Guide de bonnes pratiques bioinformatiques de séquençage ARN (RNA-Seq) en diagnostic de génétique moléculaire

Document rédigé par le groupe de travail « RNA-Seq »
de l'association BioInfoDiag

Juin 2025

BioInfoDiag

Réseau Français de Bioinformatique pour le Diagnostic

Table des matières

Préambule

Face à l'essor du séquençage de deuxième et troisième génération de l'ARN (RNA-Seq), l'association BioInfoDiag (« réseau français de bioinformatique pour le diagnostic ») a formé dès 2023 un groupe de travail dans le but de réaliser un état des lieux des méthodes existantes et éditer un document décrivant les pratiques du RNA-Seq à des fins diagnostiques, en se concentrant sur les aspects de traitement bioinformatique. Ce groupe est composé d'un panel de bioinformaticiens et biologistes évoluant dans différents centres français. Le groupe a tout d'abord identifié 3 grandes thématiques à aborder, puis s'est réuni régulièrement entre octobre 2023 et novembre 2024 avec pour objectif de :

- Décrire les différents outils disponibles
- Établir leur pertinence dans un cadre diagnostic
- Récolter quand cela était possible des retours d'expérience utilisateurs

Les résultats de ces travaux sont présentés dans ce document, qui ne se veut pas exhaustif, mais doit permettre une mise en œuvre facilitée de la bioinformatique du RNA-Seq dans un cadre de diagnostic moléculaire.

Le groupe de travail RNA-Seq BioInfoDiag.

Participants

Animation du groupe de travail

David Baux – CHU de Montpellier
Jean Muller – CHU de Strasbourg
Marie de Tayrac – CHU de Rennes

Rédacteurs principaux du document

Aurélien Perrin – CHU de Montpellier
Florent Denoual – CHU de Rennes
Perrine Brunelle – CHU de Lille
Thomas Guignard – CHU de Montpellier
Wilfrid Carré – CHU de Rennes
Abdelhakim Bouazzaoui – CHU de Rennes
Simon Cabello-Aguilar – CHU de Montpellier
Amyra Aliouat – CHU de Rennes
Christope Russo – CHU de Lille
Alexandra Lespagnol – CHU de Rennes
Laura Do Souto – CHU de Nantes
David Baux – CHU de Montpellier

Participants

Les autres participants (certains assidus !) aux différentes réunions ont alimenté les réflexions en présentant leurs propres travaux, veille bibliographique, retours d'expériences et questions diverses :

Flora Ponelle-Chachuat - CLCC de Clermont-Ferrand
Christophe Habib – CHU de Toulouse
Jennifer Chiron – CLCC Bergonié
Anne-Sophie Dénommé-Pichon – CHU de Dijon
Anne-Sophie Jourdain – CHU de Lille
Camille Benoist – Institut Curie
Fabrice Bonte – CHU de Lille
Julie Bogoin – Assistance Publique des Hôpitaux de Paris
Elise Guéret – CHU de Montpellier
Julien Buratti – Assistance Publique des Hôpitaux de Paris
Nada Maaziz – CHU de Dijon
Perrine Brunelle – CHU de Lille
Christophe Habib – CHU de Toulouse
John Rendu - CHU de Grenoble
Pierre-Antoine Rollat-Farnier - Hospices Civils de Lyon
Christel Vaché – CHU de Montpellier
Myriam Vezain – CHU du Rouen
Céline Derambure – Faculté Médecine Pharmacie de Rouen
Claude Houdayer – CHU de Rouen
Marion Larrieux – CHU de Montpellier
François Lecoquierre – CHU de Rouen
Claire Guissard – CHU de Nîmes
Jérôme Reboul – Inserm
Martin Broly – CHU de Montpellier
Charles Van Goethem – CHU de Montpellier
Luc Thomes – CHU de Lille
Eléonore Frouin – Institut Curie
Aurélien Bourdon – CLCC Bergonié
Seydi Thimbo – CHU de Lille
Laetitia Gaston – CHU de Bordeaux
Kahia Messaoudi – CHU d'Amiens
Laurence Lodé – CHU de Reims
Thérèse Commes – Inserm

Svetlana Gorokhova – Université d'Aix-Marseille
Rihab Azmani – Institut Bergonié Bordeaux
Laurent Castera – CLCC Baclesse
Benjamin Cogné – CHU de Nantes
Corentin Marco – CHU de Nimes
Rihab Azmani – CLCC Bergonié
Sylvain Mareschal – CHU de Lyon
Raphaël Leman – CLCC Baclesse
Valentin Vautrot – Université de Bourgogne
Christophe Russo – CHU de Lille

Présentations invitées

Le groupe remercie chaleureusement pour leur temps et la qualité de leurs interventions :
Jules Garreau – Université de Rennes
Ariane Mahieux – CHU de Brest
Justine Labory – INRAE
Frédéric Escudié – CHU de Toulouse
Franck Tirode – Inserm
L'équipe Bio2M notamment, Thérèse Commes, Anthony Boureux, Benoit Guibert – Inserm

Relecteurs

Enfin, ce document a été relu, corrigé et amendé avec sérieux et bienveillance par :
Sylvie Tuffery – Inserm (partie épissage)
Benjamin Cogné – CHU de Nantes
Jean Muller – CHU de Strasbourg
Marie de Tayrac – CHU de Rennes

Le groupe de travail remercie tout aussi chaleureusement Mathieu Chopelet pour l'assistance technique d'une efficacité sans faille.

