PEDECIBA BIOLOGÍA

Tesis doctoral

Mecanismos de diferenciación y auto-renovación de células madre mesenquimales a adipocitos

Autor:
Lucía Spangenberg

Supervisores:
Dr. Bruno DALLAGIOVANNA
Dr. Hugo NAYA

realizada en la

Unidad de Bioinformática
Institut Pasteur de Montevideo

17 de octubre 2013

PEDECIBA BIOLOGÍA

Resumen

INSTITUT PASTEUR de MONTEVIDEO
Unidad de Bioinformática

Mecanismos de diferenciación y auto-renovación de células madre mesenquimales a adipocitos

por Lucía Spangenberg

En esta tesis se intentan comprender más a fondo los mecanismos de regulación transcripcional y post-transcripcional de los mecanismos de diferenciación de células madre a adipocitos. Se utilizan experimentos de RNA-Seq para cuantificar la expresión génica y uso de transcriptos en muestras tanto de ARN total, como en la fracción polisomal. Comparando ambas fracciones se analizó y cuantificó el grado de regulación post-transcripcional durante la adipogénesis, encontrándose una alta cantidad de genes regulados por estos mecanismos. Se propone la poliadenilación alternativa como un mecanismo importante de regulación post-transcripcional, el cual genera transcriptos con diferentes 3'UTRs y diferentes sitios de unión a miRNAs. Incluyendo el uso diferencial de transcriptos, el largo de las 3'UTR y por lo tanto, el efecto de los miRNAs se lograron desarrollar modelos lineales capaces de explicar con mayor precisión los niveles de proteínas observados a partir del ARN mensajero. Entre otros aportes, identificamos un perfil particular de los fibroblastos, con posibles aplicaciones a su caracterización y separación. Además trabajamos sobre los procesos subyacentes a los cambios ocurridos en las tasas de asociación de los mensajeros a los polisomas y su posible implicación en el proceso de diferenciación.

Agradecimientos

Esto es un trabajo en equipo, por lo tanto quisiera agradecer a todos los jugadores.
Primero que nada, quisiera agradecer a mis tutores, Bruno Dallagiovanna y Hugo Naya, por el apoyo incondicional durante todo este período que no siempre fue fácil.
Gracias por las ideas, las discusiones, las enseñanzas y la paciencia máxima.
A Alejandro Correa, por siempre estar del otro lado del mail para contestar todas mis preguntas en tiempo real y por el buen humor que lo caracteriza (muy importante en algunas etapas).
A los UBis, cada uno de ellos, por sus aportes emocionales, académicos, alimentarios (sobre todo alimentarios).
A mis viejos amigos de toda la vida y a los que se sumaron más adelante, que me apoyaron en todo momento.
A mi familia, por la confianza total y la crítica sincera.
A mi paciente compañero de vida por todo.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Marco biológico del problema

1.1. Contexto histórico

Investigaciones médicas y científicas de mediados del 1800 ya planteaban ideas sobre células especiales con la capacidad de generar diferentes tipos celulares, particularmente en la sangre. Varias teorías fueron postuladas sobre sus características. No fue sino hasta 1963 que su existencia fue probada por dos científicos canadienses, James Till y Ernest McCulloch, los cuales aislaron células madre de médula ósea de ratón y observaron su capacidad de multiplicarse y de generar varios tipos distintos de células sanguíneas; características básicas de una célula madre como se conoce hoy en día.

A partir de ese momento, varios hitos ocurrieron a lo largo de los años. Los transplantes de médula ósea (el primero en 1968) fueron considerados los primeros transplantes de células madre. Las células madre hematopoyéticas fueron descubiertas en 1978 en el cordón umbilical humano. En 1981 los científicos Martin Evans, Matthew Kaufman, y Gail Martin derivaron las primeras células madre embrionarias de ratón y más de 15 años más tarde, Thomson y colaboradores, aislaron las primeras células madre embrionarias humanas [3].

En estos últimos 20 años, con la experiencia acumulada de décadas, los avances científicos aumentaron exponencialmente. En el 2003 Songtao Shi descubre una nueva fuente de células madre adultas en los dientes de leche [4]. En los años 2004 y 2005 el científico coreano Hwang Woo-Suk aseguró haber fabricado líneas celulares de células madre embrionarias humanas a partir de óvulos no fecundados. Sus resultados fueron publicados en artículos en la revista *Science*, sin embargo, los mismos fueron retirados al haberse demostrado su falsedad. Las células madre inducidas (iPS; induced pluripotent stem cells) de ratón fueron derivadas por los científicos japoneses Takahashi y Yamanaka en el 2006 [5]. Las mismas son células adultas diferenciadas que fueron reprogramadas para

volver a tener las características de las células madre pluripotentes. Un año más tarde estos científicos consiguen al mismo tiempo que Junying Yu y colaboradores, derivar iPS a partir de células humanas adultas ya diferenciadas (fibroblastos) [6, 7].

Las células madre mesenquimales son células adultas multipotentes, por lo que pueden generar varios tipos celulares especializados. La primera regeneración exitosa de cartílago en una rodilla humana a partir de un transplante autólogo de célula madre mesenquimal fue reportado en marzo del 2008 [8]. En el 2010 comenzó el primer ensayo clínico en humanos con células madre embrionarias, utilizándose para la reparación de las lesiones de la médula espinal. Geron fue la empresa biofarmacéutica que diseñó y llevó a cabo el ensayo clínico. Varios ensayos clínicos de la misma empresa, abarcando otras áreas, siguieron a éste. El científico Katsuhiko Hayashi a partir de células de piel de ratón logró crear células madre y utilizó estas células para crear óvulos de ratón. Éstos fueron fecundados y dieron lugar, a su vez, a ratones fértiles [9].

Los avances científicos siguen hasta el día de la fecha, captando la atención de los medios de comunicación y a través de ellos a la sociedad. El artículo en la revista *Cell* sobre la exitosa clonación de células madre embrionarias humanas (lo que intentó realizar Hwang en el 2004) publicado a principios de junio del corriente, fue publicitado por los medios con anterioridad a la publicación del artículo [10]. Existe mucha agitación y curiosidad de la comunidad científica y no científica sobre esta área de la investigación, no sólo por su enorme potencialidad y alcance, sino también, por los problemas éticos y morales que conlleva.

Las células madre tienen un increíble potencial en diferentes campos, tanto en aplicaciones médicas y farmacológicas como en el área de las ciencias básicas. Las mismas tienen el potencial de ser utilizadas, por ejemplo, en la medicina regenerativa (ingeniería de tejidos), en el desarrollo de nuevas drogas, como también en el estudio del funcionamiento normal de la célula.

1.2. Definición general de las células madre

Las células madre se caracterizan por poseer dos propiedades fundamentales, la *auto-renovación* y la *diferenciación* a otros tipos celulares.

La auto-renovación es el proceso, por el cual la célula se divide manteniendo el estado indiferenciado, generando así células similares a si misma. En un principio, se caracterizó a esa división celular como exclusivamente asimétrica, es decir, ambas células hijas no eran necesariamente idénticas entre si. Una célula hija es igual a su progenitora, mateniendo el caracter de célula madre, y la otra comienza el proceso de diferenciación, siendo de esta forma distinta a su progenitora. Más adelante, se descubrió que las células madre

también se dividen de forma simétrica [11]. A partir de una célula madre, se generan dos células madre hijas iguales a su progenitora. El cuerpo adulto a lo largo de la vida, sufre un proceso normal de pérdida de células. Las células madre (adultas) en los tejidos, a través del mecanismo de auto-renovación se aseguran de tener la cantidad necesaria de células para poder abastecer al cuerpo de células especializadas. El mecanismo de auto-renovación es vital para asegurarse un "pool" de células madre en el organismo a lo largo de la vida.

El mecanismo de diferenciación es el proceso por el cual la célula pasa de un estado menos diferenciado a uno más especializado. La diferenciación ocurre numerosas veces durante el desarrollo de un organismo multicelular, mientras el organismo se desarrolla de un simple cigoto, a un organismo complejo con varios tipos celulares y tejidos. Células madre y procesos de diferenciación existen también en el organismo adulto; éstas se encuentran en tejidos específicos, se diferencian para mantener la cantidad de células del organismo y, en caso de lesiones, para reparar las zonas dañadas. El proceso de diferenciación involucra grandes cambios en el tamaño celular, morfología, potencial de membrana, actividad metabólica, entre otros. Estos cambios se deben a variaciones en la regulación transcripcional, regulación post-transcripcional y modificaciones epigenéticas [12].

No todas las células madre generan cualquier tipo de tejido especializado. Éstas pueden clasificarse de acuerdo a su capacidad de diferenciación en *totipotentes*, *pluripotentes* y *multipotentes*. Si bien las definiciones varían frecuentemente dentro de la literatura, nos basaremos en la siguiente definición para el resto del trabajo. Las células totipotentes son aquellas que pueden generar todo tipo de tejido, incluído el cordón umbilical y la placenta. En el cuerpo humano sólo existe un estadío del desarrollo embrionario que contiene células totipontentes, y son las 8 células del cigoto. Una vez pasado este estadío, no existen más células capaces de generar todos los tipos celulares, sin restricciones. Las células *pluripotentes* son aquellas que pueden diferenciarse en cualquier tipo celular del organismo adulto, es decir, pueden formar todos los tejidos menos los del cordón umbilical y la placenta. Las células madre embrionarias pertenecen a esta categoría. Las células madre *multipotentes*, son aquellas que pueden diferenciarse en tipos específicos de tejidos, en general en tipos celulares derivados de la misma capa embrionaria. Dentro de esta categoría se encuentran también las que se denominan células madre tejido-específicas, ya que se encuentran en el organismo adulto, en tejidos específicos y son capaces de generar células específicas de ese tejido en donde se encuentran. Las células madre adultas se subclasifican a su vez de acuerdo al subtipo de tejido que pueden generar. Este trabajo de tesis utiliza células madre adultas, específicamente las *mesenquimales* que se verán en detalle más adelante.

1.3. Obtención, características y aplicaciones de las células madre embrionarias

Las células madre embrionarias tienen un enorme potencial capaz de revolucionar el sistema de salud. No solamente en el área del tratamiento de las enfermedades, sino más bien en el entendimiento del funcionamiento de la célula normal y del surgimiento y la progresión y de las enfermedades, ayudando a crear tratamientos más eficaces.

1.3.1. General

El nombre célula madre embrionaria se debe a que las mismas se derivan de las primeras etapas del desarrollo embrionario. La fecundación de un óvulo con un espermatozoide da lugar a un cigoto, una única célula con increíbles propiedades que pone en marcha al desarrollo embrionario. El cigoto se diferencia a un *blastocito* (a los 5 días posteriores a la fecundación), un cúmulo de 16 células, rodeadas por una capa celular exterior, que posteriormente dará lugar a la placenta y el cordón umbilical. Las 16 células interiores son las células madre embrionarias, las cuales pueden ser cultivadas en el laboratorio. Una

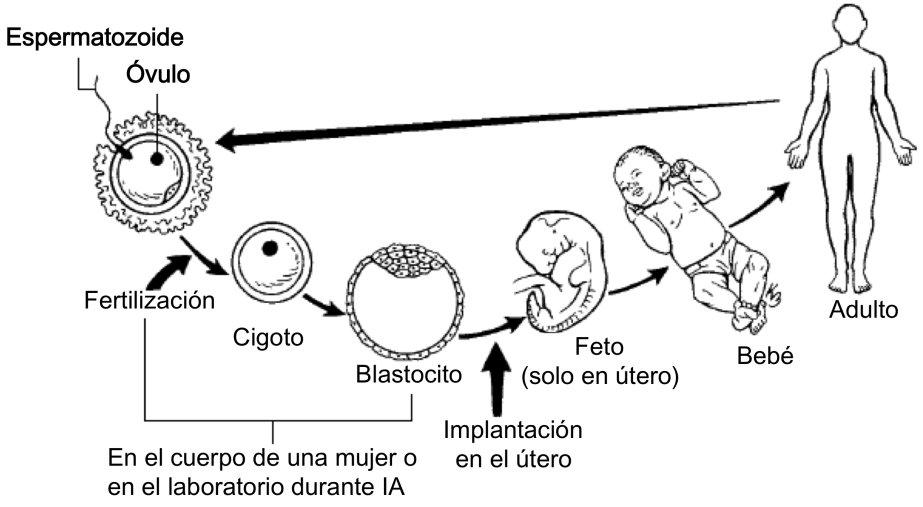


FIGURA 1.1: Obtención de células madre embrionarias a partir de blastocitos. Figura obtenida del libro "Stem cells for dummies", capítulo IV, página 54.

de las formas más usuales de obtención de estas células es el remanente de blastocitos generados para inseminaciones artificiales. Este procedimiento consiste en fecundar un óvulo con un espermatozoide y dejarlo desarrollarse hasta el estado de blastocito. Esto se realiza para varios óvulos. Un par de los blastocitos se implantan en el útero de la mujer, con el fin de iniciar un embarazo. El resto de ellos se congelan y se guardan para próximos intentos. En el caso que la pareja decida no utilizar los blastocitos restantes, dependiendo de la legislación del país, los mismos pueden ser utilizados para la investigación. La figura 1.1 muestra esquemáticamente el desarrollo desde el cigoto hasta el ser humano adulto, pasando por las diferentes etapas. Los procesos desde la fertilización hasta el blastocito pueden ocurrir tanto en el laboratorio como en el cuerpo de la mujer.
Una de las ventajas de estas células, más allá de poder diferenciarse en todos los tipos celulares en el cuerpo adulto, es la capacidad de crecer en cultivo de forma indefinida. Esta característica es extremadamente útil ya que permite generar grandes cantidades de células, vital para las aplicaciones médicas. A su vez, las células embrionarias pueden almacenarse en el laboratorio por largos períodos de tiempo, sin perder su capacidad de diferenciación. Estas células pueden permanecer auto-renovándose en el laboratorio por largos períodos, hasta que por medio de protocolos definidos, se les proporcionan las señales necesarias para que inicien el proceso de diferenciación hacia algún tipo celular particular. Existen varios tipos de protocolos, para inducir la diferenciación a diferentes tipos de tejidos, por ejemplo, cardíaco, sanguíneo, neuronal, epitelial, etc.

1.3.2. Aplicaciones de las células embrionarias

En el ámbito de la investigación básica estas células son ampliamente utilizadas para estudiar, sobre todo, el funcionamiento normal de la célula. A partir de células madre se pueden crear grandes cantidades de células especializadas, por ejemplo, motoneuronas, y estudiar en principio su funcionamiento normal: cómo funciona la motoneurona, cómo se transmiten las señales, qué genes están involucrados en qué función, etc. Lo mismo para cualquier tipo de célula que se conozca su protocolo de diferenciación. Actualmente, el estudio del funcionamiento normal de las células es su mayor campo de aplicación.
En un segundo plano, estas células se utilizaban también para estudiar modelos de enfermedad, aunque este campo está siendo absorbido actualmente por las células iPS (sección 1.4).
Los estudios de los modelos de enfermedad con embrionarias se basaban en conocer las modificaciones genéticas que causan una enfermedad particular, modificar el ADN de la célula madre para que represente el fenotipo de la misma y una vez diferenciada se obtiene una célula con el fenotipo deseado. De esta forma se pueden estudiar los mecanismos de varias enfermedades, su desarrollo, y se pueden plantear posibles tratamientos.

Muchos grupos de investigación y companías utilizaron este tipo de procedimientos para investigar enfermedades sin cura o tratamiento [13–15]. Por otro lado, muchas veces, las mutaciones específicas de la enfermedad a estudiar no son conocidas o no son fáciles de replicar en el genoma de la célula madre. Para este tipo de casos, existen otras técnicas que permiten incorporar el núcleo de una célula cualquiera de un paciente afectado (por ejemplo una célula de piel de un paciente con ALS) en un óvulo previamente enucleado. Mediante señales eléctricas, se puede lograr el comienzo de la diferenciación, sin necesidad de fecundación. Al quinto día de comenzada la diferenciación se forma un blastocito, que será la fuente de células madre embrionarias, con el fenotipo deseado. Este procedimiento se denomina *Transferencia Nuclear Celular* (TNC), y era mayormente utilizado en animales (de esta forma se clonó la oveja Dolly), ya que en humanos, hasta el momento, no se había logrado obtener células del blastocito generado. Existieron muchas controversias respecto a este tema, ya que en el 2004 y 2005 (como mencionamos previamente) hubo publicaciones alegando que un grupo de investigadores coreanos habían logrado desarrollar TNC en humanos. Finalmente se descubrió que los ensayos habían sido manipulados y las publicaciones fueron retractadas. Sin embargo, muy recientemente se publicó en la revista *Cell* la realización de este procedimiento en humanos [10]. Se esperan por lo tanto grandes avances en este tema en los próximos años.

En el ámbito de investigación médica mucho se está realizando con células madre embrionarias. Varias empresas de biotecnología están llevando a cabo ensayos clínicos, la mayoría aún en fase 1. A modo de ejemplo, la empresa biofarmacéutica californiana Geron está actualmente desarrollando 6 estudios clínicos, abordando los temas de lesiones de médula espinal, lesiones a nivel cardíaco (post-infarto), tratamientos para la diabetes tipo I, etc. Si bien alguno de los estudios no está enrolando a más pacientes (el de lesiones de médula según la empresa por motivos económicos), éstos serán controlados por los próximos 15 años. La información fue obtenida en setiembre 2013.

Si bien estas células presentan un increíble potencial no dejan de tener grandes riesgos a la hora de las aplicaciones clínicas. Las células embrionarias son altamente teratogénicas, es decir tienen tendencia a generar tumores. Por lo tanto si son utilizadas en ensayos clínicos, debe asegurarse una tasa de diferenciación altamente eficiente, para que no queden residuos de célula madre no diferenciada en el paciente. Además los métodos de purificación de las células diferenciadas deben de ser extremadamente seguros. Si existen remanentes de células embrionarias en el paciente, éste tendrá altas probabilidades de desarrollar tumores. Por estos motivos, actualmente, su foco principal de aplicación es el estudio del funcionamiento normal de la célula, mientras que los modelos de enfermedad y las aplicaciones clínicas se están estudiando más bien a través de iPS y células madre adultas.

1.3.3. Problemas éticos de las células embrionarias

La utilización de las células embrionarias para la investigación trae consigo varios problemas morales y éticos, ya que su uso implica la destrucción de un embrión. El conflicto ético radica mayormente en las diferentes creencias religiosas y filosóficas sobre la definición de qué es vida, cuándo comienza la vida humana, etc. Muchos grupos consideran el concepto de persona como un absoluto, es decir, el cigoto ya se considera vida humana, tiene alma y posee todos los derechos como ser humano. Otras personas sostienen que el proceso de convertirse en persona, es un contínuo, que no ocurre enseguida de la fecundación. El punto en el tiempo en donde el cigoto se convierte en persona depende de las diferentes religiones, creencias, filosofiías, etc.
A modo de ejemplo, el Vaticano proclama explícitamente en su declaración del 2008 (*Dignitas Personae*) su posición en contra de la investigación con células madre, por el concepto absoluto de vida humana. Sin embargo, no es una opinión general de todas las congregaciones católicas.
Por otro lado, existe a nivel de países, políticas y legislaciones que restringen la investigación con este tipo de células. Dentro de los países más conservadores en este ámbito se encuentran Alemania, Estados Unidos, Austria, Irlanda, Italia, Noruega y Polonia. Existen países con políticas más laxas, como por ejemplo, Australia, Bélgica, China, India, Israel, Japón, Corea del Sur, Suecia y Reino Unido.

1.4. Células iPS

Las células iPS (induced Pluripotent stem cells) son un tipo de células madre pluripotentes derivadas artificialmente de una célula adulta, ya diferenciada, que inicialmente no era pluripotente. Sobre la célula adulta procedente de un tejido (por ejemplo fibroblasto) se induce la expresión de varios genes exógenos, como ser Oct4, Sox2, c-Myc y Klf4, capaces de des-diferenciarla o reprogramarla. Como dice el nombre, las células resultantes de la reprogramación son pluripotentes, pudiendo dar lugar a varios tejidos específicos.
Las células iPS se derivan de células adultas por transferencia de varios genes exógenos (principalmente factores de transcripción) asociados a células madre embrionarias. Generalmente para una transferencia eficiente se utilizan retrovirus que actúan como vectores de los genes exógenos. Después de un tiempo determinado, un pequeño porcentaje de las células transfectadas comienzan a reprogramarse volviéndose morfológica y bioquímicamente similares a las células embrionarias. Las células iPS se aíslan por selección con un gen de resistencia a antibióticos y uno reportero.
Los primeros avances con las iPS surgieron en el grupo de Shinya Yamanaka en Japón

en el 2006 [5], el cual utilizó fibroblastos de ratón como célula adulta, un retrovirus como vector, el gen Fbx15 como reportero y cuatro factores de transcripción que resultaron esenciales para producir células iPS: Oct-3/4, Sox2, c-Myc, y Klf4. Sin embargo estas iPS presentaban patrones de metilación del ADN distintos a los de las células embrionarias y no producían quimeras viables. En el 2007 el mismo grupo publicó un trabajo junto con otros grupos de investigación independientes ([16, 17]), en donde se demostraba la obtención de iPS a partir de fibroblastos de ratón con capacidad de formar quimeras viables utilizando el gen Nanog en vez del Fbx15. Los patrones de metilación del ADN y la producción de ratones quiméricos viables indicaron que Nanog es un determinante importante de la pluripotencia celular [18]. En el mismo año se publicó simultáneamente por dos grupos de investigación la generación de iPS a partir de células adultas humanas (fibroblastos). Uno de los grupos utilizó como punto de partida los genes Oct3/4, Sox2, Klf4, y c-Myc y un retrovirus como vector [6], mientras que el otro utilizó los genes Oct4, Sox2, Nanog, y LIN28 con un vector de lentivirus [7].

Las aplicaciones clínicas en medicina regenerativa de estas células se ven parcialmente limitadas por la misma razón que las embrionarias: su tendencia a formar teratomas [19]. A su vez, varios de los genes que han mostrado capacidad de promover la formación de iPS están vinculados a cáncer de alguna manera. El c-Myc es incluso un oncogen, cuya omisión decrece la eficiencia de creación de iPS en un factor de 100.

Por estos motivos, las aplicaciones de las iPS se centran en estudiar modelos de enfermedad. Se pueden obtener células adultas de cualquier tejido de un paciente con una enfermedad particular y con las técnicas descritas anteriormente se pueden generar iPS con el fenotipo de interés. Éstas pueden diferenciarse y generar las células especializadas de interés. Análogo a como se procede en la Transferencia Nuclear Celular (sección 1.3.2).

A estas metodologías se le incorporan a su vez, ensayos de fármacos. Si se obtienen grandes cantidades de células con el fenotipo de alguna enfermedad, se puede analizar si fármacos potenciales tienen efecto sobre las células afectadas (por ejemplo en ALS las motoneuronas, en diabetes tipo 1 las células beta). De esta forma se realiza una preselección de los fármacos utilizando ensayos *in vitro* antes de realizar pruebas clínicas más costosas. Este tipo de encare puede aplicarse para cualquier enfermedad y cualquier tipo celular, siempre y cuando se conozca el protocolo de diferenciación de la iPS a la célula de interés.

Esfuerzos internacionales, como el proyecto StemBANCC (creado en Oxford en el 2012), están creando colecciones de líneas celulares de iPS de una variedad de enfermedades. Este proyecto financiado por varias empresas farmacéuticas y universidades tiene como meta crear 1500 lineas celulares de diferentes enfermedades para testear fármacos.

1.5. Características de las células madre adultas

Ya que las células madre embrionarias tienen un carácter controversial debido a sus problemas éticos y a su alta teratogenicidad, la comunidad científica se dedicó a investigar otro tipo de células madre con características similares. Las células madre adultas son multipotentes, por lo que pueden generar diferentes tipos celulares especializados (no todos). Si bien no tienen todas las cualidades y potencialidad de diferenciación de las células madre embrionarias, tienen un gran potencial terapéutico, no son éticamente problemáticas y son muy relevantes en el área de investigación médica y básica.

1.5.1. General

El término célula madre adulta es un poco confuso, ya que no sólo se encuentran en tejidos del cuerpo adulto, sino también en tejidos fetales. Por este motivo también se denominan células tejido-específico. Éstas, al igual que las embrionarias, tienen la capacidad de auto-renovarse y de diferenciarse. Sin embargo, en este caso, las células madre adultas pueden permanecer en auto-renovación por largos períodos de tiempo (a veces por la vida entera del organismo) y en el momento necesario se activa la diferenciación. Su función en el organismo, es asegurarse de abastecer al cuerpo de suficientes células especializadas. A modo de ejemplo, un adulto pierde a diario gran cantidad de células epiteliales, el organismo debe recuperar esas células perdidas, y lo hace mediante la diferenciación de células madre adultas epiteliales a células de la piel. Este tipo de célula madre adulta es particularmente muy activo, al igual que las células madre de la médula ósea, ya que abastecen al cuerpo de células sanguíneas y del sistema immune. La mayoría de las células madre tejido-específicas son menos activas que las anteriores, ya que la pérdida de las células especializadas es menor; los órganos como ser hígado, cerebro, entre otros, pierden menos cantidad de células diariamente.
Al igual que vimos anteriormente las células madre adultas se dividen de forma, tanto asimétrica como simétrica. En el caso de la asimétrica, una de las células hija permanece indiferenciada, y la otra comienza el proceso de diferenciación. Muchas veces, esta última, puede dividirse una o más veces, convirtiéndose ella misma en una "célula progenitora". Las divisiones adicionales de una célula progenitora permiten un "pool" de células especializadas más amplio y diverso, comparado con una única división inicial. Este es el caso de las células madre mesenquimales (las veremos a continuación) y las hematopoyéticas, que debido a una serie de divisiones adicionales dan origen a las células no sólo sanguíneas (eritrocitos, megacariocitos, monocitos, neutrófilos, etc.) sino también todas las del sistema inmune (NK, linfocitos-T, linfocitos-B, etc).

En general, las células madre adultas pueden clasificarse de acuerdo al tejido especializado que generan:

- Las mesenquimales (o estromales) generan en general células de tejidos de origen mesodérmico, como ser tejido óseo, cartilaginoso, adiposo, tendinoso.
- Las hematopoyéticas generan células sanguíneas y del sistema inmune
- Las precursoras endoteliales generan vasos sanguíneos
- Las cerebrales generan algunas neuronas del sistema olfatorio y algunas involucradas en la memoria y aprendizaje

A lo largo de los años se han descubierto más tipos de células tejido-específicas, a medida que se perfeccionaron las técnicas para encontrarlas, para extraerlas y para caracterizarlas. Se destaca que hay muy poca cantidad de células madre por tejido. En los últimos años se descubrieron células madre en el tejido cardíaco [20], pancreático [21] e intestinal [22], entre otros.
Muchas de estas células, por un lado, comparten su ubicación en el organismo y por otro, pueden aparecer en varios parénquimas diferentes. Por ejemplo, las células madres mesenquimales comparten su nicho con las células hematopeyéticas y endoteliales, en la médula ósea, y a su vez se encuentran también abundantemente en tejidos adiposos y en menos cantidad en tejido dental. En el cordón umbilical se encuentran también células madre mesenquimales (estromales), hematopoyéticas y precursores endoteliales.

1.5.2. Células madre adultas en el laboratorio

En general las células madre adultas no son tan fáciles de manipular en el laboratorio como las células madre embrionarias, ya que no tienen la capacidad de multiplicarse indefinidamente y generar mucha cantidad de células. Sobre todo las hematopoyéticas obtenidas de la médula o cordón umbilical son muy difíciles de multiplicar en el laboratorio. A su vez, estas células tejido-específicas no son fáciles de encontrar, ya que en general (hay excepciones) hay muy poca cantidad en el tejido, y se encuentran en lugares recónditos y muchas veces difíciles de acceder.
Aún cuando se conoce su ubicación en el tejido, es a veces difícil obtenerlas de un donante vivo. Por ejemplo, para obtener células madre adulta de intestino, de cerebro o de pulmón, habría que realizar un procedimiento quirúrgico invasivo en una persona sana.

1.5.3. Células madre mesenquimales (CMM)

En este trabajo nos enfocaremos en las células madre mesenquimales, también llamadas estromales. Si bien existen ambigüedades en la literatura en cuanto a la denominación y definición de estas células, en este trabajo decidimos utilizar preferentemente la primera denominación ("mesenquimales"). Como se mencionó anteriormente éstas pueden generar tejido óseo, cartilaginoso, tendinoso, adiposo, etc. La figura 1.2 muestra todos los procesos de diferenciación conocidos para las CMM. Estas células son fáciles de obtener, ya que son relativamente abundantes en el cuerpo, sobre todo las obtenidas del tejido adiposo. Dentro de las células madre adultas, las mesenquimales son unas de las más fáciles de manipular en el laboratorio.

Por este motivo y por su capacidad de generar varios tipos de tejido conectivo, el poten-

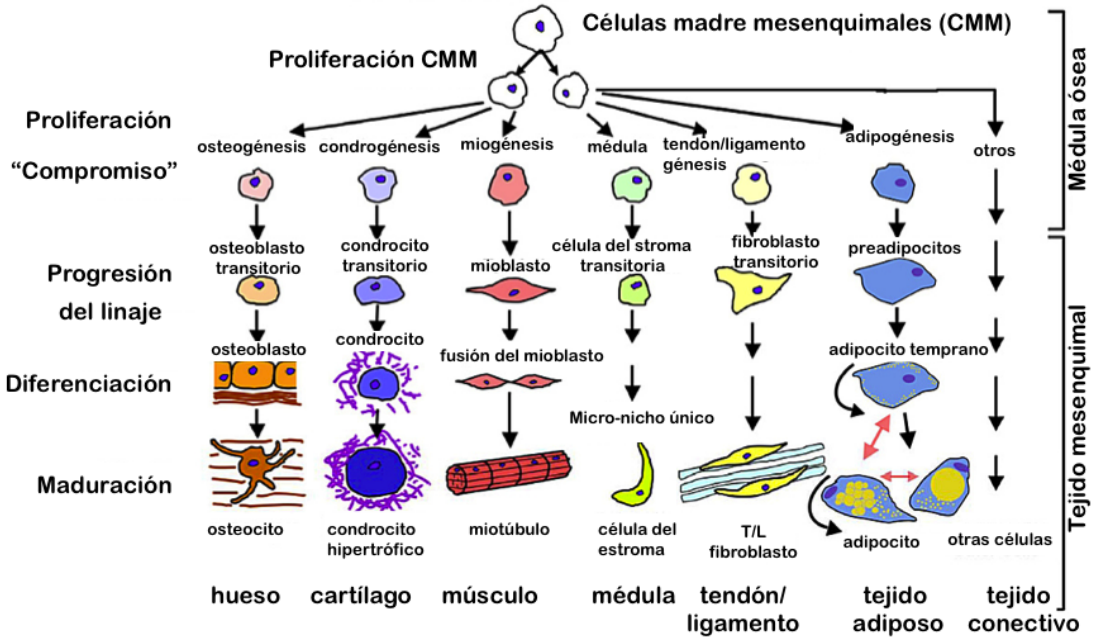


FIGURA 1.2: Diferenciación de células madre mesenquimales hacia todos los posibles tipos celulares especializados.

cial de aplicación de estas células es increíblemente amplio. Enfermedades como artritis [23] o daños de tendón [24], hueso [25] y ligamentos están siendo investigados con este tipo de células. Además, ya que su función natural es regenerar tejido conectivo se están investigando para generar injertos para diferentes tejidos.

Hasta fines del año 2012 se reportaron en el mundo 187 estudios clínicos en diferentes fases basados en CMM [26]. La mayoría se encuentra aún en fase I/II (93), pero ya 12 pasaron a la fase III. Este estudio clasifica también las diferentes enfermedades en las cuales se están focalizando estos ensayos clínicos. La mayoría de ellos se centran en

enfermedades del hueso/ligamento (37) y otros tantos en enfermedades cardíacas (27).
Si bien las células madre en general tienen un potencial increíble, las aplicaciones médicas basadas en células madre aprobadas por la FDA son sólo dos: el transplante de médula espinal para el tratamiento de leucemias (se aplica desde 1968), y los transplantes autólogos de piel para quemados. El resto de los tratamientos están en fases clínicas y no han sido aprobados hasta el día de la fecha.
Más recientemente se comenzó a tener en cuenta no sólo el potencial de diferenciación de estas células, sino que también su efecto homeostático, debido a su locación perivascular [27, 28] y su actividad parácrina [29]. La visión tradicional que se enfoca únicamente en la capacidad de diferenciación debe ampliarse para incluir el rol de modulador celular, el cual le proporciona a las CMM un escenario terapéutico aún mayor. Estas células han sido incluso denominadas "farmacias ambulantes" [30]. Estudios muestran que las CMM a raíz de una herida local se liberan de su locación perivascular, activándose y estableciendo un microambiente regenerativo mediante la secreción de moléculas bioactivas. Éstas son capaces de regular la respuesta inmune, a través de céluas dendríticas, células B y T, células killer y una variedad de células T-helper [31, 32], y de poseer un efecto trófico mediante moléculas secretadas que inhiben la apoptosis (especialmente causada por isquemias) y formación de cicatriz [33]. Por más sobre este tema referirse a los siguientes artículos [30, 34].

1.5.3.1. CMM derivadas del tejido adiposo

Si bien denominamos CMM a todos los tipos mostrados en la figura 1.2, cabe destacar que las definiciones aún son temas de discusión en la actualidad. Algunos científicos sólo consideran como CMM a las células madre que diferencian a médula ósea, descartando al resto, otros incluyen a todos los tipos.
Las CMM obtenidas de tejido adiposo (ADSCs: Adipose-derived stem cells), como también muchas de las CMM en general, se caracterizan por encontrarse en una población heterogénea de células, la cual puede contener fibroblastos, preadipocitas, macrófagos, células endoteliales, células diferenciadas, entre otras. Las ADSC pueden identificarse entre toda esa heterogeneidad por su perfil de antígenos de superficie muy similar al observado en las CMM de médula ósea. Sin embargo, las ADSC son más fáciles de obtener, ya que el tejido en donde se encuentran es más abundante y de fácil acceso.
ADSC pueden aislarse y diferenciarse *in vitro* en adipocitos maduros. Células que corresponden a ADSC contienen el siguiente perfil de antígenos: CD31−, CD34+, CD45−, CD90+, CD105−, CD146−. Esto implica por ejemplo que no poseen el marcador CD31 en la superficie, sí poseen el CD34, etc. Las células endoteliales por ejemplo, presentes

en general en la mezcla heterogénea de células, poseen la siguiente diferencia en fenotipo con las ADSC: CD34+/CD31+. Lo mismo ocurre con los macrófagos: CD34+/CD31+. En general, las células capaces de diferenciarse a adipocitos (preadipocitos, CMM) contienen el fenotipo: CD31−, CD34+, CD105− [35].
Estos fenotipos de antígenos pueden utilizarse para distinguir y separar las células ADSC en un citómetro de flujo.

1.6. Adipogénesis

Esta trabajo se basa en los estudios de diferenciación de células madre mesenquimales a adipocitos desde una perspectiva bioinformática. Aún en un marco *in silico* del problema, es necesario entender qué es lo que ocurre *in vivo* en la célula durante la adipogénesis. A continuación presentaremos muy brevemente los mecanismos básicos que llevan a la CMM a una célula adipocita y en el capítulo 3 se revisará la literatura actual para describir más en detalle los mecanismos de regulación génica y post-transcripcional relevantes en este proceso.
Adipogénesis es el proceso por el cual una CMM se diferencia a células adipocitas. Este mecanismo ocurre en dos etapas. La primera etapa se caracteriza por el compromiso de la célula CMM a un preadipocito y la segunda, precedida por una expansión clonal mitótica, es la diferenciación del preadipocito a la célula adipocita. Muchos eventos contribuyen al compromiso de una CMM al linaje de las células adipocitas, incluyendo la coordinación de una red compleja de factores de transcripción, cofactores y productos intermedios de señalización de numerosas vías.
Nuevas células grasas surgen constantemente de una población preexistente de células progenitoras indiferenciadas o de la desdiferenciación de células adipocitas a preadipocitas, y nueva diferenciación de las mismas a adipocitos. El análisis del recambio celular de los adipocitos ha demostrado que los mismos son una población dinámica y altamente regulada [36, 37]. La adipogénesis es un proceso de múltiples etapas que implica una cascada de factores de transcripción y proteínas del ciclo celular, que regulan la expresión génica y conducen finalmente al desarrollo de adipocitos. Varios reguladores positivos y negativos de esta red han sido dilucidados en los últimos años. A continuación nombraremos brevemente alguno de los procesos moleculares y celulares asociados a la adipogénesis, incluyendo factores de transcripción y cofactores.

1.6.1. Visión general de los procesos moleculares

El exceso de ingesta calórica sin gasto de energía promueve la hiperplasia adipogénica y la adiposidad. El aumento del número de adipocitos es activado por factores que

inducen la conversión de CMM a preadipocitos, los cuales diferencian posteriormente a adipocitos. Las CMM son reclutadas del estroma vascular del tejido adiposo y proveen de un recurso ilimitado de precursores de adipocitos. Los genes miembros de las familias BMP y WNT son mediadores claves del compromiso celular a preadipocitos. Estos últimos no continúan con el ciclo celular, el cual se detiene momentáneamente en la fase G1. Las CMM comprometidas al exponerse a los inductores de diferenciación, como pueden ser el factor de crecimiento insulínico 1 (IGF1), glucorcorticoides, AMP cíclico (cAMP), activan su replicación y el nuevo comienzo del ciclo celular. Este proceso se denomina expansión clonal mitótica, el cual incluye una cascada de factores de transcripción, seguido por la expresión de genes específicos de células adipocitas. Los pasos críticos en este proceso es la fosforilación del factor de transcripción CCATT enhacer binding protein β (C/EBPβ) por la MAP quinasa y GSK3β, lo que produce un cambio conformacional que le proporciona a C/EBPβ el potencial de unirse al ADN. La proteína C/EBPβ "activada" dispara a su vez la transcripción de la proteína PPARγ (peroxisome proliferator-activated receptor-γ) y C/EBPα, las cuales se coordinan para activar genes cuya expresión determina el fenotipo de adipocito.

1.6.2. Composición del tejido adiposo

El tejido adiposo se caracteriza por una marcada heterogeneidad celular: entre sus componentes celulares, se encuentran adipocitos, preadipocitos, fibroblastos, células endoteliales, las células madre mesenquimales capaces de diferenciarse en varios tipos celulares, células reguladoras T, macrófagos, etc. En general, el tejido adiposo está compuesto por aproximadamente sólo un tercio de adipocitos maduros, mientras que el resto es una combinación de los otros tipos celulares.

Existen adipocitos blancos y marrones, los cuales difieren en unas pocas propiedades significativas. Los adipocitos blancos poseen gotas lipídicas de gran tamaño, que ocupan la mayoría del volumen celular, dejando al citoplasma y núcleo en la periferia celular. Los marrones se caracterizan por las gotas lipídicas multiloculares y por alto contenido mitocondrial, los cuales se derivan de diferentes depósitos de tejido adiposo altamente vascularizado e inervado. Los adipocitos blancos, cuando son activados por hormonas lipolíticas, translocan sus lipasas citoplasmáticas sensibles a las hormonas (cytoplasmic hormone-sensitive lipase) [38] y las "adipocyte-triglyceride lipase" [39, 40] a la superficie de la gota grasa, donde ocurre la hidrólisis de los triglicéridos [41]. A su vez, varias proteínas accesorias se encuentran en la superficie de la gota lipídica, como ser perilipina, la que facilita el proceso de hidrólisis [42]. Por lo tanto, ácidos grasos derivados de triglicéridos son liberados al torrente sanguíneo para abastecer a tejidos de la periferia, especialmente al tejido esquelético, músculo cardíaco e hígado, con un combustible rico

en energía. Durante períodos de ingesta calórica excesiva, enzimas lipogénicas localizadas en el citoplasma y retículo endoplasmático sintetizan triglicéridos, los cuales son incorporados en la gota de grasa. Los genes que codifican para estas enzimas y otros genes adipogénicos son expresados en forma coordinada para producir el fenotipo de adipocito.

Comprender los mecanismos básicos de la diferenciación de CMM a adipocitos es el eje del presente trabajo, por lo que en el siguiente capítulo discutiremos las bases biológicas conocidas. Más adelante, en el capítulo 3, discutiremos un avance tecnológico cuyo impacto en la biología moderna promete ser una divisoria de aguas: las tecnologías de secuenciado masivo, o NGS por sus siglas en inglés (Next Generation Sequencing).

Capítulo 2

Regulación génica durante la adipogénesis en células madre mesenquimales

2.1. Introducción

En la sección 1.6 anterior se proporcionó un vistazo general de los integrantes fundamentales que participan en la adipogénesis. A continuación haremos una revisión bibliográfica sobre el estado actual del conocimiento de la regulación génica y las vía de señalización de las células madre mesenquimales que diferencian a adipocitos. Se detallarán las señales extracelulares, los factores de transcripción y genes relevantes. También se mencionarán brevemente ARN pequeños, como miRNAs. Los mismos, a la vez que los mecanismos de regulación post-transcripcional, serán detallados más adelante, en los capítulos introductorios a los trabajos realizados en la tesis.

2.2. Origen de los adipocitos

Los adipocitos derivan de células madre pluripotentes, específicamente las mesenquimales (CMM), las cuales pueden diferenciarse tanto en células adiposas, como miocitos, condrocitos, osteocitos, etc. [43–45]. Estas células, las cuales residen tanto en el estroma vascular del tejido adiposo como también en la médula ósea, al recibir las señales apropiadas sufren una serie de procesos que las llevan a comprometerse al linaje adipocito. El proceso del compromiso celular a este linaje lleva al surgimiento de las células preadipocitas, las cuales al inducirse entran en múltiples ciclos de mitosis, sufriendo un

proceso de expansión mitótica clonal. Luego de este mecanismo las mismas se diferencian a adipocitos maduros.

2.3. Compromiso: célula madre mesenquimal a preadipocitos

El primer paso de la diferenciación es el compromiso al linaje de adipocitos. Este reclutamiento de las CMM es impulsado por una ingesta elevada en energía y en la captación de glucosa [46] por largos períodos de tiempo. Este estado metabólico aparenta generar señales aún por identificar, que inducen a las CMM a entrar en la vía de compromiso al linaje adipocito generando el fenotipo de preadipocitos y de hiperplasia. Varios factores se han identificado que inhiben o impulsan el compromiso de CMM a preadipocitas. Estos incluyen genes pertenecientes a la familia del BMP (proteína morfogénica ósea), como ser el BMP4 y BMP2 [47, 48], WNT [49, 50] y Hh (hedgehog) [51–53]. BMP4 y BMP2 tienen un rol activador, mientras que la señalización vía Hh tiene función inhibitoria y WNT aparenta tener ambos roles, activación durante el compromiso de la CMM [50] e inhibición en la diferenciación a adipocitas [54].
Existe evidencia que indica que la determinación del linaje celular es regulado por una red de señales extracelulares, que finalmente afecta a los promotores de factores de transcripción específcos para el linaje. El equilibrio de estas moléculas de señalización es lo que determina el destino del compromiso celular, muchas veces simultáneamente promoviendo una vía e inhibiendo otra a la misma vez. Por ejemplo, Wnt10b promueve la osteogénesis y posiblemente la miogénesis, inhibiendo a la vez la adipogénesis [55]. BMP4 promueve la adipogénesis e inhibe la miogénesis [54]. PPARγ inhibe la condrogénesis y estimula la adipogénesis, mientras que Msx2 activa la osteogénesis e inhibe la adipogénisis reprimiendo la actividad transcripcional de PPARγ [56].

2.3.1. La vía de señalización BMP

Multiples señales pueden influenciar si una CMM se va a comprometer y generar preadipocitos, incluyendo factores extracelulares como ser BMP [44]. BMP4 y BMP2 están implicadas en el compromiso al linaje de adipocitos, por lo tanto si se expone una CMM a cualquiera de ambos factores los mismos generarán células preadipocitas, las cuales al ser tratadas con inductores en el momento de la detención del ciclo celular entrarán en la vía de diferenciación de adipogénesis, expresando los marcadores correspondientes y finalmente generarán el fenotipo de adipocitos [44, 47, 48]. Otros genes involucrados en

la vía del BMP, además de BMP4 y BMP2, son los receptores de los BMPs, BMPr2 y BMPr1a y los genes Smad-1, -5, -8. Los BMPs actúan mediante esos dos tipos de receptores (BMPr1 and BMPr2), los cuales forman complejos en la superficie celular con actividad de quinasa serina/treonina [57]. La unión de BMP al complejo de receptores BMPr1:BMPr2 induce la fosforilación, por lo que se activa la quinasa BMPr1. El receptor fosforila a Smad-1,-5,-8, que forma un complejo con Smad4. El mismo se trasloca al núcleo y regula expresión génica de varios genes relevantes para el compromiso celular que veremos a continuación (figura 2.1). La inhibición de la expresión del Smad4 con RNA de interferencia mostró la disrupción del proceso de compromiso celular [44]. Tres de los genes regulados por esta vía de BMP están asociados al citoesqueleto: Lox (lysyl oxidase), Tpt1 (translational controlled tumor protein 1) y αB cristalin [58]. Estudios de proteómica encontraron 8 proteínas sobreexpresadas por el efecto de BMP2 y 27 por el efecto de BMP4. 5 proteínas eran comunes a ambos BMPs y presentaban una sobreexpresión mayor a 10 veces más con respecto al control. Entre ellas se encuentran las tres proteínas del citoesqueleto anteriormente nombradas. El compromiso celular es totalmente bloqueado si se elimina el efecto del gen Lox (knockout) y es parcialmente bloqueado si se eliminan las otras dos. Los mayores cambios morfológicos celulares (las células cambian a una forma redondeada) se observan durante esta etapa del compromiso celular. La inhibición de la expresión de estos genes de citoesqueleto previenen el desarrollo de los cambios morfológicos, regenerando la organización de actina en fibras, impidiendo el compromiso de la célula a adipocitos. Estas proteínas diferencialmente reguladas determinan la habilidad de la CMM de comprometerse a la adipogénesis por medio de la regulación de la morfología celular.

2.3.2. La vía de señalización Wnt

La familia de Wnt es comprendida por glicoproteínas secretadas cuyos efectos están mediados por su receptor ("frizzled receptor") y un correceptor [59]. La proteína Wnt puede actuar mediante la vía de señalización canónica [60] y mediante la no-canónica [61]. El rol de la vía de Wnt en la adipogénesis se descubrió mediante el análisis de la expresión del gen Wnt10b. La misma desciende dramáticamente durante la diferenciación y consistentemente, la sobreexpresión forzada del gen impide la diferenciación [49], bloqueando factores de transcripción claves de la adipogénesis: PPARγ y C/EBPα. Wnt tiene un efecto dual sobre el proceso de la diferenciación. En las etapas iniciales funciona como activador [50] y en las etapas más tardías (diferenciación en sí) actúa como inhibidor (figura 2.1) [54], posiblemente mediante el efecto de otras proteínas de la familia Wnt. La vía canónica de Wnt, que funciona en los pasos iniciales del compromiso celular, incluye

la acción de la β-catenina citosólica, la cual está embebida en un complejo de "destrucción" que incluye entre otras proteínas a GSK-3β (glycogen synthase kinase-3β) [60]. En ausencia de la estimulación de Wnt, GSK-3β fosforila a la β-catenina, marcándola para ubiquitinación y degradación por el proteasoma. Por el contrario, la activación de Wnt a través de la unión con su receptor y coreceptor promueve la disociación del complejo de destrucción, permitiendo acumulación de β-catenina y posterior traslocación al núcleo [62, 63]. La acumulación de la misma en el núcleo activa la transcripción de los factores de transcripción Tcfs (T-cell factors), expresando en consecuencia varios genes relevantes como ser c-myc, Cyclina D1, etc [64] (figura 2.1).
En las etapas tardías de la diferenciacón, la vía canónica de Wnt regula el balance entre miogénesis, osteogénesis y adipogénesis, resultando en un descenso de la adipogénesis [49, 55]. Estudios observaron que la vía de la catenina puede promover tanto la miogénesis [49], como la osteogénesis [55], mientras que inhibe la diferenciación de los preadipocitos a los adipocitos [50]. Consistentemente con estos resultados, la activación de la vía Wnt promueve la miogénesis e inhibe la adipogénesis de CMM en cultivo [49, 65].

2.3.3. Las vías de señalización de Hedgehog y Rb

Se identificaron 3 ligandos de Hh en vertebrados, Sonic (Shh), Indian (Ihh) y Desert (Dhh), los cuales inician una cascada de señalización mediante receptores "Patched" (Ptch-1 y Ptch-2) [66]. La presencia de los ligandos activa a una proteína de membrana, Smo, la cual transmite la señal mediante fosforilación y traslocación nuclear de GliA [66]. La señalización vía Hh tiene un efecto inhibidor de la adipogenesis en ratón [67]. Si bien se observa una relación entre Hedgehog y adipogénesis, los mecanismos exactos aún no han sido dilucidados.
Por otra parte, la proteína del retinoblastoma Rb inhibe el ciclo celular, uniéndose y reprimiendo la actividad de E2F [68]. Frente a la hiperfosforilación de Rb (pRb), la proteína E2F se libera y promueve la transcipción de genes involucrados en el ciclo celular, sobre todo genes reguladores de la entrada a la fase S del ciclo y de progresión del mismo [69]. Estos eventos son críticos para la expansión mitótica clonal, un paso obligatorio en la diferenciación. Adicionalmente, la Rb fosforilada regula varios factores de transcripción claves para la inducción de la diferenciación [70, 71]. Dependiendo del factor de transcripción y su contexto celular pRb puede inhibir o promover la actividad de esos factores. Por ejemplo, pRb se une a Runx2 potenciando su actividad de promover la osteogenésis [72], por otro lado pRB se une a E2F para suprimir una subunidad

de PPARγ, el activador central de la adipogénesis [73, 74]. Ya que ambos eventos, osteogénesis y adipogénsis, surgen de las CMM, se ha propuesto a pRb como el factor que decide el destino de las CMM (osteocitos o adipocitos).

2.4. Diferenciación: preadipocitos a adipocitos

Una vez generados los preadipocitos, los mismos pasan por un proceso de expansión mitótica clonal, para luego continuar finalmente con la diferenciación, generando así células adipocitas.

2.4.1. Inducción de la diferenciación

Existen varios protocolos para inducir *in vitro* la diferenciación, partiendo de preadipocitos con ciclo celular detenido en G1. Alcanza con proporcionarle al medio un cocktail de inductores de estas células, que incluya alto contenido de IGF1 (factor insulínicio de crecimiento 1), dexametasona y algún agente que eleve el cAMP celular en el medio [75]. Estos inductores activan las vías de IGF1, glucocorticoides y la del cAMP, respectivamente. La inducción inicia una serie de eventos que regulan el programa de diferenciación. De 16 a 20 horas posterior a la inducción los preadipocitos sincrónicamente retoman el ciclo celular [75, 76] y sufren varias rondas de mitosis seguidas. Los mismos abandonan después el ciclo celular, pierden su morfología fibrobástica, adoptando una forma globular, acumulan triglicéridos en el citoplasma y obtienen el fenotipo y características metabólicas de los adipocitos [75]. Esta acumulación de triglicéridos está altamente correlacionada con el aumento de la tasa de lipogénesis y con la sobreexpresión de enzimas de la biosíntesis de ácidos grasos y triacilglicerol [77]. Asimismo, numerosas proteínas regulatorias características de adipocitos son expresadas, como ser receptores de insulina [78], el transportador de glucosa GLUT4 [79], leptina [80], entre otros.

2.4.2. Expansión mitótica clonal

Los preadipocitos, previamente detenidos en la fase G1 del ciclo celular, sufren dos rondas de divisón celular, denominadas expansión mitótica clonal. Este paso es necesario para la diferenciación, por lo que bloquear la divisón celular impide la diferenciación. Por ejemplo la sobreexpresión del inhibidor del ciclo celular p27 impide a la célula entrar en la fase S1, por lo que interrumpe los pasos subsecuentes de la diferenciación [81].

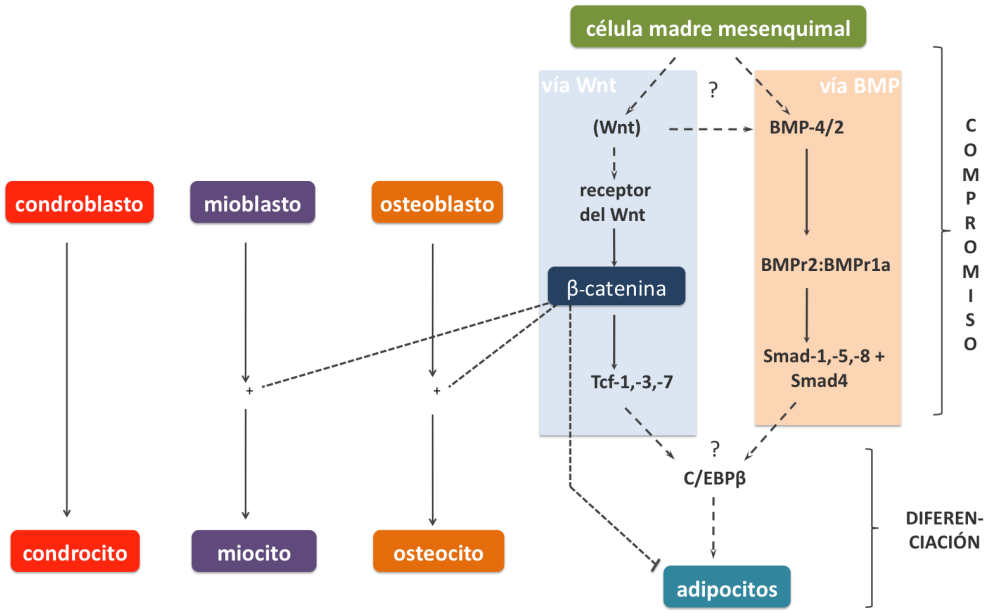


FIGURA 2.1: Esquema de eventos que llevan al compromiso de la CMM al linaje de adipocitos. Se muestran ambas vías de señalización, la del Wnt y del BMP. Wnt aparenta funcionar como activador e inhibidor en el compromiso de las CMM y la diferenciación. Las líneas punteadas indican incertidumbre. Figura basada en la revisión [1]

Inmediatamente (menos de 5 minutos) después de la inducción CREB (cAMP response element-binding protein) es fosforilada y activa la expresión de C/EBPβ [82]. Sin embargo, C/EBPβ aún no tiene capacidad de unirse al ADN. De 14 a 16 horas post inducción C/EBPβ adquiere su capacidad de unirse al ADN mediante fosforilación por parte de GSK3β, a la vez que la célula retoma el ciclo celular. A las 16-20 horas de la inducción el preadipocito entra en la fase S del ciclo. Estos eventos están relacionados con la expresión de la histona H4, la cual aumenta dramáticamente previo a la entrada a la fase S [83]. La inactivación de C/EBPβ impide la sobre expresión de H4 y detiene al ciclo celular en la fase G1 [83].

2.4.3. Factores de transcripción relevantes en la diferenciación

La inducción de la diferenciación causa el rápido aumento (en menos de 4 horas) en la expresión de las proteínas C/EBPs, sobre todo C/EBPβ y C/EBPδ. Las mismas adquieren recién a las 16 horas su capacidad de unión al ADN. Si bien existe cierta redundancia en la actividad de estos genes [84], únicamente C/EBPβ juega un rol central en la expansión mitótica clonal, ya que la inactivación de la misma impide este proceso y por lo tanto la diferenciación posterior [82]. La obtención de la capacidad de unión al ADN es mediante fosforilación secuencial. Primeramente, a las 2 horas de la inducción, una MAP quinasa fosforila la treonina 188. Más tarde (a las 16 horas)

GSK3β fosforila la treonina 179 o la serina 184. Para activar correctamente a C/EBPβ ambas fosforilaciones son necesarias [85]. Entre los blancos de los factores de transcripción C/EBPβ y C/EBPδ se encuentran los promotores de los genes C/EBPα y PPARγ (ambos son factores de transcripción relevantes en el programa de diferenciación) y el regulador de genes lipogénicos SREBP1 (sterol-regulatory-element-binding protein 1) [86, 87]. PPARγ activa el promotor del gen C/EBPα y viceversa, generando un bucle de retroalimentación positivo. A su vez, PPARγ y C/EBPα inducen la expresión de varios genes claves como ser aquellos involucrados en la sensibilidad a la insulina, lipogénesis y lipólisis, incluyendo entre ellos al GLUT4, FABP4 (fatty-acid-binding protein 4), LPL (lipasa lipoproteíca), AGPAT2 (sn-1-acylglycerol-3-phosphate acyltransferase 2), perilipina y los factores secretados adiponectina y leptina. Estudios genómicos recientes han demostrado la cooperación de PPARγ y C/EBPα en múltiples sitios de unión en regiones promotoras de varios genes, corregulando conjuntamente una gran gama de genes involucrados en el desarrollo y maduración de los adipocitos [88, 89]. Varios factores corregulan esta red central de transcripción, como ser STAT5, C/EBP proteína homóloga 10 (CHOP10) y proteínas miembros de la familia Krüppel-like factor (KLF) family [87]. Dentro de los reguladores positivos se incluyen EGR2 (early growth response-2) [90], EBF1 (early B cell factor-1) [91], KLF4 [92], entre otras. Efectos de inhibición se han descrito para las siguientes proteínas: FOXC2 (forkhead box protein C2) [93], ATO (eight-twenty-one) [94], GATA2 y GATA3 (globin transcription factors 2 y 3)[95], KLF3 [96], CTBP1 and CTBP2 (C-terminal-binding proteins 1 y 2) [97] y los factores de regulación de interferona IRF3 y IRF4 [98].

Algunos de los factores parecen tener doble funcionalidad, dependiendo el contexto actúan como promotores y o como inhibidores de la adipogénesis. Por ejemplo, estudios han demostrado que el factor de transcripción COUP-TFII (orphan nuclear receptor chicken ovalbumin upstream promoter-transcription factor II) promueve el compromiso de las células preadipocitas a células adipocitas inhibiendo la vía de señalización del gen WNT [99]. Sin embargo otros estudios han reportado que el mismo puede actuar como inhibidor de la adipogénesis reprimiendo la expresión de C/EBPα y PPARγ [100].

2.4.4. PPARγ como gen crucial

PPARγ continúa siendo siendo foco de estudio debido a su rol crucial en la adipogénesis. Muchos de los factores que influyen en la adipogénesis lo hacen mediante el efecto de este gen, ya sea de forma directa o indirecta. Por ejemplo, la sirtuína (SIRT2) inhibe la expresión de PPARγ de forma indirecta, mediante la reducción de la acetilación y fosforilación del gen FOXO1 (forkhead box O1). Esto lleva a un aumento de FOXO1 en

el núcleo, donde el mismo reprime la transcripción del gen PPARγ [101]. Otra sirtuína (SIRT1) reprime la adipogénesis actuando como co-represor directamente sobre PPARγ [102]. Varias especies de lípidos han demostrado activar al gen PPARγ actuando como ligandos endógenos [103, 104]. Sin embargo, se ha reportado recientemente que el ácido fosfatídico cíclico (cPA) puede actuar como un lípido inhibidor que se une a PPARγ y estabiliza la asociación del mismo con NCOR2, su co-represor nuclear (nuclear receptor co-repressor 2, conocido también como SMRT) [105].
La fosforilación de PPARγ proporciona otro nivel de regulación. El submódulo de quinasa del factor de transcripción general IIH (conocido también como TFIIH), fosforila a PPARγ en la serina 112 [106], la cual inhibe la función del mismo por varios mecanismos. Por un lado, se impide la reclusión de los co-activadores y por otro lado, se aumenta la unión del PPARγ con PER2 (regulador del período circadiano), el cual inhibe la unión del primero con los promotores de genes blanco relevantes [107].

2.4.5. Rol de miARNs en la adipogénesis

Se han descrito una serie de miARNs (micro ARN) que están involucrados en la diferenciación a adipogénesis. Algunos de ellos juegan un rol acelerador en el proceso [108, 109], mientras que otros aparentan inhibir el mismo [110, 111]. Por ejemplo, el miR-8 promueve la adipogénesis en la medida que inhibe la vía de Wnt [108]. Estudios observaron que la introducción de let-7 en células preadipocitas murinas (3T3-L1) inhibe la expansión clonal, y por lo tanto la diferenciación [111]. Sin embargo, no se ha podido dilucidar los mecanismos exactos de acción de estos miARNs.
Más recientemente, Zhang y colaboradores (2013) resumieron todos los miARNs involucrados en la adipogénesis conocidos hasta el momento, tanto predichos como validados, encontrando así 14 miARNs relevantes en el proceso de diferenciación [112]. En los capítulos siguientes, introductorios a cada uno de los trabajos realizados en la tesis, detallaremos más sobre los miARNs, mecanismos de acción, rol en la regulación post-transcripcional, etc.

Capítulo 3

Contexto tecnológico: secuenciación masiva

3.1. Introducción

El progreso de la ciencia va a la par del surgimiento de tecnologías revolucionarias que proveen nuevas formas de ver problemas científicos, permitiendo formular preguntas innovadoras y generar conocimiento avanzado. Así como primeramente el microscopio óptico (al rededor del 1700) y luego el electrónico (1925-1930) revolucionaron la forma de ver a la biología en su época y la PCR (desarrolada en 1983 por Kary Mullis) permitió increíbles avances en la biología molecular, el surgimiento de la secuenciación masiva revolucionó incontables áreas de la biología: la genética, genómica, transcriptómica, evolución, ecología, virología, entre tantas otras. Las posibilidades para estas metodologías son inmensas y están sólo limitadas por nuestra imaginación científica, y en muchos casos, más bien por los recursos económicos limitados. Con el paso del tiempo, las tecnologías irán bajando los costos y llegarán al alcance de más grupos científicos. Sin embargo, algunas problemáticas se mantienen hasta el día de hoy, por ejemplo, cómo almacenar y analizar inmensas cantidades de datos. Estos problemas también se irán disipando a medida que vayan madurando estas tecnologías. Estamos frente a una nueva era de la biología y medicina que nos proporcionará nuevos desafíos tanto en el área de diagnóstico médico como de investagión básica.
Durante el presente trabajo utilizamos la tecnología del SOLiD como secuenciado masivo. Sin embargo, para mejor entendimiento de la misma, se mostrarán varias de las tecnologías disponibles hoy en días con sus ventajas y sus fallas. Se culminará con un desarrollo más detallado de la tecnología utilizada.

3.2. Algunas de las tecnologías disponibles

Cada vez surgen más tecnologías de secuenciado masivo diferentes. Las mismas se diferencian en las etapas de amplificación clonal y secuenciamiento. Estas tecnologías están sujetas a contínuas modificaciones y mejoras, manteniendo siempre los mismos principios básicos.

3.2.1. Secuenciación con 454 (Roche)

La solución de interés de ADN doble hebra puede ser fragmentada aleatoriamente, o alternativamente, amplificada vía PCR. Los fragmentos resultantes de tamaño adecuado son desfosforilados y consecuentemente ligados a dos adaptadores (A y B). Sólo se amplifican fragmentos que contienen adaptador A en un extremo y en el B del otro. Estos fragmentos constituyen la biblioteca de ADN, los cuales son unidos a "microbeads" (bolitas microscópicas) a través de hibridación de primers. Las condiciones de esta reacción están dadas de forma tal de favorecer la unión de un solo fragmento de ADN (de la biblioteca) por bolita (figura 3.1 (a)). Cada "microbead" cubierta con una única molécula se coloca en una emulsión de agua y aceite, la cual actúa como microrreactor en donde se encuentra, además de esa única bolita, primers correspondientes y reactivos para PCR (figura 3.1 (b)). La amplificación por PCR dentro de la emulsión permite cubrir la totalidad de la bolita con moléculas clonales amplificadas. Cada bolita, cada microrreactor, contiene un fragmento de ADN particular (figura 3.1 (c)). Después de la amplificación se diluye la emulsión y las bolitas son enriquecidas con streptavidina. Utilizando separación magnética cada bolita se coloca en un microrreactor (pocillos dentro de una "picotitre plate"). Los fragmentos amplificados son desnaturalizados y unidos a un primer de secuenciación. La secuenciación ocurre a través del la técnica del *pirosecuenciado*, en donde la incorporación de cada nucleótido conlleva la eliminación de un fosfato. El mismo es convertido en luz por medio de ATP, y la cantidad de luz es proporcional a la cantidad de bases incorporadas.
La mayor ventaja de este método es el largo de sus "reads" (fragmento secuenciado), ya que logra tamaños de más de 500 bases. Por este motivo, la utilización de esta técnica radica mayormente en el secuenciamiento de amplicones, secuenciamiento de puentes entre "scaffolds" (porciones de genoma ya secuenciados, que aún no fueron posicionadas en el genoma total), secuenciamiento de puentes entre secuencias complejas, etc. La mayor desventaja del método se debe al pirosecuenciado, el cual tiene dificultades para secuenciar homopolímeros, llevando la tasa de error por read hasta un máximo de 1 %. Estos errores pueden mitigarse parcialmente con mayor cobertura de secuenciamiento (más cantidad de reads por base).

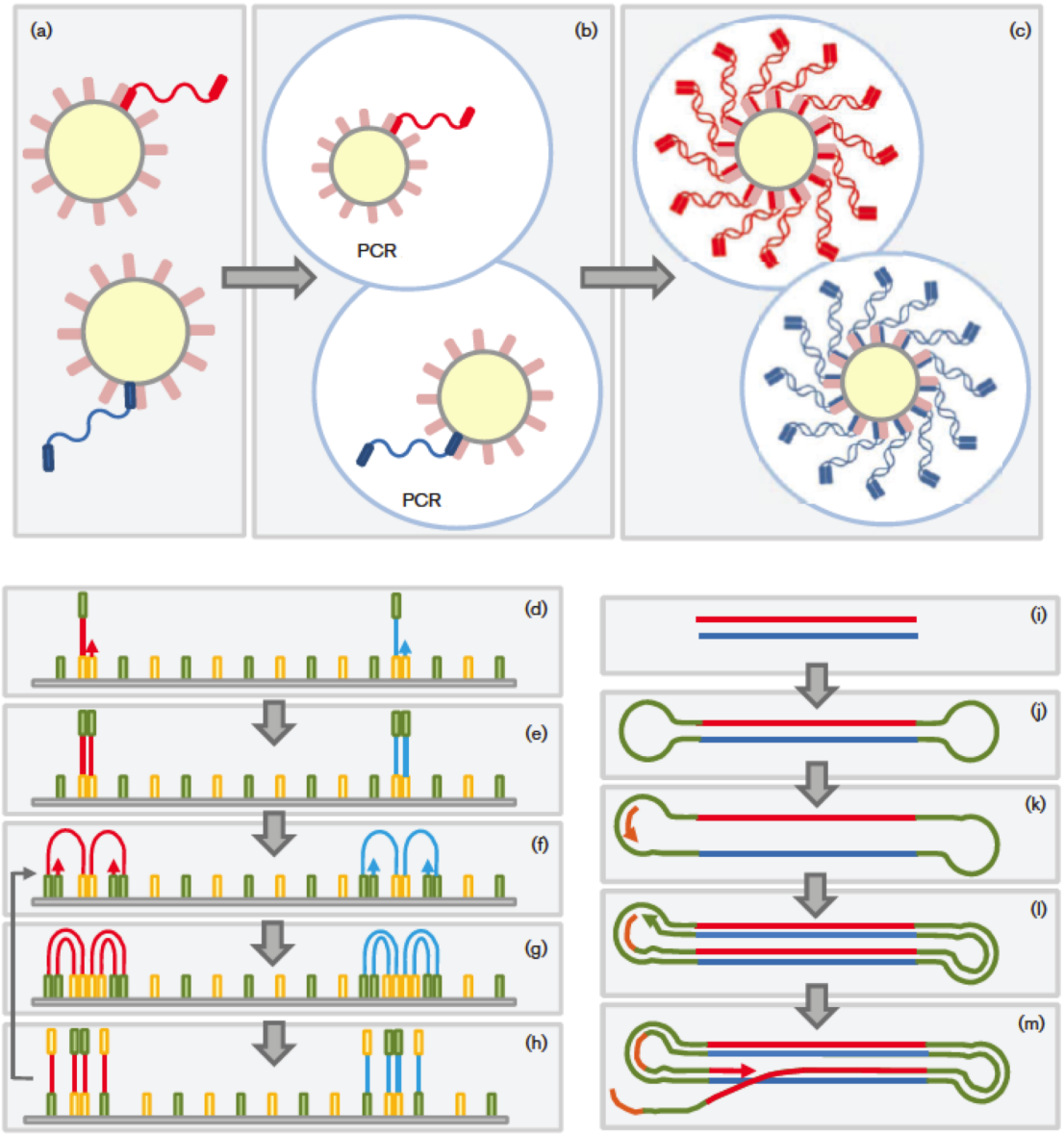


FIGURA 3.1: Principios generales de la amplificación del ADN template. (a-c) PCR de emulsión (esto es válido para 454, SOLiD y Ion Torrent). (a) Los adaptadores se utilizan para capturar moléculas de ADN template y unirlas a las "microbeads" (bolitas) via hibridación de primers. (b) Cada bolita se coloca en una emulsión controlada. (c) Amplificación del ADN por PCR dentro de la emulsión. (d-h) Amplificación en puente (válido para Illumnia). (d) Template se liga a la placa de vidrio mediante hibridización de adaptadores/primer. (e) Adaptador fijado en el vidrio funciona como primer, base de la extensión. (f) El extremo libre de cada molécula puede unirse a un segundo adaptador fijo formando un puente, que actúa como molde para una (g) segunda ronda de amplificación. (h) El resultado es cuatro moléculas lineales. Los pasos (f-h) se repiten para generar varios clusters densos de secuencias amplificadas, que se secuenciarán posteriormente. (i-m) Amplificación lineal de PacBio. Figura obtenida de [2]

3.2.2. Secuenciación de Illumina

La muestra de ADN doble hebra es fragmentada, purificada, ligada con adaptadores en ambos extremos, y desnaturalizada, resultando en un fragmento de hebra simple con dos adaptadores en los extremos. Illumina desarrolló una estructura especial para su tecnología, una "flow cell" con varios carriles (típicamente 8), capaz de facilitarle el acceso a las enzimas para reaccionar con el ADN, y a al mismo tiempo asegurando estabilidad en la unión del ADN con la superficie fija. Los fragmentos de hebra simple generados son adheridos a una placa de vidrio, que corresponde al interior de los canales de la flow cell. Esta placa contiene en su superficie oligonucleótidos que son complementarios a los adaptadores (figura 3.1 (d)). Estos oligos cumplen tanto la función de unir los fragmentos de ADN a la placa, como también la función de primers para la siguiente amplificación de los mismos (figura 3.1 (e)). La misma ocurre en la placa de vidrio por medio de una *amplificación en puente*: cada molécula tiene un extremo fijado en el vidrio y otro extremo libre. Este último puede unirse a otro adaptador fijo sobre el vidrio que esté espacialmente cerca. Este adaptador fijo funciona como primer para el próximo paso de amplificación (figura 3.1 (f)). Se incorporan nucleótidos (no marcados) y enzimas para realizar varios ciclos de PCR, generando "clusters" (agrupamientos) de secuencias de simple hebra amplificadas (figura 3.1 (g-h)). Los clusters generados, sirven como punto de partida para la secuenciación de Illumina, que está basada en el principio de *secuenciación por síntesis*, similar al tradicional secuenciado de Sanger. Sin embargo, esta técnica utiliza nucleótidos marcados con fluorescencia reversiblemente, de forma tal que cuando la enzima incorpora uno de estos nucleótidos, la reacción termina momentáneamente y con la excitación de un laser se emite una luz, correspondiente al nucleótido incorporado. Esta reacción, a su vez, elimina la marca fluorescente, y permite que continúe la síntesis. Se utilizan imágenes de alta calidad para determinar qué nucleótido se unió en dónde en cada ciclo y así obtener la secuencia de bases de los fragmentos.
Durante cada ciclo de secuenciación, todos los nucleótidos están disponibles (al igual que enzimas, etc), permitiendo una competencia natural entre ellos, simulando el comportamiento real dentro de la célula.
Los largos obtenidos con esta tecnología rondan los $\sim 50 - 250$ bases, dependiendo del secuenciador utilizado. Illumina posee varios secuenciadores, desde el potente HiSeq2500 (un máximo de 3 billones de reads, "single end"), hasta el MiSeq caracterizado por su rapidez (pero con un máximo de 15 millones de reads, "single end"). Los reads con el MiSeq llegan a las 250 bases, mientras que los de HiSeq (en los varios modelos disponibles) llegan a un largo de $100 - 150$.
Una de las desventajas de esta tecnología es su alto costo y (a no ser el MiSeq) el tiempo

necesario para realizar una corrida, que puede llegar hasta 11 días[1]. Sus aplicaciones son múltiples, desde transcriptómica, genómica, mutaciones, epigenética, etc.

