DETERMINACIÓN DE BLANCOS TRADUCCIONALES DE *PDCD4* (*PROGRAMMED CELL DEATH 4*) MEDIANTE ANÁLISIS DE DATOS GENERADOS POR SECUENCIACIÓN MASIVA DE HUELLAS POLISOMALES

TESINA DE GRADO

LICENCIATURA EN BIOQUÍMICA

Facultad de Ciencias, UdelaR

Tutor: José Sotelo-Silveira PhD

Departamento de Genética, IIBCE

Bach. Guillermo Eastman

Febrero 2012

ÍNDICE

RESUMEN

El control de la traducción en células eucariotas, en particular de la fase de iniciación es una etapa crítica para la regulación de la expresión génica. Son extensos los ejemplos donde se observa una desregulación a nivel del compartimiento traduccional en patologías severas como el cáncer. Al respecto se ha identificado un gen supresor de tumores denominado *Programmed Cell Death 4 (PDCD4)* el cual se ha propuesto que regula la expresión génica tanto a nivel transcripcional, como a nivel traduccional. Se ha propuesto, aunque no se conocen con detalles los blancos de acción, que PDCD4 como factor proteico compite con los ARNm celulares y con eIF4G por unir a eIF4A en la formación del complejo de pre-iniciación 43S de la traducción. Al interactuar PDCD4 con eIF4A podría interferir con la acción helicasa de dicho factor de iniciación, lo cual repercute en la capacidad del mismo para resolver estructuras secundarias de los ARNm traducidos de forma cap-dependiente. Dichas estructuras están presentes particularmente en los mensajeros de oncogenes.

El principal objetivo que nos propusimos fue encontrar nuevos blancos traduccionales de la proteína PDCD4, mediante el análisis de datos generados previamente por secuenciación masiva de huellas polisomales obtenidas mediante la técnica de *ribosome profiling*. Esta tecnología fue desarrollada por Ingolia *et al.* en 2009 y se basa en la secuenciación masiva de bibliotecas de ADN construidas a partir de los fragmentos de ARNm que quedan protegidos por la maquinaria traduccional de la digestión por ARNasas (huellas polisomales). Se cuenta con dichas huellas, tanto en condiciones de supresión del factor PDCD4 mediante siARN, como en condiciones normales. La cuantificación del mapeo de estas huellas polisomales y el análisis de expresión génica de ARNm nos permiten estimar el estado traduccional para cerca de 6.000 genes en modelos humanos.

A modo de preparación para el análisis de datos de secuenciación masiva, se trabajó con los datos generados en la publicación donde los autores que desarrollaron y pusieron a punto la técnica de *ribosome profiling,* la aplican por primera vez en un modelo de levaduras sometidas a privación de aminoácidos. En esta instancia previa se comprobó que nuestros métodos informáticos eran capaces de reproducir resultados

presentes en dicha publicación e incluso explorar aspectos que no fueron cubiertos en su momento por los autores.

Así pasamos al estudio detallado de la influencia traduccional de PDCD4 y la búsqueda de genes candidatos a ser sus blancos traduccionales. Para esto se realizaron alineamientos, mapeos y cálculos de eficiencia traduccional de ARNm mediante la metodología RNA-Seq. Los valores de expresión fueron normalizados, y luego de evaluar la distribución de los mismos, se compararon las dos condiciones de trabajo en búsqueda de genes que aumentaran considerablemente sus niveles traduccionales en condiciones de supresión de PDCD4. De esta forma se pudo detectar un amplio espectro de regulación traduccional (más de 400 genes) ejercido por dicho factor. La lista con dichos genes candidatos fue sometida a estudios de ontología de genes, los cuales revelaron que las funciones propuestas para PDCD4 en la literatura se verificaban, en parte, en nuestros estudios. Las principales funciones celulares-moleculares involucradas en los estudios de ontología fueron las de señalización celular, morfología, organización y desarrollo celular, así como control de la expresión génica. Muchos genes también mostraron una relación directa con la respuesta inmune mediada por interferón tipo I en donde PDCD4 podría estar implicado de forma novedosa. Dentro de los candidatos a ser blancos traduccionales de PDCD4 resaltan los ARN mensajeros de oncogenes como *RET, WNT3A* y *MLLT10*, ya que estarían regulados negativamente por PDCD4 y se trata de genes pro oncogénicos en diversas situaciones conocidas previamente.

CAPÍTULO 1

INTRODUCCIÓN

1) TRADUCCIÓN Y EXPRESIÓN GÉNICA

1.1) GENERALIDADES Y REGULACIÓN DE LA TRADUCCIÓN

Las proteínas se consideran unas de las más importantes moléculas en los procesos de la vida. Estas constituyen una importante fracción de las macromoléculas que catalizan reacciones de significancia para la vida, por ejemplo sirven de soporte estructural, transporte, reguladores y muchos más otros roles en todos los organismos, con lo cual son un factor clave a la hora de determinar el fenotipo de un individuo. Como dato puntual, las proteínas constituyen cerca del 44% del peso seco humano[1]. Por estas razones gran parte de las reservas energéticas celulares son destinadas a la síntesis proteica, un proceso sofisticado que requiere de una maquinaria biológica que abarca una larga lista de componentes. A modo de ejemplo, descontando los genes que son dispensables para el crecimiento celular en el laboratorio, puede calcularse que cerca de un 40% del total de un genoma mínimo, es necesario para un correcto funcionamiento de la maquinaria traduccional[2] (nota: de aquí al final de la sección 1.1.1 se utilizó como referencia el capítulo aquí citado, por lo que, a menos que se indique una referencia adicional, las afirmaciones presentadas fueron derivadas del mencionado capítulo).

Inicialmente, la idea central del control traduccional estuvo basada en que la expresión génica es regulada mediante la eficiencia en la utilización de ARN mensajeros (ARNm) en la síntesis de proteínas específicas. Esta noción surgió solo unos pocos años después del establecimiento del dogma central de la biología molecular[3] y muy poco después de la formulación de la hipótesis del ARN mensajero[4]. Con el transcurso del tiempo, la información aportada por los distintos avances en este campo de estudio, han permitido revelar un mecanismo de regulación traduccional que actúa de forma precisa para regular la síntesis proteica e integrando señalización muchas vías

presentes en la célula. En muchos casos este mecanismo no es considerado como se debe y puede ser pasado por alto o tenido en cuenta sólo superficialmente. Sin embargo como se explicará a lo largo de este punto, la regulación de la traducción de proteínas, en particular la regulación de la iniciación de la misma, es un proceso clave para regular la expresión génica donde cualquier desajuste en el mismo puede conducir a patologías severas como el cáncer.

1.1.1) PROPIEDADES, BLANCOS Y MECANISMOS GENERALES DEL CONTROL TRADUCCIONAL

En un proceso de múltiples pasos como la síntesis de proteínas, la regulación puede ser ejercida en varios puntos. Al respecto, los ejemplos en la regulación de la traducción se encuentran en distintos niveles, aunque existe una mayor acumulación de literatura a favor de considerar el inicio de la traducción como el punto clave de regulación. Esta observación empírica está de acuerdo con el principio biológico y lógico de que es más eficiente regular un fenómeno en sus comienzos, más que interrumpirlo en la mitad del mismo, lo que gastaría energía celular para generar solamente productos intermedios, que habría que reciclar o descartar[2]. Sin embargo, se han caracterizados casos donde la regulación ocurre en pasos posteriores a la iniciación, especialmente a nivel de la elongación[5,6,7].

De todas formas, la pregunta acerca de cuál es la fase limitante debe ser específica para cada ARNm en particular, más que a la población entera de los mismos. Una manera de discriminar cual de las fases de la traducción es la limitante en la síntesis proteica, es examinar los perfiles polisomales (ver Figura 1.1). Esta técnica representa una herramienta flexible y de alto alcance para estudiar e investigar el control traduccional. El tamaño de un polisoma en particular es determinado por tres parámetros: i) el largo de la secuencia codificante del mensajero que está traduciendo, ii) la tasa de iniciación, y iii) la tasa de elongación/terminación; siendo directamente proporcional a los dos primeros e inversamente proporcional al tercero. En promedio, se encuentra un ribosoma cada 80-100 nucleótidos en los polisomas, lo cual representa una baja densidad ribosomal en los polisomas ya que un ribosoma ocupa cerca de 30 nucleótidos del mensajero[8]. Si se supone que el proceso de iniciación es

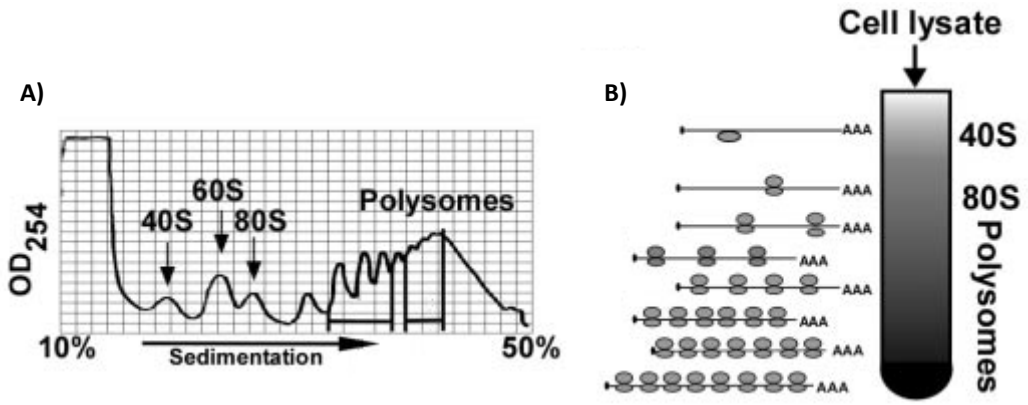


Figura 1.1: Perfil polisomal. A) Se observa el valor de absorbancia a 254nm para sucesivas fracciones de un gradiente de 10% a 50% de sacarosa, de izquierda a derecha como se muestra. Se indican las subunidades menor (40S) y mayor (60S), y el ribosoma libre (80S), así como la fracción correspondiente a los polisomas; B) Se observa el gradiente antes mencionado con las diferentes poblaciones de mensajeros y como aumenta el número de ribosomas sobre estos a medida que avanzamos hacia el fondo del tubo. Figura adaptada de Arava et al. 2005.

regulado de forma más intensa y es por lo tanto la fase limitante, esta baja densidad polisomal en los mensajeros se explica por una menor tasa de iniciación respecto de la tasa de elongación/terminación.

Bajo condiciones estables, es razonable concluir que el número de proteínas producidas sea aproximadamente igual al número de eventos de iniciación traduccionales. De todas formas, a la hora de definir los parámetros que influyen o definen la tasa traduccional de determinado ARNm, podemos encontrar principalmente cuatro parámetros que se describen brevemente a continuación:

- Cantidad de ARNm: el nivel de mensajeros en el citoplasma es determinado tanto por la tasa de transcripción, como por la proporción de transcriptos primarios que son procesados y transportados al citoplasma como mensajeros maduros. En células de mamíferos activas traduccionalmente, los mensajeros se encuentran formando parte de los polisomas, como ocurre por ejemplo con la actina[9], por lo cual es lógico pensar que la tasa de síntesis de determinada proteína es limitada por la cantidad de su ARNm. De todas formas, el total de mensajeros en el citoplasma frecuentemente aparece en exceso, donde cerca del 30% de mensajeros en células en cultivo se presenta como partículas ribonucleoproteícas libres[9,10,11]. Por esto, los niveles de ARNm, no son generalmente el factor limitante de la traducción de un ARNm.

- Abundancia de ribosomas: Los niveles celulares de ribosomas no suelen ser el factor limitante en la mayoría de las condiciones. Por ejemplo, en la mayoría de los tejidos de mamíferos y en algunas células en cultivo de rápida proliferación, los ribosomas no activos y sus subunidades constituyen aproximadamente el 20% de la población total de ribosomas, como ha sido observado en análisis por centrifugación en gradientes de sacarosa de lisados celulares de tejidos[13]. Esta observación sugiere que, más que ser el factor limitante, los niveles de ribosomas capacitan a la célula para responder rápidamente a señales extra o intracelulares que impliquen una activación traduccional.

- Actividad de la maquinaria de síntesis proteica: En aquellas células donde la tasa de síntesis proteica no es determinada por los niveles ribosomales o de ARNm, puede ocurrir que la misma sea afectada por algún componente soluble de la maquinaria el cual puede estar limitado en cantidad o actividad. Parte de este tipo de regulación involucra el estado de fosforilación de estos componentes. *A priori* este tipo de regulación afectaría la traducción de todos los mensajeros por igual, sin embargo, se ha visto que una desregulación de los primeros pasos de la traducción afecta de forma más significativa a aquellos mensajeros "débiles" que de por sí ya tienen bajas tasas de iniciación, respecto al otro tipo de mensajeros denominados "fuertes"[14]. Recíprocamente, la activación de la iniciación tiende a estimular la traducción de les mensajeros "débiles". Por esto, es de esperarse que la alteración de la actividad de los componentes que interactúan con los mensajeros y que son capaces de afectar su unión al ribosoma, genere diferentes efectos en la traducción de la población total de mensajeros[15].

- Tasa de iniciación y elongación: La tasa de iniciación puede ser inhibida si un ribosoma ya instalado abandona la región de iniciación muy lentamente, ya que un ribosoma unido al codón de iniciación ocupa unos cuantos nucleótidos "rio abajo" (*downstream)(*12-15) y cerca de la misma cantidad "rio arriba" (*upstream).* P Esta huella es de 28 a 30 nucleotidos en total. Por esto, otro ribosoma no podrá unirse al sitio de iniciación si el primero no se ha movido cerca de 10 codones abajo por el mensajero. Este ejemplo sencillo ilustra como la tasa de elongación afecta la de iniciación y repercute directamente en la síntesis proteica. Asimismo existen más casos

donde la elongación repercute en la iniciación, por ejemplo, la presencia de estructuras secundarias en los mensajeros o la presencia de codones raros o pausas ribosomales[16].

Más allá de que estos cuatro parámetros presentados definan la tasa de traducción observada para cada ARNm, esto no implica que se defina una tasa fija invariable. Por el contrario, sobre la iniciación, por ser la etapa limitante de la traducción, se aplica la mayoría del control traduccional. Esta regulación es variada y opera mediante muchos mecanismos, algunos reconocen estructuras secundarias de los mensajeros, otras actúan sobre el aparato traduccional en sí, y otras mediante proteínas que actúan en *trans* o factores de ARN.

Los blancos más comunes sobre los cuales opera el control traduccional son el mensajero, los factores de iniciación y de elongación, y el ribosoma. Brevemente respecto a los factores de iniciación se puede aportar que su actividad es generalmente modulada por los efectos de varios elementos en *cis* presentes en el 5´-UTR del mensajero, así como también existen otros factores que actúan en *trans.* Respecto a los factores de elongación, se sabe que la tasa de elongación es modulada por la fosforilación de dichos factores, particularmente a través de la regulación de la actividad del factor de elongación traduccional eEF2. Este factor junto con los otros factores de elongación, además de estar regulados por fosforilación, están sujetos a muchas otras modificaciones post-traduccionales que regulan su actividad. Respecto al ribosoma, éste es también un sitio de regulación importante donde se encuentran muchas proteínas ribosomales que son blancos de modificaciones post-traduccionales, por ejemplo fosforilaciones, metilación y ubiquitinación. Esto sugiere que una posible heterogeneidad en la población celular de ribosomas podría ser importante para la traducción de clases específicas de mensajeros.

De todas formas el punto más importante de regulación opera a través del ARN mensajero. Prácticamente todos los mensajeros maduros de eucariotas tienen un 5´-terminal m^7G[5´]ppp[5´]N *ca*p, donde N es un nucleótido cualquiera, y una cola de poli-(A) de 50 a 300 nucleótidos de largo. La caperuza o *cap* es importante pero no esencial, mientras que la cola poli-(A) solo estimula la iniciación de forma modesta. Sin embargo

en condiciones normales actúan de forma sinérgica para iniciar la traducción. Este sinergismo es mediado por interacciones entre proteínas que unen estos dos elementos de los mensajeros, haciendo que estos adopten una conformación circular. Dichas proteínas son: el complejo eIF4F y la proteína de unión a la cola poli-(A) (PABP). Los mensajeros también pueden presentar estructuras secundarias a nivel del 5´-UTR, así como distintos largos del mismo, lo cual repercute en el grado de "traducibilidad" del mensajero. Por otro lado, la síntesis proteica propiamente dicha comienza con el primer codón de iniciación, frecuentemente AUG, aunque pueden presentarse más de un codón de iniciación, o codones de iniciación no canónicos, como CUG, ACG y GUG. La elección del codón de iniciación apropiado es determinada por el contexto nucleotídico en torno al AUG, por ejemplo la secuencia GCC(A/G)CCAUGG es la óptima en células de mamíferos.

La eficiencia traduccional intrínseca para un mensajero depende de importantes elementos estructurales que actúan en *cis*, los cuales también juegan roles críticos en la regulación de la utilización de dicho mensajero. Es conveniente dividir estos elementos dentro de dos categorías, aquellos elementos que actúan a través del aparato traduccional, y aquellos cuyos efectos están mediados por factores específicos que actúan en *trans*.

En eucariotas, donde el principal mecanismo de iniciación es *cap*-dependiente, los elementos estructurales distribuidos a lo largo del mensajero determinan y modulan la eficiencia traduccional del mismo. Mensajeros que son eficientemente traducidos poseen un 5´-*cap* que es accesible a los factores de iniciación, un codón de iniciación cercano al extremo 5´, así como una cola poli(A) de tamaño adecuado. Por otra parte, la eficiencia traduccional de los mensajeros se ve disminuida por la presencia de estructuras secundarias en el 5´-UTR y/o por la presencia de codones de iniciación o uORFs *(upstream Open Reading Frame).*

Además de estas propiedades innatas, más elementos en *cis* pueden estar presentes en los mensajeros (principalmente en el 5´-UTR), permitiendo una regulación temporal de la eficiencia traduccional mediante factores que actúen en *trans.* El nivel de complejidad aumenta si consideramos la posible presencia de sitios internos de

entrada ribosomal (IRES) los cuales requieren solo algunos de los factores de iniciación canónicos, o hasta ninguno de ellos en algunos casos. La presencia de estos IRES permite al mensajero evadir los mecanismos reguladores que afectan la traducción *cap*-dependiente, lo cual permitiría que la traducción del mismo ocurra en condiciones donde la síntesis proteica ha sido inhibida, por ejemplo como ocurre en algunas infecciones virales.

Por otro lado, el 3´-UTR posee un rico catalogo de elementos en *cis* que determinan la estabilidad del mensajero y su localización en el citoplasma, y también sirven para regular el inicio de la traducción. Este tipo de control es mediado por proteínas que actúan en *trans* y por micro ARN, los cuales han sido descubiertos recientemente y poseen funciones reguladoras relevantes, que por razones de espacio no cubriremos en esta tesina.

La estabilidad de los mensajeros es un factor de suma importancia a la hora de determinar los niveles citoplasmáticos de ARNm y en cierta forma los niveles de síntesis proteica. En muchos casos, la traducción tiene un rol directo en determinar la estabilidad de los mensajeros, ya que se ha observado que puede ocurrir la degradación de los mismos de forma acoplada con su traducción cuando existen grandes pausas ribosomales (*"no-go" decay*) o fallas en la terminación de la traducción ("*nonstop" decay)*.

1.1.2) INICIACIÓN DE LA TRADUCCIÓN

El proceso de iniciación en eucariotas consiste en una serie de etapas en cada una de las cuales intervienen uno o más factores de iniciación eucariotas (eIF). Se han caracterizado al menos 13 eIF (más de 30 cadenas polipeptídicas), con varias isoformas y numerosos co-factores, cuyas funciones e importancia en la iniciación todavía no se comprende de forma completa[17,18]. Los factores canónicos son agrupados de acuerdo a la etapa en la que actúan, de todas formas, muchos factores son multifuncionales y pueden actuar en varias etapas.

El primer paso en la iniciación es el ensamblado de eIF2, GTP y el ARN de transferencia de la metionina de iniciación (Met-$tRNA_i^{Met}$) en un complejo ternario, cuya función es

transportar dicho transfer a la subunidad 40S, así como también juega un rol clave en identificar el codón de iniciación. Luego de cada ronda de iniciación, eIF2 (una GTPasa heterotrimérica) es liberado formando un complejo inactivo con GDP[19]. Como la liberación del GDP de eIF2 es muy lenta, es necesaria la presencia de un factor intercambiador de nucleótidos de guanina (GEF), llamado eIF2B, el cual acelera el remplazo de GDP por GTP al menos diez veces[20]. El cambio de GDP por GTP en eIF2, aumenta la afinidad de dicho factor por el transfer de iniciación[21,22].

Posteriormente, la interacción entre el complejo ternario con la subunidad 40S forma el complejo de pre-iniciación 43S (ver Figura 1.2), facilitado por la presencia de eIF1, eIF1A y eIF3[23,24]. También se ha implicado a eIF5 en éste proceso ya que este interactúa con las subunidades de eIF2 y eIF3, por lo que parecería ser reclutado a la subunidad 40S por estos factores. Los mecanismos por los cuales todos estos factores promueven de forma cooperativa la formación del complejo 43S incluyen interacciones directas entre eIF2 y la subunidad 40S, interacciones entre los factores, así como cambios conformacionales inducidos en la subunidad 40S[25,26]. A pesar de que la mayoría de los mensajeros celulares presenten alguna estructura secundaria en su 5´-UTR, el complejo 43S es capaz de unirlos gracias a la actividad cooperativa de eIF4F, eIF4B y posiblemente PABP. La unión del complejo al ARNm comienza con el reconocimiento del *cap* por parte de eIF4E (subunidad del complejo eIF4F). eIF4E consiste en ocho láminas β curvadas antiparalelas, apoyadas en tres largas hélices[27]. Una vez unido el *cap,* este último se mantiene entre dos residuos de triptófano en la superficie cóncava de eIF4E, mientras que contactos adicionales entre el factor y el nucleótido adyacente al *cap* estabilizan la unión de eIF4E con el mensajero. Esta unión

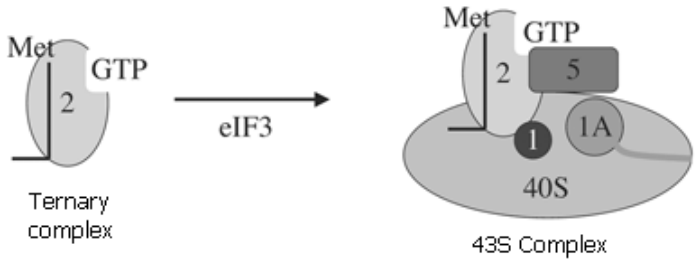


Figura 1.2: Formación del complejo 43S. El complejo ternario constituido por eIF2, GTP y Met-tRNA$_i^{Met}$, se une a la subunidad 40S formando el complejo de pre-iniciación 43S, donde participan eIF1, eIF1A, eIF3 y eIF5. Figura adaptada de Acker and Lorsch 2008.

de eIF4E al mensajero sirve a la vez para mediar la unión de los otros dos componentes del complejo eIF4F: eIF4A y eIF4G. El complejo resultante es muy estable, muestra una alta afinidad por el 5´-*cap* y se disocia del mismo en forma lenta[28]. A la vez, la unión de PABP a eIF4G, aumenta aún más la interacción entre eIF4F y el *cap,* posiblemente al inducir cambios conformacionales adicionales en eIF4E[28].

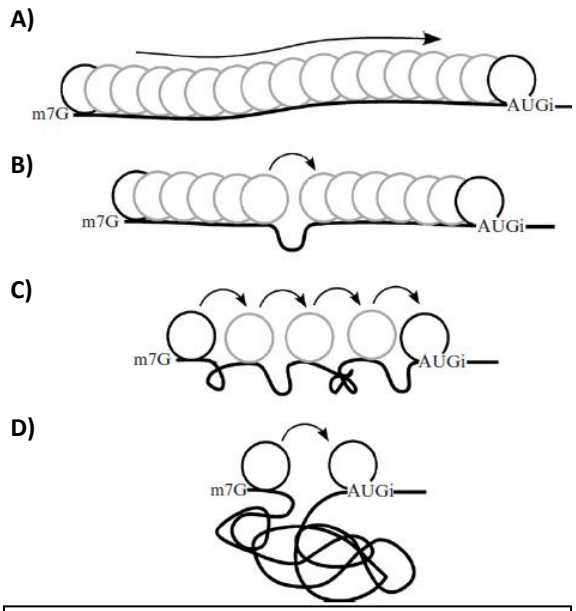


Figura 1.3: Representación esquemática de los posibles mecanismos de scanning del complejo 43S sobre el ARNm A) Búsqueda lineal; B) Salteándose algunos nucleótidos y buscando en otros; C) Buscando solo en determinadas regiones D) Pasando por alto toda la región y saltando directo al codón de iniciación. Figura adaptada de Mauro et al. 2007

Así, una vez cargado el complejo 43S en un mensajero, comienza la búsqueda o *scanning* del codón de iniciación sobre la región 5´-UTR en dirección 3´ hasta encontrar el codón de iniciación correcto en el contexto adecuado. Estudios en microscopía crioelectrónica, así como otros estudios, sugieren que eIF1 y eIF1A colaboran en mantener el complejo 43S es una conformación "abierta" competente para buscar el codón de iniciación, hasta que el correcto apareamiento de bases entre el codón y el anticodón sea logrado[29]. Iniciar la traducción en un codón incorrecto es un problema potencialmente serio para la célula ya que resulta en la producción de una proteína truncada o fuera del correcto marco de lectura. De todas formas la evolución permitió incorporar numerosos pasos de control para asegurar que sea elegido el codón de iniciación correcto.

Se ha propuesto que el movimiento o *scanning* del complejo 43S sobre el mensajero puede ocurrir de varias maneras: i) de forma lineal buscando en todos los nucleótidos; ii) salteándose algunos de ellos y buscando en el resto; iii) buscando solamente en determinadas regiones y saltearse las otras; iv) pasando por alto toda la región y saltando directo al codón de iniciación[30] (ver Figura 1.3). Sin importar por cuál de estas maneras el complejo 43S encuentra el codón de iniciación, una vez que lo hace,

se detiene y forma el complejo de iniciación 48S donde existe una interacción adecuada en el sitio P entre el codón y el anticodón[31] (ver Figura 1.4).

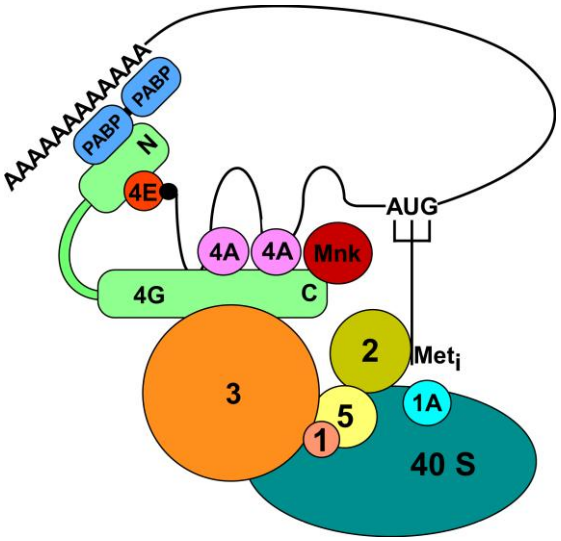


Figura 1.4: Modelo del complejo de iniciación 48S. Se muestran las interacciones entre los factores eIF1, 2, 3, 4A, 4E, 4G, 5, Mnk, PABP, ARNm y la subunidad 40S. La línea delgada representa el mensajero. Los tamaños de las proteínas son proporcionales a sus masas moleculares. Figura adaptada de Rhoads et al. 2006.

El apareamiento de bases entre el codón y el anticodón induce la hidrólisis de GTP unido a eIF2, mediada por eIF5, con la liberación de P_i lo que conduce a la disociación parcial de eIF2-GDP[32]. La forma unida a GDP del factor eIF2 posee una baja afinidad por el Met-tRNA$_i^{Met}$, lo cual lo libera del ribosoma dejando el ARNt en el sitio P del la subunidad 40S[33]. A continuación eIF5B media la subsecuente disociación de todos los factores de iniciación, salvo eIF1A, de la subunidad menor 40S. Como muchos de estos factores de iniciación se unen a la superficie correspondiente a la interface definida entre las subunidades, impiden el cargado de la subunidad mayor. En este punto, la función de eIF5B parecería ser el de remover todos esos factores para permitir el ensamblado del ribosoma completo[33]. Posteriormente la hidrólisis de GTP unido a eIF5B determina su liberación del ribosoma entero 80S (ver Figura 1.5). Aquí ocurre algo curioso, en el sentido que eIF1A permanece unido al complejo de preiniciación una vez que se ha reconocido el codón de iniciación. Esto se debe principalmente a que eIF5B presenta un dominio de unión a eIF4A, de esta forma eIF5B puede ser reclutado al complejo para ejercer su función y permitir el consecuente armado del ribosoma completo.

Una vez completada la traducción, la disociación de los ribosomas es esencial ya que una nueva iniciación solo puede ocurrir a partir de las subunidades libres. En eucariotas no se conocen factores citoplasmáticos que medien el reciclaje ribosomal, únicamente se ha observado que eIF3 tiene una actividad de disociación ARN-dte[26],

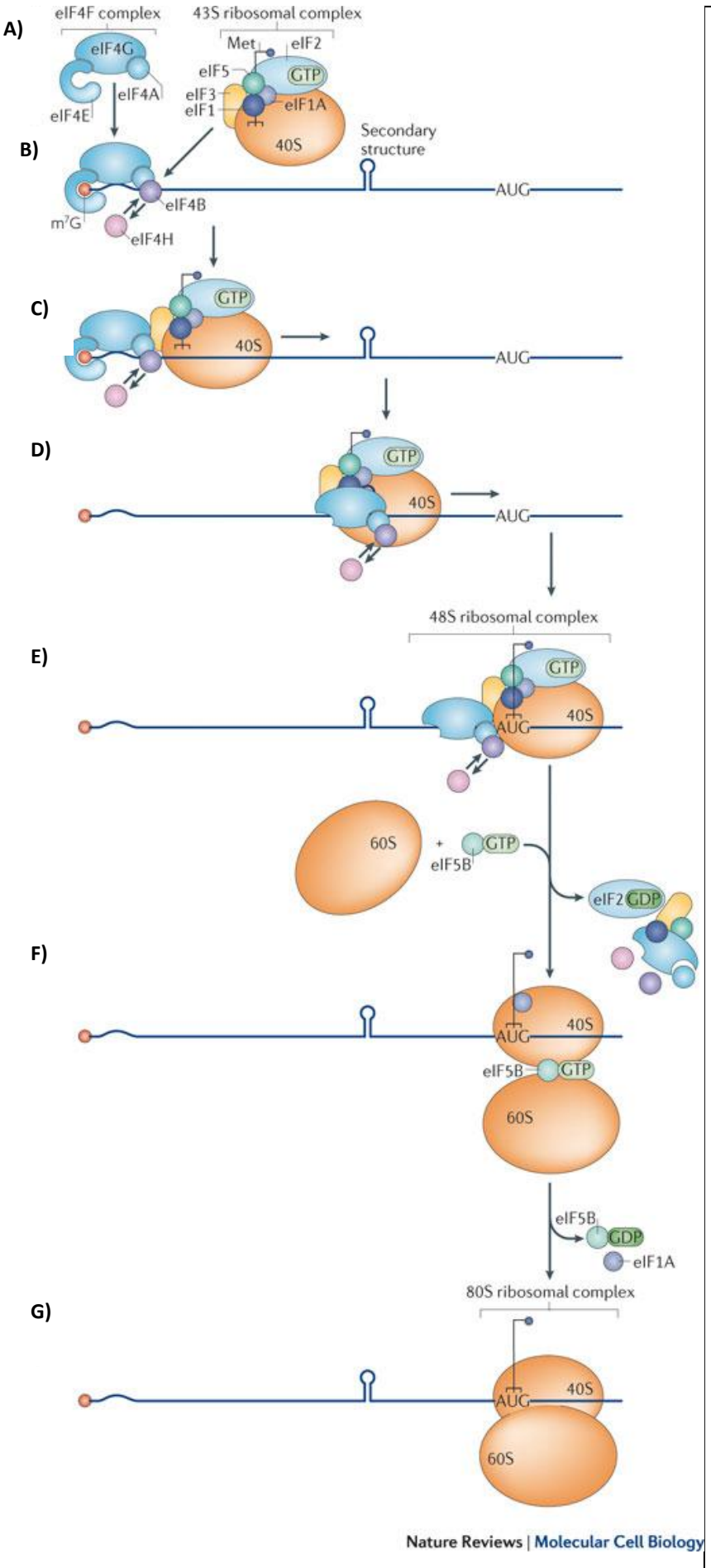


Figura 1.5: Representación esquemática de todo el proceso de iniciación de la traducción. A) En primer lugar, se forman los complejos eIF4F (compuesto por eIF4G, 4A y 4E) y el complejo 43S formado por eIF3, 1, 1A, eIF2–GTP-ARNt iniciador (Met-tRNAMeti), eIF5 todos unidos a la subunidad ribosomal 40S; B) eIF4E une el 5´-cap del mensajero. La interacción entre eIF4G y eIF3 recluta el complejo 43S. eIF4B y eIF4H comparten un sitio de unión común a eIF4A y sus interacciones son mutuamente excluyentes. La circularización del mensajero mostrada en la figura 1.4 no se muestra. C) El complejo 43S busca en dirección 5´-3´ hasta encontrar alguna estructura secundaria que lo detiene. D) eIF4A, junto con otras helicasas no canónicas y proteínas accesorias desarman las estructuras secundarias presentes en los mensajeros. eIF4B y eIF4H potencian la actividad de eIF4A. Así, el complejo 43S retoma el scanning hasta encontrar el codón de iniciación. E) El reconocimiento del codón de iniciación conduce a la formación del complejo 48S gracias a la hidrólisis de eIF2-GTP mediada por eIF5. eIF5B–GTP colabora con el reclutamiento de la subunidad mayor 60S al complejo 48S, para formar el complejo 80S desplazando eIF2–GDP, eIF1, 3, 4A, 4B, 4G, 4H y 5. F) eIF5B-GDP junto con eIF1A son desplazados del complejo ribosomal 80S. G) El ribosoma completo se ha formado correctamente y está competente para la elongación. Figura adaptada de Sonenberg et al. 2009 y Jackson et al. 2010.

la cual es mejorada por eIF1A y particularmente por eIF1. Hoy en día existe controversia acerca de cuál es el mecanismo que permite el reciclaje ribosomal y los factores partícipes varían según los autores. Sin embargo, es importante marcar que la disociación del ribosoma 80S vacío está ligada directamente con la formación del complejo de pre-iniciación 43S, el cual juega un rol clave en el inicio de la traducción.

1.1.3) eIF4A

eIF4A es un factor de iniciación altamente conservado en eucariotas[34], forma parte del complejo eIF4F y presenta funciones variadas en la iniciación de la traducción como colaborar en la formación de un complejo 43S estable y estimular la unión del complejo ternario a la subunidad ribosomal 40S[35], colaborar con la búsqueda del codón de iniciación, la selección del mismo, con el cargado de la subunidad mayor 60S, así como también junto con eIF1 reconoce errores de apareamiento entre el codón y el anticodón, y previene la hidrólisis prematura de GTP unido a eIF2 inducida por eIF5B[36].

A nivel de su estructura, eIF1A es un factor de iniciación conservado: su ortólogo procariota, IF1, es de una menor complejidad y consiste únicamente de un dominio de unión a oligonucleótidos/oligosacáridos (OB)[37]. La región central de eIF1A presenta un dominio OB de estructura similar a IF1, pero eIF1A presenta además un pequeño subdominio carboxi-terminal constituido por dos α hélices, así como dos largas colas no-estructuradas en sus extremos amino- y carboxi-terminal[24], denominadas NTT y CTT por sus siglas en inglés: *N-terminal tail* y *C-terminal tail*. Las NTT y CTT son importantes para la actividad de eIF1A y han sido implicados en la estimulación del reclutamiento ribosomal por parte del complejo ternario y la selección del codón de iniciación[24]. Sin embargo, las NTT y CTT presentan efectos opuestos en la selección del codón de iniciación: mientras que la CTT aumenta la estringencia en la selección del codón de iniciación y promueve la conformación “abierta” del complejo mientras que realiza el *scanning*, la NTT disminuye la precisión en la selección y promueve la conformación “cerrada ”[38].

eIF4A, como subunidad del complejo eIF4F, es el factor responsable del desarmado de estructuras secundarias presentes en los 5´-UTR de los mensajeros[39]. eIF4A es una

helicasa del tipo DEAD-box, ubicada dentro de la superfamilia de proteínas SFII, por la presencia de un motivo conservado de Asp-Glu-Ala-Asp (DEAD). Es posible que las proteínas del tipo DEAD-box utilicen ATP para generar cambios estructurales sobre sí mismas y así desarmar los dúplex de ARN[40]. De esta forma, eIF4A presenta una activad ATPasa ARN-dependiente y una actividad helicasa ATP-dependiente la cual no está restringida de forma direccional, con lo cual puede actuar tanto hacia el extremo 3´, como 5´ de la hebra de ARN[41].

Las diversas familias de helicasas presentes en la célula cumplen varias funciones dentro de las cuales se incluyen la biogénesis ribosomal, transcripción, *splicing*, exportación de mensajeros, maduración, degradación y traducción de transcriptos, así como el remodelado de partículas ribonucleoproteicas (RNP)[39]. Un caso donde se resalta la importancia de las helicasas es en el inicio de la traducción, dado que la amplia mayoría de los mensajeros celulares presentan 5´-UTR estructurados. Más aún, se podría decir que más de la mitad de los mensajeros transcriptos humanos presentan de moderadas a fuertes estructuras secundarias en sus 5´-UTR[39], lo cual repercute de modo directo en sus niveles de traducción. De esta forma, debe existir un complejo e integrado nivel de regulación capaz de determinar cuáles mensajeros deben ser sometidos a la acción de eIF4A para así desarmar sus estructuras y ser correctamente traducidos.

La reciente resolución de la estructura cristalina de eIF4A en levaduras ha permitido empezar a comprender mejor su función e interacciones con otros factores[42]. eIF4A presenta una estructura tipo mancuerna abierta, con dominios globulares conocidos como NTD y CTD (dominios amino- y carboxi-terminales respectivamente, distintos a las colas NTT y CTT presentadas más atrás) conectados por un fragmento de 11 residuos. Esta estructura se encuentra en equilibrio con una forma compacta en la cual los dos dominios interactúan uno con el otro. Tanto el ATP como el ARN, se unen a la forma compacta y desplazan el equilibrio hacia esta conformación. Las actividades helicasa y ATPasa de eIF4A son fuertemente estimuladas cuando éste forma parte del complejo eIF4F o cuando está asociado a otros factores como eIF4B y eIF4H[41]. eIF4B es un homólogo de eIF4H a nivel de su secuencia aminoacídica, aunque contiene dominios adicionales a nivel de los extremos amino- y carboxi-terminales. Estos dos

factores estimulan la actividad helicasa de eIF4A para resolver dúplex más largos y estables[43]. Ambos son capaces de modular la afinidad de eIF4A por ATP o ADP[44]. Recientemente se ha encontrado que eIF4B y eIF4H comparten un dominio de unión a eIF4A y sus interacciones con este factor son mutuamente excluyentes[45]. Estas dos proteínas también aumentan la afinidad de eIF4A por el ARN. A la vez, estos dos factores interactúan con el ARN mediante sus dominios con motivos de reconocimiento de ARN[46], y esta interacción podría ser importante para modular estructuras locales sobre el ARNm. Por ejemplo, es posible que tanto eIF4B como eIF4H estabilicen regiones de simple hebra en el 5´-UTR donde el complejo 40S se une inicialmente[47]. También se ha postulado que estos dos factores pueden prevenir que se vuelvan a formar las estructuras secundarias que recién han sido resueltas. A su vez, pueden favorecer una procesividad 5´ → 3´ para el movimiento de eIF4A[44] y esta misma direccionalidad para el complejo 43S[39], ya que como se dijo más atrás eIF4A como helicasa no tiene definida una procesividad direccional y puede actuar en ambas direcciones.

Como se mencionó, eIF4A es un componente clave de la iniciación de la traducción en eucariotas, sin embargo todavía no se conoce con detalle como factores auxiliares como eIF4B (ó eIF4H) y eIF4G estimulan eIF4A, así como esta estimulación contribuye al proceso de iniciación. A su vez, se han propuesto varios modelos para explicar las interacciones entre eIF4A y eIF4G en la formación del complejo eIF4F, a continuación se detallan los modelos aceptados hasta el momento. eIF4G se conoce como una proteína de *scaffolding* en el sentido que presenta varios sitios de unión a otras proteínas y le permite actuar como una plataforma en la cual convergen varios factores. Por ejemplo, tanto el eIF4G de mamíferos como de levaduras, contiene sitios de unión a PABP, eIF4A y eIF4E, mientras que sólo el eIF4G de mamíferos contiene sitios de unión adicionales para eIF3 y la quinasa Mnk1[48] (ver Figura 1.6). Como se decía, eIF4G presenta dominios de unión a eIF4A, particularmente eIF4G de mamíferos presenta tres dominios HEAT repetidos *(Huntingtin – Elongation factor 3* – Subunidad A de la fosfatasa 2A – Target de rapamicina) de unión a eIF4A, uno en el tercio central denominado eIF4Gm y otros dos hacia el extremo carboxi-terminal denominado eIF4Gc[49]. Este último es un dominio MA-3[50]. La interacción entre eIF4A y eIF4Gm

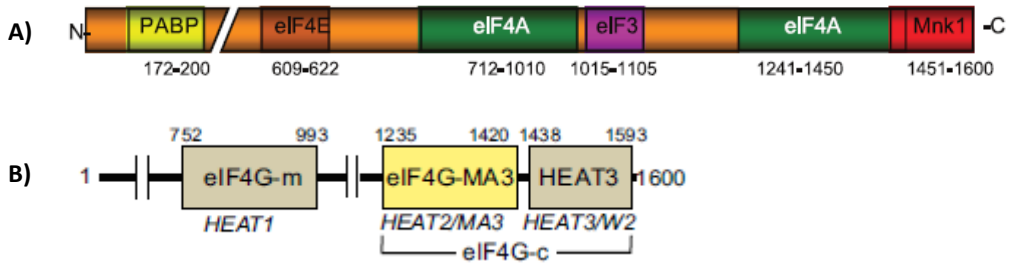


Figura 1.6: Representación estructural de eIF4G. A) En colores se observan los distintos motivos involucrados en interacciones proteína-proteína que justifican el término asignado a eIF4G como proteína de scaffolding, por la gran cantidad de interacciones que mantiene; B) Se muestran los dos principales dominios de eIF4G: eIF4G-m y eIF4G-c. Figuras adaptadas de Nielsen et al. 2010 y Suzuki et al. 2008.

es suficiente para la traducción *cap*-dependiente de mensajeros[51], y se cree que esta interacción ayuda a unir el mensajero con eIF4A y así estabilizar la orientación de los interdominios amino- y carboxi-terminales de eIF4A. También eIF4Gm ha sido implicado en interacciones tanto con el mensajero, como con eIF3[52]. Por otro lado, el dominio MA-3 eIF4Gc se piensa que jugaría un rol modulatorio, en el sentido que competiría con el mensajero por unir eIF4A y, en el caso de unirlo, estabilizaría la conformación inactiva del mismo[53]. Así, los dos dominios de unión a eIF4A de eIF4G interaccionan de formas anticooperativas.

Más allá de esto, un modelo tipo abrazadera se ha propuesto para explicar la estimulación de eIF4A por parte de eIF4G en humanos[54]. En este modelo eIF4G estabiliza una conformación particular de eIF4A que favorece su actividad. Acorde a esto, una estructura cristalina del complejo eIF4G-eIF4A en levaduras ha demostrado que eIF4A se mantiene en una conformación ligeramente abierta[55] comparada con la conocida conformación cerrada de varias ATPasas que unen ARN[39]. De esta manera, eIF4G orienta los motivos conservados de eIF4A de una forma tal que estos sean capaces de unir ARN y ATP. En esta estructura, eIF4A en ausencia de ATP, interactúa con eIF4Gm de forma débil principalmente a través su dominio carboxi-terminal (CTD). También interactúa con el dominio MA-3 eIF4Gc al cual se une fuertemente a través de sus dos dominios amino y carboxi-terminales, lo cual lo estabiliza en su conformación abierta inactiva y bloquea al menos parcialmente el sitio de unión al mensajero. Sin embargo en presencia de ATP, se rompe la interacción entre el dominio MA-3 eIF4Gc y el NTD, mientras que la interacción entre eIF4Gm y los

dos dominos NTD y CTD se estabiliza (ver Figura 1.10 parte superior). Esto favorece la conformación cerrada activa, a la cual se unen de forma cooperativa el ARN, eIF4H o eIF4B estimulando así la traducción *cap*-dependiente[49].

