, Gerstein M: **Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays.** *BMC Genomics* 2010, **11:**383.
38. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18:**1509–1517.
39. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nature methods* 2008, **5:**621–628.
40. Brandau S, Dresel A, Clos J: **High constitutive levels of heat-shock proteins in human-pathogenic parasites of the genus *Leishmania*.** *Biochem J* 1995, **310:**225–232.
41. Jardim A, Funk V, Caprioli RM, Olafson RW: **Isolation and structural characterization of the *Leishmania donovani* kinetoplastid membrane protein-11, a major immunoreactive membrane glycoprotein.** *Biochem J* 1995, **305:**307–313.
42. Leifso K, Cohen-Freue G, Dogra N, Murray A, McMaster WR: **Genomic and proteomic expression analysis of *Leishmania* promastigote and amastigote life stages: the *Leishmania* genome is constitutively expressed.** *Mol Biochem Parasitol* 2007, **152:**35–46.
43. Guerfali FZ, Laouini D, Guizani-Tabbane L, Ottones F, Ben-Aissa K, Benkahla A, Manchon L, Piquemal D, Smandi S, Mghirbi O, *et al*: **Simultaneous gene expression profiling in human macrophages infected with *Leishmania major* parasites using SAGE.** *BMC Genomics* 2008, **9:**238.
44. Mougneau E, Altare F, Wakil AE, Zheng S, Coppola T, Wang ZE, Waldmann R, Locksley RM, Glaichenhaus N: **Expression cloning of a protective *Leishmania* antigen.** *Science* 1995, **268:**563–566.
45. Folgueira C, Cañavate C, Chicharro C, Requena JM: **Genomic organization and expression of the HSP70 locus in New and Old World *Leishmania* species.** *Parasitology* 2007, **134:**369–377.
46. Quijada L, Soto M, Alonso C, Requena JM: **Analysis of post-transcriptional regulation operating on transcription products of the tandemly linked *Leishmania infantum* hsp70 genes.** *J Biol Chem* 1997, **272:**4493–4499.
47. Sterkers Y, Lachaud L, Crobu L, Bastien P, Pages M: **FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in *Leishmania major*.** *Cell Microbiol* 2011, **13:**274–283.
48. Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, Harris D, Her Y, Herzyk P, Imamura H, *et al*: **Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania.** *Genome Res* 2011, **21:**2129–2142.
49. Siegel TN, Hekstra DR, Wang X, Dewell S, Cross GAM: **Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites.** *Nucleic Acids Res* 2010, **38:**4946–4957.
50. Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, Baerlocher L, Roditi I, Ochsenreiter T: **Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*.** *PLoS Pathog* 2010, **6:**e1001037.
51. Requena JM, Quijada L, Soto M, Alonso C: **Conserved nucleotides surrounding the trans-splicing acceptor site and the translation initiation codon in *Leishmania* genes.** *Exp Parasitol* 2003, **103:**78–81.
52. Cavener DR: **Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates.** *Nucleic Acids Res* 1987, **15:**1353–1361.
53. Wahle E, Keller W: **The biochemistry of 3'-end cleavage and polyadenylation of messenger RNA precursors.** *Annu Rev Biochem* 1992, **61:**419–440.
54. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, *et al*: **Comparative genomic analysis of three *Leishmania* species that cause diverse human disease.** *Nat Genet* 2007, **39:**839–847.
55. Cohen-Freue G, Holzer TR, Forney JD, McMaster WR: **Global gene expression in *Leishmania*.** *Int J Parasitol* 2007, **37:**1077–1086.

doi:10.1186/1471-2164-14-223
Cite this article as: Rastrojo *et al.*: The transcriptome of *Leishmania major* in the axenic promastigote stage: transcript annotation and relative expression levels by RNA-seq. *BMC Genomics* 2013 **14**:223.

Hernández-Torres *et al. BMC Genomics* 2013, **14**:199
http://www.biomedcentral.com/1471-2164/14/199

BMC Genomics

RESEARCH ARTICLE Open Access

Intron retention and transcript chimerism conserved across mammals: Ly6g5b and Csnk2b-Ly6g5b as examples

Francisco Hernández-Torres[1,2], Alberto Rastrojo[1] and Begoña Aguado[1*]

Abstract

Background: Alternative splicing (AS) is a major mechanism for modulating gene expression of an organism, allowing the synthesis of several structurally and functionally distinct mRNAs and protein isoforms from a unique gene. Related to AS is the Transcription Induced Chimerism (TIC) or Tandem Chimerism, by which chimeric RNAs between adjacent genes can be found, increasing combinatorial complexity of the proteome. The *Ly6g5b* gene presents particular behaviours in its expression, involving an intron retention event and being capable to form RNA chimera transcripts with the upstream gene *Csnk2b*. We wanted to characterise these events more deeply in four tissues in six different mammals and analyse their protein products.

Results: While canonical *Csnk2b* isoform was widely expressed, *Ly6g5b* canonical isoform was less ubiquitous, although the *Ly6g5b* first intron retained transcript was present in all the tissues and species analysed. *Csnk2b-Ly6g5b* chimeras were present in all the samples analysed, but with restricted expression patterns. Some of these chimeric transcripts maintained correct structural domains from *Csnk2b* and *Ly6g5b*. Moreover, we found *Csnk2b, Ly6g5b,* and *Csnk2b-Ly6g5b* transcripts that present exon skipping, alternative 5' and 3' splice site and intron retention events. These would generate truncated or aberrant proteins whose role remains unknown. Some chimeric transcripts would encode *CSNK2B* proteins with an altered C-terminus, which could affect its biological function broadening its substrate specificity. Over-expression of human *CSNK2B, LY6G5B,* and *CSNK2B-LY6G5B* proteins, show different patterns of post-translational modifications and cell distribution.

Conclusions: *Ly6g5b* intron retention and *Csnk2b-Ly6g5b* transcript chimerism are broadly distributed in tissues of different mammals.

Background

To date, a high number of eukaryotic genomes have been sequenced. Surprisingly, it is interesting to observe that *Homo sapiens* and *Caenorhabditis elegans* genomes contain a similar number of protein-coding genes (~21.000), according to the Ensembl (http://www.ensembl.org/) database. Initially, these findings disconcerted researchers who thought the number of genes should be correlated with developmental and physiological complexity, and made them realize that other mechanisms should be involved in this evolutionary variety. Alternative splicing (AS) is a major mechanism for modulating gene expression of an organism, and enables a single gene to increase its expression capacity, allowing the synthesis of several structurally and functionally distinct mRNAs and protein isoforms from a unique gene (For reviews see [1-3]). This mechanism, which was initially described in viruses [4-6], is now known to affect 95% of all human genes [7] and has been proposed as a primary driver of the evolution of phenotypic complexity in mammals [8-10]. The human Major Histocompatibility Complex (MHC) is located on chromosome 6, and is ~4 Mb in length. It is composed by three regions, the class I and class II regions flanking the central class III region. The class III region is ~0.9 Mb in length and contains 62–64 genes and 2–3 pseudogenes, depending on the haplotype [11,12]. Previously, our group precisely defined the AS patterns of a five gene cluster from the Lymphocyte antigen-6 (LY-6) superfamily [13] and characterised the

* Correspondence: baguado@cbm.uam.es
[1]Centro de Biología Molecular Severo Ochoa (CBMSO), Consejo Superior de Investigaciones Científicas (CSIC)-Universidad Autónoma de Madrid, Madrid, Spain
Full list of author information is available at the end of the article

expression of the corresponding proteins [14] in human and mouse. LY-6 superfamily members are cysteine-rich, generally GPI-anchored, cell surface proteins, which have definite or putative immune-related roles [15]. Among these LY-6 MHC class III region genes, *Ly6g5b* showed a particular behaviour in the regulation of its expression [13,16], involving an intron retention event in human and mouse, the rarest form of alternative splicing found in metazoan species [17]. The intron retained is the first one after the initial exon and interrupts the open reading frame (ORF) just after the signal peptide introducing a premature stop codon (PSC). The presence of a PSC at this position should cause this intron-retained transcript to undergo Nonsense Mediated Decay (NMD) [18,19]. However, this transcript seemed to escape NMD and was more abundant than the correctly spliced mRNA [13,16]. In addition, findings in our laboratory showed the presence of *LY6G5B* gene exons in transcripts derived from the upstream gene *CSNK2B* [16], which encodes the *Casein Kinase II β subunit*, a ubiquitous protein kinase which regulates metabolic pathways, signal transduction, transcription, translation, and replication [20,21]. The Transcription Induced Chimerism (TIC) or Tandem Chimerism, is a phenomenon whose mechanism still remains largely unknown, although it is being promoted as a novel way to increase combinatorial complexity of the proteome [22,23]. At least 4%-5% of the tandem gene pairs in the human genome can be eventually transcribed into a single RNA sequence encoding a putative chimeric protein [22] but recent bioinformatic analyses, partially supported by experimental data, show that this phenomenon could be quite frequent [24,25]. Recently, it has been showed that these chimeras significantly exploit signal peptides and transmembrane domains, which could alter the cellular localisation of cognate proteins, and that chimeric RNAs are more tissue specific than non-chimeric transcripts [26]. Thus, this novel mechanism directly related to AS could have an important role in evolution divergence. In this regard, the majority of these studies give relevance to this phenomenon in *Homo sapiens* [24], but detailed comparative analyses among different species are, to our knowledge, not described. Here, we have deeply characterised the expression of *Ly6g5b* and *Csnk2b* transcripts independently and of the *Csnk2b-Ly6g5b* chimeric transcripts in four defined tissues among six different mammals. We conclude that *Ly6g5b* intron retention and *Csnk2b-Ly6g5b* chimerism is present in the tissues of the analysed mammalian group. In addition, we have made a comparative analysis of human CSNK2B, LY6G5B, and CSNK2B-LY6G5B protein expressions.

Results

Csnk2b transcript analysis

Canonical Csnk2kb ORF orthologue sequences from *Homo sapiens* (NM_001320), *Macaca mulata* (XM_001112478), *Sus scrofa* (XM_001928731), *Bos taurus* (NM_001046454), *Rattus norvegicus* (NM_031021) and *Mus musculus* (NM_009975) were analysed in order to find common features among them. Comparative analysis showed a total conservation rate in protein sequence among these six species, except for *Mus musculus* which presents a unique change in position 57 (V→ E) (Figure 1A). Through RT-PCR analysis, we found five different transcripts for *Csnk2b* in *Homo sapiens*, four in *Macaca mulata*, four in *Sus scrofa*, five in *Bos taurus*, two in *Rattus norvegicus* and three in *Mus musculus* (Figure 2 and Additional file 1). Only the canonical transcript sequences were on databases, except for *Bos taurus* for which BtCsnk2b-473 was also present (see Additional file 2: Table S1). We could detect the presence of the canonical *Csnk2b* transcript in each tested species and tissues (Figure 2), and it was also the isoform expressed at the highest level (data not shown). In addition, *Csnk2b* expression also generated other transcripts (Figure 2) expressed at lower levels (data not shown), but with a remarkable specificity among the analysed tissues in these six species (Figure 2). Some of them presented quite restricted expression patterns, such as the *Macaca mulata* ones that are only expressed in lung, or the *Rattus norvegicus* one, expressed only in brain. By contrast, the variants from *Sus scrofa*, *Bos taurus* and *Mus musculus* are broadly expressed. In *Homo sapiens*, the isoform which retains intron 5 is widely expressed; however the other three are only expressed in liver and lung. Through AS these *CSNK2B* transcripts present exon skipping, alternative 5' and 3' splice site and intron retention events (Figure 2) in the different species, which would generate severely truncated or aberrant proteins by using the canonical start codon.

Ly6g5b transcript analysis

Canonical *Ly6g5b* ORF orthologue sequences from *Homo sapiens* (NM_021221), *Macaca mulata* (XR_014070), *Sus scrofa* (XM_001926307), *Bos taurus* (XM_585827), *Rattus norvegicus* (NM_001001934) and *Mus musculus* (NM_148939) were compared (Figure 1B). Although some differences in amino acid sequence can be detected, a *LY-6* protein domain conservation in these *Ly6g5b* orthologues is clearly present (Figure 1B). This domain is composed of ~80 amino acids and is characterised by a conserved pattern of eight to ten cysteine residues that have a defined disulfide-bounding pattern [14]. Through RT-PCR analysis, we found four different transcripts for *Ly6g5b* in Homo sapiens, two in *Macaca mulata*, three in *Sus scrofa*, three in *Bos taurus*, four in *Rattus norvegicus* and two in *Mus musculus* (Figure 2 and Additional file 1). The majority of these transcripts were not available on databases; even the canonical sequences (see Additional file 2: Table S1). Curiously, the presence of the canonical *Ly6g5b* transcript was only detected in three of the analysed species

Hernández-Torres *et al. BMC Genomics* 2013, **14**:199
http://www.biomedcentral.com/1471-2164/14/199

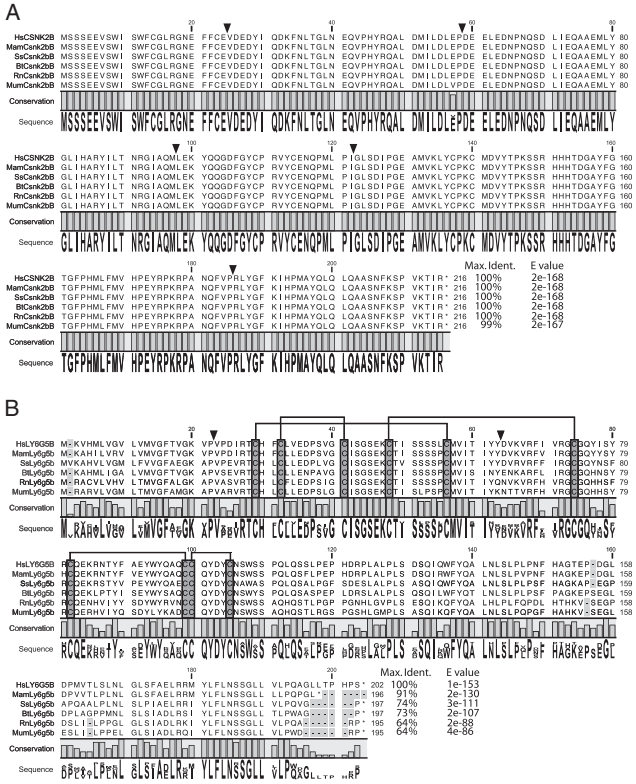


Figure 1 ***Homo sapiens*** **(Hs),** ***Macaca mulata*** **(Mam),** ***Sus scrofa*** **(Ss),** ***Bos taurus*** **(Bt),** ***Rattus norvegicus*** **(Rn) and** ***Mus musculus*** **(Mum) ORF alignments. A**) Csnk2b: N-terminal region α1-α5 (5-104), juxta-dimer interface region including zinc-finger (105-147) and α6 (163-175), C-terminal region (178-205) including interaction with CSNK2α (175-193) and dimerization (190-205) regions [27], **B**) Ly6g5b: The conserved cysteines, characteristic for Ly-6 domain, are shown highlighted in grey and their connectivity with interconnected lines. Exon-exon junctions are indicated by inverted triangles. The canonical (or theoretical canonical) sequence for each specie is only shown. The percentage of identity of each sequence respect to the human one is shown as well as the corresponding E values.

