UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

"Diseño de metodologías de enriquecimiento con transcritos primarios para detección de sitios de inicio de la transcripción en *Escherichia coli* K12. (Técnicas CAPture y pH-Switch)."

TESIS

QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIAS

PRESENTA:
Lic. Israel Aguilar Ordóñez

TUTOR PRINCIPAL

Dr. J. Enrique Morett Sánchez
Instituto de Biotecnología – UNAM

MIEMBROS DEL COMITÉ TUTOR

Dr. Federico Sánchez Rodríguez
Instituto de Biotecnología – UNAM

Dr. Pablo Vinuesa Fleischmann
Centro de Ciencias Genómicas – UNAM
Programa de Doctorado en Ciencias Biomédicas

MÉXICO, D. F. Junio, 2015

A mi Abuela, por esperarme y nunca abandonarme.

A Papá y Mamá, por estar siempre aquí.

A mi hermano, por inspirarme.

A Emma, por levantarme.

Gracias...

Al Dr. J. Enrique Morett, por la paciencia y confianza que tuvo en el proyecto, pero sobre todo por la amistad y las puertas siempre abiertas. Lo mejor que me llevo de este trabajo es su visión tan apasionada de lo que es hacer ciencia.

Al laboratorio 9, con todas las personas que pasaron por ahí en este tiempo transcurrido. Lo más importante en cualquier lugar de trabajo es la compañía. Muchas gracias por compartir los días, tardes y noches entre pipeteos, geles y risas.

A mis amigos Arlen, Yami, Getza, Brenda, Clau, Elva, Edson, Armando, Ramón, Ramses, Pancho, Panchote, Fily, Yossef, Rodrigo, y muchos más, por haber sido lo mejor que esta ciudad tuvo para ofrecerme. De todos aprendí algo, con todos compartí algo. Gracias por la felicidad que su existencia brinda a la mía.

A la Universidad Nacional Autónoma de México, por ser el centro perfecto para el desarrollo académico. Ser parte de la UNAM es la mejor oportunidad de crecer que he tenido en la vida.

Agradecimientos Técnicos.

A la Biol. Maricela Olvera Rodríguez, por su ayuda en la construcción de vectores plasmídicos y el apoyo técnico en biología molecular.

A la Ing. Leticia Olvera Rodríguez, por su ayuda en la expresión y purificación de proteínas.

Al QBC. Edson N. Cárcamo Noriega, por su ayuda en la purificación de proteínas, y aún más por las discusiones que aportaron tanto a este trabajo.

Al Biol. Filiberto Sánchez, por su ayuda en el montaje de los experimentos cromatográficos y apoyo técnico general.

Al MC. Ramón Carrasco Macías, por su ayuda en la estandarización de cultivos microbianos.

Este trabajo se llevó a cabo con el apoyo económico de CONACYT a través del proyecto 179997 a cargo del Dr. JUAN E. MORETT SANCHEZ, así como el proyecto DGAPA IN-209312 a cargo del mismo. También contribuyó el apoyo económico del Instituto de Biotecnología a través de la Partida de Practicas Escolares con cargo al presupuesto 2013. Finalmente, se agradece la beca de Ayudante de Investigador Nacional Nivel III o emérito, a través del proyecto "Identificación de TSS en *E. coli* K12 mediante secuenciación masiva de muestras enriquecidas a transcritos primarios".

CONTENIDO

I. Listado de Abreviaturas.

5' RACE	- Rapid Amplification of 5' cDNA Ends (amplifición rápida de extremos 5' cDNA)
BdRppH	- RNA 5' pirofosfohidrolasa (*Bdellovibrio bacteriovorus*).
BsRppH	- RNA 5' pirofosfohidrolasa (*Bacillus subtilis*).
cDNA	- completary DNA (DNA complementario).
CEC	- Complejo Enzimático de Capping.
DNA	- Desoxyribonucleic Acid (Ácido Desoxiribonucléico)
dNTP	- deoxynucleótido trifosfato
EcRppH	- RNA 5' pirofosfohidrolasa (*Eschericha coli*).
eIF4E	- Factor Eucariótico de Inicio de la Traducción 4E.
GOLD	- Genome OnLine Database.
GST	- Glutatión S-transferasa
GST-BdRppH	- Proteína de fusión Glutatión S-transferasa-BdRppH
GST-eIF4E	- Proteína de fusión Glutatión S-transferasa-eIF4E
kDa	- kilodaltons
m7G	- 7-metilguanosina.
mRNA	- messenger RNA (RNA mensajero).
OD_{600nm}	- Densidad óptica medida con una longitud de onda de 600 nm.
ORF	- Open Reading Frame (marco abierto de lectura).
pb	- pares de bases.
PCR	- Polymerase Chain Reaction (Reacción en Cadena de la Polimerasa).
RNA	- Ribonucleic Acid (Ácido Ribonucleico).
RNasa	- Ribonucleasa.
RNA-seq	- secuenciación masiva de RNA.
RppH	- RNA 5' pirofosfohidrolasa.
rRNA	- ribosomal RNA (RNA ribosomal).
RT	- Reverse Transcriptase (Transcriptasa Reversa).
synBd*rppH*	- gen Bd*rppH* con uso de codones optimizado para su expresión en *E. coli*.
tRNA	- transfer RNA (RNA de transferencia).
TSS	- Transcription Start Site (sitio de inicio de transcripción).

II. Índice de figuras.

1. RESUMEN

Con el advenimiento de la era de la secuenciación genómica se ha comenzado el revelador camino hacia la descripción genética detallada de los organismos que habitan la biósfera conocida. Pero no basta conocer la secuencia de DNA de un organismo para poder describirlo: la presencia de un gen en un genoma no es garantía de que éste se va a expresar. Debido a esto el estudio de la regulación de la expresión génica es el siguiente paso en ese largo camino hacia la descripción completa de los organismos a través del entendimiento de los procesos biológicos que los gobiernan.

El estudio de elementos reguladores de la transcripción por genética tradicional ha ayudado a identificar genes funcionales y a comprender sus mecanismos de regulación. Con la experiencia adquirida a lo largo de los años se han desarrollado algoritmos computacionales que a partir de una secuencia genómica pueden predecir la existencia de elementos reguladores de la expresión génica [5, 30, 43]. Este tipo de predicción *in silico* ha demostrado ser una herramienta útil aunque no del todo concluyente debido a que posterior a la predicción computacional del elemento regulador (p. ej. un promotor) se debe corroborar su funcionalidad utilizando herramientas tradicionales de biología molecular.

Una alternativa más directa para la dilucidación de elementos reguladores en un genoma es la detección experimental de todos los transcritos presentes en un organismo, lo cual nos daría en definitiva una visión global de lo que sucede *in vivo*, permitiéndonos definir qué genes se expresan en ciertas condiciones y, por compaginación con la secuencia del genoma, cuáles son sus posibles elementos reguladores. Actualmente existen plataformas de secuenciación masiva de ácidos nucleicos, tecnologías que, entre otras cosas, permiten mapear a gran escala los sitios de inicio de la transcripción (TSS, del inglés *Transcription Start Site*) en un genoma mediante la secuenciación dirigida a los extremos 5' del RNA. Para que un TSS sea fidedigno, este debe ser detectado a partir de la secuenciación de transcritos primarios, definidos estos como moléculas RNA recién sintetizadas por la RNA polimerasa, en contraste con los transcritos secundarios, aquellos que han sufrido procesos de degradación. Una vez correctamente definidos los TSS es posible atisbar computacionalmente sus promotores y demás elementos reguladores con mayor certeza,

acercándonos un poco más hacia la definición de las redes de regulación que gobiernan la expresión génica.

El mapeo de TSS en *Escherichia coli* K12 a través de RNA-seq ha probado ser efectivo pero no carece de problemas: existen ribonucleasas cuya función es la degradación de ácidos ribonucleicos, liberando productos RNA cuyos extremos 5' pudieran ser asignados erróneamente como verdaderos TSS; sumado a esto, los eventos de transcripción ubicua (*pervasive transcription*, en inglés) generan bajas cantidades de transcritos primarios a lo largo de todo el genoma, ya que la célula bacteriana es capaz de tolerar este proceso casi azaroso, el entramado transcripcional se vuelve muy complejo. De inicio, esta variedad en el ribotipo es el principal problema metodológico para la detección precisa de sitios de inicio de la transcripción.

Con la intención de contribuir a un detección más fina de TSS en *E. coli* K12, en el presente trabajo se desarrollaron dos metodologías alternas para el aislamiento de la subpoblación de RNA procariótico que contiene transcritos primarios, utilizando: 1) cromatografía de afinidad a BdRppH, una pirofosfohidrolasa involucrada en la degradación de RNA; y 2) la modificación de los transcritos primarios de *E. coli* mediante reacción *in vitro* de capping 5' para su posterior captura por cromatografía de afinidad a eIF4E, factor proteico que reconoce la estructura 5'-cap presente en transcritos primarios eucarióticos.

2. INTRODUCCIÓN

2.1 La Revolución Genómica.

En 1977 Frederick Sanger reportó por primera vez la secuencia de un genoma completo correspondiente al fago ϕX174[36]. A partir de entonces el método Sanger pasó a ser la técnica de secuenciación de DNA más usada gracias a su mayor eficiencia y facilidad de manejo comparada con otras disponibles en aquel momento. Fue hasta 1995 que se reportó la secuencia genómica del primer organismo de vida libre, la bacteria *Haemophilus influenzae*[12]. Luego en 1999 fue el turno del nemátodo *Caenorhabditis elegans*[38], el primer organismo multicelular en ser secuenciado. Posteriormente en 2001 se publicó el primer *draft* del genoma humano por dos grupos independientes[24, 40]. Cada uno de estos hitos fue la culminación de largos y laboriosos años de trabajo, puesto que si bien el método Sanger era eficiente para secuenciar fidedignamente fragmentos relativamente "largos" de DNA (hasta aprox. 700 pb) éste se veía empequeñecido cuando se enfrentaba a proyectos de secuenciación de genomas completos, los cuales requieren del procesamiento de millones de bases nucleotídicas. Con todo y esta limitación, el método Sanger se convirtió en el estándar capaz de llevar a término proyectos tan complejos como la secuenciación del genoma humano.

Treinta y siete años después del primer genoma secuenciado, hoy en día la base de datos GOLD[32] (*Genomes OnLine Database*) recopila 6651 proyectos de secuenciación completados, que abarcan organismos de 3252 géneros distintos. En total, GOLD archiva 61,853 proyectos que incluyen la secuenciación completa de organismos pertenecientes a los tres dominios (figura 1): Archaea (871 proyectos), Bacteria (37944 proyectos) y Eukaryota (6057 proyectos); mientras el resto de proyectos (16981) corresponde a estudios metagenómicos, transcriptómicos, metatranscriptómicos y epigenéticos. Tal cantidad de proyectos de secuenciación de genomas completos sigue una tendencia de incremento relacionada directamente con el desarrollo de tecnologías de secuenciación masiva (*Massively Parallel Sequencing*, en inglés) sumado a los avances en bioinformática de la primera década del siglo XXI. De continuar esta tendencia no es insensato pensar en la posibilidad futura de un catálogo genómico de la biósfera conocida.

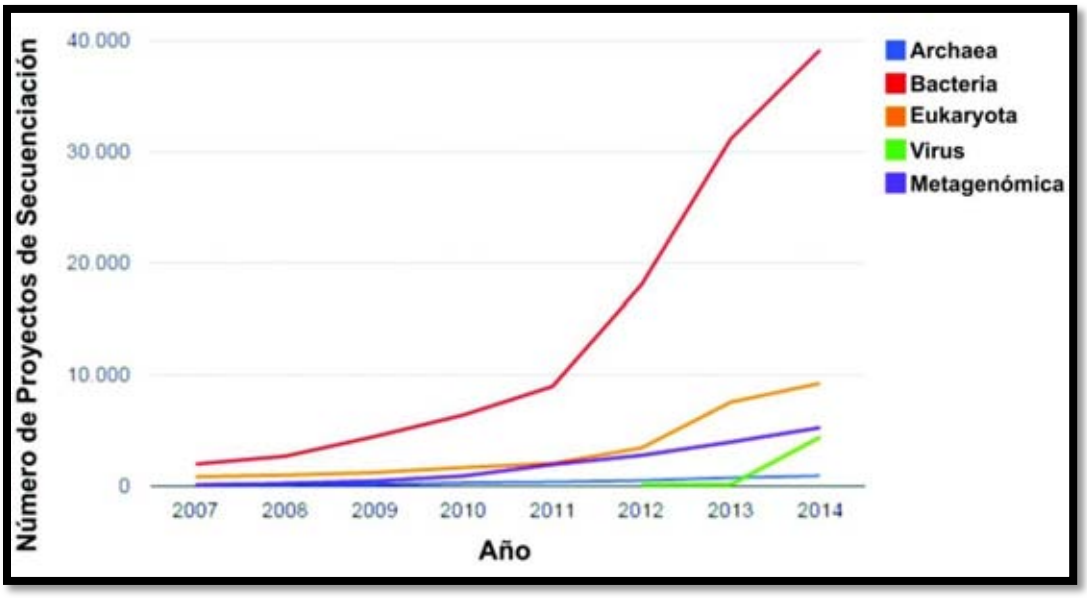


Figura 1. Proyectos de Secuenciación Genómica depositados en la base de datos GOLD. Se puede observar un importante incremento en este tipo de estudios, principalmente a finales de la primera década de este siglo. Indudablemente la tendencia se debe al constante desarrollo en las tecnologías de secuenciación de ácidos nucléicos y de manejo bioinformático de datos. (Adaptado de Reddy *et al.*[32]).

El año 2005 puede considerarse el inicio de esta revolución genómica con la publicación simultanea de los trabajos de George M. Church[37] y de Jonathan M. Rothberg[21]. La contribución de ambos fue el desarrollo de tecnologías que permitieron mejorar el rendimiento en las metodologías de secuenciación principalmente a través de la miniaturización de procesos, lo cual incrementó la capacidad de los instrumentos y redujo el costo por base secuenciada. Actualmente nos encontramos en la tercera generación de estas tecnologías y el rendimiento de las mismas continua en aumento, lo que ha llevado a la secuenciación de genomas completos a costos cada vez más accesibles. Como ejemplo basta la cifra manejada por el Beijing Genomics Institute: $999 USD / genoma humano secuenciado utilizando tecnología Complete Genomics (comunicación personal).

Gracias a este salto tecnológico ahora existe una inmensidad de datos genómicos disponibles para su análisis. ¿Pero de que serviría conocer tantos y tan diversos genomas si no entendemos cómo funcionan, qué reglas los gobiernan? Por eso en el grupo del Dr. Morett nos interesamos en el siguiente paso de la descripción de los organismos: la expresión génica, o dicho de otra manera, la exploración del transcriptoma.

2.2 Estudio del Transcriptoma a través de RNA-seq.

La regulación de la expresión génica es la base de la adaptabilidad celular de cualquier organismo. Es a través de estos procesos que podemos explicar por qué células con el mismo genoma se comportan de manera distinta dependiendo del contexto ambiental. Es decir, un gen que se expresa en cierta condición puede no expresarse con la misma intensidad en otra. Por lo anterior podemos suponer que el transcriptoma es más dinámico que el genoma. Explicar este dinamismo es el reto principal de la transcriptómica, entendiéndose ésta como el estudio de los procesos biológicos que involucran todos los tipos de moléculas RNA, sean codificantes (mRNA) o no codificantes (rRNA, tRNA, RNA reguladores y demás especies de RNA)[39].

Sin duda una de las áreas más beneficiadas por la revolución de la secuenciación masiva ha sido la transcriptómica. La cantidad extrema de datos producto de los nuevos secuenciadores ha permitido asomarse con mayor detalle a los procesos que gobiernan el entramado transcripcional de diversos organismos.

Los estudios transcriptómicos realizados con tecnologías de secuenciación masiva utilizan un conjunto de métodos colectivamente denominados RNA-seq. De manera general estos consisten en la extracción de RNA a partir de una muestra biológica de interés, que luego se convierte a cDNA utilizando una transcriptasa reversa. Este paso es necesario porque las plataformas de secuenciación masiva requieren de una biblioteca de DNA para poder funcionar. Una vez obtenido el cDNA se amplifica la cantidad de la muestra mediante PCR, obteniendo la biblioteca DNA base. De manera particular cada plataforma requiere que esta biblioteca base sea modificada con la adición de adaptadores (secuencias DNA conocidas *a priori* que servirán como anclas para que el secuenciador lea la biblioteca desde los extremos de los fragmentos); también es necesario un paso de selección de fragmentos de tamaño óptimo, el cual se acomoda a los requerimientos de los diferentes instrumentos. Una vez ensamblada esta biblioteca final el secuenciador procesa los fragmentos presentes permitiendo digitalizar la información biológica (figura 2). El resultado son millones de lecturas que computacionalmente son alineadas sobre un genoma de referencia. Así, finalmente se obtiene el panorama transcriptómico de la muestra biológica de interés.

Idealmente el proceso de construcción de bibliotecas para este tipo de experimentos debiera ser simple y con un número mínimo de pasos enzimáticos y de purificación de la muestra. Pero basta

un vistazo a la figura 2 para comprender por qué a la fecha el proceso de construcción de bibliotecas para RNA-seq es aún complejo. Cada uno de los pasos del proceso conlleva riesgos bien conocidos de introducir artificios en los resultados finales. Estos son sólo algunos de los problemas generales que se han identificado: a) el porcentaje de RNA ribosomal (el cual es indeseable para la mayoría de estudios transcriptómicos) en una muestra biológica es de hasta ~90%[6] y a pesar de que existen diversos productos comerciales estandarizados para eliminarlo de la biblioteca éstos no son tan eficientes como lo publicitado[39]; b) la mayoría de las transcriptasas reversas son capaces de utilizar como templado tanto RNA como DNA[18] fabricando cadenas cDNA "falsas" que no provienen de transcritos originales, además de mostrar tendencias a agregar ciertos nucleótidos de forma preferencial al final de las cadenas cDNA sintetizadas[44]; c) la reacción de amplificación de la biblioteca mediante PCR tiende a mostrar sesgos, amplificando preferencialmente algunas secuencias sobre otras[42]. En conjunto estos eventos erróneos pudieran contaminar gravemente los resultados ya que la cantidad de datos obtenidos por experimentos RNA-seq rodea el rango de cientos de millones de lecturas, o miles de millones de bases digitalizadas, lo cual potencia de manera importante hasta el mínimo de los sesgos.

Existen métodos bioinformáticos para atenuar el efecto de estos eventos no deseados[42], principalmente identificándolos y filtrándolos fuera del set final de datos. Sin embargo lo ideal sería evitar los eventos espurios desde la mesa de laboratorio para aumentar la calidad y cantidad de datos útiles obtenidos. Hay diversos protocolos comercialmente disponibles (Illumina, New England Biolabs, Life Technologies, Agilent, por mencionar algunos de los grandes proveedores) que pretenden cada uno a su manera y con sus respectivas patentes resolver los sesgos durante el proceso de construcción de la biblioteca RNA-seq. Las diversas alternativas existentes evidencian el amplio campo de innovación para el desarrollo de metodologías de construcción de bibliotecas RNA-seq. Esta oportunidad de innovar fue el principal impulso del trabajo presentado en esta tesis, específicamente enfocada hacia la construcción de bibliotecas para detección masiva de sitios de inicio de la transcripción.

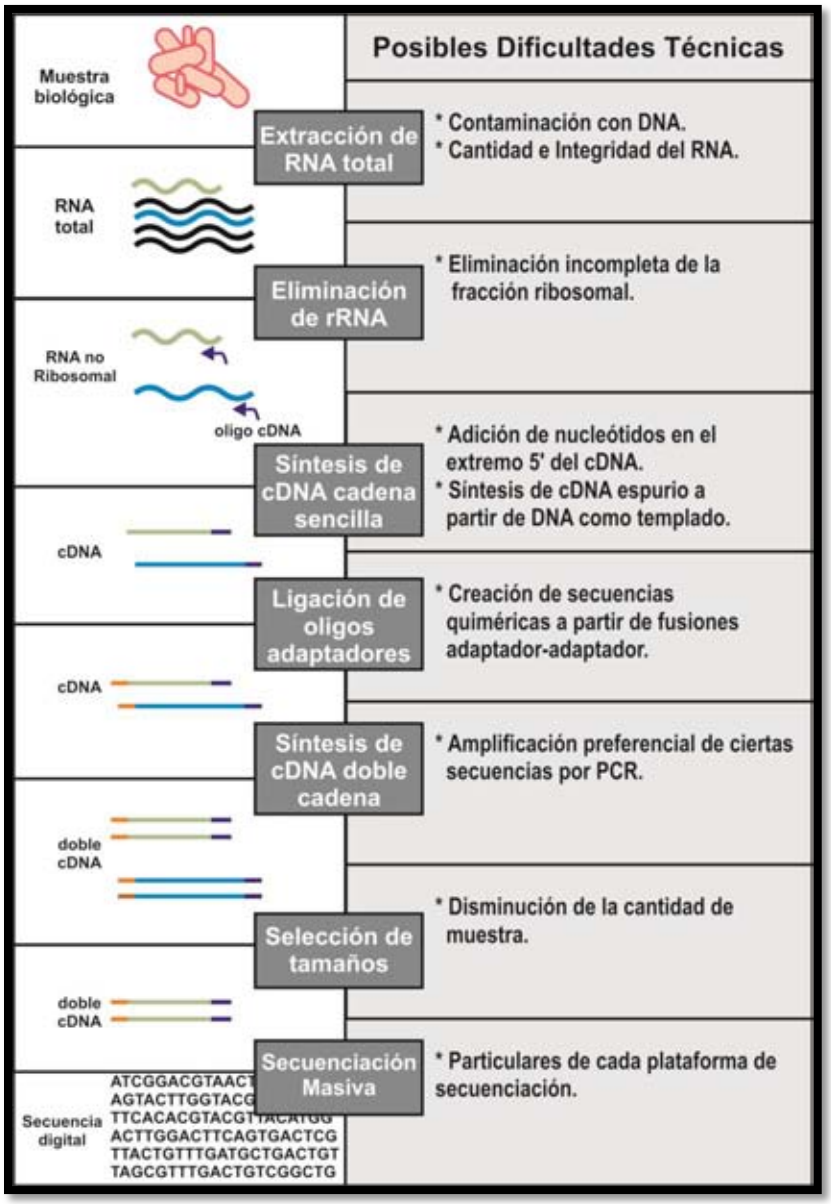


Figura 2. Proceso general de construcción de bibliotecas para RNA-seq. Del lado izquierdo se esquematiza el cambio que sufre la muestra para poder ser digitalizada. Al centro se señalan los pasos de la construcción de la biblioteca y a la derecha los posibles problemas técnicos que ocurren durante el proceso.

