

Preparación de librerías de ARN Agilent SureSelect XT HS: un flujo de trabajo mejorado y agilizado para la construcción directa de librerías de secuenciación de ARN a partir de muestras FFPE (parafinadas, del inglés Formalin-Fixed Paraffin-Embedded) o frescas

Autores

Carsten Carstens, Katherine
Felts y Sarah Johns
Agilent Technologies, Inc.

Resumen

En esta nota de aplicación presentamos un flujo de trabajo condensado y mejorado para construir librerías de secuenciación de ARN diana. El protocolo de Agilent SureSelect XT RNA Direct incluye cuatro mejoras principales:

1. Sustitución de una hibridación de enriquecimiento dirigido durante una noche por una hibridación rápida de 90 minutos.
2. Eliminación del tratamiento con uracilo desglucosilación (UDG): la especificidad de la hebra de ARN se mantiene mediante el uso de una enzima nueva.
3. Sustitución de los antiguos adaptadores SureSelect XT por los SureSelect XT HS, que permiten procesar de manera paralela ADN y ARN de la misma muestra.
4. Se proporciona un código de barras molecular (UMI) exclusivo para mejorar la eliminación de duplicados de PCR y de fragmentación.

Estos cambios han reducido el tiempo de obtención de resultados desde dos o tres días a tan solo uno. Además, analizando los datos de fusión con el uso de distintos tipos de muestras de partida (intacta, congelada en fresco, FFPE), hemos observado que este flujo de trabajo agilizado produce librerías de secuenciación de ARN de mayor complejidad y mejor rendimiento de secuenciación, con muestras de tan solo 10 ng.

Introducción

La aplicación de la secuenciación de alta productividad a la transcriptómica (secuenciación de ARN o ARN-seq) no solo ha permitido el análisis global de la expresión de los genes, sino que ha aportado información precisa sobre las variantes de corte y empalme, los transcritos de fusión, los eventos de modificación postraducción y la expresión de alelos específicos.

La manera habitual de generar librerías de ARN-seq consiste en la fragmentación de las muestras de ARN mediante la transesterificación con Mg^{2+} , la síntesis de ADNc con una primera hebra cebada con hexámeros aleatorios y la síntesis de una segunda hebra de ADNc marcada de con dUTP. Para evitar la concatenación, se eliminan los extremos del ADNc y se añade una adenina a los extremos 3' mediante polimerización de ADN sin molde. A esto le sigue el ligamiento de adaptadores de secuenciación específicos para la plataforma. Todos los pasos posteriores al ligamiento de los adaptadores son compartidos con el flujo de trabajo de secuenciación del ADN.

A pesar de que las librerías de secuenciación de ADN y ARN derivadas parecen idénticas después de ligar los adaptadores, hay diferencias concretas al comparar la secuenciación de ADN y la de ARN. La primera diferencia es el requisito de mantener la direccionalidad de los fragmentos de secuenciación para confirmar a qué hebra de ADN genómico corresponde el ARN original. Esto se suele lograr incluyendo uracilo en la reacción de síntesis de la segunda hebra, continuando con la preparación de la librería y, en un paso posterior, usando UDG para eliminar la segunda hebra. La segunda diferencia es el amplio rango dinámico de los transcritos medidos en la secuenciación del ARN, que puede ser superior en cinco órdenes de magnitud, debidos a las grandes diferencias en la abundancia relativa de los transcritos codificantes expresados y a la abundancia extremadamente alta de ARN ribosómico (ARNr) en comparación con el ARN codificante. Por tanto, el ARN-seq requiere reducir la complejidad para evitar desperdiciar lecturas de transcritos no informativos (ej.: ARNr). El método más común consiste en eliminar el ARNr disminuyendo directamente el número de ribosomas (ribodepleción) o capturar de manera específica el ARN poliadenilado (ARNm). En una muestra típica, incluso tras eliminar el ARNr, el 1% de todos los genes expresados es responsable de aproximadamente el 50% de todos los transcritos. Por lo tanto, si los objetivos de la investigación precisan el estudio de genes con una expresión baja o media, esta detección se beneficiará de una reducción de la complejidad.

Como alternativa a los métodos de ribodepleción, es posible lograr una reducción de la complejidad mediante enriquecimiento dirigido usando sondas biotiniladas (llamadas «cebos»), una estrategia habitual en la secuenciación genómica. En la secuenciación de ARN, el enriquecimiento mediante selección de cebos se asocia principalmente con muestras FFPE, donde la depleción del ARNr es decididamente irregular y el enriquecimiento con poli(A) no se puede usar a causa de la fragmentación del material de partida¹. El enriquecimiento dirigido también es ventajoso cuando solo hace falta examinar un subconjunto relativamente pequeño del transcriptoma. Un ejemplo destacado es la detección de transcritos de fusión indicadores de una fusión génica subyacente. No obstante, un supuesto en el que tan solo un subconjunto de transcritos es informativo, como el análisis de los transcritos o la detección de ediciones postranscripcionales poco frecuentes, se benefician del enriquecimiento dirigido².

Agilent ofrece el kit de preparación de librerías de Agilent SureSelect XT RNA Direct (ref. G7564A, G7564B) para construir librerías de secuenciación de ARN diana. Hemos demostrado con éxito el uso de este kit con muestras FFPE³. A continuación, se describe un flujo de trabajo mejorado y agilizado que combina componentes del kit de preparación de librerías de SureSelect XT RNA Direct y el kit enriquecimiento dirigido SureSelect XT HS (ref. G9706A) acompañado de un protocolo modificado. El flujo de trabajo se ha simplificado eliminando la liofilización del ARN de partida, quitando el paso de purificación de perlas de SPRI y sustituyendo el paso de hibridación del cebo de 24 horas por un paso de hibridación rápida de 1,5 horas. También hemos eliminado el paso de tratamiento con UDG, con lo que la especificidad se consigue mediante el uso de una enzima de PCR que discrimina el molde de ADN que contiene uracilo durante la PCR de la precaptura. Además, el uso de los adaptadores de secuenciación SureSelect XT HS añade un UMI para identificar duplicados de fragmentación y reduce las barreras prácticas para realizar un procesamiento paralelo del ADN y el ARN de la misma muestra.

Experimento

Fuentes de ARN

Se obtuvo ARN de referencia humano universal (Universal Human Reference RNA, UHRR) como material congelado en fresco de Agilent Technologies (Santa Clara, California, EE. UU. ref. 750500-41). Se obtuvo un grupo de tejidos relacionados de tumor mamario y tejido adyacente normal en forma de muestras congeladas en fresco y FFPE de la empresa CRO de muestras biológicas humanas CureLine (Brisbane, California, EE. UU., referencia personalizada). El material de ARN de referencia de fusión tumoral FFPE Seraseq v2 fue adquirido a SeraCare (Gaithersburg, Maryland, EE. UU., ref. 0710-0129).

