Identificación y caracterización de eventos de lectura de codones de parada a nivel genómico en *D. melanogaster*

TESIS DOCTORAL

Luciana Inés Escobar

Departamento de Ciencias Biológicas
Facultad de Ciencias Exactas
Universidad Nacional de La Plata

23 de Agosto de 2021

Documento maquetado con TEXIS v.1.0.

Este documento está preparado para ser imprimido a doble cara.

Identificación y caracterización de eventos de lectura de codones de parada a nivel genómico en *D. melanogaster*

Memoria que presenta para optar al título de Doctor en Ciencias

Luciana Inés Escobar

Dirigida por el Doctor

Dr. Luis Aníbal Diambra
Dr. Jorge Rafael Ronderos

Departamento de Ciencias Biológicas
Facultad de Ciencias Exactas
Universidad Nacional de La Plata

23 de Agosto de 2021

"Ser hombre es precisamente ser responsable. Es sentir, al dejar tu piedra, que estás ayudando a construir el mundo."

Antoine de Saint-Exupéry
extraído de « Terre des hommes » (Capítulo II), 1939.

Dedicado a mis padres ...

Agradecimientos

"El conocimiento no es una vasija que se llena, sino un fuego que se enciende."
Plutarco

Agradezco sincera y profundamente el apoyo y la enseñanza de todas aquellas personas, instituciones y medios que participaron haciendo posible el desarrollo de esta Tesis, complementando el proceso de una u otra forma, cediendo o construyendo las herramientas indispensables para este objetivo.

En primer lugar agradezco a mi director, el Dr. Luis A. Diambra, quien siempre me guió en forma clara, paciente y objetiva. Por brindarme un espacio de trabajo, y por su dedicación y compromiso en la tarea de guiarme a través de numerosos desafíos durante el proceso de mi formación profesional. Admiro la destacada calidad de su persona, experiencia y habilidad, por cuanto le estaré eternamente agradecida por su confianza, dado que jamás dudó en ayudarme incondicionalmente cada vez que necesité de su conocimiento y supervisión.

También deseo destacar mi gratitud a mi director conjunto, el Dr. Jorge R. Ronderos, por recibirme desde el primer momento e incentivarme a novedosos proyectos. Gracias por su fundamental aporte al trabajo, por acompañarme siempre de manera abierta y positiva, ofreciendo reiteradamente su comprensión y sabios consejos.

Al Departamento de Ciencias Biológicas de la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata (FCE-UNLP), por haber permitido realizar mis estudios de Postgrado y ofrecer las herramientas para lograr mi formación académica en esta ilustre casa de estudios.

Al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), por brindarme la posibilidad y el sustento necesarios para dedicarme a este trabajo a través del otorgamiento de una prestigiosa beca de estudio, indefectible para crecer en el extraordinario ámbito de la investigación científica.

A todos los docentes, colegas, y compañeros que han formado parte de mi carrera, poniendo siempre buena predisposición al compartir sus clases, experiencias, habilidades y anécdotas vividas, representando aspectos fundamentales en este objetivo. Obtuve de ellos una vasta cantidad de conocimientos y reflexiones, que perdurarán por siempre en mi memoria.

A las cálidas personas integrantes del CREG, por siempre dar una mano desinteresada en cada detalle que cualquiera necesitara, y por saber conformar un grupo cordial de compañerismo donde se aúnan los esfuerzos por hacer crecer a este entrañable cuerpo institucional. Quiero hacer un reconocimiento por afiliar sus voluntades a diario, por la visión crítica y el aporte constructivo en la realización de

tareas comunitarias.

A Leonardo Dettano, Cocó Villalobos, Marina Rossi y Verónica Cajen, entre mis buenos amigos, por su infinita paciencia y su fiel compañía tanto en los buenos momentos como en las dificultades que se me presentaron. Ustedes han sabido ser un refugio armonioso que ayudó a suplir la nostalgia de mis raíces. ¡Gracias de corazón por estar siempre al lado mío, por creer en mí y por cuidarme!

A mis familiares y afectos por los mimos y las distracciones, por alentarme y darme ánimos cada vez que lo necesité, a pesar de las distancias. En especial a mis padres: Mirna Cortés y Freddy Escobar, y a mis hermanos Mauro y Fernando, por su incondicional apoyo en mis decisiones. Gracias por sostenerme pacientemente en los momentos de incertidumbre, por incentivarme a ser valiente con mis aspiraciones; todo lo que soy se lo debo a ustedes. Gracias por haberme amado, enseñado y cuidado siempre... Nunca es suficiente gratitud por tanto que dan!

Y finalmente, pero no menos importante, gracias a los amigos/as de cerca o lejos que me conocen y me aprecian, por haberme soportado, comprendido, acompañado y fortalecido en diversos aprendizajes durante el camino recorrido. Cada uno de ustedes sabe cómo, cuándo y por qué. ¡Son inolvidables!

¡Gracias infinitas! Luciana.

Resumen

"La hipótesis es el principal instrumento intelectual en la investigación. Su función consiste en sugerir nuevos experimentos u observaciones y, por consiguiente, muchas veces conduce a nuevos descubrimientos aún cuando ella misma no sea correcta... Así, una hipótesis puede ser fructífera no sólo para sus proponentes, sino aún más, para desarrollar avances en ciencia."

William Ian Beardmore Beveridge

Para que las células funcionen correctamente, la información génica debe expresarse fielmente en ARN o proteínas. Un paso clave en la expresión génica es la traducción de ARNm a proteína, donde el reconocimiento de uno de los tres codones de parada (TAA, TGA o TAG) por la maquinaria traduccional es esencial para terminar la traducción en la posición correcta, y garantizar la función de la proteína sintetizada. Sin embargo, en determinadas ocasiones, los ribosomas pueden ignorar el codón de terminación y en cambio re-codificarlos continuando la traducción en la región 3' UTR, incorporando aminoácidos hasta que la maquinaria ribosomal reconoce un codón de parada posterior. En consecuencia, se sintetizan nuevos productos polipeptídicos extendidos en su extremo C-terminal que pueden adquirir un nuevo dominio funcional. Esta ausencia de reconocimiento del codón de terminación se denomina lectura del codón de parada (denotado en adelante como LCP).

Aunque existe evidencia de proteínas funcionales derivadas de eventos de LCP en diversos genomas, aún se desconocen ampliamente las condiciones que inducen este fenómeno. Por esta razón, se ha discutido la presunción de que la LCP puede constituir un error de decodificación en la traducción, debido a errores moleculares en un entorno regulador que conduce a la expresión de diferentes isoformas de proteínas. No obstante, recientes experimentos en eucariotas sugieren que los eventos de LCP son algo más generalizados de lo que se suponía previamente, y que podrían estar regulados. En este sentido, varios autores han propuesto que la LCP podría ocurrir por la acción programada de componentes moleculares específicos, aún no identificados; generando la hipótesis que la LCP es un mecanismo para ampliar la diversidad de los proteomas.

Por otro lado, las nuevas tecnologías de secuenciación han puesto de manifiesto

algunas dificultades del proceso de anotación genómica. En este sentido, se ha detectado una considerable cantidad de eventos de traducción en regiones previamente consideradas no codificantes, tanto en las extensiones de marcos de lectura por LCP, como en la traducción de pequeños marcos de lectura en 5' UTRs. Muchos de estos péptidos no convencionales descubiertos conllevan estructuras génicas que aguardan clasificación, y cuestionan la anotación existente de muchos genes conocidos. Al respecto, una anotación eficiente requiere no sólo el registro de nuevos elementos funcionales, sino además corregir la delimitación imprecisa de muchos ORFs conocidos.

En esta tesis se realizó una identificación exhaustiva de eventos de LCP en el transcriptoma de *D. melanogaster* mediante la evaluación de perfiles de densidad ribosomal, construidos a partir de datos públicos derivados del secuenciamiento de los fragmentos transcriptómicos protegidos por ribosomas (*Ribo-Seq*). Además, se realizó una estimación de la tasa de fuga ribosomal asociada a cada evento de LCP identificado. Con base en la identificación de miles de eventos de LCP, se realizó una caracterización estadística de la frecuencia de uso de nucleótidos en la región proximal al codón de parada en cada transcripto. La asociación de estas dos informaciones a través de un modelo de regresión lineal, permitió dilucidar que el contexto de nucleótidos es un factor molecular determinante en la regulación de la terminación eficiente de la traducción.

Este análisis permitió inferir la existencia de patrones que funcionan a modo de señal para la ocurrencia de la LCP. Estos son dependientes de cada codón de parada, sugiriendo la existencia de al menos dos mecanismos que alteran el proceso de terminación traduccional. Estos resultados son también evidencia de que la LCP no constituye un mero error de decodificación, sino que responde a la presencia de factores moleculares programados para la expresión diferencial de productos génicos de baja abundancia. Más allá de contribuir a la mejora en la anotación genómica de la mosca de la fruta, esta tesis profundiza en el conocimiento de los mecanismos que regulan la LCP y proporciona un avance en la comprensión de la expresión génica. En particular, este conocimiento puede ampliar el potencial de las estrategias terapéuticas para patologías genéticas causadas por mutaciones sin sentido, como la fibrosis quística.

Palabras clave: Expresión génica; Traducción; Lectura de codones de parada (LCP); Secuencias codificantes y no codificantes; RNA-Seq; Perfil de ribosomas.

Índice

Agradecimientos VII

Resumen IX

1. Introducción 1
- 1.1. La información genética . . . 1
  - 1.1.1. El concepto de gen y el dogma central . . . 2
  - 1.1.2. Cantidad de genes codificantes en los genomas . . . 5
- 1.2. Regulación de la expresión génica en procariontes . . . 6
- 1.3. Regulación de la expresión génica en eucariontes . . . 7
  - 1.3.1. Regulación a nivel de cromatina . . . 8
  - 1.3.2. Regulación a nivel transcripcional . . . 9
  - 1.3.3. Regulación a nivel postranscripcional . . . 9
  - 1.3.4. Regulación a nivel traduccional . . . 11
  - 1.3.5. Regulación a nivel postraduccional . . . 14
- 1.4. El auge de la genómica . . . 15
  - 1.4.1. LCP como mecanismo de recodificación . . . 18
  - 1.4.2. Terapias basadas en estímulo de LCP . . . 20
- 1.5. Motivación y planteo de la tesis . . . 22

2. Identificación de eventos de lectura de codones de parada 25
- 2.1. Lectura de codones de parada y la sub-anotación de genomas . . . 25
- 2.2. Los perfiles ribosomales y LCP . . . 29
- 2.3. Nuevos eventos de LCP derivados de perfiles ribosomales . . . 35
- 2.4. Confirmación de eventos de LCP por espectrometría de masa . . . 41
- 2.5. Análisis de enriquecimiento por ontología génica . . . 49
- 2.6. Conclusiones . . . 52

3. Análisis de la secuencia contexto en eventos de LCP 55
- 3.1. Factores determinantes en la lectura del codón de parada . . . 55
- 3.2. Estimación de la tasa de fuga . . . 57
- 3.3. Frecuencia de LCP en los codones de parada . . . 58
- 3.4. Análisis de la secuencia de contexto al codón de parada en LCP . . . 60
  - 3.4.1. Modelos predictivos . . . 65
  - 3.4.2. Evaluación de los modelos predictivos . . . 67

3.4.3. Conclusiones 74

4. Datos y Metodología **77**
4.1. Sobre el organismo de estudio 77
4.2. Datos utilizados 78
4.3. Pre-procesamiento 80
4.4. Cuantificación de los niveles de expresión 81
4.5. Alineamiento de las huellas ribosomales 83
4.5.1. Construcción y análisis de perfiles de ribosomas 85
4.6. Identificación de nuevos eventos de LCP a través de perfiles ribosomales 85
4.7. Confirmación de eventos de LCP por espectrometría de masa 86

5. Cierre y conclusiones **89**
5.1. Discusión 89
5.1.1. Identificación de eventos de LCP mediante perfiles ribosomales 90
5.1.2. Estudio de factores determinantes de LCP en la secuencia contexto 93
5.1.3. Conclusiones finales 95

6. ANEXOS **97**

Bibliografía **125**

Índice de figuras

1.1. Dogma Central de la Biología Molecular. 4
1.2. Procesamiento postranscripcional del pre-ARNm. 10
1.3. Factores de regulación a nivel traduccional. 12
1.4. Niveles de regulación de la expresión génica. 15
1.5. Moléculas de ARN codificantes de pequeños ORFs. 17
1.6. Terminación traduccional vs. LCP. 19
1.7. Terapias de inducción de LCP. 21

2.1. Perfil ribosomal sin LCP. 30
2.2. Perfil ribosomal con un evento simple de LCP. 31
2.3. Perfil ribosomal con múltiples eventos de LCP. 32
2.4. Perfil ribosomal de un falso evento de LCP. 34
2.5. Comparación de casos de LCP reportados. 35
2.6. Perfil ribosomal de FBtr0310464 con doble evento de LCP. 37
2.7. Perfil ribosomal de FBtr0076462 con triple evento de LCP y sintenia. 39
2.8. Perfil ribosomal de FBtr0072583 con evento de LCP y sintenia. . . . 40
2.9. Perfil ribosomal de FBtr0072343 con doble evento de LCP y sintenia. 41
2.10. Perfil ribosomal de FBtr0079297 con evento de LCP identificado y sintenia. 42
2.11. Eventos de LCP identificados por análisis de perfil ribosomal. 43
2.12. Doble evento de LCP confirmado por espectrometría de masas. . . . 45
2.13. Evento simple de LCP confirmado por espectrometría de masa. . . . 47
2.14. Evento simple de LCP confirmado por espectrometría de masa. . . . 48
2.15. Análisis segmentado de enriquecimiento por ontología génica para los genes identificados con LCP. 50
2.16. Procesos biológicos asociados a LCP. 53

3.1. Modelo de construcción de perfiles de densidad ribosomal. 58
3.2. Estimación de LCP mediante frecuencia de fuga de codones de parada 59
3.3. Frecuencia de fuga según nt 61
3.4. Divergencia en la secuencia contexto 63
3.5. Histogramas de eventos de LCP 65
3.6. Codificación numérica del contexto 68
3.7. TGA-Performance del modelo 70
3.8. TAA-Performance del modelo 72

3.9. TAG-Performance del modelo 73

4.1. Dmel 79

6.1. Ribo-Seq workflow 118
6.2. Formato SAM. 119
6.3. Formato SAM. 119
6.4. Recorte de adaptador con *Cutadapt*. 119
6.5. Algoritmo *Bowtie*. 120
6.6. TopHat workflow 121
6.7. Cufflinks workflow 122
6.8. Peptide-Shaker 123
6.9. Enrichr workflow. 124

Índice de Tablas

2.1. Lista de algunos nuevos transcriptos con LCP. 36
2.2. Fragmentos de extensiones de LCP identificadas por espectrometría de masa. 44

4.1. Estudios utilizados 80

6.1. Campos obligatorios SAM. 102
6.2. Tabla B.2 102
6.3. Tabla B.3 103

Capítulo 1

Introducción

1.1. La información genética

La información genética de cada organismo yace en su ADN, enormes moléculas compuestas por bases nitrogenadas polimerizadas en una estructura de doble hélice. Estas moléculas tienen la increíble capacidad de crear copias exactas de sí mismas, propiedad necesaria para transmitir la herencia biológica a través de las generaciones (Lodish et al., 2005; Klug et al., 2006). Esta herencia es la que permite a los organismos adquirir las características fenotípicas de sus progenitores. La información hereditaria está organizada en genes, que son las unidades que codifican la información para sintetizar proteínas y otros elementos funcionales, como por ejemplo ARNt, ARNr, miARN y ribozimas; lo que se denomina expresión génica. Básicamente, la expresión génica es un proceso altamente regulado por otras moléculas, como los factores de transcripción, que se unen al ADN en sitios regulatorios específicos. Estas uniones modulan la estructura tridimensional del ADN y determinan qué genes deben ser expresados mediante un proceso conocido como transcripción génica. De esta forma, la secuencia de nucleótidos (nt) del ADN no solo almacena la información genética de unidades funcionales en regiones codificantes (conocidas como "CDS" por sus siglas en inglés: *Coding Sequence*), sino que además contiene regiones no codificantes imprescindibles para la regulación de la expresión, es decir, determinan cuándo, cuánto y cuáles genes deben ser expresados. Podemos entonces extender la definición de genoma más allá de la simple sumatoria de genes y decir que el genoma consiste en la secuencia completa de todas las moléculas de ADN en un organismo, y pensar al genoma como un centro de cómputo de la célula, donde se almacena y recupera información en respuesta a cada estado celular o estímulo externo. La regulación de la expresión es esencial en la regulación global del metabolismo celular, dado que ningún organismo necesita todos los productos génicos de forma simultánea, ni a los mismos niveles. Incluso, la necesidad de un determinado producto génico podría cambiar con el tiempo, con lo cual el sistema de regulación sirve para ajustar estos niveles. La organización del ADN y los procesos involucrados en la regulación de la expresión génica son muy complejos, y han sido objeto de estudio de miles de investigadores en el mundo entero durante los últimos 50 años. Si bien se han hecho enormes avances, aún no ha sido completamente elucidado cómo la información lineal contenida en el ADN resulta en el repertorio de fenotipos observados en la naturaleza. Año tras año

nuevos elementos regulatorios son identificados, mostrando que aún queda camino por recorrer. Es por eso crucial el desarrollo de nuevas técnicas y estudios que permitan la identificación de nuevos actores regulatorios. Los procesos de regulación de la expresión varían ampliamente de procariotas a eucariotas, y serán brevemente detallados más adelante. Por ejemplo, en organismos procariontes, al carecer de membrana nuclear que regule el transporte molecular del núcleo hacia el citoplasma, la traducción comienza inmediatamente después de la obtención de moléculas de ARNm por el proceso de transcripción. Por otro lado, en los eucariontes, los ARN transcritos a partir de los genes deben transportarse desde el núcleo hasta el citoplasma, atravesando la membrana nuclear. Además, el genoma nuclear está organizado en cromosomas, y la expresión incluye el proceso de *splicing* que involucra el corte y empalme de secuencias con el fin de ampliar el número de productos proteicos posible, sin aumentar el número de genes y el tamaño del genoma. Sin embargo, para seguir avanzando es necesario profundizar en las nociones de gen y expresión génica.

1.1.1. El concepto de gen y el dogma central

A partir de la teoría original de Gregor Mendel sobre la determinación de caracteres físicos específicos (por ej. el color de la flor) mediante partículas hereditarias discretas (factores), el concepto de gen ha evolucionado gradualmente hacia el de unidad funcional. Desde un punto de vista actual, podemos entender el concepto mendeliano de gen como el de una unidad funcional y estructural, sujeta a transmisión, mutación y evolución, y que se distribuye ordenadamente en los cromosomas. Así, los genes y los cromosomas son los principios fundamentales de la teoría cromosómica de la herencia, que explica la transmisión de la información genética que controla los caracteres fenotípicos (Klug et al., 2006). El término "gen" (del griego "generar") fue propuesto en 1909 por el biólogo danés Wilhelm Ludvig Johannsen en referencia a la unidad física y funcional de la herencia biológica; y hacia mediados del siglo XX se consignó como la molécula de ADN que codifica una proteína (hipótesis de un gen $\rightarrow$ un polipéptido). Este es un concepto que proporciona una naturaleza molecular y estructural al gen: el gen codifica proteínas y su estructura está definida por el orden lineal de sus nucleótidos. Este concepto fue modificado posteriormente cuando se comprendió que los genes podían determinar subunidades proteicas y que existen diversos tipos de ARN involucrados en la síntesis de proteínas. Estos ARN no se traducen, demostrando que la traducción no es estrictamente necesaria para que un gen tenga una función determinada. El desarrollo de técnicas de secuenciación y clonación a fines de los '70, permitió desentrañar la estructura precisa de los genes hasta el nivel de las bases. Tales técnicas aportaron mucha información sobre la expresión de los genes, y surge el concepto de gen vinculado al "Dogma Central de la Biología Molecular" (Lodish et al., 2005; Klug et al., 2006). El Dogma Central de la Biología Molecular fue enunciado por Francis Crick en 1958, tras el descubrimiento de la estructura del ADN en 1953 (Watson y Crick, 1953) y posteriores dilucidaciones sobre cómo el ADN rige la síntesis de ARN (en el núcleo celular), el cual preside luego el ensamblaje de proteínas en los ribosomas citoplasmáticos (Crick, 1970). En resumen, el Dogma Central de la Biología Molecular establece que la expresión de genes que codifican polipéptidos fluye unidireccionalmente: del ADN al ARN, y de éste hacia las proteínas; nunca en sentido inverso. Las flechas negras en la Fig. 1.1 esquematizan este concepto.

Antes de realizarse la síntesis proteica, se inicia el proceso de transcripción, mediante el cual la secuencia nucleotídica del ADN se transcribe a ARN, conocido como ARN transcripto primario o bien ARN precursor del mensajero (pre-ARNm). Tras el procesamiento o maduración del ARNm, éste se transporta hacia el citoplasma celular, donde sirve como molde en el proceso de traducción para la síntesis de una cadena polipeptídica.

Sin embargo, la representación simplificada del Dogma Central como flujo unidireccional de información de ADN $\rightarrow$ ARN $\rightarrow$ proteína no refleja el papel de las proteínas en la síntesis de los ácidos nucleicos; y más aún su responsabilidad en la regulación de la expresión génica, el proceso global de decodificación del ADN en las proteínas que caracterizan los diversos tipos celulares. En otras palabras, la producción de proteínas requiere ADN y ARN, y a su vez, la replicación del ADN y la transcripción a ARN requieren proteínas. Esta dependencia mutua no está contemplada en el concepto inicial del Dogma Central, y demuestra que la expresión de un gen no sigue un proceso lineal de sentido único para la síntesis de productos funcionales en las células, ni para la evolución de éstos. Ciertamente, este dogma ha constituido la base de la biología a nivel celular. No obstante, la ciencia es un ente dinámico que no se asienta en dogmas. Así, han surgido una serie de excepciones que escapan a este dogma; por ejemplo, elementos subcelulares como los priones, las ribozimas y las telomerasas. Los priones son proteínas anormalmente plegadas que inducen el plegamiento anormal de proteínas correctamente plegadas. Esta propiedad le confiere una actividad catalítica de naturaleza infecciosa (Prusiner, 1997). Las ribozimas son moléculas de ARN con actividad catalítica. Un ejemplo representativo es su intervención en el proceso de traducción, donde un ARN ribosómico cataliza la transferencia del ARNt-aminoácido desde el sitio A del ribosoma al sitio P con la formación de un enlace peptídico (acción peptidiltransferasa), durante la elongación de la cadena polipeptídica. A su vez, la ribozima ARNasa P interviene en la maduración de los ARN ribosómicos y transferentes.
Por su parte, las telomerasas son ribonucleoproteínas con función enzimática que participan en el mantenimiento de la longitud de los telómeros de los cromosomas; reduciendo el desgaste progresivo del ADN que provoca la senescencia celular (Wright et al., 1996). Otro ejemplo de excepción al dogma es la transcriptasa inversa en los retrovirus, que cataliza la síntesis de ADN de doble hebra utilizando como molde ARN monocatenario (ARN $\rightarrow$ ADN) (Kunkel y Bebenek, 2000; Lodish et al., 2005; Klug et al., 2006). En procariotas, la transcriptasa inversa es codificada por una secuencia de ADN denominada retrón (retroelementos) (Lampson et al., 2005; Simon y Zimmerly, 2008), y la síntesis de ADN requiere un cebador o "primer"; mientras que los genomas eucariotas poseen tramos auto-replicantes llamados retrotransposones, abundantes en plantas y animales (Lodish et al., 2005).

Además, las ADN polimerasas controlan el proceso de replicación del ADN nuclear uniéndose a una secuencia específica (cebador), para obtener otra molécula de ADN idéntica. Las proteínas pueden incluso activar o inhibir la transcripción de un gen determinado a ARN, utilizando la molécula de ADN como molde. También es posible producir *in vitro* la traducción en ausencia de ARN, utilizando ribosomas para traducir directamente las moléculas de ADN y obtener proteínas. Por lo tanto, el fundamento central de la biología molecular propuesto por Crick tuvo que ser modificado, dado que el ARN puede replicarse y transcribirse a ADN, además de la interacción de diversas proteínas en tales procesos. De esta forma, el Dogma Central que actualmente se utiliza es mucho más complejo, como se esquematiza

Figura 1.1: Esquema moderno del Dogma Central de la Biología Molecular. Tradicionalmente, el flujo o procesamiento de la información genética en los organismos procariotas y eucariotas se explica en tres etapas principales (flechas negras): Replicación de ADN, Transcripción a ARN y Traducción a proteínas. Excepcionalmente, algunos virus utilizan mecanismos alternativos como la replicación de ARN y/o la transcripción inversa (flechas rojas), presente también en elementos retrones bacterianos y retrotransposones de eucariotas.

en la Fig. 1.1:

La característica especial en la síntesis proteica es que sólo se dispone de veinte aminoácidos (producidos a partir de la combinación de 4 bases nitrogenadas en el ARNm), que se ensamblan en un orden específico para crear una proteína particular. Esto hace de la "secuenciación" un problema crucial para el conocimiento de la enorme diversidad de proteínas en los organismos, asumiendo que la cantidad de información necesaria para producirla es considerable. Este problema se resuelve durante la traducción, donde la unidad de lectura del ARNm llamada codón (triplete de bases), interpretado por el código genético, se traduce a aminoácidos. Actualmente se sabe que los genes pueden codificar más de un polipéptido y que una proteína puede ser codificada por el conjunto de diferentes genes. Además, la existencia de genes solapantes contrastan la hipótesis de un gen $\rightarrow$ una unidad física. Y por otro lado, el *splicing* o procesamiento alternativo rebate la hipótesis de un gen $\rightarrow$ un polipéptido. Se sabe incluso que gran parte del genoma eucariótico no codifica genes funcionales, en base a la compleja organización de su secuencia caracterizada por numerosas categorías de ADN altamente y/o moderadamente repetitivo. El ADN repetitivo consiste en repeticiones en tándem agrupadas en diversas regiones del genoma, y en secuencias repetidas no agrupadas en tándem distribuidas uniformemente por el genoma. En el primero de los casos, el tamaño de cada agrupación varía entre individuos, lo que proporciona una forma de identidad bioquímica. En el segundo grupo, las secuencias pueden ser cortas o largas (como las Alu y L1, respectivamente), y son elementos transponibles (Klug et al., 2006). En conjunto, los distintos tipos de ADN repetitivo constituyen hasta el 40 % del genoma humano, donde los 20.000 a 25.000 genes funcionales estimados en las primeras publicaciones

representan menos del 5 % del genoma. Observaciones como ésta son generales en eucariotas, reflejando una característica que parece estar compartida. Por ejemplo, en el erizo de mar *Strongylocentrotus purpuratus*, los 27.350 genes que codifican proteínas ocupan menos del 10 % de su genoma (Cameron et al., 2000; Klug et al., 2006). En *Drosophila melanogaster*, sólo el 5-10 % del genoma codifica proteínas. Además del ADN repetitivo, hay gran cantidad de secuencias de ADN de copia única que parecen ser no codificantes (Klug et al., 2006). Por otra parte parte, los genes pueden ser clasificados en 2 tipos: por un lado, los codificantes de proteínas (transcritos a ARNm y luego traducidos); y por otro lado, los no-codificantes, encargados de la síntesis de ARN funcionales que no se traducen (ARNt, ARNr, miARN, etc.), comúnmente denominados ARN no codificantes. No obstante, para cualquier tipo de gen, el proceso de pasar de ADN a producto funcional se conoce como expresión génica. Los genes son susceptibles de sufrir mutaciones u otros fenómenos asociados a reorganización estructural, pudiendo en principio no ser funcionales, en cuyo caso se denominan pseudogenes. Estos surgen por duplicación y posterior divergencia de un gen funcional; y aunque no se expresan, persisten en los genomas de los seres vivos y constituyen un recurso evolutivo para la especie, siendo regiones de ADN cuasi funcionales que pueden aceptar mutaciones (y generar nuevas funciones) sin perjuicio de las funciones que ya se desarrollan en el organismo (Lodish et al., 2005).

1.1.2. Cantidad de genes codificantes en los genomas

Durante las últimas décadas, los diversos proyectos de secuenciación genómica han expandido dramáticamente nuestro conocimiento acerca de la estructura y la organización de los genomas. Sin embargo, la relación entre el genoma y la información allí codificada presenta aún numerosos enigmas sin resolver. La discusión que se inició con el primer borrador del genoma humano acerca del número de genes (que variaba de 30.000 a más de 100.000, según distintas interpretaciones) aún permanece abierta. Los 35.000 genes anotados como codificantes de proteínas fueron sorprendentemente pocos o, al menos, insuficientes considerando que diferencias tan notorias como las existentes entre una mosca y un humano, fueran tan escasas a nivel de número y conservación de genes (no de tamaño de genoma).
Con el advenimiento de las nuevas tecnologías de secuenciación masiva, se descubrió que un porcentaje mayor al que se pensaba es convertido en ARNm (Berretta y Morillon, 2009). Esta transcripción generalizada, efecto denominado en inglés *pervasive transcription*, es más notoria en eucariotas y podría constituir un nuevo nivel regulatorio. Desde este punto de vista, las nuevas tecnologías de secuenciación masiva han puesto de manifiesto que los recientemente descritos micro ARN (miRNA) (Lee et al., 1993) y los ARN no codificantes (Jacquier, 2009), cuentan con numerosos miembros, agregando así mayor complejidad al proceso de regulación de la expresión de un genoma. Existe hoy evidencia de que el número de genes, no es relativamente tan pequeño ni tan mayoritariamente conservado como se creía (Berretta y Morillon, 2009; Managadze et al., 2013; Slavoff et al., 2013). Enumerar los agentes y comprender los mecanismos implicados en las nuevas capas de regulación de la expresión, es uno de los desafíos en genómica cuando se abre una nueva era en la capacidad de secuenciación y de análisis.

1.2. Regulación de la expresión génica en procariontes

Se ha demostrado que la expresión o transcripción de los genes de organismos procariontes, como las bacterias, está regulada. Los genes cuya expresión no está regulada se denominan constitutivos, y aquellos que responden a mecanismos de regulación son llamados inducibles. En las bacterias, al igual que en organismos eucariontes, es necesario regular la expresión de los genes adaptándola a las necesidades metabólicas circunstanciales (Klug et al., 2006). Por ejemplo, existen enzimas capaces de introducir en la bacteria diferentes tipos de azúcares, y enzimas específicas capaces de metabolizar cada uno de esos tipos. Por un principio de optimización de recursos, la célula expresa los genes inducibles necesarios en cada caso. Es decir, si una bacteria vive en un medio donde la principal fuente de carbono es la lactosa, solamente se expresarían los genes necesarios para metabolizar la lactosa; mientras que los genes asociados a enzimas que degradan otros azúcares no se expresarían. Gran parte de los genes estudiados en procariontes aparecen en tándem formando agrupamientos, en donde cada uno de los genes codifica proteínas funcionalmente relacionadas, y en muchos casos, la transcripción de estos genes. A este grupo de genes con funciones relacionadas y transcritos como una unidad, se le denomina operón. Normalmente, las proteínas codificadas por los genes de un operón son enzimas que intervienen en la misma vía metabólica. El ejemplo más conocido es el operón *lac*, que codifica las enzimas necesarias para la utilización del hidrato de carbono lactosa. La transcripción de los genes de un operón es inducible o reprimible, y a menudo, el producto metabólico final de las rutas biosintéticas sirve de inductor o de represor de la expresión génica (Lodish et al., 2005; Klug et al., 2006). Los ARNm que se sintetizan a partir de un operón se denominan policistrónicos o poligénicos, y son regulados como una unidad a partir de un único promotor. El resultado de la expresión de este grupo de genes es una única molécula de ARNm, portadora de la información de varios genes, llamado mensajero policistrónico.

La expresión eficiente de la información genética en las bacterias depende de mecanismos reguladores que ejercen control positivo o negativo sobre la transcripción; necesarios para que los genomas no se transcriban continuamente (Klug et al., 2006). Los niveles de expresión de un gen están determinados por la transcripción a ARNm y por la traducción de este en los ribosomas. En procariontes, los mecanismos de regulación a nivel de la transcripción se basan en una secuencia de ADN, que generalmente precede a la región codificante, conocida como promotor. Este es el sitio donde se une la enzima ARN polimerasa para iniciar el proceso de transcripción, en el cual se sintetiza una hebra de ARN a partir del molde de ADN en la dirección 5'$\rightarrow$ 3'. Además del promotor, existen en su vecindad sitios donde factores de transcripción pueden unirse a sitios regulatorios del ADN para modular el inicio de la transcripción. Por tanto, la frecuencia con la que un gen es transcrito no sólo depende de la afinidad de la ARN polimerasa por el promotor, sino también de la medida en que las regiones regulatorias y sus factores de transcripción asociados favorezcan o no la función de la ARN polimerasa. Con este tipo de estrategias y elementos, el organismo puede activar o desactivar la transcripción de un gen u operón, como respuesta a cambios en su entorno (Campbell y Farrell, 2004; Klug et al., 2006).

La existencia de concentraciones diferentes de proteínas codificadas por un mismo

operón es explicada por la regulación a nivel de traducción del ARNm. Un primer paso de regulación traduccional está definido por el sitio de unión ribosomal, que consiste en un grupo de nucleótidos en el ARNm localizados en la región anterior al codón de iniciación. La iniciación de la traducción del ARNm depende de ello. Distintos ARNm presentan diferentes afinidades por la unión a los ribosomas. Esto resulta en que el conjunto de los mensajeros de una misma bacteria exhiba rangos de tasas de traducción de hasta tres órdenes de magnitud (Lodish et al., 2005).
Durante la traducción, la información contenida en forma de tripletes de nucleótidos (codones) del gen se decodifica en secuencias de aminoácidos de la proteína. Los cuatro nucleótidos dan origen a 64 codones diferentes, 61 de los cuales codifican para cada uno de los 20 aminoácidos, mientras que otros tres codones señalan la terminación de la traducción. Este aspecto del código genético se denomina "degeneración" de los codones (Crick et al., 1961), y los diferentes codones que codifican el mismo aminoácido se denominan codones sinónimos. Numerosos estudios han demostrado que el uso de codones sinónimos es un proceso no aleatorio. Por el contrario, durante la traducción algunos codones se usan más que otros codones sinónimos (Plotkin y Kudla, 2011). Esta característica del uso preferencial de codones en el proceso de traducción se conoce como sesgo de uso de codones o *codon usage bias* (Sharp y Li, 1987; Zeng et al., 2008). La variación en el uso de codones entre los genes proporciona un diferencial en la eficiencia traduccional y consiste en un mecanismo para regular la expresión a nivel de la traducción.

1.3. Regulación de la expresión génica en eucariontes

La expresión génica en organismos eucariotas es mucho más compleja, aunque comparten mecanismos antes mencionados con las bacterias. Sin embargo, la heterogeneidad tanto morfológica como funcional de las células en organismos pluricelulares hace necesario que existan mecanismos de control precisos de la expresión génica en las diferentes células del organismo, de modo que éstas realicen sus funciones adecuadamente. En bacterias, los sistemas de regulación de expresión de los genes son relativamente sencillos. Por otro lado, en eucariotas pluricelulares, el objetivo de la regulación de la expresión génica es garantizar la ejecución de las decisiones precisas desde la división celular o el desarrollo embrionario, que llevan a la diferenciación celular o funciones específicas de cada tipo celular. En otras palabras, garantizar que el gen correcto se active en la célula correcta, en el momento correcto.
Los organismos eucariotas tienen determinadas características que hacen que sus mecanismos de regulación génica sean diferentes de los procariotas. A nivel celular, hay mecanismos que activan porciones específicas del genoma (control positivo) y que reprimen la expresión de otros genes (control negativo). Las células eucariotas contienen mucha más información genética, repartida en varios cromosomas rodeados por la doble membrana nuclear, y su ADN está condensado en la cromatina mientras se encuentra inactivo (Klug et al., 2006). Sin embargo, sólo el 7% del ADN del genoma eucariota es codificante. El resto del genoma está constituido por secuencias, en muchos casos repetidas miles de veces, a veces en tándem, cuyo rol aún se desconoce en su gran mayoría. Además, el gen y la proteína no son colineales ya que el transcripto primario del ARNm posee intrones que se pierden durante su

maduración.
A diferencia de los procariotas, los ARNm eucariotas son en su gran mayoría monocistrónicos y no poseen operones. Así, el sistema de regulación tiene múltiples etapas. Esta multiplicidad de niveles sucesivos de regulación permiten un ajuste de la velocidad y de la intensidad de la reacción a los estímulos (Lodish et al., 2005; Klug et al., 2006). No existe, pues, un modelo general de regulación como en procariontes, sino toda una serie de posibilidades que se encadenan, desde la estructura de la cromatina hasta una regulación postraduccional, última etapa posible de regulación de la expresión. Así, la regulación de la expresión génica en organismos eucariontes puede darse a distintos niveles, que se detallarán brevemente a continuación:

- Regulación de la expresión génica a nivel de cromatina.
- Regulación de la expresión génica a nivel transcripcional.
- Regulación de la expresión génica a nivel postranscripcional.
- Regulación de la expresión génica a nivel traduccional.
- Regulación de la expresión génica a nivel postraduccional.

1.3.1. Regulación a nivel de cromatina

La organización de la cromatina en el núcleo desempeña una función importante en la regulación de la expresión génica en eucariotas (Klug et al., 2006). La cromatina está constituida por el ADN enrollado alrededor de una serie de nucleosomas, empaquetada de forma más relajada en las regiones que contienen genes activos. La condensación/descondensación de la cromatina representa el primer y más riguroso nivel de regulación. Es un nivel epigenético que determina los genes que la célula necesita transcribir y aquellos que no debe transcribir. Además de los cambios generales que ocurren en las regiones activas o potencialmente activas, ocurren cambios estructurales en sitios específicos asociados con la iniciación de la transcripción. Por ejemplo, la metilación, que tiene lugar generalmente en los dupletes CG (citosina-guanina) del ADN. La función primordial de la metilación está asociada al control de la transcripción (Lodish et al., 2005). Un gen con elevado nivel de metilación es inactivo, y no metilado es activo. Al igual que sucede con otros cambios en la cromatina (por ej. modificaciones histónicas como la acetilación o forsforilación), es consenso que la ausencia de grupos metilo esté asociada con la posibilidad de transcripción y no con el propio acto de la transcripción. Es decir, niveles altos de metilación se asocian a niveles bajos de expresión génica, mientras que si un gen se expresa, no está metilado o tiene un bajo nivel de metilación. Por eso, la metilación representa una de las diversas maneras en que los cambios genómicos pueden regular la expresión génica. En este sentido, se ha observado que las proteínas que se unen a la 5-metilcitosina podrían reclutar correpresores o histonas desacetilasas para remodelar la cromatina, o bien, complejos de remodelación de nucleosomas (Klug et al., 2006). De hecho, la metilación del ADN es uno de los mecanismos epigenéticos implicados en la regulación de la expresión génica en mamíferos. Los patrones de metilación son específicos para cada especie y tipo de tejido; y vitales para mantener el silenciamiento génico en el desarrollo normal (Nakao, 2001; Richards y Elgin, 2002; Burgers et al., 2002). Alteraciones en este proceso pueden afectar las interacciones ADN-proteína, la estructura y replicación

del ADN y su consecuente expresión, además de aumentar el riesgo de mutaciones espontáneas que derivan en patologías; lo que hace de este tópico un área activa de investigación (Costello, 2001; Paulsen y Ferguson-Smith, 2001; Hendrich, 2001).

1.3.2. Regulación a nivel transcripcional

Los genes activos de una célula eucarionte, determinan en gran parte su identidad y características. La transcripción de un gen está controlada principalmente en la iniciación, por la interacción de la ARN polimerasa con su promotor. En los eucariotas, las secuencias necesarias para la regulación de la transcripción se remontan mucho más en la dirección 5' UTR que el promotor. En esta región se encuentran una serie de elementos *CIS* sobre los cuales se fijan los factores, como por ejemplo, las cajas TATA, GC y CAAT (secuencias de nucleótidos de consenso, o canónicas). La relación entre el promotor y la ARN polimerasa está regulada por la presencia de otros elementos *CIS* denominados amplificadores o silenciadores, que tienen efecto cuando interactúan con elementos *TRANS* (proteínas reguladoras, llamadas factores de transcripción). Los sistemas reguladores *CIS-TRANS* en eucariotas pueden ser muy complejos y contar con miles de elementos regulatorios. Algunas de estas proteínas de regulación activan la transcripción y otras la inhiben (Campbell y Farrell, 2004; Lodish et al., 2005; Klug et al., 2006). Los elementos *CIS* contienen secuencias cortas y se pueden encontrar en diferentes genes relacionados, aunque no sean necesariamente idénticos.

1.3.3. Regulación a nivel postranscripcional

Incluso después de que un gen ha sido transcrito, su expresión puede ser regulada a través de modificaciones en el transcrito primario (pre-ARNm) por pequeños ARN y factores de iniciación de la traducción. Cuando un gen eucariota se transcribe en el núcleo, el pre-ARNm es considerado un mensajero inmaduro y debe pasar por ciertas modificaciones para convertirse en ARNm maduro, y así poder salir del núcleo. Dichas modificaciones incluyen la adición del casquete (CAP) en el extremo 5' UTR, la adición de múltiples adenosín monofosfatos en el extremo 3' UTR (cola poli-A o poliadenilación), y el proceso de corte y empalme o ajuste (denominado en inglés *splicing*). Estas modificaciones pueden ser potencialmente reguladas, acelerando o retrasando el producto final, o bien dando lugar a un producto diferente. Además, tales regulaciones en la expresión de algunos genes involucran procesos inherentes a la estabilidad y almacenamiento de los ARNm que serán transportados hacia el citoplasma, así como el control de su duración y traducción por microARN.

Empalme alternativo: El empalme alternativo es muy común en eucariotas. Consiste en la eliminación de intrones y la unión de exones en el pre-ARNm, dando lugar a diferentes ARNm maduros y/o proteínas del mismo transcrito de ARN. La Fig. 1.2 (paneles A y B) esquematiza el proceso de empalme alternativo en un transcripto primario, en el cual se escinden los intrones y se unen los exones, produciendo una secuencia de ARNm maduro. El empalme alternativo es un proceso controlado por proteínas reguladoras que se unen a sitios específicos en los intrones, indicando a los factores de corte y empalme los exones que deben utilizarse. Diferentes tipos de células pueden expresar distintas proteínas reguladoras, por tanto, las combinaciones alternativas de exones en cada tipo de célula conducen a la pro-

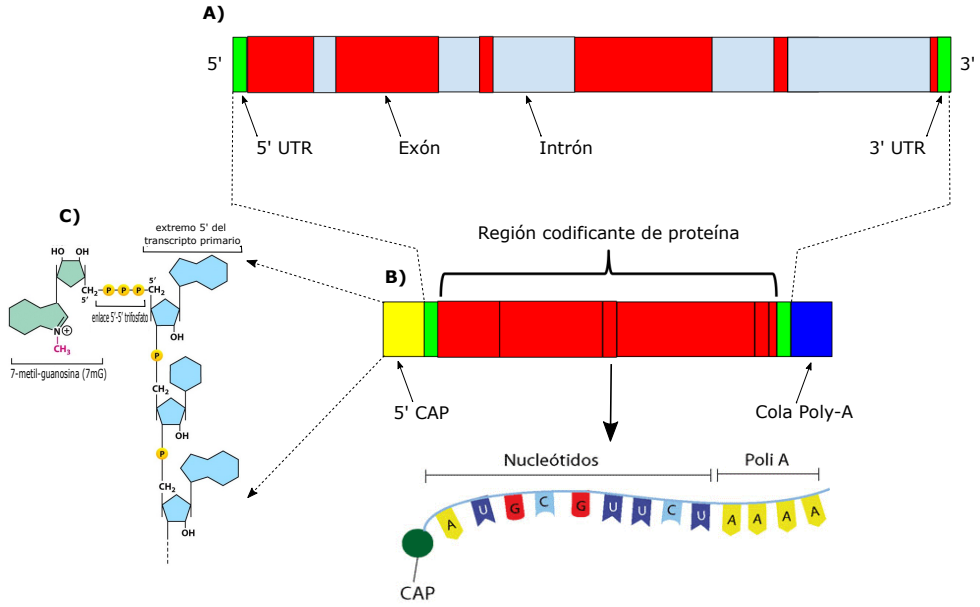

Figura 1.2: Representación esquemática del procesamiento completo del transcripto primario (pre-ARNm). **A)** Modelo de un pre-ARNm recientemente transcrito, con sus respectivos intrones y exones. **B)** Modelo de un ARN mensajero maduro (ARNm), listo para ser exportado del núcleo, tras el *splicing* y la adición de la caperuza 5' CAP y la cola 3' poly-A en los extremos UTR. La región codificante de proteína corresponde a la secuencia nucleotídica, comprendida entre los extremos UTR (lindados a su vez por la caperuza CAP y la cola poli-A). **C)** Estructura química del CAP unida al extremo 5' del ARNm mediante un enlace 5'-5' trifosfato.

ducción de diversas proteínas. El empalme alternativo puede introducir un codón de terminación o cambiar el marco de lectura de un exón (Lewin, 2000).

Adición del casquete (CAP): Luego de la eliminación de intrones del transcrito primario, y antes de que el ARNm esté listo para ser exportado, sufre otra modificación postranscripcional que consiste en adicionar un casquete o caperuza (CAP) al extremo 5' UTR mediante un enlace fosfodiéster. La estructura CAP es un ribonucleótido modificado de guanina, el 7-metil-guanosina (7mG), unido al primer nucleótido del pre-ARNm (generalmente adenina) por un enlace 5'-5' trifosfato (designado también como 7-metil guanosin trifosfato, m7GpppN, donde N es cualquier nucleótido). Esto hace que la guanosina añadida se una en sentido opuesto al del resto de la cadena polinucleotídica. En la Fig. 1.2 (panel C) se esquematiza la estructura química de la proteína CAP unida al extremo 5' UTR del ARNm mediante un enlace 5'-5' trifosfato.
El CAP es importante para el procesamiento posterior en el núcleo, protegiendo el extremo 5' UTR del transcripto inicial ante el ataque de nucleasas. Además, está implicada en el transporte del ARNm maduro hacia el citoplasma, y tiene la función de reconocer el primer codón del ARNt para iniciar la traducción.

Poliadenilación: La poliadenilación es la adición de múltiples adenosín monofosfatos en el extremo 3' UTR del ARNm. Es decir, se produce un tramo de poliadenilato conformado por bases de adenina (cola poli-A), anexadas a partir de una secuencia o señal de poliadenilación (AATAAA), situada unos 11-30 nt antes del extremo 3' UTR. El proceso de poliadenilación ocurre en alguno de varios sitios posibles, generando distintos mensajeros. La elección del sitio de poliadenilación depende probablemente de factores celulares específicos de tejido (Klug et al., 2006). La cola de poli-A es importante para permitir la exportación nuclear del ARNm, y es un determinante en la vida media de los mensajeros. En tanto el ARNm no es traducido, la cola de poli-A es gradualmente acortada, hasta que el ARNm es enzimáticamente degradado. La Fig. 1.2 representa esquemáticamente el procesamiento del transcripto primario (pre-ARNm) con la adición de la proteína CAP y la cola poli-A añadidas en los extremos 5' UTR y 3' UTR, respectivamente, además del *splicing*.

Almacenamiento de los ARNm: Además del procesamiento del transcripto primario para dar lugar al ARNm maduro, se sabe que los mensajeros maduros se almacenan en los cuerpos de procesamiento ubicados en el citoplasma. De hecho, se encuentran genes que son transcriptos pero no se observan sus productos de traducción. Sin embargo, aún se desconocen los mecanismos de regulación de este almacenamiento, siendo objeto de mucho estudio (Balagopal y Parker, 2009; Layana et al., 2012).
Por otro lado, se ha observado que después de un estímulo dado, la variación en la concentración de proteínas puede ser mucho más rápida que la velocidad de la transcripción, sugiriendo la liberación de un ARNm que estaba almacenado (Lewin, 2000).

Regulación por microARN: Existe una clase de pequeños ARN reguladores, los micro-ARNs (miARNs), que puede controlar la cantidad de ARN maduro mediante su silenciamiento o degradación (Sonenberg y Hinnebusch, 2009; Cavagnari, 2012). Además, están implicados en la regulación de varios procesos biológicos como la diferenciación celular, la proliferación, la apoptosis y el desarrollo embrionario y tisular (Ambros, 2004). Los miARN son pequeños fragmentos de cadena simple de 18-25 nucleótidos que constituyen una extensa familia de genes reguladores. Conforman un complejo de ARN-proteína dirigido hacia ARNm específicos (que poseen una secuencia complementaria al miARN). Cuando el complejo ARN-proteína se une, condiciona la traducción o determina la degradación del ARNm (Carthew y Sontheimer, 2009; Suh et al., 2011). Por otra parte, los ARN de doble hebra, exclusivos de algunas familias virales, también tienen la habilidad de interferir específicamente los mensajeros con secuencias similares, cuya función es la interferencia o el silenciamiento (Dimmock et al., 2007; Patton, 2008).

1.3.4. Regulación a nivel traduccional

La traducción de un transcripto puede aumentarse o inhibirse por diversos reguladores. Como mencionamos anteriormente, los miARN pueden inhibir la traducción de sus ARNm objetivos, pero existen otras maneras de regular la traducción en una célula.
El proceso de traducción de ARNm en eucariontes consta de cuatro etapas: iniciación, elongación, terminación y reciclamiento. Cada una de ellas es catalizada por diferentes grupos de proteínas: factores de iniciación, de elongación, y de ter-

minación, representados en el esquema de la Fig. 1.3. La regulación de las diferentes etapas de traducción permite la síntesis diferencial de proteínas específicas, resultando en cambios fisiológicos en la célula, sin que necesariamente cambie la transcripción de los genes correspondientes (Groppo y Richter, 2009).

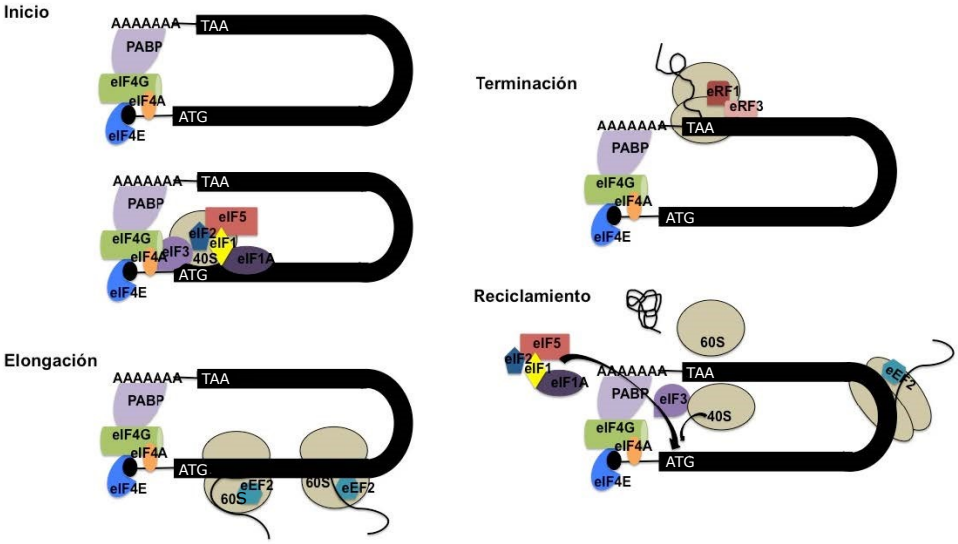

Figura 1.3: Esquema representativo de las cuatro etapas de regulación durante la traducción. **Iniciación:** el 5' CAP del ARNm se une al complejo factorial eIF4F, que junto a otros factores (eIF4G, eIF3, eIF5, eIF1, eIF1A) ensamblan el complejo de pre-inicio 43S (GTP, ARNt-Met, eIF2) y la subunidad ribosomal 40S. Esta interacción conforma el complejo de inicio 48S que reconoce el codón ATG, donde se une la subunidad ribosomal 60S y se liberan los factores de inicio. **Elongación:** comienza con la formación de la cadena polipeptídica, eEF1A lleva los ARNt al ribosoma, mientras que eEF2 promueve la traslocación. **Terminación:** al encontrar un codón de parada, eRF1 estimula la hidrólisis de la cadena peptídica. La posterior unión de eRF3-GTP permiten el desensamble del ribosoma, ARNt y factores. **Reciclamiento:** eIF3 promueve la disociación del complejo post-terminación, permitiendo la futura unión de la subunidad 40S con los factores de inicio. (Tomada de Martínez & Dinkova 2010.)

Iniciación: En primer lugar, se ensambla sobre el ARNm un complejo de proteínas y factores de inicio (eIFs). El 5' CAP del transcripto es reconocido por el factor de iniciación eucariota 4 (eIF4E), unido a la proteína de anclaje eIF4G. A éstos se une la helicasa eIF4A (que facilita el desenrollamiento de estructuras secundarias en el ARN), conformando así el complejo eIF4F. Estas interacciones facilitan la exportación nuclear del ARNm y determinan su localización citoplasmática para ser traducido, almacenado, o degradado, dependiendo de los estímulos celulares. Distintos factores del tipo eIF4E poseen especificidad para diferentes tipos de CAP y proteínas, lo que constituye un mecanismo adicional para la regulación de la

expresión genética (Fischer, 2009; Kaye et al., 2009). Luego, eIF4G recluta a la proteína PABP de unión a poly(A), estimulando la actividad de la ARN helicasa y protegiendo al transcrito de la acción de nucleasas. El complejo eIF4F junto con otros factores facilitan el ensamblaje del ribosoma sobre el ARNm, cuya ubicación correcta está asegurada por diversas proteínas auxiliares y por la unión del factor eIF2 a la subunidad pequeña del ribosoma (40S). Esta unión es controlada por la adición de un grupo fosfato a eIF2; que al fosforilarse experimenta un cambio de forma y bloquea la traducción. Otros controles de la expresión génica a nivel traduccional son la concentración citoplasmática de ARNt y de ribosomas (Sonenberg y Hinnebusch, 2009). Una vez formado el complejo de inicio 48S, éste recorre el ARNm en dirección 5'→ 3' buscando el codón de inicio ATG. Cuando el Met-ARNt reconoce el ATG, se disocian la mayoría de los factores de inicio y ocurre la unión de la subunidad ribosomal 60S para formar el ribosoma 80S, listo para comenzar la elongación del péptido. En este evento, el complejo eIF4F permanece unido al ARNm para permitir que la maquinaria de inicio de la traducción pueda reciclarse y comenzar otro evento de iniciación (Fischer, 2009; Groppo y Richter, 2009).

Elongación: la elongación de la traducción es un proceso más simple, que requiere mantener el marco de lectura, seleccionar y entregar correctamente los ARNt aminoacilados al ribosoma 80S, y formar los enlaces peptídicos. Sólo son requeridos tres factores de elongación (eEFs): eEF1A, que unido a GTP ayuda a cargar los aa-ARNt correctos al ribosoma; eEF1B, necesario para el intercambio de GDP por GTP en eEF1A; y eEF2, el cual mediante hidrólisis de GTP promueve la translocación del ribosoma exactamente tres nucleótidos sobre el ARNm. Se han observado indicios de regulación en la etapa de elongación, por ejemplo, antes de la mitosis celular. Previo a dividirse, las células reducen su velocidad de elongación, probablemente mediante la fosforilación del factor eEF2 por una quinasa específica; y reanudan la síntesis de proteínas al entrar a la fase G1 del ciclo celular (Groppo y Richter, 2009).

Terminación: la terminación de la traducción eucarionte es mediada por el factor de liberación eRF1, el cual se une al ribosoma en lugar del ARNt para reconocer cualquiera de los tres codones de parada (TAA, TAG o TGA), induciendo la hidrólisis de la proteína recién sintetizada del último ARNt. Posteriormente, el factor eRF3 unido a GTP promueve la liberación de eRF1, mediante un nuevo evento de hidrólisis (Pestova et al., 2007).
Existen vías que discriminan los ARNm con codones de parada aberrantes, sin sentido o que carecen de ellos. Se ha sugerido que el ribosoma se instala en la cola de poly(A) incrementando la terminación prematura de ribosomas río arriba, lo que resulta en una represión traduccional de un ARNm sin codón de parada (*nonSTOP mRNA*) (Groppo y Richter, 2009). También, los codones de parada pueden ser redefinidos como codones de sentido, continuando así la traducción hasta un próximo codón de parada en el ARNm. Estos casos se conocen como mecanismos de recodificación, sobre el cual nos explayaremos en la sección 1.4.1.

Reciclamiento: tras la terminación de la síntesis de proteínas y la liberación de la cadena polipeptídica, el ribosoma queda con el sitio A vacío y un ARNt desacilado en el sitio P; lo cual se conoce como complejo de post-terminación. Para que el ribosoma pueda iniciar una nueva traducción, este complejo debe disociarse para permitir la unión de la subunidad ribosomal 40S con los factores de inicio (eIFs) en el sitio donde comienza la traducción del ARNm. En bacterias, el desmontaje del complejo de post-terminación es un proceso activo catalizado por el factor de

reciclaje del ribosoma (RRF), que junto con el factor de elongación (EF-G), actúa liberando al ribosoma del ARNm. Este RRF es esencial para la viabilidad en procariontes; sin embargo, en eucariontes no existe un factor homólogo, y el mecanismo que precede al estado de terminación es diferente. En eucariotas, el factor eIF3 (en conjunto con eIF1 y eIF1A) promueve la disociación de la subunidad 60S ribosomal en el complejo de post-terminación, y de ARNt y ARNm unido a la subunidad 40S. El eIF1 interviene en la liberación del ARNt del sitio P, mientras que el eIF3 favorece la disociación de ARNm (Pisarev et al., 2007).

1.3.5. Regulación a nivel postraduccional

Las enzimas se sintetizan como precursor inactivo, y se activan mediante modificaciones bioquímicas en su estructura, que conforman un centro activo donde se ejecuta la catálisis. Tales modificaciones se producen por cortes proteolíticos o la adición de grupos químicos; como por ejemplo fosforilaciones, acetilaciones, metilaciones, ribosilaciones, glicosilaciones, entre otras. La unión de grupos prostéticos a glicoproteínas y lipoproteínas es otra forma de regulación de la expresión de los genes en eucariotas, a nivel postraduccional (Lewin, 2000). De esta forma, una proteína en su estado inactivo se puede "encender" por una modificación química simple y rápida, sin la necesidad de una nueva transcripción y traducción que demandan tiempo y energía. A veces, las modificaciones químicas también pueden determinar la ubicación de una proteína en la célula, donde se requiere su función (Klug et al., 2006). Las acuaporinas, por ejemplo, son proteínas de canal transmembrana que transportan agua, y a veces también pequeños solutos sin carga (las acuagliceroporinas). Están asociadas a la membrana de vesículas de almacenamiento intracelular, y cuando son necesarias, dichas vesículas se incorporan a la membrana basal. Las acuaporinas se regulan a partir de un cambio conformacional en la estructura de la proteína que abre o cierra el poro, y también en función de la cantidad presente en una membrana biológica, implicando la regulación a nivel transcripción/traducción (Törnroth-Horsefield et al., 2010; Kreida y Törnroth-Horsefield, 2015).
Los efectos de las modificaciones postraduccionales no son únicos y varían según la proteína: algunas son activadas o desactivadas por fosforilación, mientras que otras simplemente cambian su comportamiento (interactúan con un compañero diferente o se ubican en otra parte de la célula). Sin embargo, entre las modificaciones que sufren las proteínas se encuentra el mecanismo de ubiquitinación, consistente en la adición de una o varias moléculas de ubiquitina, una proteína pequeña, de manera covalente a proteínas blanco en residuos de lisina (Zamudio-Arroyo et al., 2012). Las proteínas modificadas por ubiquitinación son llevadas al proteosoma de la célula, donde son degradadas. De este modo, la ubiquitinación es una manera importante de controlar la persistencia de una proteína en la célula (Campbell y Farrell, 2004; Lodish et al., 2005).
En síntesis, la expresión de la información genética atraviesa distintos niveles de regulación, que involucran diversos factores moleculares y reacciones. Si bien los mecanismos de control más importantes son los que actúan a nivel transcripcional sobre las secuencias reguladoras, también se producen regulaciones durante el procesamiento o maduración del ARNm, su transporte al citoplasma o su supervivencia en el citosol. Finalmente, los controles actúan a nivel traduccional o regulando la actividad de la proteína. A continuación, la Fig. 1.4 resume los principales niveles de control de la expresión génica, explicados hasta aquí.