2.3 Detección de Sitios de Inicio de la Transcripción.

Una de las aplicaciones RNA-seq es la detección masiva de sitios de inicio de la transcripción (TSS). Particularmente en el grupo del Dr. Morett nos hemos interesado por este tipo de estudios en *Escherichia coli* K12 como un primer paso para la predicción de elementos

reguladores[25, 42] con la finalidad de contribuir al entendimiento de la red global de regulación de la expresión génica en este microorganismo modelo. La detección de TSS es una excelente aproximación hacia dicha meta puesto que una vez conocida la posición de inicio de transcripción se pueden analizar computacionalmente las regiones genómicas circundantes en busca de motivos que sugieran elementos reguladores, tales como promotores o sitios de unión de factores transcripcionales tanto conocidos como inéditos. La detección de TSS también permite delimitar transcritos atípicos como son los que ocurren dentro de marcos abiertos de lectura, transcritos antisentido y demás especies crípticas de RNA.

En este tipo de ensayos se sigue un protocolo modificado para la construcción de bibliotecas RNA-seq. Básicamente, como muestra inicial se utiliza mRNA procesado de forma tal que se concentra la subpoblación de transcritos primarios. A partir de ahí se seleccionan fragmentos correspondientes a los extremos 5'. De las lecturas obtenidas por secuenciación masiva, el primer nucleótido en la secuencia alineada contra el genoma de referencia se considera el sitio de inicio de la transcripción.

Por lo anterior la detección masiva de TSS posee una resolución a nivel de base única que la convierte en una herramienta poderosa. Pero como cualquier otra técnica relevante en biología molecular debe ser lo más precisa posible. Una detección eficiente de TSS requiere que la biblioteca cDNA sea construida tan exclusivamente como sea posible a partir de transcritos primarios, definidos como aquellos recién sintetizados por la RNA polimerasa. Sin embargo el ribotipo de un organismo es muy complejo y no todas las moléculas RNA provenientes de una muestra biológica representan verdaderos inicios de transcripción. Esto se debe a la existencia de vías de procesamiento de RNA, principalmente ribonucleasas, que liberan transcritos truncos con extremos 5' que pudieran ser falsamente detectados como TSS. Además del problema de las RNasas, existe un evento conocido como transcripción ubicua (*pervasive transcription*, en inglés), proceso que genera gran variedad de transcritos a lo largo de todo el genoma. Estos transcritos son sintetizados en baja cantidad por la célula, lo cual crea confusión en el análisis de los datos RNA-seq, puesto que por su poco nivel de expresión queda la duda de si son o no datos significativos. He ahí el problema a partir del cual concebimos este trabajo: ¿cómo es posible separar los verdaderos transcritos primarios del ruido de fondo?

3. ANTECEDENTES

3.1 El problema de las RNasas y la transcripción ubicua.

Las ribonucleasas (RNasas) son enzimas que dirigen el procesamiento, degradación y control de calidad del RNA. Se clasifican en 3 tipos: las que hidrolizan la cadena RNA internamente (endoribonucleasas), las que lo hacen a partir del extremo 5' (5' exoribonucleasas) y las que lo hacen a partir del extremo 3' (3' exoribonucleasas). Los transcritos secundarios producto de estas actividades ribonucleolíticas son fragmentos RNA con extremos 5' procesados que constituyen ruido de fondo para la detección de TSS mediante secuenciación masiva[17, 35].

Según los mecanismos enzimáticos de las RNasas presentes en *E. coli* descritas en la literatura[16] podemos prever que la principal diferencia entre estos fragmentos ruido y los transcritos primarios es la condición de sus extremos 5': sólo los transcritos primarios tienen un extremo 5' trifosfatado, mientras que los transcritos secundarios, sin importar que tipo de RNasa los haya procesado, tienen un 5' monofosfatado o un 5' hidroxilo.

El mapeo ideal de TSS requiere un primer filtro que separe los transcritos primarios del ruido biológico. Ya que el extremo 5' de las moléculas RNA es la diferencia más significativa, se puede explotar esa cualidad para separar físicamente los transcritos primarios presentes en la muestra de RNA total, proceso conocido como enriquecimiento de la muestra. Algunas de las estrategias de enriquecimiento que se han implementado a nivel de mesa de laboratorio son: 1) la eliminación del ruido mediante una 5'exoribonucleasa dependiente de extremos 5' monofosfato, que degrada los transcritos no deseados; y 2) la separación de 5' monofosfatos utilizando oligoribonucleótidos biotinilados como adaptadores que se ligan a estos extremos, marcándolos para su posterior extracción utilizando estreptavidina unida a perlas magnéticas. Ambas estrategias se enfocan en el enriquecimiento de la muestra mediante la acción directa sobre los transcritos monofosfatados.

Experimentos previos en el grupo del Dr. Morett han mostrado que los datos obtenidos mediante estas metodologías son de difícil interpretación[35]: en un primer análisis de 22 experimentos de detección de TSS en *E. coli* K12 bajo diferentes condiciones de crecimiento y enriquecimiento de la

muestra se observó un amplio margen de posibles sitios de inicio de transcripción a lo largo del genoma y un posible porcentaje importante de transcripción antisentido. Ambas observaciones concuerdan con reportes previos de nuestro grupo en este tema[25] y con el concepto de transcripción ubicua (*pervasive transcription*, en inglés), el cual sostiene que la transcripción del genoma es más laxa de lo que tradicionalmente se piensa.

Sin embargo, el análisis más profundo de los 22 experimentos reveló que la proporción de transcripción inusual (TSS correspondientes a transcritos no codificantes) disminuye conforme se es más estricto con el set de datos, particularmente cuando se toma en cuenta el Índice de Detección, definido como el número de experimentos en que se repite una posición determinada, es decir, qué inicio de transcripción está activo en *E. coli* a pesar de las diferentes condiciones de cultivo y manejo de la muestra. En la figura 3 se puede observar el siguiente efecto: cuando se cuantifican todas las posiciones presentes en al menos un experimento se detectan hasta 821,789 posibles TSS, pero conforme aumenta el número de experimentos en que está presente una misma posición disminuye drásticamente la cantidad de TSS detectados, hasta el punto en que sólo 243 posiciones concurren en los 22 experimentos. De manera similar, conforme aumenta el número de concurrencias de las posiciones, disminuye la proporción de TSS inusuales, principalmente los correspondientes a transcripción antisentido. Por supuesto 22 experimentos es una cantidad extraordinaria de exigencia para la detección certera de TSS, sin embargo ya desde un Índice de Detección de 3 se observa un cambio drástico tanto en el número de posiciones como en la proporción de transcripción antisentido (figura 3).

Lo anterior sugiere una tremenda variabilidad en el ribotipo de *E. coli* K12. Aún no hay un veredicto científico sobre si esta variabilidad en forma de transcripción ubicua constituye ruido biológico o si cumple una función celular[41]. Sin embargo, quizá antes de continuar buscando mayores significados a estos eventos inusuales habría que voltear a ver con un ojo más crítico los datos recientes en que se sustenta el concepto de transcripción ubicua, particularmente los provenientes de experimentos RNA-seq. Baste como ejemplo la experiencia obtenida por nuestro grupo: para los 22 experimentos mencionados en el párrafo anterior se prepararon las bibliotecas diligentemente siguiendo protocolos considerados estándar por otros grupos de investigación, a pesar de eso observamos una gran variabilidad en la proporción de datos desdeñables provenientes de cada experimento (figura 4). Particularmente llama la atención el hecho de que ni el proceso de eliminación de rRNA por hibridación con sondas magnéticas (RiboMinus™, de Life

Technologies) ni los protocolos de enriquecimiento basados en la 5' exoribonucleasa dependiente de monofosfatos o los adaptadores biotinilados son capaces de remover consistentemente los RNA ribosomales de la muestra inicial.

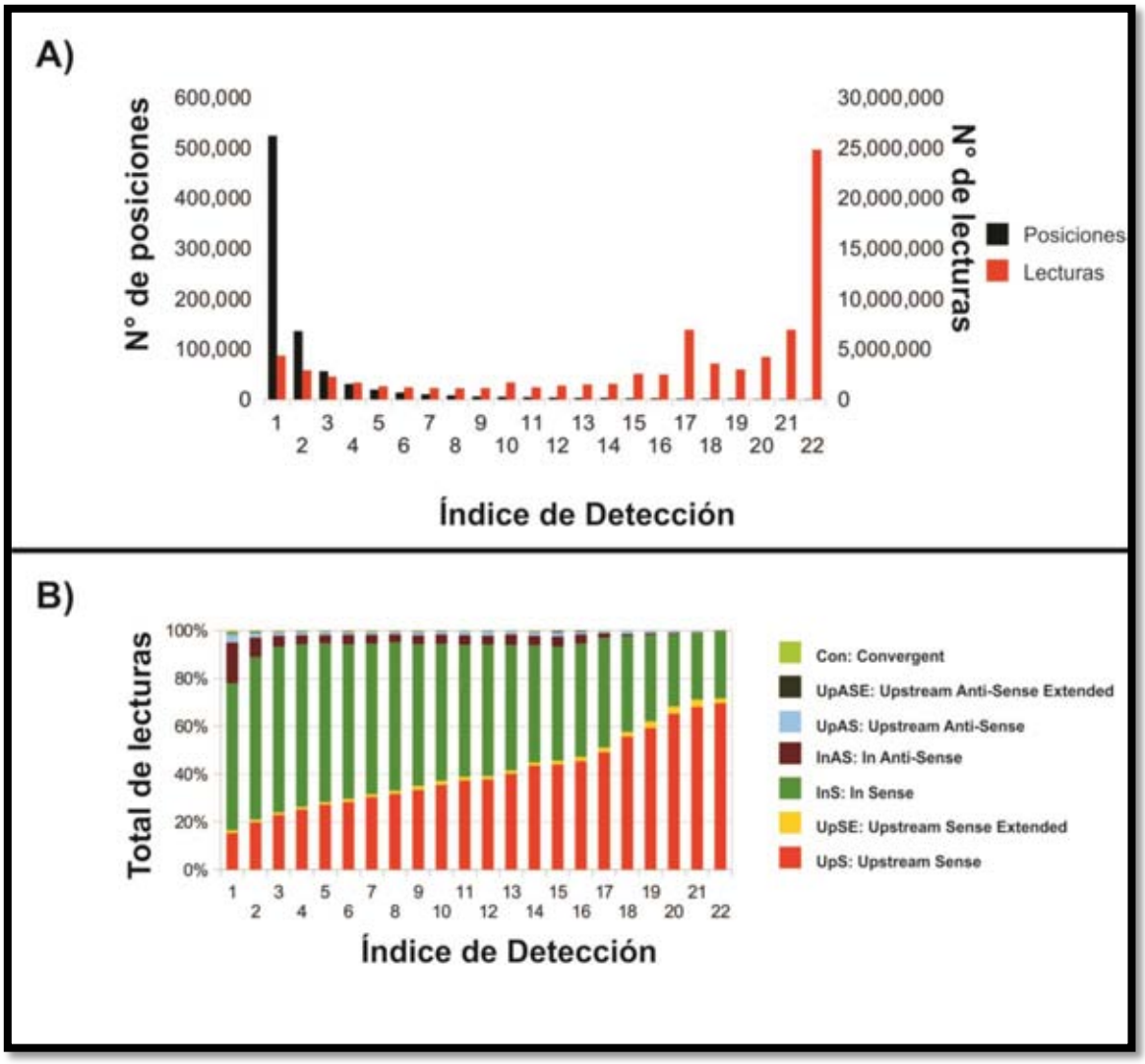


Figura 3. Transcripción ubicua en condiciones variadas de crecimiento en *E. coli* K12. A) Número de posiciones y lecturas en relación al Índice de Detección, se observan hasta aprox. 500,000 posiciones TSS presentes en sólo una de las condiciones de cultivo (Índice de Detección = 1) sondeadas de 22 librerías provenientes de *E.coli* K12 en diferentes condiciones de crecimiento y enriquecimiento de RNA. El número de TSS detectados disminuye conforme aumenta el Índice de Detección, mientras que el número de lecturas para dichos TSS aumenta. Esto sugiere que la transcripción ubicua es un evento altamente variable y poco reproducible. **B)** De manera similar, la transcripción antisentido (InAS, UpAS, UpASE) muestra mucha variabilidad entre experimentos ya que el porcentaje de TSS correspondientes a este tipo de transcritos disminuye conforme se exige un mayor Índice de Detección a las lecturas. (Datos no publicados).

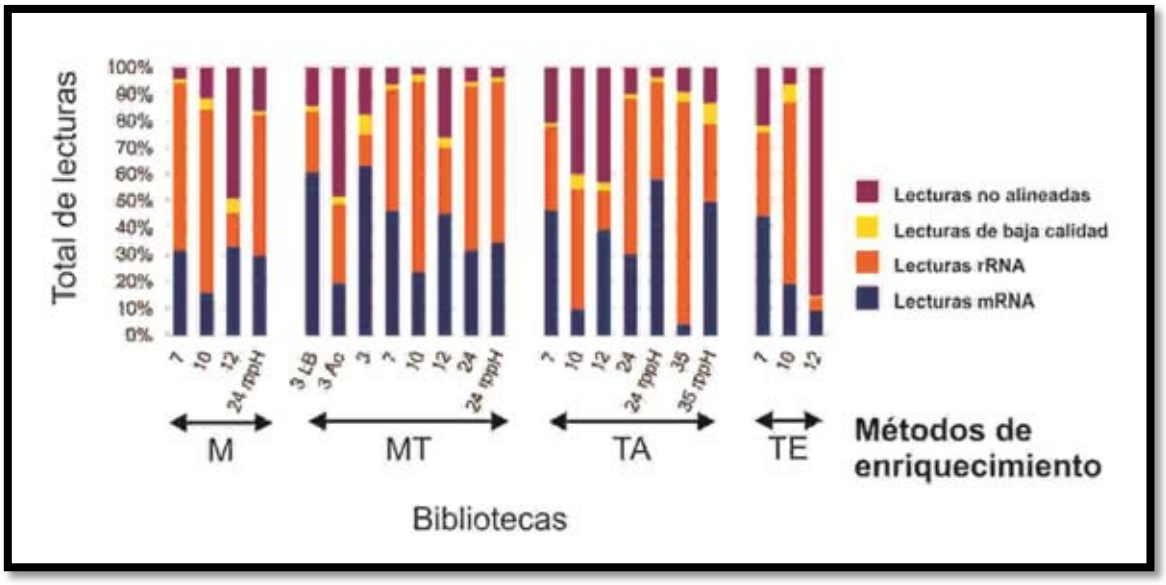


Figura 4. Eficiencia de las metodologías de enriquecimiento de RNA en *E. coli* K12. Se cuantificó el total de lecturas útiles en diferentes bibliotecas para detección masiva de TSS provenientes de muestras tratadas con distintos métodos de enriquecimiento: M = mRNA enriquecido con transcritos monofosfatados; MT = mRNA tri y mono fosfatado; TA = mRNA trifosfatado enriquecido con adaptadores biotinilados; TE = mRNA trifosfatado enriquecido con exoribonucleasa. A pesar de que todas las bibliotecas fueron tratadas para eliminar RNA ribosomal (rRNA) se siguen detectando proporciones importantes de estos junto con demás lecturas desdeñables (no alineadas o de baja calidad). (Datos no publicados).

Las observaciones anteriores ponen en cuestionamiento la eficiencia de los métodos comunes de enriquecimiento. Esto puede deberse, entre otras cosas, a que las estrategias utilizadas no desaparecen del todo el ruido de fondo por que las eficiencias de las reacciones implicadas (mediadas por una exoribonucleasa y una RNA ligasa, respectivamente) no están a la par con los requerimientos de sensibilidad de la técnica de secuenciación masiva.

Las metodologías de enriquecimiento ya mencionadas se consideran los estándares puesto que aunque el porcentaje de datos desdeñables puede llegar a ser significativo, la contraparte, los datos útiles, usualmente son considerados suficientes. Sin embargo hay que insistir en que vale la pena replantear la certeza de los datos obtenidos en este tipo de experimentos, principalmente por la variabilidad observada en la constitución de bibliotecas, como se muestra en la figura 4.

Ya que el proceso general de construcción de este tipo de bibliotecas es intrincado y conlleva varios pasos con probabilidad de introducir artificios reflejados en datos inútiles, consideramos que es imperativo explorar metodologías alternas de enriquecimiento con miras a mejorar la

consistencia de la construcción de bibliotecas RNA-seq para detección de TSS. Pensamos que una buena alternativa sería el aislamiento directo de transcritos primarios, en contraste con las metodologías de degradación o ligación de extremos 5' monofosfato. A partir de esa idea el concepto base fue diseñar una matriz cromatográfica capaz de capturar extremos 5' trifosfatados.

Bajo esa premisa realizamos una búsqueda bibliográfica de proteínas con capacidad de unión a transcritos primarios, encontrando dos cuya función depende del reconocimiento de extremos 5' trifosfatados (figura 5): eIF4E (un factor de inicio de la traducción eucariótico) y BdRppH (una pirofosfohidrolasa procariótica). En los siguientes apartados se profundizan las características que hicieron de estas proteínas las candidatas idóneas para el desarrollo de nuevas metodologías de enriquecimiento con transcritos primarios.

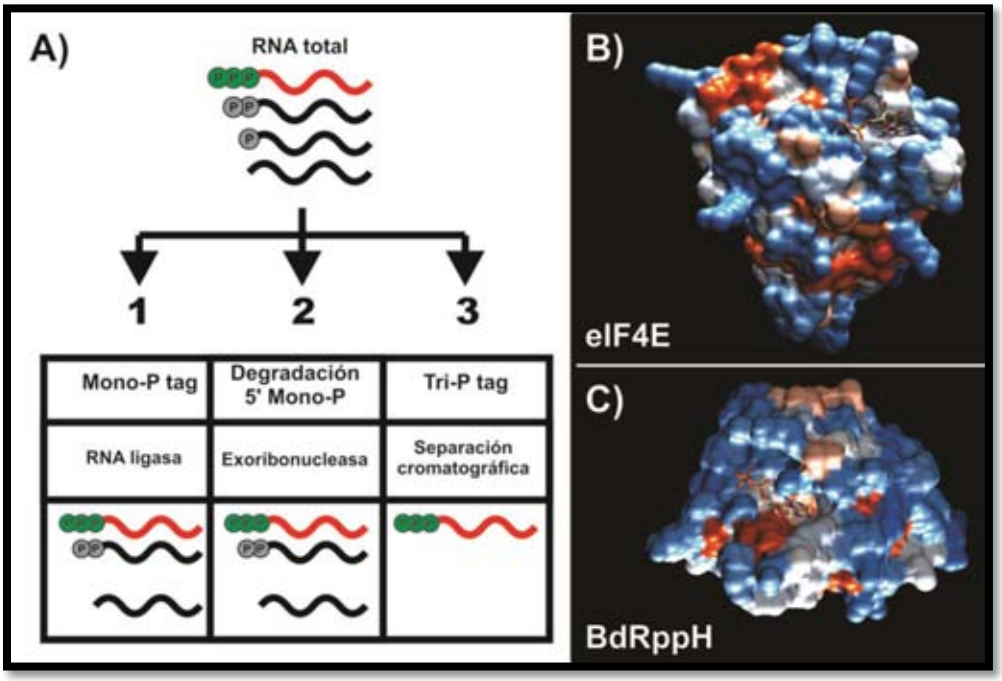


Figura 5. Estrategias de enriquecimiento de la muestra para detección de TSS. A) El RNA total extraído del microorganismo de interés puede ser procesado por diferentes métodos: 1) Mono-P tag: utilizando una RNA ligasa la muestra se somete a una reacción de ligación de adaptadores que sólo afecta transcritos monofosfatados, permitiendo su posterior eliminación; 2) Degradación 5' Mono-P: utilizando una exoribonucleasa dirigida a transcritos monofosfatados se eliminan estos de la muestra: 3) Nuestra metodología hipotética, denominada Tri-P tag, utilizando una proteína de afinidad por extremos 5' trifosfatados se podrían aislar directamente los transcritos primarios provenientes de muestras de RNA total. **B** & **C**) modelos cristalográficos de las proteínas con capacidad de reconocimiento de transcritos primarios eIF4E y BdRppH, ambas se cristalizaron unidas a un ribonucleótido trifosfatado.