1.1.4) REGULACIÓN AL INICIO DE LA TRADUCCIÓN

El control de la traducción en células eucariotas, en particular de la fase de iniciación, es crítica para la regulación de la expresión génica durante la privación de nutrientes y condiciones de estrés, desarrollo y diferenciación, funcionalidad del sistema nervioso, envejecimiento y enfermedades. Por ejemplo, en condiciones de estrés o de falta de algún nutriente, los factores de iniciación de la traducción son inactivados por determinadas modificaciones post-traduccionales, lo cual inactiva la traducción para la amplia mayoría de los mensajeros celulares (ver más adelante). Sin embargo, mecanismos especializados han evolucionado para permitir que bajo esas condiciones de estrés donde la gran parte de la traducción está reprimida, determinados mensajeros que codifican para factores de transcripción particulares, por ejemplo, sean traducidos.

Uno de los mecanismos claves del control traduccional que actúa durante condiciones de estrés, es la fosforilación de eIF2 a nivel de su serina 51 en la subunidad α (eIF2-P), lo cual repercute negativamente en el ensamblado del complejo ternario[56] (eIF2, GTP y el transfer iniciador de metionina; ver sección 1.1.2). De forma paradójica, además de reducir los niveles de iniciación de traducción generales en levaduras en condiciones de privación de aminoácidos, eIF2-P induce la activación traduccional de un activador de la transcripción en levaduras, *GCN4*, ya que puede superar los efectos inhibitorios de cuatro uORFs (*upstream open reading frames*). En detalle lo que sucede es que luego de la traducción del primer uORF (uORF1), la subunidad 40S es capaz de retomar el *scanning* y reiniciar la traducción luego de unir un complejo ternario en los siguientes uORFs que se presentan a continuación (uORF2, 3 y 4). Sin embargo, cuando los niveles del complejo ternario bajan por la privación de aminoácidos por ejemplo, la subunidad 40S luego de traducir el uORF1 no encuentra complejos ternarios disponibles y recién reinicia la traducción cuando ya superó los uORFs 2-4 y se encuentra con el codón de iniciación de *GCN4.* Este factor de transcripción activa una

batería de genes encargados de la biosíntesis de aminoácidos[57]. Otro mecanismo usado de forma extensa en eucariotas es el control de la tasa de iniciación regulando el proceso de reconocimiento del 5´-*cap* por el complejo eIF4F. Este reconocimiento puede verse afectado por problemas en la interacción entre eIF4G con eIF4E. Esta interacción es inhibida por miembros de la familia de proteínas de unión a eIF4E, denominadas 4E-BPs[58], las cuales compiten con eIF4G al compartir sitios de unión a eIF4E[59]. La unión de las 4E-BP a eIF4E es controlada por fosforilación, de modo tal que las formas hipo-fosforiladas de 4E-BP unen fuertemente a eIF4E, mientras que las formas fosforiladas presentan una interacción más débil[58]. Una de las quinasas críticas que fosforila las 4E-BPs es mTOR (*mammalian target of rapamycin*). mTOR se ubica *downstream* de la quinasa Ser/Thr en la vía de señalización PI3K/Akt por lo cual es capaz de censar e integrar señales de estimulación extracelular, disponibilidad de aminoácidos y oxígeno, así como el estado energético de la célula. De esta forma, en presencia de las adecuadas señales extracelulares anteriores, mTOR se encuentra fosforilada y activa, por lo cual determina la fosforilación de, entre otros blancos, las 4E-BPs disminuyendo su atracción por eIF4E, activando en consecuencia la traducción (ver Figura 1.7). mTOR es responsable directo o indirecto de la fosforilación de importantes sustratos relevantes para la traducción, incluyendo eIF4G y la quinasa S6, la cual fosforila eIF4B en su serina 422 y esto aumenta su interacción con eIF3[60]. La quinasa S6 también fosforila PDCD4, un supresor de tumores presentado más adelante, lo cual lo conduce a su ubiquitinación y degradación por parte del proteosoma[61]. La rapamicina es la droga más común utilizada para inhibir la vía de mTOR en variadas situaciones patológicas desde cáncer hasta en el rechazo a transplantes de órganos. Los extensos estudios realizados con esta droga y su amplio uso resaltan la importancia de la vía de mTOR en un conjunto variado de procesos y funciones celulares, que incluyen el crecimiento, envejecimiento, plasticidad sináptica, memoria, metabolismo, obesidad, cáncer, etc.[62,63,64].

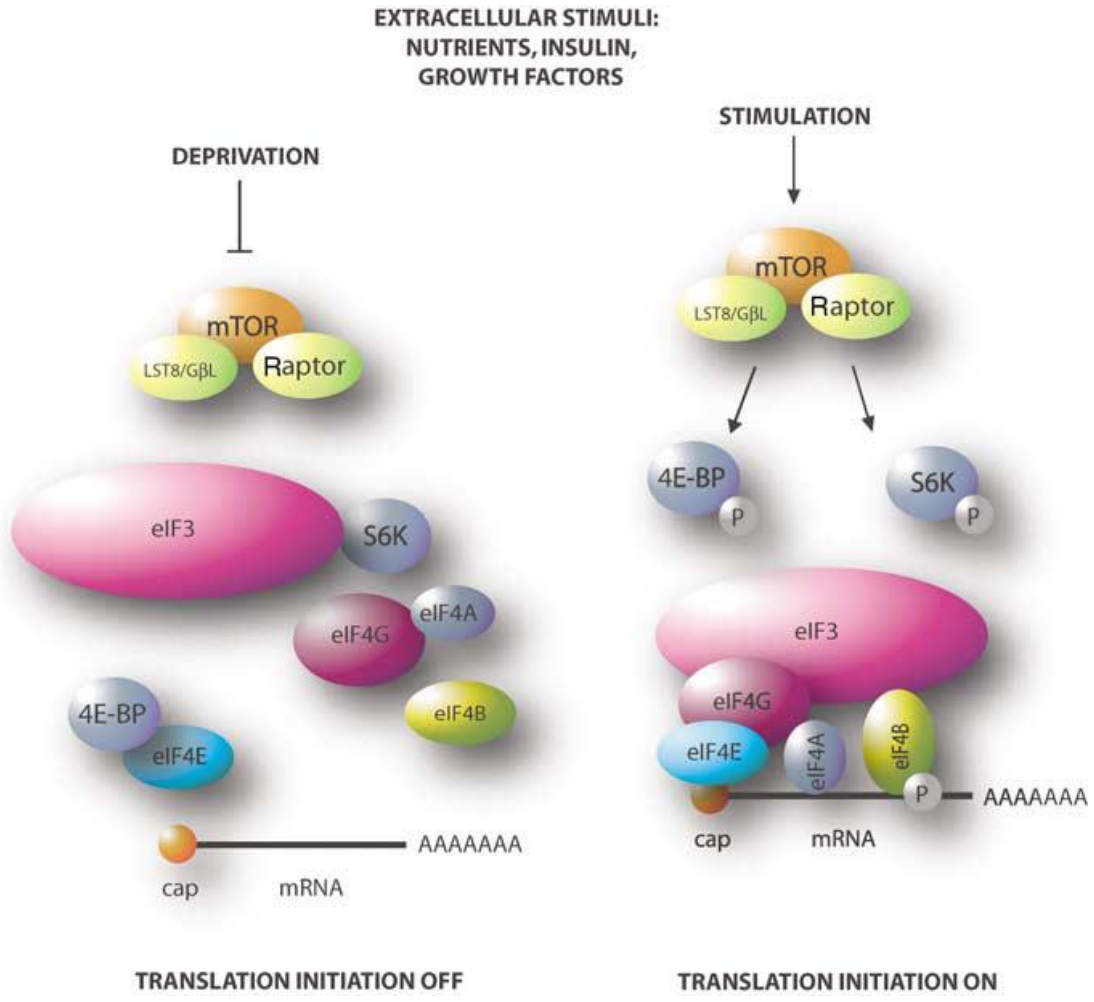


Figura 1.7: Mecanismos de activación de la traducción vía mTOR. Las señales extra-celulares (nutrientes, factores de crecimiento), activan mTOR el cual forma un complejo con otras proteínas y fosforila principalmente dos blancos: 4E-BPs y S6Ks lo cual activa la traducción por los mecanismos que se muestran (derecha). En ausencia de las apropiadas señales extra-celulares, mTOR se encuentra inactivo y por lo tanto, 4E-BPs y S6Ks se encuentran desfosforiladas, inhibiendo la traducción como se muestra (izquierda). Figura adaptada de Mamane et al. 2006.

1.2) COMPARTIMIENTO TRADUCCIONAL COMO REGULADOR DE LA EXPRESIÓN GÉNICA

Han sido varias las aproximaciones moleculares y tecnológicas que se han ido desarrollando con los objetivos de comprender los mecanismos regulatorios de la expresión génica y aproximarse a perfiles de expresión capaces de caracterizar y explicar distintas condiciones fisiológicas y/o patológicas. En este sentido, los primeros

análisis se centraron en el estudio de la transcripción y la población celular de ARNm, los cuales a groso modo se han encargado de la caracterización de la expresión génica durante las últimas dos décadas. En este contexto, los análisis por *microarrays* introducidos en la comunidad científica hace cerca de 15 años[65] revolucionaron los estudios de expresión que analizaban todo el genoma, lo cual reforzó en gran nivel el entendimiento biológico y molecular acerca de lo que ocurre en la regulación de la expresión génica, y los distintos patrones de expresión que se observan en los modelos particulares de estudio.

Sin embargo, desde una perspectiva biológica, el interés en este tipo de estudios de expresión génica se centra en determinar exhaustivamente la población de proteínas que componen el sistema de estudio, lo cual es conocido como proteoma. Sin embargo, al analizar la población total y estable de mensajeros en determinada condición celular, conjunto conocido como transcriptoma, se está todavía muy lejos de obtener una estimación que correlacione en forma adecuada con el proteoma[66,67,68,69]. Esto se debe a varios mecanismos que actúan desde que un gen se transcribe hasta que se forma una proteína estable y funcional, considerando su propia tasa de degradación. Más allá de los niveles de regulación detectados y considerados cuando se trabaja a nivel del transcriptoma (síntesis de transcriptos, estabilidad de los mismos y regulación a nivel de *splicing* alternativo), existen otros mecanismos posteriores que afectan la correlación entre transcriptoma y proteoma. Entre ellos se encuentran el transporte de mensajeros, la eficiencia y control traduccional, y la estabilidad de proteínas[70]. También existen otros mecanismos más puntuales como por ejemplo, el hecho de que no es posible una predicción exacta del producto proteico a partir de la secuencia nucleotídica del mensajero debido a la presencia de IRES (*internal ribosome entry sites)*, iniciación en codones no tradicionales (no AUG)[71] y otros mecanismos propios de la traducción, como por ejemplo la existencia de pausas ribosomales programadas que detienen la síntesis proteica momentáneamente hasta que determinada señal celular la re-active[16].

Estudios del genoma y su expresión han mostrado que la eficiencia traduccional aparece como uno de los mecanismos claves para entender las diferencias entre la cantidad estimada de ARNm y la población de proteínas estable[70]. Por esto debe ser

considerada como un factor clave en la determinación de los niveles proteicos estables.

Uno de los ejemplos que señalan a la regulación traduccional como un compartimiento de considerable importancia para determinar la expresión génica, son los estudios relacionados al factor general de iniciación traduccional eucariótico eIF4E. Este factor es la limitante del reclutamiento ribosomal y de la eficiencia traduccional y se pudo mostrar en los mencionados estudios, una relación directa entre este factor de expresión ubicua y el cáncer[72].

Con el tiempo más evidencia ha sido acumulada a favor de considerar el compartimiento traduccional como un componente clave de los mecanismos efectores de la regulación génica. Esto implica cambiar el punto de vista mediante el cual entendemos lo que ocurre en condiciones normales respecto a la expresión génica y sus mecanismos regulatorios, y también respecto a lo que ocurre cuando dichos mecanismos se desregulan causando condiciones patológicas severas como el cáncer[70]. Estas evidencias han hecho notar de forma consistente que la magnitud de la regulación originada en base a fenómenos implicados en la traducción supera ampliamente la regulación proveniente del nivel transcripcional, particularmente en aquellos estudios donde estos dos niveles se estudiaron por separado en la misma condición.

A continuación se plantea brevemente una serie de ejemplos de estudios que resaltan la significancia de lo anterior. El primero de estos ejemplos abarca estudios acerca del transcriptoma y traductoma (mensajeros asociados a maquinaria traduccional activa) en un modelo de progresión de cáncer colorrectal[73]. En este trabajo los autores mostraron que la expresión génica diferencial se observaba en promedio cuatro veces mejor a nivel traduccional que a nivel transcripcional. Los autores encontraron un incremento en la forma hiperfosforilada de 4E-BP1 en el modelo más avanzado de cáncer, lo cual resultaría en un incremento de la actividad de eIF4E, el cual activa la traduccional *cap*-dependiente. Los autores concluyen que la principal diferencia entre la condición tumoral primaria y avanzada radica en la habilidad de traducir, y no de transcribir, ARNm, gracias a la activación del factor de iniciación eIF4E. También se

observo que mientras algunos procesos eran regulados en los dos niveles, otros como la apoptosis, solamente se regulaba a nivel traduccional.

Un segundo ejemplo cubre el área de estudio del estrés celular donde se han realizado varios estudios de reclutamiento ribosomal. En éste ámbito se ha encontrado una importante regulación a nivel del reclutamiento e inicio de la traducción en condiciones donde se inducen cambios en la expresión génica mediante radiación ionizante[74]. Estos autores encuentran que diez veces más genes son afectados a nivel traduccional respecto al nivel transcripcional luego de exponer a células tumorales de cerebro humano a radiación ionizante. Otro estudio al respecto involucra distinta formas de estrés celular, por ejemplo la hipoxia en células HeLa[75]. También un estudio más reciente, analiza los cambios a nivel traduccional y transcripcional, en condiciones de privación de aminoácidos como elemento de estrés celular[76], en levaduras mediante una nueva tecnología que denominan *ribosome profiling (*ó *ribosome footprinting)*, la cual se detallará más adelante (ver sección 3.2.2), donde los autores encuentran las mismas características de regulación sesgada al control traduccional que los ejemplos anteriores.

La lista de ejemplos recopilados puede seguir extendiéndose, donde cada vez más áreas de estudio y funciones celulares son sometidas a este tipo de análisis. Sólo por mencionar algunos casos donde la regulación a nivel de la traducción ha sido demostrada como factor clave, se pueden mencionar los fenómenos de diferenciación celular[77], desarrollo celular estudiado en un modelo de embriogénesis de *Drosophila*[78], espermatogénesis en los gametos masculinos[79], aprendizaje y la memoria en mamíferos etc.[80]. También se encuentran aquellos genes críticos que requieren un control preciso, como los relacionados al crecimiento y sobrevida celular. La expresión de estos debe ser regulada cuidadosamente ya que su desregulación termina por afectar varias vías celulares que se ven generalmente desreguladas en cáncer[81]. Por último se encuentran los sistemas biológicos donde no existe control transcripcional, por ejemplo en reticulocitos, ovocitos y virus ARN, en los cuales la oportunidad de ejercer un control transcripcional es mínima o nula, por lo cual la expresión génica debe ser regulada únicamente a nivel traduccional. También, un fino control de la traducción posibilita la regulación del sitio de síntesis proteica dentro de

la célula mediante el cual se generarían gradientes de proteínas. Estos gradientes se sabe que afectan la eficiencia traduccional de otros mensajeros específicos, con lo cual se podrían generar patrones de traducción, lo cual es clave por ejemplo en el desarrollo temprano.

Estas evidencias presentadas hasta el momento apuntan claramente a considerar la traducción como un fenómeno determinante en la modulación de la expresión génica. La traducción tiene la ventaja de ser uno de los últimos actores capaces de determinar y regular la población celular proteica que explica los variados fenotipos encontrados en los distintos modelos de estudio. A la vez, esta posición privilegiada la convierte en un sitio importante de convergencia de varias vías celulares.

Respecto a este último punto, es sabido que funciones aberrantes de los componentes propios de la maquinaria traduccional subyacen una gran variedad de enfermedades humanas incluyendo ciertos tipos de cáncer y desordenes metabólicos[56]. La mayoría de los cáncer son causados por desregulación a nivel de las vías de señalización celulares que controlan el crecimiento y la proliferación celular, y estas vías también afectan la traducción. Particularmente, el cáncer se encuentra asociado a cambios aberrantes en las cantidades y actividad de los factores de iniciación, factores de regulación traduccional y ARNt[82].

2) *PDCD4* Y SUS FUNCIONES EN CÁNCER

2.1) *PDCD4* COMO GEN SUPRESOR DE TUMORES

Programmed Cell Death 4 (PDCD4) es un gen supresor de tumores el cual fue inicialmente identificado en una búsqueda de genes cuya expresión aumentaba durante la apoptosis[83]. Estudios posteriores mostraron que *PDCD4* es realmente un gen supresor de tumores involucrado en la regulación de la transcripción y traducción de ARNm, en las vías de transducción de señales celulares y en la apoptosis, entre otros[84]. Ha sido demostrado que PDCD4 suprime la transformación celular en un

modelo de progresión tumoral *in vitro* en queratinocitos de ratones[85], así como también se vio que inhibe la formación tumoral en un modelo de carcinogénesis cutánea *in vivo* también en ratones[85]. A su vez, recientemente ha sido reportada la importancia funcional de PDCD4 en células tumorales humanas[86]. Se ha observado que *PDCD4* presenta bajos niveles de expresión en numerosos tipos de cáncer humanos, como los cáncer de pulmón, mama, colon, hígado, páncreas y ovario, entre otros[86]. También una baja o nula expresión de *PDCD4* ha sido asociada a la progresión tumoral, movilidad e invasión de células tumorales y metástasis[87,88]. Más puntualmente, en un reciente trabajo se han encontrado en muestras de tumores de estroma gastrointestinal bajos niveles de PDCD4 tanto a nivel de ARNm, como a nivel proteico, respecto a muestras de tejidos normales[86]. En este mismo estudio también se han asociado niveles alterados de expresión de *PDCD4* con parámetros clínico-patológicos incluidos el grupo de riesgo, el tamaño del tumor y la mitosis. Por otro lado se ha visto que una baja expresión de *PDCD4* afecta la respuesta celular al daño al ADN[89,90].

PDCD4 codifica para una proteína muy conservada en vertebrados[84]. El gen humano está ubicado en la banda cromosómica 10q24[91] y codifica para las dos isoformas encontradas. La variante más abundante es una proteína de 469 aminoácidos mientras que la segunda variante presenta un tamaño menor. Esta segunda variante utiliza un codón de iniciación *down-stream* y un exón alternativo que se encuentra en marco de lectura[92].

PDCD4 consta de tres dominios, uno amino-terminal de unión al ARN y dos dominios MA-3 de interacción proteína-proteína, uno central (MA-3m) y otro carboxilo-terminal (MA-3c)[84]. A nivel de secuencia, la proteína cuenta con varios sitios plausibles de modificaciones post-traduccionales, como pueden ser fosforilaciones que regulan su actividad, degradación y localización subcelular (ver Figura 1.8). Respecto a ésta última, se especula que la proteína se encuentra en un continuo tráfico núcleo-citoplasma, ya que se ha reportado tanto una localización nuclear en tejidos normales y una citoplasmática en tumores[93], como también se ha observado lo opuesto: localización citoplasmática en tejidos normales y nuclear en tumores[94].

Figura 1.8: Representación estructural de PDCD4. Se observan los tres dominios de PDCD4: el dominio de unión a ARN a la izquierda, y los dos dominios de interacción proteína-proteína MA-3. También se observan los dos principales residuos plausibles de modificaciones post-traduccionales. Figura adaptada de Lankat-Buttgereit et al. 2009.

PDCD4 es una proteína expresada en forma ubicua en los tejidos, pero se ha visto que su expresión se ve disminuida en tumores y en apoptosis. Aún así dependiendo del inductor apoptótico utilizado, su expresión puede verse incrementada[83] o disminuida[95]. Esto sugiere que la expresión de dicha proteína está sujeta a una compleja red de regulación dependiente de contextos celulares específicos. Algunos de los mecanismos que controlan su expresión son conocidos: el mensajero de PDCD4 es blanco de dos microARN que regulan su traducción en condiciones fisiológicas alteradas. miRNA-21 inhibe la traducción de PDCD4 en tumores, particularmente sólidos[96]. Recientemente se ha encontrado otro microARN, miARN-199a-5p, el cual disminuye los niveles traduccionales de PDCD4 en células inducidas a apoptosis[92]. También como se comentó anteriormente la secuencia aminoacídica de PDCD4 incluye dos sitios de fosforilación particulares, las serinas 67 y 457[97] (ver Figura 1.8). Al respecto, la fosforilación de PDCD4 por Akt y S6K1 en la serina 67 desencadena su posterior degradación mediante la ubiquitin-ligasa $SCF^{\beta TrCP}$ por la vía proteosomal[61]. Los niveles proteicos de PDCD4 no son solo regulados por los mecanismos descritos anteriormente (microARN y degradación proteosomal disparada por fosforilación), sino que también existe regulación a nivel de la transcripción y mecanismos epigenéticos involucrados. Al respecto se puede aportar brevemente que se ha encontrado que el factor de transcripción v-Myb es un inductor de la transcripción de *PDCD4*[98]. Mientras que, a nivel epigenético, el grado de metilación del ADN también podría regular la expresión de *PDCD4*. Se ha visto que la desmetilación global del ADN induce su expresión[99], y también la metilación de las islas CpG hacia el extremo 5´ del gen han sido asociadas con niveles reducidos de mensajero de *PDCD4* en líneas celulares humanas de glioma[100] (ver Figura 1.9).

Respecto a las funciones de PDCD4 es sabido que interviene en la transformación neoplásica donde se manifiesta como gen supresor de tumores[84]. Esta función la desarrolla tanto a nivel transcripcional como traduccional. A nivel transcripcional se ha visto que PDCD4 regula la actividad de varios factores de transcripción como c-Jun[101], sp1[87] y p53[89], afectando entonces los niveles transcripcionales de otros varios genes.

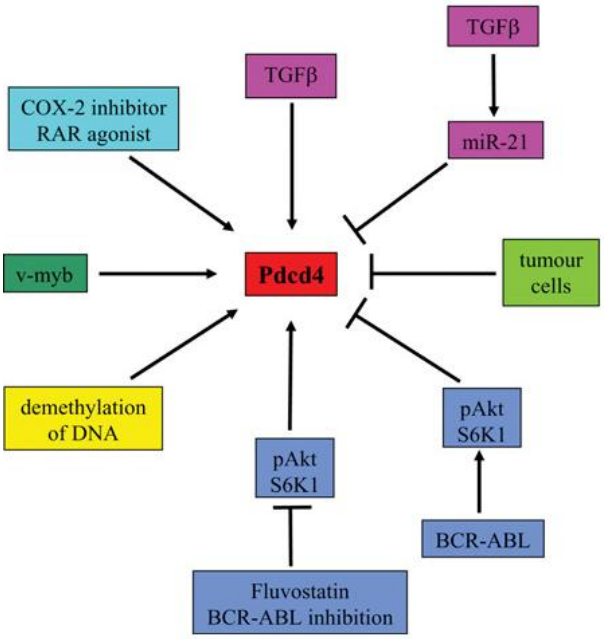


Figura 1.9: Mecanismos generales que regulan los niveles de PDCD4. T *indica inhibición y* ↑ *indica estimulación.*

Sin embargo, tanto el mecanismo por el cual actúa PDCD4 a nivel traduccional y cuáles son sus blancos traduccionales, no se conocen con certeza. Son varios los estudios que proponen distintos mecanismos para explicar la acción de PDCD4 en la maquinaria traduccional. Los trabajos más actuales proponen que son necesarias interacciones ARN-proteína y proteína-proteína para que PDCD4 reconozca estructuras secundarias en los 5´-UTR o reconozca una secuencia dentro de la región codificante e inhiba la traducción de forma específica de ciertos ARNm. Se ha visto que PDCD4 interacciona tanto con los factores de iniciación eucarióticos eIF4A y eIF4G así como con proteínas del *scaffolding* que se encuentran en el complejo de iniciación 40S y complejo de preiniciación 43S[50,102,103]. En condiciones normales eIF4G une eIF4A e interacciones adicionales activan este último factor para que manifieste su actividad de ARN helicasa. Esta permite resolver estructuras secundarias en los 5´-UTR que inhiben la traducción de los ARNm, involucrados principalmente en la oncogénesis. Esto activa la traducción *cap*-dependiente de esos ARNm. Evidencias recientes apuntan hacia el hecho de que PDCD4 compite con eIF4G y con el ARNm por la unión a eIF4A, pudiendo lograrlo en determinadas condiciones aunque sin desplazar por completo a eIF4G del complejo de iniciación en formación[53]. Esta unión de PDCD4 a eIF4A, que

parecería ser de gran estabilidad, bloquearía el sitio de unión de eIF4A al ARN, estabilizando la conformación abierta inactiva del complejo de iniciación[53] (ver Figura 1.10 y sección 1.1.3). De esta forma, la interacción de PDCD4 con eIF4A inhibe su actividad helicasa, lo cual de forma directa inhibe la traducción *cap*-dependiente de ciertos mensajeros que aún no se conocen con detalle.

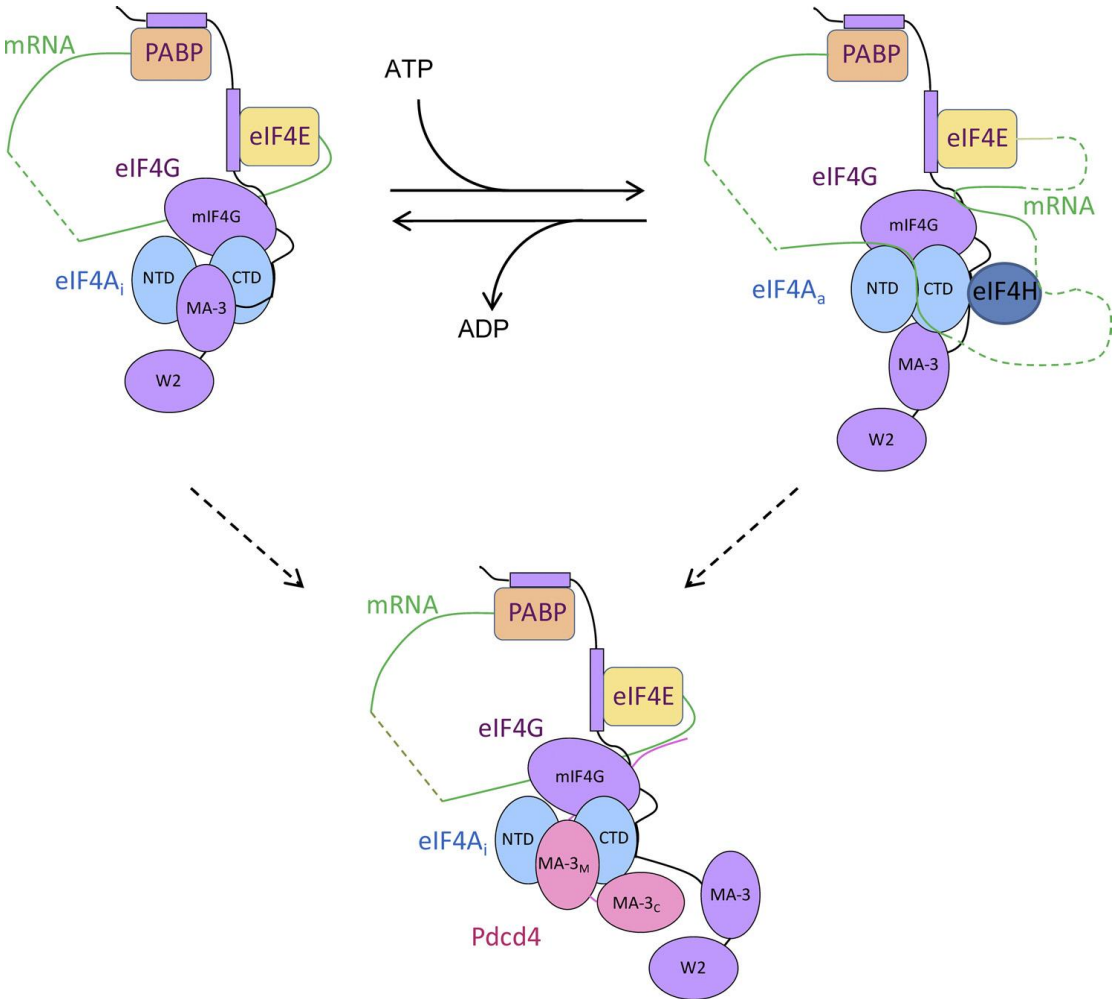


Figura 1.10: Representación esquemática del mecanismo de acción propuesto para PDCD4. En este modelo simplificado, en ausencia de ATP la interacción entre eIF4Ay el dominio MA-3 de eIF4G estabiliza la conformación abierta inactiva de eIF4A (eIF4Ai), y bloque al menos en parte el sitio de unión a ARN de eIF4A. Además esta interacción disminuye la afinidad de eIF4A por ATP y ADP cerca de tres veces, favoreciendo este estado libre de nucleótidos. En presencia de ATP, ARN, eIF4Gm, y eIF4, estos se unen de forma cooperativa a eIF4A promoviendo su conformación cerrada activa (eIF4Aa), y estimulan la actividad helicasa de eIF4A. Se ha propuesto que PDCD4 bloquea el sitio de unión a ARN de eIF4A, lo cual previene el cambio conformacional requerido para la formación del complejo activo. Interacciones adicionales entre PDCD4 y eIF4Gm podrían ayudar a mantener atrapado a eIF4A en su conformación inactiva, inactivando así todo el complejo eIF4F. Figura adaptada de Waters et al. 2011.

2.2) *PDCD4* COMO REGULADOR DE LA TRADUCCIÓN

Actualmente se conoce la existencia de tres blancos traduccionales de PDCD4. Esto implica que los ARN mensajeros de dichos blancos, son reconocidos de forma específica por PDCD4 y su traducción es inhibida. Estos tres blancos traduccionales encontrados son los mensajeros de c-Myb[104], procaspasa-3[92] y p53[105]. Genes que se encuentran implicados de forma directa en procesos tumorales y apoptóticos.

- c-Myb: El gen *c-myb* codifica para un factor de transcripción clave para la hematopoyesis[106], así como para el desarrollo normal de los linajes mieloides y eritroides y la diferenciación de células B y T[107]. La actividad de *c-myb* como factor de transcripción es sumamente variable y depende tanto de modificaciones post-traduccionales, como de los eventos de *splicing* alternativo y el importante uso de exones alternativos que ocurre en este gen. Estas variaciones pueden convertirlo en una potente oncoproteína capaz de transformar células[107], por lo cual además de su actividad como factor de transcripción también es descrito como un oncogen humano[108].

En este caso en particular, los autores mostraron mediante ensayos de *Western blot, Northern blot,* fraccionamiento subcelular y ensayos con inhibidores del proteosoma, que PDCD4 es capaz de inhibir la traducción del mensajero de *c-myb* sin afectar sus niveles transcripcionales, su estabilidad ni su exportación núcleo-citoplasma. Este resultado también fue confirmado analizando las cantidades de mensajero de *c-myb* en las distintas fracciones ribosomales, donde se vio que en presencia de PDCD4 las cantidades de mensajero de *c-myb* en la fracción polisomal disminuyen significativamente. Por otro lado, mediante ensayos de co-inmunoprecipitación y qRT-PCR para detectar el ARN co-inmunoprecipitado, los autores demostraron la asociación directa de PDCD4 con el mensajero de *c-myb in vivo.* Por último generando distintas deleciones en el mensajero de *c-myb* y aplicando los métodos anteriores de co-inmunoprecipitación y detección del ARN co-inmunoprecipitado por qRT-PCR, los autores mostraron que la región de unión de PDCD4 al mensajero de *c-myb* se encuentra dentro de la secuencia codificante de dicho mensajero[104].

- Procaspasa-3: Por otro lado, se identificó a PDCD4 como un regulador de la apoptosis ya que se demostró que era capaz de inhibir la traducción del mensajero de la procaspasa-3[92]. Las caspasas son una familia de cisteín-proteasas encargada de ejecutar la apoptosis, en las cuales convergen las distintas vías que se inician en los fenómenos de muerte celular programada[109]. Las caspasas son expresadas como zimógenos (procaspasas) de forma tal que son activadas por proteólisis limitada una vez se ha inducido la apoptosis.

En este caso los autores observaron que PDCD4 regulaba la sensibilidad a la apoptosis de las células. Para encontrar como ejercía dicha regulación, analizaron que factores apoptóticos aumentaban su expresión al silenciar PDCD4. Resultados preliminares mostraron que la pérdida de PDCD4, aumentaba la traducción del mensajero de la procaspasa-3. Para confirmar que PDCD4 inhibe la traducción del mensajero de la procaspasa-3 los autores trabajaron en un sistema libre de células donde realizan la traducción de dicho mensajero, en presencia o ausencia de PDCD4. Luego mediante *Western blot,* co-inmunoprecipitación y detección del ARN co-inmunoprecipitado por RT-PCR, se puede concluir que PDCD4 suprime la traducción de forma específica del mensajero de la procaspasa-3 y que existe una interacción directa entre ellos.

Los autores también demostraron un nuevo mecanismo de represión traduccional ejercido sobre PDCD4 y operado por un microARN, el cual es capaz de explicar los bajos niveles proteicos de PDCD4 encontrados en condiciones de estimulación apoptótica[92]. Ese nuevo microARN, miARN-199a-5p, derivó de análisis computacionales preliminares y fue confirmados mediante ensayos donde se transfectan células con dicho microARN y luego se cuantifican los niveles proteicos de PDCD4 por *Western blot.*

- p53: Por último también se demostró que el mensajero de p53 es otro blanco fisiológico de PDCD4[105]. En este caso, los autores demostraron que este efecto inhibitorio de PDCD4 depende de su habilidad de interactuar con eIF4A y está mediado por la región 5´-UTR de p53, la cual se sabe que forma una estructura secundaria estable particular[110]. A su vez, también observaron que la expresión de PDCD4 se

encontraba disminuida luego del tratamiento con agentes de daño al ADN, involucrando de forma directa a PDCD4 con la respuesta celular al daño en el ADN.

La proteína p53 es un factor clave regulador de la sobrevida y muerte celular cuyas funciones se encuentran finamente controladas en varios niveles. Ha sido bien establecido que p53 participa en el balance entre la muerte celular, la senescencia y la sobrevida celular en respuesta a varios agentes genotóxicos. A la vez hay evidencias claras de que p53 presenta roles en la regulación del metabolismo energético celular, funciones antioxidantes, autofagia, invasión y movilidad, angiogénesis, diferenciación, necrosis e inflamación[111,112].

En este caso los autores demostraron por co-inmunoprecipitación y RT-PCR para detectar el ARN co-inmunoprecipitado, que PDCD4 interacciona con el mensajero de p53. También por PCR cuantitativa en tiempo real de las distintas fracciones ribosomales demostraron que al silenciar PDCD4 la cantidad de mensajero de p53 en la fracción polisomal aumentaba. Con esto, los autores concluyen que PDCD4 se asocia al mensajero de p53 y suprime su traducción. Por otro lado, los autores también observaron que los niveles proteicos de PDCD4 disminuyen luego de tratar las células con agentes de daño al ADN como luz UV por ejemplo. Esto sugiere un modelo en el cual PDCD4 suprime la traducción de p53 en ausencia de daño al ADN, manteniendo bajos niveles de p53 en condiciones normales. En presencia de daño al ADN, los niveles de PDCD4 disminuyen lo cual incremente de forma directa los niveles de p53 necesarios para mantener la homeostasis de la célula (ver Figura 1.11).

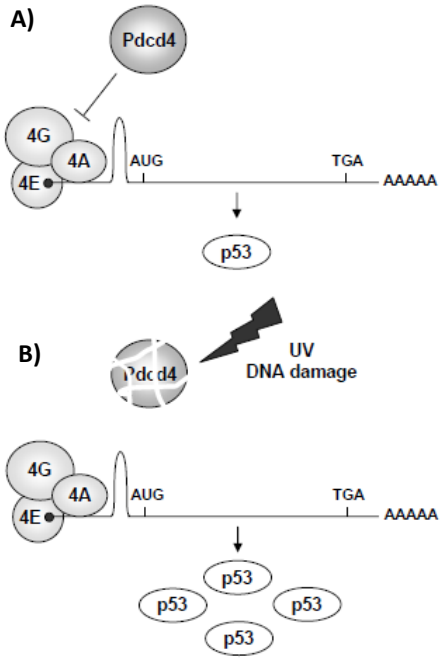


Figura 1.11: Modelo esquemático propuesto para explicar la función de PDCD4 en la inhibición de la traducción del mensajero de p53, en condiciones normales (A) y luego de la inducción de daño en el ADN (B). Figura adaptada de Wedeken et al. 2011

3) MÉTODOS DE ESTUDIO DE EXPRESIÓN GÉNICA ENFOCADOS A LA TRANSCRIPCIÓN Y TRADUCCIÓN DE ARN MENSAJEROS.

3.1) MÉTODOS CLÁSICOS

La determinación de la identidad y cantidad de proteínas producidas en una situación biológica determinada es información muy preciada para comprender diversos aspectos fisiológicos de un sistema en estudio. Con los objetivos de entender los distintos fenotipos observados, el estudio de la expresión génica implica entre otras cosas, determinar la población total de proteínas del sistema en estudio, incluyendo el estudio de los niveles de ARNm. Los métodos actuales conocidos hoy en día para la determinación de proteínas son muy variados y diversos e incluyen desde los más básicos con poca sensibilidad, como medidas espectrofotométricas de absorbancia, hasta técnicas más complejas y modernas como espectrometría de masas. Sin embargo, la identificación y caracterización de todas las proteínas individuales producidas por una célula (en el orden de varios miles proteínas) mediante espectrometría de masas puede requerir de mucho entrenamiento. También la sensibilidad de la técnica no permite siempre detectar proteínas presentes en bajas cantidades, así como tampoco es capaz de detectar pequeños cambios en los niveles de determinadas proteínas[113,114], lo cual es a veces el principal fundamento para explicar un fenotipo en particular. Por ésta razón es continuo el esfuerzo llevado a cabo para desarrollar técnicas nuevas, que permitan poder alcanzar visiones alternativas del proteoma celular. Una aproximación utilizada con frecuencia para lograr esto es la identificación masiva de los ARN mensajeros presentes en un momento dado de la vida de un sistema biológico por *microarrays* de ADN[115]. El set completo de transcriptos de una célula en un estado fisiológico y de desarrollo particular, así como sus cantidades, es conocido como transcriptoma[116] y representa una conexión importante entre la información codificada en el ADN y el fenotipo. Al respecto, la tecnología de expresión génica cuantificada por *microarrays* ha tenido resultados muy importantes acerca de como el transcriptoma cambia en función del

tipo celular y tejido, y de cómo cambia la expresión génica a lo largo del desarrollo y en fenotipos alterados[117].

Respecto a la metodología de esta técnica, los *microarrays* constan de un set de oligonucleótidos (aproximadamente de 60 nucleótidos de largo) conocidos como sondas, los cuales se encuentran inmovilizados a un soporte sólido llamado *chip.* Este *chip* es sometido a una hibridación con el conjunto de transcriptos extraídos de la muestra biológica que se está estudiando. Antes de la hibridación, estos fragmentos son marcados con determinado fluorescente. De esta forma, luego de hibridar, por complementariedad de bases y reacciones de color, puede ser detectada la presencia de algún mensajero en particular y la intensidad de luz puede ser entendida como una medida de la expresión génica (ver Figura 1.12). Así los *microarrays* de ADN son una técnica dependiente de conocimientos previos ya que se debe conocer la secuencia del transcripto cuya presencia se está cuestionando, o parte de ella, para diseñar la sonda del *chip.* De todas formas, su principal ventaja es ser una técnica cuantitativa, capaz de generar información de forma masiva[117,118].

La información generada por *microarrays* acerca de la identidad y cantidad relativa de todos los ARNm presentes en una muestra biológica en un momento dado[115], es sumamente útil en cuanto a estudiar la transcripción del genoma en una célula, pero es solo aproximado o indicativo de que proteínas van a ser traducidas a partir de dichos mensajeros observados. Esto es debido mayoritariamente, a la extensa regulación post-transcripcional ya mencionada. La disparidad entre la predicción del proteoma por parte del transcriptoma identificado y cuantificado por *microarrays* y el proteoma en sí determinado por espectrometría de masas se observa en que el coeficiente de correlación (R^2) entre ambos es, en algunos casos, tan bajo como 0,17[76].

Una mejor aproximación al estudio del proteoma celular es identificar y cuantificar los mensajeros presentes en la maquinaria traduccional activa, o sea en polisomas[119]. De esta forma se analizan solamente los transcriptos que se están traduciendo y generando proteínas, lo cual mejora la correlación con el proteoma. Sin embargo, no se han eliminado problemas o desventajas inherentes a la técnica como el

requerimiento previo de información, su deficiencia para detectar variantes de *splicing,* y el hecho de que se detecte la cantidad de ácidos nucleicos presentes en la muestra de forma indirecta, así como su menor sensibilidad respecto de las técnicas de secuenciación masiva como se comentará más adelante.

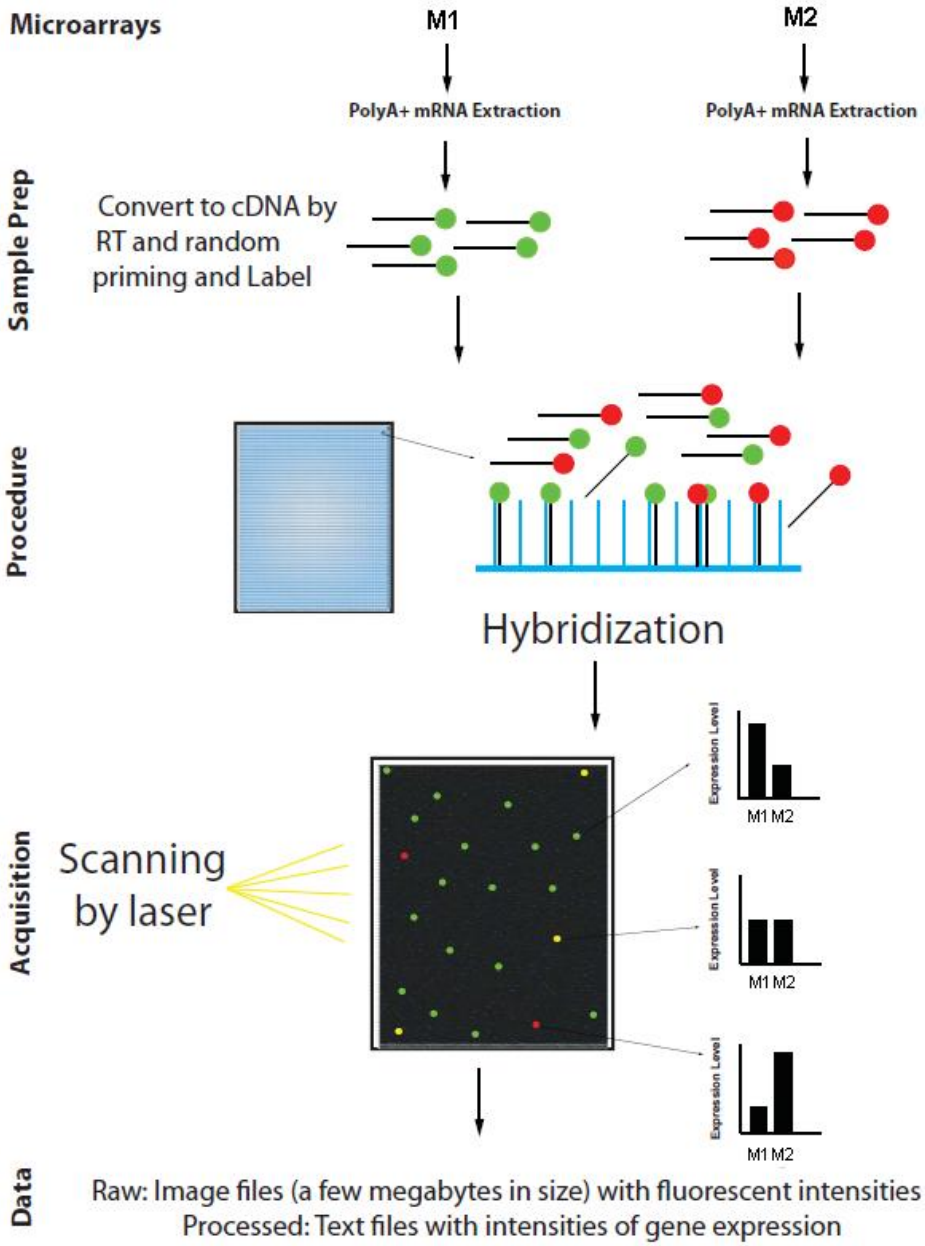


Figura 1.12: Esquema de flujo de trabajo y producción de datos en un experimento de microarrays de ADN. Estos experimentos requieren el marcado fluorescente del material a analizar, hibridaciones, lavados y un scanning con un láser adecuado para obtener medidas de expresión génica. Si la muestra consta de ARN, primero debe ser convertida a ADNc. De esta forma los genes más expresados se verán reflejados en más sitios de hibridación y mayor intensidad de fluorescencia. Figura adaptada de Malone and Oliver 2011

3.2) NUEVAS METODOLOGÍAS: SECUENCIACIÓN MASIVA DE ARNm

Las tecnologías de secuenciación masiva de moléculas de ADN permiten que un fragmento de ADN sea secuenciado repetidamente en un muy pequeño lapso de tiempo. Este procedimiento permite un incremento en la sensibilidad y precisión de la información generada[117]. Estas técnicas en particular han sido modificadas para su aplicación en al análisis del transcriptoma, lo cual se denomina RNA-Seq[120]. En lugar de recurrir a la hibridación molecular para capturar los transcriptos de interés, como ocurre en *microarrays*, en los análisis mediante RNA-Seq los transcriptos se detectan por secuenciación directa del ADN copia generado a partir de ellos. Así, las secuencias de los transcriptos son mapeadas contra un genoma de referencia, y las lecturas mapeadas son contadas para evaluar el nivel de expresión génica. De esta forma la expresión génica es entendida como el número de lecturas mapeadas en un gen o región genómica. Particularmente la expresión génica es comúnmente expresada en unidades de RPKM (*Reads Per Kilobase of exon model per Million mapped reads*). El valor de expresión génica evaluado en RPKM es una medida de la densidad de lecturas a lo largo del gen, lo cual refleja la concentración molar inicial de dicho transcripto en la muestra, ya que normaliza el número de lecturas por el largo exónico del gen y por el total de lecturas mapeadas.