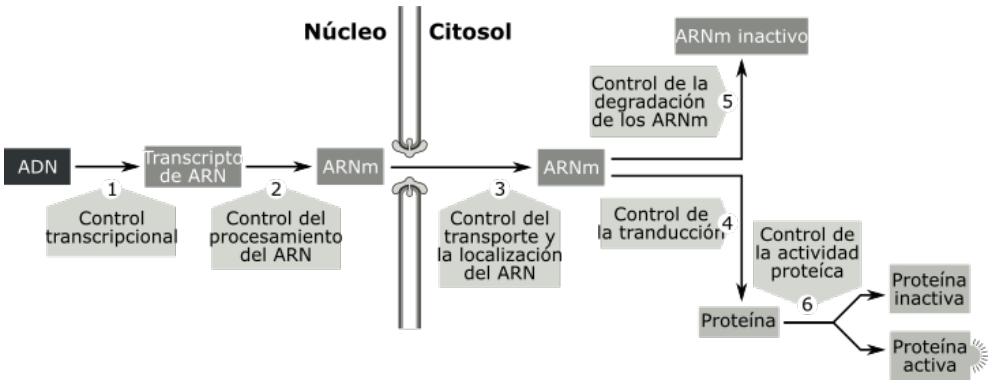

Figura 1.4: Principales niveles de regulación de la expresión génica. **1)** factores de transcripción, grado de condensación de la cromatina y grado de metilación. **2)** adición del 5'CAP y la cola poly-A, empalme alternativo. **3)** mecanismos que determinan el paso del ARNm maduro al citosol. **4)** mecanismos que determinan si el ARNm presente en el citosol es o no traducido. **5)** mecanismos que determinan la supervivencia del ARNm en el citosol. **6)** mecanismos de activación/desactivación de una proteína para la ejecución de una función determinada, así como el tiempo de supervivencia de la misma. (Tomada de Alberts *et al.* (2016) "Molecular Biology of the Cell").

1.4. El auge de la genómica

El conocimiento del genoma de un organismo es un recurso de información importante para la comprensión de la biología del mismo. El desafío siguiente al secuenciamiento de un organismo es su anotación. La anotación de un genoma tiene dos aspectos básicos: uno estructural (la delimitación física de sus componentes) y otro funcional (la asignación de una potencial función) (Rouzé et al., 1999). Estructuralmente, los genes tienen que ser asignados a regiones específicas del genoma. También se debe descifrar su estructura, determinando sus intrones, exones, sitios de inicio y de parada de la transcripción, así como sus elementos reguladores.
Históricamente el desarrollo de las técnicas de anotación ha ido a la par con el proceso de secuenciación de los genomas. Por ejemplo, la identificación de genes codificantes se realizaba mediante la búsqueda de secuencias homólogas conocidas en otras especies. Una alternativa era la búsqueda *ab initio*, que consiste en buscar signos o patrones indicadores de promotores, o regiones codificantes y distinguirlos de secuencias formadas por simple azar. Para lograr dicho objetivo, los predictores *ab initio* utilizan conjuntos de secuencias de entrenamiento que incluyen las señales que se quieren buscar y modelos probabilísticos, como los modelos ocultos de Markov. En cualquier caso, las predicciones deberán ser validadas para confirmar que efectivamente se transcriben y/o traducen (Pearson, 2006). Estas técnicas de anotación basadas en homología o métodos *ab initio* conllevan un vicio, dado que se fundamentan en aquello que es conocido, y tienen por lo tanto un sesgo que lleva a descubrir sólo genes conocidos en otros organismos, o que satisfacen señales conocidas de antemano, dejando de lado las secuencias novedosas. Sin embargo, hubo un gran salto tecnológico en las tecnologías de secuenciación, marcando un

punto de inflexión en la generación de datos y nuevos genomas. Gracias a las nuevas tecnologías de secuenciación de ADN y a los desarrollos de técnicas bioinformáticas para analizar estos datos, la genómica ha tenido un desarrollo importante en los últimos años. Hoy en día, la tarea de determinar la estructura de cada gen incluyendo sus intrones, exones, sitios de inicio y de parada de la transcripción, se lleva a cabo mediante métodos *ab initio*, pero asistidos por datos transcriptómicos obtenidos por las nuevas tecnologías de secuenciación, como el *RNA-Seq*. De este modo, es posible determinar en forma precisa los marcos abiertos de lectura que, siempre que sean transcriptos, tienen la potencialidad de ser traducidos a proteínas. La determinación de estos marcos de lectura es el tema principal a ser tratado en esta tesis. Esos marcos, que serán denotados por la sigla ORF (del inglés *Open Reading Frames*), están definidos por un codón de inicio y el primer codón de terminación encontrado en el mismo marco de lectura. En principio, por combinación al azar de los nucleótidos siempre existe una probabilidad de que este patrón sea espurio, es decir, sin un correlato proteico. Es un hecho que se observa en miles de pequeños marcos de lectura de esta naturaleza. Sin embargo, las probabilidades de encontrar un ORF espurio disminuyen con su tamaño, por eso en los procesos de anotación automatizada de genomas solo son considerados, para un subsecuente análisis de anotación, aquellos ORFs mayores a cierto tamaño (alrededor de 200 pares bases). Así, con este tipo de anotación, pequeños marcos abiertos de lectura, que denotaremos como sORF ("s" por *small*), asociados a genes funcionales, podrían escapar a los protocolos de detección, y estarían siendo descartados entre los miles de pequeños ORFs sin sentido biológico formados al azar (Basrai et al., 1997; Wang et al., 2003). En general, tampoco se anotan los ORF pequeños anidados dentro de otros más grandes. A pesar de ello, se han descubierto ORFs traducidos cadena arriba o abajo de los ORFs anotados; así como variantes de proteínas expresadas a partir de codones de inicio alternativos, tanto adentro como afuera del marco de lectura de ORFs anotados. También, se han encontrado sORFs en ARN ribosomales, y pequeñas proteínas funcionales codificadas por ARN previamente considerados no codificantes (Aspden et al., 2014; Chugunova et al., 2018; Orr et al., 2020). A partir de esto, se infiere que muchos ORFs previamente ignorados pueden ser traducidos como mecanismos reguladores o para dar proteínas bioactivas. A continuación, la Fig. 1.5 resume la visualización del conjunto de moléculas de ARN que codifican sORF.

Desafortunadamente, los sORFs son difíciles de predecir mediante los métodos bioinformáticos usuales. Sin embargo, genes que codifican péptidos de pocos aminoácidos se encuentran en transcritos mono o policistrónicos en bacterias (Hemm et al., 2008), levaduras (Basrai et al., 1997; Kastenmayer, 2006), plantas (Rohrig et al., 2002; Hanada et al., 2013) y animales (Savard et al., 2006; Galindo et al., 2007). En *Drosophila melanogaster* se encuentran más de 300 péptidos codificados a partir de sORFs con variadas funciones, tales como feromonas de apareamiento, metabolismo de energía, reguladores transcripcionales, nucleasas, quelantes de iones metálicos, hormonas, péptidos antibacterianos, reguladores de quinasas, entre otros (Kastenmayer, 2006; Ladoukakis et al., 2011). Algunos autores han sugerido que los péptidos traducidos a partir de sORFs, podrían ser el equivalente proteico de los miRNA, definiéndose de este modo una nueva capa de regulación (Brennecke et al., 2003; Ambros, 2004).

Otro de los desafíos actuales en la anotación de genomas, también asociado a los ORFs, consiste en la delimitación precisa de los mismos. Específicamente,

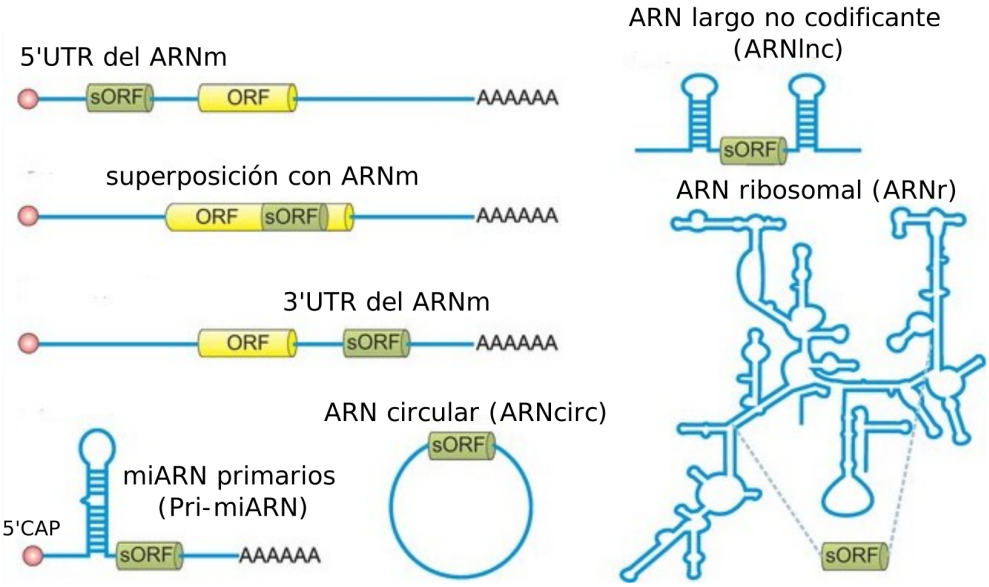

Figura 1.5: Esquema de las moléculas de ARN que codifican ORFs cortos (sORF). El círculo rojo indica la caperuza CAP en la región no traducida 5' UTR de un marco de lectura abierto (ORF). (Adaptada de Chugunova *et al.* 2017).

la biosíntesis de proteínas se completa tras el reconocimiento de uno de los tres codones de terminación de la traducción (TAA, TAG y TGA) por parte de la maquinaria ribosomal. Sin embargo, existen excepciones que pueden conducir a los ribosomas a continuar la traducción más allá del primer codón de terminación. Consecuentemente, una fracción de las proteínas resultantes incluyen aminoácidos adicionales hasta un segundo o tercer codón de terminación en el mismo marco de lectura (Robinson y Cooley, 1997; Jungreis et al., 2011; Freitag et al., 2012; Artieri y Fraser, 2014; Loughran et al., 2014; Baranov et al., 2015). Denominaremos este fenómeno como lectura del codón de parada (LCP), dado que en la literatura científica se conoce como *Stop Codon Readthrough* o *Translational Readthrough*. Se han encontrado numerosos ejemplos de lectura del codón de parada que dan lugar a una proteína extendida con función biológica alternativa (von der Haar y Tuite, 2007; Schueren y Thoms, 2016; Sapkota et al., 2019). Sin embargo, en algunas ocasiones este fenómeno, puede estar vinculado a un fallo en la terminación de la traducción programada en el codón de parada (Bidou et al., 2010; Li y Zhang, 2019). Durante los últimos años el LCP, funcional o no, ha sido identificado y documentado en varios organismos. Por ejemplo, según la base de datos FlyBase (Tweedie et al., 2009; Gramates et al., 2017) (versión FB2020_04, publicado en Agosto 18, 2020), en la mosca de la fruta existen anotados 486 transcriptos asociados con eventos de LCP.

La Fig. 1.6 esquematiza el mecanismo para la ocurrencia de LCP. La eficiencia de la terminación de la traducción depende de la competencia entre el reconocimiento del codón de terminación por factores de liberación (eRF1 y eRF3 en eucariotas), y la decodificación del codón de parada por un ARNt cognato afín que puede

emparejarse con dos de las tres bases del codón de parada. El factor eRF1 reconoce el codón de parada en el sitio A ribosomal a través de su dominio N-terminal, y desencadena la hidrólisis del peptidil-ARNt activando el centro de la peptidil transferasa del ribosoma a través de los motivos NIKS y GGQ conservados en los dominios 1 y 2, respectivamente (panel A). El reconocimiento de un codón de terminación en el sitio A por parte del factor de liberación eRF1, desencadena la terminación de la traducción y la síntesis polipeptídica. El dominio C-terminal de eRF1 está involucrado en la unión de eRF3 GTPasa, cuya actividad cataliza la escisión de peptidil-ARNt; y eRF3 actúa como un factor de corrección durante la terminación (panel B) (Bidou et al., 2010; Dabrowski et al., 2015). Por otra parte, la LCP puede ser promovida por el contexto de nucleótidos o por fármacos, provocando una recodificación del codón de parada (explicado en la sección 1.4.1). En tales casos, los ribosomas incorporan un ARNt supresor natural en el codón de parada, continuando la traducción en el mismo marco de lectura hasta el siguiente codón de parada, y resultando en la expresión de una proteína con una nueva función (Blanchet et al., 2014). Este mecanismo de LCP mediado por ARNt supresores se muestra en el panel C de la Fig. 1.6.

1.4.1. LCP como mecanismo de recodificación

Los errores de codificación ocurren azarosamente aunque en baja frecuencia, y sus productos suelen no ser funcionales. Por otro lado, en ciertas ocasiones los codones de parada pueden ser recodificados en forma programada mediante señales específicas presentes en el ARNm (Baranov et al., 2002; Namy et al., 2004; Ribas de Pouplana et al., 2014). Por ejemplo, la redefinición de los codones TAG y TGA puede conducir a la especificación de glutamina y triptófano o selenocisteína, respectivamente (Atkins et al., 2007). Cuando esta redefinición se utiliza para la expresión génica, las señales de estimulación (o señales de codificación), incrustadas en el ARNm o en su producto peptídico naciente, facilitan dicha recodificación (Gesteland, 1996; Atkins et al., 2007).
El panel C de la Fig. 1.6 anterior esquematiza un ejemplo de recodificación de codones de terminación, que resulta en LCP. Los ARNt supresores leen a través de codones de terminación naturales. Los ARNt supresores son ARNt aminoacilados con anticodones complementarios a los codones de parada en el ARNm. Al emparejarse con 2 de los 3 nucleótidos del codón, un ARNt cognato puede competir exitosamente con el factor de liberación eRF1 por el reconocimiento de un codón de parada, y ubicarse en el sitio A del ribosoma, conduciendo a la decodificación de un codón de terminación (Dabrowski et al., 2015). De esta forma, el ARNt supresor inhibe el proceso de terminación de la síntesis de proteínas, y como resultado, un aminoácido se incorpora "erróneamente" a la cadena polipeptídica y el ribosoma continúa la traducción al siguiente codón de parada en el mismo marco de lectura.
A pesar de que los eventos de recodificación estén poco identificados o mal anotados en los genomas, están presentes en todas las formas de vida y aumentan significativamente la diversidad de los polipéptidos producidos en las células. Diversos estudios han demostrado que la traducción está parcialmente regulada por LCP en virus (Valle et al., 1992; Cimino et al., 2011; Firth y Brierley, 2012), en *D. melanogaster* (Robinson y Cooley, 1997; Steneberg y Samakovlis, 2001; Beier, 2001; Sato et al., 2003; Jungreis et al., 2011; Dunn et al., 2013), en hongos (Freitag et al., 2012), plantas (Lao et al., 2009) y mamíferos (Loughran et al., 2014; Schueren et

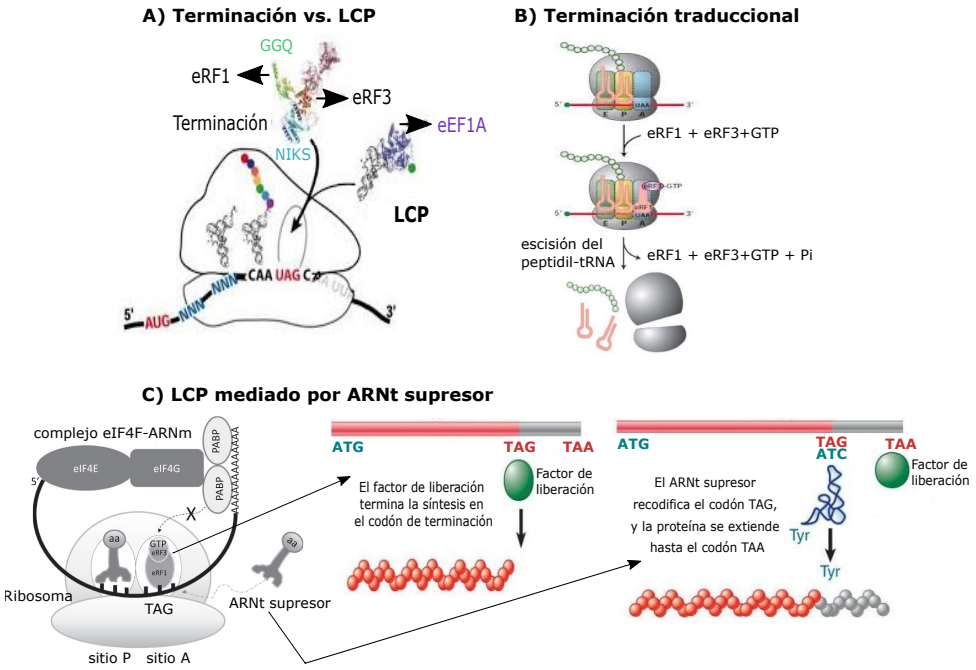

Figura 1.6: Esquema del mecanismo de la terminación eficiente de la traducción versus la lectura de codones de parada (LCP). **A)** La terminación de la traducción está mediada por los factores eRF1 y eRF3. El codón de parada (TAG) se ubica en el sitio A ribosomal, mientras que los sitios P y E están ocupados por ARNt. El ARNt supresor natural entrante se muestra unido al factor de elongación (eEF1A). **B)** Unión de eRF1 con eRF3 y GTP tras reconocer el codón de parada en el sitio A, se libera el peptidil-ARNt por hidrólisis de GTP, terminando la traducción y cambiando la conformación del ribosoma. **C)** Principio de LCP mediado por ARNt supresor. El reconocimiento de un codón de terminación (TAG) en el sitio A por parte del factor de liberación eRF1, desencadena la terminación de la traducción y la síntesis polipeptídica. En cambio, la ocupación del sitio A ribosomal por un ARNt supresor decodifica el codón TAG en un aminoácido de sentido (Tyr), y la proteína se extiende. (Adaptada de Ortolano *et al.* 2016 y Lewin's Genes X *et al.* 2011).

al., 2014). Análisis computacionales y la genómica comparativa han contribuido al conocimiento de patrones que permiten el reconocimiento de secuencias genéticas inusuales que codifican proteínas (Lin et al., 2007; Stark et al., 2007; Floquet et al., 2012; Pancsa et al., 2016). Por lo tanto, el descubrimiento de genes con lectura de codones de parada es importante para la investigación de sus funciones celulares y el mecanismo de regulación de la traducción. En este sentido, durante los últimos años se han intensificado las investigaciones centradas en los tipos de eventos de recodificación, los cuales tienen además implicaciones médicas importantes.

1.4.2. Terapias basadas en estímulo de LCP

Si bien los mecanismos moleculares que subyacen al proceso de lectura de codones de terminación de la traducción (LCP) no están del todo claros en eucariotas, la contribución de su investigación es fundamental para el desarrollo de nuevos tratamientos de enfermedades genéticas. Se sabe que aproximadamente el 11 % de las enfermedades hereditarias humanas son causadas por mutaciones sin sentido, es decir aquellas que transforman codones que determinan un aminoácido en codones de parada prematuros (CPP), llevando a una traducción truncada. Precisamente, numerosas mutaciones de este tipo han sido objeto de estudio de la terapia clínica basada en la supresión de CPP. Algunos ejemplos de enfermedades genéticas inducidas por CPP son la fibrosis quística (Rowe et al., 2011); la hemofilia (Zingman et al., 2007; Linde y Kerem, 2008); la distrofia muscular de Duchenne (Yukihara et al., 2011; Finkel et al., 2013); la retinitis pigmentosa (Guerin et al., 2008); la miopatía de Miyoshi (Wang et al., 2010); la enfermedad de Batten (Sarkar et al., 2011); el síndrome de Rett (Brendel et al., 2011; Vecsler et al., 2011); la mucopolisacaridosis tipo I (Wang et al., 2012; Kamei et al., 2013); el síndrome de Usher tipo 1 (Rebibo-Sabbah et al., 2007; Goldmann et al., 2012), etcétera. Además, se ha postulado que los CPP representan hasta un 30 % de las mutaciones carcinogénicas. Muchos cánceres están relacionados con la aparición de un CPP en un gen supresor de tumores, lo que resulta en la pérdida de la proteína o la síntesis de una proteína truncada que no puede inhibir la proliferación celular o promover la apoptosis (Bidou et al., 2017).
Los CPP alteran la expresión génica de dos maneras. Primero, los CPP terminan prematuramente la traducción, produciendo un polipéptido truncado no funcional o inestable (Bidou et al., 2017; Dabrowski et al., 2018). En segundo lugar, los CPP activan la degradación de ARNm mediada por mutaciones que resultan en un codón de parada. Este mecanismo celular, conocido por sus siglas en inglés NMD (de *Nonsense Mediated Decay*), detecta las mutaciones terminadoras y evita la expresión de proteínas truncadas (Keeling, 2016; Murtha et al., 2019). Durante la última década, los tratamientos de los trastornos genéticos ocasionados por CPP se han concentrado en una terapia farmacológica que tiende a aumentar la frecuencia con que ocurre la LCP en el CPP. Debido a que la terminación es más eficiente en los codones de parada canónicos (CPC), éstos son menos susceptibles a los fármacos inductores de estimulación de LCP; y por lo tanto, la supresión ocurre principalmente en los CPP (ver Fig. 1.7). Así, la terapia de LCP pretende restaurar la función proteica deficiente permitiendo la traducción de un polipéptido funcional de longitud completa; con el objetivo clínico de aliviar las consecuencias fenotípicas de enfermedades genéticas derivadas de CPP (Linde y Kerem, 2008; Keeling y Bedwell, 2011; Keeling et al., 2014).

Los aminoglucósidos (o aminósidos) son la clase de compuestos antibióticos mejor estudiada que estimulan la LCP con fines terapéuticos. En altas dosis bloquean la función del ribosoma bacteriano (Pfister et al., 2003; Bidou et al., 2012). Al inhibir la traducción en procariotas más eficientemente que en eucariotas, son usados en la clínica como antibióticos (Keeling et al., 2012). Por otro lado, en bajas dosis alteran la fidelidad de la traducción eucariota, aumentando la tasa de error en la traducción. En consecuencia, estos antibióticos aumentan la tasa de incorporación de aminoácidos en los codones de parada prematuros. La eficacia de los aminoglucósidos para estimular LCP y su utilidad para el tratamiento de las muta-

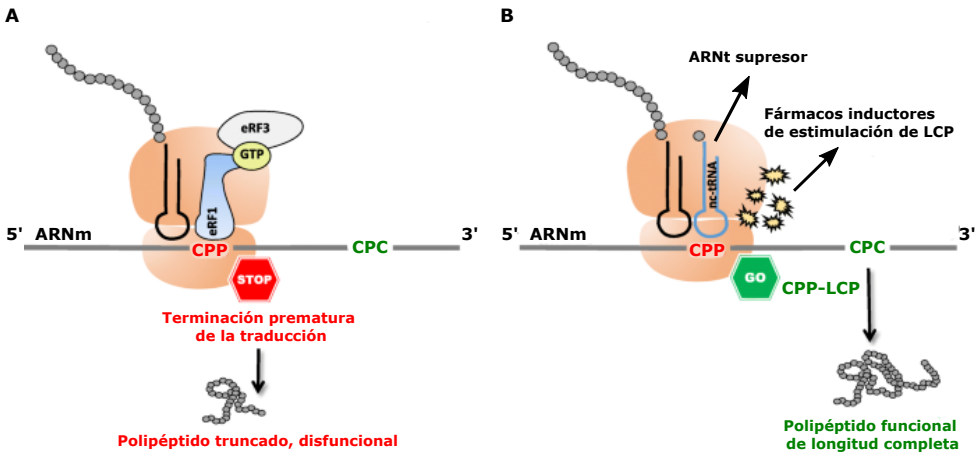

Figura 1.7: Mecanismo de inducción de LCP en codones de parada prematuros (CPP), mediante fármacos inductores. **A)** Los CPP son reconocidos por los factores de liberación (eRF1 y eRF3+GTP) antes que los CPC, ocasionando una finalización temprana de la traducción, y produciendo una proteína disfuncional. **B)** Los fármacos inductores de LCP facilitan el emparejamiento de secuencias codón-anticodón entre un ARNt supresor y el CPP, lo que inhibe la terminación prematura de la traducción. Así, los fármacos aminoglucósidos favorecen la aceptación de los ARNt supresores naturales por parte del ribosoma. (Adaptada de Dabrowski *et al.* 2018).

ciones sin sentido, se demostró por primera vez en células bronquiales humanas que expresan CFTR (regulador de la conductancia transmembrana de la fibrosis quística) con un CPP; tras el tratamiento con gentamicina se obtuvo una restauración parcial de los niveles de la proteína completa de CFTR (Howard et al., 1996). Desde entonces, numerosos estudios han investigado la efectividad del tratamiento de las mutaciones sin sentido para casi 40 enfermedades diferentes con antibióticos (Lee y Dougherty, 2012). Posteriormente, se demostró la eficacia de otros compuestos terapéuticos, utilizados como antibióticos de amplio espectro, que inducen específicamente la lectura ribosómica de mutaciones sin sentido, como son: negamicina, gentamicina, el PTC124 (conocido como ataluren), espiramicina, josamicina y tilosina (Dabrowski et al., 2018). Sin embargo, el uso clínico de compuestos inductores de LCP presenta limitaciones debido a su toxicidad (Keeling et al., 2014; Dabrowski et al., 2018). Además, los aminoglucósidos presentan una fuerte dependencia del contexto de nucleótidos que rodea el codón de parada (Keeling et al., 2014; Cridge et al., 2018). Se ha demostrado que la identidad del CPP es determinante en la eficiencia de los antibióticos para estimular LCP, siendo mayor en el caso de codones TGA, seguido por TAG y, en menor medida TAA (Manuvakhova et al., 2000); lo que sugiere que los mecanismos de LCP pueden ser distintos según el codón de parada. Incluso, las mutaciones sin sentido no son igualmente sensibles a las drogas inductoras de LCP, lo que dificulta la identificación de pacientes portadores de CPP que podrían beneficiarse de tales tratamientos. En suma, la terapia de supresión de CPP ha sido obstaculizada por falta de mayores avances en la comprensión de los

mecanismos de terminación de la traducción y de LCP.

1.5. Motivación y planteo de la tesis

Si bien el genoma de los organismos almacena la información que determina las proteínas para su desarrollo, la expresión génica depende de diversos factores biológicos y de complejos mecanismos de regulación. El desarrollo de ingeniosas herramientas en biología computacional acaecido durante los últimos años propició estudios más exhaustivos en genómica y proteómica, ampliando las fronteras de los sistemas de identificación de genes y transcritos. Esto permitió dilucidar el número de marcos abiertos de lectura *(ORFs)* codificantes de proteínas y así realizar un análisis integral de proteomas que contribuya a su anotación efectiva. A su vez, anotar un genoma conlleva la utilidad de predecir nuevas secuencias que codifiquen proteínas y ARNs estructurales. No obstante, la mayor parte de los genomas eucariotas contiene secciones no codificantes, cuyas funciones aún se encuentran indefinidas o mal caracterizadas. A pesar de ello, las técnicas actuales de secuenciamiento han abierto un nuevo campo para su estudio. En este sentido, se conocen diversos ORFs cortos (*sORF*) en regiones previamente consideradas no codificadoras de proteínas, tales como los extremos UTR o los intrones de algunos genes, e incluso en regiones conservadas. La traducción de novedosos péptidos derivados de sORF constituye una evidencia de su funcionalidad, y permite abordar análisis de función molecular, así como del contexto de su expresión y la interacción con otras biomoléculas. Sin embargo, su identificación bioinformática ha sido hasta el momento uno de los mayores impedimentos para estudios más abarcativos, llevando a una sub-anotación de los mismos. Estos descubrimientos plantean nuevas preguntas en el ámbito molecular. Por un lado, en cuanto a que la traducción no se limita sólo a transcritos codificadores de polipéptidos canónicos; sino que parece ocupar un rol mucho más generalizado, donde diversos transcriptos se expresan mediante funciones definidas en el metabolismo celular. Por otra parte, la presencia de nuevos péptidos no convencionales cuestionan los métodos de validación y anotación genómica, dado que existen nuevas estructuras génicas que aguardan clasificación, y siendo posible que muchos de los transcritos conocidos no estén correctamente anotados. De la misma manera, la anotación eficiente requiere corregir la delimitación imprecisa de los ORFs, que por defecto suelen delimitarse en el primer codón de parada (TAA, TAG o TGA). Sin embargo, se ha visto que en ocasiones esto no es así, y la traducción de la secuencia de ARNm continúa hasta otro codón de parada posterior. Este tipo de mecanismo en el proceso de traducción produce alteraciones en la cadena polipeptídica en formación, con propiedades diferentes a las predichas en base a una traducción canónica de ARNm, diversificando de la expresión génica. Particularmente, nuestro interés se centra en la lectura de codones de parada (LCP), permitiendo la biosíntesis de proteínas extendidas en su extremo carboxilo terminal. Estos casos no constituyen un mero error de traducción; dado que se sintetizan varias proteínas de importancia biológica como resultado de la lectura funcional traduccional. De hecho, como vimos en la sección previa, la LCP constituye la base de nuevas terapias para numerosas enfermedades genéticas. Por lo tanto, consideramos este enfoque muy importante y complementario para nuestro análisis de productos génicos no anotados, ya que se trata de un mecanismo de recodificación capaz de alterar el producto polipeptídico a ser expresado, expandiendo

así la plasticidad de los genomas.

En síntesis, comprender los mecanismos de terminación alternativa de la traducción puede contribuir al desarrollo de nuevas terapias para enfermedades congénitas, tales como la terapia de inducción de LCP. Esto constituye la principal motivación a abordar en el presente trabajo de investigación. En este sentido, el objetivo principal de esta tesis es la identificación de nuevos eventos de LCP y su posterior análisis. Para realizar esta tarea, utilizamos datos disponibles en bases públicas, derivados del secuenciamiento de los fragmentos transcriptómicos protegidos por ribosomas obtenidos en embriones tempranos de la mosca *Drosophila melanogaster*. Los procedimientos empleados serán detallados en el capítulo 2. Este estudio de identificación se complementa con el análisis de la tasa de incidencia de LCP para cada uno de los codones de parada, y el contexto nucleotídico circundante a éstos. Para ello, aplicamos la técnica de perfil ribosómico en combinación con un modelo de regresión lineal para el análisis del contexto del codón de parada, a fin de inferir la existencia de un patrón que funcione a modo de señal para la ocurrencia de LCP. Los procedimientos empleados serán detallados en el capítulo 3.

Capítulo 2

Identificación de eventos de lectura de codones de parada

2.1. Lectura de codones de parada y la sub-anotación de genomas

Hasta aquí hemos abordado los fundamentos sobre la regulación de la expresión génica. En este capítulo expongo un aspecto novedoso de la delimitación de muchos ORFs de *Drosophila melanogaster*. Se trata de un mecanismo de extensión de la traducción más allá del codón de parada, que puede alterar el producto final a ser expresado, dando origen a proteínas diferentes a las esperadas en base a la secuencia nucleotídica que está siendo traducida. Específicamente, la síntesis de una proteína se completa tras el reconocimiento de los codones de terminación de la traducción (TAA, TAG, TGA) mediante la interacción entre el ARNm y factores de liberación (eRF), ocurrida en el sitio A del ribosoma. En eucariotas, el eRF1 reconoce los tres codones de parada y, tras decodificarlos, facilita la liberación hidrolítica del polipéptido formado del peptidil ARNt unido al sitio P ribosomal. Finalmente, la traducción canónica finaliza cuando el ribosoma se disocia del ARNm. Sin embargo, existen excepciones a estas reacciones secuenciales que pueden conducir a los ribosomas a continuar la traducción más allá del primer codón de terminación, dando lugar a un evento de lectura del codón de parada (LCP). Estos eventos excepcionales ocurren con una cierta probabilidad, por lo tanto, una fracción de las proteínas resultantes incluyen péptidos adicionales en su extremo C-terminal (Ryoji et al., 1983; Robinson y Cooley, 1997; Jungreis et al., 2011; Freitag et al., 2012; Loughran et al., 2018). Esta fracción de proteínas extendidas será mayor cuanto mayor sea la probabilidad del ribosoma de continuar con la traducción. Cuando esta probabilidad es pequeña, los eventos podrían ser considerados como un "fallo" en la terminación y sus productos proteicos no tendrían consecuencias funcionales (alteraciones en la proteostasis) (Bidou et al., 2010; Li y Zhang, 2019). No obstante, hoy se conocen varias proteínas de importancia biológica que son el resultado de la lectura funcional del codón de parada (von der Haar y Tuite, 2007; Schueren y Thoms, 2016; Sapkota et al., 2019). En estos casos se habla de un mecanismo de LCP programada, denominado en la literatura científica como *Stop Codon Readthrough* o también *Translational Readthrough*, y se produce en niveles detectables en muchos

ARNm.

La hipótesis más aceptada de este fenómeno consiste en que los codones de terminación pueden ser decodificados como codones de sentido por un ARNt cognato que se aloja en el sitio A del centro de decodificación ribosomal. Los ARNt cognatos pueden aparearse en dos de las tres posiciones del codón. Esta interacción resulta en la competencia por el reconocimiento del codón de parada entre el factor de liberación eRF1 y el ARNt cognato; y puede inhibir el proceso de terminación de la traducción, siendo uno de los determinantes de la eficiencia en el proceso de terminación de la traducción (Dabrowski et al., 2015; Schueren y Thoms, 2016). En este caso, al ARNt cognato se le conoce como ARNt supresor (Blanchet et al., 2014; Jungreis et al., 2016; Massey, 2017). Como resultado, un aminoácido se incorpora en la cadena polipeptídica y el ribosoma continúa la traducción hasta el siguiente codón de terminación en el mismo marco de lectura.
Existen otros mecanismos de recodificación que actúan en las señales de terminación de la traducción (Firth y Brierley, 2012). En este sentido, podemos mencionar el cambio de marco ribosomal *(ribosomal frameshifting)*, en el cual una fracción de ribosomas omite algunos nucleótidos y continúa traduciendo en otro marco de lectura (Wills et al., 2008). Un ejemplo clásico es el cambio de marco programado en un codón de parada TGA para el gen de la antizima ornitina decarboxilasa (ODC-AZ) (Matsufuji et al., 1995). Además, la recodificación del codón de parada, puede involucrar un aminoácido no estándar como selenocisteína. En este caso, un ARNt de selenocisteína puede insertar un aminoácido homónimo en un codón TGA, si está presente una secuencia de inserción de selenocisteína (elemento SECIS) en la región 3' UTR (Bonetti et al., 1995; Poole, 1998). En vertebrados, este mecanismo de incorporación de selenocisteína es necesario para producir enzimas como la glutatión peroxidasa (GPX) (Chambers et al., 1986) y yodotironinas desyodinasas (Berry et al., 1991; Davey et al., 1995; Salvatore et al., 1995).

La LCP programada fue originalmente descubierta en el bacteriófago Qβ, siendo que incorpora un triptófano en el codón TGA con una eficiencia del 2 % (Weiner y Weber, 1971). También se encontró inicialmente en los virus del mosaico de tabaco y la cebada amarilla (Pelham, 1978; Brown et al., 1996). El gen *kelch* en *D. melanogaster* contiene un codón de terminación TGA dentro del marco que separa dos ORFs, y cuando es suprimido da lugar a una versión extendida de la proteína (Xue y Cooley, 1993; Bergstrom et al., 1995; Robinson y Cooley, 1997). Más recientemente, se documentó la existencia de LCP programada en algunos genes en hongos (Namy et al., 2002; Freitag et al., 2012) y eucariotas, como el gen de la β-globina en conejos y los genes *syn* y *hdc* en *D. melanogaster* (Klagges et al., 1996; Chittum et al., 1998; Steneberg y Samakovlis, 2001), para mencionar algunos de los ejemplos. Así, fuera del caso de la incorporación de selenocisteína, la LCP programada parecía tener un rol menor y extraordinario. Sin embargo, gracias a las nuevas técnicas de secuenciamiento masivo y métodos bioinformáticos, se han identificado algunos cientos de nuevos eventos de LCP programada en varios genomas de metazoos. En este sentido, se encontraron patrones de LCP programada en 283 genes que están conservados entre las 12 especies de *Drosophila* analizadas, entre ellos el gen *kelch* antes mencionado (Jungreis et al., 2011). Entre estos casos identificados, se comprobaron por expresión de GFP la existencia de 4 extensiones (cnc-RT, Sp1-RT, Abd-B-RT, z-RT), mientras que otros 7 péptidos que incluyen

extensiones fueron validados mediante espectrometría de masas (Jungreis et al., 2011). Más tarde, con un método filogenético similar, se detectaron otros 50 posibles casos de LCP programada en *Drosophila melanogaster* y 353 en el genoma de *Anopheles gambiae* (Jungreis et al., 2016).

Por otro lado, las regiones de ARNm que son traducidas activamente se pueden reconocer mediante una técnica de secuenciación de ARN de alto rendimiento *(RNA-Seq)*, que secuencia los fragmentos ocupados por ribosomas. Esta técnica, denominada *Ribo-Seq*, permite obtener el perfil ribosómico de los transcriptos, útil para cuantificar los niveles de expresión (Ingolia et al., 2009). De este modo, *Ribo-Seq* puede contribuir a la identificación de elementos funcionales en los genomas, en base a su grado de conservación. En el anexo I se exponen las características de la técnica de *Ribo-Seq*. Usando esta técnica como evidencia de traducción, se identificaron 350 eventos de LCP en cultivo de células S2 y de embriones de *D. melanogaster* (Dunn et al., 2013), entre ellos fueron 43 detectados previamente con el enfoque filogenético de (Jungreis et al., 2011). Estos hallazgos revelan eventos candidatos de LCP programada (Stark et al., 2007; Lindblad-Toh et al., 2011; Jungreis et al., 2011).

Recientemente, de los genes con LCP propuestos en estudios previos, se validaron experimentalmente cuatro genes homólogos en humanos (SACMIL, OPRK1, OPRL1, BRI3BP), junto a otros tres nuevos casos (ACP2, AQP4, MAPK10) (Loughran et al., 2014). También, se descubrió LCP programada en el ARNm del factor de crecimiento endotelial vascular A *(VEGF-A)* de mamíferos, como humanos y vacas, por ejemplo. En este caso, se decodifica el codón de terminación TGA como serina, generando una isoforma *(VEGF-Ax)* que contiene una extensión C-terminal de 22 aminoácidos (Eswarappa et al., 2014). Esto le confiere a *VEGF-Ax* una actividad antagónica a la actividad pro-angiogénica de *VEGF-A*; lo cual destaca los roles biológicos del mecanismo de LCP programada. Adicionalmente, se han desarrollado estrategias basadas en análisis de regresión lineal, para predecir transcriptos candidatos de LCP programada (Schueren et al., 2014). Estos resultados sugieren que la LCP puede ser un mecanismo de recodificación para extender proteínas más generalizado de lo que se pensaba. Su importancia fundamental radica en que ofrece una forma alternativa de expandir la plasticidad de los genomas, diversificando la expresión génica (von der Haar y Tuite, 2007; Schueren y Thoms, 2016).

Los procesos descritos anteriormente constituyen un claro ejemplo de sub-anotación de genes y de proteínas, es decir, no se conocen todos los genes sujetos al mecanismo de LCP, ni cual es la proporción entre las isoformas. Sin embargo, tampoco hay aún certezas sobre su mecanismo de acción y condiciones.
En general, la eficiencia de la terminación de la traducción varía entre los tres codones de parada. En este sentido, se sabe que el codón TGA es el más frecuente entre los candidatos con LCP programada, dado que tiene la mayor tasa de fuga ribosomal y posee, por lo tanto, la fidelidad más baja para la terminación de la traducción. El codón TAG es el segundo más común para la ocurrencia de LCP programada, mientras que TAA es el menos frecuente y tiene la menor tasa de fuga, siendo el de mayor fidelidad (Jungreis et al., 2011; Loughran et al., 2014; Cridge et al., 2018; Dabrowski et al., 2018). Sin embargo, la tasa de fuga ribosomal puede verse afectada por el contexto de nucleótidos alrededor del codón de terminación. Se ha indicado que la base inmediatamente posterior al codón de parada ejerce la

mayor influencia en la eficiencia de LCP programada, siendo la frecuencia relativa de las bases con más fugas C>T>G>A (Jungreis et al., 2011). En particular, alrededor de un tercio de los transcriptos identificados en ese trabajo terminaban con el codón TGA seguido del nucleótido C. De hecho, los genes que contienen un contexto de parada TGA-C tienen casi 10 veces más probabilidades de ser candidatos de LCP programada que los genes con otros contextos (16.4 % vs. 1.9 %) (Jungreis et al., 2011; Dabrowski et al., 2015). En efecto, tal como mencionamos anteriormente, para ciertos contextos de codones de parada asociados a LCP, un ARNt cercano puede insertar un aminoácido afín en su lugar (Bonetti et al., 1995; Poole, 1998), lo que puede resultar en un nivel de LCP superior al 5 % (Namy et al., 2001). Otros autores han señalado la posibilidad de existencia de elementos reguladores distales en el transcripto, o inclusive inducidos por agentes farmacológicos (Bidou et al., 2012). En conjunto, estos factores pueden aumentar el LCP en varios órdenes de magnitud, lo que resulta en tasas superiores al 1 % (Floquet et al., 2012; Loughran et al., 2014). Esto sugiere que en estos casos el LCP no es un error de traducción; sino un mecanismo funcional de recodificación del codón de parada para expresar dominios C-terminales de una proteína de menor abundancia, que a diferencia del empalme alternativo puede regularse a nivel de traducción (Dabrowski et al., 2015; Garofalo et al., 2019). En eucariotas, la proporción entre proteínas derivadas de LCP programada y las que no, puede ser controlada para muchos genes simultáneamente mediante la regulación de las proteínas del factor de liberación eRF1 (von der Haar y Tuite, 2007). Este control puede variar según el tejido (Robinson y Cooley, 1997) y la etapa de desarrollo (Samson et al., 1995).

Si bien hoy en día los eventos de LCP están claramente establecidos, persisten algunas preguntas y han surgido otras nuevas. Por ejemplo, no sabemos cómo distinguir claramente los eventos de LCP programados de aquellos eventos no funcionales, derivados de fallas eventuales en la terminación de la traducción. En forma más general, no conocemos aún las reglas que rigen la eficiencia de LCP programada. En este sentido, aún no se han identificado elementos regulatorios proximales o no, de estos eventos. Por lo tanto, estas preguntas serán abordadas en este capítulo. Como vimos, la LCP asume una excepción a la regla de terminación de la traducción, y constituye un mecanismo por el cual una fracción de las proteínas resultantes incluyen péptidos adicionales en su extremo C-terminal. Esta fracción de proteínas extendidas es proporcional a la frecuencia con que el ribosoma recodifica el codón de parada y continúa la traducción, lo cual denominaremos aquí como tasa de fuga ribosomal, o simplemente tasa de fuga. Nuestra hipótesis de trabajo consiste en que una fracción elevada de proteínas extendidas constituye un fuerte indicio de que estas extensiones corresponden a un evento funcional de LCP programada, y no a una mera falla en la terminación de la traducción. Por ende, la tasa de fuga ribosomal sería un determinante de la LCP programada. Como veremos en la siguiente sección, esta tasa de fuga podría, en principio, ser estimada a partir de los perfiles ribosómicos obtenidos por *Ribo-Seq*. De esta manera, en este capítulo focalizamos en identificar péptidos funcionales aún no anotados, derivados de eventos de LCP en el genoma de *D. melanogaster*.

2.2. Los perfiles ribosomales y LCP

Con el fin de identificar eventos de LCP, programados o no, examinamos los transcriptos de embriones de *D. melanogaster* utilizando la técnica de perfiles ribosomales introducida por (Ingolia et al., 2009). Esta técnica se basa en la secuenciación de los fragmentos de ARNm protegidos por ribosomas, denominados huellas ribosomales (o *footprints*). Las huellas ribosomales son mapeadas en el genoma de referencia para construir el perfil de densidad ribosomal de todos los transcriptos. Estos perfiles se definen como el número total de huellas ribosomales que se alinean al transcripto en cada posición. Así, además de medir la expresión génica a nivel traduccional, esta técnica permite examinar las ubicaciones físicas de los ribosomas asociados a cada transcripto, ayudando a determinar qué porciones de ARNm se traducen y cuáles no.
En esta tesis utilizamos los datos de *Ribo-Seq* obtenidos por Dunn y colaboradores para embriones tempranos (0-2 hs) de *Drosophila melanogaster* (disponibles en NCBI GEO, número de acceso #GSE49197) (Dunn et al., 2013). En su estudio, utilizaron estos perfiles para identificar eventos de LCP; y con ello determinaron 307 posibles eventos. Sin embargo, los autores se limitaron a estudiar una fracción del número total de transcriptos. Además, en el mapeo de las huellas no tuvieron en cuenta las junturas exón-exón, dando lugar a un menor número de huellas ribosomales mapeadas. En virtud de ello, nos propusimos realizar una búsqueda más exhaustiva de eventos de LCP. Un conjunto mayor de estos eventos nos permitirá tener mejor información para estudiar el contexto nucleotídico de estos eventos, y dilucidar si tienen un rol determinante en el mecanismo de LCP programada. Para esta tarea, removimos la secuencia de cebador de las huellas ribosomales usando el software *Cutadapt* (Martin, 2011), y descartamos todas las secuencias menores a 25 nucleótidos (nt). Para descartar aquellas lecturas alineadas con secuencias ribosómicas, las huellas de ribosomas fueron mapeadas a secuencias ribosomales con el software *Bowtie2* (Langmead et al., 2009). Este alineamiento fue realizado con el software *TopHat* (Trapnell et al., 2009), que tiene en cuenta las uniones de empalme exón-exón. Los archivos de formato SAM resultantes (del inglés: *Sequence Alignment/Map files*) fueron procesados con el software *SAMtools* y programas propios para construir el perfil de densidad de ribosomas de todos los transcriptos. Detalles de esta metodología aplicada se encuentran en el capítulo 4 (Datos y Metodología), mientras que la descripción de los softwares y formatos de archivo empleados se aportan en el anexo II.

Observamos que en general, los perfiles ribosomales obtenidos presentan una densidad sustancial de huellas ribosomales tanto en las regiones 5' UTR, como en las regiones codificantes (CDS). La densidad ribosomal no nula en la región 5' UTR contrasta con la inherente característica de que dicha región no se traduce, a excepción de algunos pequeños marcos de lectura, denominados uORF. Según Dunn *et al.*, la densidad ribosomal en los 5' UTRs podría ser atribuida a los fragmentos de ARNm protegidos por ribosomas que cubren secuencias de uORFs (Dunn et al., 2013). Sin embargo, el número de uORF reportados hasta el momento es de algunas decenas, cuando la mayoría de los transcriptos presentan una densidad ribosomal en esta región. Por este motivo, la densidad ribosomal presente en los 5' UTR no puede ser considerada como una evidencia de la traducción de estas regiones. En cambio, la gran mayoría de los transcriptos presentan una densidad ribosomal nula en el

extremo 3' UTR, más precisamente a partir de 20 nucleótidos posteriores al codón de parada anotado. La Fig. 2.1 muestra un ejemplo de este caso, el perfil ribosomal de la isoforma A del gen *Snx1* en el cromosoma 2L (transcripto FBtr0077517). Este gen cumple la función de unión de fosfatidilinositol, y esta vinculado a procesos de transporte intracelular, según datos de GenBank en NCBI (número de acceso NM_134933).

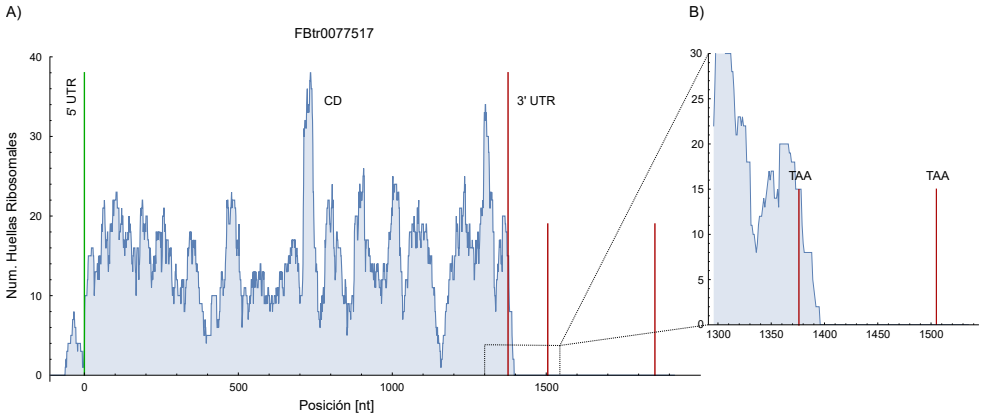

Figura 2.1: Perfil ribosomal de una isoforma del gen *Snx1* (transcripto FBtr0077517), el cual no presenta LCP. La línea vertical verde en la posición cero corresponde al codón de inicio de la traducción, delimitando el CDS del extremo 5' UTR (panel A, izquierda). Las líneas verticales rojas (panel A) indican la posición del codón de terminación anotado (TAA) en la posición 1376 nt, y un codón de parada posterior (TAA) ubicado 129 nt después. La ampliación de la región 3' UTR posterior al codón de parada anotado (panel B), muestra la ausencia de lecturas ribosomales, por cuanto la terminación traduccional canónica es eficiente.

Diferentes transcriptos pueden presentar diferentes densidades ribosomales medias. Esta diferencia se puede explicar por el nivel de expresión, es decir, los transcriptos más abundantes están asociados a mayores densidades. Sin embargo, también se puede observar en la Fig. 2.1 que los niveles de densidad ribosomal en la región codificante varían considerablemente dentro del mismo transcripto. Esto se podría explicar en base a que la eficiencia traduccional debe estar modulada para que la proteína naciente adquiera el plegamiento requerido para su función. En este sentido, los picos locales de mayor densidad suelen ser interpretados como pausas ribosomales, mientras que los valles corresponden a regiones donde los ribosomas tienen un tránsito más fluido.
Por otro lado, debido a que la iniciación y la terminación de la traducción son lentas en comparación con el alargamiento, allí se producen picos de densidad de ribosomas (Ingolia et al., 2011). En contraste, los transcriptos presentan una densidad ribosomal nula en el extremo 3' UTR, y esto refleja la terminación efectiva de la traducción. Por el contrario, la no disociación de los ribosomas del transcripto podría añadir aminoácidos adicionales después del codón de parada. La Fig. 2.2 muestra a modo de ejemplo, el perfil ribosómico correspondiente a la isoforma F

del gen *RpS15Aa* (transcripto FBtr0300828), en el cromosoma X. Este transcripto codifica la proteína ribosomal S15A, un componente estructural del ribosoma que se prevé que participe en la traducción citoplasmática. Se expresa en la cabeza y el corazón del organismo adulto, y es ortóloga de la RpS15A humana (Marygold et al., 2007).
Una ampliación del extremo 3' UTR del transcripto FBtr0300828 muestra que existe densidad ribosomal entre el codón de parada anotado y un segundo codón de parada en el 3' UTR, con una extensión de 62 aminoácidos (RNQVRIVESLYVLLRSLICRLTFYFTDHATNYVLCAQETRGENAATFIQKTNDDECSQSYAK). Este caso contrasta con la densidad ribosomal mostrada en la Fig. 2.1. Aunque este nivel de densidad puede ser mil veces inferior al registrado en la región codificante anotada, corresponde a un número sustancial de lecturas alineadas en una porción de transcripto de 170 nt. En este sentido, la presencia de una densidad ribosomal residual de este tipo ha sido considerada como una evidencia empírica de eventos de LCP (Dunn et al., 2013; Jungreis et al., 2016).

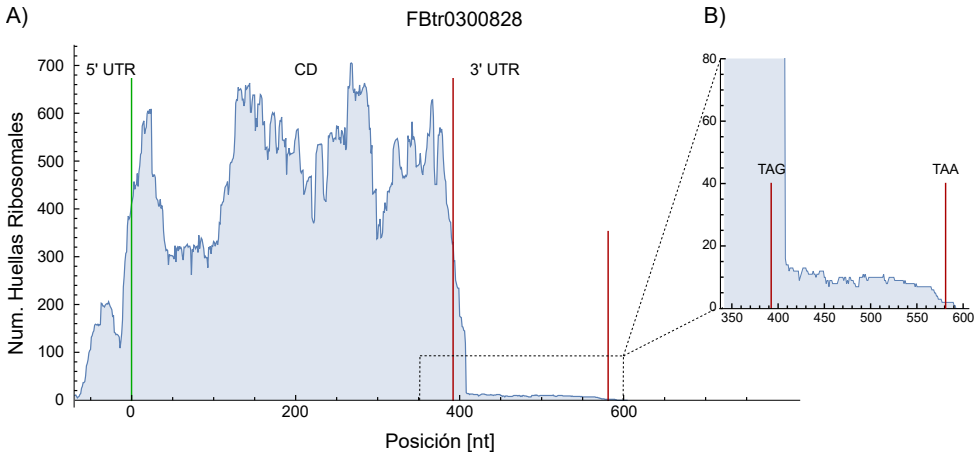

Figura 2.2: Perfil de densidad ribosomal con evidencia de LCP en la isoforma F del gen *RpS15Aa* (transcripto FBtr0300828). Este caso muestra una extensión posterior al codón de terminación anotado (TAG), ubicado en la posición 393 (nt) respecto del codón de inicio (panel A). Dicha extensión posee una densidad ribosomal de alrededor de 10 lecturas, y alcanza al segundo codón de parada (TAA) ubicado a 189 nt posteriores al codón anotado (panel B).

La Fig. 2.3 muestra el perfil ribosómico de la isoforma A del gen *Arf102F* (transcripto FBtr0089192), en el cromosoma 4. Este transcripto codifica una proteína de unión a GTP implicada en el tráfico de proteínas, y que puede modular la gemación y la eliminación de la capa de vesículas dentro del aparato de Golgi (UniProt, P40945). Este perfil presenta evidencia de múltiples eventos de LCP asociados a diferentes codones de parada, indicados con las barras verticales rojas. Este tipo de eventos múltiples ya habían sido reportados en Dunn *et al.* (2013). Sin embargo, los eventos de LCP ilustrados en las Fig. 2.2 y Fig. 2.3 son dos ejemplos que no habían sido reportados anteriormente, al igual que otros que expondremos

en esta tesis.

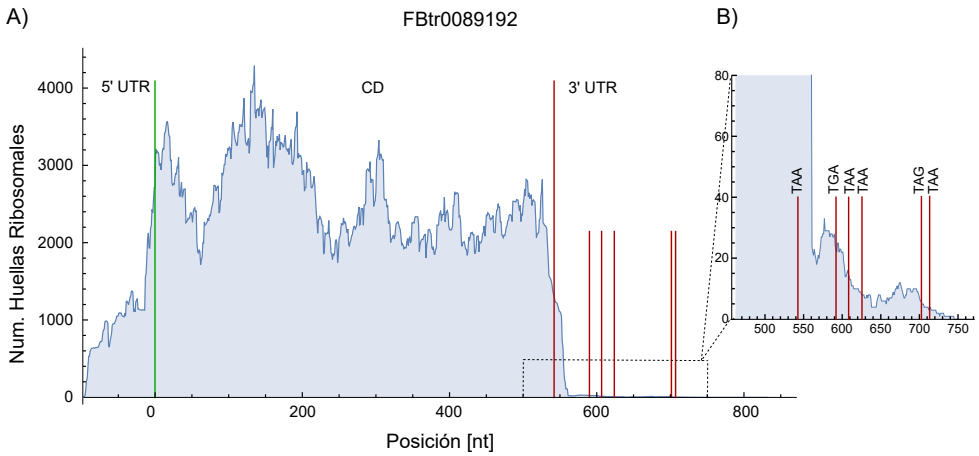

Figura 2.3: Perfil de densidad ribosomal con evidencia de múltiples eventos de LCP en la isoforma A del gen *Arf102F* (transcripto FBtr0089192). Se observa presencia de huellas ribosomales en la región posterior al codón de terminación anotado (TAA) en la posición 543 nt (panel A). La ampliación del extremo 3' UTR (panel B) refleja una extensión de la lectura hasta un quinto codón de parada.

Para nuestro análisis exhaustivo, seleccionamos todos aquellos transcriptos con RPKM mayor o igual a 2, una métrica de normalización de lecturas por kilobase por millón de lecturas secuenciadas (del inglés: *Reads Per Kilobase Million*), estimada como se detalla en el capítulo 4. De esta manera, obtuvimos la densidad ribosomal de 6739 transcriptos correspondientes a un total de 4842 genes expresados durante el estadío temprano de embriones de *D. melanogaster*. En este conjunto de 6739 transcriptos realizamos una inspección visual, asistida por computadora, de los perfiles ribosomales que exhiben una densidad ribosomal más allá del codón de parada anotado. En este análisis inicial, nos focalizamos en detectar la mayor cantidad de transcriptos cuyos perfiles tengan las características propias de aquellos con eventos de LCP. Así, seleccionamos los transcriptos cuyos perfiles ribosomales satisfacen los siguientes criterios:

- presentar una densidad ribosomal mayor a 2 huellas ribosómicas sobre el 90 % de la extensión (se entiende por extensión a la región que va desde el codón de parada anotado hasta el segundo codón de parada);

- el tamaño de la región con densidad ribosomal indicada debe ser mayor a 30 nucleótidos, contando desde la posición del codón de parada anotado.

