Note d'application

Génomique

Agilent

Trusted Answers

Préparation de librairies Agilent « SureSelect XT HS RNA » Un workflow simplifié et amélioré de la construction de librairie de séquençage d'ARN « Direct-to-Capture » à partir d'échantillon frais ou FFPE

Auteurs

Carsten Carstens, Katherine Felts et Sarah Johns

Agilent Technologies, Inc.

Résumé

Dans cette note d'application, nous présentons un workflow condensé et amélioré pour la construction de banques de séquençage d'ARN cibles. Nous avons apporté quatre améliorations importantes au protocole Agilent « SureSelect XT RNA Direct »

1. Remplacement de l'hybridation en une nuit nuit pour l'enrichissement des cibles par une hybridation rapide de 90 minutes
2. Elimination du traitement par uracile déglycosylase (UDG) (la spécificité du brin d'ARN est maintenue avec une autre enzyme)
3. Remplacement des anciens adaptateurs SureSelect XT par des adaptateurs SureSelect XT HS permettant le traitement en parallèle de l'ADN et de l'ARN d'un même échantillon
4. Fournir un code-barre moléculaire (MBC) unique afin d'obtenir une déduplication des réplicats de PCR et de fragmentation de meilleure qualité

Ces modifications ont réduit le délai de 2-3 jours à 1 journée. En outre, en analysant les données de fusion de plusieurs types d'échantillons de départ (intacts, frais congelés et FFPE), nous avons constaté que ce workflow simplifié produit des banques de séquençage d'ARN d'une plus grande complexité et améliore les performances de séquençage avec des quantités d'ARN de départ pouvant descendre jusqu'à 10 ng.

Introduction

L'application du séquençage haut débit à la transcriptomique (séquençage d'ARN ou RNA-seq) n'a pas seulement permis le profil d'expression génique globale mais a également permis d'obtenir des informations plus précises sur les variants d'épissage, les transcrits de fusion, les évènements de modifications post-traductionnelleset l'expression allélique spécifique. L'approche standard pour générer des banques RNA-seq consiste en une fragmentation de l'ARN de départ par transestérification Mg^{2+}-dépendante, la synthèse

aléatoire du premier brin d'ADNc avec une amorce hexamérique, suivie de la synthèse du second brin d'ADNc avec marquage spécifique du brin par dUTP. Afin d'éviter toute concaténation, les extrémités de l'ADNc sont émoussées et une adénine est ajoutée aux extrémités 3' par polymérisation d'ADN sans matrice. Cette étape est suivie d'une ligation d'adaptateurs de séquençage spécifiques à la plateforme. Les étapes post-ligation sont les mêmes que celles du workflow de séquençage de l'ADN.
En dépit du fait que les banques de séquençage dérivées d'ADN ou d'ARN semblent être identiques après l'étape de ligation d'adapteurs, il existe des différences spécifiques lorsque l'on compare le séquençage de l'ARN au séquençage de l'ADN. La première différence est la nécessité de garder l'orientation des fragments de séquençage afin de savoir à quel brin d'ADN génomique correspond l'ARN d'origine. Ceci est fait généralement en introduisant de l'uracile dans la réaction de synthèse du deuxième brin, en continuant la préparation de la banque, puis lors d'une étape ultérieure en utilisant une enzyme UDG pour éliminer le deuxième brin. La deuxième différence est la grande gamme dynamique des transcrits mesurés dans le séquençage d'ARN. Ceux-ci peuvent dépasser cinq ordres de grandeur en raison des grandes différences d'abondance relative des transcrits codants exprimés et de l'extrême abondance d'ARN ribosomique (ARNr) par rapport à l'ARN codant. Le séquençage RNA-seq nécessite par conséquent de réduire la complexité pour éviter de gâcher des lectures sur des transcrits n'offrant pas d'informations (tels que l'ARNr). La méthode la plus courante consiste à éliminer l'ARNr soit par appauvrissement ciblé (ribo-déplétion), soit par capture spécifique de l'ARN polyadénylé (ARNm). Dans un échantillon type, même après élimination de l'ARNr, le 1 % de gènes les plus exprimés correspondra à environ 50 % de tous les transcrits. Par conséquent, si le but est d'étudier des gènes moyennement ou peu exprimés, leur détection sera plus efficace si la complexité est réduite encore plus.

Comme alternative aux approches de déplétion ciblée, la réduction de la complexité peut aussi être obtenue par enrichissement ciblé en faisant appel à des sondes biotinylées (également appelées « amorces »), une approche couramment utilisée en séquençage génomique. Pour le séquençage de l'ARN, l'enrichissement ciblé par la sélection des amorces est surtout utilisé avec des échantillons issus de FFPE, où la déplétion de l'ARNr est connue pour être inégale et où l'enrichissement en poly(A) ne peut pas être utilisé en raison de la fragmentation du matériel source[1]. L'enrichissement ciblé est aussi utile lorsque seul un sous-ensemble relativement restreint du transcriptome doit être examiné. Un exemple bien connu est la détection de transcrits de fusion révélateurs d'un évènement sous-jacent de fusion de gènes. Cependant, tout scénario où seul un sous-ensemble de transcrits fournit des informations, tel que le profilage transcriptionnel ou la détection de modifications rares post-transcriptionnelles peut tirer parti d'un enrichissement ciblé [2].

Agilent propose le kit de préparation de banques SureSelect XT RNA Direct (réf. G7564A, G7564B) qui permet la construction de banques de séquençage d'ARN ciblé. Nous avons démontré l'utilisation efficace de ce kit sur des échantillons issus de FFPE[3]. Nous décrivons ici un workflow simplifié et amélioré grâce à la combinaison de composants du kit de préparation de banques SureSelect XT RNA Direct et du kit d'enrichissement ciblé SureSelect XT HS (réf. G9706A) conjointement à un protocole modifié. Le workflow a été simplifié en éliminant la lyophilisation de l'ARN de départ, en supprimant une étape de purification sur billes SPRI, et en remplaçant l'étape d'hybridation de l'amorce pendant 24 heures par une étape d'hybridation rapide de 90 minutes. Nous avons également éliminé l'étape de traitement par UDG, en obtenant la spécificité de brin grâce à une enzyme de PCR qui discrimine les matrices ADN contenant de l'uracile pendant la PCR de précapture. En outre, l'utilisation d'adaptateurs de séquençage SureSelect XT HS ajoute un MBC pour l'identification de duplicats de fragmentation et diminue les obstacles pratiques au traitement en tubes séparés (également appelés parallèles) de l'ADN et de l'ARN provenant d'un même échantillon.