3.2.3. Secuenciación con Ion Torrent (Life Technologies)

La secuenciación del Ion Torrent pertenece a las tecnologías de tercera generación, surgiendo en el 2010. Como en las tecnologías anteriores primeramente se fragmenta la muestra de interés y se agregan los adaptadores de secuencia conocida. Los fragmentos son amplificados mediante PCR de emulsión (figura 3.1 (a-c)), como vimos anteriormente. Para ello, se utilizan también "microbeads". La biblioteca amplificada es separada en una placa con alta densidad de pocillos, como en el 454: una única bolita por pocillo. En el Ion Torrent, estos pocillos están ubicados sobre un semiconductor sensible a los iones. Durante la extensión catalizada por la polimerasa, un enlace de hidrógeno se libera como parte de la química normal de incorporación de nucleótidos. Este ión es detectado por el semiconductor, como un pequeño cambio en el pH. Por lo tanto, a cada pocillo con la molécula de ADN y la ADN polimerasa se le provee de a un nucleótido por vez. Si el mismo, es complementario al nucleótido siguiente no pareado, el mismo se incorpora y un hidrógeno se libera, generando un cambio de pH que es registrado por el semiconductor. Si el nucleótido no es complementario, la reacción química no tiene lugar y se lava el remanente. Se pasa al siguiente nucleótido y este procedimiento se repite tantos ciclos como sea necesario.
Esta tecnología, al igual que el 454, es sensible a los homopolímeros. Si en la secuencia aparece una serie de nucleótidos repetidos, la incorporación de este nucleótido llevará a un cambio de pH que será proporcional al cambio individual de pH. Sin embargo, ya que cada medida individual tiene incertidumbre, muchas veces es difícil estimar la multiplicidad de la base de forma precisa. Por lo tanto, existe una tendencia a generar errores en la longitud de los homopolímeros, lo que se traduce en la consideración de deleciones e inserciones inexistentes.
El tamaño de los reads alcanza ~ 100 nucleótidos, la precisión para los homopolímeros de largo 5 es de 98 % y la precisión general es de 99,6 % para reads de largo 50. La ventaja de esta tecnología es la velocidad y el relativo bajo costo de la misma.

3.2.4. Secuenciación con PacBio (Pacific Biosciences)

A diferencia de las tecnologías anteriores, PacBio fija a la ADN polimerasa (no al fragmento) al fondo de los pocillos o micro celdas (denominada "Zero Mode Waveguides"

[1] Los datos anteriormente nombrados fueron obtenidos de la página de Illumina accedida el 16 de julio del 2013 (`http://www.illumina.com/systems/sequencing.ilmn`)

o ZMW en este caso) de la placa de vidrio. El ADN es fragmentado, ligado a adaptadores del tipo "hairpin", lo que quiere decir que tienen estructura de horquilla y unen ambas hebras (figura 3.1 (i-j)), generando un ADN circular de hebra simple. Un primer complementario al adaptador horquilla se une al adaptador y comienza la amplificación del fragmento. Se realizan varias copias de hebra simple de ambas hebras, la codificante y reversa (figura 3.1 (l-m)). Los fragmentos amplificados son capturados por la ADN polimerasa, fijada en la parte inferior de la micro celda (un único fragmento por micro celda). Los mismos serán secuenciados con la siguiente estrategia denominada SMRT (single molecule real time). Nucleótidos marcados con fluorocromo (a cada uno se le asigna un color) difunden al interior de la micro celda desde arriba. A diferencia de otras tecnologías, los nucleótidos contienen una molécula fluorescente enlazada al fósforo, lo que significa que el nucleótido al ser incorporado en la cadena creciente, escinde ese fósforo con el fluorocromo. Los nucleótidos marcados que difunden cerca de la polimerasa generan cierta señal de ruido, sin embargo, cuando se incorpora el nucleótido complementario a la cadena creciente, éste se mantiene mucho más tiempo en el sitio activo de la polimerasa que los nucleótidos difundidos. Esto genera una señal de color clara (muy por encima del ruido) que es captada por un detector. Después de la incorporación el fluorocromo se libera y difunde lejos del sitio activo, permitiendo a la ADN polimerasa que siga su trabajo.
Esta tecnología se denomina de tercera generación y posee ciertas ventajas frente a las de segunda generación. Por un lado la preparación de bibliotecas es muy rápida, puede durar de 4 a 6 horas en vez de días. Por otro lado, no es necesario el paso de amplificación por PCR, evitando los sesgos y errores causados por PCR. A su vez, las corridas pueden ser extremadamente rápidas, desde 30 minutos (reportado por Pacific Bioscience) hasta como máximo menos de 1 día. Esta tecnología genera los reads más largos con un promedio de 2kb y con un 5 % de reads de largo $\sim$ 4kb. Se han alcanzado largos de hasta 10kb. Algunas de las desventajas incluyen el alto costo del equipamiento y la relativa baja cantidad de secuencias en comparación con los equipos de segunda generación. Este sistema es muy útil para laboratorios clínicos, especialmente de microbiología.

3.3. Secuenciación con SOLiD (Life Technologies)

Posterior al proyecto del genoma humano (2001), 454 lanza el secuenciador masivo en el 2005, un año después Solexa libera al mercado el Genome Analyzer, seguido por el SOLiD (Sequencing by Oligo Ligation Detection) en ese momento provisto por la empresa Agencourt. Estas compañías fundadoras fueron adquiridas por otras, de forma tal que hoy en día el SOLiD es producido por Life Technologies. Dada la importancia que

esta tecnología tiene en nuestro trabajo, a continuación estudiaremos en forma detallada la química del SOLiD, las ventajas y desventajas de esta plataforma y sus aplicaciones.

3.3.1. Preparación de bibliotecas

SOLiD respalda dos tipos de bibliotecas, las "single-end" o "mate-pairs" (los reads resultantes son "pair-end"). En las primeras, las bibliotecas se componen simplemente de fragmentos de ADN, que serán secuenciados desde el principio, sólo de un lado. Las últimas, implican un fragmento de ADN que será secuenciado de ambos extremos, por lo que se obtiene información posicional, lo que permite una mayor precisión a la hora de ubicar los fragmentos en el genoma. Nos enfocaremos en la construcción de la biblioteca de fragmentos "single-end", ya que es la que utilizamos a lo largo del trabajo. De cualquier manera las fases iniciales de la construcción de la biblioteca son similares.
La muestra de ADN es fragmentada mediante sonicación o disrupción física. Se eligen los fragmentos de tamaños adecuados, que para este tipo de biblioteca pueden ir desde desde 60 a más de 200 bases, dependiendo del protocolo específico (el procedimiento de selección de tamaños puede realizarse en geles de Agarosa o PAGE). Los fragmentos son reparados en los extremos, purificados y ligados a los adaptadores, para obtener la biblioteca. Estos adaptadores, P1 y P2, se utilizan para aumentar el número de copias del blanco genómico.

3.3.2. Amplificación de bibliotecas

La amplificación de la biblioteca se realiza mediante una PCR de emulsión de la siguiente forma: los fragmentos se unen a los adaptadores P1, los que están ubicados sobre la superficie de unas "magnetic beads" (bolitas microscópicas magnéticas). Cada bolita contiene un único tipo de fragmento (figura 3.1 (a-c)). Las bolitas se colocan en unos microrreactores (pocillos) con la emulsión de agua y aceite, fragmentos de ADN, reactivos de PCR y primers. Los fragmentos son amplificados aproximadamente 30000 veces dentro de los microreactores ($\sim 10\mu m$). Los primers son complementarios al P1 y la extensión se hace hacia el adaptador P2.
Posterior a la PCR de emulsión, los fragmentos extendidos se filtran en un gradiente de glicerol para separar las bolitas con fragmentos extendidos, de las que no tienen fragmentos. Para ello, nuevas bolitas de poliestrieno cubiertas de adaptador P2 en la superficie se ponen en contacto con las bolitas de los fragmentos. Sólo aquellas con adaptador P2 permanecen para los siguientes pasos.
Los extremos 3' de los fragmentos amplificados sobre las bolitas son modificados para mejorar adherencia a una lámina de vidrio. Cada bolita se deposita sobre esa lámina en

orden aleatorio. Cada una de esas láminas puede estar dividida en varias cámaras de deposición (una, cuatro u ocho), lo que permite correr varias muestras en la misma corrida. La cantidad de millones de bolitas por cámara puede variar de acuerdo al protocolo y a la versión del equipo del SOLiD. Los valores pueden ir desde 165 millones de bolitas en una única cámara, 30 millones en cada una de las cuatro cámaras (120 millones en total), 14 en cada una de las ocho cámaras (112 millones en total), hasta un total de 700 millones de bolitas como genera el SOLiD 4 que fue utilizado en este trabajo de tesis.

3.3.3. Secuenciación por ligación

Una vez ubicadas las "magnetic beads" en sus cámaras sobre la placa de vidrio, comienza la secuenciación de los fragmentos. A diferencia de otras tecnologías, la misma ocurre mediante hibridación y ligación de fragmentos de largo 8 (8-meros), cuya secuencia general es -C-T-N-N-N-N-Z-Z-Z-. Las dos primeras posiciones están ocupadas por un dinucleótido (extremo 3'), la N implica bases degeneradas, es decir, con capacidad de parearse con cualquier base, y las Z son bases modificadas universales, las cuales no tienen preferencia de unión. El extremo 5' del 8-mero tiene ligado un fluorocromo. La especificidad de la ligación viene dada únicamente por el dinucleótido en el extremo 3', de los cuales existen 16. Cada uno de ellos posee uno de cuatro fluorocromos disponibles, implicando que cada color tiene 4 dinucleótidos asociados. El primer paso de la secuenciación comienza con la hibridación de un primer (complementario al adaptador P1). Seguidamente, la ligasa incorpora el 8-mero correspondiente a la hebra creciente, es decir, incorpora aquel fragmento con un dinucleótido complementario al que está siendo interrogado. En la incorporación del fragmento, se escinde el fluorocromo y las bases modificadas universales (Z-Z-Z), emitiendo una señal de luz fluorescente que es detectada por el secuenciador. El color de esa luz caracteriza a 4 posibles dinucleótidos. El fragmento incorporado en la hebra creciente tiene finalmente un largo de 5 (figura 3.2 (a)). A continuación la ligasa reanuda su función e incorpora el siguiente 8-mero. Se reitera este procedimiento varias veces hasta llegar al largo del read correspondiente. En 10 ciclos se obtiene un largo de read de 50 nucleótidos.
Esta primera ronda de secuenciación reporta información de las bases en posiciones 1 y 2, 6 y 7, 11 y 12, etc (figura 3.2 (b)). Este procedimiento se repite con un desfasaje del primer, por ejemplo, de una base hacia el extremo 5' (a la derecha), permitiendo reportar las bases 2 y 3, 7 y 8, 12 y 13 (figura 3.2 (b), el primer (n-3)). Este proceso de "primer resetting" se realiza en total 5 veces, con un desfasaje de -1, 0, 1, 2 y 3, para asegurarse que cada posición sea interrogada 2 veces. El desfasaje de -1 implica que el primer se coloca de forma tal que la primera base secuenciada es la última del adaptador P1, que es conocida (figura 3.2 (b) primer (n-1)).

En esta primera etapa, el resultado de la secuenciación es una serie de colores, correspondientes a los dinucleótidos interrogados, en cada uno de los 5 ciclos. Esta estructura es lo que le permite al software del secuenciador recrear la secuencia final a partir de esos colores. Cada interrogación de un dinucleótido corresponde a un color que se asocia a la posición del primer nucleótido del par. Los colores se representan con números: 0 = azul, 1 = verde, 2 = amarillo y 3 = rojo (3.2 (c), arriba). El SOLiD reporta sus reads en espacio color ("color-space"), los cuales pueden ser decodificados (no forma parte del proceso del secuenciador) posteriormente a bases ("base-space"), utilizando su sistema de decodificacíon de colores (3.2 (c), abajo). Ya que la primera base es siempre conocida (la última base del adaptador), es posible decodificar el resto de la secuencia de colores a bases, siguiendo el esquema. En general, siempre la primera base es T, de acuerdo al color que presente la segunda posición, se puede identificar la siguiente base. Si es rojo (3), la siguiente base correspondería a una A, si es amarillo (2) la siguiente sería una C, si es verde (1) una G y si fuera azul (0) sería otra T. Así se procede con los restantes colores. Con estas tres componentes, el read en espacio-color, el esquema de decodificación y la primera base, se pueden decodificar los reads (3.2 (c), derecha).
Esta técnica es la única que interroga dos veces a cada posición de la secuencia, lo que aumenta la precisión del método.

3.3.4. Resultados del secuenciador

Una vez finalizada la corrida del SOLiD (la misma puede durar desde 3 a 16 días, dependiendo de las bibliotecas y del experimento), el secuenciador reporta como resultado una serie de archivos de texto. Para cada muestra, se devuelven dos archivos, uno con la secuencia y uno con las calidades correspondientes. El archivo de secuencias es del tipo FASTA, lo que implica, que para cada fragmento (read) existe un encabezado, comenzado por ">" y seguido por identificador de read, y en la línea siguiente se encuentra el read en espacio color. Es decir, una T al comienzo (como ya se nombró más arriba) y una serie de colores representados por números del largo del read. Correspondiente a cada archivo de esos, existe un archivo de calidad que le adjudica un valor de calidad (ver siguiente sección) a cada base (color). Éste también presenta el formato FASTA. Estos archivos corresponden al punto de partida de varios posibles análisis "downstream", que veremos más adelante.
La cantidad de reads obtenidos depende del experimento, pero de acuerdo a las especificaciones del SOLiD 4[2], se logran alcanzar 1,4 billones de reads (en dos láminas en corrida simultánea), creando archivos de más de 100 Gb. En consecuencia, se precisan

[2] `http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_078637.pdf`

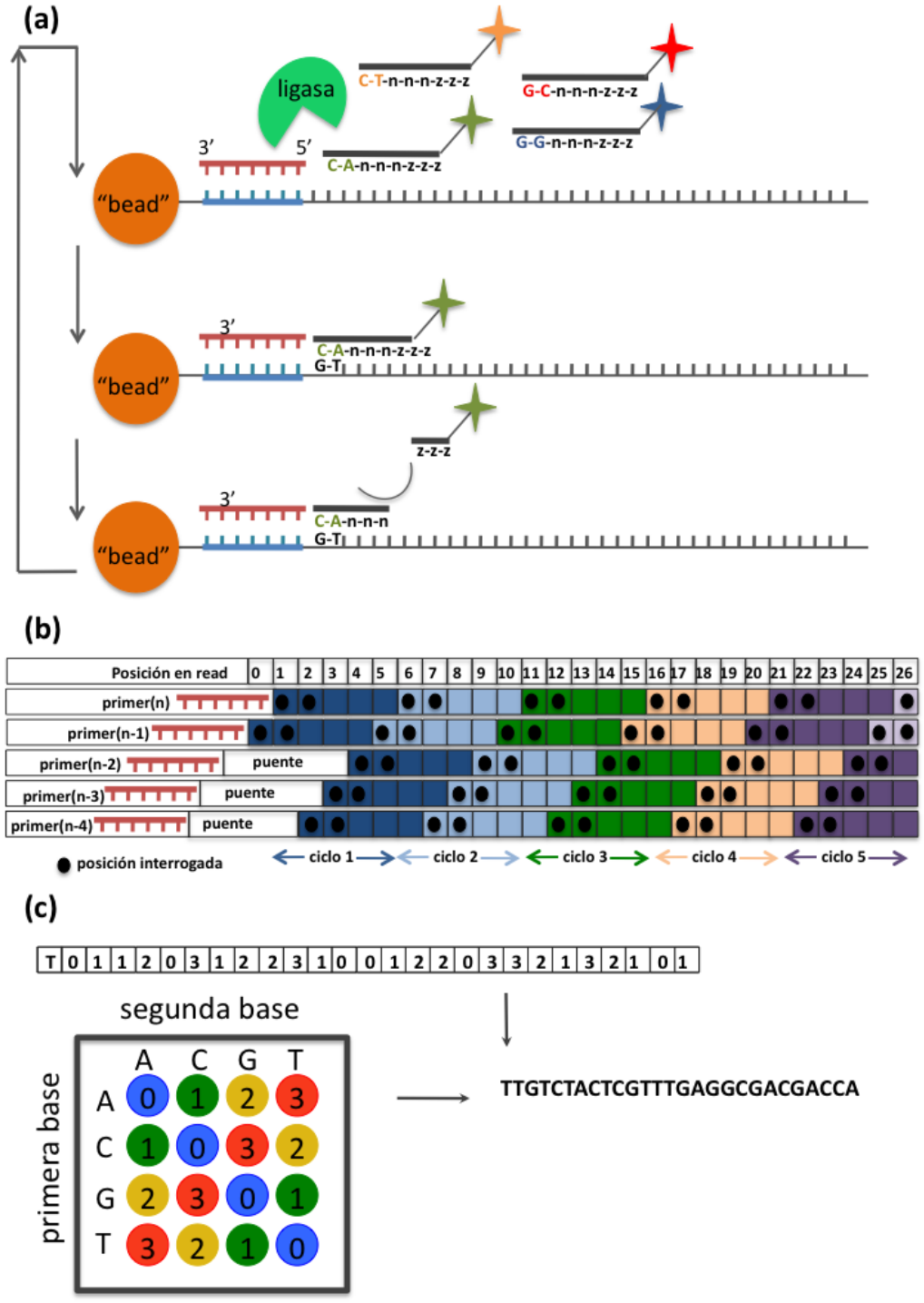


FIGURA 3.2: Química del SOLiD. (a) Ligasa incorpora fragmentos a la cadena creciente con las siguientes características: en el extremo 3' un dinucleótido, seguidamente 3 nucleótidos degenerados, se finaliza con un fluorocromo unido a 3 bases. Hay 4 colores disponibles y 16 dinucleótidos. Una vez que la ligasa une el fragmento complementario, el fluorocromo se escinde y emite una luz del color unido (señal). La ligasa vuelve a actuar, repitiendo el procedimiento. El resultado es una serie de colores que puede decodificarse a bases. (b) El primer (complementario a P1) se desfasa una posición de la posición original, repitiéndose el procedimiento de la secuenciación. El "primer reset" se repite 5 veces, para asegurar una doble interrogación de las posiciones. Conociendo la primera base del fragmento, se decodifica el read. (c) Decodificación del espacio-color: cada color tiene asociado 4 posibles dinucleótidos. Para poder decodificar a qué dinucleótido corresponde el color, se debe conocer la 1^{era} base del dinucleótido.

discos con alta capacidad de almacenamiento (por ejemplo del orden de $\sim$ 10Tb). Nuestra experiencia en experimentos de transcriptómica difiere un tanto, pero igualmente obteniendo grandes cantidades de reads (archivos de $\sim$ 10 Gb).
El largo de los reads resultantes puede variar desde 35 a 50 con el SOLiD 4. Nuevas versiones del SOLiD (550xl) logran largos desde 85 a 100 bases.

3.3.5. Calidad y precisión de acuerdo a la especificación del SOLiD 4

Como fue explicado anteriormente, la secuenciación del SOLiD también reporta valores de calidad para cada base determinada ("base-calling"). Se reportan en valores Phred (un tipo de score de calidad desarrollado por Phil Green [113]), que van desde el 0 hasta el 50. Los valores de Phred se determinan de la siguiente forma: $Q = 10 \log_{10} P$, siendo P la probabilidad de cometer un error en el "base-calling". Valores de ejemplo de Phred y sus correspondientes errores se ven en el cuadro 3.1. En general, la precisión del SOLiD es muy alta, sobre todo en las primeras bases.

Según las especificaciones del SOLiD 4, debido al doble chequeo de cada base, más

CUADRO 3.1: **Relación entre el valor de calidad de Phred y las probabilidades de error del "base-calling"**

Valor de Phred	Probabilidad de una base mal determinada	Precisión del "base-calling"
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99,9 %
40	1 in 10000	99,99 %
50	1 in 100000	99,999 %

El score de calidad Phred está relacionado logarítmicamente con las probabilidades de error.

del 80 % de las bases de una corrida tienen en general, un valor de Phred mayor a 30, lo que se refleja en una alta precisión por corrida. Esto implica que la cobertura de reads no precisa ser tan alta para varios de los análisis (sobre todo de detección de variantes genómicas), lo que lleva a menos falsos positivos (incorrecta identificación de bases o incorrecto "base-calling") y menos costos económicos y de tiempo. En resumen, de acuerdo a la ficha técnica, el SOLiD 4 logra datos más precisos, con menos corridas, y a menor costo.
Por otro lado, la calidad del read decae con los ciclos, es decir al final del read la calidad es menor que al comienzo, y esto es válido en general para todas las tecnologías de secuenciación.
A su vez, esta tecnología (sobre todo a partir del SOLiD 4 y la mejora de los reactivos)

logra secuenciar regiones genómicas difíciles a través de una cobertura más homogénea a lo largo del genoma, lo que baja la tasa de falsos positivos. El tamanño de los reads sigue planteando un problema, ya que si bien con los años esta tecnología ha mejorado los largos (de 35 a 50), los reads siguen siendo muy cortos, lo que dificulta la localización en el genoma de forma unívoca.
El rendimiento del SOLiD en cuanto a la cantidad de secuencias por corrida, ha crecido exponencialmente desde su liberación al mercado en el 2007. En sus inicios, las corridas alcanzaban un par de millones de reads, para luego en el 2010 alcanzar más de 2000 millones de reads.

3.3.6. Experiencia propia con las calidades del SOLiD

Si bien las especificaciones del SOLiD 4, que es el secuenciador utilizado en estos trabajos, realzan que aproximadamente 80 % de las bases de una corrida tienen en general un valor mayor a 30, nosotros no pudimos corroborar esos resultados. Las calidades obtenidas fueron aceptables, y posteriores análisis demostraron una buena reproducibilidad en general de los experimentos, pero esos altos valores de precisión no fueron alcanzados. La figura 3.3 (a) muestra a modo de ejemplo la calidad media de los reads por panel de una de nuestras muestras analizadas de transcriptómica (humano). Los paneles, son subsecciones de la placa de vidrio en donde hay una cantidad determinada de "beads". La señal de cada panel es captada a la vez por el scanner. El scanner recorre todos los paneles de las placas. La calidad por panel se utiliza para observar de que no haya ningún efecto espacial. El eje de la derecha de la figura 3.3 (a) muestra un gradiente de colores representando los valores de calidad en Phred. El área es en su mayoría amarilla, correspondiendo a un Phred de ~ 24 máximo, el resto del área se compone de anaranjado, correspondiente a una calidad de $\sim 20-22$. A su vez, hemos observado que algunos colores pueden tener tendencia a tener mejores o peores calidades, la figura 3.3 (b) muestra la calidad media de las bases del color 0. Ésta tiene valores entre $\sim 20-22$, la que es menor comparada con la figura 3.3 (c), en donde está representada la calidad de las bases del color 1, que comprende valores en torno a los 26. La figura 3.3 (d) muestra la calidad media a lo largo del read. Se aprecia una caída constante de la calidad a medida que nos acercamos al final del read, partiendo de valores de menos de 30 llegando hasta valores bastante menores a 20.
Este comportamiento lo comprobamos en muchas muestras analizadas, que fueron un total de 16. Los resultados siguen siendo de cualquier modo confiables y reproducibles. Esto se observa sobre todo en la clusterización de las muestras. Las mismas se agrupan perfectamente por condiciones (se verá más adelante).

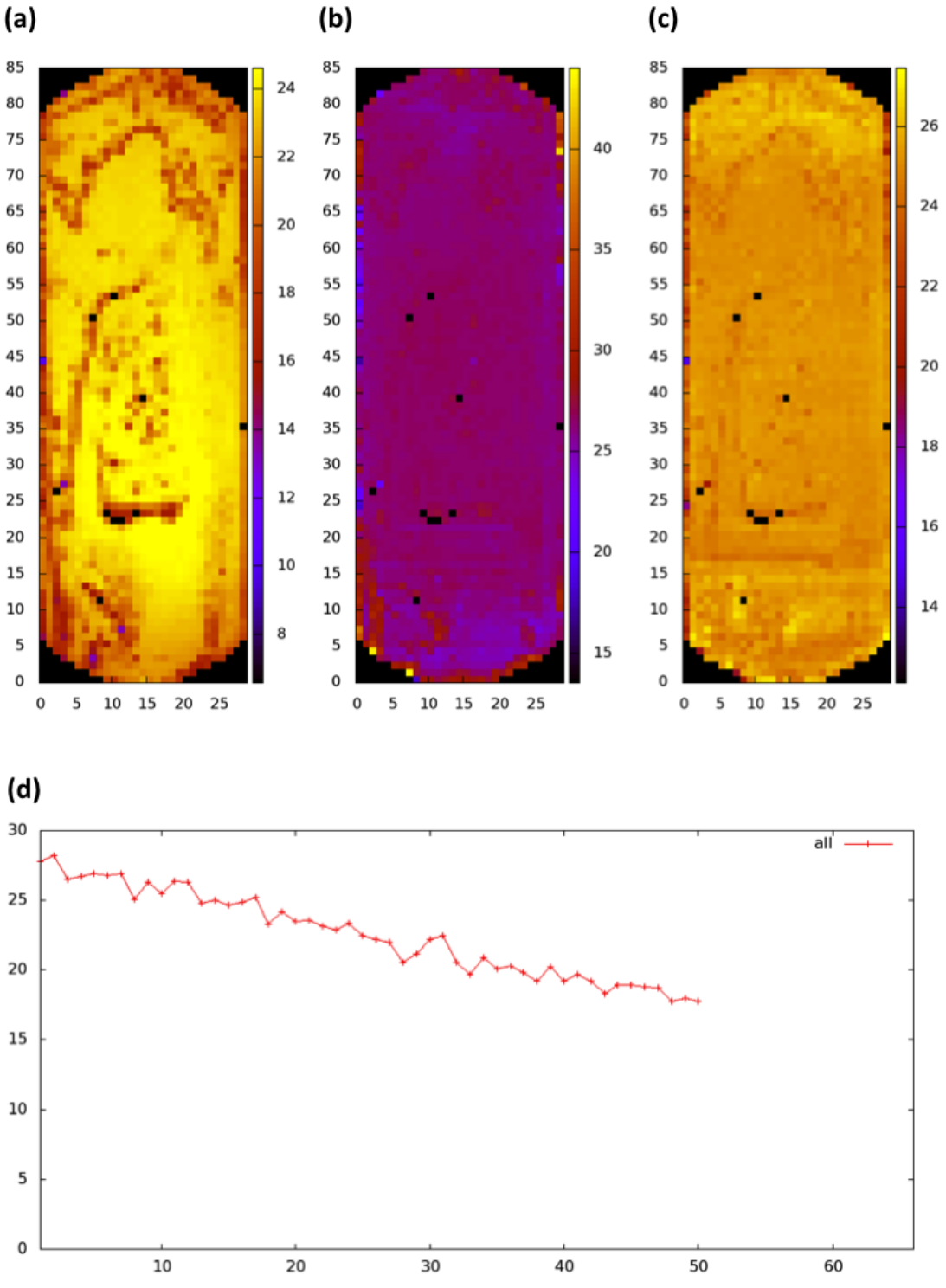


FIGURA 3.3: Calidad obtenida de una de nuestras propias muestras. (a) Calidad media por corrida por placa. A la derecha el gradiente de valores de calidades. El valor máximo de calidad observado es de 24, y solo algunas regiones presentan ese valor alto. En general, los valores oscilan entre $\sim 20 - 22$. (b) Calidad media de los reads de color 0 por placa. En general, los reads presentan calidades entre $\sim 20 - 22$. (c) Calidad media de los reads de color 1 por placa. Los valores rondan el 26. (d) Calidad media a lo largo del read. La misma decae a medida que avanzamos hacia el extremo del read.

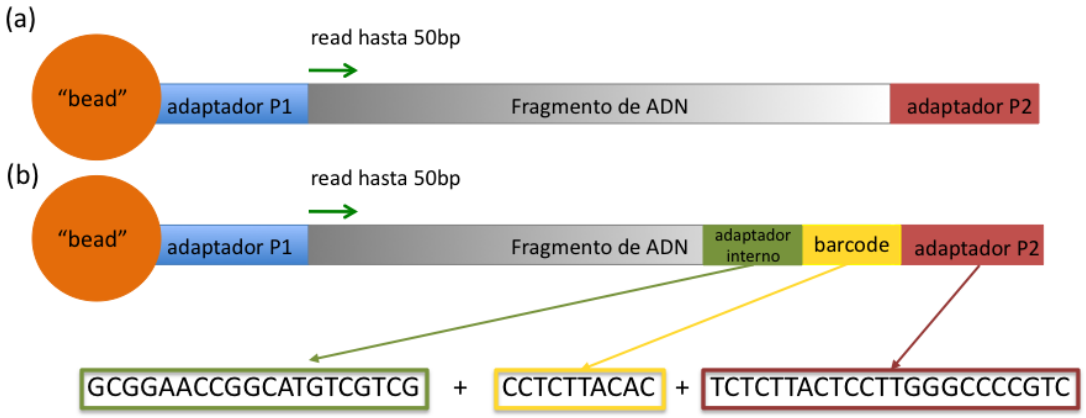


FIGURA 3.4: (a) Biblioteca sin barcode, single-end, utilizada usualmente para secuenciar genomas, transcriptomas, exomas, etc. (b) Ejemplo de biblioteca con barcode. En general se utiliza para secuenciar pequeños ARNs.

3.3.7. Multiplexado

El multiplexado de muestras es una técnica muy útil cuando se desean secuenciar regiones genómicas específicas o genomas pequeños. Esta estrategia permite la secuenciación de varias muestras simultáneamente (un pool de muestras) incrementando así el número de muestras por corrida, sin aumentar drásticamente el precio ni el tiempo necesario. Para poder utilizar el multiplexado, a cada muestra debe incorporársele un identificador único o índice ("barcode"), para poder distinguir una de otras. Este identificador se compone de una serie de nucleótidos, que se colocan previo al adaptador 3'. Dependiendo de la cantidad de muestras a multiplexar, cómo se compone la secuencia del identificador. El SOLiD 4 puede multiplexar de 16 a 96 muestras con distintas bibliotecas. La figura 3.4 compara la biblioteca sin barcodes 3.4 (a) con la biblioteca con el identificador único 3.4 (b). En este caso el largo del barcode es 10 bases, previamente se agrega un adaptador interno de largo 20, y posteriormente se encuentra el adaptador 3' P2 (de largo 23). El ejemplo es tomado de una biblioteca con 96 barcodes disponibles.

Como fue mencionado anteriormente, una de las situaciones posibles para utilizar el multiplexado es secuenciar genomas pequeños. Los genomas de gran tamaño muchas veces precisan una lámina entera, por lo que no es posible colocar varias muestras en una corrida. La otra situación y la más común en los experimentos con SOLiD es secuenciar regiones específicas del genoma, por ejemplo, el transcriptoma, pequeños ARNs, entre otros. En esos casos en los que se utiliza el multiplexado, la estructura del fragmento a secuenciar es como la mostrada en la figura 3.4 (b). En los casos que se utiliza el multiplexado, en general, los fragmentos seleccionados para secuenciar son menores que 50 (por ejemplo 22 si estamos hablando de microRNAs), por lo que el secuenciador reporta no sólo el fragmento de interés, sino que también el adaptador interno, el barcode

y eventualmente alguna porción del adaptador 3'. El SOLiD secuencia todo el fragmento hasta un largo máximo de 50 bases.

3.4. Posibles aplicaciones con el SOLiD

Son muchas las posibles aplicaciones con el SOLiD: resecuenciado de genomas, secuenciación específica de regiones genómicas, transcriptómica (esto incluye ARN pequeños, exoma, etc.), epigenomas (ChIP-Seq y estudios de metilación), estudios de mutaciones y SNPs (single nucleotide polymorphisms), etc. Cada una de estas aplicaciones presenta sus propias dificultades sobretodo a la hora de realizar las bibliotecas.

3.4.1. Resecuenciamiento de genomas

La cantidad de genomas de referencia ya secuenciados de varios organismos permite aplicaciones como el resecuenciado de genomas, es decir comparar genomas de la misma especie contra el/los genoma/s de referencia ya secuenciados, y de esta forma catalogar las variaciones, mutaciones o SNPs del genoma secuenciado y comprender sus consecuencias biológicas. Por ejemplo, se realizan resecuenciamientos de genomas de "cáncer" y se comparan con el/los sano/s para determinar las mutaciones, variaciones y poder sacar conclusiones con respecto al fenotipo de la enfermedad. Existen una cantidad de estudios de resecuenciamiento de diferentes cánceres [114–116] y de varias enfermedades, como ser atrofia multisistema [117], esclerosis múltiple [118], entre otras. Por otro lado también se realizan estudios de resecuenciamiento de cepas bacterianas (o virus), sobre todo patogénicas para la identificación de nuevas mutaciones y nuevas cepas emergentes [119, 120]. Tanto en el área de genómica humana, como en la bacteriana, han surgido varios proyectos mundiales con el fin de resecuenciar una gran cantidad de genomas de una especie. Por ejemplo el proyecto de los 1000 genomas humanos (realizado con 454, Illumina y SOLiD) para determinar las variaciones genéticas en las poblaciones[3] y el resecuenciado de 200 cepas de *Leptospira* organizado por el J. Craig Venter Institute.[4]

3.4.2. Estudios de regiones genómicas específicas

Estos estudios enriquecen una zona particular del genoma, ya sean genes específicos, o regiones codificantes o no codificantes, para estudiar una temática en particular. Existen

[3] `www.1000genomes.org`
[4] `http://gsc.jcvi.org/projects/gsc/leptospira/`

protocolos de preparación de bibliotecas diferentes para enriquecer las regiones deseadas. Varios estudios se enfocan en genes específicos en diferentes enfermedades como ser cáncer de próstata [121], rinitis juvenil y amaurosis congénita [122], ataxia telangectasia y corea acantocitosis [123], entre muchas otras. Este tipo de estudios se está utilizando cada vez más para diagnóstico clínico y ayuda a la toma de decisiones médicas.

3.4.3. Estudios de transcriptómica

Los estudios de transcriptómica implican estudiar qué porción del genoma está siendo transcripta en un momento dado bajo determinadas condiciones, en el tejido o población celular especificado. La secuenciación del transcriptoma intenta caracterizar determinadas moléculas de ARN en la célula. Dependiendo del estudio específico a realizar, se seleccionará en qué molécula de ARN se centrará la secuenciación. En el caso de análisis de expresión génica, los ARN caracterizados serán los ARN mensajeros. En el caso del estudio de pequeños ARNs, se enriquecerán los ARN de tamaños pequeños como pueden ser ARNmi (micro), ARNt (de transferencia), ARNsn (small nucleolar) o ARNsi (small interference), entre muchos otros. Dependerá del tamaño seleccionado en la creación de la biblioteca, qué tipo de ARN pequeño se analice. En general, se intenta no caracterizar los ARNr (ARN ribosomal) ya que su alta abundancia se "llevaría" toda la secuenciación. Es decir, si no se utilizan protocolos experimentales para reducir el número de moléculas de ARNr, la secuenciación resultante estará llena de éstas, sin poder estudiar los niveles de expresión de ninguna otra. Muchísimos estudios de transcriptómica han sido publicados a lo largo de estos últimos años. Varios de los mismos se enfocan en estudios de expresión génica, es decir ARNm de distintos tejidos, como ser cerebro humano [124], células madre mesenquimales humanas [125] y embrionarias de ratón [126], hígado [127], etc. Muchos estudios de transcriptómica se enfocan en pequeños ARNs en varias enfermedades, como cáncer [128], ataxias [129] o Parkinson [130], entre muchas otras.

3.4.4. Estudios de epigenética

La epigenética estudia los cambios en expresión génica o de fenotipos celulares causados por mecanismos distintos a la secuencia de ADN subyacente, los cuales pueden ser también hereditarios. La misma se enfoca en modificaciones relevantes en el genoma que no implican el cambio de nucleótidos en la secuencia. Ejemplo de estos eventos incluyen metilación de ADN y modificaciones de las histonas. Dentro de estas últimas se consideran, acetilaciones y metilaciones. En cualquiera de los casos, estas modificaciones alteran la

regulación génica sin modificar la secuencia. Estos cambios pueden permanecer por varias generaciones. Una técnica ampliamente utilizada para analizar estas modificaciones es la denominada ChIP-Seq (Chromatin-Immuno Precipitation Sequencing). Este método permite estudiar interacción de proteínas con ADN, combinando inmunoprecipitación de cromatina con secuenciado masivo, para determinar los sitios de unión de la proteína en el ADN. Muchas de sus aplicaciones consisten en generar mapas de sitios de unión de ADN con proteínas de interés como por ejemplo factores de transcripción o histonas (con modificaciones específicas). Regiones específicas del ADN que están en contacto físico directo con estos factores de transcripción o histonas modificadas (acetiladas/metiladas), pueden ser aisladas específicamente con anticuerpos (inmunoprecipitación), por lo que el proceso del ChIP genera una biblioteca de fragmentos de ADN previamente unidos a la proteína de interés *in vivo*. Estos fragmentos son secuenciados con alguna de las tecnologías de segunda generación (por ejemplo SOLiD) o de tercera generación (por ejemplo Ion Torrent) para obtener, no sólo la secuencia del sitio de unión sino también medidas cuantitativas y mapas de sitios de unión múltiples a lo largo del genoma. Estos resultados pueden integrarse con los de transcriptómica para sacar conclusiones más amplias. Por ejemplo, si se desea estudiar la trimetilación de la histona H3, porque se sostiene que puede tener efecto silenciador en los genes que presentan esta modificación en su promotor, puede realizarse un experimento de ChIP-Seq para ubicar la modificación de interés, puede obtenerse una visión cuantitativa de la abundancia de esa modificación y al combinarse con datos de expresión génica se pueden inferir las consecuencias de la modificación de interés en la expresión de los genes que la presentan. Muchos estudios utilizan esta técnica para evaluar modificaciones epigenéticas en diferentes situaciones, como activación muscular [131], en comparaciones a nivel de diferentes tejidos por ejemplo cerebro y testículo de ratón [132], en estudio de modificaciones específicas en cáncer [133], etc.

3.5. Comparación de plataformas

A continuación evaluaremos las ventajas y desventajas de las plataformas de secuenciamiento más utlizadas, basándonos en el estudio de Liu *et al* del 2012 [134]. Por un lado evaluaremos el SOLiD de Life Technologies, HiSeq 2500 de Illumina y GS FLX Titanium/GS Junior de Roche. La tabla 3.2 compara las características principales de estos tres secuenciadores.

El 454 de Roche, comenzó con largos de reads de 100 a 150 en el 2005, con una cantidad de 200 mil reads por corrida (20Mb) [135]. En el 2008 el 454 GS FLX Titanium logró largos de reads de 700, con una precisión de 99,9 % y una cantidad de secuencia de 0,7G por corrida en 24 horas. A finales del 2009 se simplificó la preparación de las

librerías y el procesamiento de los datos. Una de las mayores ventajas de Roche es su velocidad: en tan sólo 10 horas se obtiene una secuenciación. A su vez, es la tecnología con reads más largos (comparando con SOLiD e Illumina). Los costos de los reactivos siguen siendo muy altos, $12{,}56 \times 10^{-6}$ dólares por base (precios correspondientes al año 2012), solo contando los reactivos. Una de las mayores deficiencias del 454 es su tasa de error relativamente alta para secuenciar poli-bases (aparece la misma base en forma consecutiva), sobre todo cuando las mismas tienen un largo de al menos 6. La construcción de las bibliotecas puede automatizarse, y la PCR de emulsión está semiautomatizada, reduciendo al máximo la intervención humana, por lo tanto minimizando errores.
El largo original de los reads del SOLiD era de 35 y la cantidad de secuencia correspondía a 3G en el 2007. Debido a la doble interrogación de bases la precisión alcanzaba 99,85 % posterior al filtrado de los reads no mapeados. Tres años, y cinco "upgrades" del SOLiD después, sale al mercado el SOLiD 5500xl, el cual mejora el largo de los reads, la precisión y la cantidad de datos a 85, 99,99 %, y 30G por corrida, respectivamente. Una corrida completa puede completarse en una semana con un costo por base de 40×10^{-9} dólares (sólo reactivos) estimado por los usuarios del BGI (Beijing Genomics Institute, el instituto con mayor cantidad de secuenciadores de todos los tipos) y generando una inmensa cantidad de datos crudos (4TB). La deficiencia mayor sigue siendo el pequeño tamaño de los reads. El tipo de aplicaciones posibles con el SOLiD se detalló más arriba y en la parte inferior de la tabla 3.2 se compara con las otras tecnologías. Como también en las otras tecnologías, la infraestructura es cara y compleja, precisa un centro de cómputos de avanzada (clusters, aire acondicionado, personal capacitado, redes de alta velocidad). La preparación de las bibliotecas puede automatizarse también. Se puede utilizar el multiplexado que permite analizar varias muestras al mismo tiempo.
En sus comienzos Solexa (después comprada por Illumina) conseguía generar 1G de datos por corrida. Una vez que se mejoraron las polimerasas, buffers, celdas, software, se logra en el 2009 aumentar a 20G de datos con reads de 75 de largo por corrida (paired-end), 30G para reads de 100 (paired-end) y 50G por corrida para reads de 150. La última máquina de la serie GAIIx logra generar 85G de datos. A principios del 2010 Illumina saca al mercado el HiSeq 2000 que genera inmensas cantidades de datos, en un principio 200G y posteriormente 600G por corrida, la cual puede completarse en 8 días. El primer instituto en utilizar esta máquina fue BGI, y de acuerdo a sus datos la tasa de error en promedio es menor que 2 % en corridas paired-end con reads de 100 de largo. Comparado con el 454 y SOLiD, HiSeq es el más barato, costando 2×10^{-8} dólares por base (sólo reactivos, de acuerdo a los datos de BGI), la mitad que el SOLiD. El multiplexado del HiSeq permite secuenciar miles de muestras simultáneamente. La complejidad de la infraestructura informática es alta, al igual que en el resto. La preparación de las bibliotecas y la medida de la concentración de las mismas puede ser automatizada también. La mayor deficiencia es el relativamente pequeño tamaño de los reads.

CUADRO 3.2: **Comparación de cada uno de los secuenciadores de acuerdo al estudio de Liu *et al* (2012)**

	454 GS FLX	HiSeq2500	SOLiD 4
mecanismo	pirosecuenciados	secuenciación por síntesis	secuenciación por ligación
largo de read	700	50 (single-end) , 50-150 (paired-end)	50 (single-end) 50 (paired-end)
precisión	99,9 %	98 %(100 % paired-end)	99,94 % (single-end)
#de reads	1M	3G (single-end), 6G (paired-end)	$1{,}2 - 1{,}4$G
datos	0,7G	600G	120G
tiempo	24 horas	$3 - 10$ días	7 (single-end), 14 (paired-end)
ventajas	largo del read, rapidez	rendimiento	precisión
desventajas	alta tasa de error con polímeros mayor a 6, alto costo, bajo rendimiento	reads relativamente cortos	reads cortos
resecuenciado	-	sí	sí
de novo	sí	sí	-
cáncer	sí	sí	sí
array	si	si	si
↑GC	si	si	si
bacteria	sí	sí	sí
genoma grande	sí	sí	-
mutaciones	sí	sí	sí

Parte superior: características de los secuenciadores más utilizados de acuerdo a experiencia previa de varios estudios realizados sobre varias muestras[134]. Algunos de los valores pueden diferir un poco, con respecto a las especificaciones propias.
Parte inferior: aplicaciones posibles para cada tecnología.

La decisión de una plataforma es un tema delicado, para el cual no existe una solución única. Factores importantes para el apoyo a la toma de decisión es el tamaño del genoma a secuenciar, la complejidad del mismo (por ejemplo, cantidad de repetidos, contenido G+C, etc.), la precisión y cobertura necesaria. De cualquier forma, en última instancia dependerá del experimento a realizar. Si se desea ensamblar un genoma *de novo* completo, se necesitarán reads largos, si se desean obtener resultados rápidos para tomar decisiones clínicas por ejemplo, uno precisaría los secuenciadores más pequeños y rápidos. Si el estudio precisa una alta cobertura, por ejemplo para determinación de SNPs o mutaciones, SOLiD o Illumina serían una buena opción. Dependerá del experimento, los recursos económicos y de tiempo, el conjunto de plataformas a utilizar. Dentro de esa subselección, la elección es subjetiva.

3.6. Análisis de datos provenientes del SOLiD

A continuación detallaremos el "pipeline" (la serie de pasos) para el análisis de los datos de secuenciación del SOLiD.
Una vez hecha la secuenciación de las muestras de interés, los resultados se obtienen en forma de archivos de texto. La figura 3.5 muestra un esquema del pipeline, incluyendo alternativas para cada paso.

3.6.1. Salida del SOLiD

Si la corrida fue realizada con bibliotecas de barcodes, es decir varias muestras fueron analizadas a la vez, el resultado final será un único archivo con todos los (millones de) reads, que serán separados de acuerdo a sus barcodes automáticamente (ver figura 3.5, demultiplexado), generando así varios archivos, dos por cada muestra (uno con la secuenciación y otro con las calidades respectivas). Si la corrida fue simple, sin multiplexado, se obtienen directamente dos archivos por muestra.
El archivo ".csfasta", es el que contiene la secuenciación en sí, y es del tipo "fasta", más específicamente multifasta. Los archivos fasta, son archivos de texto plano con el siguiente formato:

> identificador 1
AGCTGATCGATCGATCGATCGT...

Si se concatenan varias secuencias en un mismo archivo, se obtiene un multifasta:

> identificador 1
AGCTGATCGATCGATCGATCGT
> identificador 2
ACATCATGGATCCTTCGATCAA
> identificador 3
GCTGTTGGATTTATGCGATCTG
⋮

En el caso del SOLiD los archivos se llaman csfasta (color-space fasta) y en vez de contener la secuencia en bases, las poseen en espacio color. Cada color está codificado con un número (1, 2, 3 y 4), por lo tanto el archivo csfasta se ve de la siguiente forma:

>1_41_67_F3

T2000310001300001301003110102200113011320303202000 0
>1_42_76_F3
T2000100200113201101000303302010031102002102223001 0
>1_42_91_F3
T2021233303023201122000330101300022202220100220020 0
⋮

A cada archivo csfasta le corresponde un archivo del tipo ".qual", que contiene las calidades correspondientes a esa muestra. Las calidades están representadas con los scores del tipo Phred:

>1_87_1062_F3
4 4 16 12 11 20 8 15 5 17 13 11 12 8 6 28 12 14 15 15 7 24 21 9 17 4 4 4 10 7 9
>1_95_1656_F3
6 17 18 8 4 15 13 21 23 11 10 19 25 26 17 10 24 7 24 21 21 14 22 23 9
>1_117_350_F3
7 6 4 5 9 4 4 10 15 16 7 14 13 18 15 4 16 4 18 13 4 10 4 23 17 4 4 4 15 9 29
⋮

Cada muestra por lo tanto es representada por dos archivos, csfasta y qual (ver figura 3.5, arriba). Este es el punto de partida para todos los posibles análisis bioinformáticos.

3.6.2. Análisis de calidad

Independientemente de cual sea el estudio a realizar, primeramente se debe evaluar la calidad de la secuenciación. Para ello existen varias alternativas (3.5, NGSQC, FastQC, TileQC entre varias otras), las cuales analizan en una primera etapa la calidad de los reads individuales y finalmente, dependiendo de la herramienta utilizada, varían los estudios más especializados. Estas herramientas parten de los archivos anteriormente nombrados (csfasta y qual para SOLiD, y uno análogo denominado fastq en el caso de Illumina). NGSQC (Next Generation Quality Control) [136] a modo de ejemplo, es un programa básicamente escrito en Phyton que proporciona un conjunto de medidas de calidad para detectar rápidamente una amplia variedad de problemas de calidad de datos de secuenciación, independientemente de la tecnología utilizada. Estos datos son derivados de superficies bidimensionales, ya que en esencia, se capturan imagenes emitidas en superficies, como ser paneles, "tiles" o cualquier unidad de captura de imágenes. Por este motivo NGSQC analiza la distribución espacial de colores/bases por panel, la distribución de reads mapeados al genoma con 0, 1 o 2 "mismatches" por panel, número de reads en general por panel, entre otras medidas más clásicas de calidad como ser la calidad media a lo largo del read, la calidad media por panel, etc. Estas medidas permiten observar patrones espaciales, effectos, artefactos que de otra manera serían

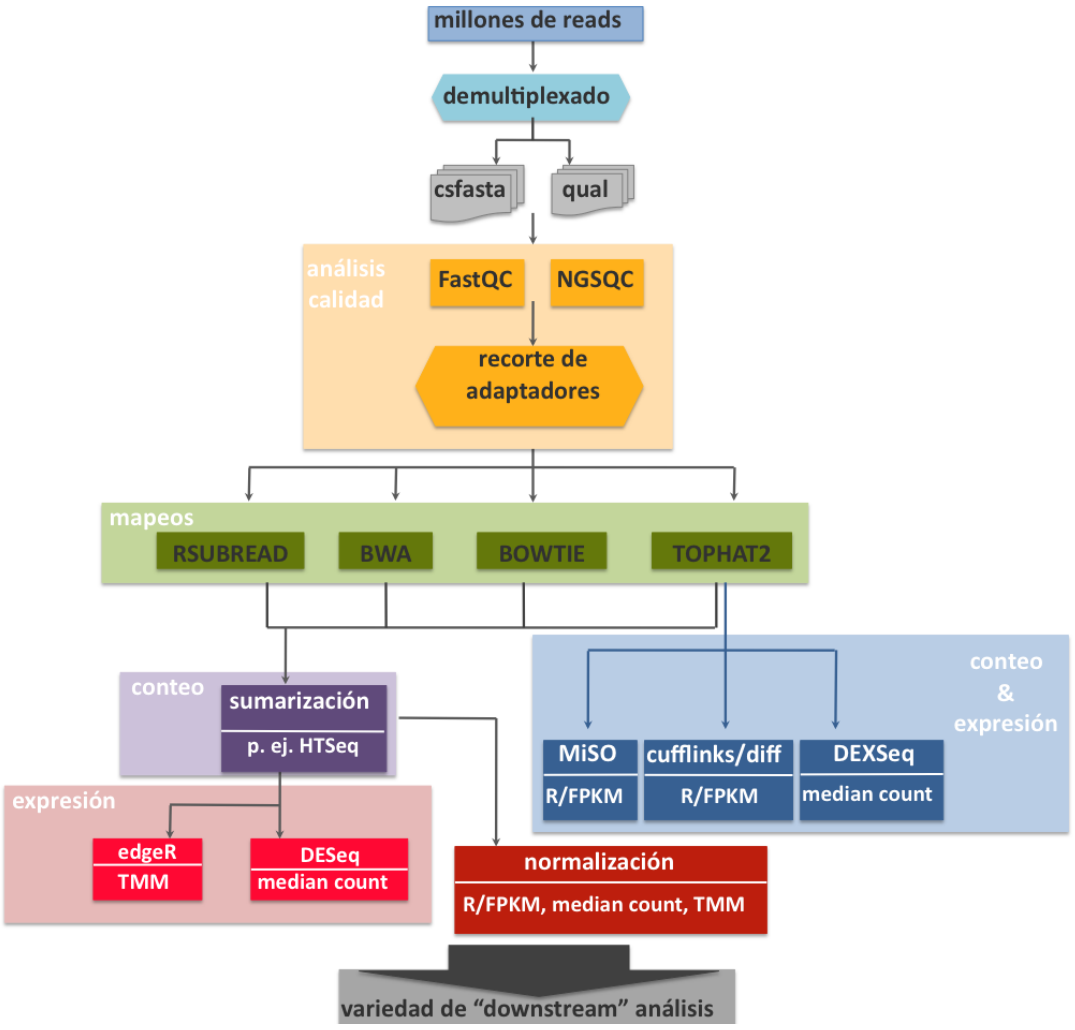


FIGURA 3.5: Posible pipeline para el análisis de datos provenientes del SOLiD

imposibles de capturar. La figura 3.3 presenta imágenes justamente de este software.
Este programa no sólo realiza un análisis de calidad a fondo, permitiendo descartar muestras parciales o totales si fuera necesario, sino que puede ayudar al investigador a evaluar si los resultados obtenidos pueden haber sido influenciados por algún problema de calidad. Aún cuando los análisis de calidad revelan una serie de problemas, es difícil de investigar los alcances de las consecuencias biológicas de los mismos. Una característica de NGSQC es la capacidad de vincular los reads relacionados con una conclusión biológica específica (por ejemplo, una lista de genes diferencialmente expresados) a los posibles problemas de calidad. El programa reubicará la lista de reads de interés en la superficie de secuenciado original para poder observar si los mismos provienen de zonas conflictivas, dudosas o de baja calidad.
Una vez analizadas las calidades de las muestras se debe decidir si se conservan todas, o se descarta alguna, y posteriormente se puede proceder con el mapeo de los reads al genoma.

3.6.2.1. Recorte de adaptadores

La figura 3.5 menciona, dentro de la casilla de calidad, el recorte de adaptadores. Este procedimiento puede observarse también como una medida de calidad, ya que muchos reads pueden ser descartados posterior a este procedimiento. El mismo es sólamente aplicable cuando se trabaja con pequeños ARNs. En estos casos, los pequeños ARNs no llegan a cubrir todos los ciclos de secuenciación (por ejemplo en un read de 50 de largo, un microARN ocuparía sólamente $\sim$ 22 nucleótidos) y el remanente se compone de adaptadores y eventualmente barcodes. Para asegurarse de que se está trabajando con los ARN pequeños de interés el extremo 3' del read debe componerse del adaptador 3' completo o porciones del mismo y eventualmente del barcode. Los reads que no presentan esos adaptadores en el extremo debe ser filtrados y no ser considerados, ya que al no presentar este adaptador no existe seguridad de que se está trabajando realmente con pequeños ARNs.

3.6.3. Mapeos

En general, llamamos mapeo (alineamiento) al procedimiento de ubicar/localizar en el genoma de referencia a los reads. Como se muestra en la figura 3.5 existen varias posibilidades a la hora de mapear o alinear. La misma muestra solo tres posibilidades muy utilizadas en la actualidad, pero existen muchas más opciones, dependiendo del experimento y pregunta de interés.
En general, todos los algoritmos en una primera etapa construyen un "índice" del genoma de referencia, es decir una estructura indexada que permite una búsqueda más eficiente de las posiciones de los reads. Este es el paso que consume más tiempo de máquina y recursos. El mismo debe de ser generado una sola vez y luego cada muestra es "mapeada" contra el "índice".
Independientemente del algoritmo que se utilice, se deben determinar ciertos parámetros relevantes que influirán el resultado del mapeo. Un parámetro importante a definir, presente en todos los programas de mapeo, es si se permiten mapeos múltiples o sólamente únicos. Es decir, en general la mayoría de los reads alinearán con una única posición en el genoma, pero una proporción considerable puede alinear en varias posiciones a la vez. Éstos, denominados multi-reads (multi-mappers) ocurren debido a las duplicaciones génicas en el genoma, a las regiones repetitivas del mismo, genes que se solapan, etc. Por lo tanto, es de esperar una cierta proporción de multi-reads. Asignarlos a un gen específico puede ser problemático y puede llevar a errores en la estimación de expresión diferencial. Algunos programas (si se elige el parámetro de mapeo múltiple) asignan aleatoriamente al read a una de sus posiciones posibles (*Rsubread* [137]), otros asignan

fracciones del mismo a cada una de las posiciones posibles (*tophat2* [138]). Esto puede resultar en diferencias de cobertura en las regiones génicas con reads únicos comparado con las regiones cubiertas por multi-reads. Por lo tanto, incluir multi-reads puede llevar a una sobre/sub estimación de la expresión de algunos genes. Por lo tanto por un lado, incluirlos en los análisis puede llevar a cierto sesgo en el cálculo de expresión diferenicial, pero por otro, la exclusión de los mismos lleva a la pérdida de información y a una subestimación de la expresión de los genes afectados por multi-reads. Esto último no es relevante si posteriormente se compara ese mismo gen en diferentes muestras. La misma subestimación estará presente en cada una de las muestras, por lo que desaparece el sesgo, lo que sí puede disminuir es la sensibilidad del test para capturar expresión diferencial. La elección está exclusivamente en el usuario, dependiendo de la aplicación resultará si es conveniente utilizar mapeos únicos o múltiples.

Otro parámetro relevante a definir es la cantidad de "mismatches" en el alineamiento. Un read puede ser mapeado al genoma de forma perfecta, es decir todas las bases del read coinciden con las correspondientes en el genoma de referencia, o puede poseer mismatches. Éstos corresponden a las bases del read que no son idénticas a las del genoma de referencia. Dependerá de la aplicación el nivel de exigencia del mapeo, ya que permitir varios mismatches por read también pueden generar incertidumbre en la posición del mismo. Existen una gran variedad de parámetros a determinar dependiendo del algoritmo que utilice el programa seleccionado. El algoritmo del *Rsubread* utiliza una subsecuencia del read, denominadas semillas, de un tamaño k (menor al largo total) para ubicar de forma más eficiente a los reads. El k es un parámetro a definir que influye los resultados y eficiencia del procedimiento.

Otra decisión importante a la hora de definir el programa a utilizar es si es de interés considerar los denominados "spliced alignments". Los mismos representan reads que caen justamente en la unión de dos exones, por lo que mapean una porción en un exón y la otra en el siguiente exón, a muchas bases de distancia en el genoma. Un programa estándar de mapeo (*BWA* [139]), utilizado más bien para transcriptómica, no lograría mapear este read. Sin embargo, otros programas más especializados intentan ubicar correctamente en el genoma estos reads, partiendo al mismo en porciones e intentando mapear cada parte en diferentes regiones génomicas apartadas entre ellas el tamaño estimado promedio de un intrón. *tophat2* es capaz de realizar este enfoque, logrando no sólo mapear la mayor cantidad de reads posibles, sino que además es posible, posterior al mapeo, estimar la expresión de transcriptos alternativos y descubrir eventos de splicing alternativos previamente desconocidos, entre otros.

Estudios recientes han comparado las distintas herramientas de mapeo para este tipo de datos [140]. Cualquiera sea el algoritmo de elección el resultado de un procedimiento de mapeo es en general un archivo de texto plano delimitado por TABs en el formato SAM (Sequence Alignment Map format) o BAM (binary SAM). El mismo consiste

en un encabezado opcional y una sección de alineamientos. El encabezado empieza con @ si está presente y la sección de alineamientos sigue a continuación. Cada línea del archivo (menos en el encabezado) representa un alineamiento, y la misma tiene 11 campos obligatorios, que contienen información relevante del mapeo como ser posición del read, hebra, cromosoma, tipo de alineamiento (reverso, en sentido codificante, etc), mismatches, una secuencia de caracteres denominada cigar string que representa el mapeo, calidades, etc. Se obtiene un archivo SAM/BAM por muestra.
Estos son tan solo ejemplos de las variables a determinar para realizar el mapeo. Una vez finalizado un procedimiento de mapeo estándar, la información debe ser sumarizada por característica ("feature") a través del procedimiento de conteo.

3.6.4. Conteos

Le llamamos "feature" (característica) tanto a genes, exones, transcriptos, como cualquier elemento de interés del que queramos evaluar la expresión. Se deben conocer las coordenadas genómicas de cada "feature" para poder realizar el conteo. Las informaciones genómicas relevantes se guardan en un archivo del tipo GFF (Generic Feature Format) o su versión 2 denominada GTF (General Transfer Format). El mismo es un archivo de texto plano que contiene informaciones del tipo nombre de la característica, ubicación genómica, hebra, cromosoma, atributos, etc. A partir de este archivo y los resultados del mapeo, se cuentan las cantidades de reads que caen por "feature". Una opción para ello puede ser el programa escrito en Python denominado htseq-counts, de la suite de HTSeq. El mismo recibe un archivo del tipo SAM y realiza el conteo. Lo que se debe definir previamente es como lidiar con reads que caen en más de una feature, o no cubre una "feature" totalmente, o caen en sitios de splicing, etc. Todo esto puede ser definido en el programa y cambiará los resultados del conteo dependiendo de la alternativa considerada. En general, este programa se utiliza para sumarizar información de reads que caen dentro de regiones codificantes de un genoma anotado. El resultado del conteo es la lista de "feature" con el correspondiente número de reads. La figura 3.5 muestra este paso en la caja denominada "conteo" (violeta) posterior a las alternativas de mapeo (verde).

3.6.5. Normalización

Para poder realizar una comparación precisa de los niveles de expresión entre diferentes muestras, los datos de conteos deben ser primeramente normalizados. Dos tipos de sesgos importantes se deben considerar para normalizar correctamente los conteos: sesgo dentro de muestra, causado mayormente por las diferencias en el largo de los transcriptos y el

sesgo entre muestras, resultado de las diferencias de tamaño de bibliotecas. Esta última se debe a las diferencias de profundidad en la secuenciación. Cada herramienta de expresión diferencial utiliza su método de normalización favorito como está indicado en la figura 3.5 (edgeR utiliza una media "trimeada" de los valores M con el método denominado TMM, DESeq utiliza la mediana de los conteos a través del método median count, cufflinks utiliza el conteo de reads por kilo base por millón denominado RPKM, etc). Uno de los métodos muy utilizados es la normalización por RPKM/FPKM (Reads o Fragments Per Kilobase of transcript per Million mapped reads), el cual ajusta los conteos crudos por el largo total del gen y por los reads totales dentro de una muestra. Por lo tanto, este cálculo tiene como resultado una normalización dentro y entre muestras simultáneamente. Estos valores son muy útiles a la hora de analizar diferencias de abundancia entre transcriptos alternativos (que surgen de diferentes variantes de splicing), ya que la corrección por el largo de diferentes variantes es esencial para este tipo de análisis. Si se compara la expresión del mismo gen entre diferentes muestras, sólo es necesaria la normalización entre muestras. En este caso, alcanzaría simplemente con normalizar los conteos por el tamaño de la biblioteca, por ejemplo dividir el número de reads por gen entre el tamaño total de la biblioteca. Sin embargo este método no es muy estable cuando algunos pocos genes se llevan la mayor parte de los reads. Por este motivo surgieron métodos similares pero más robustos, como ser TMM (trimmed mean of M-values, en la figura 3.5 en la casilla de edgeR) o median count (en la figura 3.5 en la casilla de DESeq).

3.6.6. Expresión

A partir de los datos de conteo normalizados, se pueden realizar cálculos de expresión diferencial (caja rosada en figura 3.5). Aquí existen muchas posibilidades, en la figura 3.5 sólo mencionamos dos de las más utilizadas. Cada una posee sus ventajas y desventajas, que han sido comparadas en estudios recientes [141]. Gran parte de los cálculos de expresión de nuestro trabajo han sido realizados con edgeR [142], ya que presentó ciertas ventajas para nuestro caso. Por un lado, el mismo es un paquete de R, por lo que se puede integrar fácilmente con el resto de nuestro análisis. Por otro lado, edgeR considera muestras pareadas, es decir puede tomar en cuenta el efecto del paciente: si una muestra es tomada de un paciente X y la misma se somete a una condición A (tratamiento), y a otra muestra del mismo paciente X se le aplica una condición B, ambas muestras A y B se consideran pareadas, y las comparaciones serán internas al paciente. El tipo de muestras utilizadas en nuestros análisis eran justamente de ese tipo. edgeR modela los datos de conteo Y_{g_i} (conteos de un gen g en la muestra i) utilizando una distribución binomial negativa (que puede aproximarse a una Poisson con sobredispersión):

$$Y_{g_i} \sim NB(M_i p_{g_k}, \phi_g),$$

con M_i el tamaño de biblioteca de la muestra i (número total de reads), p_{g_k} es la abundancia relativa del gen g de la muestra i, que pertenece al grupo k de "tratamiento" (o condición), ϕ_g es la dispersión del gen g. La media de la distribución corresponde al primer parámetro $\mu_{g_i} = M_i p_{g_k}$ y la varianza es $\sigma^2_{g_i} = \mu_{g_i}(1 + \mu_{g_i}\phi_g)$. El parámetro de interés para el cálculo de expresión diferencial es p_{g_k}. La reducción a la distribución de Poisson se logra con $\phi_g = 0$, correspondiendo a $\mu_{g_i} = \sigma^2_{g_i}$. Las distribuciones con media igual a la varianza presentan una sobredispersión, que es muy común en este tipo de datos de conteo. La estimación de la varianza de cada gen se realiza con máxima verosimilitud condicional, siendo la condición la cantidad total de reads por cada gen. Se utiliza una aproximación bayesiana empírica para moderar la sobredispersión, llevándola hacia un valor consenso, "tomando prestada" información de la varianza total de los genes. Una vez estimados el nivel de expresión y varianza de cada gen, el test de expresión diferencial de una condición contra otra se realiza con un test de Fisher adaptado para datos de sobredispersión [143]. El resultado de este test es el estadístico de Fisher correspondiente y un p-value que será utilizado para filtrar los genes diferencialmente expresados con cierto nivel de significancia (en general denominado α).