3.2 Proteína de Unión a Transcritos Primarios eIF4E.

En células eucariotas, tan pronto el mRNA es sintetizado en el núcleo sufre una modificación en su extremo 5'. Molecularmente la modificación consiste en una 7-metilguanosina (m^7G) ligada por un puente 5'-5'-trifosfato hacia el primer nucleótido de un transcrito primario. El complejo enzimático de capping es el complejo proteico encargado de modificar exclusivamente transcritos primarios tan pronto son sintetizados por la RNA polimerasa II[7]. No se conoce ningún otro tipo de RNA polimerasa, ya sea eucariótica o procariótica, que produzca transcritos modificados con 5'-cap. Esta modificación en los mRNA eucarióticos es importante para su correcto procesamiento post-transcripcional (splicing, transporte nuclear, traducción, y estabilidad del transcrito). En particular, una de las funciones del 5'-cap en el proceso de traducción es servir de ancla para eIF4E, una proteína de 127 aminoácidos altamente conservada filogenéticamente[29] que forma parte del complejo de inicio de la traducción.

A su vez la función de eIF4E consiste en dirigir el complejo ribosomal hacia la estructura 5'-cap de los mRNA. A partir de las estructuras cristalográficas de eIF4E unida a un ligando 5'-cap diribonucléotido (m^7GpppG) Stolarski *et al*. determinaron que el sitio de interacción ligando-proteína es un surco en la superficie de eIF4E (figura 6). Según este modelo el reconocimiento de la m^7G está mediada principalmente por una interacción de apilamiento entre dos residuos Trp56 y Trp102, interacciones tipo Watson-Crick con Glu103 y Trp102, y contactos van der Waals con Trp166. También se observan interacciones significativas entre el puente trifosfato del dinucleótido y los residuos Trp102, Trp166, Arg112, Lys162 y Arg157[29].

Stolarki *et al.* también determinaron la contribución de varios elementos estructurales del 5'-cap en el reconocimiento eIF4E-cap. Según sus resultados propusieron un mecanismo de unión en dos pasos donde el puente 5'-trifosfato es el anclaje primario de esta interacción, que posteriormente facilita el acomodamiento de la estructura m^7G dentro del surco de reconocimiento del 5'-cap. Determinaron también que la identidad del segundo nucleótido unido a la estructura del 5'-cap no influye en la interacción[29]. Si bien en estos ensayos no se utilizaron transcritos RNA modificados sino diribonucleótidos, las estructuras cristalográficas dan buena idea de lo que podría ocurrir en la interacción eIF4E-transcrito por lo menos en el primer instante del reconocimiento proteína-RNA.

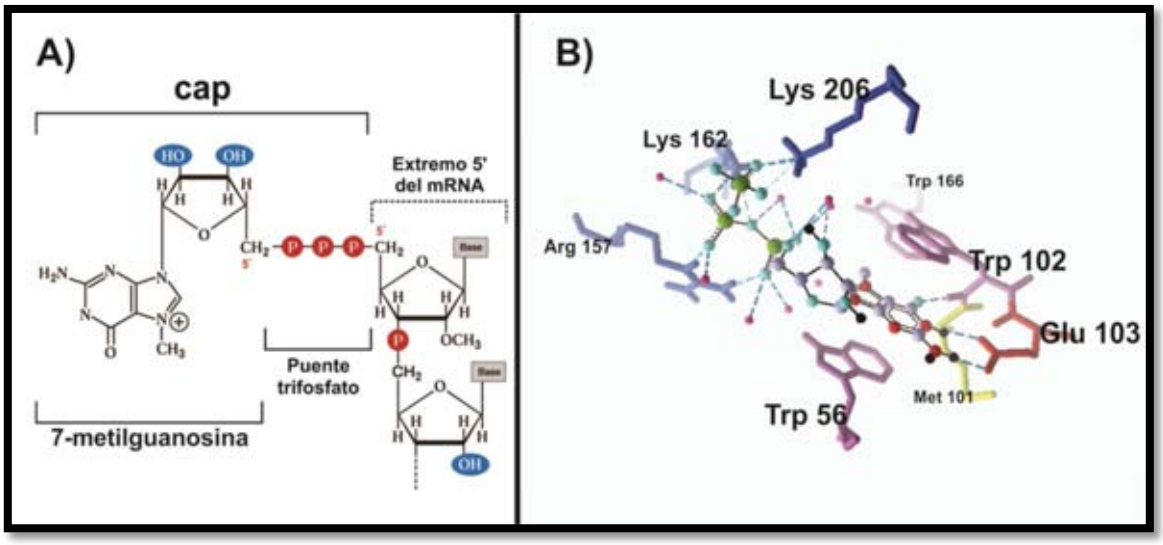


Figura 6. Estructura 5'-cap y su reconocimiento por eIF4E. A) Los extremos 5' de los transcritos primarios eucarióticos sintetizados por la RNA polimerasa II son inmediatamente modificados por el complejo enzimático de capping, el cual cataliza la unión covalente de una 7-metilguanosina con el extremo 5' del transcrito naciente a través de un puente trifosfato. (modificado de Asashima *et. al.*[1]). **B)** Contactos interatómicos en el complejo eIF4E-m^{7}GpppG determinados a partir de la estructura cristalográfica (PDB *accesion no*: 1L8B). Los puentes de hidrógeno están indicados con líneas punteadas; las moléculas de agua que intervienen en interacciones indirectas se muestran como pequeñas esferas moradas.

Una búsqueda más profunda de la bibliografía nos reveló trabajos previos y exitosos del diseño de matrices cromatográficas para aislar transcritos primarios eucarióticos utilizando eIF4E como proteína de afinidad. Primero Curt Hagedorn y luego Narry Kim reportaron el diseño de estas metodologías, con variaciones sutiles entre los protocolos empleados[8, 19]. La principal diferencia entre estos trabajos es que Hagedorn empleó una variante proteica, eIF4E(K119A), que presenta mayor afinidad por la estructura 5'-cap[13].

Por otro lado, Matsushima reportó una metodología de detección masiva de TSS aprovechando la condición 5´-cap de los transcritos primarios presentes en una línea celular de cáncer colorectal. Sin embargo en el protocolo reportado no se utilizó ningún filtro cromatográfico para enriquecer la muestra, más bien se aprovechó la cualidad protectora de la estructura 5'-cap contra la digestión enzimática de la fosfatasa alcalina bacteriana[15]. En este sentido la metodología de Matsushima también actúa principalmente sobre los transcritos secundarios, no los primarios.

Con todo lo anterior decidimos que eIF4E era una buena opción para el aislamiento directo de transcritos primarios en *E. coli*. Por supuesto el principal problema con que nos topamos fue que los transcritos procarióticos como los de *E. coli* no cuentan con la modificación 5'-cap. La solución a esto la encontramos en dos reportes: el primero Wahn *et al*., donde utilizando una reacción de capping *in vitro* se determinaron los sitios de inicio de la transcripción del precursor de rRNA 40S de *Xaenopus laevis*[33]; en el segundo reporte, de Rabinowitz *et al*., también se utiliza una reacción de capping *in vitro* para detectar los TSS en el genoma mitocondrial humano[28]. De manera similar a los transcritos de *E. coli*, tanto el precursor 40S rRNA de *X. laevis* como el RNA mitocondrial humano carecen de modificaciones 5'-cap puesto que no son sintetizados por la RNA polimerasa II. Por lo tanto la capacidad de llevar a cabo la reacción de capping *in vitro* nos sugirió la posibilidad de adecuar la muestra de RNA total procedente de *E. coli* a los requerimientos de las metodologías de enriquecimiento previamente reportadas por Hagedorn y Narry Kim.

3.3 Proteína de unión a transcritos primarios BdRppH.

La segunda opción en nuestra búsqueda por proteínas de unión a transcritos primarios resultó ser más arriesgada puesto que no existen publicaciones previas de diseños de matrices cromatográficas con ella. Encontramos reportes de una proteína con cualidades que le permitirían unirse a transcritos 5'-trifosfatados en procariotas sin necesidad de ninguna modificación elaborada en el extremo 5' del RNA: la enzima RppH, una pirofosfohidrolasa descrita originalmente en *E. coli* capaz de reconocer los extremos 5'-trifosfato en el mRNA para después convertirlos en 5'-monofosfato[10]. Irónicamente, se sospecha que esta enzima inicia la cascada de degradación *in vivo* de mRNA que genera parte del ruido de fondo causante del problema para la detección certera de TSS[2, 4].

¿Pero una cromatografía de afinidad con esta enzima no degradaría el extremo 5' de los RNA y los eluiría en vez de retenerlos? Quizá no, bajo las condiciones adecuadas. Dichas condiciones las encontramos en el trabajo de Messing *et al.* sobre BdRppH, la pirofosfohidrolasa de RNA presente en *Bdellovibrio bacteriovorus*[26].

Messing reporta la estructura cristalográfica del homodímero BdRppH unido a sustrato en condiciones de pH de 7.0, situación que inhibe la actividad hidrolítica de la enzima pero no afecta

las cargas de los aminoácidos en el sitio de unión a sustrato[26]. Según este mismo reporte la complementación con el gen *BdrppH* es capaz de restaurar el fenotipo de una mutante ΔEc*rppH* en *E. coli*. La región de identidad entre las secuencias aminoacídicas de EcRppH y BdRppH ocurre en un sitio interesante: el motivo Nudix, encargado de la catálisis del rompimiento y liberación de pirofosfato en extremos 5'-trifosfatados[23]. La figura 7 muestra el alineamiento de la región más similar en las tres RNA pirofosfohidrolasas bacterianas cristalizadas a la fecha: EcRppH de *E. coli*, BdRppH de *B. bacteriovorus,* y BsRppH de *B. subtilis*.

El modelo obtenido por cristalografía del complejo GTP-BdRppH revela una interacción entre el puente trifosfato del nucleótido y los residuos E70, E74 y G54, los tres conservados como parte del motivo Nudix. La interacción está coordinada por cationes Mg^{2+} y moléculas de agua (figura 7). La disposición observada en los contactos interatómicos es típica de proteínas Nudix al interaccionar con el sustrato[27].

Las proteínas de la superfamilia Nudix están relacionadas por dos características: el motivo de 23 residuos GXXXXXEXXXXXXXREUXEEXGU donde U puede ser I, L o V, y la presencia de un nucleótido polifosfato en sus sustratos, los cuales pueden ser compuestos nucleótido-azúcar, nucleótido-alcohol, dinucleótido-coenzima, dinucleósidos polifosfatos, deoxynucleótidos trifosfato, y RNA eucariótico o procariótico[27]. Por tanto el motivo Nudix además de un motivo de hidrólisis puede considerarse un motivo de interacción con grupos polifosfato. Aunque los demás dominios en una proteína tipo Nudix son los que confieren la especificidad del sustrato, según lo que se sabe de las 3 RNA pirofosfohidrolasas más estudiadas en bacterias el sustrato de estas enzimas parece ser RNA trifosfatado sin importar la secuencia nucleotídica del transcrito[3, 20, 26, 34].

Con estos precedentes decidimos desarrollar la no tan descabellada idea de una metodología cromatográfica para aislar transcritos primarios procarióticos utilizando la enzima BdRppH inactivada por pH en condiciones similares a las reportadas para su cristalización unida a sustrato.

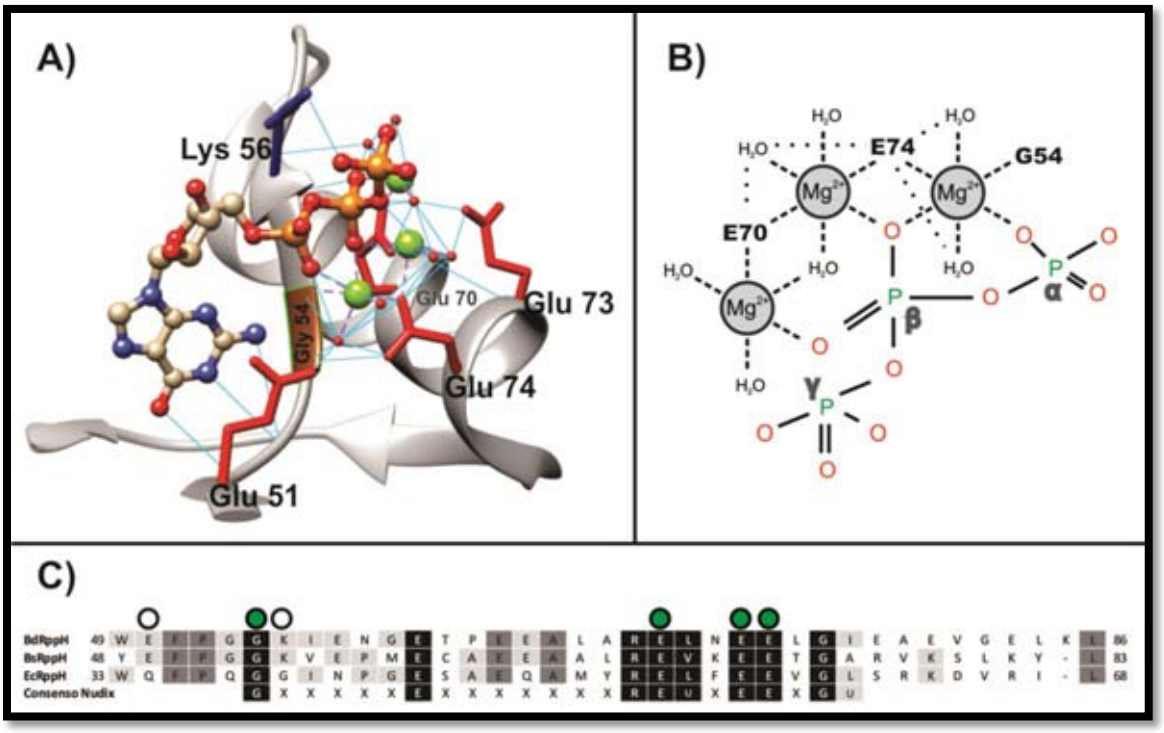


Figura 7. Reconocimiento del puente trifosfato por BdRppH. A) Interacción BdRppH-GTP (PDB *accesion no*: 3FFU). Los puentes de hidrógeno están indicados con líneas azules. De acuerdo a los datos cristalográficos BdRppH se une al puente trifosfato del sustrato a través de cationes Mg^{2+} (esferas verdes) coordinados por moléculas de agua (pequeñas bolas rojas) y los residuos E70, E74 y G54, todos presentes en el dominio Nudix. **B)** Interacción entre el puente trifostato, los cationes Mg^{2+}, los residuos del motivo Nudix y las moléculas H_2O. **C)** Alineamiento de la secuencia aminoacídica entre EcRppH, BdRppH, BsRppH y el consenso Nudix. Las secuencias pertenecen a las RNA pirofosfohidrolasas de *E. coli* (EcRppH), *B. bacteriovorus* (BdRppH) y *B. subtilis* (BsRppH). El 32-34% de identidad que comparten se concentra en una región de 35 a.a. cerca del extremo amino-terminal. Es en esta región donde las proteínas presentan el motivo Nudix, dominio encargado de la interacción con grupos polifosfatos para su posterior hidrólisis. X = cualquier aminoácido, U = I/L/V. Los círculos verdes señalan los residuos conservados del motivo Nudix implicados en la interacción con los grupos fosfato a través del Mg^{2+}; los círculos blancos señalan residuos no conservados en el motivo Nudix, pero que también están implicados en la interacción.

4. JUSTIFICACIÓN

El mapeo acertado de TSS en procariotas es una herramienta invaluable con diversas aplicaciones científicas y biotecnológicas. La secuenciación masiva de ácidos nucleicos es actualmente la mejor opción para realizar este tipo de estudios, pero se requieren diseños alternativos de metodologías de enriquecimiento con transcritos primarios para incrementar la calidad y facilitar el análisis de los datos obtenidos por este tipo de experimentos.

5. HIPÓTESIS

Es posible mapear acertadamente TSS en *E. coli* K12 utilizando transcritos primarios aislados mediante cromatografía de afinidad utilizando las proteínas de unión a extremos 5' trifosfatados eIF4E y BdRppH.

6. OBJETIVOS

- **General:**
  - Mapear fidedignamente TSS en *E. coli* K12.
- **Específicos:**
  - Validar la cromatografía de afinidad a eIF4E como un método para aislar transcritos primarios en *E. coli* K12.
  - Validar la cromatografía de afinidad a BdRppH como un método para aislar transcritos primarios en *E. coli* K12.

7. DESARROLLO EXPERIMENTAL

7.1 Purificación de la Proteína Recombinante GST-eIF4E.

El plásmido pGEXeIF4E nos fue donado amablemente por la Dr. Tzvetanka Dimitrova Dinkova, quien previamente lo ha reportado como vector de expresión de la proteína eIF4E de *Zea mays* unida covalentemente a la Glutatión S-transferasa (GST) [11]. El producto de la inducción de este vector en *E. coli* es la proteína de fusión GST-eIF4E de aproximadamente 50 kDa.

La proteína de fusión fue purificada por cromatografía de afinidad a glutatión reducido. La figura 8 muestra el análisis electroforético de las fracciones obtenidas en un experimento típico de purificación de GST-eIF4E.

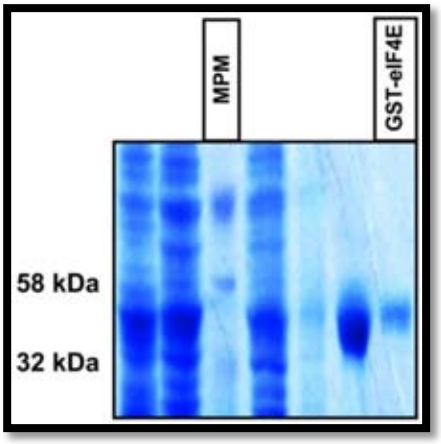


Figura 8. Purificación de GST-eIF4E. La proteína de fusión de aprox. 50 kDa se verificó mediante SDS-PAGE (acrilamida 15%, tinción Coomassie). MPM= marcador peso molecular. Los demás carriles son fracciones del proceso de purificación.

7.2 Clonación del Gen Bd*rppH* en el Vector de Expresión pGEX-4T-2.

Dado que no contábamos con ningún vector de expresión para la pirofosfohidrolasa BdRppH procedimos a construirlo como a continuación se describe.

El ORF reportado en GenBank como *bd0714* (*accession number* CAE78673) correspondiente al gen *BdrppH* de *Bdellovibrio bacteriovorus* se tomó como secuencia base para diseñar un gen de

expresión óptima en *E. coli*, cuya síntesis fue encargada a la compañía Epoch Life Science. Al extremo 5' del gen se agregaron dos sitios de corte para las enzimas de restricción BamHI y NdeI; al extremo 3' se agregó un codón de paro extra y la secuencia de corte para NotI (figura 9). La adaptación del uso de codones para expresión en *E. coli* estuvo a cargo del proveedor.

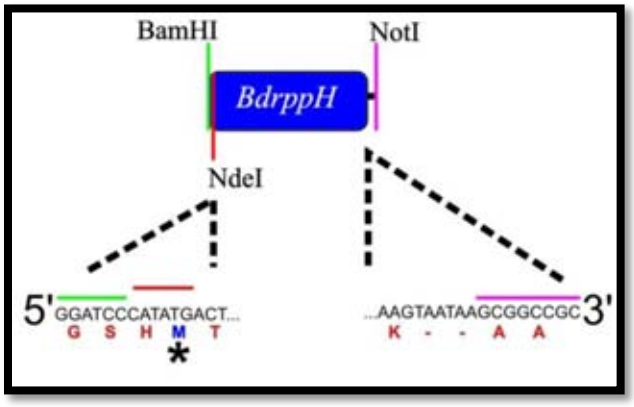


Figura 9. Rediseño de los extremos del gen *BdrppH* para facilitar su clonación posterior y la expresión óptima en *E. coli*. En 5' se agregaron los sitios de corte BamHI y NdeI, en 3' el sitio NotI; los sitios y el orden se eligieron tomando en cuenta su compatibilidad con el vector pGEX-4T-2 para futuras clonaciones. El asterisco (*) marca la metionina de inicio en la proteína.

Epoch Life Science entregó la construcción pBSK-BdrppH: un plásmido que contenía el ORF sintético basado en *BdrppH* (*synBdrppH*).

El fragmento *synBdrppH* se clonó en los sitios BamHI y NotI del vector pGEX-4T-2 (figura 10). La construcción se verificó por PCR utilizando los oligonucléotidos iGEX_Fw e iGEX_Rw (figura 11). Una de las clonas que dieron positivo a la prueba, denominada clona #4, se secuenció por el método de Sanger y el reporte reveló a *synBdrppH* insertado correctamente en pGEX-4T-2, con el ORF en fase y una identidad del 100% comparado con el gen sintetizado por Epoch Life Science. Esta construcción se denominó pGEXBdRppH y sería utilizada más adelante para purificar la proteína de fusión GST-BdRppH.