Données expérimentales

Sources d'ARN

L'ARN de référence humain universel (UHRR) a été obtenu sous forme de matériel fraîchement congelé d'Agilent Technologies (Santa Clara, CA, USA ; réf. 750500-41). Des tissus appariés de tumeur du sein et de tissu adjacent normal ont été obtenus sous forme fraîchement congelée et FFPE auprès de la SRC CureLine spécialisées dans les biospécimens humains (Brisbane, CA, USA, réf. personnalisée). L'ARN FFPE de fusion de référence de tissu tumoral v2 Seraseq a été obtenu auprès de SeraCare (Gaithersburg, MD, USA, réf. 0710-0129)

Isolation de l'ARN

Selon les besoins, l'ARN a été isolé à l'aide du kit RNeasy FFPE ou du mini kit RNeasy de Qiagen selon les instructions du fabricant (Qiagen USA, Germantown, MD, USA, réf. 73504 et 74104, respectivement). Pour un protocole plus détaillé, voir l'annexe.

Évaluation de la qualité du matériel de départ et des banques de séquençage

Les échantillons d'acides nucléiques ont été évalués sur le bioanalyseur Agilent 2100 (Agilent Technologies, réf. G2939B) soit avec le kit RNA 6000 pico Agilent (Agilent Technologies, réf. 5067-1513) pour évaluer la qualité de l'ARN, soit avec le kit ADN Agilent 1000 (Agilent Technologies, réf. 5067-1504), pour évaluer la qualité des banques de séquençage.

Autres matériels

L'actinomycine D a été obtenue auprès de Sigma (St. Louis, MO, USA, réf. A1410) et amenée à une concentration de 4 µg/µL dans une solution mère de DMSO. Les purifications sur billes SPRI ont été effectuées avec des billes AMPure XP (Beckman Coulter, Atlanta, GA, USA, réf. A63880). La capture des sondes biotinylées a été réalisée avec les billesDynabeads MyOne de streptavidine T1 (Thermo Fisher Scientific, Waltham, MA, USA, réf. 65601).

Préparation de banques d'ARN SureSelect XT HS RNA

La construction de banques de séquençage d'ARN a été réalisée à l'aide du kit « SureSelect XT RNA Direct » (Agilent Technologies, réf. G7564A) et du système d'enrichissement ciblé « SureSelect XT HS for the Illumina paired-end multiplexed sequencing library » (Agilent Technologies, réf. G9706A). Pour une description plus détaillée, voir l'annexe.

Préparation de banques d'ARN SureSelect XT RNA Direct

Les banques RNA Direct ont été générées, enrichies et séquencées selon le protocole de préparation de banques d'ARN SureSelect XT RNA Direct (voir le manuel).

Enrichissement ciblé

L'enrichissement ciblé des banques d'ARN SureSelect XT HS RNA a été réalisé à l'aide de l'exome humain complet V7 SureSelect (Agilent Technologies, réf. 5191-4029) ciblant le transcriptome codant.
Un protocole détaillé de capture d'amorce est fourni en annexe.

Séquençage et traitement des données

Les banques de séquençage ont été analysées sur le système Illumina HiSeq 4000 par séquençage « paired-end » avec un format de lecture 2 × 150. Pour l'analyse de l'expression (données non fournies), les fichiers FASTQ ont été alignés sur le transcriptome à l'aide du package STAR version 2.6.0a reconnaissant les épissages à l'aide de la version du génome hg38 comme référence. Les profils d'expression ont été générés à partir des résultats d'alignement STAR à l'aide de l'outil RSEM. Les statistiques générales de la banque (spécificité de brin, biais d'extrémités 5'-3', taux de duplication MBC-aveugle, estimations de la taille de la banque) ont été générées par le pipeline d'analyse d'ARN Picard en marquant les duplicats via l'utilisation des fichiers .bam sous-échantillonnés à 2×10^7 lectures pour obtenir des taux de duplication normalisés. Les statistiques de duplication corrigées par MBC et les estimations de la taille de la banque ont été générées avec le même pipeline mais en utilisant UmiAwareMarkDuplicatesWithMateCigar pour éliminer les duplicats de fragmentation. Les transcrits de fusion ont été comptés avec STAR-Fusion et visualisés avec l'outil FusionInspector, qui fait partie du kit d'outils d'analyse du transcriptome des cancers Trinity (CTAT)[4].