Después de seleccionar todos los transcriptos con posible LCP, cada uno de los perfiles de densidad ribosomal fue inspeccionado visualmente para descartar artefactos y elegir la isoforma más adecuada. En este sentido, excluimos de análisis

posteriores aquellos transcriptos cuya densidad ribosomal en la extensión no fuese contigua a la región codificante, y aquellos casos donde la extensión se superpone con regiones codificantes conocidas asociadas a otros transcriptos. En efecto, en estos casos el análisis de los perfiles ribosomales por sí solo, no es suficiente para distinguir si la densidad ribosomal observada se trata de la traducción de una extensión por LCP, o corresponde a la traducción canónica de una región que se solapa con la posible extensión. A continuación señalamos un ejemplo de un caso que fue previamente identificado como un evento de LCP (Jungreis et al., 2011; Dunn et al., 2013), pero que sin embargo parece representar sólo un artefacto de la metodología. En la Fig. 2.4 mostramos el perfil de densidad ribosomal asociado a la isoforma C del gen *RanBPM* (transcripto FBtr0088262), en el brazo R del cromosoma 2. Este transcripto está asociado a una proteína de 598 aminoácidos de longitud (involucrada en la regulación y organización de las células madre de la línea germinal del ovario), a la cual tanto Jungreis como Dunn le asociaron una extensión de 167 aminoácidos posterior al codón de terminación anotado. Sin embargo, como mostramos en el panel inferior de la Fig. 2.4, la supuesta extensión es consistente con la traducción canónica de una o ambas isoformas F y G del mismo gen, que tienen su codón de parada anotado 501 nt más adelante. Casos como estos fueron identificados por inspección visual y descartados de nuestro conjunto de análisis.

Así, a partir de este procedimiento, se identificaron 1303 transcriptos que presentaban una extensión con lecturas de densidad ribosomal compatibles con los criterios impuestos. Algunos de estos eventos de LCP seleccionados en el presente trabajo, han sido reportados previamente. En este sentido, Jungreis *et al.* reportaron 283 candidatos, mientras que Dunn *et al.* reportaron un total de 350 transcriptos con LCP, estudiando huellas ribosómicas provenientes tanto de embriones (los mismos datos analizados aquí) como de la línea celular S2 (Dunn et al., 2013); aunque 43 de estos ya habían sido reportados por Jungreis et al. (2011).
Por otro lado, en FlyBase se encuentran anotados 486 transcriptos, pertenecientes a 118 genes con categoría de LCP (versión FB2020_04, publicado el 18 de Agosto de 2020. http://flybase.org/stop-codon-readthrough/hitlist). De los transcriptos con LCP anotados en FlyBase, 11 de ellos coinciden con los reportados en nuestra lista de candidatos, siendo los siguientes: FBtr0330000 (gen *CG5742*), FBtr0330114 (gen *CG2126*), FBtr0330167 (gen *Ets97D*), FBtr0330236 (gen *CG30389*), FBtr0330248 (gen *hrg*), FBtr0330292 (gen *yem*), FBtr0330369 (gen *CG13601*), FBtr0330370 (gen *CG13604*), FBtr0330372 (gen *cnc*), FBtr0330382 (gen *Spps*), FBtr0330383 (gen *Tnks*).

Entre los 6739 transcriptos analizados en esta tesis, 199 transcriptos pertenecen a los eventos de LCP previamente reportados por Jungreis et al. (2011); Dunn et al. (2013). Sólo 7 de estos 199 transcriptos fueron identificados por ambos autores, específicamente son: FBtr0079019 (gen *CG4230*); FBtr0089288 (gen *Dyrk2*); FBtr0076398 (gen *vsg*); FBtr0083264 (gen *gish*); FBtr0085475 (gen *Gnpnat*); FBtr0113277 (gen *CG13604*) y FBtr0088262 (gen *RanBPM*), siendo este último caso un falso positivo. Por otro lado, no todos estos eventos de LCP previamente reportados fueron identificados en el presente estudio. A partir del conjunto de 1303 transcriptos con evidencia de LCP identificados en esta tesis, observamos que 118 coinciden con los reportados por Dunn y sólo 7 con los señalados por Jungreis. Sólo dos transcriptos fueron comunicados en todos los estudios, el FBtr0079019 (del gen *CG4230*) y el FBtr0113277 (del gen *CG13604*). Así, en resumen en este estudio es-

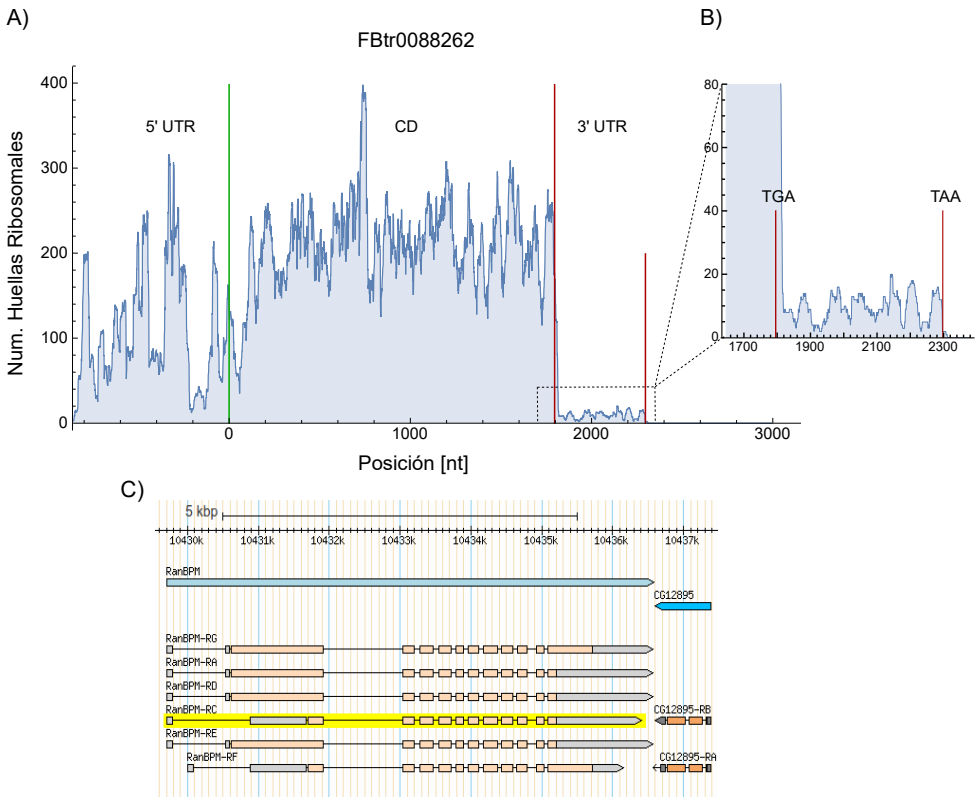

Figura 2.4: Perfil de densidad ribosomal de la isoforma C del gen *RanBPM* (transcripto FBtr0088262) mostrando un falso evento de LCP de 167 aminoácidos. En el panel inferior se observa que la supuesta extensión corresponde a la traducción canónica de otras isoformas del mismo gen.

tamos reportando un conjunto de 1176 nuevos transcriptos candidatos a presentar eventos LCP, que no fueron reportados en los estudios previos.

El diagrama de Venn en la Fig. 2.5 resume este análisis comparativo, mostrando los puntos de convergencia y diferencias sobre la cantidad de posibles eventos de LCP propuestos para evaluar en esta tesis. Cada círculo representa un conjunto de eventos de LCP sugerido por los autores. Así, el círculo amarillo representa los 1303 transcriptos seleccionados como candidatos en este estudio; mientras que los círculos verde y rojo corresponden a aquellos transcriptos con LCP reportados por Jungreis y Dunn, respectivamente, que están incluidos en el presente conjunto de análisis. Como puede apreciarse en la imagen, los elementos comunes se ubican en la intersección de los círculos.

Así, estamos reportando 1176 nuevos casos putativos de LCP, de los cuales se muestran algunos ejemplos en la 2.1. El hecho de que estamos reportando una cantidad de eventos mayor a los reportados en Dunn *et al.* usando los mismos datos de *Ribo-Seq*, puede explicarse principalmente por dos razones: examinamos un mayor número de perfiles de densidad de ribosomas asociados a transcriptos, y usamos

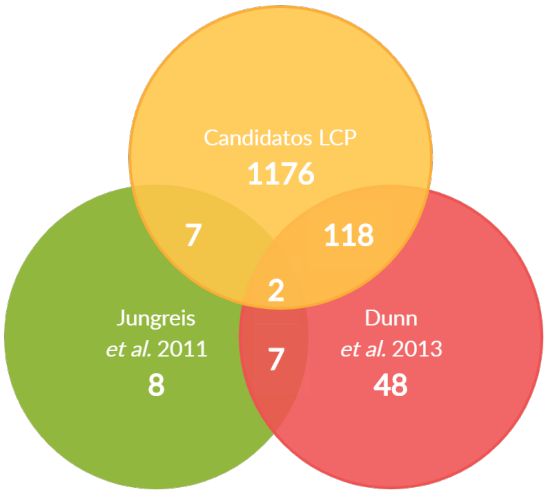

Figura 2.5: Diagrama de Venn que compara la cantidad de eventos de LCP reportados por tres diferentes estudios, en transcriptos de embriones tempranos de *D. melanogaster*. En esta tesis se pre-seleccionaron 1303 transcriptos candidatos a presentar eventos de LCP (conjunto amarillo), los cuales fueron contrastados con los reportados por Jungreis (conjunto verde) y por Dunn (conjunto rojo). Se observa que el conjunto de eventos LCP aquí propuesto comparte 118 casos con el conjunto de Dunn y 7 casos con el conjunto de Jungreis; pero sólo coincide con ambos en dos transcriptos. Exceptuando estas coincidencias, el conjunto seleccionado contiene 1176 transcriptos con posibles eventos de LCP, que no han sido reportados. Por su parte, los eventos de LCP reportados por Jungreis y Dunn son 24 y 175, respectivamente; pero sólo coinciden en 7.

diferentes métodos de alineación de huellas ribosomales (véase las secciones 4.3, 4.4 y 4.5 en el capítulo 4). Aun así, es probable que no todos los casos seleccionados en este estudio correspondan a eventos de LCP. A propósito de ello, más adelante se introducen análisis que permiten determinar eventos más fidedignos, como son: la sintenia, que compara la conservación de conexiones genéticas entre diferentes organismos; y la validación de traducción en las extensiones predichas mediante espectrometría de masas. A continuación, mostraremos algunos ejemplos de nuevos eventos de LCP asociados con estos análisis.

2.3. Nuevos eventos de LCP derivados de perfiles ribosomales

Dado que por razones de espacio no podemos mostrar todos los nuevos transcriptos con eventos de LCP identificados en nuestro estudio de perfiles ribosomales,

Cromosoma	ID del transcripto	Nombre del gen	Codón de parada	Posición de LCP	Long. (aa)	Densidad
4**	FBtr0089192	Arf102F	TAA	181	20*	11.39
4	FBtr0300555	PMCA	TAA	1256	55	15.81
X**	FBtr0300828	RpS15Aa	TAG	131	64	8.58
X	FBtr0305149	Imp	TAA	632	100	5.23
X	FBtr0074005	Top1	TAA	975	84	5.94
X	FBtr0070597	HIP-R	TAA	378	6*	4.39
X	FBtr0070141	Sec22	TAA	212	48	4.46
X	FBtr0073763	Jafrac1	TAA	195	90*	15.86
X	FBtr0071066	unc-119	TAG	266	81*	13.82
2L**	FBtr0339562	CG31673	TGA	326	49	11.59
2L**	FBtr0100268	Chrac-14	TGA	129	29	18.29
2L**	FBtr0079297	CG11070	TGA	994	52	2.56
2L**	FBtr0078997	CG3792	TAA	253	29	6.77
2L	FBtr0081263	Hakai	TAA	303	44*	5.98
2L	FBtr0273293	RpS21	TGA	82	30*	141.02
2R**	FBtr0072343	Nurf-38	TGA	339	25	10.77
2R	FBtr0072120	RabX1	TAA	262	92	4.
2R	FBtr0303464	CNBP	TAG	166	10	39.97
2R	FBtr0071935	RpS24	TAA	132	48	10.86
2R	FBtr0086502	TBCB	TGA	245	47	4.69
3L**	FBtr0076462	ghi	TAA	82	57	14.16
3L**	FBtr0072583	CG13887	TAA	229	77*	2.93
3L**	FBtr0346541	Hsp27	TAA	214	86*	157.97
3L**	FBtr0076667	Idh	TGA	417	40*	34.56
3L**	FBtr0074916	Su(z)12	TAA	856	24	6.41
3L	FBtr0075394	mbf1	TGA	146	108*	5.02
3L	FBtr0075412	fax	TAA	419	94*	3.86
3R**	FBtr0310464	Kmn2	TAG	94	28*	18.67
3R	FBtr0083005	VhaPPA1-1	TAG	213	75	3.51
3R	FBtr0084739	lili	TGA	528	67	4.87

Tabla 2.1: En esta tabla se muestran 30 ejemplos de los 1176 transcriptos mencionados como candidatos de LCP. En cada caso se informa el cromosoma al que pertenece, el código de identificación (ID) del transcripto, el nombre del gen de origen, el tipo de codón de parada y su posición (nt). Además, se indica el tamaño de la extensión encontrada según su longitud en aminoácidos (aa), y la densidad ribosomal allí detectada. Se indican con doble asterisco (**) en la primera columna aquellos casos que se ilustran en esta tesis. Aquellos casos con múltiples eventos de LCP se indican con un asterisco (*) en la sexta columna.

en esta sección vamos a mencionar a manera de ejemplo algunos de ellos. Sin embargo, todos los eventos de LCP identificados participan del análisis de la secuencia contexto en el próximo capítulo. En la Tabla 2.1 se listan 30 de los transcriptos que presentan eventos de LCP, determinados en esta tesis. Los casos que ilustramos a continuación están indicados con doble asterisco (**) en la primera columna de la tabla.

En la Fig. 2.6 se muestra el perfil de densidad ribosomal de la isoforma D del gen *Kmn2* (transcripto FBtr0310464), en el cromosoma 3R. Este gen codifica un componente esencial del complejo Ndc80 del cinetocoro Mis12-Ndc80. En el transcripto FBtr0310464 se observa doble evento de LCP, que no se encuentra reportado en FlyBase. Posee un codón de terminación TAG anotado en la posición 282 nt. También presenta un segundo y un tercer codón de parada (ambos TGA) ubicados en las posiciones 336 nt y 426 nt, respectivamente; a una distancia de 54 nt y 144 nt del primer codón de parada. Las lecturas observadas entre ellos corresponden a una primera extensión de 17 aminoácidos (DSSDRSLVLFSQYLCST), situada entre el primer y segundo codón de parada; y una segunda extensión de 11 aminoácidos más (YVCCRHSIEGD) entre el segundo y tercer codón de parada. En este caso, la extensión total hasta el tercer codón de parada tiene una longitud de 30 aminoácidos y una cobertura de alrededor de 20 lecturas de huellas ribosomales.

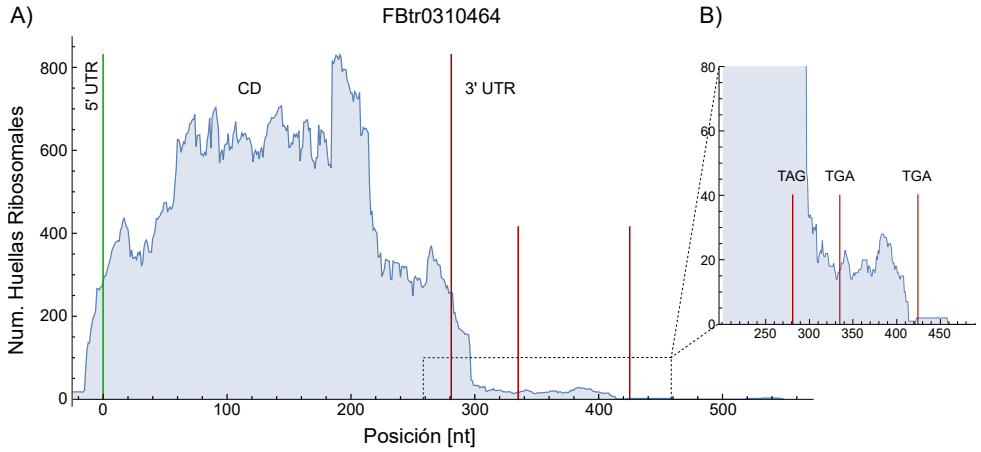

Figura 2.6: Perfil ribosomal de una isoforma del gen *Kmn2* (transcripto FBtr0310464), con doble evento de LCP. Los eventos de LCP ocurren en las posiciones 282 nt y 336 nt, correspondientes al primer y segundo codón de parada (TAG y TGA). La extensión total del péptido traducido es DSSDRSLVLFSQYLCST*YVCCRHSIEGD, posee una longitud de 30 aminoácidos y está ampliamente conservado en la familia *Drosophilidae*.

La Fig. 2.7 muestra el perfil ribosomal de la isoforma A del gen *ghiberti (ghi)* en el cromosoma 3L (transcripto FBtr0076462); que presenta tres eventos de LCP descubiertos en nuestro análisis, pero no figura en la lista de transcriptos con LCP de FlyBase. El gen *ghi* codifica una proteína homóloga a la subunidad reguladora pequeña de la serina palmitoiltransferasa en mamíferos. Las mutaciones en *ghi* afectan los primeros pasos de la biosíntesis de esfingolípidos y alteran la citocinesis

meiótica de los machos. El transcripto FBtr0076462 tiene su codón de terminación TAA anotado en la posición 246 nt. Sin embargo, la traducción parece extenderse atravesando el segundo y tercer codón de parada (TGA, 132 nt y TAA, 177 nt después); e inclusive más allá del cuarto (TAG). La primera extensión posee un tamaño de 43 aminoácidos (NTDQTMPRPIVCTVLRSNIQNSRVSVPRRDALSNPHAFIHRNQ), con un número sustancial de huellas ribosomales alrededor del segundo codón de parada. Otro aspecto interesante a tener en cuenta, es el que se aprecia claramente en la ampliación de la región 3' UTR en el perfil de densidad ribosomal de la Fig. 2.7. Allí se observa que los picos de densidad ribosomal en los codones de parada de la extensión son semejantes a los picos de densidad en los codones de parada anotados. Precisamente, esto constituye un rasgo característico de la terminación de la traducción (Ingolia et al., 2011), y como se mencionó anteriormente, una evidencia empírica de eventos de LCP (Dunn et al., 2013; Jungreis et al., 2016).

También se realizó un análisis de sintenia con otras especies de *Drosophila*. En este sentido, se buscaron secuencias peptídicas similares a la extensión, utilizando tBLASTN y la base de datos de NCBI de nucleótidos no redundantes traducidos. Para este análisis usamos como secuencia de consulta sólo la primera extensión, es decir, la correspondiente a la región entre en el codón de parada anotado y el segundo codón de parada en el mismo marco de lectura, dando la secuencia NTDQTMPRPIVCTVLRSNIQNSRVSVPRRDALSNPHAFIHRNQ. Encontramos que la secuencia extendida en cuestión está conservada con bastante similitud en otras cinco especies del género *Drosophila spp.*, como se muestra en el alineamiento múltiple de la Fig. 2.7. Este mismo argumento de conservación de la extensión es la base del criterio utilizado previamente por Jungreis *et al.* para establecer numerosos casos de LCP (Jungreis et al., 2011).

La Fig. 2.8 muestra otro ejemplo de un evento simple de LCP identificado en nuestro análisis, que también fue identificado por Dunn, pero no está reportado en FlyBase. El perfil ribosomal corresponde a la isoforma C del gen *CG13887* en el cromosoma 3L (transcripto FBtr0072583). Este gen codifica una proteína asociada al receptor B 29/31, con dominio transmembrana, involucrada en procesos de transporte intracelular de proteínas del retículo endoplasmático al Golgi. El transcripto FBtr0072583 presenta un evento de LCP en el codón de terminación anotado en la posición 687 nt (TAA). Se observa que las huellas ribosomales sobrepasan el segundo codón de parada (TAA), ubicado a una distancia de 93 nt respecto del codón de terminación anotado, dando una extensión 30 aminoácidos, THWSQVQHLNNDKTYKHGGGAAKAGTFKRL. Tras realizar la búsqueda de alineamientos de esta extensión usando tBLASTN, y un alineamiento múltiple con las secuencia encontradas, observamos que esta extensión se encuentra conservada en otras cuatro especies del género *Drosophila spp.*, como se muestra en el alineamiento múltiple de la Fig. 2.8.

El perfil ribosomal de la Fig. 2.9 muestra un evento doble de LCP identificado en nuestro análisis y no reportado en FlyBase, correspondiente a la isoforma A del gen *Nurf-38* (transcripto FBtr0072343) en el brazo R del cromosoma 2. Este gen codifica una pirofosfatasa de 38 kD, que cataliza el deslizamiento de nucleosomas dependiente de ATP y facilita la transcripción de cromatina. Aquí se observa una extensión corta de cinco aminoácidos (DVLLG) entre el codón de parada anotado en la posición 1017 nt (TGA), y el segundo codón de parada en la posición 1035 nt (TGA), ubicado sólo 18 nt después. Allí ocurre un segundo evento de LCP con ev-

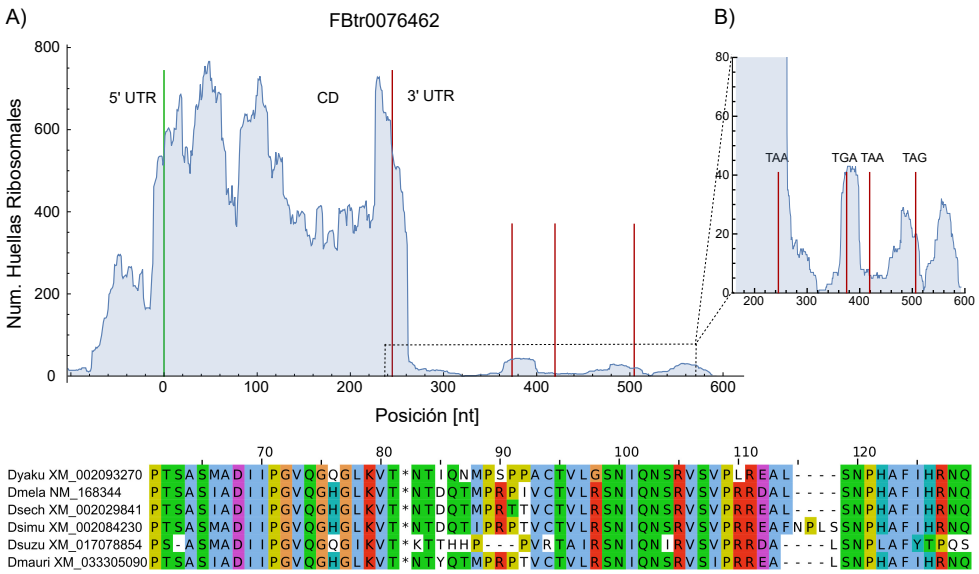

Figura 2.7: Perfil de densidad ribosomal de la isoforma A del gen *ghiberti* (transcripto FBtr0076462), con triple evento de LCP. Se observan numerosas huellas ribosomales alrededor del segundo codón de parada (TGA, posición 378 nt). La extensión de la proteína entre el codón de parada anotado y el siguiente, NTDQTMPRPIVCTVLRSNIQNSRVSVPRRDALSNPHAFIHRNQ, presenta similitud significativa con las extensiones correspondientes a otras cinco especies del mismo género.

idente cantidad de huellas ribosomales, dando lugar a una extensión de 20 aminoácidos (FHHKIHDLSVVQSNRKCLHL) que alcanza a un tercer codón de parada (TGA), ubicado en la posición 1116 nt (81 nt después del segundo). La búsqueda de extensiones similares en otros miembros de la familia con tBLASTN demuestra que esta extensión esta conservada en cuatro especies del género *Drosophila*, como se muestra en el alineamiento múltiple de la Fig. 2.9.

Finalmente, la Fig. 2.10 muestra el perfil de densidad ribosomal de otro evento de LCP identificado en nuestro estudio. Se trata del transcripto FBtr0079297, perteneciente al gen *CG11070* en el cromosoma 2L. Este gen codifica el factor de conjugación de ubiquitina E4-A. Se observa presencia de lecturas con densidad ribosomal posteriores al codón de parada anotado (TGA), ubicado en la posición 2982 nt. Aquí se produce una extensión de 16 aminoácidos de longitud (GSSEPAAAGSVAQPKS), que abarca 51 nt en la región 3' UTR, hasta interrumpirse en un segundo codón de parada (TAG), ubicado en la posición 3033 nt. Utilizando tBLASTN buscamos secuencias similares a esta extensión en la base de datos de transcriptos no redundantes de NCBI, traducidos a proteínas. En el panel inferior de la Fig. 2.10 se muestra el alineamiento múltiple correspondiente a los fragmentos de esta ubiquitina. Este análisis revela que la extensión peptídica tiene homología con las extensiones correspondientes de las especies *D. simulans*, *D. sechellia*, *D. mauritiana* y *D. yakuba*. Sin embargo, en *D. erecta* el gen homólogo *GG10427* presenta una Lisina (K) en vez del codón de parada. Esto sugiere que la extensión en

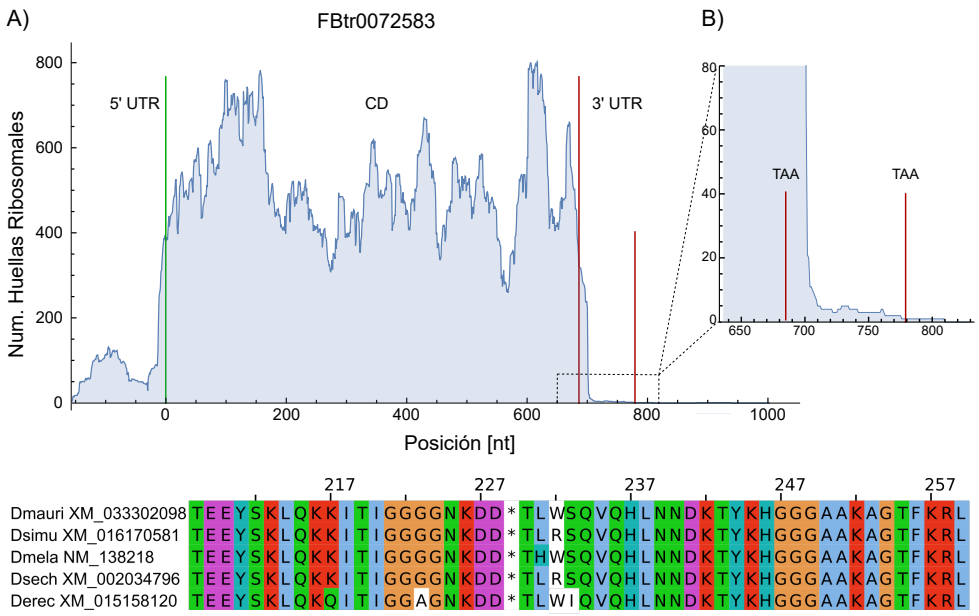

Figura 2.8: Perfil de densidad ribosomal de la isoforma C del gen *CG13887* (transcripto FBtr0072583), con un evento simple de LCP. La extensión de 30 aminoácidos presenta un elevado grado de conservación con las extensiones correspondientes a otras 4 especies del mismo género.

las especies mencionadas tiene una función biológica, ya que está incorporada a la región codificante del gen *GG10427* de *D. erecta*.

La Fig. 2.11 muestra otros dos claros ejemplos de nuevos eventos simples de LCP identificados en nuestro estudio, pero que no fueron reportados en los análisis previos realizados por Jungreis et al. (2011) y Dunn et al. (2013), y que tampoco se encuentran anotados en FlyBase. En el panel superior se muestra el perfil ribosomal asociado al transcripto FBtr0339562 (gen *CG31673*), con un evento simple de LCP, que no fue reportado en estudios previos. Este transcripto posee una primera extensión de 46 aminoácidos de longitud (EHAKFEWKKLYLYPTRIGKSILKNDIKHINVELVNRNYNGLIFFLV), que se extiende entre el codón de terminación anotado y el segundo: TGA y TAA, respectivamente. En el panel inferior se muestra una extensión en la traducción del transcripto FBtr0100268 (gen *Chrac-14*), en el cromosoma 2L. Este gen codifica a la subunidad 3 de la ADN polimerasa épsilon, perteneciente al complejo proteínico 14kD de accesibilidad de cromatina en *D. melanogaster*. En este caso el gráfico muestra un evento de LCP a partir del primer codón de terminación (TAA), anotado en la posición 387. Podemos ver que la densidad ribosomal claramente sobrepasa el segundo codón de parada TAG ubicado 87 nt después del anterior.

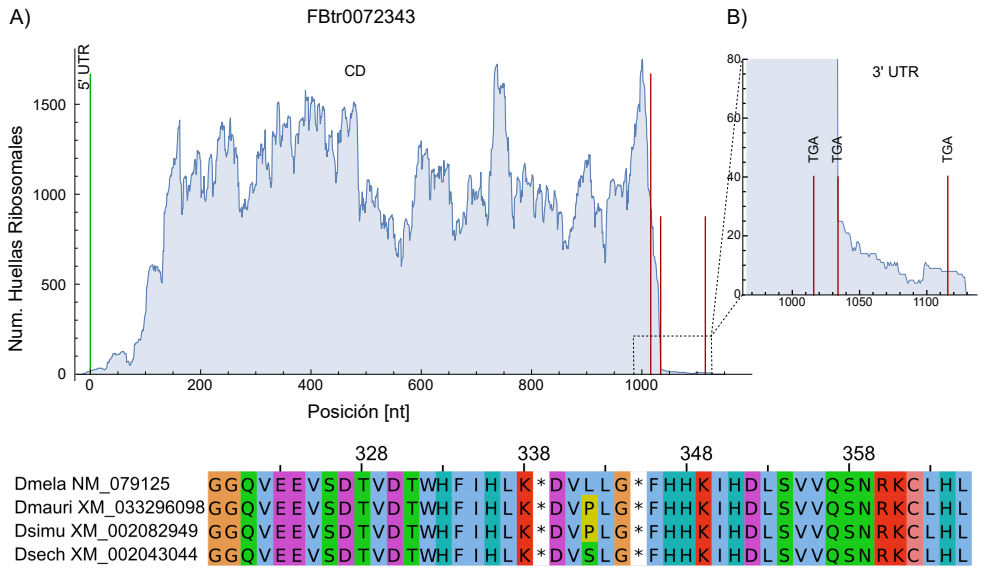

Figura 2.9: Perfil de densidad ribosomal de la isoforma A del gen *Nurf-38* (transcripto FBtr0072343), perteneciente al cromosoma 2R. En este perfil se muestra un evento doble de LCP, donde la región extendida en este caso abarca hasta el tercer codón de parada. En primera instancia, se observa un tramo corto de la extensión (DVLLG) entre el codón de terminación anotado (TGA) en la posición 1017 nt, y el segundo codón de parada (TGA) ubicado 18 nt después. La extensión continúa durante otros 81 nt hasta detenerse en el tercer codón de parada (TGA) en la posición 1116 nt, incorporando otros 20 aminoácidos (FHHKIHDLSVVQSNRKCLHL). La extensión completa identificada DVLLG*FHHKIHDLSVVQSNRKCLHL, presenta un elevado grado de identidad con las extensiones de otros miembros del género *Drosophila*.

2.4. Confirmación de eventos de LCP por espectrometría de masa

La proteómica basada en la espectrometría de masas (EM) se ha establecido como una tecnología indispensable para la identificación de proteínas y la interpretación de la información codificada en los genomas. El análisis sistemático de secuencia primaria, modificaciones postraduccionales (PTM) o interacciones de gran número de proteínas es el objetivo de la proteómica. Básicamente, un experimento típico de proteómica consta de cinco etapas:

- Las proteínas a analizar se aíslan de la muestra lisada mediante fraccionamiento bioquímico o selección por afinidad.
- Las proteínas se degradan enzimáticamente a péptidos, generalmente usando tripsina. Esto proporciona una ventaja, dado que la EM de proteínas enteras es menos sensible que la MS peptídica.

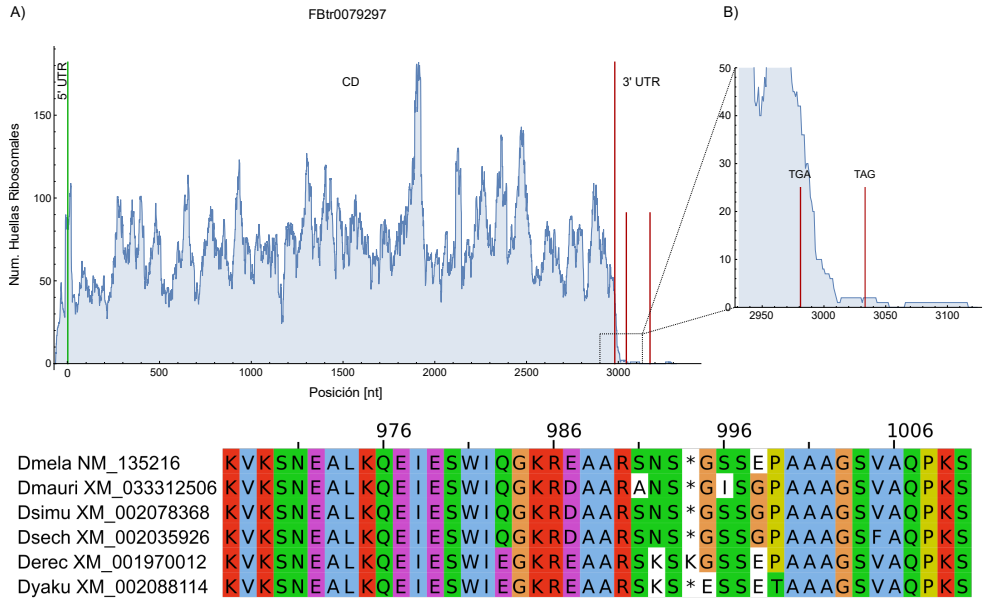

Figura 2.10: Perfil de densidad ribosomal del transcripto FBtr0079297 (gen *CG11070*), con un evento de LCP en el cual encontramos una extensión válida, GSSEPAAAGSVAQPKS, no anotada en FlyBase hasta el momento. Se remarca que la secuencia homóloga correspondiente a *D. erecta* presenta una lisina en vez del codón de parada.

- Los péptidos se separan en capilares mediante cromatografía líquida de alta performance (HPLC), y son vaporizados en pequeñas gotas altamente cargadas.
- Los péptidos cargados ingresan al espectrómetro de masas, donde se produce una fragmentación que toma el espectro de masas. La consecuencia más importante del proceso de fragmentación es que genera un espectro de masas característico y reproducible como una "huella digital" que sirve para identificar los péptidos. Los espectros de masas generalmente se registran en forma de un gráfico de barras, donde la altura determina la intensidad iónica.
- Los espectros se almacenan para compararlos con bases de datos de secuencias de proteínas, mediante programas específicos. El resultado es la identificación de los péptidos presentes en la base de datos de secuencias con los espectros de masas obtenidos para una muestra dada.

En esta tesis se utilizaron espectros obtenidos en experimentos de identificación de proteínas a gran escala y almacenados en bases de datos públicas, en particular el repositorio de *Peptide Atlas* (http://www.peptideatlas.org/repository/). En este sentido, se utilizaron datos proteómicos correspondientes a los espectros de masa de *D. melanogaster*, obtenidos mediante el uso de un espectrómetro de masas con plataforma de trampa de iones LTQ Orbitrap (Brunner et al., 2007). Los espectros aquí utilizados corresponden a embriones de 0-2 hs, es decir, el mismo tiempo de desarrollo estudiado por el perfil ribosómico en la sección previa. Los códigos de

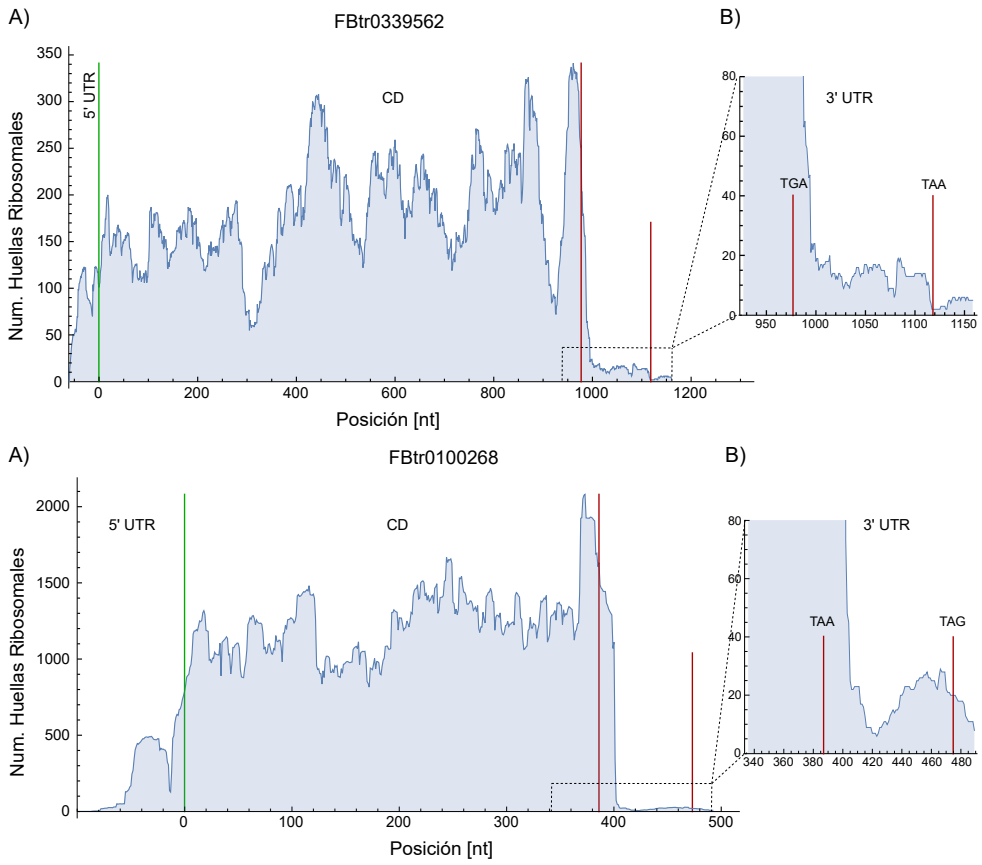

Figura 2.11: El panel superior muestra el perfil ribosomal del transcripto FBtr0339562 (gen *CG31673*), con un evento simple de LCP no reportado en FlyBase. La extensión identificada es EHAKFEWKKLYLYPTRIGKSILKNDIKHINVELVNRNYNGLIFFLV, entre el codón de terminación anotado (TGA) y el segundo (TAA). El panel inferior muestra el perfil ribosomal del transcripto FBtr0100268 (gen *Chrac-14*), mostrando un evento simple de LCP tampoco reportado en FlyBase. La extensión identificada es LRIGETIRINYPLQSLRTKLVLSINKKH, comprendida entre el codón de terminación anotado TAA y el segundo codón de parada TAG, ubicado 87 nucleótidos después.

acceso en el repositorio de *Peptide Atlas* de los espectros de masas utilizados aquí son PAe001369 y PAe001388.

A los fines de detectar proteínas de baja abundancia durante el desarrollo de la mosca de la fruta, los extractos proteicos fueron procesados con diversos protocolos detallados en Brunner *et al.* Por ejemplo, péptidos ricos en cisteína fueron etiquetados mediante el agregado de isótopos mediante la técnica ICAT (del inglés *Isotope-coded affinity tag*) (Brunner et al., 2007). Así, para la identificación de estos espectros es necesario considerar este agregado de isótopos, además de ciertas

ID del transcripto	Pos. codón de parada	Pos. inicio del frag.	ID del espectro
FBtr0078997	253	256	JK050812_06.5937.5937
FBtr0076667	417	417	050701_SG_S08.3293.3293
FBtr0084324	2250	2248	050701_SG_S16.4403.4403
FBtr0074916	856	868	SG_20050509_10.9092.9092

Tabla 2.2: En esta tabla se muestran los principales datos de algunos fragmentos espectrales identificados por *Peptide-Shaker*, que coinciden con las extensiones derivadas de LCP en los transcriptos seleccionados. Esta lista discrimina información pertinente como el ID del transcripto del cual deriva el espectro analizado, y su posición de terminación; la posición de inicio del fragmento espectral y la secuencia peptídica en cuestión; y los valores de score asignados por los motores de búsqueda que los identificaron.

modificaciones postraduccionales (PTM). En este sentido, las muestras PAe001369 y PAe001388 cuentan con dos PTM: la oxidación de la metionina (de 16 Da, considerada como PTM variable); y la carbamidometilación de cisteína (de 57 Da, considerada como PTM fija). La diferencia de masa considerada para péptidos marcados (o etiquetados) isotópicamente con reactivos ICAT es de 9 Da (Rutschow et al., 2008; Rosenthal y Harvey, 2010; Hsueh-Fen y Hsuan-Cheng, 2012).

Más allá de las extensiones encontradas mediante el análisis de perfiles ribosomales, nuestro interés se enfocó en buscar evidencia adicional de la existencia de extensiones producto de eventos de LCP. Por este motivo, se utilizó nuestra lista de transcriptos con sus extensiones traducidas a proteínas, como base de datos de secuencias contra la cual se contrastaron los espectros. Para la identificación de los espectros se utilizaron los motores de búsqueda existentes en el software *SearchGui* (http://searchgui.googlecode.com) (Vaudel et al., 2011), con los siguientes parámetros de búsqueda: A) valor esperado de corte (E-value): 0,1; B) precursor de carga: 1-4; C) tolerancia de masa del fragmento: 0,02 Da; D) precursor de la tolerancia de masa por ión: 0,5 ppm (valores: 0,1/0,5/1,0 ppm).

Los resultados de asociar los espectros de masas existentes con nuestra lista de proteínas que incluye las extensiones, fueron analizados posteriormente con el software *Peptide-Shaker* (Vaudel et al., 2015). Este programa hace un análisis de significancia tomando en cuenta una lista de péptidos con aminoácidos al azar. Los espectros resultantes más confidentes, aquellos con péptidos que contienen parte de la extensión proteica, fueron seleccionados y se presentan en la Tabla 2.2.

De esta forma, se identificaron tres espectros de masa de alta confidencia, asociados con fragmentos de las secuencias extendidas, posteriores a los codones de parada anotados. Estos resultados representan evidencia experimental acerca de las extensiones encontradas por el método de los perfiles ribosomales, mostrado en la sección anterior. Este número de péptidos identificados es pequeño frente a los 1176 posibles nuevos eventos de LCP; sin embargo, esto se debe a la pequeña concentración relativa en la muestras de las proteínas con extensiones, respecto de la cantidad de la proteína anotada.

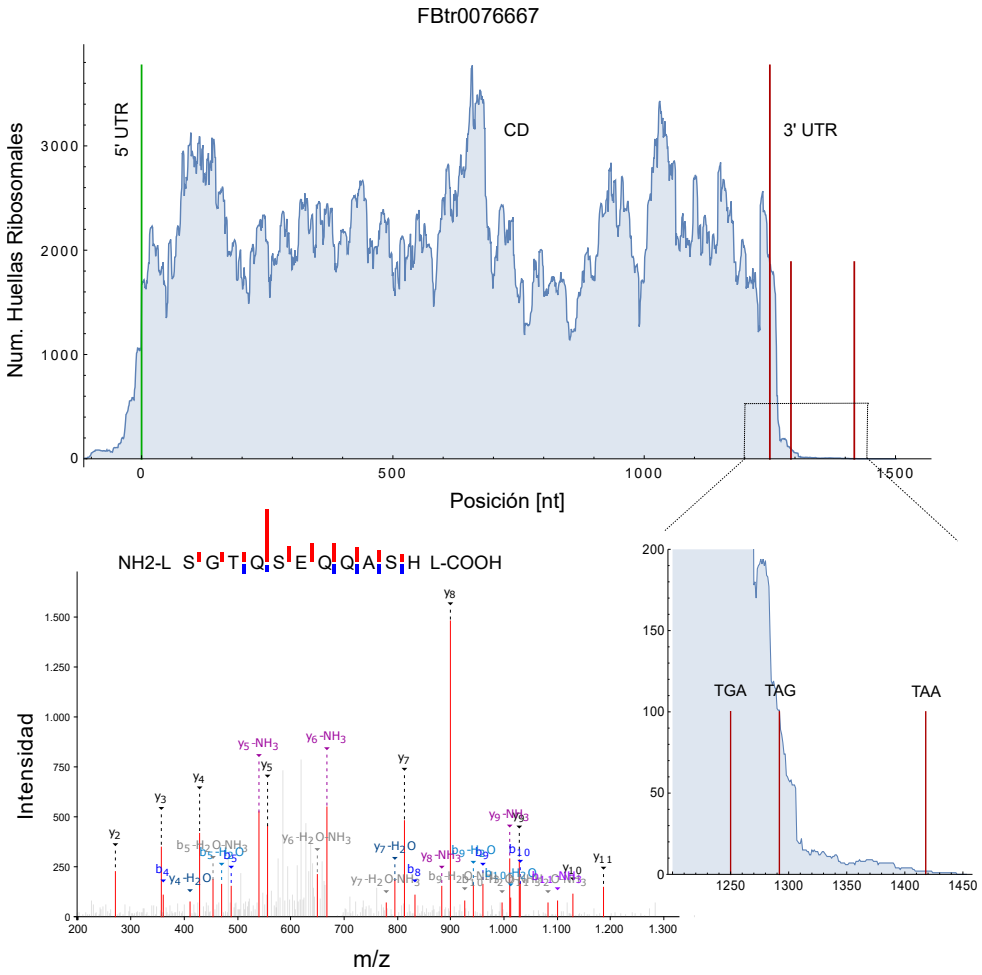

Figura 2.12: El panel superior de esta figura muestra el perfil de densidad ribosómica del transcripto FBtr0076667, una isoforma del gen *Idh*. Se observa un evento de LCP en el codón de terminación anotado (TGA) en la posición 1249 nt, con una primera extensión de 13 aminoácidos (LSGTQSEQQASHL). Esta extensión continúa a través de un segundo evento de LCP en el segundo codón de terminación (TAG) ubicado 42 nt posteriores al codón anotado, generando una extensión total de 40 aminoácidos (LSGTQSEQQASHLKTTQRCLVLVTIFTIFPFVFVDVHVLG), que alcanza a un tercer codón de terminación (TAA) ubicado 126 nt después del anterior. En el panel inferior se muestra el espectro de masa que identifica la primera extensión. La intensidad de cada ión se muestra en relación con el fragmento de la secuencia peptídica, ubicada en la margen superior izquierda.

En la Fig. 2.12 se muestra el perfil de densidad ribosomal correspondiente a la isoforma A del gen *Idh* (transcripto FBtr0076667), ubicado en el brazo L del cromosoma 3. Este gen codifica la enzima isocitrato deshidrogenasa (*Idh*) de 416

aminoácidos, que cataliza la descarboxilación oxidativa del isocitrato. El perfil ribosomal asociado presenta un evento doble de LCP. El primero de éstos ocurre en el codón de terminación anotado en la posición 1249 nt (TGA); y otro ocurre en el segundo codón de terminación (TAG), ubicado 42 nt posteriores al codón anotado. Se observa que las huellas ribosomales alcanzan el segundo codón de terminación (TAG), dando hasta allí una primera extensión de 13 aminoácidos, LSGTQSEQQASHL. Así mismo, las huellas ribosomales continúan hasta alcanzar un tercer codón de terminación (TAA) ubicado 126 nt después del anterior, generando una extensión total de 40 aminoácidos (LSGTQSEQQASHLKTTQRCLVLVTIFTIFPFVFVDVHVLG). En el panel inferior de la Fig. 2.12 se muestra el espectro de masa que identifica el péptido de 13 aminoácidos, LSGTQSEQQASHL. El ID del espectro identificado es 050701_SG_S08.3293.3293, y se encuentra en la muestra PAe001369 del repositorio *Peptide Atlas*. Esta coincidencia fue identificada con error m/z nulo por dos motores de búsqueda diferentes: el OMSSA y el MS-GF, con valores de expectación $4{,}98\times10^{-6}$ y $6{,}80\times10^{-12}$, respectivamente.

En el panel inferior de la Fig. 2.12, los picos azules representan los iones "b" de carga positiva, mientras que los picos rojos representan los iones "y" de carga negativa. Los picos observados en un espectro de fragmentos reflejan la abundancia de iones de fragmentos producidos en la celda de colisión de un espectrómetro de masas. La secuencia del péptido está determinada por la diferencia de masa entre estos picos. La secuencia de fragmentación correspondiente al espectro analizado se indica en la parte superior del espectro.

Otro ejemplo de validación por espectrometría de masas es el caso de una isoforma del gen *CG3792*, ubicado en el brazo L del cromosoma 2. La función molecular de este gen es poco precisa, pero está involucrado en el proceso biológico de síntesis de oligosacáridos y la glicosilación de proteínas. El transcripto asociado, FBtr0078997, presenta un perfil de densidad ribosomal con un evento de LCP en el codón de terminación anotado (TAA), separado por 759 nt del codón de inicio (TGA). Si bien la densidad ribosomal es apreciable solo en una fracción de la extensión, de unos 50 nt, excede los 18 nt después del codón de parada que cubren los ribosomas ubicados en TAA. Este evento de LCP da lugar a la extensión de 17 aminoácidos AVRIEVTKVETDIHREI. En la búsqueda de espectros de masa coincidentes con esta secuencia, encontramos dos espectros altamente confidentes, de los cuales sólo es mostrado en la Fig. 2.13 el espectro con ID: JK050812_06.5937.5937, de la muestra PAe001369. El panel inferior de la Fig. 2.13 muestra la secuencia de fragmentación correspondiente al péptido identificado de 14 aminoácidos, IEVTKVETDIHREI. Esta coincidencia fue identificada con error m/z de 0,02 por dos motores de búsqueda diferentes: el OMSSA y el MS-Amanda, con valores de expectación $8{,}3\times10^{-5}$ y $6{,}2\times10^{-3}$, respectivamente. Ambos espectros son casi idénticos, variando únicamente en la presencia de una carga iónica negativa en el aminoácido valina del fragmento que se muestra; mientras que en el segundo caso, esa valina no posee carga.

Finalmente, el tercer ejemplo de validación por espectrometría de masas es el caso de la isoforma A del gen *Su(z)12* (CG8013-RA), ubicado en el brazo L del cromosoma 3. Este gen es un miembro del grupo de genes *Polycomb* con productos implicados en el silenciamiento transcripcional. El gen *Su(z)12* codifica una subunidad del complejo represivo *Polycomb 2* (PRC2), donde se requiere para la actividad de la histona metiltransferasa que produce la trimetilación de la histona H3

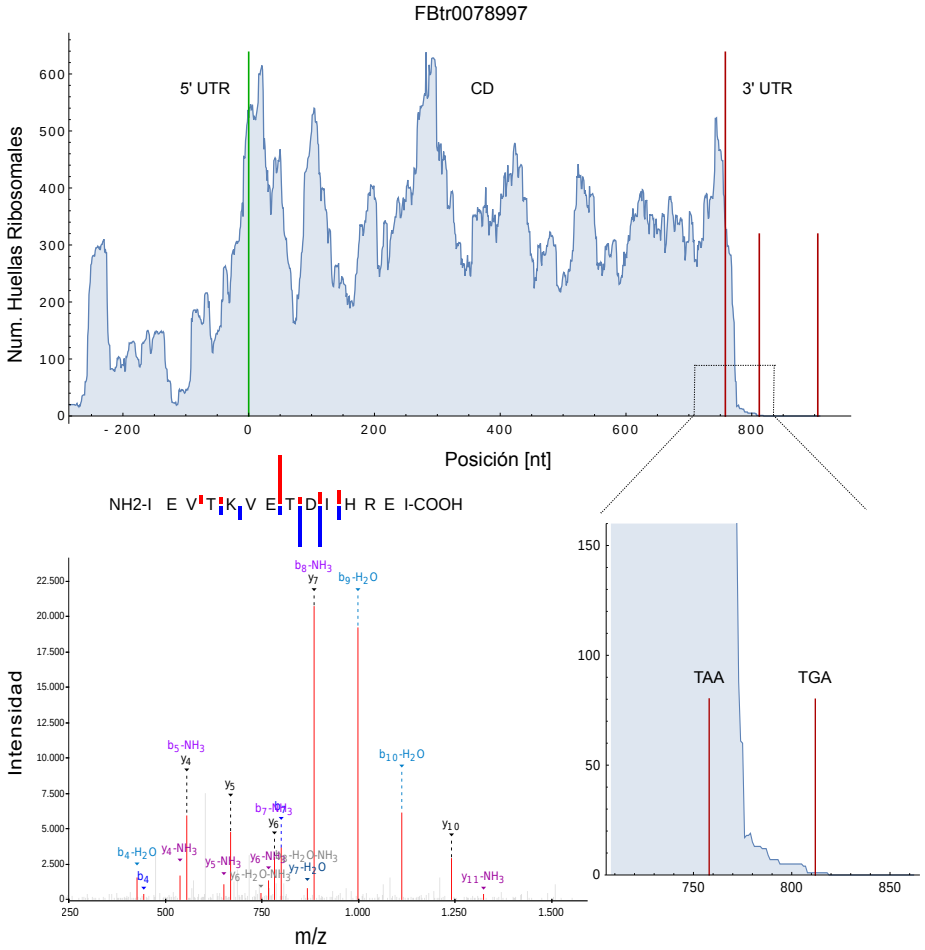

Figura 2.13: El panel superior de esta figura muestra el perfil de densidad ribosómica del transcripto FBtr0078997, una isoforma del gen *CG3792*. Se observa un único evento de LCP en el codón de terminación anotado en la posición 759 nt (TAA), generando una extensión de 17 aminoácidos (AVRIEVTKVETDIHREI) que finaliza en el segundo codón de parada (TGA), 54 nt después del primero. En el panel inferior se muestra el espectro de masas que identifica esta extensión. La intensidad de cada ión se muestra en relación con el fragmento de la secuencia peptídica, ubicada en la margen superior izquierda.

en la lisina 27 (H3-K27m3). El transcripto asociado, FBtr0074916, presenta un perfil de densidad ribosomal con un evento de LCP en el codón de terminación anotado (TGA), ubicado en la posición 2568 nt respecto del codón de inicio. La densidad ribosomal se extiende más allá del primer codón de parada, alcanzando al segundo codón de parada (TGA) en el mismo marco de lectura, tal como se muestra en el panel superior de la Fig. 2.14. El péptido correspondiente a esta extensión posee una longitud de 98 aminoácidos (ISNNTVLNKRQRYSDGSPGTGIGNGHGGGSGSG-

ANRNKSNNHSLPATSNNASSSSSNSKRAIARRRSTSERTKASGSTGGGAGGV-RTRLSVPAKYERR). El panel inferior de la Fig. 2.14 muestra el espectro altamente confidente que coincide con esta secuencia, de ID: SG_2005050-9_10.9092.9092, perteneciente a la muestra PAe001388. También se indica la secuencia de fragmentación del péptido identificado, de 24 aminoácidos: YSDGSPGTGIGNGHGGGS-GSGANR. Esta coincidencia fue identificada con una confidencia de 99 % y con error m/z 0,35 por el motor de búsqueda OMSSA, con valor de expectación $2{,}17\times10^{-5}$.

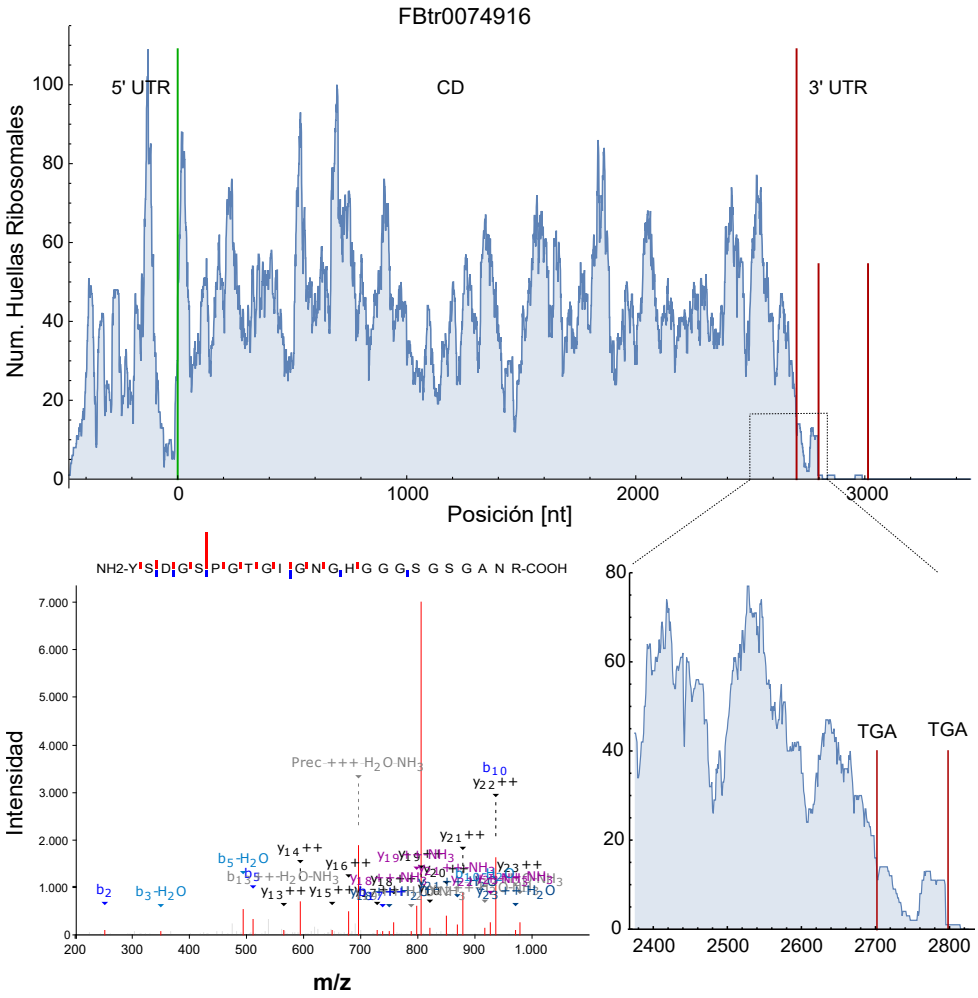

Figura 2.14: El panel superior muestra el perfil de densidad ribosomal del transcripto FBtr0074916, una isoforma del gen *Su(z)12*. Se observa un evento de LCP en el codón de terminación anotado en la posición 2568 (TGA), generando una extensión de 98 aminoácidos. En el panel inferior se muestra el espectro de masas que identifica el fragmento YSDGSPGTGIGNGHGGGS-GSGANR. La intensidad de cada ión se muestra en relación con el fragmento de la secuencia peptídica, ubicada en la margen superior izquierda.