El resto de los programas de expresión diferencial tienen otras formas de aproximar los conteos, de estimar las varianzas, y de hacer el test, etc. De cualquier manera, el resultado de todos ellos, será una lista de genes (o "features") con un valor de expresión asociado de una condición con respecto a la otra (en general expresado a través del denominado logFC), un p-value y un nivel de confianza α. Este último representa un nivel de rechazo de la hipótesis que se está testeando, la cual consiste en evaluar si el gen está diferencialmente expresado o no cambia significativamente de expresión entre condiciones. En este tipo de estudios en donde se evalúan una gran cantidad de genes (hipótesis) simultáneamente se debe corregir por comparaciones múltiples ("multiple testing"). Por ejemplo, si se fija un α de 0,05, implicando que con 5 % de probabilidad se va a rechazar la hipótesis nula (gen no está diferencialmente expresado) erróneamente, se va asignar un gen como diferencialmente expresado cuando no lo está. En el caso de analizar 100 genes, se espera que 5 estén mal clasificados. Pero en estudios reales, en donde se evalúan por lo menos 10000 genes (en general son más), se estarían clasificando erróneamente 500. Para controlar este fenómeno, existen varias técnicas de corrección, que intentan controlar lo que se denomina el "familywise error", el error general en el conjunto de hipótesis testeadas. Una de estas técnicas es la de Bonferroni [144], la cual es muy conservativa y consiste en disminuir el nivel de significancia a $\frac{\alpha}{n}$, siendo n el número de hipótesis testeadas (genes). De esta forma, al evaluar n genes, se estaría obteniendo un nivel de significancia general correspondiente al α original. Otra técnica de corrección de comparaciones múltiples es FDR por sus siglas en inglés (False Discovery Rate), también denominado Benjamini-Hochberg por sus autores [145]. Este procedimiento controla la

proporción de hipótesis nulas rechazadas erróneamente, es decir "falsos descubrimientos". FDR es un método menos estricto que los métodos que corrigen por el "familywise error", como el de Bonferroni. Este tipo de correcciones más permisivas son las más utilizadas en el contexto de nuestros datos. La tasa de falsos descubrimientos se define como

$$FDR = E(\frac{FP}{FP + TP}),$$

siendo FP los falsos positivos y TP los verdaderos positvos. Los FP son aquellos genes que fueron declarados diferencialmente expresados pero la hipótesis nula era válida, es decir no lo eran. Los TP son aquellos genes que están diferencialmente expresados y fueron clasificados como tal. Por lo tanto FDR es la proporción de los falsos positivos esperada, y este valor es el que se quiere mantener controlado, por debajo de por ejemplo 0,05.

El valor de expresión entre condiciones para cada uno de los genes es representado con el valor de logFC (log Fold Change), que compara el nivel de expresión de un gen g en la condición A y del mismo gen g en la condición B en escala logarítimica: $\log \frac{g_A}{g_B}$. Valores positivos implican mayor expresión en la condición A y valores negativos en la B. A partir de la lista de genes ("features") diferencialmente expresados pueden divergir los análisis posteriores ("downstream analysis"), dependiendo de la pregunta en cuestión. Existen análisis de redes metabólicas o de transcripción, análisis de grupos génicos de interés como GSEA (Gene Set Enrichment Analysis), entre otras tantas opciones.

La figura 3.5 muestra también una alternativa a la caja violeta de conteos. A la derecha de la misma y directamente seguido del *tophat2* se observa una caja celeste, conteos & expresión. Ésta cumple la misma función que la concatenación de las anteriores (violeta y rosada), pero con ciertas particularidades interesantes. Dentro de la suite del *tophat2*, no sólo existe la herramienta para el mapeo en sí, la cual considera spliced alignments y permite descubrir nuevos sitios de splicing, sino que se incluyen otras herramientas que realizan la estimación de los niveles de expresión (*cufflinks*) y el cálculo de expresión diferencial (*cuffdiff*). Las mismas pueden utilizarse de forma estándar, es decir, a partir de un archivo GTF/GFF pueden determinar los reads por "feature", y posteriormente, estimando varianzas entre condiciones, determinar expresión diferencial. Sin embargo, estas herramientas permiten estimar los niveles de expresión de variantes de splicing individuales, lo que es más complejo, ya que éstas pueden compartir varios exones, por lo que sólo una pequeña fracción de reads alineará contra las regiones distintas de cada variante. En general, se precisa una gran cobertura de secuenciación para obtener buenos resultados. El *cufflinks* (y también el *MiSO*) utiliza modelos estadísticos para estimar qué proporciones de reads son asignadas a qué variante determinada de splicing. El HTSeq no permitía este tipo de estimaciones, se focalizaba simplemente en las regiones codificantes. Una vez estimados los niveles de expresión de transcriptos alternativos,

cuffdiff (y también *MiSO* y *DEXSeq*) son capaces de determinar la expresión diferencial de transcriptos alternativos entre condiciones.
Los programas *cufflinks* y *cuffdiff* permiten, a su vez, descubrir nuevas "features" (transcriptos, exones), nuevos sitios de splicing alternativos, nuevos TSS (transcription start sites) a partir sólamente de los datos de secuenciación, y consecuentemente, se puede determinar expresión diferencial para todas las "features" encontradas, nuevas y ya establecidas.
Existen otro tipo de aplicaciones, no contempladas en la figura 3.5, que no precisan el cálculo de expresión, es decir, el foco no es medir genes. Por ejemplo, determinación de SNPs, análisis de mutaciones, ensamblado de genomas, entre otras.

3.6.7. microRNAs

En vez de estimar expresión de genes o transcriptos, se puede estar interesado en la expresión de ARN pequeños por ejemplo microRNAs, los cuales están teniendo cada vez más relevancia en la biología molecular. En este caso, los pasos a seguir son como los sugeridos en la figura 3.5, sobre todo en la parte de análisis de calidad, los adaptadores deben ser recortados y los reads sin adaptadores 3' deben de ser excluidos del análisis. El mapeo funciona con cualquiera de los programas planteados, sin embargo utilizar las opciones de spliced alignment por ejemplo no tendrían mucho sentido, ya que la probabilidad de que un microRNA caiga en una juntura exónica es muy baja. Los mismos se encuentran por lo general en regiones no codificantes. Por lo tanto, no habría necesidad de utilizar un *tophat2* para estos casos. Conteos, cálculos de expresión diferencial y normalización se realizan como fue detallado más arriba.

3.6.8. ChIP-Seq

Los estudios de ChIP-Seq parten de los mismos datos mencionados en la figura 3.5, y se aplican los mismos procedimientos de calidad y mapeo. Lo que varía en este caso es la forma de contar y evaluar expresión diferencial. Los estudios de ChIP-Seq evalúan expresión de segmentos específicos, aquellos segmentos que están unidos a una proteína de interés (histona, factor de transcripción, etc.). Para poder comparar abundancia de segmentos entre muestras, se debe primeramente realizar un procedimiento de identificación de picos ("peak calling"). El análisis computacional depende básicamente de la detección de estos picos, regiones del genoma en donde los reads se acumulan, sobrepasando el ruido de fondo, lo que representa la señal de unión de la proteína.
Diferentes procesamientos de los datos deben llevarse a cabo para generar una lista final

de picos. Estas listas son las que se comparan entre condiciones (por ejemplo, caso-control), para sacar conclusiones sobre la interacción ADN-proteína diferencial entre condiciones. Primeramente se debe estimar la señal de ruido ("background estimation"), identificación de picos enriquecidos, análisis estadístico de significancia de picos y eliminar artefactos. Cada uno de estos pasos puede tener parámetros que suelen ser ajustados por el usuario. Modificar estos parámetros puede afectar la lista de final de picos significativamente. Se han desarrollado una gran variedad de programas enfocados al análisis de ChIP-Seq, cada uno de los mismos tiene diferentes enfoques para cada uno de los pasos mencionados. Varias revisiones se han hecho en busca de los mejores programas de estimación y análisis de picos [146, 147].

3.6.9. Otras aplicaciones

Los análisis mencionados anteriormente son básicamente basados en estudios de abundancia (transcriptómica), tanto de genes, como de transcriptos, como de segmentos específicos, etc. Sin embargo, existe otra gama de aplicaciones, más bien genómicas, que realiza estudios de variantes genómicas, ya sean SNPs (Single Nucleotide Polymorphisms) o análisis de mutaciones. También se pueden realizar ensamblajes de genomas, transcriptomas, etc. Los pasos de trabajo para estos análisis difieren a los reportados en la figura 3.5 y se escapan al marco de este trabajo.

Con la metodología planteada en este capítulo se analizará la regulación transcripcional y post-transcripcional del proceso de diferenciación de las células madre mesenquimales a adipocitos. En una primera etapa se investigará si existe regulación post-transcripcional en el proceso de la adipogénesis y en caso afirmativo, en qué grado ocurre. Este estudio da lugar al primer artículo publicado en Stem Cell Research y desarrollado en el capítulo 4 de la tesis.
Paralelamente, debido a artículos publicados recientemente, la identidad de las células madre adultas obtenidas de tejido adiposo (las utilizadas en nuestros estudios) se ve comprometida. Se plantea que éstas pueden ser fibroblastos y no células madre. En un segundo trabajo intentamos resolver el problema de identidad de estas células. Este estudio dio lugar a un segundo artículo, el cual está en proceso de publicación en la revista Stem Cell & Development y el que se detallará en el capítulo 5 de la tesis.
Como consecuencia del primer artículo, en donde estudiamos la regulación post-transcripcional, surge un tercer trabajo en donde se propone a la poliadenilación alternativa como uno de los mecanismos responsables de dicha regulación. El estudio fue publicado en la revista PLoS One y se detallará en el capítulo 6 de la tesis.

A raíz de los resultados previos, se investiga qué otros mecanismos pueden estar involucrados en la regulación post-transcripcional de la adipogénesis. Se intentan dilucidar las diferencias sistemáticas que existen entre transcriptoma y translatoma durante este proceso, enfocándose sobre todo en los diferentes biotipos que aparecen sobre representados en las diferentes condiciones. Este último manuscrito está en preparación para ser enviado a BMC genomics y será detallado en el capítulo 7 de la tesis.
El trabajo se concluye con un último capítulo (8) de conclusiones generales.

Capítulo 4

Regulación post-transcripcional durante la adipogénesis a través de perfiles polisomales

4.1. Introducción

En este capítulo describiremos uno de los trabajos realizados durante esta tesis, basado mayormente en el análisis de los perfiles polisomales para investigar la regulación post-transcripcional en el proceso de adipogénesis a partir de datos de NGS. Mediante la comparación de la expresión génica de dos fracciones de ARN diferentes, ARN total y el ARN asociado a polisomas, se logró determinar las diferencias entre ambas y así estimar el grado de regulación post-transcripcional en este proceso. Por otro lado, un objetivo adicional de este artículo era determinar el tiempo del compromiso celular, es decir si ya a los tres días de la inducción (tiempo de toma de muestra) las células presentaban un compromiso al linaje adipocito. Utilizando las muestras de ARN polisomal y analizando la expresión diferencial de los genes se intenta determinar la activación de las vías relevantes.
Al final del capítulo se adjunta el artículo publicado en Stem Cell Research.

4.2. Perfiles polisomales y regulación post-transcripcional

Diversos análisis de comparación de ARN mensajero con niveles de proteínas en eucariotas han mostrado baja correlación entre niveles de transcripto y síntesis de proteínas. No

todo lo codificante que se está transcribiendo se traduce y lo que se traduce no está en directa proporción con la concentración de ARN mensajero, lo que indica la presencia de mecanismos de regulación post-transcripcional [148–150]. La expresión génica está siendo controlada por varios niveles adicionales de regulación. Los procesos de diferenciación celular no son la excepción. Varios estudios han ayudado a dilucidar las redes regulatorias detrás de los procesos de auto-renovación y de diferenciación [151–154]. Sin embargo, varios de ellos se han basado en el análisis de los niveles de ARN total en la célula, los cuales no reflejan de forma confiable los niveles de proteína, ni tienen en cuenta mecanismos de regulación post-transcripcional [153, 154]. Como fue sugerido por Tebaldi *et al* [150] existe un "desacoplamiento" entre el transcriptoma (ARNm total en la célula) y el translatoma (ARNm asociado a polisomas) en las células de mamíferos. Los mecanismos responsables del desacoplamiento, o de la regulación post-transcripcional, son variados y muchos desconocidos aún. Uno de ellos incluye la poliadenilación alternativa (APA) que detallaremos más adelante, en el capítulo 6.

Otro de los mecanismos es la regulación de la iniciación de la traducción, es decir la asociación de ARN mensajero a los ribosomas para formar polisomas [155, 156]. Este mecanismo puede controlarse en diferentes niveles complementarios para influenciar la formación de polisomas. Por un lado, se pueden regular los factores generales de iniciación de la traducción como ser eIF4E, eIF4G, eIF4A and PABP, que son responsables del desenrollado del ARN mensajero, de la asociación del ARNm con el ribosoma y del escaneado del mismo. Cualquier mecanismo que influencie estos factores puede comprometer la formación de los polisomas. Por otro lado, proteínas de unión al ARN (RBP: RNA binding proteins) y ARN no codificantes también pueden regular la iniciación de la traducción. Estos elementos pueden reconocer motivos de unión en las regiones 3' o 5'UTR y actuar a través de diferentes mecanismos, tanto influenciando la formación de polisomas directamente, como la establidad del ARN mensajero, la localización del mismo, etc. Los ARN no codificantes (ncRNA) también incluyen los micro ARN (miRNA), los cuales se detallarán en el siguiente capítulo en el contexto de APA.

Más recientemente se descubrieron mecanismos de regulación post-transcripcional adicionales, como ser los seudogenes. Los mismos fueron reportados como elementos regulatorios en cáncer [157], en diferenciación de células madre [158], entre otros [159]. Otros estudios recientes también han demostrado que los gránulos de ARN funcionan como moduladores clave de la regulación post-transcripcional [160, 161].

Estos son tan solo algunos de los mecanismos de regulación post-transcripcional conocidos al momento; la complejidad de las redes de regulación post-transcripcional es muy alta.

En este contexto, nos focalizamos en analizar el grado de regulación post-transcripcional en los procesos de diferenciación a adipocitos. Como se menciona más adelante en 4.4, las muestras fueron obtenidas de ARN total y de ARN asociado a polisomas. Los ARNs

asociados a polisomas deberían representar con mayor confiabilidad los niveles de proteínas, ya que los mismos están ya asociados a la maquinaria traduccional. Mediante la comparación de ARN total y polisomal se logra estimar el grado de regulación post-transcripcional en el proceso de adipogénesis.
Utilizando secuenciación masiva de varias muestras en ambas fracciones se logra comparar ambos "pools" de ARN y estimar el nivel de regulación post-transcripcional en las primeras etapas de la adipogénesis.

4.3. 3'UTR

Las 3'UTR de ARN mensajeros son regiones de gran interés debido a las funciones regulatorias controladas por las características composicionales y físicas de las mismas. Una de las características relevantes son los sitios de unión a los micro ARNs, los cuales reconocen motivos específicos y modulan la traducción, en general reprimiéndola (más detalles en el capítulo 6) [162, 163]. Existen otros sitios de unión relevantes como los correspondientes a proteínas silenciadoras [164], que también reprimen la traducción y contienen unos elementos ricos en AU (AREs: AU-rich elements). Proteínas que se unen a AREs afectan la estabilidad o la tasa de decaimiento de los transcriptos afectando el inicio de la traducción [165, 166].
El largo de las regiones 3'UTR varía mucho dentro del mismo organismo, por gen, por estadío celular, estímulos, etc. El largo promedio de las 3'UTR humanas se estima en 800 nucleótidos, mientras que, a modo de comparación, las 5'UTR tienen un largo promedio de 200 [167]. 3'UTR más largas contienen en general más elementos regulatorios, por lo que el estudio del largo es de gran interés.
En células murinas se observó que las 3'UTR de los genes presentaban una extensión progresiva durante el desarrollo embrionario [168]. Otros estudios muestran que células en estado de proliferación producen ARN mensajeros con 3'UTR más cortas, los cuales aparentan tener una vida media más larga [169]. Este estudio sugiere que los mensajeros son estabilizados por la falta de sitios de unión a micro ARNs, los cuales usualmente reprimirían la expresión. Por otro lado, Kolle *et al* en el 2012 observó una extensión de 3'UTR en ARN mensajeros de células madre embrionarias [126]. Ambos fenómenos (extensión y acortamiento) son observados en diferentes escenarios celulares. En nuestro estudio se observó que durante el proceso de diferenciación a adipocitos, los mensajeros generados presentan, tanto 3'UTR extendidas, como acortadas.

4.4. Las muestras utilizadas

Las muestras consideradas en este trabajo fueron células madre mesenquimales obtenidas de tres donadores diferentes. Las mismas son aisladas, cultivadas y caracterizadas de acuerdo a los protocolos establecidos. Se induce la adipogénesis con protocolos establecidos [170]. Las muestras se toman en el tiempo cero, antes de la diferenciación, por lo que son muestras de célula madre control (CT), y en el tiempo 72 horas posterior a la inducción. Estas últimas representan muestras inducidas (IN). Cada una de las muestras, tanto las CT como las IN, se aislaron de dos formas distintas, conformando un segundo factor de comparación. Por un lado, se obtuvo el ARN total de la muestra y por el otro lado, se aisló sólamente el ARN asociado a polisomas. Por este motivo, para cada muestra biológica se obtuvieron dos especímenes, uno corresponde al ARN total (total) y el otro a la fracción de ARN asociada a los polisomas (poli). Cada muestra fue secuenciada con el SOLiD 4 (Applied Biosystems), obteniéndose un total de 14 muestras: 8 CT (3 poli y 5 total) y 6 IN (3 poli y 3 total).
Una vez obtenidos los datos crudos, éstos fueron mapeados contra un genoma de referencia. Es este caso el mapeo (la localización de los segmentos de secuencia en el genoma) se realizó con el paquete de R denominado Rsubread, al igual que el conteo (determinar la cantidad de segmentos por característica, por ejemplo gen). Posteriormente, se llevó a cabo un análisis de calidad de las muestras, que relevó alta consistencia de los datos. Por más detalles sobre esto, referirse al artículo adjuntado a continuación (sobre todo figura 1 y la sección “Data analysis”).
A partir de estos datos se intentaron responder las siguientes preguntas.

4.5. Tiempo del compromiso celular

Las células madre mesenquimales (CMM) obtenidas de tejido adiposo presentan un excelente modelo para estudiar los procesos de diferenciación a adipogénesis. Las mismas son fáciles de obtener, a través de procedimientos poco invasivos y son relativamente fáciles de mantener y manipular en laboratorio. A su vez los protocolos para la diferenciación a adipocitos están bien determinados.
Por estas razones se utilizaron las CMM como modelo para estudiar el proceso de diferenciación a adipocitos. Una de las preguntas iniciales fue determinar el momento en el cual la CMM se compromete al linaje adipogénico. Si bien estudios han utilizado un período de tiempo de 21 días para la diferenciación [171–173], nuestras observaciones preliminares sugirieron que un tiempo menor ya es suficiente. Por este motivo se analizaron los perfiles de expresión génica de las muestras control (CT) contra las muestras inducidas (IN). Para estos análisis se tomaron en cuenta sólamente las muestras de ARN

polisomal, ya que el mismo refleja mejor el nivel de proteína de la célula (ver más adelante, 4.2).
El análisis de los genes diferencialmente expresados muestra que el proceso de adipogénesis ya está activado a nivel molecular a los 3 días de la inducción, aún cuando fenotípicamente no se observan cambios en la célula (acumulación de triglicéridos, cambio de morfología, etc). Por ejemplo, los genes cruciales reguladores de la adipogénesis, PPARγ, KLF15 y C/EBPα, se observan con una alta expresión. La vía del metabolismo lipídico también está sobre expresada, al igual que genes pertenecientes a la respuesta de insulina, incluyendo factores de crecimiento, receptores, proteínas de unión, etc.
Estos resultados sugieren un compromiso de la CMM al tercer día post inducción.

4.6. Resumen de resultados

Se realizó un agrupamiento jerárquico de las muestras para analizar la calidad de los datos. Las muestras se agruparon primero de acuerdo a condición, es decir CT e IN, y seguidamente por fracción, total y polisomal. Esto corrobora la alta calidad y consistencia de los datos.
A su vez, se realizaron análisis de ontologías GO (GO: gene ontology), tanto para los genes diferencialmente expresados en la fracción polisomal como para los de la fracción total. Los resultados mostraron grandes diferencias de los genes pertenecientes cada una de las fracciones, sobre todo en función molecular (MF: molecular function). En la fracción polisomal se observaron mayormente funciones moleculares relacionadas con procesos inflamatorios, clásicos de la diferenciación (citoquinas, interleuquinas, etc.) y actividad de óxido-reductasa, mientras que en total se observaron funciones relacionadas con la membrana y matriz extracelular. Por otro lado, genes que manifiestan cambios en ARN total (genes codifantes para ribosomas y proteínas del citoesqueleto) no muestran cambios en la fracción polisomal. Se corroboró que varios transcriptos que estaban diferencialmente expresados en total y no lo estaban en polisomal, no presentaban cambios en las proteínas. Estos resultados fueron confirmados con datos previamente publicados de proteómica y por análisis propios de Western Blot. El defasaje de funciones corrobora el desacoplamiento entre transcriptoma y translatoma, sugiriendo la presencia de mecanismos de regulación post-transcripcional que controlan la expresión génica durante la adipogénesis.
Para medir el grado de regulación post-transcripcional en este proceso, hemos cuantificado el número de genes que cambia significativamente la expresión en una fracción y no en la otra. De esta forma se obtiene que de los genes que experimentan algún cambio, 44 % cambia sólo en ARN total y no en polisomal y 13 % cambia solo en polisomal. Estos resultados implican que 57 % de los genes presentaron algún tipo de regulación

post-transcripcional (ver figura 4 del manuscrito).
Por otro lado, observamos el compromiso a nivel molecular de la célula al linaje adipocito ya a los tres días de la inducción. Esto se observó en la expresión diferencial de los genes pertenecientes a las vías relevantes en la adipogénesis: genes de la vía de la insulina, PPARγ, C/EBP, etc. Esta observación fue confirmada experimentalmente (ver figura 2 en el artículo). Las CMM fueron inducidas durante 3 días después de lo cual se retiro el medio inductor y se observo que las CMM eran capaces de completar la diferenciación celular.
Adicionalmente, se observó que la diferenciación trae consigo cambios importantes en las regiones 3'UTR de los transcriptos. Utilizando un procedimiento planteado por Kolle y colaboradores [126] para estimar el largo real de las 3'UTR basado en los reads que caen en el extremo 3' del transcripto, se observó que éstas sufren tanto extensiones como acortamientos. Se ejemplifica este fenómeno con dos genes adipogénicos claves como el FABP4 y el WNT2. El FABP4 presenta una extensión de 3'UTR de casi 1kb en las muestras inducidas con respecto a las control. El WNT2 posee el comportamiento contrario, las muestras inducidas presentan un acortamiento de las 3'UTR con respecto al control. Estos resultados se muestran en la figura 5 del artículo.
Hemos por lo tanto observado un alto grado de regulación post-transcripcional durante la diferenciación a adipocitos, a través de los experimentos realizados con el SOLiD.

4.7. Artículo

Stem Cell Research (2013) 11, 902–912

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/scr

Polysome profiling shows extensive posttranscriptional regulation during human adipocyte stem cell differentiation into adipocytes

Lucia Spangenberg[b,1], Patricia Shigunov[a,1], Ana Paula R. Abud[a], Axel R. Cofré[a], Marco A. Stimamiglio[a], Crisciele Kuligovski[a], Jaiesa Zych[a], Andressa V. Schittini[a], Alexandre Dias Tavares Costa[a], Carmen K. Rebelatto[c], Paulo R.S. Brofman[c], Samuel Goldenberg[a], Alejandro Correa[a], Hugo Naya[b], Bruno Dallagiovanna[a,*]

[a] *Instituto Carlos Chagas, Fiocruz-Paraná, Rua Professor Algacyr Munhoz Mader, 3775, Curitiba-PR 81350-010, Brazil*
[b] *Unidad de Bioinformática, Institut Pasteur Montevideo, Mataojo 2020, Montevideo 11400, Uruguay*
[c] *Núcleo de Tecnologia Celular, Pontifícia Universidade Católica do Paraná, Rua Imaculada Conceição, 1155, Curitiba-PR 80215-901, Brazil*

Received 22 December 2012; received in revised form 29 May 2013; accepted 2 June 2013
Available online 10 June 2013

Abstract Adipocyte stem cells (hASCs) can proliferate and self-renew and, due to their multipotent nature, they can differentiate into several tissue-specific lineages, making them ideal candidates for use in cell therapy. Most attempts to determine the mRNA profile of self-renewing or differentiating stem cells have made use of total RNA for gene expression analysis. Several lines of evidence suggest that self-renewal and differentiation are also dependent on the control of protein synthesis by posttranscriptional mechanisms. We used adipogenic differentiation as a model, to investigate the extent to which posttranscriptional regulation controlled gene expression in hASCs. We focused on the initial steps of differentiation and isolated both the total mRNA fraction and the subpopulation of mRNAs associated with translating ribosomes. We observed that adipogenesis is committed in the first days of induction and three days appears as the minimum time of induction necessary for efficient differentiation. RNA-seq analysis showed that a significant percentage of regulated mRNAs were posttranscriptionally controlled. Part of this regulation involves massive changes in transcript untranslated regions (UTR) length, with differential extension/reduction of the 3′UTR after induction. A slight correlation can be observed between the expression levels of differentially expressed genes and the 3′UTR length. When we considered association to polysomes, this correlation values increased. Changes in the half lives were related to the extension of the 3′UTR, with longer UTRs mainly stabilizing the transcripts. Thus, changes in the length of these extensions may be associated with changes in the ability to associate with polysomes or in half-life.
© 2013 Elsevier B.V. All rights reserved.

* Corresponding author. Fax: +55 41 33163267.
E-mail address: brunod@tecpar.br (B. Dallagiovanna).
[1] Both authors contributed equally to this work.

1873-5061/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.scr.2013.06.002

Introduction

Human adipose tissue-derived stromal cells (hASC) are readily isolated from the pools of cells resident in the vascular stroma of adipose tissue. ASCs proliferate and self-renew and, due to their multipotent nature, they can differentiate at least in vitro into several tissue-specific lineages, including the chondrogenic, osteogenic, adipogenic and miogenic lineages (De Ugarte et al., 2003; Gimble et al., 2007). Adipose tissue is ubiquitous and large quantities are easily accessible with minimal invasion procedures (Baer and Geiger, 2012). These characteristics make these cells ideal candidates for use in cell therapy. An understanding of the biological process committing the cell to differentiation into a specific cell type is essential for the successful repair of injured tissue.

Cytokines, growth factors and extracellular matrix components in the microenvironment determine stem cell fate, by regulating the switch from self-renewal to differentiation (Kratchmarova et al., 2005). However, the downstream effectors and the gene regulatory networks controlling these processes remain unclear. Gene expression profiling has provided insight into the molecular pathways involved in ASC self-renewal and differentiation (Ivanova et al., 2002; Song et al., 2006). Genome-wide analyses based on microarray hybridization and, more recently, next generation sequencing, have been carried out to assess the global expression of gene networks.

Most attempts to determine the mRNA profile of self-renewing or differentiating cells have made use of total RNA for hybridization to microarrays or RNA-Seq analysis (Jeong et al., 2007a; Menssen et al., 2011). High-throughput analyses in eukaryotes comparing mRNA and protein levels have indicated that there is no direct correlation between transcript levels and protein synthesis, suggesting a high degree of posttranscriptional regulation in eukaryote cells (Washburn et al., 2003; Keene, 2007; Tebaldi et al., 2012). This hampers the classical transcriptome-based approach to investigate controlled expression in differentiating cells. Protein abundance can be controlled and refined through the regulation of gene expression at various complementary levels. Several lines of evidence from different organisms suggest that stem cell self-renewal and differentiation are also dependent on the control of protein synthesis by posttranscriptional mechanisms (Keene, 2007; Sampath et al., 2008; Haston et al., 2009; Kolle et al., 2011). The analysis of the mRNA fraction associated to polysomes has been used as a strategy to analyze posttranscriptional mechanisms involved in the control of translation (Fromm-Dornieden et al., 2012). This posttranscriptional regulation is mediated by various molecules, such as microRNAs, noncoding RNAs and RNA binding proteins. *Trans*-acting factors recognize and bind sequences or structural elements, mostly in the untranslated regions (UTRs) of mRNAs (Mittal et al., 2009; Bar et al., 2008; Keene, 2010). Posttranscriptional control may be mediated by, amongst other things, modifications to mRNA stability or by the inhibition of transcript association with translating ribosomes.

We used adipogenic differentiation as a model, to investigate the extent to which posttranscriptional regulation controlled gene expression in hASCs. We focused on the initial steps of cell differentiation and isolated both the total mRNA fraction and the subpopulation of mRNAs associated with translating ribosomes. RNA-seq analysis showed that a significant percentage of regulated mRNAs were controlled both at the translational level and by changes to their half lives. Part of this regulation is associated with differential extension/reduction of the 3′UTR after induction.

Materials and methods

Isolation, culture, and differentiation of hASCs

Stem cells were obtained from adipose tissue from obese human donors (two males and one female, ages: 41, 52, 23). All samples were isolated, collected after informed consent had been obtained, in accordance with guidelines for research involving human subjects, and with the approval of the Ethics Committee of Fundação Oswaldo Cruz, Brazil (approval number 419/07). hASCs were isolated, cultured and characterized as previously described (Rebelatto et al., 2008). Briefly, 100 ml of adipose tissue was washed with sterile phosphate-buffered saline (PBS) (Gibco Invitrogen). A one-step digestion by 1 mg/ml collagenase type I (Invitrogen) was performed for 30 min at 37 °C during permanent shaking and was followed by filtration through first a 100- and then 40-μm mesh filter (BD FALCON, BD Biosciences Discovery Labware, Bedford, MA, USA). The cell suspension was centrifuged at 800 *g* for 10 min, and contaminating erythrocytes were removed by erythrocyte lysis buffer, pH 7.3. The cells were washed and then cultivated at a density of 1×10^5 cells/cm^2 in T75 culture flasks in DMEM-F12 (Gibco Invitrogen) supplemented with 10% FCS, penicillin (100 units/ml), and streptomycin (100 μg/ml). The medium was changed 2 days after the initial plating. The culture medium was then replaced twice each week. ASCs were subcultured after the cultures had reached 80% to 90% confluence; cells were detached by treatment with 0.25% trypsin/EDTA (Invitrogen) and were replated as passage-1 cells (the process was then continued). The characterization of the cells has been done following the minimal criteria for defining multipotent mesenchymal stromal cells as determined by the International Society for Cellular Therapy (Dominici et al., 2006). All tests were performed with cell cultures at passages 3 to 5. For adipogenic differentiation, hASCs were treated with hMSC Adipogenic Differentiation Bullet Kit (Lonza), in accordance with the manufacturer's instructions. Briefly, adipogenic differentiation was induced by 6 day cycles of induction/maintenance during 21 days. Induction medium contained the adipogenic inducers insulin, dexamethasone, indomethacin and IBMX; maintenance medium contained insulin. The medium was changed every 3 days. The degree of adipogenic differentiation was determined by assessing the cytoplasmic accumulation of triglycerides by staining with Oil Red O or Nile Red (Sigma-Aldrich), as described by Rebelatto et al. (2008). We also performed reverse transcription-polymerase chain reaction (RT-PCR) to estimate the amount of adipocyte-specific fatty acid-binding protein 4 (FABP4) mRNA. A list of the primers used is provided in Supporting Information Table S1.

Sucrose density gradient separation and RNA purification

Polysomal fractions were prepared with a modified version of the procedure described by Holetz et al. (2007). In brief,

hASC cultures at 50 to 60% confluence were treated with 0.1 mg/ml cycloheximide (Sigma-Aldrich) for 10 min at 37 °C. The cells were removed from the culture flasks with a cell scraper and resuspended in 0.1 mg/ml cycloheximide in PBS. The suspension was centrifuged (2000 ×*g* for 5 min) and the resulting pellet was washed twice with 0.1 mg/ml cycloheximide in PBS. The cells were lysed by incubation for 10 min on ice with polysome buffer (15 mM Tris–HCl pH 7.4, 1% Triton X-100, 15 mM $MgCl_2$, 0.3 M NaCl, 0.1 μg/ml cycloheximide, 1 mg/ml heparin). The cell lysate was centrifuged at 12,000 ×*g* for 10 min at 4 °C. The supernatant was carefully isolated, loaded onto 10% to 50% sucrose gradients and centrifuged at 39,000 rpm (SW40 rotor, HIMAC CP80WX HITACHI) for 160 min at 4 °C. The sucrose gradient was fractionated with the ISCO gradient fractionation system (ISCO Model 160 gradient former), connected to a UV detector for the monitoring of absorbance at 275 nm, and the polysome profile was recorded. The total and polysomal RNA fractions were extracted by a standard Trizol (Invitrogen) RNA isolation protocol.

cDNA library construction and RNA sequencing (RNA-Seq)

We subjected total and polysome-associated RNA samples to amplification with the Amino Allyl Message Amp II aRNA Amplification Kit (Ambion), to provide a template for SOLiD libraries. The cDNA libraries were prepared with the SOLiD Whole Transcriptome Analysis Kit and the purified products were evaluated with an Agilent Bioanalyzer (Agilent). Library molecules were subjected to clonal amplification according to the SOLiD Full-Scale Template Bead preparation protocol and sequenced with the SOLiD4 System (Applied Biosystems).

Data analysis

Quality control analysis was performed on the sequencing data, with NGSQC (Dai et al., 2010) software. Various quality parameters were explored visually for each sample (distribution of colors per sample/tile, genomic hit count per sample with different numbers of mismatches, sequencing read density and a quality score based on the mean values of the preceding values for each sample). All samples passed the quality control filter. Mapping and counting were performed with the R package Rsubread (Liao et al., 2013). Hierarchical clustering of the samples (log of counts plus one) was performed, to evaluate biological variability. Each sample was normalized to one million reads to account for library size. We also conducted a correspondence analysis (COA), a dimension reduction method, to the matrix of counts, to explore associations between variables. In COA it is possible to visualize samples and genes simultaneously, revealing associations between them. Genes, or samples, lying close to each other tend to behave similarly.

For the comparison of induced stem cells with undifferentiated stem cells we retained only those genes with counts of more than 1 per million in at least three conditions. In comparisons between induced stem cells and undifferentiated cells differentially expressed (DE) genes were identified with the edgeR bioconductor package (Robinson et al., 2010). This set of genes was used for GO term analysis with the goseq bioconductor package (Young et al., 2010).

The analysis of 3′UTR extension/shortening was performed according to the approach presented by Kolle et al. (2011), basically a sliding window of length w starting from last exon runs through the 3′UTR region (in steps of length s), each time determining expression of the window. If expression decays more than 50% the sliding stops and the 3′UTR ends. The parameters used in this case: w = 100 and s = 50. The analysis was restricted to the DE genes. A comparative analysis with proteomic data was performed against the murine data set of Molina et al. (2009). They have studied the proteomics of adipocyte differentiation on 3T3–L1 murine preadipocytes in four different time points (days 1, 3, 5 and 7) and determined differentially expressed genes on two sets of proteins, nuclear and secreted ones. The protein levels of both sets of genes were compared with the respective logFC determined by our differential expression analysis (in both cases polysomal and total RNA). Pearson correlation test was performed and significant results with the secreted protein set were obtained. Raw data has been submitted to the ArrayExpress repository under accession number E-MTAB-1366.

RT-qPCR

RT-PCR and real-time quantitative PCR (qPCR) were performed as previously described (Rebelatto et al., 2008). Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) transcript was used as internal control. Experiments were performed with cells from at least three donors, with technical triplicates. Student's *t*-test was used to assess the significance of differences between the cell populations. We considered p-values < 0.05 to be statistically significant.

Reduced glutathione (GSH) determination

Cells were washed twice in PBS and centrifuged at low speed. Pelleted cells were suspended in lysis buffer (15 mM Tris–HCl, 15 mM $MgCl_2$, 300 mM NaCl, 1% Triton X-100; pH 7.4) and placed on ice for 10 min. Cell lysate was centrifuged at 12,000 ×*g* for 10 min and the supernatant was used to determine the level of reduced GSH. An aliquot (150 μl) of the supernatant was added to the reaction medium (25 μl of 3 mM DTNB [5,5′-dithiobis-(2-nitrobenzoic acid)] plus 125 μl of methanol) and absorbance was determined at 412 nm on a microplate reader (BioTek®). We calculated the concentration of reduced GSH, using the molar extinction coefficient of DTNB in solution (ε = 13,600 M^{-1} cm^{-1}). Results are expressed in μM GSH per 5 × 10^5 cells. All results are expressed as means ± SEM. The significance of the differences observed was evaluated by ANOVA, with Tukey's post hoc test. $p < 0.01$ and $p < 0.05$ were considered statistically significant. Raw data is shown in the Supplemental material Dataset S3.

Estimation of mRNA decay rates

For the measurement of mRNA stability, transcription was blocked by adding actinomycin D (Sigma) to the medium at a concentration of 10 mg/ml. The mean relative transcript levels estimated by RT-qPCR at each time point after the addition of Act D were used to estimate the half-life of the

transcript from a first-order decay model, according to the equation, $\gamma = \beta_0 e^{\beta t}_1 + \varepsilon$, where γ is the mean relative amount of mRNA at time t after the addition of Act D, β_0 is the initial quantity, β_1 is a decay parameter related to half-life ($t1/2 = \pm \ln 2 / \beta_1$) and ε is an error term (Sharova et al., 2009).

Results

Adipogenesis is committed in the first days of induction

We characterized the patterns of gene expression involved in the initial steps of adipogenesis. All accepted protocols use at least three days of strong induction to promote differentiation into adipocytes so we allowed hASCs to differentiate in vitro in the presence of adipogenic induction media for 72 h. All the experiments described were performed with at least three samples of hASCs from different donors, all used in early passages. After three days of in vitro differentiation, the cells displayed no clear change in phenotype or accumulation of lipids in the cytoplasm (Supporting information Figs. S1A, B). The cells were collected for isolation of total and polysomal RNA. The hASCs had a polysome profile typical of cells with low levels of translation activity, with low concentrations of polysome complexes present throughout the gradient. In the first few days of adipogenesis, the cells displayed no significant change in overall polysome profile (Supporting information Figs. S1C, D).

We isolated polysome-associated mRNAs from the gradient fractions corresponding to polysomes. The total and polysomal mRNA fractions of hASCs were analyzed by RNA-Seq, with the SOLiD4 system. In both cases, we compared time points 0 and 72 h after adipogenic induction. The total number of reads obtained for each sample is shown in Supporting information Table S2. The reads of all samples were mapped onto the reference genome (Hg19; NCBI Build 37.64), yielding a mean mapping percentage of approximately 54%. Hierarchical clustering shows that samples cluster first as a function of conditions (control, induced) and then by RNA fraction, rather than by donor, indicating that total and polysomal RNA populations are intrinsically more characteristic than donor "idiosyncrasy" (Fig. 1A). Correspondence Analysis (COA) produced similar results (Fig. 1B), with the first axis separating samples according to fraction and the second by conditions. Most transcripts were detected in both fractions (though with different levels of expression) while a significant number of transcripts were only present in one of the RNA populations. This could reflect the existence of regulatory mechanisms which modulate the efficiency of association with polysomes (Fig. 1C). We filtered out genes with very low counts, reducing the number of genes retrieved to just over 15,000 per comparison (see Array Express E-MTAB-1366 and Supporting information Table S2). Furthermore, we identified the differentially expressed (DE) genes using this reduced set

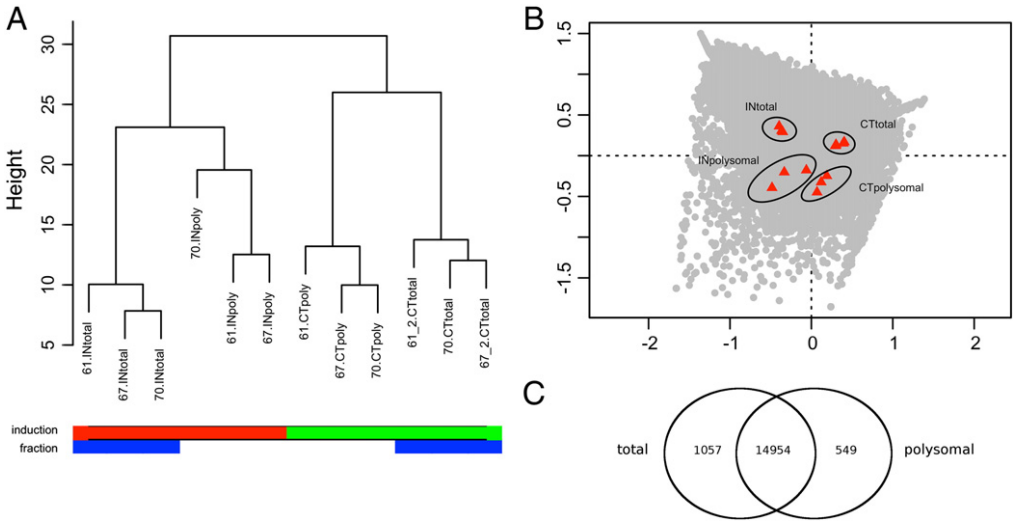


Figure 1 Results of hierarchical clustering and correspondence analysis showing the internal consistency of the data. (A): A dissimilarity based (bottom-up) hierarchical clustering was performed on the log-transformed counts of genes for the various samples. Initially, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, according to some distance measure (in this case, complete linkage approach), continuing until there is one single cluster. The process is visualized as a dendrogram: the first branching event separates control (CT) from induced (IN) samples (IN left, CT right). Subsequent branching events group the samples according to the RNA fraction ("poly" and "total"). The numbers of the samples (61, 67 and 70) correspond to the donors. The height axis represents the distance between each branching event. Condition and fraction accounted for the largest proportions of the variance in both analyses, highlighting the consistency of the experiments. (B): Correspondence analysis (COA) on the samples. The x-axis represents the first component (the one explaining the most variance 43.43%) and the y-axis the second component (representing 29.91% of the variance). Fraction (polysomal and total) appears to be represented in the y-axis and the culture conditions (CT and IN) in the x-axis. Four groups of samples are represented on the reduced (two-dimensional) space. They correspond to the four RNA populations: CT-polysomal, CT-total, IN-polysomal, IN-total. (C): Venn diagram showing the overlap between the genes detected in both conditions: polysomal and total RNA populations. 14,954 genes are common to both sets.

(edgeR, paired comparisons). We compared the control and induced states, for the polysomal RNA fraction and for total RNA. We identified 2918 DE genes in total RNA and 780 DE genes in the polysomal fraction (FDR 0.001; Supporting information Datasets S1 and S2). Gene expression values (logFC) of both RNA fractions show Spearman correlation value of 0.62 (Supporting information Fig. S2). The relative efficiency of association with polysomes was also determined in both control and induced cells (Supporting information Table S3).

Differential expression was confirmed by RT-qPCR on selected transcripts (Supporting information Fig. S3). To also confirm the differential association with polysomes we performed sucrose density fractionation of polysomes. The presence of the selected transcripts in the RNA populations present in polysomes, monosomes and ribosome-free pooled fractions was analyzed by RT-qPCR. Differential expression was associated with a shift in the association of mRNAs with the different ribosome fractions for almost all transcripts tested but one (Supporting information Fig. S4).

More detailed analysis of the DE genes in the polysomal fraction showed a strong upregulation of adipogenesis-related genes after three days of induction. Key transcriptional regulators of the onset of adipogenesis, such as PPARG, KLF15 and CEBPα, showed an increase of several fold in their transcript levels (Supporting information Dataset S2). We also detected the expression of lipid metabolism-related genes and genes of the insulin response network encoding growth factors, receptors and binding proteins. These results suggest that adipogenesis is already triggered in the first few days of induction. Standard protocols are based on continuous or alternate induction, with the medium changed every three days and differentiation allowed to occur for up to 21 days. We tested our hypothesis, by inducing the cells for only three days and then allowing them to complete differentiation without further induction. The cells differentiated fully, but their fitness differed from that of cells induced for a continuous period of 21 days, and this difference was not the same for all considered donors (Fig. 2A). This raised questions about the true minimum period of stimuli required for the commitment of the cells to differentiation. Induction kinetics showed that three days was the critical time for the induction of efficient adipogenesis, although some degree of differentiation was observed with shorter induction times (Supporting information Table S4).

We observed the upregulation mainly in the polysomal RNA fraction of several genes involved in glutathione homeostasis in the cell. It has been reported that GSH decrement results in enhanced C/EBPβ activation, which is a key event in the first phase of adipogenesis, resulting in the activation of downstream PPARγ and a more rapid acquirement of adipose phenotype in 3T3–L1 cells (Vigilanza et al., 2011). For the confirmation of these results, we measured GSH levels in both control and induced cells. We found that differentiating cells contained 30% less GSH than non-induced cells (Fig. 2B).

Polysome associated mRNAs show extensive posttranscriptional regulation

GO analysis was performed using DE genes on the polysomal and total fractions (Fig. 3). Most of the GO terms underrepresented in both fractions were involved in nucleic acid metabolism and nuclear functions (Supporting information Tables S5, S6 and S7).

We obtained 18 overrepresented GO terms (FDR 0.001) in the analysis of DE genes in the total fraction. Most of these terms are related to proteins from the extracellular space and plasma membrane, with functions involved in cell adhesion, cell signaling, receptor activity and extracellular matrix proteins.

We also obtained 18 overrepresented terms in the analysis of polysomal fraction samples. Some of the GO terms obtained were similar to those for the total fraction (e.g. receptor activity, Figs. 3A, B), whereas several others were specific to the polysomal samples (oxidoreductase activity, Fig. 3A). Furthermore, some terms for which changes in mRNA levels were observed in the total fraction (genes encoding ribosomal or actin cytoskeleton proteins) showed no difference when polysomal RNA fraction was analyzed. In particular, no difference in polysome association was observed for transcripts encoding proteins involved in the development of the nervous system and in cell differentiation, despite the increase in the

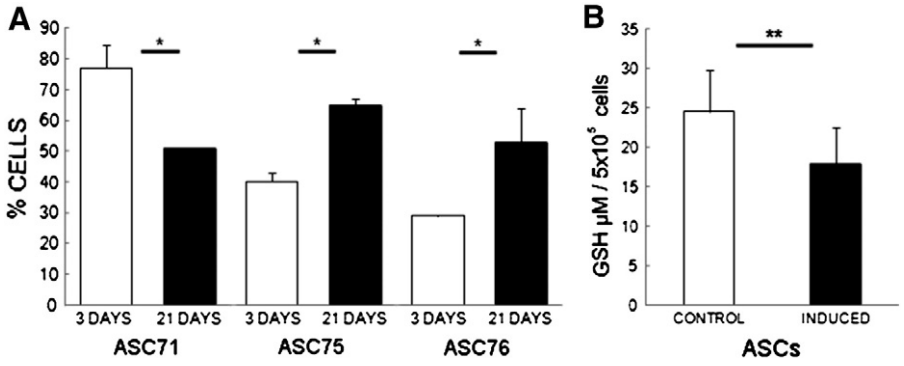


Figure 2 Morphological and metabolic changes during the differentiation process. (A): Differentiation potential analysis of ASCs: cells were maintained in differentiation-inducing conditions for 3 (white) or 21 days (black) and allowed to complete adipogenesis. Columns represent the percentage of differentiated cells in the cultures after 21 days. (B): Reduced glutathione (GSH) concentration in the induced cells (black column) was 30% lower than that in the control group (white column). The data shown are from three independent experiments carried out with ASCs from three donors. The results are expressed as means ± SD. The significance of differences between mean values was evaluated by a two-way analysis of variance (ANOVA) followed by a Tukey test. p values ≤ 0.01 and ≤ 0.05 were considered statistically significant.

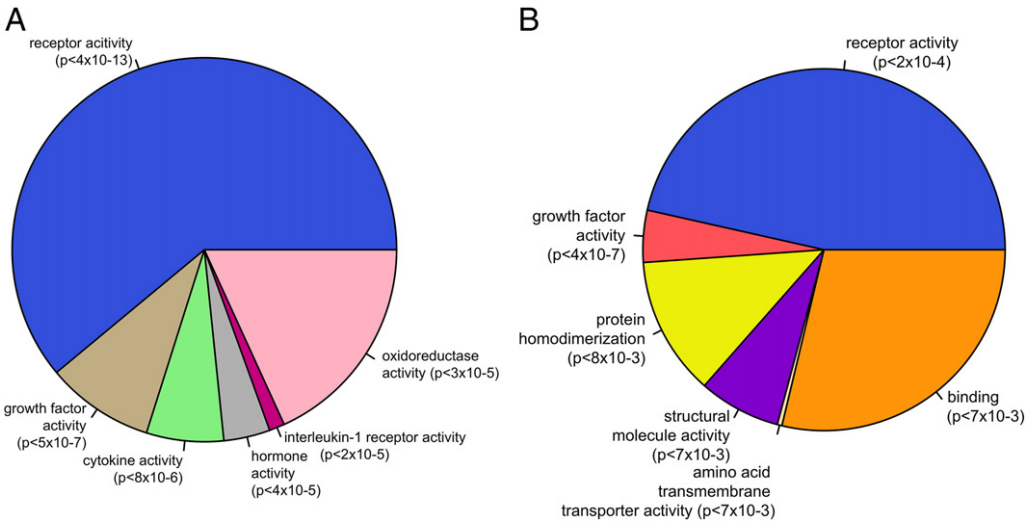


Figure 3 GO analysis of two sets of differentially expressed genes: IN vs CT in the polysomal RNA fraction (A) and IN vs CT in total RNA (B). Only overrepresented Molecular Function (MF) GO terms are shown in each pie chart in (A) and (B). For each over represented MF GO term its corresponding adjusted p-values are shown.

levels of transcripts for these proteins in the total fraction. To confirm our results we compared our data with previously published proteomic data. Proteomic analysis of human ASCs undergoing adipogenesis showed the differential expression of receptors and membrane proteins (Jeong et al., 2007b) and proteins related to oxido-reductase activity and oxidative stress response (Kim et al., 2010; Kheterpal et al., 2011). Moreover, we performed correlation analysis with the proteomic data of Molina et al. (2009) who presented a quantitative analysis (SILAC) of adipogenesis in murine 3T3–L1 cells. Molina's experiment determined protein expression levels during adipocyte differentiation at four time points (days 1, 3, 5 and 7) and for two sets of proteins: nuclear (280) and secreted (147) proteins. After determining the corresponding human orthologs we retrieved our logFC values of the comparison of IN wrt CT in the polysomal RNA fraction. No significant correlation was observed for the nuclear proteins set correlation values were not significant in each of the time points. This was expected as since we have observed a downregulation of nuclear proteins in our assays. However, the set of secreted proteins is highly correlated to our data for almost every time point and the correlation is also statistically significant (Supplementary information Fig. S5A). To further confirm our results, protein expression was also measured by western blot with protein extracts from control (CT) and induced (IN) cells. The amount of IGFBP2 protein was clearly increased in induced cells corroborating the differential expression of the insulin signaling pathway after induction. As mentioned previously, many transcripts that showed differential expression in the total fraction but not in their association with polysomes showed no changes when protein levels were analyzed (Supporting information Fig. S5B). These findings suggest that posttranscriptional mechanisms may have a key role in the control of gene expression during adipogenesis.

We compared differential expression between the total and polysomal RNA fractions, to assess the degree of posttranscriptional regulation, by an approach similar to that used by Lundberg et al. (2010). Using the differential expression values shown in Datasets S1 and S2, we compared the samples in the polysomal and total RNA fractions, using a cutoff value of $|\log FC| > 1.5$. Fig. 4A shows logFC values obtained from the comparison of induced vs. control samples of the polysomal fraction against the logFC values obtained from the comparison of total RNA samples. Most of the genes identified displayed no change in expression level during differentiation (87.65%). However, 5.25% of the transcripts displayed changes in both RNA fractions, and another 5.45% of the transcripts displayed changes in steady-state RNA levels with no change in the amount of transcript associated with translating ribosomes. Finally, 1.66% of the transcripts displayed differential mobilization to the polysomes during cell differentiation. If restricted only to changing genes (in either fraction), almost half of the genes showed differential expression in both conditions (about 43%), about 13% were DE only in the polysomal fraction and about 44% were DE only in the total fraction (Fig. 4B). Our results demonstrate the existence of posttranscriptional mechanisms regulating the association of mRNAs with translating ribosomes, irrespectively of a change in steady-state levels. Hence, we observe an extensive posttranscriptional regulation during the initial steps of adipogenesis.

Cell differentiation involves massive changes in transcript UTR length

Recent studies have reported that the transcripts expressed in human and murine embryonic stem cells have alternative 3′UTRs (Kolle et al., 2011; Ji et al., 2009). It has been suggested that the 3′UTRs of mRNAs increase in length with the progression of embryonic development in mouse embryonic stem cells (Ji et al., 2009). We considered the reads mapped onto putative extended 3′UTR to detect possible changes during the induction of adipogenesis. We found that the differentially expressed genes displayed changes in the length of their 3′UTRs during this process. By contrast to the findings in mouse cells, we observed both lengthening and shortening of the 3′UTRs of these transcripts (Dataset S4).

The 3′UTRs of mRNAs potentially contain *cis*-acting elements for the posttranscriptional regulation of gene expression. Thus,

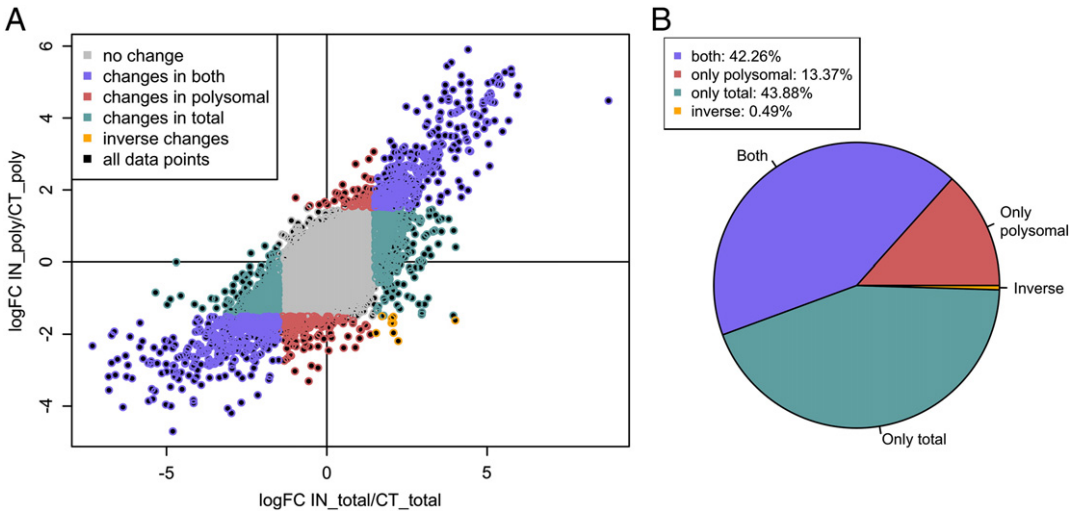


Figure 4 Differential association of mRNAs to polysomes: (A): LogFC values from different RNA fractions were compared. The logFC values (IN vs. CT) for the polysomal fraction (y-axis) were plotted against the logFC values for the total RNA fraction (x-axis). The data points are colored according to the change in the fractions. Genes displaying changes only in the polysomal fraction include genes with a |logFC| of at least 1.5 (|logFC| > 1.5) and genes with a |logFC| < 1.5 in the total fraction (light red). Changes in the total RNA fraction were associated with a |logFC| >1.5 and of |logFC| < 1.5 in the polysomal fraction (light green). Genes displaying changes in expression in both sets of conditions had high (or low) values of logFC (greater than 1.5/lower than −1.5) in both RNA fractions (violet). Inverse changes included genes with high logFC values (logFC > 1.5) in the total fraction and low values (logFC < 1.5) in the polysomal fraction, or vice versa (orange). (B): The pie chart on the right shows the percentages of genes displaying changes in expression in each of the four categories: change only in the polysomal fraction, change only in the total RNA fraction, concordant changes in the two fractions, discordant changes in the two fractions. Only the changing genes are considered.

changes in the length of these extensions may be associated with changes in the ability to associate with polysomes or in half-life. A slight correlation can be observed between the logFC of DE genes (total fraction) and the difference of 3′UTR length in control and induced samples. The correlation values are 0.28, 0.29, and 0.22 for donors 67, 61 and 70, respectively (Supporting information Table S8A). When we considered association to polysomes, these correlation values increase, 0.32, 0.35 and 0.25 respectively (Supporting information Table S8B). We also looked for a biological relationship between the different subsets of transcripts grouped by extension/shortening of the UTR and up or downregulation. Interestingly, we observed that the subset of downregulated transcripts presenting a longer 3′UTR after induction was related to the response to unfolded proteins and stress (Supplementary information Table S9).

We investigated the relationship between changes in UTR length and mRNA stability, by measuring the steady-state levels of the transcripts in control and induced cells. We focused on two examples, one extension (FABP4) and one shortening (WNT2) of transcript 3′UTRs which are differentiation markers or regulators of adipogenesis (Figs. 5A,D). Treatment of the cells with actinomycin D revealed changes in the half lives of these transcripts, which were directly related to the extension of the UTR in the two situations analyzed, with longer UTRs stabilizing the transcripts (Figs. 5 and S6). The results are summarized in Supporting information Table S10.

Discussion

Posttranscriptional regulation mechanisms are now considered to play a key role in the control of gene expression. Regulation occurs mainly through the modulation of mRNA half life or the formation of a translational initiation complex, allowing the assembly of translating polysomes (Keene, 2007).

We used adipogenesis as a cell differentiation model to study the regulation of gene expression in hASCs. The identification of mRNAs associated with polysomes could provide us with a clearer idea of which genes are actually translated into proteins in differentiating cells. By deep sequencing, we showed that adipogenesis had been triggered at the molecular level after three days of induction, although the cells did not yet display any clear phenotypic changes. Analysis of polysome associated mRNAs in the first hours of differentiation of 3T3 pre-adipocytes showed the upregulation of a discrete number of genes, mainly related with the overall control of translation (Fromm-Dornieden et al., 2012). We demonstrated, after three days of adipogenic induction, an upregulation of the expression of networks of genes involved in adipocyte differentiation (Menssen et al., 2011). Three days was identified as the minimum induction time required for the initiation of adipogenesis, as shorter induction times resulted in lower percentages of mature adipocytes. Moreover, no additional induction was required to achieve differentiation, suggesting that the cells can sustain their own differentiation, once committed. Differentiating hASCs contain high levels of expression of prolactin (PRL) (See Dataset S2). Studies with murine preadipocytes have demonstrated that fetal bovine serum (FBS), which is known to contain large amounts of prolactin (Ginsburg and Vonderhaar, 1995), is required for the efficient induction of adipogenesis and that PRL can replace FBS in the differentiation of the 3T3–L1 preadipocyte cell line (Stewart et al., 2004). PRL has also been shown to act as an adipogenesis-enhancing hormone,

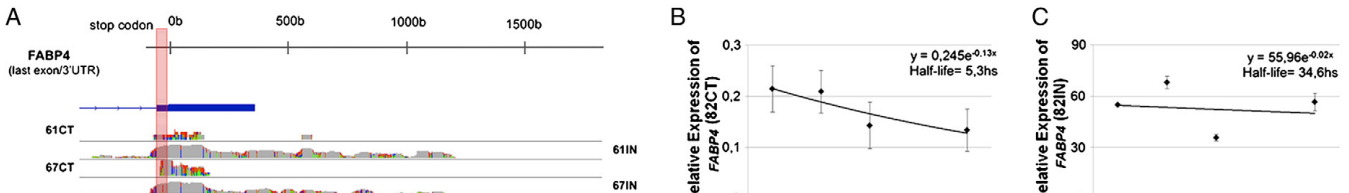


Figure 5 The coverage vectors of the 3′UTR of two genes and the half-life of the mRNA. (A): Six samples (3 CT and 3 IN) are represented as coverage vectors (log-scaled) for the *FABP4* gene. Only experiments using polysomal RNA were considered. The induced and control samples are interspersed (e.g. 61 CT and 61 IN). Induced samples of show an extension the 3′UTR (~1 kb). The red vertical bar shows the stop codon, the horizontal blue bars represent the last portion of the gene structure (last exon and 3′UTRs). (D): Same analysis for the *WNT2* gene. In this case, induced samples show a shortening of the 3′UTR, whereas control samples have a longer 3′UTR. (B–C, E–F): Half-life of *FABP4* and *WNT2* mRNAs in control (CT) and induced (IN) hASCs. hASCs were treated with Act-D for various periods, to block mRNA synthesis. Total RNA isolation, cDNA generation and real-time PCR amplification were performed as described in the text. The values shown are means ± SD of *FABP4* and *WNT2* RNA copy number per μg of total RNA, from two independent experiments carried out in triplicate.

enhancing the expression of the key transcriptional regulators of adipogenesis, C/EBPβ and PPARγ (Nanbu-Wakao et al., 2000). Interestingly, for bone marrow stromal cells, the induction of adipogenesis causes a dose-dependent increase in prolactin receptor expression, suggesting a role for prolactin and its receptor in adipocyte differentiation (McAveney et al., 1996). Studies in knockout mice have shown that an absence of PRL receptor (PRLR) signaling compromises the growth and differentiation of brown adipose tissue and that immortalized PRLR knockout preadipocytes do not differentiate into mature adipocytes, this defect being reversed by the reintroduction of PRLR (Viengchareun et al., 2008). Moreover, the production of PRL mRNA and protein increases markedly during the early differentiation of primary human preadipocytes (Hugo et al., 2006).

Clustering analysis of the total and polysomal RNA fractions showed these two fractions to be clearly different, regardless of donor origin. GO analysis revealed enrichment in different terms in the two populations. Nucleic acid metabolism-related genes were clearly downregulated in both fractions. As this category includes genes encoding regulatory proteins, further studies are required to understand the impact of this negative control on MSC differentiation. Genes encoding membrane-related and receptor proteins were upregulated in both RNA populations, suggesting that the principal changes in MSC biology during differentiation are related to the response to external stimuli and signaling.

A comparison of the total RNA and polysome-associated fractions provides a global estimate of the degree of posttranscriptional regulation, at least in terms of the control of translation initiation (Halbeisen and Gerber, 2009). Almost 60% of the differentially expressed genes showed some kind of posttranscriptional regulation. In most cases, this regulation counterbalances fluctuations in total RNA levels. Thus, transcripts increasing or decreasing in abundance in the cell are recruited to polysomes in equal amounts in differentiating cells. It was also observed a subset of transcripts that, although overexpressed in both fractions, showed a higher fold change in the polysomal RNA fraction. This has been described as a mechanism of homodirectional co-regulatory mechanism resulting in the amplification or potentiation of the positive control of gene expression (Preiss et al., 2003). However, there is a subpopulation of mRNAs that is regulated solely at the translational level.

Specific groups of related genes were found to display differential expression mostly in the polysomal fraction. In particular, oxidative stress response genes and a family of genes encoding proteins involved in the response to changes in the levels of reduced glutathione (GSH), such as glutathione peroxidases (GPxs) (Brigelius-Flohe, 1999), aldo-keto reductases (Barski et al., 2008) and metallotioneins (Nielsen et al., 2007), displayed expression patterns of this type. Oxidative stress may occur as a result of adipogenic differentiation or lipid metabolism, and the upregulation of these genes may contribute to a protective response. GPxs are involved in various reactions, such as the reduction of organic hydroperoxides (ROOH) by GSH conjugation (Arteel and Sies, 2001). We found that the concentration of reduced glutathione (GSH) was 30% lower in induced than in control cells. This higher level of activity of enzymes involved in GSH homeostasis may counteract the higher ROS concentrations in differentiated cells (Yang et al., 2012).

Posttranscriptional regulation is usually mediated by the interaction of *trans*-acting factors with the UTR regions. Thus, 3′UTRs appear to play a key role in posttranscriptional regulation, as the spatial platform bearing the sequence or structural elements in *cis* that are targeted by regulatory factors (Mignone et al., 2002). Specific modulations of UTR length in tissues have been reported, based on either the use of different polyadenylation signals or alternative splicing (Zhang et al., 2005). The length of 3′UTRs has been shown to differ between embryonic stem cells and somatic cells, in both humans and mice. In human embryonic stem cells, some transcripts have 3′UTRs several kilobases longer than the reported length (Kolle et al., 2011). Proliferating cells produce mRNAs with shorter UTRs, which have longer half-lives. It has been suggested that transcripts may be stabilized by the loss of miRNA binding sites, which usually downregulate gene expression (Sandberg et al., 2008). By contrast, in murine stem cells, the UTRs of tissue-specific transcripts increase in length following cell commitment. This extension results from the use of distal polyadenylation sites and results in an increase in the half life of the mRNA, by an unknown mechanism (Ji et al., 2009). Moreover, in brown and white adipose tissues, Ptitsyn and Gimble (2007) reported the existence of oscillatory patterns of expression of SOCS3 and JAK transcripts. Shorter and longer transcripts oscillate in opposite phases. These transcripts were generated by alternative polyadenylation in response to circadian rhythms. The relationship between alternative polyadenylation, circadian rhythms and posttranscriptional regulation mechanisms like mRNA decay needs to be determined. Nocturnin, a circadian deadenylase, enhances adipogenesis via interaction with PPAR-γ and could be a possible link between these mechanisms (Green et al., 2007; Kawai et al., 2010). However, we didn't detect Nocturnin among the DE genes which are in accordance with previous results showing that Nocturnin is not upregulated by insulin induction in bone-marrow stem cells (Kawai et al., 2010). We observed a shortening of some 3′UTRs and an extension of others following the induction of cell differentiation. Even though we recognize the limitations of our approach to estimate the actual length of the 3′UTR (biased wrt gene expression, no estimation of sampling "holes", arbitrary parameters) we obtain a first overview of the situation; many 3′UTRs are changing sizes, in both directions (extension/shortening). We found some hints supporting that the length of the UTR is directly related to the stability of the mRNA and to enhanced association to polysomes. As reported for mouse stem cells, we found a bias towards longer 3′UTRs resulting in more stable transcripts though this is not a general rule. A preliminary analysis of the alternative UTRs strongly suggests that they resulted from differential polyadenylation. Epigenetic mechanisms as histone acetylation could be involved in the choice or selection for alternative polyadenylation sites resulting in new 3′UTRs that could also regulate transcription. In 3T3 preadipocytes the 3′UTR of C/EBPβ acts as a strong enhancer element as a result of differential histone acetylation (Zhang et al., 2012). In chondrocytes, the 3′UTR of the Col2A1 gene is also a potent enhancer factor. It interacts with the promoter region through gene looping resulting in upregulation of gene expression (Jash et al., 2012). The mechanisms involved in the selection of alternative polyadenylation sites remain to be identified. It seems likely that extension of the 3′UTRs results in the

presence of new and additional regulatory elements, both positive and negative in effect.

Conclusions

Our results show that extensive posttranscriptional regulation occurs during the adipogenic differentiation of hASCs. Analysis of polysome associated transcripts showed that adipogenesis is committed after three days of induction. Differentially expressed transcripts showed else shortening or extension of their 3′UTRs. This modification in the extension of the 3′UTRs could be associated to mechanisms acting on both RNA stability and translation. The coordination of different levels of regulation ensures the efficient and correct differentiation of stem cells, and an elucidation of these mechanisms is required for an adequate understanding of the determination of stem cell fate.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.scr.2013.06.002.

Acknowledgments

This work was supported by grants from *Ministério da Saúde* and *Conselho Nacional de Desenvolvimento Cientifico e Tecnológico* — CNPq, *FIOCRUZ-Pasteur Research Program* and *Fundação Araucária*. L.S. received fellowship from ANII (Agencia Nacional de Investigación e Innovación, Uruguay); S.G., J.Z. and B.D. from *CNPq*, P.S., A.C. and M.A.S. from *FIOCRUZ*.

References

Arteel, G.E., Sies, H., 2001. The biochemistry of selenium and the glutathione system. Environ. Toxicol. Pharmacol. 10, 153–158.

Baer, P.C., Geiger, H., 2012. Adipose-derived mesenchymal stromal/stem cells: tissue localization, characterization, and heterogeneity. Stem Cells Int. 2012, 812693.

Bar, M., Wyman, S.K., Fritz, B.R., Qi, J., Garg, K.S., Parkin, R.K., Kroh, E.M., Bendoraite, A., Mitchell, P.S., Nelson, A.M., Ruzzo, W.L., Ware, C., Radich, J.P., Gentleman, R., Ruohola-Baker, H., Tewari, M., 2008. MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. Stem Cells 26, 2496–2505.

Barski, O.A., Tipparaju, S.M., Bhatnagar, A., 2008. The aldo-keto reductase superfamily and its role in drug metabolism and detoxification. Drug Metab. Rev. 40, 553–624.

Brigelius-Flohe, R., 1999. Tissue-specific functions of individual glutathione peroxidases. Free Radic. Biol. Med. 27, 951–965.

Dai, M., Thompson, R.C., Maher, C., Contreras-Galindo, R., Kaplan, M.H., Markovitz, D.M., Omenn, G., Meng, F., 2010. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics 11 (Suppl. 4), S7.

De Ugarte, D.A., Morizono, K., Elbarbary, A., Alfonso, Z., Zuk, P.A., Zhu, M., Dragoo, J.L., Ashjian, P., Thomas, B., Benhaim, P., Chen, I., Fraser, J., Hedrick, M.H., 2003. Comparison of multi-lineage cells from human adipose tissue and bone marrow. Cells Tissues Organs 174, 101–109.

Dominici, M., Le Blanc, K., Mueller, I., Slaper-Cortenbach, I., Marini, F., Krause, D., Deans, R., Keating, A., Prockop, D., Horwitz, E., 2006. Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement. Cytotherapy 8, 315–317.

Fromm-Dornieden, C., von der Heyde, S., Lytovchenko, O., Salinas-Riester, G., Brenig, B., Beissbarth, T., Baumgartner, B.G., 2012. Novel polysome messages and changes in translational activity appear after induction of adipogenesis in 3T3–L1 cells. BMC Mol. Biol. 13, 9.

Gimble, J.M., Katz, A.J., Bunnell, B.A., 2007. Adipose-derived stem cells for regenerative medicine. Circ. Res. 100, 1249–1260.