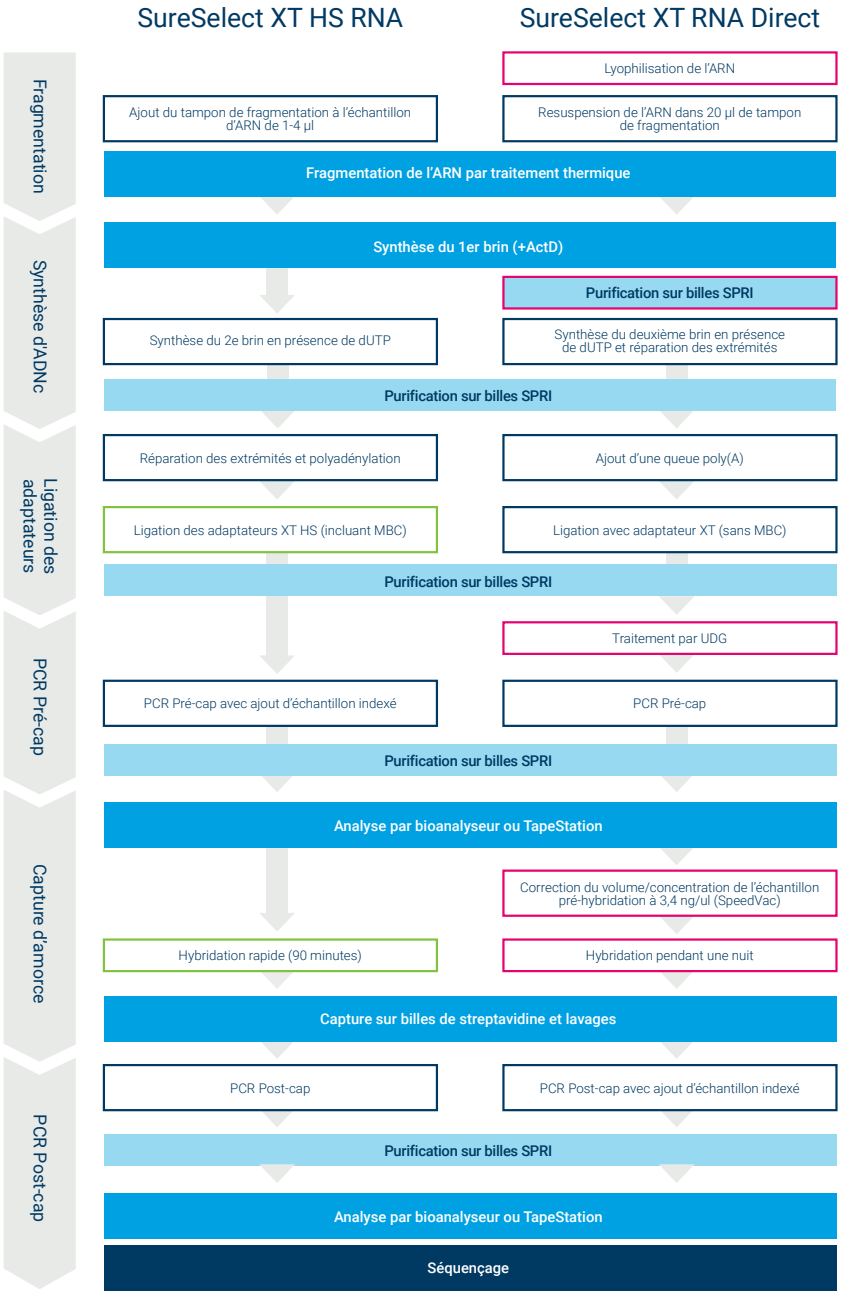


Figure 1. Comparaison du workflow Agilent SureSelect XT HS RNA (RNA XT HS) et du kit de préparation de banques Agilent SureSelect XT RNA Direct (RNA Direct). Les étapes éliminées du workflow RNA Direct sont indiquées par une bordure rose. Les étapes offrant une meilleure fonctionnalité du protocole RNA XT HS sont indiquées par une bordure verte.

Résultats et discussion

Vue générale des workflow RNA XT HS et RNA Direct

Nous souhaitions développer un workflow simplifié et amélioré pour générer des banques de séquençage d'ARN de grande qualité en combinant les composants du kit de préparation de banques d'ARN SureSelect XT RNA Direct (RNA Direct) et du kit d'enrichissement ciblé SureSelect XT HS. Ce workflow a été nommé SureSelect XT HS RNA (RNA XT HS). La Figure 1 compare ces workflow de manière succincte. Le workflow RNA XT HS offre des avantages évidents par rapport au workflow RNA Direct, grâce à l'élimination du traitement par UDG et au remplacement de l'hybridation traditionnelle en une nuit par une hybridation rapide de 90 minutes, ce qui réduit énormément le délai. Nous souhaitions également savoir s'il était possible d'éliminer l'étape de purification SPRI entre la synthèse du premier et du deuxième brin et éviter la lyophilisation de l'ARN de départ qui fait partie du workflow RNA XT HS (voir les détails en annexe).

Comparaison des performances des workflow

Afin de comparer les performances des deux processus de préparation de banques, nous avons tout d'abord généré des banques avec différentes quantités d'ARN de référence humain universel (UHRR) et de l'ARN de fusion de référence de tissu tumoral v2 Seraseq FFPE (SeraCare) comme matériel de départ représentant un échantillon intact et un échantillon FFPE idéalisé, respectivement. Les banques RNA XT HS et RNA Direct ont été enrichies avec l'exome humain complet V7 SureSelect. Enfin, ces banques enrichies ont été séquencées sur le séquenceur Illumina et les données analysées avec des pipelines personnalisés de traitement des données (voir les détails dans la section expérimentale).

Le Tableau 1 résume les statistiques globales de séquençage dérivées de l'analyse des données de séquençage RNA XT et RNA Direct. Nous avons trouvé que les banques séquencées RNA XT HS et RNA Direct ne peuvent pas être différenciées en ce qui concerne plusieurs mesures, en particulier la haute spécificité de brin (>98 %) et la faible contamination en ARNr (~0,1 %). La haute spécificité de brin des banques RNA XT HS démontre l'efficacité de l'approche utilisant le workflow RNA XT HS qui élimine le traitement par UDG et utilise plutôt une enzyme de PCR qui n'amplifie pas les matrices contenant de l'uracile. La spécificité de brin a tendance à être un peu supérieure pour les échantillons intacts que pour les échantillons FFPE, ce qui n'est pas étonnant compte tenu de la moindre qualité du matériel FFPE de départ. Néanmoins, les spécificités observées avec des tissus FFPE de départ restent très élevées (> 98 % pour le matériel FFPE) avec l'un ou l'autre protocole. Le pourcentage de contamination d'ARNr toujours faible pour les deux méthodes de préparation de banques a déjà été démontré avec des approches d'enrichissement ciblé[3].

Lorsque l'on compare les taux de cartographie des exons des banques RNA XT HS et RNA Direct, nous observons encore une fois des performances comparables (Figure 2). Une étape essentielle permettant de raccourcir le nouveau workflow RNA XT HS est le remplacement de l'hybridation traditionnelle durant la nuit par une étape d'hybridation rapide. Comme le montre la Figure 2, la capture accélérée n'affecte pas les taux de cartographie et nous avons toujours observé des taux exoniques de 90 % avec très peu de lectures cartographiées dans les régions intergéniques, quel que soit le matériel de