Existen varias cosas que los análisis del transcriptoma mediante RNA-Seq pueden hacer que los *microarrays* no. Como el RNA-Seq nos aporta un acceso directo a la secuencia, las uniones entre los distintos exones pueden ser detectadas sin un conocimiento previo de la estructura génica. También pueden ser detectados eventos de edición de ARN y el conocimiento *a priori* de los polimorfismos, puede aportar una medida directa de la expresión alelo-específica. Debido a que los *microarrays* son diseñados en base a inferencias derivadas de una secuencia genómica conocida de ante mano y se utiliza la intensidad de la luz como medida de la expresión génica, éstos no son capaces de responder las interrogantes planteadas anteriormente[117]. Por último, como el RNA-Seq nos provee un acceso directo a la secuencias, esta técnica puede ser utilizada en especies donde el genoma entero de las mismas no se encuentra disponible, mediante análisis *de novo*. En este caso la única opción en el contexto de los *microarrays* sería hibridizar las moléculas de ARN en un *chip* diseñado

para otra especie cercana, asumiendo las limitaciones generadas por la divergencia a nivel de secuencia.

Otra característica importante de los análisis mediante RNA-Seq es su capacidad de cuantificar isoformas individuales de los distintos transcriptos[121]. El *splicing* alternativo, mecanismo por el cual se generan distintas isoformas de una misma proteína, se sabe que es una importante fuente de diversidad funcional en eucariotas. Sin embargo, ha sido poco estudiado a nivel del transcriptoma principalmente por la dificultad de medidas de expresión capaces de discernir entre las distintas isoformas. Experimentos de *microarrays* para *splicing* han sido diseñados pero requieren sondas capaces de reconocer las uniones exón-exón, y esto solo puede ser generado si el gen y las distintas isoformas producidas son conocidas *a priori*[122]. En contraste, la secuenciación masiva y análisis mediante RNA-Seq nos aportan un acceso directo a las secuencias que mapean en las uniones, y esto en teoría nos permite estudiar la expresión diferencial de las isoformas para un gen y comparar su diversidad y abundancia[123]. Por ejemplo mediante análisis por RNA-Seq, se confirmaron en humanos 31.618 eventos de *splicing* conocidos y se identificaron 379 eventos nuevos[124]. También se identificaron correctamente los extremos 5´ y 3´ de los genes, y los resultados arrojados han sugerido la existencia de un gran número de regiones del genoma transcriptas, ya sea para *A. thaliana*[125], ratón[120], humano[124], *S. cerevisiae*[126] y *S. pombe*[127]. Estas nuevas regiones transcriptas, combinadas con muchas variantes de *splicing* encontradas que no se conocían, sugieren que existe una complejidad mayor en el transcriptoma respecto de la apreciada previamente.

Sin embargo, las características de un RNA-Seq no son siempre ventajosas respecto a *microarrays.* Por ejemplo la mayor ventaja actual de *microarrays* es su relativo bajo costo comparado con la secuenciación, que cuesta aproximadamente 10 veces más que los *microarrays*[117]. Otra de las principales ventajas de *microarrays* es el conocimiento del sesgo de los datos y de las diversas estrategias y diseños experimentales planteados para solucionarlos. Por su lado, las fuentes de sesgo en los datos provenientes de secuenciación masiva están todavía en una intensa búsqueda, a

la vez que se desarrollan continuamente estrategias óptimas para perfeccionar los análisis[128].

La capacidad tecnológica de secuenciar muestras de ADN ha jugado un rol más que importante en los avances de la biología molecular. Más allá del caso puntual de análisis mediante RNA-Seq para el estudio del transcriptoma, las tecnologías de secuenciación masiva han alcanzado el estudio del genoma, epigenoma y metagenoma. Dentro del genoma, se destaca la secuenciación *de novo* de grandes genomas eucarióticos[129], así como el re-secuenciado de todo el genoma humano en busca de mutaciones puntales, número de copias y variaciones estructurales, entre otros[130,131]. Respecto al epigenoma se han generado varios avances respecto a la identificación de factores de transcripción y sus blancos directos[132] y patrones de metilación del ADN[133]. Por último respecto al metagenoma, las contribuciones más importantes se han hecho en el área del ambiente[134] y en el área del microbioma humano[135].

3.2.1) *NEXT GENERATION SEQUENCING* (NGS)

Como una de las tecnologías claves en la investigación biológica, la secuenciación del ADN no solo ha incrementado su productividad en los últimos años de manera exponencial, sino que también se ha expandido a nuevas áreas con diversas aplicaciones. Esto se debe principalmente a la llegada de nuevas generaciones de plataformas de secuenciación que ofrecen un camino más rápido y barato para generar y analizar secuencias[136]. Puntualmente las tecnologías de *Next Generation Sequencing* corresponden a la segunda generación de técnicas de secuenciación de ácidos nucleicos.

A continuación se describen brevemente dos de los instrumentos correspondientes a las tecnologías de NGS, los cuales fueron utilizados para generar las secuencias que se utilizan en el capítulo 3 y 4. Estas son las plataformas de *Illumina Genome Analyzer,* y *Life Technologies SOLiD System* respectivamente. Existe una tercer plataforma disponible que corresponde a la tecnología de *Roche 454 Genome Sequencer,* no descrita aquí. Todas ellas adoptan conceptualmente similares flujos de trabajo, en el sentido de que constan de la preparación de una biblioteca de ADN, amplificación de la

misma y posterior secuenciación. Las diferencias principales entre estas tres respectan a la bioquímica y enzimática propia de la secuenciación, así como también existen variaciones en los soportes físicos utilizados donde ocurre la secuenciación[136].

- *Illumina (Solexa) Genome Analyzer*

En este caso la muestra de ADN a secuenciar es fragmentada por un mecanismo hidrodinámico de ondas sonoras, que genera fragmentos de menos de 800 pares de bases (pb) de longitud. Los fragmentos generados tienen extremos romos y son fosforilados, de esta forma un nucleótido de adenina es agregado a los extremos 3´de los fragmentos. Luego los fragmentos son ligados a adaptadores que presentan una timina saliente que reconozca la adenina incorporada anteriormente. Los fragmentos con un tamaño de entre 200 y 300 pb son seleccionados. Dichos fragmentos son sometidos a una PCR donde el diseño de los *primers* permite incorporar a los fragmentos las secuencias *P5* y *P7* que serán de utilidad en los próximos pasos. De esta forma la secuencia de interés queda flanqueada en ambos lados, primero por los adaptadores, y de forma más externa por las secuencias *P5* y *P7.* Previa desnaturalización, los fragmentos son inmovilizados por uno de sus extremos en un soporte sólido, conocido como *flow cell,* donde los fragmentos se unen al azar a la superficie. Esta superficie se encuentra densamente recubierta por los adaptadores que reconocen las secuencias *P5* y *P7*, y por sus adaptadores complementarios. De esta forma cada fragmento simple hebra es inmovilizado generando una estructura de puente ya que una vez que un extremo del fragmento se une a la superficie sólida, el otro extremo libre es unido por los adaptadores complementarios. A continuación, los adaptadores de la superficie actúan como *primers* para la amplificación por PCR: aportando mezclas adecuadas al *flow cell* que contengan los reactivos necesarios, los fragmentos de ADN son amplificados por una "PCR puente", del inglés *"bridge PCR",* proceso característico de esta tecnología[137,138]. Después de varios ciclos de PCR, se generaron cerca de 1000 copias de cada fragmento original lo cual representa una colonia o clúster, fijado a la superficie. Posteriormente, la superficie es lavada y se agrega la mezcla para la reacción de síntesis de ADN y secuenciación. Esta mezcla contiene cuatro nucleótidos de terminación reversibles acoplados a un fluoróforo particular. Luego de incorporarse a la cadena de ADN, una cámara es capaz de detectar

según el color, tanto que nucleótido fue incorporado como la posición en la superficie de la cual deriva el color identificando cada clúster de forma individual. Después, tanto el grupo de terminación en el extremo 3´, como la marca fluorescente del nucleótido recién incorporado son removidos. De esta forma se procede a la repetición de este ciclo por un número determinado de veces, entre 35 y 100 veces. Un algoritmo propio del instrumento asigna un valor de calidad a cada nucleótido lo que permite evaluar la calidad de los datos generados en cada corrida para remover aquellas secuencias de baja calidad.

En 2008 *Illumina* presentó una actualización de la tecnología presentada anteriormente, dicha actualización fue nombrada el *Genome Analyser II,* el cual ofrece una poderosa combinación entre los módulos *cBot* y *Paired-End*[139]. El *cBot* es un sistema automatizado revolucionario capaz de crear los clústers clonales a partir de una simple molécula de ADN. Por su parte, el módulo *Paired-End* introduce una simple modificación en la preparación de la librería para leer hebras simples de ADN, mediante la cual facilita la lectura de ambas hebras (*forward* y *reverse*) de cada clúster. Esto permite obtener información adicional ya que además de conocerse las secuencias de los dos extremos o *ends,* se conoce información acerca de su posicionamiento, ya que en función del protocolo establecido se puede saber que ambas lecturas se encuentran separadas por determinada cantidad de nucleótidos. Esta información adicional tiene como principal ventaja que permite un alineamiento más preciso. Una corrida típica de *Paired-End* puede alcanzar cerca de 200 millones de lecturas dobles de 75 bases.

Recientemente *Illumina* presentó su último modelo, el *HiSeq 2000.* Este instrumento maximiza las cantidades de lecturas simples y *paired-end* realizadas, alcanzado valores de 187 y 374 millones de lecturas por línea respectivamente. También duplica la cantidad de bases capaz de leer respecto a su antecesor, el *HiSeq 1000,* alcanzando las 600 gigabases[140]

- *Life Technologies SOLiD system*

Life Technologies SOLiD system es una tecnología basada en la secuenciación por ligación, que tiene sus orígenes en el 2005[141,142]. En esta tecnología, dos tipos de bibliotecas pueden ser construidas: bibliotecas de fragmentos simples o apareados. Más allá de esto, la tecnología implica que los fragmentos de las muestras sean ligados a determinados adaptadores y unidos a pequeñas perlas, como en la tecnología *454.* Luego dichos fragmentos son amplificados por una PCR en emulsión o emPCR, desnaturalizados y seleccionados (etapa de enriquecimiento). Estos fragmentos son modificados a nivel de su extremo 3´, lo cual permite su unión covalente a una lámina de vidrio, donde ocurrirá la reacción de secuenciación. El método de secuenciación es basado en una ligación secuencial con oligonucleótidos marcados[143]. En un primer paso, se procede a la hibridación de uno de los *primers* que reconoce una secuencia correspondiente al adaptador. A continuación un set de 16 oligonucleótidos (octámeros) marcados de forma fluorescente con cuatro fluorocromos distintos, compiten para unirse al *primer* de secuenciación y a la hebra de ADN a secuenciar. Estos octámeros presentan bases degeneradas (N) en sus primeros tres nucleótidos, mientras que las últimas tres bases son universales (Z). Solo la cuarta y quinta base del octámero son las interrogadas. Como existen 16 arreglos posibles de dos bases, y solo se trabaja con cuatro fluorocromos, la señal de un color da lugar a cuatro posibles dinucleótidos, sin embargo un sistema de doble interrogante elimina esta indeterminación.

En esta primera instancia, se determinan que cuatro posibles dinucleótidos están presentes en la posición definida como 1 y 2 (bases 4 y 5 del octámero) en función del color aportado por el fluorocromo particular. A continuación, se remueven tres nucleótidos del octámero desde su extremo 3´, hasta los nucleótidos recién determinados. De esta forma la marca fluorescente del octámero es removida. Luego otro ciclo de hibridación y ligación es llevado adelante donde en este caso se determinarán las bases 6 y 7 de la secuencia problema. En el ciclo posterior las bases determinadas serán las número 11 y 12, y así consecutivamente. Subsiguientes rondas de hibridación y ligación permitirán el secuenciado del fragmento de ADN cada cinco bases.

Luego de dichos ciclos, el producto de extensión sobre la hebra a secuenciar es removido y el ciclo vuelve a comenzar. En esta segunda etapa, un proceso conocido como reseteado de *primers* determina que el octámero se una en posición n-1. Luego de cinco de estas rondas, toda la secuencia fue determinada y cada base en particular fue interrogada dos veces, en dos reacciones de ligación independientes, con distintos *primers.* Por esto, este método llamado *"Two Base Encoding"*[142], es una herramienta única diseñada para identificar con mayor precisión las base (más de 99,99% de precisión[144]. Este concepto de *"Two Base Encoding"* viene acompañado del concepto de espacio color, o en inglés *color space.* Este último es el que mediante una matriz de decodificación nos permite leer la secuencia problema a partir del patrón de colores registrado por el instrumento. La combinación de la enzimología de la ligasa, con el reseteado de los *primers* y este sistema de *"Two Base Encoding"* conjugado con el espacio color, contribuyen todos en conjunto a un sistema con baja tasa de error y poco ruido[136].

3.2.2) "RIBOSOME PROFILING"

Entendiendo que una de las forma de evaluar la expresión génica es cuantificar la tasa de síntesis proteica generada a partir de cada gen en particular, una de las mejores aproximaciones para estimar dicha síntesis proteica es realizar un análisis por *microarrays* a partir de los mensajeros extraídos de la maquinaria traduccional activa o polisomas[119]. Sin embargo, dicha aproximación sufre de ciertas limitaciones de resolución y precisión. Por ejemplo la presencia de uORFs (pequeñas secuencias traducidas encontradas en los 5´-UTR de muchos genes) resulta en un ribosoma unido a un mensajero que en realidad todavía no está traduciendo la proteína codificada por dicho gen[145].

La llegada de las nuevas tecnologías de secuenciación masiva y su gran poder de resolución y sensibilidad abrieron nuevas puertas al análisis de la expresión génica utilizando el compartimiento traduccional. De esta forma Ingolia y colaboradores en el año 2009 publicaron un artículo[76] en el cual presentaron el siguiente razonamiento: la posición de un ribosoma activo en la traducción puede ser determinada de forma precisa considerando el hecho de que dicho ribosoma protege de la digestión por

nucleasas (ARNasas), un fragmento discreto o huella (~30 nucleótidos) del mensajero que está traduciendo[8]. Los avances en las tecnologías de secuenciación masiva podrían hacer posible la lectura de millones de esas cortas secuencias de ARN en paralelo, con lo cual se podría realizar un análisis completo de las huellas ribosomales (ó *ribosome footprints*) de las células en estudio. Así los autores presentaron una estrategia conocida como *ribosome profiling* ó *ribosome footprinting* basada en la secuenciación masiva de los fragmentos discretos protegidos por los ribosomas presentes en la maquinaria traduccional activa, la cual provee información de alta precisión acerca de la traducción *in vivo* con una resolución a nivel de subcodones[76] (ver Figura 1.13).

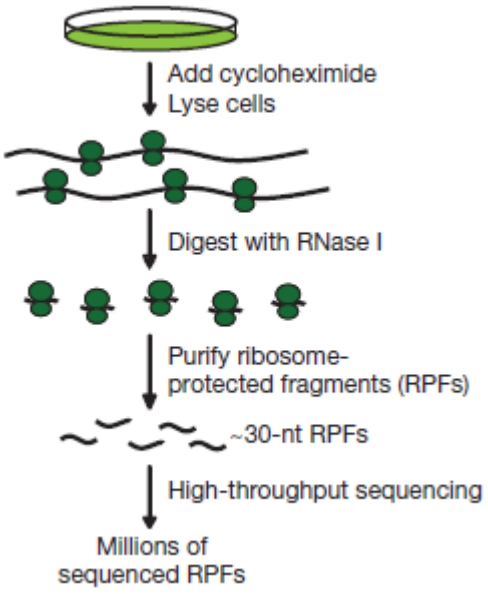


Figura 1.13: Representación esquemática de la técnica ribosome profiling. Se muestra, a grandes rasgos, el diagrama de flujo de trabajo resaltando los principales pasos. Adaptada de Guo et al. 2010.

Para poder establecer el *ribosome profiling* como una herramienta cuantitativa para evaluar la traducción, los autores debieron poner a punto tres pasos: i) la generación robusta de las huellas (*footprints*) o fragmentos de mensajeros protegidos por los ribosomas, cuya secuencia indica la posición de un ribosoma activo; ii) convertir dichas huellas de ARN en una biblioteca de moléculas de ADN listas para ser secuenciadas de forma masiva con la menor distorsión posible; y iii) lograr medir la abundancia de las diferentes huellas generadas por la secuenciación masiva. De esta forma, evaluando el número de veces que aparece un fragmento determinado, se puede obtener una medida cuantitativa de su abundancia en la biblioteca[76]. Esta medida nos permite evaluar cuantos ribosomas hay presentes sobre cada mensajero, y de aquí estimar de forma relativa la expresión proteica de cada gen.

Utilizando técnicas de mapeo y alineamiento adaptadas a procesar datos de secuenciación masiva es posible estimar los niveles de traducción para todos los ARNm

presentes en la muestra estudiada. En forma adicional, se obtiene información acerca de la composición de secuencia exacta del transcriptoma estudiado, información útil para detectar variantes génicas en el genoma.

Los resultados de Ingolia y colaboradores, son en cierta manera, impactantes: para un punto experimental único son capaces de estimar la tasa de traducción de cerca de 4000 mensajeros en levaduras. El rango dinámico de la técnica llegaría aproximadamente a detectar ARNm que son diferencialmente traducidos en dos órdenes de magnitud. Los autores fueron capaces de estudiar la regulación de la tasa de traducción frente a diferentes condiciones experimentales, observando fases de la traducción novedosas e identificando numerosas iniciaciones en codones diferentes a AUG. Un detalle no menor es que al correlacionar la tasa de traducción obtenida con resultados del proteoma de levaduras previamente publicados, logran obtener valores de R^2 tres veces mayor al obtenido simplemente al correlacionar niveles de ARNm con niveles de proteínas.

Este mismo método ha sido empleado por segunda vez con éxito para determinar con alta precisión cuál es la influencia de microARN específicos en la traducción de sus ARNm blancos, en células en cultivo humanas[146]. El resultado fue también sorprendente: el efecto principal de estos pequeños ARN reguladores es principalmente la degradación de los blancos en lugar de la inhibición de la traducción de los mismos.

Recientemente dicha metodología ha sido aplicada por tercera vez, en este caso en células madre de embrión de ratón[147]. En este trabajo los autores mejoraron la metodología con el uso de otra droga, con la cual establecieron un método para estimar la tasa de elongación. También mediante la aplicación de algoritmos computacionales lograron estimar los codones de iniciación con gran resolución. Brevemente en este nuevo artículo los autores trabajan con la droga harringtonina la cual genera la acumulación de los ribosomas en los sitios de iniciación de la traducción. Esta droga inhibe la iniciación ya que al unir solamente la subunidad mayor libre impide la formación del primer enlace peptídico, sin afectar a las subunidades mayores ya cargadas en ribosomas activos[148,149]. De esta forma se obtiene el patrón de

huellas correspondientes a los ribosomas al momento del inicio de la traducción y mediante la aplicación de algoritmos computacionales (*SVM-based machine learning*) se puede definir el uso de codones de iniciación, encontrando resultados realmente sorprendentes. Por ejemplo, el número de transcriptos con 2, 3, 4 ó 5 sitios de iniciación supera en conjunto, en más de dos veces a la cantidad de transcriptos con un solo sitio de iniciación. También la distribución de los codones de iniciación utilizados muestra que más de la mitad de dichos codones no son el canónico AUG.

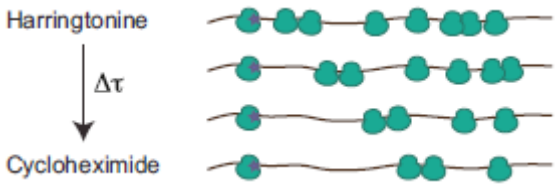


Figura 1.14: Representación esquemática del experimento diseñado para medir la tasa de elongación. Figura adaptada de Ingolia et al. 2011

También los autores monitorean la cinética de traducción o tasa de elongación, conjugando el uso de las drogas cicloheximida y harringtonina. En este experimento se detienen nuevas traducciones con el uso de harringtonina, luego se abre una pequeña ventana de tiempo (Δτ) sin drogas donde los ribosomas avanzan por el mensajero para después detener la traducción de los mismos agregando cicloheximida. Variando el lapso de tiempo en el cual ocurre la elongación se pueden obtener distintas "fotos" que pueden reconstruir una "película" de lo que ocurre con la traducción *in vivo* (ver Figura 1.14).

Actualmente existen otras dos publicaciones más donde se aplicó la tecnología de *ribosome profiling*[150,151]. En una de ellas, los autores estudiaron a alta resolución los cambios en la abundancia de mensajeros y en la producción de proteínas en un modelo de levaduras en meiosis, estudiando cada estadío meiótico en particular[151]. En el otro trabajo más reciente, los autores aplicaron la técnica para estudiar la distribución y organización subcelular de la traducción de ARNm[150].

CAPÍTULO 2

HIPÓTESIS DE TRABAJO, OBJETIVOS Y ESTRATEGIAS EXPERIMENTALES

Problemas relacionados a la biosíntesis de proteínas han sido consistentemente asociados a la transformación tumoral y a las distintas metástasis estudiadas[81,152,153]. En particular una elevada y eficiente iniciación de la traducción, resultado de procesos alterados o de una mayor abundancia de factores de iniciación disponibles, ocurre en muchos tipos de cáncer. En este contexto resulta atractiva la presencia de *PDCD4,* un gen supresor de tumores implicado en el control del inicio de la traducción. Como se comentó en la sección 2.1 del capítulo 1, *Programmed Cell Death 4 (PDCD4),* es un supresor de tumores que regula la expresión génica a nivel transcripcional y traduccional. Se ha propuesto, aunque no se conocen con detalles sus blancos de acción, que PDCD4 compite con el ARN mensajero y con eIF4G por unir a eIF4A en la formación del complejo de pre-iniciación de la traducción. Esta interacción con eIF4A podría interferir con su acción helicasa, lo cual repercute en la capacidad del mismo para resolver estructuras secundarias presentes en mensajeros que se traducen de forma *cap*-dependiente. De esta forma suena interesante detenerse en la búsqueda de dichos mensajeros blancos de la acción de PDCD4 considerando que se trata de un gen supresor de tumores y teniendo en cuenta la importancia del compartimiento traduccional en los procesos tumorales.

De esta forma la hipótesis de trabajo planteada es la siguiente: el supresor de tumores *Programmed Cell Death 4* modula la expresión de un conjunto de ARN mensajeros mediante el control de su traducción, por lo cual se denominan blancos traduccionales. Dado que se trata de un supresor de tumores, cuya expresión se ha visto disminuida en varios tumores[84], se postula que PDCD4, en condiciones normales, inhibe la traducción de ciertos genes relacionados a la activación de procesos tumorales (oncogenes), diferenciación, movilización celular, control del ciclo celular, etc. Mientras, en condiciones tumorales, la ausencia de PDCD4 o problemas respecto al

desarrollo normal de su actividad, permiten la expresión traduccional de muchos genes entre los cuales se encuentran aquellos que favorecen el desarrollo de tumores.

Teniendo en cuenta que dichos blancos traduccionales no son conocidos con detalle (excepto los tres casos reportados recientemente con relevancia en procesos tumorales[92,104,105]) el principal objetivo planteado en este trabajo fue lograr identificar grupos de genes candidatos a ser regulados justamente por PDCD4. Para lograr cumplir con esto, se plantearon otros objetivos previos, como lograr un adecuado entrenamiento en el manejo de datos generados por secuenciación masiva, así como en el manejo de las herramientas de análisis disponibles para realizar los estudios con dichos datos. Por otro lado, para lograr comprender en forma detallada en que fenómenos está implicado PDCD4, nos propusimos un tercer objetivo que implica una adaptación al manejo de herramientas disponibles para realizar análisis y estudios de ontología, para así llevarlos adelante e identificar las funciones asociadas a los genes candidatos anteriormente mencionados.

La estrategia experimental definida para estudiar dicho problema fue la aplicación de la técnica *ribosome profiling* ó *ribosome footprinting* presentada en la sección 3.2.2 del capítulo 1. Mediante esta técnica se tiene acceso directo al estudio de los cambios a nivel traduccional a alta resolución mediante la aplicación de la secuenciación masiva de regiones protegidas por los ribosomas en polisomas activos[76]. De esta manera se podría estudiar a gran escala la influencia de *PDCD4* en la traducción en modelos celulares de cáncer, mediante la supresión de dicho gen vía siARN. Luego mediante el alineamiento y mapeo (términos que de aquí en adelante serán utilizados como equivalentes) de las huellas producidas sobre todo el genoma, se procede a la cuantificación de dichas huellas y al cálculo de los niveles de expresión génica. Al comparar el anterior análisis con respecto a una condición control de referencia, se podría llegar a observar el espectro posible de regulación traduccional por parte de *PDCD4* a niveles de resolución no anticipados previamente.

Puntualmente, las aproximaciones experimentales definidas para cumplir con los objetivos presentados pueden ser separadas en dos instancias de trabajo. En una primera instancia en base a la publicación del artículo donde se describe por primera

vez el uso de la metodología *ribosome profiling (ribosome footprinting)*[76] y en base a los datos generados por dichos autores y disponibles libremente, se buscó trabajar con éstos a modo de entrenamiento en el manejo de datos de secuenciación masiva y en el uso del software disponible en el laboratorio *(CLC Genomics Workbench 4.6.1 © by CLC bio Aarhus, Denmark - www.clcbio.com).* Esta primera instancia de entrenamiento y familiarización del trabajo con datos de secuenciación masiva, satisface los requisitos definidos en el primer objetivo planteado y sirve de preparación para el pasaje a una segunda instancia de trabajo.

Dentro de esta primera instancia, la posibilidad de reproducir algunos de los resultados presentados por los autores en base a los análisis realizados por nuestra cuenta, así como comparar y verificar parámetros de expresión, plantean desafíos interesantes. En particular se trabajará con los datos generados por Ingolia *et al.* 2009. En este caso, Ingolia y colaboradores desarrollaron sus propios métodos computacionales para realizar el estudio y análisis completo de sus datos, generando un material suplementario de cerca de cien páginas. En nuestro caso se utilizarán herramientas analíticas distintas, que utilizan diferentes algoritmos tanto para realizar los alineamientos, como para realizar los complejos cálculos asociados a determinar los niveles de expresión génica de los mensajeros. Estas diferencias tornan aún más interesante y significante la posibilidad de reproducir resultados y parámetros de expresión.

El esquema de trabajo en general, sin importar cual sea el artículo del cual se extraen los datos, marca el diagrama de tareas presentado en forma esquemática en la figura 2.1. En este contexto, el flujo de trabajo a seguir determina en una primera instancia proceder a descargar las secuencias de la correspondiente base de datos a partir del número de acceso depositado en el artículo. Los bancos de datos más utilizados son el *SRA (Sequence Read Archive - www.ncbi.nlm.nih.gov/sra)* y el *GEO (Gene Expression Omnibus - www.ncbi.nlm.nih.gov/geo),* los cuales forman parte del *NCBI (National Center for Biotechnology Information).* Previo a descargar las secuencias, es aconsejable estudiar en forma profunda la información, figuras y tablas aportadas en el artículo para comprender como se generaron los datos en cuestión y así definir cuáles son los adecuados para trabajar. Lo siguiente consta en descomprimir los datos

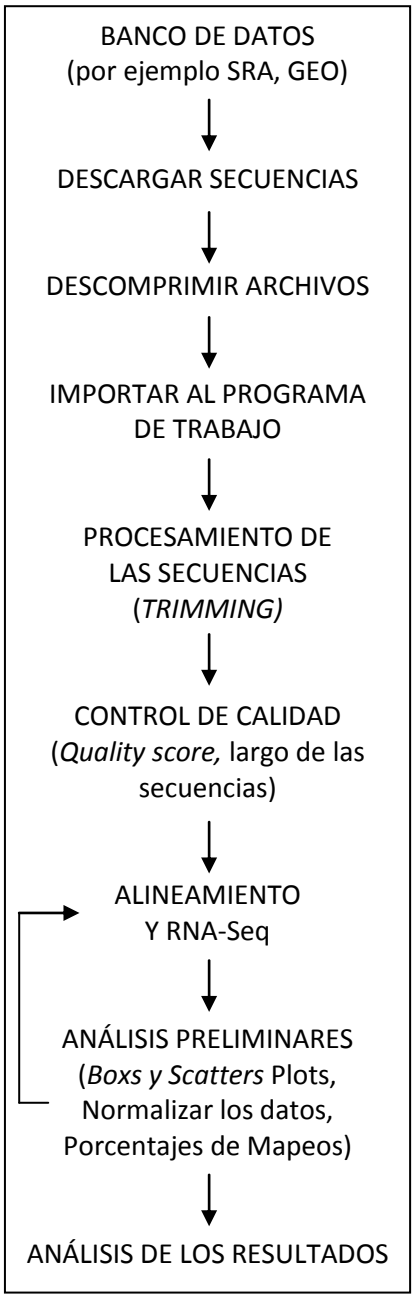


Figura 2.1: Diagrama de tareas planteado como estrategia experimental para el análisis de datos de secuenciación masiva

al formato adecuado e importarlos al software de trabajo. Así se procede a depurar y procesar las secuencias disponibles en función de algunos parámetros como los valores de calidad (*quality score)*, largo y posible presencia de adaptadores. Este proceso es denominado *trimming.* A continuación ya se está en condiciones de alinear las secuencias y realizar los cálculos correspondientes para determinar los niveles de expresión génica de ARNm, en este caso, mediante análisis mediante RNA-Seq[120]. Dichos análisis generan distintas tablas de datos con información variada, donde se incluyen los valores de expresión para cada mensajero, a partir de las cuales se construyen los análisis posteriores. En este caso, dichos análisis posteriores involucran en forma preliminar, controles de calidad de los datos y mapeos realizados mediante la evaluación de los porcentajes de lecturas que mapean contra el genoma, que son descartadas o que mapean contra los genes de ARN ribosomal, lo cual representa contaminación en la muestra. Los controles de calidad también implican la construcción de *box plots* mediante los cuales podemos estudiar la distribución global de los datos y su variabilidad. También se pueden construir *scatter plots* comparando las dos condiciones de trabajo estudiadas, para observar *a priori* cómo se comportan los datos y cuál es la tendencia global observada resultado de modificar parámetros en el sistema de estudio. Otros análisis posteriores que se llevarán a cabo son la reproducción de algún resultado presentado en el artículo original, la comparación de patrones y valores de cambio en los parámetros evaluados, así como el análisis de algún punto novedoso no cubierto previamente por el artículo. Para esto último se utilizará el *Microsoft Office Excel 2007* con el cual se pueden

manejar grandes tablas de información, construir figuras y mediante la aplicación de las variadas funciones de las cuales dispone, se puede extraer información útil y de forma rápida la cual luego se presenta a lo largo del capítulo 3 y 4. De todas formas, a lo largo del desarrollo del capítulo 3 donde se presentan los resultados producidos se detalla en forma puntual que herramienta se utilizó y bajo que parámetros, para llevar a cabo los análisis efectuados.

Respecto a la segunda instancia, esta consta del manejo y análisis de datos generados por el laboratorio, con el objetivo de extraer alguna conclusión preliminar acerca de la hipótesis de trabajo presentada. Dado que ya se cuenta con las secuencias generadas previamente mediante la aplicación de la técnica *ribosome profiling* en esta instancia se intentará cumplir con los otros dos objetivos delineados más atrás. Así las principales tareas a desarrollar en este punto son intentar responder las interrogantes acerca de quiénes son los genes regulados traduccionalmente por *PDCD4,* qué funciones cumplen dichos genes, qué características generales se observan en los mapeos de estas huellas ribosomales sobre un genoma humano y cuáles son las razones para explicar algunos patrones de mapeos particulares observados. Con respecto a la interrogante acerca de las funciones a cargo de los potenciales blancos encontrados, esta instancia se centra en el análisis de vías celulares y ontología de genes, para lo cual se utilizará el software comercial *Ingenuity Systems © Redwood city, California, USA – www.ingenuity.com),* ya que en el laboratorio se cuenta con la licencia del mismo. De todas formas, al igual que en la instancia anterior, a lo largo de la presentación de los distintos resultados encontrados se explica en forma detallada las herramientas utilizadas y sus características puntuales.

CAPÍTULO 3

ENTRENAMIENTO EN EL ANÁLISIS DE DATOS DE SECUENCIACIÓN MASIVA

La publicación de Ingolia *et al. Science* 2009 fue el artículo elegido para obtener los datos de secuenciación masiva con los cuales se procede a realizar los primeros análisis, ya que corresponde a la primera publicación donde se hace uso de la técnica *ribosome profiling* desarrollada por los propios autores. En este trabajo, los autores se propusieron estudiar los cambios a nivel transcripcional y traduccional observados en levaduras en condiciones de privación de aminoácidos. Para esto desarrollaron una metodología a la cual llamaron *ribosome profiling,* de utilidad para estudiar los cambios a nivel traduccional de los distintos mensajeros, a escala genómica. De forma paralela también realizaron estudios de RNA-Seq para evaluar los cambios a nivel transcripcional.

Respecto a la metodología de trabajo, los autores analizaron dos condiciones experimentales: una condición control (levaduras en fase logarítmica de crecimiento en medio rico) y otra condición de privación de aminoácidos (levaduras sujetas a 20 minutos de crecimiento en medio deficiente de aminoácidos). Cada condición fue sometida al estudio transcripcional mediante RNA-Seq, y traduccional por *ribosome profiling,* por separado.

A la hora de evaluar los cambios observados, los autores definieron un parámetro que denominaron eficiencia traduccional que cuantifica cuanto se traduce un mensajero de un gen en particular en función de la cantidad de transcripto presente en la célula. Dicho parámetro se calcula como el cociente entre el valor de RPKM (valor de expresión; ver sección 3.2 del capítulo 1, página 38) para el gen en cuestión según el mapeo de las lecturas derivadas de la técnica de *ribosome profiling*, y el valor de RPKM derivado del mapeo de las lecturas procedentes de la población total de mensajeros.

1) DESCARGA DE SECUENCIAS, CAMBIO DE FORMATO E IMPORTACIÓN AL PROGRAMA

En este artículo en particular los autores depositaron los datos generados en el *National Center for Biotechnology Information´s Gene Expression Omnibus* (www.ncbi.nlm.nih.gov/geo). De todas formas el acceso y la descarga de cada uno de los set de datos de secuenciación se realiza desde la base de datos *Sequence Read Archive* (SRA), que forma parte del *National Center for Biotechnology Information* (www.ncbi.nlm.nih.gov/sra). Los set de datos están organizados según el origen de la muestra (*ribosome profiling* o ARNm total), la condición de la misma (*log-phase* – levaduras en fase logarítmica de crecimiento*;* o *starvation – levaduras* creciendo en medio deficiente de aminoácidos) y se especifica el número de réplica.

Cada set de datos consta de varios archivos que se descargan en formato .lite.sra, el cual es un formato comprimido del que se puede extraer la secuencia y sus características de calidad, en formato .fastq por ejemplo (ver más adelante). Estos archivos se descargaron utilizando la tecnología *Aspera Connect* 2.7.0 disponible en la PC de trabajo en el laboratorio, para acelerar la descarga de los mismos ya que en promedio cada uno de dichos archivos .lite.sra tienen un tamaño entre los 160 y 180 Mbytes. La tecnología de transferencia *Aspera* permite eliminar fundamentalmente los cuellos de botella de las tecnologías convencionales de transferencia de archivos como FTP, HTTP Windows, por lo cual acelera drásticamente las velocidades de transferencias[154].

Una vez descargados todos los archivos se procede a convertirlos de su formato .lite.sra, a formato .fastq. El formato .fastq es un formato basado en líneas de texto utilizado para almacenar secuencias nucleotídicas y sus *quality scores* correspondientes. Originalmente fue desarrollado por el *Wellcome Trust Sanger Institute* para agrupar una secuencia FASTA y la calidad de sus datos, de todas formas recientemente se ha convertido en el formato estándar para el almacenamiento de datos de secuenciación masiva. Particularmente cada archivo consta de cuatro líneas de texto por secuencia (ver Figura 3.1). La primera línea comienza con un `@´ y sigue con la identificación de la secuencia, más una descripción opcional. La segunda línea

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Figura 3.1: Se muestra una secuencia en formato FASTA a modo de ejemplo

presenta las bases nucleotídicas descritas en letras. La tercer línea comienza con un `+´ y puede ser seguida también de la identidad de la secuencia y alguna otra descripción opcional. Por último la cuarta línea codifica los valores de calidad de la secuencia presentada, según el instrumento utilizado para la secuenciación, y debe contener el mismo número de símbolos que letras en la secuencia.

Para descomprimir los archivos a formato .fastq se copia la aplicación ejecutable *fastq-dump.exe* en la carpeta donde se encuentran los archivos a convertir (por ejemplo: `../sra/download/SRR000299/`). Dicha aplicación deriva del *SRA-Toolkit* disponible libremente. De esta forma, se ejecuta en una línea de comando del "símbolo de sistema" (Windows) el siguiente comando:

```
../sra/download/fastq-dump SRR000299
```

De esta forma se genera en la misma carpeta, un nuevo archivo con el mismo nombre, pero en formato .fastq.

Los archivos en formato .fastq están en condiciones de ser importados al software de trabajo *CLC Genomics Workbench.* Este programa permite importar datos de secuenciación masiva generados por las distintas tecnologías disponibles (ver sección 3.2.1). Las opciones son: *Roche 454, Illumina, SOLiD, Fasta, Sanger, Ion Torrent.* En este caso los autores secuenciaron haciendo uso de la máquina *Illumina Genome Analyzer II,* por lo que lo correcto es elegir la opción *Illumina.* En el proceso de importar las secuencias deben especificarse las características particulares de las secuencias, así como otros parámetros generales. Por ejemplo si se trata de secuencias tipo *paired-end* (ver sección 3.2.1); o si se desean descartar los nombres y/o los valores de calidad de las secuencias a los efectos de reducir el espacio del disco utilizado y el consumo de

memoria. También se debe definir la escala utilizada para leer los *quality score* de las secuencias. En el formato .fastq existen varias versiones distintas respecto al *quality score.* Las opciones varían según la plataforma utilizada al momento de secuenciación y las opciones generales para datos generados por la tecnología *Illumina* son: *Automatic, NCBI/Sanger Phred scores, Illumina Pipeline 1.2 and earlier, Illumina Pipeline 1.3 and 1.4* y *Illumina Pipeline 1.5 and later.* En este caso según la fecha del artículo y según información aportada en el material suplementario, se optó por elegir la tercera opción.

Debe considerarse en los sucesivos procesos de importación de datos de secuenciación masiva, la generación y disposición de carpetas ordenadas y organizadas, donde se agrupen las distintas secuencias en función de su origen y condición como se comentó anteriormente. Este proceso de importar archivos con secuencias generadas de forma masiva no supera los 5 minutos para cada condición experimental.

2) PROCESAMIENTO DE LAS SECUENCIAS, ALINEAMIENTO Y ANÁLISIS MEDIANTE RNA-Seq

Antes de proceder a los mapeos y al análisis de la expresión génica mediante RNA-Seq, las secuencias importadas deben ser depuradas y procesadas, función conocida como *trimming.* La utilidad de este proceso recae en mejorar los posteriores mapeos, ya que la gran mayoría de las secuencias que se leen en el instrumento de secuenciación masiva no reflejan exactamente el fragmento original. Esto se debe a que las secuencias leídas contienen nucleótidos de los adaptadores o nucleótidos incorporados resultado de la metodología utilizada para construir las bibliotecas. Dichos nucleótidos deben ser removidos antes de mapear las secuencias para que estas alineen en el lugar correcto. Por otro lado, también existen secuencias dentro de los archivos que presentan bajos valores de calidad, por lo cual se recomienda descartarlas antes de realizar los mapeos ya que su identidad no es confiable.

Las opciones al momento de realizar este procesamiento o *trimming* de las secuencias incluyen: i) quitar secuencias que presenten un *quality score* menor que determinado umbral; ii) descartar los extremos de las secuencias con bases determinadas de forma

ambigua (N); iii) quitar los adaptadores de las secuencias; iv) remover o seleccionar aquellas secuencias que no presenten los adaptadores; v) remover nucleótidos desde los extremos de las secuencias.

En nuestro caso se eligieron las opciones i) y v). La primera se eligió a los efectos de eliminar aquellas secuencias cuyas bases fueron determinadas de forma no confiable. La quinta opción se eligió en base a los argumentos que se presentan a continuación.

Si uno estudia con detalle el artículo encuentra en el soporte de materiales y métodos añadido al material suplementario que, si bien los autores utilizan otro software para el mapeo de las secuencias (SOAP v1.10), ellos alinean en una primera instancia únicamente los primeros 21 nucleótidos de las secuencias. Esto se determinó así por el hecho de que las secuencias con las que se cuenta en la biblioteca están compuestas por una región de largo variable del fragmento de ARN capturado, seguido de una cola poli-A, resultado de la metodología utilizada para generar dichas bibliotecas. En forma más detallada, a nivel experimental primero se generan los fragmentos protegidos por el ribosoma y los fragmentos generados por el fraccionamiento de la población total de mensajeros. Estos se seleccionan por tamaño, entre 20 y 30 nucleótidos y luego se agrega una cola poli-A. Sobre esta pueden alinear *primers* con adaptadores necesarios en los futuros pasos para la construcción de las bibliotecas[76]. De todas formas, los autores afirman que mediante estudios preliminares han demostrado que esencialmente todos los fragmentos de ARN capturados tienen al menos un largo de 21 nucleótidos, por eso la elección de alinear esos primeros 21 nucleótidos. Después de estos primeros 21 nucleótidos, una base secuenciada puede derivar ya sea del propio fragmento de ARN capturado o de la cola poli-A agregada.

Esta idea del largo variable de los fragmentos puede observarse en una de las figuras del artículo que se creyó conveniente citar (ver Figura 3.2). En dicha figura se observa la distribución de largos de los fragmentos de ARN generados discriminando según su origen: ARN ribosomal (contaminación generada durante la digestión con ARNasa de fragmentos expuestos del ribosoma), ARNm (los fragmentos derivados de una fragmentación azarosa de la población de mensajeros celulares), y los fragmentos protegidos por los ribosomas. En esta figura puede observarse como los fragmentos

presentan largos distintos según su origen, por lo cual elegir los primeros 21 nucleótidos para realizar los mapeos es adecuado, ya que es el largo mínimo que presentan todos los fragmentos.

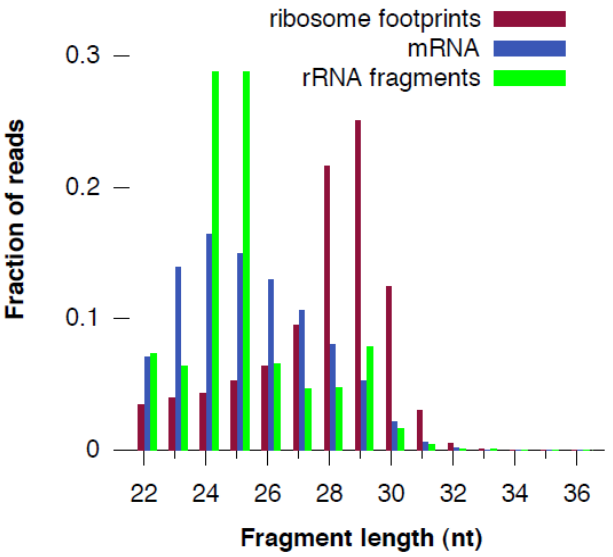


Figura 3.2: Distribución de largos de los fragmentos de ARN generados discriminados según su origen. Figura adaptada de Ingolia et al. Science 2009.