Aislamiento del ARN

Cuando fue necesario, se aisló al ARN mediante el kit FFPE RNeasy o el mini kit RNeasy de Qiagen siguiendo las instrucciones del fabricante (Qiagen USA, Germantown, Maryland, EE. UU., ref. 73504 y 74104, respectivamente). Si desea un protocolo más detallado, consulte el apéndice.

Evaluación de la calidad del material de partida y de las librerías de secuenciación

Se evaluaron muestras de ácidos nucleicos en el Sistema Bioanalyzer Agilent 2100 (Agilent Technologies, ref. G2939B) utilizando bien el kit ARN 6000 pico de Agilent (Agilent Technologies, ref. 5067-1513) para la puntuación de calidad del ARN, bien el kit ADN 1000 de Agilent (Agilent Technologies, ref. 5067-1504) para la evaluación de la calidad de las librerías de secuenciación.

Otros materiales

La actinomicina D se adquirió a Sigma (St. Louis, Missouri, EE.UU., ref. A1410) y se añadió hasta obtener una concentración de 4 $\mu\text{g}/\mu\text{l}$ en una solución madre de DMSO. Las purificaciones en perlas SPRI se llevaron a cabo con perlas AMPure XP (Beckman Coulter, Atlanta, Georgia, EE. UU., ref. A63880). La captura de las sondas biotiniladas se llevó a cabo con perlas de estreptavidina T1 Dynabeads MyOne (Thermo Fisher Scientific, Waltham, Massachusetts, EE. UU., ref. 65601).

Preparación de librerías de ARN con SureSelect XT HS

La construcción de librerías de secuenciación de ARN se llevó a cabo usando el kit SureSelect XT RNA Direct (Agilent Technologies, ref. G7564A) y el sistema de enriquecimiento dirigido SureSelect XT HS para la librería de secuenciación múltiple de extremos emparejados Illumina (Agilent Technologies, ref. G9706A). Si desea una descripción más detallada, consulte el apéndice.

Preparación de librerías SureSelect XT RNA Direct

Se generaron las librerías de ARN directo, se enriquecieron y se secuenciaron siguiendo las instrucciones del protocolo (manual) de preparación de librerías SureSelect XT RNA Direct.

Enriquecimiento dirigido

El enriquecimiento dirigido de las librerías de ARN SureSelect XT HS se realizó usando el exoma completo SureSelect Human All Exon V7 (Agilent Technologies, ref. 5191-4029), dirigido al transcriptoma codificante. En el apéndice se describe un protocolo detallado para la captura de los cebos.

Secuenciación y análisis de datos

Las librerías de secuenciación se analizaron con un Illumina HiSeq 4000 mediante secuenciación de extremos emparejados (*paired-end*) usando un formato de lectura 2 × 150 pb. En el análisis de la expresión (datos no mostrados), los archivos FASTQ se alinearon con el transcriptoma mediante la versión 2.6.0a de STAR, que puede alinear empalmes, usando la secuencia del genoma de referencia hg38. Después se generaron los perfiles de expresión a partir del resultado de alineación de STAR usando la herramienta RSEM. Las estadísticas generales de la librería (especificidad de las hebras, sesgo de extremos 5'-3', tasas de duplicación sin contar UMI, estimación del tamaño de la librería) se generaron mediante el procedimiento de análisis de ARN de Picard con duplicados marcados usando archivos .bam que se redujeron a 2 × 10⁷ lecturas para generar tasas de duplicación normalizadas. Las estadísticas de duplicación corregidas para los UMI y del tamaño estimado de las librerías se generó mediante el

mismo procedimiento, pero usando UmiAwareMarkDuplicatesWithMateCigar para eliminar los duplicados de fragmentación. Los transcritos de fusión se clasificaron con STAR-Fusion

y se visualizaron con la herramienta FusionInspector, que forma parte del kit Cancer Transcriptome Analysis Tool (CTAT) de Trinity⁴.

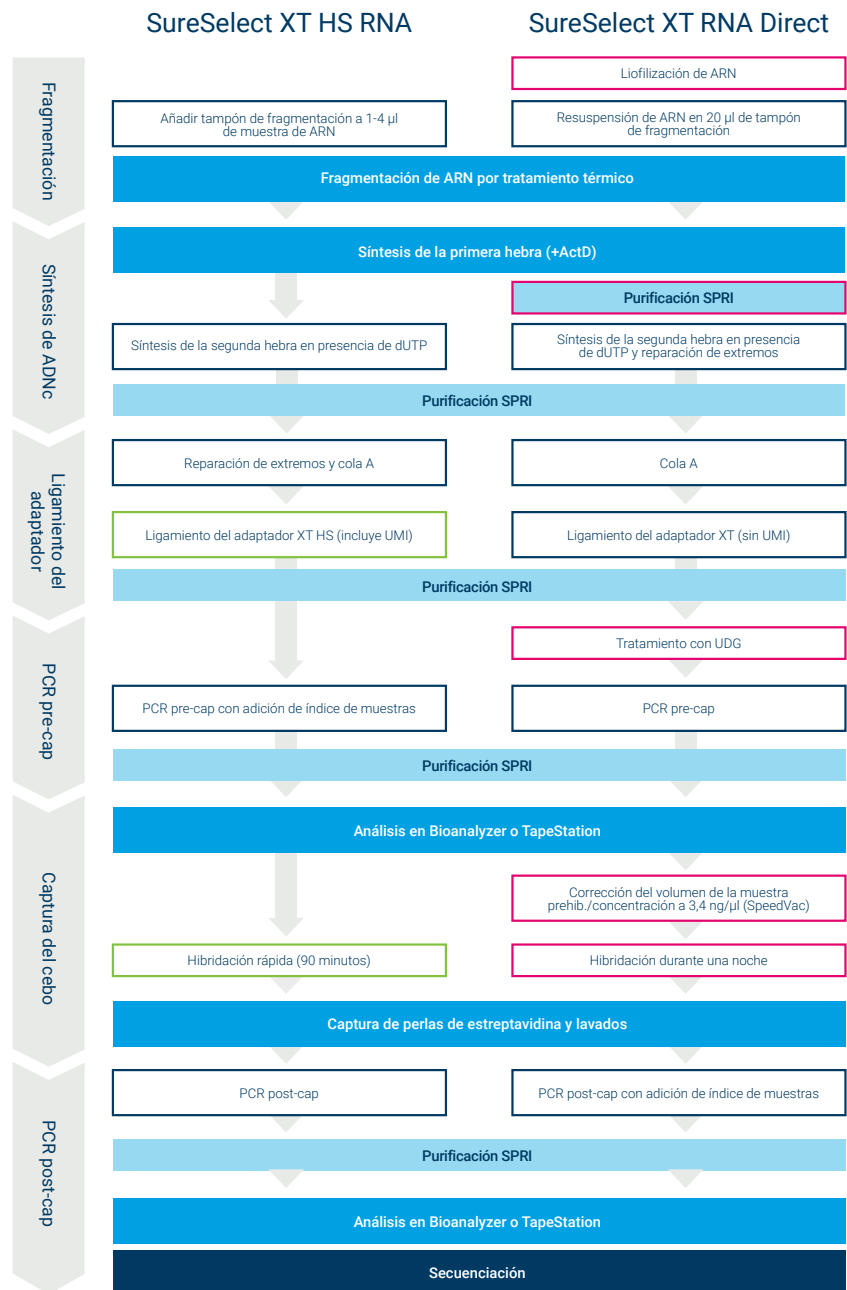


Figura 1. Comparación del flujo de trabajo de los kits de preparación de librerías Agilent SureSelect XT HS RNA (RNA XT HS) y Agilent SureSelect XT RNA Direct (RNA Direct). Los pasos eliminados del flujo se representan con un borde rosa. El borde verde representa los pasos que mejoran las prestaciones del protocolo RNA XT HS.