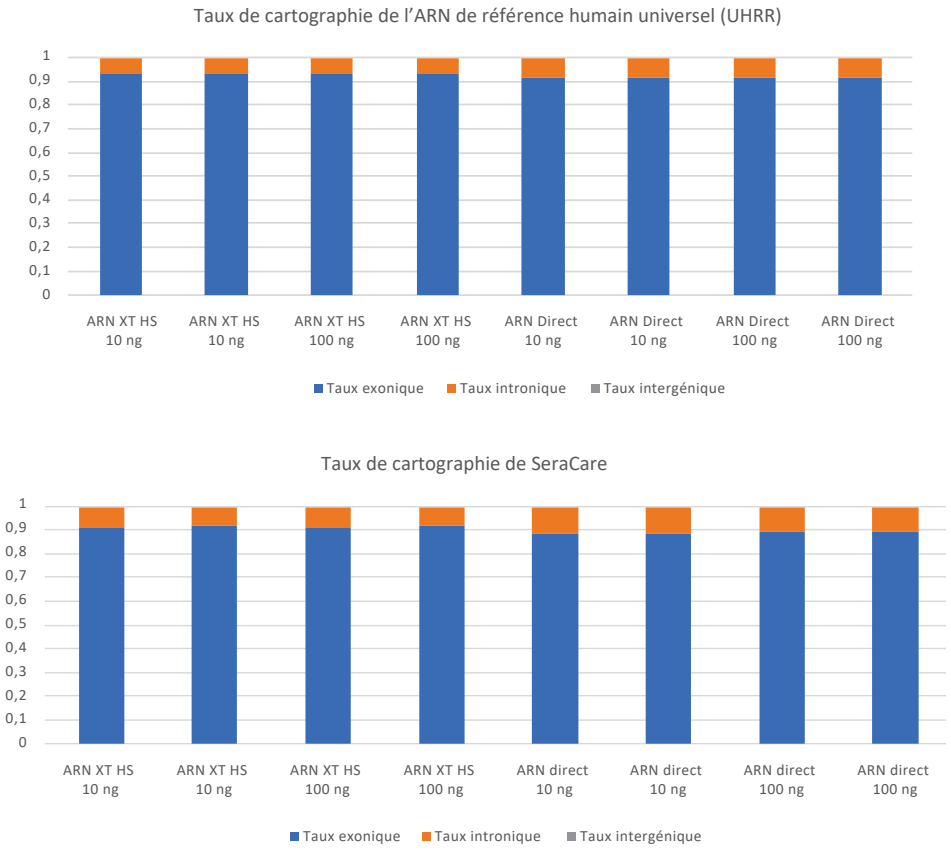


Figure 2. Comparaison du taux de cartographie des banques séquencées RNA XT HS et RNA Direct. Les banques RNA-seq ont été préparées à partir d'échantillons UHRR (A) et SeraCare (B) en utilisant le workflow RNA XT HS ou RNA Direct. Les taux de cartographie des banques de séquençage enrichies avec l'exome humain complet V7 SureSelect sont indiqués pour les séquences exoniques, introniques et intergéniques. On suppose que le taux intronique reflète le séquençage d'ARNm non traité.

départ utilisé. Les transcrits ciblés par la banque de capture V7 correspondaient à 92,9-94,1 % de toute l'expression observée quel que soit la quantité ou le type d'échantillon de départ. Les ~7 % restants sont principalement dus à la contamination de transcrits hautement exprimés tels que les gènes de mitochondries, ou à des erreurs d'annotation.

Un indicateur essentiel d'efficacité du procédé global de préparation des banques est l'estimation de la complexité initiale de la banque. Il est préférable d'avoir des banques plus grandes et plus complexes puisque leur séquençage offre plus d'informations et que la quantification est plus fiable. La complexité d'une banque est calculée en fonction de la profondeur de séquençage (en paires de lecture) et du nombre de variants uniques observés, sur la base du taux de duplication. Nous avons utilisé le pipeline Picard, qui suppose que tous les duplicats sont des duplicats de PCR, pour calculer la complexité estimée des banques. Cette approche sous-estime la véritable complexité d'une banque et il en sera question de manière plus détaillée dans la section des résultats. Les complexités de banques projetées qui en résultent sont indiquées dans le Tableau 1 et illustrées dans la Figure 3. Lorsque l'on compare les workflow RNA XT HS et RNA Direct, nous n'observons aucune différence significative entre les banques construites à partir de grandes quantités (100 ng) d'ARN de départ provenant de tissu frais ou FFPE. Cependant, lorsque la quantité d'ARN UHRR ou SeraCare de départ est réduite à 10 ng, nous observons des différences de complexité de banque notables. Tout d'abord, ces banques avec une quantité plus faible de départ sont plus petites et moins complexes que les banques avec une quantité plus importante de départ. Ceci n'est pas surprenant car plus la quantité d'ARN de départ à convertir en banque de séquençage est faible, plus la banque finale sera petite. Nous observons également que lorsque la quantité de départ est faible, le workflow RNA XT HS simplifié produit des banques qui sont 1,5 à 2 fois plus efficaces que le workflow RNA Direct. Plusieurs facteurs du workflow RNA XT HS sont responsables de ce gain d'efficacité et sont en cours d'investigation (données non fournies).

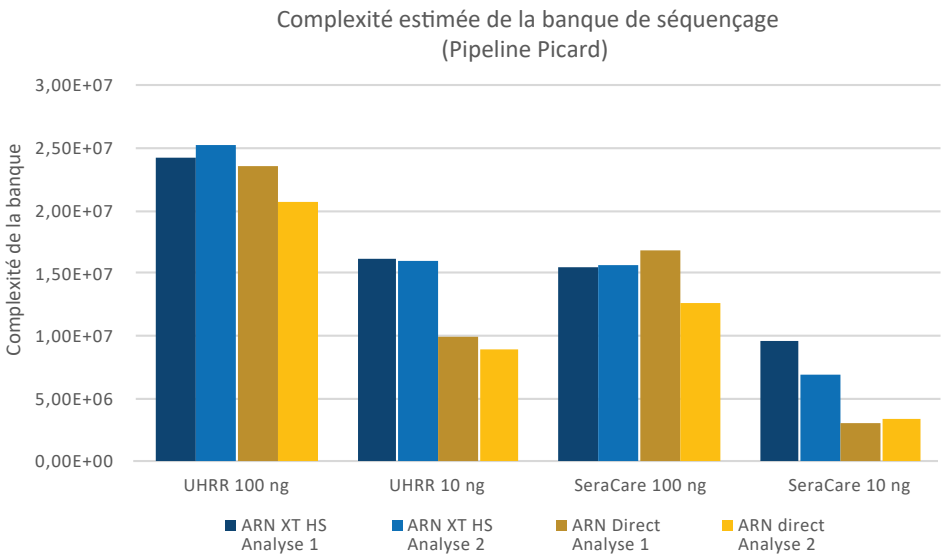


Figure 3. Différence de complexité des banques RNA XT HS et RNA Direct pour une faible quantité d'échantillon de départ. Les banques de séquençage préparées en utilisant les quantités de départ et les sources indiquées ont été séquencées sur une plateforme Illumina HiSeq 4000 par séquençage paired-end. La complexité estimée des banques a été estimée en fonction du taux de duplication observé et du nombre respectif de paires de lecture établi par l'outil d'analyse d'ARN Picard. Ce pipeline ne fait pas la différence entre les duplicats de PCR et de fragmentation.

Tableau 1. Statistiques globales de séquençage des banques SureSelect XT HS RNA et SureSelect XT RNA Direct : Les banques de séquençage d'ARN générées à partir du même matériel de départ en utilisant le protocole RNA XT HS ou RNA Direct ont été séquencées sur la plateforme Illumina HiSeq 4000. Après un sous-échantillonnage à 2×10^7 lectures, les statistiques de banque ont été générées avec l'outil d'analyse d'ARN Picard. Les statistiques avec correction MBC ont été générées avec le même pipeline après étiquetage des MBC avec UmiAwareMarkDuplicatesWithMateCigar. (A) ARN de départ de tissu fraîchement congelé (« intact ») ; (B) ARN de départ de tissu FFPE.