Note de pré-lecture :

Les différentes métriques statistiques évoquées dans le document sont explicitées ici :
https://datascientest.com/matrice-de-confusion

1 Analyse de transcrits par séquençage haut-débit

1.1 Les différentes applications en diagnostic moléculaire

La conception d'une étude de séquençage d'ARN (RNA-Seq) dans le domaine de la génétique médicale débute par **la formulation d'une question biologique**. Cette étape est déterminante puisqu'elle oriente l'ensemble de l'étude et influence le choix des méthodes expérimentales et le traitement des données subséquent.

Le type d'altérations génétiques à étudier, somatiques (présentes uniquement dans les cellules tumorales) ou constitutionnelles (affectant la lignée germinale), **conditionne le choix des méthodes expérimentales, le traitement des données et le type de contrôles à inclure.** Il en est de même avec la taille de la cohorte étudiée ou encore la capacité à recruter des échantillons témoins.

Le **délai de rendu** des résultats constitue un paramètre essentiel, surtout dans un contexte diagnostique où les analyses somatiques peuvent nécessiter une réponse plus rapide que les études de génétique constitutionnelle. De plus, la possibilité d'inclure des échantillons prénataux doit être envisagée dès le début du processus de conceptualisation.

L'état des lieux actuel de la recherche en RNA-Seq souligne la priorité accordée au développement de méthodes diagnostiques, notamment dans le contexte constitutionnel. Les **échantillons à considérer** incluent le sang, les échantillons FFPE (fixés, inclus en paraffine) et divers tissus spécifiques (peau, muscle, culture cellulaire…). La qualité de l'ARN extrait est déterminante pour une analyse qualitative. Cette étape doit s'appuyer sur des retours d'expériences des centres ayant travaillé sur des échantillons et matrices similaires, et peut nécessiter une adaptation des protocoles ainsi que des méthodes de traitement de données.

Le séquençage de transcrits issus de tissus humains ou de cellules uniques permet aujourd'hui de réaliser plusieurs types d'analyses, dont certaines sont déjà bien implémentées en diagnostic, comme la recherche de fusion de transcrits en oncologie somatique, alors que d'autres sont à l'état de recherche ou en phase de validation diagnostique comme l'expression génique différentielle pour les maladies rares.

On distingue 3 grandes catégories d'analyses en RNA-Seq à potentiel d'application diagnostique (**Figure 1**) :

- **Analyse de l'épissage :**

Le processus d'épissage alternatif peut générer différentes formes d'ARNm à partir d'un même gène. **Le transcriptome d'une cellule, d'un tissu, d'un organe varie en termes de qualité et de quantité de ces transcrits.** Le séquençage de l'ARN permet de révéler des variations d'épissage, offrant des informations sur la diversité des transcrits produits par un gène en conditions physiologiques et pathologiques. Certains de ces variants d'épissage peuvent être des altérations délétères impliquées dans le développement de pathologies.

- **Analyse des transcrits de fusion :**

La recherche de fusion de gènes est une analyse moléculaire visant à détecter des **anomalies génétiques dans lesquelles deux gènes distincts ou une séquence régulatrice et un gène fusionnent** pour former un gène hybride. Ces évènements peuvent résulter de réarrangements chromosomiques, tels que des translocations, et sont souvent associées à certains types de cancers. En génétique somatique, des kits commerciaux sont spécialement conçus pour interroger des fusions spécifiques.

- **Analyse de l'expression différentielle :**

Le séquençage de l'ARN permet de mesurer le niveau d'expression des gènes, c'est-à-dire **la quantité d'ARN messager (ARNm) produite par la transcription d'un gène particulier**. Cette quantification peut se faire au niveau du gène ou de ses isoformes. Il existe de nombreuses méthodes de quantification et il faut être particulièrement vigilant car ces données sont très sensibles à **diverses sources de variabilités** (effet « batch », manipulateur, séquenceur…).

Les études d'expression différentielle peuvent avoir des objectifs variés :

- Identifier les gènes différentiellement exprimés dans deux conditions biologiques différentes
- Identifier les gènes différentiellement exprimés dans un échantillon en particulier par rapport à une cohorte de contrôles
- Déterminer des signatures transcriptomiques caractéristiques de tissus ou de pathologies
- Etablir des classifications
- Identifier des biomarqueurs candidats