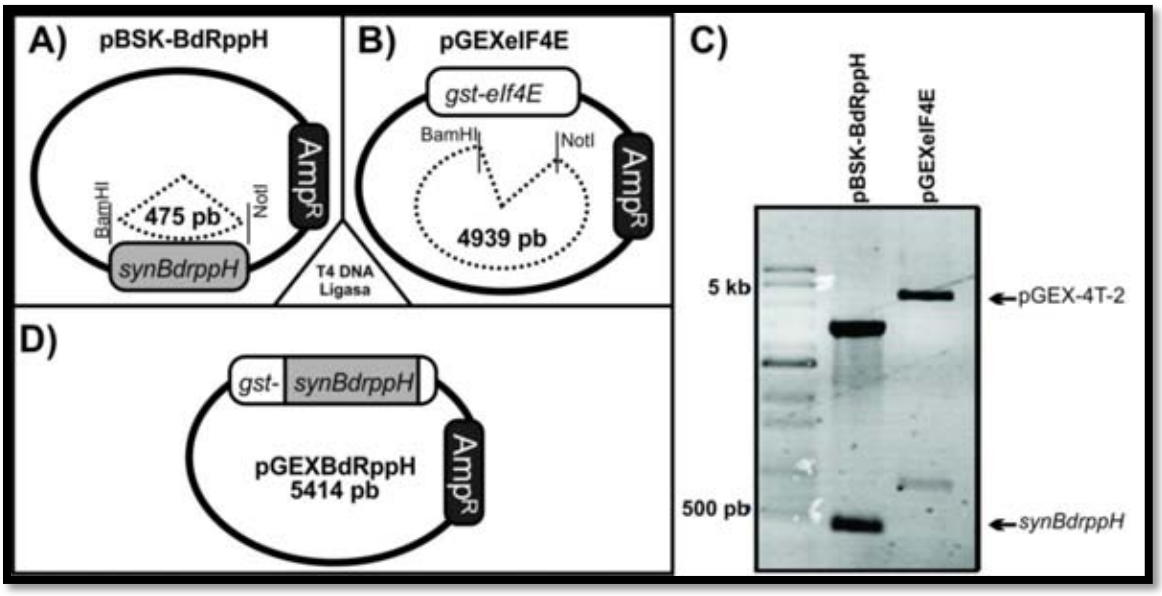


Figura 10. Construcción del plásmido pGEXBdrppH. Se digirieron los plásmidos **A)** pBSK-BdrppH y **B)** pGEXeIF4E con las enzimas BamHI y NotI. **C)** Se separaron las bandas resultantes en agarosa al 1% y se purificaron las correspondientes al fragmento *synBdrppH* (~500 pb) y al vector pGEX-4T-2 linearizado (~5 kb); la banda extra (~700 pb) en el carril pGEXeIF4E corresponde al fragmento *eIF4E* digerido. **D)** Se ligaron los fragmentos purificados utilizando T4 DNA ligasa, dando como resultado la construcción pGEXBdRppH.

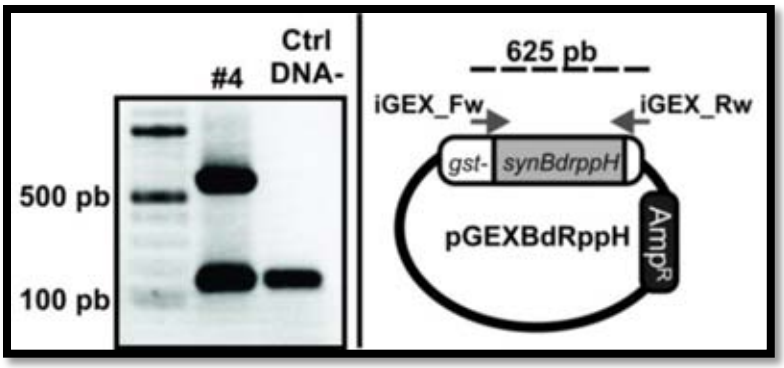


Figura 11. Análisis de PCR en colonia de la clona pGEXBdrppH. La construcción en la clona #4 se confirmó por reacción de PCR utilizando los oligos iGEX_Fw e iGEX_Rw. El producto se analizó en agarosa al 1%. Se observó el amplicón esperado de aprox. 625 pb. La banda de aprox. 100 pb resultó ser artificio experimental, como lo demuestra la reacción control sin DNA templado (Ctrl DNA-).

7.3 Purificación de la Proteína Recombinante GST-BdRppH.

La proteína de fusión GST-BdRppH fue expresada y posteriormente purificada por cromatografía de afinidad a glutatión reducido. La figura 12 muestra el análisis electroforético de las fracciones que obtuvimos en un experimento típico de purificación de GST-BdRppH.

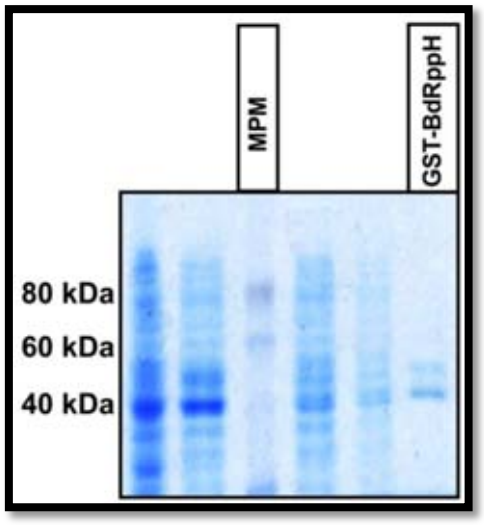


Figura 12. Purificación de GST-BdRppH. La proteína de fusión de aprox. 44 kDa se verificó mediante SDS-PAGE (acrilamida 15%, tinción Coomassie). MPM= marcador peso molecular. Los demás carriles son fracciones del proceso de purificación.

7.4 Ensayos de Enriquecimiento con Transcritos Trifosfatados.

7.4.1 Ensamblaje de las matrices de enriquecimiento.

La figura 13 muestra el resultado de la purificación simultánea de las proteínas GST-eIF4E, GST-BdRppH y GST (Glutatión-S-Transferasa expresada a partir del vector pGEX-4T-2). Observamos una banda extra de aprox. 25 kDa en las purificaciones de GST-eIF4E y de GST-BdRppH. Sospechamos que se trata del fragmento Glutatión S-transferasa liberado de la fusión pues su patrón de migración es muy similar al observado en el carril correspondiente a la proteína GST purificada.

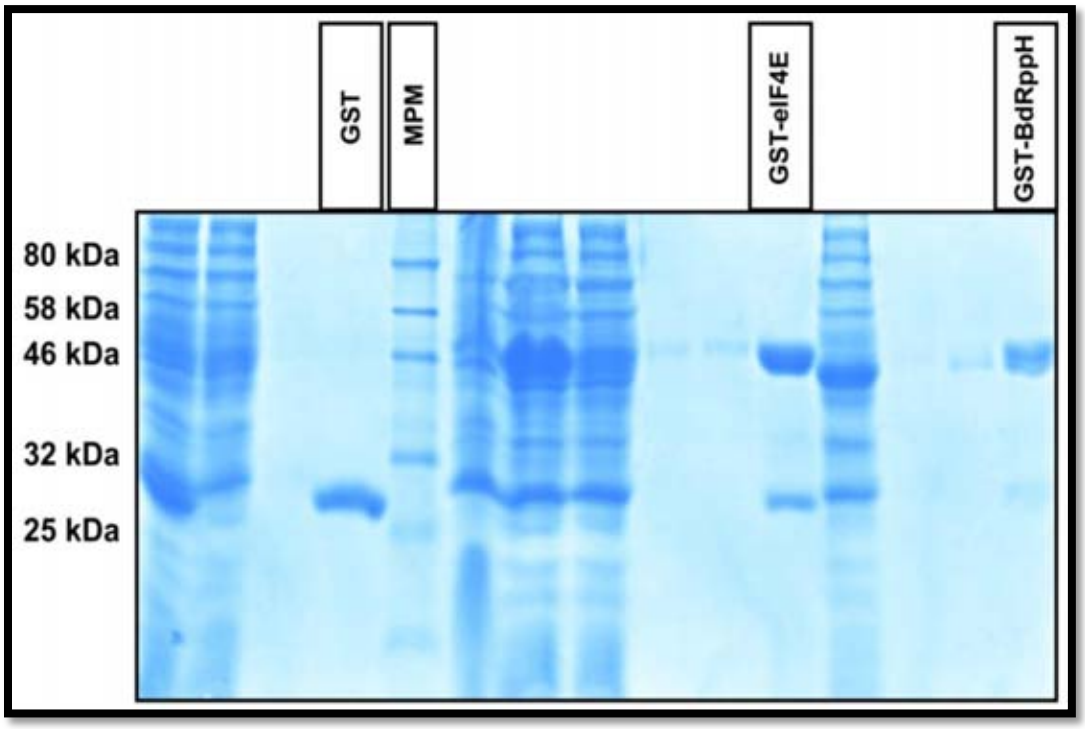


Figura 13. Purificación de proteínas de fusión GST. Se extrajeron las proteínas GST de aprox. 26 kDa, GST-eIF4E de aprox. 50 kDa, y GST-BdRppH de aprox. 44 kDa. Se verificaron mediante SDS-PAGE (acrilamida 15%, tinción Coomassie). MPM= marcador peso molecular. Los carriles donde no se señala nada son algunas fracciones del proceso de purificación de cada proteína.

El ensamblaje de las diferentes matrices se llevó acabo con cada una de las proteínas de la siguiente manera: a 1 mL de perlas de Agarosa-Glutatión se agregó 1 mg de la respectiva proteína purificada, luego se permitió la interacción durante una hora en buffer GST Bind/Wash 1X a

temperatura ambiente. El protocolo detallado se encuentra en la sección METODOLOGÍAS. Las matrices Agarosa-Glutatión-Proteína de Fusión se almacenaron a 4°C (durante no más de 4 semanas) hasta ser utilizadas.

7.4.2 Purificación de la muestra RNA.

Se extrajo RNA total utilizando el kit RNeasy (Qiagen) a partir de un cultivo de *E. coli* K12 crecido a 37 °C en LB hasta una OD_{600nm} = 0.8. El RNA se trató con DNasa I (Thermo Scientific) y la calidad de la muestra se analizó con el instrumento 2100 Bioanalyzer (Agilent). Se obtuvo una concentración de 509 ng/µl y una buena integridad de la muestra evidenciada por un valor RIN de 8.6 (figura 14).

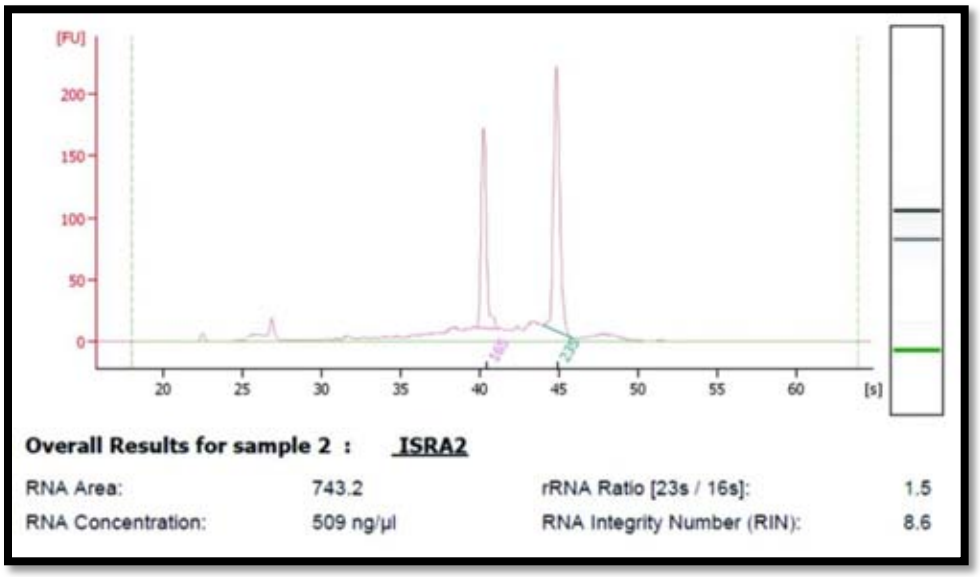


Figura 14. Purificación de la muestra de RNA. Se extrajo RNA total proveniente de *E. coli* K12 (37 °C en LB hasta una OD_{600nm} = 0.8). La muestra se revisó utilizando el instrumento 2100 Bionalyzer de Agilent. La concentración y calidad de la muestra fueron óptimos.

7.4.3 Enriquecimiento con trifosfatos a partir de RNA total.

A) Cromatografía de afinidad a eIF4E (CAPture).

Se tomaron 60 ug de RNA total de *E. coli* K12 y se sometieron a una reacción de capping *in vitro* utilizando simultáneamente los kits ScriptCap™ m^7G Capping System y ScriptCap™ 2'-O-Methyltransferase siguiendo las indicaciones del proveedor (Epicentre). Las enzimas presentes en estos kits llevan a cabo una reacción que convierte transcritos 5'-trifosfatados en transcritos 5'-cap.

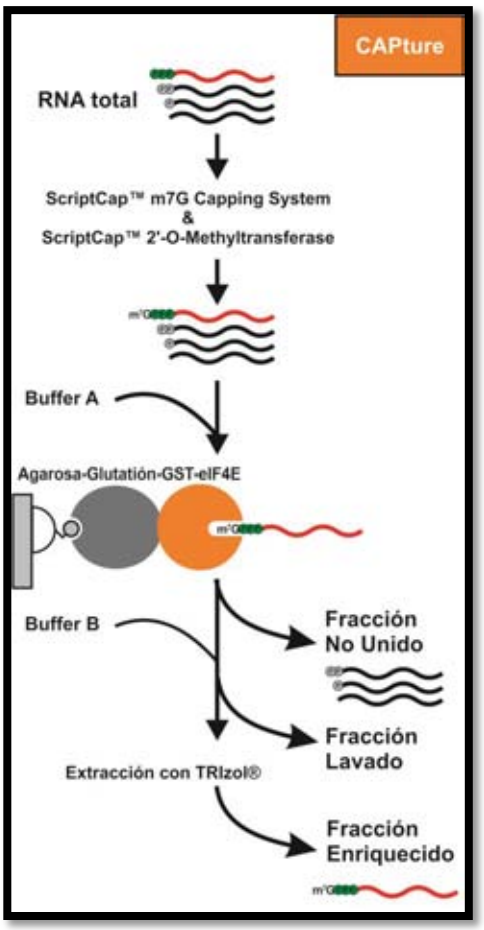


Figura 15. Metodología de enriquecimiento CAPture.

Se utilizó una alícuota de 100 µl de resina Agarosa-Glutatión-GST-eIF4E para enriquecer 60 ug de RNA CAPeado. Brevemente, el proceso fue el siguiente (figura 15): se mezcló el RNA CAPeado con la matriz, se agregó buffer de unión (buffer A) y se dejó la muestra en movimiento por inversión durante 1 hora a temperatura ambiente; luego se procedió a lavar 2 veces por centrifugación y resuspensión en buffer de lavado (buffer B); finalmente se centrifugó la muestra y se utilizó TRIzol® (Life Technologies) para extraer el RNA enriquecido, separándolo de la matriz agarosa-proteína de fusión. Las diferentes fracciones obtenidas (fracción No Unido, fracción Lavado, fracción Enriquecido) fueron precipitadas con isopropanol y procesadas para concentrar el RNA presente. Las muestras resultantes se resuspendieron en agua libre de RNasas y se almacenaron a -70 °C.

Paralelamente, se preparó otro enriquecimiento siguiendo el mismo protocolo pero utilizando 60 ug de RNA total no CAPeado. Las muestras resultantes también se resuspendieron en agua libre de RNasas y se almacenaron a -70 °C.

B) Cromatografía de afinidad a BdRppH (pH-Switch).

Una alícuota de 100 µl de resina Agarosa-Glutatión-GST-BdRppH se utilizó para enriquecer 60 ug de RNA total de *E. coli* K12. Brevemente, el proceso fue el siguiente (figura 16): se mezcló el RNA con la matriz, se agregó buffer de unión (buffer U, pH 7.0) y se dejó la muestra en movimiento por 1 hora a temperatura ambiente; luego se procedió a lavar 2 veces por centrifugación y resuspensión en buffer de lavado (buffer U, pH 7.0); después se centrifugó la muestra, se resuspendió en buffer de hidrólisis (buffer H, pH 8.5) y se dejó la mezcla en movimiento a 37 °C por 1 hora para provocar la actividad pirofosfohidrolasa de la enzima BdRppH; la muestra se centrifugó y el sobrenadante se guardó como la fracción enriquecida. Las diferentes fracciones obtenidas (fracción No Unido, fracción Lavado, fracción Enriquecido) fueron precipitadas con isopropanol y procesadas para concentrar el RNA presente. Las muestras resultantes se resuspendieron en agua libre de RNasas y se almacenaron a -70 °C.

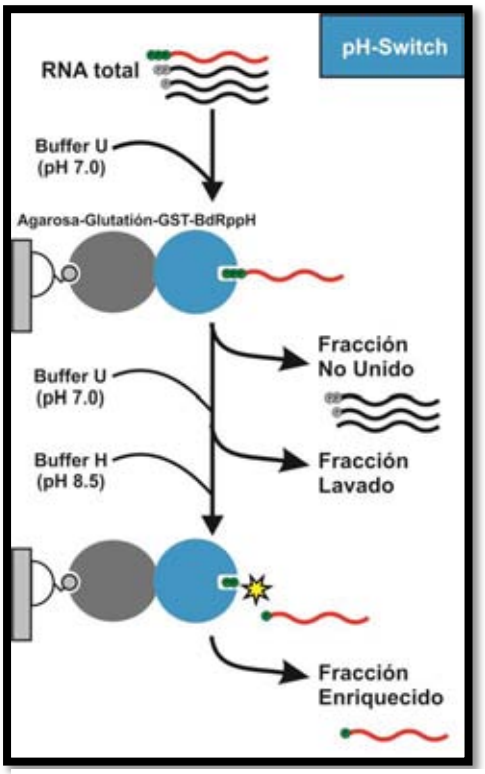


Figura 16. Metodología de enriquecimiento pH-Switch.

C) Controles negativos.

Para descartar la contaminación con RNA procedente del cultivo de *E. coli* a partir del cual se extrajeron las proteínas de fusión utilizadas en las matrices de enriquecimiento, se realizaron controles negativos utilizando 100 µl de agua libre de RNasas como muestra inicial para las metodologías CAPture y pH-Switch en experimentos separados.

También se prepararon matrices Agarosa-Glutatión-GST. Esta matriz se utilizó para enriquecer 60 ug de RNA total siguiendo las metodologías CAPture y pH-Switch. Dichas reacciones control se diseñaron para analizar la capacidad intrínseca de la proteína Glutatión S-transferasa para capturar RNA sin necesidad de eIF4E o BdRppH.

7.5 Efectividad de los Distintos Métodos de Enriquecimiento.

7.5.1 Detección de RNA en las diferentes fracciones obtenidas por las metodologías CAPture y pH-Switch.

Se comprobó la capacidad para retener RNA de las metodologías de enriquecimiento CAPture y pH-Switch. Se analizaron sus respectivas fracciones 'No Unido' (lo que desde antes del primer lavado no interaccionó con la matriz agarosa-proteína de fusión) y 'Enriquecido' (lo que se recuperó al final). El análisis se llevó a cabo con el instrumento 2100 Bionalyzer de Agilent. La figura 17 muestra los resultados. Las bandas observadas entre 1000 nt y 4000 nt en el carril 1 corresponden a los rRNA 16S (aprox. 1500 pb) y 23S (aprox. 2900 pb) presentes en la muestra de RNA total que se utilizó como material de inicio para los enriquecimientos. La banda gruesa de entre 25 nt y 200 nt presente en los demás carriles probablemente corresponde al tRNA de levadura que se utiliza en los buffers de las dos distintas metodologías para evitar interacciones inespecíficas RNA-proteína de fusión.

En cuanto a la capacidad de retención de RNA, las fracciones enriquecidas (figura 17A carriles 3 y 7) denotan la presencia de RNA en las muestras obtenidas por las metodologías CAPture y pH-Switch respectivamente. Algo interesante que observamos fue que la muestra de RNA total que no se sometió a la reacción de Capping *in vitro* también fue retenida por la matriz GST-eIF4E (17A carril 5). Las muestras correspondientes a las fracciones No Unidas (16A carriles 2, 4 y 6) también contienen RNA. Los controles negativos (figura 17B) demuestran que no hay contaminación de RNA proveniente del proceso de purificación de proteínas, puesto que no se detecta nada en los enriquecimientos donde no se utilizó RNA total de entrada (17B carriles 2-5). También se observa que la resina Agarosa-GST por sí misma no posee capacidad de retener RNA ya que la muestra inicial es recuperada inmediatamente en la fracción no unida sin importar si se utiliza la metodología CAPture (17B carril 6-8) o pH-Switch (17B carril 9-11).