(*Homo sapiens*, *Rattus norvegicus* and *Mus musculus*). Similarly to what happened for *Csnk2b*, different transcript variants are generated for *Ly6g5b* through AS. These transcripts present a remarkable specificity among the analysed tissues and species, and assuming that they could be translated into proteins starting from the first canonical start codon, only truncated or aberrant proteins could be generated by them. Nevertheless, there is an interesting feature that should be stressed, the retention of the first intron in *Ly6g5b* transcripts, giving rise to a particular isoform (exon 1, intron 1, exon 2 and exon 3) that is present in all the tissues and analysed species (Figure 2), indicating conservation, and presenting the highest expression levels (data not shown). This isoform contains a PSC after the canonical start codon, in the middle of the retained intron, and therefore should be degraded through control mechanisms like NMD [18,19]. However this seems not to be the case, as we have previously described in human [16].

Csnk2b-Ly6g5b Chimeric transcript analysis

Through RT-PCR analysis, we found ten different *Csnk2b-Ly6g5b* chimeric transcripts in *Homo sapiens*, five in

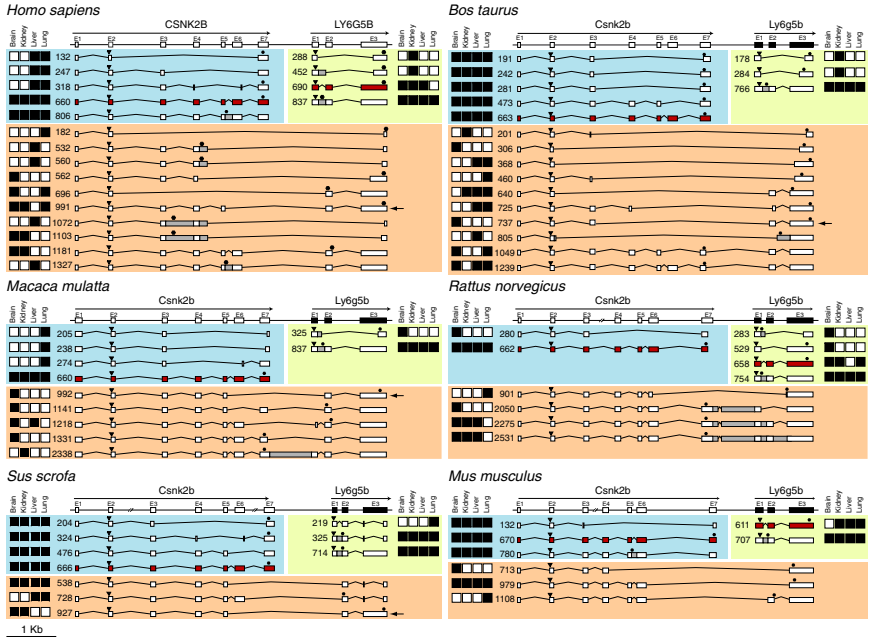


Figure 2 Schematic representation of all Csnk2b and Ly6g5b transcripts detected through nested RT-PCR on each analysed tissue for *Homo sapiens, Macaca mulata, Sus scrofa, Bos Taurus, Rattus norvegicus and Mus musculus.* White boxes represent Csnk2b, Ly6g5b and chimera Csnk2b-Ly6g5b exons. Grey boxes represent intron sequences retained on each transcript. Red boxes indicate canonical isoforms. ORFs are delimited by triangles (ATG codon) and hexagons (Stop codon). Squares at the left or right, represent the presence (black) or absence (white) of expression of each transcript on the indicated tissue. Black arrows indicate Ly6g6b-Csnk2b chimeras that carry structural domains from Csnk2b and Ly6g5b, respectively. The corresponding nucleotide and protein sequences and the accession numbers (EMBL: HE864415-HE864490) are shown in Additional file 1.

Macaca mulata, three in *Sus scrofa*, ten in *Bos taurus*, four in *Rattus norvegicus* and three in *Mus musculus* (Figure 2 and Additional file 1). Only human Csnk2b - Ly6g5b -1181 was found on databases (see Additional file 2: Table S1). As it happened with Csnk2b and Ly6g5b independent transcripts described above, AS seems to play an important role in generating these chimeras, and results in a set of transcripts that greatly vary in terms of composition and size. Indeed, exon skipping, intron retention and intergenic region retention events are present in these transcripts. The majority of the described chimeras (26/35) have a common characteristic: the total lack of the last exon (exon 7) of the upstream gene (*Csnk2b*) as well as the first exon of the downstream gene (*Ly6g5b*). There are also four chimeric transcripts that partially lack *Csnk2b* last exon (exon 7) (two in *Macaca mulata* and two in *Bos taurus*), one that partially maintains *Ly6g5b* first exon (in *Macaca mulata*) and four that retain the intergenic regions (three in *Rattus norvegicus* and one in *Macaca mulata*). Although chimeras function is still unknown, some authors defend that this kind of fusion might generate bi-functional proteins which would have the properties of both original proteins [23,26]. Assuming this, we determined the number of chimeric transcripts which conserved the ORF of both *Csnk2b* and *Ly6g5b* genes. We found such transcripts in *Homo sapiens*, *Macaca mulata*, *Sus scrofa* and *Bos taurus*, of which only *HsCsnk2b* - Ly6g5b *-991*, *MamCsnk2b-Ly6g5b-992* and *SsCsnk2b-Ly6g5b-927* (Figures 2, 3 and Additional file 1) maintain the N-terminal functional domains (alpha helices 1 to 6) from *Csnk2b* (see Figures 1A and 3) [27], such as the acidic loop (aa 55-64) and nuclear localization sequence (aa 9-14 or α1), as well as the *LY-6* structural domain (see Figures 1B and 3) [14], allowing the possibility to create potentially bi-

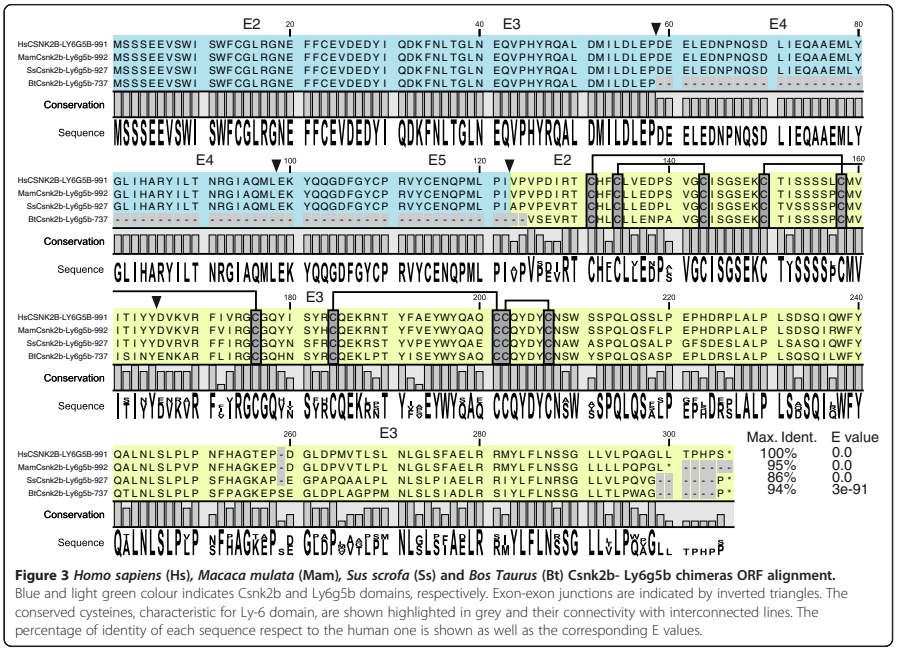


Figure 3 ***Homo sapiens*** **(Hs),** ***Macaca mulata*** **(Mam),** ***Sus scrofa*** **(Ss) and** ***Bos Taurus*** **(Bt) Csnk2b- Ly6g5b chimeras ORF alignment.** Blue and light green colour indicates Csnk2b and Ly6g5b domains, respectively. Exon-exon junctions are indicated by inverted triangles. The conserved cysteines, characteristic for Ly-6 domain, are shown highlighted in grey and their connectivity with interconnected lines. The percentage of identity of each sequence respect to the human one is shown as well as the corresponding E values.

functional proteins [26,28,29]. These particular transcripts which contain the same exon-intron structure are expressed in different tissues, but commonly in brain. *Bos taurus BtCsnk2b-Ly6g5b-737* maintains only exons 1 to 3 of *Csnk2b* and then would not encode its entire N-terminal domain [27]. We did not find this type of "bi-functional" chimeric transcripts in *Rattus norvegicus* or *Mus musculus* (Figure 2).

It is interesting to note that several (23/35) chimeric transcripts could encode *Csnk2b* truncated proteins (7/35) or with modifications in their C-terminus (16/35). The *CSNK2B* C-terminal part is involved in homodimerization and binding to CSNK2A subunit (see Figure 1) [27]. Some of these detected chimeric transcripts are generated by replacing the canonical sequence of exon 7 by sequences encoded by total or partial exon 2 or 3 of *Ly6g5b* but not corresponding to *LY-6* amino-acid sequences due to changes in the reading frames. These are chimeras *HsCsnk2b-Ly6g5b-1181, MamCsnk2b-Ly6g5b-1218, SsCsnk2b-Ly6g5b-728, MumCsnk2b-Ly6g5b-979* and *MumCsnk2b-Ly6g5b-1108*, with variable tissue distribution, except *MumCsnk2b-Ly6g5b-979* that is expressed in the four tissues. Other transcripts are altered on *Csnk2b* exon 6 or 7, lacking the zinc-finger domain, α6 and C-terminal regions of CSNK2B (see Figure 1), such as *MamCsnk2b-Ly6g5b-1141* (only expressed in brain) and *BtCsnk2b-Ly6g5b-1049* (expressed in brain and lung) which maintain the same exon-intron structure (lack of Csnk2b exon 6 and Ly6g6b exon 1), and *SsCsnk2b-Ly6g5b-538* (expressed in the four tissues) (see Figure 2). In addition, there are some chimeric transcripts that would encode a complete CSNK2B protein considering that they contain all exons (1-7) of *Csnk2b* including the stop codon and in which the Ly6g5b nucleotide sequences will act as 3´ UTRs. These are *MamCsnk2b-Ly6g5b-1331, MamCsnk2b-Ly6g5b-2338, BtCsnk2b-Ly6g5b-1239, RnCsnk2b-Ly6g5b-2050, RnCsnk2b-Ly6g5b-2275* and *RnCsnk2b-Ly6g5b-2531.* They present variable tissue distribution and exon-intron structure (see Figure 2).

Csnk2b, Ly6g5b and *Csnk2b -Ly6g5b* Chimera protein analysis

In order to analyse post-translational modifications and sub-cellular localisation of human CSNK2B, LY6G5B and CSNK2B-LY6G5B proteins, we over-expressed them in the COS7 cell line (Figure 4), using a double tag strategy. Thus, CSNK2B, LY6G5B and CSNK2B-LY6G5B proteins

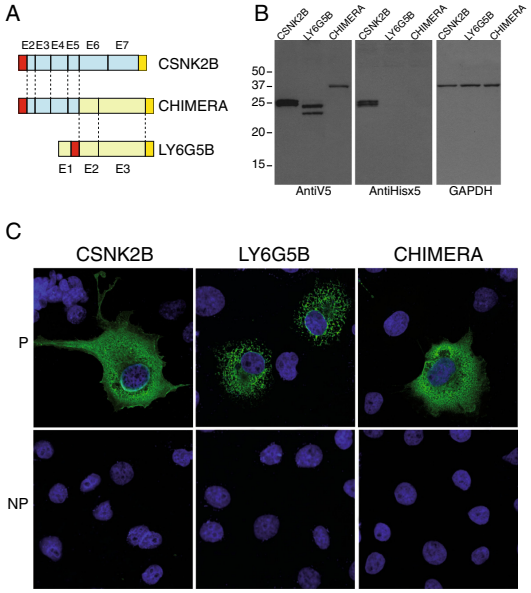


Figure 4 Human CSNK2B, LY6G5B and CSNK2B-LY6G5B chimeric protein characterisation. Schematic representation of **A**) V5 and Hisx6 double tag strategy. V5 (red box) and Hisx6 (yellow box) epitopes are shown. Exon numbers are indicated. **B**) Western-blot analysis showing in vitro expression of CSNK2B, LY6G5B and CSNK2B-LY6G5B chimera double tagged proteins. GAPDH expression is shown as internal control. **C**) Immunofluorescence experiments in permeable (P) and non permeable (NP) conditions showing CSNK2B, LY6G5B and CSNK2B-LY6G5B chimeric protein cellular distribution.

were C-terminally tagged by adding a Hisx6 tag. On the other hand, N-terminal V5 epitope was added upstream the first ATG to CSNK2B and CSNK2B-LY6G5B proteins, but due to the presence of a signal peptide in LY6G5B [13] and in order to tag the mature LY6G5B protein (Figure 4A) the V5 epitope tag was inserted after the signal peptide. Western blot analysis using anti-V5 or anti-PentaHis antibodies showed interesting results. Anti-V5 antibodies showed two close intense bands for *CSNK2B* protein of the estimated size (Figure 4B). These two bands could correspond to post-translational modifications of CSNK2B such as phosphorilation [30]. For *LY6G5B*, also two clear bands were also present, in agreement with previous results [14]. Interestingly, anti-V5 antibodies western-blot analysis showed a single discrete band of the predicted size for *CSNK2B-LY6G5B* protein, indicating lack of post-translational modifications.

On the other hand, anti-PentaHis antibodies showed similar pattern for *CSNK2B* protein to that showed by anti-V5 antibodies, but no signal of *LY6G5B* protein or *CSNK2B-LY6G5B* protein was detected (Figure 4B). This lack of detection on *LY6G5B* could be due to a C-terminal processing cleaving also the Tag. This cleavage signal sequence would also be present in *CSNK2B-LY6G5B* protein and for that also prone to be processed.

Cellular *CSNK2B* protein distribution has been described before in cytoplasm, nuclei, and other organelles [21]. Our results, through immunoflourescent confocal microscopy experiments by using anti-V5 antibodies under permeabilised conditions, showed mainly cytoplasmic cellular *CSNK2B* protein distribution, in agreement with previous data [21]. Under the same conditions, *LY6G5B* showed a protein distribution clearly related with ER pattern, and not extracellular staining, as previously described [14]. Here, for the first time, we show *CSNK2B-LY6G5B* protein distribution, which is quite similar to the one presented by *CSNK2B* and which clearly differs from the one of the *LY6G5B* protein.