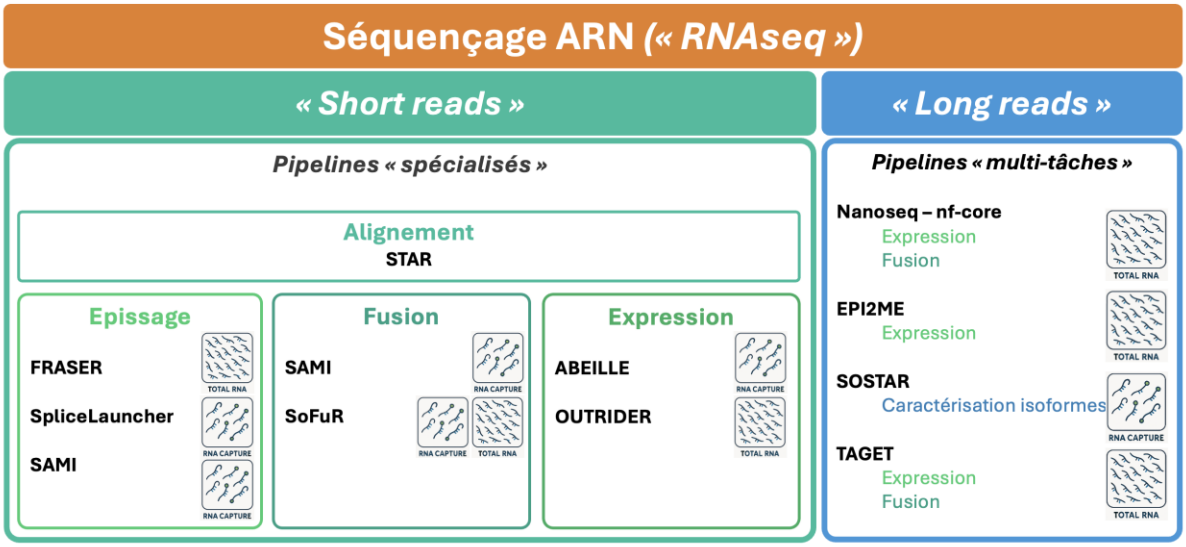

Figure 1 : Résumé des stratégies de séquençages ARN et principaux outils associés. Les logos « RNA capture » et « total RNA » ont été réalisé à l'aide de ChatGPT.

De plus, le RNA-Seq permet non seulement d'étudier l'expression monoallèlique d'un transcrit, mais également aussi à l'image du séquençage de l'ADN de détecter d'autres types d'altérations telles que des variants nucléotidiques simples (« Single Nucleotide Variation », SNV), des petites insertions/délétions (indels) ou des variations du nombre de copies (« Copy Number Variation », CNV).

Les méthodes classiquement utilisées en recherche cherchent à établir des différences entre deux conditions (atteint / non atteint, tissu 1 / tissu 2, traitement / contrôle), en comparant deux cohortes que l'on essaiera de construire statistiquement comparables. En diagnostic, **l'approche la plus fréquente consistera à comparer un échantillon individuel (le patient) à une cohorte contrôle**, constituée soit de véritables contrôles, soit d'une cohorte pseudo-contrôle composée des autres patients du laboratoire ou de l'expérience (ex : même run de séquençage). Dans ce dernier cas, il est indispensable d'**exclure de la cohorte contrôle les patients trop similaires** au patient à étudier qu'il s'agisse du phénotype, du génotype (s'il est connu), ou par des liens familiaux. Il convient également d'exclure les réplicats biologiques ou techniques.

Du point de vue biologique, la constitution d'une base de données exhaustive de données de profils transcriptomiques « normaux » est cruciale pour la référence et la caractérisation des altérations détectées. La simulation de jeux de données permet de s'assurer que le pipeline analytique est capable d'identifier les anomalies recherchées dans un jeu de données contrôle. Selon les questions posées, la conception expérimentale peut autoriser l'utilisation de **jeux de données issus de patients comme cohorte contrôle**. Ceci peut être **pertinent pour la recherche de défauts d'épissage** mais peut aussi être **un frein pour des études d'expression** de transcrits pathologiques. En effet, dans le cas de l'analyse de l'épissage, la présence d'un événement alternatif lié à une variation génétique sera observable uniquement chez un patient en cas d'événement rare. Cependant, **dans le cadre de l'étude d'expression différentielle, l'utilisation d'un ensemble de données de contrôles demeure importante**. L'emploi de données de patients peut être problématique, car la condition pathologique elle-même implique fondamentalement une signature transcriptionnelle potentiellement altérée.

De manière générale, l'important est de limiter au maximum les biais liés à la nature des échantillons, aux méthodes de préparation des échantillons et aux conditions expérimentales. Ainsi il est recommandé d'utiliser des jeux de données comparables - un même tissu, un même protocole de préparation des ARNs, une même méthode expérimentale pour le traitement de l'échantillon et des outils d'analyse identiques.

1.2 Préparation des échantillons

Le choix du tissu à analyser est déterminant pour l'analyse transcriptomique, chaque type cellulaire possédant un profil d'expression spécifique. En cancérologie, l'accès au tissu tumoral est souvent facilité, permettant d'explorer directement le transcriptome de la lésion. À l'inverse, dans les maladies rares, l'accès au tissu pathologique est souvent limité, ce qui conduit à privilégier des tissus plus accessibles ou des approches alternatives en fonction de la question diagnostique. Les biopsies musculaires ont démontré une efficacité notable dans le diagnostic des pathologies neuromusculaires, augmentant le taux de diagnostic jusqu'à 35% (Cummings et al., 2017)**. Les fibroblastes sont actuellement considérés comme le « gold standard » pour l'analyse transcriptomique** en pratique clinique, mais leur utilisation à grande échelle est contraignante. Les prélèvements sanguins sur tubes PAXgene offrent une alternative moins invasive, mais leur performance dans les analyses agnostiques reste limitée. Enfin, **la culture lymphocytaire à court terme**, récemment validée au CHU de Nantes, **représente une piste prometteuse** pour ce type d'analyse.

Les différentes alternatives testées au sein du groupe de travail et dans la littérature sont listées dans le **Tableau 1** (liste non exhaustive). Le choix du design expérimental devra être défini en étroite collaboration entre les biologistes et les bioinformaticiens.