Resultados y comentarios

Visión general de los flujos de trabajo de RNA XT HS y RNA Direct

Deseábamos desarrollar un flujo de trabajo mejorado y simplificado para generar librerías de secuenciación de ARN de alta calidad combinando componentes del kit de preparación de librerías de Agilent SureSelect XT RNA Direct (RNA Direct) con el kit de enriquecimiento dirigido SureSelect XT HS. El flujo de trabajo resultante se ha llamado SureSelect XT HS RNA (RNA XT HS). La Figura 1 muestra un resumen comparativo de estos flujos de trabajo. El flujo de trabajo de RNA XT HS posee ventajas evidentes con respecto al flujo de RNA Direct, como la eliminación del tratamiento con UDG y la sustitución de la hibridación tradicional de una noche con la hibridación rápida de 90 minutos, lo que reduce enormemente el tiempo de obtención de resultados. También deseábamos determinar si es posible eliminar el paso de purificación SPRI entre la síntesis de la primera y la segunda hebra y evitar la liofilización del ARN de partida en el flujo de trabajo RNA XT HS (los detalles pueden consultarse en el apéndice).

Comparación del rendimiento de los flujos de trabajo

Para comparar el rendimiento de los dos procesos de preparación de librerías, generamos en primer lugar librerías usando diferentes cantidades de material de entrada: ARN de referencia humano universal (UHRR) y material de referencia de ARN de fusión tumoral SeraCare FFPE v2 (SeraCare), para representar, respectivamente, una muestra intacta y una muestra FFPE idealizada. Las librerías RNA XT HS y RNA Direct se enriquecieron usando SureSelect Human All Exon V7. Por último, estas librerías enriquecidas se secuenciaron con un secuenciador Illumina y los datos se analizaron usando procedimientos específicos de análisis de datos (vea la sección Experimento para obtener los detalles).

La Tabla 1 muestra un resumen de las estadísticas globales resultantes de los datos de la secuenciación de RNA XT y RNA Direct. Los resultados muestran que las librerías de RNA XT HS son

indistinguibles de las de RNA Direct en lo que respecta a varios parámetros, los más notables son la elevada especificidad de las hebras (> 98%) y la baja contaminación de ARNr (~0,1%). La elevada especificidad de las hebras de las librerías de RNA XT HS demuestra la eficacia de eliminar el tratamiento de UDG en el flujo de trabajo RNA XT HS en favor del uso de una enzima de PCR que no amplifica los moldes que contienen uracilo. La especificidad de las hebras tiende a ser ligeramente mejor en las muestras intactas que en las FFPE, lo cual es esperable debido a la merma de calidad del material de entrada de las muestras FFPE. A pesar de todo, las especificidades observadas para FFPE con cualquiera de los protocolos sigue siendo muy elevada (> 98% para el material FFPE). El porcentaje de contaminación de ARNr

es bajo en todo momento con ambos métodos de preparación de la librería y ya se ha demostrado anteriormente en estrategias de enriquecimiento dirigido³.

Cuando se comparan las tasas de mapeo exónico entre RNA XT HS y RNA Direct, volvemos a encontrar un rendimiento similar (Figura 2). Un paso crítico en el acortamiento de la duración del nuevo flujo de trabajo RNA XT HS es la sustitución de la tradicional hibridación durante una noche por un paso de hibridación rápida. Como muestra la Figura 2, la captura acelerada no afecta a las tasas de mapeo y se han observado en todo momento tasas exónicas del 90% con muy pocas lecturas correspondientes a regiones intergénicas independientemente del material de partida. Los transcritos generados por la librería de captura V7 representaron entre

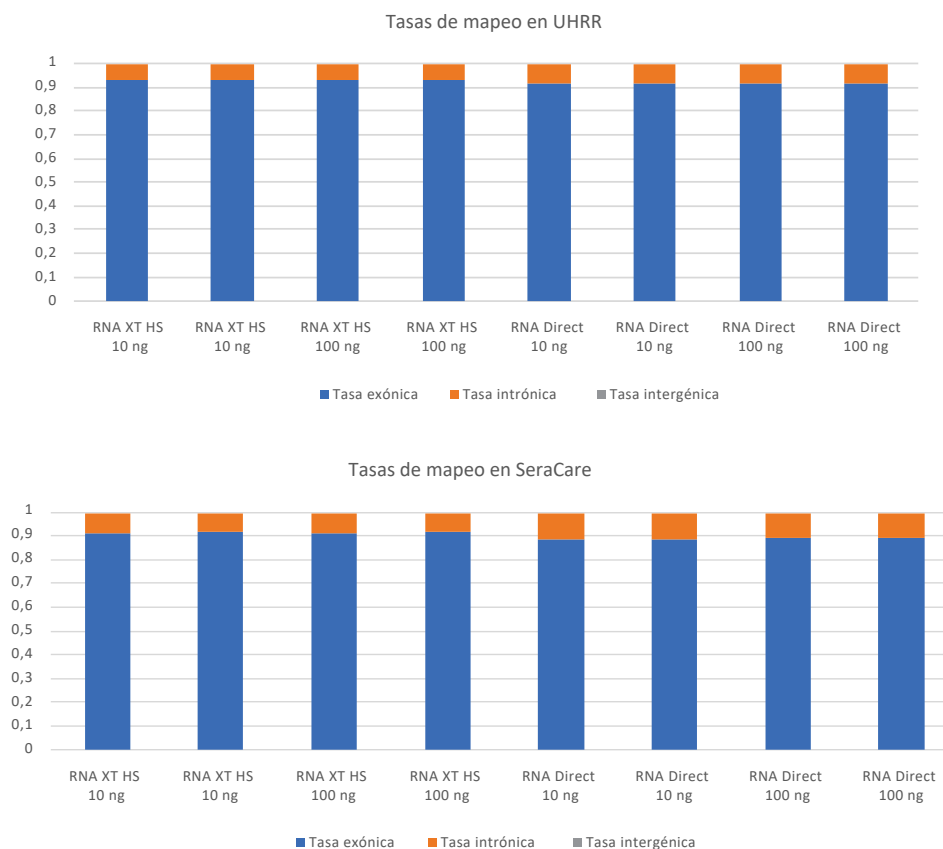


Figura 2. Comparación de la tasa de mapeo de las librerías de secuenciación de RNA XT HS y RNA Direct. Las librerías de ARN-seq se prepararon a partir de muestras UHRR (A) y SeraCare (B) usando el flujo de trabajo RNA XT HS o RNA Direct. Se muestran las tasas de mapeo de librerías de secuenciación enriquecidas del exoma Agilent SureSelect Human All Exon V7 de secuencias exónicas, intrónicas e intergénicas. Se cree que la tasa intrónica refleja la secuenciación del ARNm no procesado.

el 92,9 % y el 94,1 % de toda la expresión observada, independientemente de la cantidad de entrada o del tipo de muestra. El ~7 % restante se debe principalmente a transcritos con una elevada expresión contaminantes, como los genes mitocondriales o a errores de anotación.