Ginsburg, E., Vonderhaar, B.K., 1995. Prolactin synthesis and secretion by human breast cancer cells. Cancer Res. 55, 2591–2595.

Green, C.B., Douris, N., Kojima, S., Strayer, C.A., Fogerty, J., Lourim, D., Keller, S.R., Besharse, J.C., 2007. Loss of Nocturnin, a circadian deadenylase, confers resistance to hepatic steatosis and diet-induced obesity. Proc. Natl. Acad. Sci. U. S. A. 104, 9888–9893.

Halbeisen, R.E., Gerber, A.P., 2009. Stress-dependent coordination of transcriptome and translatome in yeast. PLoS Biol. 7, e1000105.

Haston, K.M., Tung, J.Y., Reijo Pera, R.A., 2009. Dazl functions in maintenance of pluripotency and genetic and epigenetic programs of differentiation in mouse primordial germ cells in vivo and in vitro. PLoS One 4, e5654.

Holetz, F.B., Correa, A., Avila, A.R., Nakamura, C.V., Krieger, M.A., Goldenberg, S., 2007. Evidence of P-body-like structures in *Trypanosoma cruzi*. Biochem. Biophys. Res. Commun. 356, 1062–1067.

Hugo, E.R., Brandebourg, T.D., Comstock, C.E., Gersin, K.S., Sussman, J.J., Ben-Jonathan, N., 2006. LS14: a novel human adipocyte cell line that produces prolactin. Endocrinology 147, 306–313.

Ivanova, N.B., Dimos, J.T., Schaniel, C., Hackney, J.A., Moore, K.A., Lemischka, I.R., 2002. A stem cell molecular signature. Science 298, 601–604.

Jash, A., Yun, K., Sahoo, A., So, J.S., Im, S.H., 2012. Looping mediated interaction between the promoter and 3′ UTR regulates type II collagen expression in chondrocytes. PLoS One 7, e40828.

Jeong, J.A., Ko, K.M., Bae, S., Jeon, C.J., Koh, G.Y., Kim, H., 2007a. Genome-wide differential gene expression profiling of human bone marrow stromal cells. Stem Cells 25, 994–1002.

Jeong, J.A., Ko, K.M., Park, H.S., Lee, J., Jang, C., Jeon, C.J., Koh, G.Y., Kim, H., 2007b. Membrane proteomic analysis of human mesenchymal stromal cells during adipogenesis. Proteomics 7, 4181–4191.

Ji, Z., Lee, J.Y., Pan, Z., Jiang, B., Tian, B., 2009. Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc. Natl. Acad. Sci. U. S. A. 106, 7028–7033.

Kawai, M., Green, C.B., Lecka-Czernik, B., Douris, N., Gilbert, M.R., Kojima, S., Ackert-Bicknell, C., Garg, N., Horowitz, M.C., Adamo, M.L., Clemmons, D.R., Rosen, C.J., 2010. A circadian-regulated gene, Nocturnin, promotes adipogenesis by stimulating PPAR-gamma nuclear translocation. Proc. Natl. Acad. Sci. U. S. A. 107, 10508–10513.

Keene, J.D., 2007. RNA regulons: coordination of post-transcriptional events. Nat. Rev. Genet. 8, 533–543.

Keene, J.D., 2010. The global dynamics of RNA stability orchestrates responses to cellular activation. BMC Biol. 8, 95.

Kheterpal, I., Ku, G., Coleman, L., Yu, G., Ptitsyn, A.A., Floyd, Z.E., Gimble, J.M., 2011. Proteome of human subcutaneous adipose tissue stromal vascular fraction cells versus mature adipocytes based on DIGE. J. Proteome Res. 10, 1519–1527.

Kim, J., Choi, Y.S., Lim, S., Yea, K., Yoon, J.H., Jun, D.J., Ha, S.H., Kim, J.W., Kim, J.H., Suh, P.G., Ryu, S.H., Lee, T.G., 2010. Comparative analysis of the secretory proteome of human adipose stromal vascular fraction cells during adipogenesis. Proteomics 10, 394–405.

Kolle, G., Shepherd, J.L., Gardiner, B., Kassahn, K.S., Cloonan, N., Wood, D.L., Nourbakhsh, E., Taylor, D.F., Wani, S., Chy, H.S., Zhou, Q., McKernan, K., Kuersten, S., Laslett, A.L., Grimmond,

S.M., 2011. Deep-transcriptome and ribonome sequencing redefines the molecular networks of pluripotency and the extracellular space in human embryonic stem cells. Genome Res. 21, 2014–2025.

Kratchmarova, I., Blagoev, B., Haack-Sorensen, M., Kassem, M., Mann, M., 2005. Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. Science 308, 1472–1477.

Liao, Y., Smith, G.K., Shi, W., 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 41 (10), e108.

Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Algenas, C., Lundeberg, J., Mann, M., Uhlen, M., 2010. Defining the transcriptome and proteome in three functionally different human cell lines. Mol. Syst. Biol. 6, 450.

McAveney, K.M., Gimble, J.M., Yu-Lee, L., 1996. Prolactin receptor expression during adipocyte differentiation of bone marrow stroma. Endocrinology 137, 5723–5726.

Menssen, A., Haupl, T., Sittinger, M., Delorme, B., Charbord, P., Ringe, J., 2011. Differential gene expression profiling of human bone marrow-derived mesenchymal stem cells during adipogenic development. BMC Genomics 12, 461.

Mignone, F., Gissi, C., Liuni, S., Pesole, G., 2002. Untranslated regions of mRNAs. Genome Biol. 3 (REVIEWS0004).

Mittal, N., Roy, N., Babu, M.M., Janga, S.C., 2009. Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. Proc. Natl. Acad. Sci. U. S. A. 106, 20300–20305.

Molina, H., Yang, Y., Ruch, T., Kim, J.W., Mortensen, P., Otto, T., Nalli, A., Tang, Q.Q., Lane, M.D., Chaerkady, R., Pandey, A., 2009. Temporal profiling of the adipocyte proteome during differentiation using a five-plex SILAC based strategy. J. Proteome Res. 8, 48–58.

Nanbu-Wakao, R., Fujitani, Y., Masuho, Y., Muramatu, M., Wakao, H., 2000. Prolactin enhances CCAAT enhancer-binding protein-beta (C/EBP beta) and peroxisome proliferator-activated receptor gamma (PPAR gamma) messenger RNA expression and stimulates adipogenic conversion of NIH-3 T3 cells. Mol. Endocrinol. 14, 307–316.

Nielsen, A.E., Bohr, A., Penkowa, M., 2007. The balance between life and death of cells: roles of metallothioneins. Biomark. Insights 1, 99–111.

Preiss, T., Baron-Benhamou, J., Ansorge, W., Hentze, M.W., 2003. Homodirectional changes in transcriptome composition and mRNA translation induced by rapamycin and heat shock. Nat. Struct. Biol. 10, 1039–1047.

Ptitsyn, A.A., Gimble, J.M., 2007. Analysis of circadian pattern reveals tissue-specific alternative transcription in leptin signaling pathway. BMC Bioinformatics 8 (Suppl. 7), S15.

Rebelatto, C.K., Aguiar, A.M., Moretao, M.P., Senegaglia, A.C., Hansen, P., Barchiki, F., Oliveira, J., Martins, J., Kuligovski, C., Mansur, F., Christofis, A., Amaral, V.F., Brofman, P.S., Goldenberg, S., Nakao, L.S., Correa, A., 2008. Dissimilar differentiation of mesenchymal stem cells from bone marrow, umbilical cord blood, and adipose tissue. Exp. Biol. Med. (Maywood) 233, 901–913.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Sampath, P., Pritchard, D.K., Pabon, L., Reinecke, H., Schwartz, S.M., Morris, D.R., Murry, C.E., 2008. A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation. Cell Stem Cell 2, 448–460.

Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A., Burge, C.B., 2008. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. Science 320, 1643–1647.

Sharova, L.V., Sharov, A.A., Nedorezov, T., Piao, Y., Shaik, N., Ko, M.S., 2009. Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. DNA Res. 16, 45–58.

Song, L., Webb, N.E., Song, Y., Tuan, R.S., 2006. Identification and functional analysis of candidate genes regulating mesenchymal stem cell self-renewal and multipotency. Stem Cells 24, 1707–1718.

Stewart, W.C., Baugh Jr., J.E., Floyd, Z.E., Stephens, J.M., 2004. STAT 5 activators can replace the requirement of FBS in the adipogenesis of 3 T3-L1 cells. Biochem. Biophys. Res. Commun. 324, 355–359.

Tebaldi, T., Re, A., Viero, G., Pegoretti, I., Passerini, A., Blanzieri, E., Quattrone, A., 2012. Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. BMC Genomics 13, 220.

Viengchareun, S., Servel, N., Feve, B., Freemark, M., Lombes, M., Binart, N., 2008. Prolactin receptor signaling is essential for perinatal brown adipocyte function: a role for insulin-like growth factor-2. PLoS One 3, e1535.

Vigilanza, P., Aquilano, K., Baldelli, S., Rotilio, G., Ciriolo, M.R., 2011. Modulation of intracellular glutathione affects adipogenesis in 3T3–L1 cells. J. Cell. Physiol. 226, 2016–2024.

Washburn, M.P., Koller, A., Oshiro, G., Ulaszek, R.R., Plouffe, D., Deciu, C., Winzeler, E., Yates III, J.R., 2003. Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. U. S. A. 100, 3107–3112.

Yang, S.R., Rahman, I., Trosko, J.E., Kang, K.S., 2012. Oxidative stress-induced biomarkers for stem cell-based chemical screening. Prev. Med. 54, S42–S49 (Suppl.).

Young, M.D., Wakefield, M.J., Smyth, G.K., Oshlack, A., 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 11, R14.

Zhang, H., Lee, J.Y., Tian, B., 2005. Biased alternative polyadenylation in human tissues. Genome Biol. 6, R100.

Zhang, Q., Ramlee, M.K., Brunmeir, R., Villanueva, C.J., Halperin, D., Xu, F., 2012. Dynamic and distinct histone modifications modulate the expression of key adipogenesis regulatory genes. Cell Cycle 11, 4310–4322.

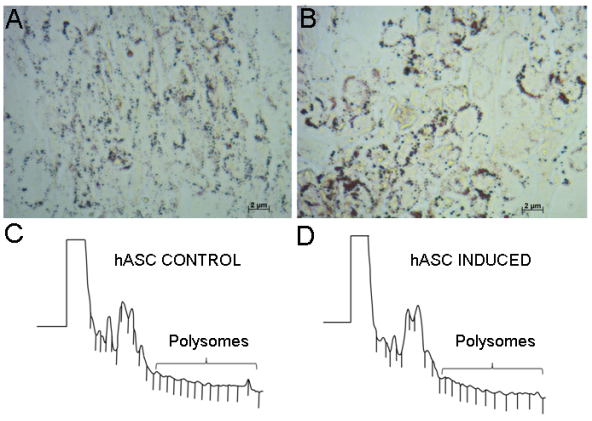


Figure S1

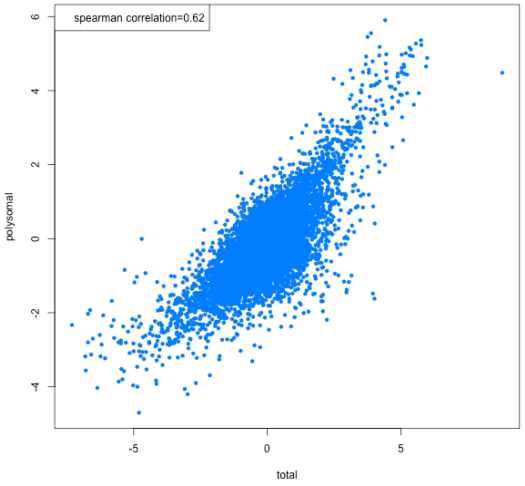


Figure S2

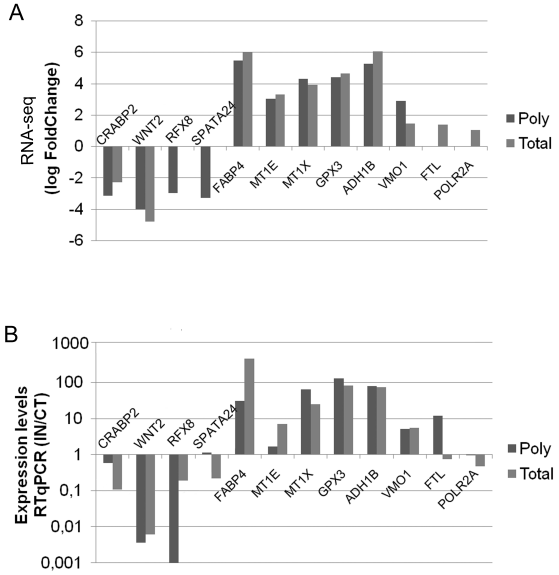


Figure S3

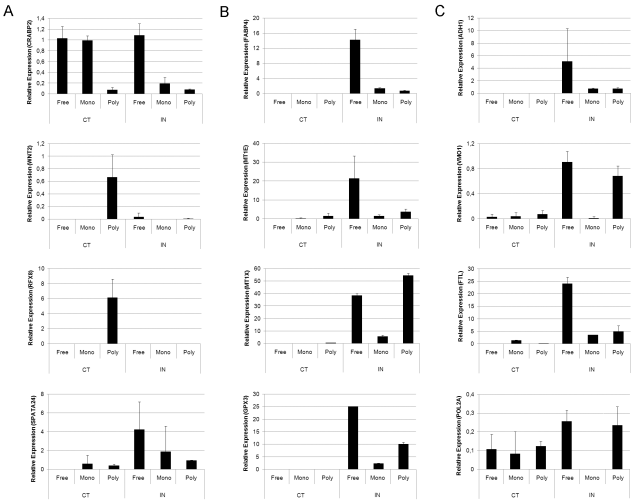


Figure S4

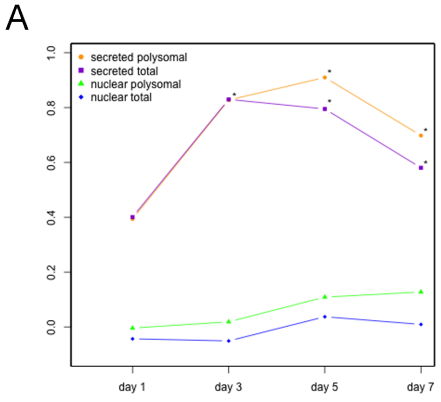


Figure S5

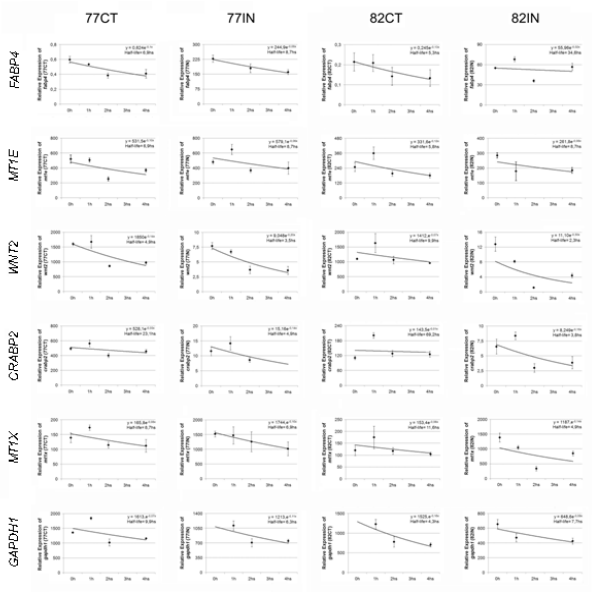


Figure S6

OVERREPRESENTED GO TERM (TOTAL: IN vs CT)				
type	**GO ID**	**p-adjusted**	**# genes***	**term**
CC	GO:0005576	2,11E-41	371	extracellularregion
CC	GO:0005886	7,50E-25	586	plasma membrane
CC	GO:0005615	6,36E-21	148	extracellularspace
CC	GO:0005887	4,83E-16	208	integral to plasma membrane
CC	GO:0005578	1,96E-10	48	proteinaceousextracellularmatrix
CC	GO:0031012	2,34E-08	23	extracellularmatrix
CC	GO:0016021	1,43E-04	840	integral tomembrane
CC	GO:0015629	3,50E-04	32	actincytoskeleton
CC	GO:0009986	2,05E-03	45	cellsurface
CC	GO:0045095	2,89E-03	13	keratinfilament
CC	GO:0030017	5,22E-03	5	sarcomere
BP	GO:0007155	1,75E-08	97	celladhesion
BP	GO:0051384	1,23E-05	13	response toglucocorticoidstimulus
BP	GO:0007165	1,52E-04	237	signaltransduction
BP	GO:0008284	1,74E-04	68	positive regulationofcellproliferation
BP	GO:0007275	2,39E-04	190	multicellularorganismaldevelopment
BP	GO:0007399	2,47E-04	76	nervoussystemdevelopment
BP	GO:0007169	4,84E-04	10	transmembrane receptor proteintyrosinekinasesignalingpathway
BP	GO:0042493	4,89E-04	50	response to drug
BP	GO:0030154	5,40E-04	97	celldifferentiation
BP	GO:0032496	7,61E-04	19	response tolipopolysaccharide
BP	GO:0007565	8,53E-04	9	femalepregnancy
BP	GO:0007267	2,05E-03	44	cell-cellsignaling
BP	GO:0006955	2,68E-03	62	immune response
BP	GO:0008152	2,72E-03	47	metabolicprocess
BP	GO:0016337	2,89E-03	16	cell-celladhesion
BP	GO:0045785	4,39E-03	3	positive regulationofcelladhesion
BP	GO:0001501	5,16E-03	37	skeletalsystemdevelopment
BP	GO:0030855	5,22E-03	6	epithelialcelldifferentiation
BP	GO:0009653	6,99E-03	18	anatomicalstructuremorphogenesis
BP	GO:0045765	7,09E-03	2	regulationofangiogenesis
BP	GO:0045766	8,13E-03	6	positive regulationofangiogenesis
MF	GO:0004872	8,10E-11	256	receptor activity
MF	GO:0008083	4,22E-07	33	growth factor activity
MF	GO:0005509	4,03E-06	120	calciumionbinding
MF	GO:0017147	2,20E-04	2	Wnt-proteinbinding
MF	GO:0004930	2,39E-04	74	G-proteincoupled receptor activity
MF	GO:0003779	3,22E-03	63	actinbinding
MF	GO:0008201	6,99E-03	19	heparinbinding
MF	GO:0015171	6,99E-03	3	amino acidtransmembranetransporteractivity
MF	GO:0005198	7,66E-03	52	structuralmoleculeactivity
MF	GO:0042803	7,88E-03	88	proteinhomodimerizationactivity
UNDERREPRESENTED GO TERMS (TOTAL: IN vs CT)				
type	**GO ID**	**p-adjusted**	**# genes***	**term**
CC	GO:0005622	<1E-100	423	intracellular
CC	GO:0005634	<1E-100	1109	nucleus
CC	GO:0005840	1,23E-04	39	ribosome
CC	GO:0005730	3,03E-04	160	nucleolus
CC	GO:0003735	5,79E-04	35	structuralconstituentofribosome

CC	GO:0005654	3,97E-03	117	nucleoplasm
CC	GO:0005739	4,09E-03	253	mitochondrion
BP	GO:0045449	4,01E-05	219	regulationoftranscription
BP	GO:0006397	1,23E-04	45	mRNAprocessing
BP	GO:0016568	1,23E-04	58	chromatinmodification
BP	GO:0006412	2,93E-04	44	translation
BP	GO:0008380	2,93E-04	52	RNA splicing
BP	GO:0006355	1,18E-03	225	regulationoftranscription, DNA-dependent
MF	GO:0003677	<1E-100	321	DNA binding
MF	GO:0003723	1<E-100	127	RNA binding
MF	GO:0008270	1,18E-03	429	zinc ionbinding
MF	GO:0046872	3,97E-03	620	metal ionbinding
MF	GO:0003676	6,02E-03	56	nucleicacidbinding

Table S6:
GO analysisofdifferentiallyexpressed genes fromthecomparison IN vs CT (total fraction). Cellularcomponent (CC), biologicalprocess (BP), molecular function (MF) are shown, includingthecorrespondingadjusted p-value, thenumberof genes falling in thatcategoryandthetermdescription.
MF terms are displayed as pie charts in figure 3B.
* Some gene overlapexistbetween GO categories.

OVERREPRESENTED GO TERM (POLYSOMAL: IN vs CT)				
type	**GO ID**	**p-adjusted**	**# genes***	**term**
CC	GO:0005576	1,16E-33	58	extracellularregion
CC	GO:0005615	1,76E-15	30	extracellularspace
CC	GO:0005887	3,24E-11	28	integral to plasma membrane
CC	GO:0005886	6,19E-09	89	plasma membrane
CC	GO:0005578	3,72E-07	9	proteinaceousextracellularmatrix
CC	GO:0016324	3,09E-04	3	apical plasma membrane
CC	GO:0031012	7,57E-04	8	extracellularmatrix
BP	GO:0007586	3,72E-07	1	digestion
BP	GO:0007584	2,93E-04	4	response tonutrient
BP	GO:0006955	3,09E-04	16	immune response
BP	GO:0007565	2,68E-05	5	femalepregnancy
BP	GO:0007267	1,48E-03	10	cell-cellsignaling
BP	GO:0007186	4,20E-03	11	G-prot.coupled receptor prot.signalingpath.
BP	GO:0042060	5,03E-03	1	woundhealing
BP	GO:0042493	5,84E-03	8	response to drug
BP	GO:0030168	9,93E-03	1	plateletactivation
MF	GO:0008083	4,75E-07	7	growth factor activity
MF	GO:0004930	1,07E-06	8	G-proteincoupled receptor activity
MF	GO:0004872	1,71E-06	39	receptor activity
MF	GO:0005125	1,15E-03	6	cytokineactivity
MF	GO:0005179	3,40E-03	4	hormoneactivity
MF	GO:0004908	8,33E-03	1	interleukin-1 receptor activity
MF	GO:0016491	9,93E-03	14	oxidoreductaseactivity
UNDERREPRESENTED GO TERMS (POLYSOMAL: IN vs CT)				
type	**GO ID**	**p-adjusted**	**# genes***	**term**
CC	GO:0005634	4,90E-06	177	nucleus
CC	GO:0005622	1,79E-03	65	intracellular

Table S7:
GO analysisofdifferentiallyexpressed genes fromthecomparison IN vs CT (polysomalfraction). Cellularcomponent (CC), biologicalprocess (BP), molecular function (MF) are shown, includingthecorrespondingadjusted p-value, thenumberof genes falling in thatcategoryandthetermdescription.
MF terms are displayed as pie charts in figure 3A.
* Some gene overlapexistbetween GO categories.

Supplementary Table S8A (total)

	logFC>1.5	logFC<-1.5	Total	chi square	r
	Patient 61				
UTR extension	242	215	457		
UTR shortening	150	314	464	3,77E-07	0,279*
Total	392	529	921		
	Patient 67				
UTR extension	226	204	430		
UTR shortening	166	322	488	2,14E-05	0.289*
Total	392	526	918		
	Patient 70				
UTR extension	264	288	552		
UTR shortening	114	216	330	0.0001528	0,218*
Total	378	504	882		

Proportions of genes having extended or shorter UTRs (with respect to control) compared with proportions of up or down regulated genes (|logFC|<1.5), only total RNA fraction considered. Chi-square test shows significant differences in the porportions in all patients. Correlation test (r column) shows the correlation value between length differences in UTR and logFC. In all cases the correlation was significant (*: *p-value*<$2.2x10^{-16}$).

Supplementary Table S8B (polysomal)

	logFC>1.5	logFC<-1.5	Total	chi square	r
	Patient 61				
UTR extension	103	72	175		
UTR shortening	71	114	185	1,56E-04	0.324*
Total	174	186	360		
	Patient 67				
UTR extension	99	70	169		
UTR shortening	70	122	192	4,18E-05	0.352*
Total	169	192	361		
	Patient 70				
UTR extension	117	101	218		
UTR shortening	48	79	127	6,236E-03	0.245*
Total	165	180	345		

Proportions of genes having extended or shorter UTRs (wrt control) compared with proportions of up or down regulated genes (|logFC|<1.5), only polysomal RNA fraction considered. Chi-square test shows significant differences in the porportions in all

patients. Correlation test (r column) shows the correlation value between length differences in UTR and logFC. In all cases the correlation was significant (*: *p-value*<$2.2x10^{-16}$).

Supplementary Table S9: GO (and other ontologies) analysis of genes presenting UTR extension/shrinkage.

genes corresponding to UTR extension with logFC>1.5

Category	p-val	# genes	Term
GO:0042562	0,0424	3	Hormone binding
GO:0017046	0,00775	3	Peptide hormone binding
KEGG:00980	0,0271	3	Metabolism of xenobiotics by cytochrome P450
KEGG:05221	0,00295	3	Acute myeloid leukemia

genes corresponding to UTR extension with logFC<1.5

Category	p-val	# genes	term
GO:0051347	0,0269	4	Positive regulation of transfer of transferase activity
GO:0033674	0,0245	4	Positive regulation of kinase activity
GO:0045860	0,0201	4	Positive regulation of protein kinase activity
GO:0035966	0,0478	3	Response to topologically incorrect protein
GO:0006986	0,037	3	Response to unfolded protein
GO:0034976	0,0183	3	Response to endoplasmic reticulum stress
GO:0035976	0,0108	3	Cellular response to topologically incorrect protein
GO:0034620	0,00744	3	Cellular response unfolded protein
GO:0006984	0,0105	3	ER-nucleus signaling pathway
GO:0030968	0,0072	3	Endoplasmic reticulum unfolded protein response

genes corresponding to UTR shrinkage with logFC>1.5

Category	p-val	# genes	term
REAC:380608	0,0499	1	5-hydroxyindole acetaldehyde to 5 hydroxyindoleacetic acid
REAC:71723	0,0499	1	Acetaldehyde + NAD+ <-> Acetat+NADH + H+

genes corresponding to UTR shrinkage with logFC<1.5

Category	p-val	# genes	term
CORUM:5816	0,0497	1	Apoptosome-procaspase 9 complex
CORUM:1248	0,0332	1	Apoptosome
REAC:114254	0,0499	1	Cytochrome C binds to Apaf-1

gProfiler was performed on the set of genes presenting UTR modifications.
We divided those genes in four sets:
-upregulated genes presenting UTR extensions
-dowregulated genes presenting UTR extension
-upregulated genes presenting UTR shrinkage
-downregulated genes presenting UTR shrinakge

Supplementary Table S10 - **Half-lives of mRNAs**

		Half-life (hours)		
mRNA		TA77	TA82	Mean
wnt2	CT	5.0	9.9	7.4 (±3.5)
	IN	3.5	2.3	2.9 (±0.81)
crabp2	CT	23.1	69.3	46.2 (±32.6)
	IN	5.0	3.9	4.4 (±0.7)
fabp4	CT	6.9	5.3	6.1 (±1.1)
	IN	8.7	34.7	21.7 (±18.3)
mt1a	CT	6.9	5.8	6.3 (±0.8)
	IN	8.7	8.7	8.7 (±0.0)

Transcript half-life values obtained by RT-qPCR.

Capítulo 5

La crisis de identidad de las células madre mesenquimales obtenidas de tejido adiposo (ADSC)

5.1. Introducción

Durante el desarrollo de los dos artículos anteriores nos encontramos con una problemática actual, que es el problema de identidad que tienen las células madre mesenquimales obtenidas de tejido adiposo (ADSC adipose derived stem cells). Las ADSC son muy parecidas a los fibroblastos, tanto morfológicamente, como molecularmente, sobre todo en los marcadores de superficie, los cuales se utilizan para distinguir y clasificar a las células.

Algunos estudios postularon que los fibroblastos poseían características similares a las células madre, como ser el potencial de diferenciación. Sin embargo, otros postulan que los mismos no pueden diferenciarse.

Este trabajo intenta aclarar si las ADSC son realmente un tipo de célula madre o se incluyen dentro de los fibroblastos con quizás alguna característica adicional. Observando características moleculares y evaluando el potencial de diferenciación de las mismas a través de experimentos de RNA-seq se intenta resolver la "crisis de identidad" de las ADSC.

Se adjunta al final del capítulo el artículo correspondiente al tema. El mismo fue enviado a la revista Stem Cell & Development.

5.2. Características de los fibroblastos

El fibroblasto es un tipo de célula residente del tejido conectivo que sintetiza matriz extracelular y colágeno, manteniendo el marco estructural (estroma) del tejido de muchos animales. Estas células proporcionan una estructura en forma de entramado a muy diversos tejidos y tienen una función relevante en la curación de heridas, siendo las células más comunes del tejido conectivo. Los mismos se derivan de células primitivas mesenquimales y pluripotentes.
Los fibroblastos y células madre mesenquimales obtenidas de tejido adiposo, sobre todo por ser ambos derivados del mismo tejido no sólo presentan la misma morfología sino que también ambos proliferan bien, adhieren al plástico y expresan los mismos marcadores de superficie. Los fibroblastos en general expresan en su superficie marcadores de ADSC, pero no expresan marcadores hematopoyéticos [174, 175]. Hasta el momento el mejor método descrito para distinguir las ADSC de los fibroblastos se basa en el análisis de las propiedades funcionales de estos dos tipos de células. Las ADSC tiene la capacidad de auto-renovarse, proliferar y diferenciarse en varios linajes celulares, mientras que los fibroblastos aparentan ser o muy limitados en esta característica o no la poseen directamente [174]. Sin embargo, incluso esta última característica ha suscitado dudas en la comunidad científica. Se ha afirmado que los fibroblastos son capaces de diferenciar a ciertos linajes celulares, pudiendo ser utilizados por lo tanto en terapia celular [176]. Recientemente, Blasi y colaboradores [175] establecieron que los fibroblastos tienen el potencial de diferenciarse, pero carecen de la capacidad anti-inflamatoria y angiogénica. Al mismo tiempo, otros estudios comparativos de ADSC y fibroblastos han observado que los mismos no poseen esa capacidad de diferenciación [174, 177]. Análisis funcionales comparativos previos entre ADSC y fibroblastos determinaron sólo 64 genes sobre-expresados en fibroblastos, una cantidad relativamente pequeña, apoyando la similitud entre ambos [177].
Por este motivo se analizó más a fondo las diferencias entre ADSC y los fibroblastos a nivel molecular por medio de experimentos de RNA-seq, entre otros.

5.3. Muestras y análisis bioinformáticos realizados

Se utilizan las mismas muestras de ADSC del experimento anterior, detalladas en 4.4. Se consideran las 3 muestras control, sólamente la fracción polisomal. Adicionalmente, se secuencian dos muestras de una línea celular de fibroblastos dermales. Los mapeos y conteos se realizan como en los capítulos anteriores 4.4. Los análisis de calidad fueron análogos también, los cuales muestran gran consistencia de los datos.
La expresión diferencial se evaluó con el paquete edgeR de R, obteniéndose 1115 genes

codificantes diferencialmente expresados (FDR< 0,0001). Se realiza un análisis de GO de los genes sobre-expresados en fibroblasto y de los sobre-expresados en célula madre (es decir, sub-expresados en fibroblasto), utilizándose para ello la herramienta en línea GOrilla. Los términos obtenidos para ADSC se relacionan con adhesión celular, señalización celular, procesos de desarrollo, etc, mientras que los términos sobre-representados en fibroblastos se relacionan con ciclo celular y proliferación. Por más detalles, referirse al artículo adjunto.

5.4. Resumen de resultados

Este artículo posee varios ensayos de laboratorio que se complementan claramente con los análisis bioinformáticos, llegando a la conclusión que, dadas nuestras condiciones de estudio, los fibroblastos no diferencian, o si lo hacen, a muy baja tasa, por lo que se concluye que las ADSC no son parte de los fibroblastos.
Las ADSC y los fibroblastos fueron inducidos a diferenciarse a tres linajes mesenquimales: adipocitos, condrocitos y osteocitos. Una vez aplicados los protocolos correspondientes, y observado las características correspondientes para evaluar diferenciación en cada linaje (células redondeadas con vacuolas enriquecidas en lípidos visualizable con Oil Red O para adipocitos, mineralización de la matriz extra-celular observable mediante Alizarin Red para osteocitos y mucopolisacáridos visualizables mediante toluidine blue), se constató que sólo las ADSC lograron diferenciar.
Para una caracterización más profunda de ambos tipos celulares, se secuencian con el SOLiD las muestras mencionas anteriormente en 5.3, ADSC comparado con fibroblasto. De los 17340 genes recuperados en el secuenciamiento, la mayor parte (15776) fue común a ambos tipos celulares, mientras que algunos genes aparentan ser específicos de ADSC (640) y de fibroblatos (924). Por otro lado, se determinan los genes diferencialmente expresados, encontrando un número 20 veces mayor que lo establecido anteriormente por otros estudios que utilizaron ARN total. Esto implica, por un lado, que existen diferencias claras entre ambos tipos celulares, y por el otro, que el hecho de utilizar ARN asociado a polisomas se ajusta más a las características fenotípicas que utilizar el ARN total. Dentro de los genes diferencialmente expresados no se encuentran la mayoría de los marcadores de superficie que han sido reportados como presentes en ambos, ADSC y en fibroblastos (CD166, β1-integrin CD29, CD44, etc.), lo que confirma nuestros resultados y las previas observaciones de la similitud entre ambos tipos celulares. El análisis de GO a partir de, por un lado la lista de genes sobre-expresados en fibroblasto, y por el otro los genes sobre-expresados en ADSC, resultó en grandes diferencias en cuanto a procesos biológicos. ADSC se caracterizó por obtener términos referentes a adhesión

celular, lo cual fue comprobado con ensayos funcionales en el laboratorio. Los fibroblastos se caracterizaron por altas tasas de proliferación, lo cual fue igualmente validado en el laboratorio. Al analizar los términos GO correspondientes a la localización celular, ADSC muestra varios términos relacionados con la membrana plasmática y la región extracelular. Al analizar proteínas de membrana, por ejemplo la CD105, un marcador bien característico de ambos tipos celulares, se observó que la misma si bien estaba sobre-expresada en ambos tipos celulares, en ADSC presenta una expresión mucho mayor. Esta diferencia alcanza para distinguir dos poblaciones mezcladas de ambos tipos, logrando separarlas por citometría de flujo con hasta un 90 % de pureza. Si se analiza el potencial de diferenciación de ambos grupos separados por CD105, se observa que el grupo con mayor cantidad de CD105, o sea las ADSC, son capaces de diferenciar, mientras que el otro grupo no posee esta característica.
Nuestros análisis de expresión génica confirman que las ADSC y fibroblastos son tipos celulares diferentes. Pueden confundirse probablemente muy fácilmente, ya que es muy difícil separarlas con alto grado de pureza.

5.5. Artículo

Polysome profiling shows the identity of human adipose-derived stromal/stem cells in detail and clearly distinguishes them from dermal fibroblasts.

Jaiesa Zych[a], Lucia Spangenberg[b], Marco A. Stimamiglio[a], Ana Paula R. Abud[a], Patrícia Shigunov[a], Fabricio Marchini[a], Crisciele Kuligovski[a], Axel R. Cofré[a], Andressa V. Schittini[a], Alessandra M. Aguiar[a], Alexandra Senegaglia[c], Paulo R.S. Brofman[c], Samuel Goldenberg[a], Bruno Dallagiovanna[a], Hugo Naya[b], Alejandro Correa[a1].

[a] Instituto Carlos Chagas, Fiocruz-Paraná. Rua Professor Algacyr Munhoz Mader, 3775. Curitiba-PR, 81350-010, Brazil.

[b] Unidad de Bioinformática, Institut Pasteur Montevideo. Mataojo 2020, Montevideo, 11400, Uruguay.

[c] Núcleo de Tecnologia Celular, Pontifícia Universidade Católica do Paraná. Rua Imaculada Conceição, 1155, Curitiba-PR, 80215-901, Brazil.

1- To whom correspondence should be addressed: e-mail: alejandro@tecpar.br Tel: (55 41) 33163237, Fax: (55 41) 33163267

BRIEF RUNNING TITLE

ADSC and fibroblast polysome profiling

ABSTRACT

Although fibroblasts and multipotent stromal/stem cells, including adipose-derived stromal cells (ADSCs), have been extensively studied, they cannot be clearly distinguished from each other. We therefore investigated the cellular and molecular characteristics of ADSCs and

fibroblasts. ADSCs and fibroblasts share several morphological similarities and surface markers, but were clearly found to be different types of cells. Contrary to previous reports, fibroblasts were not able to differentiate into adipocytes, osteoblasts or condrocytes. Polysome-bound mRNA profiling revealed that approximately 1,547 genes were differentially expressed in the two cell types; the genes were related to cell adhesion, the extracellular matrix, differentiation and proliferation. These findings were confirmed by functional analyses showing that ADSCs had a greater adhesion capacity than fibroblasts; also, the proliferation rate of fibroblasts was higher than that of ADSCs. Importantly, 185 differentially expressed genes were integral to the plasma membrane and, thus, candidate markers for ADSC isolation and manipulation. We also observed that an established marker of fibroblasts and ADSCs, CD105, was overexpressed in ADSCs at both mRNA and protein levels. CD105 expression seemed to be related to differentiation capacity, at least for adipogenesis. This study shows that ADSCs and fibroblasts are distinct cell types. These findings must be taken into account when using these two cell types in basic and therapeutic studies.

INTRODUCTION

Multipotent stromal/stem cells (MSCs), including adipose-derived stromal cells (ADSCs), and fibroblasts not only share a similar morphology but they both proliferate well and express many of the same cell surface markers. Fibroblasts usually express high levels of MSC markers and do not express hematopoietic markers [1-3]. Currently, the best approach to distinguishing between MSCs and fibroblasts is based on the analysis of their functional properties. MSCs self-renew and retain a multipotent differentiation capacity, whereas fibroblasts seem to display only limited, or no such, multipotent differentiation [2], although there is controversy in the literature over this issue. Indeed, it has been suggested that

fibroblasts are able to differentiate [1,3-7] and could be used in cell therapy [8]. Recently, Blasi et al. (2011) [3] reported that fibroblasts may differentiate but lack anti-inflammatory and angiogenic capacity. Recent studies comparing MSCs and fibroblasts, however, indicate that fibroblasts have no differentiation capacity [2,9]. Thus, a more detailed and rigorous comparative analysis of ADSCs and fibroblasts is needed to document the identity of these cell types and to determine the multipotent differentiation capacity of fibroblasts, if any. Functional genomics studies using total mRNA and miRNA have established a MSC-specific molecular signature consisting of only 64 genes and 21 miRNAs: the expression of these genes is at least 10-fold higher and of the miRNAs 2-fold higher in MSCs than fibroblasts [9]. Several studied have shown that MSCs are probably promiscuous transcribers [10-13] and, as previously stated, MSCs seem to be multi-differentiated cells at the molecular level, because they usually express markers and regulators of various differentiated cell lineages [14]. Most attempts to determine the mRNA profile of self-renewing cells have used total RNA for high throughput analyses [15,16]. Studies comparing mRNA and protein levels in eukaryotes indicate that, although transcript levels correlate with protein synthesis, the strength of the correlation is low, suggesting a high degree of post-transcriptional regulation (reviewed by [17]). These various findings suggest that studies using the total population of transcripts do not necessarily represent the identity of these cells faithfully.

In this work, we show that fibroblasts do not differentiate or differentiate only very poorly. We conducted mRNA profiling analyses and identify approximately 1547 transcripts isolated from polysomal fractions (*i.e.* associated with the translation machinery) that are differentially expressed between the two cell types. Thus, we describe 20 times more differentially expressed transcripts than previously reported. These transcripts are related to cell adhesion, the extracellular matrix and differentiation. Functional assays confirmed our findings and clearly show that fibroblasts are a terminally differentiated cell type, but which are probably

difficult to isolate with a high degree of purity. Thus, we demonstrate that dermal fibroblasts and ADSCs are functionally and molecularly different entities.

MATERIALS AND METHODS

Cell culture

Tissue samples were obtained and stem cells were isolated as previously described [10]. ADSCs were cultured in DMEM/F12 medium (Gibco Invitrogen, USA), with 10% fetal calf serum (FCS; Gibco Invitrogen, USA), 100 U/mL penicillin and 100 µg/mL streptomycin (Sigma-Aldrich, USA). The cell isolation protocols resulted in a population highly enriched (> 95 %) in adult MSCs, as defined by Dominici *et al*. [18]. Normal human adult dermal fibroblasts were obtained from the American Type Culture Collection (ATCC PCS-201-012) and were cultured in fibroblast growth basal medium, with 7.5 mM L-glutamine, 5 ng/mL rhFGF, 5 µg/mL recombinant human insulin, 1 µg/mL hydrocortisone, 50 µg/mL ascorbic acid and 2% fetal bovine serum (all from ATCC). As required for the purposes of comparison, the different cell types were cultured in DMEM/F12 medium, with 10% fetal calf serum (Gibco Invitrogen), 100 U/mL penicillin and 100 µg/mL streptomycin (Sigma-Aldrich, USA). All cultures were maintained at 37°C, in a humidified atmosphere containing 5% CO_2. The culture medium was changed every 3 or 4 days.

Phenotypic characterization by flow cytometry

Surface proteins on ADSC and fibroblasts were detected by cytofluorometry. Three samples of each cell type, between the third and fifth passages, were first incubated with purified mouse IgG (used to block Fc receptors), and then incubated with anti-CD90-FITC, anti-CD105-PE, anti-CD73-APC, anti-CD45-FITC, anti-CD34-PE, anti-HLA-DR-APC, anti-

CD31-FITC, anti-CD117-PE or anti-CD19-FITC mAb, or to the corresponding IgG matched negative controls. The samples were subsequently analyzed using a FACSCanto II apparatus (Becton Dickinson, San Jose, CA, USA). A cell gate excluding cell debris and non-viable cells was determined using forward and side scatter parameters, and was confirmed in some experiments by propidium iodide staining and immediate analysis of unfixed cells. Analyses were done after recording at least 10,000 events for each sample. Results are expressed as percentages of fluorescence-positive cells (% $cell^{+}$) and mean fluorescence intensity (MFI).

Differentiation into Mesenchymal Lineages

The capacity of ADSCs and fibroblasts to differentiate into adipocytes, osteoblasts and chondroblasts was assessed.

For adipogenic differentiation, cultures were treated with hMSC Adipogenic Differentiation Bullet Kit (Lonza) for 21 days, in accordance with the manufacturer's instructions; Oil Red O staining was used to visualize the accumulation of cytoplasmic triglycerides in the cells. Briefly, the cells were washed in PBS, fixed with 4% paraformaldehyde for 30 min and inclubated for 30 min with 0.5% Oil Red O solution (Sigma-Aldrich).

Osteogenic differentiation was induced with the hMSC Osteogenic Differentiation Bullet Kit (Lonza) for 21 days, in accordance with the manufacturer's instructions. To determine the degree of osteogenic differentiation and calcium deposition, the culture was stained with Alizarin Red S (Sigma-Aldrich). Briefly, the cells were washed in PBS, fixed with 4% paraformaldehyde for 30 min and incubated for 30 min with 2% Alizarin Red S solution, pH 4.2.

Micromass cultures and hMSC Chondrogenic Differentiation Bullet Kit (Lonza) were used to promote chondrogenic differentiation, in accordance with the manufacturer's instructions. After 21 days, chondrogenesis was visualized by toluidine blue staining. Briefly, the cells were washed in PBS, fixed with 10% formaldehyde for 1 h, dehydrated in serial ethanol

dilutions, and embedded in paraffin blocks. Sections (4 μm thick) of the paraffin blocks were stained with toluidine blue solution (Sigma-Aldrich) for histological analysis to demonstrate the presence of intracellular matrix mucopolysaccharides.

The stained cells were examined and photographed under inverted Nikon Eclipse TE300 or Eclipse E600 microscopes.

Quantification of adipocyte differentiation by Nile Red staining

Cells were washed with PBS, fixed by incubation with 4% paraformaldehyde for 10 min and washed again with PBS. They were then stained with a solution of Nile Red (Sigma-Aldrich), prepared immediately before use by diluting 1000-fold a stock solution (1 mg/mL of Nile Red dissolved in dimethylsulfoxide) in PBS, for 30 min at 4°C. The cells were washed with PBS and stained with DAPI for 20 min, washed again with PBS and photographed. Ten images of random fields were obtained at a magnification of two hundred with Nikon Eclipse TE300 fluorescence microscope. The area in pixels per nucleus was determined using ImageJ software (National Institutes of Health, Bethesda, MD, USA).

Sucrose density gradient separation and RNA purification

Polysomal fractions from hADSC and fibroblast cultures at 50 to 60% confluence were prepared according to [19]. In brief, cells were treated with 0.1 mg/mL cycloheximide (Sigma-Aldrich) for 10 min at 37°C, removed from the culture flasks with a cell scraper and resuspended in 0.1 mg/mL cycloheximide in PBS. The suspension was centrifuged (2,000 x *g* for 5 min) and the resulting pellet was washed twice with 0.1 mg/mL cycloheximide in PBS. The cells were lysed by incubation for 10 min on ice with polysome buffer (15 mM Tris-HCl pH 7.4, 1% Triton X-100, 15 mM $MgCl_2$, 0.3 M NaCl, 0.1 μg/mL cycloheximide, 1 mg/mL heparin) and the cell lysate was centrifuged at 12,000 x *g* for 10 min at 4°C. The supernatant was carefully isolated, loaded onto 10% to 50% sucrose gradients and centrifuged at 39,000 rpm (HIMAC CP80WX HITACHI) for 160 min at 4°C. The sucrose gradient was fractionated

with the ISCO gradient fractionation system (ISCO Model 160 gradient former), connected to a UV detector to monitor the absorbance at 275 nm, and the polysome profile was recorded. The polysomal RNA fractions were extracted by a standard Trizol (Invitrogen) RNA isolation protocol.

cDNA library construction and RNA sequencing (RNA-Seq)

Polysome-associated RNA samples were amplified using the Amino Allyl Message Amp II aRNA Amplification Kit (Ambion), to generate templates for SOLiD libraries. The cDNA libraries were prepared with the SOLiD Whole Transcriptome Analysis Kit (Applied Biosystems, USA) and the purified products were evaluated with an Agilent Bioanalyzer (Agilent). Library molecules were subjected to clonal amplification according to the SOLiD Full-Scale Template Bead preparation protocol and sequenced with the SOLiD4 System (Applied Biosystems).

RNAseq Data analysis

NGSQC [20] software was used for quality control analysis of sequencing data. Various quality indicators were explored visually for each sample (distribution of colors per sample/tile, genomic hit count per sample with different numbers of mismatches, sequencing read density and mean quality values for each sample). All samples passed the quality control filters. Mapping and counting were performed with the R package Rsubread [21]. Hierarchical clustering of the samples (log of counts plus one) was performed to evaluate biological variability. Each sample was normalized to one million reads to account for library size. We also conducted a correspondence analysis (COA), involving a dimension reduction method, to the matrix of counts, to explore associations between variables. COA allows samples and genes to be visualized simultaneously, revealing associations between them: Genes, or samples, lying close to each other tend to behave similarly.

For the comparison of fibroblasts with ADSCs, we analyzed only those genes with counts of more than 1 per million. Genes differentially expressed (DE) between cell types were identified with the edge R bioconductor package [22]. This set of genes was used for GO (Gene Ontology) term analysis with the Gene Ontology enRIchment anaLysis and visuaLizAtion tool (Gorilla: http://cbl-gorilla.cs.technion.ac.il/), a web server that is identifies enriched GO terms in long lists of genes. Gorilla results were then visualized using the REVIGO software (REViGO: http://revigo.irb.hr/), which summarizes long lists of GO terms with respective significance values by removing redundancy in terms.

RT-qPCR

RT-PCR and real-time quantitative PCR (qPCR) were performed as previously described [10]. The glyceraldehyde-3-phosphate dehydrogenase (GAPDH) transcript was used as an internal control. Amplifications were performed with cells from different experiments, with technical triplicates. Student's *t*-test was used to assess the significance of differences between the cell populations. We considered p-values < 0.05 to be statistically significant.

NanoLC-MS/MS analysis

Peptide mixtures from two samples (approximately $7x10^5$ cells/sample) were separated by online RP nanoscale capillary LC (nanoLC) and analyzed by ESI MS/MS. The experiments were performed with an Ultra 1D Plus (Eksigent, Dublin, CA) system connected to the LTQ-Orbitrap XL ETD mass spectrometer equipped with a nano-electrospray ion source (Thermo Scientific, Waltham, USA). Peptides were chromatographically separated in a 15-cm fused silica emitter (75 µm inner diameter) packed in-house with reversed-phase ReproSil-Pur C18-AQ 3 µm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany).

Peptide mixtures were injected onto the column at a flow rate of 250 nL/min and subsequently eluted at a flow rate of 250 nL/min with a from 5 to 40% ACN in 0.1% formic acid over 180 min. The mass spectrometer was operated in data-dependent mode to switch automatically

between MS and MS/MS (MS2) acquisition. Survey full-scan MS spectra (at 350 - 1650 *m/z* range) were acquired in the Orbitrap analyzer with resolution R = 60,000 at m/z 400 (after accumulation to a target value of 1,000,000 in the linear ion trap). The ten most intense ions were sequentially isolated and fragmented in the linear ion trap using collision-induced dissociation at a target value of 10,000. Former target ions selected for MS/MS were dynamically excluded for 90 s. Total cycle time was approximately 3 s. The general mass spectrometric conditions were: spray voltage, 2.3 kV; no sheath and auxiliary gas flow; ion transfer tube temperature, 100^{o}C; collision gas pressure, 1.3 mTorr; and normalized collision energy using wide-band activation mode 35% for MS2. Ion selection thresholds were 250 counts for MS2. An activation q = 0.25 and activation time of 30 ms were applied for MS2 acquisitions. The "lock mass" option was enabled in all full scans to improve mass accuracy for precursor ions [23].

Proteomic Data analysis

The MaxQuant platform (version 1.3.0.5) [24], which includes the algorithm Andromeda [25] for database searching, was used for Peaklist picking, and protein identification, quantification and validation. Default parameters of the software were used for all analysis steps, unless stated otherwise. Proteins were searched against a "decoy database" prepared by reversing the sequence of each entry of the human protein sequence database (containing 87,061 protein sequences, downloaded in March 29, 2010 from IPI protein database (version 3.68) and appending them to the forward sequences. This database was complemented with frequently observed contaminants (porcine trypsin, *Achromobacter lyticus* lysyl endopeptidase, and human keratins) and their reversed sequences. Search parameters specified a MS tolerance of 7 ppm, a MS/MS tolerance of 0.5 Da, and full trypsin specificity, allowing for up to two missed cleavages. Carbamidomethylation of cysteine was set as a fixed

modification and oxidation of methionines and N-terminal acetylation (protein) were allowed as variable modifications.

For validation of the identifications, a minimum peptide length of six amino acids and two peptides per protein were required. In addition, a false discovery rate (FDR) threshold of 0.01 was applied at both peptide and protein levels. When transforming peptide identifications into protein identifications, similar protein sequences (e.g. isoforms) present in the database that could not be distinguished by the experimentally detected peptides were grouped together and are referred to as protein groups.

Protein quantification was performed using a label-free approach, in which peptides eluting from each LC run are detected as three-dimensional features - retention time *versus* signal intensity (extracted ion chromatogram, XIC) *versus* mass/charge - aligned and compared across runs, as previously described [26].

Bromodeoxyuridine proliferation assay

The proliferation assay was performed according to [27]. In brief, fibroblasts and ADSCs at 70% confluence were incubated with 100 μM BrdU (Invitrogen) for 24 h. The cells were detached with trypsin and fixed by incubation with 100% ethanol for 30 min on ice. The samples were centrifuged, and the cell pellet was resuspended in 100 μL of distilled water, heated for 5 min at 95°C and rapidly chilled in an ice-water bath. The cells were incubated with an Alexa Fluor 488-conjugated anti-BrdU antibody (Invitrogen) for 30 min at room temperature. A FACSCanto II flow cytometer (BD Bioscience) and Flow Jo software (Tree Star, USA) were used for quantitative analyses of BrdU-labeled cells.

Cell adhesion assay

ADSCs and fibroblasts were tested for adhesion to 24-well culture plates for 20 and 40 min at 37˚C as follows. The cells were first washed and then added to the wells at a concentration of 2.5 x10^4 cells per well in 500 μL DMEM/F12 medium (Gibco Invitrogen) supplemented with

10% FCS. After the adhesion period (20 or 40 min), the plates were maintained for 5 min of incubation under shaking (100 rpm) and non-adherent cells were removed by washing three times with ice-cold PBS. Adherent cells were fixed (100% ethanol for 5 min) and stained (2% toluidine blue for 15 min) for counting. Four wells for each cell type (three ADSC and three fibroblast samples) were analyzed and photographed for counting labeled cells with the ImageJ software (National Institutes of Health). Five representative fields from each well were used to calculate mean values.

ADSC and fibroblast cell sorting

Cytofluorometry was used for detection and separation of ADSC and fibroblast populations. A mixed sample of each cell type, between the fourth and fifth passages, was first incubated with anti-CD105-PE mAb, and subsequently sorted using a FACSaria II apparatus (Becton Dickinson, San Jose, CA, USA). A cell gate discriminating between low and high CD105-PE fluorescence intensity was determined to separate cell populations. Doublets were excluded from the cell population using a sequential gating strategy relative to width *versus* height on side scatter (SSC) and forward scatter (FSC) dot plots. Finally, weak ($CD151^{+}$) and bright ($CD151^{++}$) CD105-positive cells were selected and sorted using the cell sorter's purity option at a rate of 2,500 events per second. Sorted populations were reanalyzed (5×10^3 events recorded) for purity and viability, and 4×10^5 cells of each population were culture-expanded for imunophenotyping and cell differentiation assays.

Statistical analysis of functional assays

Statistical significance was determined by Student's *t* test using GraphPad Prism software (San Diego, CA, USA). Values of $P < 0.05$ were considered significant. Data are expressed as means ± standard deviation.

RESULTS

The expression of cell-surface antigens proposed by Dominici et al. (2006) [18] as a minimal set for stromal/stem cell characterization was evaluated by flow cytometry in three biological samples of ADSCs and three technical samples of dermal fibroblast at passage P5 (Fig. 1A). With one exception, the two types of cells displayed similar immunophenotypes for the markers analyzed. Cells were uniformly positive for the membrane glycoprotein CD90 and the surface enzyme ecto 5´nucleotidase CD73. No detectable contamination by hematopoietic or endothelial cells was observed, as flow cytometry analysis was negative for CD19, CD31, CD34, CD45, CD117 and HLA-DR. More than 95% of the cells in both cell populations expressed the endoglin receptor CD105; however, the MFI was significantly higher for ADSCs than for fibroblasts (Fig. 1A).

Cells at passage P5 were compared for their multilineage differentiation plasticity by *in vitro* assays: specifically, differentiation to adipocytes, osteoblasts, and chondrocytes was studied. We used the presence of lipid-rich vacuoles stained with Oil Red O to analyze adipogenic induction on day 21. ADSCs presented large, rounded cells with cytoplasmic lipid-rich vacuoles, fibroblasts displayed very few differentiating cells and they had smaller intracellular lipid droplets than those observed in ADSCs (Fig. 1B). Osteogenic differentiation was assessed by the mineralization of the extracellular matrix, visualized by Alizarin Red S after 21 days of induction. Differentiation of ADSCs was readily detected. No differentiation was observed in the induced fibroblast cultures (Fig. 1B). Similar results were obtained for the chondrogenic differentiation assays. ADSCs formed aggregates that became detached, floating freely in suspension in the culture. Paraffin sections of the aggregates stained with Toluidine Blue showed a condensed structure with cuboidal cells and chondrocyte-like lacunae: the cells stained positively for Toluidine Blue, a dye specific for the highly sulfated proteoglycans of cartilage matrices. Fibroblast differentiation was minimal or undetectable

(Fig. 1B). Untreated control cultures, which were growing in standard medium without adipogenic, osteogenic or chondrogenic differentiation stimuli, did not exhibit spontaneous adipocyte or osteoblast/chondroblast formation after 21 days of cultivation (Fig. 1B).

In the culture conditions used, the mean doubling time of fibroblasts is18-24h, and the cells were sub-cultured approximately every 3 days; thus, fibroblasts at passage P10-P11 had undergone more than 30 population doublings. Consequently, fibroblasts at P10 are considered to be at a pre-senescence stage [28]. Because no evident signs of senescence were visible at P11, the adipogenic differentiation potential of fibroblasts at this passage was tested and compared to that of fibroblasts at P5. Surprisingly, there was a higher proportion of differentiated cells at P11 (Fig. S1).

For more detailed characterization of both cell lineages, gene expression patterns in each cell type were determined by studying mRNA associated with the translation machinery. The polysomes of ADSCs and fibroblasts were profiled by ultracentrifugation of cytoplasmic extracts onto sucrose density gradients (10–50%) containing cycloheximide. Starting with extracts obtained from the same number of cells, the translation activity (measured as the height of polysome peaks) was lower in ADSCs than fibroblasts (Fig. 2A). Polysome-associated mRNAs were isolated from the corresponding gradient fractions, and the mRNA fractions of ADSCs and fibroblasts were analyzed by RNA-Seq using the SOLiD4 platform. The total number of reads obtained for each sample is shown in the Supplementary Data (Table S1). The reads of all samples were mapped onto the reference genome (Hg19; NCBI Build 37.64), yielding a mean mapping percentage of approximately 55%. To decrease ambiguity, the analyses were limited to reads mapping to a unique position in the genome. Hierarchical clustering and Correspondence Analysis (COA) grouped the samples according to cell lineage (ADSCs and fibroblasts), rather donor, indicating that the cell-specific

polysomal RNA populations are intrinsically more characteristic than donor particularities (Fig. 2B).

A total of 17,340 protein-coding transcripts were sequenced. Most transcripts (15,776) were detected in both ADSC and fibroblast polysomes, albeit with different levels of expression. Nevertheless, a substantial number of transcripts were exclusively found in only one or the other RNA population: 924 in fibroblast and 640 in ADSCs. As polysomal-associated mRNA was used in this study, it was expected that most of the proteins identified in a proteomic data set of ADSCs would be represented in the mRNA population. To test whether this was the case, proteins from ADSCs from two donors of were identified by LC-MS/MS and the results compared to the RNA-Seq data for the same ADSC donors. A total of 297 proteins were identified with at least two peptides in the ADSC samples and, as expected, more than 85% of them were represented with at least five tags (after normalization) in each of the three biological samples analyzed by RNA-Seq. Moreover, of the 42 proteins (15%) not found in the RNA-Seq data, 10 were histone transcripts that lack poly-A tails and, thus, could not be detected by the RNA-Seq method used in this work (data not shown).

The differentially expressed (DE) genes were identified by paired comparisons (fibroblasts *versus* ADSCs) using the R package edgeR. We found 1,115 and 1,547 DE protein-coding genes in the polysomal fractions with FDRs < 0.0001 and < 0.001, respectively (Supplementary Dataset S1). At these FDR values, most of the genes (>98%) showed logFC values > 2 or < -2.

The DE genes in the polysomal fraction were analyzed in more detail: several of the surface markers previously described to be present in both fibroblast and ADSCs (for example, the activated leukocyte cell adhesion molecule CD166, the β1-integrin CD29, and the hyaluronate receptor CD44) were absent from the DE list (Table S2) providing some validation of our results.

Two GO analyses were performed, one for the genes upregulated in fibroblasts and one for the genes upregulated in ADSC. A target gene list and a background list were considered for these analyses: the target list corresponds to DE genes with FDR < 0.0001 and the background list included all 17,340 genes identified in this study. GOrilla identified 10,942 genes associated with a GO term. GOs with P < E-9 were extracted and used for REVIGO analysis and visualization. Sixteen overrepresented GO terms were identified for ADSCs (P < 1.9E-7), and most are related to processes such as cell adhesion, cell signaling, multicellular organism processes and developmental processes (Fig. 3A). Thirty-seven terms were found to be overrepresented in the analysis of the polysomal fraction of fibroblasts; these GO terms were similar to each other, and refer to cell cycle processes and proliferation (Fig. 3B). Three different approaches were used to confirm these findings. Firstly, the RNA-Seq data was compared with the proteomic data published by Kim et al. [29]. All the transcripts represented with at least five normalized tags in each sample of ADSCs were compared with proteins in Kim's data with at least two spectra. Of the 188 proteins present in undifferentiated ADSCs, 165 (~88%) were found to be DE by our RNA-Seq analysis. Secondly, differential expression was confirmed by RT-qPCR with two selected transcripts, one from each cell type (data not shown). Lastly, some of the overrepresented GO processes for each cell type were analyzed by functional assays. The proliferation of ADSCs was compared with that of dermal fibroblasts by studying BrdU incorporation at passage P5. Under exactly the same culture conditions (medium, temperature/CO_2, confluence and passage) fibroblasts showed significantly higher proliferation values than ADSCs: approximately 20% more cells incorporated BrdU after 24h of culture (Fig. 4A). Also, the numbers of adherent cells after 20, 40 and 60 min of culture was determined by counting the number of non-adherent cells in the culture and subtraction from the total number. At 20 min, significantly more ADSCs than fibroblasts adhered to the plastic plates (Fig. 4B). This

difference was maintained, but not statistically significant, after 40 min (Fig. 4B), and no difference was observed after 60 min of culture (data not shown). These results are in agreement with the results from the GO analysis for biological process terms obtained from RNA-Seq. Unfortunately, the number of proteins identified by the proteomic analysis was insufficient to obtain significant results in a GO analysis.

We analyzed the GO of cellular component terms: the terms associated with ADSCs concentrate mainly at the plasma membrane and extracellular location (Fig. S2). For example, CD200 has a LogFC of 6.1 in ADSCs with a FDR = 5.73E-14. Also, the endoglin protein CD105, a well-defined surface marker for ADSCs and also for fibroblasts, is significantly overexpressed in ADSCs (LogFC = 2.6). This result was confirmed by flow cytometry (see Fig. 1): more than 95% of the cells of both lineages were positive for CD105; however, the level of expression was significantly higher in ADSCs than fibroblasts (LogFC = 3.7, Table S3). We tested whether this difference would be sufficient to distinguish fibroblasts from ADSCs in a mixed population and whether it correlated with differentiation capacity. We therefore conducted a cell sorting assay with a mixed sample of ADSCs and fibroblasts. The mixed sample was stained with anti-CD105-PE mAb and, in a FSC *versus* FL2 dot plot, weakly and strongly CD105-positive cell populations were gated (Fig. 5A). There was an overlap between the CD105 expression by the two cell types impeding the definition of an unambiguous boundary between ADSC and fibroblast populations. We therefore established a cell gate dividing the total CD105-positive population into two equal halves (Fig. 5A). The percentages of $CD105^{++}$ and $CD105^{+}$ cells in each cell population before mixing are shown in figures 5 B and C, respectively. A total of $4x10^5$ cells of $CD105^{++}$ and $CD105^{+}$ cell populations were collected. By reference to the parental mixed cell population, their percentages were 92.3% and 95.6%, respectively, indicating sorting purity of more than 90% (Fig. 5D). We then subjected sorted cells to induction of adipocyte differentiation to evaluate

if these isolated populations, based on CD105 expression, display the same differentiation potential as ADSCs and fibroblasts. Sorted strongly and weakly CD105-positive populations showed, as expected, differentiation patterns resembling those exhibited by induced ADSCs and fibroblasts, respectively (Fig. 5E). The presence of rare cells with cytoplasmic lipids among induced $CD105^{+}$ cells could be attributed to two phenomena: a small number of differentiating cells with small intracellular lipid droplets as observed among induced fibroblast; and the overlap between the two cell types as concerns CD105 positive events. There were 183 genes, other than CD105 and CD200, corresponding to the "integral to plasma membrane" GO term. This means that there are more than a hundred putative cell surface markers that could be evaluated as tools or markers for ADSCs purification/manipulation (Fig. S2, Table S4).

DISCUSSION

We report data demonstrating that ADSCs and fibroblasts are different types of cells, despite sharing several morphological similarities and surface markers. They differ on the basis of hundreds of differentially expressed genes, as evaluated by the analyses of mRNAs associated with the translation machinery. We also show that, contradicting repeated previous suggestions [1,3-7], fibroblasts are not able to differentiate into adipocytes, osteoblasts or condrocytes. We believe that the residual differentiation described previously was due to the fact that it is very difficult to obtain pure populations of dermal fibroblast, as suggested also by Lennon *et al*. [30]. Thus, any differentiating "fibroblasts" observed at passage P11 may be the consequence of contaminating multipotent stromal cells overgrowing the culture as the fibroblasts themselves reach senescence, resulting in greater differentiation.

There is extensive post-transcriptional regulation in mammals and particularly in stem cells [31,32]; as a result, correlations between total mRNA and protein levels although positive, are low, ranging from $r = 0.2$ to 0.4 [19,33]. This is an important issue when analyzing mRNA

levels in cells. Indeed, the RNAs associated with polysomes represent the cellular phenotype more faithfully than analyses of total mRNA; unfortunately, total mRNA is still widely used in most stem cell studies. Moreover, it has been observed in ADSCs differentiation assays, that control cultures without differentiation inducers express several genes at the mRNA level, specific for diverse lineages such as cardiomyocytes, osteoblasts, adipocytes and condrocytes [10,34]. Additionally, promiscuous gene expression has been suggested or shown by others in multipotent stromal cells, including ADSCs [13], and also in other stem cell populations [12,35,36]; these observations clearly indicate that total mRNA populations do not provide a good molecular representation of ADSCs.

To our knowledge, the proliferation rate of fibroblasts and ADSCs has not been considered at all when comparing these two types of cells. Here, we demonstrate that proliferation-related mRNA are better represented in the polysomal fraction of fibroblasts than ADSCs and, accordingly, under the same culture conditions at early passages (P3-5) the fibroblast proliferation rate is higher than that of ADSCs.

ADSCs adhered better than fibroblasts to a plastic surface during the first 20 min of our assay, in agreement with some of the GOs overrepresented in ADSCs. Indeed, a subpopulation of multipotent stem cells derived from human bone marrow cells that exhibited enhanced adhesive and multipotent capacities has been recently identified by Bolontrade M *et al*. [37] . It would be interesting to determine whether a subpopulation of this type also exists in ADSCs and if it is the responsible for the difference to fibroblasts in adherence.

Other functional genomics studies using MSCs under standard culture conditions, *i.e.*, without differentiation inducers, and considering total mRNA for transcriptome analyses [9,38] were only able to find few dozen MSC-specific genes to establish a molecular signature for these cells. Conversely, the approach we used successfully identified several hundred DE genes, specific to each cell type, even under stringent criteria (FDR < 0.0001). More importantly,

185 genes encoded products integral to the plasma membrane; these gene products are therefore candidate, easily accessible markers for ADSC isolation and manipulation. One of these genes was identified as CD200, previously detected in stromal/stem cells isolated from bone marrow, adipose tissue and Wharton's jelly. It seems that this molecule and its receptor (CD200R, only present in myeloid-lineage cells), are responsible for two-way communication between MSCs and T-lymphocytes [39] and, probably, participate in mesenchymal stromal/stem cell immunosuppressant activity [40].

We show that an established marker of fibroblasts and ADSCs, CD105, is overexpressed in ADSCs: this was demonstrated at the mRNA level by deep sequencing and at the protein level by flow cytometry. CD105 expression level seems to be related to differentiation capacity, at least for adipogenesis. A $CD105^{++}$-enriched population differentiated three times more than a $CD105^{+}$-enriched population. Accordingly, it has been shown that down-regulation of CD105 is associated with multilineage differentiation in human umbilical cord blood-derived mesenchymal stem cells [41]. CD105, also known as endoglin, is a type I integral membrane glycoprotein. It is abundant on proliferating cells and has been identified as an accessory receptor for transforming growth factor-β (TGF-β) [42]. Nevertheless, it remains to be established whether CD105 is directly involved in multipotency.

In conclusion, the work presented here demonstrates that ADSCs and fibroblasts are distinct cell types. Although they share various morphological and immunophenotypic similarities, there are vast differences between them in terms of gene expression and functionality/biological dynamics. These findings may be relevant for any applications of these two cell types in both basic and therapeutic studies.