Tableau 1 : les différentes sources d'ARN

Tissu	Avantages	Inconvénients
Tissu d'intérêt (biopsie)	Bonne qualité de l'ARN Gènes du tissu d'intérêt exprimés	Invasif Accessibilité du tissu Pas d'inhibition du NMD
Fibroblastes (culture)	Bonne qualité de l'ARN (RIN>9) Expression différentielle possible avec données contrôles disponibles Inhibition du NMD possible Tests fonctionnels ultérieurs possibles	Invasif Plus difficile à intégrer dans une activité de diagnostic Temps de culture long (~10-15j)
Sang total (PAXgene)	Facile à obtenir Stabilisation des ARNm	Saturation des ARNm par de la globine (déplétion possible) Qualité d'ARN moyenne Résultats peu encourageants dans les études non ciblées Pas d'inhibition possible du NMD
Lymphocytes (culture)	Facile à obtenir Bonne qualité de l'ARN (RIN>9) Expression différentielle possible Temps de culture faible (48-72h) Inhibition du NMD possible	Moins étudiés que les fibroblastes (pas de données contrôles disponibles) Moins de matériel pour des études fonctionnelles post RNA-Seq

1.3 Les différentes techniques de RNA-Seq et leurs applications

La grande majorité des ARNs présents dans un isolat cellulaire ou dans un tissu **sont des ARNs ribosomaux**. **Ces ARNs ne sont pas ceux ciblés** pour l'analyse de transcrits et doivent être éliminés. Pour cela, on peut soit effectuer une **sélection positive** (par sélection des ARN polyadénylés en fixant la queue polyA ou par capture des séquences exoniques) **ou négative** (élimination des ARNr par hybridation et séparation, ribo-déplétion).

Les méthodes positives sont recommandées pour l'**analyse des gènes codants** et des transcrits résultants alors que **la ribo-dépletion** permettra aussi l'étude **des miARNs ou des ARNs non codants** en général (incluant les lncARNs). La sélection des ARNs poly-adenylés peut engendrer des biais de mesure avec une couverture relativement plus importante pour les régions proches de la queue polyA (région terminale du transcrit) par rapport à une technique de capture qui sera plus homogène (Zhao et al., 2018). La capture polyA présente une meilleure élimination de l'ADN génomique résiduel par rapport à un traitement par DNAse (Ura et al., 2022). Attention cependant à l'utilisation de sondes de capture, qui pourraient avoir un périmètre de détection limité. Par exemple, si on utilise des sondes de capture exoniques uniquement, les rétentions introniques totales ne seront pas directement visibles mais identifiées par 2 rétentions partielles en 5' et 3' de l'intron. De plus, certains outils comme IRFinder(Lorenzi et al., 2021) seront inopérants ou avec des performances fortement réduites.
L'utilisation de **technologies de capture offre une flexibilité importante** au niveau des cibles :

- RNA-Seq ciblé sur un panel de gènes d'intérêts,
- Analyse du transcriptome codant dans son ensemble (en utilisant un lot de sondes destinées à l'exome par exemple).

Un autre argument en faveur de cette technologie en diagnostic est la **possibilité d'utiliser les mêmes sondes que pour le séquençage ADN**, afin de réaliser le séquençage ARN. Cette possibilité est offerte par la plupart des fournisseurs actuels.

Enfin, valable **pour une analyse d'expression génique uniquement, le 3' RNA-Seq** se concentre sur l'analyse de l'extrémité 3' des transcrits. Cette technique est dédiée à l'analyse quantitative de l'expression génique. Elle présente plusieurs avantages, notamment celui de nécessiter beaucoup moins de quantité de séquences qu'une analyse classique pour quantifier les transcrits (environ 50 000 lectures VS 30M), mais aussi de mieux caractériser les transcrits courts. En revanche, la méthode poly-A semble pouvoir détecter plus de gènes différentiellement exprimés (Ma et al., 2019). **Le 3' RNA-Seq est la méthode utilisée en single cell RNA-Seq**.
Les avantages et inconvénients de chacune des approches en fonction des différentes applications sont résumés **Tableau 2**.

Tableau 2 : Pertinence de la méthode de capture des ARNs selon la finalité de l'analyse. Les analyses présentées dans ce tableau concernent les gènes codants.

Technique de sélection des ARNs	Épissage	Fusion	Expression	SNV/CNV calling	#échantillons par run
polyA	++	+	+	+	++
ribo déplétion	+/-	+/-	+/-	+/-	+
Capture	+	+	+	+	+++
3' RNA-Seq	-	-	++	-	++++

1.4 Apport des UMIs

Le séquençage d'ARN utilisant des UMIs (« **Unique Molecular Identifiers »**) repose sur l'ajout, dès la préparation des librairies et avant toute amplification PCR, d'une étiquette moléculaire unique à chaque molécule d'ADNc. Ces UMIs sont ensuite co-amplifiés avec les séquences d'ADNc. Lors de l'analyse bioinformatique, après le séquençage, **leur identification permet de regrouper les fragments dupliqués** portant la même étiquette et de construire une séquence consensus **(Figure 2)**. **Ce processus améliore la précision de la quantification en corrigeant les biais liés à l'amplification,** restituant ainsi fidèlement l'abondance initiale des transcrits présents dans l'échantillon (Roloff et al., 2017).