Although the *LY6G5B* protein belongs to a GPI-anchored protein family, it has not been found to be

located on the outside of the cellular membrane [14]. In addition, it is known that *CSNK2B* can be exported to the external side of the cellular membrane [31], and *CSNK2B-LY6G5B* presents *CSNK2B* domains needed for its exportation to cell surface and/or its excretion, as well as a mature Ly-6 domain (Figure 3). To know whether *CSNK2B-LY6G5B* could be on the cell surface we carried out two experimental strategies. The first one consisted of immunoflourescent confocal microscopy experiments under non-permeabilised conditions. Our results showed the absence of *CSNK2B*, *LY6G5B* as well as *CSNK2B-LY6G5B* chimeric proteins in the cell surface (Figure 4C), in COS 7 cells. The second one was to test *CSNK2B*, *LY6G5B* and/or *CSNK2B-LY6G5B* proteins presence in supernatant by western-blot experiment. It was not possible to observe expression of these proteins in the supernatant, either when loaded directly on a gel or when TCA-precipitated (data not shown).

Discussion

After the challenge supposed by the human genome sequencing, as well as of other organisms with medical or commercial interest, the determination of the transcriptome of these species is a prerequisite for fully understanding not only their molecular biology, but also for translating this information to technical applications in medicine, pharmacology and biotechnology. In this sense, high throughput technologies supported by systematic cDNA libraries sequencing has been the main approach used for transcriptome characterisation. Thus, sequencing of random transcript cDNA clones that results in short partial sequences, known as expressed sequence tags [32] (EST) or, more recently, methods for full length isolation and sequencing of random clones from cDNA libraries have been used [33,34]. In addition, in the last few years massive sequencing technologies have also been developed for transcriptome analysis [7,24,35]. All these high throughput techniques present the advantage to generate a considerable volume of information in a relative short time, but these techniques seem to be inefficient for discovering relative rare transcripts. This affirmation is supported by recent data that suggest the existence of a wealth of transcripts which had, so far, escaped detection through systematic sequencing of cDNA libraries [36]. In a recent published work [26] combining EST and RNAseq data, it has been showed that chimeras are lowly expressed transcripts. Thus, here we present that nested RT-PCR shows to be an efficient tool to discover a number of transcripts expressed from a concrete locus, not only those highly expressed but also those expressed at lower levels. In fact, 62 of 76 transcripts detected in our experiments had never been described before. This supposes 81.5% of novel sequences, showing how powerful this technique is in order to detect transcripts from a concrete locus. In this respect, our results show that current gene and transcript annotation sets might cover only a small fraction of the total transcriptional output of the human and other organism described genomes, and in this sense, a major enforce should be taken in order to detect more transcripts of a particular organism.

A detailed human and mouse *Ly6g5b* transcript analysis has been previously described by our group [13,16]. However, here we have extended this comparative analysis among different mammalian species, showing that *Ly6g5b* first intron retention is a ubiquitous and important characteristic conserved in mammals. Its apparent capacity to escape NMD together with its conservation in the mammalian group studied points to an important role for this non-coding RNA, which should be investigated. So, we could conclude that *Ly6g5b* gene presents a double expression pattern. The first one, quite similar among tissues and species, is constituted by *Ly6g5b* transcripts with first intron retention event. The second one seems to be tissue and species specific and is constituted by canonical *Ly6g5b* transcripts with a complete Ly-6 ORF as well as aberrant and non conserved transcripts, with no common pattern distribution.

For *Csnk2b* gene expression we can also describe a double expression pattern. The first one constituted by the canonical isoform, which is quite similar among tissues and species, and the second one which seems to be tissue and species specific and is constituted by aberrant and non conserved transcripts, with no common pattern distribution.

CSNK2B-LY6G5B human chimerism was initially described by Calvanesse *et al* (2008) by using different human cell lines [16]. However, here we show the first comparative analysis of this TIC event using RNA from different mammalian tissues and species. Our results show that, far to be a human characteristic, *Csnk2b-Ly6g5b* chimerism is widely conserved in mammals. Its conservation among all the species analysed in this study shows how *Csnk2b-Ly6g5b* chimerism is not a trivial event. The majority of them lack the last exon of the upstream gene as well as the first exon of the downstream gene which is consistent with previous reported data [22,23]. This eliminates the stop codon as well as the molecular targets present in the 3' UTR region of the transcripts of the upstream gene. We also agree with these authors that run-off is the most likely mechanism involved in the origin of TIC, since some chimeric transcripts detected in our study maintain the intergenic region. Other mechanisms proposed for generating chimeric transcripts, like trans-splicing, are not likely to maintain these intergenic regions.

In addition to the canonical *Csnk2b, Ly6g5* and *Csnk2b-Ly6g5b* transcripts with a coherent ORF, other transcripts detected in our study present exon skipping as well as intron retention events which allow to generate, assuming that all these transcripts could be translated into protein

by using canonical start codon, truncated or aberrant proteins. Others are non-coding RNAs. The observation that there are tens of thousands of non-coding RNA (ncRNA) expressed in mammals, and that most of the genome is transcribed, confronts and contradicts the traditional protein-centric view of genetic information and genome organisation [37], [38]. Thus, there are two opposing alternatives either the bulk of the transcription which does not yield mRNAs is 'transcriptional noise' and/or the residue of evolutionary baggage retained or accumulated within genes, or this transcription comprises another level of expression and transaction of RNA information that is important to the evolution and developmental ontogeny of the higher organisms [39]. If one assumes that all this is transcriptional noise and that all these transcripts are the result of transcriptional machinery mistakes while it is working, they should not be distributed in a specific manner.

Some authors defend that chimerism might generate bifunctional proteins having properties from both original proteins [23]. Through different analyses of our results, we could identify chimeras which maintain the ORFs of *Csnk2b* and *Ly6g5b* susceptible to form bi-functional chimeric proteins in *Homo sapiens*, *Macaca mulata*, and *Sus scrofa*. The fact that these were not found in cow, rat and mouse does not indicate functional chimera absence in these species, since they could be present in other tissues not analysed in this study. These hypothetical new chimeric proteins all carry N-terminus domains from *Csnk2b* involved in structural aspects that are required for *Csnk2b* exportation to the cellular surface [21] and/or its regulation [27], as well as Ly-6 domain amino acid sequence [13,14] at their C-terminus. However, the chimera "bi-functional" protein will affect the juxta-dimer interface region [27] containing the zinc-finger involved in the homo-dimerisation, as well as all the C-terminal domain involved in the interaction with the *CSNK2A* subunits and the crucial last 20 amino-acids also involved in the homodimerisation. In addition, the Ly-6 domain, with 10 Cysteins, could not be folded as such due to the intracellular localisation. These two facts indicate that the "bifunctional" chimera would not be formed, and would only have one function: the one of CSNK2B, although possibly binding to other kinase different to CSNK2A due to the alterations produced on the C-terminal domain commented above. Other chimeras would only be affected from exon 7, containing then the juxta-dimer interface region but differing at the very end C-terminal region and not containing the Ly-6 domain sequence due to frameshifts, and probably only affecting the binding to *CSNK2A*. It has been proposed that *CSNK2B* might bind other kinases such as Ras-1 and Mos to modify their catalytic affinity in a CK2-independent fashion. The alternative C-terminal ends, generated by the chimeric transcripts, could increment the binding repertoire of *CSNK2B* to other kinases or non-kinase proteins converting *CSNK2B* in even a wider "wild-card" regulator subunit than previously proposed [27].

In addition, we have found chimeric transcripts that would encode a complete CSNK2B protein, but with different 3´UTR due to the Ly6g5b sequence, in *Macaca mulata*, *Bos taurus* and *Rattus norvegicus*. These transcripts could have different mRNA localisations or stabilities which could have altered protein functional implications [40,41]. It has been described that some 3´UTR contain "localization elements" or "zip codes" which target mRNAs to specific subcellular sites.

Conclusions

Alternative splicing has an important role in *Csnk2b* and *Ly6g5b* gene expression. *Ly6g5b* intron retention and *Csnk2b-Ly6g5b* chimera transcripts are present in many tissues of different mammals. The data and analysis we have performed should serve as a valuable resource for further characterizing the possible functional role of TIC and the mechanisms that affect it.

Methods

Computational analysis

Csnk2b and *Ly6g5b* orthologue ORF sequences *Homo sapiens*, *Macaca mulata*, *Sus scrofa*, *Bos taurus*, *Rattus norvegicus* and *Mus musculus* were obtained from the National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/BLAST/) [42]. Multiple alignments of sequences were performed with ClustalW2 software (http://www.ebi.ac.uk/Tools/msa/clustalw2/) [43,44]. Protein sequences analyses were carried out using Pfam (http://www.sanger.ac.uk/resources/databases/pfam.html) [45], SMART (http://smart.embl-heidelberg.de/) and InterProScan (http://www.ebi.ac.uk/InterProScan/) [46-49] databases. The sequencing results were evaluated using the BLAST algorithm at the NCBI web page and the MultAlin alignment software (http://bioinfo.genotoul.fr/multalin/multalin.html) [50]. Primers were designed using the Primer3 software (www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi).

mRNA extraction and retrotranscription

Homo sapiens, *Macaca mulata*, *Sus scrofa*, *Bos taurus*, *Rattus norvegicus* and *Mus musculus* brain, kidney, liver and lung tissue total RNAs were obtained from BioChain® (USA) http://www.biochain.com through one of their Europe distributor "AMS" http://www.amsbio.com (UK). One μg of total RNA from each tissue was used for oligo-dT primed cDNA synthesis which was performed using the ImProm-II(TM) Reverse Transcription System (Promega) in a 20 μl reaction volume following the manufacturer's instructions.

Table 1 Nested-PCR primers

		CSNK2B		LY6G5B	
		First Round	Second Round	First Round	Second Round
H.sapiens	**F**	HsCSNK2BE1f01	Hs-MamCSNK2BE1-2f01	HsLY6G5BE1f01	HsLY6G5BE1f02
		CGTTCCCTGGAAGTAGCAA	GCTGACGTGAAGATGAGCAG	GAGCATGGTCACAGGAAGGT	CATCTCCCCAGAATTCCAAA
	R	Hs-SsCSNK2BE7r02	Hs-Mam-RnCSNK2BE7r01	HsLY6G5BE3r02	Hs-MamLY6G5BE3r01
		CCCACCACAATAACGACTCC	TCAGCGAATCGTCTTGACTG	CGGAGGCCTAAGAAATCACA	TGTTTTCAGAGAGGGCAGTG
M.mulata	**F**	MamCsnk2bE1iso1f01	Hs-MamCsnk2bE1-2f01	MamLy6g5bE1f01	MamLy6g5bE1f02
		ACCCTCCCCAATTTCCACT	GCTGACGTGAAGATGAGCAG	TCACAGGAAGGTGGGGTTT	CGTCTCCCCAGAATTCCATA
	R	Mam-BtCsnk2bE7r02	Hs-Mam-RnCsnk2bE7r01	MamLy6g5bE3r02	Hs-MamLy6g5bE3r01
		TCCCACCACAATAATGACTCC	TCAGCGAATCGTCTTGACTG	CAACAGAGGGAGGCCTAAGA	TGTTTTCAGAGAGGGCAGTG
S.scrofa	**F**	SsCsnk2bE1f01	SsCsnk2bE1-2f01	SsLy6g5bE1f01	SsLy6g5bE1f02
		TCCTTGGAAGCAGAAACTCC	CGCTGAAGTGAAGATGAGCA	CTCAGGGATCACCCCTCTC	TCACGTGCTCGTAGGTATGC
	R	Hs-SsCsnk2bE7r02	SsCsnk2bE7r01	SsLy6g5bE3r02	SsLy6g5bE3r01
		CCCACCACAATAACGACTCC	GGGAATCAGCGAATTGTCTT	GAATGCTGGGGTCTTAGGG	CCGGGAAGGATGAGATGTTA
B.taurus	**F**	BtCsnk2bE1f01	BtCsnk2bE1-2f01	BtLy6g5bE1f01	BtLy6g5bE1f02
		GAAGCAGAAACTCCCCTTCC	CCGACGTGAAGATGAGCAG	GTCAGAACCACCCTGCAGTT	CTCTCCCCAGAAGTCCATGA
	R	Mam-BtCsnk2bE7r02	BtCsnk2bE7r01	BtLy6g5bE3r02	BtLy6g5bE3r01
		TCCCACCACAATAATGACTCC	GGAATCAGCGAATCGTCTTC	GGGGTGTCAGGAGTGAGATG	AATCTATCTGCGACGGGAAA
R.norvegicus	**F**	Rn-MumCsnk2bE1f01	Rn-MumCsnk2bE1-2f01	RnLy6g5bBE1f01	RnLy6g5bE1f02
		GTTCCTTGGAAGCACAGCTC	CCGCGGACATAAAGATGAGT	AGTGTGATCCCAGGAAGGTG	CTACTCCACGGGAGTTGCTC
	R	RnCsnk2bE7r02	Hs-Mam-RnCsnk2bE7r01	RnLy6g5bE3r02	RnLy6g5bE3r01
		GTGTCACAGGCAGAGGAGGT	TCAGCGAATCGTCTTGACTG	CCATGGAGAGCAGAAGGAAG	CAGAGCAGAATCTGGGAAGG
M.musculus	**F**	Rn-MumCsnk2bE1f01	Rn-MumCsnk2bE1-2f01	MumLy6g5bE1f01	MumLy6g5bE1f02
		GTTCCTTGGAAGCACAGCTC	CCGCGGACATAAAGATGAGT	ACTGCCTGTCAAACCCATTC	TCCCCCAAATTCCATAATGA
	R	MumCsnk2bE7r02	MumCsnk2bE7r01	MumLy6g5bE3r02	MumLy6g5bE3r01
		AGCAGAGGAATGGTGGTGTC	GTGGGCAATCAGCGAATAGT	GAGTGTTCACAGACCGCAGA	GGGGAACACATCAGGGTCTA

Name and sequence (5'-3' orientation). F, forward primer; R, reverse primer.

Nested RT-PCR

In order to detect Csnk2b, Ly6g5b and Csnk2b-Ly6g5b transcripts, first and second rounds of nested RT-PCR were performed with the primers indicated on Table 1 by using GoTaq® Green Master Mix (Promega) in a 20 μl reaction volume. For the first round of PCR 1 μl of cDNA was used in each reaction. Amplification conditions were: 95°C for 5 min followed by 40 cycles of 95°C for 30 s, 60°C for 30 s and 72°C for 90 s followed by 72°C for 10 min. For the second round of PCR, 1 μl from the first round product diluted 1/25 was used in each PCR reaction. Amplification conditions were the same as before but only 30 cycles were used. PCR products were separated by size through electrophoresis, purified from gel (Wizard® SV Gel and PCR Clean-Up System, Promega) and finally cloned into pGEM-T Easy Vector (Promega). Sequencing of PCR products and plasmids were carried out by the sequencing service of the Instituto de Investigaciones Biomédicas "Alberto Sols" (Madrid, Spain. http://www.iib.uam.es).