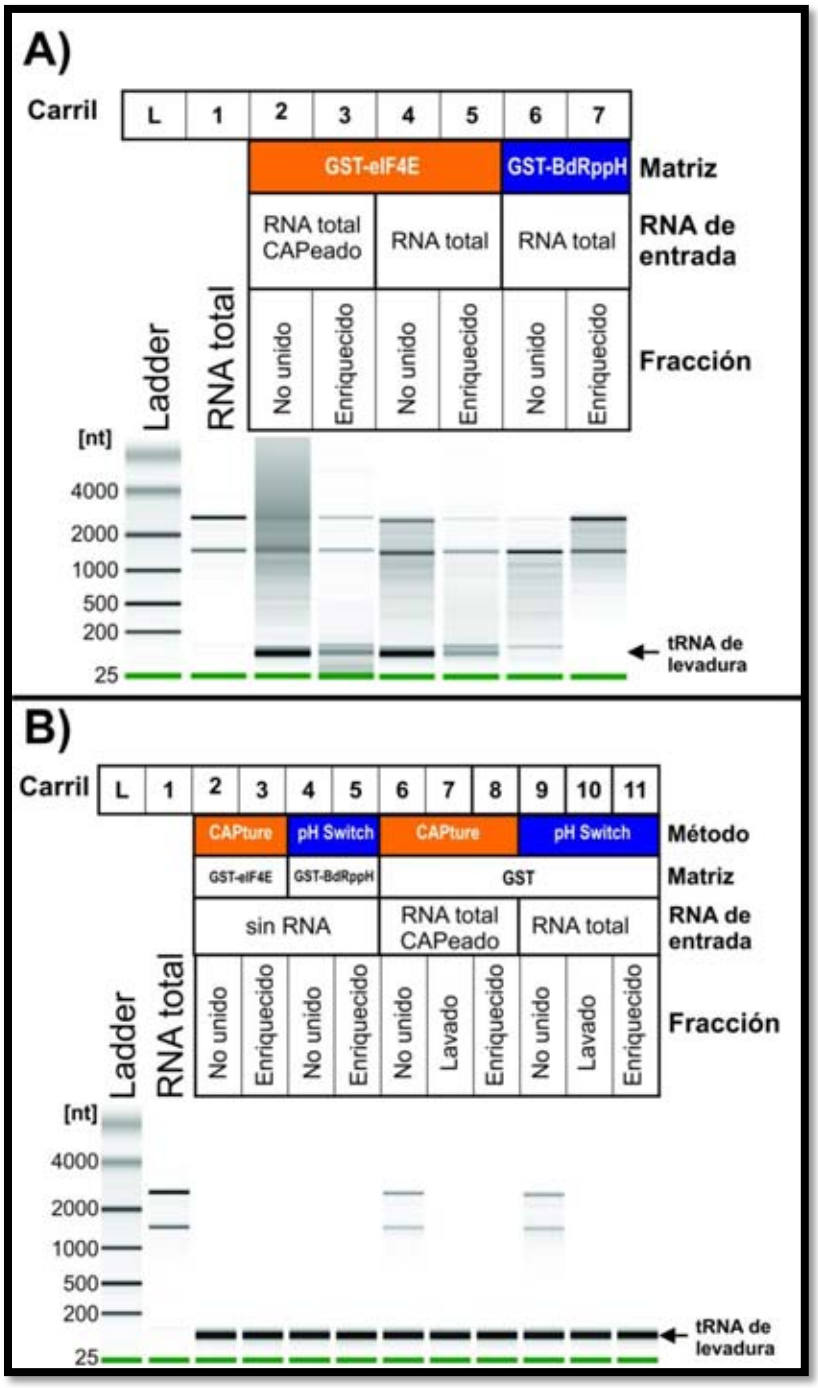


Figura 17. Capacidad de las técnicas CAPture y pH Switch para capturar RNA. **A)** las fracciones enriquecidas utilizando las matrices basadas en las proteínas eIF4E (carril 3 y 5) y BdRppH (carril 7) muestran la presencia de RNA. **B)** en las fracciones enriquecidas a partir del experimento control donde no se utilizó RNA como muestra inicial (carriles 2-5) no se detecta la existencia de RNA. Tampoco se detecta la presencia de RNA en las fracciones enriquecidas con matrices Agarosa-GST bajo ninguna de las condiciones, ya sea la técnica CAPture (carril 8) o la técnica pH Switch (carril 11).

7.5.2 Detección de sitios de inicio de transcripción.

Para verificar la utilidad de las metodologías desarrolladas se realizaron ensayos 5’ RACE. Brevemente, la preparación de las bibliotecas cDNA para cada una de las diferentes fracciones obtenidas en los ensayos de enriquecimiento se realizó de la siguiente manera (figura 18): 50 ng de RNA fueron sometidos a la reacción de transcripción reversa con la enzima Superscript™ II (Invitrogen) utilizando el oligonucleótido sRNAP1 que incluye una secuencia de 6 nucleótidos al azar en su extremo 3’ (5’-caagcagaagacggcatacgannnnnn-3’). Estas primeras bibliotecas cDNA de cadena sencilla se sometieron a una reacción de poliadenilación con la enzima Terminal Transferasa (Roche). Luego se generó la segunda cadena cDNA utilizando el oligonucleótido RACE1 (5’-gactcgagtcgacatcgattttttttttttttttt-3’) en una reacción de PCR con 5 ciclos de amplificación. Finalmente, la biblioteca cDNA de doble cadena fue reamplificada por PCR utilizando los oligos sRNAPCR1 (5’-caagcagaagacggcatacga-3’) y RACE2 (5’-gactcgagtcgacatcga-3’). El protocolo detallado se encuentra en la sección METODOLOGÍAS.

Figura 18. Construcción de bibliotecas cDNA para ensayos 5’ RACE.

A partir de estas bibliotecas DNA doble cadena se sondeó mediante PCR en busca de amplicones correspondientes a extremos 5’ de distintos transcritos previamente detectados por nuestro grupo en experimentos de secuenciación masiva[25, 35]. Para estos primeros ensayos se utilizaron condiciones especiales de PCR durante el proceso de construcción de bibliotecas, duplicando el número estándar de ciclos de replicación recomendado para ese tipo de experimentos: en el paso de síntesis de segunda cadena de cDNA se utilizaron 10 ciclos de PCR, en el paso de reamplificación de la biblioteca por PCR se utilizaron 20 ciclos, y finalmente en el paso de detección dirigida de TSS por

5’ RACE se utilizaron 35 ciclos de PCR. Estas condiciones bruscas de amplificación fueron utilizadas para compensar los diferentes niveles de expresión de los genes blanco, con la intención de detectar hasta los de mínima expresión. Algunos de los genes sondeados fueron: *rpsB, nirB, xylA, fruA, csrB, galE, degS, ppa, secA, gatY, ompA, nrdD*. Además de las diferencias en niveles de expresión, estos genes fueron elegidos porque previamente nuestro grupo los había detectado ya sea por ensayos 5’ RACE o por secuenciación masiva. El perfil 5’ RACE obtenido de cada una de las bibliotecas enriquecidas fue analizado por separación electroforética (figura 19). De acuerdo a trabajos previos en el grupo, se esperaban bandas de entre 500 y 150 pb. Sólo el perfil 5’ RACE correspondiente a la biblioteca enriquecida mediante la técnica CAPture a partir de RNA total no CAPeado mostró un patrón de bandeo definido (figura 19A), sin embargo ninguna de las bandas extraídas a partir de este gel resultaron en secuencias completamente legibles. Los perfiles 5’ RACE correspondientes a las bibliotecas CAPture a partir de RNA CAPeado (figura 19B) y pH-Switch (figura 19C) mostraron patrones difusos de bandeo y las secuencias obtenidas fueron también ilegibles.

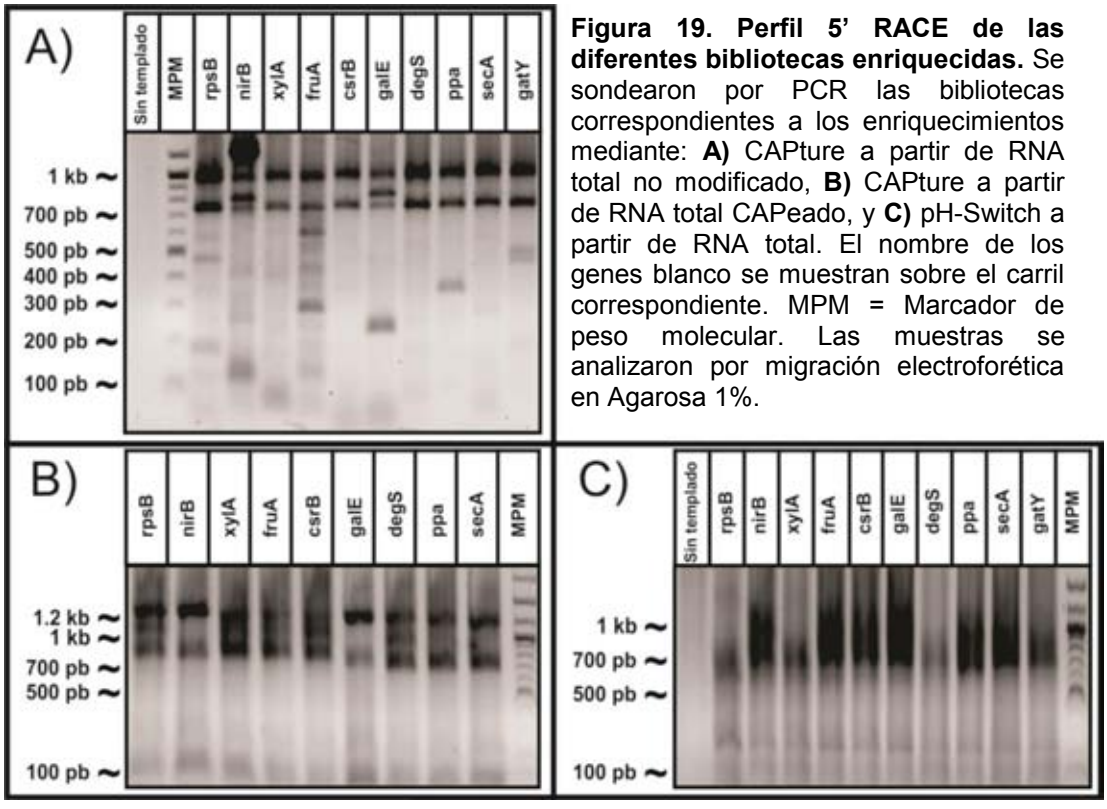


Figura 19. Perfil 5’ RACE de las diferentes bibliotecas enriquecidas. Se sondearon por PCR las bibliotecas correspondientes a los enriquecimientos mediante: **A)** CAPture a partir de RNA total no modificado, **B)** CAPture a partir de RNA total CAPeado, y **C)** pH-Switch a partir de RNA total. El nombre de los genes blanco se muestran sobre el carril correspondiente. MPM = Marcador de peso molecular. Las muestras se analizaron por migración electroforética en Agarosa 1%.

La mala calidad de los perfiles 5’ RACE obtenidos fue atribuida a las condiciones extremas de amplificación utilizadas durante la construcción de las bibliotecas. Creemos que el alto número de ciclos de amplificación favorece la creación de quimeras a partir de la unión entre los oligos adaptadores sRNAP1 y RACE1 debido a la naturaleza laxa de sus extremos 3’.

Para sortear estos problemas, utilizando el gen ppa como prueba, se afinaron las condiciones de amplificación de la siguiente manera: se utilizaron sólo 5 ciclos para la síntesis de la primera cadena cDNA, 15 ciclos para la síntesis de segunda cadena de cDNA, y 20 ciclos para la detección dirigida de TSS por 5’ RACE. El tiempo de extensión y la temperatura de alineamiento en el paso de detección dirigida de TSS se ajustaron hasta las condiciones más óptimas mediante PCR en gradiente. La figura 20 muestra un ensayo 5’ RACE más limpio si se compara con el carril correspondiente a ppa en la figura 19A.

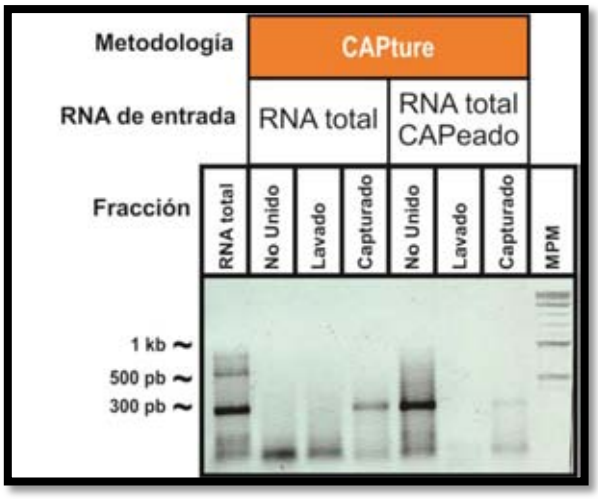


Figura 20. Detección de ppa por 5’ RACE. Se sondearon las distintas fracciones de las bibliotecas enriquecidas por la metodología CAPture en busca de un amplicón de aprox. 250 pb. Las condiciones más específicas de amplificación con que se llevó a cabo este ensayo permitieron obtener resultados más limpios en comparación con los observados en la figura 19A y 19B.

Siguiendo estos protocolos de refinamiento, determinamos las condiciones óptimas para la detección por 5’ RACE de los blancos OmpA y rRNA23S. Estos dos blancos fueron escogidos debido a que *ompA* es el único gen de su unidad transcripcional, por lo cual esperamos que el transcrito original sintetizado a partir de los promotores ompAp1 u ompAp2 sea un transcrito primario[9, 35]; por su parte el rRNA 23S proviene del procesamiento de un transcrito mayor que originalmente contiene los otros rRNAs (16S y 5S) además de algunas secuencias tRNA. Aunque en *E. coli* existen 7 operones distintos que codifican para el precursor de los rRNAs, todos estos transcritos precursores son procesados por RNasas para producir los rRNAs maduros. Debido a este paso de maduración podemos considerar el rRNA 23S un transcrito secundario no trifosfatado[14]. Con lo anterior fue posible cuestionar más a detalle la capacidad de nuestros métodos de

enriquecimiento para discernir entre transcritos primarios y productos del procesamiento de RNA. La figura 21 muestra las secuencias de los oligonucleótidos utilizados para el ensayo y la longitud del amplicón esperado según el contexto genómico de estos genes anotados en RegulonDB[35].

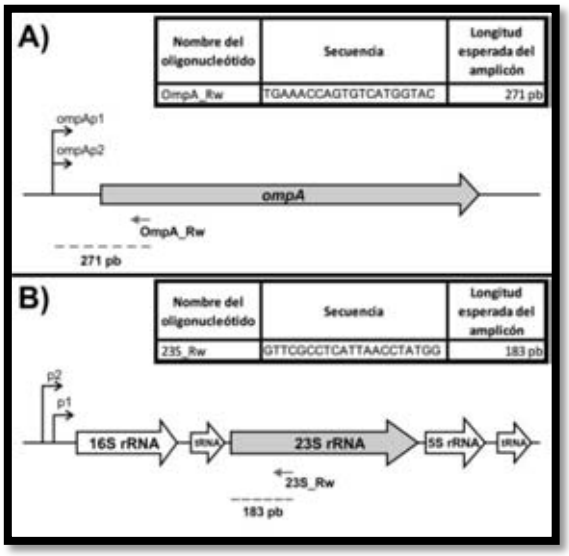


Figura 21. Unidades transcripcionales de los genes *ompA* y 23S rRNA en *E. coli* K12. A) *ompA* es el único gen presente en su unidad transcripcional, por tanto se espera que el oligo OmpA_Rw produzca un amplicón de aprox. 271 pb al extenderse hasta cualquiera de los dos TSS reportados (transcritos a partir de ompAp1 u ompAp2). **B)** los genes que codifican para 23S rRNA (*rrlA*, *rrlB*, *rrlC*, *rrlD*, *rrlE*, *rrlG*, *rrlH*) forman cada uno parte de su propia unidad transcripcional, la cual tras ser procesada libera el 23S rRNA maduro. Se espera que el oligo 23S_Rw produzca un amplicón de 183 pb al extenderse hasta el extremo 5' del 23S rRNA maduro.

Los productos PCR provenientes de los enriquecimientos se analizaron por migración electroforética en agarosa 1%. A continuación las figuras 22, 23 y 24 muestran los amplicones detectados en cada fracción de los diferentes enriquecimientos. Como control, se utilizó RNA total (no enriquecido) proveniente de *E. coli* K12.

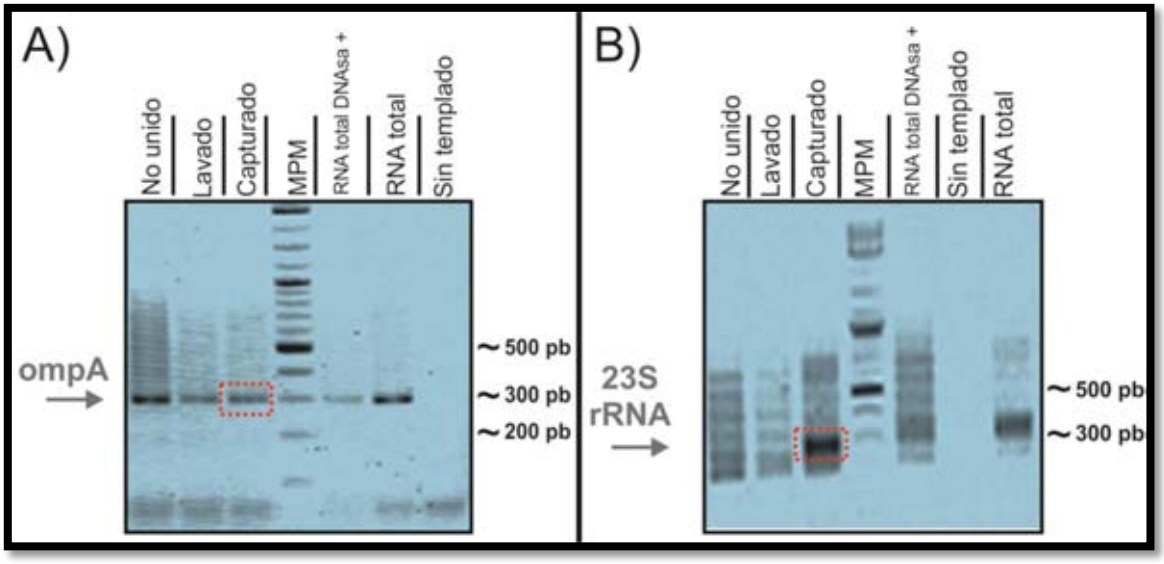


Figura 22. Enriquecimientos a partir de RNA total utilizando la técnica pH-Switch. Las diferentes fracciones se sondearon por 5' RACE teniendo como blanco los transcritos **A)**

ompA y **B)** 23S rRNA. Los amplicones señalados con un cuadro punteado presentes en la fracción capturada fueron purificados y secuenciados.

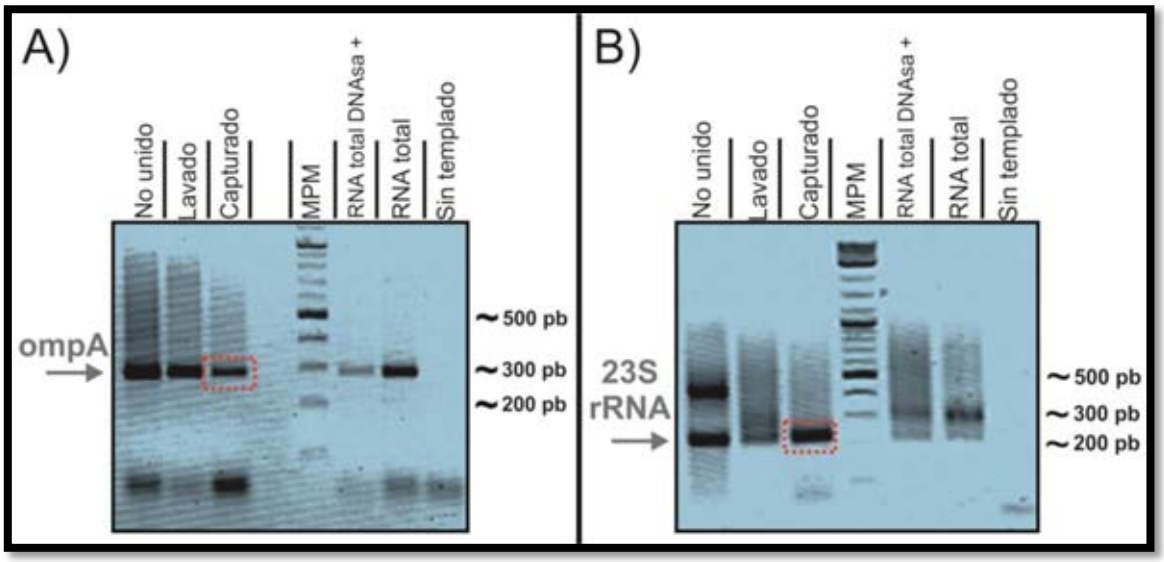


Figura 23. Enriquecimientos a partir de RNA total CAPeado utilizando la técnica CAPture. Las diferentes fracciones se sondearon por 5' RACE teniendo como blanco los transcritos **A)** ompA y **B)** 23S rRNA. Los amplicones señalados con un cuadro punteado presentes en la fracción capturada fueron purificados y secuenciados.