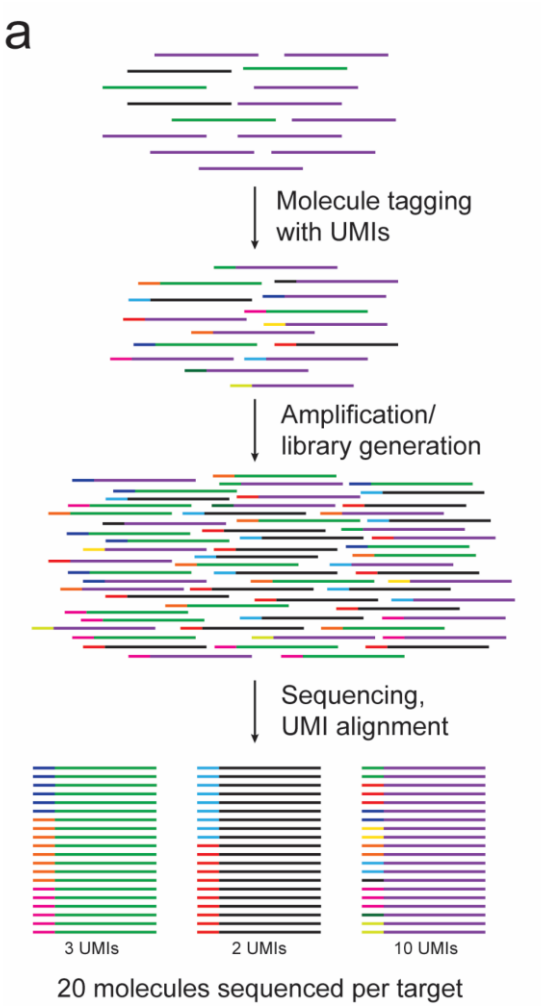

Figure 2 : L'utilisation des UMI dans les librairies de séquençage de seconde génération (Roloff et al., 2017)

Les **UMIs offrent une solution efficace pour réduire le biais d'amplification PCR** garantissant ainsi une quantification précise des ARN initiaux. Leur utilisation améliorerait la détection des variations rares (améliore la limite de détection) et réduirait les faux positifs dus aux erreurs de séquençage et biais de préparation des librairies.

Bien que largement adoptée, l'intégration des UMIs entraîne une complexité technique accrue ainsi qu'un coût de séquençage supplémentaire. À ce jour, aucune directive n'a été établie quant à leur utilisation, et **une évaluation exhaustive des bénéfices associés selon les différentes applications fait encore défaut** (Bieler et al., 2023).

La gestion bioinformatique de ces UMIs doit suivre les étapes suivantes :

- Extraction des UMIs
- Construction de familles de séquences ayant les mêmes UMIs
- Génération d'une séquence consensus pour toutes les séquences appartenant à la même famille d'UMI.

Plusieurs outils bioinformatiques spécialisés gèrent ces étapes. Le package UMI-tools (Smith et al., 2017) propose diverses méthodes de déduplication : i) unique et percentile qui regroupent uniquement les UMIs identiques et ii) « cluster », « adjacency » et « directional » qui connectent les UMIs selon leur distance d'édition via des approches basées sur les réseaux. L'outil fgbio (https://github.com/fulcrumgenomics/fgbio) de Fulcrum Genomics permet de regrouper les lectures selon des critères définis et de générer des séquences consensus. Des pipelines complets comme zUMIs (Parekh et al., 2018) ou Alevin (Srivastava et al., 2019) automatisent l'ensemble du processus en intégrant des corrections d'erreurs adaptées aux données « bulk » ou « single-cell RNA-Seq ». Le choix de ces outils et de leurs paramètres impacte directement la qualité des données et la fidélité de la quantification des ARN.

1.5 Apport des données de lectures orientées (« stranded reads »)

Une des limitations du protocole standard de RNA-Seq est **la perte de l'information sur le brin d'origine** de chaque transcrit. La synthèse d'ADNc double brin aléatoirement amorcé, suivie de l'ajout d'adaptateurs pour le séquençage de nouvelle génération, entraîne cette perte d'information. Cela rend **difficile la détermination précise de l'expression des gènes chevauchants**, c'est-à-dire ceux ayant des coordonnées génomiques partiellement chevauchantes mais transcrits à partir de brins opposés.

Le **séquençage d'ARN orienté (dit « stranded ») permet de surmonter cette limitation**. Parmi les protocoles existants, la méthode d'élimination du second brin d'ADN par le dUTP a été identifiée comme l'une des meilleures en termes de simplicité et de qualité des données (**Figure 3**) (Levin et al., 2010).

En préservant l'orientation du brin d'ARN d'origine, cette technique permet d'approfondir l'analyse de l'expression génique et de mieux comprendre les mécanismes régissant la régulation des gènes. Elle est particulièrement utile pour l'analyse des gènes qui se chevauchent et pour la quantification précise de l'expression des gènes (Zhao et al., 2015).

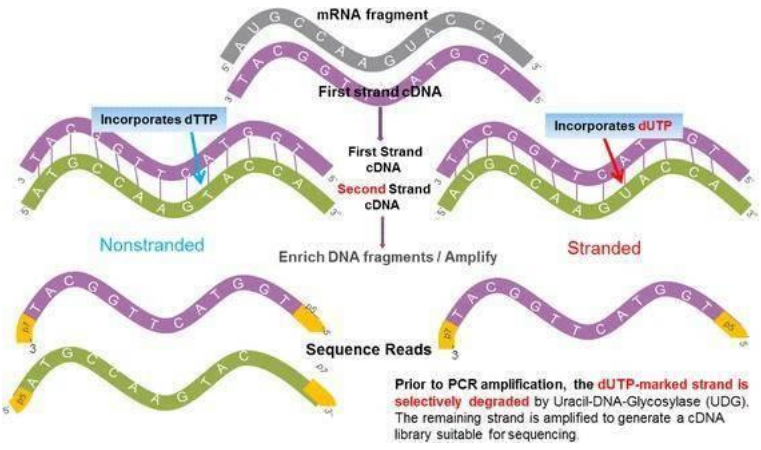

Figure 3 : Non-stranded versus stranded RNA-Seq protocol (Zhao et al., 2015)**.**