A. ARN de référence humain universel (UHRR)

Protocole	SureSelect XT HS RNA				SureSelect XT RNA Direct			
Échantillon de départ (ng)	100	100	10	10	100	100	10	10
Lectures analysées (millions)	1,9	1,9	1,9	1,9	1,9	1,9	1,9	1,9
Taux d'ARNr (%)	0,1 %	0,1 %	0,1 %	0,1 %	0,1 %	0,1 %	0,1 %	0,1 %
Spécificité de brin (%)	99,1 %	99,1 %	99,0 %	99,0 %	99,1 %	99,2 %	99,1 %	99,1 %
Taux de duplication (%)	17,3 %	16,7 %	24,3 %	24,7 %	17,9 %	20,0 %	35,8 %	38,7 %
Estimation de la taille de la banque (x 10^6)	24,2	25,2	16,2	15,9	27,5	23,6	10,3	9,2
Taux de duplication avec correction MBC (%)	7,1 %	6,4 %	17,1 %	17,8 %	S.o.	S.o.	S.o.	S.o.
Estimation de la taille de la banque avec correction MBC (x 10^6)	65,9	72,8	24,6	23,6	S.o.	S.o.	S.o.	S.o.

B. ARN de fusion de référence de tissu tumoral v2 Seraseq FFPE (SeraCare)

Protocole	SureSelect XT HS RNA				SureSelect XT RNA Direct			
Échantillon de départ (ng)	100	100	10	10	100	100	10	10
Lectures analysées (millions)	18,6	18,6	18,7	18,6	19	18,8	18,9	18,9
Taux d'ARNr (%)	0,1 %	0,2 %	0,1 %	0,1 %	0,1 %	0,2 %	0,1 %	0,1 %
Spécificité de brin (%)	98,6 %	98,6 %	98,6 %	98,7 %	98,9 %	98,9 %	98,9 %	98,9 %
Taux de duplication (%)	24,8 %	24,7 %	36,0 %	45,0 %	23,5 %	29,4 %	69,1 %	67,2 %
Estimation de la taille de la banque (x 10^6)	15,5	15,5	9,6	6,9	18,6	13,3	3	3,3
Taux de duplication avec correction MBC (%)	13,2 %	13,7 %	29,3 %	38,8 %	S.o.	S.o.	S.o.	S.o.
Estimation de la taille de la banque avec correction MBC (x 10^6)	32,5	31,2	12,7	8,66	S.o.	S.o.	S.o.	S.o.

Impact de codes-barres moléculaires uniques (MBC) sur les banques de séquençage de l'ARN

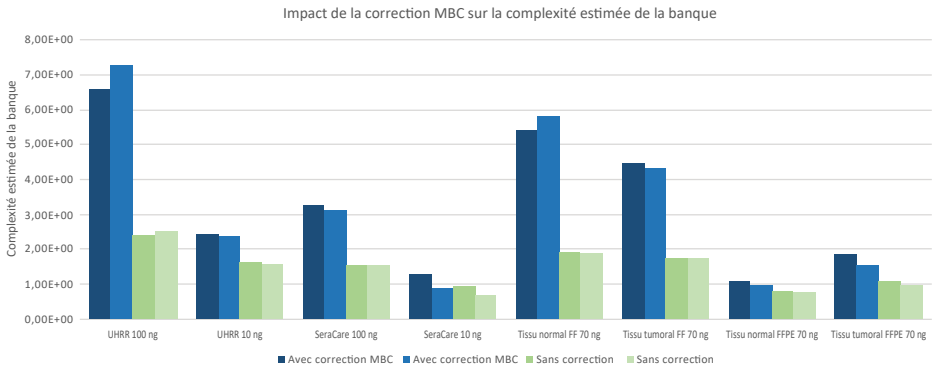
Comme il a été mentionné plus haut, la mesure exacte de la complexité des banques est essentielle pour évaluer l'efficacité d'un protocole de préparation de banque. Les estimations de complexité de banque ci-dessus dépendent de l'hypothèse que les paires de lecture ayant un même début et une même fin sont des duplicats de PCR dérivés de la même molécule de la banque d'origine. Cependant, des duplicats peuvent également provenir de la fragmentation aléatoire, donnant deux fragments indépendants avec les mêmes extrémités. Puisque les duplicats de fragmentation sont véritablement des membres indépendants de la banque de séquençage, la complexité d'une banque ne doit être déterminée qu'en fonction des vrais duplicats de PCR.

Comparé au séquençage de l'ADN, le séquençage de l'ARN peut donner lieu à un nombre plus élevé de duplicats de fragmentation du fait que l'expression élevée de certains gènes augmente la probabilité de générer des duplicats de fragmentation de manière aléatoire. Ceci induit plus d'erreurs d'estimations de complexité des banques. Le workflow RNA XT HS présente l'avantage potentiel de construire des banques avec des adaptateurs XT HS contenant des MBC uniques de 10 pb. Nous avons émis l'hypothèse que cet MBC pourrait être utilisé pour distinguer les duplicats de PCR et de fragmentation. Nous avons ré-analysé les données de séquençage RNA XT HS ci-dessus avec un pipeline de traitement des données modifié qui utilise les données de MBC (note : les banques RNA Direct ne comportent pas de MBC et n'ont donc pas été incluses dans cette nouvelle analyse). Les résultats de cette analyse sont rapportés dans le Tableau 1 et la Figure 4.

Comme le montre la Figure 4, la correction pour les duplicats de fragmentation donne des estimations de complexité de banque plus élevées, notamment pour les quantités plus importantes de départ où la complexité d'une banque est sous-estimée d'un facteur de 3 environ si l'on ne tient pas compte des duplicats de fragmentation. On attend un biais plus élevé observé pour les banques plus grandes et les quantités de départ plus importantes correspondantes, puisqu'il y a plus de chances d'observer des duplicats de fragmentation si plus de molécules dérivées d'une même séquence de codage sont traitées dans des banques.

Nous voulions ensuite étudier l'impact de la correction MBC pour un ensemble d'échantillons en situation réelle. Nous avons donc généré des banques RNA XT HS en utilisant un ensemble apparié de tissu de tumeur mammaire et de tissu normal adjacent tous deux conservés sous forme d'échantillons fraîchement congelés et FFPE. Ces banques ont été enrichies, séquencées et analysées comme ci-dessus avec les banques RNA XT HS UHRR et SeraCare. Les statistiques globales de ces banques sont données dans le Tableau 2 et la Figure 4. Nous observons une excellente spécificité de brin et un taux d'ARNr dans les données de séquençage du tissu tumoral/normal fraîchement congelé et FFPE qui est comparable à nos banques UHRR et SeraCare. Nous observons également que les banques obtenues à partir de tissu FFPE sont significativement