2.5. Análisis de enriquecimiento por ontología génica

Una vez identificados los transcriptos con eventos de LCP, se procedió a asociar las proteínas con su función por ontología. Las ontologías pueden definirse como una representación formal de los tipos, propiedades y relaciones entre las entidades en un dominio específico del conocimiento. Las ontologías son creadas para organizar la información y limitar la complejidad. En el dominio de la biología existen diversas ontologías para organizar el conocimiento acerca de genes, proteínas, dominios. En el caso de genes existen tres categorías ontológicas que están establecidas y reguladas por una organización llamada *Gene Ontology (GO)* (http://www.geneontology.org) (Ashburner et al., 2000; The Gene Ontology Consortium, 2017). Las categorías GO están compuestas por: proceso biológico (vías y procesos compuestos por las actividades de múltiples productos génicos); función molecular (acción desarrollada a nivel molecular por el producto de un gen); y componente celular (las ubicaciones relativas a las estructuras celulares en las que un producto de un gen desarrolla una función). Así, los genes de los diversos organismos pueden estar o no asociados a algunos de los términos existentes en cada una de estas tres categorías. Los términos GO se usan comúnmente para la anotación gen/producto genético. Dado un conjunto de genes, es posible determinar cuáles términos GO, relativos a cada una de las categorías, están enriquecidos en esa lista. De esta manera, el análisis de enriquecimiento (ANEN) de los términos GO permite inferir la naturaleza biológica de un conjunto de genes de entrada, comparándolo con el universo de todos los genes anotados.

En esta sección, se utilizó el ANEN con el objetivo de determinar si los genes que presentan LCP están asociados a algún proceso biológico particular, o si desarrollan alguna función. Dado que el codón de terminación podría indicar una señal adicional al propio fin de la traducción, el ANEN fue hecho en forma independiente para cada uno de los codones de parada asociados a eventos de LCP programado. Para ello, se listaron los transcriptos con LCP con fuerte señal de densidad ribosomal en la extensión, excluyendo así los eventos de LCP que podrían ocurrir muy ocasionalmente, en tres grupos según el codón de parada anotado (TAA, TAG o TGA). Cada grupo fue evaluado según el software *FlyEnrichr* (amp.pharm.mssm.edu/FlyEnrichr) (Chen et al., 2013; Kuleshov et al., 2016), introduciendo como entrada únicamente los nombres de los genes asociados a los transcriptos agrupados según el tipo de codón de terminación TAA, TAG y TGA (127, 114 y 93 genes respectivamente). La salida de *FlyEnrichr* consiste en tablas de términos GO, asociadas a cada categoría, con su código GO, p-valor, p-valor ajustado y el z-score. La Fig. 2.15 muestra un resumen de las principales anotaciones GO para las categorías de procesos biológicos, componentes celulares y funciones moleculares representadas en los conjuntos de genes estudiados.

En la Fig. 2.15-A se representa la categoría de procesos biológicos para los tres codones de parada. En el conjunto de genes con codón de terminación TAA (barras rojas) se observa que las anotaciones GO más sobresalientes corresponden a genes vinculados a la traducción citoplasmática (GO:0002181) y la traducción en general (GO:0006412). Estos procesos poseen un p-valor ajustado de $3{,}84\times10^{-15}$ y $1{,}26\times10^{-12}$, respectivamente. Con una representación mucho menor, se observan otros dos procesos relativos también a la traducción: como el alargamiento traduc-

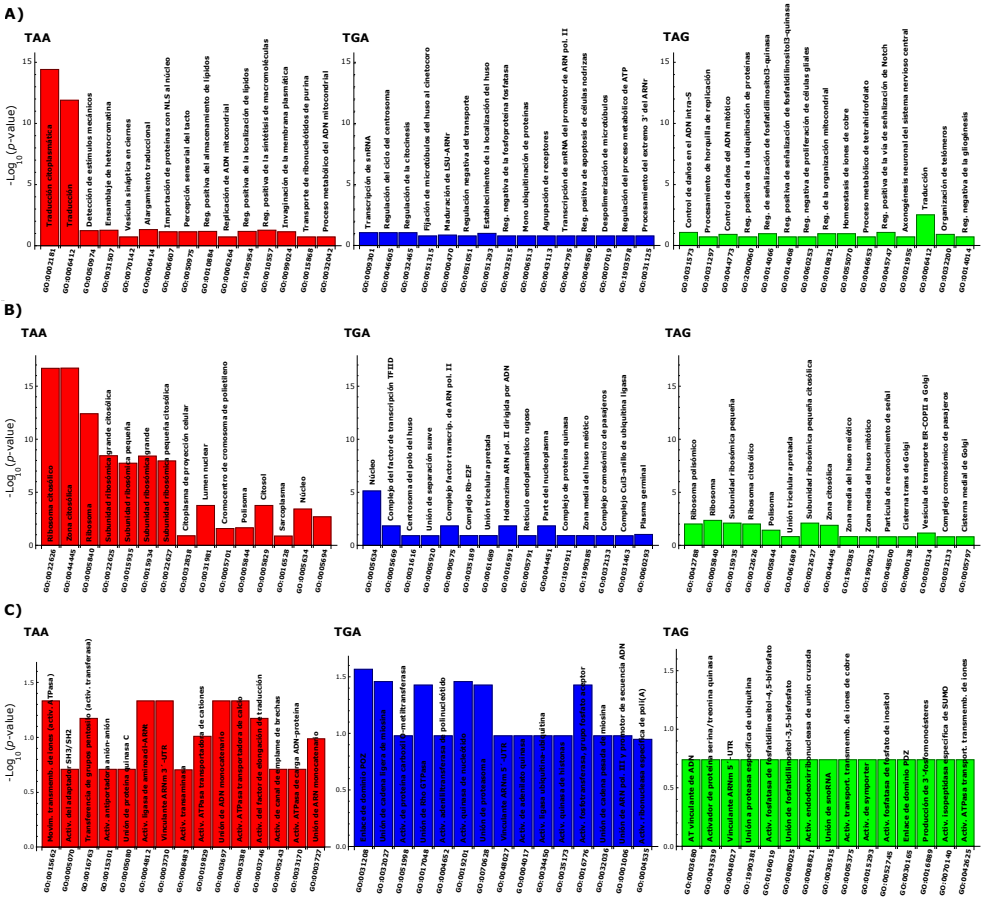

Figura 2.15: Resumen de términos GO de según las categorías de procesos biológicos (A), componentes celulares (B) y funciones moleculares (C) para los genes analizados. Estos genes representan los conjuntos de transcriptos con presencia de LCP, separados en función del codón de terminación: TAA (barras rojas), TGA (barras azules) y TAG (barras verdes). La altura de cada barra representa el valor de tomar $-\log_{10}$ del p-valor ajustado de cada término GO.

cional (GO:0006414, p-valor ajustado: $4{,}78\times10^{-2}$) y la regulación positiva de la biosíntesis de macromoléculas (GO:0010557, p-valor ajustado: $4{,}64\times10^{-2}$). En función de esta observación se puede concluir que los transcriptos que terminan en TAA y que presentan LCP están asociados con el proceso biológico de traducción. Por otro lado, en el caso de los genes con codón de terminación TGA (barras azules), no observamos procesos que estén sobrerrepresentados en forma significativa, es decir con p-valor ajustado menor a $5{,}0\times10^{-2}$. En cuanto al conjunto de genes con codón de terminación TAG (barras verdes), el único proceso biológico sobrerrepresentado significativamente es el de traducción (GO:0006412, p-valor ajustado: $3{,}04\times10^{-3}$) en forma similar al caso de codón de terminación TAA.

En la Fig. 2.15-B se representa la categoría de componentes celulares. En el conjunto de genes con codón de terminación TAA (barras rojas) se observa que las dos anotaciones GO más preponderantes corresponden a componentes vinculados a ribosomas eucarióticos 80S ubicados en el citoplasma (GO:0022626, p-valor ajustado: $2{,}07\times10^{-17}$). Este componente esta relacionado con el proceso de traducción. En relación al conjunto de genes con codón de terminación TGA (barras azules), la anotación GO más prominente concierne al núcleo (GO:0005634, p-valor ajustado: $7{,}41\times10^{-6}$), probablemente vinculados al complejo de transcripción. En lo referente al conjunto de genes con codón de terminación TAG (barras verdes), los componentes celulares con mayor cantidad de anotaciones GO corresponden a ribosomas (GO:0005840 y GO:0022626), otra vez coincidentes con aquellos terminados en TAA, aunque en menor abundancia en el grupo TAG.

Por último, la Fig. 2.15-C representa las anotaciones GO para la categoría de función molecular. El conjunto de genes con codón de terminación TAA (barras rojas) muestra que hay una leve preponderancia de cinco funciones moleculares, que comparten el mismo p-valor ajustado ($4{,}66\times10^{-2}$). Tres de los cuales estarían relacionados con el proceso de traducción: la actividad de ligasa de aminoacil-ARNt (GO:0004812); la unión no covalente de ADN monocatenario (GO:0003697); y el vinculante de la región 3' UTR del ARNm (GO:0003730). En lo referente al conjunto de genes con codón de terminación TGA (barras azules), el único término GO en la categoría de función molecular es el vinculado al enlace no covalente de dominio BTB/POZ (GO:0031208), con p-valor ajustado $2{,}71\times10^{-2}$. Se trata de un dominio estructural de interacción proteína-proteína que se encuentra en muchos factores de transcripción. En segundo lugar, y con igual p-valor ajustado ($3{,}49\times10^{-2}$), las anotaciones GO más sobresalientes corresponden a la unión no covalente de cadena ligera de un complejo de miosina (GO:0032027), y la actividad quinasa de nucleósido monofosfato (GO:0019201). Un tercer grupo de tres anotaciones funcionales destacadas y con igual p-valor ajustado ($3{,}74\times10^{-2}$) incluye a: 1) la unión de proteínas de la familia Rho-GTPasas (GO:0017048) implicadas en la señalización; 2) la unión de proteasoma (GO:0070628), un gran complejo proteico de múltiples subunidades que cataliza la degradación de proteínas; y 3) la actividad fosfotransferasa en la catálisis de la transferencia de un grupo que contiene fósforo (en un compuesto donante) a un grupo fosfato como aceptor (GO:0016776). Por último, el conjunto de genes con codón de terminación TAG (barras verdes) no refleja funciones moleculares sobrerrepresentadas. Los términos GO en este conjunto comparten el mismo p-valor ajustado ($1{,}82\times10^{-1}$), el cual no resulta significativo para este análisis. En conclusión, en este caso no observamos funciones moleculares que estén sobrerrepresentados en forma significativa en ninguno de los tres codones de parada.

A partir de los resultados obtenidos del análisis de ANEN para las tres categorías ontológicas asociadas a cada tipo de codón de terminación, se observa que los principales términos GO de genes que presentan LCP vinculados al codón TAA (conjunto rojo) corresponden al proceso biológico traduccional, mientras que este aspecto no estuvo ampliamente representado en los codones TGA y TAG.

Para visualizar mejor esta diferencia, se utilizaron los resultados obtenidos con el software *FlyEnrichr* relativos a la categoría GO de procesos biológicos, solo de los genes asociados con codones de terminación TAA y TGA en la plataforma *REVIGO* (revigo.irb.hr/) (Supek et al., 2011). Esta plataforma toma las listas de términos GO y elimina los términos redundantes. Los términos GO restantes se

pueden visualizar en diagramas de dispersión basados en similitudes semánticas. En la Fig. 2.16 podemos observar el diagrama de dispersión obtenido con *REVIGO* para los términos GO vinculados a procesos biológicos asociados a los transcriptos con LCP y con codón de parada TAA (Fig. 2.16-A, discos rojos), y aquellos con codón de parada TGA (Fig. 2.16-B, discos azules). Cada círculo representa un cluster de términos GO, después de la reducción de redundancia, ubicado en un espacio bidimensional de similitudes semánticas de los términos GO (Supek et al., 2011). El tamaño de cada disco indica la frecuencia del término GO en la base de datos GO. Observando los diagramas de dispersión en la Fig. 2.16, podemos ver que los principales clústeres asociados a los codones de parada TAA y TGA ocupan lugares diferentes en el espacio semántico, lo que indica que los procesos biológicos de los transcriptos con LCP identificados en este estudio dependen fuertemente del codón de parada. Los clústeres más destacados tienen asociado el número GO y se listan en Fig. 2.16-C, donde el color identifica el tipo de codón de parada. Así, el proceso más evidenciado en el conjunto TAA es el metabolismo de proteínas celulares (GO:0044267, frecuencia: 18,54 %). En segundo término se ubica la biosíntesis de macromoléculas celulares (GO:0034645, frecuencia: 16,24 %); y en tercer lugar, el desarrollo del tejido nervioso (GO:0007399, frecuencia: 14,72 %). En cuanto a la traducción citoplasmática en el conjunto TAA, el término correspondiente (GO:0002181) posee una frecuencia de 9,84 %. Este cluster comprende 29 genes codificantes de proteínas ribosomales. Por su parte, el conjunto TGA no refleja términos GO vinculados al proceso traduccional. El proceso biológico más frecuente es la regulación del metabolismo proteico (GO:0051246, frecuencia: 5,97 %). El segundo término de mayor frecuencia corresponde al transporte mediado por vesículas unidas a membranas aceptoras (GO:0016192, frecuencia: 5,66 %). En tercer lugar, se encuentra el procesamiento de ARN (GO:0006396, frecuencia: 4,91 %), involucrado en la conversión de uno o más transcritos de ARN primarios en moléculas de ARN maduras.

En síntesis, el análisis de enriquecimiento de los términos GO permite conocer si un rol biológico dado esta sobrerrepresentado en un grupo de genes. Así, mediante el análisis de ANEN aplicado al conjunto de transcriptos seleccionados con LCP, se obtuvo información sobre las categorías ontológicas asociadas a cada grupo de genes, según el tipo de codón de terminación traduccional. Se observó que las anotaciones GO más frecuentes en la categoría de procesos biológicos se vinculan por un lado a la traducción, principalmente en el conjunto de genes con codón TAA, y en menor medida en el de codón TAG. En sintonía con esta observación los principales términos GO en la categoría de componentes celulares corresponden a ribosomas en el caso del conjunto TAA, así como a ribosomas y polisomas en el conjunto TAG; mientras que para el conjunto TGA predominan el núcleo y complejos de factores de transcripción. Por otro lado, no se observaron similitudes significativas entre los grupos de genes evaluados, por lo que se puede inferir que los productos derivados de transcriptos con LCP están asociados a diferentes tipos de procesos biológicos dependiendo del tipo de codón de parada que se recodifica.

2.6. Conclusiones

La evaluación de las extensiones mediante perfiles ribosomales permitió caracterizar 1176 transcriptos con una significativa densidad de ribosomas, atribuible a

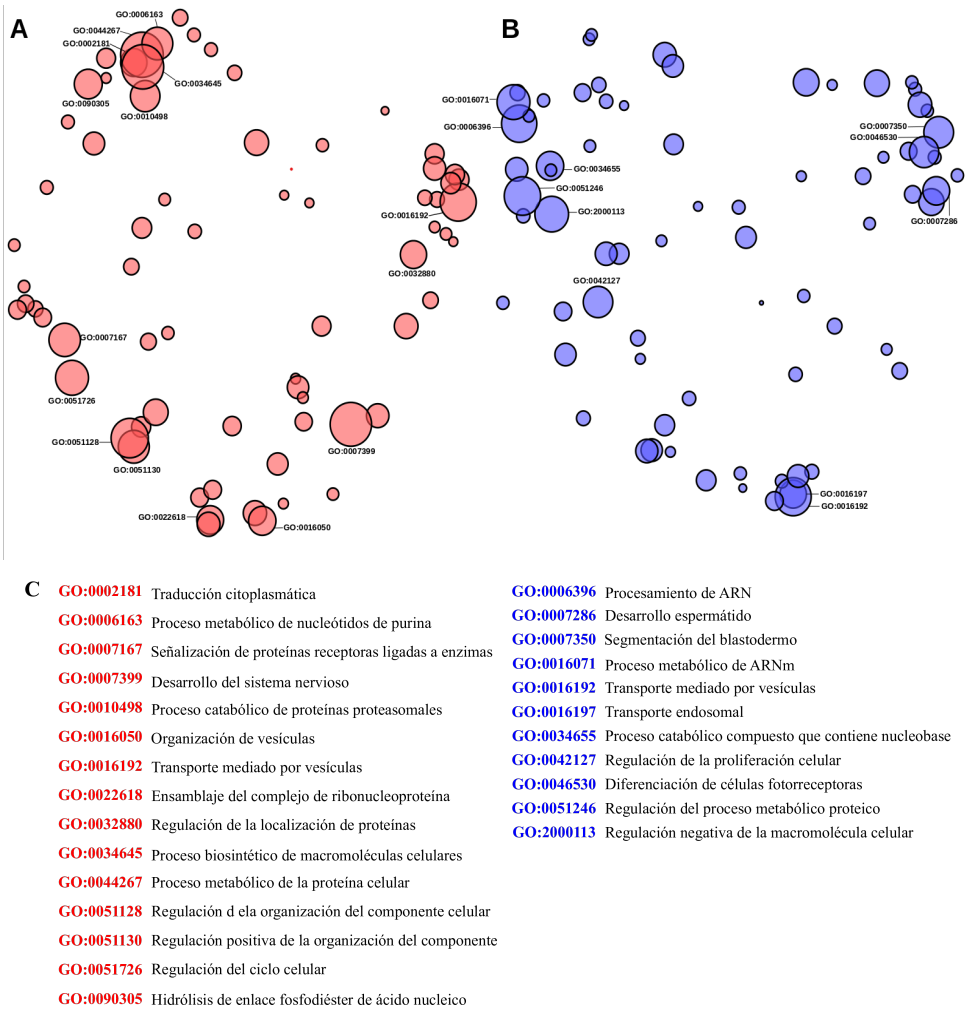

Figura 2.16: Diagrama de dispersión adaptado del resultado de *REVIGO*, que visualiza en forma de clúster los principales procesos biológicos asociados a los genes con LCP encuadrados en un espacio semántico (X,Y). Los términos GO están agrupados según el tipo de codón de parada anotado: TAA en rojo (panel A) y TGA en azul (panel B). El tamaño de cada cluster es directamente proporcional a la frecuencia del término GO relacionado. El panel C lista los números GO de los clústeres más frecuentes para cada conjunto.

probables eventos de LCP. Esta cantidad representa un 17,45 % del conjunto de ARNm analizados (6739) en embriones tempranos de *D. melanogaster*; habiéndose identificado muchos más eventos de LCP que los predichos en análisis previos (Jungreis et al., 2011; Dunn et al., 2013). En este trabajo hemos evidenciado incluso la existencia de nuevos eventos de LCP tanto simples, como dobles y triples.

Se observó que los picos de densidad ribosomal en los codones de parada de la extensión son semejantes a los picos de densidad en los codones de parada anotados, lo que indica una menor velocidad de lectura por parte de los ribosomas, típicamente más concentrados en el inicio y la terminación de la traducción (Ingolia et al., 2011). La densidad ribosomal observada en las extensiones demuestra que si bien son traducidos, lo hacen con baja frecuencia en comparación a los traducidos canónicamente, dando lugar a la baja abundancia de las proteínas extendidas. Así, la síntesis de proteínas extendidas contribuiría a ampliar la diversidad del proteoma, ya que las isoformas de proteínas que albergan diferentes extensiones o truncamientos C-terminales pueden cumplir funciones diferentes respecto de las conocidas para las proteínas canónicas.

Se ha demostrado que los estudios de conservación sinténica, entre otros, constituyen un filtro importante para la validación de nuevos genes (Jungreis et al., 2011); dado que la ocurrencia de sintenia garantiza que la similitud de secuencias es debida a la homología, y corrobora la conservación en función de la funcionalidad. En este sentido, la comparación sinténica mostró que la secuencia proteica extendida está conservada con similitud significativa en otras especies del género *Drosophila*, en concordancia con el criterio utilizado por Jungreis *et al.,* (2011). En el caso del transcripto FBtr0079297 (gen *CG11070*), además de la homología observada entre varias especies, se observa que en el gen homólogo *GG10427* en *D. erecta* la extensión peptídica está incorporada a la región codificante, el cual difiere en la presencia de una lisina (K) como se observa en la Fig. 2.10. Por lo tanto, debido a su conservación a nivel de secuencia, pensamos que el rol funcional de estas extensiones también está conservado en esta familia de moscas, lo que sugiere que se encuentran involucradas en procesos fisiológicos.

Por otro lado, la espectrometría de masas proporciona evidencia experimental acerca de la identidad de los productos proteicos. Con el fin de confirmar los eventos de LCP observados por el método de perfiles ribosomales, los espectros de embriones de *Drosophila* aquí utilizados (Brunner et al., 2007) fueron contrastados con las extensiones traducidas a proteínas de los transcriptos con LCP, mediante los softwares *SearchGui* y *Peptide-Shaker*. Se identificaron tres espectros de alta confidencia asociados a fragmentos peptídicos (LSGTQSEQQASHL, IEVTKVETDIHREI, YSDGSPGTGIGNGHG-GGSGSGANR) que coinciden con las extensiones de los transcriptos FBtr0076667 (gen de la enzima isocitrato deshidrogenasa, *Idh*), FBtr0078997 (gen *CG3792*, involucrado en la síntesis de oligosacáridos) y FBtr0074916 (gen *Su(z)12*, implicado en la actividad metiltransferasa de histonas), respectivamente. Si bien estos tres péptidos identificados son muy pocos frente a los 1176 posibles nuevos eventos de LCP, la diferencia en la cantidad se debe a la pequeña concentración relativa en las muestras de proteínas con extensiones, respecto a la cantidad de proteína anotada. Estos resultados representan una evidencia extra de las extensiones encontradas en los perfiles ribosomales, ya que señalan la expresión de péptidos traducidos a partir de eventos de LCP.

Capítulo 3

Análisis de la secuencia contexto en eventos de LCP

3.1. Factores determinantes en la lectura del codón de parada

En el capítulo previo se analizó la presencia de eventos de LCP en transcriptos de ARNm de embriones de *D. melanogaster* que presentan una extensión C-terminal de su secuencia codificante. Para ello se utilizaron datos transcriptómicos *(Ribo-Seq)* (Ingolia et al., 2009), logrando identificar 1176 nuevos casos putativos de LCP. Las numerosas evidencias de expresión de péptidos derivados de LCP indican que este fenómeno ocurre comúnmente en los genomas (Jungreis et al., 2011; Firth y Brierley, 2012; Dunn et al., 2013; Loughran et al., 2014). Sin embargo, la supresión o no reconocimiento del codón de terminación traduccional podría constituir un error de decodificación, generando diferentes isoformas de proteínas y contribuyendo a la inexactitud del proteoma celular (Bidou et al., 2010; Li y Zhang, 2019; Palma y Lejeune, 2021). La supresión de codones se produce de forma natural con baja frecuencia: $10{\times}10^{-4}$ en células de mamíferos (Manuvakhova et al., 2000) o 0,3 % en levaduras (Namy et al., 2001); aunque esta eficiencia puede estar influenciada por diversas variables.

La maquinaria de decodificación traduccional de los codones de ARNm está también regulada por el sesgo de uso de codones (Quax et al., 2015; Diambra y Diambra LA., 2017; Hanson y Coller, 2018), así como por los ARNt y sus diversas modificaciones (Cantara et al., 2011; Chan et al., 2012). Se ha establecido que las modificaciones postranscripcionales de los ARNt favorecen la interacción codón-anticodón en el ribosoma, mientras que su ausencia puede ocasionar menor fidelidad y un mayor reconocimiento de codones por ARNt cognatos (Grosjean y Westhof, 2016; Blanchet et al., 2018).

Por otro lado, si bien el complejo de terminación de la traducción posee mayor estabilidad energética en el reconocimiento del codón de terminación, en determinadas condiciones o con una tasa muy baja, los ARNt cognatos pueden competir por esa posición reconociendo dos de las tres bases del codón, dando lugar a un evento de LCP (Rajon y Masel, 2011; Floquet et al., 2012). Por ejemplo, se sabe que en células humanas los codones TAG y TAA comparten los mismos ARNt cognatos,

diferentes de los del codón TGA. Así, durante la LCP se incorporan aminoácidos específicos en cada caso (glutamina-tirosina-lisina y arginina-cisteína-triptofano, respectivamente) (Roy et al., 2015). También se ha reportado la incorporación de otros aminoácidos durante la LCP, en base a una mutación sin sentido y su contexto de nucleótidos (Xue et al., 2017). En este sentido, se cree que ciertas modificaciones en el entorno nucleotídico alrededor del codón de parada pueden influir en el proceso de terminación de la traducción (Meyer et al., 2012; Li et al., 2014); pero aún no se comprende el mecanismo exacto de la recodificación de los codones de parada por los ARNt cognatos, y cómo esto afecta la funcionalidad de la proteína sintetizada.

En suma, según la presencia de elementos reguladores o moléculas promotoras de los mismos, se reconocen tres tipos diferentes de LCP. El primer tipo son los eventos de LCP no programados, que ocurren a niveles basales en ausencia de dichos factores, tanto en codones de parada fisiológicos como en codones de parada prematuros (CPP), y podrían considerarse un error de traducción (Rajon y Masel, 2011; Floquet et al., 2012). En cambio, la presencia de elementos reguladores en algunos ARNm específicos vuelve admisible la LCP programada, diferenciándose así el segundo tipo. Por último, el tercer tipo de LCP es el inducido por moléculas como aminoglucósidos, que promueven la lectura de CPP favoreciendo el reclutamiento de ARNt cognatos. Los CPP son más sensibles a la LCP que los codones de parada normales, ya que surgen por mutación en un entorno de nucleótidos seleccionado para promover la traducción, y no su terminación (Dabrowski et al., 2018; Cridge et al., 2018; Palma y Lejeune, 2021).

Si bien los niveles basales de LCP varían según la identidad del codón de parada (siendo mayor en TGA y menor en TAA), este mecanismo puede ser influenciado por los elementos que actúan en *cis* o *trans* en el contexto de nucleótidos cercano al codón de parada (Howard et al., 2000; Manuvakhova et al., 2000; Bidou et al., 2004; Floquet et al., 2012; Baranov et al., 2015). Se ha demostrado que algunos nucleótidos inmediatamente posteriores al codón de parada modulan la eficiencia de la LCP programada, según el enriquecimiento de bases A/T y G/C, respectivamente (Wangen y Green, 2020). Otros estudios indican que determinadas secuencias de consenso tanto cadena arriba como abajo del codón de parada pueden promover la LCP (Beier, 2001; Namy et al., 2001; Harrell, 2002; Xue et al., 2014). Por ejemplo, la secuencia de nucleótidos entre las posiciones -6 a +9 influye en la tasa de LCP, que es activada particularmente por las posiciones -1 y +4; mientras que la secuencia consenso T-stop-C induce LCP eficiente (Floquet et al., 2012). Por otra parte, algunos elementos *trans*, por ejemplo, tanto la ribonucleoproteína A2/B1 como el miARN Let7a interactúan con nucleótidos ubicados unos 10 pb cadena abajo del codón de parada para inducir la LCP programada del ARNm de los genes VEGFA y Ago1, respectivamente (Eswarappa et al., 2014; Singh et al., 2019).

Aunque la caracterización de los eventos LCP requiere muchas veces de estudios moleculares que permitan dilucidar su mecanismo de acción, existen también esfuerzos por enfoques complementarios de biología de sistemas para investigar las condiciones en las que se produce este fenómeno de recodificación del codón de parada para expresar dominios C-terminales de proteínas de baja abundancia. En este sentido se puede mencionar los modelos de regresión *in silico* (Schueren et al., 2014; Loughran et al., 2014; Stiebler et al., 2014), que permiten no solo la predicción de eventos de LCP sino también la influencia de las secuencia contexto al codón de terminación. Se basa en la regresión lineal entre los valores de lectura experimental y sus respectivas secuencias representadas en un espacio vectorial multidimension-

al. Los coeficientes de regresión resultantes describen la influencia ejercida sobre la LCP por todos los nucleótidos en todas las posiciones del contexto al codón de terminación (Schueren et al., 2014). Este modelo fue aplicado en 57 genes humanos candidatos a presentar eventos de LCP, seis de los cuales fueron probados experimentalmente (Schueren et al., 2014; Loughran et al., 2014; Stiebler et al., 2014).

En este sentido, el objetivo es identificar los elementos *cis* que inducen la LCP y determinan la expresión de un compuesto extendido. Para eso indagamos la relación entre la frecuencia con que el ribosoma suprime el codón de parada, o tasa de fuga ribosomal y los nucleótidos lindantes al codón de parada, con dos herramientas. En primer lugar, un análisis estadístico de las frecuencias de uso de nucleótidos en la secuencia contexto. Posteriormente, nos propusimos desarrollar un modelo predictivo para explicar la influencia de nucleótidos observada en secuencias contexto de diferente tamaño, y la tasa de fuga ribosomal.

3.2. Estimación de la tasa de fuga

En esta sección introducimos una herramienta para la estimación de la tasa de fuga ribosomal a partir de los perfiles ribosomales elaborados previamente. Para eso, tendremos en cuenta la densidad ribosomal acumulada en la región de 30 nt alrededor del codón de parada, denotada por δ_{CD}, y la densidad ribosomal acumulada en la región de 30 nt asociada a la extensión C-terminal δ_{ext}. Estas regiones están indicadas en celeste en el panel B de la Fig. 3.1. La motivación para elegir la primera región consiste en que los ribosomas protegen fragmentos de 28 a 30 nt de longitud y que el codón correspondiente al sitio-A del ribosoma se encuentra en la posición 14-16 (Fig. 3.1-C). Por lo tanto, consideramos la región que abarca los 14 nt precedentes al codón de parada y los 16-nt posteriores a él (Fig. 3.1-B) como referente de los ribosomas en la posición del codón de parada. Para calcular δ_{ext}, consideramos la región que abarca los nucleótidos 29 a 58 posteriores al codón de parada. De esta manera, se reducen las posibilidades de contar huellas ribosomales del final de la región codificante como parte de la extensión. Definimos la tasa de fuga ribosomal, denotada por la letra y, como el cociente de la densidad ribosomal acumulada en la región de 30 nt asociada a la extensión C-terminal δ_{ext}, y la densidad ribosomal en la región de codificación δ_{CD}; tal como se resume en la siguiente expresión:

$$y = \frac{\delta_{ext}}{\delta_{CD}}. \tag{3.1}$$

La densidad ribosomal en la extensión puede ser afectada por codones de parada subsiguientes, dando lugar a una densidad mayor (se puede ver en Fig. 3.1-B). En estos casos es difícil asegurar que la densidad ribosomal observada se deba a un mayor número de ribosomas "fugados", o si es debido al hecho de que estos tienen un tránsito más lento en la zona a causa del segundo codón de parada. Por este motivo, la estimación de la tasa ribosomal por este método podría verse afectada. Por eso, para el cálculo de la tasa de fuga ribosomal adoptamos un criterio adicional, seleccionando aquellos transcriptos donde la distancia entre el codón de parada anotado y el próximo codón de parada es igual o mayor a 30 nucleótidos, aún cuando la extensión con densidad ribosomal sea mayor a los 30 nt. Además de

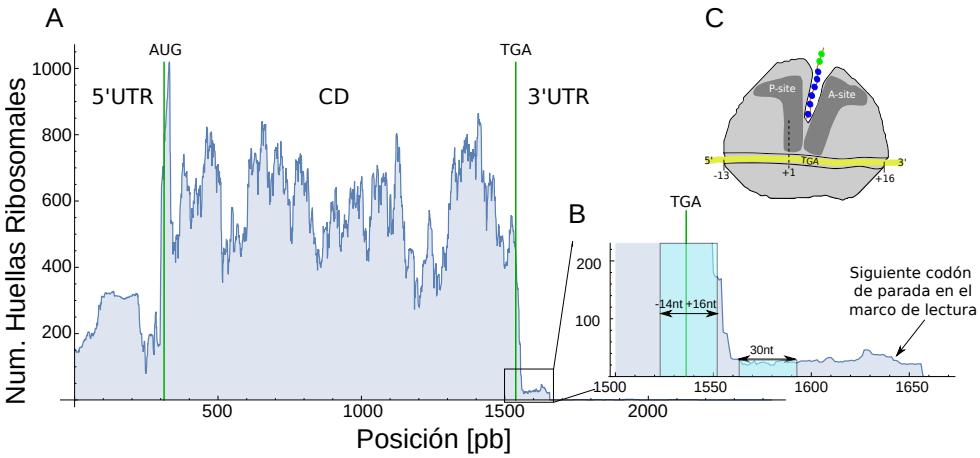

Figura 3.1: Modelo de construcción de perfiles de densidad de ribosomas a partir de la tasa de LCP asociada a cada codón de parada anotado. El panel A representa un perfil de densidad ribosomal típico, distinguiendo los extremos 5' y 3' UTR de la región codificante (CD) delimitada por los codones de inicio ATG y de terminación TGA, señalados por líneas verticales verdes. El panel B muestra una ampliación de la región del codón de terminación y la extensión C-terminal contigua a éste; indicando en color celeste la densidad ribosomal acumulada durante 30 nt asociados a éstas regiones: δ_{CD} y δ_{ext}. El panel C esquematiza la ubicación del sitio-A del ribosoma respecto del codón de terminación TGA.

esta restricción metodológica, impusimos un criterio más estricto para seleccionar un conjunto de transcriptos con eventos de LCP más fidedigno. En este sentido, del conjunto de 1303 transcriptos identificados anteriormente, seleccionamos 1036 casos donde la tasa de fuga ribosomal asociada (y) es mayor que 0.005. De este nuevo conjunto, 241 transcriptos tienen TGA como codón de parada anotado, 453 tienen TAA, mientras que 342 transcriptos están asociados al codón TAG. En la próxima sección haremos un análisis estadístico inicial acerca de los nucleótidos vecinos al codón de parada y la tasa de fuga ribosomal asociada.

3.3. Frecuencia de LCP en los codones de parada

Entre los 6739 transcriptos examinados en esta tesis, el codón TGA está presente en el 23 % de los casos, siendo el menos común de los tres. El codón TAA esta presente en alrededor del 41 % de los transcriptos, siendo el más común de los tres, mientras que el codón TAG esta presente en el 36 % de los casos. Si evaluamos la frecuencia de los codones de parada en el conjunto de transcriptos con LCP, esta resulta en 23, 44 y 33 % para los codones TGA, TAA y TAG, respectivamente. Es decir, la frecuencia de los codones de parada asociada a LCP es similar a la frecuencia de uso de codones de parada en la totalidad de los transcriptos. Este resultado parece estar en aparente contradicción con lo reportado por Jungreis et

al. (2011), donde se reporta que TGA tiene más probabilidades de presentar LCP. Sin embargo, un resultado similar se puede obtener imponiendo un criterio relativo a la tasa de fuga más estricto, como veremos en adelante.

La Fig.3.2 representa los histogramas de frecuencia para cada codón de parada, en función de la tasa de fuga. Podemos ver que para tasas pequeñas ($< 20 \times 10^{-3}$) la mayoría de los eventos se distribuyen de manera casi uniforme, mientras que para tasas de fuga más altas existe un sesgo al codón TGA. Si los eventos de LCP programada están asociados a altas tasas de fuga ribosomal, el resultado de la Fig. 3.2 estaría en concordancia con Jungreis et al. (2011), quien sostiene que la LCP programada utiliza mayormente el codón TGA. También, en forma similar a estudios previos, se analizó la posible incidencia del nucleótido ubicado en las posiciones adyacentes tanto anterior como posterior al codón de parada.

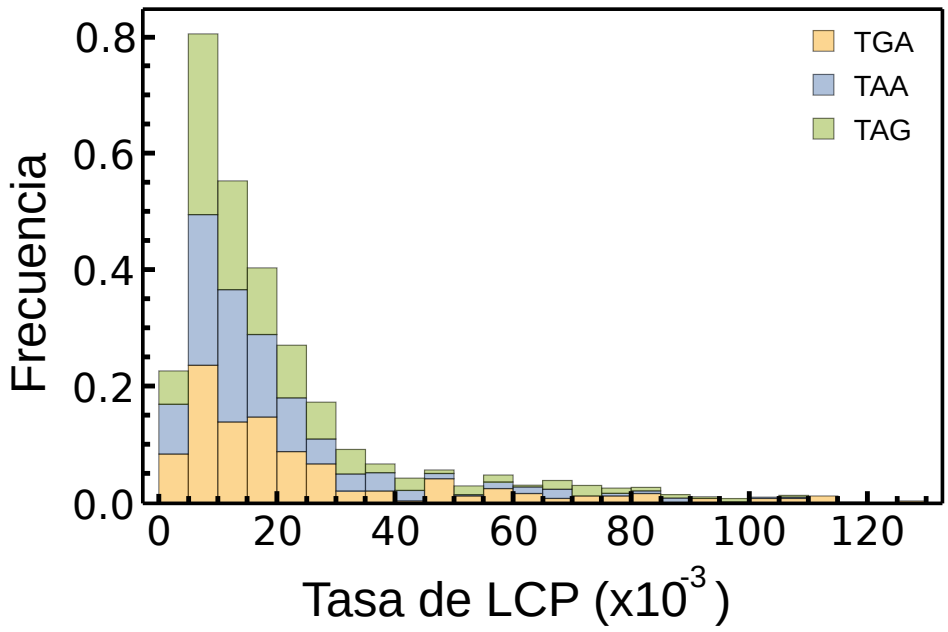

Figura 3.2: Estimación de la tasa de LCP mediante histogramas de frecuencia de la tasa de fuga (10^{-3}) para cada codón de parada TGA, TAA y TAG, representados por los segmentos en colores amarillo, azul y verde, respectivamente.

A propósito de ello, se evaluó la incidencia sobre la tasa de LCP para el nucleótido adyacente al codón de parada, comparando el efecto según el tipo de nucleótido presente (A, T, C y G) y su ubicación anterior o posterior al codón de parada. Así, los paneles de la izquierda en la Fig. 3.3 representan los histogramas para los 12 cuartetos "N" -codón de parada en función de la tasa de fuga, mientras que los paneles de la derecha muestran los histogramas para los 12 cuartetos codón de parada- "N". Para el caso del codón TGA (paneles superiores), observamos que existe una preferencia de nucleótidos C y G tanto en la posición precedente como en la que sucede al codón TGA. Esto coincide con el hecho de que TGA-C es

uno de los contextos de 4-nt utilizados con menor frecuencia en transcriptos con terminación eficiente (Jungreis et al., 2011). También observamos que la fracción de eventos de LCP con mayor tasa de fuga es mayor en el caso del codón TGA que para los otros dos codones de parada. En cuanto al codón TAA (paneles intermedios), observamos que no hay una preferencia marcada de nucleótidos en la posición precedente al codón de parada, observándose una leve inclinación hacia el nucleótido C, seguida por los nucleótidos G y T en proporciones similares. En cambio, se observa una mayor incidencia de los nucleótidos G y A en la posición +4 contigua al codón de parada. Por su parte, el codón TAG (paneles inferiores) refleja una marcada preferencia del nucleótido C ubicado como antecesor, similar al caso de TGA; mientras que el nucleótido sucesor más representado es G. En base a estas diferencias observadas en las frecuencias de uso de nucleótidos en el contexto 4-nt del codón de parada, nuestros resultados no concuerdan con las frecuencias relativas sugeridas por Jungreis et al. (2011), en las que los nt posteriores al codón de parada con más fugas serían C>T>G>A, y por tanto, TGA-C podría ser un contexto de codón de parada permeable a LCP. Mas bien, el presente análisis indica una mayor preferencia de la base C cadena arriba del codón de parada, seguida de las bases G/T (paneles izquierdos); mientras que cadena abajo al codón de parada, la base en mayor proporción es G, seguida de A.

3.4. Análisis de la secuencia de contexto al codón de parada en LCP

El análisis de la secuencia de contexto mostrado en la sección previa tiene razones casi históricas, en el sentido que hasta ahora el contexto de 4-nt es el único tipo de estudio de contexto que aparece en la literatura, y la sección previa tiene como principal sentido la comparación con resultados anteriores. Este tipo de estudio, no considera por ejemplo si el *bies* observado es propio de los transcriptos con LCP o no, dado que no hay una comparación con las frecuencias observadas en transcriptos sin LCP. Además, no se puede descartar la hipótesis de que nucleótidos distales al codón de parada jueguen un papel importante en los eventos de LCP. En este sentido, sólo el trabajo realizado por Schueren et al. (2014) considera en un modelo predictivo de eventos de LCP, una región de contexto de 15 nucleótidos.

En esta sección iremos un paso más adelante en el análisis del contexto al codón de parada. En este sentido, calculamos la frecuencia de uso de los nucleótidos en una amplia región de contexto del codón de parada en transcriptos que presentan eventos de LCP, y las comparamos con las frecuencias correspondientes a las mismas posiciones pero calculada sobre un gran conjunto de transcriptos sin eventos de LCP. Este último conjunto será nuestro control. La comparación entre las frecuencias obtenidas sobre los distintos conjuntos se realizó a través de una medida de divergencia conocida como divergencia de Kullback-Leibler:

$$D(r) = \sum_i p_i(r) \log\left(\frac{p_i(r)}{p_i^*(r)}\right),$$

Esta medida cuantifica las diferencias entre la frecuencia de uso de cada nucleótido i en la posición r, calculada en el conjunto de transcriptos con LCP, denotado por $p_i(r)$, y la frecuencia de uso correspondiente calculada sobre los transcriptos que no

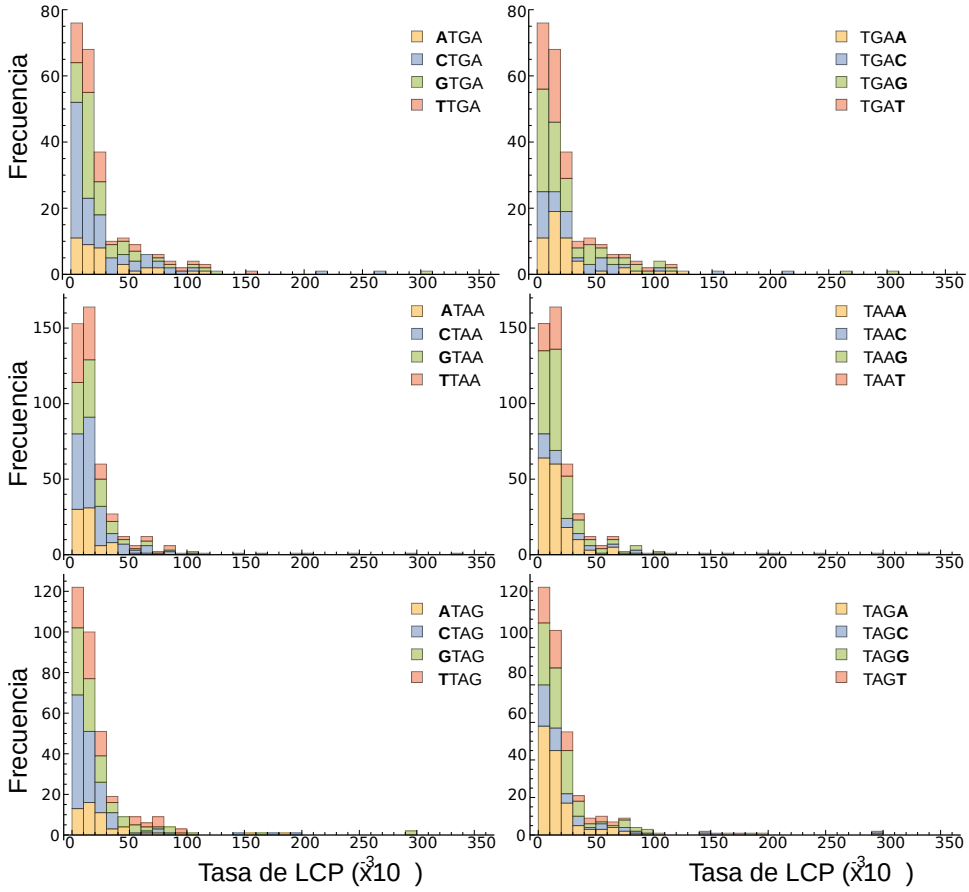

Figura 3.3: Histogramas de frecuencia de la tasa de fuga en función del nucleótido adyacente anterior y posterior al codón de parada. Los paneles superiores muestran histogramas asociados al codón de parada TGA, los paneles medios muestran histogramas asociados al codón de parada TAA, y los inferiores muestran histogramas asociados al codón de parada TAG. Los nucleótidos adyacentes A, C, G y T están identificados por colores amarillo, verde, azul y rojo, respectivamente.

presentan LCP, denotados por $p_i(r)^*$. En este contexto, valores bajos de divergencia denotan una frecuencia de uso similar en ambos grupos de transcriptos, mientras que un valor alto indica un sesgo en el uso de nucleótidos vinculado a eventos de LCP. Como región de contexto consideramos una secuencia de 70 nucleótidos de longitud, 49-nt antes del codón de parada y 18-nt después de él. Realizamos este estudio para cada uno de los tres codones de parada por separado, asumiendo que los mecanismos de LCP programada para cada uno de ellos podrían ser diferentes. Para una mejor ponderación del efecto de la tasa de fuga en el análisis, dividimos los transcriptos que presentan LCP en dos conjuntos:

- transcriptos con tasa de fuga superior a $3{\times}10^{-4}$ (denotados por TS1), y

- transcriptos con tasa de fuga superior a 20×10^{-4} (denotados por TS2).

A fin de identificar los elementos que inducen la LCP, se utilizaron los valores de divergencia calculados para los nucleótidos en cada posición dentro de la región de contexto. Observamos que en algunas posiciones existe un aumento en la divergencia cuando se calcula sobre una conjunto de transcriptos con mayor tasa de fuga. Este aumento corresponde a una mayor sesgo en la frecuencia de uso de un nucleótido en esa posición. En este sentido, esperamos que las posiciones en que la divergencia es alta sean más probables de ejercer influencia en la LCP.

En la Fig. 3.4 podemos ver los valores de divergencia resultantes para cada posición del contexto de nucleótidos analizado. Los puntos azules corresponden a los valores de divergencia calculados utilizando transcriptos pertenecientes a TS1, mientras que los puntos amarillos corresponden a valores calculados utilizando transcriptos TS2, asociados a elevada tasa de fuga ribosomal. En adelante, todas las posiciones harán referencia a la posición del codón de parada. En el caso del codón TGA (panel superior de la Fig. 3.4), observamos que en la posición +4 (adyacente a TGA) existe una divergencia significativa. Esta se debe mayormente a la diferencia en la frecuencia de uso de los nucleótidos C y T entre el grupo con alta tasa de fuga (TS2) y su respectivo control. Estas frecuencias en los transcriptos asociados a altas tasas de fuga son 23.65 % y 19.35 %, respectivamente; cambiando a 15.34 y 26.52 % en el grupo control. Podemos ver entonces que el uso del nucleótido C aumenta un 35 % en el grupo de transcriptos con alta tasa de fuga, estando en acuerdo con los estudios previos (Jungreis et al., 2011). Este aumento es compensado por una disminución del uso del nucleótido T en este grupo. Sin embargo, otras posiciones corriente abajo presentan altos valores de divergencia (por ejemplo +9, +10, +11, +19 y +20), inclusive algunas de ellas están asociadas a un valor de divergencia mayor a la posición proximal usualmente estudiada (+1). Por ejemplo, la frecuencia de uso de los nucleótidos A y C en el grupo control en la posición +11 es 27.4 % y 23.8 %, respectivamente. Sin embargo, estos nucleótidos presentan una preferencia de uso muy diferente en el grupo TS2, los porcentajes cambian a 14 % y 32.3 %, respectivamente. Es decir, un sesgo mayor al observado en la posición +1. Por otro lado, la posición -1 presenta un valor de divergencia despreciable. Sin embargo, la posición -2 tiene una divergencia levemente mayor a la posición +1, allí la frecuencia de uso del nucleótido A se incrementa de 40.7 % hasta 53.8 %. Existen, además, otras posiciones distales, como en -12, -21, -40 y -47 con elevada divergencia. En particular, en la posición con mayor divergencia (-12) encontramos que la frecuencia de uso de los nucleótidos A, T y G, en el grupo control, es de 30.9 %, 16.7 % y 30 %, respectivamente. La frecuencia de uso de estos nucleótidos en TS2 cambia a 48.4 %, 9.6 % y 20.4 %, respectivamente, mostrando grandes diferencias por presencia u omisión de nucleótidos. También observamos un grupo de posiciones contiguas (entre -18 y -12) donde los valores de divergencia obtenidos de transcriptos pertenecientes a TS2 (puntos amarillos) son significativamente mayores que los obtenidos de transcriptos pertenecientes a TS1 (puntos azules); lo que sugiere alguna influencia en los eventos LCP de estas posiciones en conjunto. Además, nuestro análisis muestra que hay posiciones distantes, como -40 y -47, donde se observa un fuerte sesgo de uso en nucleótidos. En particular, en la posición -47, G y T presentan frecuencias de 29 % y 16.1 %, respectivamente, en TS2 y cambian a 19.3 % y 25.3 %, respectivamente, en el grupo control.

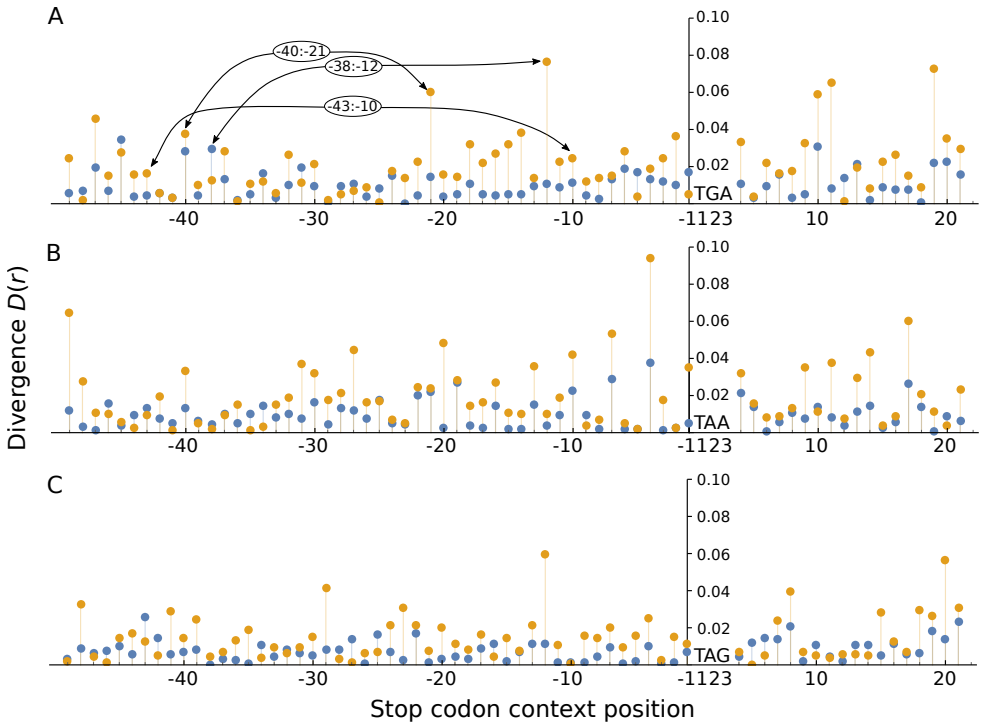

Figura 3.4: Divergencia $D(r)$ de la frecuencia de uso de nucleótidos en la secuencia contexto a cada codón de parada. Los puntos azules y amarillos corresponden a los valores de divergencia calculados para los conjuntos de transcriptos asociados a tasas de fuga ribosomal moderadas (TS1) y elevadas (TS2), respectivamente. Los nucleótidos se representan el eje horizontal como posiciones en la secuencia de contexto al codón de parada.

En el caso de transcriptos con codón de parada TAA (panel central de la Fig. 3.4), la posición -1 tiene valores de divergencia mucho mayores que los observados en el caso de transcriptos con TGA, con frecuencias de 41.7 % (nucleótido C) y 13.4 % (nucleótido A) para el más y el menos utilizados en TS2. En esa misma posición, pero en el grupo control respectivo, estas frecuencias de uso son 32.5 % y 19.8 %, respectivamente. Por otro lado, en el grupo de transcriptos TS2 los nucleótidos más y menos utilizados en la posición +4 fueron G y C, con frecuencias de 41.7 % y 13.4 %, respectivamente. Estas frecuencias en el grupo control cambian a 31.6 % y 11.5 %, respectivamente. A diferencia del codón TGA, la frecuencia de uso de la base C es relativamente baja en el contexto del codón TAA. También observamos un alto valor de divergencia en la posición -4, que está asociada a un fuerte sesgo de uso en el nucleótido C, el cual presenta una frecuencia del 48 % en los transcriptos asociados con una mayor tasa de fuga, contra 28 % observado en el grupo control. Otra característica interesante revelada por el análisis de divergencia de los transcriptos con LCP, utilizando TAA como codón de parada, es que existe un notable *bies* en el uso de nucleótidos en las posiciones -1, -4, -7, -10, -13, -16, -19, -31, -40 y -49, que corresponden a la tercera base de los respectivos codones.

Todas estas posiciones están asociadas a un alto contenido de nucleótidos C y G. De hecho, el contenido de GC en estas posiciones es: 69.3, 71.6, 70.0, 72.4, 72.4, 70.9, 70.1, 74.8, 66.1 y 78.7 %, respectivamente, muy por encima del 50 % esperado en una hipótesis de uso uniforme de codones. Estos valores son consistentemente más altos que el contenido promedio de GC observado en esas posiciones (60.1 %), computado entre los transcriptos del grupo control que terminan en el codón de parada TAA. Este patrón de uso diferencial de nucleótidos contrasta notablemente con el patrón encontrado en el caso de TGA, lo que indicaría que los mecanismos de LCP operando sobre transcriptos con codón TGA podrían no ser los mismos que aquellos asociados a codones TAA. También es notable la preferencia de uso de nucleótidos observada en la posición -49, la posición más distal analizada aquí. Allí, la frecuencia de uso del nucleótido G representa el 44.9 % de los transcriptos en TS2, mientras que el nucleótido A representa solo el 6.3 % en TS2.

En el caso de transcriptos con codón de parada TAG, observamos en general valores de divergencia más bajos que en los casos anteriores, con algunas excepciones en las posiciones -48, -29, -12, +8 y +20 (panel inferior de la Fig. 3.4). Ambas posiciones inmediatamente adyacentes al codón TAG no presentan ningún sesgo importante de nucleótidos con respecto a su respectivo control. Los resultados obtenidos por el análisis de divergencia mostrado arriba sugieren que las posiciones distales tendrían un papel en la LCP tan importante, o más, que los nucleótidos adyacentes estudiados hasta el momento (Jungreis et al., 2011; Dunn et al., 2013; Dabrowski et al., 2015).

Además de correlacionar la composición de nucleótidos con la tasa de LCP en las posiciones de un contexto grande al codón de parada, también analizamos si existe correlación entre los nucleótidos en todos los pares de posiciones dentro del contexto estudiado. Esto corresponde a 2211 pares de posiciones posibles. Para ello, calculamos la divergencia de Kullback-Leibler considerando pares de posiciones según la fórmula de MacKay y Mac Kay (2003):

$$D(r,s) = \sum_{(i,j)pairs} p_{i,j}(r,s) \log\left(\frac{p_{i,j}(r,s)}{p_i(r)p_j(s)}\right),$$

donde $p_{i,j}(r,s)$ es la frecuencia de los nucleótidos i y j en las posiciones r y s respectivamente, mientras que $p_i(r)$ es la frecuencia del nucleótido i en la posición r. Esta medida cuantifica la correlación estadística de uso de nucleótidos en dos posiciones diferentes. En otras palabras, la divergencia entre la probabilidad conjunta p_{ij} de nucleótidos i y j, y la probabilidad esperada asumiendo independencia estadística, es decir, $p_i \times p_j$. Grandes valores de divergencia indican que los nucleótidos aparecen de manera concertada en determinados pares de posiciones y no al azar, es decir, que los patrones estarían asociados con valores elevados de divergencia. Se calculó entonces la divergencia de los 2211 pares para los tres conjuntos de transcriptos TS2 correspondientes a los tres codones de parada en forma independiente. En la Fig. 3.5 se pueden ver los histogramas obtenidos con los valores de divergencia para los tres codones de parada. En cada gráfico están indicados en color amarillo aquellos pares de posiciones en el cuantil 0.985 más alto. A simple vista se puede ver que los valores de divergencia correspondientes a los codones de parada TAA y TAG no superan 0.12, mientras que los correspondientes a TGA presentan varios pares que superan el valor de 0.15.

Para analizar estos pares de posiciones también tuvimos en cuenta el sesgo de las posiciones individuales observado en la Fig. 3.4, es decir, nos concentramos en

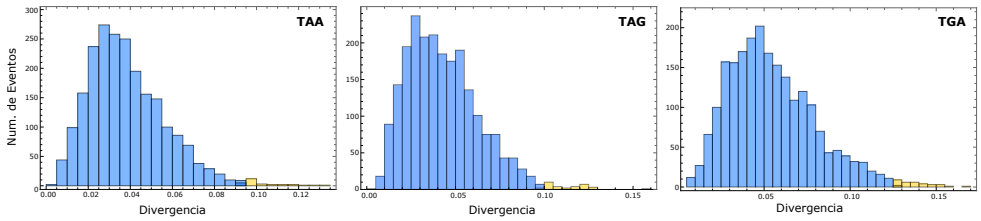

Figura 3.5: Histograma de Divergencia D(r) de nucleótidos en eventos de LCP asociados a los codones de parada TAA, TAG y TGA. Los pares de posiciones asociados a altos valores de divergencia se indican en color amarillo, sugiriendo que los nucleótidos correspondientes representan un patrón de señalización para la ocurrencia de eventos de LCP.

aquellos pares que además de presentar una divergencia alta, las posiciones que los conforman también presentan una elevada divergencia. El panel superior de la Fig. 3.4 muestra algunos pares de posiciones relevantes en el caso del codón de parada TGA, que presentan dos nucleótidos de forma concertada incluso en posición distal, como los pares -40:-21, -38:-12, -43:-10, -16:-15 y -21:-10, para mencionar los más importantes. Por ejemplo, los pares C:G y G:G aparecen en el 32.2 % de los transcriptos en las posiciones -40:-21. Quizás, el caso más interesante son las posiciones adyacentes -16:-15 dentro del grupo de posiciones -18 y -12 identificado en el análisis de divergencia previo. En este caso, la probabilidad de encontrar los pares de nucleótidos C:A o G:G suma 1/3, cuando el valor esperado por casualidad es 1/8. Otros pares de posiciones adyacentes que presentan nucleótidos de forma concertada están en la posición 19:20, corriente abajo del codón de parada. En este lugar, la aparición de los nucleótidos C:A y G:C representa el 30,1 % de los casos, casi el doble del valor esperado, suponiendo la independencia (17,3 %). En el caso del codón de parada TAA, encontramos que los nucleótidos G:C están representados en exceso en las posiciones -21:-4, respectivamente, con una frecuencia del 22 %. No identificamos la dependencia estadística entre los nucleótidos en la tercera posición en diferentes codones, lo que indica en este caso, que el contenido de GC en la tercera posición del codón, y no un patrón específico, estaría asociado a una alta tasa de LCP. Para el caso del codón de parada TAG no observamos una correlación significativa entre pares de posiciones con elevado valor de divergencia. Por completitud, se puede mencionar el par de posiciones -13:-12, el cual evidencia una preferencia por los pares de nucleótidos G:G, C:T y G:C, que representa el 42.6 % de los casos. En las posiciones -9:-8 la probabilidad de encontrar los pares de nucleótidos G:A y A:A suma el 33 %; mientras que en las posiciones -5:-4, la probabilidad de encontrar los pares de nucleótidos A:G y G:C suma un 29,5 %. Sin embargo, en estos casos sólo la posición -12 presenta un rol de interés.