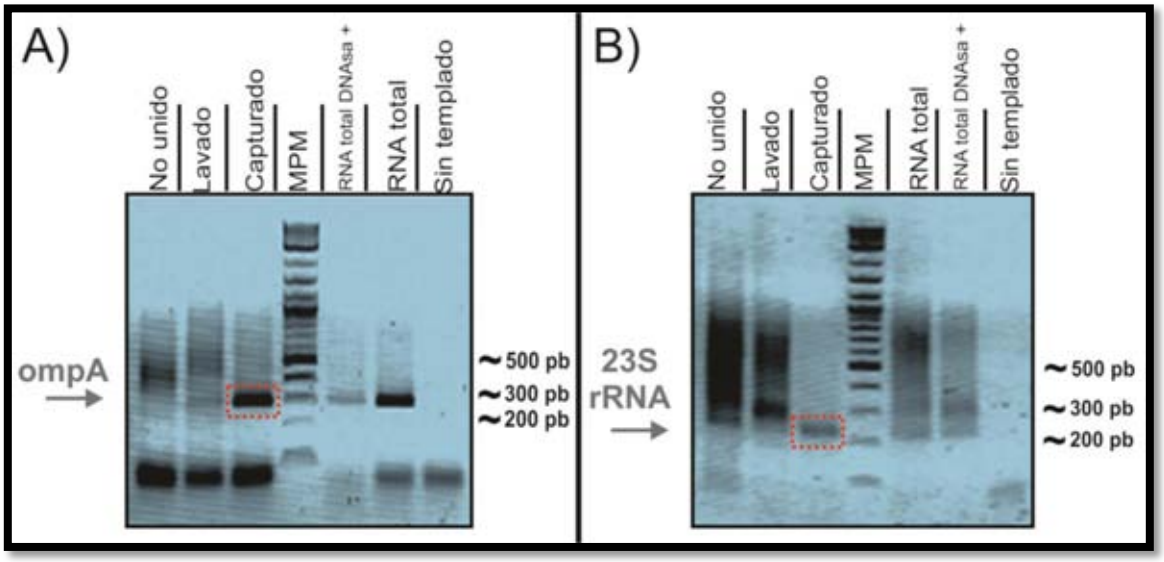


Figura 24. Enriquecimientos a partir de RNA total no modificado utilizando la técnica CAPture. Las diferentes fracciones se sondearon por 5' RACE teniendo como blanco los transcritos **A)** ompA y **B)** 23S rRNA. Los amplicones señalados con un cuadro punteado presentes en la fracción capturada fueron purificados y secuenciados.

En el caso de ompA logramos obtener bandas definidas en las fracciones enriquecidas de los tres tratamientos. Los amplicones correspondían en su migración electroforética con los presentes en las muestras control de RNA total no enriquecido tanto tratado con DNasa como no tratado. Por otro lado, para 23S rRNA también obtuvimos bandas definidas en las fracciones enriquecidas, pero estas no correspondían del todo en su patrón de migración electroforética con las presentes en las muestras control. Cabe recordar que en *E. coli*, 23S rRNA está presente en 7 transcritos policistrónicos diferentes, lo cual podría explicar el patrón difuso en los controles de RNA total. Nuestros resultados sugieren que la madeja de transcritos que contienen la secuencia 23S rRNA originalmente presentes en la muestra de RNA total fueron afectados por las técnicas de enriquecimiento de tal manera que una isoforma del transcrito fue filtrada del resto.

Para profundizar en el análisis se purificaron y secuenciaron los amplicones presentes en las fracciones capturadas con la intención de determinar la secuencia exacta del extremo 5' de los transcritos blanco. Las secuencias obtenidas se compararon con las regiones genómicas de *ompA* y *rrlA* (uno de los 7 genes codificantes para 23S rRNA) en *E. coli* K12. En los tres métodos de enriquecimiento empleados, los transcritos ompA capturados coincidieron en la posición de inicio de transcripción a 1 nucleótido río abajo del TSS más próximo reportado en RegulónDB. La figura 25 muestra el alineamiento de las secuencias ompA obtenidas.

Figura 25. Alineamiento de las secuencias obtenidas por 5' RACE contra ompA a partir de los enriquecimientos con las técnicas CAPture y pH-Switch. *ompA* = región genómica de *E. coli* K12 que codifica para el transcrito primario ompA; CAP1 = secuencia obtenida a partir del enriquecimiento por la técnica CAPture utilizando RNA total CAPeado; CAP2 = secuencia obtenida a partir del enriquecimiento por la técnica CAPture utilizando RNA total no CAPeado. pH-S= secuencia obtenida a partir del enriquecimiento por la técnica pH-Switch utilizando RNA total. Los asteriscos (*) señalan TSS previamente reportados en RegulonDB; +1 señala el TSS detectado en nuestros experimentos.

Respecto a 23S rRNA, no nos fue posible obtener una secuencia útil en el caso del enriquecimiento por la técnica pH-Switch, aún tras dos repeticiones del experimento. Las secuencias obtenidas fueron crípticas pues no alineaban con ninguna región del genoma de *E. coli* ni parecían ser un

constructo quimérico producto de la fusión de los oligonucleótidos utilizados durante la construcción de la biblioteca. En contraste, los amplicones enriquecidos por la metodología CAPture utilizando RNA CAPeado, y CAPture utilizando RNA no modificado sí resultaron en secuencias que fueron tan legibles como inesperadas puesto que coincidían exactamente con el extremo 5’ de un transcrito 23S rRNA procesado, es decir, monofosfatado.

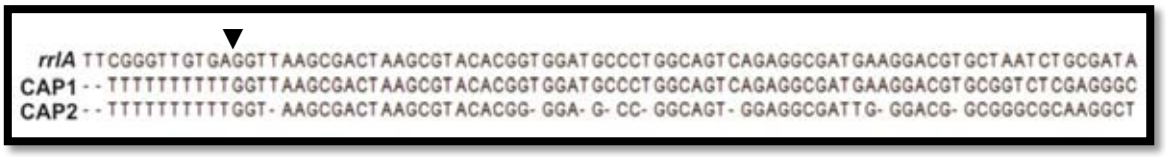


Figura 26. Alineamiento de las secuencias obtenidas por 5’ RACE contra 23S rRNA a partir de los enriquecimientos con la técnica CAPture. *rrlA* = región genómica de *E. coli* K12 que codifica para 23S rRNA; CAP1 = secuencia obtenida a partir del enriquecimiento por la técnica CAPture utilizando RNA total CAPeado; CAP2 = secuencia obtenida a partir del enriquecimiento por la técnica CAPture utilizando RNA total no modificado. La punta de flecha (▼) señala el extremo 5’ de un transcrito 23S rRNA típicamente procesado. La secuencia obtenida de las bibliotecas enriquecidas por la técnica pH-Switch no alineó con ninguna región del genoma de referencia y por tanto no se muestra.

8. DISCUSIÓN

Capacidad de las matrices para retener RNA.

De acuerdo con lo observado las matrices Agarosa-GST-BdRppH y Agarosa-GST-eIf4E son capaces de retener RNA. Los controles negativos mostrados en la figura 17B demuestran que el RNA presente en los enriquecimientos no proviene de los cultivos de *E. coli* a partir de los que se purificaron las proteínas de fusión ya que en los experimentos donde se utilizó agua libre de RNasas como muestra inicial (figura 17B carriles 2 a 5) no se detecta presencia de RNA. Además se determinó que la capacidad de retención de RNA es una propiedad intrínseca de las proteínas BdRppH y eIF4E como se demostró al utilizar matrices Agarosa-GST para enriquecer una muestra de RNA total utilizando ambas metodologías (figura 17B carriles 6 y 11). En estos ensayos sólo observamos RNA en la fracción no unida, que se recupera inmediatamente en el primer lavado. Esto significa que la porción GST en las proteínas de fusión utilizadas no influye en los procesos de enriquecimiento mediados por nuestras metodologías.

La matriz Agarosa-GST-eIF4E es capaz de retener RNA no CAPeado.

Quizá la observación más llamativa de este trabajo es que utilizando la técnica CAPture se puede enriquecer RNA bacteriano sin necesidad de modificarlo con 5'-cap. Esto lo demuestra el ensayo donde utilizamos una matriz Agarosa-GST-eIF4E para enriquecer RNA total no modificado proveniente de *E. coli* (figura 17A carril 5), a partir del cual obtuvimos una muestra de RNA óptima para realizar ensayos 5' RACE (figuras 20, 23 y 24) que coinciden con extremos 5' previamente reportados para ompA y 23S RNA (figuras 25 y 26). El principal inconveniente de este dato es que no concuerda del todo con lo hecho por Hagedorn[8], quien al utilizar RNA CAPeado y RNA no CAPeado sintetizado *in vitro* detectó que el RNA no modificado se recuperó gradualmente en las primeras fracciones de lavado de la cromatografía y no fue retenido en la fracción final. Sin embargo, a esto tenemos dos observaciones: primero, en el experimento de Hagedorn no se reporta la concentración exacta de RNA utilizado específicamente en el ensayo que referimos;

segundo, al purificar su proteína de fusión no se reportó la utilización de RNasa para eliminar los posibles transcritos arrastrados a partir del cultivo de *E. coli* del cual se extrajo la proteína. En conjunto ambas observaciones se complementan para explicar lo observado por nosotros ya que existe la posibilidad de que los transcritos arrastrados desde el cultivo de *E. coli* en el experimento de Hagedorn compitieran por los sitios de unión a eIF4E con el RNA no CAPeado sintetizado *in vitro* utilizado como control de reconocimiento específico eIF4E-5'cap, asumiendo que este último se encontrara en menor cantidad molar. Cabe señalar también que en el experimento de Hagedorn el RNA no modificado no se recupera inmediatamente en la fracción no unida, sino hasta algunos pasos de lavado después. Con lo anterior en cuenta podemos especular que si bien eIF4E tiene mucha más afinidad por transcritos con extremos 5'-cap, también es capaz de retener extremos 5'-no modificados en ausencia de extremos 5'-cap. Esta idea encuentra cierto sustento en el trabajo de Stolarski, quien analizó el comportamiento cinético de la unión entre eIF4E y distintos análogos de la estructura 5'-cap, reportando que tanto análogos 5'-trifosfatados no modificados como la estructura 5'-cap (m^7GMP) podían unirse a eIF4E, aunque con una diferencia en la constante de asociación de 1 orden de magnitud, siendo mayor para la estructura 5'-cap[29]. Lamentablemente en sus experimentos Stolarski utilizó monoribonucléotidos y diribonucleótidos, pero no transcritos sintetizados *in vitro*, lo cual hace que las observaciones no sean completamente comparables.

Especificidad de la interacción entre las matrices y los transcritos primarios.

En su concepción las matrices diseñadas en este trabajo se idealizaron como un método definitivo para asilar exclusivamente transcritos trifosfatados. Sin embargo, como se observa en las figuras 17, 22, 23 y 24 no nos fue posible deshacernos completamente del amplicón 23S rRNA, proveniente de un transcrito considerado monofosfatado. Esto lo confirmamos al secuenciar dichos amplicones presentes en las diferentes fracciones enriquecidas, observando que el extremo 5' de los fragmentos enriquecidos correspondía exactamente al primer nucleótido presente en el 23S rRNA procesado (figura 26), por lo menos en el caso de los enriquecimientos mediados por eIF4E. En cuanto al enriquecimiento mediado por BdRppH, la secuencia obtenida por 5' RACE contra 23S rRNA no alineó con ninguna región del genoma de *E. coli* K12. Tampoco podemos considerar esa secuencia críptica como un artificio porque no está presente en la reacción control donde no pusimos templado para la PCR. Por lo anterior no podemos aseverar que BdRppH sea

capaz de desdeñar transcritos monofosfatados, puesto que en la muestra enriquecida existe algo que interacciona con la secuencia específica del oligonucleótido correspondiente a 23S rRNA.

Nuevamente hay que tener en cuenta que hasta el 90% del RNA total en cultivos de *E. coli* es rRNA[6], y del 10% restante no existen reportes de cuanto corresponde a transcritos primarios. Por tanto la captura de 23S por nuestras metodologías podría deberse a un exceso de rRNA en comparación con los transcritos trifosfatados que nos interesan. Para solventar esto se requeriría ajustar las metodologías escalando la relación molar proteína:RNA, con la intención de evitar un exceso de sitios de unión libres en la columna cromatográfica que pudieran interaccionar con el exceso de rRNA.

Reconocimiento de puentes trifosfato.

Respecto a eIF4E existen artículos detallados del mecanismo de interacción con la estructura 5'-cap. Principalmente el grupo de Stolarski ha trabajado en ello y según sus resultados el puente trifosfato es parte importante del reconocimiento inicial entre eIF4E y el ribonucleótido[29], además de que algunos de los residuos en el sitio de unión de la proteína deben estar cargados eléctricamente de manera óptima para poder acomodar las cargas negativas de los grupos en el puente trifosfato[45].

No hay reportes que exploren cuestiones similares en BdRppH, pero de acuerdo a los cristales de proteínas unidas a GTP tanto en BdRppH (PDB *accesion no*: 3FFU) como en BsRppH (PDB *accesion no*: 4JZT) el puente trifosfato se encuentra en contacto estrecho con la proteína, particularmente con residuos conservados del motivo Nudix, específicamente la primer glicina y el primer y segundo glutamato presentes en el motivo.

En conjunto los reportes sugieren que en proteínas que reconocen transcritos primarios, como es el caso de eIF4E y BdRppH, el puente trifosfato en los transcritos blanco es esencial para una interacción exitosa. Sin embargo, existe aún la duda de qué es lo que confiere la especificidad a los transcritos procesados por este tipo de proteínas. Para el caso de BsRppH se sospecha de preferencias por ciertas bases al inicio del transcrito, sin embargo la preferencia, en palabras de los autores, "es apenas modesta"[31].

En cuanto a eIF4E existen diversas familias e isoformas de esta proteína en los organismos eucarióticos. Se sabe que las isoformas se expresan en distintas condiciones celulares, y que entre isoformas varía la capacidad de favorecer la traducción de diversos transcritos particulares[22], aunque *in vitro* todas las isoformas son capaces de unirse a la estructura 5'-cap, sugiriendo una vez más que la especificidad de reconocimiento de transcritos blanco yace fuera del motivo conservado de unión a ligando. Sin embargo, por lo menos parte de la especificidad de los blancos se asocia directamente con la presencia de la estructura 5'-cap en sus extremos, asumiendo que eIF4E actúa solamente sobre transcritos sintetizados por la RNA pol II. Nuestros resultados sugieren modestamente la posibilidad de que eIF4E sea capaz de interaccionar con transcritos no CAPeados, bajo las condiciones adecuadas. Por supuesto, esto último es apenas una sospecha y requiere de un análisis que escapa a los objetivos de este trabajo.

9. CONCLUSIÓN.

Tanto la técnica CAPture como la técnica pH-Switch son capaces de enriquecer muestras de RNA total. Aunque falta por definir cuál de las dos metodologías es más específica y por tanto eficiente para llevar a cabo el trabajo, hasta ahora la candidata más atractiva es la técnica CAPture utilizando RNA total no modificado. Lo consideramos así puesto que dicha metodología es menos laboriosa y por tanto menos brusca con la muestra de RNA. Sin embargo, es muy importante realizar ajustes a la técnica para mejorar su especificidad.

10. PERSPECTIVAS

Determinar las eficiencias de enriquecimiento de las técnicas CAPture y pH-Switch.

La principal motivación de este trabajo fue la creación de una metodología más efectiva de enriquecimiento para la detección masiva de sitios de inicio de la transcripción. Ello implica que nuestras metodologías propuestas deben ser comparadas experimentalmente con otras consideradas estándar. Para esto, el experimento ideal sería uno en el que se utilice una mezcla de cantidades conocidas de dos tipos de transcritos sintetizados *in vitro*, un subgrupo monofosfatado y otro trifosfatado, marcados radiactivamente para hacer más fina su cuantificación y detección. La mezcla de transcritos sería sometida a los diferentes métodos de enriquecimiento, tanto los propuestos en este trabajo como los considerados estándar. Las diferentes fracciones obtenidas serían entonces analizadas para determinar de manera cuantitativa cuál metodología es más eficiente en la captura de transcritos trifosfatados.

Aplicar las técnicas CAPture y pH-Switch en proyectos de secuenciación masiva.

Como se mostró en este trabajo, los ensayos 5’ RACE son un método confiable para evaluar metodologías de enriquecimiento. De manera similar, las bibliotecas enriquecidas podrían ser analizadas por PCR en tiempo real, con resultados cuantitativos (algo que 5’ RACE no permite hacer directamente). Sin embargo, dado la visión masiva hacia la que aspira nuestro trabajo, lo el siguiente paso lógico es utilizar las metodologías descritas para estudios RNA-seq. De manera símil a lo descrito en el párrafo anterior, lo ideal sería aplicar las técnicas en estudios de varios organismos con transcriptomas previamente definidos, que permitan hacer una mejor comparación de nuestras metodologías respecto a las estándar para poder definir los pros y contras de su aplicación. Aunque nuestras metodologías están apenas en una fase inicial, las plataformas de secuenciación masiva permitirían evaluarlas de manera más fina y cuantificable.

11. METODOLOGÍAS

11.1 Cepas, medios de cultivo, oligonucleótidos, buffers y plásmidos.

CEPA	GENOTIPO
Escherichia coli BL21	F^- *dcm ompT hsdS*(r_B^- m_B^-) *gal* [*malB*$^+$]$_{K-12}$(λ^S)
Escherichia coli K12 MC1061	F^- Δ(ara-leu)7697 [*araD*139]$_{B/r}$ Δ(*codB*-lacI)3 *galK*16 *galE*15 λ^- e14$^-$ *mcrA*0 *relA*1 *rpsL*150(strR) *spoT*1 *mcrB*1 *hsdR*2(r^-m^+)

MEDIO DE CULTIVO	CONSTITUCIÓN
LB (Lysogeny Broth)	Gramos por litro de agua destilada: Cloruro de sodio 5 g, Extracto de levadura 5 g, Peptona de caseína 10 g. (pH 7.2 ± 0.2).

OLIGONUCLEÓTIDO	SECUENCIA 5' – 3'
degS	CGAAATAATCAGATCGTTGACC
fruA	CCGCCCATGTCAGTACACATC
gatY	TCATGACTGAGCGCACGCCA
iGEX_Fw	GGCGACCATCCTCCAAAATC
iGEX_Rw	GGCATCCGCTTACAGACAAG
ompA	TGAAACCAGTGTCATGGTAC
ppa	GTCCAGAGACAGGGTGTGGT
RACE1	GACTCGAGTCGACATCGATTTTTTTTTTTTTTTT
RACE2	GACTCGAGTCGACATCG
rpsB	GTCGCGCATGGAAACAGTTG
rRNA_23S_rev_183	GTTCGCCTCATTAACCTATGG
sRNAP1	CAAGCAGAAGACGGCATACGANNNNNN
sRNAPCR1	CAAGCAGAAGACGGCATACGA
xylA	TCCGCCCCGTTCCAGCAGAA

BUFFERS.

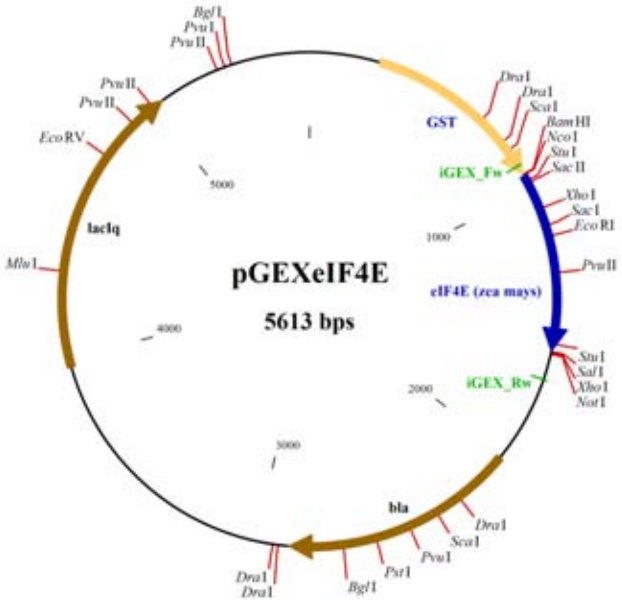
Debido al importante enfoque en el manejo de RNA que conllevan los experimentos reportados en este trabajo, todos los buffers (excepto TBE) en la siguiente lista fueron preparados con agua grado Milli-Q tratada con DEPC para disminuir en todo lo posible la actividad RNasa presente en las soluciones requeridas.