ACKNOWLEDGMENTS

This work was supported by grants from Ministério da Saúde and Conselho Nacional de Desenvolvimento Cientifico e Tecnológico — CNPq, FIOCRUZ-Pasteur Research Program and Fundação Araucária. L.S. received fellowship from ANII (Agencia Nacional de Investigación e Innovación, Uruguay); S.G., J.Z. and B.D. from CNPq; P.S. from FIOCRUZ; and A.C. from Fundação Araucária.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

1. Lorenz K, M Sicker, E Schmelzer, T Rupf, J Salvetter, M Schulz-Siegmund and A Bader. (2008). Multilineage differentiation potential of human dermal skin-derived fibroblasts. Exp Dermatol 17:925-32.
2. Alt E, Y Yan, S Gehmert, YH Song, A Altman, S Gehmert, D Vykoukal and X Bai. (2011). Fibroblasts share mesenchymal phenotypes with stem cells, but lack their differentiation and colony-forming potential. Biol Cell 103:197-208.
3. Blasi A, C Martino, L Balducci, M Saldarelli, A Soleti, SE Navone, L Canzi, S Cristini, G Invernici, EA Parati and G Alessandri. (2011). Dermal fibroblasts display similar phenotypic and differentiation capacity to fat-derived mesenchymal stem cells, but differ in anti-inflammatory and angiogenic potential. Vasc Cell 3:5.
4. Brohem CA, CM de Carvalho, CL Radoski, FC Santi, MC Baptista, BB Swinka, AUC de, LR de Araujo, RM Graf, IH Feferman and M Lorencini. (2013). Comparison between fibroblasts and mesenchymal stem cells derived from dermal and adipose tissue. Int J Cosmet Sci.
5. Lysy PA, F Smets, C Sibille, M Najimi and EM Sokal. (2007). Human skin fibroblasts: From mesodermal to hepatocyte-like differentiation. Hepatology 46:1574-85.
6. Huang HI, SK Chen, QD Ling, CC Chien, HT Liu and SH Chan. (2010). Multilineage differentiation potential of fibroblast-like stromal cells derived from human skin. Tissue Eng Part A 16:1491-501.
7. Haniffa MA, XN Wang, U Holtick, M Rae, JD Isaacs, AM Dickinson, CM Hilkens and MP Collin. (2007). Adult human fibroblasts are potent immunoregulatory cells and functionally equivalent to mesenchymal stem cells. J Immunol 179:1595-604.
8. Osonoi M, O Iwanuma, A Kikuchi and S Abe. (2011). Fibroblasts have plasticity and potential utility for cell therapy. Human cell 24:30-34.
9. Bae S, JH Ahn, CW Park, HK Son, KS Kim, NK Lim, CJ Jeon and H Kim. (2009). Gene and microRNA expression signatures of human mesenchymal stromal cells in comparison to fibroblasts. Cell Tissue Res 335:565-73.
10. Rebelatto CK, AM Aguiar, MP Moretao, AC Senegaglia, P Hansen, F Barchiki, J Oliveira, J Martins, C Kuligovski, F Mansur, A Christofis, VF Amaral, PS Brofman, S Goldenberg, LS Nakao

and A Correa. (2008). Dissimilar differentiation of mesenchymal stem cells from bone marrow, umbilical cord blood, and adipose tissue. Exp Biol Med (Maywood) 233:901-13.
11. Shigunov P, J Sotelo-Silveira, C Kuligovski, AM de Aguiar, CK Rebelatto, JA Moutinho, PS Brofman, MA Krieger, S Goldenberg, D Munroe, A Correa and B Dallagiovanna. (2012). PUMILIO-2 is involved in the positive regulation of cellular proliferation in human adipose-derived stem cells. Stem Cells Dev 21:217-27.
12. Tondreau T, L Lagneaux, M Dejeneffe, M Massy, C Mortier, A Delforge and D Bron. (2004). Bone marrow-derived mesenchymal stem cells already express specific neural proteins before any differentiation. Differentiation 72:319-26.
13. Zipori D. (2004). Mesenchymal stem cells: harnessing cell plasticity to tissue and organ repair. Blood Cells Mol Dis 33:211-5.
14. Jiang Y, B Vaessen, T Lenvik, M Blackstad, M Reyes and CM Verfaillie. (2002). Multipotent progenitor cells can be isolated from postnatal murine bone marrow, muscle, and brain. Experimental hematology 30:896-904.
15. Menssen A, T Haupl, M Sittinger, B Delorme, P Charbord and J Ringe. (2011). Differential gene expression profiling of human bone marrow-derived mesenchymal stem cells during adipogenic development. BMC Genomics 12:461.
16. Jeong JA, KM Ko, S Bae, CJ Jeon, GY Koh and H Kim. (2007). Genome-wide differential gene expression profiling of human bone marrow stromal cells. Stem Cells 25:994-1002.
17. Keene JD. (2007). RNA regulons: coordination of post-transcriptional events. Nat Rev Genet 8:533-43.
18. Dominici M, K Le Blanc, I Mueller, I Slaper-Cortenbach, F Marini, D Krause, R Deans, A Keating, D Prockop and E Horwitz. (2006). Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement. Cytotherapy 8:315-7.
19. Spangenberg L, P Shigunov, AP Abud, AR Cofre, MA Stimamiglio, C Kuligovski, J Zych, AV Schittini, AD Costa, CK Rebelatto, PR Brofman, S Goldenberg, A Correa, H Naya and B Dallagiovanna. (2013). Polysome profiling shows extensive posttranscriptional regulation during human adipocyte stem cell differentiation into adipocytes. Stem Cell Res 11:902-912.
20. Dai M, RC Thompson, C Maher, R Contreras-Galindo, MH Kaplan, DM Markovitz, G Omenn and F Meng. (2010). NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics 11 Suppl 4:S7.
21. Liao Y and W Shi. (2012). Seed-and-vote: the next generation read alignment paradigm. submitted.
22. Robinson MD, DJ McCarthy and GK Smyth. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-40.
23. Olsen JV, B Blagoev, F Gnad, B Macek, C Kumar, P Mortensen and M Mann. (2006). Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127:635-48.
24. Cox J and M Mann. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26:1367-72.
25. Cox J, N Neuhauser, A Michalski, RA Scheltema, JV Olsen and M Mann. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 10:1794-805.
26. Luber CA, J Cox, H Lauterbach, B Fancke, M Selbach, J Tschopp, S Akira, M Wiegand, H Hochrein, M O'Keeffe and M Mann. (2010). Quantitative proteomics reveals subset-specific viral recognition in dendritic cells. Immunity 32:279-89.
27. Zych J, MA Stimamiglio, AC Senegaglia, PR Brofman, B Dallagiovanna, S Goldenberg and A Correa. (2013). The epigenetic modifiers 5-aza-2'-deoxycytidine and trichostatin A influence adipocyte differentiation in human mesenchymal stem cells. Braz J Med Biol Res 46.

28. Makpol S, A Zainuddin, KH Chua, YA Yusof and WZ Ngah. (2012). Gamma-tocotrienol modulation of senescence-associated gene expression prevents cellular aging in human diploid fibroblasts. Clinics (Sao Paulo) 67:135-43.
29. Kim J, YS Choi, S Lim, K Yea, JH Yoon, DJ Jun, SH Ha, JW Kim, JH Kim, PG Suh, SH Ryu and TG Lee. (2010). Comparative analysis of the secretory proteome of human adipose stromal vascular fraction cells during adipogenesis. Proteomics 10:394-405.
30. Lennon DP, SE Haynesworth, DM Arm, MA Baber and AI Caplan. (2000). Dilution of human mesenchymal stem cells with dermal fibroblasts and the effects on in vitro and in vivo osteochondrogenesis. Dev Dyn 219:50-62.
31. Cheung TH and TA Rando. (2013). Molecular regulation of stem cell quiescence. Nat Rev Mol Cell Biol 14:329-40.
32. Mathieu J and H Ruohola-Baker. (2013). Regulation of Stem Cell Populations by microRNAs. Adv Exp Med Biol 786:329-51.
33. Miranda HC, RH Herai, CH Thome, GG Gomes, RA Panepucci, MD Orellana, DT Covas, AR Muotri, LJ Greene and VM Faca. (2012). A quantitative proteomic and transcriptomic comparison of human mesenchymal stem cells from bone marrow and umbilical cord vein. Proteomics 12:2607-17.
34. Rebelatto CK, AM Aguiar, AC Senegaglia, CM Aita, P Hansen, F Barchiki, C Kuligovski, M Olandoski, JA Moutinho, B Dallagiovanna, S Goldenberg, PS Brofman, LS Nakao and A Correa. (2009). Expression of cardiac function genes in adult stem cells is increased by treatment with nitric oxide agents. Biochem Biophys Res Commun 378:456-61.
35. Miyamoto T, H Iwasaki, B Reizis, M Ye, T Graf, IL Weissman and K Akashi. (2002). Myeloid or lymphoid promiscuity as a critical step in hematopoietic lineage commitment. Dev Cell 3:137-47.
36. Parekkadan B, AL Fletcher, M Li, MY Tjota, A Bellemare-Pelletier, JM Milwid, JW Lee, ML Yarmush and SJ Turley. (2012). Aire controls mesenchymal stem cell-mediated suppression in chronic colitis. Mol Ther 20:178-86.
37. Bolontrade MF, L Sganga, E Piaggio, DL Viale, MA Sorrentino, A Robinson, G Sevlever, MG Garcia, G Mazzolini and OL Podhajcer. (2012). A specific subpopulation of mesenchymal stromal cell carriers overrides melanoma resistance to an oncolytic adenovirus. Stem Cells Dev 21:2689-702.
38. Jaager K, S Islam, P Zajac, S Linnarsson and T Neuman. (2012). RNA-seq analysis reveals different dynamics of differentiation of human dermis- and adipose-derived stromal stem cells. PLoS One 7:e38833.
39. Najar M, G Raicevic, F Jebbawi, C De Bruyn, N Meuleman, D Bron, M Toungouz and L Lagneaux. (2012). Characterization and functionality of the CD200-CD200R system during mesenchymal stromal cell interactions with T-lymphocytes. Immunol Lett 146:50-6.
40. Pietila M, S Lehtonen, E Tuovinen, K Lahteenmaki, S Laitinen, HV Leskela, A Natynki, J Pesala, K Nordstrom and P Lehenkari. (2012). CD200 positive human mesenchymal stem cells suppress TNF-alpha secretion from CD200 receptor positive macrophage-like cells. PLoS One 7:e31671.
41. Jin HJ, SK Park, W Oh, YS Yang, SW Kim and SJ Choi. (2009). Down-regulation of CD105 is associated with multi-lineage differentiation in human umbilical cord blood-derived mesenchymal stem cells. Biochem Biophys Res Commun 381:676-81.
42. Cheifetz S, T Bellon, C Cales, S Vera, C Bernabeu, J Massague and M Letarte. (1992). Endoglin is a component of the transforming growth factor-beta receptor system in human endothelial cells. J Biol Chem 267:19027-30.

FIGURES

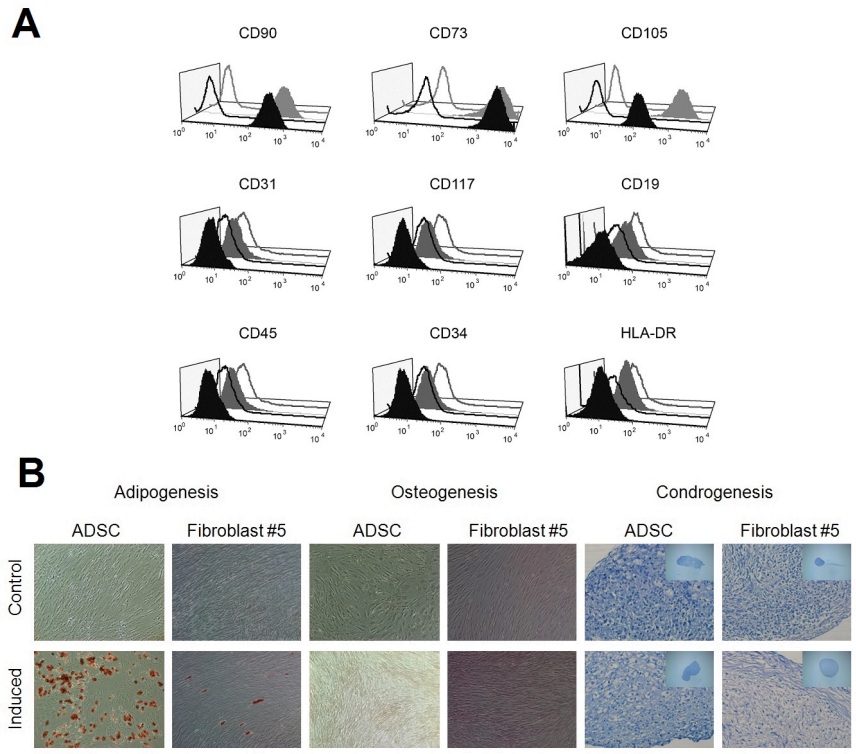


Figure 1. Immune and functional phenotypes of ADSCs and fibroblasts. **(A)** Analysis of immune phenotypes by flow cytometry. ADSCs (black lines) and fibroblasts (gray lines) were labeled with antibodies against the indicated antigens, and analyzed by flow cytometry. Representative histograms are displayed. Isotype controls are shown as empty lines/areas and the solid histograms indicate reactivity with the antibody. **(B)** Differentiation of ADSCs and fibroblasts. Both cell types at passage P5 were incubated for 21 days in the presence of specific agents inducing differentiation into adipocytes, osteoblasts and chondrocytes (see materials and methods for media composition). Differentiation into the adipocyte lineage was demonstrated by staining with Oil Red O; Alizarin Red S staining shows mineralization of the extracellular matrix; Toluidine Blue shows the deposition of proteoglycans and lacunae. Untreated control cultures without adipogenic, osteogenic or chondrogenic differentiation stimuli are shown on the upper line of photographs. The bar indicates 200 μm.

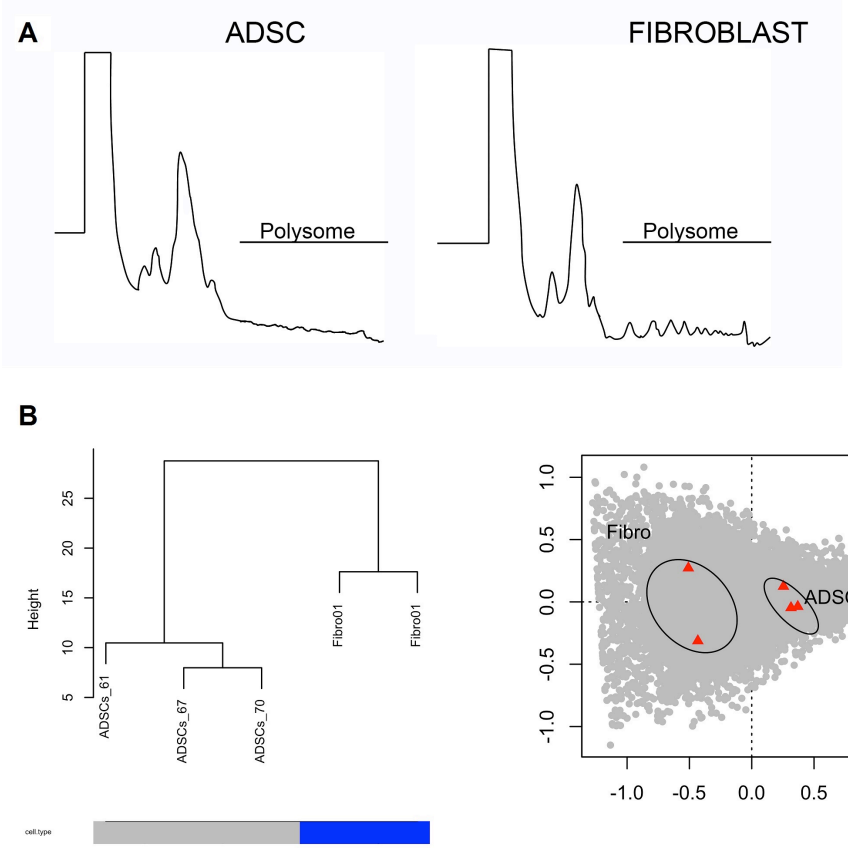


Figure 2. Polysome profiles and internal consistency of RNA-Seq data. **(A)** Polysome profile analysis of ADSCs (left) and fibroblasts (right) obtained by fractionation of cytoplasmic extracts in sucrose gradients 10-50%. **(B)** Hierarchical clustering and correspondence analysis showing the internal consistency of the data. Left panel: dissimilarity based (bottom-up) hierarchical clustering was performed on the log-transformed counts of genes for the various samples. Initially, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, according to some distance measure (in this case, a complete linkage approach), continuing until there is one single cluster. The process is visualized as a dendrogram: the first branching event separates fibroblast (Fibro) from adipose-derived stromal cell (ADSC) samples. The numbers of the samples correspond to the donors. The height axis represents the distance between each branching event. Cell type accounted for the largest proportions of the variance in both analyses, evidence of the consistency of the experiments. Right panel: Correspondence

analysis (COA) with the samples. The x-axis represents the first component (that explaining the most variance x %) and the y-axis the second component (representing x % of the variance).

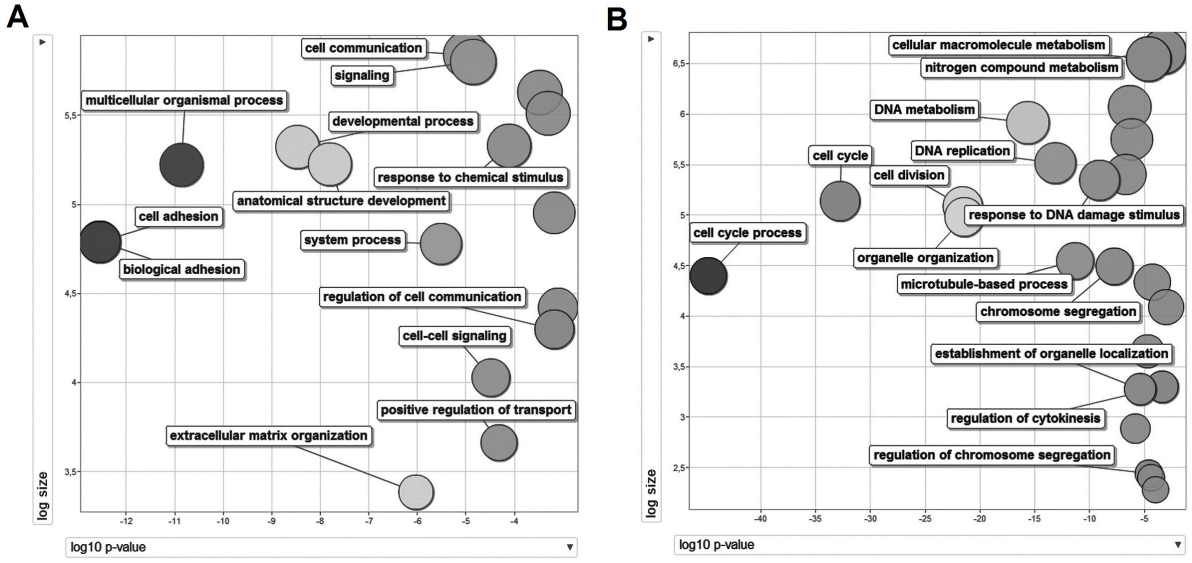


Figure 3. GO analyses, one for the genes upregulated in ADSCs **(A)** and one for the genes upregulated in fibroblasts **(B)**. REVIGO visualization of the GO analyses performed with GOrilla is shown. For these analyses, a target gene list and a background list were used: the target list is composed of genes represented with FDR < 0.0001; the background list included all the 17340 genes identified in this study. The log10 of the p-value of each GO after REVIGO analyses is plotted on the x-axis and the log 10 of the size of GOs on the y-axis.

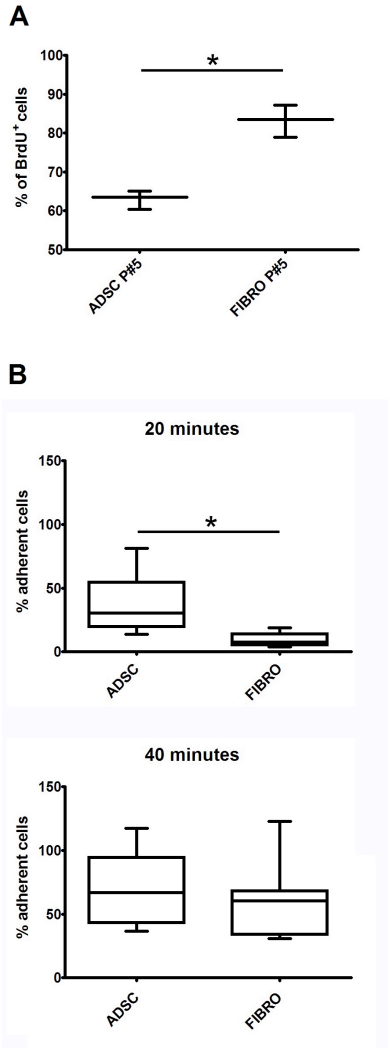


Figure 4. Comparison between ADSC and fibroblast proliferation and adhesion. **(A)** Proliferation was assessed by BrdU incorporation over 24 hours. The mean numbers of BrdU-positive cells for three donors of ADSCs and for technical triplicates of fibroblasts. BrdU, bromodeoxyuridine. **(B)** Adhesion assay. Number of adherent cells after 20 and 40 minutes in culture was evaluated by counting the number of non-adhered cells in the culture and subtracting. *$P < 0.05$.

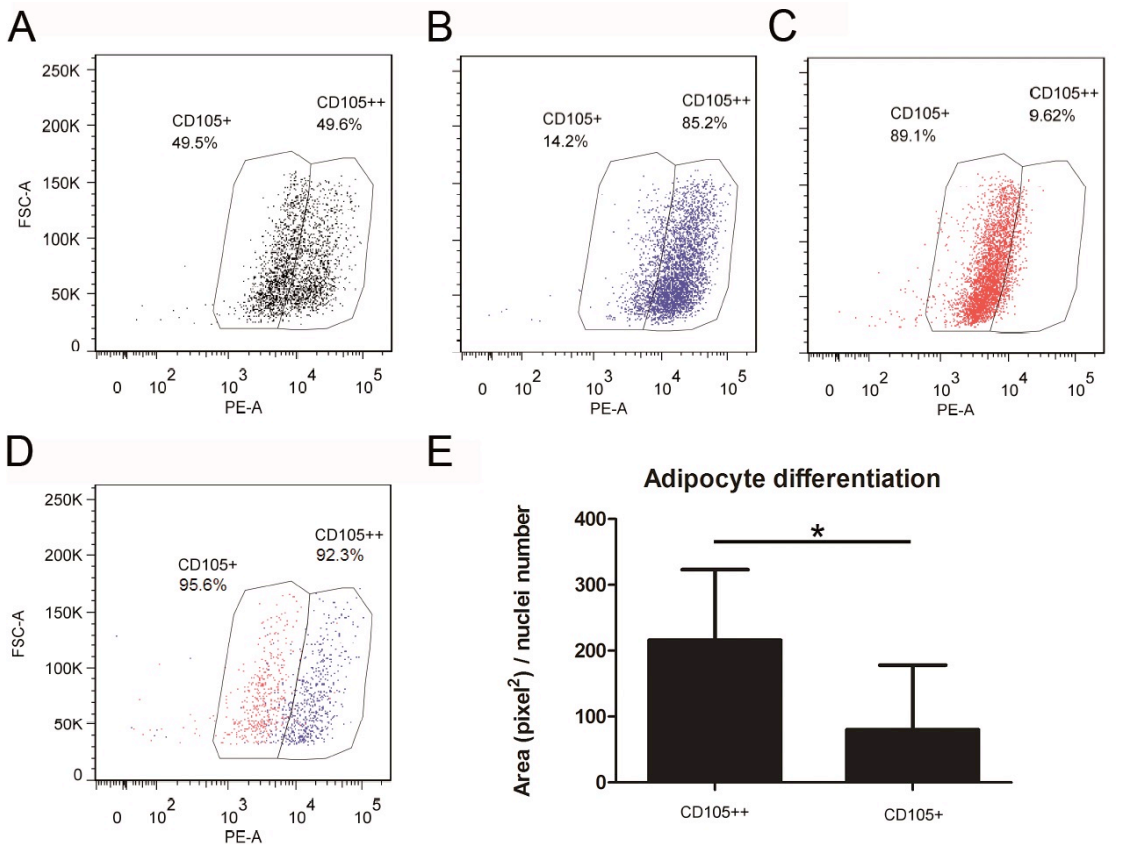


Figure 5. Isolation and differentiation of weakly and strongly CD105-positive cells from a mixed population of ADSCs and fibroblasts. **(A)** Representative FSC and FL2 plot of the 1:1 mixed sample of ADSCs and fibroblasts. CD105 expression by **(B)** ADSCs and **(C)** fibroblasts before sorting. **(D)** Representative overlay dot plot of sorted $CD105^{+}$ and $CD105^{++}$ cells. The percentages shown are according to the gating strategy defined in **(A)** and are referenced to the parental mixed cell population. **(E)** Sorted $CD105^{+}$ and $CCD105^{++}$ cells induced to differentiate into adipocytes. Columns represent the quantification of Nile Red stained area in $pixel^2$ per nuclei number. The significance of differences between mean values was evaluated with Student's *t*-test. $*P < 0.05$.

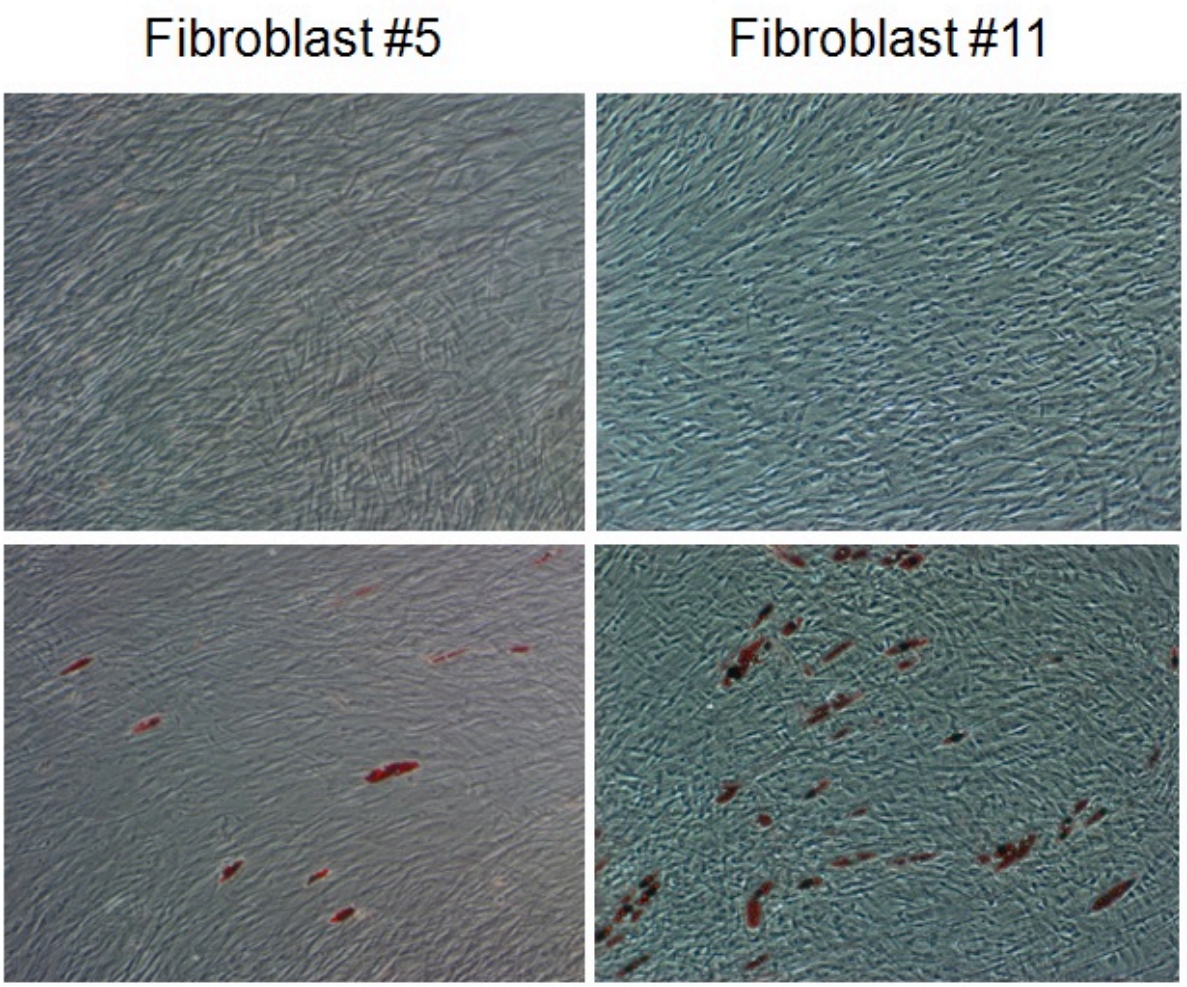


Figure S1. Fibroblast cultures at passages P5 and P11 induced to differentiate into adipocytes after 21 days. Untreated control cultures are shown on the upper line of photographs. The bar indicates 200 μm.

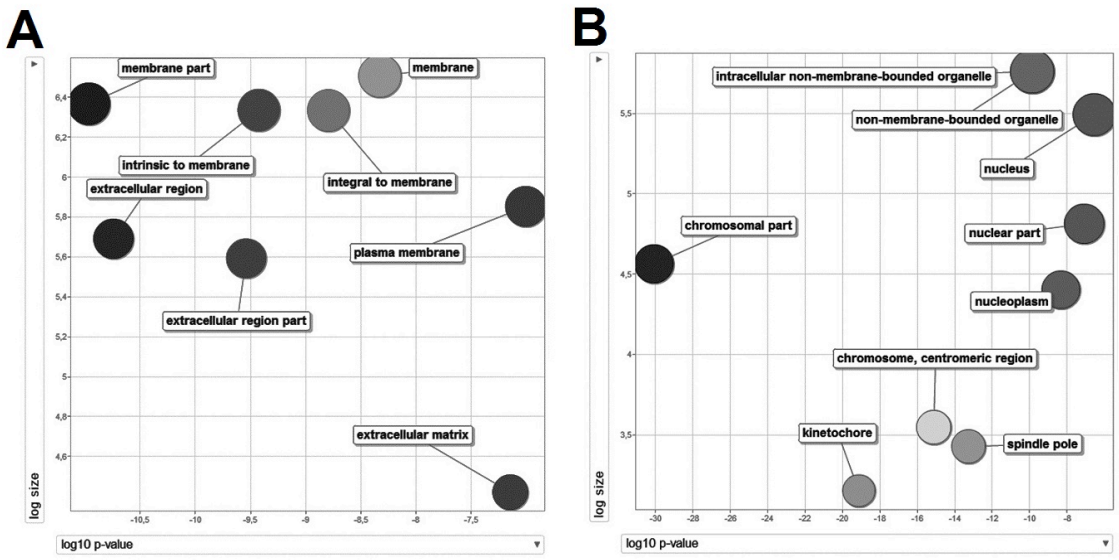


Figure S2. GO analyses for cellular component terms in ADSCs and fibroblasts. REVIGO visualization of the GO analyses with GOrilla is shown. The log10 of the p-value of each GO after REVIGO analyses is plotted on the he x-axis and the log 10 of the size of GOs on the y-axis. ADSC terms are relevant mainly to the plasma membrane and extracellular compartment.

TABLES

Table S1. RNA-Seq data for polysome-associated mRNA of ADSCs and fibroblasts.

Table S2. Previously described surface markers of fibroblasts and ADSCs. The presence of these markers did not differ between the two cell types such that they were not differentially expressed (not statistically significant).

Table S3. Comparison between sequencing and cytometry data for the surface markers CD90, CD73 and CD105. LogFC and *P*-value of the ADSC/fibroblast ratio are shown in the columns. MFI: mean fluorescence intensity.

Table S4. List of genes corresponding to "integral to plasma membrane" GO terms.

DATASET

Dataset 1. List of genes differentially expressed between ADSCs and fibroblasts with FDR<0.001 and FDR<0.0001. RNAseq results obtained from sequencing mRNAs purified from polysomal fractions.

Capítulo 6

El rol de la poliadenilación alternativa en la adipogénesis

6.1. Introducción

En este capítulo presentaremos otro aspecto trabajado durante la tesis, el cual postula a la poliadenilación alternativa (APA: alternative polyadenylation) como mecanismo relevante en la regulación post-transcripcional durante los primeros pasos de la adipogénesis. Primeramente se describirán brevemente los mecanismos involucrados en APA y seguidamente mencionaremos la metodología utilizada para detectar la influencia del mismo en la adipogénesis.
Al final del capítulo se adjunta el artículo publicado en la revista PLoS One.

6.2. Regulación post-transcripcional a través de APA

Como mencionamos anteriormente existen varios mecanismos de regulación post-transcripcional en las células, dentro de los cuales se incluye la poliadenilación alternativa. La poliadenilación es simplemente la adición de una cola poli-A a una molécula de ARN, es decir es una porción de secuencia del ARN que contiene (casi) exclusivamente adeninas. El proceso de la poliadenilación comienza cuando finaliza la transcripción de un gen. El extremo 3' del ARN mensajero recién sintetizado es cortado por un complejo de proteínas, el cual es también responsable de sintetizar la cola poli-A localizada en el extremo 3' del mismo. En algunos genes, sin embargo, estas mismas proteínas pueden adicionar la cola poli-A en muchas posiciones posibles alternativas, generando así múltiples transcriptos a partir de un único gen, de forma similar al proceso de splicing alternativo [178].

6.2.1. Mecanismo molecular

La maquinaria de poliadenilación en el núcleo actúa sobre productos de la ARN polimerasa II, como ser ARN mensajero precursor. En este contexto, un complejo multiproteico corta el extremo 3' del ARN mensajero recientemente producido y poliadenila el extremo producido por dicha escisión. Ambas reacciones están estrechamente acopladas. Este multicomplejo proteico está formado principalmente por: CPSF (cleavage/polyadenylation specificity factor), CstF (Cleavage Stimulation Factor), PAP (polyadenylate polymerase), CFI y CFII (cleavage factors) y más recientemente se identificó a la propia ARN polimerasa II como partícipe de este mecanismo.
La escisión ("cleavage") es catalizada por la enzima CPSF [179] y ocurre aproximadamente a 10-30 nucleótidos de su sitio de unión. Este sitio está caracterizado por presentar la secuencia AAUAAA, sin embargo, existen variantes de la misma que interactúan con CPSF en forma más débil [180]. Estudios *in vitro* mostraron que CPSF sola interactúa muy débilmente con el ARN [181]. Dos proteínas adicionales son necesarias para la unión con el ARN: CstF y CFI. La primera reconoce una región rica en GU más adelante ("downstream") del sitio de unión de CPSF [182]. CFI reconoce un tercer sitio de unión, caracterizado por el motivo UGUAA, y además puede reclutar a CPSF aún en la ausencia de AAUAAA, generando sitios alternativos de poliadenilación [183, 184]. La señal de poliadenilación, el motivo reconocido por el complejo CPSF, varía entre grupos de eucariotas. La mayoría de los sitios de poliadenilación en humano contienen el clásico motivo AAUAAA.
La escisión del preARN mensajero ocurre en general previamente a la terminación de la transcripción (la polimerasa II sigue transcribiendo posterior al sitio de escisión). Las proteínas necesarias están interaccionando a su vez con la polimerasa II (CPSF, CstF) [185]. Cuando el RNA es escindido comienza la poliadenilación, la que es catalizada por la enzima PAP. La misma es la encargada de sintentizar la cola poli-A al final del transcripto. Cuando la secuencia de adeninas sintetizadas alcanza los $\sim$ 250 nucleótidos, PAP no logra seguir en contacto con CPSF, por lo que culmina la poliadenilación. Esto determina el largo de la cola poli-A [186, 187]. CPSF señaliza a la ARN polimerasa II que culmine la transcripción. Cuando la misma alcanza la secuencia de terminación AAUAAA (correspondiente a TTATTT en el ADN), la transcipción es finalizada [188]. La maquinaria de la poliadenilación está a su vez físicamente unida al espliceosoma, el complejo responsable de remover los intrones del preARN mensajero.
Una proteína que reconoce poli-As (poly(A)-binding protein) interactúa con esta región y promueve la exportación del ARN mensajero afuera del núcleo hacia el citoplasma, inhibiendo a su vez la degradación [189]. Los ARNs no exportados son degradados por exosomas [190]. Las proteínas poly(A)-binding pueden reclutar varias proteínas que afectan la traducción, una de las mismas incluye a eIF4G, la cual recluta a su vez a la

subunidad 40S del ribosoma. No obstante, la proteína poly(A)-binding no es necesaria para la traducción de todos los ARNs [191].

6.2.2. Poliadenilación alternativa y sus efectos "downstream"

Muchos genes codificantes poseen más un sitio de poliadenilación, es decir sitios de poliadenilación alternativa (APA). La figura 6.1 muestra dos ejemplos de APA. Los sitios de poliadenilación pueden localizarse en el extremo 3' del ARN mensajero (figura 6.1 A), generando ARN mensajeros que difieren únicamente en el largo de la región 3'UTR. Por otro lado, los mismos pueden ubicarse entre exones (lo cual es menos frecuente), por lo que se generan dos ARN mensajeros del mismo gen que difieren en su número de exones (figura 6.1 B), generando proteínas diferentes.
La elección del sitio de poliadenilación se ve influenciada por estímulos extracelulares que, en general, dependen de la expresión de proteínas que forman parte de la maquinaria de poliadenilación [192, 193]. Por ejemplo, en el caso de un tipo de células del sistema inmune, los macrófagos, la expresión de una subunidad de CstF aumenta en respuesta a la presencia de lipopolisacáridos en la superficie bacteriana. Este hecho (el aumento de expresión de CstF) resulta en la elección de sitios alternativos de poliadenilación (más débiles), generando en este caso transcriptos más cortos. Los mismos carecen de algunos elementos regulatorios, lo que los hace tener vidas medias más largas y producir mayor cantidad de la proteína que codifican [192]. Por otro lado, proteínas de unión al ARN (RBP) adicionales (no necesariamente las pertenecientes a la maquinaria de poliadenilación) pueden influenciar la elección del sitio de poliadenilación [193, 194], a la vez que la metilación del ADN en sitios cercanos a la señal de poliadenilación [195].
microARNs son moléculas de ARN no codificante reguladoras que reconocen sitios de unión en las regiones 3'UTR de transcriptos. Su función es, en general, reprimir la expresión génica a través de diferentes mecanismos posibles (ver siguiente sección 6.3). APA genera transcriptos con diferentes largos de 3'UTR (6.1 A) [196], por lo que genera transcriptos posiblemente con diferentes sitios de unión a micro ARNs. El largo de la 3'UTR juega un rol relevante en el destino del mensajero. El mismo puede terminar siendo traducido, si la 3'UTR no presenta sitios de unión para ciertos micro ARNs adecuados, porque por ejemplo su 3'UTR es más corta debido a APA, o el mismo, puede terminar siendo degradado si su 3'UTR más larga presenta sitios de unión blanco para ciertos micro ARNs presentes.

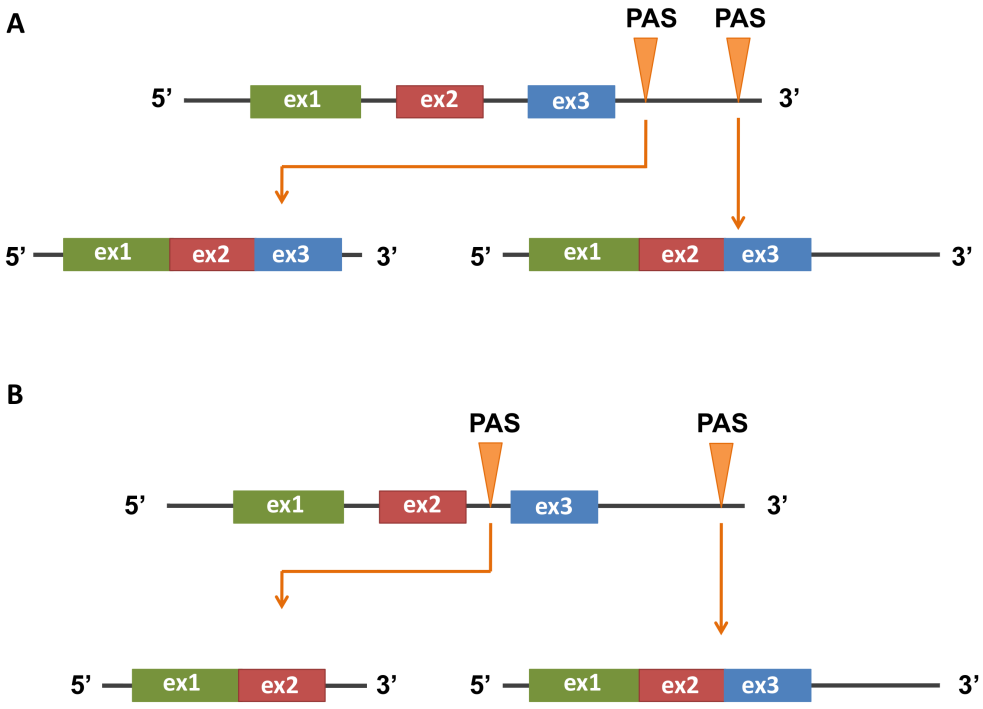


FIGURA 6.1: Esquema representativo de la poliadenilación alternativa. A) Se muestran dos sitios de poliadenilación diferentes. Dependiendo de cual se utilice el ARN mensajero resultante, con sus largos de 3'UTR correspondientes. B) Pueden existir sitios de poliadenilación entre exones, por lo que partiendo de un mismo pre ARN mensajero, se pueden obtener diferentes ARN mensajeros, con diferente número de exones.

6.3. miARNs como protagonistas

Los micro ARNs (miRNA) son moléculas de ARN monocatenarias, no codificantes y pequeñas, variando entre 20 y 25 nucleótidos de largo. Las mismas poseen una función regulatoria, tanto en la regulación transcripcional como post-transcripcional.

Los miRNAs son moléculas producidas a partir de sus propios genes de ADN o a partir de intrones, pero no son codificadas a proteínas [197]. El gen del miRNA tiene una secuencia cuya longitud supera el tamaño final del propio miRNA. Ésta incluye la secuencia del miRNA como también una región complementaria a la misma. Esto provoca, que durante la transcripción del miRNA se apareen ambas regiones complementarias y se forme una estructura de horquilla. Esta secuencia con su estructura se denomina pri-miRNA (miRNA primario), la cual posteriormente por la acción de una enzima llamada Drosha, es cortada en las bases de la horquilla. De esta forma se genera el pre-miRNA (ARN doble cadena conteniendo sólo las dos regiones complementaria), el cual es transportado del núcleo al citoplasma por una exportina. Una vez que el pre-miRNA se encuentra en el citoplasma es fragmentado por la enzima Dicer, que lo corta hasta la longitud final de 21 a 25 nucleótidos.

Información más detallada sobre los miRNAs pueden encontrarse en las revisiones sugeridas a continuación, tanto sobre la biogénesis de los miRNA [198], como sobre características generales y formas de acción [199, 200], como también miRNAs específicamente en células madre [201], y en el proceso de adipogénesis [112].

6.3.1. Función de los miRNAs

La función de los miRNAs si bien ha sido estudiada a fondo estos últimos años, aún queda mucho por descifrar [202]. Los miRNAs se unen a la región 3'UTR de los ARN mensajeros blanco ("targets") a través de una complementariedad imperfecta, en general, reprimiendo la traducción y estabilidad del mismo [203]. Sin embargo, se observó que ocasionalmente, un tipo particular de miRNAs activan la traducción de mensajeros blancos y regulan su estabilidad [204]. Si bien la complementariedad del miRNA con la 3'UTR es imperfecta, una región denominada "seed" (semilla) de $\sim 6-8$ nucleótidos de largo en el extremo 5' del miRNA (de la base 2 a la 7) es el determinante de la especificidad del target (ARN mensajero) [205]. Un miRNA puede tener como "target" diversos ARN mensajeros, similarmente un ARN mensajero puede ser regulado por varios miARNs [206].

Como ya fue mencionado, en general los miRNAs actúan como inhibidores de la expresión génica. Los mismos, una vez en el citoplasma, posterior a la acción de la enzima Dicer, son incorporados en un complejo ribonucleoproteico denominado RISC (RNA-induced silencing complex) que contiene una proteína miembro de la familia Argonauta. Guiado por la complementariedad entre el miRNA y el mensajero blanco, la inhibición de la expresión génica mediada por el complejo RISC se puede dividir en tres mecanismos: i) escisión sitio-específica, ii) degradación potenciada del mensajero y iii) inhibición de la traducción. El primer proceso, es el menos usual, se restringe sólo a miRNAs con complementaridad perfecta con la 3'UTR blanco, lo que se denomina RNAi (RNA interference). Los otros dos, los cuales están asociados a la complementaridad imperfecta mencionada anteriormente, son los principales mecanismos de acción de los miRNA en mamíferos. La degradación potenciada sugiere que el miRNA acelera la degradación del mensajero blanco promoviendo la vía de decaimiento del ARN 5'-3' (5'-to-3' mRNA decay pathway) [207]. En esta vía, el mensajero es deadenilado y luego se le retira el extremo 5' (5' cap) y finalmente es degradado por una exonucleasa XRN1 [208, 209]. La inhibición de la traducción puede darse a muchos niveles, tanto en la iniciación (translation initiation), en la elongación y en la terminación. Estos mecanismos aún no han sido dilucidados al detalle [210], sin embargo existe cada vez más evidencia que apunta a una inhibición en la iniciación [211]

No está completamente claro si ambos procesos son excluyentes o están inter relacionados. En varios casos aparentan ser independientes, pero algunos estudios muestran que pueden actuar complementariamente [212].

6.4. Cambios en niveles de proteinas no se reflejan en mensajero

Como ya fue mencionado anteriormente, la correlación entre niveles de proteínas y niveles de ARN mensajero es baja, debido a los mecanismos de regulación post-transcripcional. Por este motivo, modelos lineales básicos, que predicen el nivel de proteína (o los cambios del mismo entre dos condiciones) a partir únicamente de los niveles de ARN mensajero en la célula (o los cambios del mismo entre dos condiciones, por ejemplo: log fold change; logFC) son poco precisos, resultando en correlaciones bajas entre ambas magnitudes. Estudios en humanos y levadura han encontrado una relación lineal entre las abundancias de proteína y ARN en escala logarítmica [213–215]. Se puede construir un modelo lineal entre ambas variables (en escala logarítmica):

$$\ln(protein) = b_1 \cdot \ln(mRNA) + b_0,$$

en donde b1 representa la eficiencia traduccional como función de la cantidad de mensajero y b_0 el intercepto, el cual puede acomodar todos los efectos aditivos en escala logarítmica. Haciendo cuentas se puede estimar la cantidad de proteína como sigue:

$$protein = e^{b_1 \cdot \ln(mRNA) + b_0}$$

En general, en estudios de expresión diferencial se comparan dos condiciones:

$$protein_x = e^{b_1 \cdot \ln(mRNA_x) + b_0} \text{ y } protein_y = e^{b_1 \cdot \ln(mRNA_y) + b_0},$$

las cuales se combinan generalmente en un cociente denominado log fold-change (logFC) de la siguiente forma:

$$\ln \frac{protein_x}{protein_y} = \ln \frac{e^{b_1 \cdot \ln(mRNA_x) + b_0}}{e^{b_1 \cdot \ln(mRNA_y) + b_0}} = \ln(e^{b_1 \cdot \ln(mRNA_x) + b_0}) - \ln(e^{b_1 \cdot \ln(mRNA_y) + b_0})$$
$$= (b_1 \cdot \ln(mRNA_x) + b_0) - (b_1 \cdot \ln(mRNA_y) + b_0) = b_1 \cdot (\ln(mRNA_x) - \ln(mRNA_y))$$

$$\Leftrightarrow \ln \frac{protein_x}{protein_y} = b_1 \cdot \ln \frac{mRNA_x}{mRNA_y} \Leftrightarrow logFC_{protein} = b_1 \cdot logFC_{mRNA}$$

Esta ecuación y el hecho que b_1 ha sido estimado como menor que 1, demuestra que existe una compresión de rangos de tamaño $b_1$en logFC si se considera logFC_{mRNA} como predictor. El tamaño de este coeficiente es la eficiencia traduccional (en escala logarítmica) como función de la cantidad de mRNA.

6.5. Modelos lineales para comprobar el efecto de APA en la regulación post-transcipcional

Se desea mejorar este modelo, de forma tal de mejorar la varianza explicada por el mismo, incluyendo los efectos de regulación post-transcripcional. Como ya fue discutido, el mecanismo de regulación post-transcripcional APA produce transcriptos con 3'UTRs de diferentes largos, implicando diferencias en los sitios de unión a miRNAs. 3'UTR más largas tienden a tener más sitios de unión a miRNAs, por el contrario, las más cortas pueden carecer de ellas. En este contexto, el uso diferencial de transcriptos implica "sensibilidad" diferencial a miRNAs. Se propone en este trabajo a APA como uno de los mecanismos principales de regulación post-transcripcional, sobre todo por causa del efecto de miRNAs en las 3'UTR de tamaños diferentes. En el manuscrito anterior (capítulo 4), se determinó la correlación entre $\text{logFC}_{proteinas}$ y logFC_{mRNA} utilizando datos de proteómica de Molina *et al* [216] de SILAC, obteniendo para muchos casos correlaciones muy bajas. En este trabajo ajustamos modelos lineales que incorporan el efecto de los miRNA y el uso diferencial de transcriptos en la predicción del logFC de proteínas para mejorar la varianza explicada por el modelo. Para ello se incluye el efecto de los sitios de unión de los miRNAs en los respectivos transcriptos, lo cual puede modelarse de la siguiente manera:

$$\ln(protein) = b_0 + b_1 \cdot \ln(mRNA) + b_2 \cdot miR,$$

en donde miR es la proporción de reads asignados a un transcripto en particular (para un gen en particular) que tiene un sitio de unión para ese miRNA. Por lo tanto, como fue mostrado anteriormente, el mejor predictor para la cantidad de proteína sería:

$$protein = e^{b_0 + b_1 \cdot \ln(mRNA) + b_2 \cdot miR}$$

Nuevamente, considerando dos condiciones:

$$\frac{protein_x}{protein_y} = \frac{e^{b_0 + b_1 \cdot \ln(mRNA_x) + b_2 \cdot miR_x}}{e^{b_0 + b_1 \cdot \ln(mRNA_y) + b_2 \cdot miR_y}} = \frac{e^{b_1 \cdot \ln(mRNA_x)}}{e^{b_1 \cdot \ln(mRNA_y)}} \cdot \frac{e^{b_2 \cdot miR_x}}{e^{b_2 \cdot miR_y}}$$

.

$$\frac{protein_x}{protein_y} = \left(\frac{mRNA_x}{mRNA_y}\right)^{b_1} \cdot e^{b_2 \cdot (miR_x - miR_y)}$$

$$\ln \frac{protein_x}{protein_y} = b_1 \cdot \ln \frac{mRNA_x}{mRNA_y} + b_2 \cdot (miR_x - miR_y)$$

La diferencias de los sitios de unión de transcriptos alternativos entre ambas condiciones ($miR_x - miR_y$) pueden determinarse de la siguiente forma. Cada gen expresa diferentes transcriptos en diferentes proporciones. A su vez, cada transcripto particular (dependiendo de su 3'UTR resultante) contiene sitios de unión para varios miRNAs. Considerando todos los miRNA de humanos y los mensajeros blancos conocidos, se puede pesar la presencia de cada miRNA en cada transcripto por el uso diferencial del mismo en cada condición. Es decir, si un gen_X, que genera un transcripto_A y transcripto_B, se utiliza en proporciones 20 % y 80 % en la condición inducido (IN), respectivamente y 60 % y 40 % en control (CT), las diferencias (IN−CT) en las proporciones serían: −40 % para A y 40 % para B. Los miRNAs presentes en A tendrán menos efecto que los que actúen en B en la condición IN y los presentes en B mayor efecto. Se pondera, por lo tanto, la presencia de los miRNAs por los coeficientes correspondientes −0,4 y 0,4 respectivamente para A y B. De esta forma, cada transcripto obtiene un vector de presencia de miRNAs ponderado por su diferencia de uso en ambas condiciones. Finalmente, se sumarizan los vectores de presencia ponderada por gen. De esta forma cada gen obtiene el uso promedio (ponderado) de cada miRNA (y un 0 si el miRNA no tiene como blanco ese gen). A partir de estos datos para todos los genes, conjuntamente con los datos de expresión para cada gen (RNA-seq) y la correspondiente proteína (SILAC), se puede implementar el modelo lineal mencionado. Por más detalles referirse al artículo, sobre todo la figura 6.
Primeramente, la covariable $miR_x - miR_y$ implica un único miRNA incluído en el modelo. Se determina un modelo para cada miRNA conocido. Posteriormente, se incorporan de a dos miRNAs. Finalmente, se incluyen de a 3, 4 y hasta 5. El conjunto de modelos generados se puede evaluar con el Bayesian Information Criterion (BIC), una medida de ajuste de modelos. Se seleccionan los mejores 5 y se determina la varianza explicada por esos modelos.
De acuerdo a nuestros resultados, la incorporación del efecto de los miRNAs de este modo aumenta considerablemente la varianza explicada en muchos casos a más del doble comparado con el modelo base.

6.6. Resumen de resultados

En este trabajo por un lado, se analizó el uso diferencial de transcriptos durante la diferenciación a adipogénesis utilizando las mismas muestras definidas anteriormente en 4.4. Se analizan los largos de 3'UTR con el fin de encontrar una tendencia general en la diferenciación. Es decir, por ejemplo, si durante la adipogénsis los transcriptos expresados tienden a tener 3'UTRs más largas o más cortas (con respecto al control, célula madre mesenquimal sin diferenciar). Nuestros análisis muestran que existe una tendencia en expresar 3'UTR más largas durante la diferenciación, por lo que los transcriptos son más susceptibles al efecto de los miRNAs. Observamos una diferencia media de 18 bases en el largo de transcriptos (inducido-control). Encontramos a su vez más genes que presentan 3'UTR más largas (6608) que más cortas (5931), y esta diferencia en número es estadísticamente significativa (p-val$< 1 \times 10^{-8}$).
Por otro lado comparamos las diferencias de expresión de ARN mensajero y proteínas utilizando

datos de proteómica previamente publicados. El modelo lineal básico que predice el cambio en el nivel de proteína, entre inducido y control, ($\log FC_{proteina}$) únicamente a partir de los cambios en nivel de expresión del ARN mensajero ($\log FC_{mRNA}$) tienen en general un ajuste muy pobre en los datos considerados, resultando en correlaciones muy bajas entre ambas medidas. En este trabajo proponemos modelos lineales que incorporan como covariable el efecto de miRNAs y el uso diferencial de transcriptos, obteniendo de esta forma correlaciones mucho más altas. Por último, identificamos una serie de miRNAs relevantes en la adipogénesis.
A su vez observamos que genes previamente descritos como relevantes en los procesos de diferenciación (los genes pertenecientes a la denominada Plurinet) están enriquecidos en 3'UTR más largas en la diferenciación.

6.7. Artículo

OPEN ACCESS Freely available online

PLOS ONE

Role of Alternative Polyadenylation during Adipogenic Differentiation: An *In Silico* Approach

Lucía Spangenberg[1], Alejandro Correa[2], Bruno Dallagiovanna[2], Hugo Naya[1,3]*

1 Bioinformatics Unit, Institut Pasteur Montevideo, Montevideo, Uruguay, **2** Instituto Carlos Chagas, Fiocruz-Paraná, Curitiba, Paraná, Brazil, **3** Departamento de Producción Animal y Pasturas, Facultad de Agronomía, Universidad de la República

Abstract

Post-transcriptional regulation of stem cell differentiation is far from being completely understood. Changes in protein levels are not fully correlated with corresponding changes in mRNAs; the observed differences might be partially explained by post-transcriptional regulation mechanisms, such as alternative polyadenylation. This would involve changes in protein binding, transcript usage, miRNAs and other non-coding RNAs. In the present work we analyzed the distribution of alternative transcripts during adipogenic differentiation and the potential role of miRNAs in post-transcriptional regulation. Our *in silico* analysis suggests a modest, consistent, bias in 3′UTR lengths during differentiation enabling a fine-tuned transcript regulation via small non-coding RNAs. Including these effects in the analyses partially accounts for the observed discrepancies in relative abundance of protein and mRNA.

Citation: Spangenberg L, Correa A, Dallagiovanna B, Naya H (2013) Role of Alternative Polyadenylation during Adipogenic Differentiation: An *In Silico* Approach. PLoS ONE 8(10): e75578. doi:10.1371/journal.pone.0075578

Editor: Qiong Wu, Harbin Institute of Technology, China

Received July 2, 2013; **Accepted** August 14, 2013; **Published** October 15, 2013

Copyright: © 2013 Spangenberg et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by grants from Ministério da Saúde and Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq, FIOCRUZ-Pasteur Research Program and Fundação Araucária. Lucía Spangenberg received a fellowship from ANII (Agencia Nacional de Investigación e Innovación, Uruguay); Bruno Dallagiovanna was supported by CNPq, Hugo Naya by FIOCRUZ-Pasteur and Alejandro Correa received a fellowship from Fundação Araucária. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: naya@pasteur.edu.uy

Introduction

Mesenchymal stem cells (MSCs) are able to differentiate to mutiple cell types including those in bone, ligament, muscle and connective tissue [1] among others and are thus the focus of stem cell-based therapies. Tissue engineering [2], therapy for degenerative and autoimmune diseases [3,4] and cardiac tissue repair [5,6] are some of the areas of focus in adult stem cell research. Although much progress has been made, the regulatory processes controlling MSC differentiation remains poorly understood. Adipose derived human MSCs are easily isolated from pools of cells resident in vascular stroma of adipose tissue. Since adipose tissue is ubiquitous and easily accessible with minimally invasive procedures [7], it is an ideal resource for research and development of cell-based therapy. Understanding MSC commitment to differentiation to a specific cell type is essential for the successfully repair or regeneration of injured tissues. The switch from self-renewal to differentiation is regulated by many factors including cytokines, growth factors and extracellular matrix components present in a given microenvironment [8]. Nevertheless, the transcriptional and post-transcriptional regulatory processes remain not fully understood.

Gene expression analysis has provided great insights into the regulatory networks determining self-renewal and differentiation processes [9,10]. Deep sequencing techniques have also played a key role in clarifying the complex mechanisms involved. Regulation is at both the transcriptional [11] and post-transcriptional [12,13] levels. Also non-coding elements are involved [14] in the regulatory machinery [15]. In order to address post-transcriptional regulation, many groups are focusing on sequencing mRNAs associated to translating polysomes and comparing them with total RNA [12,13,16].

Expression analysis with deep sequencing methods enables the distinction of alternative transcripts of the same gene. In this context, the focus is shifted from analyzing genes as an entity (represented by a single canonical transcript) towards an alternative transcript usage model, non-coding RNAs (e.g., miRNAs), alternative splicing, 3′UTR switching, polyadenylation [17,18], etc. Alternative polyadenylation (APA) results in subpopulations of transcripts differing in 3′UTR length, which makes them more or less susceptible to the regulation by miRNAs (shorter 3′UTR might have fewer miRNA binding sites) [19,20]. A recent study has shown a role for APA in muscle stem cell development. The Pax3 protein represses differentiation in that transcripts can be targeted by mir-206. Boutet *et al.* [18] showed that different muscle tissues process Pax3 transcripts differently through APA, in which transcripts were differentially targeted by miR-206 based on 3′UTR length. In turn, different Pax3 protein levels result in functional changes in muscle stem cell behavior. Other groups assessed this type of mechanism in a global way, analyzing 3′UTRs length patterns of all genes in different scenarios. Sandberg *et al.* showed a global shortening of 3′UTRs in proliferating murine CD4+ T lymphocytes [21], and Kolle *et al* showed human embryonic stem cells to have extended 3′UTRs. The latter study also found alternative gene model usage [13]. In addition, Ji and collaborators reported that mouse genes tend to express longer 3′UTRs during the progression of embryonic development [22].

In the present work, we focus on post-transcriptional regulation during adipogenesis, specifically analyzing transcript usage differ-

ences based on 3′UTR length. We analyze data previously obtained using RNAseq [12] to study the initial phases of adipocyte differentiation of adipose-derived human mesenchymal stem cells (hASCs). Total mRNA (total) and mRNAs associated with translating ribosomes (polysomal fraction) were sequenced at two time points: 0 and 3 days after induction. We found that 3′UTRs tended to be longer after cells were induced, thereby potentially providing more miRNA binding sites. A mean difference of 18 bases in transcript length was found in induced versus control conditions. In our previous study, based on a subset of the proteomic data of Molina *et al.* [23], we found a low correlation between protein and corresponding mRNA changes. Standard linear models predicting changes in protein levels based only on mRNA changes were inaccurate. Here, we propose linear models that incorporate the effect of miRNAs on protein changes, which substantially improve the correlation between protein and mRNA change. Furthermore, our linear models indicate several miRNAs that could potentially be involved in post-transcriptional regulation of genes relevant for adipogenesis. Moreover, we also observed that genes previously described as involved in the differentiation process (Plurinet genes [24]) are enriched in longer 3′UTR in the induced condition.

Results

1 Global analysis of differential transcript usage

Previous studies have shown that the use of alternative polyadenylation sites, which generates transcripts with varying 3′UTR length (shorter or longer), are associated with cells having higher proliferation rates [21,22] (those generally having shorter 3′UTR), with cells undergoing differentiation [13] (longer 3′UTR) and with post-transcriptional regulation events in general. We determined alternative transcript usage by comparing the proportions of FPKM of each transcript for IN (induced samples, differentiating cells) vs. CT (control samples, undifferentiated cells). Analysis was done with total and polysomal fractions (see 2), however, total RNA was analyzed in greater detail to more accurately recover all alternative transcripts. Transcripts destabilized by miRNA are not expected to be associated with polysomes.

A preliminary global analysis of our data showed that the average 3′UTR length, weighted by the proportion of transcripts used for each gene, differed under IN compared with CT conditions. The mean difference was 18 bases, and 11 bases when outliers were excluded. In this context, we defined outliers as 3′UTRs with an average difference between conditions (IN–CT) longer than 1 kb. We excluded extreme values to avoid a bias in the determination of the mean (only for these calculations). Both lengths (18 and 11) are sufficient for generation of an additional miRNA binding site (see Discussion). Extension of 3′UTR regions was found in 6608 genes ($IN-CT>0$, weighted by the proportion of transcripts), whereas 5931 had shorter 3′UTRs ($IN-CT<0$). As such, we observed a tendency for longer 3′UTR under IN conditions compared with CT ($p<1*10^{-8}$, Wilcoxon test). We tested our data using the Cochran-Mantel-Haenszel (CMH) statistic, as in Fu *et al.* [25] to assess the significant of the differences observed. Since several genes have more than two transcripts and the length of the 3′UTR is a quantitative variable, the linear trend alternative to independence test [26] is more accurate than a standard χ^2 test. CMH determines a trend value for each gene, based on a Pearson correlation, with a corresponding p-value. In our setup, a positive correlation is observed if there is a tendency for longer 3′UTRs under IN conditions and a negative correlation for longer 3′UTR in CT. From the 16832 genes tested, 5952 displayed a negative trend, 6675 a positive trend and 4205 showed no trend. Tendencies are based on the calculated correlation values needed for the CMH test. Furthermore, 182 genes were significant at an FDR <0.01. Of the significant 182 genes, 114 had a positive correlation value and 68 a negative one. This difference is again significant ($p<1*10^{-3}$, Wilcoxon test). In summary, we found that there is a modest but consistent tendency to use alternative transcripts with longer 3′UTR under IN conditions compared with CT in our dataset.

Trends observed in polysomal fractions were similar to those in total RNA fractions, however, the number of genes were smaller: 5340 genes had a negative trend (length $CT>IN$), 6152, a positive trend (length IN>CT) and 4210 no trend ($p<3.6*10^{-14}$). These trend results are also based on the correlation values used for the CMH test. Differences in the distribution of gene trends for total and polysomal fractions were significant ($p<5*10^{-4}$), but were relatively small considering the large numbers compared. Of 92 significant genes at FDR <0.01, 51 had positive correlation values and 41 negative values. A number of significant genes, each having at least 20 nucleotides of 3′UTR length difference between conditions, were found in both total and polysomal fractions (positives and negatives). The overlap list of negative genes includes: ARL6IP5, COL1A2, RPL23, CD59, THBS1, TMED9, SPARC and MFAP5, and the positive list includes: DCN, BRK1, OSTC, PEBP1, BNIP3L, SAR1A and LSM6.

The observed mean difference in whole transcript lengths between conditions was 20 bases, considering all 3′UTRs, and 12 bases without outliers (defined as before). Interestingly, the correlation between trend statistic for total and polysomal fractions was very low, $r=0.06$ ($p<1.6*10^{-13}$), pointing towards important differences in post-transcriptional regulation.

2 Large fold change differences between mRNA and proteins

Large differences can be observed between mRNA and protein products in eukaryotic cells. This is due to various types of post-transcriptional regulation including tRNA and ribosome availability, regulation by small non-conding RNAs and transcripts nucleotide composition. However, in general a reasonably good agreement (in logarithmic base) is expected [27,28]. We previously correlated protein fold changes (in mouse) determined by SILAC (Molina *et al.* [23]) and our human RNAseq data [12]. We found a relatively high correlation between our RNAseq data and a subset of Molinas data, consisting of a group of secreted proteins. However, we were unable to find a high correlation with the entire dataset, which also included nuclear proteins. Using the same data set, we addressed the reasons behind the low correlations observed between mRNA and protein fold changes. In brief, our RNAseq dataset consists of two sets: RNAseq of total RNA (total) and of polysome associated RNA (polysomal). The samples were hADS cells taken at time point 0 (control; CT) and three days after adipogenesis induction (induced; IN). Molina *et al.* measured 3T3-L1 murine stem cell protein levels at different time points during adipogenesis: day 0, 1, 3, 5 and 7. Ideally, such comparisons would be more appropriate comparing experiments from the same species, however, Molina's dataset was the most suitable available for comparison with our RNAseq analysis (see Materials and Methods). To the best of our knowledge, studies on adipogenesis comparing different species have not been reported. However, embryonic stem cell pluripotency is established and maintained by a largely conserved regulatory network in eutherian mammals [29]. Other studies have shown conserved genes and pathways

involved in mammary gland development in human and mouse similarly governing cell-fate decisions and differentiation processes [30].

A linear model for $\log FC_{protein}$ values (log fold change of protein, e.g. $log(\frac{day5}{day0})$) versus our $\log FC_{mRNA}$ values (log fold change values of mRNA, $log(\frac{IN}{CT})$) was fit for each time point in the experiment of Molina *et al.*, and residuals analyzed. Such differences (residuals of the corresponding linear model) were very large for several genes. Fig. 1 shows the differences in logFC for each time point (day 1, 3, 5 and 7) in the secretome dataset (nuclear in Fig. S1). Only those genes with the greatest differences are shown, and both RNA fractions are considered (A polysomal and B total). Genes clustered into two groups: negative differences ($logFC_{protein} \leq logFC_{mRNA}$) are shown at the bottom of Fig. 1 (green) and positive differences above (red). The large differences suggest post-transcriptional regulation of several genes potentially by small non-coding RNAs, especially miRNAs. Linear models were constructed taking into account alternative transcript usage between conditions and characterizing miRNA binding sites involved. We discuss these results in the next subsections.

3 Alternative transcripts and miRNAs help explain protein fold changes

We analyzed the effect of miRNAs targeting 3′UTR of alternative transcripts in the fold change of proteins by linear models. The base model included only $\log FC_{mRNA}$ (and the intercept) as predictor variable for the $\log FC_{protein}$. miRNA target sites were then included in order to increase the variance explained by the model. Of the 147 secreted proteins analyzed by Molina *et al.*, 111 genes were represented in our expression dataset (total and polysomal RNA), and of the 280 nuclear proteins, 214 were found in our set. In addition, we determined the relative transcript abundance per gene in our dataset (using *cuffdiff*, see 5). Once we established the miRNAs targeting those transcripts (weighting by transcript usage) and the logFC values for each gene, we predicted the effect of each miRNA on protein level. Hereinafter, when we mention models "including/considering miRNAs", we are referring to models, which incorporate the effect of the differences in miRNA target sites. First, models including each miRNA individually were constructed (694 miRNAs in total), then all combinations of two to five miRNAs were included in models. The best models were selected based on BIC (Bayesian Information Criterion).

Table 1 shows linear model results for secreted and nuclear proteins with polysomal and total RNA fractions. The base model (the effect of $\log FC_{mRNA}$ on protein change without considering any miRNAs) is shown, as well as two single miRNA models (per comparison: polysomal/total and secreted/nuclear) and the best model by BIC (including one or two miRNAs).