3.4.1. Modelos predictivos

En el estudio de la sección previa se determinó qué posiciones dentro del contexto del codón de parada presentan un sesgo entre los grupos de transcriptos con altas tasas de LCP y aquellos otros que no la presentan. Estas posiciones tendrían un rol importante en la determinación de los eventos de LCP. En este sentido,

sugerimos que una configuración de ciertos nucleótidos en estas posiciones aumenta las probabilidades con que un ribosoma dado interprete la señal de parada de otra manera; es decir, aumentando su tasa de fuga. Una forma de corroborar el rol de las posiciones determinadas es a través de modelos predictivos que, en base a una secuencia contexto, nos determine con una cierta probabilidad la tasa de fuga ribosomal asociada. Esta idea no es nueva en el área de LCP. De hecho, recientemente ha sido aplicada por Schueren et al. (2014), donde los autores proponen un modelo predictivo de eventos de LCP considerando una región de contexto de 15 nucleótidos (6 nt antes y 6 nt después del codón de parada). Sin embargo, en ese trabajo no hay una cuantificación de cuáles nucleótidos/posiciones tienen un rol importante. Tampoco se analiza la existencia de un patrón que oficie como señal para la ocurrencia de LCP. Estos autores consideran los nucleótidos en las 15 posiciones como un todo, y utilizan un modelo de regresión lineal para predecir la tasa de fuga. Para calcular los parámetros del modelo utilizaron un conjunto de sólo 66 transcriptos con LCP, cuyas tasa de fuga asociada tienen valores conocidos (Floquet et al., 2012). Obviamente, con un conjunto de entrenamiento tan pequeño solo se puede aspirar a modelos que consideren secuencias de contexto pequeñas. Dado que nuestro análisis de huellas ribosomales nos permitió identificar y estimar las respectivas tasas de fuga de más de mil secuencias, es posible aspirar a modelos más sofisticados con mayor poder predictivo, como los detallados en esta sección.

En este sentido, elaboramos un modelo para cada codón de parada que además toma en cuenta el contenido nucleotídico en posiciones individualmente relevantes. El criterio de relevancia está determinado por el nivel de divergencia calculado para cada posición. Así, proponemos un modelo aditivo en el cual la tasa de fuga depende del tipo de nucleótido en las posiciones relevantes. Este modelo para la tasa y se puede escribir de la siguiente manera:

$$y = a_0 + \sum_{i}^{N} b_i x_i. \tag{3.2}$$

Como consideramos un contexto de 70 nucleótidos de longitud, 49-nt antes del codón de parada y 18-nt después de él, y dado que el codón de parada esta implícito en cada modelo, se tiene $N = 67$ entradas. En el modelo aditivo descrito en la Ec. 3.2, las variables x_i son numéricas. Sin embargo, los datos disponibles corresponden a cuatro símbolos posibles, los nucleótidos A, C, T y G, en 67 posiciones. Para poder implementar este modelo, debemos entonces representar los nucleótidos en las posiciones individuales, seleccionadas mediante números por medio de una codificación. En nuestro caso, los cuatro nucleótidos pueden ser determinados por solo dos valores de la siguiente manera: (`A` $\rightarrow \{1, 1\}$, `C` $\rightarrow \{-1, 1\}$, `T` $\rightarrow \{1, -1\}$ y `G` $\rightarrow \{-1, -1\}$). Con esta codificación podemos transformar una secuencia contexto en un vector binario de N elementos, según se indica en la Fig. 3.6. Así, a modo de ejemplo, tomemos una secuencia de 5 nucleótidos ATGGT, podría recodificarse por un vector de 10 componentes $\{1, 1, 1, -1, -1, -1, -1, -1, 1, -1\}$. Cada componente de este vector tiene asociado en el modelo un parámetro b_i, representado por el segundo término de la Ec. 3.2 . El primer término a_0 corresponde al intercepto del modelo lineal. Los parámetros del modelo serán determinados a partir de las secuencias contexto asociadas a cada codón de parada en forma independiente; es decir, tendremos un modelo para cada codón de parada. En consecuencia, si el modelo incorpora los 67 nucleótidos de la secuencia contexto, entonces el número de

parámetros a determinar sería de 135. En general, un modelo con muchos parámetros a determinar puede tener un poder predictivo deficiente si no contamos con un conjunto de datos lo suficientemente grande. Por eso, es conveniente para el cálculo de los parámetros tener un número de secuencias contexto mucho mayor al número de parámetros a determinar. Por esta razón, proponemos en el modelo incluir sólo a las posiciones y los pares de posiciones seleccionadas. Nuestra hipótesis de trabajo es que las posiciones que presentan un sesgo en el uso de nucleótidos son las que probablemente afectan al proceso de LCP. Para seleccionar las posiciones y pares de posiciones, usamos las medidas de divergencia evaluadas en la sección previa. En este sentido, seleccionamos como entrada de nuestro modelo las posiciones con divergencia mayor al percentil 66.6 %. Luego, la performance predictiva obtenida de este modelo será comparada con modelos que toman en cuenta los nucleótidos en diferentes posiciones.

Con el fin de simplificar la notación para el procedimiento de estimación de parámetros, notamos que la Eq. 3.2 se puede reescribir como $y = \mathbf{w} \cdot \mathbf{v}$, donde $\mathbf{w}$ es un vector de $N + 1$ dimensiones que incluye todos los coeficientes a determinar (es decir, $\mathbf{w} = (a_0, b_i)$), y $\mathbf{v}$ es el vector binario definido previamente, al cual se le agrega un 1 como primer elemento, asociado al coeficiente a_0. Para estimar los coeficientes del modelo $\mathbf{w}$, disponemos de un conjunto de M secuencias contexto, cuya información se codifica en los vectores binarios de la misma forma, dando lugar a una matriz X, que corresponde a la información de entrada del modelo. Cada una de estas secuencias tiene asociado su respectivo valor de tasa de fuga (y), estimada por la Ec.3.1 a partir de los perfiles ribosómicos detallados en el capítulo previo. Estas tasas de fuga se corresponden con la salida del modelo, es decir, la magnitud que se pretende predecir. Los M pares de entrada-salida definen el conjunto de entrenamiento, es decir, la información necesaria para el proceso de inferencia de los valores de los parámetros, y lo representamos por $D = \{\mathbf{X}, \mathbf{y}\}$. Donde $\mathbf{X}$ es una matriz $M \times N$ extraída de las secuencias, en que los nucleótidos están codificados en 1 y -1, como se muestra en la Fig. 3.6. Las columnas de la matriz $\mathbf{X}$, corresponden a las posiciones utilizadas; mientras que las filas indican cada una de las secuencias contexto. El vector $\mathbf{y}$ corresponde a los M valores de la tasa de fuga asociados a las secuencias contexto.

Para la estimación de los coeficientes del modelo, utilizamos la regresión de mínimos cuadrados basada en la descomposición de valores singulares (SVD) de la matriz X^T (^{T} denota la transpuesta de la matrix), es decir, $X^T = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T$, donde $\mathbf{U}$ es una matriz unitaria $M \times N$ de vectores propios izquierdos, $\mathbf{S}$ es una matriz diagonal $N \times N$ que contiene los valores propios $\{s_1, \ldots, s_N\}$, y $\mathbf{V}$ es una matriz unitaria $N \times N$ vectores propios de la derecha. Por lo tanto, la solución con la norma L_2 más pequeña viene dada por $\mathbf{w} = \mathbf{y} \cdot \mathbf{U} \cdot \mathrm{diag}(s_j^{-1}) \cdot \mathbf{V}^T$, y $\mathbf{w} \cdot \mathbf{v}$ corresponde a la tasa de fuga predicha por el modelo para un vector de características de secuencia $\mathbf{v}$.

3.4.2. Evaluación de los modelos predictivos

Como se mencionó anteriormente, se construyó en forma independiente un modelo predictivo para cada uno de los codones de parada. Además, estudiamos para cada uno de los respectivos codones diferentes modelos que tienen en cuenta diferentes tamaños de la secuencia contexto. En este sentido, en forma similar al trabajo de Schueren et al. (2014), se consideran modelos que incluyen la información de los

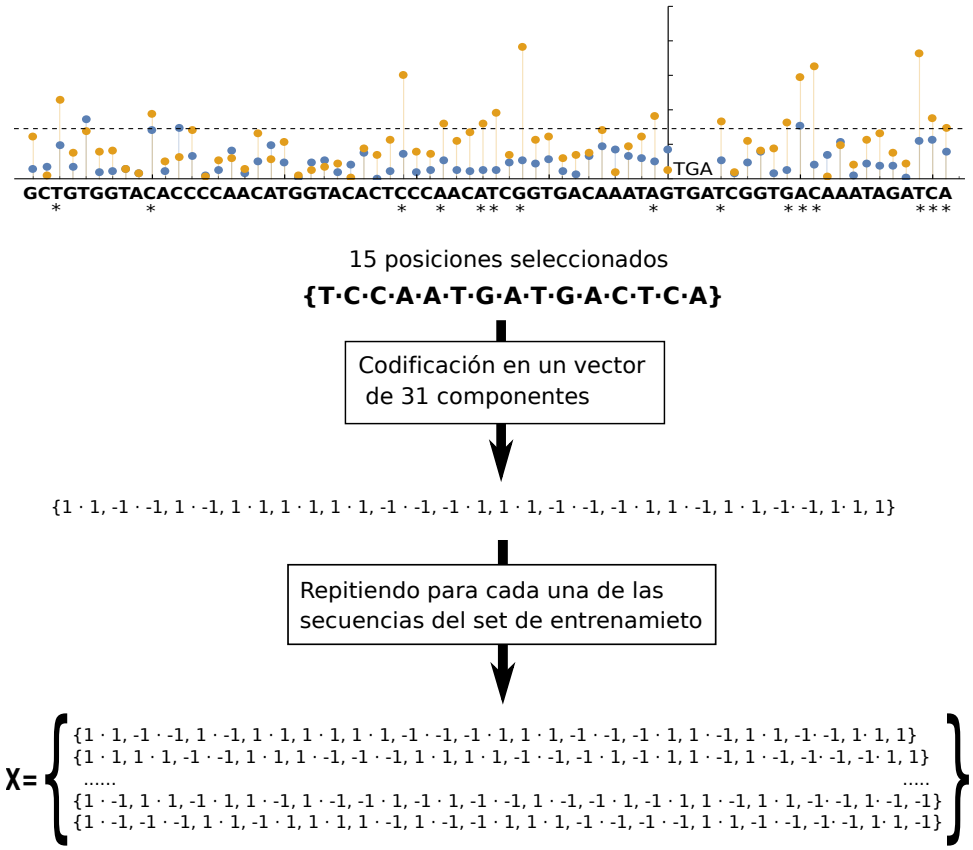

Figura 3.6: En este diagrama ilustrativo se explica cómo las secuencias de contexto se codifican numéricamente. Se seleccionan los nucleótidos en las posiciones indicadas por $*$. Cada nucleótido es codificado por un par de 1 o -1. El procedimiento es realizado para todos las secuencias contexto y se construye la matriz $\mathbf{X}$.

nucleótidos en 12 posiciones contiguas al codón de parada, 6 delante y 6 después de él, indicados por una llave negra. También incluimos modelos que consideran un contexto más amplio, e incluimos la información de los nucleótidos en 29 posiciones contiguas al codón de parada (21 delante y 8 después), indicados por una llave roja. Por otro lado, también consideramos modelos que solo incluyen las posiciones con alto valor de divergencia, sin importar si son o no contiguas al codón de referencia. Estas posiciones son indicadas con flechas negras en el panel superior de las Figs. 3.7, 3.8 y 3.9. Para cada caso, se determinaron los parámetros del modelo por regresión lineal, usando SVD como se indicó previamente. Luego, se evaluó el poder predictivo del modelo usando un conjunto de secuencias de evaluación particular para cada codón de parada. Estas secuencias de evaluación del modelo incluyen transcriptos anotados en la base de datos FlyBase que presentan LCP, y que no fueron utilizados para la determinación de los parámetros de los modelos. Además, el conjunto de evaluación de cada codón de parada incluye 50 secuencias

que no presentan LCP para el mismo codón. Si bien el modelo predice la tasa de fuga asociada con una secuencia contexto dada, la evaluación de la *performance* del mismo fue hecha a través del cálculo de la fracción de falsos positivos y falsos negativos, y no por la diferencia entre la tasa de fuga experimental y la predicha por el modelo. Es decir, falsos positivos son aquellas secuencias que no están asociadas a LCP, aunque el modelo predice una tasa de fuga positiva. Por otro lado, los falsos negativos corresponden a aquellas secuencias que presentan LCP (tasa de fuga experimental mayor que 0), pero que el modelo predice una tasa de fuga negativa. Por supuesto, el objetivo es obtener modelos que minimicen ambos tipos de errores.

En la Fig. 3.7 se pueden observar las fracciones de falsos positivos y falsos negativos obtenidas para cuatro modelos diferentes desarrollados para el codón TGA, que utilizan distintas posiciones en la secuencia de contexto. El primer par de barras corresponde al modelo que sólo toma en cuenta 6 posiciones contiguas a ambos lados del codón, en forma similar al trabajo de Schueren et al. (2014). Este modelo identifica la existencia de eventos de LCP en casi un 80 % de los transcriptos que presentan este fenómeno. Sin embargo, tiene una elevada fracción de falsos negativos. Esta fracción disminuye a la mitad cuando extendemos el modelo a 29 posiciones. Por otro lado, la inclusión de las 29 posiciones empeora la predicción de los transcriptos que presentan LCP. Esto indica que la relación entre el número de secuencias en el conjunto de entrenamiento y el número de parámetros a determinar (número de posiciones usadas) tiene un impacto considerable en la *performance* del modelo. Es decir, que si se mantiene fija la cantidad de secuencias en el conjunto de entrenamiento, aumentar el número de posiciones usadas en el modelo no garantiza un mejor poder predictivo. Teniendo esto en cuenta, es interesante considerar la evaluación de un modelo que incorpore sólo las posiciones más informativas, disminuyendo así la cantidad de parámetros a determinar. En este sentido, el modelo que utiliza sólo las 18 posiciones óptimas, indicadas por flechas negras, tiene una tasa de falsos positivos y de falsos negativos más baja que los otros modelos.

Por otro lado, el cuarto modelo predictivo mostrado en el par de barras a la derecha de la Fig. 3.7, muestra la *performance* del modelo obtenido con secuencias que presentan codón de terminación TAA, evaluado con las secuencias de codón de terminación TGA. Se puede apreciar que este modelo tiene capacidad de predicción aún en un grupo de transcriptos diferente, sugiriendo que existen características comunes subyacentes a los eventos de LCP entre estos dos grupos de transcriptos.

De manera similar, en la Fig. 3.8 se pueden observar las fracciones de falsos positivos y falsos negativos obtenidas para cuatro modelos correspondientes al codón TAA, que utilizan distintas posiciones en la secuencia contexto. El primer par de barras corresponde al modelo inspirado en el trabajo de Schueren et al. (2014), que solo toma en cuenta 6 posiciones contiguas a ambos lados del codón TAA. Este modelo identifica la presencia de eventos de LCP en casi un 80 % de los transcriptos candidatos. Sin embargo, tiene una elevada fracción de falsos negativos. Al igual que en el caso del codón TGA, esta fracción disminuye a la mitad cuando extendemos el modelo a 29 posiciones. Por otro lado, la inclusión de las 29 posiciones incrementa en 0.2 % la predicción de los transcriptos que presentan LCP, aunque la proporción de dicho incremento es menor que el caso anterior. Nuevamente, esto indica que la relación entre el número de secuencias en el conjunto de entrenamiento y el número de parámetros a determinar (número de posiciones usadas) altera la *performance* del modelo, sin garantizar un mayor poder predictivo. En este caso, para disminuir

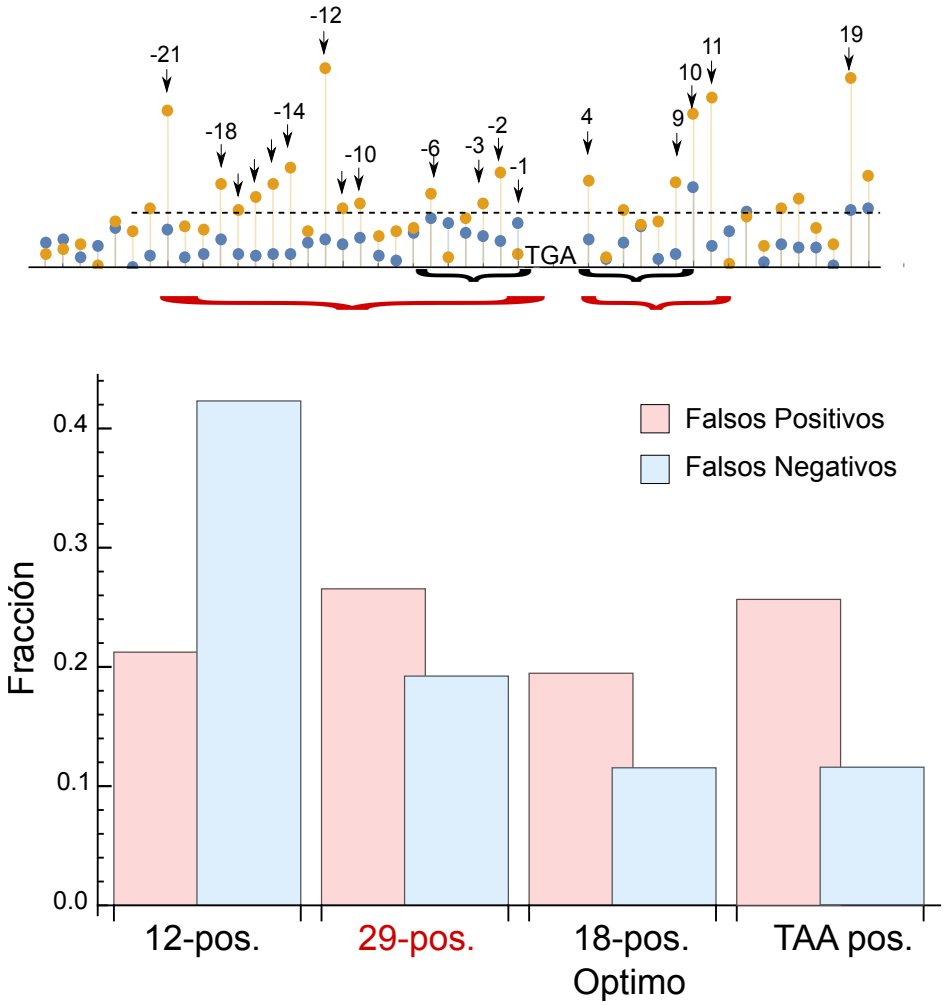

Figura 3.7: El panel superior ilustra las posiciones de nucleótidos en la secuencia contexto, utilizadas por tres modelos lineales diferentes para predecir LCP en transcriptos con codón de parada TGA. Las llaves negras indican las 12 posiciones consideradas en Schueren et al. (2014); las llaves rojas consideran un modelo extendido con 29 posiciones contiguas al codón de parada; las flechas negras consideran solo 18 posiciones seleccionadas por tener asociados los mayores valores de divergencia. En el panel inferior, las barras muestran la fracción de predicciones erróneas: los falsos positivos (en rosa) y los falsos negativos (en celeste), para los tres modelos indicados en el panel superior y un cuarto modelo discutido en el texto principal.

la cantidad de parámetros a determinar, se evaluó el modelo incorporando sólo las 13 posiciones más informativas (óptimas), indicadas por flechas negras. Este modelo tiene una tasa de falsos positivos y de falsos negativos notablemente más baja

que los otros modelos, y ambas tasas son significativamente menores que en el modelo de las 18 posiciones óptimas para el codón TGA. Además, esta tasa de falsos negativos es la menor de todas, comparando las posiciones óptimas evaluadas para cada codón de parada.

El cuarto modelo predictivo mostrado en el par de barras a la derecha de la Fig. 3.8, muestra la *performance* del modelo obtenido con secuencias con codón de terminación en TGA evaluado con las secuencias con codón de terminación en TAA. En comparación con el modelo opuesto (Fig. 3.7), se observa una reducción en la fracción de falsos positivos y un aumento en la fracción de falsos negativos. No obstante, este modelo resulta insuficiente para precisar un nivel de predicción confiable, dado que el número de parámetros a determinar varía según las secuencias correspondientes a cada tipo de codón de parada.

Por último, en la Fig. 3.9 se pueden observar las fracciones de falsos positivos y falsos negativos obtenidas de cuatro modelos para el codón TAG, que utilizan distintas posiciones en la secuencia de contexto. El primer par de barras corresponde al modelo inspirado en el trabajo de Schueren et al. (2014), que solo toma en cuenta 6 posiciones contiguas a ambos lados del codón TAG. Este modelo identifica la presencia de eventos de LCP en casi un 80 % de los transcriptos candidatos. Sin embargo, presenta una elevada fracción de falsos negativos. Esta fracción disminuye en forma considerable cuando extendemos el modelo a 29 posiciones. Por otro lado, la inclusión de las 29 posiciones incrementa en 0.15 % la predicción de los transcriptos que presentan LCP, aunque la proporción de dicho incremento es la menor de los tres casos. La reducción de falsos negativos ligada a un leve incremento de falsos positivos aquí observada, indicaría que la relación entre el número de secuencias en el conjunto de entrenamiento y el número de parámetros a determinar (número de posiciones usadas) alteran favorablemente la *performance* del modelo. Sin embargo, la evaluación del modelo incorporando sólo las posiciones más informativas muestra un mejor poder predictivo que los otros modelos, a la vez que reduce la cantidad de parámetros a determinar. En este caso, el modelo que sólo usa las 25 posiciones óptimas, indicadas por flechas negras, muestra una tasa de falsos positivos y de falsos negativos apreciablemente más baja que los otros modelos. Se observa incluso que la tasa de falsos positivos es la menor en comparación con todos los modelos aplicados; mientras que la tasa de falsos negativos es la mayor entre los tres modelos que usan las posiciones óptimas en el contexto al codón de parada.

El par de barras a la derecha de la Fig. 3.9, muestra la *performance* del modelo obtenido con secuencias con codón de terminación en TAA evaluado con las secuencias con codón de terminación en TAG. En comparación con el mismo modelo aplicado en el codón de terminación TGA (Fig. 3.7), se observa una reducción en la fracción de falsos positivos y un aumento en la fracción de falsos negativos. Nuevamente, esta comparación muestra la capacidad de predicción entre dos grupos de transcriptos con diferente codón de terminación, sugiriendo la existencia de posibles características comunes vinculadas a los eventos de LCP. Sin embargo, una predicción fehaciente mediante este tipo de modelo requeriría el uso de parámetros (posiciones más informativas) comunes entre las secuencias contrastadas.

Estos resultados indican que la tasa con la cual ocurren los eventos de LCP podría estar regulada por un contexto mayor a los propuestos por los estudios previos (Schueren et al., 2014; Jungreis et al., 2011). Además, es importante aclarar la relación entre número de parámetros y número de secuencias. Si bien es claro que cuanto mayor sea el tamaño del conjunto de entrenamiento mejor será la

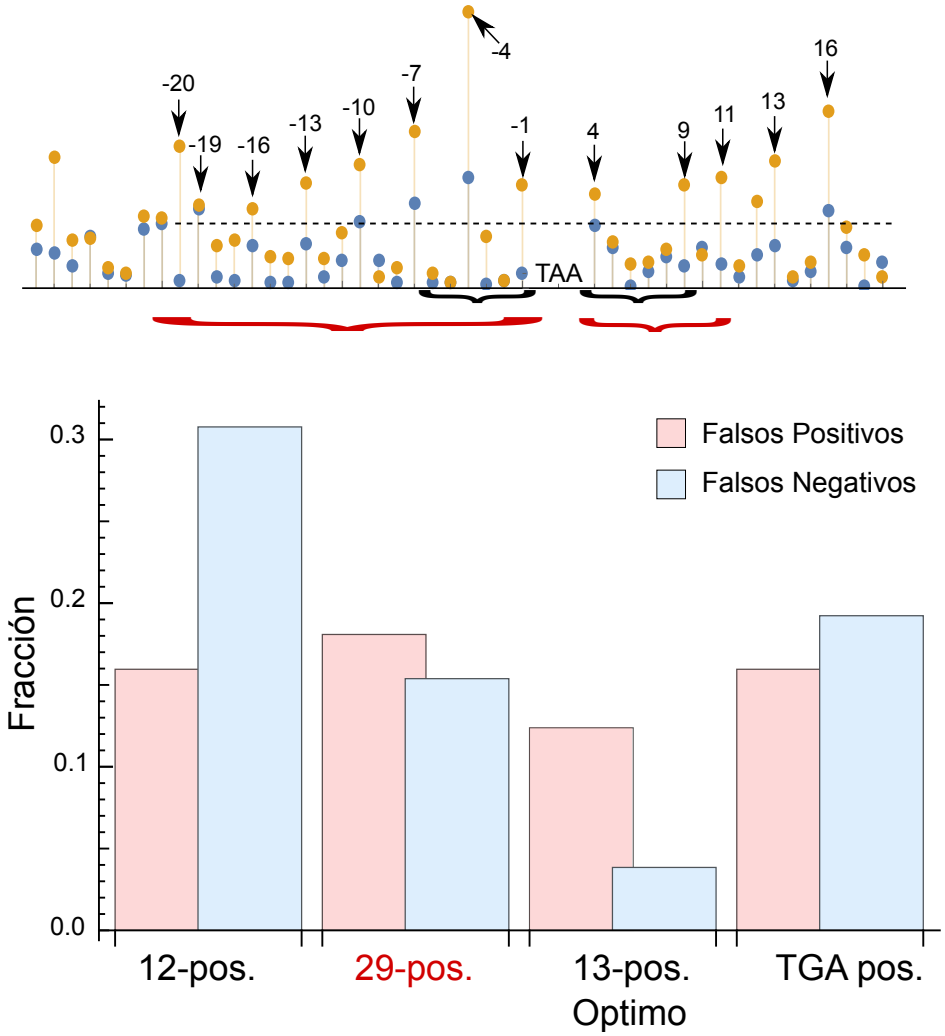

Figura 3.8: El panel superior ilustra las posiciones de nucleótidos en la secuencia contexto, utilizadas por tres modelos lineales diferentes para predecir LCP en transcriptos con codón de parada TAA. Las llaves negras indican las 12 posiciones consideradas en Schueren et al. (2014); las llaves rojas consideran un modelo extendido con 29 posiciones contiguas al codón de parada; las flechas negras consideran 13 posiciones seleccionadas por tener asociados los mayores valores de divergencia. En el panel inferior, las barras muestran la fracción de predicciones erróneas: los falsos positivos (en rosa) y los falsos negativos (en celeste), para los tres modelos indicados en el panel superior y un cuarto modelo discutido en el texto principal.

parametrización del modelo, no es así con relación al número de parámetros. Finalmente, mostramos que la divergencia puede ser utilizada como un criterio para

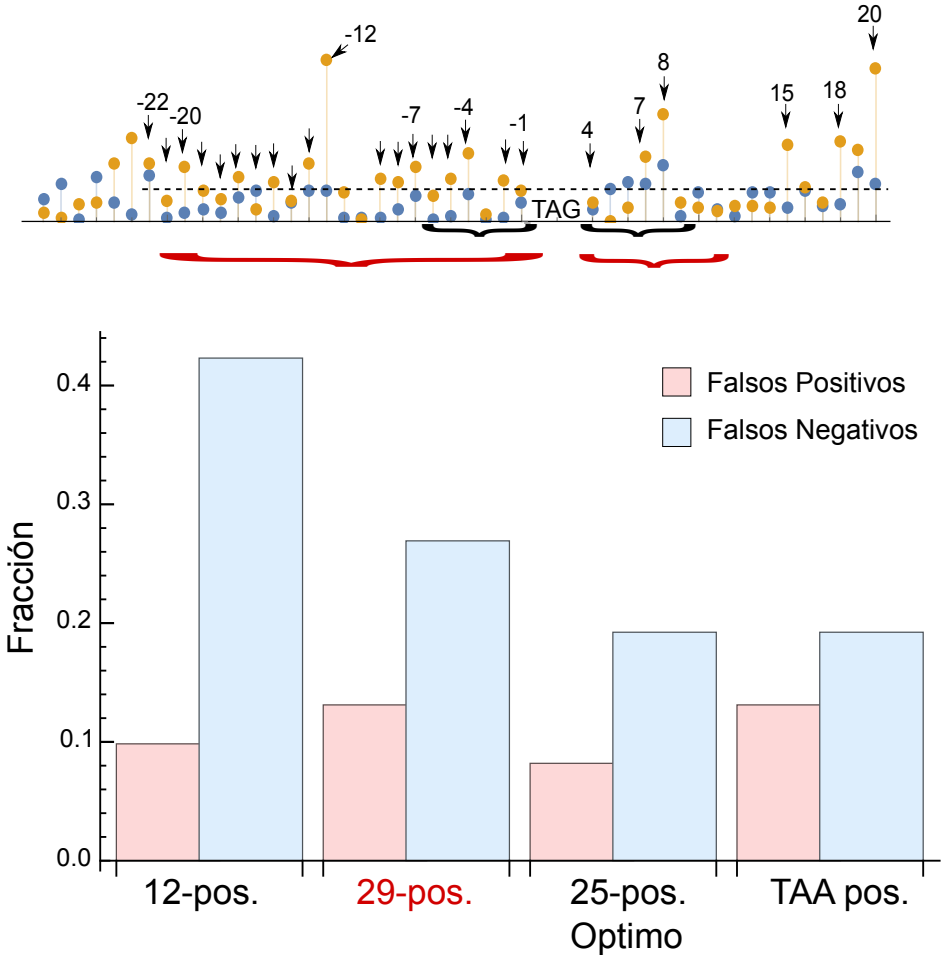

Figura 3.9: El panel superior ilustra las posiciones de nucleótidos en la secuencia contexto, utilizadas por tres modelos lineales diferentes para predecir LCP en transcriptos con codón de parada TAG. Las llaves negras indican las 12 posiciones consideradas en Schueren et al. (2014); las llaves rojas consideran un modelo extendido con 29 posiciones contiguas al codón de parada; las flechas negras consideran 25 posiciones seleccionadas por tener asociados los mayores valores de divergencia. En el panel inferior, las barras muestran la fracción de predicciones erróneas: los falsos positivos (en rosa) y los falsos negativos (en celeste), para los tres modelos indicados en el panel superior y un cuarto modelo discutido en el texto principal.

seleccionar las posiciones más informativas en el modelado.

3.4.3. Conclusiones

En este capítulo definimos la tasa de fuga ribosomal (y) de una forma que se puede estimar a partir de los perfiles de huellas ribosomales. En base a esto, se determinó dicha tasa para los 6739 transcriptos observados. El análisis de eventos de LCP en los codones de parada demostró una incidencia de 23, 33 y 44 % para los codones TGA, TAG y TAA, respectivamente. Estos porcentajes son similares a los porcentajes de uso de codones de parada en el total de transcriptos estudiados, indicando que los eventos de LCP no podrían ser atribuidos a errores de decodificación traduccional de un codón de terminación específico, sino que necesariamente serían causados por un mecanismo molecular implícito (Firth et al., 2011; Dunn et al., 2013; Blanchet et al., 2014; Dabrowski et al., 2015). Sin embargo, también observamos que los eventos de LCP asociados con TGA tienen una tasa de fuga más elevada, particularmente con TGA-C. Este resultado confirma hallazgos anteriores (Loughran et al., 2014; Jungreis et al., 2016; Cridge et al., 2018). Este aumento de la representatividad ribosómica, acompañada de una mayor incidencia de un determinado nucleótido inmediatamente posterior al codón de parada, podría indicar eventos programados de LCP. De ser así, esto representaría una mayor producción de péptidos bioactivos de baja abundancia, y no simples errores traduccionales.

En cuanto a la evaluación de la influencia de los nucleótidos adyacentes al codón de parada sobre la LCP, hemos evidenciado que existe un mayor sesgo al codón TGA bajo altas tasas de fuga ribosomal, con abundante presencia de nucleótidos C y G. De manera similar, el codón TAG mostró preferencia por los mismos nucleótidos, aunque con C como antecesor y G como sucesor (C-TAG, TAG-G). EL codón TAA reflejó una leve tendencia a la presencia del nucleótido C, seguida por G y T como nucleótidos antecesores inmediatos; mientras que la posición adyacente posterior se vió ocupada por G y A. Este resultado concuerda sólo parcialmente con las observaciones de Jungreis *et al.* (2011, 2016) sobre el uso preferencial del codón TGA (TGA>TAG>TAA) y las frecuencias relativas de contextos (C>T>G>A). El autor atribuye mayor grado de conservación en TGA como primer codón de parada respecto de TAG y TAA en *Drosophila spp.*, y propone que el nucleótido C en la posición 4-nt incrementa la tasa de fuga ribosomal durante la traducción.

En la presente tesis, se complementa el estudio de la influencia de los nucleótidos adyacentes con un análisis más exhaustivo de la ocurrencia de LCP en los contextos de los codones de parada, a fin de identificar patrones que puedan predecir los eventos de LCP. Esta identificación fue hecha con la medida de divergencia de Kullback-Leibler, que indica la presencia de sesgo en el uso de nucleótidos en cada posición. En este sentido encontramos también elevados valores de divergencia en diversas posiciones distales cadena arriba, muchas de las cuales son propias de cada codón de terminación. Por ejemplo, el codón TGA presenta alta divergencia en las posiciones -2 y -12, con preferencia de uso del nucleótido A (53.8 % y 48.4 % respectivamente) o del nucleótido G (29 %) en la posición -47. Estas posiciones en la secuencia de contexto al codón de parada, con fuerte sesgo de uso de determinado nucleótido, podrían señalar una marcada influencia para la ocurrencia de eventos programados de LCP. Además, nuestro estudio demostró la presencia de alto contenido de nucleótidos C y G (entre 66 y 78.7 %) en la tercera base de los codones de numerosas posiciones; lo que resulta superior a la proporción observada en las mismas posiciones de transcriptos con codón de terminación TAA del grupo control (60 %), y a la frecuencia esperada por uso uniforme de codones (50 %). Este

notable sesgo difiere ampliamente del observado en TGA, indicando que los mecanismos de LCP operarían bajo patrones de uso diferencial de nucleótidos específicos en posiciones distales para cada tipo de codón de terminación, algo que no estaba contemplado en estudios previos (Jungreis et al., 2011; Dunn et al., 2013; Dabrowski et al., 2015). En el caso del codón TAG, la divergencia en el uso de nucleótidos resultó ser generalmente menor que para TAA y TGA, y no presentó sesgos en las posiciones adyacentes.

Por otra parte, la divergencia de Kullback-Leibler calculada entre los nucleótidos en los 2211 pares de posiciones dentro de cada secuencia contexto analizada en transcriptos con alta tasa de LCP, reflejó que existe una correlación entre pares de posiciones con divergencia superior a la media. En el caso del codón de parada TGA, se observó una frecuencia de 32.2 % en posiciones -40:-21 de los pares de nucleótidos C:G y G:G. En el caso del codón de parada TAA, los nucleótidos G y C están sobre representados también en la tercera base de los codones; sin embargo no se obtuvo correlación significativa en la frecuencia de pares entre estas posiciones. Este resultado sugiere que, en el caso de TAA, los eventos de LCP no serían inducidos por un patrón nucleotídico específico y extenso, sino que estarían asociados a un alto contenido de GC en la tercera base de los codones en posiciones con elevada divergencia.

Finalmente, el modelo de regresión *in silico* propuesto en este trabajo permitió desarrollar un método para la predicción de genes con LCP en genomas. En base al nivel de divergencia de los nucleótidos en tres secuencias contexto de diferente tamaño, nuestro modelo es capaz de comparar la influencia de tales nucleótidos sobre la tasa de fuga de cada codón de parada en forma independiente, a fin de corroborar el rol de las posiciones más relevantes identificadas en los eventos de LCP analizados. El modelo de 12 posiciones contiguas a cada codón de parada identifica eventos de LCP en el 80 % de los transcriptos evaluados, pero presenta elevada fracción de falsos negativos en todos los casos. El modelo de 29 posiciones contiguas a cada codón de parada reduce considerablemente la cantidad de falsos negativos obtenida por el modelo anterior, pero aumenta la fracción de falsos positivos. Esto demuestra que el tamaño de las secuencias contexto utilizadas en función de un número fijo de secuencias de entrenamiento, no garantiza un poder predictivo fiable. En cambio, el modelo que incluye únicamente las posiciones con alto valor de divergencia asociadas al contexto de cada codón de parada con LCP (contiguas o no a éste), resultó eficaz al reducir notoriamente la tasa de falsos positivos y negativos respecto de los otros dos modelos. Así, el uso de las posiciones más informativas en cuanto al nivel de divergencia dentro de una secuencia contexto constituye un criterio novedoso y útil para el desarrollo de herramientas computacionales como la que aquí se presenta. En este sentido, y en conjunto con las ventajas combinadas de otros métodos experimentales, nuestro modelo predictivo basado en identificación de sesgo de uso de nucleótidos ha demostrado ser óptimo para intentar facilitar la predicción de eventos de LCP. Consecuentemente, un estudio más amplio y profundo de la relación entre las posiciones de nucleótidos más destacados permitiría corroborar y anotar el conjunto completo de productos génicos sometidos a LCP.

Capítulo 4

Datos y Metodología

4.1. Sobre el organismo de estudio

La identificación de los eventos de traducción no convencionales propuesta en esta tesis requirió de la integración de datos genómicos, transcriptómicos y proteómicos. Por este motivo, se utilizó un organismo del que se cuenta con abundante información disponible y cuyo genoma está anotado en detalle. Estos requisitos se satisfacen ampliamente en el caso de la mosca de la fruta *Drosophila melanogaster*. A esta especie se le considera un organismo modelo porque, entre otras razones, es versátil de manipulación genética; resultando en uno de los organismos eucariontes genéticamente más conocidos y ampliamente utilizados en diversos experimentos de biología molecular (Pierce, 2009).

Además, como otros insectos, comparte diversos procesos biológicos con mamíferos. De hecho, los insectos presentan numerosas semejanzas y homologías en las vías de señalización con vertebrados (Otero-Moreno et al., 2016; Alzugaray et al., 2019); por lo que han sido utilizadas exitosamente para estudiar el control del metabolismo, el crecimiento y la proliferación entre otros diversos procesos (Rubin, 1988; Smith et al., 2014; Owusu-Ansah y Perrimon, 2014). *Drosophila*, ofrece una serie de ventajas prácticas sobre otras especies en lo que a propósitos de investigación genética se refiere. Entre otras, presentan un ciclo de vida corto, son fáciles de criar y manipular en el laboratorio, requieren poco espacio para su almacenamiento y un costo reducido.

El genoma de *D. melanogaster* tiene un tamaño aproximado de 180 MB (1 MB equivale a un millón de pares de bases) y contiene alrededor de 13.600 genes (Adams, 2000). Fue publicado originalmente en marzo de 2000 en la revista *Science* por la corporación *Celera Genomics* (Pennisi, 2000), y luego anotado en la base de datos FlyBase (https://flybase.org/). Es uno de los genomas eucarióticos pluricelulares más pequeños, representa tan sólo el 6 % del tamaño del genoma humano (3.000 Mb). Sin embargo, es un tamaño típico si lo comparamos con el de otras especies de dípteros (ej., el de *Anopheles gambiae* tiene 260 Mb) (Powell, 1997).

El complemento cromosómico de *D. melanogaster* consta de 4 pares de cromosomas: un par sexual X/Y (cromosoma I) y 3 pares de autosomas (II, III y IV). Los cromosomas Y y IV son pequeños y telocéntricos. La mayoría de los genes se

localizan en los cromosomas X, II y III, que son grandes y metacéntricos. En cuanto a la organización molecular del genoma de *D. melanogaster*, se han descrito tres componentes principales: ADN de secuencia única, que representa el 67 % del total; ADN moderadamente repetitivo (12 %) y ADN altamente repetitivo (21 %) (Hartl et al., 1995). También pueden encontrarse secuencias relacionadas entre sí como por ejemplo pseudogenes o secuencias parálogas (Powell, 1997).

El genoma, el transcriptoma y el proteoma de *D. melanogaster* han sido estudiados y caracterizados en diferentes etapas de su ciclo de vida (Zeitouni et al., 2007; Cammarato et al., 2011). Actualmente se dispone de amplias colecciones de líneas mutantes accesibles a la comunidad científica. Se estima que cerca del 75 % de los genes humanos vinculados con enfermedades tienen su homólogo en el genoma de *Drosophila spp.*, y el 50 % de las secuencias proteicas de la mosca tiene homólogos en mamíferos (Reiter et al., 2001). Dada la gran conservación de genes en relación con mamíferos, este género constituye un modelo extensamente utilizado para el estudio genético de enfermedades tales como diabetes (Rulifson, 2002), cáncer (Bier, 2005), Parkinson (Ambegaokar et al., 2010), Alzheimer (Fernandez-Funez et al., 2015), obesidad (Diop y Bodmer, 2012), enfermedades cardio-vasculares (Wolf et al., 2006) y diferentes tipos de adicciones (McClung y Hirsh, 1998); entre otras. Esta mosca también se usa en estudios de mecanismos del envejecimiento y estrés oxidativo, sistema inmunitario, adicción a drogas, etcétera.

Las especies del género Drosófila (*Drosophila: Drosophilidae*) son insectos dípteros braquíceros, de distribución cosmopolita. Poseen un ciclo vital de alrededor de 10 días; y la expectativa de vida promedio del adulto es de 70 días a 25^{o}C, de modo que se pueden estudiar muchas generaciones en un corto período de tiempo (Ashburner et al., 2005). La oviposición puede efectuarse a diario, llegando a producir entre 400 y 500 huevos en 10 días. El ciclo de vida está integrado por 4 estados: huevo, larva (3), pupa y adulto; tal como se esquematiza en la Fig. 4.1.

La naturaleza del estudio propuesto en esta tesis hace necesaria la utilización de la técnica de *Ribo-Seq*, que consiste en el secuenciamiento de los fragmentos de transcriptos protegidos por ribosomas, además de datos genómicos, transcriptómicos y proteómicos; que serán detallados en las próximas secciones. Dadas estas restricciones, logramos ubicar una ventana temporal en el ciclo vital de la mosca: el embrión temprano de 0-2 horas; el cual cuenta con abundante información disponible públicamente.

4.2. Datos utilizados

Durante el desarrollo de esta tesis, se utilizó el genoma de *D. melanogaster* dos Santos et al. (2015). En particular, la versión 6.03 (dmel_r6.03_FB2014_06) en formato .FASTA, así como el archivo de las anotaciones en formato .GTF correspondientes a la misma versión; ambos obtenidos de la base de datos FlyBase. En el archivo .GTF se encuentran listadas las anotaciones del genoma, es decir las posiciones que delimitan físicamente los genes, exones e intrones conocidos del genoma. Esta versión consiste en 480803 genes. El genoma utilizado en nuestro estudio fue reducido en términos de que se eliminaron todos aquellos fragmentos de secuencias que no estaban apropiadamente vinculados a una posición exacta en un cromosoma. Además, se desconsideró el cromosoma Y por contener sólo 16 genes y ser enteramente heterocromático. Por lo tanto, el genoma reducido utilizado corresponde a

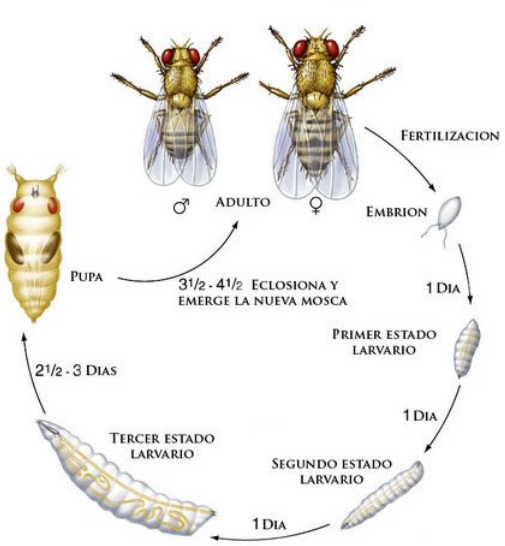

Figura 4.1: Ciclo de Vida de *Drosophila melanogaster*. La imagen representa el dimorfismo sexual entre machos y hembras, y los seis estadíos de desarrollo comprendidos entre el huevo fertilizado, los tres estadíos larvarios, y la pupa donde transcurre la metamorfosis holometábola hasta la emergencia del organismo adulto.

un archivo fasta con las secuencias de ambos brazos de los cromosomas 2 y 3 (2R, 2L, 3R y 3L), además de los cromosomas 4 y X.

Este genoma reducido fue utilizado para mapear, como veremos más adelante, lecturas de fragmentos de ARNm de dos tipos diferentes: fragmentos de ARNm protegidos por ribosomas (denominados *Ribo-Seq*) y fragmentos de ARNm "libres" (denominados *RNA-Seq*). La técnica de *Ribo-Seq* es utilizada para analizar perfiles de densidad de ribosomas y brinda evidencia de traducción de las secuencias encontradas en estos fragmentos, denominadas huellas ribosómicas (o *footprints*). Esta técnica permite analizar la traducción a nivel de genoma completo. Fue desarrollado por Ingolia et al. (2009). En esta tesis utilizamos los datos crudos, es decir sin procesamiento, obtenidos por Dunn et al. (2013) y disponibles en *Short Read Archive* (SRA) del NCBI (https://www.ncbi.nlm.nih.gov/sra). El conjunto completo de estos datos corresponde al proyecto con ID SRP028243 en SRA (o la serie GSE49197 en GEO), y consta del secuenciamiento de 12 muestras transcriptómicas de *D. melanogaster*. Dichos secuenciamientos fueron realizados en un equipo *Illumina HiSeq 2000*. Seis de estas muestras corresponden a datos de *Ribo-Seq* de embriones tempranos de 0-2 horas de *D. melanogaster*; cuatro muestras corresponden a datos de *Ribo-Seq* de un cultivo de células S2, un linaje derivado de embriones de *Drosophila* en etapa tardía. Además, esta serie contiene otras muestras de embriones tempranos de 0-2 hs correspondientes a *RNA-Seq*. En esta tesis, consideramos las lecturas de *Ribo-Seq* y *RNA-Seq* derivadas de las muestras de embriones tempranos de 0-2 hs, listados en la Tabla 4.1.

Para digerir los fragmentos de ARNm no protegidos por ribosomas se utiliza

ID del Exp.*	Réplica	ID de las corridas*	# de secuencias	# de bases
		SRR942868	54.857.951	$2{,}7\times10^9$
SRX327686	A	SRR942869	144.718.737	$8{,}2\times10^9$
		SRR942870	66.618.455	$3{,}3\times10^9$
SRX327687	B	SRR942871	146.555.910	$8{,}4\times10^9$
SRX327688**	A	SRR942872	88.948.007	$4{,}4\times10^9$
SRX327689**	B	SRR942873	80.431.415	$4{,}0\times10^9$
SRX327690	A	SRR942874	27.730.622	$1{,}4\times10^9$
SRX327691	B	SRR942875	28.089.260	$1{,}4\times10^9$
SRX327692	A	SRR942876	32.237.241	$1{,}6\times10^9$
SRX327693	B	SRR942877	40.195.868	$2{,}0\times10^9$

Tabla 4.1: Los asteriscos (*) se refieren al identificador de experimentos y corridas en SRA. Los asteriscos dobles (**) corresponden a *RNA-Seq*, los restantes corresponden a *Ribo-Seq*. Todas las secuencias obtenidas son de 50 pb no pareadas.

la enzima nucleasa micrócica (MNase) que digiere ácidos nucleicos, y se aísla la fracción de monosomas mediante gradientes de sacarosa. Luego, se extrae el ARN de las fracciones de monosomas usando fenol y cloroformo. Posteriormente, se seleccionaron los fragmentos por tamaño de 28-34 meros mediante purificación en gel. En las muestras de *RNA-Seq*, el ARN se extrajo con Trizol y se seleccionaron aquellos fragmentos con poli(A) usando gránulos con oligo dT (25) Dynabeads (ThermoFisher). A continuación, todos los fragmentos se desfosforilaron, se amplificaron por PCR y se sometieron a 50-57 ciclos de secuenciación en un equipo *Illumina HiSeq 2000*. El resultado son lecturas simples, es decir, no pareadas, de 50 pares de bases de tamaño.

4.3. Pre-procesamiento

Estas secuencias están en archivos en formato de archivo binario .SRA, propia del *Sequence Read Archive*, y deben ser transformados para su procesamiento al formato de archivo .FASTQ. Para esta transformación de formato de archivo utilizamos el comando *fastq-dump* del software *SRAtoolkit* (versión 2.8.2), disponible en el sitio de NCBI (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software). Luego, se eliminaron los adaptadores del extremo 3' UTR (CTGTAGGCACCATCAAT) de las lecturas de los archivos .FASTQ mediante el software *Cutadapt* (versión 1.9.1) (Martin, 2011). Además, dado que las lecturas demasiado cortas se alinean repetidas veces, o bien, que el software de alineación las corta para realizar alineamientos muy poco confiables; se descartaron todas las lecturas cuya longitud después del recorte del adaptador haya quedado menor a 25 pb. Para realizar esta tarea, se incluyó en el comando de *Cutadapt* una longitud mínima de las lecturas. El siguiente comando ejemplifica el procedimiento con un archivo tipo SRR_file.fastq:

./cutadapt -a CTGTAGGCACCATCAAT -m 25 -o trim_file.fastq SRR_file.fastq

donde *file* corresponde al ID de cada corrida, -a especifica que elimina el adaptador

CTGTAGGCACCATCAAT del extremo 3', y -m 25 especifica la eliminación de las lecturas que resultaran menores a 25 pb después de la eliminación del adaptador.

Como estamos interesados solamente en secuencias codificantes de proteínas, también se eliminaron las lecturas derivadas de secuencias no-codificantes conocidas. Esto fue realizado utilizando el software de alineamiento de lecturas cortas *Bowtie2* (versión 2.2.6) (Langmead y Salzberg, 2012), disponible en http://bowtie-bio.sourceforge.net/bowtie2. Para eso se construye un índice de secuencias que incluye todas las secuencias de ARN ribosomal (ARNr), de transferencia (ARNt), y pequeños ARN nucleares. Este índice se construye con el siguiente comando:

./bowtie2-build trsnRNA.fasta RNADB_index

donde el archivo trsnRNA.fasta contiene las secuencias de ARN mencionadas. Luego de elaborar el índice RNADB_index, procesamos cada uno de los archivos .FASTQ correspondientes tanto a *RNA-Seq* como *Ribo-Seq*, con el siguiente comando:

./bowtie2 -q –quiet –un select_file.fastq -x RNADB_index- -U trim_file.fastq -S tmp.sam

En este caso, el programa retiene en el archivo "select_file.fastq" solo aquellas lecturas que no se alinean contra el índice RNADB_index y que serán usadas en los siguientes análisis. Por otro lado, los alineamientos obtenidos son almacenados en el archivo tmp.sam, que será descartado de otros análisis. La eliminación de las lecturas ribosomales mencionada se realizó a todos los secuenciamientos en la Tabla 4.1.

4.4. Cuantificación de los niveles de expresión

Para cuantificar los niveles de expresión génica se requiere contabilizar la cantidad de las lecturas de *RNA-Seq* previamente seleccionadas que se alinean a cada uno de los genes en el genoma de referencia, en este caso el genoma de *D. melanogaster* reducido, mencionado anteriormente. El proceso de identificar la región del genoma que se alinea con la lectura se conoce como el procedimiento de mapeo. En el trabajo de Dunn et al. (2013) el alineamiento se realizó con el programa *Bowtie*. Este programa no reconoce el sitio de empalme de dos exones en las lecturas, por lo que ofrece un alineamiento muy pobre de este tipo de lecturas. En esta tesis, la alineación de las lecturas al genoma reducido se realizó mediante los softwares *TopHat2* y *Cufflinks*. Si bien *TopHat2* utiliza *Bowtie2* como motor de alineamiento, este es capaz de dividir una lectura con empalme exón-exón en dos fragmentos y mapear cada uno de los fragmentos a los diferentes exones de un gen. Por esta razón, cuando se trata de organismos que tienen splicing alternativo, este software ofrece un alineamiento más fidedigno que *Bowtie2* y es capaz de identificar y cuantificar las diferentes isoformas de un mismo gen.

El mapeo con *TopHat2* también requiere la construcción de un índice con las secuencias genómicas. Este índice fue construido utilizando el comando *bowtie2-build*, del siguiente modo:

./bowtie2-build dmel-chrom234X-r6.03.fasta dmel-genome603

Este proceso genera seis archivos de salida ("dmel-genome603.*.ebwt") que conformaron la base de datos indexada, construida a partir del genoma reducido conformado con las secuencias de los cromosomas 2R, 2L, 3R, 3L, 4 y X.

El alineamiento al índice "dmel-genome603" de las secuencias seleccionadas, correspondientes a las dos réplicas de *RNA-Seq* (SRX327688 y SRX327689), se realizó mediante el siguiente comando:

./tophat2 -p 4 -o ./dir-out -G dmel-all-r6.03.gtf –no-novel-juncs dmel-genome603 select_SRX327688.fastq,select_SRX327689.fastq

El término "dmel-all-r6.03.gtf" indica el nombre del archivo en formato .GTF donde están las anotaciones del genoma. En particular, en este procedimiento es necesaria la delimitación de cada uno de los exones de cada transcripto. El término "–no-novel-juncs" especifica al programa que no considere nuevas junturas exón-exón durante el análisis, sólo las anotadas en el .GTF.

El alineamiento con *TopHat2* crea un archivo de salida denominado "accepted-hits.bam" en el directorio indicado como "dir-out". El archivo .BAM es una versión binaria, y por lo tanto más compacta, del formato .SAM. Estos archivos, .SAM o .BAM, contienen toda la información resultante de los alineamientos encontrados por *TopHat2*; tales como la secuencia mapeada, su posición en el cromosoma, el sentido de la cadena o la calidad del alineamiento. Estos archivos son convertibles unos en otros con las herramientas disponibles en *SAMtools*. Para realizar procesamientos subsecuentes, este archivo de salida requiere ser indexado utilizando el comando *samtools index* de la siguiente manera:

./samtools index accepted-hits.bam

Esta indexación genera un archivo de extensión .BAI. Después de alinear las lecturas de *RNA-Seq* al genoma con *TopHat2* y generar el archivo de salida, se requiere un análisis cuantitativo del mapeo, es decir, la determinación y conteo de las isoformas de cada transcripto. Para esta tarea, se empleó el software *Cufflinks* (versión 2.2.1) (Trapnell et al., 2012). Este programa organiza las lecturas alineadas y ensambla las isoformas en forma parsimoniosa. Luego, estima la abundancia relativa de estas transcripciones en función de cuántas lecturas son compatibles con cada isoforma. Este procedimiento se realizó utilizando el siguiente comando:

./cufflinks -p 8 -o ./cufflink-out accepted-hits.bam

Este método genera, en la carpeta indicada "cufflinks-out", cuatro archivos de salida denominados "isoforms.fpkm-tracking", "transcripts.gtf", "genes.fpkm-tracking" y "skipped.gtf". El archivo "isoforms.fpkm-tracking": contiene los valores estimados de expresión de cada isoforma. El archivo "transcripts.gtf" es un archivo en formato .GTF que contiene información como la posición en el genoma o sobre los transcriptos ensamblados, el cual será utilizado en el paso siguiente. Por su parte, "genes.fpkm-tracking", en forma similar al archivo de las isoformas, contiene los valores estimados de expresión de cada gen. Finalmente, "skipped.gtf" guarda algunos transcriptos cuyo número de lecturas mapeadas excede o no alcanzan ciertos parámetros. En general, un gen con demasiadas lecturas mapeadas no es analizado

y es reportado aquí.

Los niveles de expresión informados por *Cufflinks* están en unidades FPKM (del inglés: *Fragments Per Kilobase Million*), que suelen ser comparables entre muestras; pero que en determinadas situaciones, la aplicación de un nivel adicional de normalización puede eliminar las fuentes de sesgo en los datos. *Cufflinks* incluye un programa, *Cuffnorm*, que normaliza un conjunto de muestras para que estén en escalas lo más similares posible, lo que puede mejorar los resultados que se obtienen con otras herramientas posteriores. Este procedimiento se realizó utilizando el siguiente comando:

./cuffnorm -p 8 transcripts.gtf accepted-hits.bam

donde el archivo "transcripts.gtf" es el generado en el paso previo por *Cufflinks*. *Cuffnorm* genera un conjunto de archivos de texto delimitados por tabulaciones, que contienen niveles de expresión normalizados para cada gen o transcripción, en el experimento.

La cuantificación de los niveles de expresión por *RNA-Seq* se utilizó para descartar de los análisis subsiguientes a transcriptos pobremente expresados.

4.5. Alineamiento de las huellas ribosomales

En la primera parte de este estudio, se alinearon las lecturas de *RNA-Seq* al genoma de *D. melanogaster* para descartar del posterior análisis, aquellos transcriptos que no cuentan con la expresión suficiente para elaborar y analizar los perfiles ribosomales que identifican eventos de LCP. En esta sección se detalla la metodología para el mapeo de las huellas ribosomales y la construcción de los perfiles ribosomales de cada transcripto. En forma similar a la cuantificación de los niveles de expresión génica a partir de las lecturas de *RNA-Seq*, la alineación de los fragmentos protegidos por ribosomas se realizó mediante el software *TopHat2*, usando el índice del genoma reducido "dmel-genome603" previamente construido, con el siguiente el comando:

./tophat2 -p 4 -o ./dir-out -G dmel-all-r6.03.gtf –no-novel-juncs dmel-genome603 select_file.fastq

para cada uno de los archivos "select_file.fastq" de Ribo-Seq, donde "file" es uno de los identificadores: SRR942868, SRR942869, SRR942870, SRR942871, SRR942874, SRR942875, SRR942876, SRR942877. Como el número de lecturas totales es muy grande y resulta difícil manipular archivos de gran tamaño, optamos por la estrategia de dividir las lecturas según cromosomas y por sentido de lectura. Esto se realizó con un comando del *kit* de herramientas de *SAMtools*, que extrae de un archivo .BAM los alineamientos separados por cromosomas y por sentido. Esto fue realizado mediante el uso de dos comandos:

./samtools view -F 20 accepted-hits.bam cromo >cromo-fw_file.bam
./samtools view -f 0x10 accepted-hits.bam cromo >cromo-rv_file.bam

El primer comando selecciona los alineamientos en el sentido *forward* (5'-3') del

archivo .BAM en un cromosoma dado, indicado por la opción "cromo". El segundo comando selecciona aquellos alineamientos en el sentido contrario *reverse*. La opción f o F del comando *samtools view* filtra los alineamientos deseados de la columna 2 del archivo .SAM, mientras que la columna 3 indica en qué cromosoma fue mapeada la lectura Este procedimiento es repetido para cada archivo y para cada uno de los seis cromosomas considerados, es decir que "cromo" toma valores 2L, 2R, 3L, 3R, 4 y X.