Expression constructs

The full-length coding sequence of human *CSNK2B*, *LY6G5B* and *CSNK2B-LY6G5B* were amplified from the previously cloned pGEM-T Easy vectors with the specific primers showed on Table 2 by using *Pfu* DNA Polymerase (Promega). Amplification conditions were: 95°C for 5 min followed by 35 cycles of 95°C for 30 s, 60°C for 30 s and 74°C for 3 min followed by 74°C for 10 min. PCR products were cloned in *Bam*HI/*Age*I cloning sites present in pcDNA3.1/V5-His-TOPO plasmid (Invitrogen), removing V5 epitope and in frame with the His tag creating *CSNK2B*-His, *CSNK2B-LY6G5B*-His and exon 2-exon 3 *LY6G5B*-His plasmids. To create a V5Tag-pcDNA3.1/Zeo vector we designed long primers containing the V5 epitope sequence (Table 2). These primers were hybridised (100°C for 4 min followed by 1 min cycles decreasing 0,5°C per cycle from 100°C to 4°C) to create the dsV5 Tag sequence, which was cloned in *NotI*/*Eco*RV cloning sites present in pcDNA3.1/Zeo(-) plasmid (Invitrogen) ending with the V5-Tag-pcDNA3.1/Zeo vector. Then, ATG-Stop *CSNK2B*-His and *CSNK2B-LY6G5B*-His and exon 2-exon 3 *LY6G5B*-His coding sequences were amplified with primers showed on Table 2 by using *Pfu* DNA Polymerase (Promega) under the same PCR conditions as above and cloned in *Bam*HI/*Hin*dIII cloning sites present in the new V5-Tag-pcDNA3.1/Zeo vector in frame with the N-terminal

Table 2 Expression constructs primers

Long primers for V5 epitope cloning in pcDNA3.1/Zeo(–)	
Name	Sequence
MCV5_F01	GGG***GCGGCCGC***ATGGGAAAGCCGATCCCAAACCCTCTATTAGGTCTGGACTCCACC***GGATCC***TGA***GATATC***GGG
MCV5_R01	CCC***GATATC***TCA***GGATCC***GGTGGAGTCCAGACCTAATAGAGGGTTTGGGATCGGCTTTCCCAT***GCGGCCGC***CCC
Primers for CSNK2B, LY6G5B and CSNK2B-LY6G5B Chimera ORFs cloning in pcDNA3.1/V5-His-TOPO	
Name	Sequence
HsCSNK2BORF_F01	CCC***GGATCC***ATGAGCAGCTCAGAGG
HsCSNK2BORF_R01	CCC***ACCGGT***GCGAATCGTCTTGAC
HsLY6G5BORF_F01	CCC***GGATCC***ATGAAGGTCCATATGC
HsLY6G5BORF_R01	CCC***ACCGGT***GGAAGGGTGAGGTGTC
Primers for final construct in pcDNA3.1/Zeo(–)	
Name	Sequence
HsCSNK2BORF_F01	CCC***GGATCC***ATGAGCAGCTCAGAGG
HsLY6G5BORFE2_F02	CCC***GGATCC***GTTCCTGTTCCCGACATC
Hisx6HindIIIR_01	CCC***AAGCTT***TCAATGGTGATGGTGATGATG
HsLY6G5BPS_F01	GGG***TCTAGA***ATGAAGGTCCATATGC
HsLY6G5BPS_R01	CCC***GCGGCCGC***TCTTTCCTACTGTGAA

Restriction enzyme target sequences are in bold and italics.

V5 epitope creating V5-*CSNK2B*-His, V5-*CSNK2B-LY6G5B*-His and V5-exon 2-exon 3 *LY6G5B*-His. Finally, LY6G5B exon 1 was amplified from LY6G5B-His plasmid using the primers showed on Table 2 and *Pfu* DNA Polymerase (Promega) under the same PCR conditions described above and cloned in *Xba*I/*Not*I cloning sites present in V5-exon 2-exon 3 LY6G5B-His plasmid to finally create an exon1-V5-exon 2-exon 3 *LY6G5B*-His vector. Sequencing of PCR products and plasmids were carried out by the sequencing service of the Instituto de Investigaciones Biomédicas "Alberto Sols" (Madrid, Spain. http://www.iib.uam.es).

Transfections and western blot analyses

Transfections of COS-7 cells were carried in 24-well plates, with 0.5 μg of plasmid per well, by using TransIT® COS Transfection kit (Mirus-BioNova) following the manufacturer´s instructions. For Western Blot analyses, two days after transfection, cells were harvested in Laemmli´s SDS sample buffer. Samples were resolved on SDS 12% (w/v) polyacrylamide gels and the proteins were transferred onto nitrocellulose membranes (Protran). After blocking the membrane 30 minutes in PBST (PBS-0,05% (v/v) Tween) containing 5% (w/v) skimmed milk powder (blocking solution), the blot was first incubated for 90 min either in a 1/2000 dilution of mouse anti-Hisx5 (Qiagen), a 1/10000 dilution of mouse anti-V5 (Sigma) or a 1/5000 dilution of mouse anti-GAPDH monoclonal antibodies in blocking solution, washed three times for 10 min with PBST, and then incubated for 30 min in a 1:5000 dilution of peroxidase-conjugated anti-mouse IgG antibody (Sigma) in PBST. Membrane was washed twice for 10 min with PBST followed by a final wash in PBS for 10 min. Bound antibodies were detected by ECL Plus Western Blotting (Amershan). Secreted proteins were TCA-precipitated from the culture media after the addition of 1 volume of TCA to 4 volumes of media, incubated 10 min at 4°C, washed with 200 μl cold acetone, resuspended in Laemmli´s SDS sample buffer, and analysed by SDS-PAGE followed by Western blotting, as described above.

Immunofluorescences

Transfections were performed as described above. Two days after transfection, cells were fixed for 15 min at room temperature with 4% (v/v) paraformaldehyde in PBS. Following fixation, cells were washed three times for 15 min with PBS, treated with ammonium chloride 50 mM for 30 min and permeabilised for 20 min with PBS containing 0.1% (v/v) Triton X-100 (wash solution). Cells were blocked for 30 min in wash solution with 5% (w/v) BSA (Bovine Serum Albumin, blocking solution) and then incubated for 1 h with 1:400 dilution of mouse anti-V5 monoclonal antibody (Sigma) in blocking solution. After that, cells were washed three times for 15 min with wash solution and then incubated with a secondary antibody coupled to Alexa 488 (anti mouse IgG, Invitrogen) in a 1:500 dilution for 1 h in blocking solution. Cells were washed three times for 15 min with wash solution and nucleic acids were stained with To-Pro3 in a 1:500 dilution in blocking solution. Finally, cells were washed three times for 15 min with wash solution, mounted with Prolong Gold Antifade Reagent (Molecular Probes) and photographed with a Leika confocal microscope. Non-permeabilised immunofluorescence experiments were

carried out under the same conditions described above but avoiding Triton X-100 addition.

Additional files

Additional file 1: List of all the nucleotide and protein sequences of the mRNAs described on this article with their corresponding names and accession numbers.

Additional file 2: Table S1. Table that contains all the different spliced isoforms detected in all the species studied in this work indicating whether they were previously described (deposited on databases) and, in the case they were found on databases the access number of the corresponding sequence or of the ESTs are indicated.

Abbreviations

RT-PCR: Reverse transcription polymerase chain reaction; *CDS*: Coding sequence; *AS*: Alternative splicing; *MHC*: Major Histocompatibility Complex; *ORF*: Open reading frame; *PSC*: Premature stop codon; *NMD*: Nonsense mediated decay; *TIC*: Transcription induced chimerism; *EST*: Expressed sequence tag; *cDNA*: Complementary DNA; *mRNA*: Messenger RNA; *ncRNA*: Non-coding RNA; *TCA*: Trichloroacetic acid; *ER*: Endoplasmic reticulum.

Competing interest

The authors declare that they have no competing interests.

Authors' contributions

FH-T and AR designed and carried out the experiments, and analysed and interpreted the data. FH-T drafted the manuscript. BA conceived the study, participated in its design and data interpretation, and completed the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to Fernando Carrasco from the Genomics Service at CBMSO for his technical expertise and advice. This work was supported by grants from the Ministerio de Educación y Ciencia (BFU2005-03683), Ministerio de Ciencia e Innovación (BFU 2008-03126, BFU2009-09117), Comunidad de Madrid (GR/SAL/0670/2004, and 200620 M078) and Fundación Ramón Areces. F Hernández-Torres was funded by Genoma España and A Rastrojo by a postgraduate fellowship (FPU) from the Ministerio de Educación y Ciencia. B Aguado held a Programa Ramón y Cajal (MEC) contract and an Amarouto (Comunidad Madrid-Fundación Severo Ochoa) contract. We acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI). The CBMSO receives an institutional grant from Fundación Ramón Areces.

Author details

[1]Centro de Biología Molecular Severo Ochoa (CBMSO), Consejo Superior de Investigaciones Científicas (CSIC)-Universidad Autónoma de Madrid, Madrid, Spain. [2]Present address: Experimental Biology Department, Universidad de Jaén, Jaén, Spain.

Received: 24 July 2012 Accepted: 13 March 2013
Published: 22 March 2013

References

1. Villate O, Rastrojo A, Lopez-Diez R, Hernandez-Torres F, Aguado B: **Differential splicing, disease and drug targets.** *Infect Disord Drug Targets* 2008, **8**(4):241–251.
2. Irimia M, Blencowe BJ: **Alternative splicing: decoding an expansive regulatory layer.** *Curr Opin Cell Biol* 2012, **24**(3):323–332.
3. Kalsotra A, Cooper TA: **Functional consequences of developmentally regulated alternative splicing.** *Nat Rev Genet* 2011, **12**(10):715–729.
4. Chow LT, Gelinas RE, Broker TR, Roberts RJ: **An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA.** *Cell* 1977, **12**(1):1–8.
5. Gelinas RE, Roberts RJ: **One predominant 5'-undecanucleotide in adenovirus 2 late messenger RNAs.** *Cell* 1977, **11**(3):533–544.
6. Berget SM, Moore C, Sharp PA: **Spliced segments at the 5' terminus of adenovirus 2 late mRNA.** *Proc Natl Acad Sci USA* 1977, **74**(8):3171–3175.
7. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF: *Schroth GP*. Burge CB: Alternative isoform regulation in human tissue transcriptomes. Nature; 2008.
8. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.** *Science* 2003, **302**(5653):2141–2144.
9. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**(6822):860–921.
10. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al*: **The sequence of the human genome.** *Science* 2001, **291**(5507):1304–1351.
11. Xie T, Rowen L, Aguado B, Ahearn ME, Madan A, Qin S, Campbell RD, Hood L: **Analysis of the gene-dense major histocompatibility complex class III region and its comparison to mouse.** *Genome Res* 2003, **13**(12):2621–2636.
12. The MHC sequencing consortium: **Complete sequence and gene map of a human major histocompatibility complex.** *Nature* 1999, **401**(6756):921–923.
13. Mallya M, Campbell RD, Aguado B: **Transcriptional analysis of a novel cluster of LY-6 family members in the human and mouse major histocompatibility complex: five genes with many splice forms.** *Genomics* 2002, **80**(1):113–123.
14. Mallya M, Campbell RD, Aguado B: **Characterization of the five novel Ly-6 superfamily members encoded in the MHC, and detection of cells expressing their potential ligands.** *Protein Sci* 2006, **15**(10):2244–2256.
15. Stroncek DF, Caruccio L, Bettinotti M: **CD177: A member of the Ly-6 gene superfamily involved with neutrophil proliferation and polycythemia vera.** *J Transl Med* 2004, **2**(1):8.
16. Calvanese V, Mallya M, Campbell RD, Aguado B: **Regulation of expression of two LY-6 family genes by intron retention and transcription induced chimerism.** *BMC Mol Biol* 2008, **9**:81.
17. Kim E, Magen A, Ast G: **Different levels of alternative splicing among eukaryotes.** *Nucleic Acids Res* 2007, **35**(1):125–131.
18. Lejeune F, Maquat LE: **Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells.** *Curr Opin Cell Biol* 2005, **17**(3):309–315.
19. Conti E, Izaurralde E: **Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species.** *Curr Opin Cell Biol* 2005, **17**(3):316–325.
20. Jakobi R, Voss H, Pyerin W: **Human phosvitin/casein kinase type II. Molecular cloning and sequencing of full-length cDNA encoding subunit beta.** *Eur J Biochem* 1989, **183**(1):227–233.
21. Rodriguez FA, Contreras C, Bolanos-Garcia V, Allende JE: **Protein kinase CK2 as an ectokinase: the role of the regulatory CK2beta subunit.** *Proc Natl Acad Sci USA* 2008, **105**(15):5693–5698.
22. Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, Antonarakis SE, Guigo R: **Tandem chimerism as a means to increase protein complexity in the human genome.** *Genome Res* 2006, **16**(1):37–44.
23. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, Sorek R: **Transcription-mediated gene fusion in the human genome.** *Genome Res* 2006, **16**(1):30–36.
24. Nacu S, Yuan W, Kan Z, Bhatt D, Rivers CS, Stinson J, Peters BA, Modrusan Z, Jung K, Seshagiri S, *et al*: **Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples.** *BMC Med Genomics* 2011, **4**:11.
25. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, *et al*: **Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions.** *Genome Res* 2007, **17**(6):746–759.
26. Frenkel-Morgenstern M, Lacroix V, Ezkurdia I, Levin Y, Gabashvili A, Prilusky J, Del Pozo A, Tress M, Johnson R, Guigo R, *et al*: **Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts.** *Genome Res* 2012, **22**(7):1231–1242.
27. Bolanos-Garcia VM, Fernandez-Recio J, Allende JE, Blundell TL: **Identifying interaction motifs in CK2beta–a ubiquitous kinase regulatory subunit.** *Trends Biochem Sci* 2006, **31**(12):654–661.
28. Kumar-Sinha C, Kalyana-Sundaram S, Chinnaiyan AM: **SLC45A3-ELK4 chimera in prostate cancer: spotlight on cis-splicing.** *Cancer discovery* 2012, **2**(7):582–585.
29. Zhang Y, Gong M, Yuan H, Park HG, Frierson HF, Li H: **Chimeric transcript generated by cis-splicing of adjacent genes regulates prostate cancer cell proliferation.** *Cancer discovery* 2012, **2**(7):598–607.