The variance explained by the models increases substantially when the effect of specific miRNAs is incorporated. For example, for polysomal secreted proteins, the base model explains 15.5% of the variance, while 32.1% is explained by the two-miRNA model. The effect of miR-130b and miR-558 on the $\log FC_{mRNA}$ more accurately reflects the observed protein logFC. These miRNAs may have an important regulatory role in adipogenesis. Similar results were obtained with the remaining datasets. In addition, we also found that variances explained by polysomal fraction models (secretome and nuclear) were in general higher than those using total RNA (Table 1). This can be explained by the reduced effect on mRNA destabilization in polysomal mRNAs (they are already associated with polysomes). Finally, Table 2 shows all miRNAs that were significant at an FDR < 0.05 in single miRNA models at day 5, in the different datasets. Several of these miRNAs (underlined in the table) were previously found to be involved in adipogenesis [31]. To assess the possibility that our results were due to random sampling on the miRNA matrix, we performed a

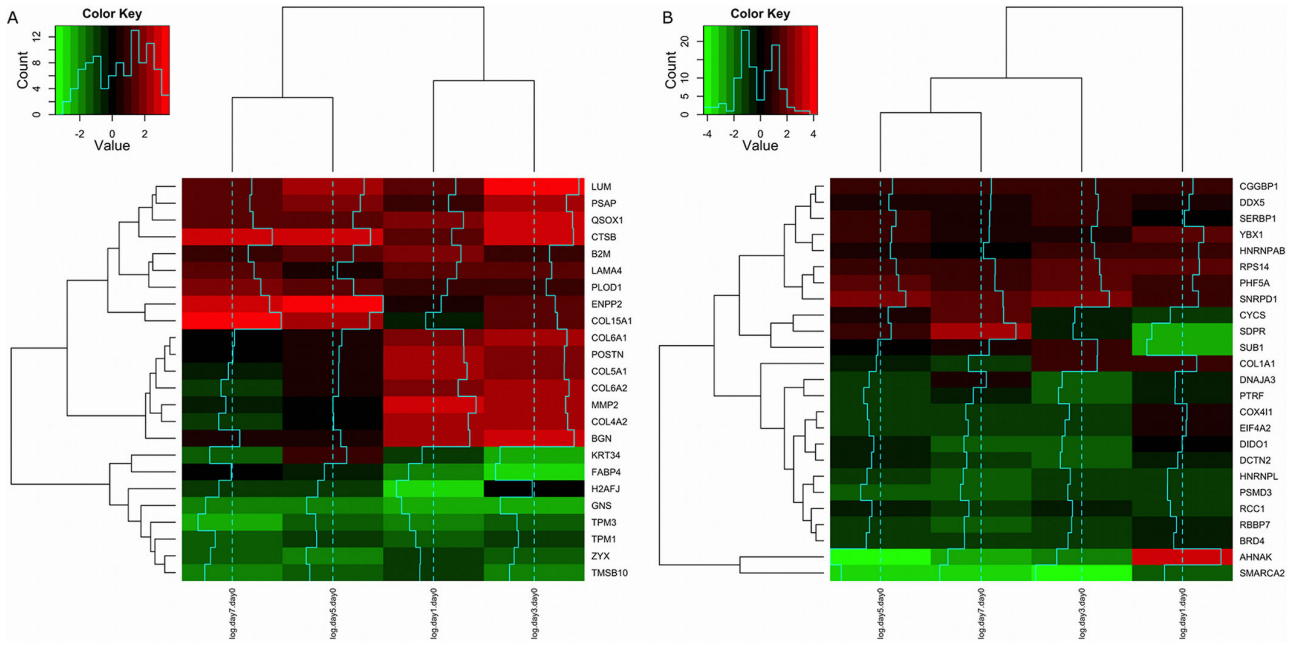


Figure 1. Heatmap of the residuals of the model $\log FC_{protein} \sim \log FC_{mRNA}$. Protein levels (logFC) of the set of secreted proteins are compared against the logFC of our data set and the residuals of the linear model analyzed; polysomal fraction (A) and total fraction (B). All time points are considered: day 1, 3, 5 and 7 (dendrogram on the top). Genes are on the rows (dendrogram on the left). Only data for genes with large absolute residuals are shown.
doi:10.1371/journal.pone.0075578.g001

Table 1. Linear model results for secreted and nuclear proteins at day 5.

SECRETOME				NUCLEAR			
Polysomal RNA				**Polysomal RNA**			
logFC	$microRNA_1$	$microRNA_2$	$adjustedR^2$	**logFC**	$microRNA_1$	$microRNA_2$	$adjustedR^2$
0.377†	-	-	0.155	0.0978	-	-	0.004
0.332†	miR-130a†	-	0.279	0.0952	miR-185*†	-	0.202
0.337†	miR-130b†	-	0.299	0.0831	miR-20b*†	-	0.175
0.322†	miR-130b†	miR-558†	0.321	0.105	miR-16-2*†	miR-185*†	0.300
Total RNA				**Total RNA**			
logFC	$microRNA_1$	$microRNA_2$	$adjustedR^2$	**logFC**	$microRNA_1$	$microRNA_2$	$adjustedR^2$
0.209†	-	-	0.100	0.0612	-	-	0.002
0.203†	miR-130a†	-	0.232	0.0410	miR-372†	-	0.191
0.205†	miR-19a†	-	0.228	0.0463	miR-106b†	-	0.181
0.205†	miR-150*†	-	0.239	0.0339	miR-523†	-	0.206

Results for applying linear models to the data at day 5 secreted and nuclear proteins. Both RNA fractions are considered. For each subtable (e.g. secretome-polysomal) the first row shows the results for a linear model without considering microRNA effect (the standard model: $logFC_{protein}$ vs. $logFC_{mRNA}$). The 2nd and 3rd row represent the values for univariate models, including the effect of only one miRNA. We selected the two most significant miRNAs. The last row shows the (multivariate) best model as determined by the BIC value. In several cases the best model is not multivariate, especially since BIC penalizes the number of parameters.
†means a significance level of $<1*10^{-3}$.
doi:10.1371/journal.pone.0075578.t001

bootstrap analysis as described in section 6. Our results ruled out this possibility, since for all significant miRNAs, much less than 5% of random models had explained variances comparable to BIC-selected "true" models. Fig. 2 shows an analysis indicating how many times each miRNA wins, comparing explained variances using "true" over random models (R^2 values are color coded). The miRNAs that win at least 95% of the times generate the best fitting models (more variance explained), and are shown in red. These miRNAs (red) could be distinguished from those winning in random models (>5%). Explained variances for both miRNA groups were compared and the differences were found to be statistically significant ($p<1*10^{-20}$, Kruskal-Wallis test); miRNAs winning in true models (>95% of the times) usually explain much more variance than miRNAs winning in random models (see Fig. S2).

4 Consequences of including miRNAs and alternative transcripts

While the effect of $logFC_{mRNA}$ is significant for the secretome set (both fractions), it is not for the nuclear set (both RNA fractions), as shown in Table 1. Significant $logFC_{mRNA}$ coefficients are higher for polysomal than for total RNA, which is expected since polysomal RNA reflects protein levels more accurately. Fig. 3 summarizes results for the best BIC models for the log-fold change in secreted proteins on day 5 with respect to day 0, for polysomal and total RNA. Fig. 3, (A) and (C) show the distribution of genes when comparing $logFC_{mRNA}$ with $logFC_{protein}$ not including the miRNA effect (base model). Fig. 3, (B) and (D) show the model including the effect of miR130b and miR-558 (polysomal) and miR-150* (total). While the base model performs poorly in predicting behavior of several genes (colored dots), in that they deviate from the predicted model line, our model shifts them towards a more expected position. In addition, among the shifted genes several established adipogenesis genes were found: FABP4, FABP5, LPL and ADIPOQ.

The coefficient for $logFC_{mRNA}$ is low in the base model for both RNA fractions, ca. 0.377 (polysomal) and 0.209 (total). This coefficient decreases even more in our models. This indicates a range compression comparing protein fold-change with mRNA fold-changes (in log-log scale). This might be unexpected, however, translational efficiency (the number of protein molecules produced per mRNA molecule) may decay with the number of transcripts (see Appendix S1 (B) for more details). In fact, several studies have shown a decrease in translational efficiency [27,28,32], observed as a linear trend in the dot plot of absolute protein quantity vs mRNA quantity. Furthermore, as we show in Appendix S1 (A), the slope of this relation (1 indicating no decreasing translational efficiency with mRNA quantity, and 0 a complete decrease) is identical to the coefficient of $logFC_{mRNA}$ in the linear models we have fit here.

Regulatory features we found help to explain protein level changes seen during adipogenesis, even though we used a limited data set. For this reason, in addition to analyzing significant miRNAs acting as predictor variables in protein-mRNA logFC relationships, we also analyzed the distribution of all miRNAs in all genes (with RNAseq data) having alternative transcripts.

5 Multiple miRNA functioning together in regulation

Evidence shows that multiple miRNAs may act together to co-regulate specific genes for normal function [33–36]. We investigated co-occurrence of miRNAs in our data set, and found established as well as novel regulatory correlations between them.

Table 2. Significant miRNAs at day 5 as obtained from the linear univariate model.

	Polysomal RNA	Total RNA
secreted	miR-103,miR-107,miR-130a,miR-130b	miR-103,miR-107,miR-130a
	miR-142-3p,miR-144,miR-148a	miR-130b,miR-142-3p,miR-144
	miR-148b,miR-150*,miR-152,miR-15a	miR-150*,miR-152,miR-15a
	miR-15b,miR-16,miR-190b,miR-195	miR-190b,miR-19a,miR-19b
	miR-19a,miR-220c,miR-28-3p,miR-29a	miR-210,miR-220c,miR-26a
	miR-29b,miR-29b-2*,miR-29c,miR-301a	miR-26b,miR-27a*,miR-28-3p
	miR-301b,miR-302a,miR-302d,miR-338-5p	miR-29a,miR-29b,miR-29b-2*
	miR-33a,miR-33a*,miR-33b,miR-340	miR-29c,miR-301a,miR-301b
	miR-486-5p,miR-509-5p,miR-510,miR-551b*	miR-338-5p,miR-33a,miR-33a*
	miR-553,miR-558,miR-569,miR-574-5p	miR-340,miR-361-5p
	miR-589*,miR-628-5p,miR-633,miR-672	miR-486-5p,miR-509-5p
	miR-768-3p,miR-768-5p,miR-891b	miR-510,miR-551b*,miR-553
		miR-558,miR-569,miR-574-5p
		miR-575,miR-582-3p,miR-587
		miR-589*,miR-604,miR-607
		miR-628-5p,miR-672
		miR-768-3p,miR-768-5p,miR-891b
nuclear	miR-143*,miR-16-2*,miR-185*,miR-20b*	miR-100,miR-106b,miR-10b*,miR-185*
	miR-346,miR-372,miR-378*,miR-587	miR-193a-5p,miR-222*,miR-28-5p
		miR-372,miR-433,miR-507
		miR-523,miR-548b-3p,miR-551b
		miR-576-5p,miR-621,miR-885-5p

Set of significant miRNAs in each data set. Underlined miRNAs correspond to those found in Zhang *et al.* (revision on miRNAs involved in adipogenesis) [31].
doi:10.1371/journal.pone.0075578.t002

In addition to the co-occurrence in the linear models described before, we now explored the correlation of miRNA occurrences in the different transcripts analyzing the presence/absence matrix of miRNAs by transcript, weighted by transcript usage differences between IN and CT inside genes (see subsection 7). Based on the total RNA fraction (reflects status of all transcripts, e.g. before degradation) we observed some miRNA pairs with significant correlations. We describe four of the cases found in our study. In all cases, we restricted our analysis to transcripts (rows in the matrix) in which at least one miRNA (of the two per comparison) is present and we compared the correlations obtained with the presence/absence matrix (1 and 0) with the values obtained with the matrix weighted by transcripts usage. First, the presence/absence matrix for miR-204 and miR-211 target sites was considered and a correlation was determined $r=0.044$ ($p>0.15$). When using the weighted matrix, we obtained a correlation value of $r=0.76$ ($p<1*10^{-15}$). Similarly, for transcripts targeted by miR-17 and miR-93, the correlation using the presence/absence matrix was $r=0.24$ ($p<1*10^{-15}$), whereas the correlation with the weighted matrix was$r=0.91$ ($p<1*10^{-15}$). For transcripts targeted by miR-17 and miR-20a a negative correlation is observed using the presence/absence matrix ($r=-0.79$, p-value$<1*10^{-15}$), however considering weighted data a significant positive correlation is observed ($r=0.57$, $p<1*10^{-15}$). Pair miR-34 and miR-449 presents a negative correlation in both cases ($r=-0.23$, $p<1*10^{-15}$ and $r=-0.79$ presence/absence matrix, $p<1*10^{-15}$ for our weighted data).

6 Alternative transcripts in relevant genes from other sources: PluriNet genes

The PluriNet is a protein-protein network with 299 members common to pluripotent stem cells based on gene expression profiles of 150 human cell samples. Such molecular network is believed to be involved in the differentiation and self-renewal of pluripotent stem cells [24].

We investigated 3′UTR length distribution of PluriNet transcripts for IN vs CT conditions. Similar trends were observed for the total and polysomal fraction. We found that positive differences correspond to longer 3′UTR under IN conditions, and negatives the converse situation (zero indicates no differences), when considering the weighted differences in length (as determined in 4). We first ranked all genes by 3′UTR length differences, and identified PluriNet genes within the ranking. As shown in Fig. 4, PluriNet genes accumulated near small negative differences but distributed evenly for all positive values. Of the 299 Plurinet genes, 216 were found in our dataset. 123 had positive differences in length (3′UTR longer in IN) and 88 negative (3′UTR longer in CT) with 5 having no differences. GO analysis of the 88 negative genes resulted in the following over-represented terms: metabolism of non-coding RNA ($p<1.1*10^{-2}$), snRNP assembly ($p<1.1*10^{-2}$), loading and methylation of Sm proteins onto SMN complexes ($p<1.1*10^{-2}$), RC complex during G2/M-phase of cell cycle ($p<1.55*10^{-2}$). In the set of positive correlated genes, one enriched term was found: nuclear part ($p<2.08*10^{-3}$).

Interestingly, according to the Cochran-Mantel-Haenszel statistic (with FDR<0.01) the following PluriNet genes showed

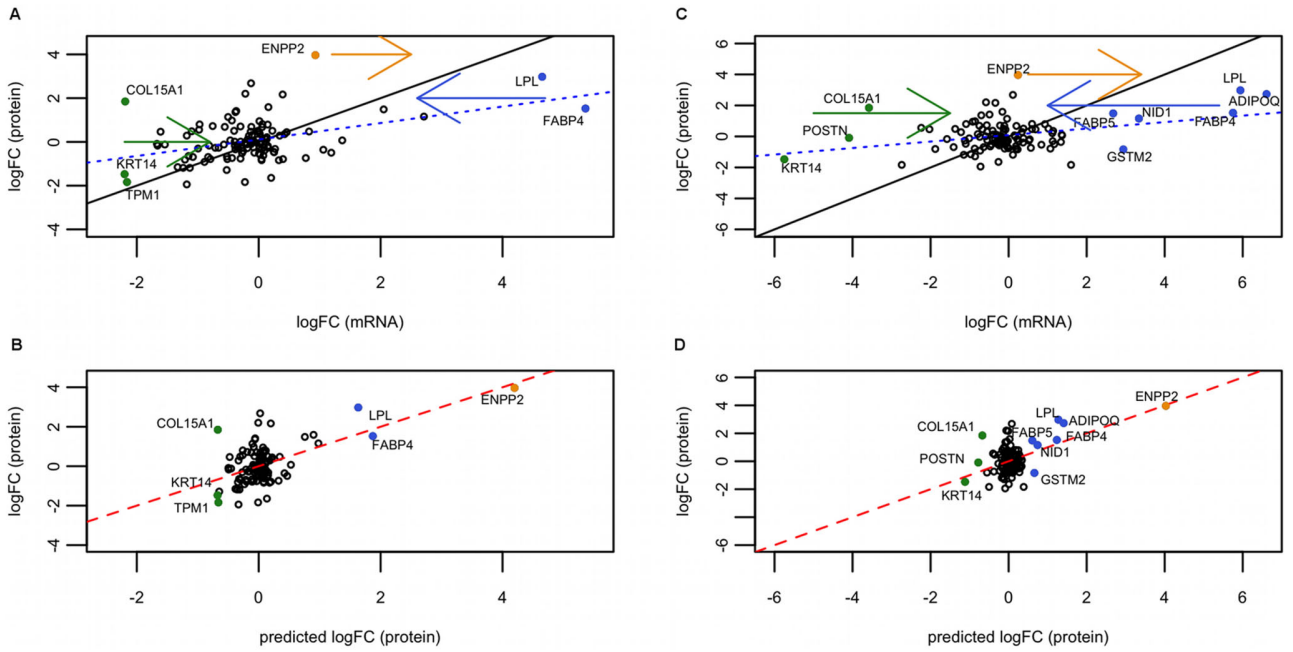


Figure 2. Bootstrap to asses our results for each RNA fraction and each protein set. Bootstrap results for total RNA fractions are shown in A (nuclear) and B (secretome). Polysomal fraction is shown in C (nuclear) and D (secretome). For each such pair of conditions, we performed a bootstrap analysis as explained in 0.6. For each miRNA we permute the values of the genes and calculate the explained variance from the resulting linear model. This procedure is repeated 1000 times. The y-axis represents how many times the "true" miRNA wins over the random model. The x-axis represents all miRNAs. The colors, from red to green, represent the explained variance from the current "true" model. It can be observed that the miRNAs win almost all times (the larger bars, almost reaching 1), explain the larger variance, and hence produce the best models (red).
doi:10.1371/journal.pone.0075578.g002

significant 3′UTR length differences between IN vs CT: PSMA3, PSMA4, PSME3, proteasome assembly (subunits and activator), HSPA8 (heat shock 70 kDa protein 8), SNRPF (small nuclear ribonucleoprotein polypeptide F), SUMO1 (small ubiquitin-like modifier which promotes SUMOylation), TMEM258 (transmembrane protein 258) and SNRPE (small nuclear ribonucleoprotein polypeptide E). Only SNRPE had a positive correlation, while the others had a negative correlation.

Discussion

We previously showed important differences in mRNAs changes comparing polysomal and total fractions during adipogenesis [12]. Furthermore, mRNA changes were poorly correlated with observed protein changes during differentiation [23]. Altogether, these results point to a very important role for post-transcriptional regulation in adipogenesis. To gain deeper insight into the mechanisms involved, we explored the differences observed in alternative transcript usage focusing on differences in the 3′UTR regions. These are relevant since they have well-known regulatory features, particularly involving small non-coding RNAs. An example showing how different miRNA binding sites can be generated in the 3′UTR of alternative transcripts is shown in Fig. 5. The gene illustrated is RER1, which is one of the significant genes in the polysome fraction in this study having alternative transcripts during adipogenesis. As indicated longer 3′UTRs may have additional miRNA binding sites.

Our results show that significant differences in transcript isoforms arise by APA during adipogenesis. A trend towards longer 3′UTR was observed in both RNA fractions, total (18/11 bases) and polysomal (20/12 bases). We proposed that this small differences in length were still sufficient for the generation of new miRNA binding sites. We tested this, by analyzing the pairwise differences between the 3′UTR length of transcripts and the corresponding differences in miRNA binding sites, for each gene. Our preliminary analysis showed that for the differences of interest (20, 18, 12 and 11 bases), out of the 16937 genes analyzed, 1235, 1204, 1132 and 1112 genes, respectively, differed in at least one miRNAs binding site.

The difference in the total RNA fraction is also consistent with the number of genes displaying a positive trend (3′UTR length IN>CT), which is significantly higher than those showing a negative trend. Regarding trend-length differences comparing IN and CT conditions, 182 genes showed statistically significant trends (FDR <0.01): 114 had a positive correlation value and 68 a negative value. Very similar trends were also observed for correlation values in the polysomal fraction. Two adipogenesis relevant genes, FABP4 and WNT2, appeared to exhibit APA and differential 3′UTR length during differentiation in our previous study [12] by visually inspection. Here we confirmed these results by analytical methods. In our earlier work, the FABP4 gene exhibited a much longer 3′UTR under IN compared with CT conditions. The WNT2 gene in contrast showed the opposite behavior having a longer 3′UTR under CT conditions. Results obtained in this study showed a (positive) difference of 103 bases and a significant correlation value of ~ 0.10 for the FABP4 gene, and for WNT2, a (negative) difference of -407 bases and a significance correlation value of ~ -0.48.

A protein-protein network was previously described for pluripotent stem cells (Plurinet) [24]. Construction of the network was based on gene expression profiles for 299 human proteins. We analyzed the distribution of differences in 3′UTR length for Plurinet genes having expression values in our dataset (216 in 299). As shown in Fig. 4, the distribution of length differences

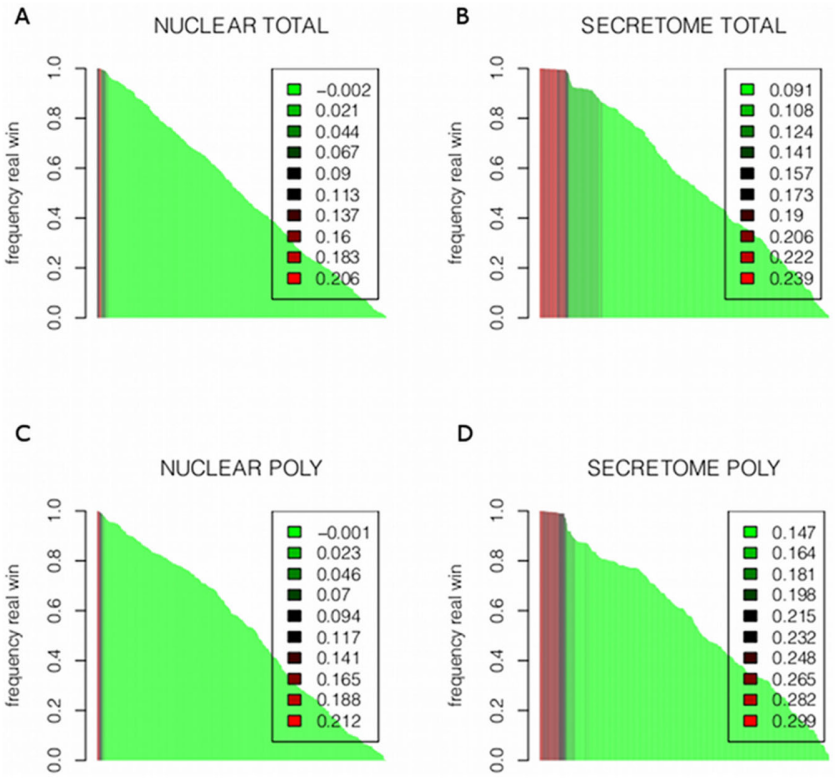


Figure 3. Linear models for day 5 secreted proteins represented graphically. (A, B) Polysomal fraction, (C, D) total RNA. (A) and (C): plot representing $\log FC_{mRNA}$ against $\log FC_{protein}$. The dashed blue line is the best fitting line of the base model, $logFC_{protein}$ against $logFC_{mRNA}$. The straight black line is the identity line (so you get an idea of the real coefficient of the model). The colored full dots are genes, which are moved after applying the model with miRNAs. Hence, they represent genes that are better explained by our model. The arrows indicate the direction of the movement. (B) and (D): plot representing our linear model including miRNA effect. In this case, the best (multivariate) model is shown: miR-130b and miR-558 (polysomal) and miR-150* (total). Full dots are the genes that were corrected by our model, being now closer to the protein prediction line of the model (red full line). Black identity line concurs with the red line. Note that the abscissas of (A) and (C) seem to have a compression of range with respect to the plots below, (B) and (D). This is not a compression, since they are different x-axis: (A) and (C) hold $\log FC_{mRNA}$ values, while (B) and (D) $\log FC_{protein}$.
doi:10.1371/journal.pone.0075578.g003

substantially deviates from the behavior of all genes. In particular, genes with much longer 3′UTRs in control cells compared with induced cells were under represented. Additionally, we found an enrichment of the term "metabolism of non-coding RNA" among genes with 3′UTR length CT>IN, which could be associated with post-transcriptional regulation.

The dataset of Molina et al., was analyzed to understand the potential role for APA in protein changes [23]. Even though the cell line used by these authors was murine, this dataset was the most suitable available to compare with our RNAseq experiment. Several studies indicated a reasonable conservation in regulatory networks between human and mouse [29,30]. Comparing differences between $\log FC_{protein}$ and the predicted protein quantity according to the $\log FC_{mRNA}$ ($\log FC_{predicted}$), some large residuals (gene differences) were observed using this dataset (Fig. 1). Adipogenic relevant genes FABP4, GNS, TPM1, TPM3, KRT34, TMSB10 and ZYX were among genes with larger negative differences, i.e, $\log FC_{protein} < logFC_{predicted}$. On the other hand, residuals with positive differences ($\log FC_{protein} > logFC_{predicted}$), include LUM, PSAP, QSOX1, COL15A1, POSTN, ENPP2 and LPL (total RNA fraction). In addition, we have found that the observed differences (residuals) do not correlate significantly with the absolute magnitude of change in mRNA. As such the differences can't be explained by the expected compression of range (see section Appendix S1 (A)).

Clear differences were observed in APA isoform usage comparing IN and CT conditions, as well as differences between predicted fold change (by mRNA) and observed protein fold change for some genes. To further investigate this discrepancy we compared explained variances of base models just including $\log FC_{mRNA}$ as predictive variable, against different models that incorporate miRNAs target site differences between transcripts as co-variables. The rationale behind including these miRNAs is to account for their potential effect on destabilizing or inhibiting translation resulting in discordance between the observed proteins and the mRNA levels. We have shown that hMSCs use their transcripts differentially during adipogenesis. We were able to test whether presence of miRNA binding sites is associated with change in the fate of specific transcripts by incorporating preferences for alternative transcripts (with alternative 3′UTR length) in our analyses. As summarized in Table 1, differences in explained variance were striking (even after adjusting for model complexity) when the effects of different miRNAs were introduced in the models. As expected, polysomal $\log FC_{mRNA}$ was higher correlated with $\log FC_{protein}$ than the corresponding correlation in total RNA. This can be seen in the explained variances of both datasets, i.e., secreted and nuclear proteins. More surprising,

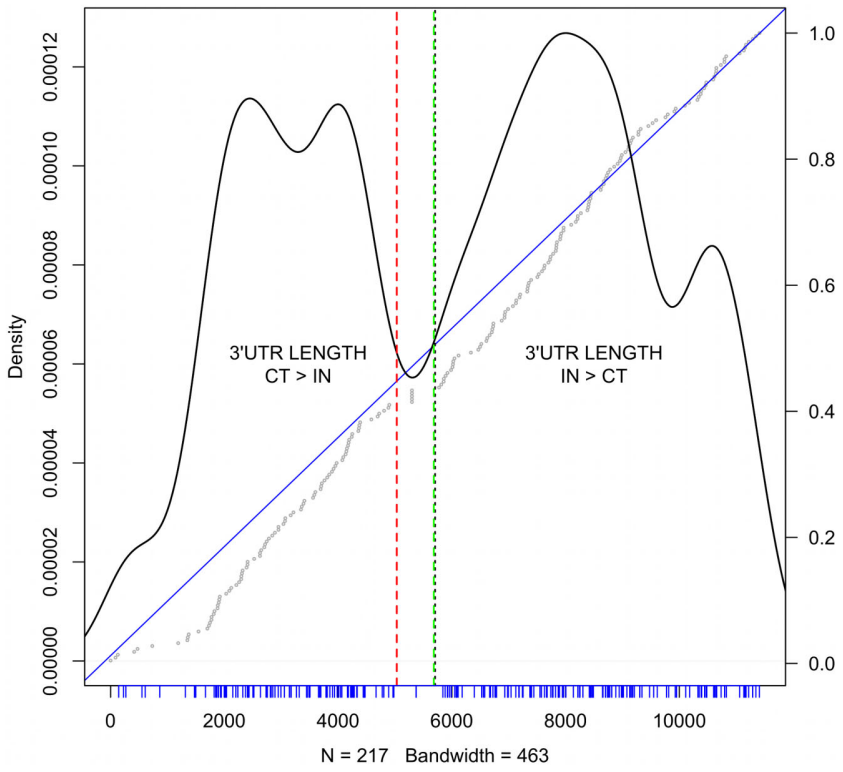


Figure 4. 3′UTR differences for PluriNet genes. On the x-axis one observes the ranking of 3′UTR lengths as determined in section 1 of all genes used for logFC calculations in the total RNA fraction. The ranking of genes belonging to the PluriNet are shown as densities (y-axis on the left). Negative lengths (CT>IN) lie to the left of the red dashed line. Positive values are to the right of the green dashed line. The wide space between those lines correspond to genes with no differences in 3′UTR length. The median of the rankings is represented as a doted black line. Tick marks in blue represent the ranking positions of the PluriNet genes. On top of the density plot the cumulative distribution of rankings is shown. The straight blue line has slope 1 and intersect 0. Gray dots represent the cumulative ranking of the PluriNet genes. The y-axis to the right indicates the meassure of this cumulative ranking. An under-representation of PluriNet genes with high negative values and a slight over-representation of positive values is observed. Moreover, only marginal PluriNet genes are presenting values of 0.
doi:10.1371/journal.pone.0075578.g004

however, is that changes in nuclear proteins were very poorly correlated with changes in mRNAs (the coefficient for $\log FC_{mRNA}$ was never significant, even in absence of other co-variables). While several reasons might account for this, mechanisms involving protein translocation could be collaborating to this lack of correlation.

A range compression of $\log FC_{protein}$ compared with $\log FC_{mRNA}$ can be seen in the slope of Fig. 3 (A and C) and in coefficients for $\log FC_{mRNA}$ in Table 1. If translational efficiency decreases with increased mRNA levels (competition for scarce resources, e.g., ribosomes) in such a way that a linear trend is observed in log-log scale when plotting amounts of protein vs mRNA, the observed range compression would be expected (see section Appendix S1 (B)). In fact, this trend was observed in several studies [27,27,32] and a coefficient of ~ 0.50 for *Saccharomyces cerevisiae* was determined [32]. We calculated a coefficient of ~ 0.35 for comparisons with the secretome dataset, a reasonable estimate. We may be underestimating this coefficient since our comparisons and analyses are between species (mouse and human). Moreover, as we are only considering up to 214 genes, our coefficient may not correspond to a global scenario in the cell. Finally, even though a significant improvement in explained variances is found by incorporating miRNAs in models, the small changes in $\log FC_{mRNA}$ coefficients indicate that the improvement in performance is basically obtained by adjusting the prediction of ''poorly-behaved'' genes. In addition, the linear models presented here also reveal several genes whose regulation might be explained by specific miRNAs included in the models. In particular, we observed that the following genes were better fit by miRNA-models than the base model: ENPP2, LPL, FABP4, KRT14, TPM1, COL15A1 (polysomal RNA) and ENPP2, LPL, ADIPOQ, FABP5, FABP4, NID1, GSTM2, COL15A1, POSTN, KRT14 (total RNA). In the case of polysomal RNA, miR-130b and miR-558 were the miRNAs included in the model, whereas miR-150* was the co-variable in the model considering total RNA. It is worth mentioning, that we are only considering presence of miRNA binding sites, the expression levels of the miRNAs themselves is not included in our work.

Table 2 lists all significant miRNAs for which one-miRNA models were constructed, and also indicates which are previously mentioned as relevant for adipogenesis according to the revision of Zhang et al. [31]. In particular, we found 8 significant miRNAs of the 23 previously identified. Additionally, we found several miRNAs involved in other differentiation processes not described by Zhang et al. These include miR-142-3p, miR-16 and miR-15a which are associated with (TPA)-induced differentiation of human leukemia cells (HL-60) to monocyte/macrophage-like cells [37]. Also, miR-144 was implicated in erythroid differentiation [38] and miR-148a, miR-26, miR-378, miR-486 and miR-29 were identified in skeletal myogenic differentiation [39], and miR-10

	RER1-			
miRNA	004	001	201	002
218			1	1
96*	1			1
197	1			
92b	1			
212	1			
220	1	1		
323-5p	1			
126	1		1	1
887	1			1
502-5p		1	1	1
616*		1	1	1
30d*		1		
30a*		1	1	1
623		1	1	
22*		1		
220b		1		
188-5p		1		
623		1		1
424*		1		
582-5p		1		
651		1	1	1
548b-3p		1		
548d-5p		1		
548b-5p		1		
548c-5p		1		
616*		1		
938		1		1
424*				1

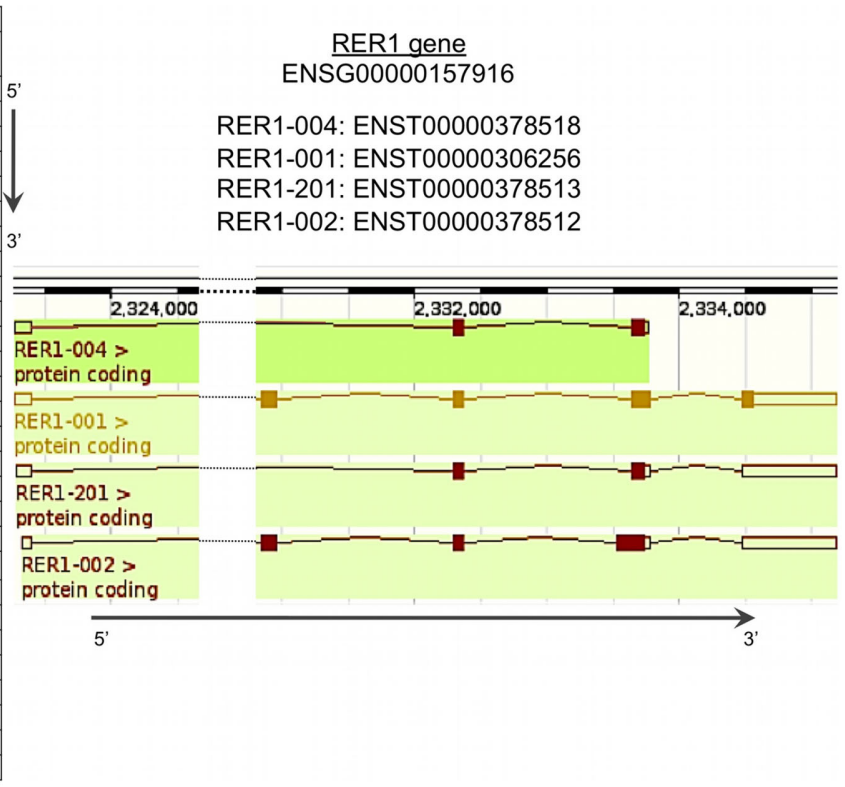


Figure 5. An example of how different microRNAs binding sites arise from alternative transcripts. The table shows the presence of the miRNAs in the transcripts. The longer the 3′UTR the more binding sites are seen.
doi:10.1371/journal.pone.0075578.g005

was involved in endodermal differentiation [40]. Hence, miRNAs identified using our *in silico* analysis were previously found to be involved in several differentiation processes (including adipogenesis) by experimental methods.

Co-occurrence of miRNAs is not unusual; several miRNAs have been found to work together in gene regulation. Based on differences observed in alternative transcript usage, we explored miRNA co-occurrence in adipogenesis. We have found several strong associations in our presence/absence matrix weighted by differences in transcripts usage. Here we discuss some examples. Our primary analysis shows a statistically significant, but relatively trivial (since they are homologous) co-ocurrence of miR-204 and miR-211, whose common target is the Runx2 gene. miR-204/211 inhibits expression of Runx2, which inhibits osteogenesis and promotes adipogenesis of mesenchymal progenitor cells and bone marrow stromal cells [33]. We also observed a highly significant association of miRNA pair miR-17 and miR-93. They belong to the family including miR-17-5p, miR-20a, miR-93, and miR-106a, are differentially expressed in developing mouse embryos and have a controlling function in stem cell differentiation [41]. They are also key regulators of induced pluripotent stem cells and play a role in reprogramming efficiency of such cells [34]. On the other hand, miR-34 and miR-449 are negatively correlated in our data set implying that the presence of one results in the absence of the other. Both miRNAs belong to the same family; miR-449a, b and c are strong inducers of cell death, cell cycle arrest and cell differentiation; miR-34 is activated with expression of p53 protein and miR-449 is induced by E2F1, a cell cycle regulatory transcription factor. They are responsible for an asymmetric feedback loop that keeps the balance between E2F and p53 functions. miR-449 helps to ensure normal cell function but is also involved in maintaining a close interaction between cell differentiation and tumor suppression [35].

In summary, in the present work we found interesting and consistent differences in transcript isoforms used during adipogenesis. We found that, in general, induced cells had longer 3′UTRs compared with undifferentiated hMSCs. Furthermore, we characterized these differences by identifying genes whose transcripts had important differences in miRNAs target sites. Additionally, we demonstrated that by incorporating the effect of several miRNAs and alternative transcript usage in linear models, we were able to substantially improve prediction of $\log FC_{protein}$ over the base model that only includes $\log FC_{mRNA}$. We need to expand our dataset by obtaining more accurate proteomic data to further corroborate our findings. Our results indicate that post-transcriptional regulation plays a key role in differentiation.

Materials and Methods

1 Ethics statement

Samples were isolated and collected after obtention of written informed consent, agreeing with guidelines for research involving human subjects, and with the approval of the Ethics Committee of Fundação Oswaldo Cruz, Brazil (approval number 419/07), as previously mentioned in [12].

2 Sample description

We used samples described by Spangenberg *et al.* [12]. Raw data is available under the accession number E-MTAB-1366 in the ArrayExpress repository. Stem cells were obtained from

Table 3. Mapping statistics of RNA-seq.

donor	condition	raw data	reads for mapping	mapped	unmapped	junctions	%
61	CT_poly	15105571	15041140	8026275	8462131	38127	53.6
61	IN_poly	18367050	18280311	10057455	10359395	32762	55.2
67	CT_poly	40032577	39862820	19398037	23642317	40995	48.8
67	IN_poly	148700586	147993700	55932659	103696209	37807	37.8
67	CT_total	8883206	8845973	4436690	5185133	39802	50.6
70	CT_poly	17473812	17403946	9415117	9862766	39336	54.3
70	IN_poly	32280923	32151368	16831536	19327229	32614	52.5
70	CT_total	121079661	120741759	58016204	71585612	39275	48.1
61	IN_total	31667090	31573343	16498438	18437894	57296	52.4
67	IN_total	27685080	27615016	13794780	16630549	53312	50.1
70	IN_total	60059063	59886179	31756854	34628161	47819	53.1
61	CT_total	22644356	22584805	11745550	12531817	54097	52.2
67	CT_total	34358013	34267292	19408351	19849852	32832	56.7

Mapping data of SOLiD runs. Following data is shown: donor number, condition considered (CT or IN, and polysomal or total RNA), number of raw reads obtained from the sequencing process, number of reads considered for mapping, number of mapped reads, unmapped reads, and the percentage of mapped reads.
doi:10.1371/journal.pone.0075578.t003

adipose tissue of three obese human donors. hASCs were isolated, cultured and characterized as previously described [42]. Briefly, adipogenesis was induced with 6 day-cycles of induction/maintenance over 21 days. Induction medium contained the adipogenic inducers insulin, dexamethasone, indomethacin and IBMX; maintenance medium contained insulin. Medium was changed every 3 days. The degree of adipogenic differentiation was determined by assessing cytoplasmic accumulation of triglycerides by staining with Oil Red O or Nile Red (Sigma-Aldrich). Samples were taken at time point 0 (control samples, CT) and then after three days (induced samples, IN).

A total of 13 samples were sequenced with SOLiD4 System (Applied Biosystems), 7 CT (2 polysomal-associated RNA and 6 total RNA samples) and 6 IN (3 polysomal-associated RNA and 3 total RNA). Table 3 shows an overview of samples. The proteomic data used in this study is from Molina *et al.* [23]. They quantified two sets of 3T3-L1 murine proteins with SILAC: 280 nuclear and 147 secreted proteins, with a total of 427 proteins. These were analyzed during adipogenesis (at day 0, 1, 3, 5 and 7).

While our RNA-seq data is from human donors, nevertheless we decided to compare it against murine proteomic data. Of course, this assumes a high conservation at protein level between this two organisms in the involved networks, a fact relatively supported by recent studies [29,30]. Furthermore, at transcriptional level, some studies have shown that a conservation is also seen for several genes [43].

3 Primary analysis of SOLiD RNA-seq samples

Table 3 summarizes results of the mapping procedure with *tophat2* and *cuffdiff*. We obtained a median of 52% mapped reads in the 13 samples. Information on transcript usage for 62134 ensembl gene ids was obtained from *cuffdiff* for total and polysomal RNA samples. These were filtered according to the quality status of transcripts, because the low number of reads might compromise determination of FPKM. After filtering we obtained 61381 for both sets, polysomal and total RNA. From those genes, 21647 have annotated 3′UTRs according to ensembl annotation, corresponding to 74803 transcripts.

4 Summarizing transcript differences

We calculated the relative frequency of each transcript for each condition (IN and CT), and weighted the transcript 3′UTR length by the differences in frequency (we did this for each gene). To assess the significance of the differences observed above, we tested our data using the Cochran-Mantel-Haenszel statistic, a test of linear trend alternative to independence [26], which is more sensitive than a standard χ^2 test if a linear trend holds. Additionally, for each gene we calculated and analyzed the Pearson-r distribution between 3′UTR length and condition (CT = 0, IN = 1) [26].

5 Mapping and annotation

13 samples were mapped onto the reference genome (hg19 GR37p2) using *tophat2* [44]. *cufflinks* [45] v2.1.1 was then used for transcript assembly. Determination of isoform abundance was done with *cuffdiff* v2.1.1. The annotation file used for counting was based on the genome version Hg19 Gr37p10 (August 2012), downloaded from the ensembl. The 3′UTR annotation file was also created from the ensembl (version Hg19Gr37p10, 15 August 2012) human gff annotation file. The miRNA target information considered is the one included in the R package microRNA, from Gentleman and Falcon [46], which is also based on ensembl. Currently, it contains a total of 694 miRNAs targeting a total of 34507 transcripts.

Mapping, gene expression assessment and differential expression determination in our earlier work was performed using the *Rsubread* and *edgeR* R packages.

6 Linear model for correlation of microRNAs with protein levels

We developed a linear model approach to show the influence of miRNAs targeting 3′UTR regions of transcripts on respective protein expression levels.

Our starting point is data generated from *cuffdiff* software. An abundance normalized measure, FPKM, is first obtained for each transcript isoform which represents the number of fragments per kilobase per million fragments falling on each feature (e.g.,

Transcripts	IN_{prop}	CT_{prop}	$Prop_{IN-CT}$	$miRNA_1$	$miRNA_2$	...	$miRNA_{694}$
$isoform_A$	0.3	0.6	−0.3	1	0	0	0
$isoform_B$	0.4	0.1	0.3	1	1	0	1
$isoform_C$	0.1	0.2	−0.1	0	1	1	0
$isoform_D$	0.2	0.1	0.1	1	0	0	0
$gene_X$	1	1	0	0.1	0.2	−0.1	0.3
$isoform_A$	0.6	0.2	0.4	0	0	1	1
$isoform_B$	0.05	0.1	−0.05	1	0	1	0
$isoform_C$	0.35	0.7	−0.35	1	0	0	1
$gene_Y$	1	1	0	−0.4	0	0.35	0.05
⋮	...	...	...	...	...	...	...

↓

Gene	$logFC_{pro}(3)$	$logFC_{pro}(5)$	$logFC_{pro}(7)$	$logFC_{mRNA}$	$miRNA_1$	$miRNA_2$	...
$gene_X$	2.1	3.2	2.1	1.9	0.1	0.2	...
$gene_Y$	−1.2	−1.4	2.9	2.4	−0.4	0	...
$gene_Z$	−0.9	2.4	3.9	5.4	0.6	−0.1	...
⋮	...	...	...	...	...	...	...

Figure 6. Representative table for constructing the model. For each gene we determined the proportion of FPKM in each sample and calculated the differences ($Prop_{IN} - Prop_CT$). Furthermore, we determined the miRNAs targeting transcripts (inside 3′UTRs). A total of 694 were considered. The isoform has a 1 in $miRNA_1$ if that miRNA is present in that transcript, a 0 otherwise. For each *miRNA* (eg. $miRNA_1$) corresponding to one gene (e.g. $gene_X$), the $Prop_{IN-CT}$ vector is multiplied by the presence/absence vector of $miRNA_1$ (with assigned 1 s and 0 s). The intermediate result is, thus, a vector having the respective $Prop_{IN-CT}$ value if $miRNA_i$ was present in the isoform and 0 otherwise ($\vec{v}=\{-0.3,0.3,0,0.1\}$). The resulting vector $\vec{v}$ is summed giving a total value for $miRNA_1$ for $gene_X$ ($sum(\vec{v})=0.1$). This represents the mean weighted usage of the miRNA in that specific gene. Larger positive values indicate that the miRNA is used more (appears more often) in IN than in CT. Larger negative values represent a higher usage in CT (values around 0 indicate same usage in both). The same procedure is done for each miRNA (so a vector of 694 values is assigned to $gene_X$) and for each gene. The gene wise table below in addition to showing the resulting values calculated above, also shows the other data needed for the model; the $logFC_{protein}$ values (at day 3, 5 and 7, from Molina *et al.*) and the respective $logFC_{mRNA}$ values (our data).
doi:10.1371/journal.pone.0075578.g006

transcript). A FPKM value is calculated for each condition and each transcript, which allows determination of differential isoform usage. The proportion of each transcript isoform for each gene was determined under all conditions based on the FPKM values. Proportions in control samples are subtracted from the proportions in induced samples (IN) to determine the differences in isoform usage. Differences in proportions of each isoform for each gene ($Prop_{IN-CT}$) and the presence of miRNA binding sites in transcript 3′UTRs (represented as 1 s in Fig. 6) were determined. The $Prop_{IN-CT}$ value is multiplied by the corresponding miRNA binding site present and the resulting vector is summed for a given gene (Fig. 6). This results in one value for each miRNA binding site for each gene, which represents a weighted mean for usage of that miRNA for that gene. Large positive values (closer to 1) are miRNAs highly used in IN samples, large negative values (closer to −1) are those most used in CT. In other words, values closer to 1 correspond to miRNAs targeting transcripts preferentially used in IN samples, and those with values closer to −1 are preferentially used in CT. Note that a given miRNA might have several binding sites in a given 3′UTR, nevertheless we considered one or more sites as either present or absent with no multiplicity value assigned. This is still a matter open for discussion, since several studies have shown cooperative effects in the past [47–50], while others suggested the opposite behavior in large and comprehensive human and mouse datasets [18,51]. We have also run our analysis considering the cooperative effect, obtaining conceptually similar results (data not shown). However, for simplicity reasons, we decided to consider the simplest model accepted and used the present/absent values. Since such values are determined for each gene and for each miRNA, results can be presented in a table with #of genes × #of microRNAs. For each day d (1, 3, 5 and 7), miRNA i and assuming $e \sim N(0,1)$, we applied following model:

$$logFC_{prot_d} = logFC_{mRNA} + microRNA_i + e_{d,i},$$

so we can determine the effect of each microRNA on protein level.

The possibility that significant miRNAs coefficients arise by chance was assessed by bootstrap analysis. We randomly assigned the existing values to genes for each miRNA, and calculated the explained variance from the linear model. We repeated this procedure 1000 times. The proportion of times the variance explained by the random model was larger than the "true" model was determined for each miRNA for the four datasets (nuclear, secreted vs total, polysomal). We arbitrarily set a threshold of 5% (times the random wins over the "true") for each dataset and compared the explained variances of the two groups (random vs. "true") using the Kruskal-Wallis test.

7 Determining significative correlation for co-occurring microRNAs

Co-occurrence of miRNAs was investigated to demonstrate regulatory effects. We analyzed the complete presence/absence table of miRNAs in human (downloaded from the *microRNA* R package). This table contains all transcripts analyzed (34507) in which 1 is assigned if $microRNA_i$ is present in that transcript, and a 0 if not, for all miRNAs considered (694). We compared pairwise correlations for all miRNAs based on that information and the same in our weighted data set. This means, we also determined the correlation of miRNAs, but weighted by proportion of the transcripts used. If a transcript with a given miRNA is used

only 40% of the time by the gene, the miRNA value assigned would be 0.4, and not a simple 1.

Not all entries were used for each pairwise correlation; we eliminate all entries in which both miRNAs had values of 0, i.e., pairwise-zero entries. Several of such entries exists, since not every transcript has either one of the miRNAs considered (in most cases, they have neither). With such strategy we have compared the correlations found by the presence/absence table, and the ones obtained by our weighted filtered data.

Supporting Information

Figure S1 Heatmap of the residuals of the model $\mathbf{logFC}_{protein} \sim \mathbf{logFC}_{mRNA}$ of nuclear proteins. Protein levels (logFC) of the set of nuclear proteins are compared against the logFC of our data set and the residuals of the linear model analyzed; polysomal fraction (A) and total fraction (B). All time points are considered: day 1, 3, 5 and 7 (dendrogram on the top). Genes are on the rows (dendrogram on the left). Only data for genes with large absolute residuals are shown.
(TIFF)

Figure S2 Box plot to show the distribution of random and ''true'' models in the bootstrap. All comparisons are shown (polysomal-secreted, polysomal-nuclear, total-secreted, total-nuclear). For each such dataset, bootstrap was performed, and two groups were determined. Low-Random group holds models in which ''true'' miRNAs data won over random sampling of the miRNA values at least 95% of the time. The High-Random group corresponds to miRNAs in which random sampling of miRNA values produce models that are better than the ''true'' more than 5% of the time.
(TIFF)

Appendix S1 (A) Range compression is observed in protein log fold-change (in our data), when $\log FC_{mRNA}$ is considered as predictor. The size of this effect is the translational efficiency (in log-log scale) as a function of the quantity of mRNA. (B) Messenger exponential decay with alternative target miRNA sites. We show that the basic assumption underlying the way in which we modeled the effect of miRNAs is an exponential decay of mRNA as a function of differential target sites.
(PDF)

Acknowledgments

We are indebted to Tamara Fernandez for helpful discussions on the manuscript. We are also grateful to Paul Gill for comments on the manuscript and correcting the language.

Author Contributions

Conceived and designed the experiments: LS AC BD HN. Analyzed the data: LS HN. Contributed reagents/materials/analysis tools: LS AC BD HN. Wrote the paper: LS HN.

References

1. Pittenger MF (1999) Multilineage potential of adult human mesenchymal stem cells. Science 284: 143–147.
2. Rosenbaum AJ, Grande DA, Dines JS (2008) The use of mesenchymal stem cells in tissue engineering: A global assessment. Organogenesis 4: 23–27.
3. Tae SK, Lee SH, Park JS, Im GI (2006) Mesenchymal stem cells for tissue engineering and regenerative medicine. Journal of Cellular Physiology 1: 341–347.
4. Uccelli A, Mancardi G, Chiesa S (2008) Is there a role for mesenchymal stem cells in autoimmune diseases? Autoimmunity 41: 592–595.
5. Boyle AJ, McNiece IK, Hare JM (2010) Mesenchymal stem cell therapy for cardiac repair. Methods In Molecular Biology 660: 65–84.
6. Jain M, Pfister O, Hajjar RJ, Liao R (2005) Mesenchymal stem cells in the infarcted heart. Coronary Artery Disease 16: 93–97.
7. Baer PC, Geiger H (2012) Adipose-derived mesenchymal stromal/stem cells: tissue localization, characterization, and heterogeneity. Stem cells international 2012: 812693.
8. Kratchmarova I, Blagoev B, Haack-Sorensen M, Kassem M, Mann M (2005) Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. Science 308: 1472–1477.
9. Ivanova NB, Dimos JT, Schaniel C, Hackney JA, Moore KA, et al. (2002) A stem cell molecular signature. Science 298: 601–604.
10. Song L, Webb NE, Song Y, Tuan RS (2006) Identification and functional analysis of candidate genes regulating mesenchymal stem cell self-renewal and multipotency. Stem Cells 24: 1707–1718.
11. Jääger K, Islam S, Zajac P, Linnarsson S, Neuman T (2012) RNAseq analysis reveals different dynamics of differentiation of human dermis- and adipose-derived stromal stem cells. PLoS ONE 7: e38833.
12. Spangenberg L, Shigunov P, Abuda APR, Cofré AR, Stimamiglio MA, et al. (2013) Polysome profiling shows extensive posttranscriptional regulation during human adipocyte stem cells differentiation into adipocytes. Stem Cell Research 1: 341–347.
13. Kolle G, Shepherd JL, Gardiner B, Kassahn KS, Cloonan N, et al. (2011) Deep-transcriptome and ribonome sequencing redefines the molecular networks of pluripotency and the extracellular space in human embryonic stem cells. Genome Research 21: 2014–25.
14. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, et al. (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Research 18: 610–621.
15. Koh W, Sheng C, Tan B, Lee Q, Kuznetsov V, et al. (2010) Analysis of deep sequencing microRNA expression profile from human embryonic stem cells derived mesenchymal stem cells reveals possible role of let-7 microRNA family in downstream targeting of hepatic nuclear factor 4 alpha. BMC Genomics 11: S6.
16. Fromm-Dornieden C, Von Der Heyde S, Lytovchenko O, Salinas-Riester G, Brenig B, et al. (2012) Novel polysome messages and changes in translational activity appear after induction of adipogenesis in 3T3-L1 cells. BMC Molecular Biology 13: 9.
17. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.
18. Boutet SC, Cheung TH, Quach NL, Liu L, Prescott SL, et al. (2012) Alternative polyadenylation mediates microRNA regulation of muscle stem cell function. Cell Stem Cell 10: 327–336.
19. Liaw HH, Lin CC, Juan HF, Huang HC (2013) Differential microRNA regulation correlates with alternative polyadenylation pattern between breast cancer and normal cells. PLoS One 8: e56958.
20. Di Giammartino DC, Nishida K, Manley JL (2011) Mechanisms and consequences of alternative polyadenylation. Molecular Cell 43: 853–866.
21. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB (2008) Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. Science 320: 1643–1647.
22. Ji Z, Lee JY, Pan Z, Jiang B, Tian B (2009) Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proceedings of the National Academy of Sciences of the United States of America 106: 7028–7033.
23. Molina H, Yang Y, Ruch T, Kim JW, Mortensen P, et al. (2009) Temporal profiling of the adipocyte proteome during differentiation using a 5-plex silac based strategy. J Proteome Res 8: 48–58.
24. Müller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, et al. (2008) Regulatory networks define phenotypic classes of human stem cell lines. Nature 455: 401–405.
25. Fu Y, Sun Y, Li Y, Li J, Rao X, et al. (2011) Differential genome-wide profiling of tandem 3′UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Research 21: 741–747.
26. Agresti A (2007) An Introduction to Categorical Data Analysis. John Wiley and Sons, 400 pp.
27. Stevens SG, Brown CM (2013) *In silico* estimation of translation efficiency in human cell lines: potential evidence for widespread translational control. PLoS One 8: e57625.
28. Tuller T, Kupiec M, Ruppin E (2007) Determinants of protein abundance and translation efficiency in S. *cerevisiae*. PLoS Computational Biology 3: 10.
29. Fernandez-Tresguerres B, Caon S, Rayon T, Pernaute B, Crespo M, et al. (2010) Evolution of the mammalian embryonic pluripotency gene regulatory network. Proceedings of the National Academy of Sciences of the United States of America 107: 19955–19960.
30. Lim E, Wu D, Pal B, Bouras T, Asselin-Labat ML, et al. (2010) Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. Breast cancer research BCR 12: R21.
31. Zhang R, Wang D, Xia Z, Chen C, Cheng P, et al. (2013) The role of microRNAs in adipocyte differentiation. Frontiers of medicine 7: 223–230.
32. Futch B, Latter GI, Monardo P, McLaughlin CS, Garrels JI (1999) A sampling of the yeast proteome. Molecular and celular biology 19: 7357–7368.

33. Huang J, Zhao L, Xing L, Chen D (2010) MicroRNA-204 regulates Runx2 protein expression and mesenchymal progenitor cell differentiation. Stem cells Dayton Ohio 28: 357–64.
34. Li Z, Yang CS, Nakashima K, Rana TM (2011) Small RNA-mediated regulation of iPS cell generation. the The European Molecular Biology Organization Journal 30: 823–834.
35. Liz M, Klimke A, Dobbelstein M (2011) MicroRNA-449 in cell fate determination. Cell Cycle 10: 2874–2882.
36. Meenhuis A, Van Veelen PA, De Looper H, Van Boxtel N, Van Den Berge IJ, et al. (2011) MiR- 17/20/93/106 promote hematopoietic cell expansion by targeting sequestosome 1-regulated pathways in mice. Blood 118: 916–925.
37. Kasashima K, Nakamura Y, Kozu T (2004) Altered expression profiles of microRNAs during TPA-induced differentiation of HL-60 cells. Biochemical and Biophysical Research Communications 322: 403–410.
38. Fu YF, Du TT, Dong M, Zhu KY, Jing CB, et al. (2009) MiR-144 selectively regulates embryonic alpha-hemoglobin synthesis during primitive erythropoiesis. Blood 113: 1340–1349.
39. Zhang J, Ying ZZ, Tang ZL, Long LQ, Li K (2012) MicroRNA-148a promotes myogenic differentiation by targeting the ROCK1 gene. The Journal of Biological Chemistry.
40. Tzur G, Levy A, Meiri E, Barad O, Spector Y, et al. (2008) MicroRNA expression patterns and function in endodermal differentiation of human embryonic stem cells. PLoS ONE 3: 14.
41. Foshay KM, Gallicano GI (2009) miR-17 family miRNAs are expressed during early mammalian development and regulate stem cell differentiation. Dev Biol 326: 431–433.
42. Rebelatto CK, Aguiar AM, Moreto MP, Senegaglia AC, Hansen P, et al. (2008) Dissimilar differentiation of mesenchymal stem cells from bone marrow, umbilical cord blood, and adipose tissue. Experimental biology and medicine Maywood NJ 233: 901–913.
43. Zambelli F, Pavesi G, Gissi C, Horner DS, Pesole G (2010) Assessment of orthologous splicing isoforms in human and mouse orthologous genes. BMC Genomics 11: 534.
44. Trapnell C, Pachter L, Salzberg SL (2009) Tophat: discovering splice junctions with RNAseq. Bioinformatics 25: 1105–1111.
45. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNAseq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28: 511–515.
46. Gentleman R, Falcon S (2012) microRNA: Data and functions for dealing with microRNAs. R package version 1.16.0.
47. Doench JG, Petersen CP, Sharp PA (2003) siRNAs can function as miRNAs. Genes & Development 17: 438–442.
48. Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, et al. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Molecular Cell 27: 91–105.
49. Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J, et al. (2007) Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. RNA 13: 1894–1910.
50. Bartel DP (2009) Review microRNAs:target recognition and regulatory functions. Cell 136: 215–233.
51. Hu Z (2009) Insight into microRNA regulation by analyzing the characteristics of their targets in humans. BMC Genomics 10: 594.

Appendix

A Range compression effect

Studies in human and yeast found a linear trend for logarithm protein abundance versus logarithmic mRNA abundance [1–3]. A reasonable fit for a linear model between both (log-transformed) variables is obtained in which protein abundance can be described as:

$$\ln(protein) = b_1 \cdot \ln(mRNA) + b_0,$$

where the coefficient b_1 represents the translational efficiency as function of the quantity of mRNA, and b_0 is the intercept that can accommodate all other additive effects in logarithmic scale (multiplicative in the original scale), as presented by Stevens and Brown [3] or Tuller *et al.* [2]. Following this, the best predictor of the quantity of protein is:

$$protein = e^{b_1 \cdot \ln(mRNA) + b_0}$$

In differential expression experiments we usually compare two conditions, x and y. Hence:

$$protein_x = e^{b_1 \cdot \ln(mRNA_x) + b_0}$$

and

$$protein_y = e^{b_1 \cdot \ln(mRNA_y) + b_0}$$

Comparisons in this field are usually made through a $log-$ratio to obtain:

$$\frac{protein_x}{protein_y} = \frac{e^{b_1 \cdot \ln(mRNA_x) + b_0}}{e^{b_1 \cdot \ln(mRNA_y) + b_0}}$$

and after taking the log values on both sides:

$$\ln \frac{protein_x}{protein_x} = \ln \frac{e^{b_1 \cdot \ln(mRNA_x) + b_0}}{e^{b_1 \cdot \ln(mRNA_y) + b_0}} = \ln(e^{b_1 \cdot \ln(mRNA_x) + b_0}) - \ln(e^{b_1 \cdot \ln(mRNA_y) + b_0})$$

$$= (b_1 \cdot \ln(mRNA_x) + b_0) - (b_1 \cdot \ln(mRNA_y) + b_0) = b_1 \cdot (\ln(mRNA_x) - \ln(mRNA_y))$$

$$\Leftrightarrow \ln \frac{protein_x}{protein_y} = b_1 \cdot \ln \frac{mRNA_x}{mRNA_y}$$

This shows a range compression of size b_1 in protein log fold-change (the data we are considering), when logFC_{mRNA} is considered as predictor. Furthermore, the size of this effect (coefficient b_1) is the translational efficiency (in log-log scale) as a function of the quantity of mRNA. Additionally, as b_1 was derived from a log-log regression it is scale invariant, with effect of the scale represented by b_0 that is removed in log fold-change comparisons.

B Messenger exponential decay with alternative target miRNA sites

While model comparisons is beyond the scope of this study, we show that the basic assumption underlying the way in which we modeled the effect of miRNAs is an exponential decay of mRNA as function of differential target sites. If we assume that the linear log-log relationship between protein and mRNA holds, we can introduce the effect of a miRNA as:

$$\ln(protein) = b_0 + b_1 \cdot \ln(mRNA) + b_2 \cdot miR,$$

where miR is the proportion of reads assigned to a given mRNAs (for a given gene) that have a recognition site for this miRNA. As shown previously, now the best predictor of the quantity of protein is:

$$protein = e^{b_0 + b_1 \cdot \ln(mRNA) + b_2 \cdot miR}$$

When we compare two conditions, x and y as in the previous subsection, we have:

$$\frac{protein_x}{protein_y} = \frac{e^{b_0 + b_1 \cdot \ln(mRNA_x) + b_2 \cdot miR_x}}{e^{b_0 + b_1 \cdot \ln(mRNA_y) + b_2 \cdot miR_y}} = \frac{e^{b_1 \cdot \ln(mRNA_x)}}{e^{b_1 \cdot \ln(mRNA_y)}} \cdot \frac{e^{b_2 \cdot miR_x}}{e^{b_2 \cdot miR_y}}$$

.

Rearranging the terms, we can write this relation as:

$$\frac{protein_x}{protein_y} = \left(\frac{mRNA_x}{mRNA_y}\right)^{b_1} \cdot e^{b_2 \cdot (miR_x - miR_y)}$$

which is an exponential function of the differences in alternative target sites between the two conditions x and y. In the common form of log fold-changes, we have:

$$\ln \frac{protein_x}{protein_y} = b_1 \cdot \ln \frac{mRNA_x}{mRNA_y} + b_2 \cdot (miR_x - miR_y),$$

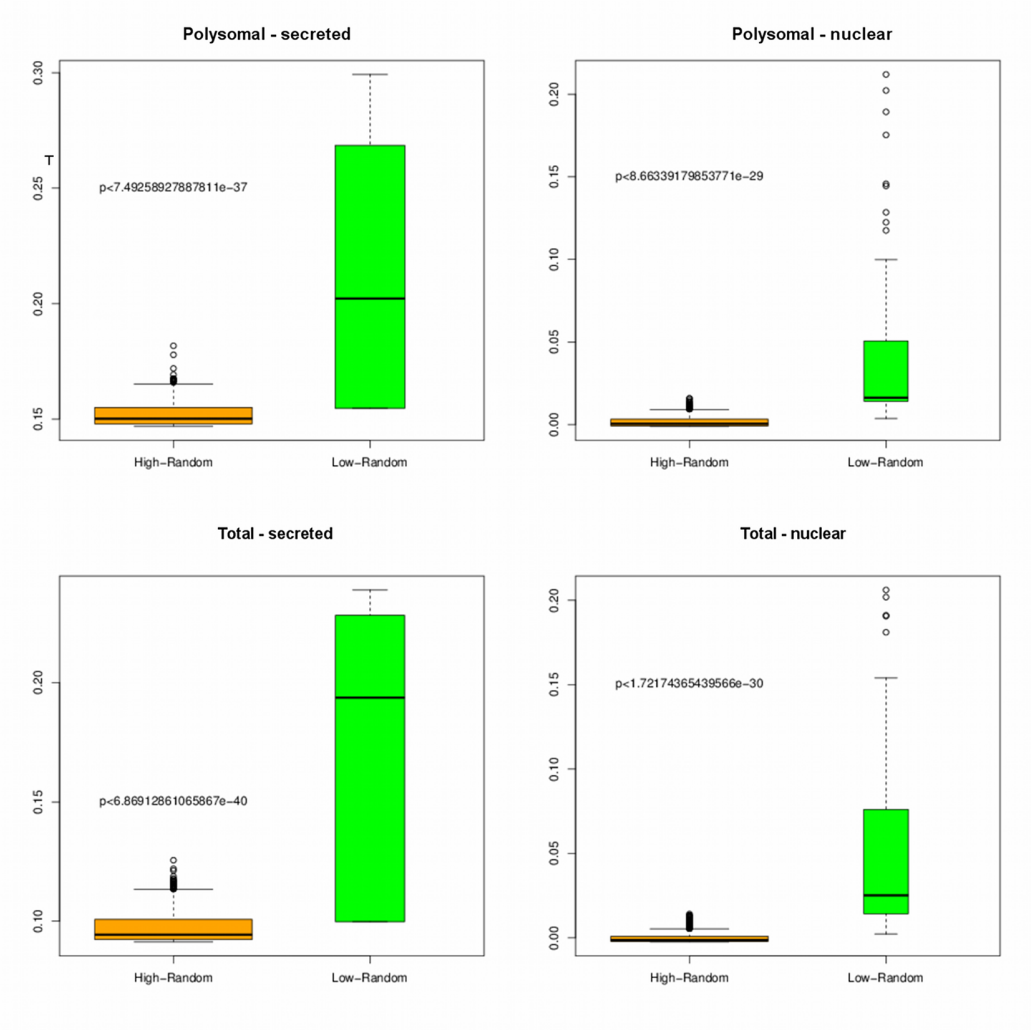
that is, the model we have used in our setting.

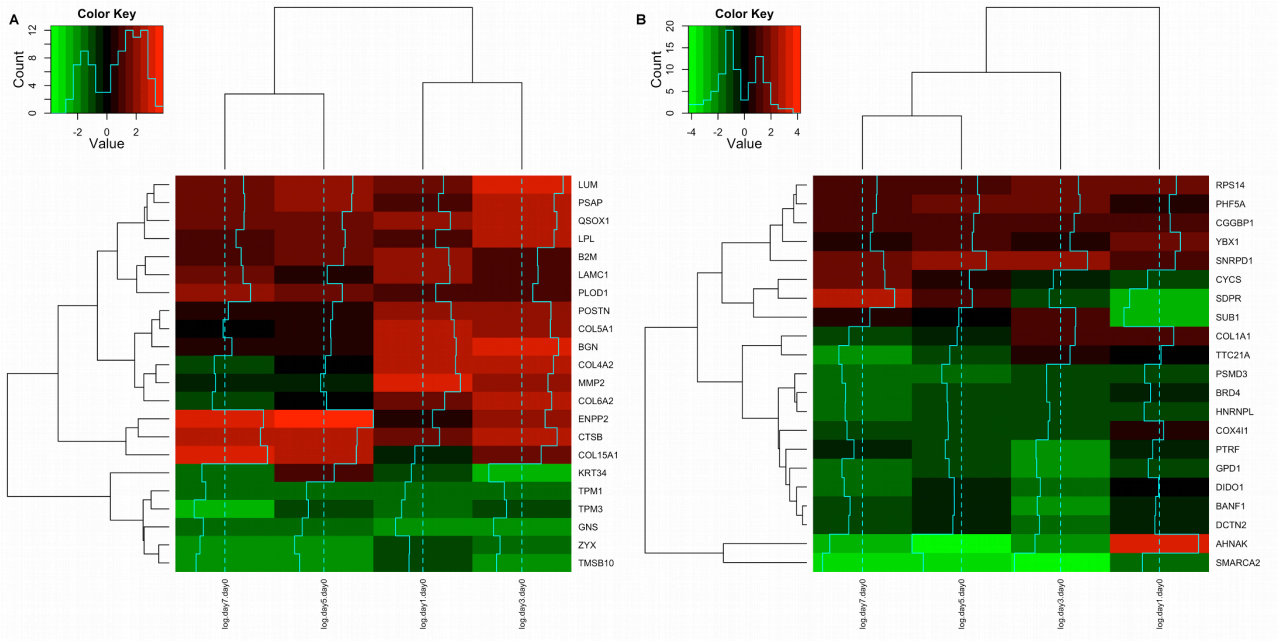
References

1. Futch B, Latter GI, Monardo P, McLaughlin CS, Garrels JI (1999) A sampling of the yeast proteome. Molecular and celular biology 19: 7357–7368.

2. Tuller T, Kupiec M, Ruppin E (2007) Determinants of protein abundance and translation efficiency in S. *cerevisiae*. PLoS Computational Biology 3: 10.

3. Stevens SG, Brown CM (2013) *In silico* estimation of translation efficiency in human cell lines: potential evidence for widespread translational control. PLoS One 8: e57625.