Un indicador clave de la eficiencia del proceso global de preparación de la librería es la complejidad estimada de la librería inicial. Son preferibles las librerías más extensas y, por tanto, más complejas, ya que la secuenciación de estas ofrece más información y mayor confianza en la cuantificación. La complejidad de la librería se calcula a partir de la profundidad de la secuenciación (en pares de bases leídos) y el número de variantes únicas observadas, derivado de la tasa de duplicación. Para calcular las complejidades estimadas de las librerías se ha utilizado el procedimiento Picard, que asume que todos los duplicados son duplicados de PCR. Este enfoque subestima la complejidad real de la librería y se tratará con más detalle en la sección de resultados. Los resultados de complejidad de las librerías se indican en la Tabla 1 y se representan en la Figura 3. Cuando se comparan los flujos de trabajo de RNA XT HS y de RNA Direct, no se encuentran diferencias entre las bibliotecas construidas a partir de grandes cantidades (100 ng) de ARN de entrada fresco o FFPE. Sin embargo, cuando el material de partida se redujo a 10 ng, ya sea de UHRR o de SeraCare, se encontraron diferencias notables en la complejidad de la librería. En primer lugar, estas librerías formadas con menor cantidad de entrada son más pequeñas y menos complejas. Esto es lo esperado, ya que cuanto menor sea el ARN de entrada que se intenta convertir en una librería de secuenciación, menor será la librería final. También encontramos que, con cantidades bajas, el flujo agilizado de RNA XT HS produjo librerías de 1,5 a 2 veces más eficaces que con el flujo RNA Direct. Existen múltiples factores del flujo de trabajo RNA XT HS que están siendo investigados (no se muestran datos) y que podrían ser la causa de este aumento de la eficiencia.

Complejidad estimada de las librerías de secuenciación (procedimiento Picard)

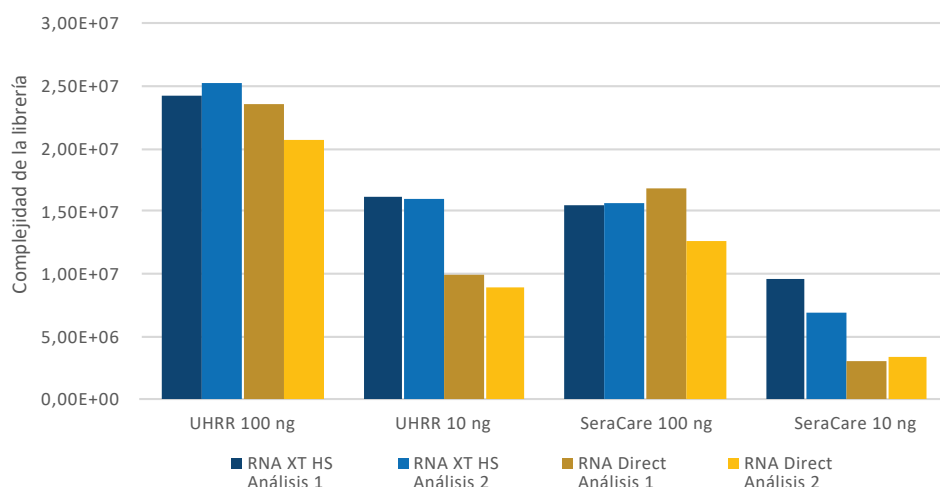


Figura 3. Diferencias de complejidad entre las librerías RNA XT HS y RNA Direct con poca cantidad de muestra de partida. Las librerías de secuenciación preparadas usando las fuentes y las cantidades de muestra indicadas se secuenciaron con una plataforma Illumina HiSeq 4000 mediante secuenciación de extremos emparejados. Los datos de complejidad de las librerías se determinaron basándose en las tasas de duplicación observadas y en el número correspondiente de pares de leídos determinados con la herramienta de análisis de ARN Picard. Este procedimiento no diferencia entre duplicados de PCR y de fragmentación.

Tabla 1. Estadísticas globales de las librerías de secuenciación SureSelect XT HS RNA y SureSelect XT RNA Direct: Las librerías de secuenciación de ARN generadas a partir del mismo material de partida, usando ya sea el protocolo RNA XT HS o el RNA Direct se secuenciaron con una plataforma Illumina HiSeq 4000. Después de reducir el tamaño a 2×10^7 lecturas, se generaron las estadísticas de las librerías mediante la herramienta de análisis de ARN Picard. Las estadísticas corregidas para los UMI se generaron con el mismo procedimiento después de marcar los UMI con UmiAwareMarkDuplicatesWithMateCigar. (A) ARN de partida congelado en fresco («intacto»); (B) muestra FFPE.

A. ARN de referencia humano universal (UHRR)

Protocolo	SureSelect XT HS RNA				SureSelect XT RNA Direct			
	100	100	10	10	100	100	10	10
Muestra de partida (ng)	100	100	10	10	100	100	10	10
Lecturas analizadas (millones)	1,9	1,9	1,9	1,9	1,9	1,9	1,9	1,9
Tasa de ARNr (%)	0,1%	0,1%	0,1%	0,1%	0,1%	0,1%	0,1%	0,1%
Especificidad de las hebras (%)	99,1%	99,1%	99,0%	99,0%	99,1%	99,2%	99,1%	99,1%
Tasa de duplicación (%)	17,3%	16,7%	24,3%	24,7%	17,9%	20,0%	35,8%	38,7%
Tamaño estimado de la librería ($\times 10^6$)	24,2	25,2	16,2	15,9	27,5	23,6	10,3	9,2
Tasa de duplicación corregida para los UMI (%)	7,1%	6,4%	17,1%	17,8%	NA	NA	NA	NA
Tamaño estimado de la librería corregido para los UMI ($\times 10^6$)	65,9	72,8	24,6	23,6	NA	NA	NA	NA