Por esto se decidió cortar las secuencias desde su extremo 3´ hasta que queden solo esos primeros 21 nucleótidos en cuestión. Sin embargo también se decidió cortarlas hasta los primeros 28 nucleótidos, que es en teoría la cantidad de nucleótidos de los fragmentos protegidos por la maquinaria traduccional de la acción de ARNasas durante el experimento de *ribosome profiling*[8] (ver Figura 3.2). Esta posibilidad nos permite en principio, no solo poder analizar y comparar los mapeos generados a partir de secuencias de 21 y 28 nucleótidos de largo en busca de diferencias y similitudes, sino que también nos permite visualizar la reproducibilidad de la técnica. Esto último se plantea en el sentido de que los mapeos realizados a partir de las secuencias derivadas de los ensayos de *ribosome profiling* recortadas a 21 o a 28 nucleótidos, no deberían variar mucho en sus porcentajes de mapeos. Esto se debe a que al ser las secuencias (en teoría) de 28 nucleótidos, al recortarlas a 21 nucleótidos se quitan de forma segura bases que se sabe que derivan del fragmento de ARN capturado. Aún así, esto último no se espera para los fragmentos de 21 y 28 nucleótidos derivados de las muestras de ARNm total, ya que dichos fragmentos fueron generados por una ruptura al azar de la población de mensajeros totales. En este caso, no existe ninguna razón por la cual se pueda afirmar que dichos fragmentos deben tener un largo común, sino que por el contrario es lógico pensar que la población de fragmentos presente un rango de longitud sumamente variable. Esto determina que los mapeos realizados en este caso sí sean sensibles al largo del fragmento que se mapee.

Este proceso de *trimming* tarda entre 5 y 10 minutos en función de la cantidad de archivos presentes en cada condición (ver Tabla 3.1), ya que como se puede observar deben removerse nucleótidos y secuencias de baja calidad, de un total de entre 15 y 20 millones de secuencias por condición.

Una vez que finalizan todos los procesos de *trimming*, se cuenta con todas las secuencias de cada condición con un largo de 21 y 28 nucleótidos listas para mapearse y cuantificar la expresión génica. Así se procede a realizar un RNA-Seq para el conjunto

Tabla 3.1: Nombre, descripción y tamaño de los archivos en megabases (Mbases), descargados a partir de los datos generados por Ingolia et al. Science 2009. Origen "ARNm total" implica que la muestra deriva de una extracción de mensajeros totales, mientras que origen "RFP", significa que se secuenciaron los fragmentos protegidos por el ribosoma, como describe la técnica ribosome profiling. La condición "control" deriva de la muestra en fase logarítmica de crecimiento, mientras que la condición "no aac." deriva de la muestra sujeta a privación de aminoácidos por 20 minutos.

origen	condición	réplica	archivo (.lite.sra)	cantidad de secuencias y su largo		tamaño (Mbases)
ARNm total	control	1	SRR014385	4.700.324	36 (nt)	169,2
		2	SRR014386	4.912.441	36	176,8
			SRR014387	1.762.132	36	63,4
			SRR028774	6.343.497	35	222,0
	no aac.	1	SRR014382	4.135.727	36	148,9
		2	SRR014383	4.915.502	36	177,0
			SRR014384	5.743.864	35	201,0
RFP	control	1	SRR014374	4.601.558	36	165,7
			SRR014375	4.737.798	36	170,6
			SRR014376	4.221.683	36	152,0
		2	SRR014377	4.237.384	36	152,5
			SRR014378	4.211.315	36	151,6
			SRR014379	4.758.893	35	166,6
			SRR014380	5.322.596	35	186,3
			SRR014381	5.468.170	35	191,4
	no aac.	1	SRR014368	4.485.329	36	161,5
			SRR014369	4.588.611	36	165,2
		2	SRR014370	4.323.096	36	155,6
			SRR014371	4.223.635	36	152,1
			SRR014372	5.222.229	35	182,8
			SRR014373	5.072.716	35	177,5
			TOTALES	**97.988.500**		**3.489,7**

de archivos que definen cada condición en particular, sin considerar las réplicas, mapeando contra el genoma de referencia de levaduras. En el programa disponible en el laboratorio, la herramienta RNA-Seq puntualmente es capaz de realizar los mapeos de todas las secuencias y posteriormente cuantificar la expresión génica en función de la cantidad de mapeos realizados sobre cada gen.

Respecto al algoritmo de mapeo utilizado por el programa, debe considerarse que existen dos tipos de algoritmos de alineamientos. Estos discriminan en función de largo de las secuencias que mapean: uno aplica para secuencias cortas (menos de 56 nucleótidos), y otros para largas (las lecturas pueden tener cualquier largo hasta 8000 pb). En nuestro caso se trabaja con el primero. Particularmente los algoritmos de mapeo de secuencias cortas realizan alineamientos locales contra la secuencia de referencia. La ventaja de trabajar con alineamientos locales frente a alineamientos globales es que los extremos pueden ser automáticamente descartados si existen demasiados errores de secuenciación allí[155].

Al igual que Maq, Soap y otros programas similares que realizan mapeos de secuencias cortas, por defecto se utiliza un alineamiento “sin espacios” (en inglés *ungapped alignment*). De esta forma cada alineamiento realizado presenta un score que evalúa la calidad del mismo. Para esto se definen los scores para las coincidencias (+1) y para las discrepancias (-2). Aceptar o rechazar un alineamiento implica evaluar el score del mismo para cada lectura en función de un score límite definido. Por ejemplo, si el score límite es 8 por debajo de la longitud de las secuencias, y se trabaja con secuencias de 20 nucleótidos de largo, esto implica que se aceptarán mapeos con scores mayores o iguales a 12. De esta forma se toleran mapeos con hasta dos discrepancias y dos nucleótidos de los extremos sin alinear (nótese que considerar una discrepancia implica no sumar 1 y restar 2, por lo que se pierden 3 puntos en total en el score final). También se tolerarían mapeos con 8 nucleótidos de los extremos sin alinearse, ó con un solo error en el mapeo y hasta 5 nucleótidos sin alinear, etc. (ver Figura 3.3).

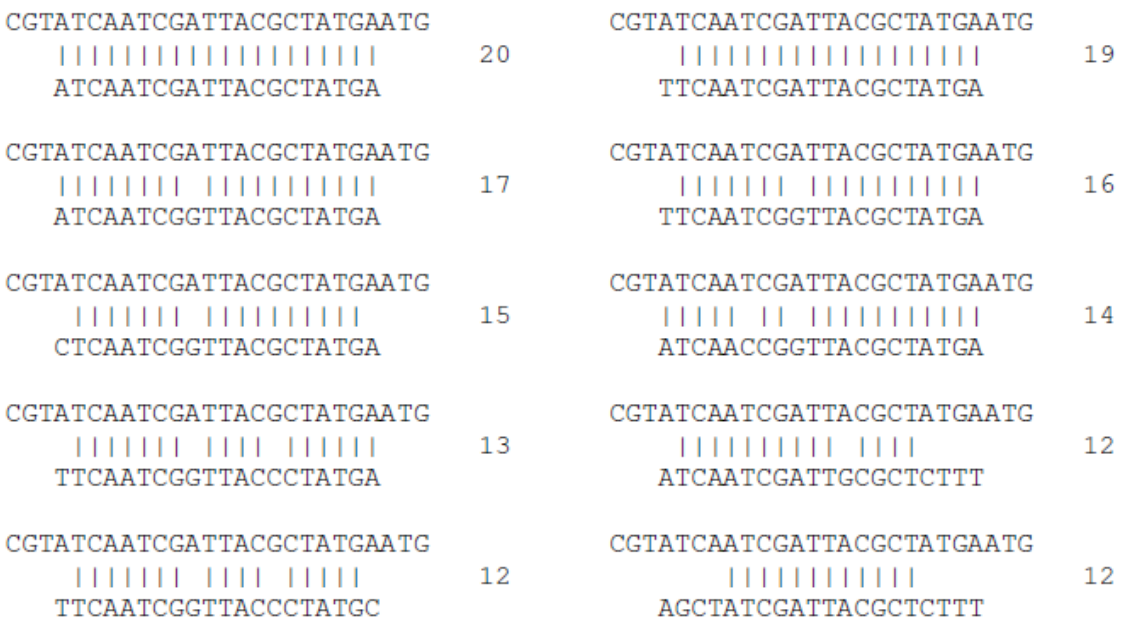


Figura 3.3: Ejemplos de alineamientos permitidos para una lectura de 20 pb de largo con un score límite de 8. Los scores asignados a cada alineamiento se presentan a la derecha del correspondiente alineamiento. Figura adaptada de "White paper on reference assembly in CLC Assembly", disponible en www.clcbio.com

El score límite definido se calcula en función del costo de aceptar un error en el mapeo (definido en 2), a través de la siguiente fórmula[155]:

score límite = 3 x (1 + costo de un error) − 1

De esta forma se define el score límite de 8 utilizado en el ejemplo anterior. Esta fórmula asegura que más allá del valor asignado al costo de aceptar un error y más allá del largo de las secuencias mapeadas, siempre se aceptarán alineamientos mejores que uno con tres discrepancias.

Al momento de realizar el mapeo y análisis por RNA-Seq debe definirse el genoma contra el cual las secuencias serán mapeadas, en este caso se utiliza el genoma de referencia de levaduras (*Saccharomyces cerevisiae* S288c) con las correspondientes anotaciones para identificar cada uno de los genes. También deben definirse varios parámetros específicos del alineamiento, entre ellos: i) número máximo de errores aceptado (este parámetro solo se encuentra disponible si trabajamos con secuencias cortas como ocurre en nuestro caso), definido en 2. ii) número máximo de veces que una lectura mapea inespecíficamente en la referencia, antes de ser descartada; definido en 10. En este caso si una lectura mapea 10 o menos veces en el genoma de

referencia será asignada a un gen en particular en función de la cantidad de lecturas génicas únicas que presente el gen, normalizado por el largo exónico de dicho gen. Esto método asigna más de estas lecturas inespecíficas a los genes de más alta expresión. El software también tiene en cuenta que una lectura puede mapear más de una vez si mapea contra un mismo exón común presente en varios tipos distintos de transcriptos, por esto el programa tiene en cuenta el número de transcriptos anotados en el genoma de referencia para no descartar lecturas de este tipo. iii) alineamiento hebra-específico, si se activa esta opción las lecturas solo serán mapeadas en su orientación *forward.* En nuestro caso esta opción se encuentra desactivada. iv) tipo de organismo, se debe especificar si se trata de un organismo procariota o eucariota a los efectos de que el software conozca si la secuencia de referencia tiene intrones o no. v) hallazgo de exones, el programa da la posibilidad de reportar nuevos exones en función de los siguientes parámetros: a) nivel de expresión relativa requerida, representa el porcentaje de expresión que, al menos, debe tener el nuevo exón respecto del gen que lo contiene; b) número mínimo de lecturas, mientras que la opción anterior refería a un número relativo a la expresión global del gen, aquí se especifica un valor absoluto; c) largo mínimo del nuevo exón en nucleótidos.

Existen otros dos parámetros por definir, pero que solo aplican para lecturas largas, y en todos estos casos se trabaja con secuencias cortas, menores a 56 nucleótidos. Dichos parámetros son: i) la fracción mínima del largo de la secuencia que debe alinear, por ejemplo por defecto este valor es de 0,9 lo que significa que al menos el 90% de las bases deben alinear con la referencia. ii) la fracción mínima de similitud, la cual representa cuan exacta debe ser la similitud en el mapeo. Esto significa, por ejemplo, que si se utilizan los valores por defecto de estos últimos dos parámetros, 0,9 y 0,8 respectivamente, esto significa que al menos un 90% de la lectura debe alinear con un mínimo del 80% de similitud con la referencia para incluir la lectura.

Respecto a las opciones de salida de los resultados, este programa crea un reporte donde se especifica información variada relacionada al mapeo de secuencias sobre cada gen en particular, donde se presenta el valor de expresión génica calculado medido en RPKM (*Reads Per Kilobase of exon model per Million mapped reads*).

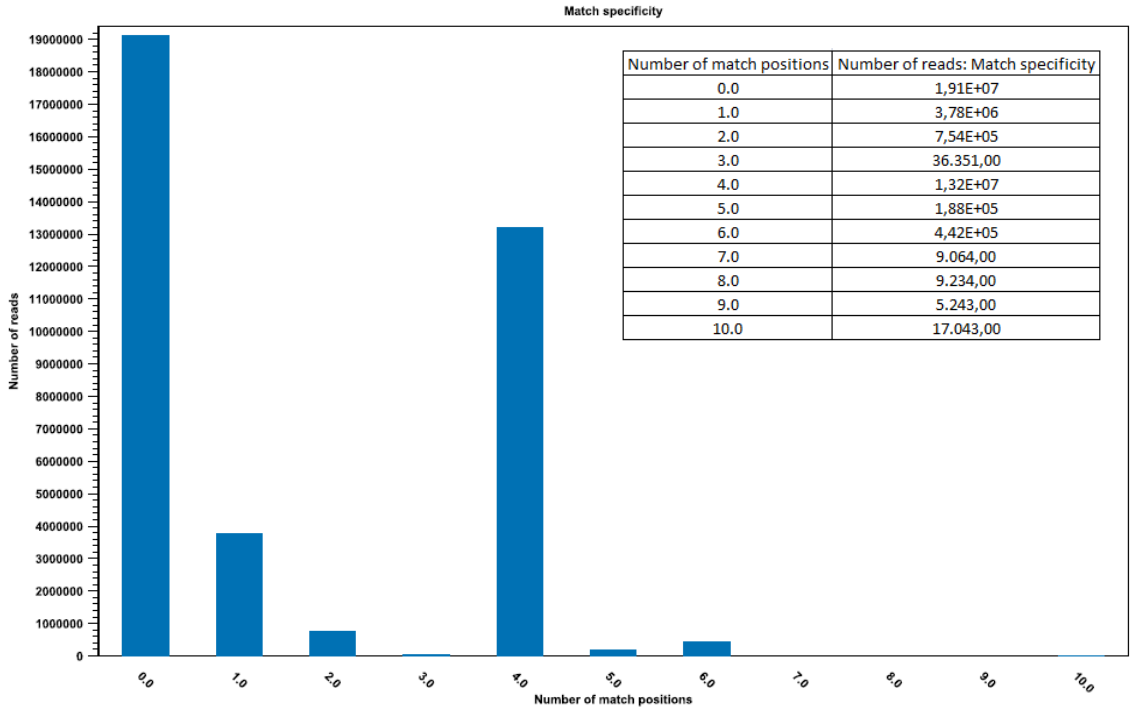
Tabla 3.2: Duración de los distintos RNA-Seq. Se especifica el origen y la condición de cada RNA-Seq respetando el mismo criterio presentado en la Tabla 1 a partir de los datos generados por Ingolia et al. Science 2009. También se especifica el largo de los fragmentos utilizados para realizar los mapeos.

origen	condición	largo de los fragmentos	duración (minutos)
ARNm total	control	21	70
		28	80
	no aac.	21	70
		28	70
RFP	control	21	100
		28	40
	no aac.	21	80
		28	20
		TOTAL	**8 hs 50 minutos**

El valor de RPKM es una medida de la densidad de lecturas a lo largo de un gen, lo cual refleja la concentración molar inicial de dicho transcripto en la muestra. También normaliza el número de lecturas por el largo exónico del gen y por el total de lecturas mapeadas[120] (ver sección 3.2 del capítulo 1).

Estos análisis por RNA-Seq suelen tardar unos cuantos minutos, en la Tabla 3.2 se especifica la duración de cada una de las corridas.

Culminados los análisis por RNA-Seq, la cantidad de información generada requiere de diversos tipos de análisis *a posteriori* para su mayor comprensión. Estos análisis se irán presentando de aquí en adelante. En una primera instancia se pueden estudiar los reportes generados por los RNA-Seq recientemente culminados. Dichos reportes presentan una amplia información respecto al alineamiento y mapeo realizado. Por ejemplo detallan los archivos utilizados, así como la cantidad y el largo de secuencias presentes en cada archivo. También presentan información acerca de la referencia utilizada y su tamaño en bases nucleotídicas, así como el número de genes identificados presentes en dicha referencia y distintos gráficos tipo histogramas donde se detallan el número de transcriptos por gen, el número de exones por gen, etc. A la



Number of match positions	Number of reads: Match specificity
0.0	1,91E+07
1.0	3,78E+06
2.0	7,54E+05
3.0	36.351,00
4.0	1,32E+07
5.0	1,88E+05
6.0	4,42E+05
7.0	9.064,00
8.0	9.234,00
9.0	5.243,00
10.0	17.043,00

Figura 3.4: Histograma de mapeos específicos. Se muestra mediante un histograma y sus correspondientes valores numéricos en una tabla, la cantidad de lecturas que mapean de forma específica de cero a diez veces en el mapeo realizado. A modo de ejemplo, existen 36.351 lecturas que mapean tres veces en la referencia. Las lecturas derivan de la condición de privación de aminoácidos luego de aplicar la técnica de ribosome profiling.

vez se presentan las cantidades de secuencias mapeadas y descartadas en forma de tablas con información general y detallada, y un gráfico interesante también del tipo histograma, donde se presenta cuantas lecturas tienen un número de mapeos específicos determinado (ver Figura 3.4).

Teniendo a disposición los valores y porcentajes de los distintos mapeos realizados en los análisis llevados a cabo por Ingolia y colaboradores en el trabajo en cuestión (accesibles en el material suplementario de dicho artículo) se procedió a comparar dichos valores con los que derivaron de los análisis propios, ya sea para los fragmentos de 21 y 28 nucleótidos de largo con los que se contaba, en las cuatro condiciones experimentales de trabajo. Dichas comparaciones se muestran mediante gráficas de barras en las figuras 3.5, 3.6, 3.7 y 3.8.

Al momento de interpretar los resultados mostrados en las distintas figuras anteriores, se debe tener en cuenta que los resultados obtenidos son distintos respecto del origen

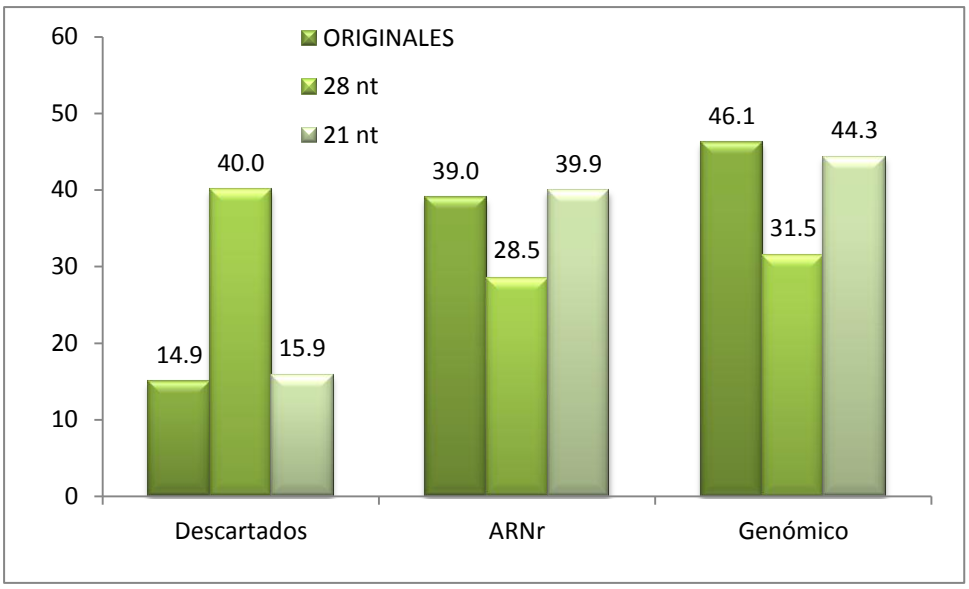


Figura 3.5: Comparación de los porcentajes de mapeos propios respecto a los presentados por Ingolia et al. 2009. Se presenta la condición y el origen de las secuencias. En este caso las mismas derivan de la condición control y provienen de una población total de mensajeros. La primer barra del gráfico presenta los porcentajes de mapeo aportados por el trabajo de Ingolia et al. 2009, y a continuación se presentan los porcentajes propios resultado de mapear las secuencias descargadas de dicho trabajo y recortadas previamente a largos de 28 y 21 nucleótidos respectivamente (ver texto).Se distinguen los porcentajes de secuencias descartadas, que mapearon contra ARN ribosomal (contaminación) y que mapearon contra el genoma.

de la muestra, sea esta proveniente de la población total de mensajeros o de los fragmentos protegidos por los ribosomas (*ribosome profiling*).

En el primer caso, se observa como los análisis efectuados son sensibles al largo de las secuencias mapeadas (ver Figuras 3.5 y 3.6). Por ejemplo, en la figura 3.5 se presentan los porcentajes de secuencias mapeadas para la condición control, partiendo de la población total de mensajeros. Los valores presentados por Ingolia y colaboradores en su artículo revelan que cerca de un 40% de sus lecturas mapearon en regiones correspondientes al ARN ribosomal, por lo que se consideran contaminación. Por otro lado, cerca del 15% de las secuencias fueron descartadas por características intrínsecas o porque no mapearon contra el genoma de levaduras de referencia utilizado. El restante 45% aproximadamente de las secuencias leídas mapearon efectivamente contra el genoma utilizado.

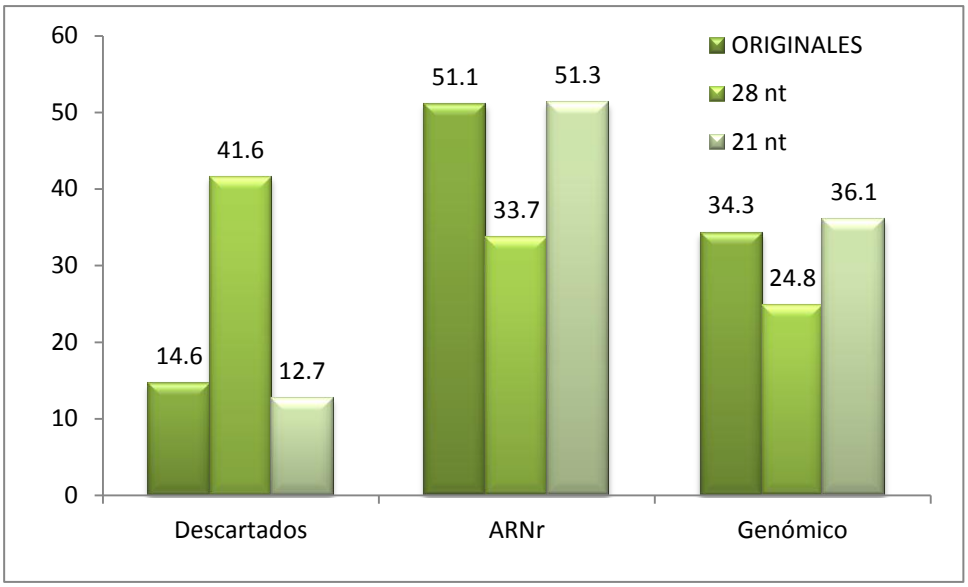


Figura 3.6: Comparación de los porcentajes de mapeos propios respecto a los presentados por Ingolia et al. 2009. Figura análoga a la figura 3.5, solo que en este caso se presentan los porcentajes de mapeos efectuados en base a la condición de privación de aminoácidos. Las secuencias se generaron también a partir de una población de mensajeros totales.

Cuando se realizó el mapeo en el laboratorio con las mismas secuencias recortadas a largos de 28 y 21 nucleótidos de largo los resultados fueron los esperados. El mapeo de las secuencias de 21 nucleótidos se asimiló claramente al mapeo descrito por Ingolia y colaboradores. Al trabajar con ese largo de secuencia aconsejado por los autores, los porcentajes de lecturas que eran descartadas y que mapeaban contra el genoma o ARN ribosomales fueron similares (14,9% y 15,9%; 46,1% y 44,3%; y 39,0% y 39,9% respectivamente; ver Figura 3.5). En cambio al realizar el mapeo partiendo de secuencias de largo mayor (28 nucleótidos) los porcentajes no fueron nada parecidos. La razón que explica estas diferencias se basa en que al trabajar con secuencias de este largo mayor se están incluyendo en las lecturas nucleótidos que no derivan de los fragmentos de ARN originales, sino que son resultados de la metodología utilizada para producir las bibliotecas. Por esto, el porcentaje correspondiente a las lecturas descartadas y que no mapean contra el genoma es ampliamente mayor en el caso del análisis con secuencias de 28 bases de longitud respecto a los estadísticos aportados por Ingolia (40,0% y 14,9% respectivamente). También disminuyó el número de lecturas que alinean tanto contra el genoma, como contra los ARN ribosomales (de 46,1% a 31,5% y de 39,0% a 28,5% respectivamente, ver Figura 3.5).

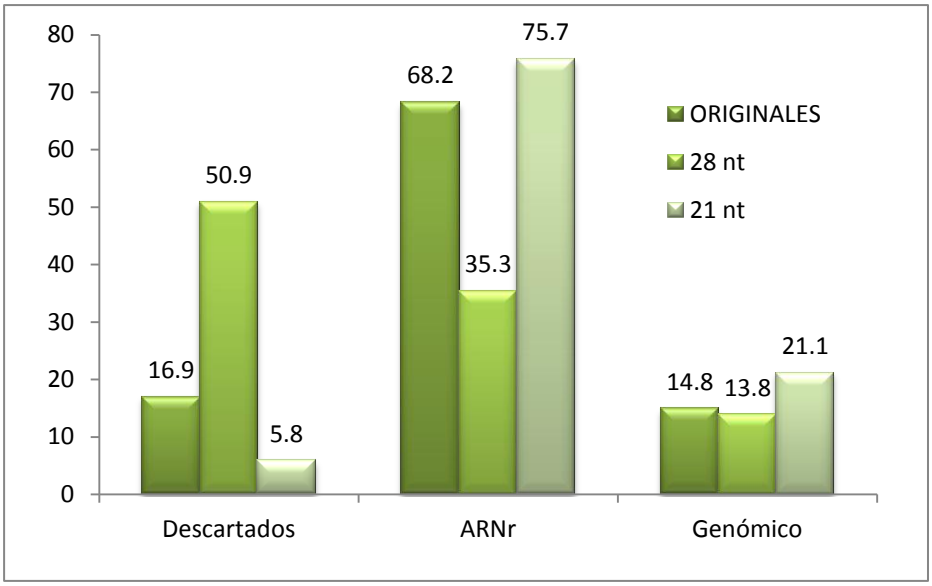


Figura 3.7: Comparación de los porcentajes de mapeos propios respecto a los presentados por Ingolia et al. 2009. Figura análoga a la figura 3.5, solo que en este caso se presentan los porcentajes de mapeos efectuados en base a secuencias derivadas de la condición control y que provienen de los fragmentos protegidos por los ribosomas (ribosome

El mismo patrón de resultados se obtiene para los mapeos realizados con las secuencias provenientes de la condición de privación de aminoácidos originados a partir también de una muestra de mensajeros totales (ver Figura 3.6).

Por otro lado se encuentran los porcentajes de mapeos generados en base a las secuencias derivadas de los fragmentos protegidos por los ribosomas, presentados en las en forma gráfica en las figuras 3.7 y 3.8. En estos dos casos los resultados obtenidos y la interpretación de los mismos es distinta a lo explicado hasta el momento, con los fragmentos derivados de la población total de mensajeros.

Los fragmentos que se mapearon en estas dos condiciones experimentales (control y privación de aminoácidos) derivaron del ensayo de *ribosome profiling,* por lo que se espera que dichos fragmentos presenten un largo promedio de 28 nucleótidos[8] (ver Figura 3.2). En estas condiciones hipotéticas de trabajo, las secuencias de 28 bases de largo no verán afectado su mapeo sensiblemente como en el caso anterior, ya que es esperable que los fragmentos en cuestión presenten dicho largo. De esta forma los mapeos con secuencias de 28 y 21 nucleótidos contra el genoma, en teoría deberían ser similares entre ellos y similares a la vez a los mapeos descritos por Ingolia y colaboradores.

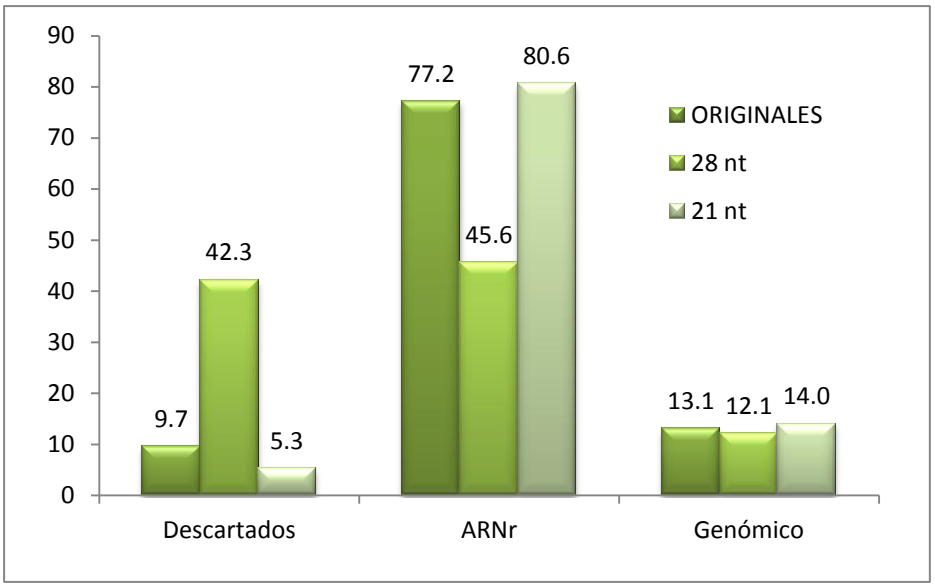


Figura 3.8: Comparación de los porcentajes de mapeos propios respecto a los presentados por Ingolia et al. 2009. Figura análoga a la figura 3.5, solo que en este caso se presentan los porcentajes de mapeos efectuados en base a secuencias derivadas de la condición de privación de aminoácidos y que provienen de los fragmentos protegidos por los ribosomas (ribosome footprinitng).

Justamente el comportamiento anterior es el observado. En las figuras 3.7 y 3.8 se puede observar que los porcentajes de lecturas que mapean contra al genoma en el caso de Ingolia *et al.* y en los análisis propios (ya sea con secuencias de 21 o 28 nucleótidos de largo) son muy parecidos: 16,9%, 21,1% y 13,8% para los fragmentos derivados de la condición control (ver Figura 3.7); y 13,1%, 14,0% y 12,1% para los fragmentos derivados de la condición de privación de aminoácidos (ver Figura 3.8). Aquí también se aprecia claramente como los porcentajes de mapeo correspondientes a los fragmentos de menor tamaño son mayores respecto a los fragmentos de mayor tamaño. Esto ocurre simplemente por una cuestión de probabilidad, es más probable que se mapee un fragmento de 21 nucleótidos que uno de 28, simplemente por su largo.

Al observar estas últimas dos figuras también ocurre algo interesante respecto a los porcentajes de mapeos en ARN ribosomales y en los valores de lecturas descartadas o que no corresponden al genoma. El comportamiento que se observa aquí es que al realizar los mapeos con los fragmentos más largos la cantidad de lecturas descartadas o que no mapean es considerablemente mayor al valor aportado por Ingolia *et al.* Este

comportamiento se acompaña de una disminución considerable del número de lecturas que corresponden a ARN ribosomales. En cambio en los mapeos realizados con los fragmentos de 21 nucleótidos de largo, los estadísticos se asemejan mejor respecto a los descritos en el artículo.

Lo que se piensa que ocurre en este caso es que ese desbalance es generado por un mismo efecto: dado que los fragmentos provenientes de contaminación con ARN ribosomal presentan un tamaño menor (aproximadamente 24-25 nucleótidos, ver Figura 3.2), al recortar las secuencias a un largo amplio de 28 bases, estos fragmentos mantienen bases derivadas de la metodología utilizada para generar las bibliotecas y por esto no mapean y son descartadas. Por otro lado, al cortar en mayor forma las secuencias y mapear solamente los primeros 21 nucleótidos de estas, los fragmentos de ARN ribosomal pierden todas las bases de origen metodológico (por llamarlo de alguna forma) y pueden mapear correctamente, lo que disminuye por efecto directo la cantidad de secuencias descartadas o no mapeadas.

Respecto a lo anterior puede observarse en las figuras 3.7 y 3.8 que cuando se trabaja con fragmentos de 28 nucleótidos de largo, la disminución porcentual, respecto a los valores aportados por Ingolia y colaboradores, en las lecturas asignadas a ARN ribosomal en las dos condiciones experimentales es similar al aumento porcentual en las secuencias descartadas o que no mapean contra el genoma (33,0% con 36,1%, y 31,6% con 32,6%). Ya que el porcentaje de secuencias que mapean contra el genoma se mantiene constante, lo anterior significa entonces que al cortar las secuencias a largos de 28 nucleótidos, una gran cantidad de secuencias de origen ribosomal son descartadas. Mientras que si se trabaja con las secuencias de 21 nucleótidos de largo, esas secuencias de origen ribosomal que antes eran descartadas ahora sí mapean en los correspondientes genes de ARN ribosomal.

A modo de ilustrar lo anterior, en la figura 3.9 puede observarse de forma gráfica la presentación mediante la cual el programa utilizado devuelve el mapeo realizado a partir de las secuencias importadas y el genoma de referencia elegido. En este caso se utilizaron las secuencias derivadas de la condición control (cortadas a largos de 21 y 28 nucleótidos) a partir de los fragmentos protegidos por el ribosoma, y el gen mostrado

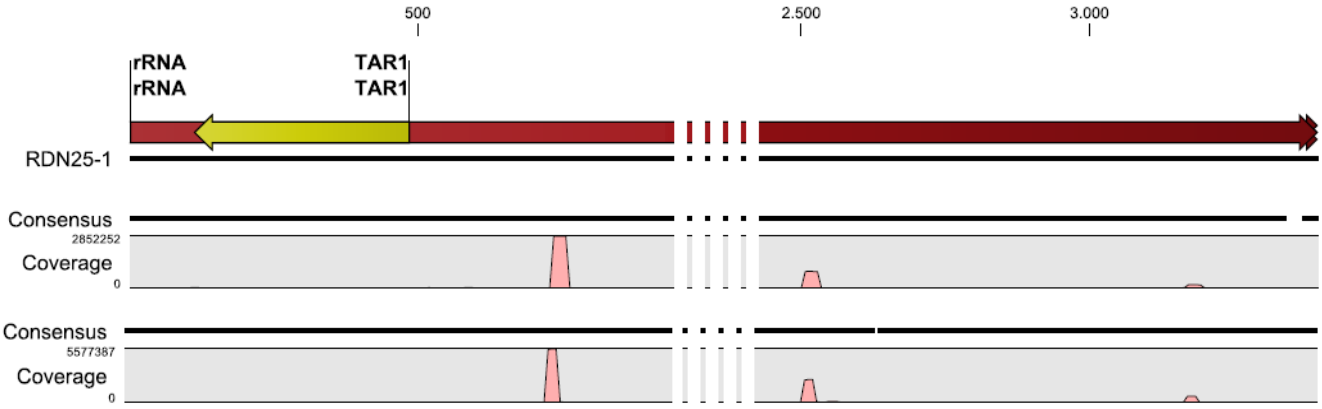


Figura 3.9: Representación gráfica del mapeo de las lecturas sobre el gen de ARN ribosomal RDN25-1. En la parte superior se observa el gen y una escala indicando la cantidad de nucleótidos. Abajo se observa el mapeo realizado con las lecturas de 28 nucleótidos de largo y por último, más abajo, el mismo mapeo pero con las lecturas de 21 nucleótidos de longitud. En los dos mapeos pueden observarse las regiones cubiertas por lecturas y los picos característicos de contaminación como se comentó (ver texto). A la izquierda de los mapeos se presenta la cobertura máxima alcanzada en dichos mapeos, a la cual se hace referencia en el texto.

es unos de los genes ribosomales de levaduras, *RDN25-1*. Resultado de aplicar la técnica *ribosome profiling* existe una importante contaminación con ARN ribosomal que es a veces difícil de solucionar. Se ha observado en base a los experimentos realizados en el laboratorio y a la experiencia formada en el trabajo con esta metodología, que dicha contaminación suele observarse como un pico importante dentro de los genes ribosomales, de muy alta cobertura. En la figura 3.9 puede observarse claramente ese patrón de un solo pico, y puede observarse que en el caso del mapeo con secuencias de 28 nucleótidos de largo, el mapeo tiene una cobertura mucho menor respecto al mismo mapeo con secuencias de 21 nucleótidos. La extensa diferencia entre ambos valores de cobertura (2.725.153 lecturas) es una prueba clara a favor del argumento presentado antes.

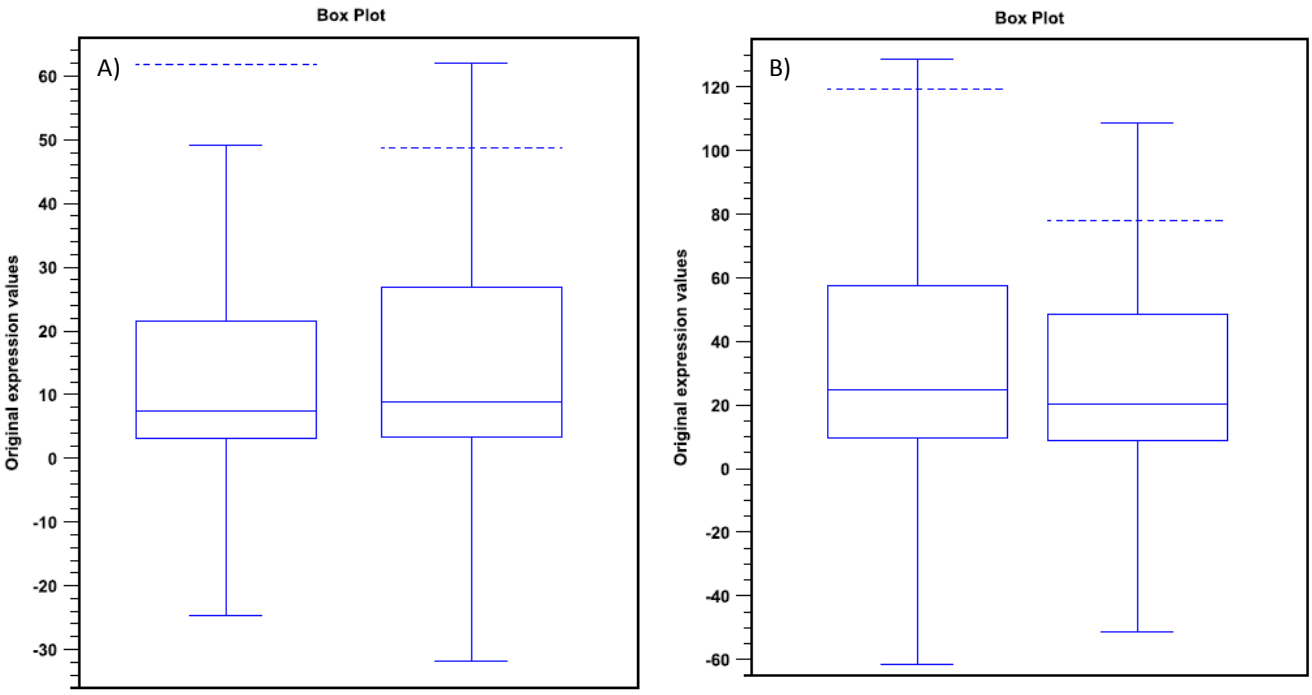
A la hora de efectuar análisis posteriores, se eligieron los mapeos realizados en base a las secuencias de 21 nucleótidos para los dos casos, ya sea para los fragmentos derivados de la población total de mensajeros para el estudio del transcriptoma, como para los fragmentos derivados de la técnica *ribosome profiling* para el estudio del traductoma. La razón por la cual se tomó esta decisión es que estos mapeos fueron los que mejor se acercaron a los mapeos presentados por Ingolia *et al.* en base a los

porcentajes de mapeos presentados. Si los porcentajes de mapeos difieren, se afecta el número total de lecturas mapeadas de la biblioteca, lo cual afecta de forma directa el valor de RPKM, ya que se modifica uno de sus parámetros involucrados en el cálculo.

3) CONTROL DE CALIDAD, ANÁLISIS PRELIMINARES Y ESTUDIO DE LOS PATRONES DE MAPEO

En una primera instancia se procedió a evaluar los cambios generales en los niveles de transcripción y traducción celulares de los genes de levaduras sujetas a la privación de aminoácidos. Se sabe que la privación de aminoácidos en levaduras produce sustanciales cambios en las tasas de transcripción y traducción[156], así como una disminución general en la iniciación de la traducción[156]. De la misma forma que se presenta en el artículo, se compararon medidas de expresión a nivel transcripcional y traduccional para el conjunto de genes que presentaban 128 ó más lecturas por gen. Dicho valor fue elegido en base a criterios estadísticos por los autores de forma tal que los genes con menos de 128 lecturas eran descartados, ya que no alcanzaban el nivel necesario para afirmar que están siendo detectados por la técnica. Esto se debe a que, en genes con menos de 128 lecturas, la variación entre réplicas supera ampliamente la variación biológica y dicha variación inter-réplicas se atribuye a efectos tanto propios de la partición binomial, como de la estadística de conteo.

Se presentan en las figuras 3.10 y 3.11 los *box plot* y *scatter plot* de las distintas condiciones experimentales trabajadas. Los esquemas tipo *box plots* y *scatter plots* son útiles para realizar controles de calidad en las muestras. Estas herramientas permiten tanto observar la distribución global de los datos, como analizar la variabilidad y similitud entre los set de datos, en busca de diferencias sistemáticas entre las muestras. Puntualmente en la figura 3.10 se presentan los *box plot* para las condiciones control (normal, levaduras en fase logarítmica de crecimiento) y sin aminoácidos, ya sea para el traductoma resultado de aplicar la metodología *ribosome profiling* (ver Figura 3.10A) o para el transcriptoma resultado de extraer la población de mensajeros totales (ver Figura 3.10B). En ambos casos se puede ver que tanto la mediana, como la media de los valores de RPKM disminuyen en la condición de privación de aminoácidos. Para el caso del estudio traduccional, resultado de la



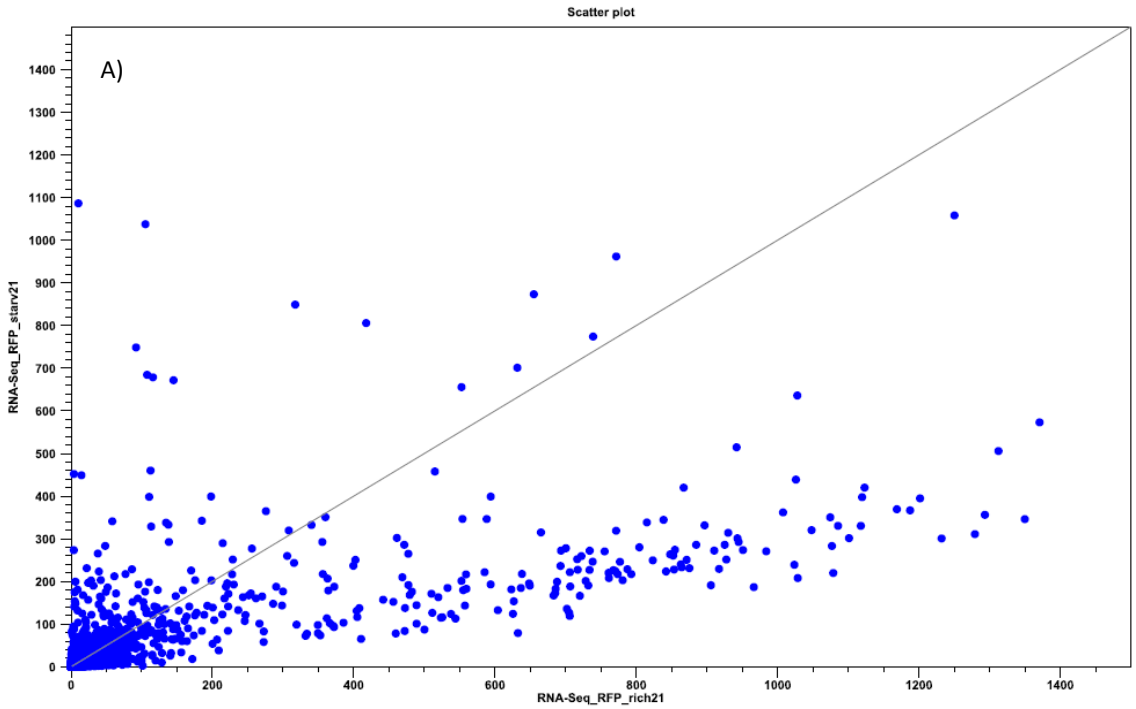
CONDICIÓN:	CONTROL	SIN AÁC.
Mínimo	0	0
Mediana	7,46	8,9
Media	61,78	48,67
Máximo	11.112,86	14.211,78

CONDICIÓN:	CONTROL	SIN AÁC.
Mínimo	0	0
Mediana	24,73	20,76
Media	119,27	78,22
Máximo	60.435,58	37.075,72

Figura 3.10: Box plot con los valores de expresión génica calculados mediante los mapeos y análisis por RNA-Seq, para las condiciones control (levaduras en fase logarítmica de crecimiento en un medio rico) y privación de aminoácidos, ya sea para el traductoma (A), como para el transcriptoma (B). Se muestran las cajas características de los box plot donde los límites inferior y superior de las cajas son el percentil 25 y 75; mientras que el largo de los bigotes es 1,5 veces el alto de la caja. Se muestra mediante una línea continua el valor proyectado de la mediana, y mediante una línea punteada el valor de la media. En las tablas de más abajo se pueden reconocer los valores más importantes de los box plot: el mínimo, mediana, media y el máximo. Tanto visualmente, como mediante los valores presentados en las tablas, puede observarse como las tasas de traducción y transcripción disminuyen resultado de crecer las levaduras en un medio deficiente en aminoácidos.

privación de aminoácidos se observa una distribución de los valores de expresión con mayor variabilidad, puesto que aumenta el alto de la caja y el largo de los bigotes. Para el caso del estudio a nivel transcripcional ocurre lo opuesto, los valores de expresión en la condición de privación de aminoácidos son menos variables.