Polysomal - secreted
p<7.49258927887811e-37
High−Random
Low−Random
Polysomal - nuclear
p<8.66339179853771e-29
High−Random
Low−Random
Total - secreted
p<6.86912861065867e-40
High−Random
Low−Random
Total - nuclear
p<1.72174365439566e-30
High−Random
Low−Random


A
Color Key
Count
Value
LUM
PSAP
QSOX1
LPL
B2M
LAMC1
PLOD1
POSTN
COL5A1
BGN
COL4A2
MMP2
COL6A2
ENPP2
CTSB
COL15A1
KRT34
TPM1
TPM3
GNS
ZYX
TMSB10
log.day7.day0
log.day5.day0
log.day1.day0
log.day3.day0
B
Color Key
Count
Value
RPS14
PHF5A
CGGBP1
YBX1
SNRPD1
CYCS
SDPR
SUB1
COL1A1
TTC21A
PSMD3
BRD4
HNRNPL
COX4I1
PTRF
GPD1
DIDO1
BANF1
DCTN2
AHNAK
SMARCA2
log.day7.day0
log.day5.day0
log.day3.day0
log.day1.day0

Capítulo 7

Fracción de ARN polisomal contra ARN total

7.1. Introducción

En este capítulo se describe otro de los trabajos realizados durante la tesis. El mismo se basa en los datos de secuenciación descritos anteriormente 4.4. El objetivo de este estudio es comparar las fracciones de ARN total y ARN asociado a polisomas, separadamente por un lado, dentro de las muestras de células madre mesenquimal (es decir las muestras control; CT) y por el otro en las muestras inducidas. Se intentan observar las diferencias basales existentes entre ambas fracciones con el fin de poder comprender más a fondo las diferencias sistemáticas entre transcriptoma y translatoma. Nos enfocamos por un lado en el estudio de biotipos, es decir qué biotipo está más sobre/sub-representado en qué fracción de ARN. Los sesgos en las distribuciones pueden brindar información sobre los procesos subyacentes y más allá de los resultados obvios esperables descubrir excepciones a un paradigma de la biología molecular en proceso de cambio.
Por otro lado intentamos cuantificar lo que denominamos el coeficiente de expansión de la fracción polisomal. Si bien la fracción polisomal es una parte del ARN total, cuando las muestras polisomales se secuencian, se expanden en cantidad. Se pierde la relación entre total y polisomal, obteniendo por ejemplo conteos mayores de genes en polisomal que en total. Por lo tanto, se intenta estimar el factor de expansión, para luego estimar el coeficiente de asociación a polisomas de cada gen de forma precisa. Estos coeficientes brindan información sobre qué genes están siendo activamente traducidos y cuales más probablemente no lo están siendo. Analizar las funciones de los mismos ayuda a entender el proceso de adipogénesis. Además, estimar el coeficiente de expansión permite obtener una separación (estimada) de una fracción polisomal y una no-polisomal (mayormente citoplasmática no asociada), las que a diferencia de polisomal/total son ortogonales y por lo tanto directamente comparables.
El artículo que ha sido adjuntado al final del capítulo no ha sido enviado a la revista aún. Se planea enviarlo a BMC Genomics a la brevedad.

7.2. Tests de Fisher para identificar biotipos relevantes

Para evaluar biotipos relevantes, es decir que cambien de representación entre total y polisomal, se analizó la diferencia de ordenamiento (ranking) de expresión entre polisomal y total ($\text{ranking}_{polisomal} - \text{ranking}_{total}$) dentro de cada biotipo. Primeramente, se determina a partir de los conteos normalizados de cada muestra (tres donadores en polisomal contra tres donadores en total) para las muestras de célula madre (control; CT) y lo mismo para las muestras inducidas (IN), la diferencia de rankings entre polisomal y total. Seguidamente, para cada biotipo se determinan cuántos genes presentan, por un lado diferencias de ranking negativas y por el otro positivas. Los genes con diferencias positivas son aquellos que presentan alto ranking (tienden a mayor expresión) en polisomal y bajo ranking (tienden a baja expresión) en total. Los genes con diferencias negativas presentan el comportamiento opuesto: menor expresión en polisomal y una tendencia a mayor expresión en total. Por lo tanto, para cada biotipo se obtienen dos números: i) cantidad de genes con $\text{ranking}_{polisomal} - \text{ranking}_{total} > 0$ y ii) cantidad de genes con $\text{ranking}_{polisomal} - \text{ranking}_{total} < 0$. Comparando ambas proporciones se puede determinar si un biotipo tiende a estar más "expresado" en polisoma o en "total". El test de Fisher evalúa la significancia de las proporciones observadas en los dos grupos i) e ii) para cada uno de los biotipos considerados. Esto se calcula para las muestras CT y para las IN, separadamente. La tabla 1 del artículo muestra los resultados del test para cada biotipo. El sentido de la comparación es i) vs ii). Se observan que los "odds-ratios" (OR) son mayores que 1 para los genes codificantes, en ambos casos CT e IN. Esto implica que hay mayor proporción de genes con mayor expresión en polisomal que en total dentro de ese biotipo, lo cual es esperable, ya que los genes codificantes precisan de asociación al polisoma para su traducción. Con este mismo comportamiento se ven los transcriptos no codificantes "sense intronic", snoRNA, miscellaneous RNA (sin categoría definida). Los biotipos que presentan OR$<$ 1, por lo que están sobre representados en diferencias negativas (genes con mayor expresión en total), son los siguientes: miRNAs, lincRNA, transcriptos antisenso y transcriptos procesados.

7.3. Rank Product

Se realizó un análisis de Rank Product para evaluar cuáles genes presentan expresión diferencial entre las fracciones (total y polisomal). Este test estadístico no paramétrico intenta evaluar si el rango (ranking) de los genes en una condición es significativamente diferente a la otra condición. El mismo se basa en calcular el "rank product" por medio de la media geométrica para cada condición, es decir para cada gen g se calcula:

$$RP(g) = (\Pi_{i=1}^{k} r_{g,i})^{\frac{1}{k}},$$

siendo k la cantidad de réplicas (en este caso 3), $r_{g,i}$ es el rango del gen g en la réplica i. Se utilizan test de permutaciones para determinar que tan probable es que un RP se observe por azar (bootstrap), por lo tanto, se obtiene un p-value para cada RP. El mismo puede ser corregido por testeos múltiples. Seguidamente se comparan los RPs de ambas condiciones y se evalúa si la diferencia es significativa.

El Rank Product lo utilizamos también para evaluar si existe tasa de asociación diferencial a polisomas en IN con respecto a CT. En vez de evaluar expresión, se evalúa la tasa estimada de asociación (ver siguiente sección).

7.4. Estimación del coeficiente de expansión y el *h realizado*

Las comparaciones realizadas son entre diferentes fracciones de ARN, total y polisomal. En circunstancias ideales, si el ARN total corresponde realmente al total en la célula, todo el ARN polisomal debería estar contenido en el total. Sin embargo, encontramos en nuestros datos de secuenciación, conteos de genes en la fracción polisomal, que no tienen conteos, no se detectan, en total. Esto puede ser debido a varias fuentes de error durante el proceso de secuenciado y la metodología (comparar conteos entre fracciones inclusivas directamente). Además, el secuenciado de ARN total y del polisomal ocurren separadamente, por lo que la relación real entre polisomal y total no se mantiene (ni se conoce). Ya que el proceso de secuenciado explota al máximo la muestra, se obtienen aproximadamente la misma cantidad de reads en total y en polisomal. Se pierde por lo tanto en el proceso, la verdadera relación polisomal a total. Por este motivo intentamos estimarla, llamándole a la relación coeficiente de expansión (*EC*). *EC* representa el numero por el cual debemos multiplicar los conteos de total para poder obtener la verdadera relación $\frac{POLISOMAL}{TOTAL}$. A su vez, cada gen (ARN mensajero) tiene su propia tasa de asociación al polisoma, que denominamos h_i. Valores de h altos implican alta asociación al polisoma, mientras que bajos implican baja tasa de asociación. Investigar el h de los genes durante la adipogénesis, nos puede brindar información acerca de los mecanismos post-transcripcionales que influencian a los genes. Para estimar estos parámetros *EC* y h utilizamos el programa JAGS (Just Another Gibbs Sampler), el cual estima modelos bayesianos jerárquicos utilizando simulaciones basadas en cadenas Markov Monte Carlo (MCMC). Nuestro modelo se define de la siguiente manera:

$$POLISOMAL_i \sim N(EC \times h \times (TOTAL_i + y_i), \frac{1}{\tau}),$$

siendo i los genes, y_i un error con distribución normal con media 0 y precisión τ. Las priors para cada uno de los parámetros son las siguientes:
$EC \sim dgamma(3, 0{,}5)$, $h \sim dbeta(1{,}0, 1{,}0)$, $\tau \sim \frac{1}{\sigma^2}$, $\sigma \sim dunif(0, 10)$.
Ya que estamos trabajando en un marco bayesiano, se estiman las modas de las distribuciones a posteriori para obtener un valor numérico representativo del parámetro. Una vez estimado *EC*, se puede obtener el *h realizado* para cada gen como:

$$h_{realizado_i} = \frac{seudo - conteos_{POLISOMAL_i}}{seudo - conteos_{TOTAL_i}}$$

.

7.5. Resumen de resultados

Como muestra la tabla 1 del artículo adjunto, los biotipos siginificativos que están sobre-representados en las diferencias de ranking positivas son los genes condificantes, lo cual es esperable, y los ARN no codificantes "sense intronic" (sólo en CT), snoRNA, y miscellaneous RNA (sólo en IN). Los snoRNAs son pequeños ARN no codificantes que modifican entre otros a ARN ribosomales. Algunos estudios han observado snoRNAs asociados a polisomas [217, 218]. Un estudio de RNA-Seq en levadura ha encontrado ARN no codificantes intrónicos asociados a polisomas [218]. Los biotipos con OR$<$ 1, por lo que están sobre-representados en las diferencias de ranking negativas (mayor expresión en ARN total), incluyen los siguientes: lincRNA, miRNA, seudogenes, transcriptos antisentido (sólo en IN) y transcriptos procesados (sólo en CT). La aparición de los miRNA en este grupo, si bien es en cierta medida esperada, habla probablemente a favor de que, de los dos mecanismos mayoritarios para la inhibición de expresión mediada por miRNAs, el mecanismo de degradación del mensajero estaría primando sobre el de bloqueo de la traducción. Estudios muestran que muchos de los ARN largos no codificantes (como ser lincRNAs y antisenso) tienen localización nuclear por lo que no tienden a estar asociados a polisomas [219].

Al analizar los genes diferencialmente expresados con Rank Product, entre polisomal y total (para CT) y lo mismo para IN, se obtienen 48 genes sobre-expresados en la fracción total en CT y 142 genes para IN. Los resultados del análisis de términos GO de estos 142 genes se puede observar en la Tabla 2 del artículo. Muchos términos están relacionados con la regulación del proceso neuronal. Este tipo de marcadores neuronales y factores de transcripción relacionados con las funciones neurales se observan normalmente en células madre no diferenciadas.

Los coeficientes *EC* para CT e IN fueron estimados, obteniéndose como resultado 2,89 para CT y 1,29 para IN. Esto implica que las diferencias entre total y polisomal son mucho más grandes en CT que durante la inducción (aproximadamente el doble). El proceso de inducción causa un movimiento de los genes hacia los polisomas, para su activa traducción. A partir de los *EC* calculados se estiman los *h realizados* para cada gen. A su vez, las diferencias de tasa de asociación entre IN y CT se estiman mediante el Rank Product. Se obtienen 150 genes sobre-traducidos en IN cuyas funciones se pueden observar en la Tabla 3 del artículo adjunto. Todas las funciones están relacionadas con la diferenciación, sobre todo de células mesenquimales. Esto implica que los genes altamente asociados a polisomas son los relevantes para el proceso de diferenciación.

En el artículo presentado en el capítulo 4 se determinó el grado de regulación post-transcripcional evaluando los genes que cambiaban exclusivamente en una fracción de ARN y no en la otra. Por ejemplo, se evaluó cuántos genes cambiaban su expresión en la fracción polisomal (logFC= $\frac{IN}{CT}$) y no lo hacían en el ARN total (grupo P), por otro lado se determinaron los genes que cambiaban en ARN total y no lo cambiaban su expresión en la fracción polisomal (grupo T). Por último también se determinaron los genes que cambiaban consistentemente en ambas fracciones (grupo B). Se analizó la distribución de los *h realizados* para cada uno de estos grupos. Para lograr una mejor comparación con los datos de nuestro trabajo anterior, se evalúa el cociente en log ($log_2 \frac{h_{IN}}{h_{CT}}$) de los *h realizados*. La figura 3 del artículo muestra las distribuciones para los tres grupos. Los resultados son consistentes, presentando un posible escenario de lo que está aconteciendo durante la inducción. Por ejemplo, los genes del grupo P, que están sobre-expresados en IN, presentan altas tasas de asociación a polisomas. A nivel de ARN total, CT e IN presentan

niveles similares de ARN, sin embargo, la tasa de asociación alta en IN, hace que aumente la cantidad de mensajero en el polisoma, por lo que se observan cambios a nivel de la fracción polisomal. Igualmente para el grupo de genes T. Análogamente para el grupo T. Más detalles en la discusión del artículo.

7.6. Artículo

RESEARCH

Systematic differences between translatome and transcriptome during adipogenesis: total vs. polysomal RNA

Lucía Spangenberg[1], Alejandro Correa[2], Bruno Dallagiovanna[2] and Hugo Naya[1,3*]

*Correspondence:
naya@pasteur.edu.uy
[1]Institut Pasteur de Montevideo, Uruguay, Mataojo 2020, Montevideo, Uruguay
Full list of author information is available at the end of the article

Abstract

Translational control of gene expression is a mechanism that affects the relationship between mRNA level and protein level. An "uncoupling" of the transcriptome with respect to the translatome is observed and expected. Sequencing studies are focusing on RNA associated to polysomes as a more accurate measure of protein levels. However, with this approach the functions and regulatory mechanisms of a very interesting RNA fraction remains unveiled. Here, we focus on the differentiation process of mesenchymal stem cells to adipocytes and studied the differences between total RNA and the polysomal RNA fraction. We estimate an "expansion coefficient", which corresponds to the expansion in sequencing efforts that should be applied to the total fraction to accurately represent the true relation between total/polysomal RNA quantities. After that, we calculated for each gene its polysomal association rate, "h", which gives a hint about the translation efficiency at gene level. Moreover, we investigated the Biotypes distributions in both fractions, finding interesting biases in particular ones, such as snRNA, protein coding, pseudogenes, among others.

Results: Our results show that there is an expansion coefficient of ~ 2.9 in the polysomal fraction for non-differentiating cells and ~ 1.3 for adipocyte induced cells. We incorporate this factor to the comparisons of polysomal association gene-wise and found that genes presenting higher polysomal association rates were relevant to the differentiation process. We also observed that genes changing exclusively in the polysomal fraction and are upregulated in induced samples (wrt to control) present high association to polysomes, while downregulated genes present a high association in control samples. Analogous behavior can be seen in genes changing exclusively in total RNA. On the other hand by studying the biases in biotypes distributions we found that several biotypes are over-represented in the polysomal fraction rather than total RNA, such as protein coding genes, antisense, pseudogenes, etc.

Conclusion: We investigated the differences between total RNA and polysomal RNA fraction in mesenchymal stem cells and during the processes of adipogenic differentiation and found systematic differences regarding gene expression, gene association rates and the corresponding biotypes between both fractions.

Keywords: polysomal association; RNA-Seq; adipogenesis

Background

Adipocyte derived stem cells (hASCs) are multipotent cells that are able to proliferate, self-renew and differentiate into several specific lineages, such as chondrogenic, osteogenic, adipogenic and myogenic lineages. Adipose tissue is ubiquitous

and large quantities can be extracted using minimal invasive procedures. Such characteristics make hASC ideal candidates for use in cell therapy. Understanding the commitment of these cells to differentiation to a specific cell type is essential for the successful repair or regeneration of injured tissues. Gene expression analyses, especially using deep sequencing techniques have provided great insights into the regulatory networks determining self-renewal and differentiation processes [1, 2]. Many of those studies were based on total mRNA levels in the cell. However, evidence suggests that differentiation is also dependent on the control of protein synthesis by post-transcriptional regulation mechanisms. Translational control of gene expression is a mechanism that affects the relationship between mRNA level and protein level. As suggested by Tebaldi *et al* [3], an uncoupling of the transcriptome (total mRNA in the cell) with respect to the translatome (mRNA associated to polysomes) is observed and is expected after stimuli in mammalian cells. This deregulation is also observed in stem cell differentiation [4], cancer [5, 6] and other diseases. Such post-transcriptional mechanisms might include alternative polyadenylation sites (APA), that generate alternative transcripts with different 3'UTR length, which might affect, among other characteristics, miRNA susceptibility (fewer/more miRNA binding sites), hence potential degradation or translational repression of the mRNA [7]. Another mechanism might be the regulation of translation initiation, i.e. ribosomes associating to messenger ribonucleoprotein particles (mRNP) to form polysomes [8, 9]. This mechanism can be controlled at different levels to influence polysome formation: i) by regulating general initiation factors such as eIF4E, eIF4G, eIF4A and PABP, which are proteins responsible for mRNA circularization and ribosome scanning and ii) by the effect of RNA binding proteins (RBP) or non-coding RNAs, which might recognize sequences on 3' or 5'UTR regions and are able to regulate polysome formation, mRNA stabilization, mRNA localization, etc. Non-coding RNA (ncRNA) include also the above mentioned miRNAs. A further post-transcriptional mechanism could involve the pseudogenes. Recent studies have observed regulatory functions of several pseudogenes in cancer [10], in stem cell differentiation [11], among others [12].

In this context, many groups have been recently focusing on analyzing mRNA associated to translating polysomes instead of total RNA [13, 14, 4], since it should reflect more accurately the protein level in the cell. In a recent work of ours, we found a high degree of post-transciptional regulation in the first stages of adipogenesis, by analyzing RNAseq data of total mRNA and RNA associated to polysomes [4].

In the present study, we described the main differences in gene expression, biotype usage and association rates between polysomal fraction and total RNA, so we might be able to have a better picture about the post-transcriptional regulation mechanisms during adipogenesis.

Materials and Methods

Sample description and primary analysis of SOLiD RNA-seq

We used samples described by Spangenberg *et al.* [4]. Raw data is available in the ArrayExpress repository under the accession number E-MTAB-1366. As described before [4] stem cells were obtained from adipose tissue of three obese human donors.

Human adipocyte derived stem cells (hASCs) were obtained, cultured and characterized as previously described [15]. The induction to adipogenic differentiation was achieved by 6 day-cycles of induction/maintenance over 21 days and the degree of adipogenic differentiation was established by determining cytoplasmic accumulation of triglycerides by staining with Oil Red O or Nile Red (Sigma-Aldrich). Samples were taken at time point 0 (hASC samples or control samples, CT) and then after three days (induced samples, IN). At each time point, two types of samples were extracted for sequencing. One comprehends total RNA in the cell, and the other one only the fraction associated to polysomes. Altogether, a total of 13 samples were sequenced with SOLiD4 System (Applied Biosystems), 7 CT (2 polysomal-associated RNA and 6 total RNA samples) and 6 IN (3 polysomal-associated RNA and 3 total RNA).
An overview of the quality analysis of the samples can be seen in [4]. The mapping procedure was performed with Rsubread [16] using as reference genome version hg19 GR37p2. Gene expression determination and differential expression analysis was done with edgeR [17], using genome annotation version Hg19 Gr37p10 (August 2012), downloaded from ensembl. Biotype information was also obtained from ensembl (also based on Gr37p10).

Biotype analysis

In order to have a general view about the regulatory mechanisms involved during differentiation we decided to investigate the biotypes over/under-represented during the process. For this we considered expression ranking differences between polysomal and total RNA fractions ($\text{rank}_{polysomal} - \text{rank}_{total}$). First, we ranked the gene expression values (in normalized number of reads) in each patient and fraction (polysomal and total). Second, we calculated the differences of ranking (polysomal−total) for each patient and we kept only those genes that are consistent between patients (genes that have the same sign in the difference in each of the three patients). Control samples (polysomal compared with total) and induced samples were separately analyzed.
Proportion of different biotypes were determined in each of the following groups: i) $\text{rank}_{polysomal}$-$\text{rank}_{total} > 0$ and $\text{rank}_{polysomal}$-$\text{rank}_{total} < 0$ for control and ii) $\text{rank}_{polysomal}$-$\text{rank}_{total} > 0$ and $\text{rank}_{polysomal}$-$\text{rank}_{total} < 0$ for induced.
In order to assess significance of the determined proportions Fisher test was performed for each biotype. In both cases i) and ii) negative ranking differences correspond to genes with low expression in the polysomal fraction (small ranking) and a high expression in the total RNA (higher ranking), hence $\text{ranking}_{polysomal}$-$\text{ranking}_{total}$ is a negative value. OR< 1 (the proportions in Fisher test are $\frac{rk_{dif}>0}{rk_{dif}<0}$) for a particular biotype means that it is over-represented in negative ranking differences, hence genes belonging to this biotype tend to be at higher rankings of expression in total and rather at lower rankings in polysomal; they are shifting their behavior from being the most expressed (in total) to be less expressed in polysomal. OR> 1 means that the biotype is over-represented in positive ranking differences, including genes that are more expressed in polysomal (higher ranking) and tend to be less expressed in total (lower ranking).
We applied fisher tests for all biotypes (32 biotypes as defined in ensembl), in control and in polysomal. In order to compare induced with control samples to get an

idea of the differences in biotypes usage during differentiation, we compared the ORs obtained in each case by z-score statistics, obtaining for each comparison a p-value. See table 1 for results.

Rank Product analysis

The Rank Product is a non-parametric test applied to detect differentially expressed genes in replicated samples. It determines genes that are consistently found among the most upregulated genes in a number of replicate experiments in each condition. Rank Product assumes that under the null hypothesis the order of all genes is random and the probability of finding a specific gene among the top r of a list of n elements is p= $\frac{r}{n}$. Multiplying these probabilities leads to the definition of the rank product RP = $\Pi_i \frac{r_i}{n_i}$, where r_i is the rank of the gene in the i^{th} list and n_i is the total number of genes in the i^{th} list. We use the R package RankProd [18] to determined the RP for each gene in each class (total and polysomal) on the normalized counts data and to estimate the FDR in each case. As a result we obtained a list of genes that are over-represented in total and polysomal.
Those identified genes were further analyzed. On the one side a GO analysis was performed and on the other side Fisher tests on the biotypes were applied.

Estimation of proportions "*EC*" and "*h*"

For each biological sample, we compared to different RNA fractions, polysomal and total. In ideal circumstances (if total RNA is truly the total RNA in the cell) all polysomal RNA should be contained in the total sample. However, we have found gene counts in polysomal with no counts in the total fraction. While several error source can be envisaged, including isolation techniques, the most evident source of inconsistencies come from direct counting comparison methodology. Moreover, since the sequencing is performed separately, total and polysomal, the "true" polysomal/total RNA proportion is not conserved or even known. This means, that even though the amount of total RNA is much larger than polysomal, the result of the sequencing (the amount of reads and recovered genes) should be approximately the same, since we exploit the sequencing process to the maximum. We are losing in the process the real proportions of total and polysomal RNA. For this reason we tried to estimate what we called the "expansion coefficient" (*EC*), the number by which the total counts should be multiplied to arrive at the "true" proportions of polysomal/total RNA.
In addition, each gene (mRNA) has a different rate of association to polysomes, which we named "h_i" (denoting the i^{th} gene). If h_i is high, a relatively high proportion of the available mRNA is being translated. By the contrary, if h_i is low, most mRNA is not actually involved in protein production. Additionally, an error term should be introduced to account for sampling differences in the counting process. Ideally, all parameters should be estimated from the data, but for simplicity reasons and due to the low number of samples in comparison to the number of parameters to estimate, we decided to consider a general h for all genes and estimate more accurately *EC*. Departing from this estimate and after multiplying the total pseudo-counts (PC) by the *EC*, we calculated the "realized h_i" for each gene ($\frac{polysomal_{PC_i}}{EC * total_{PC_i}}$).

For the estimations we made use of JAGS (Just Another Gibbs Sampler), which is a program for the analysis of Bayesian hierarchical models using Markov Chain Monte Carlo (MCMC) simulation, similar to BUGS. Our model to be estimated is defined as follows:

$$POLYSOMAL_i \sim N(EC \times h \times (Total_i + y_i), 1/\tau)$$

with i standing for each gene, y_i a normally distributed error with mean 0 and precision τ. The priors for each parameters are:
$EC \sim dgamma(3, 0.5)$, $h \sim dbeta(1.0, 1.0)$, $\tau \sim \frac{1}{\sigma^2}$, $\sigma \sim dunif(0, 10)$.
In order to evaluate convergence of the chains we applied the Geweke test. We work only on convergent chains. Distributions mode were estimated using the function mlv ("modeest" R package) using Parzen method with gaussian kernel and visually adjusting the bandwidth. Estimates for EC and h are analyzed and discussed in the results section.

Results

Biased distributions within biotypes

In order to characterize the differences in biotypes in total and polysomal RNA fraction, we analyzed the ranking differences within each biotype. Are there some specific biotypes over-represented in positives (or negatives) ranking differences? Are the genes shifting their behavior within a specific biotype? Biased distributions might give us information about the regulation mechanisms underlying differentiation.
We analyzed the distribution of rank differences within biotypes and observed some biases for particular biotypes. In order to assess the significance of the observed biases Fisher tests in biotypes were performed.
Table 1 shows the Fisher tests results for ranking differences in CT and IN samples. The first three columns hold the results for CT samples and the following three have the IN results. The results includes odds ratio (OR) and a corresponding p-value; the sense of the comparison is positive against negative. The last column holds the p-value corresponding to the test comparing both ORs, IN against CT. The comparisons are made with z-score statistics. Only Fisher-significant biotypes (in either fraction) are shown. As explained above an OR>1 in this case means an over-representation in positive ranking differences. Genes belonging to this biotype are shifting their behavior from being low expressed in total to be highly associated to polysomes. We observed OR >1 in both conditions CT and IN, for protein coding genes, sense-intronic, snoRNA and miscelaneous RNA. As expected, OR<1 in both conditions were observed in pseudogenes, antisense, processed transcripts, lincRNA and miRNA. Genes belonging to these biotypes are over-represented in negative differences, hence they are expressed in total but they are not reaching polysomes efficiently. Only snRNA presents an OR<1 in CT and OR>1 in IN, shifting its behavior.
Figure 1 A and B shows a cumulative rank sum of ranking differences for control and induced samples, respectively. Each genes' rank difference contributes with a small proportion to the total ranking difference. When considering the genes belonging to

its specific biotype, we observed the contribution of each biotype to the total rank difference. For example, protein-coding genes (in control; CT and in induced; IN), are below the null-distribution line (all genes considered, regardless of biotype) in large negative rank differences, while they are above the null-distribution line in smaller negative and all positive differences. This means, that protein-coding genes tend to be more associated to polysomes (more expression in polysomal, larger ranking) than expressed in the total fraction (less expression in total, smaller ranking). Furthermore, miRNAs (green) tend to be slightly above the null-distribution line in large negative differences and far below the null-line in all positive values (in CT and IN). As expected, we observed that miRNAs are more expressed in total RNA than in polysomal fraction (over-represented in negative differences and under-represented in positive differences). The same happens for all non-coding elements: lincRNA, snoRNA, snRNA, antisense, etc. Pseudogenes present (in both cases, CT and IN) a more ambiguous behavior since it runs almost along the null-line.

Differentially expressed genes according to Rank Product

Rank Product test on total vs polysomal normalized counts was separately applied for control (CT) and induced (IN) samples. As polysomal and total samples aren't independent (polysomal is an unknown fraction of total) classical differential expression methodolgies could led to a very biased view. While Rank Product has similar biases we used it here just as a first approach and to compare results with a similar analysis on h values. We identified 48 genes upregulated in total RNA (wrt polysomal) in CT and 142 in IN. No upregulated genes were found in polysomal, which was expected due to the asymmetry in gene composition between the two fractions.

Analyzing the GO biological process of upregulated genes in IN we found several terms related to neurological regulation processes (Table 2). No significant biological processes were found in CT. The reason might be that very few protein coding genes (only 10) were found among these 48 differentially expressed genes in CT and GO analysis rely on functions of protein coding genes. The rest were non-coding RNAs: antisense (11), lincRNA (6), miRNA (1), miscellaneous RNA (2), processed transcript (3), pseudogene (10), sense intronic (2), snRNA (1) and snoRNA (2). The distribution of protein coding and non-coding RNAs in IN was similar, but since we obtained more genes in general, we were able to find significant GO terms.

Translational efficiency: EC and h estimations

The estimations of EC and h were performed for CT and IN, separately. Since we worked on a bayesian framework, the results are posterior distributions of the parameters. Figure 2 A shows the posterior distributions of parameter EC for CT (blue) and IN (red). Figure 2 B shows the posteriors for h in both conditions. Since we have not enough data to estimate one individual h for each gene, we estimated a single general parameter h for all genes. In order to determine such individual h parameters (so we can analyze the association rate of all genes to polysomes) we determined an "*realized h*" for each gene (as explained in Estimation of proportions

"*EC*" and "*h*"). The distributions of *realized h* are in Figure 2 C for CT, and D for IN. In order to obtain a value a representative value summarizing the posteriors, we determined the mode for parameter *EC*. We obtained an *EC* coefficient of 2.81 for CT and 1.29 for IN. These results suggest that there is an increase in the proportion of RNA being translated, or in other words a higher association to polysomes (lower *EC* reflects a higher similarity between both fractions).

For each gene and each donor we have a *realized h* value, representing the association rate to polysomes. We then determined "differentially translated" genes, which are those genes having different h_i under IN conditions wrt CT. We use the non-parametric Rank Product test (RankProd R package) to detect such genes. On the one hand, we found 150 up-translated genes in IN, meaning that those genes present higher *h*'s in IN. On the other hand, we found 252 genes that are down-translated in IN, hence those genes have lower *h*'s in IN, or higher *h*'s in CT. The function of these two sets of genes were analyzed via GO analysis using GOrilla server [19]. Significant Biological Process terms were found for genes up-translated in IN and the terms can be seen in table 3. They are mainly involved in various differentiation processes, meaning that the genes whose association rates to polysomes increases the most during induction are the ones relevant in such process. Not only should one expect an increase in their transcription levels, but also an increase in association to the translational machinery. We found no relevant GO term for upregulated genes in CT.

In order to further investigate the differential association rates we analyze the change of *h* inside each biotype. Biotypes that strongly change rates between IN and CT should be significant in a test that evaluates ranking differences inside biotypes, such as Wilcoxon Rank Sum test. We observed that protein coding genes, pseudogenes, processed transcripts and lincRNA are significantly changing between conditions. Table 4 shows the results. We combined the p-values for each sample in a single p-value using the Fisher Method [20] and we also computed a combined median for all samples (last column). This should give the direction of the change. Since all combined median are positive these biotypes are changing towards higher association rates in induced samples.

In a previous work [4] in an attempt to assess the degree of post-transcriptional regulation during adipogenesis, we analyzed the genes that showed changes in one fraction but presented no changes in the other one. A gene was considered to change expression if $|logFC| > 1.5$ (logFC= $log\frac{IN}{CT}$) and no change was considered when $|logFC| < 1.5$. For example, we observed 13% of the analyzed genes presenting changes exclusively in the polysomal fraction (no change in total RNA) and we observed 44% of the genes with changes exclusively in total RNA, which might be explained by the large proportion of genes that are not supposed to be translated (pseudogenes, miRNAs, etc). As expected, several genes (42%) presented changes in both fractions simultaneously. We were now interested in analyzing the *h* distribution of all these gene sets. Figure 3 shows the distribution of each set of genes, separately for positive and negative logFC values. The distributions observed are log ratios of *h*'s values in IN compared to CT, $log_2\frac{h_{IN}}{h_{CT}}$. The distribution of *h* log-odds ratios for genes only changing in polysomal fraction with positive logFC (light red) are clearly towards the large values of *h*'s distribution. Such genes are over-expressed in IN only in polysomal fraction and present similar RNA levels (CT

and IN) in total RNA. The high association rate to the translational machinery in IN (high h-log ratios) as observed in the figure 3, might be the responsible for the shift in expression ($logFC > 1.5$ in polysomal). The same for the genes only changing in total with negative logFC, they lie also towards the larger positive h values. In this case, genes are downregulated in IN, hence upregulated in CT (in total). Nevertheless no change is observed in polysomal. The high association rate of the genes in the IN conditions in total RNA might be the responsible for the balance of RNA levels in polysomal fraction, hence no changes are observed. There are two sets of genes lying on the left side of h's distributions, hence showing a high association rate in CT (wrt to IN): the ones exclusively changing in polysomal with $logFC < 0$ and the ones changing in total with $logFC > 0$. The first group presents similar RNA levels in total and a shift towards CT in polysomal. Such shift can be explained due to the high association rate of CT. The second group presents similar RNA levels in polysomal and a shift towards IN in the polysomal fraction. This might also be explained due to the high association rate of CT in total.
When considering the set of genes changing in both fractions, we observed an h's distribution centered around ~ 1, for both negative and positive logFCs. This concurs with our rationale since we are not expecting a differential association rate for the genes in this category, in either condition, IN or CT. The changes in the total fraction are also observed in the polysomal fraction (and in the same sense), so no additional regulation mechanisms is expected. The fact that both such distributions are centered around ~ 1 and not zero (since this would mean the same h for IN and CT), is also expected, since *realized* h tend to be larger in IN than in CT, approximately twice as large, due to the estimates of EC (EC of IN is 1.29 and is approximately $\sim \frac{1}{2}$ as large as CT's EC 2.89, resulting in twice as large values of *realized* h for IN, hence $log_2(2) = 1$).

Discussion

When analyzing the biased biotypes distribution in ranking differences we observed several categories having a tendency to highly associate to polysomes (OR$>$ 1). Protein coding genes were among them (in both conditions CT and IN), which was expected since they need to bind to the translational machinery in order to create the proteins. Falling into the same category as protein coding genes, we found snoRNA, miscellaneous RNA (only in IN) and sense intronic transcripts (only in CT). snoRNAs are small non-coding RNAs that modify post-transcriptionally other RNAs, such as snRNA, tRNA and especially rRNA. Several snoRNAs that methylate or pseudouridylate rRNA can be found associating to polysomes [21, 22]. Sense intronic transcripts are non-coding transcripts found in introns of coding or non-coding genes. A study has shown that this kind of long non conding RNA is found to be associated to ribosomes via deep sequecing analysis in yeast [22]. We found also several uncharacterized RNAs with a tendency to associate to polysomes during induction (miscellaneous RNAs), which might be having a regulatory function as well.
The biotypes with low association to polysomes included miRNA (in both conditions), lincRNA (in both), pseudogenes (in both), antisense (only in IN), processed transcripts (only in CT). Even though miRNAs might be expected to be depleted

in polysomal fraction, this gives a hint about the mechanisms of action preferred by them. Usually two ways of action are accepted for the inhibition of translation by miRNAs: degradation of transcript or translation blockage [23]. Since we are not finding them on the polysomes, degradation should be the most likely mechanism. lincRNA are long intergenic non coding RNA and have been shown to be involved in cell differentiation and maintenance of pluripotency state [24]. They tend to act mainly at transcriptional level [25, 26, 24], hence it concurs with the under-representation in the polysomal fraction. A large percentage of ncRNAs are nuclear-enriched with unknown function [27]. Antisense RNAs belong to this nuclear-enriched category and may form sense-antisense pairs by pairing with a protein-coding gene on the opposite strand to regulate epigenetic silencing, transcription and mRNA stability [28, 29, 30]. Since its location is mainly nuclear the tendency should be to find such non conding RNAs in total RNA than in the polysomal fraction. Specific antisense RNAs have been shown to have a post-trascriptional function during processes of differentiation (epithelial-mesenchymal transition) [29], which could explain the reason why the under representation in the polysomal fraction was only observed in IN. Pseudogenes were significant in both conditions, CT and IN. This should be expected since usually they are not supposed to be actively translated (hence associated to polysomes) [31, 32].
When analyzing the cumulative distribution of the differences, one observes in both conditions (CT and IN) that protein coding genes are well above the null-distribution line for all positives and small negative ranking differences, which is in agreement with what we found with Fisher tests.

In order to further investigate the differences in polysomal association, we tried to determine differentially "expressed" genes (or more precisely, differentially translated genes) in polysomal wrt total, for CT and IN separately. We acknowledge that a direct calculation of differentially translated genes between polysomal and total is not accurate due to the inclusion of the polysomal fraction into the total. For this reason, we first attempted to use an R package called ANOTA, which considers this inclusion of samples [33]. We were not able to find highly significant results with this approach (data not shown). Hence, we used a more conventional, non-parametric method (Rank Product) to directly calculate differentially "translated' genes between polysomal against total, in order to detect some of them whose signal is so clear that it overrides the inclusion effect. We found 142 in IN and 48 in CT that were upregulated in total RNA. No genes were found upregulated in polysomal, which is expected; genes that are highly associated to polysomes, should also have relative high levels in total RNA. Analyzing the function of the 142 genes found to be over-represented in the total fraction during induction we observed biological processes involved in the regulation of neurological processes. Studies have shown that adipose-derived human stem cells are able to differentiate also to ectodermal lineages *in vitro*, being able to give rise to neurons [34, 35]. Nevertheless, studies have shown that several of the commonly used neuro-marker genes are already expressed by undifferentiated human mesenchymal stem cells. Furthermore, mRNA for some of the neural-related transcription factors were also found to be strongly expressed [36, 37]. We observed neural related genes being over expressed in total

RNA during induction, which might imply that in CT those genes were being expressed, associating to polysomes, but on induction to adipogenesis, they tend to remain in the total fraction, not associating to polysomes.

So far we found differences in biotype usages and differentially expressed genes in the polysomal RNA fraction compared to the total RNA. In order to better quantify the differential association of all genes to polysomes, in stem cells as well as during adipogenesis, we estimated the *EC* coefficient and the realized h value for each gene. First, it is important to notice the differences between *EC* for IN (1.29) and *EC* for CT (2.89). This implies, that the differences between total and polysomal fraction are much larger in CT than during induction; about twice as much ($\frac{2.89}{1.29} = 2.24$). The induction process causes a shift of genes towards the polysomes, so they can be actively translated. We further analyzed the observed h for CT and IN. Figure 2 C and D shows the distribution of h for each condition, in which it can be seen that CTs distribution is more enriched in lower values, while INs distribution is more homogenous. This is also in agreement with the fact that there is a general shift towards the polysomes during induction. The high bar corresponding to $h = 0$ means that for those genes no polysomal information were found. We determined the "h-differentially translated" genes between IN and CT and analyzed the biological function of those 150 up-translated genes in IN via GOrilla server. We obtained several functions related to mesenchymal differentiation and again some functions involved in neurogenic differentiation. This implies that genes associating strongly to polysomes are relevant features for completion of differentiation. We found no significant biological process among the "h-translated" genes in CT. This might be due to the fact, that out of the 252 genes only 103 are protein coding (the rest corresponds to non-coding RNAs) and those 103 are genes belonging to more general functions, not characteristic to differentiation processes.
When investigating how h's are changing within biotypes during induction (results in Table 4), we found that the biotypes changing the most are protein-coding (expected), followed by pseudogenes, lincRNA and lastly processed transcript. Along this study, we found several hints about the involvement of this non-coding elements in the adipogenesis. In this case, we found that genes belonging to these biotypes tend to increase the association rate to polysomes during differentation.
Lastly, we compared previously published results with our h's distribution. In our previous work we were able to determined a group of genes changing in the polysomal fraction but not changing in the total fraction (P), an other group of genes changing only in total RNA (T), and a further group with genes changing consistently in both fractions (B). Here, we analyzed the log h-distribution in each group. As mentioned before (in section "Translational efficiency: *EC* and h estimations") we were expecting a relative homogenous distribution around 1 for the genes belonging to B, since transcripts levels in total are also reflected in the polysomal fraction; there is no necessity of differential association of transcripts to polysomes to increase/decrease the effect. However, for genes belonging to group P and are upregulated in IN, the changes in the polysomal fraction might be due to the high association rates and not due transcripts levels in total. The log odds h-distribution shows a clear shift towards larger h values. For genes in P but showing downregulation in IN (upregulation in CT), the log-odds h distribution shows a shift towards

the larger negative values, which is expected since log-odds h is calculated as the ratio $log_2(\frac{h_{IN}}{h_{CT}})$, implying a higher association rate to polysomes in CT. The same rationale holds here: genes in P, which present no change in transcripts levels in total, are changing expression in polysomal due to higher association rates. Gene group T presents the same consistent results as P. The upregulated genes in T (IN vs CT) present lower log-odds h values, hence they are associating more in CT equating both transcripts levels (CT and IN) in the polysomal fraction. The same rationale holds for genes downregulated in T. This results show consistency with previously published results, and offers a potential explanation of the changes observed in either one fraction.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section . . .

Acknowledgements
Text for this section . . .

Author details
[1]Institut Pasteur de Montevideo, Uruguay, Mataojo 2020, Montevideo, Uruguay. [2]Instituto Carlos Chagas, Fiocruz-Paraná, Curitiba, Rua Prof. Algacyr Munhoz Mader 3775, Curitiba, Brasil. [3]Departamento de Producción Animal y Pasturas, Facultad de Agronoma, Universidad de la Republica, Avenida Gral. Eugenio Garzón 780, Montevideo, Uruguay.

References

1. Ivanova, N.B., Dimos, J.T., Schaniel, C., Hackney, J.A., Moore, K.A., Lemischka, I.R.: A stem cell molecular signature. Science **298**(5593), 601–604 (2002)
2. Song, L., Webb, N.E., Song, Y., Tuan, R.S.: Identification and functional analysis of candidate genes regulating mesenchymal stem cell self-renewal and multipotency. Stem Cells **24**(7), 1707–1718 (2006)
3. Tebaldi, T., Re, A., Viero, G., Pegoretti, I., Passerini, A., Blanzieri, E., Quattrone, A.: Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. BMC Genomics **13**(1), 220 (2012)
4. Spangenberg, L., Shigunov, P., Abuda, A.P.R., Cofré, A.R., Stimamiglio, M.A., Kuligovski, C., Zych, J., Schittini, A.V., Dias Tavares Costa, A., Rebelattoc, C.K., Brofman, P.R., Goldenberg, S., Correa, A., Naya, H., Dallagiovanna, B.: Polysome profiling shows extensive posttranscriptional regulation during human adipocyte stem cells differentiation into adipocytes. Stem Cell Research **1**(2), 341–347 (2013)
5. Luyimbazi, D., Akcakanat, A., McAuliffe, P.F., Zhang, L., Singh, G., Gonzalez-Angulo, A.M., Chen, H., Do, K.-A., Zheng, Y., Hung, M.-C., et al.: Rapamycin regulates stearoyl coa desaturase 1 expression in breast cancer. Molecular Cancer Therapeutics **9**(10), 2770–2784 (2010)
6. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., Selbach, M.: Global quantification of mammalian gene expression control. Nature **473**(7347), 337–342 (2011)
7. Spangenberg, L., Correa, A., Dallagiovanna, B., Naya, H.: Role of alternative polyadenylation during adipogenic differentiation: an *in silico* approach. PLoS One (2013)
8. Groppo, R., Richter, J.D.: Translational control from head to tail. Current Opinion in Cell Biology **21**(3), 444–451 (2009)
9. Jackson, R.J., Hellen, C.U.T., Pestova, T.V.: The mechanism of eukaryotic translation initiation and principles of its regulation. Nature Reviews Molecular Cell Biology **11**(2), 113–127 (2010)
10. Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., Pandolfi, P.P.: A coding-independent function of gene and pseudogene mrnas regulates tumour biology. Nature **465**(7301), 1033–1038 (2010)
11. Lin, H., Shabbir, A., Molnar, M., Lee, T.: Stem cell regulatory function mediated by expression of a novel mouse oct4 pseudogene. Biochem Biophys Res Commun **355**, 111–116 (2007)
12. Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L., Carter, D.R.F.: Pseudogenes: pseudo-functional or key regulators in health and disease? Rna New York Ny **17**(5), 792–798 (2011)
13. Kolle, G., Shepherd, J.L., Gardiner, B., Kassahn, K.S., Cloonan, N., Wood, D.L.A., Nourbakhsh, E., Taylor, D.F., Wani, S., Chy, H.S., et al.: Deep-transcriptome and ribonome sequencing redefines the molecular networks of pluripotency and the extracellular space in human embryonic stem cells. Genome Research **21**(12), 2014–25 (2011)
14. Fromm-Dornieden, C., Von Der Heyde, S., Lytovchenko, O., Salinas-Riester, G., Brenig, B., Beissbarth, T., Baumgartner, B.G.: Novel polysome messages and changes in translational activity appear after induction of adipogenesis in 3T3-L1 cells. BMC Molecular Biology **13**(1), 9 (2012)
15. Rebelatto, C.K., Aguiar, A.M., Moretão, M.P., Senegaglia, A.C., Hansen, P., Barchiki, F., Oliveira, J., Martins, J., Kuligovski, C., Mansur, F., et al.: Dissimilar differentiation of mesenchymal stem cells from bone marrow, umbilical cord blood, and adipose tissue. Experimental biology and medicine Maywood NJ **233**(7), 901–913 (2008)

16. Liao, Y., Smith, G.K., Shi, W.: The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research **41**(10), 108 (2013)
17. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. Nucleic Acids Research **26**, 139–140 (2010)
18. Hong, F., with contribution from Rainer Breitling, B.W., Smith, C., Battke, F.: RankProd: Rank Product Method for Identifying Differentially Expressed Genes with Application in Meta-analysis. (2011). R package version 2.30.0
19. Eden, E., Navon, R., Steinfeld, I., Lipson, D., Yakhini, Z.: Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. BMC Bioinformatics **10**(1), 48 (2009)
20. Fisher, R.A.: Combining independent tests of significance. American Statistician **2**(5), 30 (1948)
21. Tycowski, K.T., Shu, M.D., Steitz, J.A.: A mammalian gene with introns instead of exons generating stable rna products. Nature **379**(6564), 464–466 (1996)
22. Zywicki, M., Bakowska-Zywicka, K., Polacek, N.: Revealing stable processing products from ribosome-associated small rnas by deep-sequencing data analysis. Nucleic Acids Research **40**(9), 1–12 (2012)
23. Bartel, D.P.: Review micrornas : Target recognition and regulatory functions. Cell **136**(2), 215–233 (2009)
24. Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al.: lincrnas act in the circuitry controlling pluripotency and differentiation. Nature **477**(7364), 295–300 (2011)
25. Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al.: Widespread transcription at neuronal activity-regulated enhancers. Nature **465**(7295), 182–187 (2010)
26. øOrom, U.A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., et al.: Long non-coding rnas with enhancer-like function in human cells. Cell **143**(1), 359–361 (2010)
27. Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., et al.: Rna maps reveal new rna classes and a possible function for pervasive transcription. Science **316**(5830), 1484–1488 (2007)
28. Hastings, M.L., Ingle, H.A., Lazar, M.A., Munroe, S.H.: Post-transcriptional regulation of thyroid hormone receptor expression by cis-acting sequences and a naturally occurring antisense rna. The Journal of Biological Chemistry **275**(15), 11507–13 (2000)
29. Beltran, M., Puig, I., Pena, C., Garcia, J.M., Alvarez, A.B., Pena, R., Bonilla, F., De Herreros, A.G.: A natural antisense transcript regulates zeb2/sip1 gene expression during snail1-induced epithelial-mesenchymal transition. Genes Dev **22**(6), 756–769 (2008)
30. Ebralidze, A.K., Guibal, F.C., Steidl, U., Zhang, P., Lee, S., Bartholdy, B., Jorda, M.A., Petkova, V., Rosenbauer, F., Huang, G., et al.: Pu.1 expression is modulated by the balance of functional sense and antisense rnas regulated by a shared cis-regulatory element. Genes & Development **22**(15), 2085–2092 (2008)
31. Mighell, A.J., Smith, N.R., Robinson, P.A., Markham, A.F.: Vertebrate pseudogenes. FEBS Letters **468**(2-3), 109–114 (2000)
32. Balakirev, E.S., Ayala, F.J.: Pseudogenes: are they "junk" or functional dna? Annual Review of Genetics **37**(1), 123–51 (2003)
33. Larsson, O., Sonenberg, N., Nadon, R.: Anota: ANalysis Of Translational Activity (ANOTA). (2011). R package version 1.4.0
34. Anghileri, E., Marconi, S., Pignatelli, A., Cifelli, P., Galié, M., Sbarbati, A., Krampera, M., Belluzzi, O., B., B.: Neuronal differentiation potential of human adipose-derived mesenchymal stem cells. Stem Cells & Development **17**(5), 909–916 (2008)
35. Franco Lambert, A.P., Fraga Zandonai, A., Bonatto, D., Cantarelli Machado, D., Pêgas Henriques, J.a.A.: Differentiation of human adipose-derived adult stem cells into neuronal tissue: does it work? Differentiation research in biological diversity **77**(3), 221–228 (2009)
36. Chen, Y., Teng, F.Y.H., Tang, B.L.: Coaxing bone marrow stromal mesenchymal stem cells towards neuronal differentiation: progress and uncertainties. Cellular and molecular life sciences CMLS **63**(14), 1649–1657 (2006)
37. Montzka, K., Lassonczyk, N., Tschöke, B., Neuss, S., Führmann, T., Franzen, R., Smeets, R., Brook, G.A., Wöltje, M.: Neural differentiation potential of human bone marrow-derived mesenchymal stromal cells: misleading marker gene expression. BMC Neuroscience **10**(16), 16 (2009)

Figures

Figure 1 Distribution of ranking differences within biotype. (A) Cumulative rank sum of ranking differences in CT samples. The x-axis represents the ranking differences. Each gene is represented as a dot colored according to its biotype. For each gene its contribution (proportion) to the total ranking difference within the considered biotype is represented in the y-axis. A dotted black line represents the contribution of each gene to the ranking difference regardless of the biotype (null distribution). Protein-coding genes (red) are below the null-distribution line for negative differences (under-represented in negative values), and are clearly above de null-line for positive rank differences (over-represented in positive differences). An opposite behavior can be seen among the other biotypes. (B) The same plot for induced samples. The same behavior of protein-coding genes is observed.

Figure 2 EC **and** h **estimations.** A) shows the distribution of estimated EC for CT (blue) and IN (red) samples. B) shows the distribution of estimated h for CT (blue) and IN (red) samples. C) shows the $realized\ h$ for CT and D) shows the same for IN. Large frequency bars around 0 implies a large number of genes having $realized\ h$ equal to zero. This is the case for genes having no counts in the polysomal fraction.

Figure 3 Distribution of log ratios of h **values of genes changing exclusively in one fraction or changing in both.** We considered log ratios of h values, calculated as follows $log_2(\frac{h_{IN}}{h_{CT}})$ to resemble our previous analysis [4]. Three groups of genes are shown: in red tones the ones changing only in the polysomal fraction ($|logFC| > 1.5$). In a lighter red the ones with $logFC > 1.5$ and in a darker red the ones with $logFC < 1.5$. In green tones the genes only changing in total are shown, with light green showing genes with $|logFC > 1.5|$ and dark green representing genes with $logFC < -1.5$. In blue tones are the genes that change in both fractions. Light blue represents positive logFC values and dark blue negative ones.

Table 1 Fisher tests results. Significance of the different proportions of each biotype (BT) in the groups of genes with ranking differences > 0 and < 0 were evaluated via Fisher test. OR stands for the resulting odds-ratio and the p-value corresponds to the significance of the Fisher test. This is performed for control and induced samples (eg. OR CT, OR IN). We assess the significance of the difference of both ORs (CT and IN) with a z-score test. p-values are displayed in the last column and significant ones at 0.05 confidence are distinguished with "*".

	RANKINGS DIFFERENCES VALUES (rank POLY−rank TOTAL)				
BT	**OR CT**	**p-value CT**	**OR IN**	**p-value IN**	comparison IN vs CT (p-value)
protein coding	1.438	2.2×10^{-16}	1.398	$2.20\times^{-16}$	0.24196
pseudogenes	0.720	2.2×10^{-16}	0.633	$2.20\times^{-16}$	0.00453*
antisense	0.944	0.3742	0.849	0.00925	0.00657*
processed transcript	0.788	2.81×10^{-3}	0.932	0.402	0.06811
lincRNA	0.819	0.000788	0.742	$6.41\times^{-7}$	0.11702
miRNA	0.607	$4.07\times^{-7}$	0.769	$9.13\times^{-3}$	0.04648*
sense intronic	1.671	7.84×10^{-3}	2.0215	0.2451	0.2451
snRNA	0.709	7.63×10^{-5}	1.0745	0.431	0.0004*
snoRNA	1.316	0.00608	1.258	0.04048	0.37448
misc RNA	1.143	0.1614	1.265	0.01242	0.22363

Table 2 GO Analysis Results from RankProd

GO Term	Description	p-value	Genes
0044702	single organism reproductive process	$3.09e^{-4}$	MSH4,SLC6A4,NPFF, SYCP2,SYNGAP1,MYCBPAP
0006836	neurotransmitter transport	$4.25e^{-4}$	RAB3A,SLC6A4,SLC6A13
0001505	regulation of neurotransmitter levels	$4.25e^{-4}$	RAB3A,SLC6A4,SLC6A13
0031644	regulation of neurological system process	$6.6e^{-4}$	RAB3A,SLC6A4,NPFF,SYNGAP1
0023061	signal release	$8.1e^{-4}$	RAB3A,NPFF,SLC6A13
0044057	regulation of system process	$8.23e^{-4}$	INHA,RAB3A,SLC6A4,NPFF, SYNGAP1

Table 3 GO Analysis results of "h-differentially expressed" genes according to RankProd test. Only induced samples presented significant GO Biological Process terms.

GO Term	Description	p-value	Genes
0021675	nerve development	4.57E-5	SALL1, PAX2,NGFR
0021879	forebrain neuron differentiation	$1.1e^{-4}$	SALL1, ERBB4
0021889	olfactory bulb interneuron differentiation	$1.1e^{-4}$	SALL1, ERBB4
0010092	specification of organ identity	$2.2e^{-4}$	$PAX2, PAX3$
0003337	mesenchymal to epithelial transition involved in metanephros morphogenesis	$2.2e^{-4}$	SALL1, PAX2
0021953	central nervous system neuron differentiation	$2.66e^{-4}$	SALL1,ERBB4,PAX3
0060231	mesenchymal to epithelial transition	$3.65e^{-4}$	SALL1, PAX2
0021545	cranial nerve development	$5.45e^{-4}$	SALL1, PAX2

Tables

Additional Files

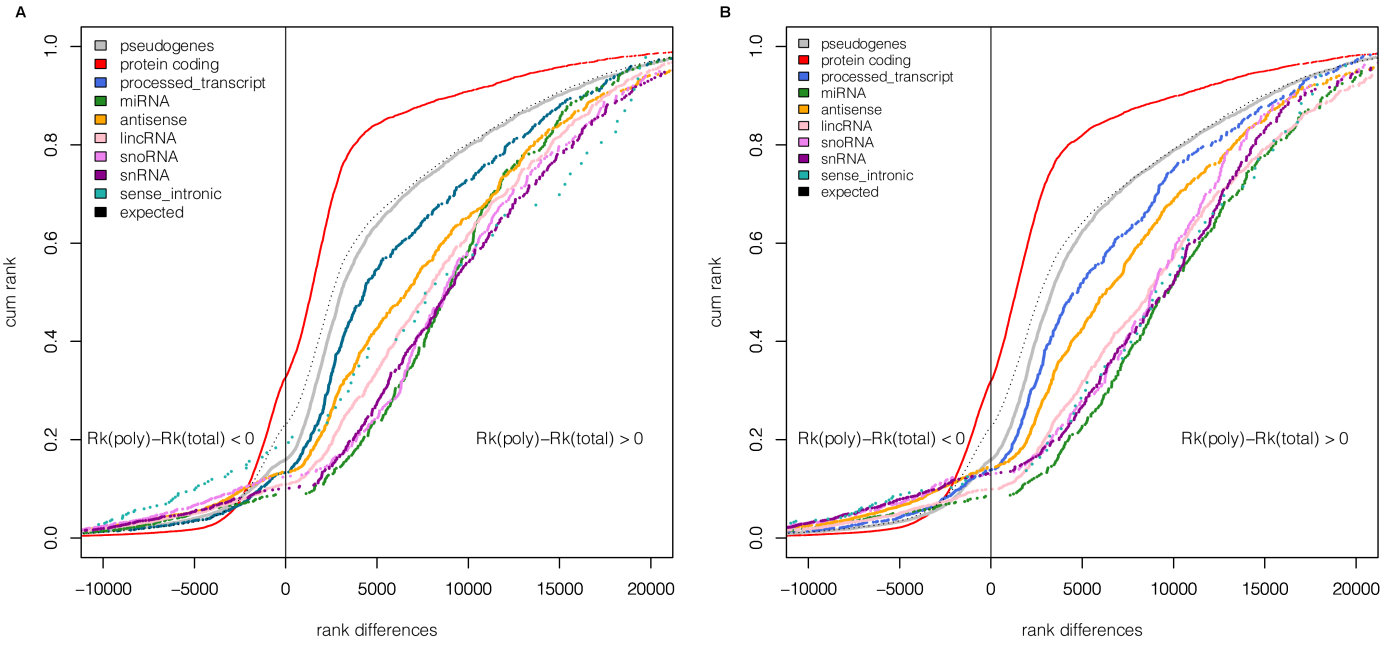
Additional file 1 — Table with values for fisher tests

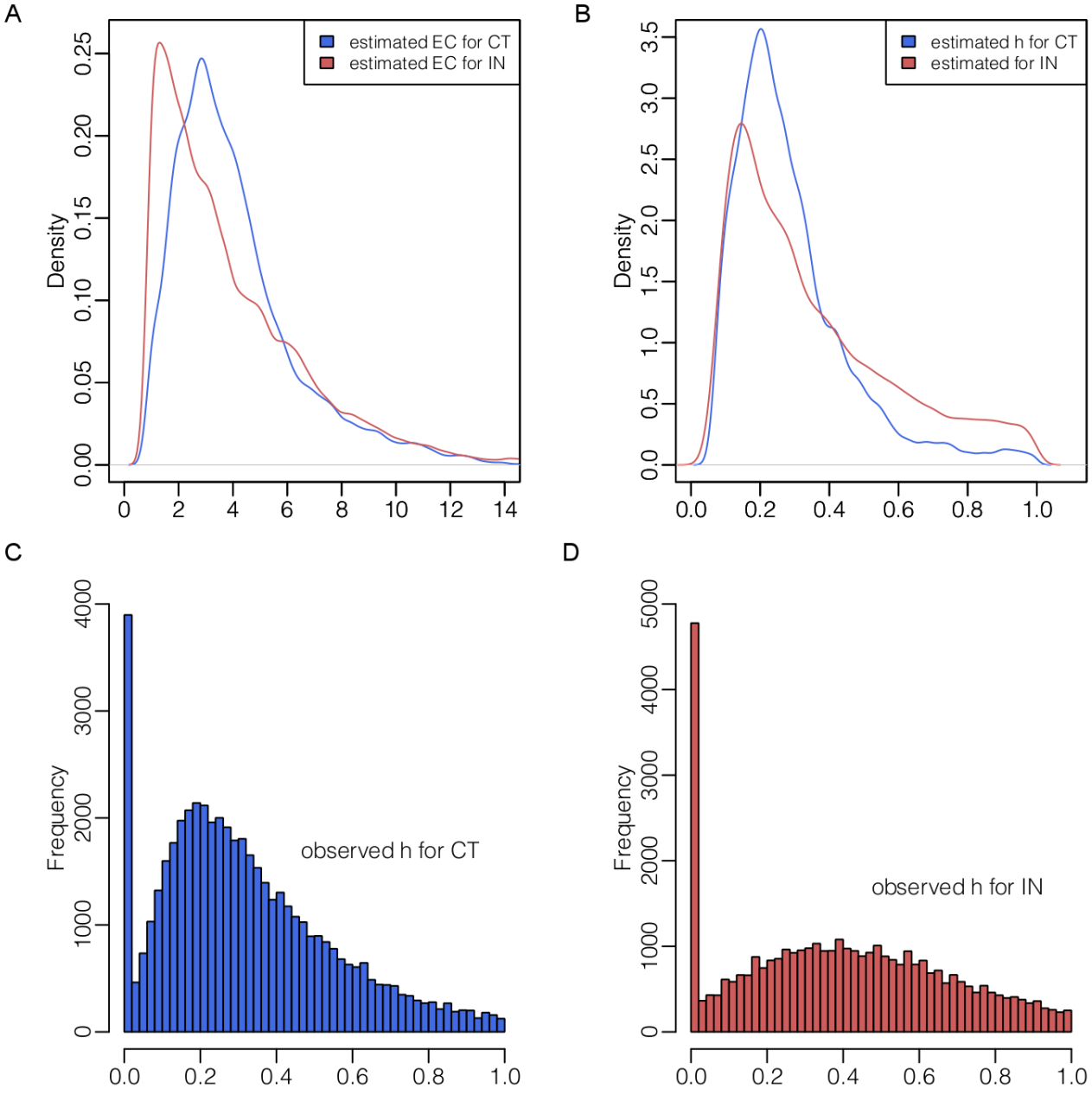
Additional file 2 — Sample additional file title

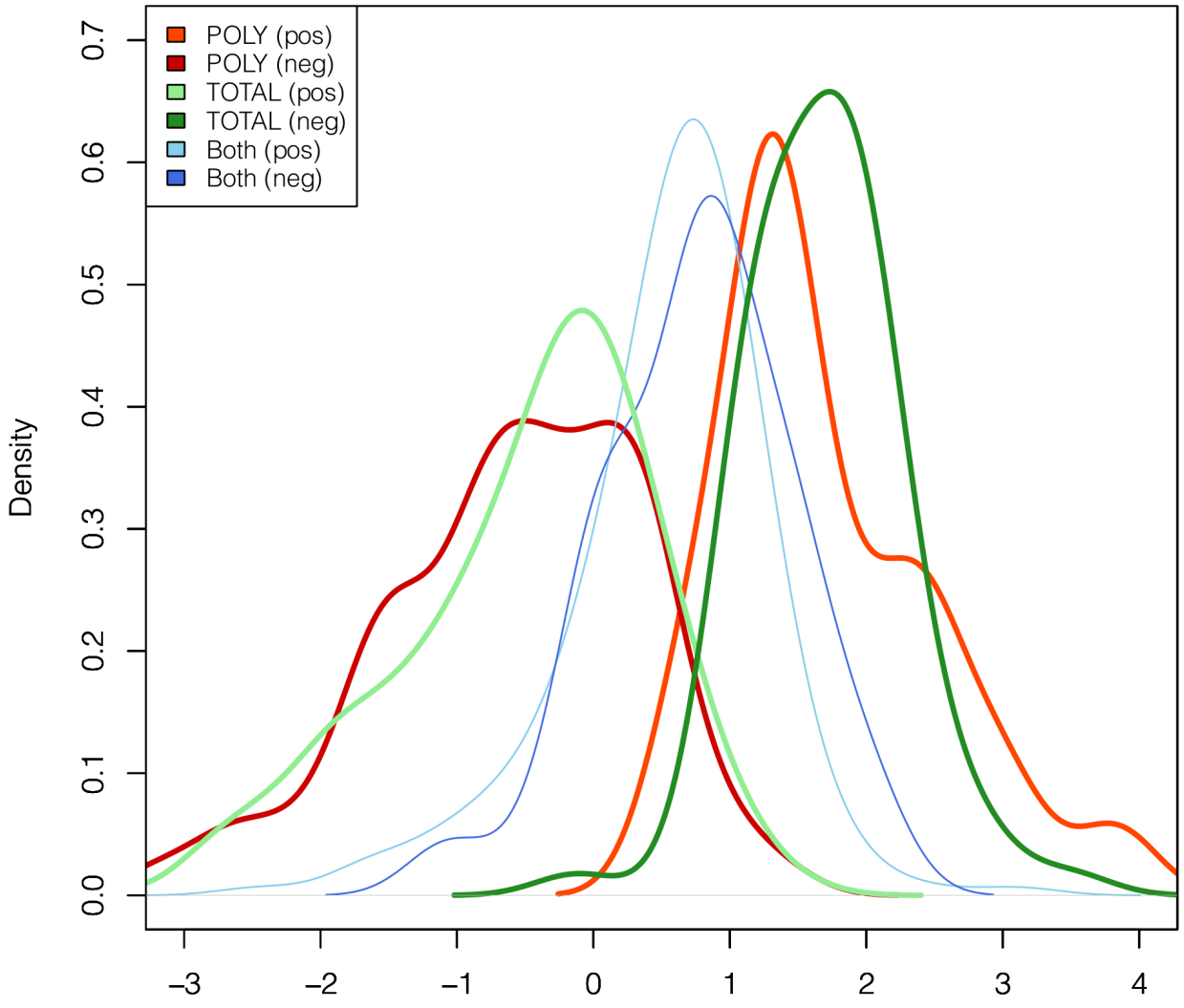
Additional file descriptions text.

Table 4 Wilcoxon Test for the ranking differences of h between IN and CT inside biotypes (BT).

BT	Donor 1t	Donor 2	Donor 3	combined p-value	combined median
protein coding	$2.2e^{-16}$	$2.2e^{-16}$	$2.2e^{-16}$	$6.34e^{-44}$	0.192
pseudogene	$2.2e^{-16}$	$2.2e^{-16}$	$1.95e^{-13}$	$4.95e^{-41}$	0.108
lincRNA	$8.24e^{-7}$	$2.4e^{-5}$	0.295	$2.11e^{-9}$	0.0361
processed transcript	$2.58e^{-7}$	$2.68e^{-8}$	0.0375	$1.77e^{-13}$	0.07788


A
B
pseudogenes
protein coding
processed_transcript
miRNA
antisense
lincRNA
snoRNA
snRNA
sense_intronic
expected
cum rank
rank differences
Rk(poly)−Rk(total) < 0
Rk(poly)−Rk(total) > 0
0.0
0.2
0.4
0.6
0.8
1.0
−10000
−5000
0
5000
10000
15000
20000


A
estimated EC for CT
estimated EC for IN
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
12
14
B
estimated h for CT
estimated for IN
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
C
Frequency
0
1000
2000
3000
4000
observed h for CT
0.0
0.2
0.4
0.6
0.8
1.0
D
Frequency
0
1000
2000
3000
4000
5000
observed h for IN
0.0
0.2
0.4
0.6
0.8
1.0


POLY (pos)
POLY (neg)
TOTAL (pos)
TOTAL (neg)
Both (pos)
Both (neg)
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−3
−2
−1
0
1
2
3
4

Capítulo 8

Conclusiones generales

8.1. Introducción

Durante la tesis se intentaron estudiar los mecanismos de regulación durante la diferenciación de células madre a adipocitos. Se utilizaron datos de RNA-Seq provenientes del SOLiD para analizar el proceso de la adipogénesis, tanto a nivel de expresión génica, como a nivel de regulación transcripcional y post-transcripcional. En cada etapa, con cada trabajo individual, logramos abordar un aspecto particular, permitiéndonos un comprendimiento más a fondo de la temática global. Pudimos comprobar la existencia de regulación post-transcripcional en el proceso estudiado, logrando cuantificar el grado de la misma. Se verificó un compromiso celular a nivel molecular al linaje de adipocitos a los 3 días de la inducción. La regulación post-transcripcional está compuesta, en general, por una orquesta de pequeños mecanismos altamente coordinados e interrelacionados que determinan el destino de la célula. Se propone a la poliadenilación alternativa (APA) como uno de los principales mecanismos de la regulación post-transcripcional. Analizando el uso diferencial de transcriptos, con sus respectivas 3'UTRs y comparando los cambios en niveles de expresión a nivel de ARN mensajero con los cambios a nivel de proteínas se validó el rol de APA en la regulación post-transcripcional.
Por otro lado, se abordó una temática actual relacionada con el posible potencial de los fibroblastos de diferenciación. Varios estudios le han adjudicado capacidad de diferenciación a los mismos, mientras que otros refutan esa teoría. Nuestros análisis de expresión génica a partir de datos de RNA-Seq se unen la postura de estos últimos. Los fibroblastos aparentan ser células ya diferenciadas, sin potencialidad de diferenciación.
A continuación, veremos más al detalle cada una de las conclusiones parciales.

8.2. Regulación post-transcipcional a través de perfiles polisomales

Nuestros resultados muestran que casi un 60 % de los genes experimentan algún tipo de regulación post-transcripcional en el proceso de la adipogénesis a partir de células madre mesenquimales (CMM). Nuestros análisis de expresión génica basados en los transcriptos asociados a polisomas muestran que el fenotipo de adipocitos está ya presente a nivel molecular a los 3 días de la inducción. Si bien el protocolo de inducción fue el estándar, en cuanto al medio de cultivo, tratamiento de células, etc., el tiempo de duración fue acortado: se deja de inducir a las 72 horas. Este tiempo es suficiente para comprometer a la célula a nivel molecular, aún cuando morfológicamente no hay cambios apreciables. Las vías relevantes, como ser la de la insulina se observa sobre-expresada. A su vez, genes relevantes en la adipogénesis como ser PPARγ, KLF15, C/EBPα se encuentran entre los genes diferencialmente expresados. Marcadores de adipogénesis se observan sobre-expresados como ser el FABP4 y el WNT, entre otros. Todo indica que la maquinaria molecular ha llevado a la célula a comprometerse con el linaje de adipocitos.
Cabe destacar que la calidad de los datos de secuenciación fue muy alta, lo que se observa en el clustering jerárquico (dendrograma) presentado en el primer artículo (figura 1), al final del capítulo 4. Con la confianza depositada en los datos, se pudieron sacar conclusiones sobre la existencia de la regulación post-transcripcional al comparar expresión génica en la fracción de ARN total (muestras inducidas vs. controles) con la correspondiente expresión en la fracción polisomal (muestras inducidas vs. controles). Primeramente, se determinan los genes diferencialmente expresados en la condición inducido con respecto a control en cada una de las fracciones por separado, y seguidamente se analizaron las funciones biológicas que poseen estos dos conjuntos de genes. En la figura 3 del mismo artículo se pueden observar las funciones sobre-representadas en cada conjunto, destacándose que si bien un gran porcentaje de genes presenta funciones concordantes en ambos conjuntos (polisomal y total), muchas funciones aparentan ser específicas de la fracción. La actividad inflamatoria, paracrina, aparenta dominar las funciones de los genes diferencialmente expresados en la fracción polisomal, en concordancia con la función de las CMM durante la inducción. El conjunto de genes de la fracción total presenta, sin embargo, funciones relacionadas más con lo estructural. Estas diferencias sugieren la existencia de algún mecanismo post-transcripcional, ya que los genes diferencialmente expresados en total no necesariamente son los mismos que llegan a la fracción polisomal. Algún mecanismo entre la transcripción y la asociación a la máquina traduccional tiene efecto sobre los transcriptos, cambiándo su abundancia entre un conjunto y otro. Ya que se estableció la existencia de estos mecanismos, se quiere determinar el grado de regulación post-transcripcional que se ocurre en la adipogénesis. Se encontró que casi un 60 % de los genes están de alguna forma sometidos a algún tipo de mecanismo de regulación post-transcripcional. Esto se observa sobre todo en la figura 4 del mismo artículo, en donde considerando sólo los genes que presentan algún cambio, el 44 % presentan cambios exclusivamente en la fracción total y 13 % únicamente en polisomal. Se concluye por lo tanto, que el proceso de la adipogénesis presenta un alto grado de regulación post-transcripcional.
Para tener una pista de qué mecanismos pueden estar actuando sobre estos genes, analizamos algunas regiones 3'UTR de éstos. Las 3'UTR de los transcriptos diferencialmente expresados revelan tanto extensiones como acortamientos, sugiriendo que alguno de los mecanismos de

regulación post-transcripcional activos en este proceso, podrían estar actuando a nivel de los elementos regulatorios en estas regiones, funcionando de esta forma como un "fine-tunning" en la adipogénesis. La coordinación de varios de estos mecanismos finos de regulación es vital para un correcto y eficiente funcionamiento del proceso de diferenciación. Por este motivo el completo entendimiento de la regulación post-transcripcional en los procesos de diferenciación con claves para determinar el destino celular.