B. Material de referencia de ARN de fusión tumoral FFPE v2 (SeraCare)

Protocolo	SureSelect XT HS RNA				SureSelect XT RNA Direct			
	100	100	10	10	100	100	10	10
Muestra de partida (ng)	100	100	10	10	100	100	10	10
Lecturas analizadas (millones)	18,6	18,6	18,7	18,6	19	18,8	18,9	18,9
Tasa de ARNr (%)	0,1%	0,2%	0,1%	0,1%	0,1%	0,2%	0,1%	0,1%
Especificidad de las hebras (%)	98,6%	98,6%	98,6%	98,7%	98,9%	98,9%	98,9%	98,9%
Tasa de duplicación (%)	24,8%	24,7%	36,0%	45,0%	23,5%	29,4%	69,1%	67,2%
Tamaño estimado de la librería ($\times 10^6$)	15,5	15,5	9,6	6,9	18,6	13,3	3	3,3
Tasa de duplicación corregida para los UMI (%)	13,2%	13,7%	29,3%	38,8%	NA	NA	NA	NA
Tamaño estimado de la librería corregido para los UMI ($\times 10^6$)	32,5	31,2	12,7	8,66	NA	NA	NA	NA

Impacto de los códigos de barras moleculares (UMI) en las librerías de secuenciación de ARN

Como se ha mencionado anteriormente, la evaluación de la eficiencia del proceso de preparación de la librería requiere una medición precisa de la complejidad de la librería. Las complejidades estimadas arriba suponen que los pares leídos con el mismo inicio y el mismo final son duplicados de PCR originados a partir de la misma molécula de la librería original. No obstante, también pueden aparecer duplicados debidos a fragmentación aleatoria, que dan lugar a dos fragmentos independientes con los mismos extremos. Habida cuenta de que los duplicados de fragmentación son miembros realmente independientes de la librería de secuenciación, la determinación de la complejidad de la librería solo se debería basar en los duplicados reales de la PCR.

La secuenciación de ARN puede dar lugar a un número mayor de duplicados de fragmentación que la secuenciación de ADN debido a la elevada expresión de ciertos genes, que puede aumentar la probabilidad de aparición de duplicados de fragmentación. Esto da lugar a estimaciones poco precisas de la complejidad. Una posible ventaja del flujo de trabajo RNA XT HS es que las librerías se construyen con adaptadores XT HS que contienen UMI de 10 pb. Nuestra hipótesis es que este UMI se podría usar para diferenciar los duplicados de PCR y de fragmentación. Hemos vuelto a analizar los datos de secuenciación anteriores de RNA XT HS usando un procedimiento de análisis de datos que tiene en cuenta los datos de UMI (nota: las librerías de RNA Direct no incluyen un UMI y, por ende, no se incluyen en este segundo análisis). El resultado de este análisis queda registrado en la Tabla 1 y en la Figura 4.

Tal y como se muestra en la Figura 4, la corrección de los resultados de los duplicados de fragmentación da lugar a estimaciones más altas de la complejidad de las librerías, especialmente en presencia de mayores cantidades de entrada, donde la complejidad de la librería queda subestimada hasta 3 veces si no se toman en consideración los duplicados de fragmentación. La tendencia al alza observada en las librerías más grandes y con las cantidades de partida más elevadas

es previsible ya que aumenta la probabilidad de observar duplicados de fragmentación si se procesan más moléculas derivadas de la misma secuencia codificadora.

Deseábamos conocer el impacto de la corrección de UMI en una serie de ejemplos reales. Por este motivo, generamos librerías de RNA XT HS usando un grupo de tejidos relacionados de tumor mamario y tejido adyacente normal tanto que se almacenaron y trataron como muestras congeladas en fresco (FF, del inglés *fresh-frozen*) y muestras FFPE. Estas

librerías se enriquecieron, secuenciaron y analizaron como se ha descrito con las librerías RNA XT HS con UHRR y SeraCare. El resumen de los resultados estadísticos de estas librerías se puede consultar en la Tabla 2 y la Figura 4. Podemos observar una especificidad de las hebras y una tasa de ARNr en los datos de secuenciación de los tejidos tumorales/normales congelados en fresco y FFPE comparables con nuestras librerías con UHRR y SeraCare. También hemos observado que las librerías FFPE son significativamente más pequeñas que las

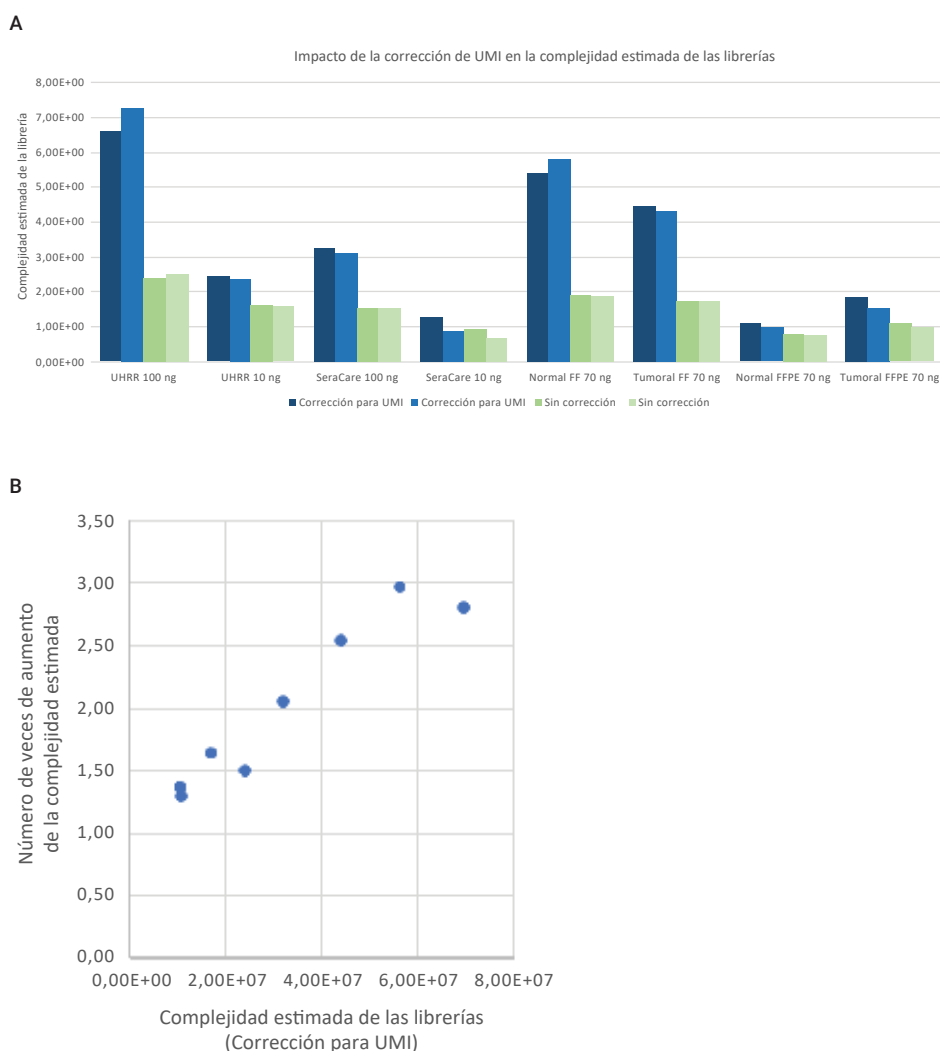


Figura 4. Impacto de la corrección de los códigos de barras moleculares (UMI) en la estimación del tamaño de las librerías. Los tamaños de las librerías de RNA XT HS se calcularon mediante con eliminación de duplicados de Picard (sin corrección) o bien después de usar los UMI (con corrección para UMI). La eliminación de duplicados de Picard supone que todos los duplicados son duplicados de PCR, mientras que la corrección para UMI diferencia entre los duplicados de PCR y de fragmentación. **A)** Complejidad estimada de la librería para varias librerías de secuenciación RNA XT HS derivadas de una serie de muestras y entradas diferentes. **B)** Las diferencias entre la complejidad estimada con o sin la corrección para los UMI son más pronunciadas en las librerías más grandes y, por tanto, con cantidades de partida mayores.