30. Ackerman P, Glover CV, Osheroff N: **Stimulation of casein kinase II by epidermal growth factor: relationship between the physiological activity of the kinase and the phosphorylation state of its beta subunit.** *Proc Natl Acad Sci USA* 1990, **87**(2):821–825.
31. Rodriguez F, Allende CC, Allende JE: **Protein kinase casein kinase 2 holoenzyme produced ectopically in human cells can be exported to the external side of the cellular membrane.** *Proc Natl Acad Sci USA* 2005, **102**(13):4718–4723.
32. Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC: **Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library.** *Nat Genet* 1993, **4**(4):373–380.
33. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, *et al*: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309**(5740):1559–1563.
34. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, *et al*: **Functional annotation of a full-length mouse cDNA collection.** *Nature* 2001, **409**(6821):685–690.
35. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL: **Targeted RNA sequencing reveals the deep complexity of the human transcriptome.** *Nat Biotechnol* 2012, **30**(1):99–104.
36. Djebali S, Kapranov P, Foissac S, Lagarde J, Reymond A, Ucla C, Wyss C, Drenkow J, Dumais E, Murray RR, *et al*: **Efficient targeted transcript discovery via array-based normalization of RACE libraries.** *Nat Methods* 2008, **5**(7):629–635.
37. Ecker JR, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, Segal E: **Genomics: ENCODE explained.** *Nature* 2012, **489**(7414):52–55.
38. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS: **lncRNAdb: a reference database for long noncoding RNAs.** *Nucleic Acids Res* 2011, **39**(Database issue):D146–D151.
39. Mattick JS, Makunin IV: **Non-coding RNA.** *Hum Mol Genet* 2006, **15**(Spec No 1):R17–R29.
40. Holt CE, Bullock SL: **Subcellular mRNA localization in animal cells and why it matters.** *Science* 2009, **326**(5957):1212–1216.
41. Martin KC, Ephrussi A: **mRNA localization: gene expression in the spatial dimension.** *Cell* 2009, **136**(4):719–730.
42. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC Bioinformatics* 2009, **10**:421.
43. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, *et al*: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23**(21):2947–2948.
44. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R: **A new bioinformatics analysis tools framework at EMBL-EBI.** *Nucleic Acids Res* 2010, **38**(Web Server issue):W695–W699.
45. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38**(Database issue):D211–D222.
46. Bucher P, Karplus K, Moeri N, Hofmann K: **A flexible motif search technique based on generalized profiles.** *Comput Chem* 1996, **20**(1):3–23.
47. Scordis P, Flower DR, Attwood TK: **FingerPRINTScan: intelligent searching of the PRINTS motif database.** *Bioinformatics* 1999, **15**(10):799–806.
48. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**(9):755–763.
49. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**(17):3389–3402.
50. Corpet F: **Multiple sequence alignment with hierarchical clustering.** *Nucleic Acids Res* 1988, **16**(22):10881–10890.

doi:10.1186/1471-2164-14-199
Cite this article as: Hernández-Torres *et al.*: Intron retention and transcript chimerism conserved across mammals: Ly6g5b and Csnk2b-Ly6g5b as examples. *BMC Genomics* 2013 **14**:199.

Infectious Disorders - Drug Targets **2008**, *8*, 000-000

Differential Splicing, Disease and Drug Targets

O. Villate, A. Rastrojo, R. López-Díez, F. Hernández-Torres and B. Aguado*

Centro de Biología Molecular Severo Ochoa (CBMSO), CSIC, Campus Cantoblanco, 28049 Madrid, Spain.

Abstract: Genome complexity and diversity can be due to Alternative Splicing (AS), a process by which one gene can generate multiple mRNA isoforms and then several proteins. This is part of a normal process of variation on an individual, and when it is disrupted or modified, may trigger disease. To date, there are many pathologies described due to the effects of altered splicing isoforms, and effort is focused on the description of new ones. The design of drug target has to consider splicing, as in many occasions, a drug might have effect on different isoforms, instead of on the particular one implicated in the pathology. Interestingly, the strategies used to alter splicing can be used to modify a form towards the canonical one, or towards an aberrant one, when the latter one has a beneficial effect on the individual. Here we describe differential splicing, diseases produced by alterations on the mRNA isoforms, and drugs or methods used to restore these alterations.

Keywords: Alternative splicing, miRNA, siRNA, mRNA isoform, antisense oligonucleotide, SSOs.

INTRODUCTION

Alternative Splicing (AS) is a major mechanism for modulating the gene expression of an organism, and enables a single gene to increase its coding capacity, allowing from an unique gene the synthesis of several structurally and functionally distinct mRNA and protein isoforms (Fig. (**1**)). This is of great relevance, considering that the number of genes among different species is quite similar. For example, the human genome contains ~25.000-30.000 genes and the *C elegans* genome ~27.000 according to Ensembl (http://www.ensembl.org/) databases. However, the number of transcripts and proteins are quite different with nearly 50.000 reference mRNA sequences on Ensembl human databases and with ~29.000 for *C elegans*. These numbers are expected to be even greater in relation to protein sequences. The result of AS is the introduction of variable segments from particular genes (Fig. (**1**)), within otherwise identical mRNAs. In humans, about 80% of this variability falls within Open Reading Frame (ORFs), greatly expanding the human proteome [1, 2], and the 20% that falls within untranslated regions affects *cis*-elements that control mRNA stability, translation efficiency (including miRNA binding sites), and mRNA localization. Furthermore, one-third of AS events introduce premature termination codons (PTCs), which in many cases cause mRNA degradation by nonsense-mediated decay (NMD) [2, 3]. Therefore, regulation of AS controls the temporal and spatial expression of functionally diverse isoforms, on-off regulation by NMD, or other post-transcriptional regulatory responses. Hence, the knowledge of AS and its regulation is necessary for understanding gene expression in different tissues and species, considering that their mRNA isoforms can also show specificity on function. This mechanism is part of a normal process of variation, generating diversity and complexity genetics and, when this process is disrupted or altered, may trigger disease.

THE SPLICING CODE

In general, control and regulation of splicing are mediated by two molecular elements: *cis* and *trans*-elements. Together, *cis* and *trans* elements make up what is now recognized as the "Splicing Code" resumed on the Fig. (**2**). *Cis*-elements could be defined as small necessary sequences present at mRNA level which can be identified by the splicing machinery (spliceosome) to perform an adequate splicing process. *Trans*-elements include the proteins (spliceosome) which are involved in processes required to identify intron/exon boundaries and catalysis of the cut and paste reactions that remove introns and join exons in a correct order.

Cis-Elements Involved in Splicing

3', 5', and Branch Splice Sites

The typical human gene contains an average of eight exons. Internal exons average 145 nucleotides in length, and introns average more than 10 times this size and can be much larger [4, 5]. Exons are defined by three short and degenerate classical splice-site sequences 3' and 5' splice sites at the intron/exon borders, and the branch splice site within introns (Fig. (**2**)). Components of the basal splicing machinery bind to the classical splice-site sequences and promote assembly of the multicomponent splicing complex known as the spliceosome. These consensus splice sites are relatively easy to identify from alignments of exon–intron boundary sequences (Fig. (**2**)).

Splicing Enhancers and Repressors

In addition to the splice sites, exons are defined by other *cis*-acting regulatory elements, which can be divided into four functional categories: **1)** *Exonic Splicing Enhancers* (ESEs), **2)** *Exonic Splicing Silencers* (ESSs), **3)** *Intronic Splicing Enhancers* (ISEs) also known as intronic activators of splicing (IASs) and **4)** *Intronic Splicing Silencers* (ISSs) (Fig. (**2**)). At exons sequence level, ESEs activate exons recognition and promote their inclusion in mature transcripts, whereas ESSs repress inclusion in mature transcripts [6, 7]. In the other hand, at intron sequence level, ISEs activate inclusion to the adjacent exons [6, 8, 9], whereas ISSs inhibit exon definition by recruit splicing repressors that directly bind and occlude critical *cis*-acting elements of regulated exons [6, 10] or by recruiting repressors to binding sites that flank regulated exons creating a zone of silencing [6, 10].

*Address correspondence to this author at the Centro de Biología Molecular Severo Ochoa (CBMSO), CSIC, Campus Cantoblanco, 28049 Madrid, Spain; E-mail: baguado@cbm.uam.es

1871-5265/08 $55.00+.00

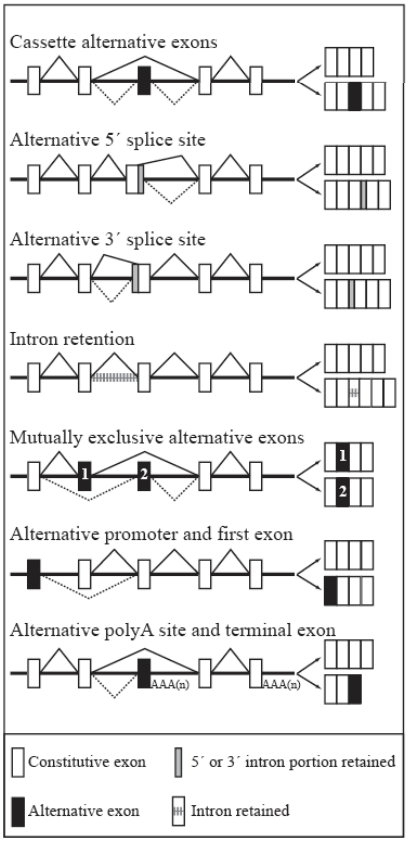


Fig. (1). Different types of possible Alternative Splicing processes. Modified from Blencowe, B. J. 2006 [74].

***Trans*-Elements Involved in Splicing**

The spliceosome is composed of five small nuclear ribonucleoproteins (snRNPs) and more than other 150 proteins, including kinases, phosphatases and helicases, many of which are required for spliceosomal function, as well as associated proteins such as mRNA-export factors and transcription factors [2, 11]. The branch and 5′ splice sites (Fig. (**2**)) serve as binding places for the RNA components of U2 and U1 small nuclear ribonucleoproteins (snRNPs), respectively. This RNA:RNA base pairing determines the precise joining of exons at the correct nucleotides. Exons and introns contain diverse sets of enhancer and suppressor elements that refine *bona fide* exon recognition. At exon level, some ESEs bind SR proteins and recruit and stabilize binding of spliceosome components such as U2AF, other ESEs bind different splicing enhancer proteins which contribute to the correct splicing process. In addition, some ESSs bind protein components of heterogeneous nuclear ribonucleoproteins (hnRNP) to repress exon usage. At intron level, some ISEs and ISS bind auxiliary splicing factors that are not normally associated with the spliceosome to regulate AS (Fig. (**2**)).

SPLICING CODE ALTERATIONS AND DISEASE

The mechanisms causing altered splicing can involve disruption of either *cis*-elements, within the affected gene, or *trans*-elements that are required for normal splicing or splicing regulation. The distinction between *cis*- and *trans*-elements effects has important mechanistic implications. Effects in *cis* have a direct impact on the expression of only one gene, whereas effects in *trans* have the potential to affect the expression of multiple genes. Therefore, a failure or deregulation of either *trans* or *cis*-elements could flow in a pathology.

***Cis*-Elements Alterations**

Cis-elements mutations can affect non-coding and coding regions. Mutations located in non-coding regions, such those affecting 5′ and 3′ splice sites, branch sites or polyadenylation signals, are frequently the cause of hereditary disease. Mutations on the ORF of a single gene can generate synonymous, nonsense or missense mutations.

Synonymous single-nucleotide polymorphisms (SNPs) located in coding regions (cSNPs), although seemingly

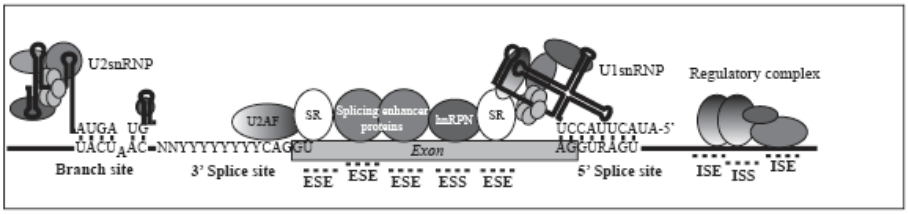


Fig. (2). The Splicing Code. *Cis*-elements sequences (Branch site, 3' and 5' splice sites, ESEs, ESSs, ISEs and ISSs) within and around introns and exons are required for recognition and regulation by *trans*-elements (spliceosome). Modified from Wang, G. S. *et al* 2007 [2].

translationally silent, could have a profound influence on AS. In fact, cSNPs can disrupt (or eventually create) ESE's and ESS's; create new splice sites or strengthen cryptic ones; alter pre-mRNA secondary structures important for exon-definition; and, conceivably, modify the pausing architecture of a gene, causing changes in RNAPol II processivity [12, 13], which might in turn affect splice site choice.

These defects are not exclusive of cSNPs: nonsense and missense mutations as well as exonic deletions or insertions can affect AS in similar ways. In fact, it has recently been proposed that 60% of mutations that cause disease do so by disrupting the splicing code [2, 14], rather than by the predicted disruption of the protein reading frame. Table (**1**) summarizes examples of hereditary disorders caused by exonic point mutations that affect AS. Nonsense mutations may provoke premature termination codons, which in the case of being early on the ORF, this mRNA normally is targeted for degradation by NMD, which involves an mRNA quality-control step [12, 15, 16], disabling the mutated mRNA. Missense mutations generate an amino-acid change, which might have functional relevant implications, such as different protein localization, substrate affinity or ligand binding.

Trans-Elements Alterations

Mutations and non-genetics alterations of factors required for splicing have been implicated in the pathophysiology of human disease [5, 6]. In fact, they seem to be

Table 1. Diseases Due to Aberrant Splicing

Disease		OMIN number	Elements alterations	Comments	Reference
Acute intermittent porphyria		#176000	*cis*-element	Nucleotide substitutions (C to G) in exon 3 of porphobilinogen deaminase	[75]
β-thalassemia		#603902	*cis*-element	Mutations in the intron 2 of *β-globin* that generates a cryptic 3′ acceptor site	[55]
Bilateral periventricular nodular heterotopia (BPNH)		#300049	*cis*-element	Missense mutation in exon 6 (G to C) of *Filamin A* (*FLNA*) gene	[76]
Cancer	Breast and ovarian cancer	+113705	*cis*-element	Nonsense mutation of *BRCA1* in exon 18 (G to T)	[41]
	Leukemias and sarcomas	*605221	*trans*-element	Interaction of FUS-interacting protein 1 (FUSIP1/SRp38/SRRp40) with FUS is disrupted	[77]
	Liposarcomas, acute myeloid leukemia (AML)	*137070	*trans*-element	Translocation of *FUS* gene (*TLS*)	[78, 79]
	Hepatocellular carcinoma	*604739	*trans*-element	Autoantibodies to the splicing factor HCC1	[80]
	Papillary renal cell carcinoma	*605199	*trans*-element	Translocation of *SFPQ* (splicing factor proline- and glutamine-rich) and fused with the *TFE3* gene.	[81]
Cystic fibrosis (CF)		#219700	*cis*-element	Four nonsense mutations of *CFTR* gene in exons 3 (G to U, and C to T), 11 (C to T) and 20 (G to A)	[24]
Fanconi anemia (FA)		#227650	*cis*-element	Nonsense mutations of *FANCG* in exon 8 (C to T)	[82]
Hemophilia A		+306700	*cis*-element	Nonsense mutations of *Factor VIII* in exons 19 (G to T) and 22 (C to T)	[24]
Metachromatic leukodystrophy		#250100	*cis*-element	Missense mutation of *Arylsulfatase A* in exon 8 (C to T)	[83]
Myotonic dystrophy 1 (DM1)		#160900	*trans*-element	Sequestration of CUG binding proteins (CUGBP1) by trinucleotide repeat disorders (CUG)n>50 in *DMPK* gene.	[84]
			trans-element	Sequestration of MBNL1 (MBNL) proteins	[85, 86]
Retinitis pigmentosa		#268000	*trans*-element	More than thirty conserve mutations in genes involved in snRNP function	[28, 29]
Spinal muscle atrophy (SMA)		#253300	*cis*-element	Deletion complete of *SMN1*	[35]
			cis-element	Nonsense mutations of *SMN1* in exon 3 (G to A)	[87]
			cis-element	Nucleotide substitutions (C to T) in *SMN2* disrupts an ESE in exon 7 that causes exon skipping	[39]

implicated in a high range of pathologies, including cancer, blindness, and muscular distrophies [6]. *Trans*-elements splicing mutations can affect the function of the basal splicing machinery or factors that regulate AS. Mutations that affect the basal splicing machinery have the potential to affect splicing of all pre-mRNAs, whereas mutations that affect a regulator of AS will affect only the subset of pre-mRNAs that are targets of the regulator. Table (**1**) summarizes examples of splicing *trans*-element factors associated with human diseases.