8.3. Rol de la poliadenilación alternativa en adipogénesis

En el artículo anterior determinamos la existencia de mecanismos de regulación post-transcripcional durante la adipogénesis y detectamos cierta influencia de las 3'UTR en los mismos. En este estudio intentamos investigar más a fondo en esta temática, proponiendo un mecanismo particular de regulación post-transcripcional plausible que aparenta jugar un rol primordial en la regulación post-transcripcional: la poliadenilación alternativa. Este mecanismo genera a partir de un gen, transcriptos con diferentes largos de 3'UTR, haciéndolos sensibles diferencialmente a miRNAs.
A partir del análisis del uso diferencial de transcriptos se observan tendencias generales en los largos de las 3'UTR: en la inducción los genes tienden a utilizar transcriptos con 3'UTR más largas. Si bien la diferencia media entre condiciones (inducido y control) es relativamente baja, 18 bases para la fracción total y 12 para la fracción polisomal, alcanza para generar nuevos sitios de unión a miRNAs. Cabe destacar que esos números son medias, existiendo varios genes con diferencias mucho mayores. Por otro lado, la cantidad de genes con 3'UTRs más largas en inducido que en control corresponde a 6608 comparado con 5931 en la fracción total, siendo esta diferencia en números significativa (Wilcoxon, $p< 1 \times 10^{-8}$). El mismo comportamiento se observa en la fracción polisomal. Por lo tanto, se observó que existen diferencias en la distribución de los largos de las 3'UTR, en general más largo en inducido que en control. Seguidamente, se determina qué genes presentan diferencias de 3'UTR estadísticamente significativas. Para ello analizamos nuestros datos con el estadístico de Cochran-Mantel-Haenszel (CMH), el cual está comentado en la sección de materiales y métodos en el artículo del capítulo 5. CMH determina un valor de tendencia para cada gen, basado en la correlación de Pearson asignando un p-valor correspondiente. Se observa una correlación positiva si existe una tendencia de generar 3'UTR más largas en la inducción y una negativa si existen 3'UTR más largas en control. De un total de 16832 genes analizados (total), 5952 presentan una tendencia negativa, 6675 una positiva y los restantes 4205 ninguna tendencia. 182 genes fueron estadísticamente significativos (FDR$<$ 0,01), presentando 114 un correlación positva y 68 una negativa. Similares resultados se observan en la fracción polisomal. En conclusión, se observa una tendencia modesta pero consistente en utilizar transcriptos alternativos con 3'UTR más largas durante la inducción.
Adicionalmente, se ha observado en otros estudios y en el nuestro que los cambios de expresión a nivel de proteínas no se corresponden directamente con los cambios en el ARN mensajero. Comparamos datos cuantitativos de expresión de proteínas (SILAC) con nuestros datos de RNAseq pudiendo constatar grandes diferencias a nivel del logFC. Para poder explicar la existencia de estas diferencias, las cuales indican mecanismos de regulación post-transcripcional, introducimos modelos lineales que tengan en cuenta el uso diferencial de transcriptos, por lo tanto los largos de las 3'UTR y por consecuencia los diferentes sitios de unión de los miRNAs. Los modelos lineales

incluyen como covariable no solamente al logFC a nivel de mensajero (logFC_{mRNA}) como sería el modelo base, sino que también incluyen el uso diferencial de transcriptos con los miRNAs correspondientes que caen en las 3'UTRs. Se incluyen por tanto, el efecto del uso diferencial de transcripto, es decir las proporciones reales utilizadas de cada transcripto alternativo (más detalles en la figura 6 y en la sección "Alternative transcripts and miRNAs help explain protein fold changes" del artículo adjuntado en el capítulo 5) y la presencia de sitios de unión conocidos para las 3'UTR alternativas. Es decir, cada gen obtiene un vector de presencia/ausencia de sitios de unión de miRNAs, pesados por el uso real de sus transcriptos. Cabe destacar que muchos de los miRNAs tienen varios sitios de unión en la misma 3'UTR, sin embargo se considera en este trabajo simplemente la presencia o ausencia, sin incluir multiplicidad. Se evalúan primeramente los modelos generados al introducir un único miRNA cada vez en el mismo, es decir obteniendo modelos sólo con dos covariables (logFC_{mRNA} y el miRNA). Seguidamente, se evalúan modelos con dos miRNAs, y así sucesivamente hasta 5. Todos estos modelos son evaluados por BIC. Se utilizó esta medida para la selección de modelos ya que la misma penaliza a su vez por número de parámetros. Se podría también haber utilizado otra, como ser el AIC, que también penaliza por el número de parámetros. Los cinco modelos seleccionados por BIC fueron evaluados con los datos, obteniendo resultados interesantes. La inclusión del efecto de los miRNAs pesados por el uso de transcriptos alternativos implicó una mejora significativa en la varianza explicada por el modelo (ver tabla 1 del artículo). En muchos casos, la varianza explicada es más del doble respecto al modelo base, lo que nos indica que la predicción de la variable observada, $\text{logFC}_{proteinas}$ es más precisa al incluir estas nuevas covariables.

El comportamiento de muchos genes no lograba ser explicado por el modelo base, el mismo no predice con precisión a partir del logFC_{mRNA} su $\text{logFC}_{proteinas}$. Sin embargo, al aplicar el nuevo modelo, el comportamiento de muchos de ellos se ve ampliamente mejorado. Casualmente, los genes que fueron más "corridos" por el modelo nuevo (ver figura 3 del artículo) son genes relevantes en la adipogénesis, como ser el FAPB4, LPL, KRT14, etc. Los miRNAs incluídos en el modelo, en este caso miR-130b and miR-558 para la fracción polisomal y miR-150* para la total, están siendo justamente los causantes del mejor ajuste de los datos, por lo que aparentan ser relevantes también en el proceso de diferenciación. Si bien varios de los miRNAs que mejoran la varianza explicada al ser incluídos en el modelo, han sido descritos previamente como relevantes en la adipogénesis, lo cual fortalece la confianza a nuestra metodología, varios de los miRNAs restantes han sido vinculados a procesos de diferenciación en general u a otros procesos, por lo que valdría la pena una investigación más a fondo.

Por otro lado, ya que estamos trabajando con células madre, nos pareció importante analizar la Plurinet. Ésta es un conjunto de genes que han sido vinculados con los procesos de diferenciación de células madre embrionarias y con la manutención del carácter pluripotente de las mismas, a nivel experimental. Esta red consiste en 299 genes de los cuales pudimos recuperar la gran mayoría en nuestro juego de datos. Analizamos la distribución de los largos de las 3'UTRs de estos genes en particular y observamos la misma tendencia general: las 3'UTR tienen una tendencia a ser más largas en el estado inducido (figura 4 del artículo).

Para evaluar la robustez de nuestros análisis y la confiabilidad de nuestra metodología realizamos un análisis de bootstrap, con el fin de corroborar que por azar no se logran estos mismos o mejores resultados. Se permutan los miRNAs y se recalcula el modelo, junto con la varianza explicada por el mismo, repitiendose este procedimiento 1000 veces. Una explicación más detallada

se puede obtener en la sección "Linear model for correlation of microRNAs with protein levels" del artículo. Si el modelo random le "gana" la mayoría de las veces al modelo real, entonces los resultados obtenidos pueden haber sido simplemente por azar. Ya que este no es el caso, podemos afirmar que nuestros resultados fueron consistentes y robustos.
En este estudio hemos encontrado una tendencia modesta pero consistente en los largos de las 3'UTR, observándose extensiones durante la inducción a la adipogénesis. Se han desarrollado modelos lineales que toman en cuenta el efecto de APA obteniendo resultados más consistentes con la realidad. Con los avances en la biología molecular y sobre todo en el área de regulación, los análisis están cambiando su foco de atención: ya no son los genes los principales actores sino que los transcriptos son ahora los protagonistas ("gene-centric vs. transcript-centric approaches").

8.4. El problema de caracterización de los fibroblastos

Los fibroblastos han pasado por un problema de identidad en estos últimos años. En un principio se consideraban células ya diferenciadas, posteriormente se observó que tenían potencial de diferenciación, y estudios más recientes alegaron que esto no era cierto. En este artículo mediante los experimentos de RNA-Seq de muestras de ARN asociadas a polisomas logramos en gran detalle determinar los genes diferencialmente expresados entre ambos tipos celulares, y gracias a ello, ver las discordancias entre estos tipos celulares. Las diferencias observadas a nivel de expresión génica y en los resultados de los anásis de GO para determinar enriquecimiento de funciones biológicas, aportaron evidencia a la hipótesis de que los fibroblastos no poseen el potencial de diferenciación como las células madre. Ensayos de diferenciación en el laboratorio corroboraron estos resultados.
La gran confusión entre ambos tipos celulares surge por la problemática de que ambas presentan los mismos marcadores de superficie (por lo menos identificados hasta ahora), lo que hace muy dificultosa la separación de las mismas con altos grados de pureza. En este estudio encontramos un marcador de superficie (CD105) previamente descrito, que presenta una gran diferencia de expresión entre ambos tipos celulares, por lo que utilizamos el nivel de expresión del mismo para lograr separar con más precisión las dos poblaciones. Utilizando este marcador y su expresión se lograron identificar dos poblaciones celulares, las cuales se diferenciaban en su capacidad de diferenciación.
Se llega a la conclusión de que los fibroblastos son células ya diferenciadas, por tanto, sin capacidad de diferenciación.

8.5. Fracción polisomal contra fracción total: diferencias y congruencias

El objetivo de este estudio fue comparar las fracciones de ARN total y ARN asociado a polisomas, para ambas condiciones estudiadas: control e inducción a adipocitos, con el fin de comprender más a fondo las diferencias sistemáticas entre transcriptoma y translatoma. Nos enfocamos primeramente en el estudio de biotipos basado en diferencia de ranking de expresión. Los sesgos en

las distribuciones brindan información sobre los procesos subyacentes.
Se observan genes codificantes, snoRNAs y ARN "sense intronic" como sobre-representados en diferencias positivas, es decir genes con tendencia a estar sobre-expresados en polisomas. Por otro lado, varios ARNs no codificantes como ser miRNA, lincRNA, seudogenes, entre otros se observan como con baja tendencia a asociarse a polisomas. Estos resultados, que en cierta forma eran esperados, muestran a su vez un conjunto importante de excepciones al paradigma de la biología molecular que serán analizados específicamente en trabajos futuros.
Ya que la verdadera relación polisomal a total se pierde por el proceso de secuenciado, intentamos estimarla (coeficiente de expansión; *CE*) utilizando un modelo estadístico en un contexto bayesiano. A su vez, se estimó para cada gen (ARN mensajero) su propia tasa de asociación al polisoma, denominado h_i *realizado*. Los valores obtenidos de *CE* para CT e IN, indican que en el proceso de inducción los genes en general tienden a asociarse más al polisoma. La inducción ocasiona un "shift" hacia los polisomas, para que los genes relevantes en la diferenciación puedan ser traducidos eficientemente.
Al analizar las diferencias de tasa de asociación (*h realizados*) entre IN y CT se obtienen una serie de genes con altos h's, sobre-expresados en IN, cuyas funciones están relacionadas con la diferenciación, sobre todo de células mesenquimales. Esto apoya la idea de que los genes altamente asociados a polisomas son los relevantes para el proceso de diferenciación.
Finalmente, observamos que el comportamiento de los genes que sólo cambian su expresión exclusivamente en una fracción (ya sea o polisomal o total), puede ser explicado mediante su tasa de asociación al polisoma.

8.6. Conclusión final

En este trabajo de tesis hemos observado que existe un alto grado de regulación post-transcripcional durante el proceso de adipogénesis. Por un lado, hemos encontrado evidencia de diferencias importantes entre el ARNm observado a nivel global y el asociado a los polisomas. Al mismo tiempo, también reportamos diferencias significativas en el largo de las regiones 3'UTR entre las células no diferenciadas y las células inducidas, proponiendo un mecanismo posible en este tipo de regulación, como ser la poliadenilación alternativa. Entre otros aportes marcamos un perfil particular de los fibroblastos, con posibles aplicaciones a su caracterización y separación. Finalmente, comenzamos a detallar los procesos subyacentes a los cambios ocurridos en las tasas de asociación de los mensajeros a los polisomas y su posible implicación en el proceso de diferenciación. En conjunto, creemos haber realizado un modesto aporte, aunque novedoso, al esclarecimiento de los procesos de diferenciación celular.

Bibliografía

[1] Qi Qun Tang and M Daniel Lane. Adipogenesis: From stem cell to adipocyte. *Annual Review of Biochemistry*, 81(March):1–22, 2012. URL `http://www.ncbi.nlm.nih.gov/pubmed/22463691`.

[2] Alan D Radford, David Chapman, Linda Dixon, Julian Chantrey, Alistair C Darby, and Neil Hall. Application of next-generation sequencing technologies in virology. *Journal of General Virology*, pages 1–31, 2012.

[3] J A Thomson, J Itskovitz-Eldor, S S Shapiro, M A Waknitz, J J Swiergiel, V S Marshall, and J M Jones. Embryonic stem cell lines derived from human blastocysts. *Science*, 282 (5391):1145–1147, 1998.

[4] Masako Miura, Stan Gronthos, Mingrui Zhao, Bai Lu, Larry W Fisher, Pamela Gehron Robey, and Songtao Shi. Shed: Stem cells from human exfoliated deciduous teeth. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10): 5807–5812, 2003.

[5] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, 2006.

[6] Kazutoshi Takahashi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–872, 2007.

[7] Junying Yu, Maxim A Vodyanik, Kim Smuga-Otto, Jessica Antosiewicz-Bourget, Jennifer L Frane, Shulan Tian, Jeff Nie, Gudrun A Jonsdottir, Victor Ruotti, Ron Stewart, and et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318 (5858):1917–1920, 2007.

[8] Christopher J Centeno, Dan Busse, John Kisiday, Cristin Keohan, Michael Freeman, and David Karli. Increased knee cartilage volume in degenerative joint disease using percutaneously implanted, autologous mesenchymal stem cells. *Pain Physician*, 11(3):343–353, 2008.

[9] K Hayashi, S Ogushi, S Kurimoto, S Shimamoto, H Ohta, and M Saitou. Offspring from oocytes derived from in vitro primordial germ cell-like cells in mice. *Science*, 338(6109): 971–975, 2012.

[10] Masahito Tachibana, Paula Amato, Michelle Sparman, Nuria Marti Gutierrez, Rebecca Tippner-Hedges, Hong Ma, Eunju Kang, Alimujiang Fulati, Hyo-Sang Lee, Hathaitip Sritanaudomchai, and et al. Human embryonic stem cells derived by somatic cell nuclear transfer. *Cell*, pages 1–11, 2013.

[11] Sean J Morrison and Judith Kimble. Asymmetric and symmetric stem-cell divisions in development and cancer. *Nature*, 441(7097):1068–1074, 2006.

[12] Ryan Lister, Mattia Pelizzola, Yasuyuki S Kida, R David Hawkins, Joseph R Nery, Gary Hon, Jessica Antosiewicz-Bourget, Ronan O?Malley, Rosa Castanon, Sarit Klugman, and et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 471(7336):68–73, 2011.

[13] A J Joannides and S Chandran. Human embryonic stem cells: an experimental and therapeutic resource for neurological disease. *Journal of the neurological sciences*, 265(1-2): 84–88, 2008.

[14] A Varanou, CP Page, and SL Minger. Human embryonic stem cells and lung regeneration. *Br J Pharmacol.*, pages 316–325, 2008.

[15] Xiaoming He, Satoshi Imanishi, Hideko Sone, Reiko Nagano, Xian-Yang Qin, Jun Yoshinaga, Hiromi Akanuma, Junko Yamane, Wataru Fujibuchi, and Seiichiroh Ohsako. Effects of methylmercury exposure on neuronal differentiation of mouse and human embryonic stem cells. *Toxicology Letters*, 212(1):1–10, 2012.

[16] N Maherali, R Sridharan, W Xie, J Utikal, S Eminli, K Arnold, M Stadtfeld, R Yachechko, J Tchieu, R Jaenisch, K Plath, and K. Hochedlinger. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell*, 1 (1):55–70, 2007.

[17] Marius Wernig, Alexander Meissner, Ruth Foreman, Tobias Brambrink, Manching Ku, Konrad Hochedlinger, Bradley E Bernstein, and Rudolf Jaenisch. In vitro reprogramming of fibroblasts into a pluripotent es-cell-like state. *Nature*, 448(7151):318–324, 2007.

[18] Keisuke Okita, Tomoko Ichisaka, and Shinya Yamanaka. Generation of germline-competent induced pluripotent stem cells. *Nature*, 448(7151):313–7, 2007.

[19] Paul S Knoepfler. Deconstructing stem cell tumorigenicity: A roadmap to safe regenerative medicine. *Stem Cells*, 27(5):1050–1056, 2009.

[20] Georgina M Ellison, Daniele Torella, Santo Dellegrottaglie, Claudia Perez-Martinez, Armando Perez De Prado, Carla Vicinanza, Saranya Purushothaman, Valentina Galuppo, Claudio Iaconetti, Cheryl D Waring, and et al. Endogenous cardiac stem cell activation by insulin-like growth factor-1/hepatocyte growth factor intracoronary injection fosters survival and regeneration of the infarcted pig heart. *Journal of the American College of Cardiology*, 58(9):977–986, 2011.

[21] Fang-Xu Jiang and Grant Morahan. Pancreatic stem cells: From possible to probable. *Stem Cell Reviews*, 2011.

[22] Nick Barker, Johan H Van Es, Jeroen Kuipers, Pekka Kujala, Maaike Van Den Born, Miranda Cozijnsen, Andrea Haegebarth, Jeroen Korving, Harry Begthel, Peter J Peters, and et al. Identification of stem cells in small intestine and colon by marker gene lgr5. *Nature*, 449(7165):1003–1007, 2007.

[23] Marie Maumus, Christian Jorgensen, and Daniéle Noël. Mesenchymal stem cells in regenerative medicine applied to rheumatic diseases: role of secretome and exosomes. *Biochimie*, pages 1–6, 2013.

[24] Lawrence V Gulotta, Salma Chaudhury, and Daniel Wiznia. Stem cells for augmenting tendon repair. *Stem cells international*, 2012.

[25] SunMI Palumbo and Wan-Ju Li. Osteoprotegerin enhances osteogenesis of human mesenchymal stem cells. *Tissue Eng Part A.*, 2013.

[26] Shihua Wang, Xuebin Qu, and Robert Chunhua Zhao. Clinical applications of mesenchymal stem cells. *Journal of hematology oncology*, 5(1):19–28, 2012.

[27] Arnold I Caplan. All mscs are pericytes? *Cell stem cell*, 3(3):229–230, 2008.

[28] Mihaela Crisan, Solomon Yap, Louis Casteilla, Chien-Wen Chen, Mirko Corselli, Tea Soon Park, Gabriella Andriolo, Bin Sun, Bo Zheng, Li Zhang, and et al. A perivascular origin for mesenchymal stem cells in multiple human organs. *Cell stem cell*, 3(3):301–313, 2008.

[29] Arnold I Caplan and James E Dennis. Mesenchymal stem cells as trophic mediators. *Journal of Cellular Biochemistry*, 98(5):1076–1084, 2006.

[30] Arnold I Caplan and Diego Correa. The msc: an injury drugstore. *Cell stem cell*, 9(1): 11–15, 2011.

[31] Smita S Iyer and Mauricio Rojas. Anti-inflammatory effects of mesenchymal stem cells: novel concept for future therapies. *Expert Opinion on Biological Therapy*, 8(5):569–581, 2008.

[32] Ben J Jones and Steven J McTaggart. Immunosuppression by mesenchymal stromal cells: from culture to clinic. *Experimental Hematology*, 36(6):733–741, 2008.

[33] Nora G Singer and Arnold I Caplan. Mesenchymal stem cells: mechanisms of inflammation. *Annual Review of Pathology*, 6(November 2010):457–478, 2011. URL `http://www.ncbi.nlm.nih.gov/pubmed/21073342`.

[34] Lindolfo Da Silva Meirelles, Arnold I Caplan, and Nance Beyer Nardi. In search of the in vivo identity of mesenchymal stem cells. *Stem Cells*, 26(9):2287–2299, 2008.

[35] Coralie Sengenés, Karine Lolméde, Alexia Zakaroff-Girard, Rudi Busse, and Anne Bouloumié. Preadipocytes in the human subcutaneous adipose tissue display distinct features from the adult mesenchymal and hematopoietic stem cells. *Journal of Cellular Physiology*, 205(1):114–122, 2005.

[36] Kirsty L Spalding, Erik Arner, Pål O Westermark, Samuel Bernard, Bruce A Buchholz, Olaf Bergmann, Lennart Blomqvist, Johan Hoffstedt, Erik Näslund, Tom Britton, and et al. Dynamics of fat cell turnover in humans. *Nature*, 453(7196):783–787, 2008.

[37] Yazmín Macotela, Brice Emanuelli, Marcelo A Mori, Stephane Gesta, Tim J Schulz, Yu-Hua Tseng, and C Ronald Kahn. Intrinsic differences in adipocyte precursor cells from different white fat depots. *Diabetes*, 61(7):1691–9, 2012.

[38] J J Egan, A S Greenberg, M K Chang, S A Wek, M C Moos, and C Londos. Mechanism of hormone-stimulated lipolysis in adipocytes: translocation of hormone-sensitive lipase to the lipid storage droplet. *Proceedings of the National Academy of Sciences of the United States of America*, 89(18):8537–8541, 1992.

[39] Maryam Ahmadian, Marcia J Abbott, Tianyi Tang, Carolyn S S Hudak, Yangha Kim, Matthew Bruss, Marc K Hellerstein, Hui-Young Lee, Varman T Samuel, Gerald I Shulman, Yuhui Wang, Robin E Duncan, Chulho Kang, and Hei Sook Sul. Desnutrin/atgl is regulated by ampk and is required for a brown adipose phenotype. *Cell Metabolism*, 13(6):739–748, 2011.

[40] Maryam Ahmadian, Yuhui Wang, and Hei Sook Sul. Lipolysis in adipocytes. *The international journal of biochemistry cell biology*, 42(5):555–559, 2010.

[41] C Londos, D L Brasaemle, C J Schultz, D C Adler-Wailes, D M Levin, A R Kimmel, and C M Rondinone. On the control of lipolysis in adipocytes. *Annals Of The New York Academy Of Sciences*, 892:155–168, 1999.

[42] C Londos, J Gruia-Gray, D L Brasaemle, C M Rondinone, T Takeda, N K Dwyer, T Barber, A R Kimmel, and E J Blanchette-Mackie. Perilipin: possible roles in structure and metabolism of intracellular neutral lipids in adipocytes and steroidogenic cells. *International journal of obesity and related metabolic disorders journal of the International Association for the Study of Obesity*, 20 Suppl 3(3):S97–S101, 1996.

[43] Dimas T Covas, Rodrigo A Panepucci, Aparecida M Fontes, Wilson A Silva, Maristela D Orellana, Marcela C C Freitas, Luciano Neder, Anemari R D Santos, Luiz C Peres, Maria C Jamur, and Marco A Zago. Multipotent mesenchymal stromal cells obtained from diverse human tissues share functional properties and gene-expression profile with cd146+ perivascular cells and fibroblasts. *Experimental Hematology*, 36(5):642–54, 2008.

[44] Haiyan Huang, Tan-Jing Song, Xi Li, Lingling Hu, Qun He, Mei Liu, M Daniel Lane, and Qi-Qun Tang. Bmp signaling pathway is required for commitment of c3h10t1/2 pluripotent stem cells to the adipocyte lineage. *Proceedings of the National Academy of Sciences of the United States of America*, 106(31):12670–12675, 2009.

[45] Ching-Shwun Lin, Zhong-Cheng Xin, Chun-Hua Deng, Hongxiu Ning, Guiting Lin, and Tom F Lue. Defining adipose tissue-derived stem cells in tissue and in culture. *Histology and histopathology*, 25(6):807–15, 2010.

[46] P R Shepherd, L Gnudi, E Tozzo, H Yang, F Leach, and B B Kahn. Adipose cell hyperplasia and enhanced glucose disposal in transgenic mice overexpressing glut4 selectively in adipose tissue. *The Journal of Biological Chemistry*, 268(30):22243–22246, 1993.

[47] Robert R Bowers, Jae Woo Kim, Tamara C Otto, and M Daniel Lane. Stable stem cell commitment to the adipocyte lineage by inhibition of dna methylation: Role of the bmp-4 gene. *Proceedings of the National Academy of Sciences of the United States of America*, 103(35):13022–13027, 2006.

[48] Qi-Qun Tang, Tamara C Otto, and M Daniel Lane. Commitment of c3h10t1/2 pluripotent stem cells to the adipocyte lineage. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9607–11, 2004.

[49] S E Ross, N Hemati, K A Longo, C N Bennett, P C Lucas, R L Erickson, and O A MacDougald. Inhibition of adipogenesis by wnt signaling. *Science*, 289(5481):950–953, 2000.

[50] Robert R Bowers and M Daniel Lane. Wnt signaling and adipocyte lineage commitment. *Cell cycle Georgetown Tex*, 7(9):1191–1196, 2008.

[51] B K Zehentner, U Leser, and H Burtscher. Bmp-2 and sonic hedgehog have contrary effects on adipocyte-like differentiation of c3h10t1/2 cells. *DNA and Cell Biology*, 19(5):275–281, 2000.

[52] S Spinella-Jaegle, G Rawadi, S Kawai, S Gallea, C Faucheu, P Mollat, B Courtois, B Bergaud, V Ramez, A M Blanchet, and et al. Sonic hedgehog increases the commitment of pluripotent mesenchymal cells into the osteoblastic lineage and abolishes adipocytic differentiation. *Journal of Cell Science*, 114(Pt 11):2085–94, 2001.

[53] Geertje Van Der Horst, Hetty Farih-Sips, Clemens W G M Löwik, and Marcel Karperien. Hedgehog stimulates only osteoblastic differentiation of undifferentiated ks483 cells. *Bone*, 33(6):899–910, 2003.

[54] Sona Kang, Christina N Bennett, Isabelle Gerin, Lauren A Rapp, Kurt D Hankenson, and Ormond A Macdougald. Wnt signaling stimulates osteoblastogenesis of mesenchymal precursors by suppressing ccaat/enhancer-binding protein alpha and peroxisome proliferator-activated receptor gamma. *The Journal of Biological Chemistry*, 282(19):14515–24, 2007.

[55] Christina N Bennett, Kenneth A Longo, Wendy S Wright, Larry J Suva, Timothy F Lane, Kurt D Hankenson, and Ormond A MacDougald. Regulation of osteoblastogenesis and bone mass by wnt10b. *Proceedings of the National Academy of Sciences of the United States of America*, 102(9):3324–3329, 2005.

[56] Yucheng XU, Yan Larry Zhou, Robin L Erickson, Ormond A Macdougald, and Malcolm L Snead. Physical dissection of the ccaat/enhancer-binding protein alpha in regulating the mouse amelogenin gene. *Biochemical and Biophysical Research Communications*, 354(1): 56–61, 2007.

[57] C Heldin, K Miyazono, and P ten Dijke. Tgf-beta signalling from cell membrane to nucleus through smad proteins. *Nature*, 390:465–471, 1997.

[58] HY Huang, LL Hu, TJ Song, X Li, and Q He. Involvement of cytoskeleton-associated proteins in the commitment of c3h10t1/2 pluripotent stem cells to adipocyte lineage induced by bmp2/4. *Mol. Cell. Proteomics*, 10:1–8, 2011.

[59] Yuko Komiya and Raymond Habas. Wnt signal transduction pathways. *Organogenesis*, 4 (2):68–75, 2008.

[60] T Reya and H Clevers. Wnt signalling in stem cells and cancer. *Nature*, 434:843–850, 2005.

[61] Constantinos Christodoulides, Claire Lagathu, Jaswinder K Sethi, and Antonio Vidal-Puig. Adipogenesis and wnt signalling. *Trends in endocrinology and metabolism TEM*, 20 (1):16–24, 2009.

[62] Ken M Cadigan and Yan I Liu. Wnt signaling: complexity at the surface. *Journal of Cell Science*, 119(Pt 3):395–402, 2006.

[63] Akira Kikuchi, Hideki Yamamoto, and Shosei Kishida. Multiplicity of the interactions of wnt proteins and their receptors. *Cellular Signalling*, 19(4):659–671, 2007.

[64] L A Davis and N I Zur Nieden. Mesodermal fate decisions of a stem cell: the wnt switch. *Cellular and Molecular Life Sciences*, 65(17):2658–2674, 2008.

[65] Jennifer A Kennell, Erin E O'Leary, Brian M Gummow, Gary D Hammer, and Ormond A MacDougald. T-cell factor 4n (tcf-4n), a novel isoform of mouse tcf-4, synergizes with beta-catenin to coactivate c/ebpalpha and steroidogenic factor 1 transcription factors. *Molecular and Cellular Biology*, 23(15):5366–5375, 2003.

[66] M Michael Cohen. The hedgehog signaling network. *American journal of medical genetics Part A*, 123A(1):5–28, 2003.

[67] Aaron W James, Philipp Leucht, Benjamin Levi, Antoine L Carre, Yue Xu, Jill A Helms, and Michael T Longaker. Sonic hedgehog influences the balance of osteogenesis and adipogenesis in mouse adipose-derived stromal cells. *Tissue engineering Part A*, 16(8):2605–2616, 2010.

[68] M Bamshad, C K Song, and T J Bartness. Cns origins of the sympathetic nervous system outflow to brown adipose tissue. *American Journal of Physiology*, 276(6 Pt 2):R1569–R1578, 1999.

[69] Deborah L Burkhart and Julien Sage. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nature Reviews Cancer*, 8(9):671–82, 2008.

[70] QQ Tang and MD Lane. Activation and centromeric localization of ccaat/enhancer bindingproteins during the mitotic clonal expansion of adipocyte differentiation. *Genes Dev.*, 13:2231–2241, 1999.

[71] Michael Korenjak and Alexander Brehm. E2f-rb complexes regulating transcription of genes important for differentiation and development. *Current opinion in genetics development*, 15(5):520–527, 2005.

[72] D M Thomas, S A Carty, D M Piscopo, J S Lee, W F Wang, W C Forrester, and P W Hinds. The retinoblastoma protein acts as a transcriptional coactivator required for osteogenic differentiation. *Molecular Cell*, 8(2):303–316, 2001.

[73] Lluis Fajas, Rebecca L Landsberg, Yolande Huss-Garcia, Claude Sardet, Jacqueline A Lees, and Johan Auwerx. E2fs regulate adipocyte differentiation. *Developmental Cell*, 3 (1):39–49, 2002.

[74] L Fajas, V Egler, R Reiter, J Hansen, K Kristiansen, M Debril, S Miard, and J Auwerx. The retinoblastoma-histone deacetylase 3 complex inhibits ppar? and adipocyte differentiation. *Developmental Cell*, 3(6):903–910, 2002.

[75] A K Student, R Y Hsu, and M D Lane. Induction of fatty acid synthetase synthesis in differentiating 3t3-l1 preadipocytes. *The Journal of Biological Chemistry*, 255(10):4745–50, 1980.

[76] L A Davis and N I Zur Nieden. Mesodermal fate decisions of a stem cell: the wnt switch. *Cellular and Molecular Life Sciences*, 65(17):2658–2674, 2008.

[77] O A MacDougald and M D Lane. Transcriptional regulation of gene expression during adipocyte differentiation. *Annual Review of Biochemistry*, 64(1):345–373, 1995.

[78] B C Reed and M D Lane. Insulin receptor synthesis and turnover in differentiating 3t3-l1 preadipocytes. *Proceedings of the National Academy of Sciences of the United States of America*, 77(1):285–289, 1980.

[79] K H Kaestner, R J Christy, J C McLenithan, L T Braiterman, P Cornelius, P H Pekala, and M D Lane. Sequence, tissue distribution, and differential expression of mrna for a putative insulin-responsive glucose transporter in mouse 3t3-l1 adipocytes. *Proceedings of the National Academy of Sciences of the United States of America*, 86(9):3150–4, 1989.

[80] C S Hwang, T M Loftus, S Mandrup, and M D Lane. Adipocyte differentiation and leptin expression. *Annual Review of Cell and Developmental Biology*, 13:231–259, 1997.

[81] Y M Patel and M D Lane. Mitotic clonal expansion during preadipocyte differentiation: calpain-mediated turnover of p27. *The Journal of Biological Chemistry*, 275(23):17653–17660, 2000.

[82] Jiang-Wen Zhang, Qi-Qun Tang, Charles Vinson, and M Daniel Lane. Dominant-negative c/ebp disrupts mitotic clonal expansion and differentiation of 3t3-l1 preadipocytes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(1):43–47, 2004.

[83] You-You Zhang, Xi Li, Shu-Wen Qian, Liang Guo, Hai-Yan Huang, Qun He, Yuan Liu, Chun-Gu Ma, and Qi-Qun Tang. Transcriptional activation of histone h4 by c/ebp? during

the mitotic clonal expansion of 3t3-l1 adipocyte differentiation. *Molecular Biology of the Cell*, 22(13):2165–2174, 2011.

[84] T Tanaka, N Yoshida, T Kishimoto, and S Akira. Defective adipocyte differentiation in mice lacking the c/ebpbeta and/or c/ebpdelta gene. *the The European Molecular Biology Organization Journal*, 16(24):7432–7443, 1997.

[85] Qi-Qun Tang, Mads Grønborg, Haiyan Huang, Jae-Woo Kim, Tamara C Otto, Akhilesh Pandey, and M Daniel Lane. Sequential phosphorylation of ccaat enhancer-binding protein β by mapk and glycogen synthase kinase 3β is required for adipogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 102(28):9766–9771, 2005.

[86] Victoria A Payne, Wo-Shing Au, Christopher E Lowe, Shaikh M Rahman, Jacob E Friedman, Stephen O?Rahilly, and Justin J Rochford. C/ebp transcription factors regulate srebp1c gene expression during adipogenesis. *The Biochemical journal*, 425(1):215–223, 2010.

[87] Ursula A White and Jacqueline M Stephens. Transcriptional factors that promote formation of white adipose tissue. *Molecular and Cellular Endocrinology*, 318(1-2):10–14, 2010.

[88] Martina I Lefterova, Yong Zhang, David J Steger, Michael Schupp, Jonathan Schug, Ana Cristancho, Dan Feng, David Zhuo, Christian J Stoeckert, X Shirley Liu, and et al. Ppargamma and c/ebp factors orchestrate adipocyte biology via adjacent binding on a genome-wide scale. *Genes & Development*, 22(21):2941–2952, 2008.

[89] Ronni Nielsen, Thomas Askov Pedersen, Dik Hagenbeek, Panagiotis Moulos, Rasmus Siersbaek, Eva Megens, Sergei Denissov, Michael Børgesen, Kees-Jan Francoijs, Susanne Mandrup, and et al. Genome-wide profiling of ppargamma:rxr and rna polymerase ii occupancy reveals temporal activation of distinct metabolic pathways and changes in rxr dimer composition during adipogenesis. *Genes & Development*, 22:2953–2967, 2008.

[90] Z Chen, J Torrens, A Anand, B Spiegelman, and J Friedman. Krox20 stimulates adipogenesis via c/ebp[beta]-dependent and -independent mechanisms. *Cell Metab.*, 1:93–106, 2005.

[91] Maria A Jimenez, Peter Akerblad, Mikael Sigvardsson, and Evan D Rosen. Critical role for ebf1 and ebf2 in the adipogenic transcriptional cascade. *Molecular and Cellular Biology*, 27(2):743–757, 2007.

[92] Kivanç Birsoy, Zhu Chen, and Jeffrey Friedman. Transcriptional regulation of adipogenesis by klf4. *Cell Metabolism*, 7(4):339–347, 2008.

[93] K E Davis, M Moldes, and S R Farmer. The forkhead transcription factor foxc2 inhibits white adipocyte differentiation. *J. Biol. Chem*, 279(41):42453–42461, 2004.

[94] Justin J Rochford, Robert K Semple, Matthias Laudes, Keith B Boyle, Constantinos Christodoulides, Claire Mulligan, Christopher J Lelliott, Sven Schinner, Dirk Hadaschik, Meera Mahadevan, and et al. Eto/mtg8 is an inhibitor of c/ebp? activity and a regulator of early adipogenesis. *Molecular and Cellular Biology*, 24(22):9863–9872, 2004.

[95] Qiang Tong, Judy Tsai, and Guo Tan. Interaction between gata and the c / ebp family of transcription factors is critical in gata-mediated suppression of adipocyte differentiation. *Society*, 25(2):706–715, 2005.

[96] Nancy Sue, Briony H A Jack, Sally A Eaton, Richard C M Pearson, Alister P W Funnell, Jeremy Turner, Robert Czolij, Gareth Denyer, Shisan Bao, Juan Carlos Molero-Navajas, and et al. Targeted disruption of the basic krüppel-like factor gene (klf3) reveals a role in adipogenesis. *Molecular and Cellular Biology*, 28(12):3967–3978, 2008.

[97] Briony H A Jack and Merlin Crossley. Gata proteins work together with friend of gata (fog) and c-terminal binding protein (ctbp) co-regulators to control adipogenesis. *The Journal of Biological Chemistry*, 285(42):32405–32414, 2010.

[98] Jun Eguchi, Qing-Wu Yan, Dustin E Schones, Michael Kamal, Chung-Hsin Hsu, Michael Q Zhang, Gregory E Crawford, and Evan D Rosen. Interferon regulatory factors are transcriptional regulators of adipogenesis. *Cell Metabolism*, 7(1):86–94, 2008.

[99] Luoping Li, Xin Xie, Jun Qin, George S Jeha, Pradip K Saha, Jun Yan, Claire M Haueter, Lawrence Chan, Sophia Y Tsai, and Ming-Jer Tsai. The nuclear orphan receptor coup-tfii plays an essential role in adipogenesis, glucose homeostasis, and energy metabolism. *Cell Metabolism*, 9(1):77–87, 2009.

[100] M Okamura, H Kudo, K I Wakabayashi, T Tanaka, A Nonaka, A Uchida, S Tsutsumi, I Sakakibara, M Naito, T F Osborne, and et al. Coup-tfii acts downstream of wnt/beta-catenin signal to silence ppargamma gene expression and repress adipogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14):5819–5824, 2009.

[101] Enxuan Jing, Stephane Gesta, and C Ronald Kahn. Sirt2 regulates adipocyte differentiation through foxo1 acetylation/deacetylation. *Cell Metabolism*, 6(2):105–114, 2007.

[102] Frédéric Picard, Martin Kurtev, Namjin Chung, Acharawan Topark-Ngarm, Thanaset Senawong, Rita Machado De Oliveira, Mark Leid, Michael W McBurney, and Leonard Guarente. Sirt1 promotes fat mobilization in white adipocytes by repressing ppar-gamma. *Nature*, 429(6993):771–776, 2004.

[103] Toshimasa Itoh, Louise Fairall, Kush Amin, Yuka Inaba, Attila Szanto, Balint L Balint, Laszlo Nagy, Keiko Yamamoto, and John W R Schwabe. Structural basis for the activation of ppar? by oxidized fatty acids. *Nature Structural Molecular Biology*, 15(9):924–931, 2008.

[104] Peter Tontonoz and Bruce M Spiegelman. Fat and beyond: the diverse biology of ppargamma. *Annual Review of Biochemistry*, 77(1):289–312, 2008.

[105] Tamotsu Tsukahara, Ryoko Tsukahara, Yuko Fujiwara, Junming Yue, Yunhui Cheng, Huazhang Guo, Alyssa Bolen, Chunxiang Zhang, Louisa Balazs, Fabio Re, and et al. Phospholipase d2-dependent inhibition of the nuclear hormone receptor ppargamma by cyclic phosphatidic acid. *Molecular Cell*, 39(3):421–432, 2010.

[106] Katja Helenius, Ying Yang, Jukka Alasaari, and Tomi P Mäkelä. Mat1 inhibits peroxisome proliferator-activated receptor gamma-mediated adipocyte differentiation. *Molecular and Cellular Biology*, 29(2):315–323, 2009.

[107] Benedetto Grimaldi, Marina Maria Bellet, Sayako Katada, Giuseppe Astarita, Jun Hirayama, Rajesh H Amin, James G Granneman, Daniele Piomelli, Todd Leff, and Paolo Sassone-Corsi. Per2 controls lipid metabolism by direct regulation of ppar?. *Cell Metabolism*, 12(5):509–520, 2010.

[108] Jennifer A Kennell, Isabelle Gerin, Ormond A MacDougald, and Ken M Cadigan. The microrna mir-8 is a conserved negative regulator of wnt signaling. *Proceedings of the National Academy of Sciences of the United States of America*, 105(40):15417–15422, 2008.

[109] Qiang Wang, Yan Chun Li, Jinhua Wang, Juan Kong, Yuchen Qi, Richard J Quigg, and Xinmin Li. mir-17-92 cluster accelerates adipocyte differentiation by negatively regulating tumor-suppressor rb2/p130. *Proceedings of the National Academy of Sciences of the United States of America*, 105(8):2889–2894, 2008.

[110] Sang Yun Kim, A Young Kim, Hyun Woo Lee, You Hwa Son, Gha Young Lee, Joo-Won Lee, Yun Sok Lee, and Jae Bum Kim. mir-27a is a negative regulator of adipocyte differentiation via suppressing ppargamma expression. *Biochemical and Biophysical Research Communications*, 392(3):323–328, 2010.

[111] Tingwan Sun, Mingui Fu, Angie L Bookout, Steven A Kliewer, and David J Mangelsdorf. Microrna let-7 regulates 3t3-l1 adipogenesis. *Molecular endocrinology Baltimore Md*, 23 (6):925–931, 2009.

[112] Rong Zhang, Di Wang, Zhuying Xia, Chao Chen, Peng Cheng, Hui Xie, and Xianghang Luo. The role of microRNAs in adipocyte differentiation. *Frontiers of medicine*, 7:223–230, 2013.

[113] B Ewing, L Hillier, M C Wendl, and P Green. Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Research*, 8(3):175–185, 1998.

[114] Hege G Russnes, Nicholas Navin, and James Hicks. Review series insight into the heterogeneity of breast cancer through next-generation sequencing. *Journal of Clinical Investigation*, 121(10):3810–8, 2011.

[115] Roland Wittler and Cedric Chauve. Consistency-based detection of potential tumor-specific deletions in matched normal/tumor genomes. *BMC Bioinformatics*, 12 Suppl 9(Suppl 9):S21, 2011.

[116] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Genome Biology*, 458(5):719–724, 2008.

[117] T Matsukawa, H Ishiura, Y Fukuda, Y Ichikawa, H Date, B Ahsan, Y Nakahara, Y Momose, Y Takahashi, A Iwata, J Goto, Y Yamamoto, M Komata, K Shirahige, K Hara, A Kakita, M Yamada, H Takahashi, O Onodera, M Nishizawa, H Takashima, R Kuwano,

H Watanabe, M Ito, Sobue Gen, H Soma, I Yabe, H Sasaki, M Aoki, K Ishikawa, H Mizusawa, K Kanai, T Hattori, S Kuwabara, K Arai, S Koyano, Y Kuroiwa, K Hasegawa, T Yuasa, K Yasui, K Nakashima, H Ito, MV Hananosato, Y Izumi, R Kaji, T Kato, S Kusunoki, Y Osaki, M Horiuchi, T Kondo, S Murayama, N Hattori, M Yamamoto, M Murata, W Satake, T Toda, A Dürr, A Brice, A Filla, T Klockgether, U Wüllner, G Nicholson, S Gilman, CW Shults, CM Tanner, WA Kukull, VM Lee, U Parkinson, E Masliah, PA Low, P Sandroni, JQ Trojanowski, U Parkinson, L Ozelius, T Foroud, and S. Tsuji. Mutations in coq2 in familial and sporadic multiple-system atrophy. *N Engl J Med*, 369(3):233–244, 2013.

[118] Adrian Cortes, Judith Field, Evgeny A Glazov, Johanna Hadler, Jim Stankovich, and Matthew A Brown. Resequencing and fine-mapping of the chromosome 12q13-14 locus associated with multiple sclerosis refines the number of implicated genes. *Human molecular genetics*, 22(11):2283–2292, 2013.

[119] Shima Khoshraftar, Stacy Hung, Sadia Khan, Yunchen Gong, Vibha Tyagi, John Parkinson, Mohini Sain, Alan M Moses, and Dinesh Christendat. Sequencing and annotation of the ophiostoma ulmi genome. *BMC genomics*, 14(1):162, 2013.

[120] CU Köser, JM Bryant, J Becq, ME Török, MJ Ellington, MA Marti-Renom, AJ Carmichael, J Parkhill, GP Smith, and SJ. Peacock. Whole-genome sequencing for rapid susceptibility testing of m. tuberculosis. *N. Engl. J. Med.*, 396(3):290–292, 2013.

[121] H Beltran, R Yelensky, G M Frampton, K Park, S R Downing, T Y Macdonald, M Jarosz, D Lipson, S T Tagawa, D M Nanus, and et al. Targeted next-generation sequencing of advanced prostate cancer identifies potential therapeutic targets and disease heterogeneity. *Eur Urol*, 2012.

[122] X Wang, H Wang, V Sun, HF Tuan, V Keser, K Wang, H Ren, I Lopez, JE Zaneveld, S Siddiqui, S Bowles, A Khan, J Salvo, SG Jacobson, A Iannaccone, F Wang, D Birch, JR Heckenlively, GA Fishman, EI Traboulsi, Y Li, D Wheaton, RK Koenekoop, and R Chen. Comprehensive molecular diagnosis of 179 leber congenital amaurosis and juvenile retinitis pigmentosa patients by targeted next generation sequencing. *J. Med. Genet.*, 2013.

[123] Z Chen, J L Wang, B S Tang, Z F Sun, Y T Shi, L Shen, L F Lei, X M Wei, J J Xiao, Z M Hu, and et al. Using next-generation sequencing as a genetic diagnostic tool in rare autosomal recessive neurologic mendelian disorders. *Neurobiol Aging*, 2013.

[124] N A Twine, C Janitz, M R Wilkins, and M Janitz. Sequencing of hippocampal and cerebellar transcriptomes provides new insights into the complexity of gene regulation in the human brain. *Neurosci Lett*, 2013.

[125] Lucía Spangenberg, Patricia Shigunov, Ana Paula R Abud, Axel R Cofré, Marco A Stimamiglio, Crisciele Kuligovski, Jaiesa Zych, Andressa V Schittini, Alexandre Dias Tavares Costa, Carmen K Rebelatto, PR Brofman, Samuel Goldenberg, Alejandro Correa, Hugo Naya, and Bruno Dallagiovanna. Polysome profiling shows extensive posttranscriptional regulation during human adipocyte stem cell differentiation into adipocytes. *Stem cell research*, 11(2):902–912, 2013.

[126] G Kolle, J L Shepherd, B Gardiner, K S Kassahn, N Cloonan, D L A Wood, E Nourbakhsh, D F Taylor, S Wani, H S Chy, and et al. Deep-transcriptome and ribonome sequencing redefines the molecular networks of pluripotency and the extracellular space in human embryonic stem cells. *Genome Research*, 21(12):2014–25, 2011.

[127] Amy Webb, Audrey C Papp, Jonathan C Sanford, Kun Huang, Jeffrey D Parvin, and Wolfgang Sadee. Expression of mrna transcripts encoding membrane transporters detected with whole transcriptome sequencing of human brain and liver. *Pharmacogenetics and genomics*, 23(5):269–78, 2013.

[128] Helena Persson, Anders Kvist, Natalia Rego, Johan Staaf, Johan Vallon-Christersson, Lena Luts, Niklas Loman, Goran Jonsson, Hugo Naya, Mattias Hoglund, and et al. Identification of new microrna in paired normal and tumor breast tissue suggests a dual role for the erbb2/her2 gene. *Cancer Research*, 71(1):78–86, 2011.

[129] María Isabel Alvarez-Mora, Laia Rodriguez-Revenga, Irene Madrigal, Francisca Torres-Silva, Elisabet Mateu-Huertas, Esther Lizano, Marc Friedlander, Eulália Martí, Xavier Estivill, and Montserrat Milá. Microrna expression profiling in blood from fragile x-associated tremor/ataxia syndrome patients. *Genes brain and behavior*, 2013.

[130] L Soreq, N Salomonis, M Brinstein, DS Greenberg, Z Israel, H Bergman, and H Soreq. Small rna sequencing-microarray analyses in parkinson leukocytes reveal deep brain stimulation-induced splicing changes that classify brain region transcriptomes. *Front. Mol. Neuroscience*, 2013.

[131] R Blum and BD Dynlacht. The role of myod1 and histone modifications in the activation of muscle enhancers. *Epigenetics*, 2013.

[132] Peng Cui, Wanfei Liu, Yuhui Zhao, Qiang Lin, Daoyong Zhang, Feng Ding, Chengqi Xin, Zhang Zhang, Shuhui Song, Fanglin Sun, and et al. Comparative analyses of h3k4 and h3k27 trimethylations between the mouse cerebrum and testis. *Genomics proteomics bioinformatics Beijing Genomics Institute*, 10(2):82–93, 2012.

[133] Jia-Yi Yao, Lei Zhang, Xin Zhang, Zhi-Ying He, Yue Ma, Li-Jian Hui, Xin Wang, and Yi-Ping Hu. H3k27 trimethylation is an early epigenetic event of p16ink4a silencing for regaining tumorigenesis in fusion reprogrammed hepatoma cells. *The Journal of Biological Chemistry*, 285(24):18828–18837, 2010.

[134] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012:1–11, 2012.

[135] Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, 2008.

[136] Manhong Dai, Robert C Thompson, Christopher Maher, Rafael Contreras-Galindo, Mark H Kaplan, David M Markovitz, Gil Omenn, and Fan Meng. Ngsqc: cross-platform quality analysis pipeline for deep sequencing data. *BMC genomics*, 11 Suppl 4:S7, 2010.

[137] Y Liao, GK Smyth, and W Shi. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10):108, 2013.

[138] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[139] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[140] Nuno A Fonseca, Johan Rung, Alvis Brazma, and John C Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, 2012.

[141] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, 14:91, 2013.

[142] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140, 2010.

[143] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics Oxford England*, 9(2):321–332, 2008.

[144] Hervé Abdi. The bonferonni and Šidák corrections for multiple comparisons. *Cognition*, 1:1–9, 2007.

[145] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*, 57(1):289–300, 1995.

[146] Shirley Pepke, Barbara Wold, and Ali Mortazavi. Computation for chip-seq and rna-seq studies. *Nature Methods*, 6(11 Suppl):S22–32, 2009.

[147] Elizabeth G Wilbanks and Marc T Facciotti. Evaluation of algorithm performance in chip-seq peak detection. *PLoS ONE*, 5(7):12, 2010.

[148] Michael P Washburn, Antonius Koller, Guy Oshiro, Ryan R Ulaszek, David Plouffe, Cosmin Deciu, Elizabeth Winzeler, and John R Yates. Protein pathway and complex clustering of correlated mrna and protein expression analyses in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 100(6):3107–3112, 2003.

[149] Jack D Keene. Rna regulons: coordination of post-transcriptional events. *Nature Reviews Genetics*, 8(7):533–543, 2007.

[150] Toma Tebaldi, Angela Re, Gabriella Viero, Ilaria Pegoretti, Andrea Passerini, Enrico Blanzieri, and Alessandro Quattrone. Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. *BMC Genomics*, 13(1):220, 2012.

[151] Natalia B Ivanova, John T Dimos, Christoph Schaniel, Jason A Hackney, Kateri A Moore, and Ihor R Lemischka. A stem cell molecular signature. *Science*, 298(5593):601–604, 2002. URL `http://www.ncbi.nlm.nih.gov/pubmed/12228721`.

[152] Lin Song, Nicole E Webb, Yingjie Song, and Rocky S Tuan. Identification and functional analysis of candidate genes regulating mesenchymal stem cell self-renewal and multipotency. *Stem Cells*, 24(7):1707–1718, 2006. URL `http://www.ncbi.nlm.nih.gov/pubmed/16574750`.

[153] Ju Ah Jeong, Kyung-Min Ko, Sohyun Bae, Choon-Ju Jeon, Gou Young Koh, and Hoeon Kim. Genome-wide differential gene expression profiling of human bone marrow stromal cells. *Stem Cells*, 25(4):994–1002, 2007.

[154] Adriane Menssen, Thomas Häupl, Michael Sittinger, Bruno Delorme, Pierre Charbord, and Jochen Ringe. Differential gene expression profiling of human bone marrow-derived mesenchymal stem cells during adipogenic development. *BMC Genomics*, 12(1):461, 2011.

[155] Rachel Groppo and Joel D Richter. Translational control from head to tail. *Current Opinion in Cell Biology*, 21(3):444–451, 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19285851`.

[156] Richard J Jackson, Christopher U T Hellen, and Tatyana V Pestova. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biology*, 11(2):113–127, 2010. URL `http://www.ncbi.nlm.nih.gov/pubmed/20094052`.

[157] Laura Poliseno, Leonardo Salmena, Jiangwen Zhang, Brett Carver, William J Haveman, and Pier Paolo Pandolfi. A coding-independent function of gene and pseudogene mrnas regulates tumour biology. *Nature*, 465(7301):1033–1038, 2010.

[158] H Lin, A Shabbir, M Molnar, and T Lee. Stem cell regulatory function mediated by expression of a novel mouse oct4 pseudogene. *Biochem Biophys Res Commun*, 355:111–116, 2007.

[159] Ryan Charles Pink, Kate Wicks, Daniel Paul Caley, Emma Kathleen Punch, Laura Jacobs, and David Raul Francisco Carter. Pseudogenes: pseudo-functional or key regulators in health and disease? *Rna New York Ny*, 17(5):792–798, 2011.

[160] Ujwal Sheth and Roy Parker. Decapping and decay of messenger rna occur in cytoplasmic processing bodies. *Science*, 300(5620):805–808, 2003. URL `http://www.sciencemag.org/content/300/5620/805.abstract`.

[161] Paul Anderson and Nancy Kedersha. Rna granules: post-transcriptional and epigenetic modulators of gene expression. *Nature Reviews Molecular Cell Biology*, 10(6):430–6, 2009. URL `http://www.ncbi.nlm.nih.gov/pubmed/19461665`.

[162] Merav Bar, Stacia K Wyman, Brian R Fritz, Junlin Qi, Kavita S Garg, Rachael K Parkin, Evan M Kroh, Ausra Bendoraite, Patrick S Mitchell, Angelique M Nelson, and et al. Microrna discovery and profiling in human embryonic stem cells by deep sequencing of small rna libraries. *Stem Cells*, 26(10):2496–2505, 2008.

[163] Nitish Mittal, Nilanjan Roy, M Madan Babu, and Sarath Chandra Janga. Dissecting the expression dynamics of rna-binding proteins in posttranscriptional regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106(48): 20300–20305, 2009.

[164] Prabha Sampath, Barsanjit Mazumder, Vasudevan Seshadri, and Paul L Fox. Transcript-selective translational silencing by gamma interferon is directed by a novel structural element in the ceruloplasmin mrna 3? untranslated region. *Molecular and Cellular Biology*, 19(10):6898–6905, 2003.

[165] Andrea Thiele, Yoshikuni Nagamine, Sunna Hauschildt, and Hans Clevers. Au-rich elements and alternative splicing in the beta-catenin 3?utr can influence the human beta-catenin mrna stability. *Experimental Cell Research*, 312(12):2367–2378, 2006.

[166] Lucy W Barrett, Sue Fletcher, and Steve D Wilton. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and molecular life sciences CMLS*, pages 1–22, 2012.

[167] Graziano Pesole. mrna untranslated regions (utrs). *Wiley Online Library*, (2000):1–5, 2005.

[168] Zhe Ji, Ju Youn Lee, Zhenhua Pan, Bingjun Jiang, and Bin Tian. Progressive lengthening of 3? untranslated regions of mrnas by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences of the United States of America*, 106(17):7028–7033, 2009. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2669788&tool=pmcentrez&rendertype=abstract`.

[169] Rickard Sandberg, Joel R Neilson, Arup Sarma, Phillip A Sharp, and Christopher B Burge. Proliferating cells express mrnas with shortened 3? untranslated regions and fewer microrna target sites. *Science*, 320(5883):1643–1647, 2008. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2587246&tool=pmcentrez&rendertype=abstract`.

[170] C K Rebelatto, A M Aguiar, M P Moretão, A C Senegaglia, P Hansen, F Barchiki, J Oliveira, J Martins, C Kuligovski, F Mansur, and et al. Dissimilar differentiation of mesenchymal stem cells from bone marrow, umbilical cord blood, and adipose tissue. *Experimental biology and medicine Maywood NJ*, 233(7):901–913, 2008.

[171] Ichiro Sekiya, Benjamin L Larson, Jussi T Vuoristo, Jian-Guo Cui, and Darwin J Prockop. Adipogenic differentiation of human adult stem cells from bone marrow stroma (mscs). *Journal of bone and mineral research the official journal of the American Society for Bone and Mineral Research*, 19(2):256–264, 2004. URL `http://www.ncbi.nlm.nih.gov/pubmed/14969395`.

[172] Yang Wang, You Wang, Yunfeng Rui, Lin Du, Tingting Tang, and Kerong Dai. In vitro study on multiple differentiation potential of swine synovium-derived mscs. *Zhongguo xiu fu chong jian wai ke za zhi Zhongguo xiufu chongjian waike zazhi Chinese journal of reparative and reconstructive surgery*, 23(6):737–741, 2009.

[173] Mette Hemmingsen, Søren Vedel, Peder Skafte-Pedersen, David Sabourin, Philippe Collas, Henrik Bruus, and Martin Dufva. The role of paracrine and autocrine signaling in the early phase of adipogenic differentiation of adipose-derived stem cells. *PloS one*, 8(5):e63638, 2013.

[174] Eckhard Alt, Yasheng Yan, Sebastian Gehmert, Yao-Hua Song, Andrew Altman, Sanga Gehmert, Daynene Vykoukal, and Xiaowen Bai. Fibroblasts share mesenchymal phenotypes with stem cells, but lack their differentiation and colony-forming potential. *Biology of the cell under the auspices of the European Cell Biology Organization*, 103(4):197–208, 2011.

[175] Antonella Blasi, Carmela Martino, Luigi Balducci, Marilisa Saldarelli, Antonio Soleti, Stefania E Navone, Laura Canzi, Silvia Cristini, Gloria Invernici, Eugenio A Parati, and et al. Dermal fibroblasts display similar phenotypic and differentiation capacity to fat-derived mesenchymal stem cells, but differ in anti-inflammatory and angiogenic potential. *Vascular cell*, 3(1):5, 2011.

[176] Makoto Osonoi, Osamu Iwanuma, Akihito Kikuchi, and Shinichi Abe. Fibroblasts have plasticity and potential utility for cell therapy. *Human cell official journal of Human Cell Research Society*, 24(1):30–34, 2011.

[177] Sohyun Bae, Jung Hoon Ahn, Chae Woon Park, Hye Kyung Son, Keun-Soo Kim, Nam-Kyu Lim, Choon-Ju Jeon, and Hoeon Kim. Gene and microrna expression signatures of human mesenchymal stromal cells in comparison to fibroblasts. *Cell and Tissue Research*, 335(3):565–573, 2009.

[178] Nick J Proudfoot, Andre Furger, and Michael J Dye. Integrating mrna processing with transcription. *Cell*, 108(4):501–512, 2002.

[179] M Davila Lopez and T Samuelsson. Early evolution of histone mrna 3' end processing. *RNA*, 14(1):1–10, 2007.

[180] Carol S Lutz. Alternative polyadenylation: A twist on mrna 3' end formation. *ASC Chemical Biology*, 3(10):609–617, 2008.

[181] E Wahle and W Keller. The biochemistry of polyadenylation. *Trends in Biochemical Sciences*, 21(7):247–250, 1996.

[182] E. Beaudoing, S Freier, JR Wyatt, JM Claverie, and D Gautheret. Patterns of variant polyadenylation signal usage in human genes. *Genome Research*, 10(7):1001–1010, 2000.

[183] K Venkataraman. Analysis of a noncanonical poly(a) site reveals a tripartite mechanism for vertebrate poly(a) site recognition. *Genes & Development*, 19(11):1315–1327, 2005.

[184] K Venkataraman. An interaction between U2AF 65 and CF Im links the splicing and 3' end processing machineries. *The EMBO Journal*, 25(20):4854–4864, 2006.

[185] Kira Glover-Cutter, Soojin Kim, Joaquin Espinosa, and David L Bentley. Rna polymerase ii pauses and associates with pre-mrna processing factors at both ends of genes. *Nature Structural & Molecular Biology*, 15(1):71–8, 2008.

[186] E Wahle. Poly(a) tail length control is caused by termination of processive synthesis. *The Journal of Biological Chemistry*, 270(6):2800–2808, 1995.

[187] Nicolas Viphakone, Florence Voisinet-Hakil, and Lionel Minvielle-Sebastia. Molecular dissection of mrna poly(a) tail length control in yeast. *Nucleic Acids Research*, 36(7):2418–2433, 2008.

[188] Anita Nag, Kazim Narsinh, and Harold G Martinson. The poly(a)-dependent transcriptional pause is mediated by cpsf acting on the body of the polymerase. *Nature Structural & Molecular Biology*, 14(7):2800–2808, 2007.

[189] J M Coller and M P Wickens. mrna stabilization by poly(a) binding protein is independent of poly(a) and requires translation. *Genes & Development*, 12(20):3226–3235, 1998.

[190] N Siddiqui, D A Mangus, T C Chang, J M Palermino, A B Shyu, and K. Gehring. Poly(a) nuclease interacts with the c-terminal domain of polyadenylate-binding protein domain from poly(a)-binding protein. *Journal of Biological Chemistry*, 282(34):25067–25075, 1998.

[191] N K Gray, J M Coller, K S Dickson, and M Wickens. Multiple portions of poly(a)-binding protein stimulate translation in vivo. *The EMBO Journal*, 19(17):4723–4733, 1998.

[192] S A Shell, C Hesse, Jr SM Morris, and C Milcarek. Elevated levels of the 64-kda cleavage stimulatory factor (cstf-64) in lipopolysaccharide-stimulated macrophages influence gene expression and induce alternative poly(a) site selection. *Journal of Biological Chemistry*, 280(48):39950–39961, 2005.

[193] Sven Danckwardt, Anne-Susan Gantzert, Stephan Macher-Goeppinger, Hans Christian Probst, Marc Gentzel, Matthias Wilm, Hermann-Josef Gröne, and Peter Schirmacher. p38 mapk controls prothrombin expression by regulated rna 3? end processing. *Molecular Cell*, 41(3):298–310, 2011.

[194] Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, and Anthony C Schweitzer. Hits-clip yields genome-wide insights into brain alternative rna processing. *Molecular Cell*, 456(7221):464–469, 2008.

[195] A J Wood, R Schulz, K Woodfine, K Koltowska, C V Beechey, J Peters, D Bourc'his, and R J Oakey. Regulation of alternative polyadenylation by genomic imprinting. *Genes & Development*, 22(9):1141–1146, 1998.

[196] Y Shen, G Ji, B J Haas, X Wu, J Zheng, G J Reese, and Q Q Li. Genome level analysis of rice mrna 3'-end processing signals and alternative polyadenylation. *Nucleic Acids Research*, 36(9):3150–3161, 2008.

[197] M Lagos-Quintana, R Rauhut, W Lendeckel, and T Tuschl. Identification of novel genes coding for small expressed rnas. *Science*, 294(5543):853–858, 2001.

[198] Xuezhong Cai, Curt H Hagedorn, and Bryan R Cullen. Human micrornas are processed from capped, polyadenylated transcripts that can also function as mrnas. *Rna New York Ny*, 10(12):1957–1966, 2004.

[199] Yimei Cai, Xiaomin Yu, Songnian Hu, and Jun Yu. A brief review on the mechanisms of mirna regulation. *Genomics Proteomics Bioinformatics*, 7(4):147–154, 2009.

[200] Sun Wei, Julie Li Yi-Shuan, Huang Hsien-Da, Y-J Shyy John, and Chien Shu. microrna: A master regulator of cellular processes for bioengineering systems. *Annual Review of Biomedical Engineering*, 12:1–27, 2010.

[201] Vamsi K Gangaraju and Haifan Lin. Micrornas: key regulators of stem cells. *Nature Reviews Molecular Cell Biology*, 10(2):116–25, 2009.

[202] Ana Eulalio, Eric Huntzinger, and Elisa Izaurralde. Getting to the root of mirna-mediated gene silencing. *Cell*, 132(1):9–14, 2008.

[203] Tariq M Rana. Illuminating the silence: understanding the structure and function of small rnas. *Nature Reviews Molecular Cell Biology*, 8(1):23–36, 2007.

[204] Shobha Vasudevan, Yingchun Tong, and Joan A Steitz. Switching from repression to activation: micrornas can up-regulate translation. *Science*, 318(5858):1931–1934, 2007.

[205] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, 2005.

[206] Nikolaus Rajewsky. microrna target predictions in animals. *Nature Genetics*, 38 Suppl (June):S8–13, 2006.

[207] Eric Huntzinger and Elisa Izaurralde. Gene silencing by micrornas: contributions of translational repression and mrna decay. *Nature Reviews Genetics*, 12(2):99–110, 2011.

[208] M Arribas-Layton, D Wu, J Lykke-Andersen, and H Song. Structural and functional control of the eukaryotic mrna decapping machinery. *Biochim. Biophys. Acta*, 1829:580–589, 2013.

[209] T Nishihara, L Zekri, J E Braun, and E Izaurralde. mirisc recruits decapping factors to mirna targets to enhance their degradation. *Nucleic Acids Research*, pages gkt619–, 2013.

[210] Shuo Gu and Mark A Kay. How do mirnas mediate translational repression? *Silence*, 1 (1):11, 2010. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2881910&tool=pmcentrez&rendertype=abstract`.

[211] Sergej Djuranovic, Ali Nahvi, and Rachel Green. A parsimonious model for gene regulation by mirnas. *Science*, 331(6017):550–553, 2011.

[212] Witold Filipowicz, Suvendra N Bhattacharyya, and Nahum Sonenberg. Mechanisms of post-transcriptional regulation by micrornas: are the answers in sight? *Nature Reviews Genetics*, 9(2):102–14, 2008.

[213] B Futch, G I Latter, P Monardo, C S McLaughlin, and J I Garrels. A sampling of the yeast proteome. *Molecular and celular biology*, 19(11):7357–7368, 1999.

[214] Tamir Tuller, Martin Kupiec, and Eytan Ruppin. Determinants of protein abundance and translation efficiency in S. *cerevisiae*. *PLoS Computational Biology*, 3(12):10, 2007. URL `http://dx.plos.org/10.1371/journal.pcbi.0030248`.

[215] S G Stevens and C M Brown. *In silico* estimation of translation efficiency in human cell lines: potential evidence for widespread translational control. *PLoS One*, 8(2):e57625, 2013. URL `http://www.ncbi.nlm.nih.gov/pubmed/23460887`.

[216] Henrik Molina, Yi Yang, Travis Ruch, Jae-Woo Kim, Peter Mortensen, Tamara Otto, Anuradha Nalli, QiQun Tang, M. Daniel Lane, Raghothama Chaerkady, and Akhilesh Pandey. Temporal profiling of the adipocyte proteome during differentiation using a 5-plex silac based strategy. *J Proteome Res.*, 8(1):48–58, 2009.

[217] K T Tycowski, M D Shu, and J A Steitz. A mammalian gene with introns instead of exons generating stable rna products. *Nature*, 379(6564):464–466, 1996.

[218] Marek Zywicki, Kamilla Bakowska-Zywicka, and Norbert Polacek. Revealing stable processing products from ribosome-associated small rnas by deep-sequencing data analysis. *Nucleic Acids Research*, 40(9):1–12, 2012.

[219] Philipp Kapranov, Jill Cheng, Sujit Dike, David A Nix, Radharani Duttagupta, Aarron T Willingham, Peter F Stadler, Jana Hertel, Jörg Hackermüller, Ivo L Hofacker, and et al. Rna maps reveal new rna classes and a possible function for pervasive transcription. *Science*, 316(5830):1484–1488, 2007.