de muestras en fresco congeladas, lo cual es un fenómeno bien documentado. Cuando hemos aplicado la corrección para los UMI en nuestros datos de secuenciación FFPE, hemos visto mejoras en la complejidad estimada de las librerías de entre 0,3 y 3,5 veces y, por lo tanto, se han recuperado datos que se hubieran perdido de haber supuesto que todos los duplicados son debidos a la PCR, en lugar de diferenciar entre duplicados de PCR y de fragmentación. Esto demuestra que, para aprovechar al máximo la información de secuenciación que ofrecen las muestras de ARN FFPE, es crucial identificar con precisión los duplicados de PCR y los de fragmentación en estas muestras.

Detección de fusiones génicas

El uso de fragmentos dirigidos de ARN-seq ha demostrado ser útil para detectar fusiones génicas, sobre todo con muestras complicadas (FFPE). Hemos querido determinar si nuestro flujo RNA XT HS podría funcionar en estos casos en las condiciones de enriquecimiento con sondas de exones (en comparación con un panel específico para fusión) y lograr la máxima longitud de la secuencia. Como se ha mencionado anteriormente, sometimos a nuestras librerías RNA XT HS de todas las muestras (UHRR, SeraCare, congeladas en fresco y FFPE) a un enriquecimiento dirigido usando SureSelect Human All Exon V7 y las secuenciamos en el modo de extremos emparejados con una longitud de lectura de 2×150 pb. Los datos de secuenciación de los tejidos tumorales/normales congelados en fresco y FFPE se analizaron con STAR-Fusion y se visualizaron con FusionInspector.

El resultado de FusionInspector detectó una serie de posibles fusiones, incluidos falsos positivos tales como recombinaciones VDJ, lo que indica la presencia de células inmunitarias en algunas de las muestras. Los recombinantes VDJ se han eliminado de la Tabla 3. El análisis de los datos muestra 15 posibles fusiones génicas presentes en nuestras muestras. En 12 de los 15 compañeros de fusión existía una relación registrada con anterioridad con fusiones génicas asociadas a tumores, aunque no en la combinación observada. Una de las fusiones, FCHSD2-FAM168A (resaltada en gris en la Tabla 3), ha sido reconocida anteriormente como asociada

Tabla 2. Estadísticas globales de secuenciación de librerías RNA XT HS de muestras congeladas en fresco y FFPE Agilent SureSelect Librerías de secuenciación de ARN generadas a partir de un grupo de tejidos relacionados de tumor mamario y tejido adyacente normal, en forma de muestras congeladas en fresco y muestras FFPE siguiendo el protocolo RNA XT HS. Las librerías se secuenciaron con la plataforma Illumina HiSeq 4000. Después de reducir el tamaño a 2×10^7 lecturas, se generaron las estadísticas de las librerías mediante la herramienta de análisis de ARN Picard, con y sin marcado de los UMI.

Tipo de muestra	Congelada en fresco				FFPE			
	Normal	Normal	Tumoral	Tumoral	Normal	Normal	Tumoral	Tumoral
Muestra de partida (ng)	70	70	70	70	70	70	70	70
Lecturas analizadas (millones)	18,8	18,8	18,7	18,8	18,6	18,6	18,7	18,7
Tasa de ARNr (%)	0,2%	0,2%	0,2%	0,2%	0,2%	0,2%	0,1%	0,1%
Especificidad de las hebras (%)	99,0%	99,0%	99,1%	99,2%	98,4%	98,4%	98,8%	98,8%
Tasa de duplicación (%)	21,0%	21,0%	23,0%	23,0%	41,0%	43,0%	33,0%	35,0%
Tamaño estimado de la librería ($\times 10^6$)	19	18,8	17,3	17,3	7,9	7,4	10,9	9,8
Tasa de duplicación corregida para los UMI (%)	8,0%	8,0%	10,0%	10,0%	33,0%	35,0%	22,0%	25,0%
Tamaño estimado de la librería corregido para los UMI ($\times 10^6$)	53,9	58,2	44,7	43,1	11	9,93	18,6	15,4

Tabla 3. Detección de fusiones génicas en muestras congeladas en fresco y FFPE normales y tumorales. Se secuenciaron y analizaron las librerías de secuenciación RNA XT HS usando el procedimiento STAR-Fusion. Las posibles fusiones detectadas en algunas muestras se indican en cada fila junto con los datos que respaldan la detección de la fusión correspondiente. Las lecturas de uniones y lecturas de apoyo se expresan en forma de lecturas por millón de pares de bases únicas leídas de la librería. Cada columna representa la media de dos repeticiones técnicas. La fusión resaltada en color gris está identificada en la base de datos TGCA como fusión asociada con el cáncer de mama y que puede ser una mutación impulsora de este tumor. Los eventos de recombinación VDJ observados indican la presencia de células inmunitarias en las muestras y fueron eliminados del análisis. La detección se basa en un análisis de $0,7-1,4 \times 10^7$ pares de bases únicas leídas por cada muestra.

Posible fusión	Congelada en fresco				FFPE			
	Normal		Tumoral		Normal		Tumoral	
	Unión	Soporte	Unión	Soporte	Unión	Soporte	Unión	Soporte
MYLK-LPAR6	0,00	0,00	8,86	0,86	0,00	0,00	1,68	0,00
FCHSD2-FAM168A	0,00	0,04	5,60	0,21	0,00	0,00	2,79	0,04
FAM157A-RB1	0,00	0,00	2,59	0,04	0,00	0,00	0,91	0,00
RP4-565E6.1-HYDIN	0,00	0,00	1,47	0,00	0,00	0,00	1,33	0,00
CDR2-FRG1	0,00	0,00	0,42	0,00	0,00	0,00	0,48	0,00
SLC7A5-RP11-645C24.2	0,00	0,04	0,33	0,20	0,00	0,00	0,09	0,04
PHRF1-TXNDC5	0,00	0,00	2,78	0,17	0,00	0,00	0,00	0,00
NUDT1-AC004840.8	0,00	0,00	1,71	0,00	0,00	0,00	0,00	0,00
MPZL1-RCS1	0,00	0,00	1,34	0,06	0,00	0,00	0,00	0,00
CCDC66-SLMAP	0,00	0,00	0,96	0,00	0,00	0,00	0,00	0,00
UBE2Q2-C15orf27	0,00	0,00	0,77	0,06	0,00	0,00	0,00	0,00
POLR2J-AC004980.9	0,00	0,00	0,00	0,00	11,53	0,00	0,00	0,00
POLR2J-UPK3B	0,00	0,00	0,00	0,00	3,96	0,00	0,00	0,00
RP11-634B7.4-TRIM58	0,36	0,00	0,00	0,00	0,00	0,00	0,00	0,00
RB1-MYLK-AS1	0,00	0,00	0,16	0,03	0,00	0,00	0,00	0,00