Pathologies Due to AS Mis-Regulation

There are an increasing number of diseases generated by alterations on the splicing code, and many more are expected to be described as we increase our knowledge on the different AS isoforms of the human genome. Some of the better described are:

Myotonic Dystrophy

Myotonic Dystrophy Type 1 (DM1) is the most common form of adult muscular dystrophy. It is an autosomal dominant neuromuscular disease associated with CTG repeat expansion in the 3' untranslated region of the DM protein kinase (*DMPK*) gene. A key molecular feature of DM1 is the misregulation of developmentally regulated AS for a subset of genes, such that embryonic and neonatal splicing patterns are retained in adult myotonic dystrophy tissues [2, 17-19]. Disease symptoms such as myotonia and insulin resistance result from the inappropriate expression of embryonic proteins in adult tissues. This is one of the best examples of disease caused by alteration of *trans*-elements which are involved in altered splicing of many other non-related genes. The expanded CUG-RNA disrupts normal postnatal AS transitions that are regulated by two families of proteins: the MBNL and CELF families. The best characterized members of these families are MBNL1 and CUGBP1, which were first identified as CUG-repeat RNA binding proteins. For the genes tested, these two proteins regulate the same splicing events antagonistically by binding to separate regulatory elements. The mechanism of normal postnatal transitions seems to involve a loss of nuclear CUGBP1 owing to decreased protein expression [2, 20, 21], and a gain of nuclear MBNL1 activity owing to translocation from the cytoplasm [2, 20]. The activities of both proteins are disrupted in DM1 by the toxic RNA. Nuclear MBNL1 is depleted as a result of its sequestration into RNA foci, whereas CUGBP1 steady-state levels increase owing to phosphorylation and increased protein half-life [2, 22]. Mouse models support a primary role for loss of MBNL1 function as well as a gain of CUGBP1 function in the splicing abnormalities.

Cystic Fibrosis

Cystic fibrosis (also known as CF, Mucoviscoidosis, or Mucoviscidosis) is a hereditary disease that affects the exocrine (mucus) glands of the lungs, liver, pancreas, and intestines, causing progressive disability due to multisystem failure abnormalities. CF is one of the most common life-shortening, childhood-onset inherited diseases. In the United States, 1 in 3,900 children is born with CF. It is most common among Europeans and Ashkenazi Jews; where one in twenty-two people of European descents are carriers of one gene for CF, making it the most common genetic disease in these populations. The disease is linked with the disruption of the Cystic Fibrosis Transmembrane Conductance Regulator (*CFTR*) gene. This gene encompasses approximately 180,000 base pairs on the long arm of chromosome 7. The most common mutation is a deletion (ΔF508) of three nucleotides that results in a loss of the amino acid phenylalanine (F) at the 508th position on the protein [23]. However, currently more than 1000 disease-associated mutations in *CFTR* gene have been described in the coding sequence, messenger RNA splice signals, and other regions. Focusing our attention in those mutations that affect RNA splice signals, scientific findings have unraveled the presence of SNPs that affect *cis*-elements involved in splicing all over the *CFTR* gene. In fact, there are examples of mutations that affect ESEs elements in exons 3, 11 and 20 [24] and examples of mutations that affect intronic elements, as is the case of mutation 3849+10 kb C3T mutation in intron 19, which lead to inclusion of a cryptic exon of 84-nucleotides [25].

Retinitis Pigmentosa

Retinitis pigmentosa (RP) is one of the most common forms of inherited retinal degeneration [26] affecting 1 in 4000 people worldwide. This disorder is characterized by the progressive loss of photoreceptor cells and may eventually lead to blindness [27]. RP is caused by conserve mutations (more than thirty) in genes involved in snRNP function [2, 28, 29] such as HPRP3, PRPF31 and PRPC8, which encode proteins required for proper assembly and function of the U4-U5-U6-snRNP of the spliceosome component [30-32]. The failure of this component generates a fault in *trans*-elements of the splicing phenomenon.

Spinal Muscular Atrophy

Spinal muscular atrophy (SMA) is an autosomal recessive neuromuscular disorder characterized by degeneration of the anterior horn cells of the spinal cord motor neurons leading to symmetrical muscle weakness and atrophy. SMA is the second most common lethal autosomal recessive disease in Caucasians after cystic fibrosis [33], specially on childhood. This disease is caused by homozygous disruption of the survival motor neuron 1 (*SMN1*) gene by deletion, conversion, or mutation [34]. *SMN1* gen product is involved in snRNP assembly. SMA occurs due to the complete deletion of *SMN1* in 96% of cases [33, 35], because the deletion of this gene prevents assembly of the U1 RNP complexes in the cytoplasm the results is a global defects in the binding of *trans*-elements to the pre-mRNA molecule [36], and loss of snRNP production has been directly linked with this disease [37].

The duplicated gene *SMN2* is transcribed when *SMN1* is deleted but the *SMN2* gene does not completely compensate for the loss of *SMN1* function. *SMN2* may contain as well nucleotide substitutions which do not alter the protein coding sequence although one of the nucleotide substitutions (C to U) disrupts an ESE in exon 7 that causes a exon skipped in the majority of SMN2 mRNAs [38, 39]. As a result SMN2 mRNA encodes a truncated protein missing the C-terminal 16 residues with the consequent loss of functionality. Thus,

low but some expression of *SMN2* generates a less severe disease.

Cancer

The development of many cancer diseases is associated with splicing alterations. Cancer-specific alterations in splice site selection affect genes controlling cellular proliferation (e.g., *FGFR2, p53, MDM2, FHIT*, and *BRCA1*), cellular adhesion, invasion (e.g., *CD44, Ron*) angiogenesis (e.g., *VEGF*), and apoptosis (e.g., *Fas*, *Bcl-x*, and *caspase-2*). The development or progression of cancer can be attributed to *cis*-element mutations within a single gene or *trans*-element mutations that affect most gene encoding. *Cis*-element mutations that affect the splicing of proto-oncogenes, tumour suppressors and DNA repair genes can have multiple roles in cancer initiation and progression [40], for example a *BRCA1* nonsense mutation causes exon skipping by the loss of a functional ESE element [41]. But most cancer associated splicing alterations affect *trans*-elements. These splicing factors most commonly associated with cancer belong to the SR protein family, as a result of changes in SR protein phosphorylation [42, 43]. Other examples of *trans*-element factors that affect AS in cancer include ASF/SF2 and PTB.

SPLICING AND DRUGS

Because alterations in RNA splicing can cause many different diseases [5], characterization of these splice specific alterations can provide new therapeutic targets. Information on AS has been accumulated at a rapid rate during the last years, but the core drug discovery processes still entail techniques that cannot distinguish between splice variants. If the phenomenon of AS is ignored, drug discovery process is exposed to only a fraction of the actual proteomic world and therefore misses many potential protein targets [44].

AS has big impact on drug development and on diagnostic applications. Splice variants may have a different function due to different regulatory properties and/or structural changes that create new domains. Moreover, soluble variants with therapeutic or disease-related functions may be naturally occurring in specific tissues, so they may be candidates for drug targets. In this context, a drug may target various splice variants causing side effects, so an effective drug is needed to target specifically the splice variant of interest. The recognition of the importance of splicing has proved that splicing reactions are potential therapeutic targets.

Mutations causing human diseases may affect splice sites as well as regulatory sequences leading to the production of defective or altered proteins [6]. Thus, targeting either the mutated sequences or the factors that bind them may prove to be a valuable strategy to correct aberrant splicing, and many different approaches, from conventional small-molecules drugs to RNA-based gene therapy have been used. RNA-based strategies offer a series of novel therapeutic applications, including altered processing of the target pre-mRNA transcript, reprogramming of genetic defects through mRNA repair, and the targeted silencing of allele- or isoform-specific gene transcripts.

Targeting Protein Isoforms

The first approach in the therapy of a particular aberrant splicing disease is the specific inhibition/blockage of the altered protein with a specific drug. An example of this approach can be cyclo-oxigenases (COX) enzymes which function is the catalyses of the main reaction of prostaglandin synthesis. There are two genes encoding COX enzymes (*COX1* and *COX2*), and it has been recently discovered that these genes are able to carry out AS [45-48]. *COX1* was considered a constitutive gene which product is related to the synthesis of physiologically relevant prostanoids such as those that regulate the stomach mucosa and platelets aggregation. By contrast *COX2* was thought to be an inducible gene in response to inflammation, fever or injury [49-51]. Non-steroidal anti-inflammatory drugs (NSAIDs), such as Phenacetin, have been largely used to trait the inflammation, pain and fever. These drugs are thought to inhibit the two well known products of *COX1* and *COX2* genes and scientists are trying to improve these drugs to reduce secondary effects due to their broad range of action. Therefore, the discovery of various alternatively spliced products of these two genes open a new point of view in the use of specific treatments for pain/inflammation. Now, it is possible to study the implication of each splice variant in specific pain pathways and thus it will be possible to inhibit a specific kind of pain avoiding undesired side effects. For example, COX-3, a recent discovered isoform produced by AS (intron 1 reteined) of *COX1* gene, is thought to be implicated in the control of fever, because of its brain distribution and the inhibition by some NSAIDs fever specific (Fig. (**3**)).

Another therapy strategy to correct aberrant splice variants is the use of specific antibodies against proteins regions which can be controlled by AS due to exon insertion/exclusion or intron retention leading to the modification of some domains in the final protein. This particular domain or region can be useful in the design of antibodies which can discriminate one splice variant from the others. For example, the human *CD44* gene encodes type 1 transmembrane glycoproteins involved in cell to cell and cell to matrix interactions. This gene undergoes AS generating at least 20 different proteins that may suffer many post-translational modifications like glycosylation (N- or O-glycosylation) and phosphorylation. Some CD44 isoforms decorated with heparin sulphate side chains bind growth factors and can promote growth factor receptor-mediated signaling [52, 53]. Of special interest is the splice variant CD44v6 (generated by exon inclusion) which is over-expressed in many different tumors. The use of specific antibodies against the domain encoded by exon v6 in combination with radionuclides (e.g. 186 Re) made radiotherapy more specific in killing tumor cells disabling side-effects (Fig. (**3**)) [52-54].

Targeting Specific mRNA

RNA targeting is emerging as a powerful alternative to conventional DNA gene replacement therapies for the treatment of genetic disorders. The potential of such approaches ranges from elimination of the mRNA in

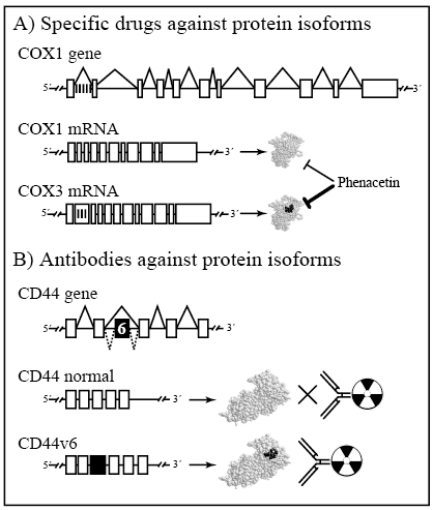


Fig. (3). **Therapeutic approaches against protein isoforms**. A) COX3 enzyme is generated by AS of the *COX1* gene due to intron 1- retention. B) CD44v6 protein is a splice variant that contain a specific domain (v6). Specific antibodies linked with radionuclides against this specific domain may kill tumour cells that over-expressed this protein.

question, to modification of the mature mRNA product by the removal or addition of natural elements or exons, and to repair the mRNA transcript by the addition of foreign mRNA elements to create a chimeric gene product [55]. RNA interference (RNAi) and antisense oligonucleotides, which depend on RNase-H mechanism, induce the degradation of mRNA, whereas steric-blocker oligonucleotides physically prevent or inhibit the progression of splicing or the translation machinery.

RNA Interference (RNAi)

RNAi has become a powerful tool in functional and medical genomic research through directed post-transcriptional gene silencing. The discovery that 21-23 nucleotide RNA duplexes, known as small interfering RNAs (siRNAs), can knockdown the homologous mRNAs in mammalian cells has revolutionized many aspects of drug discovery including down-regulation of disease-associated splicing isoforms. In addition, RNAi-mediated silencing of splicing regulators has the potential to define the complex network of AS regulation and to analyze gene function [56].

RNAi has been used for targeting disease-linked splicing isoforms. For example, in the case of *Bcl-x*, a member of *Bcl-2* gene family, that undergoes AS to generate two isoforms, Bcl-xL and Bcl-xS (Fig. (**4**)). Bcl-xL is antiapoptotic while forced over-expression of Bcl-xS sensitizes cells to a variety of antineoplastic agents and radiation. Bcl-xL

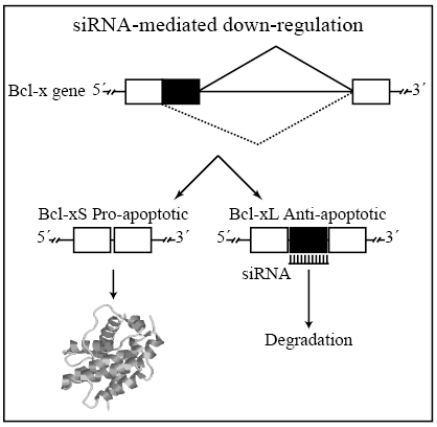


Fig. (4). RNA interference (RNAi)-mediated down-regulation of a splicing isoform. *Bcl-x* undergoes AS to generate two major isoforms, Bcl-xL (anti-apoptotic) and Bcl-xS (pro-apoptotic). A siRNA which recognizes the specific sequence of Bcl-xL induces that RNA to degradation.

specific small interfering RNA down-regulates Bcl-xL protein and inhibit the proliferation of 5-fluorouracil and tumor necrosis factor-related apoptosis-inducing ligand (TRAIL)-resistant cells [57].

Antisense Oligonucleotides

Traditionally, antisense oligonucleotides have been employed to down-regulate gene transcription. On the basis of mechanism of action, two classes of antisense oligonucleotide can be discerned: (*a*) the RNase H-dependent oligonucleotides, which induce the degradation of mRNA; and (*b*) the steric-blocker oligonucleotides, which physically prevent or inhibit the progression of splicing or the translational machinery [58]. Oligonucleotide-assisted RNase H-dependent reduction of targeted RNA expression can be quite efficient, reaching 80-95% down-regulation of protein and mRNA expression. In contrast to the steric-blocker oligonucleotides, RNase H-dependent oligonucleotides can inhibit protein expression when targeted to virtually any region of the mRNA. Most steric-blocker oligonucleotides are efficient only when targeted to the 5'- or AUG initiation codon region. Steric blockade of translation can be demonstrated by the arrest of the polypeptide chain elongation, as shown by Dias *et al.* in 1999 [58]. The optimal use of antisense oligonucleotides in the treatment of disease requires the resolution of problems relating to the effective design and efficient target delivery.