Por su parte en la figura 3.11 se presentan los *scatter plot* resultado de graficar el valor de expresión génica en RPKM, en la condición de privación de aminoácidos en función del mismo valor pero en la condición normal. En este caso también se presentan dos *scatter plot*, uno derivado del estudio del perfil de expresión del traductoma y otro del



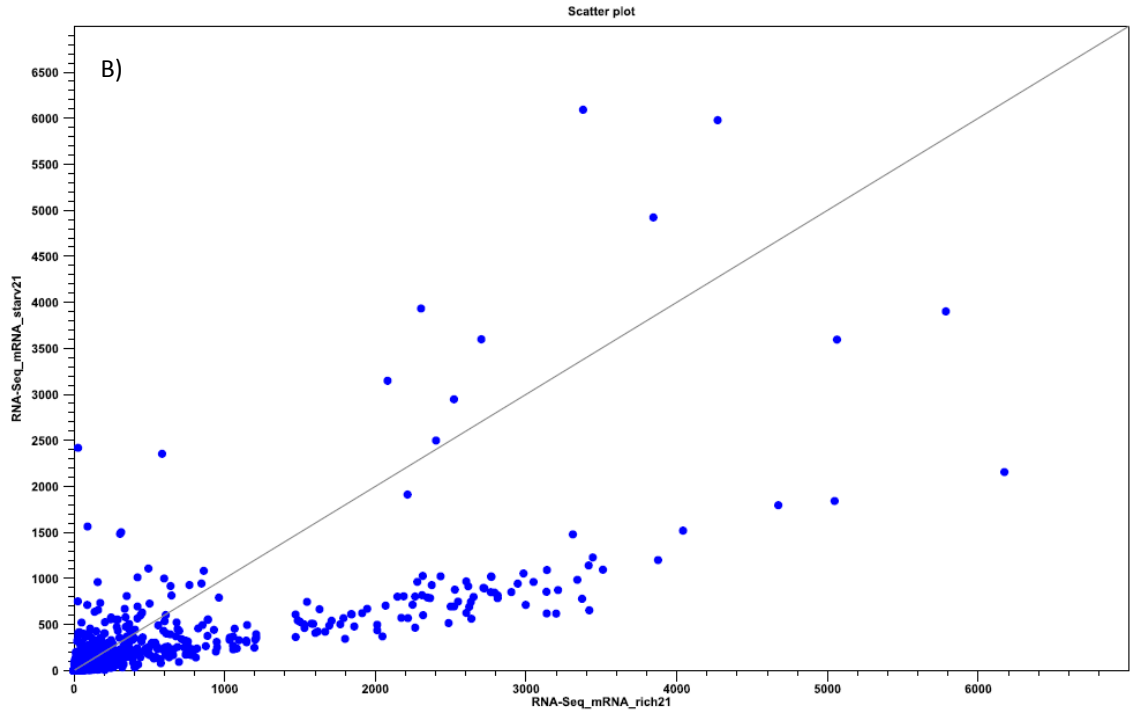
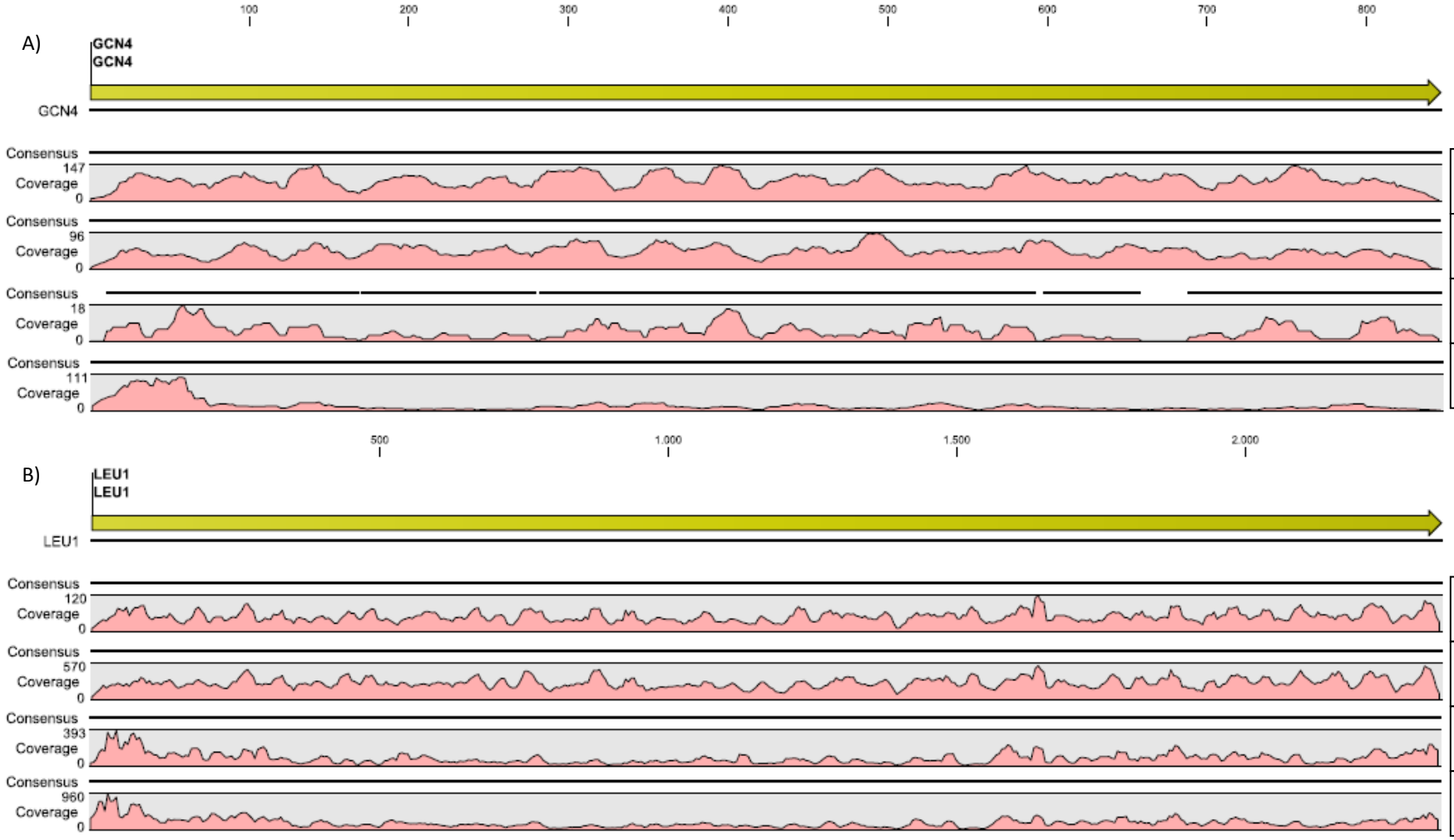


Figura 3.11: Scatter plot donde se comparan la condición control contra la problema en función de los valores de expresión génica (RPKM) de la totalidad de genes descritos en levaduras, para las dos sistemas de estudio, traductoma (A) y transcriptoma (B). En ambos scatter, en el eje vertical se grafica el valor de expresión en RPKM de los genes para la condición de privación de aminoácidos (condición problema), mientras que en el eje horizontal se grafica lo mismo pero para la condición control (levaduras en crecimiento en medio rico). También se muestra mediante una línea la recta y=x. Puede observarse claramente como una gran cantidad de genes disminuyen su nivel de traducción y transcripción, y quedan por lo tanto por debajo de la recta descrita anteriormente. Los puntos que quedan por encima de la recta son casos particulares y fueron identificados individualmente como genes de respuesta a condiciones de estrés.

transcriptoma (ver Figura 3.11A y 3.11B respectivamente). En este caso ambos *scatter plot* se realizaron a partir del total de los 6.281 genes de levaduras, ya que en cada condición en particular el número de genes que presenta más de 128 lecturas por gen es distinto en cada caso. Para realizar este tipo de gráficos el número de genes debe ser el mismo, por eso se partió del total de los genes presentes en el genoma de referencia de levaduras. En esta figura puede observarse claramente como una gran masa de puntos tiende a ubicarse por debajo de la línea *y=x*, ya que la mayoría de los genes disminuye su tasa de traducción o transcripción en condiciones de privación de aminoácidos. Sin embargo también pueden observarse algunos puntos que tienden a ubicarse por encima de dicha recta. Dichos puntos representan los genes de repuesta a condiciones de estrés que sirven de ayuda para que la célula sobreviva en un medio deficiente en aminoácidos. También puede observarse una gran masa de puntos acumulada hacia el origen ya que al trabajar con el total de los genes de levaduras, muchos de ellos tienen nula o baja expresión lo que acumula puntos en dicha región.

Resultado de estudiar la bibliografía disponible[156], se encontraron varios genes puntuales que pareció interesante estudiar debido a que eran regulados de forma opuesta a la mayoría de los genes. Esto significa que en lugar de verse a nivel transcripcional y/o traduccional, estos aumentaban en ambas tasas (ó en alguna de ellas) resultado de la privación de aminoácidos. Estos genes corresponden la mayoría a enzimas de biosíntesis de aminoácidos y a factores de transcripción particulares que aseguran la sobrevida de la célula en condiciones donde faltan los aminoácidos. Algunos ejemplos de estos genes son *GCN4, TRP1* y *LEU1*. En la figura 3.12 se muestran la densidad de lecturas mapeadas con el valor máximo de cobertura alcanzado en un gen y el valor calculado de expresión evaluado en RPKM, en las distintas condiciones. Se muestran algunos de los genes anteriores y otros donde los niveles de transcripción y traducción disminuyen de acuerdo a lo que ocurre de forma global. En dicha figura puede observarse como para ambas condiciones, control y privación de aminoácidos, los patrones de mapeos son similares. Sin embargo los valores de cobertura máxima alcanzada y los valores de expresión en cada caso son distintos. A modo de ejemplo se cita el caso de *GCN4* (ver Figura 3.12A; para una descripción de las funciones de este gen ver sección 1.1.4 del capítulo 1, página 21). En este caso, resultado del mapeo de



269,47 RPKM
170,63 RPKM
6,48 RPKM
27,37 RPKM
157,35 RPKM
961,08 RPKM
113,37 RPKM
328,88 RPKM

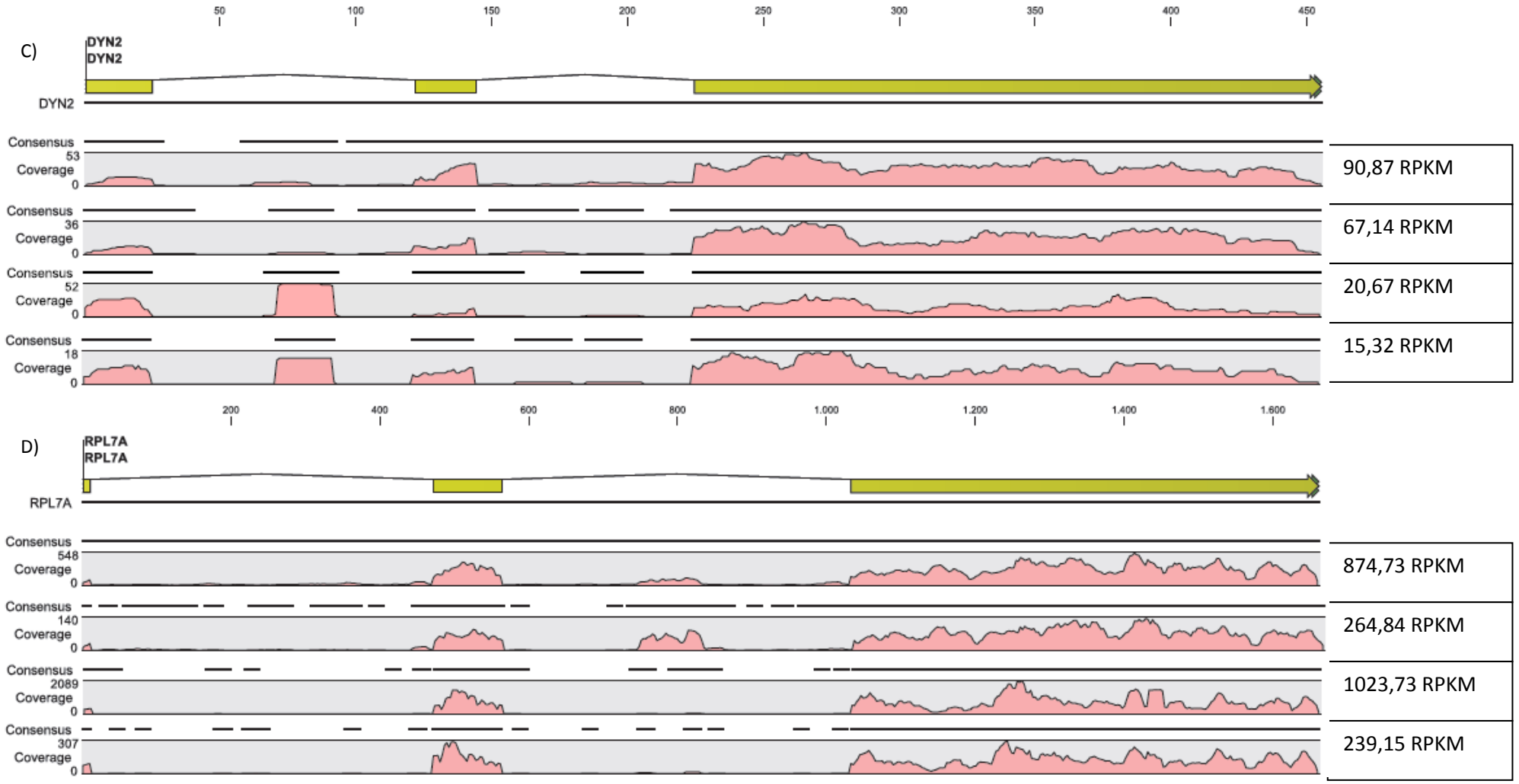


Figura 3.12: Representación gráfica de los mapeos realizados en las distintas condiciones sobre determinados genes de interés. Se muestra para cada gen estudiado, en la parte superior, la estructura génica del mismo donde se presenta de forma esquemática la secuencia codificante del gen, identificando el patrón de exones que define el mensajero maduro, para los casos donde existen intrones. Abajo se muestran los distintos mapeos en el siguiente orden: i) mapeo de los fragmentos derivados de la población total de mensajeros (estudio de expresión a nivel del transcriptoma) en la condición control; ii) lo mismo pero para la condición de privación de aminoácidos; iii) mapeo de los fragmentos protegidos por el ribosoma (estudio de expresión a nivel del traductoma) para la condición control; iv) lo mismo pero para la condición de privación de aminoácidos. Puede observarse como los mapeos alcanzan principalmente las regiones codificantes y apenas algunas zonas inter-exónicas (intrones). Más allá de analizar los mapeos cualitativamente a través de la forma del patrón de mapeo, los mismos pueden ser analizados cuantitativamente observando el valor de cobertura máximo alcanzado, presentado a la izquierda de los mapeos, y observando el valor de expresión calculado en RPKM presentado a la derecha de los mapeos. Los genes elegidos fueron: GCN4 (A), LEU1 (B), DYN2 (C) y RPL7A (D). Los dos primeros representan genes que se activan en respuesta a condiciones de estrés, como es la privación de aminoácidos (Hinnebusch et al. 1988), lo cual se observa de forma clara en la figura. Los otros dos son genes "comunes" donde se puede apreciar la baja en las tasas de transcripción y traducción.

los fragmentos que provenían de la población total de mensajeros, en la condición control la cobertura máxima alcanzada es de 147 mapeos (valor de expresión: 269,47 RPKM), mientras que en la condición de privación de aminoácidos la máxima es de 96 (valor de expresión: 170,63 RPKM). Lo opuesto ocurre con los mapeos realizados en base a los fragmentos protegidos por el ribosoma, donde en la condición control la cobertura máxima alcanzada es de 18 (valor de expresión: 6,48 RPKM), mientras que en la condición de privación de aminoácidos la máxima alcanza el valor de 111 (valor de expresión: 27,37 RPKM). Esto indica que resultado de la quita de aminoácidos del medio de crecimiento, las levaduras disminuyen la tasa de transcripción de *GCN4* pero aumentan en gran forma su tasa de traducción. Para el caso del gen *LEU1,* el cual codifica para una enzima encargada de catalizar el segundo paso en la biosíntesis de la leucina, se observa un aumento en la expresión tanto a nivel del transcriptoma, como del traductoma.

Por otro lado, puede observarse en los mapeos correspondientes a los genes *DYN2* y *RPL7A* (ver Figuras 3.12C y 3.12D), un descenso tanto en la cobertura máxima alcanzada en los mapeos, como en el valor de expresión calculado. Esto se debe a que se tratan de genes “comunes” donde la respuesta global a la privación de aminoácidos que se observa es la disminución en las tasas transcripcional y traduccional. En estos dos casos, también puede observarse una región de mapeo fuera de los exones mostrados. Estos casos pueden corresponder a la detección de intrones, tanto a nivel transcripcional (más claro en la Figura 3.12D) como a nivel traduccional (más visible en la Figura 3.12C)

4) REPRODUCCIÓN DE RESULTADOS, CONTRASTE DE LOS ANÁLISIS Y EXTENSIÓN DE LOS ESTUDIOS ABARCADOS EN EL ARTÍCULO ORIGINAL

En una primera instancia, se procedió a evaluar la reproducibilidad de la técnica estudiando la correlación entre las dos réplicas presentadas por los investigadores en el artículo, para cada condición biológica estudiada. Para esto se compararon mediante un gráfico los valores expresión en RPKM de las dos réplicas, derivados del

mapeo de los fragmentos originados de la condición control, ya sea por la técnica *ribosome profiling,* o derivados de la población total de mensajeros (ver Figura 3.13). En la figura, al igual que en las figuras originales presentadas por Ingolia y colaboradores (ver figura 3.14), pueden distinguirse los genes con menos de 128 lecturas por gen (considerados no detectados) y aquellos con 128 o más lecturas (detectados). Los genes que no presentaban ninguna lectura en las dos réplicas fueron descartados: 243 en el caso de los fragmentos protegidos por el ribosoma y 139 en el caso de los fragmentos derivados de la población total de mensajeros.

En la figura 3.13 puede observarse una buena correlación de valores de expresión entre las réplicas, ya sea para los fragmentos protegidos por el ribosoma (ver Figura 3.13A), como para los fragmentos derivados de la población total de mensajeros (ver Figura 3.13B). Los altos valores de los índices de correlación R^2 se asemejan bastante a los presentados por Ingolia y colaboradores en su publicación (ver Figuras 3.13 y 3.14). En dicha figura también puede observarse como para los genes con menos de 128 lecturas, indicados en verde, la correlación no es tan buena y existe una mayor variación inter-réplicas respecto de los genes detectados indicados en azul. A la vez puede conocerse el número de genes que superan la cantidad de lecturas marcada como límite de confianza estadístico y los que no. Se puede notar que la cantidad de genes detectados con confianza estadística por la técnica *ribosome profiling* es bastante menor que la cantidad de genes detectados cuando los fragmentos se originan de la población total de mensajeros. Esto se debe a que la gran mayoría de los genes se transcriben pero no necesariamente se traducen.

Otra figura que se consideró relevante para reproducir representa un plot donde se grafica el logaritmo en base dos del cambio en la eficiencia traduccional, en función del logaritmo en base dos del cambio en la cantidad de mensajero. La eficiencia traduccional fue descrita al principio de esta sección y representa cuanto se traduce un mensajero en función de su cantidad. Cuando se hace referencia a un cambio en la eficiencia traduccional o a un cambio en la cantidad de mensajero, dicho cambio representa el cociente del valor en cuestión en la condición de privación de aminoácidos, entre el mismo valor en la condición control. En este plot (ver Figura 3.15) se puede observar la repuesta traduccional y transcripcional a la privación de

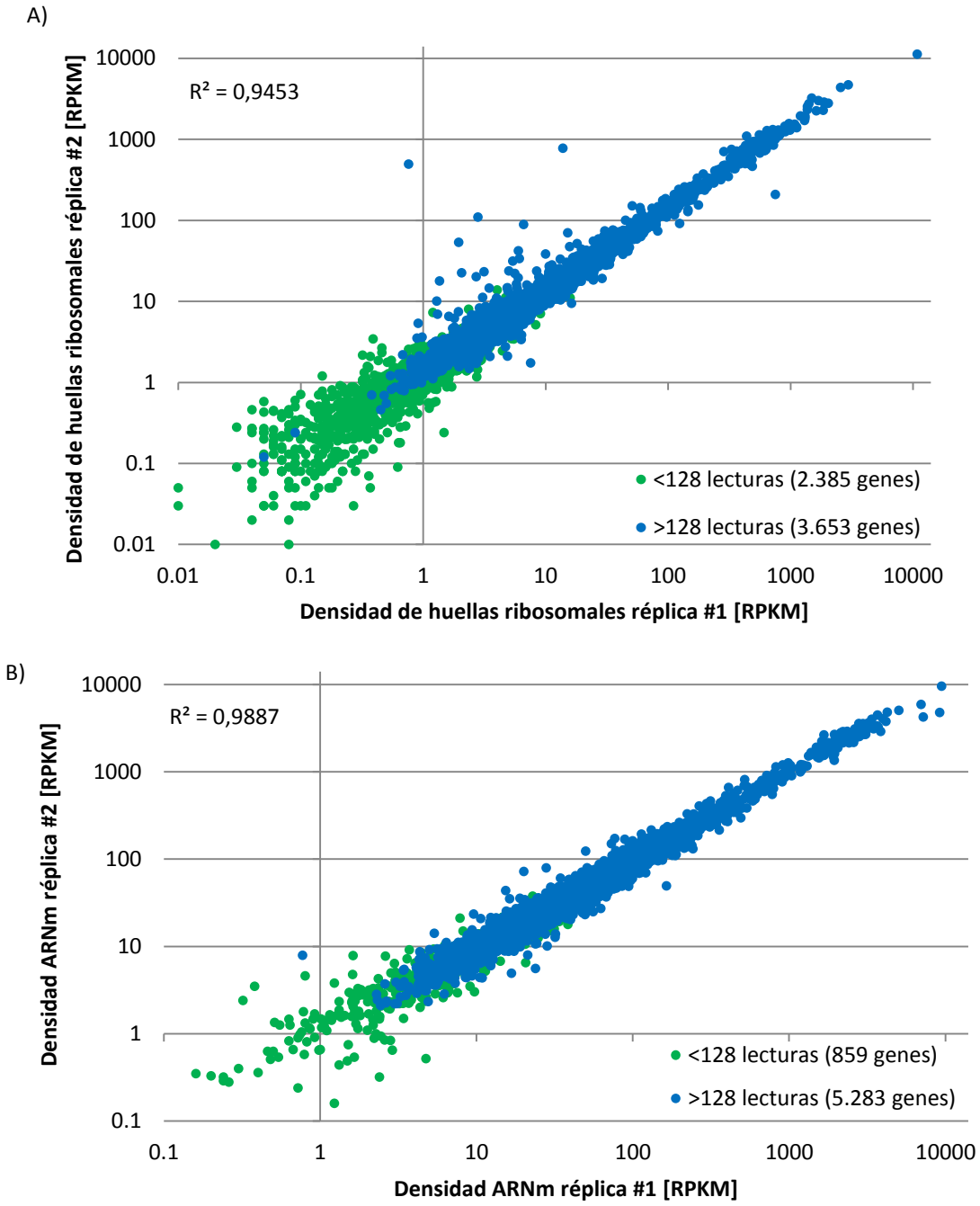


Figura 3.13: Correlación entre las dos réplicas presentadas por los investigadores, para los mapeos realizados tanto con los fragmentos derivados del ensayo de ribosome profiling (A), como con los fragmentos derivados de la población total de mensajeros (B). En ambos gráficos se representa el valor de expresión o densidad calculado por el análisis mediante RNA-Seq en unidades de RPKM, para las dos réplicas, representadas en cada eje. En colores se distinguen los genes con menos de 128 lecturas que se consideran no detectados y por lo tanto son descartados (en verde), y los genes con 128 o más lecturas (en azul) que son los que sí superan el valor establecido como límite de confianza estadística (ver texto). En ambos casos también se observa una buena correlación de las réplicas.

aminoácidos. En primera instancia, se observa a grandes rasgos que la gran masa de puntos se centra en el origen ya que no presentan considerables cambios en sus tasas transcripcionales y traduccionales, aunque existen muchos genes donde se identifica

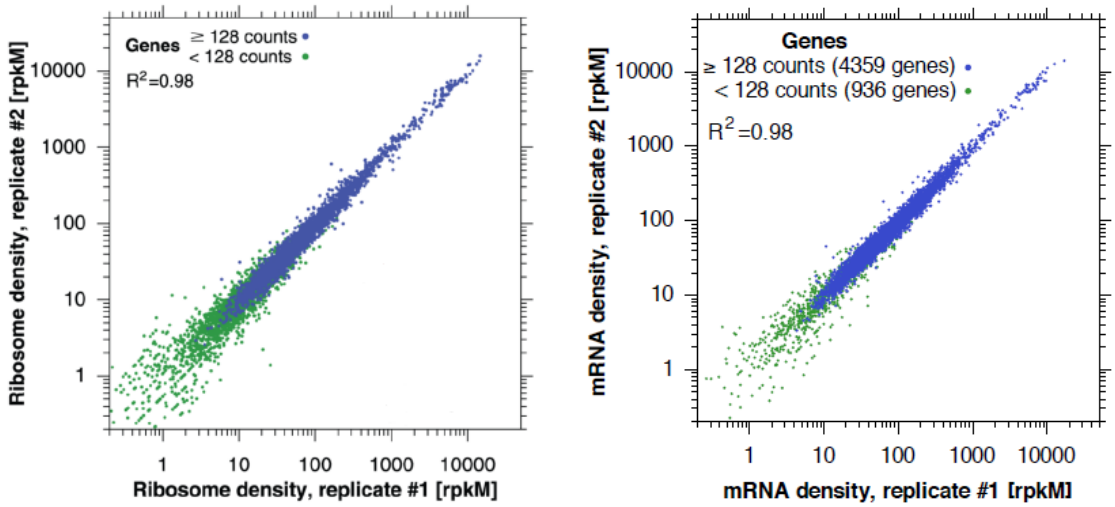


Figura 3.14: Correlación entre las distintas réplicas utlizadas. Figuras originales presentadas en la publicación de Ingolia et al. Sciencie 2009 (figuras 2D y S5A respectivamente). Se presentan a los efectos de comprarlas con las figuras presentadas en la figura 3.13. Para una descripción detallada ver texto al pie de dicha figura.

una respuesta interesante. Al respecto, los cuadrantes más interesantes son el superior izquierdo y el inferior derecho (identificados como SI e ID en la figura). En el primero, los genes que ahí aparezcan han reducido su tasa transcripcional pero han aumentado su eficiencia traduccional. En el segundo, la población de genes presentes ha aumentado su tasa transcripcional pero reducido su eficiencia traduccional. No llama la atención que *GCN4* se encuentre cercano al cuadrante superior izquierdo, ya que se mostró anteriormente (ver Figura 3.12A), su tasa transcripcional disminuye pero su eficiencia traduccional aumenta. En la Figura 3.16 se presenta la figura original presente en el trabajo de Ingolia *et al. Science* 2009 a los efectos de compararla con la figura anterior. Pueden observarse patrones similares en ambas figuras, así como una distribución similar de los genes coloreados que representan los genes responsables de la biogénesis ribosomal y una ubicación también similar para el gen *GCN4.*

También pareció interesante reproducir un resultado donde los autores muestran mediante un gráfico la mediana del valor de eficiencia traduccional, como función del largo de la secuencia codificante de los distintos genes. Para este gráfico los autores agrupan los genes en función del largo en codones de su secuencia codificante y calculan la mediana de la eficiencia traduccional de dicho conjunto de genes. De esta forma se construyó el gráfico original presentado en la figura 3.17. Esta figura

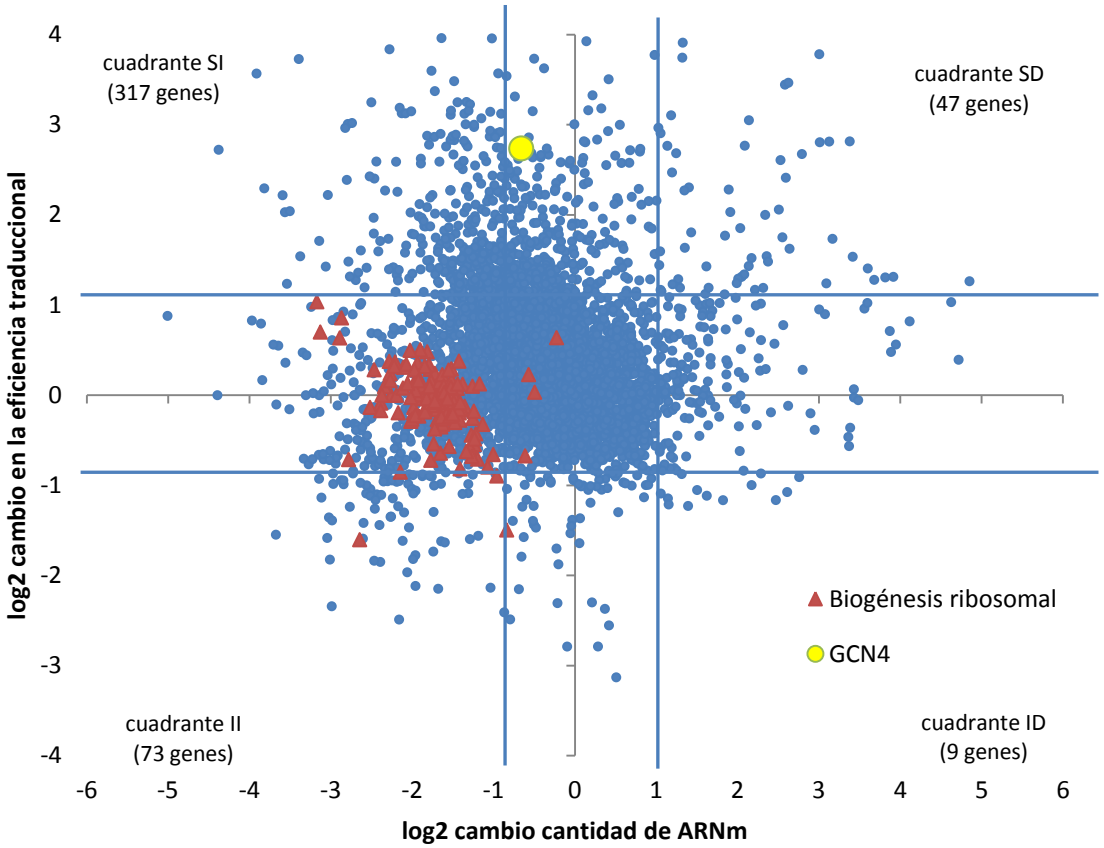


Figura 3.15: Respuesta traduccional a la privación de aminoácidos. Se muestran los cambios en la cantidad de mensajeros (eje horizontal), así como los cambios en la eficiencia traduccional (eje vertical). Se observa como la mayoría de puntos (genes) se mantienen cerca del origen, sin grandes cambios en sus tasas de transcripción y traducción; aunque se llega a apreciar una leve tendencia hacia la izquierda (disminución en la tasa transcripcional) en la masa de puntos. Se distinguen los genes asociados a la biogénesis ribosomal, GCN4 y los cuatro cuadrantes definidos.

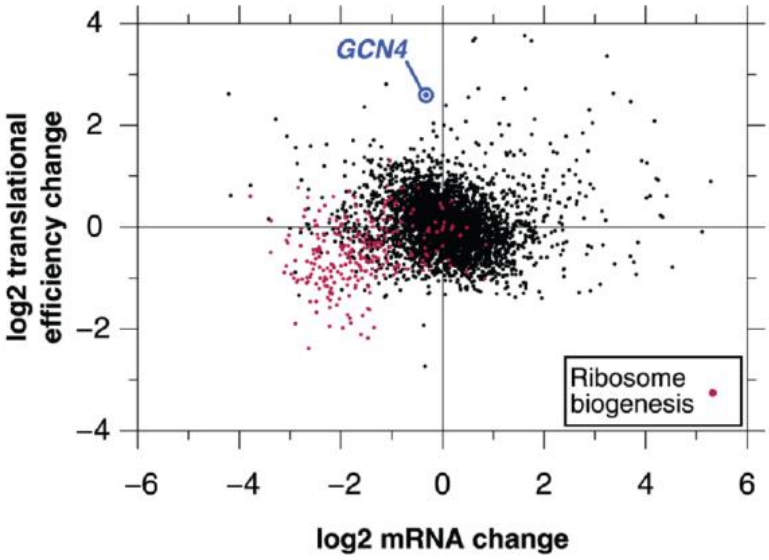


Figura 3.16: Respuesta traduccional a la privación de aminoácidos. Figura original presentada en la publicación de Ingolia et al. Sciencie 2009 (figura 4A). Se presentan a los efectos de comprarlas con las figuras presentadas en la figura 3.15. Para una descripción detallada ver texto al pie de dicha figura

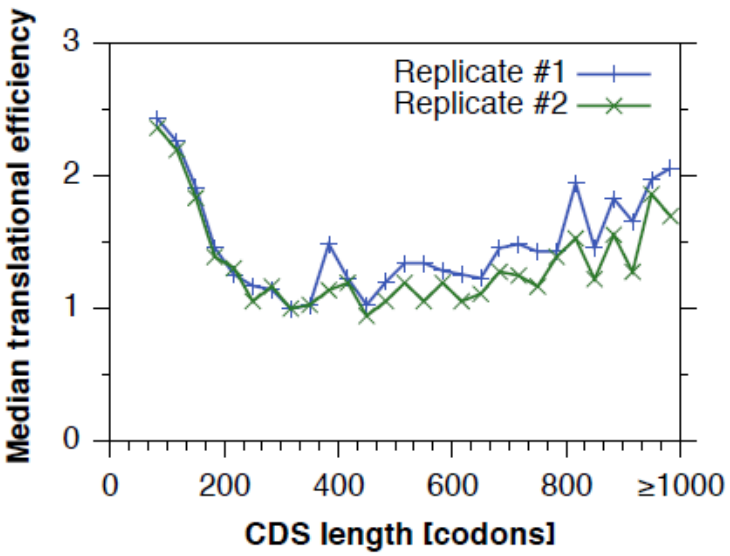


Figura 3.17: Eficiencia traduccional en función del largo (medido en codones) de la secuencia codificante de los distintos genes. Las medidas fueron tomadas del análisis de expresión por RNA-Seq (medido en RPKM), de los fragmentos protegidos por el ribosoma en condiciones de crecimiento normal. Se presenta la figura original presentada en la publicación de Ingolia et al. Science 2009 (figura S10). Se distinguen las dos réplicas con las que se trabaja en el artículo.

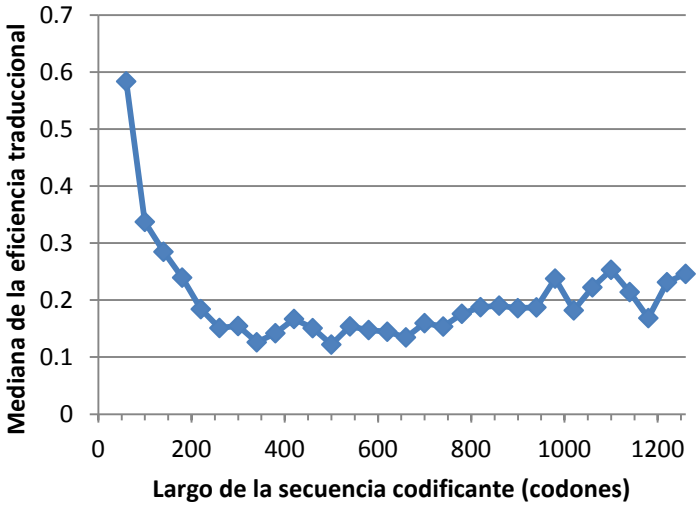


Figura 3.18: Eficiencia traduccional en función del largo (medido en codones) de la secuencia codificante de los distintos genes. Figura análoga a la figura 3.17. A grandes rasgos se observa el mismo patrón de gráfico en esta figura respecto a la anterior.

representa una componente clave para mostrar lo que ocurre *in vivo* respecto de la traducción de mensajeros en levaduras. En esta figura puede observarse como mensajeros más pequeños tienen una eficiencia traduccional mayor. Esto se debe a que los mensajeros más cortos tienden a tener mayor densidad ribosomal[119], ya que como observan Ingolia y colaboradores, existe una acumulación de ribosomas en los primeros 30 a 40 codones de un mensajero hasta los codones 100 a 200, después de los cuales la densidad se vuelve uniforme hasta el codón de terminación. Este exceso de ribosomas en el principio de los mensajeros se explica a raíz del uso de una droga que frena la elongación, ya que los eventos de iniciación no se detienen y más ribosomas se acumulan en los codones de iniciación sin poder avanzar. Esto último fue confirmado por Ingolia *et al.* en otra publicación posterior[147]. Más recientemente se

ha planteado la idea también de que una mayor concentración natural de ribosomas en el principio de los mensajeros de levaduras, sería una estrategia general conservada tipo “rampa”, con la cual se podría minimizar los atascamientos y colisiones entre ribosomas más adelante en los mensajeros[157].

Al momento de reproducir esta resultado mediante la construcción del gráfico correspondiente (ver Figura 3.18), en un principio se debieron quitar aquellos genes (1.453 genes) que codifican para ARNnc (ARN no codificantes). De esta forma, los genes se agruparon según el largo de su secuencia codificante medido en codones y se calculó la mediana de los valores de eficiencia traduccional. Los agrupamientos de mensajeros se realizaron de la siguiente manera: <60 codones; 61-100 y así consecutivamente sumando de a 40 codones; y >1260 codones. Puede observarse, comparando las figuras 3.17 y 3.18, que el patrón general en ambos gráficos es el mismo, comienza en valores altos que luego disminuyen y se estabilizan. Sin embargo a largos de mensajeros mayores se observa un leve aumento y una mayor variación en los valores de eficiencia traduccional.

Para profundizar en la reproducción de resultados, se decidió por examinar los valores de *fold change* en la eficiencia traduccional y en la cantidad de ARN mensajero, para aquellos genes que presentaban una respuesta considerable a la privación de aminoácidos. Se trabajó con las tablas presentadas en el material suplementario del artículo original, donde se mostraban aquellos genes que resultado de la privación de aminoácidos, presentaban un *fold change* mayor a 2 en cualquier dirección. También se construyeron tablas propias con aquellos genes en los que se detectaba el mismo cambio. Analizando dichas tablas mediante las aplicaciones y funciones disponibles en *Microsoft Office Excel 2007,* se construyó un diagrama de Venn (ver Figura 3.19), el cual mostró que para 235 genes consultados derivados de los datos de Ingolia y colaboradores, 212 estaban presentes en las tablas propias y sólo 23 no figuraban. Puede observarse en dicho diagrama que nuestros estudios son más permisivos en la detección de genes con respuestas significativas a la privación de aminoácidos, ya que detectamos 1207 genes mientras que Ingolia y colaboradores reportan 235. Actualmente se están estudiando las posibles razones de estas diferencias para entender las diferentes sensibilidades. De todas formas el hecho de detectar en los

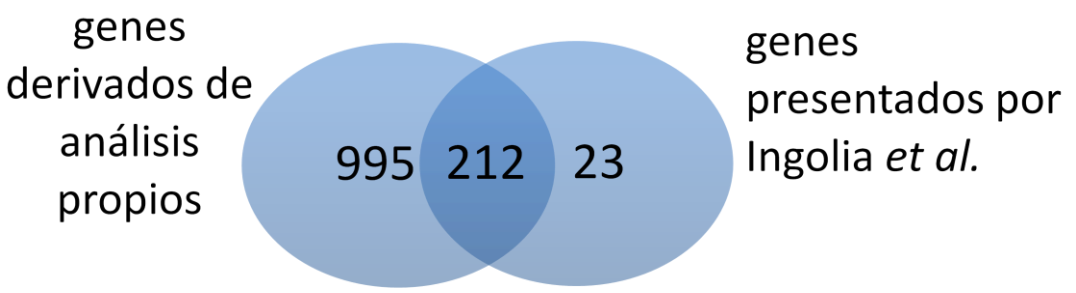


Figura 3.19: Diagrama de Venn mostrando la comparación de la identidad de los genes con valores de fold change mayores a 2 en cualquier dirección. Se comparan los genes presentados por Ingolia et al. y los genes derivados de los análisis propios realizados.

análisis propios la gran mayoría de los genes detectados por Ingolia y colaboradores es relevante pues está a favor respecto de validar la similitud entre ambos estudios.

Una comparación más cuantitativa de estos genes en los cuales se observa una respuesta positiva o negativa a la privación de aminoácidos, se realizó evaluando la correlación entre aquellos valores de *fold change* apreciados por Ingolia y colaboradores y los calculados en el laboratorio. De esta forma se construyeron distintas gráficas donde el eje horizontal correspondía por ejemplo, a los valores de *fold change* en la eficiencia traduccional presentados en el artículo original para los genes regulados positivamente (*fold change* > 2). Mientras que el eje vertical correspondía también a los valores de *fold change* en la eficiencia traduccional para esos mismos genes regulados de forma positiva, pero con los valores derivados de nuestros análisis (ver Figura 3.20). Lo mismo se graficó para los genes regulados de forma negativa (*fold change* < 0,5). También se realizaron gráficas del mismo estilo solo que en lugar de graficar los *fold change* en la eficiencia traduccional, se graficaron los *fold change* en la cantidad de mensajero, también para los genes regulados de forma positiva y negativa. De esta forma se construyeron cuatro gráficas de correlación cuyos valores de coeficientes de correlación se muestran en la tabla 3.3.

En dicha tabla pueden observarse valores cercanos a uno para los coeficientes de correlación lo cual también apoya la posibilidad de validar como similares los análisis propios y los realizados por Ingolia y colaboradores con sus propios métodos computacionales.

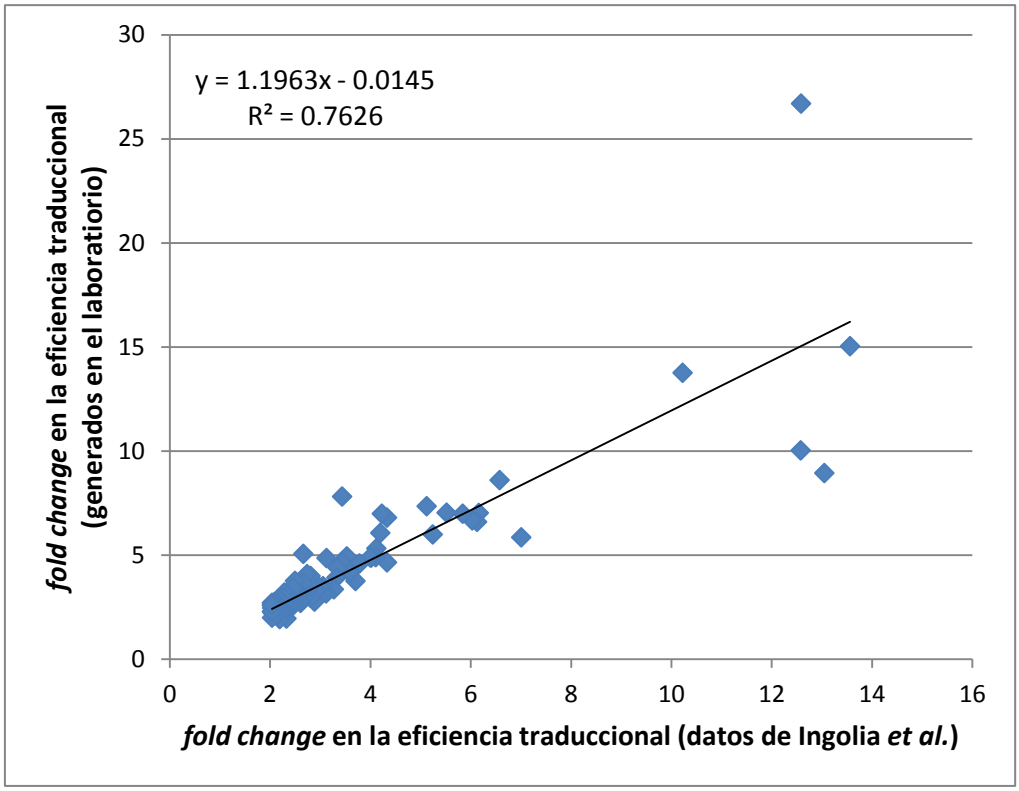


Figura 3.20: Correlación entre los valores de fold change calculados en los análisis propios respecto a los presentados por Ingolia et al. En este caso se muestran los valores de fold change en la eficiencia traduccional para aquellos genes donde se observa un aumento en la traducción (fold change > 2). En el gráfico se presenta el valor de R^2 como parámetro para evaluar la correlación.