BUFFER	**CONSTITUCIÓN**
CAP-A 1X	10 mM K_2HPO_4, KCl 100 mM, EDTA (sal dibásica) 2mM, Glicerol 5 %, Triton X-100 0.005 %, 1.3 % Poli-vinil alcohol (98-99 %) hidrolizado, DTT 6 mM, 20 U/mL RNAsin, 100 ug/mL tRNA de levadura. (pH 8.0)
CAP-B 1X	10 mM K2HPO4, KCl 100 mM, EDTA (sal dibásica) 2mM, Glicerol 5 %, Triton X-100 0.005 %, 1.3 % Poli-vinil alcohol (98-99 %) hidrolizado, DTT 6 mM, 20 U/mL RNAsin. (pH 8.0)
GST Bind/Wash 1X	4.3 mM Na2HPO4, 1.47 mM KH2PO4, 137 mM NaCl, 2.7 mM KCl. (pH 7.3).
GST Elution 1X	100 mM glutatión reducido, 50 mM Tris-HCl. (pH 8.0).
pHS-A 1X	0.05 M acetato de sodio, 0.6 M acetato de amonio, 10 mM $MgCl_2$, 20 U/µl Rnasin (Promega). (pH 7.0)
pHS-B 1X	50 mM Tris, 5 mM $MgCl_2$, 1 mM DTT, 1 U/µl Rnasin (Promega). (pH 8.5)
TBE 1X	890 mM Tris-borato, 890 mM ácido bórico, 20 mM EDTA. (pH 8.2 - 8.4).

PLÁSMIDOS.

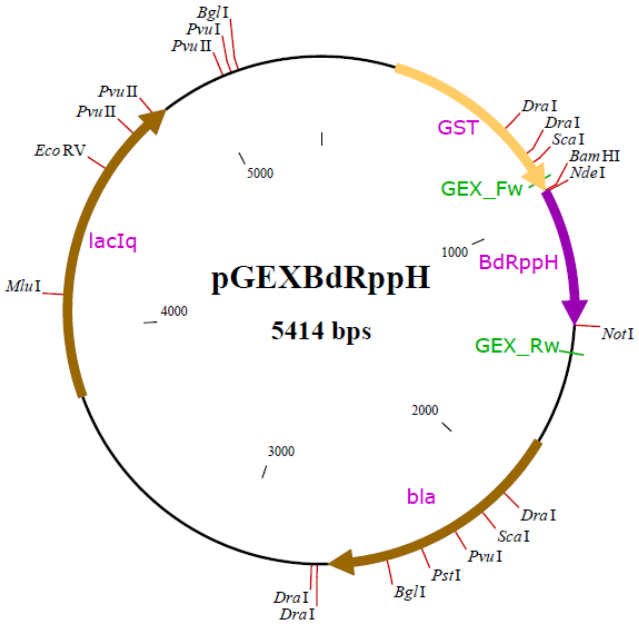
- **pGEXeIf4E**

Regalo de la Dr. Tzvetanka Dimitrova Dinkova. Derivado del plásmido pGEX-4T-2 contiene el gen eIF4E de *Zea mays* insertado en los sitios BamHI y NotI (no hay observaciones sobre si el uso de codones fue optimizado para la expresión de la proteína en *E. coli,* pero esta construcción fue utilizada con éxito previamente[18]). El gen quedó bajo el control del promotor tac y el operador *lac*, y está fusionado a una etiqueta de Glutation-S-transferasa (GST) en el amino terminal; el gen GST está interrumpido en su carboxilo terminal y no contiene codones de paro, así que la fusión es continua con eIF4E. El gen *bla* confiere resistencia a Ampicilina (100 ug/mL). Para más información sobre la expresión de proteínas en este vector consulte el manual de los vectores pGEX de GE Healtcare.



- **pGEXBdRppH**

Derivado del plásmido pGEX-4T-2 contiene el gen *BdrppH* (optimizado con el uso de codones para expresión en *E. coli*) insertado en los sitios BamHI y NotI. El gen quedó bajo el control del promotor *tac* y el operador *lac*, y está fusionado a una etiqueta de Glutation-S-transferasa (GST) en el amino terminal; el gen GST está interrumpido en su carboxilo terminal y no contiene codones de paro, así que la fusión es continua con el inserto. El gen *bla* confiere resistencia a Ampicilina (100 ug/mL). Para más información sobre la expresión de proteínas en este vector consulte el manual de los vectores pGEX de GE Healtcare.



11.2 Construcción del vector pGEXBdRppH.

- EXTRACCIÓN DE DNA PLASMÍDICO.

El vector pGEXeIF4E y el vector pBSK-GS51453 (vector en que la compañía Epoch Life Sciences entregó clonado el ORF sintético *synBdRppH*) fueron transformados en *E. coli* DH5-Alpha y posteriormente cultivados en LB ampicilina (100 ug/µl) durante aprox. 15 horas (overnight). A partir de ahí se extrajo DNA plasmídico utilizando el kit de purificación en columna High Pure Plasmid Isolation siguiendo las indicaciones del proveedor (Roche).

- DIGESTION DE DNA.

Los vectores fueron sometidos a una doble digestión enzimática con las enzimas BamHI y NotI (ambas del proveedor New England Biolabs). Se utilizaron reacciones de 50 µl para digerir 2 ug de DNA con 1 µl de cada enzima incubando 1 hora a 37 °C.

- PURIFICACIÓN DE DNA A PARTIR DE GELES DE AGAROSA.

Los productos de las digestiones enzimáticas fueron separados electroforéticamente en agarosa 1% con buffer TBE. A partir de ahí se escindieron manualmente las bandas correspondientes al vector pGEX-4T-2 (aprox. 4939 pb) y al ORF *synBdRppH* (aprox. 475 pb). Ambas fueron purificadas utilizando el kit High Pure PCR Product Purification siguiendo las indicaciones del proveedor (Roche).

- LIGACIÓN DE FRAGMENTOS DE DNA.

El vector pGEX-4T-2 y el inserto *synBdRppH* fueron sometidos a una reacción de ligación enzimática con la enzima T4 DNA ligasa siguiendo las indicaciones del proveedor (Life Technologies). El volumen de reacción fue de 20 µl y la relación molar inserto:vector se mantuvo en 3:1. El producto de la ligación se transformó en *E. coli* DH5-Alpha y se cultivó overnight en medio LB sólido adicionado con ampicilina (100 ug/mL) para su posterior monitoreo.

- CONFIRMACIÓN DE LA CONSTRUCCIÓN.

Las clonas producto del paso anterior fueron monitoreadas por PCR en colonia utilizando los oligonucleótidos iGEX_Fw e iGEX_Rw. Se seleccionaron algunas clonas que mostraron el patrón de ampificación correcto (una banda de aprox. 625 pb) y se secuenciaron en la Unidad de Síntesis y Secuenciación de ADN del IBT-UNAM. El resultado de la secuenciación confirmó a la clona #4 como la construcción correcta. Dicha clona se marcó como pGEXBdRppH, se cultivó overnight en LB ampicilina (100 ug/µl) y después se almaceno en glicerol a -70 °C.

11.3 Purificación de las proteínas GST, GST-eIF4E y GSTBdRppH.

Se purificaron los vectores pGEX-4T-2, pGEXeIF4E y pGEXBdRppH a partir de cultivos overnight de *E. coli* DH5-Alpha a 37 °C utilizando el kit High Pure Plasmid Isolation de Roche. Los vectores se utilizaron para purificar las proteínas de fusión GST, GST-eIF4E y GST-BdRppH respectivamente. El protocolo general fue el siguiente:

- EXPRESIÓN DE LA PROTEÍNA.

1. El vector purificado fue transformado en la cepa de expresión *E. coli* BL21. El producto de la transformación fue cultivado overnight en medio LB sólido adicionado con ampicilina (100 ug/mL) a 37 °C.
2. Una colonia producto de la transformación fue seleccionada y se cultivó overnight en LB ampicilina (100 ug/mL).
3. A partir del cultivo overnight se hizo una dilución 1:100 en 100 mL de medio LB ampicilina (100 ug/mL) y se incubó a 37 °C hasta alcanzar una OD_{600nm} = 0.6.
4. Alcanzada la OD_{600nm} = 0.6, se indujo la expresión de la proteína agregando IPTG hasta una concentración final de 0.2 mM.
5. El cultivo inducido se incubó overnight a 20 °C.

- PURIFICACIÓN DE LA PROTEÍNA.

6. Las células fueron recolectadas por centrifugación a 7700 x g durante 10 minutos a 4 °C, el sobrenadante desechado, y la pastilla celular resuspendida en buffer frío GST Bind/Wash 1X.
7. El tubo conteniendo la muestra fue colocado en hielo y se lisó el cultivo utilizando un sonicador marca Branson.
8. El cultivo lisado se separó por ultracentrifugación a 12,000 x g durante 10 minutos a 4 °C. Se desechó el precipitado celular. El sobrenadante que contenía las proteínas de fusión fue separado en un tubo nuevo (pruebas preliminares revelaron que no había presencia importante de la proteína de fusión en cuerpos de inclusión).
9. Se agregó RNasa A (Roche) al sobrenadante a una concentración final de 0.1 mg/mL. La reacción se incubó a 37 °C durante 30 minutos.

10. Se vertió el sobrenadante en una columna de 4 mL de resina *GST-Bind*™, previamente preparada de acuerdo a las indicaciones del proveedor (Novagen). Se dejó fluir la muestra por gravedad con una velocidad de flujo de 40 mL por hora.
11. Se lavó la columna con 40 mL de buffer GST Bind/Wash 1X.
12. Se eluyó la muestra con 12 mL de buffer GST Elution 1X.
13. La elución se mantuvo a 4 °C mientras se verificaba la presencia de la proteína purificada por SDS-PAGE en un gel de acrilamida 15%.
14. La muestra se dializó en una membrana Spectra/Por® MWCO 12-14000 sumergida en 1 Lt de buffer GST Bind/Wash 1X. La diálisis se llevó a cabo en agitación a 4 °C durante 12 horas, después de las cuales se remplazó el buffer por buffer fresco GST Bind/Wash 1X y se continuó la agitación a 4 °C durante otras 6 horas.
15. La muestra se pasó de la membrana de diálisis hacia un tubo falcón de 15 mL a 4 °C.

- CUANTIFICACIÓN Y ALMACENAJE.

16. Se cuantificó la cantidad de proteína por un ensayo de Bradford en microplaca (Bio-Rad).
17. Se agregó glicerol al stock de proteína hasta una concentración final de 20%.
18. Finalmente la proteína purificada se almacenó a -20 °C por no más de 2 semanas hasta ser utilizada.

11.4 Purificación de RNA total.

1. La cepa *E. coli* K12 MC1061 se cultivó en medio LB a 37 °C hasta una OD_{600}= 0.8.
2. A partir de ese cultivo se extrajo RNA total utilizando el kit RNAeasy® siguiendo las indicaciones del proveedor (Qiagen). La muestra se almacenó en H_2O grado Milli-Q tratada con DEPC.
3. Para eliminar la contaminación por DNA se agregó DNasa I hasta una concentración de 1 u/ug de RNA total siguiendo las indicaciones del proveedor (Thermo Scientific).
4. Se cuantificó la muestra final de RNA tratado con DNasa I en un instrumento 2100 Bioanalyzer (Agilent).

11.5 Método CAPture para enriquecimiento con transcritos primarios procarióticos.

*La metodología aquí presentada es una adaptación de los métodos propuestos por Hagedorn[8] y Narry Kim[9].

- PREPARACIÓN DE LA MATRIZ agarosa-glutation-GST-eIF4E.

1. Se centrifugaron 400 µl de resina agarosa-glutatión (GST-Bind, Novagen) a 1,000 x g durante 5 minutos; se tiró el sobrenadante, y se resuspendió por pipeteo gentil en 1 ml de buffer Bind/Wash 1X.
2. Se centrifugó a 1,000 x g durante 5 minutos y se tiró el sobrenadante. Se resuspendió la resina en 200 µl de buffer Bind/Wash 1X.
3. Se agregaron 400 ug de proteína GST-eIF4E y se mezclaron por inversión en un rotador de tubos durante 2 horas a temperatura ambiente.
4. Se centrifugó la matriz agarosa-glutation-GST-eIF4E a 1,000 x g durante 5 minutos y se tiró el sobrenadante. Luego se resuspendió la muestra por inversión en 2 ml de buffer Bind/Wash 1X.
5. Se repitió el paso anterior.
6. Finalmente se separó la matriz agarosa-glutatión-GST-eIF4E en dos alícuotas de 1 ml cada una, las cuales se almacenaron a 4 °C por no más de 3 semanas.

- CAPPING *IN VITRO* DEL RNA PROCARIÓTICO.

1. Una muestra de 60 ug de RNA total tratado con DNasa se sometió a una reacción de capping *in vitro* utilizando simultáneamente los kits ScriptCap™ m^7G Capping System y ScriptCap™ 2'-O-Methyltransferase siguiendo las indicaciones del proveedor (Epicentre). La reacción se incubó a 37 °C durante 1 hora.
2. La muestra se mezcló con 0.1 volúmenes de acetato de sodio 3M, 3 volúmenes de etanol 100%, y 1 µl de glucógeno grado RNA (Life technologies). Se mezcló vigorosamente por vortex y se guardó la muestra a -70°C durante dos horas.
3. Se centrifugó la muestra a 12,000 x g durante 30 minutos a 4 °C. Se removió y desechó el sobrenadante.

4. Se agregaron 500 µl de una solución fría etanol 75%. Se mezcló vigorosamente por vortex. Posteriormente se centrifugó la muestra a 12,000 x g durante 5 minutos a 4 °C. Se removió y desechó el sobrenadante.

5. Se repitió el paso 4 una vez más, cuidando remover la mayor cantidad de etanol sin molestar la pastilla de RNA.

6. La muestra se dejó secar al aire hasta eliminar todo rastro de solución.

7. La pastilla de RNA se resuspendió en 30 µl de H_2O libre de RNasas.

- ENRIQUECIMIENTO DE LA MUESTRA.

1. Se tomó una alícuota de 1 ml de la matriz agarosa-glutatión-GST-eIF4E. La alícuota se centrifugó a 5,000 x g durante 10 segundos a temperatura ambiente. Se removió con cuidado el sobrenadante, tratando de perder la menor cantidad posible de matriz durante el pipeteo.

2. Se agregaron 500 µl de buffer CAP-A 1X para resuspender gentilmente la matriz por inversión. Se centrifugó la muestra a 3,800 x g durante 10 segundos a temperatura ambiente y se desechó el sobrenadante.

3. Se repitió 2 veces el paso anterior. Hasta este momento, el volumen de la matriz cromatográfica se mantuvo en 100 µl aproximadamente.

4. Una muestra de 60 ug de RNA total capeado *in vitro* se desnaturalizaron por calor a 70 °C durante 10 minutos primero, luego se enfriaron en hielo durante 2 minutos.

5. Se agregó buffer CAP-A 1X a los 60 ug de RNA desnaturalizado hasta llevar la muestra a un volumen de 250 µl. Esta muestra se sumó a los 100 µl de matriz cromatográfica del paso 3, y ambas se mezclaron gentilmente por inversión en un rotador de tubos durante 1 hora a temperatura ambiente.

6. La muestra se centrifugó a 3,800 x g durante 10 segundos. Se removió el sobrenadante. (Este sobrenadante corresponde a la fracción “No Unida” de la cromatografía. A esta fracción sobrenadante se agregaron 0.1 volúmenes de acetato de sodio 3M, 3 volúmenes de etanol 100% y 1 µl de glucógeno grado RNA. Luego se guardó a -70 °C en un tubo separado para su posterior análisis).

7. A la matriz cromatográfica se agregaron 500 µl de buffer Cap-B 1X adicionado con guanosín difosfato (GDP) y se dejó mezclar por inversión en un rotador de tubos durante 5 minutos a temperatura ambiente. La muestra se centrifugó a 3,800 x g durante 10 segundos. Se removió el sobrenadante y se guardó en un tubo separado para su posterior análisis.

8. Se repitió el paso 7 cuatro veces más. (Los 5 sobrenadantes se mezclaron en un mismo tubo y se marcaron como la fracción “Lavado” de la cromatografía. A esta fracción se agregaron 0.1 volúmenes de acetato de sodio 3M, 3 volúmenes de etanol 100% y µl de glucógeno grado RNA. Luego se guardó a -70 °C en un tubo separado para su posterior análisis).

9. A la matriz cromatográfica se agregaron 300 µl de H_2O tratada con DEPC y se resuspendió gentilmente por pipeteo hasta homogenizar la muestra.

10. Se agregaron 400 µl de TRIzol® (Life Technologies) y se incubó a temperatura ambiente durante 5 minutos.

11. Se agregaron 80 µl de cloroformo y se mezcló la muestra vigorosamente por vortex. Posteriormente se incubó 3 minutos a temperatura ambiente.

12. Se centrifugó a 12,000 x g durante 5 minutos a 4 °C. La muestra se separará en 3 fases.

13. Cuidadosamente se recuperó la fase superior y se pasó a otro tubo. (Esta fase recuperada corresponde a la fracción “Enriquecido” de la cromatografía. A esta fracción sobrenadante se agregaron 0.1 volúmenes de acetato de sodio 3M, 3 volúmenes de etanol 100% y 1uL de glucógeno grado RNA. Luego se guardó a -70 °C en un tubo separado para su posterior análisis).

*Todas las fracciones cromatográficas obtenidas se almacenaron a -70 °C hasta ser utilizadas.

- PRECIPITACIÓN DEL RNA PRESENTE EN LAS FRACCIONES DE LA CROMATOGRAFÍA.

Para solubilizar el RNA presente en las 3 fracciones provenientes de la cromatografía (“No Unido”, “Lavado”, “Enriquecido”) conservadas en etanol 100% a -70 °C, las muestras se trataron de la siguiente manera:

1. Se centrifugó la muestra a 12,000 x g durante 30 minutos a 4 °C. Se removió y desechó el sobrenadante.

2. Se agregaron 500 µl de una solución fría etanol 75%. Se mezcló vigorosamente por vortex. Posteriormente se centrifugó a 12,000 x g durante 5 minutos a 4 °C. Se removió y desechó el sobrenadante.

3. Se repitió el paso 4 una vez más, cuidando remover la mayor cantidad de etanol sin molestar la pastilla de RNA.

4. La muestra se dejó secar al aire hasta eliminar todo rastro de solución.

5. La pastilla de RNA se resuspendió en 50 µl de H_2O libre de RNasas.

6. Se utilizó 1 µl de la muestra para cuantificar el RNA utilizando el instrumento 2100 Bioanalyzer de Agilent.

*Todas las fracciones RNA se almacenaron a -70 °C hasta ser utilizadas para la construcción debibliotecas cDNA (apartado 11.6).

11.6 Método pH-Switch para enriquecimiento con transcritos primarios procarióticos.

- PREPARACIÓN DE LA MATRIZ agarosa-glutation-GST-BdRppH.

1. Centrifugar 400 ml de resina agarosa-glutatión (GST-Bind, Novagen) a 1,000 x g durante 5 minutos; tirar el sobrenadante, y resuspender por pipeteo gentil en 1 ml de buffer Bind/Wash 1X.

2. Centrifugar a 1,000 x g durante 5 minutos y tirar el sobrenadante. Resuspender la resina en 200 µl de buffer Bind/Wash 1X.

3. Agregar 400 ug de proteína GST-BdRppH y dejar mezclar por inversión en un rotador de tubos durante 2 horas a temperatura ambiente.

4. Centrifugar la matriz agarosa-glutation-GST-BdRppH a 1,000 x g durante 5 minutos, tirar el sobrenadante, y resuspender por inversión en 2 ml de buffer Bind/Wash 1X.

5. Repetir el paso anterior.

6. Separar la matriz agarosa-glutatión-GST-BdRppH en dos alícuotas de 1 ml cada una. Almacenar las alícuotas a 4 °C por no más de 3 semanas.

- ENRIQUECIMIENTO DE LA MUESTRA.

1. Se tomó una alícuota de 1 ml de la matriz agarosa-glutatión-GST-BdRppH. La alícuota se centrifugó a 5,000 x g durante 10 segundos a temperatura ambiente. Se removió con cuidado el sobrenadante, tratando de perder la menor cantidad posible de matriz durante el pipeteo.

2. Se agregaron 500 µl de buffer pHS-A 1X para resuspender gentilmente la matriz por inversión. Se centrifugó la muestra a 3,800 x g durante 10 segundos a temperatura ambiente y se desechó el sobrenadante.

3. Se repitió 2 veces el paso anterior.

4. 60 ug de RNA total se desnaturalizaron por calor a 70 °C durante 10 minutos primero, luego se enfriaron en hielo durante 2 minutos.

5. Se agregó buffer pHS-A a los 60 ug de RNA desnaturalizado hasta llevar la muestra a un volumen de 250 µl. Esta muestra se sumó a los 100 µl de matriz cromatográfica del paso 3, y ambas se mezclaron gentilmente por inversión en un rotador de tubos durante 1 hora a temperatura ambiente.

6. La muestra se centrifugó a 3,800 x g durante 10 segundos. Se removió el sobrenadante. (Este sobrenadante corresponde a la fracción "No Unida" de la cromatografía. A esta fracción se agregaron 0.1 volúmenes de acetato de sodio 3M, 3 volúmenes de etanol 100% y 1 µl de glucógeno grado RNA. Luego se guardó a -70 °C en un tubo separado para su posterior análisis).