2 Contrôle qualité

Tout comme pour le séquençage d'ADN (DNA-Seq), **la qualité de l'échantillon, de la préparation de la librairie et du traitement bioinformatique est déterminante dans la réussite des analyses**. Certaines métriques de qualité seront communes aux deux types de séquençage DNA-Seq et RNA-Seq, d'autres seront spécifiques au RNA-Seq, voire au sous-type d'analyse (expression génique ou recherche de transcrits de fusion par exemple). On distinguera de manière non exhaustive un certain nombre d'évaluations à réaliser sur les données brutes ou alignées :

- **Traitement des données brutes** (fichiers FastQ) :
  - Qualité de séquençage standard (nombre de lectures produites par échantillon, qualité q30, …)
  - % après « trimming » des adaptateurs de séquençage et nettoyage des séquences de mauvaise qualité
  - Gestion des UMIs le cas échéant
- **Traitement des séquences alignées** (fichiers BAM ou CRAM) :
  - % de séquences alignées, taux de duplicats
  - % de lectures ADN, ARN et ambiguës pour qualifier l'échantillon de départ
  - La taille moyenne des lectures ARN
  - Nombre de gènes à zéro lectures et/ou pourcentage de couverture des gènes ciblés par le panel expérimental (évaluation du non spécifique avec le off target) : ces indicateurs reflètent la qualité de la matrice initiale
  - Distribution des lectures sur les gènes (exoniques, introniques, intergéniques)
  - Couverture des gènes de ménage dans le cas du RNA-Seq total ou couverture de gènes témoins dans le cadre les panels ciblés, utilisée pour qualifier la qualité d'expérimentation.

On pourra se référer à la littérature sur le sujet, par exemple :(Zhou et al., 2018) ou encore aux rapports fournis par certains outils : SQANTI3/example/SQANTI3_QC_output/UHR_chr22_SQANTI3_report.pdf at master · ConesaLab/SQANTI3 · GitHub

3 Alignement

L'étape d'alignement est commune à la plupart des types d'analyse, elle consiste à **déterminer la localisation génomique dont sont issues les lectures** à partir des fichiers au format FASTQ. Elle est souvent précédée d'une étape préliminaire qui aura pour but d'éliminer les séquences utilisées pour identifier la provenance des lectures (index).

3.1 Performances des différents algorithmes

Une publication récente (Fenn et al., 2023) a comparé 11 outils d'alignement en utilisant un jeu de données simulées de lectures courtes (76pb) avec une complexité croissante. Ils ont évalué la précision (ou valeur prédictive positive, TP/(TP+FP)) de chacun des outils ainsi que la fraction de lectures non alignées (« unmapped reads »). Les performances sont comparables entre les jeux de données de complexité simple et élevée, avec les volumes de 50M à 200M de lectures : **STAR et HISAT2 sont les outils les plus performants** comme le montre la **Figure 4**.

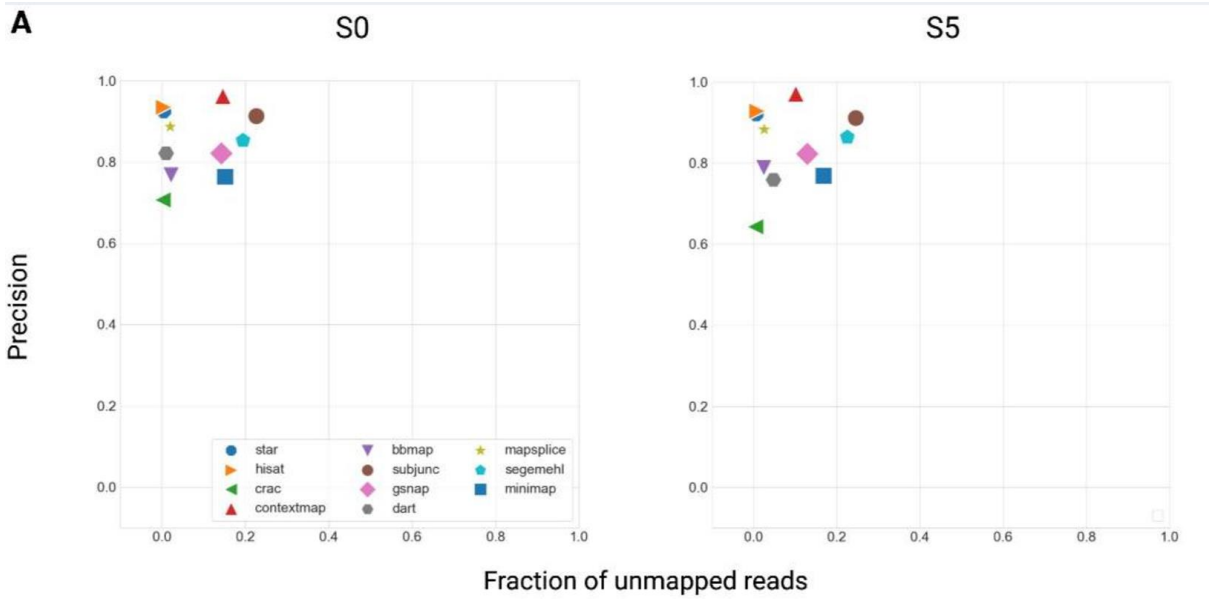

Figure 4 : Précision et fraction de lectures non alignées pour les algorithmes d'alignement « splice-aware » avec les jeux de données S0 (complexité minimale) et S5 (complexité maximale) (Fenn et al., 2023)**.** S0 correspond au jeu de données le plus simple (1 évènement par transcrit, pas d'évènements complexes, …) et S5 au jeu le plus complexe.