al cáncer de mama y puede llegar a ser una mutación impulsora de este tumor⁵. En el resto de fusiones específicas, al menos un compañero de fusión (normalmente ambos) figura en la base de datos TCGA y está asociado con la formación de tumores. Hemos encontrado el transcrito de fusión A:B y tanto A como B aparecen en la base de datos de tumores. Sin embargo, no aparece la fusión A:B. De todos modos sí hay ejemplos de transcritos A:C y D:B. Estos datos sugieren que las librerías RNA XT HS enriquecidas con un exoma permiten a los investigadores identificar fusiones a nivel de transcriptoma completo, incluso en muestras complicadas.

Conclusión

Con el fin de establecer un método más eficiente para la construcción de librerías de ARN-seq dirigido, hemos propuesto una mejora del flujo de trabajo actual SureSelect XT RNA Direct sirviéndonos de componentes del kit de preparación de librerías SureSelect XT RNA Direct junto con el kit de enriquecimiento dirigido SureSelect XT HS. El nuevo flujo de trabajo SureSelect XT HS RNA incorpora varias mejoras significativas:

- Se ha sustituido la liofilización del ARN de partida por la adición directa del tampón de fragmentación a la muestra de ARN.
- Se ha eliminado el paso de purificación de perlas SPRI tras la síntesis de la primera hebra de ADNc.
- Se ha utilizado una enzima de PCR alternativa que no amplifica los moldes que contienen uracilo, con lo que deja de ser necesario el tratamiento con UDG para mantener la especificidad de las hebras.
- Se ha modificado el procedimiento de captura de la sonda («cebo»), de un paso de hibridación de 24 horas a un protocolo de hibridación rápida de 1,5 horas.

- Se han sustituido los adaptadores de secuenciación XT por los adaptadores XT HS, lo que permite usar UMI para estimar de manera más precisa la complejidad de las librerías y «recuperar» lecturas de secuenciación.

Hemos observado que el flujo de trabajo RNA XT HS reduce el tiempo requerido para el enriquecimiento dirigido y la preparación de la librería, de 2-3 días a 1-2 días.

Este nuevo flujo de trabajo optimizado produce librerías de ARN-seq que resultan indistinguibles de las librerías generadas con el protocolo SureSelect XT RNA Direct en lo relativo a especificidad de las hebras, tasas de ARNr y tasas de mapeo con grandes cantidades de material de partida. Con cantidades más pequeñas (10 ng) hemos demostrado que el flujo de trabajo agilizado no afecta al rendimiento global y, de hecho, parece ser más eficiente. También hemos observado que incluir UMI en el diseño perfeccionado del adaptador (XT HS) permite ahora detectar los duplicados de fragmentación. Esta nueva posibilidad de identificar los duplicados de fragmentación permite mejorar los resultados de la secuenciación evitando la pérdida de lecturas que ocurre al usar los métodos habituales de duplicación de inicio y parada.

Cuando ensayamos muestras «de la vida real», encontramos que el flujo de trabajo RNA XT HS siguió produciendo datos de alta calidad incluso con muestras FFPE. El análisis preliminar de las fusiones génicas detectó posibles fusiones en nuestras muestras tumorales congeladas en fresco y FFPE, lo que permite pensar en usar el método RNA XT HS para secuenciar ARN de muestras FFPE. Aunque no mostramos datos de expresión génica o empalmes, nuestros análisis preliminares indican que el ARN-seq diana puede usarse en el análisis de la expresión génica global, detección de variantes de empalme, detección de expresión de variantes y análisis de expresión de alelos específicos. Por último, el desarrollo del flujo de trabajo SureSelect XT HS está en sintonía con el kit de enriquecimiento dirigido DNA SureSelect XT HS, con lo que se facilita la posibilidad de secuenciar en paralelo ARN y ADN de

la misma muestra. Esto podría resultar ventajoso en numerosas aplicaciones, como la investigación multiómica general y el procesado paralelo de muestras para análisis TMB-MSI y detección de fusiones.

Abreviaturas

FFPE: fijado en formol e incluido en parafina; TPM: transcritos por millón de kilobases; nt: nucleótido; UHRR: ARN de referencia humano universal; UMI: código de barras molecular; UDG: uracilo ADN desglucosilación.

Referencias

- 1) Cieslik, M., *et al.* The Use of Exome Capture RNA-Seq for Highly Degraded RNA with Application to Clinical Cancer Sequencing. *Genome Res.* **2015**, 25, 1372–1381.
- 2) Mittempergher, L., *et al.*, MammaPrint and BluePrint Molecular Diagnostics Using Targeted RNA Next-Generation Sequencing Technology. *J. Mol. Diagn.* **2019**, 21, 808–823.
- 3) Jones, J. C.; Alex Siebold, A.; Lucas, A. B. SureSelect XT RNA Direct Protocol Provides Simultaneous Transcriptome Enrichment and Ribosomal Depletion of FFPE RNA. Nota de aplicación de Agilent Technologies, número de publicación 5991-8119EN, **2017**.
- 4) Haas, B., *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. bioRxiv. 120295 (**2017**).
- 5) Hu, X., *et al.* TumorFusions: An Integrative Resource for Cancer-Associated Transcript Fusions. *Nucleic Acids Res.* **2018**, 4, 46(D1), D1144-D1149.

Apéndice

1. Protocolo detallado para la construcción de librerías de secuenciación de ARN

Preparación de las muestras de ARN

Se aisló el ARN total a partir de cortes de FFPE usando el kit RNeasy FFPE de Qiagen siguiendo las instrucciones del fabricante. El ARN procedente de tejido congelado se aisló usando el minikit RNeasy de Qiagen. Todas las muestras de ARN total se analizaron en un Sistema Bioanalyzer Agilent 2100 usando un kit ARN 6000 pico de Agilent. Las muestras se calentaron a 80 °C durante 2 minutos antes de cargarlas en el chip. Los valores RIN y DV200 se calcularon mediante el software del sistema Bioanalyzer. Estos parámetros de calidad de la muestra de las muestras de ARN total ensayadas están registradas en la Tabla S1.

Preparación del ADNc usando reactivos del kit Agilent SureSelect XT RNA Direct

Nota: Se preparó de antemano una solución patrón de 4 µg/µl de actinomicina-D en DMSO y se almacenó congelada en alícuotas de un solo uso (3 µl) a -20 °C.

Nota: La mezcla de fragmentación, la mezcla maestra de la primera hebra, las mezclas de la enzima de la segunda hebra y de los oligos del kit Agilent SureSelect XT RNA Direct se descongelaron en hielo y se mezclaron en vórtex durante 5 segundos a alta velocidad y después se centrifugaron brevemente antes de su uso.