Todos los archivos .BAM, resultantes del paso previo, correspondientes a un cromosoma dado y a un sentido dado, se unen para poder obtener en el próximo paso el correspondiente perfil ribosomal. Esta unión de los alineamientos se hace con la herramienta *merge* de *SAMtools*. Así, por ejemplo, los alineamientos de todos los *Ribo-Seq* correspondientes al cromosoma 2L en el sentido 5'-3' se obtienen con el siguiente comando:

./samtools merge 2L-fw_merged.bam 2L-fw_SRR942868.bam 2L-fw_SRR942869.bam
2L-fw_SRR942870.bam 2L-fw_SRR942871.bam 2L-fw_SRR942874.bam
2L-fw_SRR942875.bam 2L-fw_SRR942876.bam 2L-fw_SRR942877.bam

El paso previo para obtener los perfiles de densidad ribosomal, consiste en ordenar por posición cromosómica cada una de las lecturas alineadas. Así, cada uno de los archivos "cromo-sentido_merged.bam", (en el ejemplo previo sería el archivo "2L-fw_merged.bam") debe ser sorteado con el siguiente comando:

./samtools sort -o cromo-sentido_sorted.bam cromo-sentido_merged.bam

Finalmente, para obtener el perfil de huellas ribosomales, las lecturas que resultaron mapeadas a cada cromosoma en un sentido dado deben ser compiladas en un formato de archivo apropiado, denominado .WIG. Este formato informa el número de lecturas que han sido mapeadas a cada posición en un cromosoma con resolución de un nucleótido. Para realizar esta compilación se utilizaron los comandos *mpileup* del *kit* de *SAMtools*, tal como se muestra a continuación:

./samtools mpileup cromo-sentido_sorted.bam -O -o cromo-sentido_sorted.wig

Este archivo de salida .WIG consta de columnas separadas por tabulación, y cada línea representa la acumulación de lecturas en una única posición genómica. Las primeras tres columnas corresponden al nombre del cromosoma, la posición y el nucleótido de referencia en esta posición. Las columnas restantes muestran los datos de acumulación de lecturas mapeadas en la posición indicada en la columna 2. La información de interés es el número de lecturas que cubren esta posición, y está indicada en la cuarta columna. Esta información permitió construir el perfil de densidad ribosómica de cada transcripto de interés, que consiste en el número de lecturas alineadas en cada posición del transcripto.

Utilizando un programa elaborado en el software *Mathematica*, se construyeron los gráficos de perfiles de ribosomas de cada transcripto cuyo nivel de expresión, obtenido por el procesamiento previo de los datos de *RNA-Seq*, sea mayor a 2 RPKM. En esta instancia, se indicó el inicio y fin de los marcos de lectura para distinguir las regiones 5' UTR y 3' UTR de la región codificante, según lo indicado en el archivo de anotaciones .GTF. Además, se indicaron las posiciones de los codones

de parada posteriores, dentro del mismo marco de lectura.

Para evitar las variantes de un mismo gen producidas por los fenómenos de *splicing* alternativo, se descartaron todas las variantes de *splicing* que no registraron lecturas mapeadas.

4.5.1. Construcción y análisis de perfiles de ribosomas

El análisis de los perfiles de ribosomas consistió en construir gráficos que visualicen la cantidad de lecturas de huellas de ribosomas a lo largo de todo el transcripto, tanto en la región codificante (CDS) como en las regiones no traducidas UTR. La delimitación de dichas regiones fue señalada con barras ubicadas en las posiciones de los nucleótidos correspondientes al codón de inicio y de parada. Particularmente, se seleccionaron aquellos casos que presentaban alta densidad ribosomal en la región 3' UTR, inmediatamente posterior al codón de terminación anotado.

Para elaborar los perfiles de densidad ribosomal de cada uno de los transcriptos seleccionados, se empleó el software *Mathematica*. En este programa se configuraron los parámetros necesarios para la construcción de la gráfica del perfil, introduciendo para ello los datos correspondientes a cada uno de los transcriptos. De esta forma, se obtuvo una imagen de la cantidad de lecturas de huellas ribosomales medida en RPKM.

La gráfica de perfiles de densidad ribosómica que se muestra para un transcripto dado, consiste en un plano cartesiano donde la cantidad de huellas ribosomales medidas se indican en el eje Y, mientras que la ubicación física en la secuencia del transcripto se define sobre el eje X según la posición a nivel de pares de bases nitrogenadas (pb). Además, se señalan los codones de inicio (posición cero) y de terminación de la traducción con líneas verticales verdes y rojas, respectivamente. De este modo quedan delimitadas las regiones codificantes (CDS) y las regiones no traducidas 5' UTR (izquierda) y 3' UTR (derecha).

4.6. Identificación de nuevos eventos de LCP a través de perfiles ribosomales

La identificación de los eventos de LCP fue realizada a través del análisis de los perfiles ribosomales de todos los transcriptos, obtenidos tras el mapeo de las lecturas correspondientes a las huellas ribosomales previamente secuenciadas. Dicho mapeo se realizó con el software *TopHat2*, generando un archivo de alineamiento .SAM (detallado en la sección 4.4).

En primera instancia, con el objeto de identificar el mayor número posible de eventos de LCP, se examinaron los perfiles de ribosomas de todos los transcriptos de experimentos con embriones de *D. melanogaster*. En este sentido, se seleccionaron 1303 casos que presentaban una densidad de ribosomas significativa más allá de los codones de terminación anotados. Esta colección de eventos conformó la materia pria para la segunda parte de nuestro estudio, que consistió en el establecimiento de los factores determinantes del LCP.

Para discriminar posibles secuencias de traducción, se tuvieron en cuenta los siguientes aspectos:

Por un lado, la región 5' UTR suele presentar una densidad ribosomal menor que en la región codificante, aunque no es nula como la esperada para una región que no se traduce. Diversos autores indican que esta densidad en la región 5' UTR no es considerada como indicativo de traducción de esa región (Jungreis et al., 2011; Dunn et al., 2013). Sin embargo, en la región 3' UTR pueden encontrarse situaciones diferentes. Por ejemplo, existen casos donde la densidad de ribosomas es nula, con total ausencia de lecturas mapeadas en esa región. En estos casos, entonces, la terminación de la traducción es eficiente en el primer codón de parada anotado. Por el contrario, en otros casos puede existir densidad de ribosomas entre el codón de parada anotado y un segundo codón de parada en el extremo 3' UTR. Aunque este nivel de densidad ribosomal es muy inferior al registrado en la región codificante, corresponde a un número sustancial de lecturas alineadas a una porción de transcripto. Este tipo de patrón de densidad de ribosomas ha sido considerado como un marcador confiable de eventos de LCP (Dunn et al., 2013; Jungreis et al., 2016). De acuerdo con estos autores, en esta tesis se considera la existencia de la densidad ribosomal como un marcador de eventos de LCP.

4.7. Confirmación de eventos de LCP por espectrometría de masa

El espectro de masas es una gráfica de histograma que representa la distribución de iones frente a su relación masa/carga (m/z) en una molécula. Se considera un análisis químico que determina la intensidad de señal de los iones, separados en función de la masa. El eje X de un espectro de masas representa una relación entre la masa de un ión dado y el número de cargas elementales que lleva. Esto se escribe como el estándar IUPAC (m/z) para denotar la cantidad formada dividiendo la masa de un ion por la unidad de masa atómica unificada y por su número de carga (valor absoluto positivo) (Todd, 1995). El eje Y del espectro representa la intensidad de la señal de los iones, que se correlaciona con la abundancia relativa; y se indica con picos (barras). Los picos reflejan la abundancia de iones para cierta masa (peso atómico específico), por lo que picos altos indican mayor presencia de un aminoácido dado. El número que aparece próximo a cada pico indica la masa del pico, y a la izquierda se mide su intensidad relativa. Al pico más alto (de máxima intensidad) se le denomina pico base y corresponde al ión más estable.

En la secuenciación de péptidos *de novo*, la secuencia identificada se fragmenta usando los iones "b" e "y" cargados individualmente; según la Nomenclatura de Fragmentación de Péptidos (Roepstorff y Fohlman, 1984; Biemann, 1988). De este modo, los picos de fragmentos que se extienden desde la parte frontal del péptido (el extremo amino) se denominan "iones b". El fragmento que contiene sólo el aminoácido amino terminal se denomina "ion b1". El fragmento que contiene los dos primeros aminoácidos amino terminales se denomina "ion b2", y así sucesivamente. De manera similar, los grupos de iones de fragmentos de péptidos que se extienden desde la parte reversa del péptido (el terminal C), se denominan "iones y".

Para la interpretación de los resultados de la identificación proteómica se utilizó el software *Peptide-Shaker* (http://peptide-shaker.googlecode.com) (Vaudel et al., 2015). La asignación de secuencias peptídicas a los espectros MS/MS depende generalmente de aproximaciones computacionales, en las que a cada pareja péptido-espectro se les denomina PSM (del inglés: *Peptide to Spectrum Match*, o Asignación

Péptido-Espectro). En este caso, el espectro correspondiente al PSM seleccionado se muestra en la parte inferior derecha con la anotación de fragmentos de iones (personalizable por el usuario). Los tres *subplots* ubicados por encima del espectro facilitan la evaluación de calidad del *match*:

- se muestra la intensidad de cada fragmento de ion en relación con la secuencia peptídica para iones hacia adelante (azul, hacia abajo) y hacia atrás (rojo, hacia arriba);
- un histograma de las intensidades de los picos anotados (verde) y no anotados (gris);
- el error del fragmento de ion m/z se traza contra el pico m/z para los iones hacia adelante (azul) y hacia atrás (rojo).

Esto último permite la detección directa de problemas de calibración. El usuario puede incluso personalizar la anotación de espectro y la pantalla de visualización, como el color y la apariencia de los picos, beneficiarse de la anotación avanzada y exportar los gráficos en calidad de publicación (Vaudel et al., 2011). Por defecto, sólo se muestran en primer plano y en rojo los picos actualmente anotados por un fragmento de ion. Los otros picos se muestran en gris y no son seleccionables, a menos que se indique en las opciones del menú del espectro. En lo referente a la anotación, *Peptide-Shaker* anota automáticamente el espectro con los iones más probables, dado el espectro y la secuencia de péptidos identificados. Esto incluye la selección automática de pérdidas neutrales. Sin embargo, el usuario puede anular la anotación predeterminada y seleccionar diferentes tipos de iones. En cuanto a la fragmentación de la secuencia peptídica, la misma se representa usando los iones "b" e "y" cargados individualmente, con picos azules y rojos, respectivamente. Al posar el mouse sobre un pico, se muestra el tipo y número de iones del fragmento (ej.: y5). Por otro lado, el histograma de intensidad representa la distribución de intensidad de los picos anotados (en verde) y de los picos no anotados (en gris), con intervalos de intensidad en el eje X y frecuencia en el eje Y. En un espectro bien anotado, la mayoría de los picos de alta intensidad deberían ser anotados, mientras que la mayoría de los picos no anotados deberían tener bajas intensidades. Finalmente, el gráfico de error de masa estándar muestra la relación de carga-masa m/z en el eje X y el error de masa (en Da) en el eje Y.

Capítulo 5

Cierre y conclusiones

5.1. Discusión

En la expresión génica los genes son leídos por un conjunto de enzimas, siendo generalmente la síntesis de proteínas el producto final. No obstante, el "control de calidad" de la expresión génica opera post transcripcionalmente y en varios niveles. En particular, mediante un variado conjunto de mecanismos regulatorios, la traducción de los transcriptos de ARNm es regulada con precisión para garantizar la síntesis diferencial de proteínas y su concentración celular. Por lo visto en esta tesis y trabajos previos, la regulación de la liberación de las proteínas maduras, cuando el ribosoma alcanza el primer codón de terminación podría agregarse a este conjunto de mecanismos de regulación.

Hemos visto que durante la LCP el codón de parada anotado es leído por el ribosoma y en lugar de escindir la cadena peptídica, continúa traduciendo en la región 3' UTR del transcripto hasta alcanzar el siguiente codón de parada. Como resultado, se sintetizan isoformas proteicas extendidas en su extremo C-terminal, con efectos potencialmente alterados en su función. En la actualidad, la LCP se emplea como terapia biomédica basada en el uso de aminoglucósidos, para morigerar el efecto de mutaciones sin sentido que causan enfermedades. El uso de estos fármacos permite a las células ignorar la señal de parada prematura (CPP) y continuar la producción de una proteína completa (Keeling et al., 2014; Blanchet et al., 2014; Bidou et al., 2017; Dabrowski et al., 2018). Si bien existe abundante evidencia de LCP como un fenómeno común en la expresión génica, las pruebas relacionadas a sus consecuencias funcionales son escasas (von der Haar y Tuite, 2007; Jungreis et al., 2011; Eswarappa et al., 2014; Dabrowski et al., 2015; Schueren y Thoms, 2016; Loughran et al., 2018).

Respecto a la identificación de genes que experimentan LCP, se han empleado diversos enfoques. Por un lado, el enfoque filogenético en el que, mediante estudios comparativos, se probó que la codificación alternativa del codón de parada se conserva evolutivamente (Lin et al., 2007, 2011; Jungreis et al., 2011, 2016; Loughran et al., 2018; Garofalo et al., 2019; Sapkota et al., 2019). Más recientemente, se utilizó el análisis de perfiles de densidad de ribosomas (Ingolia et al., 2009, 2011; Dunn et al., 2013) y los modelos de regresión *in silico* (Schueren et al., 2014; Loughran et al., 2014; Stiebler et al., 2014). Estos experimentos revelaron la existencia de una lectura generalizada de codones de parada en eucariotas, con niveles que varían

entre los distintos tejidos (Dunn et al., 2013; Artieri y Fraser, 2014; Jungreis et al., 2016). Estas demostraciones condujeron a la "hipótesis adaptativa" que plantea que la LCP es un mecanismo regulado, de importancia en la generación de diversidad en los proteomas (Dunn et al., 2013; Stiebler et al., 2014; Van Damme et al., 2014; Baranov et al., 2015; Pancsa et al., 2016). En contraposición, otros autores han propuesto que la LCP deviene principalmente de errores moleculares que se producirían durante el proceso de traducción, resultando perjudicial para el mantenimiento de la fracción de proteínas con actividad normal y dando como resultado además, un posible incremento de proteínas tóxicas o disfuncionales, siendo generalmente no adaptativa (Bidou et al., 2010; Li y Zhang, 2019).

La hipótesis del error molecular radica en que la presión negativa de la selección natural en detrimento de los procesos de LCP en un gen, se intensificaría con la concentración del ARNm correspondiente, y en consecuencia, la tasa de LCP debería disminuir de acuerdo al incremento de la expresión génica. Por el contrario, la hipótesis adaptativa no predice esta correlación negativa, sino que propone que la tasa de LCP en un determinado gen dependería de la función específica que tendría la proteína elongada. Por lo tanto, dado que las evidencias de traducción de proteínas extendidas son consistentes y variables entre especies, resulta indispensable conocer qué patrones estructurales presentes en las secuencias promueven este tipo de recodificación. Además, el estudio de los nucleótidos próximos al codón recodificado permite predecir qué genes son susceptibles a desarrollar LCP, y así conocer el potencial de variabilidad a nivel de los proteomas que este mecanismo puede generar.

En esta tesis se proporciona evidencia de patrones estructurales en las secuencias que promueven este tipo de recodificación del codón de parada, abonando la hipótesis de la recodificación funcional mediante LCP programada en *Drosophila melanogaster*.

Partiendo de la hipótesis de que la presencia de una fracción elevada de proteínas extendidas en su extremo C-terminal constituye un fuerte indicio de que estas extensiones corresponden a un evento funcional de LCP programada, y no a un fallo en la terminación de la traducción, nos propusimos identificar nuevos eventos de LCP. Además, adoptamos como segunda hipótesis que el mecanismo de LCP puede ser influenciado por determinados patrones nucleotídicos en la secuencia de contexto del codón de parada. En este sentido, el objetivo se centró en identificar los elementos que actúan en *cis* e inducen la expresión de un compuesto extendido mediante LCP.

5.1.1. Identificación de eventos de LCP mediante perfiles ribosomales

Con el objetivo de contrastar estas hipótesis y obtener evidencias del mecanismo de regulación del proceso de recodificación, se procesaron y analizaron los datos de *ARN-Seq* y *Ribo-Seq* obtenidos a partir de embriones de *D. melanogaster*, de acuerdo a lo detallado en el capítulo 4 (Datos y metodología). Las lecturas de fragmentos de ARNm capturados por ribosomas (*footprints*) se mapearon contra el genoma completo de *D. melanogaster* usando el software *TopHat*, y se cuantificó la densidad de ribosomas asociada a lo largo de cada transcripto. Esta técnica permite identificar con alta precisión qué regiones de ARNm se traducen, y cuantificar la traducción a nivel de genoma completo. En esta tesis se elaboraron perfiles de densidad ribosomal

para 6739 transcriptos. Posteriormente, se estimó la tasa de fuga ribosomal asociada a cada codón de parada. A fin de reconocer los posibles factores que inducen la LCP, se analizó la influencia de los nucleótidos en diferentes posiciones cercanas al codón de parada original, determinando aquellos que presentan un sesgo de uso entre grupos de transcriptos con altas tasas de LCP y aquellos que no la presentan. Finalmente, mediante un enfoque de modelos predictivos, y usando la información de distintas posiciones de la secuencia contexto, se evaluó el rol de tales posiciones en función de la tasa de fuga.

La evaluación exhaustiva de las regiones 3' UTR de dichos perfiles permitió identificar y caracterizar 1176 transcriptos con probables eventos de LCP, lo que equivale a un 17,45 % de los ARNm analizados. Se evidenció incluso la existencia de dobles y triples eventos de LCP sobre diferentes codones de terminación en un mismo transcripto. Se reconocieron, además, presuntas extensiones de LCP que no eran tales, entre las que se encuentra la secuencia del transcripto FBtr0088262, erróneamente caracterizado previamente por Jungreis et al. (2011) y Dunn et al. (2013). La cantidad de nuevos eventos de LCP encontrados en este trabajo supera los 350 candidatos reportados originalmente en embriones y células S2 de *D. melanogaster* (Dunn et al., 2013), entre los que se incluyen 43 casos previamente detectados en base al enfoque filogenético (Jungreis et al., 2011).

5.1.1.1. Análisis de sintenia y validación por espectrometría de masa

Para profundizar y ampliar el análisis, buscamos además otras formas de evidenciar los eventos de LCP señalados por el estudio de los perfiles ribosómicos. En primera instancia, las extensiones peptídicas comprendidas entre el codón de parada anotado y el siguiente codón de parada en el mismo marco de lectura, fueron sometidas a un análisis de sintenia para evaluar el grado de conservación filogenética. La comparación por alineamientos múltiples de las secuencias extendidas por LCP reflejó un elevado porcentaje de similitud con secuencias correspondientes a diferentes especies del género *Drosophila*, especialmente *D. yakuba*, *D. sechellia*, *D. simulans*, *D. suzukii*, *D. mauritiana* y *D. erecta*. Esto demuestra que dichas extensiones se encuentran conservadas en el genoma del género *Drosophila*. Incluso, presentamos un caso en el cual la extensión peptídica está presente en la región codificante de un gen homólogo, sugiriendo que la función biológica también se encuentra conservada. Así, la homología existente entre las secuencias extendidas en *Drosophila* indicaría que la traducción de las mismas confiere funciones concretas que podrían modificar y/o regular diferentes funciones de las proteínas traducidas canónicamente. Además, la expresión coordinada de transcriptos con LCP provenientes de genes relacionados en el estadio temprano de embriones de *Drosophila*, sugiere la posibilidad de la presencia de genes vecinos que comparten secuencias reguladoras, cuya funcionalidad es ampliamente desconocida hasta ahora. En este sentido, estudios anteriores establecieron que la búsqueda computacional de homología con dominios conocidos de proteínas sugiere la función codificante de proteínas en las regiones con LCP (Sato et al., 2003; Jungreis et al., 2011).

En concreto, nuestros resultados se condicen con los hallazgos de Jungreis et al. (2011), confirmando la utilidad del criterio establecido para establecer casos de LCP en base al estado de conservación de la extensión. Dado que la conservación sinténica garantiza la homología entre las secuencias contrastadas, constituye una

evidencia de su funcionalidad y un aporte valioso para la validación de nuevos genes.

El análisis de las extensiones a través de bases de datos obtenidas por medio de espectrometría de masa permitió la confirmación experimental de tres nuevos casos de LCP, determinando además, que este mecanismo de traducción alternativa es eficiente para producir proteínas funcionales, y que las mismas son detectables a nivel de espectros caracterizados. De manera similar a lo realizado previamente (Jungreis et al., 2011), las extensiones evaluadas por espectrometría de masa en el presente estudio fueron contrastadas con espectros provenientes de extractos proteicos obtenidos por Brunner et al. (2007) (disponibles en la base de datos *Peptide Atlas*). Los péptidos validados en este trabajo corresponden a fragmentos de proteínas similares caracterizadas en *D. melanogaster*, tales como la isocitrato deshidrogenasa (*Idh*), enzimas asociadas al metabolismo de oligosacáridos y a la metilación de histonas. Las secuencias peptídicas de dichos fragmentos reflejaron un alto grado de identidad con secuencias asociadas al mismo gen codificante de la proteína convencional. Con respecto a los fragmentos peptídicos identificados por espectrometría, tanto en la extensión del transcripto como en el espectro correspondiente, no se encontraron similitudes significativas. En este sentido, se infiere que dichos fragmentos de secuencia no estarían conservados entre especies del mismo género. Sin embargo, ampliando la información con secuencias de otras especies de *Drosophila*, podría evaluarse si existe conservación a nivel de género.

5.1.1.2. Análisis de enriquecimiento genético

Al analizar la ontología de la lista de genes vinculados a los transcriptos candidatos a ser blancos traduccionales de LCP, se observaron funciones biológicas asociadas principalmente al proceso de traducción. En particular, existe una marcada tendencia de genes con LCP en el codón TAA que muestran actividad en procesos relacionados con la traducción y la actividad de ribosomas. Este fenómeno se observa también en genes con el codón TAG, aunque en menor proporción. El análisis ontológico de expresión diferencial determinó además, que los genes vinculados a transcriptos con LCP se asocian a procesos biológicos diferentes en función del tipo de codón de parada recodificado. En este sentido, el proceso de traducción se vio ampliamente representado en transcriptos con codón de terminación TAA.

Como futuro análisis complementario, se sugiere comparar la ontología génica de los candidatos aquí analizados con la lista de 483 genes asociados a LCP anotados en FlyBase, a fin de corroborar si existen similitudes en las funciones afectadas por este mecanismo en ambos conjuntos. Dado que los transcriptos susceptibles de sufrir LCP presentados en esta tesis no están aún contemplados en FlyBase, creemos que este singular fenómeno no es atípico en los genomas eucariotas, y coincidimos con la opinión de otros autores en que su ejecución programada posibilita el incremento de la versatilidad funcional de los proteomas y controla la proporción de isoformas de proteínas con diferente localización o función (Beier, 2001; Touriol et al., 2003; von der Haar y Tuite, 2007; Namy y Rousset, 2010; Dunn et al., 2013; Ribas de Pouplana et al., 2014).

5.1.2. Estudio de factores determinantes de LCP en la secuencia contexto

Además de las estrategias de genómica comparativa, varias líneas de evidencia sostienen que la recodificación de proteínas observada durante la traducción es el resultado de la lectura del codón de parada, en lugar de mecanismos alternativos. Al respecto, en el trabajo de Jungreis *et al.* (2011) se proporciona una estimación independiente del número de genes con LCP en *D. melanogaster*. A partir de su análisis de la información de la secuencia de 283 transcriptos susceptibles a LCP, propusieron que este fenómeno podría ser inducido por un sesgo en la distribución de nucleótidos en el marco de lectura cadena abajo del codón de parada, en lugar de ser consecuencia del azar u otros mecanismos. En este sentido, postulan también que el contexto 4-nt del codón de parada admite la fuga del ribosoma y la continuidad de la traducción. En base a que los contextos de terminación eficiente son los más comunes en genes de mayor expresión, surge la premisa de que la frecuencia de diferentes contextos podría ser un indicador de la eficiencia en la terminación de la traducción (Bonetti et al., 1995).

En una segunda instancia de la investigación desarrollada en esta tesis, se profundizó en el estudio de la secuencia de contexto asociada al codón de terminación en cada transcripto. Si bien está comprobado que es posible reconocer eventos genuinos de LCP funcional en algunos ARNm (Palma y Lejeune, 2021), existe discordancia sobre cómo y cuándo se producen. Por este motivo, en este trabajo nos propusimos analizar la influencia de nucleótidos ubicados en diferentes posiciones de la secuencia de contexto.

La evaluación de las regiones 3' UTR mediante perfiles de densidad ribosomal nos permitió identificar 1176 transcriptos con eventos probables de LCP. En función de la tasa de fuga ribosomal ($y > 0.005$) estimada sobre 1036 candidatos, se determinó que estos eventos ocurren con una frecuencia de 23, 33 y 44 % para los codones de parada TGA, TAG y TAA, respectivamente. De hecho, se observó que la frecuencia de uso de codones de parada asociados a LCP es similar en la totalidad de los transcriptos, indistintamente de si presentan o no este tipo de eventos. Este resultado contrasta con lo reportado por Jungreis et al. (2011), para quien el codón TGA es el que se omite más frecuentemente. Sin embargo, esto puede explicarse en base al bajo índice RDI (*Repertoire Dissimilarity Index*) de dicho codón, el que podría ser atribuido a la presencia de errores de traducción. Dicho índice RDI permite la cuantificación de la variación media entre repertorios en la utilización de segmentos de genes (López-Santibáñez-Jácome et al., 2019).

El estudio de la frecuencia de uso de nucleótidos adyacentes al codón de parada determinó que existe un sesgo de uso preferencial hacia los nucleótidos C y G, tanto en la posición anterior como posterior a los codones TGA y TAG, lo que coincide parcialmente con observaciones previas (Jungreis et al., 2011) que señalan que TGA-C es uno de los contextos de 4-nt utilizados con menor frecuencia en transcriptos con terminación eficiente. Por otro lado, de acuerdo con hallazgos anteriores (Loughran et al., 2014; Cridge et al., 2018; Jungreis et al., 2016), observamos que los eventos de LCP asociados al codón TGA presentan una densidad ribosomal más elevada, indicando una mayor tasa de fuga ribosomal, particularmente con TGA-C. La correlación entre este incremento en la densidad ribosómica y la frecuencia de uso del nucleótido inmediatamente posterior al codón de parada, sugiere que los

eventos de LCP son programados, y que están causados por un mecanismo implícito asociado a la secuencia nucleotídica, en lugar de tratarse de meros errores durante la decodificación traduccional (Firth et al., 2011; Dunn et al., 2013; Blanchet et al., 2014; Dabrowski et al., 2015).

Con el fin de identificar patrones que puedan predecir eventos de LCP, en el presente trabajo se evaluó exhaustivamente la influencia de los nucleótidos contenidos en una amplia región de contexto respecto del codón de parada. Este análisis fue realizado con la medida de divergencia de Kullback-Leibler, la cual indica la presencia de sesgos en el uso de los nucleótidos en cada posición. Mediante este desarrollo, encontramos altos valores de divergencia en varias posiciones distales cadena arriba del codón de parada, muchas de las cuales son específicas de cada uno de éstos. Por ejemplo, el codón TGA muestra elevada divergencia en las posiciones -2 y -12, con una preferencia por el uso del nucleótido A (53.8 % y 48.4 % respectivamente). Las posiciones con un fuerte sesgo en el uso de un determinado nucleótido podrían indicar una marcada influencia en la ocurrencia de eventos programados de LCP. Además, nuestro estudio demostró una importante frecuencia de uso para los nucleótidos G o C (alrededor del 75 %) en la tercera base de varios codones en transcriptos con el codón TAA. Esto supera la proporción observada en las mismas posiciones en el grupo control (60 %), y la frecuencia esperada por el uso uniforme de los codones (50 %). Este notable sesgo difiere ampliamente del observado en el codón TGA, lo que indica que los mecanismos de LCP operarían bajo patrones de uso diferencial de nucleótidos específicos en posiciones distales para cada tipo de codón de terminación; algo que no estaba contemplado en estudios previos (Jungreis et al., 2011; Dunn et al., 2013; Dabrowski et al., 2015).

En el caso de los transcriptos con el codón de parada TAG, la divergencia en el uso de nucleótidos resultó ser generalmente menor que para los codones TAA y TGA, y no mostró sesgos importantes en las posiciones adyacentes.

En el transcurso de esta investigación propusimos un conjunto de modelos de regresión que difieren en el tamaño de la secuencia de contexto utilizada para hacer la predicción. El objetivo fue relacionar la influencia de tales nucleótidos en la tasa de fuga de cada codón de parada en forma independiente, con el fin de corroborar el rol de las posiciones más relevantes identificadas por el análisis de divergencia de Kullback-Leibler. El modelo de regresión lineal nos permite la identificación de genes con LCP en genomas completos (Schueren et al., 2014). Este modelo fue inicialmente desarrollado en el transcriptoma humano, bajo la hipótesis de que la LCP es influenciada por el contexto de los seis nucleótidos anteriores y posteriores al codón de terminación. Con el objetivo de identificar la ubicación de nucleótidos que puedan inducir la ocurrencia de eventos programados de LCP, el modelado propuesto en nuestro trabajo compara la influencia de los nucleótidos en tres contextos de tamaño diferente. En este sentido, el modelo que tiene en cuenta las 12 posiciones contiguas a cada codón de parada (6 anteriores y 6 posteriores) identificó eventos de LCP en el 80 % de los transcriptos evaluados, pero presentó una elevada fracción de falsos negativos en todos los casos. El modelo que usa una secuencia de contexto de 29 posiciones contiguas al codón de parada redujo considerablemente la cantidad de falsos negativos obtenida por el modelo anterior, pero incrementa la fracción de falsos positivos. Esto demuestra que el tamaño de las secuencias de contexto utilizadas en función de un número fijo de secuencias de entrenamiento no garantiza un poder predictivo fiable. Por otro lado, el modelo que sólo incluye las posiciones con un alto valor de divergencia asociado al contexto de cada codón de

parada con LCP (contiguo o no), fue efectivo al reducir significativamente la tasa de falsos positivos y negativos con respecto a los otros dos modelos. De esta manera, el uso de las posiciones más informativas sobre el nivel de divergencia dentro de una secuencia de contexto, constituye un criterio novedoso y útil para el desarrollo de herramientas computacionales como la que aquí se presenta.

Finalmente, nuestros resultados demuestran que el análisis de divergencia puede ser utilizado como criterio para seleccionar las posiciones más informativas durante el desarrollo del modelo. Nuestros resultados indican también que la velocidad a la que ocurren los eventos de LCP podría estar regulada por un contexto mayor que los propuestos por estudios anteriores (Jungreis et al., 2011; Schueren et al., 2014). También es importante tener en cuenta la relación entre el número de parámetros y el número de secuencias. Aunque está claro que cuanto mayor sea el tamaño del conjunto de entrenamiento, mejor será el ajuste de parámetros del modelo, este no es el caso en relación con el número de parámetros.

En conclusión, este estudio extenso y en profundidad de la relación entre la mayoría de las posiciones de nucleótidos y la tasa de fuga ribosómica, permite no sólo anotar un conjunto más grande de genes sometidos a LCP, sino que también revela mecanismos implícitos relacionados con la secuencia de nucleótidos, que regulan los eventos de LCP.

5.1.3. Conclusiones finales

En esta tesis se presenta la mayor cantidad de eventos de LCP identificados en transcriptos de embriones de *D. melanogaster* hasta la fecha, así como la expansión de los modelos propuestos para la evaluación de la secuencia de contexto al codón de terminación. Además, se sugiere el cálculo de la frecuencia de uso de los codones de parada y la divergencia en el sesgo de uso de los nucleótidos involucrados en este mecanismo, como método para evaluar la forma en que se afecta la tasa de LCP.

En conjunto, hemos aplicado análisis genómicos, transcriptómicos, filogenéticos y moleculares para aportar a la comprensión de la recodificación de codones de terminación durante el proceso activo de traducción. Este trabajo ha permitido una integración de las técnicas de identificación de eventos de LCP, a la vez que ha ampliado significativamente las fronteras del conocimiento sobre la regulación intramolecular de este mecanismo, posibilitando además el planteo de nuevas preguntas e hipótesis para proyectos futuros.

Los datos analizados y los resultados obtenidos durante la presente investigación apoyan firmemente la hipótesis planteada. De hecho, se demuestra que la mayoría de los eventos de LCP generan productos funcionales y de carácter adaptativo, presentando signos de conservación en la región traducida. Por lo tanto, los resultados presentados respaldan las predicciones en favor de la LCP como un mecanismo de lectura programado y funcional del codón de terminación de la traducción (von der Haar y Tuite, 2007; Jungreis et al., 2011; Dunn et al., 2013; Schueren y Thoms, 2016); y en consecuencia, rechazan la afirmación sobre la teoría del error postulada por Bidou *et al.* (2010) y Li *et al.* (2019).

Capítulo 6

ANEXOS

ANEXO A: Técnica de perfil ribosómico para el análisis de transcriptomas

En este anexo se resume el contexto del desarrollo de la técnica de perfil de densidad de ribosomas, conocida en la literatura científica como *Ribosome Profiling* o *Ribo-Seq*. Esta técnica, empleada en el presente trabajo de tesis, identifica en forma precisa las secuencias de ARNm que son leídas por los ribosomas durante el proceso de síntesis de proteínas. Se basa en el análisis de datos de secuenciación de nueva generación (*NGS*), donde la cantidad de cada proteína sintetizada se indica por el número de ribosomas unidos simultáneamente a cada molécula de ARNm.

Secuenciación de nueva generación: *ARN-Seq* y *Ribo-Seq*

Los estudios tradicionales de expresión génica *in vivo* eran realizados con *microarrays* basados en hibridación del ARN aislado de polisomas, así como el perfil de traducción a través de la purificación por afinidad de ribosomas etiquetados con epítopo. No obstante, aunque estos métodos son útiles y complementarios, los *microarrays* pueden incluir inconvenientes como artefactos de hibridación cruzada, mala cuantificación de genes baja/altamente expresados, y la necesidad de conocer la secuencia a *priori*. Debido a estos problemas técnicos, la transcriptómica transicionó hacia métodos basados en la secuenciación a escala genómica; progresando desde el secuenciamiento de Sanger de librerías de etiquetas de secuencias expresadas (*expressed sequence tags, EST*), hasta los métodos químicos basados en etiquetas (por ej., análisis en serie de la expresión génica) (Wang et al., 2009).

Actualmente, un eficiente protocolo es la secuenciación de nueva generación (*Next Generation Sequencing -NGS of cDNA-*), también llamado secuenciación por escopeta *whole transcriptome shotgun sequencing (WTSS)*, empleado para analizar el transcriptoma celular en continuo cambio. En particular, la secuenciación de traducción de ARNm activa (*ARN-Seq*) revela la presencia y cantidad de ARNm en una muestra biológica en un momento determinado; permitiendo identificar los posibles transcriptos codificantes de proteínas, incluso aquellos no anotados. En concreto, facilita la capacidad de analizar transcriptos de genes con empalme alternativo, modificaciones post-transcripcionales, fusiones génicas, mutaciones/SNPs y

cambios en la expresión génica. Además de las transcripciones de ARNm, *ARN-Seq* puede examinar diferentes poblaciones de ARN para incluir el ARN total (ARN pequeño, miARN, ARNt) y analizar perfiles ribosómicos. Esta técnica también sirve para determinar los límites exón/intrón y verificar o corregir los extremos 5'y/o 3' UTR previamente anotados. Así, la secuenciación *NGS* ha permitido avances en la caracterización de genes, transcriptos y su traducción a escala genómica. La aplicación de renovados protocolos de *ARN-Seq*, como el perfil de densidad de ribosomas (*Ribo-Seq*), ha revelado que puede transcribirse hasta el 85 % de un genoma de mamífero, en contraste con el 4 % de los genes anotados como codificantes de proteínas (Mumtaz y Couso, 2015); ampliando así nuestra comprensión del potencial codificador de proteínas de un genoma.

Perfil de densidad ribosomal: *Ribo-Seq* y *Poly-Ribo-Seq*

El perfil de densidad de ribosomas (del inglés: *Ribosome Profiling*), también llamado secuenciación de huellas de ribosomas (*Ribosome Footprinting*) o *Ribo-Seq*, es una técnica desarrollada por Joan Steitz y Marilyn Kozak hace 50 años; cuyo método fue adaptado por Nickolas Ingolia y Jonathan Weissman para trabajar con secuenciación *NGS* (Ingolia et al., 2009). Históricamente, el enfoque de *Ribo-Seq* se basa en el descubrimiento de que el ARNm dentro de un ribosoma puede aislarse mediante el uso de nucleasas que degradan regiones de ARNm no protegidas. Además, involucra la preparación de bibliotecas de secuenciación de alto rendimiento y el análisis de datos en forma similar al *ARN-Seq*. Pero a diferencia de este, que secuencia todo el ARNm presente en una muestra dada, el *Ribo-Seq* sólo captura secuencias de ARNm unidas a ribosomas y protegidas por nucleasas durante el proceso de decodificación de la traducción. El objetivo es determinar la posición de los ribosomas activos en todos los ARNm en una célula; y así determinar cuáles se traducen activamente en un momento particular, lo cual se denomina translatoma (Ingolia, 2014). Una técnica relacionada que también ayuda a determinar qué ARNm son activamente traducidos, es el protocolo de purificación por afinidad del ribosoma de traducción (TRAP) (Heiman et al., 2014). TRAP no implica la huella de ribosoma, pero proporciona información específica del tipo de célula.

En síntesis, *Ribo-Seq* presenta una serie de utilidades para proporcionar información sobre la expresión génica global. Sus principales ventajas son:

- **Análisis de todo el genoma**, sin requerir conocimiento previo del ARN o los marcos de lectura abierta (ORF).

- **Identificación de regiones de ARNm traducidas:** lo que permite corroborar regiones codificantes con resolución de un solo nucleótido (nt), revelando una imagen instantánea con la ubicación precisa de los ribosomas en el ARNm. Además, al usar drogas específicas, el perfil de densidad ribosomal puede identificar regiones iniciadoras de la traducción en ARNm o regiones de elongación (Michel y Baranov, 2013). Incluso pueden observarse sitios de pausa dentro del transcriptoma en codones específicos (Buskirk y Green, 2017; Andreev et al., 2017). Estos sitios de traducción lenta o pausada se demuestran por un aumento en la densidad de ribosomas, y dichas pausas podrían vincular proteínas específicas con sus funciones dentro de la célula

(Ingolia, 2014).

- **Observación del plegado de péptidos nacientes:** el acoplamiento de perfiles ribosomales con el factor de transcripción vinculante ChIP (cromatina-inmunoprecipitación) podría dilucidar cómo y cuándo se pliegan las proteínas recién sintetizadas (Ingolia, 2014). Usando las huellas proporcionadas por *Ribo-Seq (footprints)*, se puede purificar ribosomas específicos asociados con factores como chaperonas; pausar el ribosoma en puntos específicos -permitiéndole traducir un polipéptido a lo largo del tiempo-, y exponer los diferentes puntos a una chaperona. Precipitar utilizando ChIP purifica estas muestras y puede mostrar en qué momento se está plegando el péptido (Ingolia, 2014).

- **Medición de la síntesis proteica y sus reguladores:** *Ribo-Seq* también cuantifica las frecuencias relativas con las que se traducen diferentes regiones de transcriptos, por lo que refleja la tasa de síntesis de proteínas más de cerca que los niveles de ARNm. Esto se logra alterando inicialmente las proteínas que se unen al ARNm y utilizando perfiles ribosomales para medir la diferencia en la traducción (Andreev et al., 2017). Estos ARNm alterados se pueden asociar con proteínas cuyos sitios de unión ya han sido mapeados en el ARN, para indicar la regulación (Ingolia, 2014; Andreev et al., 2017).

Estos usos permiten identificar la ubicación de los sitios de inicio de la traducción para determinar secuencias no-ATG, así como la distribución de los ribosomas en un ARNm y su cinética de traducción (Weiss y Atkins, 2011; Ingolia, 2014; Buskirk y Green, 2017). De hecho, esta técnica ha promovido investigaciones sobre la conexión entre los sesgos de codones y las tasas de traducción; así como de las formas alternativas de *splicing* para proporcionar una cobertura más amplia del transcriptoma. También permite evaluar el potencial de codificación de ORFs en el genoma, especialmente aquellos no anotados o clasificados como regiones no codificantes en el transcriptoma (*ncRNA*), incluidos los extremos 5'y 3' UTR y los ARN largos (*lncRNA*) (Van Damme et al., 2014; Mumtaz y Couso, 2015). Así, el uso de perfiles ribosómicos ha ampliado el conocimiento de proteomas de diversos organismos (Ingolia et al., 2009, 2011; Chew et al., 2013; Bazzini et al., 2014; Duncan y Mata, 2014; Smith et al., 2014), permitiendo nuevos análisis de traducción a nivel de genoma completo y la identificación de eventos programados y generalizados de lectura de codones de parada (LCP) en eucariotas (Dunn et al., 2013), no predichos por métodos filogenéticos.

No obstante, la aplicación del perfil ribosómico fuera de secuencias traducidas canónicamente puede conducir a conclusiones diferentes (Chew et al., 2013; Guttman et al., 2013). En este sentido, las desventajas más importantes de este método son que la iniciación desde varios sitios dentro de un solo transcripto dificulta la definición de todos los ORFs, y que no proporciona la cinética de la elongación de la traducción. El problema radica en que la huella ribosómica no siempre infiere a la traducción, ya que entre ellas pueden encontrarse enlaces ribosoma-ARNm no productivos, así como subunidades ribosomales 40S; los cuales no conducen a una traducción productiva (Aspden et al., 2014; Wilson y Masel, 2011). Por este motivo, a fin de distinguir eventos genuinos de traducción y descartar falsos positivos,

éste método ha sido refinado en un protocolo alternativo denominado *Poly-Ribo-Seq* para mejorar la naturaleza cualitativa y cuantitativa de los resultados. Este último incluye una mejora en la base bioquímica del *Ribo-Seq* para el estudio de sORFs y ORFs canónicos más largos, que en lugar de realizar perfiles de todos los ARNm unidos a ribosomas, los efectúa sobre fracciones polisomales. De este modo, los ARNm unidos a ribosomas múltiples y, por lo tanto, traducidos activamente, pueden ser aislados y distinguirse de los ARNm unidos a ribosomas simples y esporádicos, o bien de subunidades ribosomales; supuestamente no productivos (Aspden et al., 2014). Cabe destacar que el perfil ribosómico es diferente del perfil de polisomas, y aunque ambos estudian la asociación de ARNm a ribosomas para análisis de translatomas, los datos que generan tienen niveles de especificidad muy diferentes.

Procedimiento para la elaboración de perfiles ribosomales:

En genética y bioquímica, la secuenciación significa determinar la estructura primaria de un biopolímero no ramificado. La secuenciación da como resultado una representación simbólica lineal, conocida como una secuencia que resume sucintamente gran parte de la estructura a nivel atómico de la molécula secuenciada (Ingolia et al., 2009). El perfil ribosómico (*Ribo-Seq*), aísla el ARN que está procesando el ribosoma para monitorear el proceso de traducción. A continuación, se describe el procedimiento para la elaboración de perfiles de densidad ribosomal, y se esquematiza en la Fig. 6.1.

Procedimiento:

- Lisar las células o tejidos y aislar las moléculas de ARNm unidos a los ribosomas (complejos de ARNm-ribosomas).
- Inmovilizar complejos. Comúnmente esto es realizado con la enzima cicloheximida, pero se pueden emplear otros productos químicos. La adición de harringtonina (un alcaloide que inhibe la síntesis de proteínas) hace que los ribosomas se acumulen con precisión en los codones de inicio y ayuda a su detección. También es posible prescindir de los inhibidores de traducción en condiciones de lisis incompetentes para la traducción.
- Usando enzimas ribonucleasas, digerir el ARN no protegido por ribosomas.
- Aislar los complejos de ARNm-ribosoma utilizando centrifugación de gradiente de densidad de sacarosa, o columnas cromatográficas especializadas.
- Purificar la mezcla con fenol/cloroformo para eliminar proteínas.
- Seleccionar el tamaño de los fragmentos de ARNm previamente protegidos.
- Ligar el adaptador 3á los fragmentos.
- Restar contaminantes conocidos de ARNr (opcional).
- Transcripción inversa de ARN a ADNc utilizando la enzima transcriptasa inversa.
- Amplificar las hebras de manera específica, utilizando dNTPs (desoxi-Nucleósidos trifosfatos).

- Obtener las secuencias de lectura (*Sequence reads*).
- Método de secuenciación - Biblioteca de ADNc. Alinear la secuencia resultante con la secuencia genómica para determinar el perfil traduccional.

Anexo B: Descripción de softwares y formatos de archivo utilizados

Este apartado hace una revisión de una serie de herramientas computacionales del contexto de la secuenciación de nueva generación, que hemos utilizado para desarrollar este trabajo de tesis. Se resumen las características y mecanismos de operación de los programas empleados, entre los que podemos destacar herramientas para indexar genomas, búsquedas de alineamientos en genomas, herramientas para realizar análisis transcriptómicos cualitativos y/o cuantitativos.

Anexo B.1: Formatos de archivo

B.1.1 SAM

SAM son las siglas del formato de archivo *Sequence Alignment/Map* (Li et al., 2009), actualmente el más utilizado para almacenar datos de alineación. Es un formato de alineación genérico para almacenar alineaciones de lectura contra secuencias de referencia, que admite lecturas cortas y largas (hasta 128 Mbp) producidas por diferentes plataformas de secuenciación. Es de estilo flexible, tamaño compacto, eficiente en acceso aleatorio y es el formato en el que se publican las alineaciones del Proyecto 1000 Genomas. Ya sea re-secuenciación, transcriptoma o epítome, casi todos los procesos generarán archivos SAM/BAM como paso intermedio, seguido de un proceso de análisis especial. Se trata de un formato delimitado por tabuladores, consistente en una sección de encabezado (opcional) y una sección de alineamientos. De estar presente, el encabezado (header) se ubica antes de los alineamientos. Las líneas de encabezado deben comenzar con @, mientras que las líneas de alineamientos no. Cada línea de alineamiento contiene once campos obligatorios para la información básica del alineamiento, además de un número variable de campos opcionales para información adicional. Como ejemplo, la Fig. 6.2 muestra la posible situación de comparación del genoma leído y de referencia. Las lecturas r001/1 y r001/2 representan datos finales emparejados, r003 es una lectura quimérica y r004 es el resultado de la alineación después de que se interrumpe la secuencia original.

Después de utilizar un software de comparación especial, como BWA, BOWTIE2, etc., el archivo SAM correspondiente es el siguiente, como se muestra en la Fig. 6.3:

Para interpretar el formato SAM, es necesario comprender los siguientes términos y conceptos:

Encabezado del formato SAM: Cada línea del encabezado debe comenzar con el caracter seguido de un codigo formado por dos letras. En el encabezado, cada línea debe estar tabulada excepto las líneas que comiencen por. Con respecto a los campos de cada línea, deben seguir la siguiente estructura: Etiqueta:Valor, donde Etiqueta es una cadena de dos letras que define el contenido y el formato de Valor.

Nº	Campo	Tipo	Exp Regular/Rango	Descripción
1	QNAME	cadena	[!-?A-]1, 255	Nombre de la consulta
2	FLAG	entero	[0, 2 16 - 1]	Bandera de opciones
3	RNAME	cadena	* |[! - () + - <>-][!-]*	Referencia del nombre de la secuencia
4	POS	entero	[0, 2 29 - 1]	Posición de la primera base más a la izquierda
5	MAPQ	entero	[0, 2 8 - 1]	Calidad del Mapeo
6	CIGAR	cadena	* |([0 - 9] + [M IDN SHP X =])+	Cadena CIGAR
7	RNEXT	cadena	* | = |[! () + - <>-][!-]*	Nombre de referencia del siguiente fragmento
8	PNEXT	entero	[0, 2 29 - 1]	Posición el siguiente fragmento
9	TLEN	entero	[-2 29 + 1, 2 29 - 1]	Longitud de la plantilla
10	SEQ	cadena	* |[A - Za - z = .]+	Fragmento de secuencia
11	QUAL	cadena	[!-]+	Calidad de la secuencia

Tabla 6.1: En esta tabla se muestran los campos obligatorios del formato SAM.

Bit	Descripción
0x1	La plantilla tiene múltiples fragmentos de secuenciación
0x2	Cada fragmento se alinea correctamente de acuerdo al alineador
0x4	Fragmento no mapeado/asignado
0x8	Siguiente fragmento en la plantilla no está mapeado
0x10	SEQ está inversamente complementada
0x20	SEQ del siguiente fragmento en la plantilla ha sido invertida
0x40	Primer fragmento en la plantilla
0x80	Último fragmento en la plantilla
0x100	Alineamiento secundario
0x200	No ha pasado los controles de calidad
0x400	Duplicado PCR o duplicado óptico

Tabla 6.2: Bits de la etiqueta FLAG.

Hay que indicar que las Etiquetas con letras minúsculas están reservadas para los usuarios finales.

Campos obligatorios de la sección de alineamientos: La segunda parte de un archivo SAM registra específicamente los resultados de alineación de cada lectura. Cada línea que representa a un alineamiento tiene 11 campos que son obligatorios. Estos campos siempre deben aparecer en el mismo orden, aunque sus valores pueden ser 0 o * (dependiendo del campo) si la información correspondiente no está disponible. La Tabla 6.1 resume los campos obligatorios del formato SAM.

- **QNAME:** Nombre de la consulta. Se considera que las lecturas y/o fragmentos que tienen idéntico QNAME provienen de la misma plantilla. Un QNAME con * indica que la información no está disponible.
- **FLAG:** Bandera de opciones. Cada bit se explica en la Tabla 6.2:
- **RNAME:** Nombre de la secuencia de referencia de la alineación. Si @SQ está presente en el encabezado, RNAME (si no *) debe estar presente en una de las etiquetas SQ-SN. Si RNAME es *, no podemos hacer suposiciones acerca de POS y CIGAR.

Op	BAM	Descripción
M	0	Coincidencia en la alineación (puede ser una coincidencia o no)
I	1	Inserción en la referencia
D	2	Borrado de la referencia
N	3	Se omite la región de la referencia
S	4	Recorte suave (recorta las secuencias presentes en la SEQ)
H	5	Recorte duro (recorta las secuencias no presentes en la SEQ)
P	6	Relleno
=	7	Coincidencia en la secuencia
X	8	Desajuste en la secuencia

Tabla 6.3: Operaciones CIGAR.

- **POS:** Posición de la primera base más a la izquierda de la asignación. La primera base en una secuencia de referencia tiene la coordenada 1. POS vale 0 en el caso de tener una lectura no asignada sin coordenada. Si POS es 0, no podemos hacer suposiciones acerca de RNAME y CIGAR.
- **MAPQ:** Calidad del Mapeo. Es igual a -10 log 10 P r, redondeado al entero más próximo. Un valor de 255 indica que la calidad no está disponible.
- **CIGAR:** Cadena CIGAR. Las operaciones CIGAR se muestran en la Tabla 6.3 (si no está disponible usamos *).

  - **H** sólo puede estar presente en la primera/última operación.
  - **S** sólo puede tener operaciones H entre ellos y el final de la cadena CIGAR.
  - Para alineamientos entre mRNA y genoma, una operación N representa un intrón. Para otro tipo de alineamientos, la interpretación de N no está definida.
  - La suma de las longitudes de las operaciones M/I/S/=/X debe ser igual a la longitud de SEQ.

- **RNEXT:** Referencia al nombre del siguiente fragmento de secuencia en la plantilla. Para el caso del último fragmento, su siguiente sería el primer fragmento de la secuencia. Si está incluida la etiqueta @SQ, RNEXT debe estar presente en alguna de las etiquetas SQ-SN (si no debe aparecer * o =). Este campo viene denotado por * cuando la información no está disponible, y con -çuando RNEXT es idéntico a RNAME. Si este campo no es -z el siguiente fragmento en la plantilla tiene una asignación primaria, este campo es idéntico a RNAME del siguiente fragmento. Si el siguiente fragmento tiene múltiples asignaciones primarias, no podemos asumir nada acerca de RNEXT y PNEXT. Si RNEXT es *, no podemos asumir nada acerca de PNEXT y el bit 0x20.
- **PNEXT:** Posición del siguiente fragmento dentro de la plantilla. Cuando no hay información no disponible aparecerá un 0. Este campo es igual al campo POS del siguiente fragmento. Si PNEXT es 0, no podemos hacer suposiciones acerca de RNEXT y el bit 0x20.
- **TLEN:** Longitud de la plantilla. Si todos los fragmentos se asignan a la misma referencia, TLEN es igual al número de bases desde la base más a la izquierda hasta la más a la derecha.

- **SEQ:** Fragmento de secuencia. Este campo puede valer * cuando la secuencia no se ha almacenado. Si tenemos contenido diferente a *, entonces la longitud de la cadena debe ser igual a la suma de las longitudes de las operaciones M/I/S/=/X en CIGAR. Un signo = indica que la base es idéntica a la base de referencia.
- **QUAL:** Hace referencia a la calidad que tiene cada una de las componentes de la secuencia. Para más información ver la sección que habla del formato FASTQ.

Campos opcionales en la sección de alineamientos: Excepto por la información necesaria en las 11 columnas anteriores, las otras columnas son información adicional personalizada por un software de comparación diferente, que se llama "Etiqueta" (Tag). El formato de la etiqueta es generalmente Etiqueta:Tipo:Valor (Tag:Type:Value), donde la etiqueta es una cadena de dos caracteres de la forma [A-Za-z][A-Za-z0-9]. Cada etiqueta sólo puede aparecer una única vez dentro de una línea de alineamiento. Una etiqueta que esté en letras minúsculas está reservada para usuarios finales. En cada campo opcional, "Tipo" es una letra sola (sensible a mayúsculas y minúsculas) la cual define el formato del "Valor". El siguiente ejemplo, NM:i:4 MD:Z:1C53C22G69G1 AS:i:136 XS:i:0, introduce algunas de las etiquetas más comunes; donde:
- NM: editar distancia
- MD: posiciones / bases no coincidentes
- AS: puntuación de alineación
- BC: secuencia de código de barras (secuencia de código de barras)
- XS: puntuación de alineación subóptima

B.1.2 BAM

El formato BAM (Binary Alignment/Map) es la versión binaria comprimida del formato SAM, y mantiene exactamente la misma información. BAM está comprimido por la biblioteca genérica BGZF (Li et al., 2009). Es una representación compacta e indexable de las secuencias de alineamientos. Muchas herramientas de NGS trabajan con los formatos SAM/BAM, por lo que es un formato importante. Al ser BAM un formato binario, no es posible visualizarlo con un editor de texto. Para poder ver el contenido, debemos de convertir el archivo BAM a un archivo SAM, que sí es posible verlo con un editor de texto plano. El comando para convertir un archivo BAM a un archivo SAM es el siguiente:

./samtools view -S -b -o archivo.bam archivo.sam

Para más información, el formato BAM se detalla en el sitio web `http://genome.ucsc.edu/goldenPath/help/bam.html`.

B.1.3 FASTQ

La extensión de archivo FASTQ es un formato de texto plano para almacenar tanto secuencias biológicas (principalmente de nucleótidos), como los valores de calidad asociados a cada base. Tanto la base como la calidad están codificados con

un único carácter ASCII (*American Standard Code for Information Interchange*). Consiste en una extensión del formato FASTA, desarrollado para almacenar la salida de instrumentos de secuenciación de alto rendimiento.

Formato: generalmente, un archivo FASTQ usa 4 líneas por cada secuencia:

- La primera línea contiene la información de la secuencia, comienza con un caracter "@" (encabezado o *Header*) y le sigue un identificador de secuencia y opcionalmente una descripción (como en el caso del formato FASTA). Si son datos crudos contiene información del secuenciador que identifica esa secuencia y el par de lecturas; si son datos ya procesados en SRA contiene una descripción de la secuencia.
- La línea 2 contiene todo el contenido de la secuencia sin procesar (las bases nucleotídicas).
- La línea 3 comienza con el caracter "+", y opcionalmente es seguido por el identificador de la secuencia que encontramos en la línea 1 (*Header*).
- La línea 4 codifica los valores de calidad de secuenciación asociados a cada base de la secuencia de la línea 2. Debe contener el mismo número de símbolos que letras en la secuencia, ya que cada letra o símbolo representa a una base de la secuencia codificado en formato ASCII.

Un ejemplo de archivo FASTQ es el siguiente:

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65

La información de calidad se codifica en ASCII porque esto permite codificar un valor de calidad en un solo caracter. Por lo tanto las líneas 2 y 4 tienen el mismo número de caracteres. El byte que representa la calidad va de 0x21 (calidad más baja, "!" en ASCII) a 0x7e (calidad más alta, " " en ASCII). Para más información acceda al sitio web `http://www.phrap.com/phred/`.

B.1.4 GFF

El formato de archivo GFF o "Formato de características generales" (del inglés *General/Genomic Feature Format*) sirve para representar características genéticas con una función anotada (gen, ARNm, ARNr, ARNt, etc). En este formato, cada línea está asociada a una característica y consta de nueve campos obligatorios (columnas) separados por tabulaciones, más líneas de definición de pista opcionales. Si estos campos se separan por espacios en blanco, en vez de tabuladores, pueden aparecer errores a la hora de mostrar la información. Todos los campos excepto el final de cada línea deben contener un valor, las columnas "vacías" deben indicarse con un "."
Para más información acerca del formato GFF, acceder al sitio oficial `http://www.sanger.ac.uk/resources/software/gff`.

A continuación se incluye una breve descripción de los campos del formato GFF:

< *seqname* > El nombre de la secuencia o cromosoma. Es importante que este campo sea uno que se use dentro de Ensembl, es decir, un nombre de cromosoma estándar o un identificador de Ensembl como un ID de andamio, sin ningún contenido adicional como especies o ensamblaje.
< *source* > Nombre del programa que generó dicha información, o la fuente de datos (base de datos o nombre del proyecto).
< *feature* > El nombre del tipo de característica o función (por ej., gen, variación, similitud). Algunos ejemplos de características estándar son "CDS", "start codon", "stop codon" y "exón".
< *start* > Posición de inicio en la secuencia. La primera base se numera como 1.
< *stop* > Última posición en la secuencia.
< *score* > Puntuación entre 0 y 1000, un valor de punto flotante. Hace referencia a la tonalidad de gris que se usará para mostrar la información.
< *strand* > Orientación de la hebra: "+" (*forward*) o "-" (*reverse*).
< *frame* > Si la característica es un "exón", *frame* será un número entre 0 y 2 que representa el marco de lectura de la primera base: "0" indica que la primera base de una característica es la primera base de un codón, "1" indica que la segunda base es la primera base de un codón, y así sucesivamente. En caso contrario el valor deberá ser "."
< *group* > Todas las líneas de un mismo grupo son enlazadas juntas en un mismo elemento.