A



B

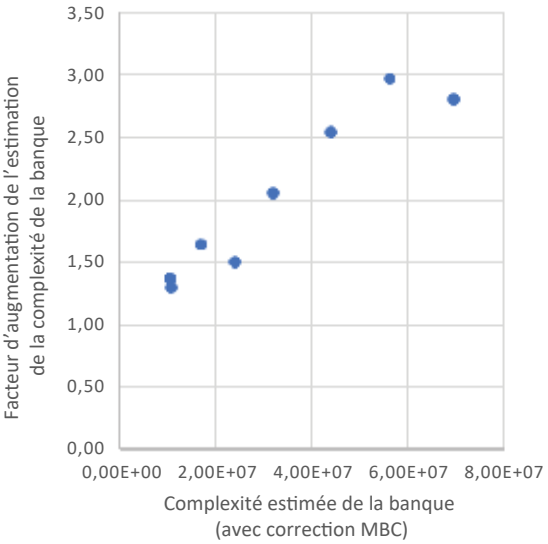


Figure 4. Impact de la correction par des codes-barres moléculaires uniques (MBC) sur l'estimation de la taille des banques. La taille des banques RNA XT HS a été calculée soit par déduplication Picard (sans correction), soit après utilisation de MBC (avec correction MBC). La déduplication Picard suppose que tous les duplicats sont des duplicats de PCR alors que la correction MBC fait la distinction entre les duplicats de PCR et de fragmentation. **A)** Complexité de banque estimée pour plusieurs banques de séquençage RNA XT HS obtenues à partir de toute une gamme d'échantillons et de quantités de départ. **B)** Les différences de complexité des banques corrigées et non corrigées par MBC sont plus prononcées pour les banques plus grandes et donc pour les quantités de départ plus importantes correspondantes.

plus petites que celles de tissus fraîchement congelés, un phénomène souvent rapporté. Lorsque nous avons analysé nos données de séquençage à partir de tissu FFPE avec correction MBC, nous avons noté une amélioration d'un facteur de 0,3-3,5 de l'estimation de la complexité et avons donc récupéré des données qui auraient été perdues si nous avions supposé que tous les duplicats étaient des duplicats de PCR plutôt que des duplicats de PCR et de fragmentation. Cela montre que pour optimiser les informations de séquençage obtenues à partir d'échantillons d'ARN FFPE, il est essentiel que les duplicats de PCR et de fragmentation soient correctement identifiés dans ces échantillons.

Détection de fusions de gènes

La RNA-seq ciblée a montré son utilité pour détecter des évènements de fusion de gènes, notamment avec les échantillons les plus récalcitrants (FFPE). Nous voulions savoir si notre approche RNA XT HS pourrait être utilisée dans ce but dans des conditions d'enrichissement avec des sondes d'exome (par rapport à un panel spécifique de la fusion) et en maximisant la longueur des séquences. Comme mentionné plus haut, nos banques RNA XT HS de tous les types d'échantillons (UHRR, SeraCare, fraîchement congelés et FFPE) ont subi un enrichissement ciblé avec l'exome humain complet V7 SureSelect et les banques enrichies ont été séquencées en tant que banques à extrémités appariées avec une longueur de lecture de 2 × 150. Les données de séquençage de tumeur/tissu normal fraîchement congelé(e) et FFPE ont été analysées avec STAR-Fusion et visualisées sur FusionInspector.

FusionInspector a détecté une gamme de fusions supposées, y compris des faux positifs tels que des évènements de recombinaison de type VDJ, indiquant la présence de cellules immunitaires dans certains échantillons. Les recombinants VDJ ont été éliminés du Tableau 3. Le traitement des données montre 15 éventuelles fusions de gènes présentes dans nos échantillons, avec 12 des 15 partenaires de fusion signalés comme étant impliqués dans des fusions de gènes liées au cancer, mais pas dans le

Tableau 2. Statistiques globales de séquençage de banques Agilent SureSelect RNA XT HS obtenues à partir de tissus fraîchement congelés et FFPE. Banques de séquençage de l'ARN générées à partir d'un ensemble de tissu tumoral et de tissu adjacent normal sous forme d'échantillons fraîchement congelés et FFPE en utilisant le protocole RNA XT HS. Les banques ont été séquencées sur la plateforme Illumina HiSeq 4000. Après un sous-échantillonnage à 2×10^7 lectures, les statistiques de banque ont été générées avec l'outil d'analyse d'ARN Picard avec ou sans étiquetage du MBC.

Type d'échantillon	**Fraîchement congelé**				**FFPE**			
Source	Normale	Normale	Tumorale	Tumorale	Normale	Normale	Tumorale	Tumorale
Échantillon de départ (ng)	70	70	70	70	70	70	70	70
Lectures analysées (millions)	18,8	18,8	18,7	18,8	18,6	18,6	18,7	18,7
Taux d'ARNr (%)	0,2 %	0,2 %	0,2 %	0,2 %	0,2 %	0,2 %	0,1 %	0,1 %
Spécificité de brin (%)	99,0 %	99,0 %	99,1 %	99,2 %	98,4 %	98,4 %	98,8 %	98,8 %
Taux de duplication (%)	21,0 %	21,0 %	23,0 %	23,0 %	41,0 %	43,0 %	33,0 %	35,0 %
Estimation de la taille de la banque (x 10^6)	19	18,8	17,3	17,3	7,9	7,4	10,9	9,8
Taux de duplication avec correction MBC (%)	8,0 %	8,0 %	10,0 %	10,0 %	33,0 %	35,0 %	22,0 %	25,0 %
Estimation de la taille de la banque avec correction MBC (x 10^6)	53,9	58,2	44,7	43,1	11	9,93	18,6	15,4

Tableau 3. Détection de fusions de gènes dans des échantillons fraîchement congelés et FFPE de tissu normal/tumoral. Les banques de séquençage RNA XT HS ont été séquencées puis analysées pour détecter des fusions de gènes avec le pipeline STAR-Fusion. Les fusions supposées détectées dans un ou plusieurs échantillons sont énumérées dans chaque rangée avec les preuves de détection de la fusion. Les lectures de jonction et les lectures de confirmation sont exprimées en lectures par million de paires de lectures uniques de la banque. Chaque colonne représente la moyenne de deux réplicats techniques. La fusion surlignée en gris est identifiée dans la base de données TGCA comme étant une fusion associée au cancer du sein et pourrait être une mutation donnant un avantage sélectif à ce type de tumeur. Les évènements de recombinaison VDJ observés, indiquant la présence de cellules immunitaires dans les échantillons, ont été éliminés de l'analyse. La détection est basée sur l'analyse de $0{,}7–1{,}4 \times 10^7$ paires de lecture uniques pour chaque échantillon.