Concernant les temps de traitement, HISAT2 et STAR se comportent bien pour l'indexation du génome et pour les étapes d'alignement à proprement parler (**Figure 5**).

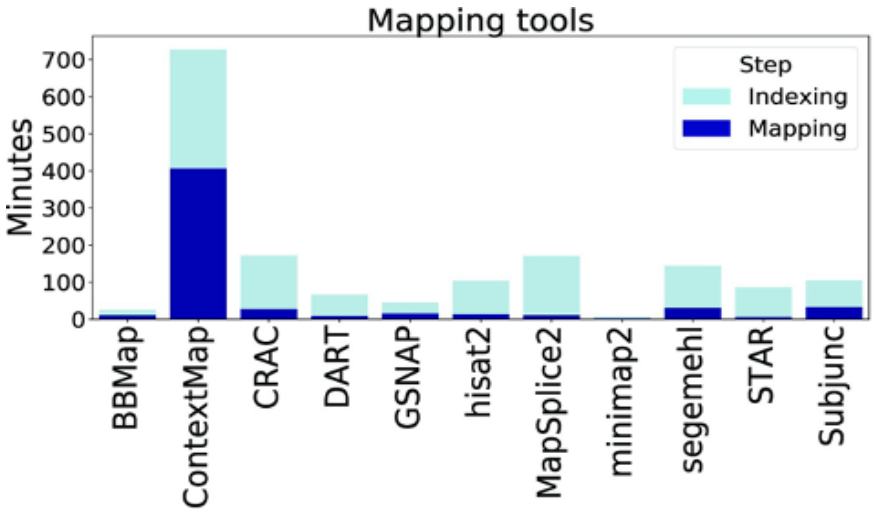

Figure 5 : Temps d'exécution en minutes des outils d'alignements « splice-aware » et des outils de détection d'évènements d'épissage alternatifs (Fenn et al., 2023)**.**

3.2 STAR

STAR a l'avantage d'être basé sur un **algorithme spécialement conçu pour l'analyse des données RNA-Seq**. En plus du fichier d'alignement, il génère également un fichier récapitulatif des jonctions d'épissage.

A noter que pour la détection des évènements d'épissage, STAR est plus performant en mode « 2-pass » qui correspond à 2 alignements successifs, en utilisant les jonctions identifiées lors du premier passage comme annotations pour le second (plus de précisions dans la documentation de l'outil : https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf, p 19).

Une bonne vulgarisation de l'algorithme d'alignement de STAR se trouve ici : https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

STAR, en plus de pouvoir efficacement aligner les lectures courtes issues notamment de séquençage Illumina, est **utilisable pour les lectures longues**, et peut être utilisé en aval pour une analyse de l'épissage mais aussi des transcrits de fusion. En pratique STAR est assez gourmand et utilise 32 Go de RAM par défaut (Dobin et al., 2013).

3.3 HISAT2

Le second algorithme d'alignement performant et rapide est HISAT2 (Kim et al., 2019a) **(Figure 4).** L'originalité de l'algorithme de HISAT2 tient dans l'utilisation d'une **représentation du génome de référence sous forme de graphe**, en se basant sur le génome linéaire classique et en y ajoutant des variants (SNVs et petites indels) en tant que chemins alternatifs. **HISAT2 n'est pas dédié à l'alignement d'expériences de RNA-Seq**, il fonctionne aussi avec des données issues de séquençage génomique classique. Il est à noter cependant qu'actuellement HISAT2 ne traite pas les lectures longues ni les variants de structures (« Structural Variations », SV).

En pratique, **STAR est l'algorithme le plus utilisé** et est recommandé pour l'utilisation de plusieurs outils récents d'analyse de l'épissage (FRASER2 (Scheller et al., 2023), SpliceLauncher (Leman et al., 2020)) notamment pour pouvoir utiliser les fichiers de description des jonctions.

4 Le séquençage d'ARN pour des analyses d'évènements d'épissage

4.1 Les mécanismes d'épissage

L'épissage de l'ARN pré-messager est un processus nucléaire au cours duquel les introns d'une molécule de pré-ARNm naissante transcrite par l'ARN polymérase II à partir de l'ADN génomique sont excisés, et les exons sont ligués pour former une molécule d'ARNm mature **(Figure 6A**) (Sharp, 1994).

4.1.1 Le spliceosome et l'épissage

Le spliceosome, un complexe macromoléculaire, **catalyse la réaction d'épissage** lors de deux réactions de transestérifications successives en recrutant séquentiellement plusieurs sous-unités ribonucléoprotéiques sur les sites consensus d'épissage en 5' et 3' de l'intron (**Figure 6B**).