7. A la matriz cormatográfica se agregaron 500 µl de buffer pHS-B y se dejó mezclar por inversión en un rotador de tubos durante 5 minutos a temperatura ambiente. La muestra se centrifugó a 3,800 x g durante 10 segundos. Se removió el sobrenadante y se guardó en un tubo separado para su posterior análisis.

8. Se repitió el paso 7 dos veces más. (Los 3 sobrenadantes se mezclaron en un mismo tubo y se marcaron como la fracción "Lavado" de la cromatografía. A esta fracción se agregaron 0.1 volúmenes de acetato de sodio 3M, 3 volúmenes de etanol 100% y µl de glucógeno grado RNA. Luego se guardó a -70 °C en un tubo separado para su posterior análisis).

9. Se agregaron 500 µl de buffer pHS-H y se dejó mezclar por inversión en un rotador de tubos durante 1 hora a 37 °C para permitir que se lleve a cabo la reacción de pirofosfohidrólisis.

10. La muestra se centrifugó a 3,800 x g durante 10 segundos. Se removió el sobrenadante y se guardó en un tubo separado para su posterior análisis. (Esta fase recuperada corresponde a la fracción "Enriquecido" de la cromatografía. A esta fracción se agregaron 0.1 volúmenes de acetato de sodio 3M, 3 volúmenes de etanol 100% y 1ul de glucógeno grado RNA. Luego se guardó a -70 °C en un tubo separado para su posterior análisis).

*Todas las fracciones cromatográficas obtenidas se almacenaron a -70 °C hasta ser utilizadas.

- PRECIPITACIÓN DEL RNA PRESENTE EN LAS FRACCIONES DE LA CROMATOGRAFÍA.

Para solubilizar el RNA presente en las 3 fracciones provenientes de la cromatografía ("No Unido", "Lavado", "Enriquecido") conservadas en etanol 100% a -70 °C se trataron de la siguiente manera:

1. Se centrifugó la muestra a 12,000 x g durante 30 minutos a 4 °C. Se removió y desechó el sobrenadante.

2. Se agregaron 500 µl de una solución fría etanol 75%. Se mezcló vigorosamente con vortex. Posteriormente se centrifugó a 12,000 x g durante 5 minutos a 4 °C. Se removió y desechó el sobrenadante.

3. Se repitió el paso 4 una vez más, cuidando remover la mayor cantidad de etanol sin molestar la pastilla de RNA.

4. La muestra se dejó secar al aire hasta eliminar todo rastro de solución.

5. La pastilla de RNA se resuspendió en 50 µl de H_2O libre de RNasas.

6. Se utilizó 1 µl de la muestra para cuantificar el RNA utilizando el instrumento 2100 Bioanalyzer de Agilent.

*Todas las fracciones RNA se almacenaron a -70°C hasta ser utilizadas para la construcción de bibliotecas cDNA (apartado 11.6).

11.7 Construcción de bibliotecas cDNA.

- SÍNTESIS DE PRIMERA CADENA cDNA.

1. 500 ng de RNA total se sometieron a una reacción de transcripción reversa utilizando la enzima Superscript II siguiendo las indicaciones del proveedor (Invitrogen) en un volumen final de 20 µl. Se utilizaron 100 ng del oligonucleótido random sRNAP1.

2. Se purificó la muestra cDNA utilizando el kit High Pure PCR Product Purification siguiendo las indicaciones del proveedor (Roche). El volumen final de elución fue 50 µl.

- POLIADENILACIÓN.

1. 1 ug de cDNA de cadena sencilla se sometieron a una reacción de poliadenilación utilizando la enzima Transferasa Terminal (TdT) siguiendo las indicaciones del proveedor (New England Biolabs) en un volumen final de 50 µl. Se utilizaron 0.5 µl 10 mM dATP y la muestra se incubo a 37°C por 30 minutos.

2. Se purificó la muestra cDNA utilizando el kit High Pure PCR Product Purification siguiendo las indicaciones del proveedor (Roche). El volumen final de elución fue 50 µl.

- SÍNTESIS DE SEGUNDA CADENA cDNA.

1. Se utilizaron 20 ng de cDNA-poliadenilado como templado para generar la cadena complementaria. La reacción se llevó a cabo con la enzima Taq DNA polimerasa recombinante producida en nuestro laboratorio, utilizando el oligonucleótido RACE1. Las condiciones de reacción fueron las siguientes:

Buffer 10X	5 µl
dNTPs [2.5 uM]	4 µl
Oligo RACE1 [10 uM]	2.5 µl
cDNA templado	20 ng
H_20	hasta llevar la reacción a 48 µl
Taq Polimerasa (5 U/µl)	2 µl
Volumen Final	50 µl

	95 °C	**2 minutos**
5 ciclos	**95 °C**	**45 segundos**
	45 °C	**2 minutos**
	72 °C	**30 segundos**
	72 °C	**5 minutos**

2. Se purificó la muestra cDNA utilizando el kit High Pure PCR Product Purification siguiendo las indicaciones del proveedor (Roche). El volumen final de elución fue 50 µl.

- REAMPLIFICACIÓN DE LA BIBLIOTECA cDNA.

1. Se utilizaron 20 ng de cDNA doble cadena como templado para generar la biblioteca final para detección de sitios de inicio de transcripción, Taq DNA polimerasa recombinante producida en nuestro laboratorio, y los oligonucleótidos RACE2 y sRNAPCR-1. Las condiciones de reacción fueron las siguientes:

Buffer 10X	5 µl
dNTPs [2.5 uM]	4 µl
Oligo RACE2 [10 uM]	2.5 µl
Oligo sRNAPCR-1 [10 uM]	2.5 µl
cDNA templado	20 ng
H_20	hasta llevar la reacción a 48 µl
Taq Polimerasa (5 U/µl)	2 µl
Volumen Final	50 µl

	95 °C	**2 minutos**
15 ciclos	**95 °C**	**45 segundos**
	45 °C	**45 segundos**
	72 °C	**15 segundos**
	72 °C	**5 minutos**

*Los tiempos de extensión durante el programa de PCR fueron cortos porque en nuestra experiencia vimos que esto reducía la cantidad de amplicones espurios en ensayos posteriores de RACE 5'.

2. Se purificó la biblioteca cDNA final utilizando el kit High Pure PCR Product Purification siguiendo las indicaciones del proveedor (Roche). El volumen final de elución fue 50 µl.

*Todas las bibliotecas correspondientes a las diferentes fracciones de los procesos de enriquecimiento CAPture y pH-Switch fueron almacenadas a -20 °C hasta ser utilizados para los ensayos 5' RACE.

11.8 Ensayos 5’ RACE.

1. La detección de sitios de inicio de la transcripción se llevó a cabo mediante ensayos RACE 5’ utilizando oligonucleótidos específicos dependiendo del gen blanco. Las condiciones de reacción fueron las siguientes:

Buffer 10X	5 µl
dNTPs [2.5 uM]	4 µl
Oligo RACE2 [10 uM]	2.5 µl
Oligo específico [10 uM]	2.5 µl
cDNA templado	20 ng
H_20	hasta llevar la reacción a 48 µl
Taq Polimerasa (5 U/µl)	2 µl
Volumen Final	50 µl

	95 °C	**2 minutos**
30 ciclos	**95 °C**	**45 segundos**
	45 °C	**45 segundos**
	72 °C	**30 segundos**
	72 °C	**2 minutos**

*Es recomendable realizar previamente ensayos de PCR en gradiente de temperatura para determinar las condiciones óptimas de amplificación del ensayo RACE 5'. Se requiere de una amplificación lo más específica posible para continuar con el siguiente paso.

2. El producto de la reacción de PCR fue separado por electroforesis en un gel de agarosa 1% con buffer TBE. La banda correspondiente al amplicón esperado fue escindida del gel y purificada utilizando el kit High Pure PCR Product Purification siguiendo las indicaciones del proveedor (Roche). El volumen final de elución fue 50 µl.

3. El amplicón purificado fue enviado a secuenciar por el método Sanger y alineado contra el genoma de referencia de *E. coli* K12.

12. BIBLIOGRAFÍA

1 Asashima, M., Ikeuchi, M. and Ishiura, S. (2011). et al. 2011. Chapter 8.3: Post- transcriptional Modification. A comprehensive approach to LIFE SCIENCE. Web textbook. The University of Tokio. [Available Online at: http://csls-text3.c.u-tokyo.ac.jp/inactive/08_03.html].

2 Bail, S. and Kiledjian, M. (2009). Tri- to be Mono- for Bacterial mRNA Decay. Structure, 17(3), pp.317-319.

3 Belasco, J. (2010). All things must pass: contrasts and commonalities in eukaryotic and bacterial mRNA decay. Nature Reviews Molecular Cell Biology, 11(7), pp.467-478.

4 Carpousis, A. (2007). The RNA Degradosome of Escherichia coli : An mRNA-Degrading Machine Assembled on RNase E. Annu. Rev. Microbiol., 61(1), pp.71-87.

5 Chakravarty, A., Carlson, J., Khetani, R. and Gross, R. (2007). A novel ensemble learning method for de novo computational identification of DNA binding sites. BMC Bioinformatics, 8(1), p.249.

6 Chen, Z. and Duan, X. (2011). Ribosomal RNA Depletion for Massively Parallel Bacterial RNA-Sequencing Applications. Methods in Molecular Biology, pp.93-103.

7 Cho, E., Takagi, T., Moore, C. and Buratowski, S. (1997). mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. Genes & Development, 11(24), pp.3319-3326.

8 Choi, Y. and Hagedorn, C. (2003). Purifying mRNAs with a high-affinity eIF4E mutant identifies the short 3' poly(A) end phenotype. Proceedings of the National Academy of Sciences, 100(12), pp.7033-7038.

9 Cole, S., Bremer, E., Hindennach, I. and Henning, U. (1982). Characterisation of the promoters for the ompA gene which encodes a major outer membrane protein of Escherichia coli. MGG Molecular & General Genetics, 188(3), pp.472-479.

10 Deana, A., Celesnik, H. and Belasco, J. (2008). The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal. Nature, 451(7176), pp.355-358.

11 Dinkova, T., Márquez-Velázquez, N., Aguilar, R., Lázaro-Mixteco, P. and Sánchez de Jiménez, E. (2011). Tight translational control by the initiation factors eIF4E and eIF(iso)4E is required for maize seed germination. Seed Science Research, 21(02), pp.85-93.

12 Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J. and al., e. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 269(5223), pp.496-512.

13 Friedland, D., Wooten, W., LaVoy, J., Hagedorn, C. and Goss, D. (2005). A Mutant of Eukaryotic Protein Synthesis Initiation Factor eIF4E K119A Has an Increased Binding Affinity for both m 7 G Cap Analogues and eIF4G Peptides †. Biochemistry, 44(11), pp.4546-4550.

14 Gutgsell, N. and Jain, C. (2009). Coordinated Regulation of 23S rRNA Maturation in Escherichia coli. Journal of Bacteriology, 192(5), pp.1405-1409.

15 Hashimoto, S., Qu, W., Ahsan, B., Ogoshi, K., Sasaki, A., Nakatani, Y., Lee, Y., Ogawa, M., Ametani, A., Suzuki, Y., Sugano, S., Lee, C., Nutter, R., Morishita, S. and Matsushima, K. (2009). High-Resolution Analysis of the 5'-End Transcriptome Using a Next Generation DNA Sequencer. PLoS ONE, 4(1), p.e4108.

16 Kaberdin, V., Singh, D. and Lin-Chao, S. (2011). Composition bacteria. Journal of Biomedical Science, 18(1), p.23.

17 Kroger, C., Dillon, S., Cameron, A., Vogel, J. and Hinton, J. (2012). The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. Proceedings of the National Academy of Sciences, 109(20), pp.E1277-E1286.

18 Kulpa, D. (1997). Determination of the site of first strand transfer during Moloney murine leukemia virus reverse transcription and identification of strand transfer-associated reverse transcriptase errors. The EMBO Journal, 16(4), pp.856-865.

19 Lee, Y., Kim, M., Han, J., Yeom, K., Lee, S., Baek, S. and Kim, V. (2004). MicroRNA genes are transcribed by RNA polymerase II. The EMBO Journal, 23(20), pp.4051-4060.

20 Luciano, D., Hui, M., Deana, A., Foley, P., Belasco, K. and Belasco, J. (2012). Differential Control of the Rate of 5'-End-Dependent mRNA Degradation in Escherichia coli. Journal of Bacteriology, 194(22), pp.6233-6239.

21 Margulies, M., Egholm, M., Altman, W., Attiya, S., Bader, J., Bemben, L., Berka, J., Braverman, M., Chen, Y., Chen, Z., Dewell, S., Du, L., Fierro, J., Gomes, X., Godwin, B., He, W., Helgesen, S., Ho, C., Irzyk, G., Jando, S., Alenquer, M., Jarvie, T., Jirage, K., Kim, J., Knight, J., Lanza, J., Leamon, J., Lefkowitz, S., Lei, M., Li, J., Lohman, K., Lu, H., Makhijani, V., McDade, K., McKenna, M., Myers, E., Nickerson, E., Nobile, J., Plant, R., Puc, B., Ronan, M., Roth, G., Sarkis, G., Simons, J., Simpson, J., Srinivasan, M., Tartaro, K., Tomasz, A., Vogt, K., Volkmer, G., Wang, S., Wang, Y., Weiner, M., Yu, P., Begley, R. and Rothberg, J. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature.

22 Martínez-Silva, A., Aguirre-Martínez, C., Flores-Tinoco, C., Alejandri-Ramírez, N. and Dinkova, T. (2012). Translation Initiation Factor AteIF(iso)4E Is Involved in Selective mRNA Translation in Arabidopsis Thaliana Seedlings. PLoS ONE, 7(2), p.e31606.

23 McLennan, A. (2005). The Nudix hydrolase superfamily. Cellular and Molecular Life Sciences, 63(2), pp.123-143.

24 McPherson, J., Marra, M., Hillier, L., Waterston, R., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E., Wilson, R., Fulton, R., Kucaba, T., Wagner-McPherson, C., Barbazuk, W., Gregory, S., Humphray, S., French, L., Evans, R., Bethel, G., Whittaker, A., Holden, J., McCann, O., Dunham, A., Soderlund, C., Scott, C., Bentley, D., Schuler, G., Chen, H., Jang, W., Green, E., Idol, J., Maduro, V., Montgomery, K., Lee, E., Miller, A., Emerling, S., Kucherlapati, R., Gibbs, R., Scherer, S., Gorrell, J., Sodergren, E., Clerc-Blankenburg, K., Tabor, P., Naylor, S., Garcia, D., de Jong, P., Catanese, J., Nowak, N., Osoegawa, K., Qin, S., Rowen, L., Madan, A., Dors, M., Hood, L., Trask, B., Friedman, C., Massa, H., Cheung, V., Kirsch, I., Reid, T.,

Yonescu, R., Weissenbach, J., Bruls, T., Heilig, R., Branscomb, E., Olsen, A., Doggett, N., Cheng, J., Hawkins, T., Myers, R., Shang, J., Ramirez, L., Schmutz, J., Velasquez, O., Dixon, K., Stone, N., Cox, D., Haussler, D., Kent, W., Furey, T., Rogic, S., Kennedy, S., Jones, S., Rosenthal, A., Wen, G., Schilhabel, M., Gloeckner, G., Nyakatura, G., Siebert, R., Schlegelberger, B., Korenberg, J., Chen, X., Fujiyama, A., Hattori, M., Toyoda, A., Yada, T., Park, H., Sakaki, Y., Shimizu, N., Asakawa, S., Kawasaki, K., Sasaki, T., Shintani, A., Shimizu, A., Shibuya, K., Kudoh, J., Minoshima, S., Ramser, J., Seranski, P., Hoff, C., Poustka, A., Reinhardt, R. and Lehrach, H. (2001). A physical map of the human genome. Nature, 409(6822), pp.934-941.

25 Mendoza-Vargas, A., Olvera, L., Olvera, M., Grande, R., Vega-Alvarado, L., Taboada, B., Jimenez-Jacinto, V., Salgado, H., Juárez, K., Contreras-Moreira, B., Huerta, A., Collado-Vides, J. and Morett, E. (2009). Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in E. coli. PLoS ONE, 4(10), p.e7526.

26 Messing, S., Gabelli, S., Liu, Q., Celesnik, H., Belasco, J., Piñeiro, S. and Amzel, L. (2009). Structure and Biological Function of the RNA Pyrophosphohydrolase BdRppH from Bdellovibrio bacteriovorus. Structure, 17(3), pp.472-481.

27 Mildvan, A., Xia, Z., Azurmendi, H., Saraswat, V., Legler, P., Massiah, M., Gabelli, S., Bianchet, M., Kang, L. and Amzel, L. (2005). Structures and mechanisms of Nudix hydrolases. Archives of Biochemistry and Biophysics, 433(1), pp.129-143.

28 Montoya, J., Christianson, T., Levens, D., Rabinowitz, M. and Attardi, G. (1982). Identification of initiation sites for heavy-strand and light-strand transcription in human mitochondrial DNA. Proceedings of the National Academy of Sciences, 79(23), pp.7195-7199.

29 Niedzwiecka, A., Marcotrigiano, J., Stepinski, J., Jankowska-Anyszka, M., Wyslouch-Cieszynska, A., Dadlez, M., Gingras, A., Mak, P., Darzynkiewicz, E., Sonenberg, N., Burley, S. and Stolarski, R. (2002). Biophysical Studies of eIF4E Cap-binding Protein: Recognition of mRNA 5′ Cap Structure and Synthetic Fragments of eIF4G and 4E-BP1 Proteins. Journal of Molecular Biology, 319(3), pp.615-635.

30 Pedersen, A., Baldi, P., Chauvin, Y. and Brunak, S. (1999). The biology of eukaryotic promoter prediction—a review. Computers & Chemistry, 23(3-4), pp.191-207.

31 Piton, J., Larue, V., Thillier, Y., Dorleans, A., Pellegrini, O., Li de la Sierra-Gallay, I., Vasseur, J., Debart, F., Tisne, C. and Condon, C. (2013). Bacillus subtilis RNA deprotection enzyme RppH recognizes guanosine in the second position of its substrates. Proceedings of the National Academy of Sciences, 110(22), pp.8858-8863.

32 Reddy, T., Thomas, A., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E. and Kyrpides, N. (2014). The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Research, 43(D1), pp.D1099-D1106.

33 Reeder, R., Sollner-Webb, B. and Wahn, H. (1977). Sites of transcription initiation in vivo on Xenopus laevis ribosomal DNA. Proceedings of the National Academy of Sciences, 74(12), pp.5402-5406.

34 Richards, J., Liu, Q., Pellegrini, O., Celesnik, H., Yao, S., Bechhofer, D., Condon, C. and Belasco, J. (2011). An RNA Pyrophosphohydrolase Triggers 5'-Exonucleolytic Degradation of mRNA in Bacillus subtilis. Molecular Cell, 43(6), pp.940-949.

35 Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muniz-Rascado, L., Garcia-Sotelo, J., Weiss, V., Solano-Lira, H., Martinez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernandez, S., Alquicira-Hernandez, K., Lopez-Fuentes, A., Porron-Sotelo, L., Huerta, A., Bonavides-Martinez, C., Balderas-Martinez, Y., Pannier, L., Olvera, M., Labastida, A., Jimenez-Jacinto, V., Vega-Alvarado, L., del Moral-Chavez, V., Hernandez-Alvarez, A., Morett, E. and Collado-Vides, J. (2012). RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Research, 41(D1), pp.D203-D213.

36 Sanger, F., Air, G., Barrell, B., Brown, N., Coulson, A., Fiddes, J., Hutchison, C., Slocombe, P. and Smith, M. (1977). Nucleotide sequence of bacteriophage ϕX174 DNA. Nature, 265(5596), pp.687-695.

37 Shendure, J. (2005). Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. Science, 309(5741), pp.1728-1732.

38 The *C. elegans* Sequencing Consortium, (1998). Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. Science, 282(5396), pp.2012-2018.

39 van Vliet, A. (2010). Next generation sequencing of microbial transcriptomes: challenges and opportunities. FEMS Microbiology Letters, 302(1), pp.1-7.

40 Venter, J. (2001). The Sequence of the Human Genome. Science, 291(5507), pp.1304-1351.

41 Wade, J. and Grainger, D. (2014). Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Micro, 12(9), pp.647-653.

42 Weiss, V., Medina-Rivera, A., Huerta, A., Santos-Zavaleta, A., Salgado, H., Morett, E. and Collado-Vides, J. (2013). Evidence classification of high-throughput protocols and confidence integration in RegulonDB. Database, 2013(0), pp.bas059-bas059.

43 Werner, T. (1999). Models for prediction and recognition of eukaryotic promoters. Mammalian Genome, 10(2), pp.168-175.

44 Zajac, P., Islam, S., Hochgerner, H., Lönnerberg, P. and Linnarsson, S. (2013). Base Preferences in Non-Templated Nucleotide Incorporation by MMLV-Derived Reverse Transcriptases. PLoS ONE, 8(12), p.e85270.

45 Zuberek, J., Jemielity, J., Jablonowska, A., Stepinski, J., Dadlez, M., Stolarski, R. and Darzynkiewicz, E. (2004). Influence of Electric Charge Variation at Residues 209 and 159 on the Interaction of eIF4E with the mRNA 5‘ Terminus †. Biochemistry, 43(18), pp.5370-5379.