Reprogramming the Splicing

The reprogramming of splicing is a new form of therapy that modifies mRNA without directly changing the sequence of the gene. This strategy can be sub-divided into three different approaches: drugs against splicing regulators,

modifying the processing of the mRNA, and altering mRNA sequence (*trans*-splicing).

Drugs Against Splicing Regulators

The use of small-molecule drug therapy is an attractive approach to modifying splicing patterns because of the relative ease of delivery and dosage control [2]. The most common regulators are the serine-arginine-rich RNA binding proteins (SR proteins) and the heterogeneous nuclear ribonucleoproteins (hnRNPs). They modulate splice-site choice by interacting with components of the splicing machinery and binding to the exonic and intronic *cis*-element signals [59]. The phosphorylation status of SR proteins affects their RNA binding specificity, protein-protein interactions and intracellular distribution, so small molecules that affect the activities of these enzymes can be used to alter splicing patterns (Fig. (**5**)) [2].

Numerous studies have reported several approaches allowing correction of aberrant splicing events by targeting the splicing regulators whose binding is affected by the mutation. Blanchette and colleagues [60] tackled the identification of targets of four splicing regulators in *D melanogaster* with a splicing-sensitive microarray, while Soret *et al*. [61] screened for chemical compounds that directly bind to SR proteins in human cells and interfere with spliceosomal assembly.

Members of the SR protein family are thought to play a major role in the regulation of HIV-1 pre-mRNA splicing. To express key viral proteins, HIV-1 uses a combination of several alternative 5´and 3´splice sites to generate more than 40 different mRNAs. The choice of these sites depends on specific interaction between HIV pre-mRNA sequences and *trans*-element factors, SR proteins and hnRNPs [62]. Bakkour *et al* [63] showed that an indole derivative (IDC16) that interferes with exonic splicing enhancer activity of the SR protein splicing factor SF2/ASF suppresses the production of key viral proteins, thereby compromising subsequent synthesis of full-length HIV-1 pre-mRNA and assembly of infectious particles. IDC16 can efficiently block HIV-1 viral production in peripheral blood mononuclear cells, or PBMCs, or macrophages infected with different laboratory strains or clinical isolates from patients resistant to anti-HIV multitherapies.

Modifying the Splice-Process of the Pre-mRNA

By using modified DNA or RNA oligonucleotides it is possible to alter an exon skipping/inclusion caused by a silent mutation, which has altered the consensus splice site-sequences or the enhancers/inhibitors-sequences, thus guiding the spliceosome to the right splice variant [64]. These oligonucleotides have been named Splice-Switching Oligonucleotides (SSOs). Unlike antisense down-regulation of gene expression via RNAse H or RNA interference degradation pathways, SSOs modulate AS of targeted pre-mRNA, up-regulating expression of desirable protein isoforms, while down-regulating undesirable isoforms. SSOs that block aberrant splice sites can restore normal splicing, whereas those targeting alternative splice sites can switch splicing patterns from detrimental to beneficial isoforms or produce non-functional mRNAs that lead to gene knockdown [65].

Exon skipping is an approach that uses SSOs to modulate splicing by hiding specific sites essential for exon inclusion from the splicing machinery. An example of this approach is the induction of the expression of the recent discovered splice variant of the TNF receptor 2 gene (*TNFR2*). In inflammatory diseases like collagen-induced arthritis (CIA) or TNF-α induced hepatitis has been discovered a weak over-expression of the novel soluble-splice variant of the *TNFR2* lacking the exon 7 of the normal receptor. Therefore, the soluble protein can sequester some of the TNF-α responsible of many of the symptoms of these pathologies [66]. Graziewicz *et al* [66] have developed an efficient SSO capable of induce exon 7 skipping of *TNFR2* and therefore increasing the amount of the soluble receptor that can block TNF-α signal (Fig. (**6A**)). These SSOs were much more effective in reducing TNF-α effect than the drugs commonly used for the treatment of such inflammatory disease.

Actually, SSO exon skipping is currently one of the most promising therapeutic approaches for Duchenne muscular

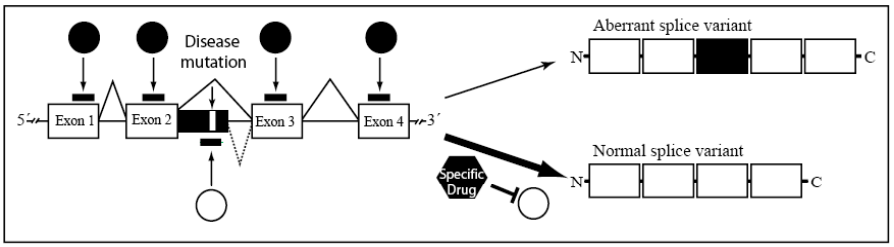


Fig. (5). Rescuing aberrant splicing with small-molecule drugs. A gene that is normally spliced with four exons is represented. The filled circles represent splicing regulators that act at the sites depicted by the black bars to promote splicing. The filled rectangle represents part of an intron with a mutation (white vertical line) creating a binding site for a regulator (white circle) which activates a cryptic 5' splice site, leading to the splicing of an additional sequence into the final mRNA and the production of a defective protein. The therapy consists of a specific drug that abolishes binding of the new regulator and restores normal splicing.

dystrophy (DMD)[67]. The disease is caused by mutations in the *DMD* gene that abolish the production of functional dystrophin. *DMD* deletions and duplications mainly occur in two hot spot regions, the major hot spot region (involving exon 45 to exon 53) and the minor hot spot region (between exon 2 and exon 20). The most notable example is exon 51 skipping. In DMD the open reading frame is disrupted, by deletion of exons 48-50 (the most common mutation), resulting in a premature stop codon and a truncated dystrophin. Specific SSOs hybridize to exon 51 and hide this exon to the splicing machinery, resulting in the splicing of exon 51 with its flanking intron. In this case, the approach restores the ORF generating a shorter but functional dystrophin protein. A successful first-in-man trial has recently been completed [68-70].

Using this approach it is also possible to block a cryptic splicing promote by an intron mutation. This is the case of β-thalassemia; *β-globin* gene has a mutation in the intron 2 and because of this the spliceosome use a cryptic 3' acceptor that produces the inclusion of a new exon [55]. SSOs can modify this wrong splicing blocking the cryptic acceptor sequence and therefore inducing the spliceosome to make the right splicing (Fig. (**6B**)) [55, 71]. In this way, it has also been recently described exon inclusion using this strategy. This is the case of SMA caused by a deletion of the *SMN1* gene. The severity of SMA is compensated for the expression of its paralogous gene *SMN2*. In many cases, *SMN2* gene has a mutation in an ESE element that produces a SMN2 mRNA lacking exon 7 and therefore a truncated protein. Using SSOs that masks the mutated ESE of SMN2 pre-mRNA, and introducing a consensus one, the recovery of the normal splice has been demonstrated in cell lines and patient-derived cells (Fig. (**6C**)).

Altering mRNA Sequence Trans-Splicing

Another abnormality on mRNA splicing, apart of mutations in *cis* and *trans*-elements is the phenomenon known as *trans*-splicing. *Trans*-splicing is a natural process, although rare in mammals, which involves splicing between two separately transcribed mRNAs such that a composite transcript is produced. Manipulation of this process offers the potential for induction of isoform switching or the correction of mutations by conversion to a wild gene product. There are two common methodologies, spliceosome mediated RNA *trans*-splicing (SMaRT) and ribozyme mediated *trans*-splicing (Fig. (**7**)) [55].

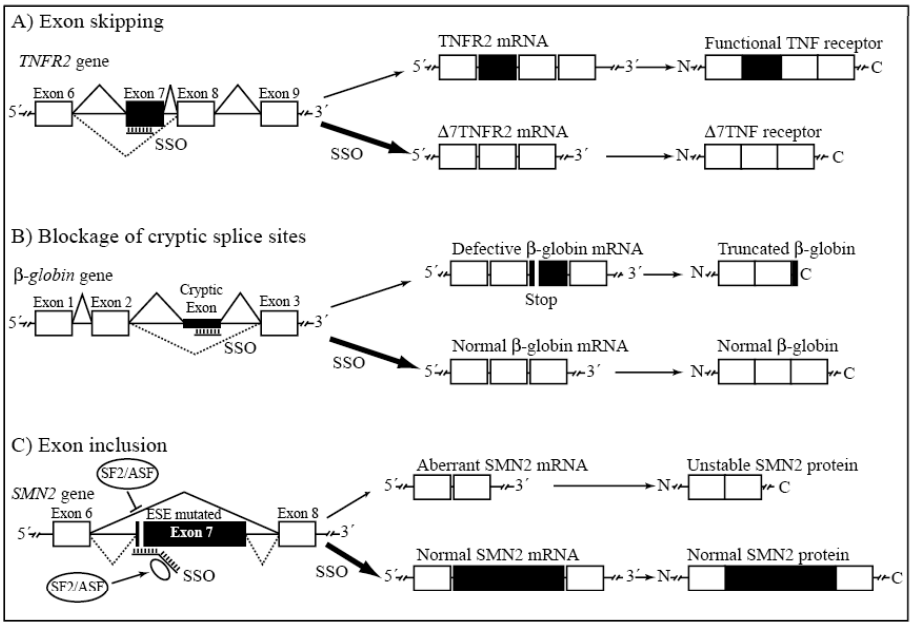


Fig. (6). SSOs methods to restore normal or desired splicing. A) TNFR2 exon 7 skipping to over-express the natural soluble isoform of the receptor which reduce TNF-α level in inflammatory sites. B) Blockage of the cryptic exon created by an intron-mutation in *β-globin* gene using a SSO that restores the normal splice variant. C) The excluded exon 7 of the *SMN2* gene may be included using a SSO that carries the ESE-consensus sequence for SF2/ASF protein.

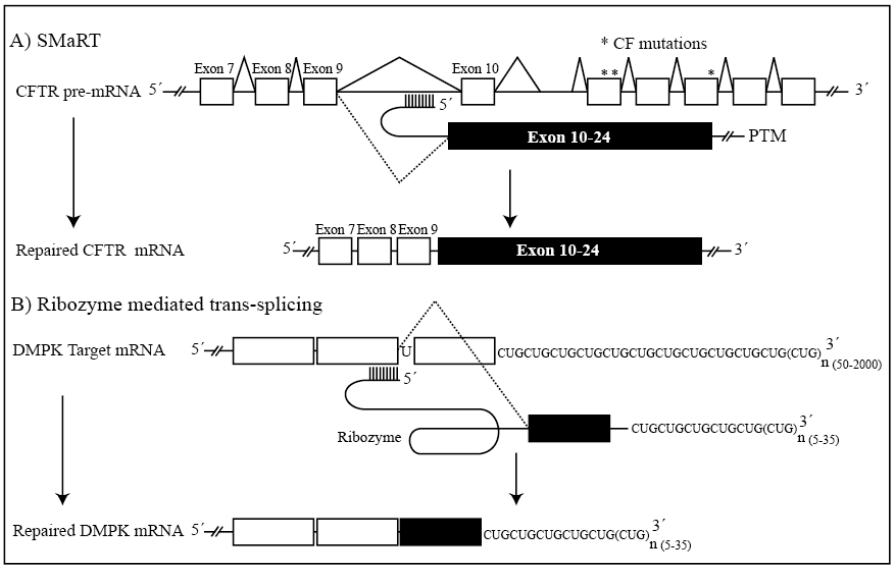


Fig. (7). RNA *Trans*-splicing. (A) Correction of CF mutations in the *CFTR* gene using SMaRT. A PTM containing a binding domain, splicing domain and a coding domain incorporating exons 10-24 of wild-type *CFTR* mRNA binds to intron 9 of *CFTR* pre-mRNA containing disease-causing mutations (stars). SMaRT removes the mutant pre-mRNA so the reprogrammed transcript allows synthesis of a functional protein. (B) Ribozyme-mediated *trans*-splicing and its application to correct trinucleotide repeat expansions in myotonic dystrophy. Ribozymes containing a reduced number of CUG repeats are targeted to the mutant *DMPK* transcript. Binding of the ribozyme allows *trans*-splicing and smaller CTG repeat expansion produce a non-toxic *DMPK* mRNA transcript.

SMaRT: An engineered pre-mRNA *trans*-splicing molecule (PTM) binds to target pre-mRNA in the nucleus such that it triggers *trans*-splicing in a process mediated by the spliceosome. The PTM has a 5´ binding sequence specific for the target, a splicing region that contains motifs necessary for the *trans*-splicing reaction to occur and a coding domain that includes the new or modified genetic information that will reprogram the target (Fig. (**7**)). Functional correction using spliceosome-mediated *trans*-splicing has been reported in several preclinical disease models, including cystic fibrosis, haemophilia A and X-linked immunodeficiency [55, 72]. SMaRT has several advantages over conventional gene therapy. As the gene is repaired rather than introduced, the spatial and temporal expression of the gene should be controlled by endogenous regulation. Other advantage is that PTM constructs are easily accommodated in current vector systems. As repair will only occur where the target transcript is expressed, adverse effects would not be anticipated in cells that were nonspecifically targeted during delivery. The main disadvantage is that a single PTM, in most cases, would not be able to address all the mutations in an affected population.

Ribozyme-mediated trans-splicing: Ribozyme-mediated *trans*-splicing consists of a 5´ guide sequence complementary to the target sequence, the ribozyme domain, and a 3´terminal exon that is to be *trans*-spliced (Fig. (**7**)). Following binding, the ribozyme catalyses *trans*-splicing between the 3´ exons of the ribozyme and the 5´ target mRNA. In DM1, increased levels of trinucleotide repeat CUG expansion in the 3´untranslated region of the dystrophia myotonica-protein kinase (*DMPK*) gene, are responsible for the clinical condition. A specifically designed ribozyme was used to reduce the lenght of expansions from high to low CUG repeats, at the 3´end of DMPK transcripts [73]. As well as correcting disease-causing mutations, *trans*-splicing ribozymes have the potential to create chimeric gene transcripts by splicing foreign cDNA to a targeted mRNA [55].

ACKNOWLEDGEMENTS

The laboratory is supported by grants from the Ministerio de Educación y Ciencia (Plan Nacional BFU2005-03683), the Comunidad de Madrid, the Fundación Ramón Areces and Genoma España. We also thank the Fundación Ramón Areces for Departmental support. BA holds a Ramón y Cajal Programme Fellowship and OV a FPI studentship from the Comunidad de Madrid.

REFERENCES

[1] Modrek, B.; Lee, C. *Nat. Genet.*, **2002**, *30*, 13.
[2] Wang, G.S.; Cooper, T.A. *Nat. Rev. Genet.*, **2007**, *8*, 749.

[3] Lewis, B.P.; Green, R.E.; Brenner, S.E. *Proc. Natl. Acad. Sci. USA*, **2003**, *100*, 189.

[4] 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.; Raymond, C.; 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. *Nature*, **2001**, *409*, 860.

[5] Faustino, N.A.; Cooper, T.A. *Genes Dev.*, **2003**, *17*, 419.

[6] Garcia-Blanco, M.A.; Baraniak, A.P.; Lasda, E.L. *Nat. Biotechnol.*, **2004**, *22*, 535.