El valor de coeficiente de correlación presentado en la tabla 3.3 difiere del presentado en la figura 3.20, pues fue re-calculado quitando los 3 posibles "*outliers*" presentes en la muestra.

Por otro lado, la posibilidad de utilizar bases de datos de acceso libre donde se analiza la ontología de genes, motivó estudiar las funciones de los distintos genes en los cuales se observaba una respuesta interesante a la privación de aminoácidos. Los genes seleccionados en este caso corresponden a los distintos cuadrantes de la figura 3.15. Este análisis en particular no fue realizado en los estudios presentados en el artículo original, donde solo se estudió la ontología de los genes con una respuesta traduccional positiva o negativa a la privación de aminoácidos. Los estudios de ontología permiten entender en que grandes procesos celulares se encuentran implicados la cantidad de genes seleccionados para el estudio, de una manera rápida, sin tener que analizar de forma individual cada gen, y con un respaldo estadístico de

Tabla 3.3: Se presentan los distintos valores de los coeficientes de correlación, resultado de comparar los valores de fold change calculados a partir de los análisis propios, respecto a los presentados por Ingolia et al. Se discrimina si el valor de fold change refiere a un cambio en la eficiencia traduccional, o en la cantidad del mensajero. También se especifica si dicho cambio es positivo (fold change > 2) o negativo (fold change < -2).

Fold change respecto a:	Genes regulados de forma:	Coeficiente de correlación (R^2)
Eficiencia traduccional	positiva	0,893
	negativa	0,856
Cantidad de mensajero	positiva	0,941
	negativa	0,995

confianza aportado por los algoritmos patentados que desarrollan los distintos programas de estudio de ontología de genes.

De esta forma se generaron distintas tablas incluyendo los genes de cada uno de los cuatro cuadrantes de la figura 3.15, definiéndolos a partir de las rectas *x=1, x=-1, y=1* e *y=-1* (estas rectas pueden observarse trazadas en la figura 3.15). Estas rectas representan los valores de *fold change* de 2 y -2, a nivel transcripcional (eje horizontal) y de la eficiencia traduccional (eje vertical). Esto implica por ejemplo, que al seleccionar los genes del cuadrante superior izquierdo (cuadrante SI en la figura), se seleccionarán aquellos que presenten un valor de *fold change* en la cantidad de mensajero menor a -2 y un valor de *fold change* en la eficiencia traduccional mayor a 2. Así se estarán evaluando las funciones de los genes que responden a la privación de aminoácidos aumentando su eficiencia traduccional pero disminuyendo en la cantidad de mensajero. De forma análoga se construyeron las tablas con genes para los otros tres cuadrantes. Estas tablas se estudiaron en la plataforma de acceso libre *DAVID*[158,159,160] *(Database for Annotation, Visualization and Integrated Discovery v6.7)* la cual fue desarrollada por el *LIB (Laboratory of Immunopathogenesis and Bioinformatics, SAIC-Frederick)* para dar apoyo al *NIAID (National Institute of Allergy and Infectious Diseases, NIH).* Como se observa en la figura 3.15, el cuadrante superior izquierdo es el más interesante de analizar pues es el que presenta aquellos genes que responden a la privación de aminoácidos disminuyendo en sus niveles transcripcionales, pero aumento sus niveles traduccionales (317 genes). Por su parte, el cuadrante inferior derecho presenta tan solo 9 genes por lo que no fue considerado para los estudios de ontología, ya que

contar con 9 genes no es suficiente para desarrollar estudios de ontología fundamentados desde un punto de vista estadístico.

En la tabla 3.4 se presentan las principales funciones y procesos celulares derivados del estudio de ontología para los cuadrantes a excepción del inferior derecho (genes con aumento a nivel transcripcional y descenso a nivel traduccional). El límite establecido para seleccionar las funciones celulares fue presentar un valor de p mayor a 0,05 considerando el sólido soporte matemático que existe atrás de este tipo de análisis[161].

Tabla 3.4: Procesos celulares derivados del estudio de ontología. Se presentan los tres cuadrantes estudiados y se especifica el número de genes asociada a cada proceso y el porcentaje que dicho número representa en el total de genes asociados a esa función o proceso celular. Se muestra también el valor de p asociado a detectar la presencia de cada función o proceso celular.

PROCESOS, ACTIVIDADES Y COMPARTIMENTOS CELULARES IMPLICADOS	# genes	%	valor de p
GENES CON AUMENTO TRANSCRIPCIONAL Y TRADUCCIONAL (CUADRANTE SUPERIOR DERECHO)			
Actividad Oxidorreductasa	15	21,7	2,90E-07
Biosíntesis de glucógeno	4	5,8	6,10E-05
Actividad Hexosiltransferasa	5	7,2	1,40E-04
Biosíntesis de Metionina	5	7,2	1,70E-04
NAD	7	10,1	3,00E-04
Biosíntesis de Aminoácidos	7	10,1	5,10E-04
Piridoxal Fosfato	5	7,2	1,40E-03
Actividad Glicosiltransferasa	6	8,7	1,60E-03
NADP	6	8,7	1,70E-03
Actividad Liasa	5	7,2	7,50E-03
Actividad Transferasa	14	20,3	1,40E-02
Fosforilación de Proteínas	37	53,6	2,60E-02
Biosíntesis de Serina	2	2,9	4,20E-02
GENES CON DESCENSO TRANSCRIPCIONAL Y AUMENTO TRADUCCIONAL (CUADRANTE SUPERIOR IZQUIERDO)			
Esporulación	23	7,6	2,00E-10
Meiosis	24	7,9	1,00E-09
Biogénesis/Degradación de la Pared Celular	17	5,6	3,90E-06
Glicoproteínas	40	13,2	2,30E-04
Unión al ADN	33	10,9	5,50E-04
Proteínas de Transmembrana	98	32,3	9,70E-04
Señalización Celular	27	8,9	2,00E-03

Metabolismo Lipídico	8	2,6	2,30E-03
Regulación de la transcripción	36	11,9	4,20E-03
Reparación del ADN	16	5,3	4,30E-03
Daño al ADN	16	5,3	8,60E-03
Biosíntesis de Timina	4	1,3	9,50E-03
Secreción	10	3,3	1,20E-02
Transcripción	36	11,9	1,50E-02
Retículo Endoplasmático	25	8,3	3,00E-02
Pared Celular	8	2,6	3,10E-02
Nucleosoma	3	1,0	3,90E-02
Membrana Celular	16	5,3	4,10E-02
Núcleo	91	30,0	4,20E-02
Glicosilación	6	2,0	4,50E-02
Transporte	48	15,8	4,50E-02
GENES CON DESCENSO TRANSCRIPCIONAL Y TRADUCCIONAL (CUADRANTE INFERIOR IZQUIERDO)			
Biogénesis Ribosomal	24	53,3	4,50E-27
Procesamiento del ARN	18	40,0	4,70E-17
Núcleo	36	80,0	6,30E-14
Fosforilación de Proteínas	37	82,2	2,30E-08
Unión al ARN	12	26,7	3,60E-06
Ribonucleopartículas	9	20,0	5,20E-04
Actividad Metiltransferasa	5	11,1	1,00E-03
S-Adenosil-Metionina	4	8,9	2,80E-03
Unión a Nucleótidos	5	11,1	1,50E-02
Actividad Helicasa	4	8,9	2,20E-02

En la tabla 3.4 puede observarse una separación entre las funciones celulares asociadas a los distintos cuadrantes. Particularmente se observan mayores diferencias entre los cuadrantes correspondientes a los genes que aumentan a nivel de su eficiencia traduccional y los que por el contrario disminuyen dicha eficiencia (cuadrantes superior izquierdo y superior derecho respecto al inferior izquierdo). Respecto a las funciones asociadas a los genes que disminuyen a nivel de su eficiencia traduccional y también a nivel transcripcional, se observa que se representan principalmente funciones relacionadas a la biogénesis ribosomal y procesamiento de ARN. Esto es en parte lógico ya que la célula en condiciones de bajos recursos energéticos tiende a minimizar la producción de proteínas, proceso celular que más energía consume (ver capítulo 1).

Por su parte las funciones celulares asociadas a los genes que aumentan su eficiencia traduccional en respuesta a la privación de aminoácidos son distintas. Puede verse un sesgo hacia la presencia de genes cuyas funciones están asociadas al metabolismo celular de aminoácidos. Esto también es lógico en el sentido de que la célula tiende a producir metabólicamente los aminoácidos ya que estos no se encuentran disponibles en el medio. Genes que clasifican dentro de las funciones celulares de biosíntesis de aminoácidos se observan tanto en los cuadrantes superior derecho, como izquierdo. Esto implica que en los genes encargados de la biosíntesis de aminoácidos se observa un sostenido aumento a nivel traduccional, más allá de que los niveles transcripcionales de estos genes puedan bajar o subir.

Dentro de las funciones asociadas a los genes con aumento traduccional se observan varios genes asociados a diversos tipos de metabolismos celulares. Esto también es de esperar pues en condiciones de bajos recursos energéticos, además de disminuir los niveles de síntesis proteica, se activan los procesos que producen energía como el metabolismo lipídico, y la producción de reservas energéticas como es la síntesis de glucógeno.

Se destaca también la presencia de varios procesos celulares asociados a la esporulación dentro de los genes que aumentan su eficiencia traduccional, pero disminuyen sus niveles transcripcionales. Dichos genes que se activan en estas condiciones preparan a las levaduras al posible pasaje al estado de espora, dadas las condiciones de estrés por las que atraviesan[162]. Por ejemplo, la presencia de procesos celulares como la meiosis y la biogénesis/degradación de la pared celular, están relacionados directamente al pasaje a estadío de espora en *Saccharomyces cerevisae*[163]. Al respecto se sabe que la meiosis es necesaria para reducir la ploidía de la célula y producir núcleos haploides que se puedan encapsular durante la formación de la espora[162]. Sin embargo, la formación de esporas requiere de una división celular particular donde las células hijas se forman dentro del citoplasma de la célula madre. Este proceso requiere la generación de dos estructuras celulares particulares: un compartimiento membranoso dentro del citoplasma que defina la membrana plasmática de cada espora, así como una extensa pared, característica de las esporas, que las proteja de amenazas ambientales[163].

Por último se analizó la ontología de aquellos genes donde no se observaba un cambio significativo a nivel transcripcional, pero en los cuales se detectaba un apreciable aumento en la traducción. Dichos genes pertenecen a la franja central superior comprendida entre las rectas *x=1* y *x=-1*, considerando los genes con valores de *y>1.* Esto implica genes con un fold change en la eficiencia traduccional mayor a 2, pero con un fold change en la cantidad de transcripto comprendido entre 1 y -1 (ver Figura 3.15). En dicha franja se presentan 485 genes, dentro de los cuales se destaca la presencia de *GCN4* (ver sección 1.1.4 del capítulo 1, página 21). El estudio de la ontología de dichos genes realizado de forma análoga a los anteriores, arrojó la información presentada en la tabla 3.5. En dicha tabla pueden observarse funciones

Tabla 3.5: Tabla análoga a la tabla anterior, mostrando en este caso los procesos celulares derivados del estudio de ontología realizado a partir de los genes de la franja central superior donde se detecta un significativo cambio en la eficiencia traduccional, pero no se aprecia un cambio a nivel transcripcional.

PROCESOS, ACTIVIDADES Y COMPARTIMENTOS CELULARES IMPLICADOS	# genes	%	valor de p
GENES CON SIN CAMBIOS TRANSCRIPCIONAL Y AUMENTO TRADUCCIONAL (FRANJA CENTRAL SUPERIOR)			
Transporte	86	18,9	2,20E-05
Transporte de azúcares	11	2,4	3,10E-05
Membrana Mitocondrial Interna	29	6,4	4,60E-05
Perixosoma	14	3,1	5,50E-05
Actividad Oxidorreductasa	36	7,9	9,00E-05
Mitocondria	75	16,5	2,30E-04
Glicproteínas	54	11,9	2,60E-04
Autofagia	10	2,2	5,30E-04
Membrana	164	36,0	6,40E-04
Proteínas de Transmembrana	138	30,3	1,70E-03
Esporulación	16	3,5	2,20E-03
Respuesta al Estrés	14	3,1	5,50E-03
Coenzima A	7	1,5	6,60E-03
Meiosis	16	3,5	1,10E-02
Unión a Metales	62	13,6	1,20E-02
Metabolismo de Carbohidratos	8	1,8	1,30E-02
Biogénesis del Perixosoma	5	1,1	1,40E-02
Conjugación	4	0,9	1,50E-02
Metabolismo de Ácidos Grasos	5	1,1	1,80E-02
Cadena Respiratoria	6	1,3	3,80E-02

análogas a las asignadas a los genes donde también se detectaba un aumento traduccional, con aumento o descenso a nivel transcripcional. Dichas funciones son por ejemplo la esporulación, genes involucrados en la respuesta al estrés, la presencia de genes asignados al metabolismo de carbohidratos y ácidos grasos utilizado para producir energía, etc. Esta tendencia a observar de forma repetida dichas funciones celulares marca la importancia de la regulación traduccional que existe en este tipo de fenómenos. Más allá del cambio positivo, negativo o despreciable que pueda existir a nivel transcripcional para muchos genes, se observa un fuerte aumento a nivel traduccional para aquellos encargados de la producción de aminoácidos, generación de recursos energéticos y hasta de la activación de mecanismos involucrados a la esporulación, como camino alternativo de las levaduras para sobrevivir a las condiciones de estrés por las que están atravesando.

De esta forma se puede observar que, mediante el análisis de ontología realizado a partir de los genes en los cuales se observaba una respuesta traduccional y/o transcripcional significativa a la privación de aminoácidos, a grandes rasgos se obtuvieron los resultados esperados. Los análisis de ontología efectuados en la plataforma *DAVID* nos permiten darle sentido a las listas de genes que obtenemos producto del estudio de expresión diferencial. Así, dado el respaldo informático y matemático de dicha plataforma[158,159,161] podemos comprender de forma confiable cuales son los procesos, funciones y compartimentos subcelulares involucrados en nuestra lista de genes. Más allá de que en este caso el resultado obtenido es el esperado, esta poderosa herramienta de análisis de ontología permite plantearse nuevos desafíos e intentar responder nuevas interrogantes, sin importar si el resultado a obtener será novedoso o no.

CAPÍTULO 4

ESTUDIO DE LA INFLUENCIA TRADUCCIONAL DE *PDCD4*

Los principales objetivos planteados en esta instancia de trabajo fueron la construcción de listas con genes candidatos a ser regulados de forma traduccional por parte de *PDCD4* así como el estudio de la ontología de dichos genes candidatos. Esto implica a la vez, evaluar esos genes desde las vías de señalización celular que forman parte y como interaccionan con otras vías. También los objetivos involucran un análisis de los patrones generales de mapeos observados sobre dichos genes candidatos, así como otro análisis especial para genes con mapeos con características particulares. A su vez de forma más específica los objetivos comprenden realizar estudios generales para normalizar los datos y analizar a nivel global el alcance y resolución de la metodología aplicada a este tipo de problema biológico.

A continuación se describe brevemente la metodología experimental aplicada para generar las huellas ribosomales con las que se cuenta en el laboratorio. En esta etapa experimental se trabajó con una línea celular derivada de tejido de mama humano, T47D, la cual es usada frecuentemente para realizar experimentos con tejidos de mama. Esta línea celular deriva de un cáncer de mama, sin embargo estudios preliminares mostraron que dicha línea no presenta patrones alterados respecto a la expresión de *PDCD4.*

Específicamente se trabajó con dos condiciones por paralelo, una denominada siPDCD4, fue transfectada con un ARN de silenciamiento (siARN) contra el gen *PDCD4* a los efectos de silenciar la expresión de este factor en sistema de estudio. La otra condición utilizada como control, fue transfectada con un siARN control que no silencia ningún gen (*scrambled)*. A ambas condiciones, se les aplica la técnica *ribosome profiling* para generar millones de fragmentos de ARN que quedan protegidos por los ribosomas de la digestión con ARNasas. En este caso, por tratarse de un experimento piloto, no se cuenta con una extracción de ARN total a los efectos de estudiar la

expresión del transcriptoma. Por esta razón debe tenerse presente que a lo largo de los siguientes análisis y las distintas interpretaciones de los resultados arrojados, en ningún caso los resultados están normalizados según los cambios en la expresión del transcriptoma, solo se observarán cambios a nivel de expresión del traductoma. Esta limitante restringe de forma considerable la capacidad de extraer conclusiones precisas respecto al problema biológico aquí estudiado. Sin embargo, estudiar el problema sólo desde el punto de vista traduccional, no implica que no se puedan extraer conclusiones preliminares respecto a genes donde se observa un sostenido cambio traduccional, quedando dichos genes sujetos a análisis posteriores más exhaustivos.

1) ESTUDIOS PRELIMINARES, ALINEAMIENTOS Y ANÁLISIS MEDIANTE RNA-Seq

En el laboratorio ya se contaba con las secuencias de los fragmentos derivados de la técnica *ribosome profiling,* para las dos condiciones de trabajo como se explicó anteriormente. Dichas secuencias también se encontraban procesadas, esto significa que se les había aplicado un *trimming,* donde se eliminaron secuencias con un bajo valor de *quality score,* menor a 0,05. También fueron removidos en el mismo *trimming* nucleótidos propios de los adaptadores utilizados en la metodología. Por último también se quitaron aquellas secuencias que, producto de remover los nucleótidos correspondientes a los adaptadores, presentaban un largo menor a los 18 nucleótidos.

Analizando la información disponible en los archivos puede determinarse que, en un principio se contaba con 150.372.590 secuencias de 35 nucleótidos de largo cada una, correspondientes al estudio de la condición control, y 177.972.812 secuencias también de 35 nucleótidos de largo pero correspondientes a la condición siPDCD4. Esto significa que se secuenciaron 5,26 y 6,23 gigabases (Gb) respectivamente en el estudio realizado. Pueden apreciarse las diferencias respecto al primer trabajo publicado por Ingolia y colaboradores, donde en ese se secuenciaron un total de 3,49 Gb para realizar el estudio completo de las dos condiciones de trabajo, tanto a nivel del transcriptoma como a nivel del traductoma (ver Tabla 3.1). Esta cantidad apreciablemente menor se explica teniendo en cuenta el considerable menor tamaño

que presenta el genoma de levaduras respecto al genoma humano (un poco más de 12 millones de pares de bases contra casi 3.100 millones en humanos). Por ejemplo, el genoma de levaduras consta de poco más de 6.000 genes mientras que el de humanos presenta más de 35.000. Esto determina que para abarcar lo extenso del genoma humano deba ser necesario generar una mayor cantidad de secuencias, mientras que el genoma de levaduras es cubierto por una cantidad menor. También la diferencia se explica en parte por los avances realizados respecto a las tecnologías de secuenciación masiva, su poder y alcance. Sin embargo estos avances suben de uno a dos órdenes solamente la capacidad de los secuenciadores.

De todas formas, luego de procesar y descartar las secuencias mediante el *trimming* descrito, solamente quedan disponibles un total de 85.546.463 secuencias con un largo promedio de 25,4 nucleótidos para la condición control. Mientras que para la condición siPDCD4, quedan disponibles para trabajar un total de 78.740.756 secuencias con un largo promedio de 24,4 nucleótidos.

A partir de dichas secuencias se realizan los mapeos contra el genoma humano de referencia disponible en el laboratorio (GRCh37.p2) y los correspondientes cálculos de los valores de expresión génica mediante un análisis por RNA-Seq, para cada condición de trabajo estudiada. Los parámetros de mapeo elegidos para correr estos análisis fueron los mismos que los elegidos en los correspondientes RNA-Seq realizados al momento de reproducir los análisis de Ingolia *et al. Science* 2009. La gran cantidad de secuencias disponibles para mapear aumentó considerablemente los tiempos de duración de los mapeos: ambos análisis consumieron más de 9 horas.

En una primera instancia, luego de finalizados los dos mapeos y análisis mediante RNA-Seq, se evaluó la cantidad de secuencias presentes originadas por contaminación con ARN ribosomal producto de la metodología utilizada, a los efectos de comparar dichos porcentajes con los valores derivados del trabajo de Ingolia *et al. Science* 2009. En la tabla 4.1 se presentan los porcentajes de mapeos, para las dos condiciones de trabajo estudiadas. En este caso se pueden reconocer importantes diferencias respecto a los valores presentados en la reproducción de los análisis de Ingolia y colaboradores, mostrados en las figuras 3.7 y 3.8 (ver capítulo 3). La diferencia que más llama la

Tabla 4.1: Se presentan los porcentajes de mapeo y cantidad de secuencias que alinean contra el genoma y contra los genes correspondientes al ARN ribosomal (contaminación), así como también la cantidad de secuencias descartadas. Se discriminan las dos condiciones experimentales de trabajo: control y siPDCD4.

Descripción	CONTROL		siPDCD4	
Total	85.546.463	*100,0*	78.740.756	*100,0*
descartados y no genómico	42.231.634	*49,4*	40.596.849	*51,6*
no descartados	43.315.388	*50,6*	38.143.907	*48,4*
ARNr	**6.854.835**	***8,0***	**4.616.605**	***5,9***
Genómico	**36.460.553**	***42,6***	**33.527.302**	***42,6***

atención es la considerable baja en el porcentaje de lecturas asignadas a contaminación con ARN ribosomal y el correspondiente aumento en las lecturas que mapean en el genoma. Los porcentajes de contaminación con fragmentos de ARN ribosomal disminuyen desde valores del 70% hasta valores entre el 6% y 8%, mientras que los porcentajes de lecturas que mapean en el genoma ascienden desde valores cercanos al 20% hasta valores de más del 40%.

Es preciso notar que el descenso en el número de lecturas derivadas de la contaminación con ARN ribosomal genera un aumento en el porcentaje de lecturas que mapean contra el genoma. Esto se debe a que al aplicar la técnica de *ribosome profiling,* el número de lecturas generadas que mapean contra el genoma no varía sustancialmente. De esta forma el valor del porcentaje de lecturas que mapean contra el genoma depende del total de lecturas mapeadas, y como en este caso existen menos lecturas derivadas de la contaminación con ARN ribosomal, ese total disminuye y el porcentaje aumenta.

La principal razón que explica el descenso en la cantidad de contaminación con fragmentos de ARN ribosomal, es una modificación en uno de los pasos a realizar en la metodología de *ribosome profiling.* El cambio introducido respecta al método por el cual se extraen los polisomas, a partir de los cuales se procede a la digestión con ARNasas. En el primer trabajo publicado por Ingolia y colaboradores, estos primero realizan la digestión con ARNasa y luego extraen los ribosomas mediante centrifugación en gradiente continuo de sacarosa. Por este método, una vez finalizada

la centrifugación el gradiente se fracciona en pequeñas porciones, y en función de los valores de absorbancia de dichas fracciones se eligen las correspondientes a los monosomas. La modificación introducida permite primero extraer los polisomas utilizando una centrifugación en gradiente discontinuo o colchón de sacarosa, y luego realizar la digestión. De esta forma, los monosomas y subunidades libres de los ribosomas quedan en la porción superior de menor densidad de sacarosa, mientras que los polisomas alcanzan el fondo del tubo ya que logran atravesar la interface entre las dos soluciones de sacarosa de distinta densidad. Así se minimiza la contaminación con subunidades libres y monosomas, la cual genera la gran cantidad de lecturas que derivan de fragmentos de ARN ribosomal. Esta modificación en la metodología fue posteriormente aplicada por Ingolia y colaboradores, en el tercer artículo que publican utilizando la técnica de *ribosome profiling*[147].

A continuación, el siguiente análisis planteado fue evaluar la sensibilidad de la técnica aplicada a este estudio en particular. En este caso, el criterio de confianza estadístico definido para considerar a un gen como detectado fue el valor de 150 lecturas por gen como mínimo. En este contexto se construyó la figura 4.1 donde se grafica el número de genes en función de la cantidad de lecturas que mapean sobre ellos. De esta forma, por ejemplo, se puede apreciar que existen 2.660 genes que presentan 500 o más lecturas en ambas condiciones de trabajo. De esta forma, en función del criterio definido, se puede observar que existen más de 6.000 genes que presentan más de 150 lecturas en las dos condiciones (ver Figura 4.1). En estas condiciones se estarían detectando más de 6.000 genes presentes en la maquinaria traduccional en una condición de estudio particular en células eucariotas de humanos. Por esta razón, se puede decir que esta tecnología de *ribosome footprinting* es una técnica muy sensible.

Lo siguiente en estos análisis preliminares de los datos fue evaluar cual de las normalizaciones disponibles en el programa era la más adecuada para aplicar a los datos. Es preciso marcar que en este caso se optó por realizar dichas normalizaciones ya que no se contaba con la información apropiada acerca de la expresión y cambios a nivel del transcriptoma. De esta forma se buscó mediante las herramientas para normalizar datos disponibles en el programa, acercar los valores de expresión calculados en cada condición para que ambos set de datos tengan distribuciones

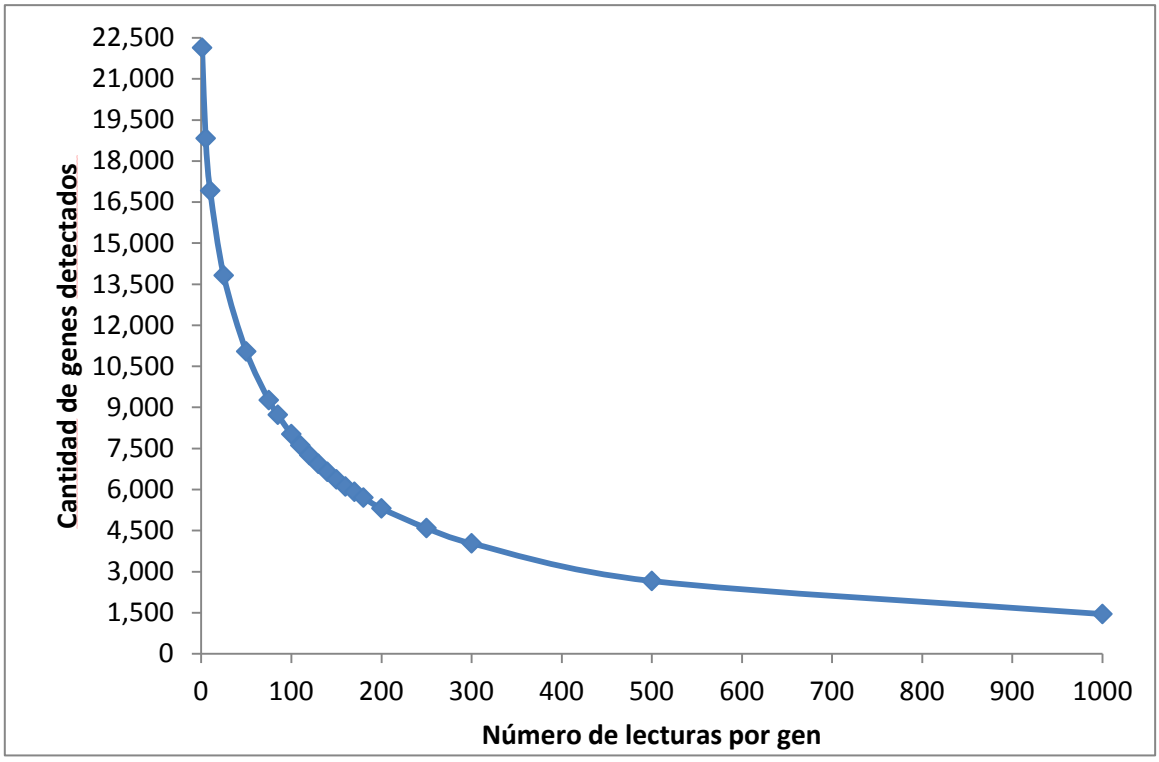


Figura 4.1: Sensibilidad de la técnica. El gráfico muestra que cantidad de genes son detectados en ambas condiciones experimentales, fijando como umbral distintos valores asignados en el eje horizontal. De esta forma definiendo un límite de 150 lecturas por gen, se estarían detectando más de 6.300 genes.

estadísticamente similares y así poder compararlos. Estos procesos complementan la normalización que ofrece trabajar con medidas de RPKM como valores de expresión, ya que dicho valor ofrece un grado de normalización al dividir la cantidad de lecturas mapeadas sobre un gen entre el largo exónico de dicho gen, y entre el total de lecturas mapeadas en todo el mapeo.

Las posibilidades que ofrece el programa a la hora de normalizar datos son la normalización por cuantiles (*Quantil normalization)* y la normalización por escalado (*Scaling normalization)*. De esta forma en la normalización por escalado, los valores de los set de datos se multiplican por una constante y se elige cual parámetro se desea mantener fijo, si la mediana o la media[164]. Si se elige la opción de normalización por cuantiles, las distribuciones empíricas de los valores de expresión de los distintos set de datos son utilizadas para calcular una distribución en común. Para esto se grafican los valores de los distintos cuantiles en diferentes ejes con el objetivo de generar una recta diagonal perfecta. En los casos en que ocurre una desviación en dicha diagonal

los valores de los cuantiles en dichos puntos son sustituidos por el promedio de los cuantiles[164], de esta forma se generan nuevos set de datos normalizados.

Una vez normalizados los datos, se intentó determinar mediante distintos enfoques, cuál de las dos normalizaciones aplicadas era la más adecuada, para continuar con los análisis posteriores utilizando los datos normalizados por el método seleccionado. Para esto, en una primera instancia se evaluó la expresión de distintos genes *housekeeping*, que en teoría no deberían diferir en sus niveles de expresión en las dos condiciones de trabajo: control y siPDCD4[165]. Los genes *housekeeping* elegidos para dicho análisis fueron: γ-actina (ACTG), β-tubulina (TUBB), gliceraldehído-3-fosfato deshidrogenasa (GAPDH), ARN-polimerasa II (POLR2A) y la fosfoglicerato quinasa 1 (PGK1). En la figura 4.2 se puede observar mediante un gráfico de barras las diferencias en los niveles de expresión traduccional de dichos genes, entre la condición control y siPDCD4, ya sea para los valores normalizados por escalado como por cuantiles. Las diferencias en los valores de expresión se presentan como porcentajes respecto del valor de expresión en la condición control. Se puede observar que en cuatro de los cinco genes mostrados, el método de normalización por cuantiles presenta menos diferencias en los valores de expresión.

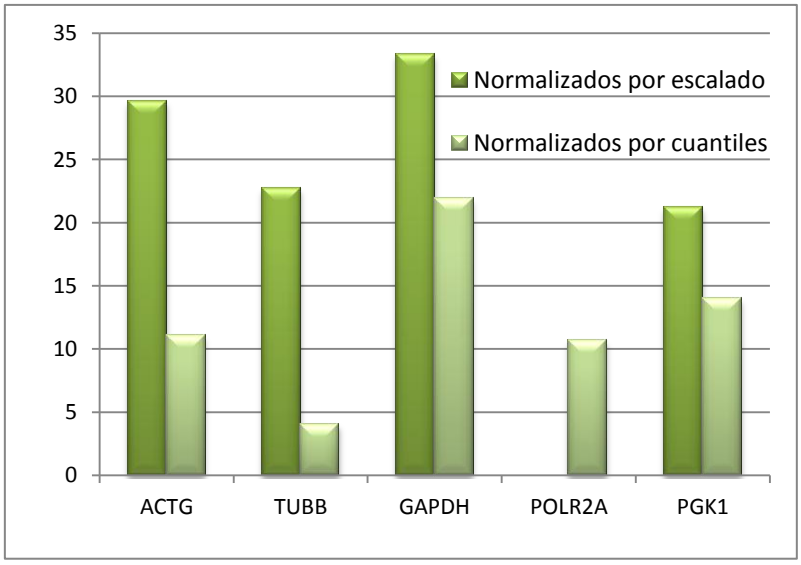


Figura 4.2: Diferencias en los niveles de expresión de ARNm entre las dos condiciones de estudio, para cinco genes housekeeping, normalizando los valores por los métodos de escalado y cuantiles. La diferencia en los valores de expresión es porcentualizada respecto al valor de expresión en la condición control y dicho porcentaje es el que se grafica.

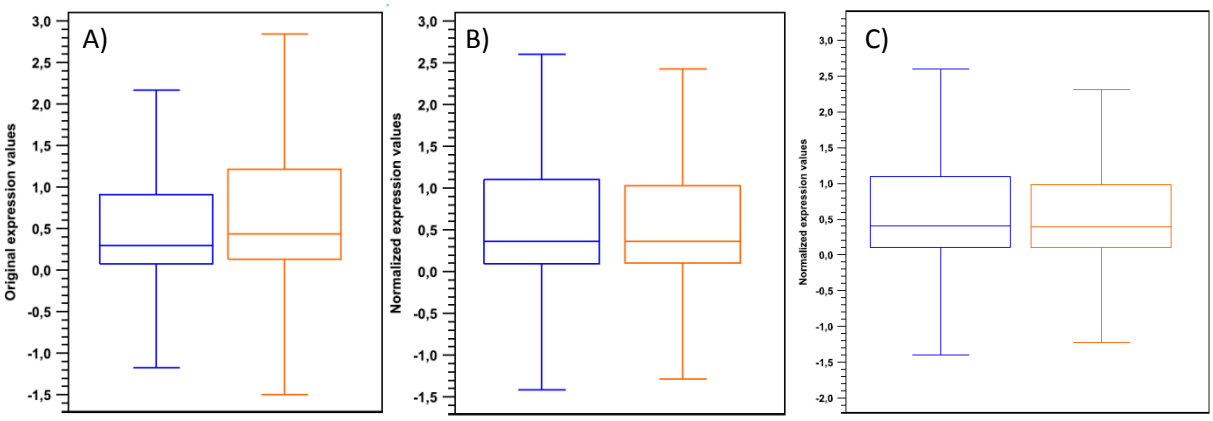


Figura 4.3: Se muestran los box plots construidos a partir de los datos crudos (A), normalizados por cuantiles (B) y normalizados por escalado (C). En cada caso se muestra la condición control (a la izquierda, en azul) y la condición siPDCD4 (a la derecha, en naranja).

Otra forma de evaluar cual de los dos métodos es el más adecuado para normalizar estos set de datos es realizar un control de calidad, estudiando la distribución general de los datos y su variabilidad mediante la construcción de *box plots.* De esta forma se construyeron los *box plots* generados a partir los datos crudos y normalizados por uno y otro método (ver Figura 4.3). Como puede observarse en la figura 4.3 mediante este enfoque no se logra distinguir bien las bondades de cada método de normalización, sino que más bien se indica como ambas estrategias normalizan justamente los datos y generan nuevos valores con distribuciones similares y por lo tanto comparables. Esto se observa, por ejemplo en el hecho de que la forma de las cajas para las dos condiciones, normalizadas por cualquiera de los dos métodos, presentan formas muy similares (ver Figura 4.3B y 4.3C). Por su parte, también se puede observar que el set de datos con los valores originales o crudos presenta diferencias en la distribución de sus datos entre las dos condiciones estudiadas, ya que sus diagramas de caja difieren en su forma (ver Figura 4.3A). Basados entonces en la efectiva normalización llevada a cabo por ambos métodos (ver Figura 4.3), y por sus efectos en los valores de expresión de distintos genes *housekeeping* (ver Figura 4.2), se optó por elegir el método de normalización por cuantiles como el más adecuado para los posteriores análisis a realizar.

Previo al análisis de la expresión diferencial y búsqueda de genes candidatos a ser regulados traduccionalmente por parte de *PDCD4,* se analizarán los patrones globales de mapeos generados en los dos alineamientos y análisis por RNA-Seq para ambas

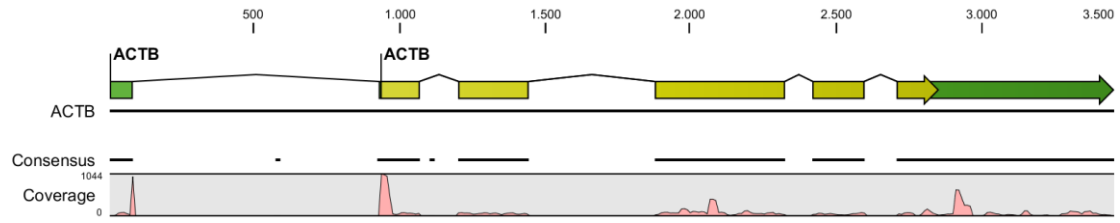


Figura 4.4: Representación del mapeo de huellas ribosomales sobre el mensajero de la β-actina. Se observa su estructura exónica, donde a la izquierda y derecha en verde se presentan el 5´-UTR y 3´-UTR respectivamente. En amarillo se observa la secuencia codificante. Puede apreciarse el fuerte pico resultado de la acumulación de ribsosomas a la altura del codón de iniciación, cerca del nucleótido número 1.000 indicado en la escala mostrada en la parte superior. Se destaca también la presencia de huellas en las regiones no traducidas.

condiciones de trabajo. En los mapeos realizados se observan ciertas características distinguibles que fueron reportadas antes en los distintos artículos publicados donde se aplicó la técnica de *ribosome profiling* en cuestión. Por ejemplo se observa una fuerte acumulación de huellas en el codón de iniciación lo cual se observa en un fuerte pico en el gráfico de mapeo definido para un mensajero (ver Figura 4.4). Este fuerte pico suele explicarse considerando que la droga utilizada para detener la elongación, no detiene los fenómenos de iniciación que continúan de forma normal[147]. Otra característica observada es la presencia de huellas ribosomales en las regiones no traducidas de los mensajeros, tanto el 5´-UTR como el 3´-UTR (ver Figura 4.4). Esta característica que sin lugar a dudas llama la atención, ya había sido reportada por Ingolia *et al. Science* 2009, y en este caso los autores explican que dichas huellas corresponden a ribosomas activos que se encargan de traducir pequeños marcos de lecturas abiertos que generan péptidos con funciones reguladoras diversas, o que simplemente representan ser extensiones amino-terminales de la proteína global. Respecto a la presencia de huellas ribosomales en la región 3`-UTR de los genes, esta puede explicarse considerando el modelo estructural planteado en la figura 1.4 del capítulo 1 donde se observa la conformación circular del mensajero al momento de la formación del complejo de iniciación 48S. En este modelo, los eventos de reciclado ribosomal implican que el ribosoma podría avanzar sobre el 3`-UTR hasta alcanzar la región de iniciación y de esta forma generar un nuevo evento traduccional.

También si se observan los gráficos de los mapeos con detenimientos pueden llegarse a distinguirse pequeños picos dentro de la región codificante del mensajero (ver Figura 4.4). Estos picos podrían corresponderse con pausas ribosomales, aunque se necesitan más estudios para poder afirmar esto con certeza. Por último también se puede notar que sobre el codón STOP no se aprecia ningún pico en el mapeo (ver Figura 4.4). Esto difiere con lo observado en experimentos donde no se utiliza ninguna droga para detener la elongación, donde si se aprecia un fuerte pico sobre la región del codón STOP[147]. Sin embargo dicha acumulación no se debe a que la elongación está activa, ya que en presencia de un análogo no hidrolizable de GTP, se observa el mismo pico sobre la región del codón STOP[147].

Por último también se compararon algunos patrones de mapeos para ambas condiciones en aquellos genes donde se observaba un cambio en la expresión traduccional (ver Figura 4.5). En estos casos puede observarse de forma sencilla como, al igual que en los mapeos representados con los datos de Ingolia *et al. Science* 2009, la forma global del mapeo o patrón de huellas es muy similar en las dos condiciones, solamente que el valor de cobertura máxima alcanzada cambia en gran forma, modificando así el valor de expresión génica evaluado en RPKM (ver Figura 4.5). En este caso se eligió el gen *PRIC285,* el cual codifica para un factor de transcripción que es un co-activador de los PPARs (*Peroxisome Proliferator Activated Receptors;* refPPAR). Este gen en particular tiene significancia estadística y está relacionado a procesos y funciones celulares involucradas en el cáncer (ver más adelante páginas 106 y 110).

2) ESTUDIOS DE EXPRESIÓN DIFERENCIAL DE ARNm: BÚDQUEDA DE GENES CANDIDATOS A SER REGULADOS TRADUCCIONALMENTE POR *PDCD4*

Hasta ahora se ha evaluado la sensibilidad de la metodología aplicada a este problema biológico en particular, se han analizado los porcentajes de mapeo, interpretando y justificando los valores de lecturas que mapearon contra el genoma, así como las lecturas derivadas de contaminación con ARN ribosomal. También se realizó un control

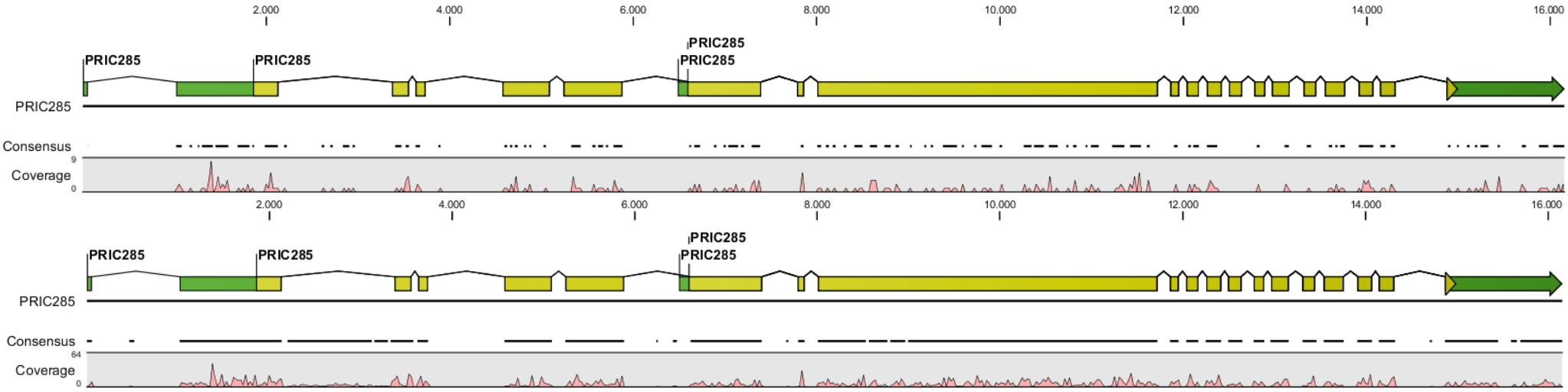


Figura 4.5: Se muestran los mapeos de huellas polisomales sobre el gen PRIC285 para la condición control (arriba) y la condición siPDCD4 (abajo). Puede observarse como el patrón de mapeo es el mismo en los dos casos, sólo que el valor máximo de cobertura alcanzado sube de 9 a 64 aumentando así el valor de expresión génica calculado considerablemente.

de calidad de los datos presentes en las dos condiciones de trabajo y se evaluó cual de los distintos algoritmos disponibles en el programa de análisis era el más adecuado para normalizar los datos y volverlos más comparables. Por su parte, también se analizaron patrones de mapeo globales, explicando cual era la interpretación respecto a lo que ocurría en cada caso. De esta forma, a continuación se procede al análisis de expresión diferencial en búsqueda de genes que varíen de forma significativa su tasa de traducción al silenciar el gen *PDCD4.* Dicho análisis de expresión diferencial compara gen a gen los valores de expresión, evaluados en RPKM, en las distintas condiciones de trabajo. De esta forma se calcula la tasa de cambio o *fold change* para cada gen. Dicho valor representa el cociente entre el valor de expresión medido en RPKM de la condición siPDCD4, entre el mismo en la condición normal. En los casos en que este cociente sea menor a 1, se presenta el valor inverso, pero con el signo opuesto, a los efectos de interpretar mejor los valores al visualizarlos.

Con respecto a los tres blancos traduccionales reportados para PDCD4, analizando nuestros resultados, no se puede confirmar de forma confiable que dicha regulación exista en nuestro sistema de estudio. En los tres casos, los genes presentaban una baja cobertura de lecturas lo cual genera bajos niveles de expresión en ambas condiciones. De todas formas, los valores de *fold change* calculados para los tres genes son mayores que uno, aunque estos casos están en el límite de la sensibilidad de la técnica por lo cual no se puede afirmar con certeza que se observa un aumento traduccional de los blancos reportados (ver Tabla 4.2). Aún así, estas discrepancias no son resultados

Tabla 4.2: Expresión diferencial de los blancos traduccionales de PDCD4 reportados. Se muestra el total de lecturas mapeadas contra los genes c-myb, casp3 y tp53 en ambas condiciones de trabajo, así como el valor de fold change calculado en los análisis.

	total de lecturas condición control	total de lecturas condición siPDCD4	*fold change*
c-myb	148	181	**1,33**
Casp3	99	119	**1,04**
tp53	182	227	**1,49**

desalentadores dado que las diferencias pueden ser entendidas por trabajar en distintos sistemas de estudio y distintos modelos celulares donde la expresión particular de cada gen varía de forma considerable.