Aquí vemos un ejemplo de un fragmento de archivo GFF:

track name=regulatory description=TeleGene(tm) Regulatory Regions chr22 TeleGene enhancer 1000000 1001000 500 + . touch1 chr22 TeleGene promoter 1010000 1010100 900 + . touch1 chr22 TeleGene promoter 1020000 1020000 800 - . touch2

B.1.5 GTF

Las siglas del formato de archivo GTF significan "Formato de Transferencia de Genes" (del inglés *Gene Transfer Format*). Toma prestadas características de GFF, pero añade una estructura adicional que garantiza una definición por separado y un formato de nombre. La estructura es como un archivo GFF, por lo que los campos son:

<seqname><source><feature><start><end><score><strand><frame>[attributes] [comments]

Aquí vemos un ejemplo simple con 3 exones traducidos, donde el orden de las filas no es relevante:

381 Twinscan CDS 380 401 . + 0 gene_id "001"; transcript_id "001.1";
381 Twinscan CDS 501 650 . + 2 gene_id "001"; transcript_id "001.1";
381 Twinscan CDS 700 707 . + 2 gene_id "001"; transcript_id "001.1";
381 Twinscan start_codon 380 382 . + 0 gene_id "001"; transcript_id "001.1";

381 Twinscan stop_codon 708 710 . + 0 gene_id “001”; transcript_id “001.1”;

Los espacios en blanco en el ejemplo son para facilitar la lectura. En GTF, los campos deben estar separados por tabuladores, no por espacios en blanco.
A continuación, se incluye una definición más detallada de cada uno de los campos de un registro dentro de un archivo GTF:

< *seqname* > El nombre de la secuencia. Por lo general, este es el Id del cromosoma o el Id del contig.
< *source* > Source debe ser una columna con una etiqueta única que indique de dónde vienen las anotaciones (por lo general de un programa de predicción o una base de datos pública).
< *feature* > Se requieren alguna de las siguientes características: “CDS”, “start codon” o “stop codon”. Las opciones “5 UTR”, “3 UTR”, “inter”, “inter CNS”, “intron CNS” y “exon” son opcionales. Todas las demás serán ignoradas. Para más informacion acerca de estas características visite el sitio oficial de GTF: `http://mblab.wustl.edu/GTF22.html`.
< *start* >< *end* > Entero que representa las coordenadas de inicio y fin relativas a la secuencia nombrada por <seqname>. <start>debe ser menor o igual que <end>. Los valores de <start>y <end>que se extiendan fuera de la secuencia de referencia son técnicamente aceptables, pero se desalientan.
< *score* > El campo puntuación indica el grado de confianza en la existencia de la característica así como en las coordenadas. El valor de este campo no tiene escala global, pero puede tener importancia relativa cuando el campo <source>indica la predicción del programa usado para crear dicha anotación. Puede ser un número en punto flotante o entero, y no es necesario, puede ser reemplazado por un punto.
< *strand* > Contenido de la hebra.
< *frame* > Un 0 indica que la característica comienza con un codón en la primera base del sentido 5’. Un 1 significa que hay una base extra (la tercera base de un codón) antes del primer codón entero, y un 2 significa que hay dos bases extras (las bases segunda y tercera del codón) antes del primer codón. Tenga en cuenta que para la hebra inversa, la base más cerca de la 5és la coordenada <end>.
atributes Una lista separada por punto y coma de pares etiqueta-valor, que proporciona información adicional sobre cada característica. Los nueve campos tienen los mismos dos atributos obligatorios al final del registro:
- gene id value: Identificador único que identifica al gen. Si está vacío, no hay gen asociado al registro.
- transcript id value: Identificador único que identifica al transcrito. Si está vacío, no hay transcrito asociado al registro.
comments Los comentarios comienzan con una almohadilla (“#”) y continúan hasta el final de la línea. Pueden aparecer en cualquier parte del archivo, incluyendo el final de un registro.

Anexo B.2: Softwares utilizados

B.2.1 SRA toolkit

El archivo de lectura de secuencias SRA (del inglés: *Sequence Read Archive* -anteriormente conocido como *Short Read Archive*-) es una base de datos bioin-

formática que proporciona un repositorio público de datos de secuenciación de ADN, especialmente las "lecturas cortas" generadas por la secuenciación de alto rendimiento, que suelen tener menos de 1000 pares de bases en longitud. Se trata de un formato de intercambio definido por NCBI (*National Center for Biotechnology Information*, `https://www.ncbi.nlm.nih.gov/`) para datos de secuenciación de nueva generación (*NGS*). Para NCBI, la sigla SRA significa "Secuencia de Referencia" (*RefSeqGene, RefSeq*). Cada experimento de metadata SRA (número SRX de acceso a SRA) es un resultado de secuenciación único para una muestra específica. A diferencia del repositorio público GEO (*Gene Expression Omnibus*, `https://www.ncbi.nlm.nih.gov/geo/`), que contiene archivos de datos de secuencia procesados, el formato SRA contiene datos de secuencias sin procesar de tecnologías NGS, incluidas 454, *IonTorrent*, *Illumina*, SOLiD, *Helicos* y *Complete Genomics*. Además de los datos de secuencias sin procesar, SRA almacena información de alineación en forma de ubicaciones leídas en una secuencia de referencia. Por este motivo, en caso de enviar datos a NCBI, se requiere convertir cualquier formato en el que estén (fastq, bam, etc.) a formato SRA usando una de las herramientas de carga. Luego, cualquier persona puede descargar los datos de NCBI y extraerlos en uno de los diferentes formatos que desee (ABI, fasta/qual, fastq). El kit de herramientas SRA (*SRA toolkit*, `http://ncbi.github.io/sra-tools/`) es útil para operar directamente en ejecuciones SRA. Este software consiste en una colección de herramientas y bibliotecas para usar datos en los archivos de lectura de secuencia de INSDC, y puede descargarse desde el sitio `https://www.ncbi.nlm.nih.gov/sra`. Además de SRA, los archivos de extensión SDK de NCBI generan herramientas de carga y descarga con sus respectivas bibliotecas para crear nuevas ejecuciones y acceder a las existentes.

B.2.2 SAM tools

SAMtools es una biblioteca y un paquete de software para analizar y manipular alineaciones cortas de lectura de secuencias de ADN en los formatos SAM (*Sequence Alignment/Map*), BAM (*Binary Alignment/Map*) y CRAM (archivo columnar comprimido), alojado en *GitHub* y disponible en http://samtools.sourceforge.net (Li et al., 2009). Estos archivos se generan como salida mediante alineadores de lectura corta como BWA. *SAMtools* proporciona utilidades para el procesamiento de alineaciones, incluida la clasificación, fusión, indexación, extracción de datos y generación de alineaciones en un formato por posición; así como el llamador de variantes y la visualización de alineaciones. Puede convertir desde otros formatos de alineación, ordenar y fusionar alineaciones, eliminar duplicados de PCR, generar información por posición en el formato pileup, llamar a SNP y variantes cortas, y mostrar alineaciones en un formato de texto. *SAMtools* permite trabajar directamente con un archivo BAM comprimido, ya que este formato resulta más eficiente que SAM para el software). Además, dada la complejidad del formato de un archivo SAM/BAM, que contiene lecturas, referencias, alineaciones, información de calidad y anotaciones especificadas por el usuario, *SAMtools* reduce el esfuerzo necesario para usar archivos SAM/BAM al ocultar detalles de bajo nivel. *SAMtools* también puede abrir archivos en servidores FTP o HTTP remotos si el nombre del archivo comienza con "ftp://", "http://", etc. *SAMtools* comprueba el directorio de trabajo actual para el archivo de índice, y puede descargar el índice en caso de ausencia; pero no recupera todo el archivo de alineación a menos que se le solicite.

Cada comando tiene su propia página de manual, el cual puede descargarse desde http://www.htslib.org/doc/samtools.html.

B.2.3 Cutadapt

Los amplicones son fragmentos de ADN formados como producto de eventos de amplificación (natural o artificial), cuyas lecturas comienzan con una secuencia de cebador. Las colas de Poli-A son útiles para extraer ARN de una muestra, pero a menudo no se desea que estén presentes en las lecturas a utilizar, siendo necesaria una limpieza de datos para el procesamiento de una muestra. A su vez, las secuencias de nucleótidos individuales (lecturas) producidas por las máquinas de secuenciación actuales alcanzan longitudes superiores a 100 pb. Al secuenciar ARN, las máquinas inician desde el adaptador ligado al extremo 3' UTR de cada molécula durante la preparación de la biblioteca. En consecuencia, las lecturas resultantes suelen ser más largas que el ARN secuenciado, dado que contienen la secuencia de la molécula de interés y también partes del adaptador 3'. Este adaptador debe eliminarse durante el análisis de dichos datos, antes de realizar un mapeo.

Otras herramientas de mapeo de lectura, como SOAP (versión 1) (Li et al., 2008), MAQ (Li et al., 2008) y *Novoalign* (http://novocraft.com/) tambien recortan adaptadores, pero sólo para uso del respectivo programa. Así, estos programas resultan difíciles de usar o no ofrecen las funciones requeridas, en particular el soporte para datos de espacio de color. Como alternativa fácil de usar, la herramienta de línea de comandos independiente *Cutadapt* (`https://cutadapt.readthedocs.io/en/stable`) (Martin, 2011), ayuda con estas tareas de recorte al encontrar y eliminar las secuencias de adaptadores, cebadores, colas de poli-A y otros tipos de secuencias no deseadas en lecturas de secuenciación de alto rendimiento. Este software admite variedad de formatos de archivo producidos por los secuenciadores, como datos 454 o *Illumina*, y puede recortar correctamente las lecturas del espacio de color (ej., secuenciadores SOLiD).

Entre otras características útiles, *Cutadapt* ofrece dos algoritmos de recorte de adaptadores, y es compatible con FASTQ, FASTA. Produce resultados en formato FASTA o FASTQ, y detecta automáticamente la compresión gzip de los archivos de entrada o salida. *Cutadapt* también puede modificar y filtrar lecturas de varias maneras, dado que las secuencias de adaptador pueden contener caracteres comodín IUPAC. Además, se admiten las lecturas de extremos emparejados e incluso los datos del espacio de color. Si se desea, también permite simplemente demultiplexar los datos de entrada, sin eliminar las secuencias del adaptador.
En cuanto a su implementación, *Cutadapt* está escrito principalmente en Python, pero para mayor velocidad, el algoritmo de alineación se implementa en C como un módulo de extensión de Python. El programa fue desarrollado en Ubuntu Linux, pero probado también en Windows y Mac OS X, y funciona en otras plataformas para las que Python está disponible. El software *Cutadapt*, incluido su código fuente, viene predeterminado con un extenso conjunto de pruebas automatizadas y está disponible para descargar en http://code.google.com/p/cutadapt/, bajo los términos de la licencia del MIT.

La Fig. 6.4 esquematiza el recorte de un adaptador usando *Cutadapt*.

B.2.4 Bowtie

Bowtie es un programa ultrarrápido de mapeo de lecturas cortas contra secuencias largas de referencia, y que posee una gestión de memoria eficiente (http://bowtie-bio.sourceforge.net/index.shtml) (Langmead et al., 2009; Langmead y Salzberg, 2012). Consta de un sistema dirigido a alinear grandes conjuntos de cadenas cortas de ADN, como lecturas de 50 y hasta cientos o miles de caracteres, con genomas relativamente largos (por ej., mamíferos) a un ritmo de 35 millones de lecturas por hora. *Bowtie* indexa el genoma de referencia usando una técnica de data-compression basada en la transformada de Burrows-Wheeler (*"The Burrows-Wheeler transform"*) (Burrows y Wheeler, 1994), para mantener en la memoria una pequeña huella del mismo. Esta estructura de datos de memoria eficiente permite a *Bowtie* buscar lecturas en un genoma de mamífero empleando alrededor de 2 GB de memoria. Por ejemplo, para el caso del genoma humano, el índice suele ocupar unos 2,2 GB para la alineación sin parejas, o 2,9 GB para la alineación de parejas finales o una alineación por espacio de colores (Alemán Ramos, 2011). Para conseguir una mayor velocidad en la alineación se pueden emplear múltiples procesadores en simultáneo. Además, *Bowtie* produce alineaciones en formato estándar de SAM, lo que le permite trabajar con otras herramientas externas que soportan SAM, incluyendo *SAMtools consensus*, SNP e *Indel Callers*.

Bowtie se ejecuta a través de la línea de comandos y es multiplataforma. Los sistemas operativos soportados son: Windows, Mac OS X, Linux, y Solaris. El programa *Bowtie* puede obtenerse del sitio oficial `http://bowtie-bio.sourceforge.net/`, donde pueden descargarse tanto los archivos fuente como los binarios ejecutables para cada plataforma, según las versiones de arquitecturas Intel i386 y x86_64 (para procesadores de 32 y 64 bits respectivamente). *Bowtie* también se usa como base para otras herramientas como son: *TopHat*, un mapeador de empalme *RNA-Seq*; *Cufflinks*, para montaje de isoformas y cuantificación; *Lighter*, para corrección rápida de errores; *Crossbow*, una herramienta computacional de genotipado en la nube para re-secuenciación de datos a gran escala; y *Myrna*, otra herramienta informática en la nube para el cálculo de expresión diferencial de genes en grandes conjuntos de secuencias de ARN. Es necesario aclarar que *Bowtie* no es una herramienta de alineación de propósito general como son *MUMmer*, *BLAST* ó *Vmatch*. El funcionamiento óptimo de *Bowtie* consiste en alinear lecturas cortas en genomas grandes, aunque admite también secuencias de referencia arbitrariamente pequeñas, como amplicones (fragmentos de ADN formados como producto de eventos de amplificación natural o artificial), y lee cadenas de hasta 1024 bases. *Bowtie* está diseñado para ser extremadamente rápido con conjuntos de lecturas cortas donde:

- muchas de las lecturas cortas tienen al menos una alineación, que es válida,
- muchas de las lecturas son de calidad relativamente alta,
- el número de alineaciones reportadas por cada lectura es pequeño (cercano a 1).

La Fig. 6.5 muestra un esquema básico de las tres fases delalgoritmo de *Bowtie*. Si se desea que *Bowtie* soporte *multithreading* (opción "-p"), deberemos tener instalada en el sistema la librería "pthreads". Para compilar *Bowtie* sin "pthreads" (es decir, desactivar "-p") deberemos usar el comando *make* BOWTIE PTHREADS=0. Para compilar *Bowtie* desde los archivos fuente se requiere un entorno similar a GNU

que contenga GCC, GNU Make y otras herramientas básicas para compilar código. En las plataformas Linux y Mac es posible instalar *Bowtie* fácilmente siguiendo las instrucciones de la documentación del proyecto. Para el caso de Windows, desde el proyecto *Bowtie* se recomienda usar MinGW (o Cygwin), que emulan un sistema GNU. En dicho caso es necesario instalar MSYS.

B.2.5 TopHat

TopHat es un programa que trata de alinear secuencias cortas de ARN dentro de un genoma para identificar las uniones de empalme exon-exon (Trapnell et al., 2009). Se basa en la alineación ultra rápida de lecturas cortas de *Bowtie* para su funcionamiento, por lo que su ejecución requiere tener los ejecutables de *Bowtie* en la variable *path* (bowtie, bowtie-inspect, bowtie-build). Otro requisito necesario para el correcto funcionamiento de *TopHat* es tener la versión 2.4 o superior de Python. Además, las alineaciones de salida de *TopHat* usan el formato BAM, por lo que se necesita instalar *SAMtools*. *TopHat* es un software de fuente abierta y gratuito, disponible en `http://tophat.cbcb.umd.edu`. Corre únicamente en Linux y OS X, para hacerlo funcionar en Windows se requiere usar algún tipo de emulador.

Tradicionalmente, el método estándar para determinar la secuencia de genes transcritos ha sido la captura y secuenciamiento del ARNm usando etiquetas de secuencia expresada (expressed sequence tags (ESTs) (Adams et al., 1993), o secuencias de ADN complementario largo (ADNc), usando tecnología convencional de secuenciamiento de Sanger. Durante la última década, el método *RNA-Seq* desarrollado para la secuenciación del ARNm en una célula, aportó importantes ventajas sobre el secuenciamiento tradicional de EST. Al utilizar tecnologías de secuenciamiento de última generación (NGS), el protocolo de *RNA-Seq* puede muestrear el ARNm con menor sesgo, generando millones de fragmentos de secuencia corta en una sola ejecución, comúnmente 25-50 nt versus varios cientos de nucleótidos con tecnologías más antiguas (Trapnell et al., 2009). Estos fragmentos, o "lecturas" (*reads*), aportan mucha más data por experimento, útil como medida directa del nivel de expresión génica. Así, los experimentos de *RNA-Seq* no sólo capturan el transcriptoma, sino que reemplazan las técnicas de *microarrays* y aportan mayor resolución en medidas de expresión a un costo comparable (Marioni et al., 2008). Un paso crítico en los análisis con *RNA-Seq* es el mapeo de lecturas *NGS* al genoma de referencia (en caso de que el transcriptoma esté incompleto), lo cual destaca dos grandes objetivos: 1) la identificación de nuevos transcritos en regiones cubiertas en el mapeo; y 2) la estimación de abundancia de transcriptos a partir de la profundidad de cobertura en el mapeo. Hasta entonces, las estrategias convencionales de mapeo incluían procedimientos de alineación diseñados para localizar lecturas *Illumina* o SOLID en exones conocidos del genoma. Pero debido a que las lecturas de *RNA-Seq* son cortas y suelen expandirse de los límites del exón, parte de ellas no se mapean a la referencia, pudiendo determinar variantes de nuevos empalmes de genes. Sin embargo, los softwares utilizados para alinear los datos de *RNA-Seq* con un genoma se basan en uniones de empalme conocidas y no pueden identificar nuevas uniones. Este problema se resolvió mediante el concatenamiento de exones adyacentes conocidos, y luego creando fragmentos de secuencias sintéticas a partir de los transcriptos empalmados. Aquellas lecturas que no se alinearon al genoma pero que mapearon con estos fragmentos sintéticos representan evidencia de sitios de empalme entre exones conocidos (Trapnell et al., 2009). Incluso pueden detec-

tarse sitios de empalme ab initio mediante la identificación de lecturas que abarcan uniones de exones, pero esto representa rigurosos cambios computacionales, especialmente con lecturas cortas. En este sentido, *TopHat* es un eficiente algoritmo de mapeo de secuencias cortas de transcriptos, diseñado para alinear las lecturas de un experimento *RNA-Seq* con un genoma de referencia sin depender de sitios de empalme conocidos. Es decir, *TopHat* identifica sitios de empalme ab initio mediante el mapeo a gran escala de lecturas *RNA-Seq*; alineando primero las lecturas sin uniones (aquellas contenidas dentro de los exones) al genoma de referencia, a fin de identificar las uniones de corte y empalme exón-exón (Trapnell et al., 2009). Al haber sido construido sobre el programa de mapeo ultra-rápido de lecturas cortas *Bowtie*, el pipeline de *TopHat* es mucho más rápido que los sistemas anteriores: cartografía de casi 2,2 Millones de lecturas por hora de CPU, lo cual es suficiente para procesar experimentos de RNA-Seq en menos de un día en una computadora estándar. A continuación, la Fig. 6.6 ilustra el flujo de trabajo de *TopHat* utilizando *Bowtie*.

TopHat encuentra uniones sin utilizar anotaciones de referencia. En primer lugar, mapea las secuencias cortas de ARN en el genoma; luego identifica exones potenciales, ya que muchas secuencias cortas se alinearán de manera continua en el genoma. Usando este mapeo inicial, *TopHat* construye una base de datos con todas las uniones posibles, y a continuación, mapea cada una de las lecturas con su posible unión para confirmarlas. Actualmente, las máquinas de secuenciación producen lecturas cortas de 100 o más pares de bases. Sin embargo, algunos exones son más cortos que 100 pb, por lo que pueden perderse en el mapeo inicial. Para solucionar este inconveniente, *TopHat* fracciona todas las lecturas de entrada en fragmentos más pequeños, y los mapea de forma independiente. Finalmente, vuelve a unir los segmentos para poder producir las alineaciones. A la hora de generar la base de datos de la uniones, *TopHat* hace uso de tres posibles fuentes de evidencias:

- La primera fuente son los emparejamientos de "islas de cobertura", que se localizan en distintas regiones de la pila de lecturas de la asignación inicial. Las "islas" vecinas son, a menudo, colocadas juntas en el transcriptoma; así, *TopHat*, busca la forma de unirlas con un intrón.

- La segunda fuente sólo es utilizada cuando *TopHat* se ejecuta con lecturas de extremos emparejados. Cuando se lee un par proveniente de diferentes exones de un transcripto, por lo general, se asignan bastante separados en el genoma. Frente a una situación de este tipo, *TopHat* trata de "cerrar" la brecha entre ambos mediante la búsqueda de subsecuencias del genoma que sean çompañeras"de una longitud total igual a la distancia de la brecha. Después, los intrones de la subsecuencia se añadirán a la base de datos que genera *TopHat*.

- La tercera, y más fuerte fuente de evidencias, se produce cuando dos segmentos de la misma lectura se asignan lejos uno de otro, o cuando falla el mapeo de un segmento interno. Con lecturas largas (de más de 75 pares de bases), los intrones "GT-AG", "GC-AG" y "A-AC" se encuentran al principio. Con cadenas cortas, *TopHat* sólo informa alineamientos con los intrones "GT-AG".

B.2.6 Cufflinks

Cufflinks es un software que ensambla alineamientos de lecturas de *RNA-Seq* en transcriptos, calcula estimaciones de su abundancia y prueba la expresión diferencial y la regulación del transcriptoma (`http://cufflinks.cbcb.umd.edu`) (Trapnell et al., 2012). *Cufflinks* se desarrolló originalmente como parte de un esfuerzo de colaboración entre el Laboratorio de Biología Matemática y Computacional, dirigido por Lior Pachter en UC Berkeley, el grupo de genómica computacional de Steven Salzberg en el Instituto de Medicina Genética de la Universidad Johns Hopkins y el laboratorio de Barbara Wold en Caltech. El proyecto ahora es mantenido por el laboratorio de Cole Trapnell en la Universidad de Washington. *Cufflinks* es proporcionado bajo la licencia Boost aprobada por OSI. Se ejecuta tanto en máquinas con sistema operativo Linux como Mac OS X. Está implementado en C++ y hace un uso substancial de las librerías Boost, así como de la librería de gráficos LEMON, que fue lanzada por *Egervári Research Group on Combinational Optimization (EGRES)*.

La secuenciación de ADNc de alto rendimiento (*RNA-Seq*) puede revelar nuevos genes y variantes de empalme, y ayuda a descubrir transcriptos en simultáneo y su estimación de abundancia, cuantificando así la expresión de todo un genoma en un solo ensayo. Sin embargo, esto requeriría algoritmos que no estén restringidos por anotaciones genéticas previas y que tengan en cuenta la transcripción y el empalme alternativos, tales como el software *Cufflinks*. Este programa ayuda a comprender la flexibilidad regulatoria y la complejidad de los transcriptomas, mejorando la anotación genómica (Trapnell et al., 2010).

Para ensamblar los transcriptos, *Cufflinks* construye un pequeño conjunto de transcritos a partir de las lecturas observadas en un experimento de secuencias de ARN. Esto lo realiza reduciendo el problema de ensamblaje a un problema de emparejamiento máximo en grafos bipartitos. En esencia, *Cufflinks* implementa una demostración constructiva del Teorema de Dilworth (Dilworth, 1950) mediante la construcción de una relación de cobertura en las lecturas a alinear, y encontrando un camino de cobertura mínimo en el grafo acíclico dirigido de la relación. Aunque *Cufflinks* trabaja muy bien con lecturas de ARN desemparejadas, está diseñado para trabajar con lecturas emparejadas. El algoritmo de ensamblaje se encarga explícitamente de manejar lecturas de extremos emparejados mediante el tratamiento de un par determinado como un objeto único en la relación de cobertura. La demostración del Teorema de Dilworth indica que se encuentra una cardinalidad máxima en el grafo bipartito del cierre transitivo del DAG (Grafo Dirigido Acíclico). Sin embargo, no necesariamente es única la máxima cardinalidad encontrada, lo que refleja el hecho de que debido al tamaño limitado de los fragmentos de ADNc, no podemos saber con certeza que los resultados de los eventos de *splicing* alternativo van de la mano de los mismos transcritos (Alemán Ramos, 2011). *Cufflinks* trata de encontrar el conjunto correcto de transcriptos mediante la realización de un emparejamiento de coste mínimo. El costo de asociar los eventos de splicing está basado en la puntuación de "porcentaje de uniones" introducido en Wang et al. (2008). Este emparejamiento es ampliado a un camino mínimo de cobertura del DAG, cada camino representa un transcrito diferente. El algoritmo se basa en las ideas que hay detrás del algoritmo *ShoRAH* para la estimación de la abundancia de halotipos en las poblaciones virales (Eriksson et al., 2008). El ensamblador también toma prestadas algunas ideas introducidas en el algoritmo PASA para anotar genomas de *EST* y

plena evidencia de la longitud del ARNm, descrito en Haas et al. (2003).

Herramientas del paquete *Cufflinks*:

Cufflinks reúne transcriptos, estima su abundancia y prueba la expresión y regulación diferencial en muestras de *RNA-Seq*. Acepta las lecturas alineadas de *RNA-Seq* y ensambla las alineaciones en un parsimonioso conjunto de transcriptos. Luego, *Cufflinks* estima la abundancia relativa de estos transcriptos en función de cuántas lecturas son compatibles con cada uno, teniendo en cuenta los sesgos en los protocolos de preparación de la biblioteca. El paquete de herramientas de *Cufflinks* incluye diversos componentes que informan el resultado en varios formatos de archivo diferentes. Si bien algunos de estos son estándar, otros son únicos, y se describen a continuación:

Cufflinks: *Cufflinks* es tanto el nombre de un conjunto de herramientas como un programa dentro de ese conjunto. El programa ensambla transcriptomas de datos *RNA-Seq* y cuantifica su expresión.

Cuffcompare: Después de ensamblar un transcriptoma a partir de una o más muestras, probablemente querrá comparar su ensamblaje con las transcripciones conocidas. Incluso si no hay un transcriptoma de "referencia" para el organismo que está estudiando, es posible que desee comparar los transcriptomas ensamblados a partir de diferentes bibliotecas de *RNA-Seq*. *Cuffcompare* lo ayuda a realizar estas comparaciones y evaluar la calidad de su ensamblaje.

Cuffmerge: Cuando tiene varias bibliotecas *RNA-Seq* y ha ensamblado transcriptomas de cada una de ellas, le recomendamos que combine estos ensamblajes en un transcriptoma maestro. Este paso es necesario para un análisis de expresión diferencial de las nuevas transcripciones que ha ensamblado. *Cuffmerge* realiza este paso de fusión.

Cuffquant: La cuantificación de la expresión de genes y transcripciones en muestras de *RNA-Seq* puede ser computacionalmente costosa. *Cuffquant* le permite calcular los perfiles de expresión de genes y transcripciones y guardar estos perfiles en archivos que puede analizar posteriormente con *Cuffdiff* o *Cuffnorm*. Esto puede ayudarlo a distribuir su carga computacional en un clúster y se recomienda para análisis que involucren más de un puñado de bibliotecas.

Cuffdiff: La comparación de los niveles de expresión de genes y transcripciones en experimentos de *RNA-Seq* es un problema difícil. *Cuffdiff* es una herramienta altamente precisa para realizar estas comparaciones, y puede decirle no solo qué genes están regulados hacia arriba o hacia abajo entre dos o más condiciones, sino también qué genes están empalmados diferencialmente o están experimentando otros tipos de regulación a nivel de isoforma.

Cuffnorm: A veces, todo lo que desea hacer es normalizar los niveles de expresión de un conjunto de bibliotecas *RNA-Seq* para que estén todos en la misma escala, facilitando los análisis posteriores como el agrupamiento. Los niveles de expresión informados por los Gemelos en unidades FPKM generalmente son comparables entre muestras, pero en ciertas situaciones, la aplicación de un nivel adicional de normalización puede eliminar las fuentes de sesgo en los datos. *Cuffnorm* normaliza un conjunto de muestras para que estén en escalas tan similares como sea posible, lo que puede mejorar los resultados que obtiene con otras herramientas posteriores.

Flujo de trabajo de *Cufflinks*:

El conjunto de herramientas *Cufflinks* se puede usar para realizar varios tipos de análisis diferentes para los experimentos *RNA-Seq*. La suite de *Cufflinks* incluye una serie de programas diferentes que trabajan juntos para realizar estos análisis. El flujo de trabajo completo, que realiza todos los tipos de análisis que puede ejecutar *Cufflinks*, se resume en el gráfico de la Fig. 6.7, extraído del manual de uso del programa, disponible en el sitio web oficial: http://cole-trapnell-lab.github.io/cufflinks/manual/. El lado izquierdo ilustra el flujo de trabajo clásico de *RNA-Seq*, que incluye el mapeo de lectura con TopHat, el ensamblaje con *Cufflinks* y la visualización y exploración de los resultados con CummeRbund. Se introdujo un nuevo worfklow más avanzado con la versión 2.2.0 de *Cufflinks*, y se muestra a la derecha. Ambos siguen siendo compatibles. Para mayor detalle sobre el flujo de trabajo clásico se recomienda leer el documento de protocolo.

B.2.7 SearchGui y Peptide-Shaker

La identificación y cuantificación de proteínas por espectrometría de masa es una técnica estándar en el campo de la proteómica, que se basa en los motores de búsqueda para realizar las identificaciones de los espectros adquiridos (Vaudel et al., 2011). Un paso vital consiste en analizar espectros de masa generados experimentalmente para identificar las secuencias peptídicas subyacentes para su posterior mapeo a las proteínas de origen (Barsnes y Vaudel, 2018). En este sentido, existe una interfaz de usuario gráfica fácil de usar, liviana y de código abierto llamada *SearchGUI* (`http://searchgui.googlecode.com`) (Vaudel et al., 2011), para configurar y ejecutar los motores de búsqueda OMSSA y X!Tandem disponibles simultáneamente; así como para toda búsqueda e identificación en proteómica y los motores de novo más frecuentemente utilizados. La interfaz de la línea de comandos para *SearchGUI*, conocida como *SearchCLI*, hace posible ejecutar todos los motores de búsqueda y algoritmos de novo compatibles con *SearchGUI* utilizando una sola línea de comandos. *SearchCLI* busca archivos de espectro de acuerdo con los parámetros de búsqueda usando X!Tandem, MS-GF+, MS Amanda, MyriMatch, Comet, Tide, Andromeda, OMSSA, Novor y DirecTag. Tenga en cuenta que los espectros deben proporcionarse en el formato de Archivo Genérico de Mascot (mgf). Para la conversión de archivos de espectro, se recomienda usar *msConvert*, parte de *ProteoWizard*. *SearchGUI* se ha convertido en un componente central en numerosos flujos de trabajo de bioinformática (Barsnes y Vaudel, 2018). Disponible de forma gratuita con la licencia Apache2 permitida, *SearchGUI* es compatible con Windows, Linux y OS Vaudel2011.

La proteómica de shotgun se basa en la asignación de un gran número de espectros de péptidos teóricos derivados de una base de datos de secuencias. Se han desarrollado varios motores de búsqueda para esta tarea, cada uno con sus propias ventajas e inconvenientes. La búsqueda llevada a cabo mediante comparación de péptidos en concordancia de espectro, genera dos archivos que contienen los péptidos concordados por OMSSA y X!Tandem para cada espectro, llamados péptidos con comparaciones de espectro (PSMs). A partir de estos, necesitamos encontrar los péptidos y proteínas identificados. Esa es la tarea de *Peptide-Shaker* (`http://peptide-shaker.googlecode.com`) (Vaudel et al., 2015), disponible también en forma gratuita. *Peptide-Shaker* consiste en un motor de búsqueda de plataforma

independiente, recomendado para la visualización, análisis e identificación de péptidos y proteínas resultantes de múltiples motores de búsqueda. En *Peptide-Shaker*, las proteínas se marcan con cuatro colores: verde, amarillo, naranja y rojo. Es necesario tener en cuenta que las proteínas en un grupo proteico se pueden unir a diferentes genes o cromosomas. La elección de uno u otro puede, por lo tanto, impactar fuertemente la interpretación biológica de los resultados. *Peptide-Shaker* siempre elige la proteína con la mejor evidencia según la base de datos *UniProt*; por cuanto se recomienda el uso de éstas para beneficiarse de una anotación de proteínas completa.

La Fig. 6.8 muestra la visualización de resultados de *Peptide-Shaker* para el análisis de péptidos por espectrometría de masa. La base de datos de secuencia se selecciona en SearchGUI y se configuran los parámetros de búsqueda, lo que permite buscar datos con múltiples algoritmos de software de identificación. Los resultados de la búsqueda se procesan, combinan, interpretan y muestran en *Peptide-Shaker*.

B.2.8 FlyEnrichr

El análisis de enriquecimiento es un método popular para analizar conjuntos de genes y/o proteínas generados por experimentos de genoma completo. Los productos génicos expresados diferencialmente en las células de los organismos necesitan ser analizados para conocer sus funciones colectivas, a fin de extraer nuevos conocimientos. Una vez que se generan listas no sesgadas de genes o proteínas a partir de tales experimentos, estas listas se utilizan como entrada para el enriquecimiento informático con listas existentes creadas a partir de conocimientos previos organizadas en bibliotecas de conjuntos de genes. El software *FlyEnrichr* (Chen et al., 2013; Kuleshov et al., 2016) es un conjunto de herramientas para el análisis de enriquecimiento de conjuntos de genes de las moscas de *Drosophila*. Se trata de una aplicación integrada del software interactivo y colaborativo de acceso abierto *Enrichr*, disponible gratuitamente en (`https://amp.pharm.mssm.edu/FlyEnrichr/enrich`). *Enrichr* incluye diversas bibliotecas de conjuntos de genes para organismos modelo, disponibles para su análisis y descarga. Este software también puede integrarse en cualquier herramienta que realice análisis de listas de genes, de modo que constituye un enfoque alternativo para clasificar términos enriquecidos, y varios enfoques de visualización interactiva para mostrar resultados de enriquecimiento utilizando la biblioteca de JavaScript (como se esquematiza en la Fig. 6.9) (Chen et al., 2013). En general, *Enrichr* es un recurso completo para conjuntos de genes seleccionados y un motor de búsqueda que acumula conocimiento biológico para futuros descubrimientos biológicos (Kuleshov et al., 2016).

B.2.9 Mathematica

Wolfram Mathematica es un sistema informático técnico moderno, utilizado en áreas científico-técnicas de ingeniería, matemática y computación; incluidas las redes neuronales, el aprendizaje automático, el procesamiento de imágenes, la geometría, la ciencia de datos, las visualizaciones y otros. Fue concebido en 1988 por Stephen Wolfram y desarrollado por la compañía *Wolfram Research* de Champaign, Illinois (`https://www.wolfram.com`) (Wolfram, 2005). Se trata de un sistema de álgebra computacional, que integra un poderoso lenguaje de programación multipropósito, denominado *Wolfram Language* (Auerbach, 2014). Posee bibliotecas

integradas para varias áreas de la computación técnica que permiten el cálculo simbólico, la manipulación de matrices, las funciones de trazado y varios tipos de datos, la implementación de algoritmos, la creación de interfaces de usuario y la interfaz con programas escritos en otros lenguajes de programación. El software *Mathematica* utiliza la interfaz de cuaderno de Wolfram, el cual le permite organizar todo lo que haga en cuadernos enriquecidos que incluyen texto, código ejecutable, gráficos dinámicos, interfaces de usuario y más. *Mathematica* puede utilizarse como una calculadora convencional, aunque con una importante precisión en el cálculo. Las operaciones se realizan en forma exacta o bien aproximada (con el grado de precisión que queramos).

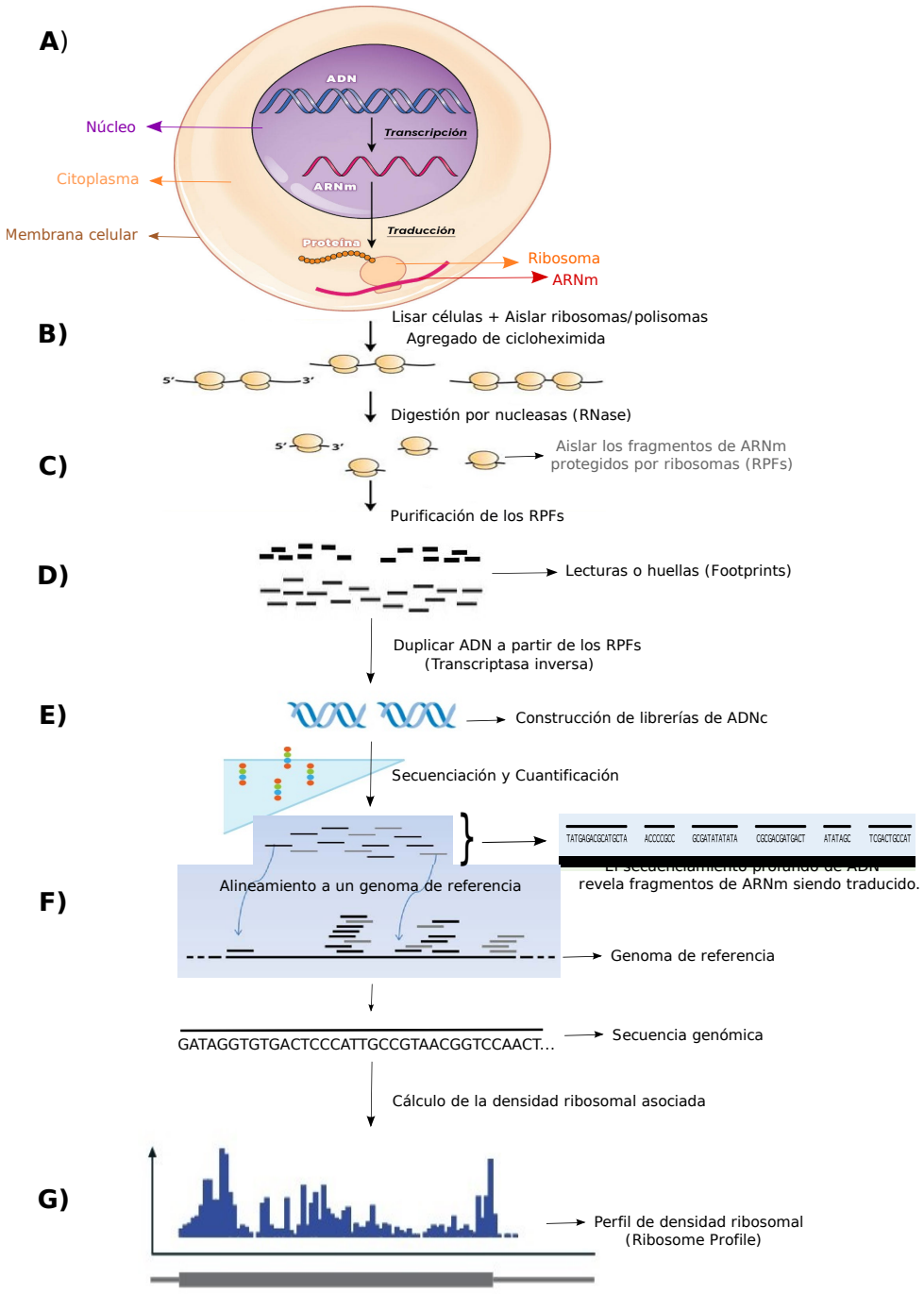

Figura 6.1: Esquema del protocolo de *Ribo-Seq* o *Poly-Ribo-Seq* para la elaboración de perfiles de densidad ribosomal. **A)** Ilustración de los procesos de transcripción del ADN (en el núcleo) y de traducción citoplasmática del ARN a ARNm por parte de los ribosomas. **B)** El lisado de tejido e inhibición de la biosíntesis proteica por medio de la enzima cicloheximida genera extractos celulares en los cuales los ribosomas se han detenido fielmente, a lo largo de cada ARNm que se está traduciendo. **C)** Los ARNm que no están unidos a ribosomas se someten a digestión por enzimas ribonucleasas (*RNase*). **D)** Mediante un proceso de purificación y recuperación, se aíslan los fragmentos de ARNm unidos a ribosomas (*RPFs*), y se descarta el ARNr. **E)** Conversión cuantitativa (transcripción inversa) de los fragmentos de ARNm en una biblioteca/librería de ADNc, que puede analizarse mediante secuenciación de alto rendimiento contra un genoma de referencia. **F)** La secuenciación