[7] Del Gatto-Konczak, F.; Olive, M.; Gesnel, M.C.; Breathnach, R. *Mol. Cell Biol.*, **1999**, *19*, 251.

[8] Del Gatto-Konczak, F.; Bourgeois, C.F.; Le Guiner, C.; Kister, L.; Gesnel, M.C.; Stevenin, J.; Breathnach, R. *Mol. Cell Biol.*, **2000**, *20*, 6287.

[9] Del Gatto, F.; Plet, A.; Gesnel, M.C.; Fort, C.; Breathnach, R. *Mol. Cell Biol.*, **1997**, *17*, 5106.

[10] Valcarcel, J.; Singh, R.; Zamore, P.D.; Green, M.R. *Nature*, **1993**, *362*, 171.

[11] Jurica, M.S.; Moore, M.J. *Mol. Cell*, **2003**, *12*, 5.

[12] Caceres, J.F.; Kornblihtt, A.R. *Trends Genet*, **2002**, *18*, 186.

[13] Roberts, G.C.; Gooding, C.; Mak, H.Y.; Proudfoot, N.J.; Smith, C.W. *Nucleic Acids Res.*, **1998**, *26*, 5568.

[14] Lopez-Bigas, N.; Audit, B.; Ouzounis, C.; Parra, G.; Guigo, R. *FEBS Lett.*, **2005**, *579*, 1900.

[15] Maquat, L.E.; Carmichael, G.G. *Cell*, **2001**, *104*, 173.

[16] Iborra, F.J.; Jackson, D.A.; Cook, P.R. *Science*, **2001**, *293*, 1139.

[17] Ranum, L.P.; Cooper, T.A. *Annu. Rev. Neurosci.*, **2006**, *29*, 259.

[18] Cho, D.H.; Tapscott, S.J. *Biochim. Biophys. Acta.*, **2007**, *1772*, 195.

[19] Osborne, R.J.; Thornton, C.A. *Hum. Mol. Genet.*, **2006**, *15 Spec No 2*, R162.

[20] Lin, X.; Miller, J.W.; Mankodi, A.; Kanadia, R.N.; Yuan, Y.; Moxley, R.T.; Swanson, M.S.; Thornton, C.A. *Hum. Mol. Genet*, **2006**, *15*, 2087.

[21] Ladd, A.N.; Stenberg, M.G.; Swanson, M.S.; Cooper, T.A. *Dev. Dyn.*, **2005**, *233*, 783.

[22] Kuyumcu-Martinez, N.M.; Wang, G.S.; Cooper, T.A. *Mol. Cell*, **2007**, *28*, 68.

[23] Rowe, S.M.; Miller, S.; Sorscher, E.J. *N. Engl. J. Med.*, **2005**, *352*, 1992.

[24] Liu, H.X.; Cartegni, L.; Zhang, M.Q.; Krainer, A.R. *Nat. Genet.*, **2001**, *27*, 55.

[25] Chiba-Falek, O.; Kerem, E.; Shoshani, T.; Aviram, M.; Augarten, A.; Bentur, L.; Tal, A.; Tullis, E.; Rahat, A.; Kerem, B. *Genomics*, **1998**, *53*, 276.

[26] Hartong, D.T.; Berson, E.L.; Dryja, T.P. *Lancet*, **2006**, *368*, 1795.

[27] Farrar, G.J.; Kenna, P.F.; Humphries, P. *Embo. J.*, **2002**, *21*, 857.

[28] Briese, M.; Esmaeili, B.; Sattelle, D.B. *Bioessays.*, **2005**, *27*, 946.

[29] Mordes, D.; Luo, X.; Kar, A.; Kuo, D.; Xu, L.; Fushimi, K.; Yu, G.; Sternberg, P., Jr.; Wu, J.Y. *Mol. Vis.*, **2006**, *12*, 1259.

[30] Chakarova, C.F.; Hims, M.M.; Bolz, H.; Abu-Safieh, L.; Patel, R.J.; Papaioannou, M.G.; Inglehearn, C.F.; Keen, T.J.; Willis, C.; Moore, A.T.; Rosenberg, T.; Webster, A.R.; Bird, A.C.; Gal, A.; Hunt, D.; Vithana, E.N.; Bhattacharya, S.S. *Hum. Mol. Genet.*, **2002**, *11*, 87.

[31] Vithana, E.N.; Abu-Safieh, L.; Allen, M.J.; Carey, A.; Papaioannou, M.; Chakarova, C.; Al-Maghtheh, M.; Ebenezer, N.D.; Willis, C.; Moore, A.T.; Bird, A.C.; Hunt, D.M.; Bhattacharya, S.S. *Mol. Cell*, **2001**, *8*, 375.

[32] McKie, A.B.; McHale, J.C.; Keen, T.J.; Tarttelin, E.E.; Goliath, R.; van Lith-Verhoeven, J.J.; Greenberg, J.; Ramesar, R.S.; Hoyng, C.B.; Cremers, F.P.; Mackey, D.A.; Bhattacharya, S.S.; Bird, A.C.; Markham, A.F.; Inglehearn, C.F. *Hum. Mol. Genet.*, **2001**, *10*, 1555.

[33] Wirth, B. *Hum Mutat.*, **2000**, *15*, 228.

[34] Lunn, M.R.; Wang, C.H. *Lancet*, **2008**, *371*, 2120.

[35] Rodrigues, N.R.; Owen, N.; Talbot, K.; Ignatius, J.; Dubowitz, V.; Davies, K.E. *Hum. Mol. Genet.*, **1995**, *4*, 631.

[36] Rossoll, W.; Kroning, A.K.; Ohndorf, U.M.; Steegborn, C.; Jablonka, S.; Sendtner, M. *Hum .Mol. Genet.*, **2002**, *11*, 93.

[37] Winkler, C.; Eggert, C.; Gradl, D.; Meister, G.; Giegerich, M.; Wedlich, D.; Laggerbauer, B.; Fischer, U. *Genes Dev.*, **2005**, *19*, 2320.

[38] Cartegni, L.; Krainer, A.R. *Nat Genet*, **2002**, *30*, 377.

[39] Lorson, C.L.; Hahnen, E.; Androphy, E.J.; Wirth, B. *Proc. Natl. Acad. Sci. USA*, **1999**, *96*, 6307.

[40] Srebrow, A.; Kornblihtt, A.R. *J. Cell Sci.*, **2006**, *119*, 2635.

[41] Mazoyer, S.; Puget, N.; Perrin-Vidoz, L.; Lynch, H.T.; Serova-Sinilnikova, O.M.; Lenoir, G.M. *Am. J. Hum. Genet.*, **1998**, *62*, 713.

[42] Karni, R.; de Stanchina, E.; Lowe, S.W.; Sinha, R.; Mu, D.; Krainer, A.R. *Nat. Struct. Mol. Biol.*, **2007**, *14*, 185.

[43] Ghigna, C.; Giordano, S.; Shen, H.; Benvenuto, F.; Castiglioni, F.; Comoglio, P.M.; Green, M.R.; Riva, S.; Biamonti, G. *Mol. Cell*, **2005**, *20*, 881.

[44] Levanon, E.Y.; Sorek, R. *Targets*, **2003**, *2*,

[45] Kowalski, M.L.; Borowiec, M.; Kurowski, M.; Pawliczak, R. *Allergy*, **2007**, *62*, 628.

[46] Censarek, P.; Steger, G.; Paolini, C.; Hohlfeld, T.; Grosser, T.; Zimmermann, N.; Fleckenstein, D.; Schror, K.; Weber, A.A. *Thromb. Haemost*, **2007**, *98*, 1309.

[47] Kis, B.; Snipes, J.A.; Gaspar, T.; Lenzser, G.; Tulbert, C.D.; Busija, D.W. *Inflamm. Res.*, **2006**, *55*, 274.

[48] Schneider, C.; Boeglin, W.E.; Brash, A.R. *Biochem. J.*, **2005**, *385*, 57.

[49] Warner, T.D.; Mitchell, J.A. *Proc. Natl. Acad. Sci. USA*, **2002**, *99*, 13371.

[50] Vane, J.R. *J. Physiol. Pharmacol.*, **2000**, *51*, 573.

[51] Mitchell, J.A.; Warner, T.D. *Br. J. Pharmacol.*, **1999**, *128*, 1121.

[52] Bennett, K.L.; Jackson, D.G.; Simon, J.C.; Tanczos, E.; Peach, R.; Modrell, B.; Stamenkovic, I.; Plowman, G.; Aruffo, A. *J. Cell Biol.*, **1995**, *128*, 687.

[53] Bennett, K.L.; Modrell, B.; Greenfield, B.; Bartolazzi, A.; Stamenkovic, I.; Peach, R.; Jackson, D.G.; Spring, F.; Aruffo, A. *J. Cell Biol.*, **1995**, *131*, 1623.

[54] Heider, K.H.; Kuthan, H.; Stehle, G.; Munzert, G. *Cancer Immunol. Immunother*, **2004**, *53*, 567.

[55] Wood, M.; Yin, H.; McClorey, G. *PLoS Genet.*, **2007**, *3*, e109.

[56] Gaur, R.K. *Biotechniques*, **2006**, *Suppl*, 15.

[57] Zhu, H.; Guo, W.; Zhang, L.; Davis, J.J.; Teraishi, F.; Wu, S.; Cao, X.; Daniel, J.; Smythe, W.R.; Fang, B. *Mol. Cancer Ther.*, **2005**, *4*, 451.

[58] Dias, N.; Stein, C.A. *Mol. Cancer Ther.*, **2002**, *1*, 347.

[59] Yeo, G.W. *Genome. Biol.*, **2005**, *6*, 240.

[60] Blanchette, M.; Green, R.E.; Brenner, S.E.; Rio, D.C. *Genes Dev.*, **2005**, *19*, 1306.

[61] Soret, J.; Bakkour, N.; Maire, S.; Durand, S.; Zekri, L.; Gabut, M.; Fic, W.; Divita, G.; Rivalle, C.; Dauzonne, D.; Nguyen, C.H.; Jeanteur, P.; Tazi, J. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 8764.

[62] Freed, E.O.; Mouland, A.J. *Retrovirology*, **2006**, *3*, 77.

[63] Bakkour, N.; Lin, Y.L.; Maire, S.; Ayadi, L.; Mahuteau-Betzer, F.; Nguyen, C.H.; Mettling, C.; Portales, P.; Grierson, D.; Chabot, B.; Jeanteur, P.; Branlant, C.; Corbeau, P.; Tazi, J. *PLoS Pathog.*, **2007**, *3*, 1530.

[64] Resina, S.; Kole, R.; Travo, A.; Lebleu, B.; Thierry, A.R. *J. Gene Med.*, **2007**, *9*, 498.

[65] Kurreck, J. *Eur. J. Biochem.*, **2003**, *270*, 1628.

[66] Graziewicz, M.A.; Tarrant, T.K.; Buckley, B.; Roberts, J.; Fulton, L.; Hansen, H.; Orum, H.; Kole, R.; Sazani, P. *Mol. Ther.*, **2008**,

[67] Yin, H.; Lu, Q.; Wood, M. *Mol. Ther.*, **2008**, *16*, 38.

[68] Aartsma-Rus, A.; van Ommen, G.J. *Rna*, **2007**, *13*, 1609.

[69] van Deutekom, J.C.; Janson, A.A.; Ginjaar, I.B.; Frankhuizen, W.S.; Aartsma-Rus, A.; Bremmer-Bout, M.; den Dunnen, J.T.; Koop, K.; van der Kooi, A.J.; Goemans, N.M.; de Kimpe, S.J.; Ekhart, P.F.; Venneker, E.H.; Platenburg, G.J.; Verschuuren, J.J.; van Ommen, G.J. *N. Engl. J. Med.*, **2007**, *357*, 2677.

[70] Arechavala-Gomeza, V.; Graham, I.R.; Popplewell, L.J.; Adams, A.M.; Aartsma-Rus, A.; Kinali, M.; Morgan, J.E.; van Deutekom, J.C.; Wilton, S.D.; Dickson, G.; Muntoni, F. *Hum. Gene Ther.*, **2007**, *18*, 798.

[71] Kole, R.; Williams, T.; Cohen, L. *Acta. Biochim. Pol.*, **2004**, *51*, 373.

[72] Garcia-Blanco, M.A. *Prog. Mol. Subcell. Biol.*, **2006**, *44*, 47.

[73] Phylactou, L.A.; Darrah, C.; Wood, M.J. *Nat. Genet*, **1998**, *18*, 378.

[74] Blencowe, B.J. *Cell*, **2006**, *126*, 37.

[75] Buratti, E.; Dork, T.; Zuccato, E.; Pagani, F.; Romano, M.; Baralle, F.E. *Embo. J.*, **2001**, *20*, 1774.

[76] Tsuneda, S.S.; Torres, F.R.; Montenegro, M.A.; Guerreiro, M.M.; Cendes, F.; Lopes-Cendes, I. *J. Mol. Neurosci.*, **2008**, *35*, 195.

[77] Yang, L.; Embree, L.J.; Hickstein, D.D. *Mol. Cell Biol.*, **2000**, *20*, 3345.

[78] Crozat, A.; Aman, P.; Mandahl, N.; Ron, D. *Nature*, **1993**, *363*, 640.

[79] Ichikawa, H.; Shimizu, K.; Hayashi, Y.; Ohki, M. *Cancer Res.*, **1994**, *54*, 2865.

[80] Imai, H.; Chan, E.K.; Kiyosawa, K.; Fu, X.D.; Tan, E.M. *J. Clin. Invest*, **1993**, *92*, 2419.

[81] Clark, J.; Lu, Y.J.; Sidhar, S.K.; Parker, C.; Gill, S.; Smedley, D.; Hamoudi, R.; Linehan, W.M.; Shipley, J.; Cooper, C.S. *Oncogene*, **1997**, *15*, 2233.

[82] Yamada, T.; Tachibana, A.; Shimizu, T.; Mugishima, H.; Okubo, M.; Sasaki, M.S. *J Hum Genet.*, **2000**, *45*, 159.

[83] Hasegawa, Y.; Kawame, H.; Ida, H.; Ohashi, T.; Eto, Y. *Hum Genet*, **1994**, *93*, 415.

[84] Roberts, R.; Timchenko, N.A.; Miller, J.W.; Reddy, S.; Caskey, C.T.; Swanson, M.S.; Timchenko, L.T. *Proc. Natl. Acad. Sci. USA*, **1997**, *94*, 13221.

[85] Kanadia, R.N.; Johnstone, K.A.; Mankodi, A.; Lungu, C.; Thornton, C.A.; Esson, D.; Timmers, A.M.; Hauswirth, W.W.; Swanson, M.S. *Science*, **2003**, *302*, 1978.

[86] Miller, J.W.; Urbinati, C.R.; Teng-Umnuay, P.; Stenberg, M.G.; Byrne, B.J.; Thornton, C.A.; Swanson, M.S. *Embo J.*, **2000**, *19*, 4439.

[87] Sossi, V.; Giuli, A.; Vitali, T.; Tiziano, F.; Mirabella, M.; Antonelli, A.; Neri, G.; Brahe, C. *Eur. J. Hum. Genet.*, **2001**, *9*, 113.