1. Las muestras de ARN total se prepararon en un volumen de 4 µl de nucleasas sin agua. Las cantidades de partida variaron entre los experimentos y se indican en la presentación de los resultados.

Nota: Se permiten volúmenes pequeños de ARN de partida (< 4 µl), pero no se recomiendan volúmenes de ARN de partida superiores (>4 µl).

2. Se añadió la mezcla de fragmentación a la muestra de ARN hasta alcanzar un volumen final de muestra de 20 µl.
3. La muestra de ARN se fragmentó calentándola en un SureCycler 8800 (o un termociclador equivalente) usando las condiciones basadas en los

Tabla S1. Parámetros de calidad de la muestra y condiciones de fragmentación.

Descripción de la muestra	RIN	DV200	Fragmentación
ARN de referencia humano universal (UHRR)	9,2	94%	94 °C, 8 min
Material de referencia de ARN de fusión tumoral Sera-seq FFPE v2	2	54%	94 °C, 3 min, 65 °C, 2 min
Tejido mamario normal congelado	6,3	94%	94 °C, 8 min
Tejido mamario tumoral congelado	5	88%	94 °C, 8 min
Tejido mamario normal FFPE	2,1	48%	65 °C, 5 min
Tejido mamario tumoral FFPE	2	47%	65 °C, 5 min

Tabla S2. Mezcla de reacción de la primera hebra.

Reactivo	Volumen para una reacción	Volumen para ocho reacciones + 10 % de exceso
Mezcla maestra de la primera hebra ARN-seq	8 µl	70,4 µl
Actinomicina D (120 ng/µl)	0,5 µl	4,4 µl
Total	8,5 µl	74,8 µl

Tabla S3. Parámetros del protocolo de lavado SPRI

Volumen de perlas AMPure	105 µl (1,8x volumen)
Tiempo de incubación de las perlas	5 minutos
Lavado con etanol 70% (Realizar dos veces)	200 µl
Tiempo de secado a 37 °C	1-2 min o menos
Volumen de elución	50 µl nucleasas sin agua

parámetros de calidad de la muestra de ARN individual recomendadas en el protocolo del kit Agilent SureSelect XT RNA Direct (G9691). Los parámetros de fragmentación usados para las muestras se indican en la Tabla S1. Tras la fragmentación, las muestras se almacenaron en hielo hasta que estuvieran listas para proceder con el paso de síntesis de la primera hebra.
4. Se diluyó una disolución madre de 4 µg/µl de actinomicina-D en DMSO con agua hasta obtener una concentración de 120 ng/µl (3 µl de actinomicina-D + 97 µl de agua).
5. Se preparó una mezcla de reacción a granel de ocho muestras para la síntesis de la primera hebra (Tabla S2). La mezcla de reacción se mezcló en vórtex y se mantuvo en hielo hasta ser usada.
6. Se añadieron 8,5 µl de la mezcla de reacción de la primera hebra a cada muestra fragmentada de 20 µl en hielo. Las muestras se mezclaron en vórtex y se centrifugaron brevemente.

7. Las mezclas de reacción de 28,5 µl se incubaron en un SureCycler 8800 preprogramado durante 10 minutos a 25 °C, después 40 minutos a 37 °C y finalmente se almacenaron a 4 °C (o en hielo) hasta proceder con la síntesis de la segunda hebra.
8. Los tubos de mezcla tanto de la segunda hebra + enzima de reparación de extremos como los de la segunda hebra de ARN-seq + oligos de reparación de extremos se agitaron en vórtex antes de su uso.
9. Se añadieron 25 µl de la segunda hebra + enzima de reparación de extremos (tapón azul) a la mezcla de reacción de 28,5 µl de la primera hebra.
10. Inmediatamente después se añadieron 5 µl de la mezcla de la segunda hebra + oligos de reparación de extremos (tapón amarillo).
11. Las muestras se taparon, se mezclaron en vórtex, se centrifugaron brevemente y luego se volvieron a poner en hielo.

12. Las mezclas de reacción de 58,5 µl se incubaron en un SureCycler 8800 preprogramado durante 60 minutos a 16 °C y después se almacenaron a 4 °C (o en hielo) hasta proceder con la purificación con SPRI.
13. Después de la síntesis de la segunda hebra, el ADNc se purificó con SPRI usando el protocolo descrito en la Tabla S3.

Preparación de las librerías de ADNc SureSelect XT HS usando reactivos del kit de preparación de librerías SureSelect XT HS

14. La reparación de extremos, la adición de la cola dA y el ligamiento del adaptador XT HS del ADNc se llevaron a cabo usando los reactivos y las instrucciones descritas en el sistema de enriquecimiento dirigido SureSelect XT HS para el protocolo y el kit de secuenciación múltiple de extremos emparejados (G9702).

Nota: El protocolo SureSelect XT HS se siguió a partir del Paso 3 en la página 27 con las excepciones y modificaciones descritas en los pasos del 15 al 20 de este protocolo.

15. En el paso 6 de la página 34 del protocolo SureSelect XT HS. La amplificación por PCR previa a la captura (PreCap) de las librerías se llevó a cabo siguiendo las instrucciones en ese protocolo con las modificaciones relativas al número de ciclos de PCR indicados a continuación:
 - a. UHRR, Seraseq™ v2 durante 12 ciclos de PCR (ARN de alta calidad)
 - a. Muestras de cuadrante mamario durante 14 ciclos de PCR (ARN de baja calidad)
16. Las librerías PreCap fueron analizadas en el Bioanalyzer o la TapeStation para averiguar la producción y la distribución de los tamaños de las moléculas de la librería.

Enriquecimiento dirigido y secuenciación de librerías de ADNc SureSelect XT HS

17. En el paso 1 de la página 46 del protocolo SureSelect XT HS. El enriquecimiento dirigido se llevó a cabo usando 200 ng de librería PreCap como material de entrada y se usaron reactivos de hibridación rápida SureSelect XT HS. Para la hibridación se usaron sondas de exoma de 5 µl de SureSelect Human All Exon V7.
18. La captura de las perlas de estreptavidina y los lavados posteriores se realizaron de acuerdo con el protocolo.
19. La amplificación por PCR posterior a la captura (PostCap) se llevó a cabo según el protocolo, usando 12 ciclos de PCR para todas las muestras. Las librerías PostCap fueron analizadas en el Bioanalyzer o la TapeStation para averiguar la producción y la distribución de los tamaños de las moléculas de la librería.
20. Todas las librerías se secuenciaron en una plataforma Illumina HiSeq 4000 a una longitud de lecturas de 2 × 150 pb.

www.agilent.com

**Solo para uso en investigación.
Prohibido su uso en procedimientos diagnósticos.**

Esta información está sujeta a cambios sin previo aviso.

PR7000-2381
© Agilent Technologies, Inc. 2019, 2020
Impreso en EE. UU., 29 de enero de 2020
5994-1644ES