A continuación se procede a aplicar un test estadístico a los efectos de poder analizar e interpretar los resultados con cierta certeza estadística. Aún así la posibilidad de aplicar herramientas estadísticas a este estudio se encuentra bastante limitado pues no se cuenta con réplicas. De todas formas, existe un test estadístico que si se puede aplicar en estas condiciones, se trata del test de Kal o test Z[166]. Este test se basa principalmente en la aproximación que puede hacerse de la distribución binomial a la distribución normal, teniendo en cuenta el gran número de lecturas secuenciadas y mapeadas[166]. A su vez al trabajar con proporciones, más que con un número bruto de lecturas mapeadas, este test es también aplicable a situaciones donde la suma de las lecturas mapeadas en las dos muestras es diferente entre las muestras, como ocurre en este caso. Al aplicar este test estadístico se adiciona información a los set de datos ya disponibles respecto a la diferencia en las proporciones ("*Proportions difference*") entre las dos condiciones, un valor de *fold change* acerca de cuantas veces más grande es la proporción en la condición siPDCD4 respecto de la condición control, y por último los valores estadísticos resultado de aplicar el test Z y el correspondiente p-valor del test a dos colas. De esta forma si se volviera a aplicar la misma técnica, se espera que aquellos genes que presenten un valor de p menor a 0,05 presenten el mismo patrón de cambio en su expresión. Resultado de aplicar dicho test se observaron solamente 46 genes con un valor de p igual o menor a 0,05.

Una alternativa más gráfica para visualizar los resultados generados por el anterior test, es utilizar un gráfico tipo *Volcano plot.* Este es un tipo especial de gráfico donde se muestra la relación entre el p-valor y la magnitud de la diferencia de expresión entre las muestras. Particularmente, en el *Volcano plot* se grafica en el eje vertical, el opuesto del logaritmo en base diez del p-valor, mientras que en el eje vertical se presenta el logaritmo en base dos del valor de *fold change* (ver Figura 4.6). De esta forma los puntos que representen genes con una alta significancia estadística serán ubicados hacia lo más alto del gráfico, mientras que aquellos genes que cambien en gran medida su expresión serán representados hacia la derecha (si aumentan su

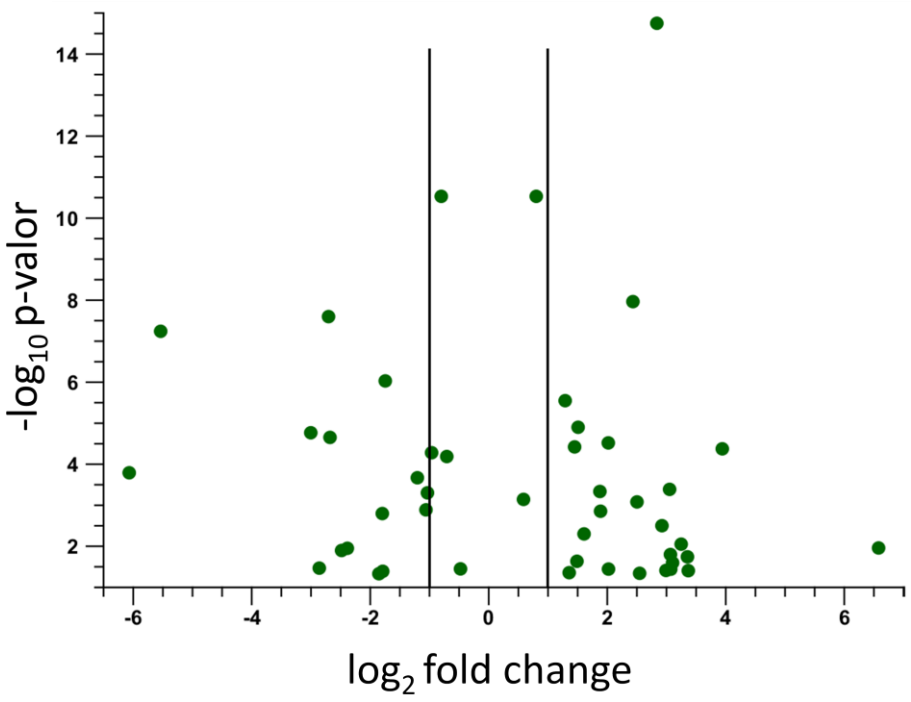


Figura 4.6: Volcano plot. Se muestra la relación entre el p-valor (representado como -log_{10} p-valor en el eje vertical) y la magnitud del fold change (representado como log_2 fold change en el eje horizontal). Se muestran solo los genes con un valor de p menor a 0,05. Las rectas verticales en x=1 y x=-1 indican los valores de fold change de 2 y -2 respectivamente, mencionados en el texto.

expresión) o hacia la izquierda del gráfico (si disminuyen). En la figura 4.6 se presenta el *Volcano plot* para el set de datos comparando las condiciones siPDCD4 y control. Se presentan solamente los genes con valores de p menores a 0,05 y las líneas verticales indican los valores de *fold change* correspondientes a 2 y -2. Pueden identificarse de esta forma más de 40 genes con valores de p menores a 0,05 (significancia estadística) y valores de *fold change* mayores a 2, o menores a -2. Junto con esta figura, se presenta la tabla 4.3 donde se especifica la identificación característica de aquellos genes (ID) donde se observa un aumento en la expresión. En la tabla 4.3 también se presenta el valor de *fold change* y el p-valor asociado a cada uno de estos genes. Aquí aparecen varios casos que parecen ser artefactos (*ACSBG1, SAP30, RTP2, MNDA, TXNIP*). Dichos casos, en un principio son detectados por los excesivos valores de *fold change* que se observan, como ocurren en el caso de *ACSBG1.* También son detectados considerando que en las dos condiciones de trabajo los mapeos realizados son iguales, solo que aumenta únicamente y de forma exagerada el mapeo sobre una región discreta del mensajero, aumentando así la cobertura máxima alcanzada y el valor de

Tabla 4.3: Se describen los genes mostrados en la figura 4.6, solamente los que presentan un fold change mayor a 2. Se especifica el nombre (ID) de cada gen así como su valor de fold change y su p-valor

ID	*Fold Change*	p-valor
ACSBG1	95,53	0,01
SAP30	10,31	0,04
OAS2	10,22	0,02
RTP2	9,5	8,90E-03
PRIC285	8,58	0,03
IFIT1	8,41	0,04
ISG15	8,38	0,02
MX1	8,29	4,09E-04
UBE2L6	7,95	0,04
MNDA	7,59	3,15E-03
FAM69C	7,14	1,78E-15
KCNA1	5,84	0,05
IFI6	5,66	8,29E-04
MLXIP	5,4	1,09E-08
TXNIP	4,06	0,04
IRF9	4,05	3,01E-05
SLC33A1	3,7	1,40E-03
OAS3	3,67	4,61E-04
DZIP3	3,05	5,01E-03
GPR77	2,84	1,25E-05
ATF2	2,8	0,02
RRH	2,73	3,79E-05
C15orf43	2,56	0,04
DDR1	2,45	2,82E-06

expresión de RPKM calculado. Esto puede ser causado por la presencia de algún repetido o por mapeos inespecíficos que generan estas diferencias de expresión que no son verídicas. En lista mostrada en la tabla 4.3 también pueden observarse varios genes relacionados a la respuesta inmune celular, los cuales al silenciar *PDCD4* aumentan a nivel traduccional. Entre ellos se destacan *ISG15, OAS2*, *OAS3*, *IFIT1*, *MX1*, *IFI6* e *IRF9*. Una posible explicación para entender la presencia de dicho grupo de genes involucra la vía de mTOR, en la cual participa PDCD4, y la activación de la respuesta celular antiviral vía interferón tipo I, causada por la activación traduccional del mensajero de *IRF7*. En este contexto ya se han involucrado mecanismos que se encuentran *downstream* en la vía de mTOR como responsables directos de la activación de la respuesta vía IFN mediante la activación traduccional del factor de transcripción *IRF7*[167] (ver Figura 4.7). Esos mecanismos particularmente son, como se muestra en la figura 4.7, la fosforilación de las 4E-BPs y de la quinasa S6K. La fosforilación de dichos blancos permite a su vez tanto la fosforilación de IRF7 mediante un complejo formado a partir del reconocimiento del virus vía TLR9, así como la activación de su traducción gracias a la presencia de complejos eIF4F activos. Por otro lado, se ha visto que el mensajero de *IRF7* presenta un muy estructurado 5´-UTR evolutivamente conservado[168]. También se ha demostrado que dicha estructura impide su correcta traducción[168]. Por esto, en la hipótesis planteada, se supone que al silenciar *PDCD4* vía siARN, se permite que se formen más complejos eIF4F activos capaces de

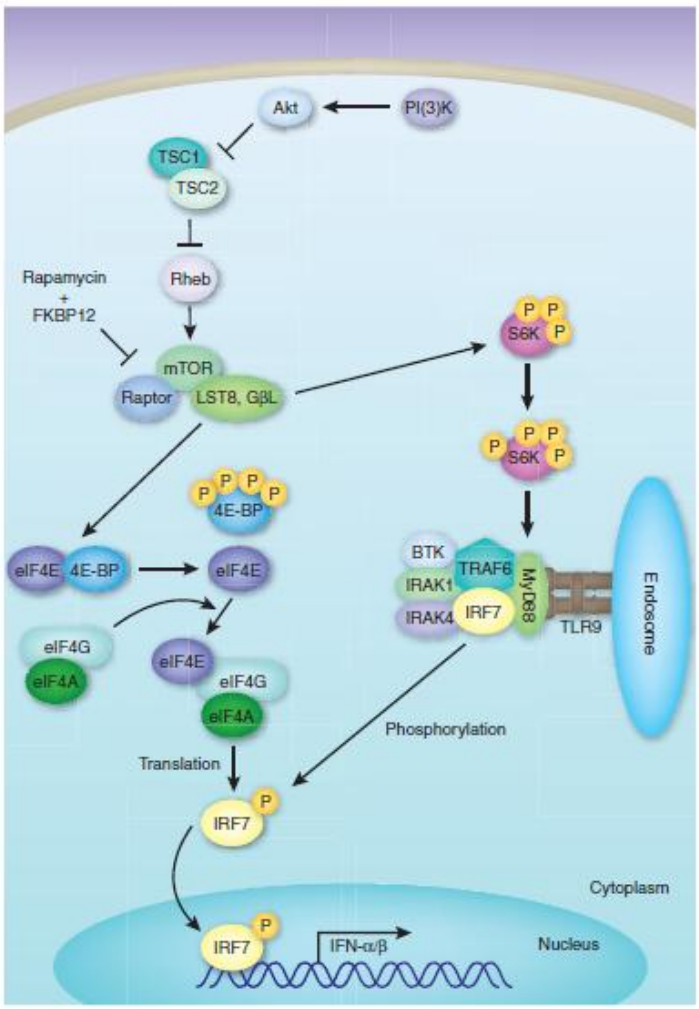


Figura 4.7: Esquema de la activación transcripcional de la respuesta vía interferón. Esta es disparada por la activación traduccional de IRF7 como se muestra. Figura adaptada de Costa-Mattioli & Sonenberg 2008.

desenvolver la estructura secundaria presente en el mensajero de *IRF7*, ya que eIF4A no estaría compitiendo con PDCD4. De esta forma, al activarse la traducción del mensajero de *IRF7,* se desencadena la transcripción de una batería de genes relacionados con la respuesta celular antiviral vía interferón tipo I a cargo de *IRF7*[169]. Esta hipótesis no descarta que el intermediario para la activación de la respuesta antiviral vía interferón, no sea otro factor de transcripción que también regule la expresión de los genes activados en dicha respuesta. Este podría ser el caso, solo hipotéticamente, de *IRF9* el cual aparece con un *fold change* significativo estadísticamente en la lista de la tabla 4.3. En el marco de esta hipótesis, el aumento a nivel traduccional de genes relacionados a la respuesta inmune en condiciones de

siPDCD4, se explica dada la activación traducción de uno o más genes que regulan la respuesta tipo interferón.

Existen también genes en la lista presentada en la tabla 4.3 relacionados de alguna forma con los procesos tumorales y el cáncer. Respecto a esto, a continuación se citan, modo de ejemplo, los genes *UBE2L6* y *ATF2*. El primero codifica para ubiquitin-ligasa E2 y ha sido relacionado directamente en modelos de cáncer de mama[170]. En este modelo la ubiquitinación forma parte junto con los procesos de internalización y degradación de receptores activados, de la regulación de la señalización por factores de crecimiento. En este caso, una actividad aberrante en cualquiera de dichos procesos puede estar implicada en el cáncer[171]. Se ha visto también que UBE2L6 es la enzima E2 para una proteína tipo ubiquitina denominada ISG15 que también aparece en la lista de la tabla 4.3[172]. El segundo de los ejemplos mencionados respecta al gen *ATF2* el cual codifica para un factor de transcripción al cual se le han descrito actividades como oncogen y como gen supresor de tumores[173].

Un camino más directo para analizar las funciones celulares y estudiar la ontología de extensas listas de genes, es el uso del programa comercial *Ingenuity Systems ©* presentado en el capítulo 2. Esta herramienta de análisis agrupa los genes según las funciones que se han descrito para cada uno de ellos y nos informa acerca de cuáles procesos celulares globales están cubiertos por los genes de nuestras listas. De esta forma uno puede analizar las principales funciones celulares y vías en las cuales intervienen los genes que se están estudiando. A la vez el software genera redes biológicas donde uno también puede observar que relaciones existen y de qué tipo son, entre los genes en estudio y también con otros genes complementarios que forman parte de la red que se está estudiando.

En este caso la lista con genes donde se observaba un aumento a nivel traduccional resultado de silenciar *PDCD4* presentaba un total de 452 genes seleccionados en función de los valores de *fold change* en la expresión génica y el total de lecturas que presentaban (ver Tabla 4.4 al final del capítulo). Para generar esta lista no se tuvo en cuenta el valor de p derivado de la aplicación del test de Kal, ya que se priorizó la significancia biológica entendida en valores de *fold change,* antes que la significancia

estadística evaluada en valores de p. En el marco de la hipótesis de trabajo presentada, dichos genes representan los candidatos a ser regulados de alguna forma por *PDCD4.* Esto se debe a que al silenciar PDCD4, un supresor de la traducción *cap*-dependiente, se observa un aumento en la expresión a nivel traduccional de dichos genes presentados en la tabla 4.4. De esta forma los análisis realizados en el *Ingenuity Systems ©* generaron la información y los resultados mostrados en la tabla 4.5 y en las figuras 4.8 a 4.11. En la tabla 4.5 se observan en primer lugar las principales funciones asociadas a las redes biológicas detectadas junto con un score asignado a cada función. Dicho score es calculado en función de la cantidad de genes presentes en cada red biológica, así como también en función de los valores de *fold change* de dichos genes, los cuales también son incorporados en el análisis de ontología, mediante algoritmos patentados. Aquí se identifican funciones relacionadas de forma clara con los procesos tumorales en lo que respecta, por ejemplo, a funciones como señalización celular, reparación del ADN, expresión génica, muerte celular, desarrollo celular, etc. Todas estas funciones eran de esperarse que fueran detectadas en los análisis dados los procesos con los cuales se ha identificado a PDCD4[84]. En segundo lugar dentro de la tabla 4.5, se observan las principales enfermedades y desordenes asociados a los genes incluidos en el análisis. En este caso, el resultado es un poco más difícil de interpretar dada la gran complejidad que abarca cada tipo de desorden y enfermedad representada. Se observa como los desordenes genéticos son el casillero que presenta la mayor cantidad de moléculas asociadas.

Más resultados interesantes se presentan en la tercer parte de la tabla 4.5. Aquí se observan las principales funciones celulares y moleculares derivadas del estudio, junto con el número de moléculas asociadas a cada función y un p-valor que refleja la probabilidad de que dichas funciones no hayan sido representadas por azar, si se introdujera en el análisis una lista aleatoria de genes (también por algoritmos patentados). En este caso las funciones presentadas parecen estar claramente relacionadas a las funciones descritas de PDCD4. Considerando la hipótesis de trabajo planteada en el capítulo 2, dado que PDCD4 es un gen supresor de tumores que inhibe la traducción de ciertos mensajeros de forma específica, es de esperarse que al silenciar la expresión de este gen, se observe la activación de un conjunto de genes

Tabla 4.5: Se muestran los principales resultados obtenidos del estudio de ontología realizado a partir de lista de genes de la tabla 4.4.

FUNCIONES ASOCIADAS A LAS REDES BIOLÓGICAS		Score
1	Desordenes Gastrointestinales, Movimiento y Organización Celular	44
2	Señalización Celular, Replicación, Recombinación y Reparación del ADN, y Metabolismo de AN	36
3	Expresión Génica, Muerte y Desarrollo Celular	35
4	Muerte y Desarrollo Celular y Respuesta Antimicrobiana	32
5	Expresión Génica, Metabolismo de Aminoácidos y Modificaciones Post-traduccionales	30

ENFERMEDADES Y DESORDENES	p-valor	# moléculas
Enfermedades Gastrointestinales	3,48E-02	164
Desordenes en el Sistema Endócrino	1,76E-02	130
Enfermedades Metabólicas	3,48E-02	137
Desordenes Genéticos	3,48E-02	220
Enfermedades Inflamatorias	3,48E-02	127
FUNCIONES CELULARES Y MOLECULARES	p-valor	# moléculas
Señalización e Interacción Celular	1,32E-05	43
Morfología Celular	1,33E-04	33
Organización y Estructuración Celular	9,10E-04	37
Compromiso Celular	9,10E-04	16
Desarrollo Celular	9,10E-04	27
PRINCIPALES VÍAS CELULARES	p-valor	# moléculas
Señalización por Interferón	2,95E-04	5/36
Activación de IRF por Receptores Citosólicos	8,15E-04	6/72
Función de JAK1, JAK2 y TYK2 en la Respuesta vía IFN	8,24E-03	3/27
Función de la Familia de Quinasas JAK en la Respuesta vía Citoquina IL-6	9,25E-03	3/27
Señalización por Receptor de Glutamato	1,77E-02	4/69

relacionados a procesos tumorales que activan funciones celulares que luego se desarrollan de forma aberrante en los tumores. En este caso, todas las funciones presentadas en esta tercer parte de la tabla 4.5 se asocian a procesos tumorales en el sentido de que se desarrollan de forma alterada en dichos procesos. Por ejemplo una

interpretación posible a las funciones que allí se presentan es considerar que en un proceso tumoral se pierden las interacciones célula-célula y con la matriz extracelular, cambia la morfología de las células, éstas se desarrollan de forma exagerada y desorganizada donde no existe un compromiso celular a activar la apoptosis por más que estén dadas las condiciones para su activación.

Por último, en cuarto lugar de la tabla, se presentan las principales vías celulares representadas en el estudio. En este caso existe una considerable presencia de las vías relacionadas a la respuesta celular inmune y al desarrollo de una respuesta vía interferón. Las posibles razones que explican la activación de dicha respuesta ya fueron presentadas antes (ver página 106).

Por otro lado, como resultados del análisis y estudio de ontología de los genes en los cuales se observaba una respuesta positiva traduccional, se presentan también una serie de figuras donde se muestran las redes biológicas generadas en dicho estudio. En este caso, el programa utilizado agrupa las moléculas codificadas por los genes en estudio, según su clasificación dentro de los grupos definidos en la primera parte de la tabla 4.5 y construye redes biológicas considerando las variadas interacciones reportadas en la literatura. De esta forma, se observa por ejemplo en la figura 4.8 la red biológica definida para las funciones de señalización celular; replicación, recombinación y reparación del ADN y metabolismo de los ácidos nucleicos. Mediante símbolos se define la identidad de cada molécula (ver Figura 4.12) y en escalas de colores se codifica el valor de *fold change* de cada gen, de manera que cuanto más oscuro es el color, mayor es el *fold change.* En este caso en particular se destaca en azul la presencia del gen *RET.* Este gen fue inicialmente reportado como una proteína oncogénica encontrada en ensayos de transformación celular[174] y actualmente constituye un oncogen involucrado en varios estudios relacionados con el cáncer[174,175]. La presencia de este oncogen en esta red biológica en particular, con un valor de *fold change* de 2,050 apunta a que se trata de un posible blanco traduccional de PDCD4, dado que se trata de un gen que activo produce un desarrollo tumoral, cuya expresión a nivel traduccional en condiciones donde PDCD4 está ausente aumentan respecto a condiciones normales.

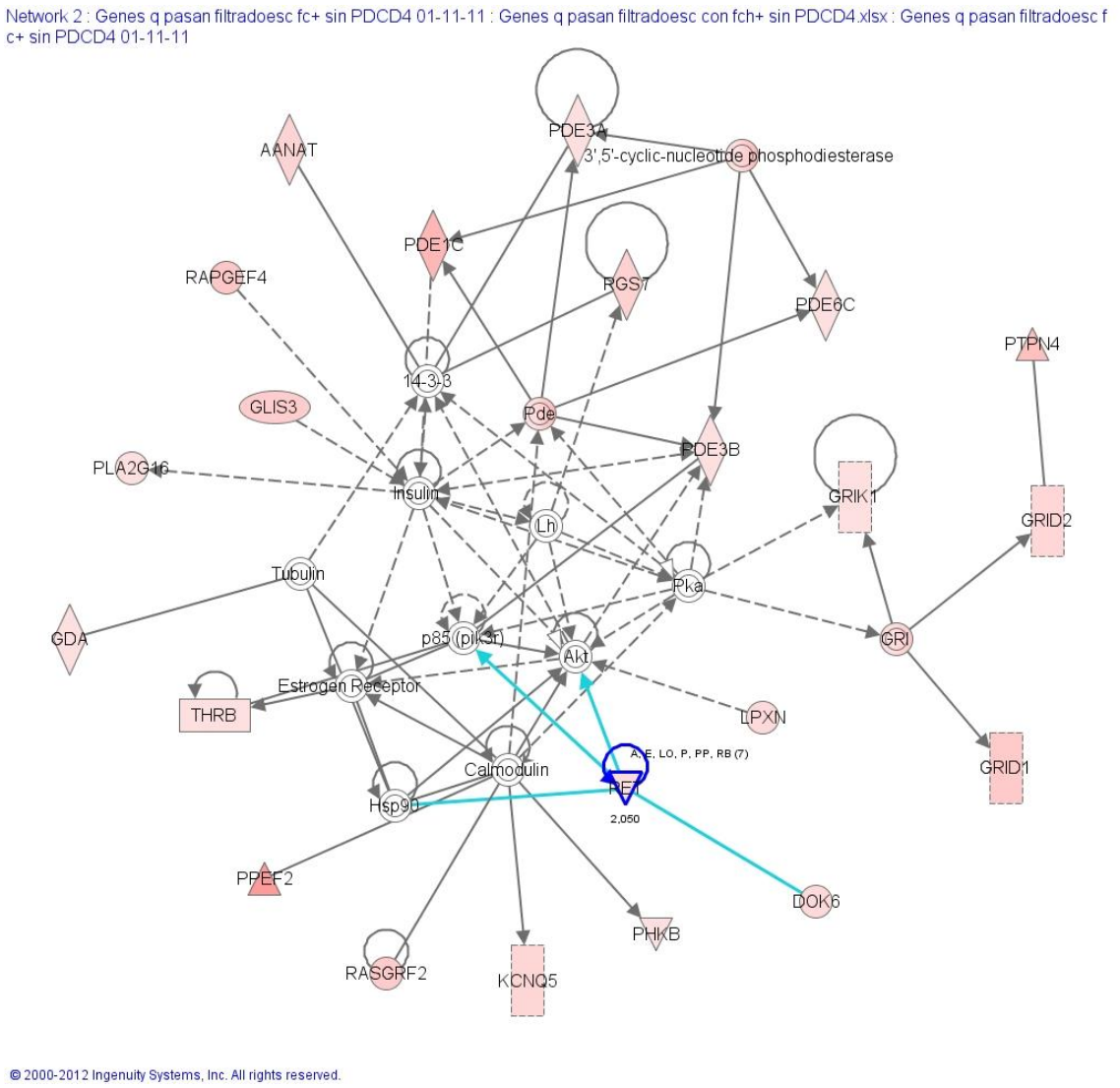


Figura 4.8: Red biológica asignada a las funciones de Señalización celular, Replicación, Recombinación y Reparación del ADN, y Metabolismo de Ácidos Nucleicos. Se resalta en azul el oncogen RET (ver texto). Por código de símbolos y flechas ver Figura 4.12

En la figura 4.9 se muestra la red biológica construida a partir de los genes asignados al tercer grupo de funciones definidas en la primera parte de la tabla 4.5, se trata entonces de los genes cuyas funciones se asignan al control de la expresión génica, desarrollo y muerte celular. En este caso se presenta la red biológica indicando la localización subcelular de las distintas moléculas. La red está organizada en compartimentos donde se distinguen el nuclear, citoplasmático, otro asignado a la membrana plasmática, uno correspondiente al espacio extracelular y otro bajo el nombre de “desconocido” donde se ubican las moléculas cuya localización no se conoce con certeza.

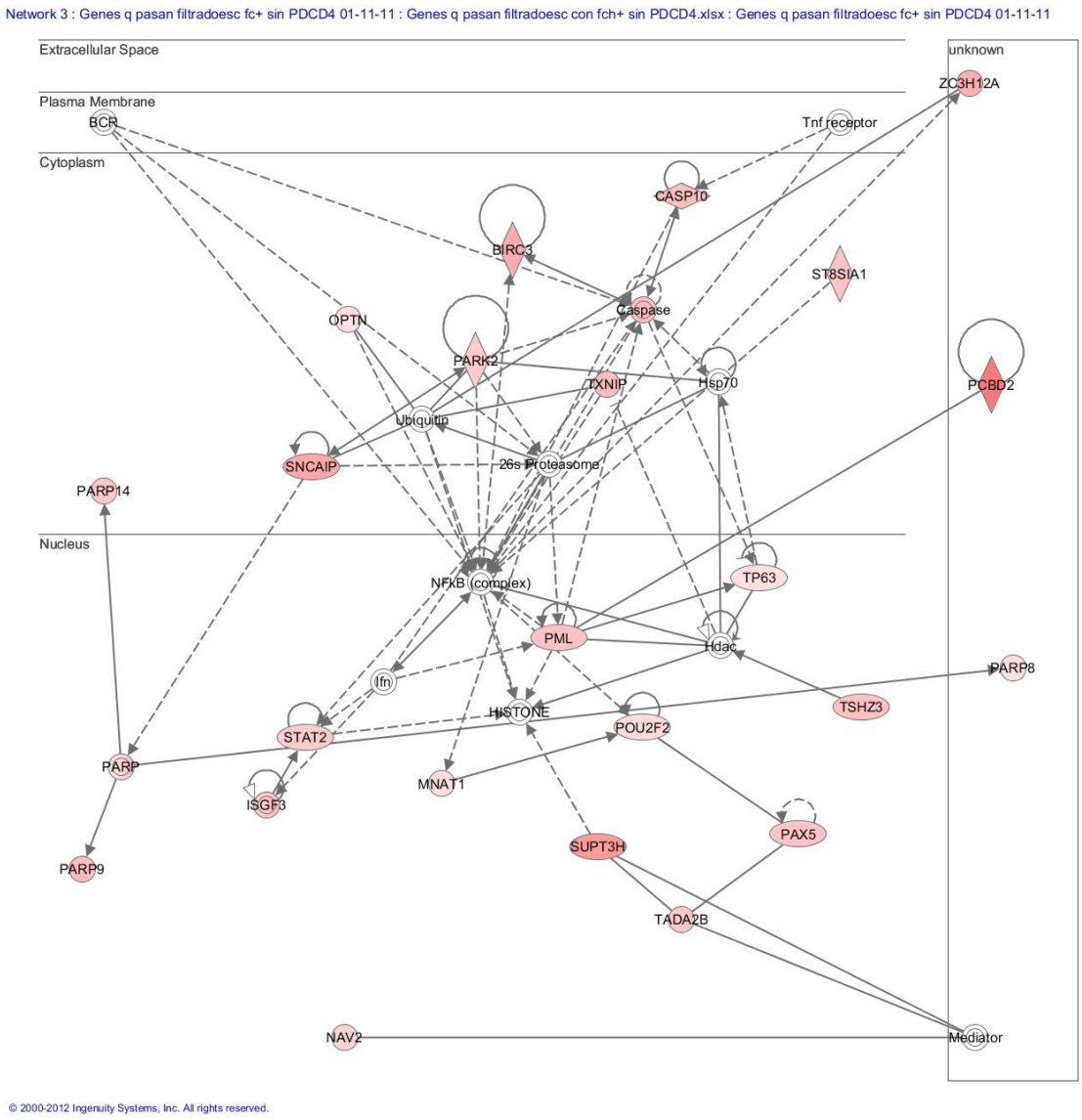


Figura 4.9: Red biológica asignada a las funciones de Expresión génica, Muerte y Desarrollo Celular. Se distinguen los distintos compartimientos subcelulares. Por código de símbolos y flechas ver Figura 4.12

Respecto a la figura 4.10, en esta se presenta la red biológica donde se relacionan los genes cuyas funciones se agrupan dentro de los procesos de muerte y desarrollo celular y respuesta antimicrobiana. En este caso, al igual que en la figura 4.8, se resalta la presencia del gen *MLLT10.* Este gen, que codifica para un factor de transcripción, ha sido asociado a procesos de arreglos cromosómicos y fusiones características de leucemias[176]. De esta forma, con un valor de *fold change* de 1,270 se presenta otro gen con una relación directa con el cáncer en la lista de genes candidatos.

Más allá del caso anterior cuya validez es discutible dado su bajo valor de *fold change*, en la figura 4.11, donde se presenta la red biológica asociada a los procesos de

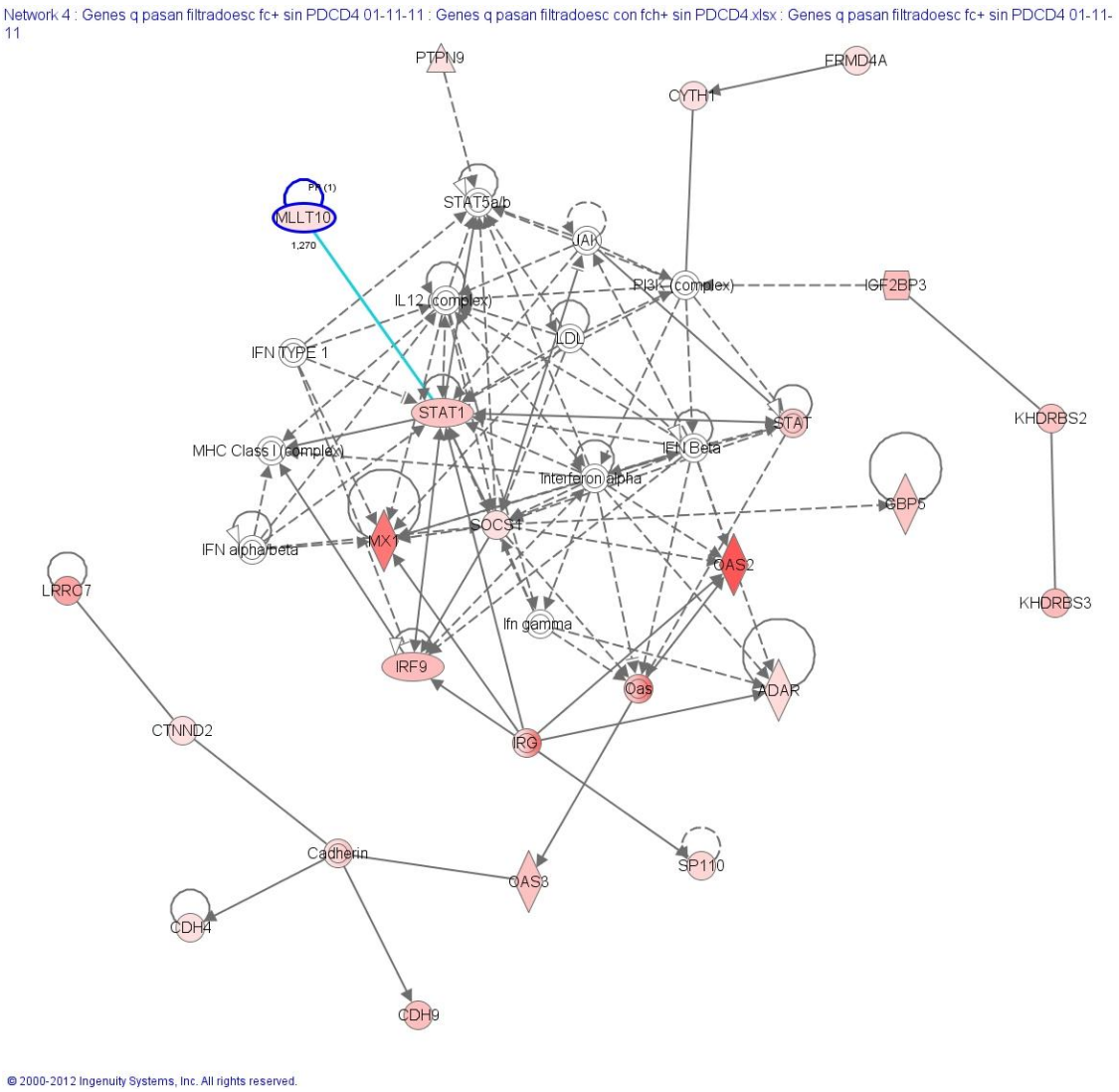


Figura 4.10: Red biológica asignada a las funciones de Muerte y Desarrollo Celular y Respuesta Antimicrobiana. Se resalta en azul el gen MLLT10 asociado a leucemias (ver texto). Por código de símbolos y flechas ver Figura 4.12

expresión génica, metabolismo aminoacídico y modificaciones post-traduccionales, aparece otro gen que llama la atención. Se trata de *WNT3A* que aparece marcado también en azul. Con un valor de *fold change* de 4,64 se trata de un gen implicado de forma directa en el desarrollo de tumores. La familia de genes *WNT* consiste en un grupo de genes relacionados estructuralmente que codifican para proteínas de señalización[177]. Estas proteínas han sido implicadas en el desarrollo de tumores y en el control del destino de células madre durante la embriogénesis[178,179], lo cual junto con el alto valor detectado de *fold change,* destaca la presencia de dicho gen como candidato de posible blanco traduccional de PDCD4.

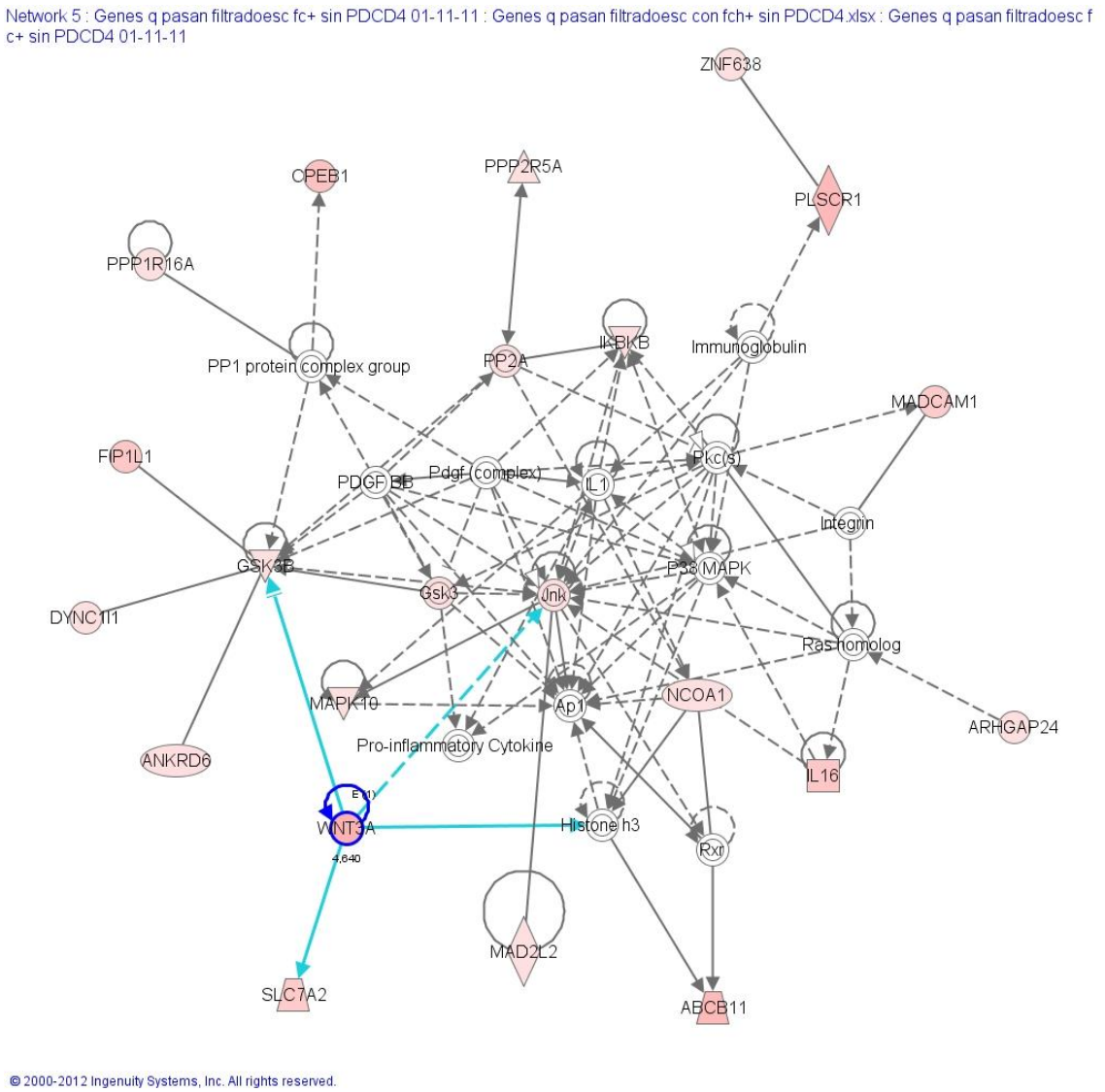


Figura 4.11: Red biológica asignada a las funciones de Expresión Génica, Metabolismo de Aminoácidos y Modificaciones Post-traduccionales. Se resalta en azul el gen WNT3A implicado en el desarrollo de tumores (ver texto). Por código de símbolos y flechas ver Figura 4.12

A modo de conclusión se puede decir entonces que las funciones descritas anteriormente para PDCD4 respecto a su rol como gen supresor de tumores y el control que ejerce en la muerte celular programada, han sido también encontradas en los estudios de ontología realizados. Resultado de silenciar dicho gen mediante ensayos de siARN en un modelo celular de cáncer de mama, se extrajo una amplia lista de genes en los cuales se observaba un aumento en su expresión traduccional mediante la aplicación de la técnica *ribosome footprintng*. Aceptando la hipótesis de trabajo planteada, se supone que el hecho de silenciar la expresión de este gen supresor de tumores, genera un aumento en la expresión traduccional de un conjunto

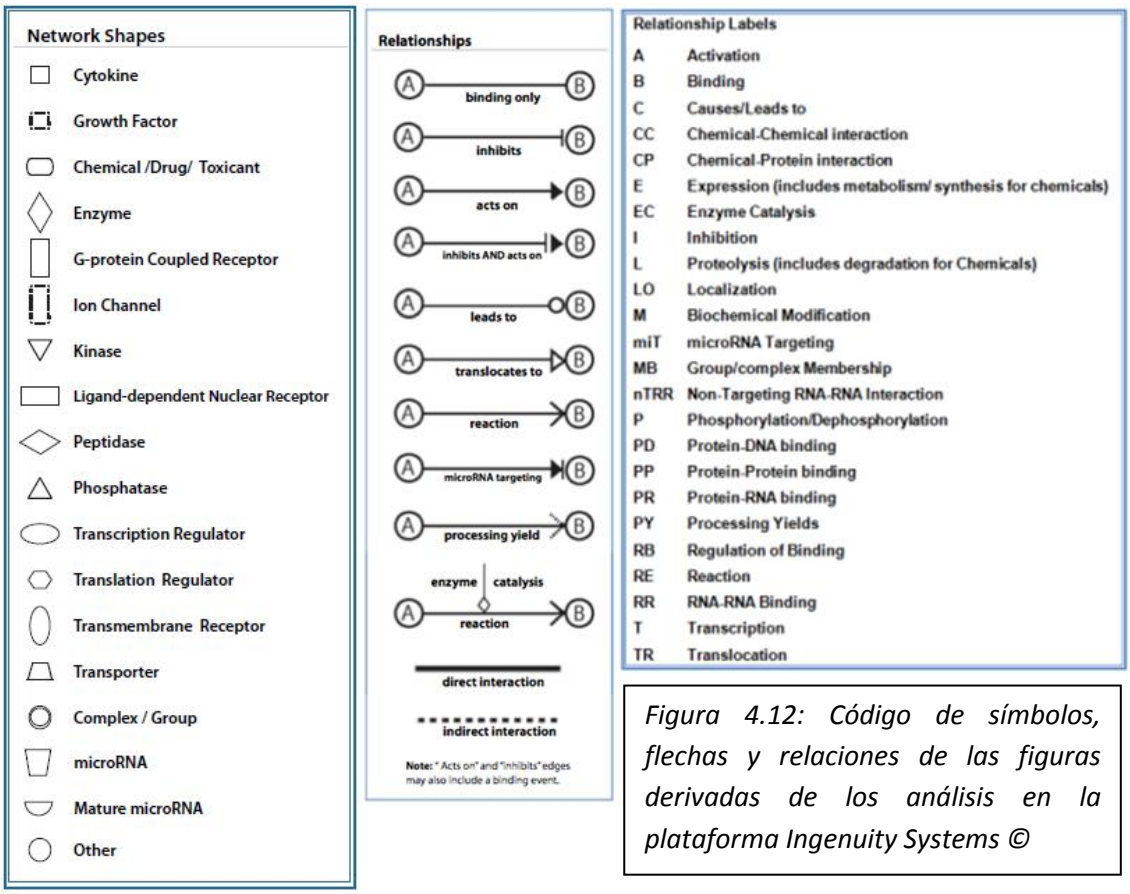


Figura 4.12: Código de símbolos, flechas y relaciones de las figuras derivadas de los análisis en la plataforma Ingenuity Systems ©

particular de mensajeros, dado que en condiciones normales PDCD4 inhibe la traducción de éstos. De esta forma es de esperar que al silenciar la expresión de PDCD4, parte de esos genes en los cuales se observa un aumento en la traducción, estén relacionados con el desarrollo de tumores. En particular se destacaron los casos puntuales de tres de estos genes presentes en los análisis de ontología efectuados (*RET, WNT3A* y *MLLT10*). Sin duda que son muchos más los genes involucrados que pueden ser analizados en forma detallada, de todas formas quedan planteados desafíos por delante acerca de la confirmación mediante experimentos puntuales de estos casos señalados. Estos experimentos, como los realizados en los tres casos donde se notificó la regulación traduccional por parte de PDCD4 sobre mensajeros puntuales, permitirían demostrar que efectivamente existe una asociación directa entre PDCD4 y los mensajeros de los genes aquí nombrados. También se podría confirmar el hecho de que es la ausencia puntual de PDCD4 lo que dispara su traducción y no otros mecanismos que puedan estar afectando la respuesta observada.

Tabla 4.4: Lista extendida de genes candidatos a ser regulados traduccionalmente por PDCD4. Estos genes fueron elegidos según su valor fold change y la cantidad de lecturas que presentaban. Con esta lista de genes se realizaron los estudios de ontología. Nota: los valores de infinito se generan pues la expresión en la condición control es cero y al dividir entre cero, el cociente vale infinito. Representan entonces, aquellos genes que en condiciones normales no se expresan pero en ausencia de PDCD4 sí.