	Fraîchement congelé				**FFPE**			
	Normale		Tumorale		Normale		Tumorale	
Fusion supposée	Jonction	Assistance	Jonction	Assistance	Jonction	Assistance	Jonction	Assistance
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--RCSD1	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

cas de la combinaison observée. L'une des fusions, FCHSD2-FAM168A (surlignée en gris dans le Tableau 3), a précédemment été identifiée comme étant associée au cancer du sein et pourrait être une mutation donnant un avantage sélectif à ce type de tumeur[5]. Pour toutes les autres fusions spécifiques aux tumeurs, au moins un partenaire de fusion (habituellement deux) est mentionné dans la base de données TCGA et est associé à la formation de tumeurs. Nous trouvons le transcrit de fusion A:B. A et B sont présents dans la base de données des tumeurs. Cependant, A:B n'y est pas répertorié. Mais il y a des exemples de transcrits de fusion A:C et D:B. Ces données suggèrent que les banques RNA XT HS enrichies avec un exome permettent aux chercheurs d'identifier des fusions au niveau du transcriptome entier, même avec les échantillons les plus récalcitrants.

Conclusion

Afin de permettre la construction plus efficace de banques RNA-seq ciblées, nous avons cherché à améliorer le workflow actuel SureSelect XT RNA Direct en utilisant les composantes du kit de préparation de banque SureSelect XT RNA Direct et du kit d'enrichissement ciblé SureSelect XT HS. Le nouveau workflow SureSelect XT HS RNA comporte plusieurs améliorations significatives :

- La lyophilisation de l'ARN de départ a été remplacée par l'ajout direct du tampon de fragmentation à l'échantillon d'ARN.
- L'étape de purification sur billes SPRI après la synthèse du premier brin d'ADNc a été éliminée.
- Une enzyme de PCR alternative a été utilisée pour ne pas amplifier les matrices contenant de l'uracile, éliminant ainsi le besoin d'un traitement par UDG pour garder la spécificité de brin.
- La procédure de capture de la sonde (« amorce ») est passée d'un protocole d'hybridation de 24 heures à une hybridation rapide de 90 minutes.
- Les adaptateurs de séquençage XT ont été remplacés par les adaptateurs XT HS permettant l'utilisation d'un MBC et donc une estimation plus précise de la complexité de la banque et la « récupération » de lectures de séquençage.

Nous avons observé que le workflow RNA XT HS réduit le délai de construction de la banque et d'enrichissement ciblé de 2–3 jours à 1–2 jours. Le nouveau workflow simplifié permet d'obtenir des banques RNA-seq que l'on ne peut pas distinguer des banques générées par le protocole SureSelect XT RNA Direct en termes de spécificité de brin, de taux d'ARNr et de taux de cartographie pour des quantités de départ importantes. Pour les quantités de départ moindres (10 ng), nous avons trouvé que le workflow plus court n'affectait pas les performances globales et semblait même plus efficace. Nous avons également observé que l'inclusion de MBC dans la conception améliorée des adaptateurs (XT HS) permet dorénavant l'identification de duplicats de fragmentation. Cette identification renforcée de duplicats de fragmentation améliore les résultats de séquençage en évitant la perte de lectures rencontrée avec les méthodes de duplication standard de type start-stop.

Lorsque nous avons testé des échantillons en situation réelle, nous avons observé que le workflow RNA XT HS continuait à produire des données de grande qualité même avec les échantillons FFPE. L'analyse préliminaire de fusion de gènes a détecté des fusions de gènes potentielles dans nos échantillons de tumeur fraîchement congelés et FFPE, indiquant la possibilité d'une utilisation du protocole RNA XT HS en situation réelle pour le séquençage d'ARN provenant d'échantillons FFPE. Bien que nous n'ayons pas présenté de données d'expression génique ou d'épissage, nos analyses préliminaires indiquent que la méthode RNA-seq ciblée peut être utilisée pour des analyses d'expression des gènes globale, la détection de variants d'épissage, la détection de l'expression de variants et l'analyse de l'expression allèle-spécifique. Enfin, le développement du workflow SureSelect RNA XT HS est basé sur le kit d'enrichissement ciblé SureSelect RNA XT HS, ce qui facilite le séquençage de l'ADN et de l'ARN d'un même échantillon en parallèle. Ceci pourrait être utile pour plusieurs applications, y compris l'analyse multiomique en général et le traitement en parallèle d'échantillons pour l'analyse TMB-MSI et la détection de fusions.

Abréviations

FFPE, formalin-fixed, paraffin embedded (fixé au formol et inclus dans la paraffine) ; TPM, transcrits par million de kilobases ; nt, nucléotide ; UHRR, universal human reference RNA (ARN de référence humain universel) ; MBC, molecular barcode (code-barre moléculaire) ; UDG, uracile déglycosylase ;

Références

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. Note d'application Agilent Technologies, publication numéro 5991-8119EN, **2017**.

4) Haas, B., *et al.* STAR-Fusion: Détection précise des transcripts de fusion provenant de bioRxiv du RNA-seq Fast and Accurate Fusion Transcript Detection from RNA-Seq. bioRxiv. 120295 (**2017**).

5) Hu, X., *et al.* Tumor Fusions: An Integrative Resource for Cancer-Associated Transcript Fusions. *Nucleic Acids Res.* **2018**, 4, 46(D1), D1144-D1149.

Annexe

1. Protocole détaillé pour la construction de banques de séquençage de l'ARN

Préparation des échantillons d'ARN

L'ARN total de copeaux de FFPE a été isolé avec le kit Qiagen RNeasy FFPE selon les instructions du fabricant. L'ARN de tissu congelé a été isolé avec le mini kit Qiagen RNeasy. Tous les échantillons d'ARN total ont été analysés sur un bioanalyseur Agilent 2100 avec le kit RNA 6000 pico Agilent. Les échantillons ont été chauffés jusqu'à 80 °C pendant 2 minutes avant de les charger sur la puce. Les valeurs RIN et DV200 ont été calculées avec le logiciel du bioanalyseur. Ces mesures de la qualité de tous les échantillons d'ARN total testés sont enregistrées dans le Tableau S1.

Tableau S1. Mesures de qualité des échantillons et conditions de fragmentation

Description de l'échantillon	RIN	DV200	Fragmentation
ARN de référence humain universel (UHRR)	9,2	94 %	94 °C, 8 min
ARN de fusion de référence de tissu tumoral v2 Seraseq FFPE	2	54 %	94 °C, 3 min, 65 °C, 2 min
Tissu mammaire normal congelé	6,3	94 %	94 °C, 8 min
Tissu mammaire tumoral congelé	5	88 %	94 °C, 8 min
Tissu mammaire normal FFPE	2,1	48 %	65 °C, 5 min
Tissu mammaire tumoral FFPE	2	47 %	65 °C, 5 min

Tableau S2. Mélange de réaction premier brin.