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                      CAGCGGCAT
```

Figura 6.2: Ejemplo de lecturas en el archivo de formato .SAM

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Figura 6.3: Ejemplo de archivo SAM, forma de almacenamiento después de la comparación de secuencias.

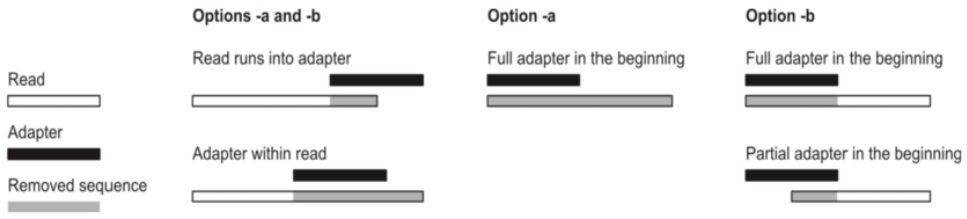

Figura 6.4: Se muestran todas las posibles alineaciones entre el adaptador y la secuencia de lectura. Las opciones "-a" y "-b" indican dos modos de recorte diferentes: 3'para "-a", y ambos 3'y 5'para "-b". (Extraído de Stanchev *et al.* 2016).

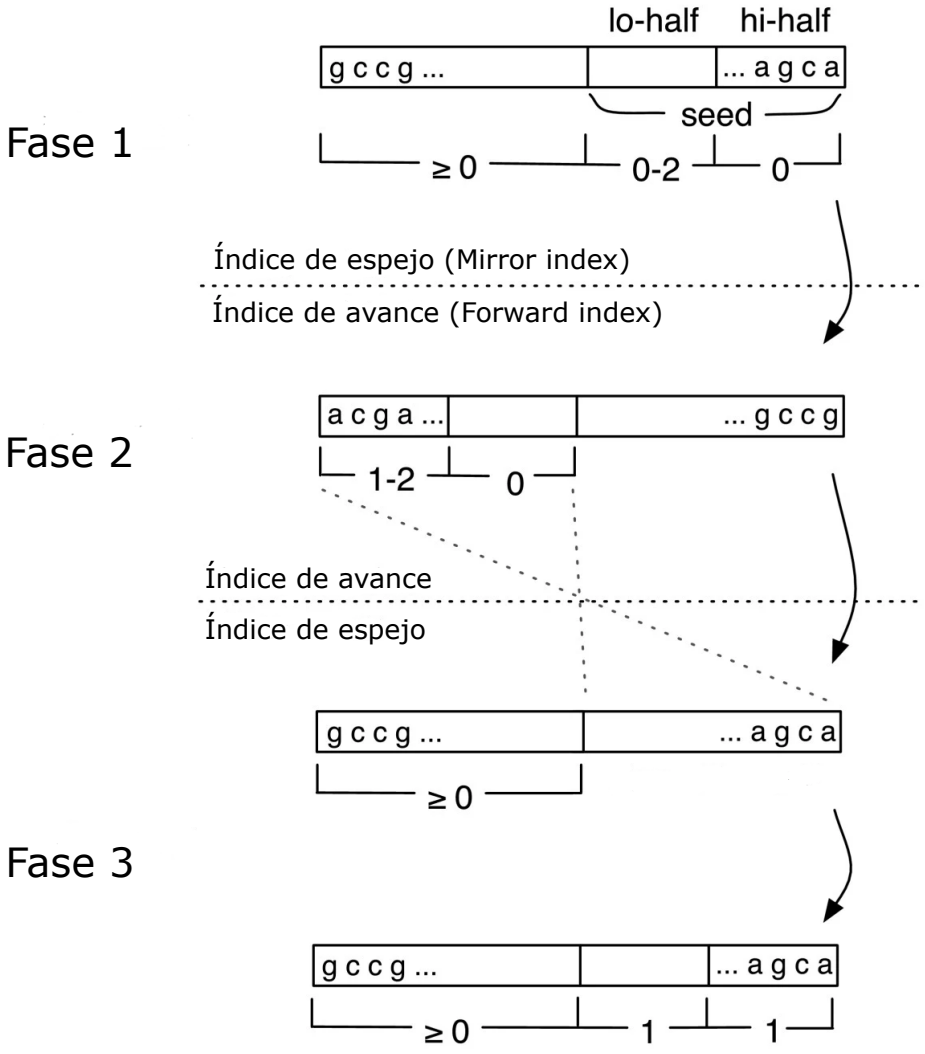

Figura 6.5: Un enfoque de tres fases encuentra alineaciones para los casos de dos desajustes 1 a 4 al tiempo que minimiza el retroceso. La fase 1 usa el índice espejo e invoca el alineador para encontrar alineaciones para los casos 1 y 2. Las fases 2 y 3 cooperan para encontrar alineaciones para el caso 3: la fase 2 encuentra alineaciones parciales con desajustes solo en la mitad superior, y la fase 3 intenta extender esas alineaciones parciales en alineaciones completas. Finalmente, la fase 3 invoca al alineador para encontrar alineaciones para el caso 4. (Extraído de Langmead *et al.* 2009).

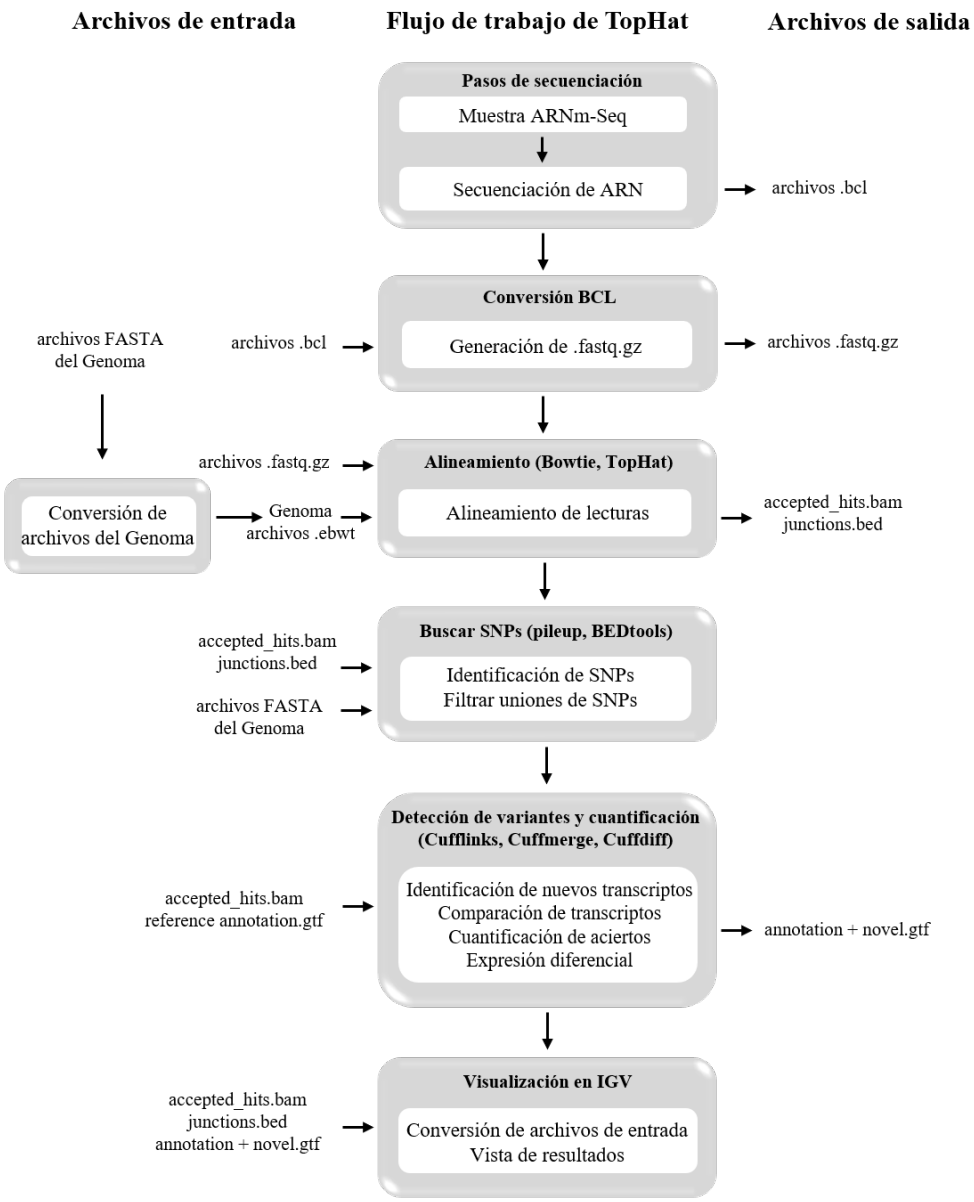

Figura 6.6: Esquema del pipeline para el análisis de *RNA-Seq* con *TopHat* y componentes de software utilizados en este protocolo. *Bowtie* forma el núcleo algorítmico de *TopHat*, que alinea millones de lecturas de secuencia de ARN con el genoma por hora de CPU. Las alineaciones de lectura de *TopHat* son ensambladas por *Cufflinks* y su programa de utilidad asociado para producir una anotación del transcriptoma del genoma. *Cuffdiff* cuantifica este transcriptoma en múltiples condiciones utilizando las alineaciones de lectura de *TopHat*. *CummeRbund* ayuda a los usuarios a explorar y visualizar rápidamente los datos de expresión génica producidos por *Cuffdiff*, incluidos genes y transcritos expresados diferencialmente. (Extraído de Trapnell *et al.* 2009, 2012).

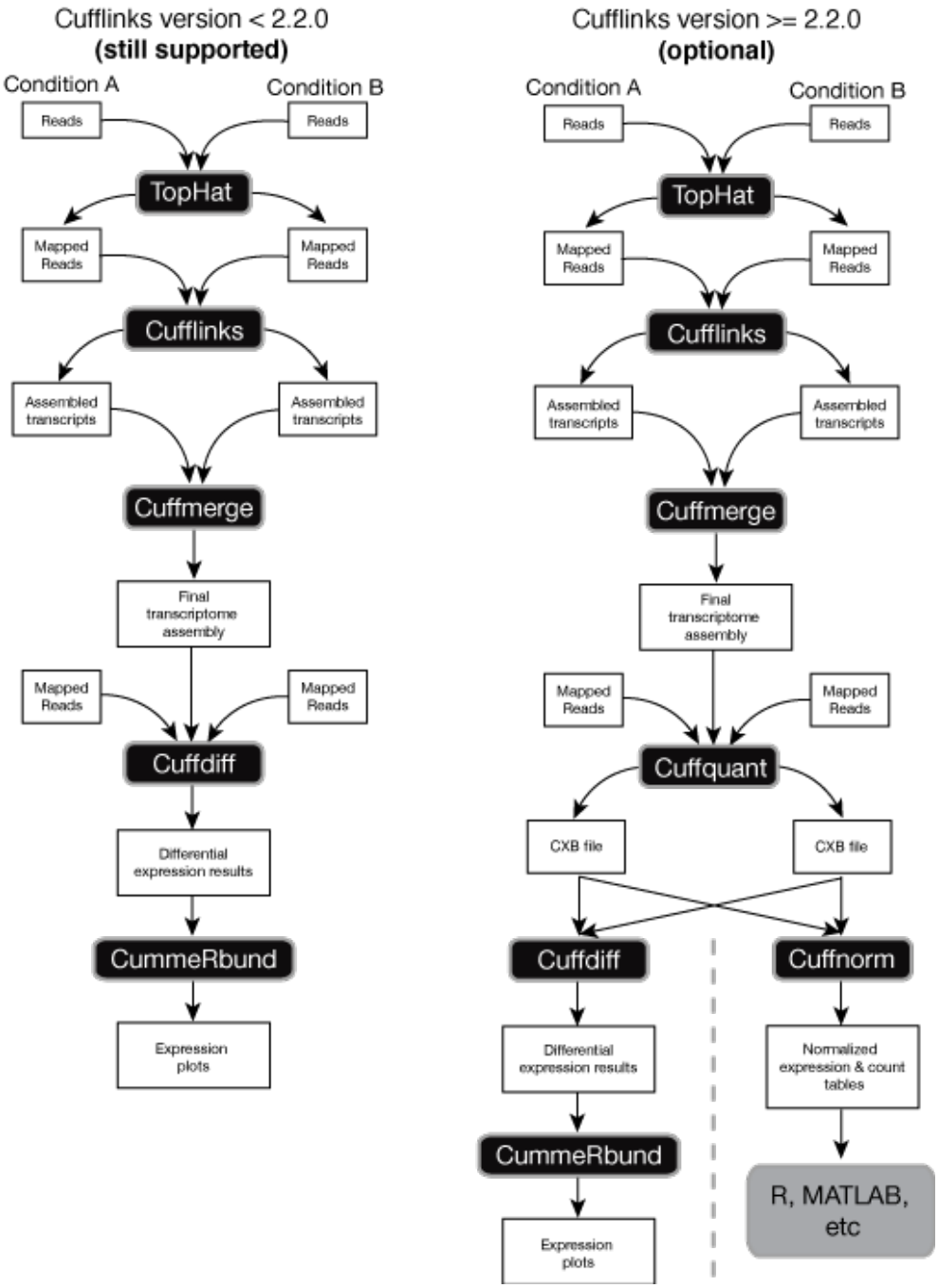

Figura 6.7: Esquema general del *pipeline* para el análisis de *RNA-Seq* con el software *Cufflinks*. Este protocolo comienza con lecturas de RNA-Seq sin procesar y concluye con la visualización lista para la publicación del análisis. En un experimento que involucra dos condiciones, las lecturas se asignan primero al genoma con TopHat. Las lecturas de cada réplica biológica se mapean de forma independiente. Estas lecturas mapeadas se proporcionan como entrada a Cufflinks, que produce un archivo de transfrags ensamblados para cada réplica. Los archivos de ensamblaje se fusionan con la anotación del transcriptoma de referencia en una anotación unificada para su análisis posterior. Esta anotación combinada se cuantifica en cada condición por Cuffdiff, que produce datos de expresión en un conjunto de archivos tabulares. Estos archivos se indexan y visualizan con CummeRbund para facilitar la exploración de genes identificados por Cuffdiff como genes expresados diferencialmente, empalmados o regulados transcripcionalmente. **FPKM:** fragmentos por kilobase de transcripción por millón de fragmentos mapeados. (Extraído

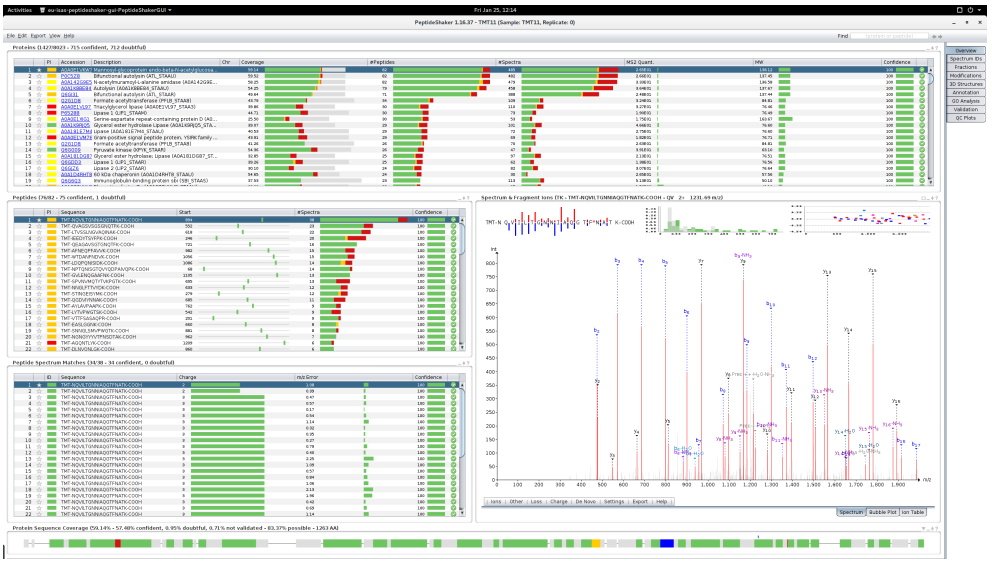

Figura 6.8: Pestaña de descripción general de resultados en *Peptide-Shaker*, que muestra todas las proteínas en el conjunto de datos, junto con los péptidos y las coincidencias de péptido-espectro (PSM) para el péptido/proteína seleccionado. Se muestra una representación gráfica de la coincidencia de espectro seleccionada (abajo a la derecha) con la anotación de fragmentos de iones, y una representación visual de la cobertura de la secuencia de proteínas (abajo). (Extraído de Vaudel *et al.* 2015).

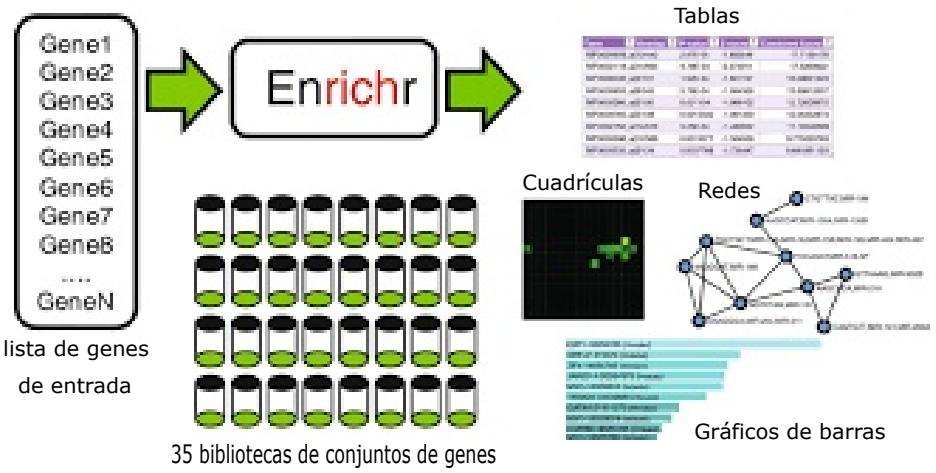

Figura 6.9: Esquema del *pipeline* original de *Enrichr*. El software *Enrichr* recibe listas de genes humanos o de ratón como entrada. Utiliza más de 45 bibliotecas de conjuntos de genes para calcular el enriquecimiento. Los resultados del enriquecimiento se muestran de forma interactiva como gráficos de barras, tablas, cuadrículas de términos con los términos enriquecidos resaltados y redes de términos enriquecidos.(Tomado de Chen et al. (2013) y Kuleshov et al. (2016)).

Bibliografía

Y así, del mucho leer y del poco dormir,
se le secó el celebro de manera que vino
a perder el juicio.

Miguel de Cervantes Saavedra

Adams, M. The Genome Sequence of Drosophila melanogaster. *Science*, vol. 287(5461), páginas 2185–2195, 2000.

Adams, M. D., Soares, M. B., Kerlavage, A. R., Fields, C. y Venter, J. C. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature Genetics*, vol. 4(4), páginas 373–380, 1993.

Alemán Ramos, A. *Análisis Genómico a Través de Herramientas Informáticas Aplicadas a Datos de Secuenciación de Nueva Generación.*. Tesis Doctoral, Universidad de Sevilla, 2011.

Alzugaray, M. E., Bruno, M. C., Villalobos Sambucaro, M. J. y Ronderos, J. R. The Evolutionary History of The Orexin/Allatotropin GPCR Family: from Placozoa and Cnidaria to Vertebrata. *Scientific Reports*, vol. 9(1), página 10217, 2019.

Ambegaokar, S. S., Roy, B. y Jackson, G. R. Neurodegenerative models in Drosophila: Polyglutamine disorders, Parkinson disease, and amyotrophic lateral sclerosis. *Neurobiology of Disease*, vol. 40(1), páginas 29–39, 2010.

Ambros, V. The Functions of Animal microRNAs. *Nature*, vol. 431(7006), páginas 350–355, 2004.

Andreev, D., O'Connor, P., Loughran, G., Dmitriev, S., Baranov, P. y Shatsky, I. Insights Into the mechanisms of Eukaryotic Translation Gained With Ribosome Profiling. *Nucleic Acids Research*, vol. 45(2), páginas 513–526, 2017.

Artieri, C. G. y Fraser, H. B. Evolution at two levels of gene expression in yeast. *Genome Research*, vol. 24(3), páginas 411–421, 2014.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris,

M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. y SHERLOCK, G. Gene Ontology: tool for the unification of biology. *Nature Genetics*, vol. 25(1), páginas 25–29, 2000.

ASHBURNER, M., GOLIC, K. y HAWLEY, R. *Drosophila: a Laboratory Handbook*. Cold Spring Harbor Press; 2nd Edition (12 Dec. 2011), 2005.

ASPDEN, J., EYRE-WALKER, Y., PHILLIPS, R., AMIN, U., MUMTAZ, M., BROCARD, M. y COUSO, J. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife*, vol. 3, página e03528, 2014.

ATKINS, J. F., WILLS, N. M., LOUGHRAN, G., WU, C.-Y., PARSAWAR, K., RYAN, M. D., WANG, C.-H. y NELSON, C. C. A case for "StopGo": Reprogramming translation to augment codon meaning of GGN by promoting unconventional termination (Stop) after addition of glycine and then allowing continued translation (Go). *RNA*, vol. 13(6), páginas 803–810, 2007.

AUERBACH, D. Language Barriers: Wolfram Mathematica Language. Stephen Wolfram's new programming language: Can he make the world computable?. *Slate Magazine*, 2014.

BALAGOPAL, V. y PARKER, R. Polysomes, P bodies and stress granules: states and fates of eukaryotic mRNAs. *Current Opinion in Cell Biology*, vol. 21(3), páginas 403–408, 2009.

BARANOV, P. V., ATKINS, J. F. y YORDANOVA, M. M. Augmented genetic decoding: Global, local and temporal alterations of decoding processes and codon meaning. *Nature Reviews Genetics*, vol. 16(9), páginas 517–529, 2015.

BARANOV, P. V., GESTELAND, R. F. y ATKINS, J. F. Recoding: translational bifurcations in gene expression. *Gene*, vol. 286(2), páginas 187–201, 2002.

BARSNES, H. y VAUDEL, M. SearchGUI: a Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *Journal of Proteome Research*, vol. 17(7), páginas 2552–2555, 2018.

BASRAI, M., HIETER, P. y BOEKE, J. Small Open Reading Frames: Beautiful Needles in the Haystack. *Genome Research*, vol. 7(8), páginas 768–771, 1997.

BAZZINI, A. A., JOHNSTONE, T. G., CHRISTIANO, R., MACKOWIAK, S. D., OBERMAYER, B., FLEMING, E. S., VEJNAR, C. E., LEE, M. T., RAJEWSKY, N., WALTHER, T. C. y GIRALDEZ, A. J. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal*, vol. 33(9), páginas 981–993, 2014.

BEIER, H. Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs. *Nucleic Acids Research*, vol. 29(23), páginas 4767–4782, 2001.

BERGSTROM, D. E., MERLI, C. A., CYGAN, J. A., SHELBY, R. y BLACKMAN, R. K. Regulatory autonomy and molecular characterization of the Drosophila out at first gene. *Genetics*, vol. 139(3), páginas 1331–46, 1995.

BERRETTA, J. y MORILLON, A. Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO reports*, vol. 10(9), páginas 973–982, 2009.

BERRY, M. J., BANU, L. y LARSEN, P. R. Type I iodothyronine deiodinase is a selenocysteine-containing enzyme. *Nature*, vol. 349(6308), páginas 438–440, 1991.

BIDOU, L., ALLAMAND, V., ROUSSET, J.-P. y NAMY, O. Sense from nonsense: therapies for premature stop codon diseases. *Trends in Molecular Medicine*, vol. 18(11), páginas 679–688, 2012. ISSN 14714914.

BIDOU, L., BUGAUD, O., BELAKHOV, V., BAASOV, T. y NAMY, O. Characterization of new-generation aminoglycoside promoting premature termination codon readthrough in cancer cells. *RNA Biology*, vol. 14(3), páginas 378–388, 2017.

BIDOU, L., HATIN, I., PEREZ, N., ALLAMAND, V., PANTHIER, J.-J. y ROUSSET, J.-P. Premature stop codons involved in muscular dystrophies show a broad spectrum of readthrough efficiencies in response to gentamicin treatment. *Gene Therapy*, vol. 11(7), páginas 619–627, 2004.

BIDOU, L., ROUSSET, J.-P. y NAMY, O. Translational errors: from yeast to new therapeutic targets. *FEMS Yeast Research*, vol. 10(8), páginas 1070–1082, 2010.

BIEMANN, K. Contributions of mass spectrometry to peptide and protein structure. *Biomedical & Environmental Mass Spectrometry*, vol. 16(1-12), páginas 99–111, 1988.

BIER, E. Drosophila, the golden bug, emerges as a tool for human genetics. *Nature Reviews Genetics*, vol. 6(1), páginas 9–23, 2005.

BLANCHET, S., CORNU, D., ARGENTINI, M. y NAMY, O. New insights into the incorporation of natural suppressor tRNAs at stop codons in Saccharomyces cerevisiae. *Nucleic Acids Research*, vol. 42(15), páginas 10061–10072, 2014. ISSN 1362-4962.

BLANCHET, S., CORNU, D., HATIN, I., GROSJEAN, H., BERTIN, P. y NAMY, O. Deciphering the reading of the genetic code by near-cognate tRNA. *Proceedings of the National Academy of Sciences*, vol. 115(12), páginas 3018–3023, 2018.

BONETTI, B., FU, L., MOON, J. y BEDWELL, D. M. The Efficiency of Translation Termination is Determined by a Synergistic Interplay Between Upstream and Downstream Sequences inSaccharomyces cerevisiae. *Journal of Molecular Biology*, vol. 251(3), páginas 334–345, 1995. ISSN 00222836.

BRENDEL, C., BELAKHOV, V., WERNER, H., WEGENER, E., GÄRTNER, J., NUDELMAN, I., BAASOV, T. y HUPPKE, P. Readthrough of nonsense mutations in Rett syndrome: evaluation of novel aminoglycosides and generation of a new mouse model. *Journal of Molecular Medicine*, vol. 89(4), páginas 389–398, 2011.

BRENNECKE, J., HIPFNER, D., STARK, A., RUSSELL, R. y COHEN, S. Bantam Encodes a Developmentally Regulated microRNA that Controls Cell Proliferation and Regulates the Proapoptotic Gene hid in Drosophila. *Cell*, vol. 113(1), páginas 25–36, 2003.

BROWN, C. M., DINESH-KUMAR, S. y MILLER, W. A. Local and Distant Sequences Are Required for Efficient Readthrough of the Barley Yellow Dwarf Virus PAV Coat Protein Gene Stop Codon. *Journal of Virology*, vol. 70(9), páginas 5884–5892, 1996.

BRUNNER, E., AHRENS, C. H., MOHANTY, S., BAETSCHMANN, H., LOEVENICH, S., POTTHAST, F., DEUTSCH, E. W., PANSE, C., DE LICHTENBERG, U., RINNER, O., LEE, H., PEDRIOLI, P. G. A., MALMSTROM, J., KOEHLER, K., SCHRIMPF, S., KRIJGSVELD, J., KREGENOW, F., HECK, A. J. R., HAFEN, E., SCHLAPBACH, R. y AEBERSOLD, R. A high-quality catalog of the Drosophila melanogaster proteome. *Nature Biotechnology*, vol. 25(5), páginas 576–583, 2007.

BURGERS, W. A., FUKS, F. y KOUZARIDES, T. DNA methyltransferases get connected to chromatin. *Trends in Genetics*, vol. 18(6), páginas 275–277, 2002.

BURROWS, M. y WHEELER, D. Technical report 124. *Palo Alto, CA: Digital Equipment Corporation*, 1994.

BUSKIRK, A. y GREEN, R. Ribosome Pausing, Arrest and Rescue in Bacteria and Eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372(20160183), 2017.

CAMERON, R. A., MAHAIRAS, G., RAST, J. P., MARTINEZ, P., BIONDI, T. R., SWARTZELL, S., WALLACE, J. C., POUSTKA, A. J., LIVINGSTON, B. T., WRAY, G. A., ETTENSOHN, C. A., LEHRACH, H., BRITTEN, R. J., DAVIDSON, E. H. y HOOD, L. A sea urchin genome project: Sequence scan, virtual map, and additional resources. *Proceedings of the National Academy of Sciences*, vol. 97(17), páginas 9514–9518, 2000.

CAMMARATO, A., AHRENS, C. H., ALAYARI, N. N., QELI, E., RUCKER, J., REEDY, M. C., ZMASEK, C. M., GUCEK, M., COLE, R. N., VAN EYK, J. E., BODMER, R., O'ROURKE, B., BERNSTEIN, S. I. y FOSTER, D. B. A Mighty Small Heart: The Cardiac Proteome of Adult Drosophila melanogaster. *PLoS ONE*, vol. 6(4), página e18497, 2011.

CAMPBELL, M. y FARRELL, S. *BioquÃmica (4ta Edicion)*. Cengage Learning Editores S.A., 2004.

CANTARA, W. A., CRAIN, P. F., ROZENSKI, J., MCCLOSKEY, J. A., HARRIS, K. A., ZHANG, X., VENDEIX, F. A. P., FABRIS, D. y AGRIS, P. F. The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Research*, vol. 39(Database), páginas D195–D201, 2011.

CARTHEW, R. y SONTHEIMER, E. Origins and Mechanisms of miRNAs and siRNAs. *Cell*, vol. 136(4), páginas 642–655, 2009.

CAVAGNARI, B. Regulación de la Expresión Génica: cómo Operan los Mecanismos Epigenéticos. *Archivos Argentinos de Pediatria*, vol. 110(2), páginas 132–136, 2012.

CHAMBERS, I., FRAMPTON, J., GOLDFARB, P., AFFARA, N., MCBAIN, W. y HARRISON, P. R. The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded by the 'termination'codon, TGA. *The EMBO journal*, vol. 5(6), páginas 1221–7, 1986.

CHAN, C. T., PANG, Y. L. J., DENG, W., BABU, I. R., DYAVAIAH, M., BEGLEY, T. J. y DEDON, P. C. Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. *Nature Communications*, vol. 3(1), página 937, 2012.

CHEN, E. Y., TAN, C. M., KOU, Y., DUAN, Q., WANG, Z., MEIRELLES, G., CLARK, N. R. y MAÁYAN, A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, vol. 14(1), página 128, 2013.

CHEW, G.-L., PAULI, A., RINN, J. L., REGEV, A., SCHIER, A. F. y VALEN, E. Ribosome profiling reveals resemblance between long non-coding RNAs and 5'leaders of coding RNAs. *Development (Cambridge, England)*, vol. 140(13), páginas 2828–34, 2013.

CHITTUM, H. S., LANE, W. S., CARLSON, B. A., ROLLER, P. P., LUNG, F.-D. T., LEE, B. J. y HATFIELD, D. L. Rabbit β-Globin Is Extended Beyond Its UGA Stop Codon by Multiple Suppressions and Translational Reading Gaps. *Biochemistry*, vol. 37(31), páginas 10866–10870, 1998.

CHUGUNOVA, A., NAVALAYEU, T., DONTSOVA, O. y SERGIEV, P. Mining for Small Translated ORFs. *Journal of Proteome Research*, vol. 17(1), páginas 1–11, 2018.

CIMINO, P. A., NICHOLSON, B. L., WU, B., XU, W. y WHITE, K. A. Multifaceted Regulation of Translational Readthrough by RNA Replication Elements in a Tombusvirus. *PLoS Pathogens*, vol. 7(12), página e1002423, 2011.

COSTELLO, J. F. Methylation matters. *Journal of Medical Genetics*, vol. 38(5), páginas 285–303, 2001.

CRICK, F., BARNETT, L., BRENNER, S. y WATTS-TOBIN, R. General Nature of the Genetic Code for Proteins. *Nature*, vol. 192(4809), páginas 1227–1232, 1961.

CRICK, F. H. C. Central Dogma of Molecular Biology. *Nature*, vol. 227(5258), páginas 561–563, 1970.

CRIDGE, A. G., CROWE-MCAULIFFE, C., MATHEW, S. F. y TATE, W. P. Eukaryotic translational termination efficiency is influenced by the 3ñucleotides within the ribosomal mRNA channel. *Nucleic Acids Research*, vol. 46(4), páginas 1927–1944, 2018.

DABROWSKI, M., BUKOWY-BIERYLLO, Z. y ZIETKIEWICZ, E. Translational readthrough potential of natural termination codons in eucaryotes–The impact of RNA sequence. *RNA biology*, vol. 12(9), páginas 950–8, 2015.

DABROWSKI, M., BUKOWY-BIERYLLO, Z. y ZIETKIEWICZ, E. Advances in therapeutic use of a drug-stimulated translational readthrough of premature termination codons. *Molecular Medicine*, vol. 24(1), página 25, 2018.

DAVEY, J. C., BECKER, K. B., SCHNEIDER, M. J., GERMAIN, D. L. S. y GALTON, V. A. Cloning of a cDNA for the Type II Iodothyronine Deiodinase. *Journal of Biological Chemistry*, vol. 270(45), páginas 26786–26789, 1995.

DIAMBRA, L. A. y DIAMBRA LA. Differential bicodon usage in lowly and highly abundant proteins. *PeerJ*, vol. 5:e3081, 2017.

DILWORTH, R. A Decomposition Theorem for Partially Ordered Sets. *Annals of Mathematics, Second Series*, vol. 51(1), páginas 161–166, 1950.

DIMMOCK, N. J., EASTON, A. J. y LEPPARD, K. *Introduction to Modern Virology*. John Wiley & Sons, 6th edición, 2007.

DIOP, S. B. y BODMER, R. Drosophila as a model to study the genetic mechanisms of obesity-associated heart dysfunction. *Journal of Cellular and Molecular Medicine*, vol. 16(5), páginas 966–971, 2012.

DUNCAN, C. D. S. y MATA, J. The translational landscape of fission-yeast meiosis and sporulation. *Nature structural & molecular biology*, vol. 21(7), páginas 641–7, 2014.

DUNN, J. G., FOO, C. K., BELLETIER, N. G., GAVIS, E. R. y WEISSMAN, J. S. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife*, vol. 2, 2013.

ERIKSSON, N., PACHTER, L., MITSUYA, Y., RHEE, S., WANG, C., GHARIZADEH, B., RONAGHI, M., SHAFER, R. y BEERENWINKEL, N. Viral Population Estimation Using Pyrosequencing. *PLoS Computational Biology*, vol. 4(5), 2008.

ESWARAPPA, S. M., POTDAR, A. A., KOCH, W. J., FAN, Y., VASU, K., LINDNER, D., WILLARD, B., GRAHAM, L. M., DICORLETO, P. E. y FOX, P. L. Programmed Translational Readthrough Generates Antiangiogenic VEGF-Ax. *Cell*, vol. 157(7), páginas 1605–1618, 2014.

FERNANDEZ-FUNEZ, P., DE MENA, L. y RINCON-LIMAS, D. E. Modeling the complex pathology of Alzheimer's disease in Drosophila. *Experimental Neurology*, vol. 274, páginas 58–71, 2015.

FINKEL, R. S., FLANIGAN, K. M., WONG, B., BÖNNEMANN, C., SAMPSON, J., SWEENEY, H. L., REHA, A., NORTHCUTT, V. J., ELFRING, G., BARTH, J. y PELTZ, S. W. Phase 2a Study of Ataluren-Mediated Dystrophin Production in Patients with Nonsense Mutation Duchenne Muscular Dystrophy. *PLoS ONE*, vol. 8(12), página e81302, 2013.

FIRTH, A. E. y BRIERLEY, I. Non-canonical translation in RNA viruses. *Journal of General Virology*, vol. 93(Pt_7), páginas 1385–1409, 2012.

FIRTH, A. E., WILLS, N. M., GESTELAND, R. F. y ATKINS, J. F. Stimulation of stop codon readthrough: frequent presence of an extended 3'RNA structural element. *Nucleic Acids Research*, vol. 39(15), páginas 6679–6691, 2011.

FISCHER, P. M. Cap in hand: Targeting eIF4E. *Cell Cycle*, vol. 8(16), páginas 2535–2541, 2009.

FLOQUET, C., HATIN, I., ROUSSET, J.-P. y BIDOU, L. Statistical Analysis of Readthrough Levels for Nonsense Mutations in Mammalian Cells Reveals a Major Determinant of Response to Gentamicin. *PLoS Genetics*, vol. 8(3), página e1002608, 2012.

FREITAG, J., AST, J. y BÖLKER, M. Cryptic peroxisomal targeting via alternative splicing and stop codon read-through in fungi. *Nature*, vol. 485(7399), páginas 522–525, 2012.

GALINDO, M., PUEYO, J., FOUIX, S., BISHOP, S. y COUSO, J. Peptides Encoded by Short ORFs Control Development and Define a New Eukaryotic Gene Family. *PLoS Biology*, vol. 5(5), página e106, 2007.

GAROFALO, R., WOHLGEMUTH, I., PEARSON, M., LENZ, C., URLAUB, H. y RODNINA, M. V. Broad range of missense error frequencies in cellular proteins. *Nucleic Acids Research*, vol. 47(6), páginas 2932–2945, 2019.

GESTELAND, R. F. Recoding: Dynamic Reprogramming of Translation. *Annual Review of Biochemistry*, vol. 65(1), páginas 741–768, 1996.

GOLDMANN, T., OVERLACK, N., MÖLLER, F., BELAKHOV, V., VAN WYK, M., BAASOV, T., WOLFRUM, U. y NAGEL-WOLFRUM, K. A comparative evaluation of NB30, NB54 and PTC124 in translational readtrough efficacy for treatment of an USH1C nonsense mutation. *EMBO Molecular Medicine*, vol. 4(11), páginas 1186–1199, 2012.

GRAMATES, L. S., MARYGOLD, S. J., DOS SANTOS, G., URBANO, J.-M., ANTONAZZO, G., MATTHEWS, B. B., REY, A. J., TABONE, C. J., CROSBY, M. A., EMMERT, D. B., FALLS, K., GOODMAN, J. L., HU, Y., PONTING, L., SCHROEDER, A. J., STRELETS, V. B., THURMOND, J. y ZHOU, P. FlyBase at 25: looking to the future. *Nucleic Acids Research*, vol. 45(D1), páginas D663–D671, 2017.

GROPPO, R. y RICHTER, J. D. Translational control from head to tail. *Current Opinion in Cell Biology*, vol. 21(3), páginas 444–451, 2009.

GROSJEAN, H. y WESTHOF, E. An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Research*, vol. 44(17), páginas 8020–8040, 2016.

GUERIN, K., GREGORY-EVANS, C., HODGES, M., MOOSAJEE, M., MACKAY, D., GREGORY-EVANS, K. y FLANNERY, J. G. Systemic aminoglycoside treatment in rodent models of retinitis pigmentosa. *Experimental Eye Research*, vol. 87(3), páginas 197–207, 2008.

GUTTMAN, M., RUSSELL, P., INGOLIA, N. T., WEISSMAN, J. S. y LANDER, E. S. Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell*, vol. 154(1), páginas 240–251, 2013.

VON DER HAAR, T. y TUITE, M. F. Regulated translational bypass of stop codons in yeast. *Trends in Microbiology*, vol. 15(2), páginas 78–86, 2007.

HAAS, B., DELCHER, A., MOUNT, S., WORTMAN, J., SMITH, R., HANNICK, L., MAITI, R., RONNING, C., RUSCH, D., TOWN, C., SALZBERG, S. y WHITE, O. Improving the Arabidopsis Genome Annotation Using Maximal Transcript

Alignment Assemblies. *Nucleic Acids Research*, vol. 31(19), páginas 5654–5666, 2003.

HANADA, K., HIGUCHI-TAKEUCHI, M., OKAMOTO, M., YOSHIZUMI, T., SHIMIZU, M., NAKAMINAMI, K., NISHI, R., OHASHI, C., IIDA, K., TANAKA, M., HORII, Y., KAWASHIMA, M., MATSUI, K., TOYODA, T., SHINOZAKI, K., SEKI, M. y MATSUI, M. Small Open Reading Frames Associated With Morphogenesis are Hidden in Plant Genomes. *Proceedings of the National Academy of Sciences*, vol. 110(6), páginas 2395–2400, 2013.

HANSON, G. y COLLER, J. Codon optimality, bias and usage in translation and mRNA decay. *Nature Reviews Molecular Cell Biology*, vol. 19(1), páginas 20–30, 2018.

HARRELL, L. Predominance of six different hexanucleotide recoding signals 3óf read-through stop codons. *Nucleic Acids Research*, vol. 30(9), páginas 2011–2017, 2002.

HARTL, D. L., LOZOVSKAYA, E. R. ET AL. *The Drosophila genome map: a practical guide.*. 1995.

HEIMAN, M., KULICKE, R., FENSTER, R. J., GREENGARD, P. y HEINTZ, N. Cell type-specific mRNA purification by translating ribosome affinity purification (TRAP). *Nature Protocols*, vol. 9(6), páginas 1282–1291, 2014.

HEMM, M., PAUL, B., SCHNEIDER, T., STORZ, G. y RUDD, K. Small Membrane Proteins Found by Comparative Genomics and Ribosome Binding Site Models. *Molecular Microbiology*, vol. 70(6), páginas 1487–1501, 2008.

HENDRICH, B. Human diseases with underlying defects in chromatin structure and modification. *Human Molecular Genetics*, vol. 10(20), páginas 2233–2242, 2001.

HOWARD, M., FRIZZELL, R. A. y BEDWELL, D. M. Aminoglycoside antibiotics restore CFTR function by overcoming premature stop mutations. *Nature Medicine*, vol. 2(4), páginas 467–469, 1996.

HOWARD, M. T., SHIRTS, B. H., PETROS, L. M., FLANIGAN, K. M., GESTELAND, R. F. y ATKINS, J. F. Sequence specificity of aminoglycoside-induced stop codon readthrough: Potential implications for treatment of Duchenne muscular dystrophy. *Annals of Neurology*, vol. 48(2), páginas 164–169, 2000.

HSUEH-FEN, J. y HSUAN-CHENG, H. *Systems biology: Applications in cancer-related research.*. WORLD SCIENTIFIC, New Jersey, 2012. ISBN 978-981-4324-45-8.

INGOLIA, N. Ribosome Profiling: New Views of Translation, from Single Codons to Genome Scale. *Nature Reviews Genetics*, vol. 15(3), páginas 205–213, 2014.

INGOLIA, N., GHAEMMAGHAMI, S., NEWMAN, J. y WEISSMAN, J. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, vol. 324(5924), páginas 218–223, 2009.

INGOLIA, N. T., LAREAU, L. F. y WEISSMAN, J. S. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell*, vol. 147(4), páginas 789–802, 2011.

JACQUIER, A. The Complex Eukaryotic Transcriptome: Unexpected Pervasive Transcription and Novel Small RNAs. *Nature Reviews. Genetics*, vol. 10(12), páginas 833–844, 2009.

JUNGREIS, I., CHAN, C. S., WATERHOUSE, R. M., FIELDS, G., LIN, M. F. y KELLIS, M. Evolutionary Dynamics of Abundant Stop Codon Readthrough. *Molecular biology and evolution*, vol. 33(12), páginas 3108–3132, 2016.

JUNGREIS, I., LIN, M. F., SPOKONY, R., CHAN, C. S., NEGRE, N., VICTORSEN, A., WHITE, K. P. y KELLIS, M. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. *Genome Research*, vol. 21(12), páginas 2096–2113, 2011.

KAMEI, M., KASPERSKI, K., FULLER, M., PARKINSON-LAWRENCE, E. J., KARAGEORGOS, L., BELAKHOV, V., BAASOV, T., HOPWOOD, J. J. y BROOKS, D. A. Aminoglycoside-Induced Premature Stop Codon Read-Through of Mucopolysaccharidosis Type I Patient Q70X and W402X Mutations in Cultured Cells. 2013.

KASTENMAYER, J. Functional Genomics of Genes With Small Open Reading Frames (sORFs) in S. cerevisiae. *Genome Research*, vol. 16(3), páginas 365–373, 2006.

KAYE, N. M., EMMETT, K. J., MERRICK, W. C. y JANKOWSKY, E. Intrinsic RNA Binding by the Eukaryotic Initiation Factor 4F Depends on a Minimal RNA Length but Not on the m 7 G Cap. *Journal of Biological Chemistry*, vol. 284(26), páginas 17742–17750, 2009.

KEELING, K. Nonsense Suppression as an Approach to Treat Lysosomal Storage Diseases. *Diseases*, vol. 4(4), página 32, 2016.

KEELING, K. M. y BEDWELL, D. M. Suppression of nonsense mutations as a therapeutic approach to treat genetic diseases. *Wiley Interdisciplinary Reviews: RNA*, vol. 2(6), páginas 837–852, 2011.

KEELING, K. M., WANG, D., CONARD, S. E. y BEDWELL, D. M. Suppression of premature termination codons as a therapeutic approach. *Critical Reviews in Biochemistry and Molecular Biology*, vol. 47(5), páginas 444–463, 2012.

KEELING, K. M., XUE, X., GUNN, G. y BEDWELL, D. M. Therapeutics Based on Stop Codon Readthrough. *Annual Review of Genomics and Human Genetics*, vol. 15(1), páginas 371–394, 2014.

KLAGGES, B. R. E., HEIMBECK, G., GODENSCHWEGE, T. A., HOFBAUER, A., PFLUGFELDER, G. O., REIFEGERSTE, R., REISCH, D., SCHAUPP, M., BUCHNER, S. y BUCHNER, E. Invertebrate Synapsins: A Single Gene Codes for Several Isoforms in Drosophila. *The Journal of Neuroscience*, vol. 16(10), páginas 3154–3165, 1996.

KLUG, W., CUMMINGS, M. y SPENCER, C. *Conceptos de GenÃ©tica (8va Edicion)*. Editorial Pearson Educacion S.A. (Madrid), 2006.

KREIDA, S. y TÖRNROTH-HORSEFIELD, S. Structural insights into aquaporin selectivity and regulation. *Current Opinion in Structural Biology*, vol. 33, páginas 126–134, 2015.

KULESHOV, M. V., JONES, M. R., ROUILLARD, A. D., FERNANDEZ, N. F., DUAN, Q., WANG, Z., KOPLEV, S., JENKINS, S. L., JAGODNIK, K. M., LACHMANN, A., MCDERMOTT, M. G., MONTEIRO, C. D., GUNDERSEN, G. W. y MAÁYAN, A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, vol. 44(W1), páginas W90–W97, 2016.

KUNKEL, T. A. y BEBENEK, K. DNA Replication Fidelity. *Annual Review of Biochemistry*, vol. 69(1), páginas 497–529, 2000.

LADOUKAKIS, E., PEREIRA, V., MAGNY, E., EYRE-WALKER, A. y COUSO, J. Hundreds of Putatively Functional Small Open Reading Frames in Drosophila. *Genome Biology*, vol. 12(11), página R118, 2011.

LAMPSON, B., INOUYE, M. y INOUYE, S. Retrons, msDNA, and the bacterial genome. *Cytogenetic and Genome Research*, vol. 110(1-4), páginas 491–499, 2005.

LANGMEAD, B. y SALZBERG, S. Fast Gapped-Read Alignment with Bowtie 2. *Nature Methods*, vol. 9(4), páginas 357–359, 2012.

LANGMEAD, B., TRAPNELL, C., POP, M. y SALZBERG, S. Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. *Genome Biology*, vol. 10(3), página R25, 2009.

LAO, N. T., MALONEY, A. P., ATKINS, J. F. y KAVANAGH, T. A. Versatile Dual Reporter Gene Systems for Investigating Stop Codon Readthrough in Plants. *PLoS ONE*, vol. 4(10), página e7354, 2009.

LAYANA, C., FERRERO, P. y RIVERA-POMAR, R. Cytoplasmic Ribonucleoprotein Foci in Eukaryotes: Hotspots of Bio(chemical)Diversity. *Comparative and Functional Genomics*, vol. 2012, páginas 1–7, 2012.

LEE, H.-L. R. y DOUGHERTY, J. P. Pharmaceutical therapies to recode nonsense mutations in inherited diseases. *Pharmacology & Therapeutics*, vol. 136(2), páginas 227–266, 2012.

LEE, R., FEINBAUM, R. y AMBROS, V. The C. elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complementarity to lin-14. *Cell*, vol. 75(5), páginas 843–54, 1993.

LEWIN, B. *Genes VII.*. Oxford University Press; New Edition (2000, 2002), 2000.

LI, C. y ZHANG, J. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. *PLOS Genetics*, vol. 15(5), página e1008141, 2019.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. y DURBIN, R. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics*, vol. 25(16), páginas 2078–2079, 2009.

LI, R., LI, Y., KRISTIANSEN, K. y WANG, J. SOAP: short oligonucleotide alignment program. *Bioinformatics*, vol. 24(5), páginas 713–714, 2008.

LI, Y., WANG, X., LI, C., HU, S., YU, J. y SONG, S. Transcriptome-wide N 6 -methyladenosine profiling of rice callus and leaf reveals the presence of tissue-specific competitors involved in selective mRNA modification. *RNA Biology*, vol. 11(9), páginas 1180–1188, 2014.

LIN, M. F., CARLSON, J. W., CROSBY, M. A., MATTHEWS, B. B., YU, C., PARK, S., WAN, K. H., SCHROEDER, A. J., GRAMATES, L. S., ST. PIERRE, S. E., ROARK, M., WILEY, K. L., KULATHINAL, R. J., ZHANG, P., MYRICK, K. V., ANTONE, J. V., CELNIKER, S. E., GELBART, W. M. y KELLIS, M. Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Research*, vol. 17(12), páginas 1823–1836, 2007.

LIN, M. F., JUNGREIS, I. y KELLIS, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, vol. 27(13), páginas i275–i282, 2011.

LINDBLAD-TOH, K., GARBER, M., ZUK, O. y LIN, M. F. E. A. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, vol. 478(7370), páginas 476–482, 2011.

LINDE, L. y KEREM, B. Introducing sense into nonsense in treatments of human genetic diseases. *Trends in Genetics*, vol. 24(11), páginas 552–563, 2008.

LODISH, H., BERK, A., MATSUDAIRA, P., KAISER, C., KRIEGER, M., SCOTT, M., ZIPURSKY, S. y DARNELL, J. *BiologÃa Celular y Molecular (5ta Edicion)*. Editorial MÃ©dica Panamericana, 2005.

LÓPEZ-SANTIBÁÑEZ-JÁCOME, L., AVENDAÑO-VÁZQUEZ, S. E. y FLORES-JASSO, C. F. The Pipeline Repertoire for Ig-Seq Analysis. *Frontiers in Immunology*, vol. 10, página 899, 2019.

LOUGHRAN, G., CHOU, M.-Y., IVANOV, I. P., JUNGREIS, I., KELLIS, M., KIRAN, A. M., BARANOV, P. V. y ATKINS, J. F. Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Research*, vol. 42(14), páginas 8928–8938, 2014.

LOUGHRAN, G., JUNGREIS, I., TZANI, I., POWER, M., DMITRIEV, R. I., IVANOV, I. P., KELLIS, M. y ATKINS, J. F. Stop codon readthrough generates a C-terminally extended variant of the human vitamin D receptor with reduced calcitriol response. *Journal of Biological Chemistry*, vol. 293(12), páginas 4434–4444, 2018.

MACKAY, D. J. y MAC KAY, D. J. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, first edit edición, 2003.

MANAGADZE, D., LOBKOVSKY, A., WOLF, Y., SHABALINA, S., ROGOZIN, I. y KOONIN, E. The Vast, Conserved Mammalian lincRNome. *PLoS Computational Biology*, vol. 9(2), página e1002917, 2013.

MANUVAKHOVA, M., KEELING, K. y BEDWELL, D. M. Aminoglycoside antibiotics mediate context-dependent suppression of termination codons in a mammalian translation system. *RNA*, vol. 6(7), página S1355838200000716, 2000.

MARIONI, J., MASON, C., MANE, S., STEPHENS, M. y GILAD, Y. RNA-Seq: an Assessment of Technical Reproducibility and Comparison With Gene Expression Arrays. *Genome Research*, vol. 18(9), páginas 1509–1517, 2008.

MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, vol. 17(1), página 10, 2011.

MARYGOLD, S. J., ROOTE, J., REUTER, G., LAMBERTSSON, A., ASHBURNER, M., MILLBURN, G. H., HARRISON, P. M., YU, Z., KENMOCHI, N., KAUFMAN, T. C., LEEVERS, S. J. y COOK, K. R. The ribosomal protein genes and Minute loci of Drosophila melanogaster. *Genome Biology*, vol. 8(10), página R216, 2007.

MASSEY, S. E. The identities of stop codon reassignments support ancestral tRNA stop codon decoding activity as a facilitator of gene duplication and evolution of novel function. *Gene*, vol. 619, páginas 37–43, 2017.

MATSUFUJI, S., MATSUFUJI, T., MIYAZAKI, Y., MURAKAMI, Y., ATKINS, J. F., GESTELAND, R. F. y HAYASHI, S.-I. Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. *Cell*, vol. 80(1), páginas 51–60, 1995.

MCCLUNG, C. y HIRSH, J. Stereotypic behavioral responses to free-base cocaine and the development of behavioral sensitization in Drosophila. *Current Biology*, vol. 8(2), páginas 109–112, 1998.

MEYER, K. D., SALETORE, Y., ZUMBO, P., ELEMENTO, O., MASON, C. E. y JAFFREY, S. R. Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3-UTRs and near Stop Codons. *Cell*, vol. 149(7), páginas 1635–1646, 2012.

MICHEL, A. y BARANOV, P. Ribosome Profiling: A Hi-Def Monitor for Protein Synthesis at the Genome-Wide Scale. *Wiley Interdisciplinary Reviews: RNA*, vol. 4(5), páginas 473–490, 2013.

MUMTAZ, M. y COUSO, J. Ribosomal Profiling Adds New Coding Sequences to the Proteome. *Biochemical Society Transactions*, vol. 43, páginas 1271–1276, 2015.

MURTHA, K., HWANG, M., PECCARELLI, M. C., SCOTT, T. D. y KEBAARA, B. W. The nonsense-mediated mRNA decay (NMD) pathway differentially regulates COX17, COX19 and COX23 mRNAs. *Current Genetics*, vol. 65(2), páginas 507–521, 2019.

NAKAO, M. Epigenetics: interaction of DNA methylation and chromatin. *Gene*, vol. 278(1-2), páginas 25–31, 2001.

NAMY, O., DUCHATEAU-NGUYEN, G. y ROUSSET, J.-P. Translational readthrough of the PDE2 stop codon modulates cAMP levels in Saccharomyces cerevisiae. *Molecular Microbiology*, vol. 43(3), páginas 641–652, 2002.

NAMY, O., HATIN, I. y ROUSSET, J.-P. Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO reports*, vol. 2(9), páginas 787–793, 2001. ISSN 1469221X.

NAMY, O. y ROUSSET, J.-P. Specification of Standard Amino Acids by Stop Codons. En *Recoding: Expansion of Decoding Rules Enriches Gene Expression.* (editado por G. R. e. Atkins J.), capítulo Chapter, páginas 79–100. Springer, New York., Springer, New York., 2010.

NAMY, O., ROUSSET, J.-P., NAPTHINE, S. y BRIERLEY, I. Reprogrammed Genetic Decoding in Cellular Gene Expression. *Molecular Cell*, vol. 13(2), páginas 157–168, 2004.

ORR, M. W., MAO, Y., STORZ, G. y QIAN, S.-B. Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Research*, vol. 48(3), páginas 1029–1042, 2020.

OTERO-MORENO, D., PEÑA-RANGEL, M. y RIESGO-ESCOVAR, J. Crecimiento y Metabolismo: la Regulacion y la Via de la Insulina desde la Mosca de la Fruta, Drosophila melanogaster. *Revista Especializada en Ciencias QuÃmico-Biológicas*, vol. 19(2), páginas 116–126, 2016.

OWUSU-ANSAH, E. y PERRIMON, N. Modeling Metabolic Homeostasis and Nutrient Sensing in Drosophila: Implications for Aging and Metabolic Diseases. *Disease Models & Mechanisms*, vol. 7(3), páginas 343–350, 2014.

PALMA, M. y LEJEUNE, F. Deciphering the molecular mechanism of stop codon readthrough. *Biological Reviews*, vol. 96(1), páginas 310–329, 2021.

PANCSA, R., MACOSSAY-CASTILLO, M., KOSOL, S. y TOMPA, P. Computational analysis of translational readthrough proteins in Drosophila and yeast reveals parallels to alternative splicing. *Scientific Reports*, vol. 6(1), página 32142, 2016.

PATTON, J. T. *Segmented double-stranded RNA viruses: structure and molecular biology.* Horizon Scientific Press, 2008. ISBN 978-1-904455-21-9.

PAULSEN, M. y FERGUSON-SMITH, A. C. DNA methylation in genomic imprinting, development, and disease. *The Journal of Pathology*, vol. 195(1), páginas 97–110, 2001.

PEARSON, H. What is a gene? *Nature*, vol. 441(7092), páginas 398–401, 2006.

PELHAM, H. R. Leaky UAG termination codon in tobacco mosaic virus RNA. *Nature*, vol. 272(5652), páginas 469–471, 1978. ISSN 0028-0836.

PENNISI, E. GENOMICS:Fruit Fly Genome Yields Data and a Validation. *Science*, vol. 287(5457), páginas 1374a–1374, 2000.

PESTOVA, T., LORSCH, J. y HELLEN, C. Translational Control in Biology and Medicine. En *Translational Control in Biology and Medicine.* (editado por J. W. H. Michael B. Mathews, Nahum Sonenberg), páginas 87–128. CSHL Press, Cold Spring Harbor, NY., New York, 2007. ISBN 978-087969767-9.

PFISTER, P., HOBBIE, S., VICENS, Q., BATTGER, E. C. y WESTHOF, E. The molecular basis for a-site mutations conferring aminoglycoside resistance: Relationship between ribosomal susceptibility and x-ray crystal structures. *ChemBioChem*, vol. 4(10), páginas 1078–1088, 2003.

PIERCE, B. A. *Genetics: A Conceptual Approach (3rd Edition)*. Editorial Panamericana, 2009.

PISAREV, A. V., HELLEN, C. U. y PESTOVA, T. V. Recycling of Eukaryotic Posttermination Ribosomal Complexes. *Cell*, vol. 131(2), páginas 286–299, 2007.

PLOTKIN, J. y KUDLA, G. Synonymous But Not the Same: the Causes and Consequences of Codon Bias. *Nature Reviews Genetics*, vol. 12(1), páginas 32–42, 2011.

POOLE, E. Translational termination in Escherichia coli: three bases following the stop codon crosslink to release factor 2 and affect the decoding efficiency of UGA-containing signals. *Nucleic Acids Research*, vol. 26(4), páginas 954–960, 1998. ISSN 13624962.

POWELL, J. R. *Progress and prospects in evolutionary biology: the Drosophila model.*. Oxford University Press., 1997.

PRUSINER, S. B. Prion Diseases and the BSE Crisis. *Science*, vol. 278(5336), páginas 245–251, 1997.

QUAX, T. E., CLAASSENS, N. J., SÖLL, D. y VAN DER OOST, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell*, vol. 59(2), páginas 149–161, 2015.

RAJON, E. y MASEL, J. Evolution of molecular error rates and the consequences for evolvability. *Proceedings of the National Academy of Sciences*, vol. 108(3), páginas 1082–1087, 2011.

REBIBO-SABBAH, A., NUDELMAN, I., AHMED, Z. M., BAASOV, T. y BEN-YOSEF, T. In vitro and ex vivo suppression by aminoglycosides of PCDH15 nonsense mutations underlying type 1 Usher syndrome. *Human Genetics*, vol. 122(3-4), páginas 373–381, 2007.

REITER, L., POTOCKI, L., CHIEN, S., GRIBSKOV, M. y BIER, E. A Systematic Analysis of Human Disease-Associated Gene Sequences in Drosophila melanogaster. *Genome Research*, vol. 11(6), páginas 1114–1125, 2001.

RIBAS DE POUPLANA, L., SANTOS, M. A., ZHU, J.-H., FARABAUGH, P. J. y JAVID, B. Protein mistranslation: friend or foe? *Trends in Biochemical Sciences*, vol. 39(8), páginas 355–362, 2014.

RICHARDS, E. J. y ELGIN, S. C. Epigenetic Codes for Heterochromatin Formation and Silencing. *Cell*, vol. 108(4), páginas 489–500, 2002.

ROBINSON, D. N. y COOLEY, L. Examination of the function of two kelch proteins generated by stop codon suppression. *Development (Cambridge, England)*, vol. 124(7), páginas 1405–17, 1997.

ROEPSTORFF, P. y FOHLMAN, J. Letter to the editors. *Biological Mass Spectrometry*, vol. 11(11), páginas 601–601, 1984.

ROHRIG, H., SCHMIDT, J., MIKLASHEVICHS, E., SCHELL, J. y JOHN, M. Soybean ENOD40 Encodes Two Peptides that Bind to Sucrose Synthase. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99(4), páginas 1915–1920, 2002.

ROSENTHAL, N. y HARVEY, R. *Heart Development and Regeneration.*. Academic Press, European Molecular Biology Laboratory, Monterotondo (Rome), Italy, academic press edición, 2010. ISBN 9780123813329.

ROUZÉ, P., PAVY, N. y ROMBAUTS, S. Genome Annotation: which tools do we have for it? *Current Opinion in Plant Biology*, vol. 2(2), páginas 90–95, 1999.

ROWE, S. M., SLOANE, P., TANG, L. P., BACKER, K., MAZUR, M., BUCKLEY-LANIER, J., NUDELMAN, I., BELAKHOV, V., BEBOK, Z., SCHWIEBERT, E., BAASOV, T. y BEDWELL, D. M. Suppression of CFTR premature termination codons and rescue of CFTR protein and function by the synthetic aminoglycoside NB54. *Journal of Molecular Medicine*, vol. 89(11), páginas 1149–1161, 2011.

ROY, B., LESZYK, J. D., MANGUS, D. A. y JACOBSON, A. Nonsense suppression by near-cognate tRNAs employs alternative base pairing at codon positions 1 and 3. *Proceedings of the National Academy of Sciences*, vol. 112(10), páginas 3038–3043, 2015.

RUBIN, G. Drosophila melanogaster as an Experimental Organism. *Science*, vol. 240(4858), páginas 1453–1459, 1988.

RULIFSON, E. J. Ablation of Insulin-Producing Neurons in Flies: Growth and Diabetic Phenotypes. *Science*, vol. 296(5570), páginas 1118–1120, 2002.

RUTSCHOW, H., YTTERBERG, A. J., FRISO, G., NILSSON, R. y VAN WIJK, K. J. Quantitative Proteomics of a Chloroplast SRP54 Sorting Mutant and Its Genetic Interactions with CLPC1 in Arabidopsis. *Plant Physiology*, vol. 148(1), páginas 156–175, 2008.

RYOJI, M., HSIA, K. y KAJI, A. Read-through translation. *Trends in Biochemical Sciences*, vol. 8(3), páginas 88–90, 1983.

SALVATORE, D., LOW, S. C., BERRY, M., MAIA, A. L., HARNEY, J. W., CROTEAU, W., ST GERMAIN, D. L. y LARSEN, P. R. Type 3 Iodothyronine deiodinase: cloning, in vitro expression, and functional analysis of the placental selenoenzyme. *Journal of Clinical Investigation*, vol. 96(5), páginas 2421–2430, 1995.

SAMSON, M. L., LISBIN, M. J. y WHITE, K. Two distinct temperature-sensitive alleles at the elav locus of Drosophila are suppressed nonsense mutations of the same tryptophan codon. *Genetics*, vol. 141(3), páginas 1101–11, 1995.

DOS SANTOS, G., SCHROEDER, A., GOODMAN, J., STRELETS, V., CROSBY, M., THURMOND, J., EMMERT, D., GELBART, W. y FLYBASE CONSORTIUM. FlyBase: Introduction of the Drosophila melanogaster Release 6 Reference Genome

Assembly and Large-scale Migration of Genome Annotations. *Nucleic Acids Research*, vol. 43, páginas D690–D697, 2015.

SAPKOTA, D., LAKE, A. M., YANG, W., YANG, C., WESSELING, H., GUISE, A., UNCU, C., DALAL, J. S., KRAFT, A. W., LEE, J.-M., SANDS, M. S., STEEN, J. A. y DOUGHERTY, J. D. Cell-Type-Specific Profiling of Alternative Translation Identifies Regulated Protein Isoform Variation in the Mouse Brain. *Cell Reports*, vol. 26(3), páginas 594–607.e7, 2019.

SARKAR, C., ZHANG, Z. y MUKHERJEE, A. B. Stop codon read-through with PTC124 induces palmitoyl-protein thioesterase-1 activity, reduces thioester load and suppresses apoptosis in cultured cells from INCL patients. *Molecular Genetics and Metabolism*, vol. 104(3), páginas 338–345, 2011.

SATO, M., UMEKI, H., SAITO, R., KANAI, A. y TOMITA, M. Computational analysis of stop codon readthrough in D.melanogaster. *Bioinformatics*, vol. 19(11), páginas 1371–1380, 2003.

SAVARD, J., MARQUES-SOUZA, H., ARANDA, M. y TAUTZ, D. A Segmentation Gene in Tribolium Produces a Polycistronic mRNA that Codes for Multiple Conserved Peptides. *Cell*, vol. 126(3), páginas 559–569, 2006.

SCHUEREN, F., LINGNER, T., GEORGE, R., HOFHUIS, J., DICKEL, C., GÄRTNER, J. y THOMS, S. Peroxisomal lactate dehydrogenase is generated by translational readthrough in mammals. *eLife*, vol. 3, 2014.

SCHUEREN, F. y THOMS, S. Functional Translational Readthrough: A Systems Biology Perspective. *PLoS Genetics*, vol. 12(8), páginas 1–12, 2016.

SHARP, P. y LI, W. The codon Adaptation Index: a Measure of Directional Synonymous Codon Usage Bias, and its Possible Applications. *Nucleic Acids Research*, vol. 15(3), páginas 1281–1295, 1987.

SIMON, D. M. y ZIMMERLY, S. A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Research*, vol. 36(22), páginas 7219–7229, 2008.

SINGH, A., MANJUNATH, L. E., KUNDU, P., SAHOO, S., DAS, A., SUMA, H. R., FOX, P. L. y ESWARAPPA, S. M. Let-7a-regulated translational readthrough of mammalian AGO 1 generates a micro RNA pathway inhibitor. *The EMBO Journal*, vol. 38(16), 2019.

SLAVOFF, S., MITCHELL, A., SCHWAID, A., CABILI, M., MA, J., LEVIN, J., KARGER, A., BUDNIK, B., RINN, J. y SAGHATELIAN, A. Peptidomic Discovery of Short Open Reading Frame-Encoded Peptides in Human Cells. *Nature Chemical Biology*, vol. 9(1), páginas 59–64, 2013.

SMITH, W., THOMAS, J., LIU, J., LI, T. y MORAN, T. From Fat Fruit Fly to Human Obesity. *Physiology & Behavior*, vol. 136(3), páginas 15–21, 2014.

SONENBERG, N. y HINNEBUSCH, A. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell*, vol. 136(4), páginas 731–745, 2009.

STARK, A., LIN, M. F., KHERADPOUR, P., PEDERSEN, J. S., PARTS, L., CARLSON, J. W., CROSBY, M. A., RASMUSSEN, M. D., ROY, S., DEORAS, A. N., RUBY, J. G., BRENNECKE, J., HODGES, E., HINRICHS, A. S., CASPI, A., PATEN, B., PARK, S.-W., HAN, M. V., MAEDER, M. L., POLANSKY, B. J., ROBSON, B. E., AERTS, S., VAN HELDEN, J., HASSAN, B., GILBERT, D. G., EASTMAN, D. A., RICE, M., WEIR, M., HAHN, M. W., PARK, Y., DEWEY, C. N., PACHTER, L., KENT, W. J., HAUSSLER, D., LAI, E. C., BARTEL, D. P., HANNON, G. J., KAUFMAN, T. C., EISEN, M. B., CLARK, A. G., SMITH, D., CELNIKER, S. E., GELBART, W. M. y KELLIS, M. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, vol. 450(7167), páginas 219–232, 2007.

STENEBERG, P. y SAMAKOVLIS, C. A novel stop codon readthrough mechanism produces functional Headcase protein in Drosophila trachea. *EMBO reports*, vol. 2(7), páginas 593–7, 2001.

STIEBLER, A. C., FREITAG, J., SCHINK, K. O., STEHLIK, T., TILLMANN, B. A. M., AST, J. y BÖLKER, M. Ribosomal Readthrough at a Short UGA Stop Codon Context Triggers Dual Localization of Metabolic Enzymes in Fungi and Animals. *PLoS Genetics*, vol. 10(10), página e1004685, 2014.

SUH, J., RIVEST, A., NAKASHIBA, T., TOMINAGA, T. y TONEGAWA, S. Entorhinal Cortex Layer III Input to the Hippocampus Is Crucial for Temporal Association Memory. *Science*, vol. 334(6061), páginas 1415–1420, 2011.

SUPEK, F., BOŠNJAK, M., ŠKUNCA, N. y ŠMUC, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE*, vol. 6(7), página e21800, 2011.

THE GENE ONTOLOGY CONSORTIUM. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, vol. 45(D1), páginas D331–D338, 2017.

TODD, J. F. Recommendations for nomenclature and symbolism for mass spectroscopy. *International Journal of Mass Spectrometry and Ion Processes*, vol. 142(3), páginas 209–240, 1995.

TÖRNROTH-HORSEFIELD, S., HEDFALK, K., FISCHER, G., LINDKVIST-PETERSSON, K. y NEUTZE, R. Structural insights into eukaryotic aquaporin regulation. *FEBS Letters*, vol. 584(12), páginas 2580–2588, 2010.

TOURIOL, C., BORNES, S., BONNAL, S., AUDIGIER, S., PRATS, H., PRATS, A.-C. y VAGNER, S. Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biology of the Cell*, vol. 95(3-4), páginas 169–178, 2003.

TRAPNELL, C., PACHTER, L. y SALZBERG, S. TopHat: Discovering Splice Junctions with RNA-Seq. *Bioinformatics*, vol. 25(9), páginas 1105–1111, 2009.

TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. y PACHTER, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, vol. 7(3), páginas 562–578, 2012.

TRAPNELL, C., WILLIAMS, B. A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M. J., SALZBERG, S. L., WOLD, B. J. y PACHTER, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, vol. 28(5), páginas 511–515, 2010.

TWEEDIE, S., ASHBURNER, M., FALLS, K., LEYLAND, P., MCQUILTON, P., MARYGOLD, S., MILLBURN, G., OSUMI-SUTHERLAND, D., SCHROEDER, A., SEAL, R. y ZHANG, H. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Research*, vol. 37(1 (Database)), páginas D555–D559, 2009.

VALLE, R. P., DRUGEON, G., DEVIGNES-MORCH, M.-D., LEGOCKI, A. B. y HAENNI, A.-L. Codon context effect in virus translational readthrough A study in vitro of the determinants of TMV and Mo-MuLV amber suppression. *FEBS Letters*, vol. 306(2-3), páginas 133–139, 1992.

VAN DAMME, P., GAWRON, D., VAN CRIEKINGE, W. y MENSCHAERT, G. N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. *Molecular & cellular proteomics : MCP*, vol. 13(5), páginas 1245–61, 2014.

VAUDEL, M., BARSNES, H., BERVEN, F., SICKMANN, A. y MARTENS, L. SearchGUI: an Open-Source Graphical User Interface for Simultaneous OMSSA and X!Tandem Searches. *Proteomics*, vol. 11(5), páginas 996–999, 2011.

VAUDEL, M., BURKHART, J. M., ZAHEDI, R. P., OVELAND, E., BERVEN, F. S., SICKMANN, A., MARTENS, L. y BARSNES, H. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology*, vol. 33(1), páginas 22–24, 2015.

VECSLER, M., BEN ZEEV, B., NUDELMAN, I., ANIKSTER, Y., SIMON, A. J., AMARIGLIO, N., RECHAVI, G., BAASOV, T. y GAK, E. Ex Vivo Treatment with a Novel Synthetic Aminoglycoside NB54 in Primary Fibroblasts from Rett Syndrome Patients Suppresses MECP2 Nonsense Mutations. *PLoS ONE*, vol. 6(6), página e20733, 2011.

WANG, B., YANG, Z., BRISSON, B. K., FENG, H., ZHANG, Z., WELCH, E. M., PELTZ, S. W., BARTON, E. R., BROWN, R. H. y SWEENEY, H. L. Membrane blebbing as an assessment of functional rescue of dysferlin-deficient human myotubes via nonsense suppression. *Journal of Applied Physiology*, vol. 109(3), páginas 901–905, 2010.

WANG, D., BELAKHOV, V., KANDASAMY, J., BAASOV, T., LI, S.-C., LI, Y.-T., BEDWELL, D. M. y KEELING, K. M. The designer aminoglycoside NB84 significantly reduces glycosaminoglycan accumulation associated with MPS I-H in the Idua-W392X mouse. *Molecular Genetics and Metabolism*, vol. 105(1), páginas 116–125, 2012.

WANG, E., SANDBERG, R., LUO, S., KHREBTUKOVA, I. y ZHANG, L. NIH Public Access. *Nature*, vol. 456(7221), páginas 470–476, 2008.

WANG, J., LI, S., ZHANG, Y., ZHENG, H., XU, Z., YE, J., YU, J. y WONG, G. Vertebrate Gene Predictions and the Problem of Large Genes. *Nature Reviews. Genetics*, vol. 4(9), páginas 741–9, 2003.

WANG, Z., GERSTEIN, M. y SNYDER, M. RNA-Seq: a Revolutionary Tool for Transcriptomics. *Nature Reviews Genetics*, vol. 10(1), páginas 57–63, 2009.

WANGEN, J. R. y GREEN, R. Stop codon context influences genome-wide stimulation of termination codon readthrough by aminoglycosides. *eLife*, vol. 9, 2020.

WATSON, J. D. y CRICK, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, vol. 171(4356), páginas 737–738, 1953.

WEINER, A. M. y WEBER, K. Natural Read-through at the UGA Termination Signal of Qβ Coat Protein Cistron. *Nature New Biology*, vol. 234(50), páginas 206–209, 1971.

WEISS, R. y ATKINS, J. Translation Goes Global. *Science*, vol. 334(6062), páginas 1509–1510, 2011.

WILLS, N. M., O'CONNOR, M., NELSON, C. C., RETTBERG, C. C., HUANG, W. M., GESTELAND, R. F. y ATKINS, J. F. Translational bypassing without peptidyl-tRNA anticodon scanning of coding gap mRNA. *The EMBO Journal*, vol. 27(19), páginas 2533–2544, 2008.

WILSON, B. A. y MASEL, J. Putatively Noncoding Transcripts Show Extensive Association with Ribosomes. *Genome Biology and Evolution*, vol. 3, páginas 1245–1252, 2011.

WOLF, M. J., AMREIN, H., IZATT, J. A., CHOMA, M. A., REEDY, M. C. y ROCKMAN, H. A. From The Cover: Drosophila as a model for the identification of genes causing adult human heart disease. *Proceedings of the National Academy of Sciences*, vol. 103(5), páginas 1394–1399, 2006.

WOLFRAM, S. Simple solutions; the iconoclastic physicist's mathematica software nails complex puzzles. *Business Week, October*, vol. 3, 2005.

WRIGHT, W. E., PIATYSZEK, M. A., RAINEY, W. E., BYRD, W. y SHAY, J. W. Telomerase activity in human germline and embryonic tissues and cells. *Developmental Genetics*, vol. 18(2), páginas 173–179, 1996.

XUE, F. y COOLEY, L. Kelch encodes a component of intercellular bridges in Drosophila egg chambers. *Cell*, vol. 72(5), páginas 681–693, 1993.

XUE, X., MUTYAM, V., TANG, L., BISWAS, S., DU, M., JACKSON, L. A., DAI, Y., BELAKHOV, V., SHALEV, M., CHEN, F., SCHACHT, J., J. BRIDGES, R., BAASOV, T., HONG, J., BEDWELL, D. M. y ROWE, S. M. Synthetic Aminoglycosides Efficiently Suppress Cystic Fibrosis Transmembrane Conductance Regulator Nonsense Mutations and Are Enhanced by Ivacaftor. *American Journal of Respiratory Cell and Molecular Biology*, vol. 50(4), páginas 805–816, 2014.

XUE, X., MUTYAM, V., THAKERAR, A., MOBLEY, J., BRIDGES, R. J., ROWE, S. M., KEELING, K. M. y BEDWELL, D. M. Identification of the amino acids inserted during suppression of CFTR nonsense mutations and determination of their functional consequences. *Human Molecular Genetics*, vol. 26(16), páginas 3116–3129, 2017.

YUKIHARA, M., ITO, K., TANOUE, O., GOTO, K., MATSUSHITA, T., MATSUMOTO, Y., MASUDA, M., KIMURA, S. y UEOKA, R. Effective Drug Delivery System for Duchenne Muscular Dystrophy Using Hybrid Liposomes Including Gentamicin along with Reduced Toxicity. *Biological & Pharmaceutical Bulletin*, vol. 34(5), páginas 712–716, 2011.

ZAMUDIO-ARROYO, J. M., PEÑA-RANGEL, M. T. y RIESGO-ESCOVAR, J. R. La ubiquitinacion: un sistema de regulacion dinámico de los organismos. *TIP Revista Especializada en Ciencias Químico-Biologicas*, vol. 15(2), páginas 133–141, 2012.

ZEITOUNI, B., SÉNATORE, S., SÉVERAC, D., AKNIN, C., SÉMÉRIVA, M. y PERRIN, L. Signalling Pathways Involved in Adult Heart Formation Revealed by Gene Expression Profiling in Drosophila. *PLoS Genetics*, vol. 3(10), página e174, 2007.

ZENG, J., ALHAJJ, R. y DEMETRICK, D. Effectiveness of Applying Codon Usage Bias for Translational Initiation Sites Prediction. En *2008 IEEE International Conference on Bioinformatics and Biomedicine*, páginas 121–126. IEEE, 2008.

ZINGMAN, L. V., PARK, S., OLSON, T. M., ALEKSEEV, A. E. y TERZIC, A. Aminoglycoside-induced Translational Read-through in Disease: Overcoming Nonsense Mutations by Pharmacogenetic Therapy. *Clinical Pharmacology & Therapeutics*, vol. 81(1), páginas 99–103, 2007.

"Todo investigador sufre las decepciones, los largos meses pasados en una dirección equivocada, los fracasos. Pero los errores son también útiles, porque bien analizados, pueden conducir al éxito. Y para el investigador no existe alegría comparable a la de un descubrimiento, por pequeño que sea..."

Alexander Fleming

"La ciencia no sabe de países, porque el conocimiento le pertenece a la humanidad y es la antorcha que ilumina al mundo. La ciencia es el alma de la prosperidad de las naciones y la fuente de todo progreso."

Louis Pasteur