ID	Fold Change	ID	Fold Change	ID	Fold Change
HPD	∞	ANKDD1B	∞	ITGA8	∞
MAL	∞	SH3RF3	∞	FREM3	∞
MS4A5	∞	ANKDD1A	∞	FAM19A4	∞
TMEM61	∞	PKDCC	∞	FLJ37396	∞
HTN3	∞	GUCA1A	∞	LOC100129848	∞
FAM70B	∞	ADAMTS3	∞	C8A	∞
SCEL	∞	IQCJ	∞	ZFP2	∞
GJA1	∞	BSPH1	∞	HNF4A	∞
CD200	∞	SPATA21	∞	KCNJ6	∞
HIGD1C	∞	PYHIN1	∞	ANKRD55	∞
OOSP1	∞	GAP43	∞	SLCO6A1	∞
PLSCR2	∞	FSIP1	∞	CCDC33	∞
SMARCD3	∞	ZNF536	∞	HMGCLL1	∞
ZCWPW2	∞	CDH15	∞	CCDC160	∞
LOC100127983	∞	LOC100129654	∞	PAMR1	∞
TSPAN8	∞	MARCH11	∞	KCNJ3	∞
ST6GALNAC5	∞	CDK15	∞	CCDC102B	∞
GABRR2	∞	NTNG2	∞	LOC647166	∞
C4orf45	∞	SYT16	∞	SCGN	∞
SGCG	∞	ENPP3	∞	CRYM	∞
SLC2A9	∞	CDO1	∞	NR1I2	∞
C10orf107	∞	ZBBX	∞	ADAM23	∞
C4orf51	∞	XKR3	∞	MOBP	∞
TMEM150C	∞	EDAR	∞	LOC100132891	∞
HTR2C	∞	DPT	∞	NKAIN3	∞
F10	∞	CHRND	∞	WIPF1	∞
LOC100131818	∞	PLA1A	∞	HPGDS	∞
NAALADL1	∞	C4orf22	∞	ARHGAP25	∞
CYLC2	∞	NPSR1	∞	TRPM6	∞
AMPH	∞	CCDC3	∞	SLC5A5	∞
C1orf228	∞	ACRBP	∞	VWC2	∞
LOC643339	∞	EDIL3	∞	FAM154A	∞
TACR3	∞	PRR4	∞	GABRG3	∞
TNFSF13B	∞	ELMO1	∞	C12orf26	∞
HCRTR2	∞	LOC728637	∞	DMRT1	∞

ID	Fold Change	ID	Fold Change	ID	Fold Change
FRMD5	∞	APBA2	5,36	ZNF781	3,83
EMR3	∞	UNC5A	5,29	COL11A1	3,81
MEOX2	∞	SNCAIP	5,19	RTN1	3,81
TMPRSS11A	∞	FGF12	5,19	FAM118A	3,80
CES5A	∞	IGLON5	5,11	MGAT5B	3,78
UPP2	∞	BIRC3	5,06	ENTPD1	3,75
CLIC2	∞	C7orf46	4,97	HRH1	3,73
MUSK	∞	TNR	4,93	SLC33A1	3,70
GPA33	∞	PTPRT	4,91	VPS54	3,70
ARMC3	∞	OTUD7A	4,91	NLRP9	3,68
CHRM2	∞	STXBP6	4,80	OAS3	3,67
CDH18	∞	CCDC108	4,79	CPEB1	3,67
NEXN	∞	COL22A1	4,76	ATG4A	3,65
PPP2R2B	∞	ANKS3	4,74	KCNT1	3,65
LAMP3	19,30	ZC3H12A	4,74	PML	3,64
CPA6	12,44	SYN2	4,65	TSHZ3	3,63
DDX60L	11,76	WNT3A	4,64	COL14A1	3,62
TRANK1	10,64	CIB2	4,58	C1QTNF7	3,61
OAS2	10,22	NTN4	4,49	ST8SIA1	3,59
LHFP	9,53	C12orf42	4,38	STAT1	3,58
IKZF3	9,32	TMEM90B	4,38	PAX5	3,58
PRIC285	8,58	HSD17B7	4,37	KCNK12	3,57
MX1	8,29	PDE1C	4,37	SMOC2	3,51
PCBD2	8,07	HPSE2	4,36	TSPAN5	3,51
TMEM108	7,92	FAM134B	4,21	TMEM231	3,50
KCNAB1	7,26	TMEM154	4,20	DCDC2C	3,50
FAM69C	7,14	ABCB11	4,17	TSGA10	3,50
CELF5	6,34	TSNARE1	4,14	USP49	3,46
GALNT5	6,34	PLSCR1	4,11	SEMA3A	3,46
PRDM11	6,14	SERAC1	4,10	S1PR2	3,45
ADAD1	6,14	KHDRBS3	4,09	GBP5	3,45
PPEF2	6,09	TXNIP	4,06	AASS	3,43
SUPT3H	6,07	WDR25	4,06	GMEB1	3,42
CYP39A1	5,90	IRF9	4,05	ASTN1	3,42
CCDC81	5,89	C7orf60	4,04	KHDRBS2	3,42
RUNDC2B	5,86	PEX7	4,00	LAMA1	3,42
KCNA1	5,84	PARP9	3,98	RAPGEF4	3,39
GLDC	5,58	PHEX	3,93	SLC47A2	3,39
TMEM163	5,52	IGF2BP3	3,92	ULK4	3,35
SLC6A11	5,51	CASP10	3,88	MAGEC3	3,34
LRRC7	5,44	ZNFX1	3,85	TADA2B	3,33
MLXIP	5,40	CDH9	3,85	FIP1L1	3,31
RGSL1	5,39	PTPN4	3,84	MUC7	3,31

ID	Fold Change	ID	Fold Change	ID	Fold Change
IL16	3,31	KCNQ5	2,50	C2orf39	2,04
MPPED1	3,29	SP110	2,48	CNTLN	2,03
ASZ1	3,28	DDR1	2,45	GRM1	2,02
RAB28	3,26	GRID2	2,43	CCDC91	2,01
PARP14	3,25	NRXN3	2,43	MTUS1	2,01
COL13A1	3,20	RALYL	2,42	CACNA2D3	2,01
RNF111	3,20	LARGE	2,40	SLC4A4	2,01
CNTNAP2	3,20	ADAMTS19	2,38	AFF2	2,01
ADAMTSL3	3,19	BRIP1	2,36	CSNK1G1	2,00
GRID1	3,18	ABCA10	2,35	ZNF571	1,99
SOBP	3,18	KCNT2	2,33	MYOM1	1,98
ZFP42	3,16	C9orf3	2,32	SLC26A5	1,95
SLC7A2	3,16	NPAS3	2,32	PHACTR3	1,94
EFHC2	3,16	MSH3	2,31	WDR70	1,92
SLC16A7	3,14	PARD3B	2,30	LOC100128355	1,92
CCDC109B	3,13	SLC24A4	2,30	ANKRD6	1,92
RASGRF2	3,13	PLD5	2,26	LRRTM4	1,92
TPRG1	3,13	LAMA2	2,26	FGF14	1,91
PARK2	3,12	ARHGAP6	2,26	C15orf33	1,89
MADCAM1	3,08	LPXN	2,22	PTPN9	1,89
FGF13	3,07	DOK6	2,22	C1orf112	1,88
TACR1	3,06	STXBP5L	2,22	MAPK10	1,88
DZIP3	3,05	OFCC1	2,21	SLC35F1	1,88
STAT2	3,05	COL21A1	2,19	OPTN	1,87
OTOG	3,05	GTDC1	2,19	PEX5L	1,86
GLIS3	3,00	ANKRD45	2,19	GDA	1,85
PPFIA2	2,96	ZC4H2	2,19	NCOA1	1,84
IMMP1L	2,92	ADAR	2,18	C3orf20	1,81
IMMP2L	2,84	ZNF3	2,16	GPR158	1,80
IGSF21	2,77	CNBD1	2,16	C20orf94	1,79
TPTE2	2,76	KLF12	2,14	SPATA5	1,78
RRH	2,73	LSAMP	2,13	NELL2	1,77
RGS7	2,71	FMN1	2,10	CTNNA3	1,76
NTRK2	2,69	CACNA2D2	2,09	KLHL1	1,75
TMEM117	2,68	ANO3	2,09	SOCS1	1,74
COL25A1	2,64	CNDP1	2,08	FBXO32	1,74
EPHB1	2,64	MYO5B	2,07	FUT10	1,74
NAV2	2,62	FAM49A	2,07	TP63	1,74
ITGB8	2,62	POU6F2	2,06	SLC7A5	1,73
AANAT	2,60	RET	2,05	RALGAPA2	1,73
C15orf43	2,56	SPTY2D1	2,04	NBEA	1,73
PTPRG	2,53	RASSF3	2,04	MYO3A	1,73
GPR112	2,52	POU2F2	2,04	AKAP6	1,73

ID	Fold Change	ID	Fold Change	ID	Fold Change
IKBKB	1,72	NLGN4X	1,46	PCDH15	1,13
ZNF438	1,72	PHACTR1	1,45	C2orf86	1,12
GRIK1	1,72	ZNF521	1,43	AGK	1,11
ZRANB3	1,72	EPHB2	1,42		
STX11	1,70	TGS1	1,42		
TTF2	1,70	NELL1	1,40		
ZBTB20	1,70	CTNND2	1,40		
MYO9A	1,68	CSMD1	1,39		
TULP1	1,68	PPP2R5A	1,36		
CLIP4	1,68	CNTN5	1,36		
ADAMTS2	1,67	NTM	1,36		
ROR2	1,66	WNK2	1,35		
MYCBP2	1,65	MARCH6	1,35		
ARHGEF33	1,65	SIPA1L2	1,35		
PRR3	1,65	ACSL3	1,33		
PARP8	1,65	PLA2G16	1,33		
SAMD12	1,65	C22orf13	1,33		
ARHGAP10	1,64	MYO3B	1,33		
C1orf201	1,64	THRB	1,32		
C7orf58	1,64	ASXL3	1,32		
MGLL	1,63	PDE3A	1,31		
DNAH14	1,63	MAD2L2	1,28		
TANC2	1,63	AGAP1	1,28		
ARHGAP24	1,63	OLFML2A	1,27		
PPP1R16A	1,62	MLLT10	1,27		
SH2D4B	1,62	HTR7	1,24		
CDH4	1,61	SLCO5A1	1,24		
CCDC141	1,60	MKL1	1,23		
TYW1B	1,60	SASH1	1,21		
GALNTL6	1,60	MNAT1	1,20		
PDE6C	1,59	VPS8	1,18		
NINL	1,59	SEC22A	1,18		
FOXP2	1,59	KIN	1,17		
ARHGEF15	1,59	ZMIZ1	1,16		
FAM190A	1,55	PHKB	1,16		
RPS6KC1	1,53	CLASP1	1,16		
PDE3B	1,52	CYTH1	1,15		
VEZT	1,50	HS2ST1	1,15		
FUT8	1,50	FRMD4A	1,15		
DYNC1I1	1,49	ZNF638	1,14		
MTUS2	1,47	KIAA1468	1,14		
C5orf36	1,47	GSK3B	1,13		
COL23A1	1,47	USP53	1,13		

REFERENCIAS

[1] S.D. Davidson, R. Passmore, J.F. Brock, Human nutrition and dietetics, (1973).

[2] M. Mathews, N. Sonenberg, J. Hershey, Translational Control in Biology and Medicine, in: M. Mathews, N. Sonenberg, J. Hershey, NY, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2007pp. 1-40.

[3] F.H. Crick, On protein synthesis, Symposia Of The Society For Experimental Biology. 12 (1958) 138-163.

[4] F.C. Jacob, J. Monod, Genetic regulatory mechanisms in the synthesis of proteins, J Mol Biol. 3 (1961) 318-356.

[5] M.D. Roper, W.D. Wicks, Evidence for acceleration of the rate of elongation of tyrosine aminotransferase nascent chains by dibutyryl cyclic AMP, Proc. Natl. Acad. Sci. USA. 75 (1978) 140-144.

[6] C.G. Proud, Control of the elongation phase of protein synthesis, in: N. Sonenberg, Translation Control Of Gene Expression, New York, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2000pp. 719-739.

[7] L. Gehrke, J. Ilan, Regulation of messenger RNA translation at the elongation step during estradiol-induced vitellogenin synthesis in avian liver, in: J. Ilan, Translational Regulation Of Gene Expression, New York, Plenum Press, 1987pp. 165-186.

[8] J.A. Steitz, Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA, Nature. 6 (1969) 957-964.

[9] T. Endo, B. Nadal-ginard, Three Types of Muscle-Specific Gene Expression in Fusion-Blocked Rat Skeletal Muscle Cells: Translational Control in EGTA-Treated Cells, Cell. 49 (1987) 515-526.

[10] A.J. Kinniburgh, M.D. Mcmullex, T.E. Martii, Distribution of Cytoplasmic Poly(A+)RNA Sequences in Free Messenger Ribonucleoprotein and Polysomes of Mouse Ascites Cells, J Mol Biol. 132 (1979) 695-708.

[11] T. Geoghegan, S. Cereghini, G. Brawermant, Inactive mRNA-protein complexes from mouse sarcoma-180 ascites cells, Proc. Natl. Acad. Sci. USA. 76 (1979) 5587-5591.

[12] A.J. Ouellette, C.P. Ordahl, J. Van Ness, R.A. Malt, Mouse kidney nonpolysomal Messenger ribonucleic acid: Metabolism, coding function, and translational activity, Biochemistry. 21 (1982) 1169-1177.

[13] S. Penman, K. Sherrer, Y. Becker, J.E. Darnell, Polyribosomes in normal and poliovirus-infected HeLa cells and their relationship to messanger-RNA, PNAS. 49 (1963) 654-662.

[14] H.F. Lodish, Translational control of protein synthesis, Annu. Rev. Biochem. (1976).

[15] T. Godefroy-colburns, R.E. Thach, The Role of mRNA Competition in Regulating Translation IV. KINETIC MODEL, Biological Chemistry. 256 (1981) 11762-11773.

[16] A. Deana, J.G. Belasco, Lost in translation: the influence of ribosomes on bacterial mRNA decay, Genes & Development. 19 (2005) 2526-2533.

[17] R.J. Jackson, C.U. Hellen, T.V. Pestova, The mechanism of eukaryotic translation initiation and principles of its regulation, Nature. 10 (2010) 113-127.

[18] A. Roll-mecak, B.S. Shin, T.E. Dever, S.K. Burley, Engaging the ribosome: Universal IFs of translation, Trends Biochem. Sci. 26 (2001) 705-709.

[19] P. Raychaudhuri, A. Chaudhuris, U. Maitra, Formation and Release of Eukaryotic Initiation Factor 2.GDP Complex during Eukaryotic Ribosomal Polypeptide Chain Initiation Complex Formation, The Journal Of Biological Chemestry. 260 (1985) 2140-2145.

[20] J. Nika, S. Rippel, E.M. Hannig, Biochemical Analysis of the eIF2 ɴʟ □ Complex Reveals a Structural Function for eIF2 ␣ in Catalyzed Nucleotide Exchange, Biochemistry. 276 (2001) 1051-1056.

[21] F. Erickson, E.M. Hannig, Ligan interactions with eukariotic translation initiation factor 2: role of the subunit gamma, EMBO Journal. 15 (1996) 6311-6320.

[22] L.D. Kapp, J.R. Lorsch, GTP-dependent Recognition of the Methionine Moiety on Initiator tRNA by Translation Factor eIF2, (2004) 923-936.

[23] R. Benne, J.W. Hershey, The Mechanism of Action of Protein Synthesis Initiation Factors from Rabbit Reticulocytes, The Journal Of Biological Chemestry. 253 (1978) 3078-3087.

[24] J.L. Battiste, T.V. Pestova, C.U. Hellen, G. Wagner, The eIF1A Solution Structure Reveals a Large RNA-Binding Surface Important for Scanning Function, Molecular Cell. 5 (2000) 109-119.

[25] L. Válasek, K.H. Nielsen, A.G. Hinnebusch, Direct eIF2- eIF3 contact in the multifactor complex is important for translation initiation in vivo, The EMBO Journal. 21 (2002) 5886-5898.

[26] V.G. Kolupaeva, A. Unbehaun, I.B. Lomakin, C.U. Hellen, T.V. Pestova, Binding of eukaryotic initiation factor 3 to ribosomal 40S subunits and its role in ribosomal dissociation and anti-association, RNA. 11 (2005) 470-486.

[27] J. Marcotrigiano, A. Gingras, N. Sonenberg, S.K. Burley, Cocrystal Structure of the Messenger RNA 5′ Cap-Binding Protein (eIF4E) Bound to 7-methyl-GDP, Cell. 89 (1997) 951-961.

[28] T.V. Haar, P.D. Ball, J.E. Mccarthy, Stabilization of Eukaryotic Initiation Factor 4E Binding to the mRNA 5□-Cap by Domains of eIF4G, The Journal Of Biological Chemestry. 275 (2000) 30551-30555.

[29] L.A. Passmore, T.M. Schmeing, D. Maag, D.J. Applefield, M.G. Acker, M.A. Algire, et al., The eukaryotic translation initiation factors eIF1 and eIF1A induce an open conformation of the 40S ribosome, Mol. Cell. 26 (2007) 41-50.
[30] V.P. Mauro, S.A. Chappell, J. Dresios, Analysis of Ribosomal Shunting During Translation Initiation in Eukaryotic mRNAs, in: Methods In Enzymology, 429 ed., 2007pp. 323-354.
[31] Y. Yu, A. Marintchev, V.G. Kolupaeva, A. Unbehaun, T. Veryasova, S. Lai, et al., Position of eukaryotic translation initiation factor eIF1A on the 40S ribosomal subunit mapped by directed hydroxyl radical probing, Nucleic Acid Research. 37 (2009) 5167-5182.
[32] M.A. Algire, D. Maag, J.R. Lorsch, Release from eIF2, Not GTP Hydrolysis, Is the Step Controlled by Start-Site Selection during Eukaryotic Translation Initiation, Molecular Cell. 20 (2005) 251-262.
[33] M.G. Acker, J.R. Lorsch, Mechanism of ribosomal subunit joining during eukaryotic translation initiation, Biochemical Society Transactions. 36 (2008) 653-657.
[34] N.C. Kyrpides, C.R. Woese, Universally conserved translation initiation factors, Proc. Natl. Acad. Sci. USA. 95 (1998) 224-228.
[35] T.V. Pestova, J.R. Lorsch, C.U. Hellen, The mechanism of translation initiation in eukaryotes, in: M.B. Mathews, N. Sonenberg, J.W. Hershey, Translational Control In Biology And Medicine, NY, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2007pp. 87-128.
[36] T.V. Pestova, V.G. Kolupaeva, The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection, Genes & Dev. 16 (2002) 2906-2922.
[37] M. Sette, P.V. Tilborg, R. Spurio, R. Kaptein, M. Paci, C.O. Gualerzi, et al., The structure of the translational initiation factor IF1 from E.coli contains an oligomer-binding motif, The EMBO Journal. 16 (1997) 1436-1443.
[38] C.A. Fekete, S.F. Mitchell, V.A. Cherkasova, A. Eld, A. D., M.a, et al., N-and C-terminal residues of eIF1A have opposing effects on the delity of start codon selection, EMBO Journal. 26 (2007) 1602-1614.
[39] A. Parsyan, Y. Svitkin, D. Shahbazian, C. Gkogkas, P. Lasko, W.C. Merrick, et al., mRNA helicases: the tacticians of translational control, Nature. 12 (2011) 235-245.
[40] F. Liu, A. Putnam, E. Jankowsky, ATP hydrolysis is required for DEAD-box protein recycling but not for duplex unwinding, PNAS. 51 (2008) 20209-20214.
[41] J.A. Grifot, R.D. Abramson, C.A. Satler, W.C. Merrick, RNA-stimulated ATPase Activity of Eukaryotic Initiation Factors, Society. 259 (1984) 8648-8654.
[42] J.M. Caruthers, E.R. Johnson, D.B. Mckay, Crystal structure of yeast initiation factor 4A, a DEAD-box RNA helicase, PNAS. 2000 (2000).
[43] G.W. Rogers, N.J. Richter, W.F. Lima, W.C. Merrick, Modulation of the helicase activity of eIF4A by eIF4B, eIF4H, and eIF4F, J. Biol. Chem. 276 (2001) 30914-30922.
[44] A. Marintchev, K.A. Edmonds, B. Marintcheva, E. Hendrickson, M. Oberer, C. Suzuki, et al., Topology and regulation of the human eIF4A/4G/4H helicase complez in translation initiation, Cell. 136 (2009) 447-460.
[45] N. Rozovsky, A.C. Butterworth, M.J. Moore, Interactions between eIF4AI and its accessory factors eIF4B and eIF4H, RNA. 14 (2008) 2136-2148.
[46] T. Naranda, W.B. Strong, J. Menaya, B.J. Fabbri, J.W. Hershey, Two structural domains of initiation factor eIF-4B are involved in binding to RNA, J. Biol. Chem. 269 (1994) 14465-14472.
[47] L. Lindqvist, H. Imataka, J. Pelletier, Cap-dependent eukaryotic initiation factor-mRNA interactions probed by cross-linking, RNA. 14 (2008) 960-969.
[48] K.H. Nielsen, M.A. Behrens, Y. He, C.L. Oliveira, L.S. Jensen, S.V. Hoffmann, et al., Synergistic activation of eIF4A by eIF4B and eIF4G, Nucleic Acid Research. 39 (2011) 2678-2689.
[49] L.C. Waters, S.L. Strong, E. Ferlemann, O. Oka, F.W. Muskett, V. Veverka, et al., Structure of the Tandem MA-3 Region of Pdcd4 Protein and Characterization of Its Interactions with eIF4A and eIF4G, Journal Of Biological Chemistry. 286 (2011) 17270 -17280.
[50] N. LaRonde-LeBlanc, A.N. Santhanam, A.R. Baker, A. Wlodawer, N.H. Colburn, Structural Basis for Inhibition of Translation by the Tumor Suppressor Pdcd4, Molecular And Cellular Biology. 27 (2007) 147-156.
[51] E.D. Gregorio, T. Preiss, M.W. Hentze, Translation driven by an eIF4G core domain in vivo, EMBO Journal. 18 (1999) 4865-4874.
[52] H. Imataka, N. Sonenberg, Human eukaryotic translation initiation factor 4G (eIF4G) possesses two separate and independent binding sites for eIF4A, Mol. Cell. Biol. 17 (1997) 6940-6947.
[53] C. Suzuki, R.G. Garces, K.A. Edmonds, S. Hiller, S.G. Hyberts, A. Marintchev, et al., PDCD4 inhibits translation initiation by binding to eIF4A using both its MA3 domains, PNAS. 105 (2008) 3274-3279.
[54] M. Oberer, A. Marintchev, G. Wagner, Structural basis for the enhancement of eIF4A helicase activity by eIF4G, Genes Dev. 19 (2005) 2212-2223.
[55] M. Bumann, A.E. Oberholzer, C. Bieniossek, H. Trachsel, P. Schu, M. Altmann, et al., Crystal structure of the yeast eIF4A-eIF4G complex: An RNA-helicase controlled by protein– protein interactions¨, PNAS. 105 (2008) 9564-9569.
[56] N. Sonenberg, A.G. Hinnebusch, Regulation of translation initiation in eukaryotes: mechanisms and biological targets, Cell. 136 (2009) 731-745.

[57] A. Hinnebusch, T.E. Dever, K. Asano, Mechanism of translation initiation in the yeast Saccharomyces cerevisiae, in: M. Mathews, N. Sonenberg, J. Hershey, Translational Control In Biology And Medicine, NY, Cold Spring Harbor, Cold Spring Harbor Laboratory Press, 2007pp. 225-268.
[58] B. Raught, A.C. Gingras, Signaling to translation initiation, in: M. Mathews, N. Sonenberg, J.W. Hershey, Translational Control In Biology And Medicine, NY, Cold Spring Harbor, Cold Spring Harbor Laboratory Press, 2007pp. 369-400.
[59] J. Marcotrigiano, I.B. Lomakin, N. Sonenberg, T.V. Pestova, S.K. Burley, A Conserved HEAT Domain within eIF4G Directs Assembly of the Translation Initiation Machinery, Molecular Cell. 7 (2001) 193-203.
[60] M.K. Holz, B.A. Ballif, S.P. Gygi, J. Blenis, mTOR and S6K1 Mediate Assembly of the Translation Preinitiation Complex through Dynamic Protein Interchange and Ordered Phosphorylation Events, Cell. 123 (2005) 569-580.
[61] N. Dorrello, Valerio, A. Peschiaroli, D. Guardavaccaro, N. Colburn, N. Sherman, M. Pagano, S6K1- and BTRCP-Mediated Degradation of PDCD4 Promotes Protein Translation and Cell Growth, Science. 314 (2006) 467-471.
[62] R.J. Dowling, I. Topisirovic, B.D. Fonseca, N. Sonenberg, Dissecting the role of mTOR: Lessons from mTOR inhibitors, Biochimica Et Biophysica Acta. 1804 (2010) 433-439.
[63] N. Hay, N. Sonenberg, Upstream and downstream of mTOR, Genes Dev. 18 (2004) 1926-1945.
[64] Y. Mamane, E. Petroulakis, O. LeBacquer, N. Sonenberg, mTOR, translation initiation and cancer, Oncogene. 25 (2006) 6416-6422.
[65] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, et al., Expression monitoring by hybridization to high-density oligonucleotide arrays, Nature. 14 (1996).
[66] D. Baek, J. Villén, C. Shin, F.D. Camargo, S.P. Gygi, D.P. Bartel, The impact of microRNAs on protein output, Nature. 455 (2008) 64-71.
[67] S.P. Gygi, Y. Rochon, B.R. Franza, R. Aebersold, Correlation between Protein and mRNA Abundance in Yeast, Molecular And Cellular Biology. 19 (1999) 1720-1730.
[68] M. Selbach, B. Schwanhausser, N. Thierfelder, Z. Fang, R. Khanin, N. Rajewsky, Widespread changes in protein synthesis induced by microRNAs, Nature. 455 (2008) 58-63.
[69] N. Sonenberg, A.G. Hinnebusch, New Modes of Translational Control in Development, Behavior, and Disease, Molecular Cell. 28 (2007) 721-729.
[70] O. Larsson, R. Nadon, Gene expression - time to change point of view?, Biotechnol. Genet. Eng. Rev. 25 (2008) 77-92.
[71] O. Namy, J. Rousset, S. Napthine, I. Brierley, Reprogrammed Genetic Decoding in Cellular Gene Expression, Molecular Cell. 13 (2004) 157-168.
[72] A. Lazaris-karatzas, K.S. Montine, N. Sonenberg, Malignant transformation by a eukariotic initiation factor subunit that binds to mRNA 5´ cap, Nature. 345 (1990).
[73] A. Provenzani, R. Fronza, F. Loreni, A. Pascale, M. Amadio, A. Quattrone, Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis, Carcinogenesis. 27 (2006) 1323-1333.
[74] X. Lu, L.D. Peña, C. Barker, K. Camphausen, P.J. Tofilon, Radiation-Induced Changes in Gene Expression Involve Recruitment of Existing Messenger RNAs to and away from Polysomes, Cancer Research. 66 (2006) 1052-1061.
[75] J.D. Blais, V. Filipenko, M. Bi, H.P. Harding, D. Ron, C. Koumenis, et al., Activating Transcription Factor 4 Is Translationally Regulated by Hypoxic Stress, Molecular And Cellular Biology. 24 (2004) 7469-7482.
[76] N.T. Ingolia, S. Ghaemmaghami, J.R. Newman, J.S. Weissman, Genome-wide Analysis in Vivo of Translation with Nucleotide Resoultion Using Ribosome Profiling, Science. 324 (2009) 218-223.
[77] H. Kitamura, T. Nakagawa, M. Takayama, Y. Kimura, A. Hijika, O. Ohara, Post-transcriptional effects of phorbol 12-myristate 13-acetate on transcriptome of U937 cells, FEBS Letters. 578 (2004) 180-184.
[78] X. Qin, S. Ahn, T. Speed, G. Rubin, Global analyses of mRNA translational control during early Drosophila embryogenesis, Genome Biology. 8 (2007) 1-18.
[79] N. Iguchi, J.W. Tobias, N.B. Hecht, Expression profiling reveals meiotic male germ cell mRNAs that are translationally up- and down-regulated, PNAS. 103 (2006) 7712-7717.
[80] E. Klann, J. Richter, Translational control of synaptic plasticity and learning and memory, in: M. Mathews, N. Sonenberg, J.W. Hershey, Translational Control In Bioloy And Medicine, NY, Cold Spring Harbor, Cold Spring Harbor Laboratory Press, 2007pp. 485-506.
[81] V.A. Polunovsky, P.B. Bitterman, The Cap-Dependent Translation Apparatus Integrates and Amplifies Cancer Pathways, Rna Biology. 3 (2006) 10-17.
[82] R. Schneider, N. Sonenberg, Translational control in cancer development and progression, in: M. Mathews, N. Sonenberg, J.W. Hershey, Translational Control In Biology And Medicine, NY, Cold Spring Harbor, Cold Spring Harbor Laboratory Press, 2007pp. 401-432.
[83] K. Shibahara, M. Asano, Y. Ishida, T. Aoki, T. Koike, T. Honjo, Isolation of a novel mouse gene MA-3 that is induced upon programmed cell death, Gene. 166 (1995) 297-301.
[84] B. Lankat-buttgereit, R. Goke, The tumor suppressor Pdcd4: recent advances in the elucidation of function and regulation, Biol Cell. 101 (2009) 309-317.

[85] J.L. Cmarik, H. Min, G. Hegamyer, S. Zhan, M. Kulesz-martin, H. Yoshinaga, et al., Differentially expressed protein Pdcd4 inhibits tumor promoter-induced neoplastic transformation, PNAS. 96 (1999) 14037-14042.
[86] L. Ding, X. Zhang, M. Zhao, Z. Qu, S. Huang, M. Dong, et al., An essential role of PDCD4 in progression and malignant proliferation of gastrointestinal stromal tumors, Medical Oncology. (2011).
[87] J.H. Leupold, H.S. Yang, N.H. Colburn, I. Asangani, S. Post, H. Allgayer, Tumor suppressor Pdcd4 inhibits invasion/intravasation and regulates urokinase receptor (u-PAR) gene expression via Sp-transcription factors, Oncogene. 26 (2007) 4550-4562.
[88] Q. Wang, Z. Sun, H.S. Yang, Downregulation of tumor suppressor Pdcd4 promotes invasion and activates both β-catenin/Tcf and AP-1-dependent transcription in colon carcinoma cells, Oncogene. 27 (2008) 1527-1535.
[89] N. Bitomsky, N. Wethkamp, R. Marikkannu, K.H. Klempnauer, siRNA-mediated knockdown of Pdcd4 expression causes upregulation of p21 Waf1/Cip1 expression, Oncogene. 27 (2008) 4820-4829.
[90] P. Singh, R. Marikkannu, N. Bitomsky, K.H. Klempnauer, Disruption of the Pdcd4 tumor suppressor gene in chicken DT40 cells reveals its role in the DNA-damage response, Oncogene. 28 (2009) 3758-3764.
[91] H. Soejima, O. Miyoshi, H. Yoshinaga, Z. Masaki, I. Ozaki, S. Kajiwara, et al., Assignment of the programmed cell death 4 gene (PDCD4) to human chromosome band 10q24 by in situ hybridization, Cytogenet. Cell Genet. 87 (1999) 113-114.
[92] K. Eto, S. Goto, W. Nakashima, Y. Ura, S. Abe, Loss of programmed cell death 4 induces apoptosis by promoting the translation of procaspase-3 mRNA, Cell Death And Differentiation. (2011) 1-9.
[93] S. Matsuhashi, Y. Narisawa, I. Ozaki, T. Mizuta, Expression patterns of programmed cell death 4 protein in normal human skin and some representative skin lesions, Exp. Dermatol. 16 (2007) 179-184.
[94] Y.H. Wen, X. Shi, L. Chiriboga, S. Matsahashi, H. Yee, O. Afonja, Alterations in the expression of PDCD4 in ductal carcinoma of the breast, Oncol. Rep. 18 (2007) 1387-1393.
[95] Y. Onishi, S. Hashimoto, H. Kizaki, Cloning of the TIS gene suppressed by topoisomerase inhibitors, GENE. 215 (1998) 453-459.
[96] I.A. Asangani, S.A. Rasheed, D.A. Nikolova, J.H. Leupold, N.H. Colburn, S. Post, et al., MicroRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pdcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer, Oncogene. 27 (2008) 2128-2136.
[97] A. Palamarchuk, A. Efanov, V. Maximov, R.I. Aqeilan, C.M. Croce, Y. Pekarsky, Akt Phosphorylates and Regulates Pdcd4 Tumor Suppressor Protein, Cancer Research. 65 (2005) 11282-11286.
[98] U. Schlichter, O. Burk, S. Worpenberg, K.H. Klempnauer, The chicken Pdcd4 gene is regulated by v-Myb, Oncogene. 20 (2001) 231-239.
[99] H. Fan, Z. Zhao, Y. Quan, J. Xu, J. Zhang, W. Xie, DNA methyltransferase 1 knockdown induces silenced CDH1 gene reexpression by demethylation of methylated CpG in hepatocellular carcinoma cell line SMMC-7721, European Journal Of Gastroenterology & Hepatology. 19 (2007) 952-961.
[100] F. Gao, X. Wang, F. Zhu, Q. Wang, X. Zhang, C. Guo, et al., PDCD4 gene silencing in gliomas is associated with 5´CpG island methylation and unfavourable prognosis, J. Cell. Mol. Med. 13 (2009) 4257-4267.
[101] N. Bitomsky, M. Bohm, K.H. Klempnauer, Transformation suppressor protein Pdcd4 interferes with JNK-mediated phosphorylation of c-Jun and recruitment of the coactivator p300 by c-Jun, Oncogene. 23 (2004) 7484-7493.
[102] P.G. Loh, H. Yang, M.A. Walsh, Q. Wang, X. Wang, Z. Cheng, et al., Structural basis for translational inhibition by the tumour suppressor Pdcd4, EMBO Journal. 28 (2009) 274-285.
[103] L. Wedeken, J. Ohnheiser, B. Hirschi, N. Wethkamp, K. Klempnauer, Association of Tumor Suppressor Protein Pdcd4 With Ribosomes Is Mediated by Protein-Protein and Protein-RNA Interactions, Genes & Cancer. 1 (2010) 293-301.
[104] P. Singh, L. Wedeken, L.C. Waters, K. Klempnauer, Pdcd4 directly binds the coding region of c-myb mRNA and suppresses its translation, Oncogene. (2011).
[105] L. Wedeken, P. Singh, K. Klempnauer, The tumor suppressor protein Pdcd4 inhibits the translation of p53 mRNA, J Biol Chem. 286 (2011) 42855-42862.
[106] M.L. Mucenski, K. McLain, A.B. Kier, S.H. Swerdlow, C.M. Schreiner, T.A. Miller, et al., A Functional c-myb Gene Is Required for Normal Murine Fetal Hepatic Hematopoiesis, Cell. 65 (1991) 677-689.
[107] Y.E. Zhou, J.P. O´ Rourke, J.S. Edwards, S.A. Ness, Single Molecule Analysis of c-myb Alternative Splicing Reveals Novel Classifiers for Precursor B-ALL, PLoS ONE. 6 (2011).
[108] Y. Zhou, c-MYB alternative splicing: a novel biomarker in leukemia, Albuquerque, New Mexico, The University of New Mexico, 2011.
[109] A. Strasser, L.O. Connor, V.M. Dixit, Apoptosis Signaling, Annu. Rev. Biochem. 69 (2000) 217-245.
[110] J. Mosner, T. Mummenbrauer, C. Bauer, G. Sczakiel, F. Grosse, W. Deppert, Negative feedback regulation of wild-type p53 biosynthesis, EMBO Journal. 14 (1995) 4442-4449.
[111] K. Bensaad, K.H. Vousden, p53: new roles in metabolism, Trends In Cell Biology. 17 (2007) 286-291.
[112] K.H. Vousden, K.M. Ryan, p53 and metabolism, Nature Reviews - Cancer. 9 (2009) 691-700.
[113] E. Lundberg, L. Fagerberg, D. Klevebring, I. Matic, T. Geiger, J. Cox, et al., Defining the transcriptome and proteome in three functionally different human cell lines¨, Molecular Systems Biology. 6 (2010) 1-10.

[114] N. Nagaraj, J.R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, et al., Deep proteome and transcriptome mapping of a human cancer cell line, Molecular Systems Biology. 60 (2011).
[115] P.O. Brown, D. Botstein, Exploring the new world of the genome with DNA microarrays, Nature Genetics Supplement. 21 (1999) 33-37.
[116] A. Jacquier, The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs, Nat Rev Genet. 10 (2009) 833-844.
[117] J.H. Malone, B. Oliver, Microarrays, deep sequencing and the true measure of the transcriptome, BMC Biology. 9 (2011).
[118] M. Reimers, Making Informed Choices about Microarray Data Analysis, PLoS Computational Biology. 6 (2010) 1-7.
[119] Y. Arava, Y. Wang, J.D. Storey, C.L. Liu, P.O. Brown, D. Herschlag, Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae, PNAS. 100 (2003) 3889-3894.
[120] A. Mortazavi, B.A. Williams, K. Mccue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nature Methods. 5 (2008) 621-628.
[121] H. Richard, M.H. Schulz, M. Sultan, A. Nurnberg, S. Schrinner, D. Balzereit, et al., Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments, Nucleic Acid Research. 38 (2010).
[122] M. Telonis-scott, A. Kopp, M.L. Wayne, S.V. Nuzhdin, L.M. McIntyre, Sex-Specific Splicing in Drosophila: Widespread Occurrence, Tissue Speci city and Evolutionary Conservation, Genetics. 181 (2009) 421-434.
[123] A. Agarwal, D. Koppstein, J. Rozowsky, A. Sboner, L. Habegger, W. Hillier, et al., Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays, BMC Genomics. 11 (2010).
[124] R.D. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T.J. Pugh, et al., Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing, Biotechniques. 45 (2008).
[125] R. Lister, R.C. O´Malley, J. Tonti-Filippini, B.D. Grogory, C.C. Berry, A.H. Millar, et al., Highly Integrated Single-Base REsolution Maps of the Epigenome in Arabidopsis, Cell. 133 (2008) 523-536.
[126] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, et al., The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing, Science. 320 (2008) 1344-1349.
[127] B.T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, et al., Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution, Nature. 453 (2008) 1239-1243.
[128] P.L. Auer, R.W. Doerge, Statistical Design and Analysis of RNA Sequencing Data, Genetics. 185 (2010) 405-416.
[129] R. Li, J. Wang, The sequence and de novo assembly of the giant panda genome, Nature. 463 (2010) 311-317.
[130] D.R. Bentley, Whole-genome re-sequencing, Curr Opin Genet Dev. 16 (2006) 545-552.
[131] Q. Xia, Y. Guo, Z. Zhang, D. Li, Z. Xuan, Z. Li, et al., Complete Resequencing of 40 Genomes Reveals Domestication Events and Genes in Silkworm (Bombyx), Science. 326 (2009) 433-436.
[132] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-Wide Mapping of in Vivo Protein-DNa Interactions, Science. 316 (2007) 1497-1502.
[133] J.F. Costello, M. Krzywinski, M.A. Marra, A first look at entire human methylomes, Nature Biotechnology. 27 (2009) 1130-1132.
[134] R.A. Edwards, B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D.M. Peterson, et al., Using pyrosequencing to shed light on deep mine microbial ecology, BMC Genomics. 7 (2006).
[135] J.J. Qin, J. Wang, A human gut microbial gene catalogue established by metagenomic sequencing, Nature. 464 (2010) 59-65.
[136] X. Zhou, L. Ren, Q. Meng, Y. Li, Y. Yu, J. Yu, The next-generation sequencing technology and application, Protein & Cell. 1 (2010) 520-536.
[137] C. Adessi, G. Matton, G. Ayala, G. Turcatti, J. Mermod, P. Mayer, et al., Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms, Nucleic Acid Research. 28 (2000).
[138] M. Fedurco, A. Romieu, S. Williams, I. Lawrence, G. Turcatti, BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies, Nucleic Acid Research. 34 (2006).
[139] http://www.illumina.com/systems/genome_analyzer.ilmn, (2011).
[140] http://www.illumina.com/systems/sequencing.ilmn, (2012).
[141] J. Shendure, G.J. Porreca, N.B. Reppas, X.X. Lin, J.P. McCutcheon, A.M. Rosenbaum, et al., Accurate multiplex polony sequencing of an
evolved bacterial genome, Science. 309 (2005) 1728-1732.
[142] K. McKeman, A. Bianchard, L. Kotler, G. Costa, Reagents, methods, and libraries for bead-based sequencing, US
patent Application 20080003571. (2006).
[143] J.N. Housby, E.M. Southern, Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides, Nucleic Acids Res. 26 (1998) 4259-4266.
[144] http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html, 2012.

[145] H.A. Meijer, A.A. Thomas, Control of eukaryotic protein synthesis by upstream open reading frames in the 5´-untranslated region of an mRNA, Biochem J. 267 (2002).
[146] H. Guo, N.T. Ingolia, J.S. Weissman, D.P. Bartel, Mammalian microRNAs predominantly act to decrease target mRNA levels, Nature. 466 (2010) 835-841.
[147] N.T. Ingolia, L.F. Lareau, J.S. Weissman, Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes, Cell. 147 (2011) 789-802.
[148] M. Fresno, A. Jimenez, D. Vázquez, Inhibition of Translation in Eukaryotic Systems by Harringtonine, J. Biochem. 72 (1977) 323-330.
[149] F. Robert, M. Carrier, S. Rawe, S. Chen, S. Lowe, J. Pelletier, Altering Chemosensitivity by Modulating Translation Elongation, PLoS ONE. 4 (2009).
[150] D.W. Reid, C.V. Nicchitta, Genome-scale ribosome footprinting identifies a primary role for endoplasmic reticulum-bound ribosomes in the translation of the mRNA transcriptome, J. Biol. Chem. (2011).
[151] G.A. Brar, M. Yassour, N. Friedman, A. Regev, N.T. Ingolia, J.S. Weissman, High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling, Sciencexpress. (2011).
[152] M. Livingstone, E. Atas, A. Meller, N. Sonenberg, Mechanisms governing the control of mRNA translation, Physical Biology. 7 (2010).
[153] F. Meric, K.K. Hunt, Translation Initiation in Cancer: A Novel Target for Therapy, Molecular Cancer Therapeutics. 1 (2002) 971-979.
[154] www.asperasoft.com/en/technology/fasp_overview_1/fasp_technology_overview_1, (2011).
[155] http://clcbio.com/files/whitepapers/white_paper_on_reference_assembly_on_the_CLC_Assembly_Cell.pdf, (2012).
[156] A.G. Hinnebusch, Mechanisms of Gene Regulation in the General Control of Amino Acid Biosynthesis in Sacchatomyces cerecisiae, 52 (1988) 248-273.
[157] T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, et al., An evolutionarily conserved mechanism
for controlling the efficiency of protein translation, Cell. 141 (2010) 344-354.
[158] D.W. Huang, B.T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, et al., DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists, Nucleic Acid Research. 35 (2007) 169-175.
[159] G.J. Dennis, B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, et al., DAVID: Database for Annotation, Visualization, and Integrated Discovery, Genome Biology. 4 (2003).
[160] D.W. Huang, B.T. Sherman, R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucleic Acid Research. 37 (2009) 1-13.
[161] D.W. Huang, B.T. Sherman, Q. Tan, J.R. Collins, G. Alvord, J. Roayaei, et al., The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists, Genome Biology. (2007).
[162] I. Piekarska, J. Rytka, B. Rempola, Regulation of sporulation in the yeast Saccharomyces cerevisiae, Acta Biochimica Polonia. 57 (2010) 241-250.
[163] A.M. Neiman, Sporulation in the Budding Yeast Saccharomyces cerevisiae, Genetics. 189 (2011) 737-765.
[164] B.M. Bolstad, R.A. Irizarry, M. Astrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics. 19 (2003) 185-193.
[165] J.H. Bullard, E. Purdom, K.D. Hansen, S. Dudoit, Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, BMC Bioinformatics. 11 (2010).
[166] A.J. Kal, A.J. Zonneveld, V. Benes, M.V. Berg, M.G. Koerkamp, K. Albermann, et al., Dynamics of Gene Expression Revealed by Comparison of Serial Analysis of Gene Expression Transcript Profiles from Yeast Grown on Two Different Carbon Sources, Molecular Biology Of The Cell. 10 (1999) 1859 -1872.
[167] W. Cao, S. Manicassamy, H. Tang, S.P. Kasturi, A. Pirani, N. Murthy, et al., Toll-like receptor–mediated induction of type I interferon in plasmacytoid dendritic cells requires the rapamycin-sensitive PI(3)K-mTOR-p70S6K pathway, Nature Immunology. 9 (2008) 1157-1164.
[168] R. Colina, M. Costa-Mattioli, R.J. Dowling, M. Jaramillo, L. Tai, C.J. Breitbach, et al., Translational control of the innate immune response through IRF-7, Nature. 452 (2008) 323-329.
[169] K. Honda, H. Yanai, H. Negishi, M. Asagiri, M. Sato, T. Mizutani, et al., IRF-7 is the master regulator of type-I interferon-dependent immune responses, Nature. 434 (2005) 772-777.
[170] M.K. Tripathi, G. Chaudhuri, Down-regulation of UCRP and UBE2L6 in BRCA2 knocked-down human breast cells, Biochem Biophys Res Commun. 328 (2005) 43-48.
[171] S. Lipkowitz, Ubiquitin mediated degradation of growth factor receptors in the pathogenesis and treatment of cancer, Breast Cancer Research. 5 (2003) 8-15.
[172] C. Zhao, S.L. Beaudenon, M.L. Kelley, M.B. Waddell, W. Yuan, B.A. Schulman, et al., The UbcH8 ubiquitin E2 enzyme is also the E2 enzyme for ISG15, an IFN- ␣␤ NL -induced ubiquitin-like protein, PNAS. 101 (2004) 7578-7582.
[173] A. Bhoumik, Z. Ronai, ATF2 A transcription factor that elicits oncogenic or tumor suppressor activities, Cell Cycle. 7 (2008) 2341-2345.

[174] M. Takahashi, J. Ritz, G.M. Cooper, Activation of a novel human transforming gene, ret, by DNA rearrangement, Cell. 42 (1985) 581-588.
[175] A. Morandi, I. Plaza-Menacho, C.M. Isacke, RET in breast cancer: functional and therapeutic implications, Trends Mol Med. 17 (2011) 149-157.
[176] C. Moreiro, A. Rapella, E. Tassano, C. Rosanda, C. Panarello, MLL-MLLT10 fusion gene in pediatric acute megakaryoblastic leukemia, Leuk Res. 29 (2005) 1223-1226.
[177] http://www.ncbi.nlm.nih.gov/gene/89780, (2012).
[178] M. Katoh, M. Katoh, WNT Signaling Pathway and Stem Cell Signaling Network, Clinical Cancer Research. 13 (2007) 4042-4045.
[179] T.P. Rao, M. Kühl, An update Overview on Wnt Signaling PAthways, Circulation Research. 106 (2010) 1798-1806.