Réactif	Volume pour une réaction	Volume pour huit réactions + excès de 10 %
Master mix RNA-seq premier brin	8 µL	70,4 µl
Actinomycine D (120 ng/µL)	0,5 µL	4,4 µl
Total	**8,5 µL**	**74,8 µl**

Tableau S3. Paramètres du protocole de lavage SPRI.

Volume de billes AMPure	105 µL (volume 1,8X)
Durée d'incubation des billes	5 minutes
Lavage à l'éthanol 70 % (deux fois)	200 µl
Temps de séchage à 37 °C	1–2 min maximum
Volume d'élution	50 µl d'eau exempte de nucléase

Préparation de l'ADNc avec les réactifs du kit Agilent SureSelect XT RNA Direct

Remarque : Une solution mère de 4 µg/µL d'actinomycine D dans du DMSO a été préparée d'avance et conservée congelée à -20 °C en aliquotes à usage unique (3 µL).

Remarque : Le mélange de fragmentation, le master mix de premier brin, l'enzyme de deuxième brin et les mélanges d'oligonucléotides du kit SureSelect XT RNA Direct ont été décongelés sur glace et agités au vortex pendant 5 secondes à vitesse élevée, puis centrifugés brièvement avant utilisation.

1. Les échantillons d'ARN total ont été préparés dans un volume de 4 µL d'eau exempte de nucléases. Les quantités de départ variaient d'une expérience à l'autre et sont indiquées dans la présentation des résultats.

Remarque : De plus petits volumes d'échantillons d'ARN de départ (< 4 µL) sont permis mais il est déconseillé d'utiliser de plus grands volumes (>4 µl) d'échantillons d'ARN de départ.

2. Le mélange de fragmentation a été ajouté à l'échantillon d'ARN pour obtenir un volume final de 20 µL.
3. L'échantillon d'ARN a été fragmenté en le chauffant dans un cycleur thermique SureCycler 8800 (ou équivalent) dans des conditions basées sur les mesures de qualité de chaque échantillon d'ARN selon les recommandations du protocole du kit SureSelect XT RNA Direct (G9691). Les paramètres de fragmentation utilisés pour chaque échantillon sont indiqués dans le Tableau S1. Après la fragmentation, les échantillons ont été conservés sur glace jusqu'à l'étape de synthèse du premier brin.
4. Une solution mère de 4 µg/µL d'actinomycine D dans du DMSO a été diluée dans de l'eau jusqu'à obtenir une solution de 120 ng/µL (3 µL d'actinomycine D + 97 µL d'eau).
5. Un mélange de réaction suffisant pour 8 échantillons a été préparée pour la synthèse du premier brin (Table S2). Le mélange de réaction a été agité au vortex et gardé sur glace jusqu'à utilisation.
6. 8,5 µL de mélange de réaction pour le premier brin ont été ajoutés à chaque 20 µL d'échantillon fragmenté sur glace. Les échantillons ont été agités au vortex puis centrifugés brièvement.
7. Les réactions de 28,5 µL ont été incubées dans un thermocycleur SureCycler 8800 préprogrammé pendant 10 minutes à 25 °C, puis 40 minutes à 37 °C et ensuite conservées 4 °C (ou sur glace) jusqu'à l'étape de synthèse du deuxième brin.
8. Les tubes de mélange deuxième brin + mélange d'enzymes de réparation des extrémités et le deuxième brin RNA-seq + oligonucléotides de réparation des extrémités ont été agités au vortex avant utilisation.
9. 25 µL de mélange deuxième brin + enzymes de réparation des extrémités (capuchon bleu) ont été ajoutés aux 28,5 µL de réaction premier brin sur glace.
10. Immédiatement après, 5 µL de mélange deuxième brin + oligonucléotides de réparation des extrémités (capuchon jaune) ont été ajoutés.
11. Les tubes d'échantillons ont été rebouchés, agités au vortex, centrifugés brièvement, puis remis sur glace.

12. Les réactions de 58,5 µL ont été incubées dans un thermocycleur SureCycler 8800 pendant 60 minutes à 16 °C puis conservées à 4 °C (ou sur glace) jusqu'à l'étape de purification SPRI.
13. Après la synthèse du deuxième brin, l'ADNc a été purifiée sur billes SPRI par le protocole résumé dans le Tableau S3.

Préparation de banques d'ADNc SureSelect XT HS avec les réactifs du kit de préparation de banques SureSelect XT HS

14. La réparation des extrémités, l'addition de queue poly(dA) et la ligation d'adaptateurs XT HS de l'ADNc ont été effectuées avec les réactifs et selon les instructions détaillées du système d'enrichissement ciblé SureSelect XT HS pour le protocole et le kit de séquençage multiplexe à extrémités appariées Illumina (G9702).

Remarque : Le protocole SureSelect XT HS a été suivi à partir de l'étape 3 page 27 avec les exceptions et les modifications mentionnées aux étapes 15 à 20 de ce protocole.

15. À l'étape 6 page 34 du protocole SureSelect XT HS. L'amplification PCR de pré-capture des banques a été effectuée selon les instructions du protocole en modifiant le nombre de cycles de PCR comme suit :
    a. UHRR, Seraseq™ v2 pour 12 cycles de PCR (ARN de grande qualité)
    b. Échantillons mammaires en quadruple pour 14 cycles de PCR (ARN de moindre qualité)
16. Les bibliothèques PreCap ont été évaluées sur le Bioanalyseur ou la TapeStation en termes de rendement et de distribution de taille des molécules.

Enrichissement ciblé et séquençage des banques d'ADNc SureSelect XT HS

17. À l'étape 1 page 46 du protocole SureSelect XT HS. L'enrichissement ciblé a été réalisé avec 200 ng de banque PreCap de départ en utilisant les réactifs d'hybridation rapide SureSelect XT HS. 5 µL de sondes d'exome humain complet V7 SureSelect ont été utilisés pour l'hybridation.
18. La capture par billes de streptavidine et les lavages ultérieurs ont été effectués conformément au protocole.
19. L'amplification PCR post-capture a été réalisée selon le protocole avec 12 cycles de PCR pour tous les échantillons. Les bibliothèques PostCap ont été évaluées sur le Bioanalyseur ou la TapeStation en termes de rendement et de distribution des tailles des molécules.
20. Toutes les banques ont été séquencées sur une plateforme Illumina HiSeq 4000 avec une longueur de lecture de 2 × 150.

www.agilent.com

Destiné à la recherche uniquement.
Ne pas utiliser à des fins de diagnostic.

Ces informations peuvent être modifiées sans préavis.

PR7000-2381

Imprimé aux États-Unis, 29 janvier 2020
5994-1644FR

Agilent
Trusted Answers