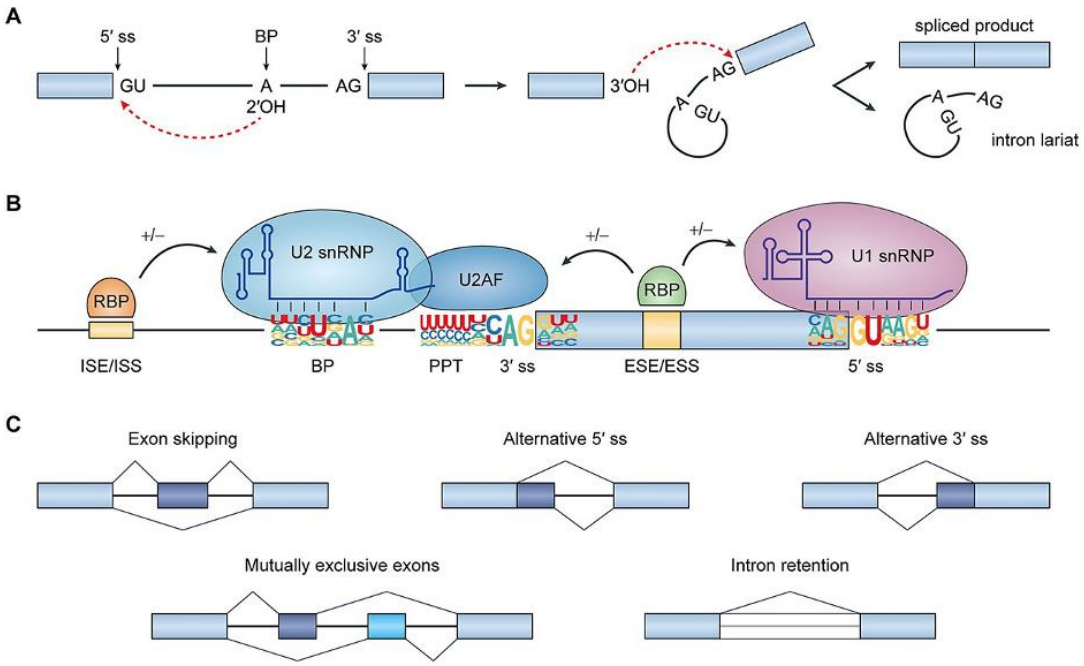

Figure 6 : L'épissage du pré-ARNm et sa régulation. A) Au cours de l'épissage, les introns sont éliminés des transcrits du pré-ARNm par un processus en deux étapes. Dans la première étape, le groupe 2'-hydroxyle de l'adénosine du point de branchement effectue une attaque nucléophile sur le site d'épissage 5', ce qui coupe l'exon en amont de l'intron et génère un intron lasso intermédiaire. Dans la deuxième étape, le groupe 3'-hydroxyle du site d'épissage 5' attaque le site d'épissage 3', ce qui entraîne la ligature des deux exons et la libération de l'intron lasso. **B)** L'épissage des pré-ARNm est régulé par un vaste réseau d'interactions protéine-ARN impliquant des éléments en *cis* le long du pré-ARNm ainsi que des facteurs en *trans* qui reconnaissent ces éléments. Les principaux signaux d'épissage comprennent le site d'épissage 5' (qui comporte un dinucléotide GU conservé), le site d'épissage 3' (qui comporte un dinucléotide AG conservé), la séquence du point de branchement et le tractus polypyrimidine. La snRNP U1 reconnaît le site d'épissage 5' et la snRNP U2 reconnaît la séquence du point de branchement. Les protéines U2AF s'engagent dans le site d'épissage 3' et le tractus polypyrimidine. D'autres éléments régulateurs en *cis* (activateurs et inhibiteurs d'épissage) modulent encore l'épissage par des interactions avec des RBP agissant en *trans*. **C)** Cinq modèles de base de l'épissage alternatif. Abréviations : ss, site d'épissage ; OH, hydroxyle; BP, point de branchement ; PPT, tract polypyrimidine; ESE,"exonic splicing enhancer"; ESS, "exonic splicing silencer"; ISE, "intronic splicing enhancer"; ISS, "intronic splicing silencer"; RBP, "RNA-binding protein"; snRNP, "small nuclear ribonucleoprotein"; U2AF, U2 "auxiliary factor" (Wang et al., 2023)

L'épissage peut produire différents transcrits à partir d'un même gène. On distingue :

- **L'épissage constitutif :** mécanisme d'élimination des introns de l'ARN pré-messager et de ligation des exons dans l'ordre initial sur le génome.
- **L'épissage alternatif :** processus biologique essentiel où les exons et les introns peuvent être inclus ou exclus selon différentes combinaisons pour permettre l'expression de différentes isoformes d'ARNm et de protéines à partir d'un transcrit primaire. Les différents types d'épissage alternatifs sont décrits **Figure 6C**.

L'épissage alternatif, source de complexité du protéome, est finement régulé en fonction des conditions biologiques, par exemple le stade de développement, la spécificité tissulaire/cellulaire, ou encore la différentiation, en réponse à des contraintes environnementales…

Trois principaux sites sont utilisés au cours de l'épissage (**Figure 7A**) :

- **Site donneur :** site d'épissage situé au début d'un intron (extrémité 5' de l'intron)
- **Point de branchement :** site situé en amont du site accepteur (environ 40 pb de l'extrémité 3' de l'intron)
- **Site accepteur :** site d'épissage situé à l'extrémité d'un intron (extrémité 3' de l'intron)

Ces sites conservés sont nécessaires au processus d'épissage de l'ARN. La grande majorité des introns commencent par le dinucléotide GT et finissent par le dinucléotide AG (introns GT-AG). Ils représentent 99,24% des introns humains et sont épissés par le **spliceosome majeur** constitués des snRNP U1, U2, U4, U5 et U6. Ils sont dits **Introns de type U2**. Le spliceosome majeur est capable également d'épisser les introns de type GC-AG (0,7% des introns). Il existe une catégorie différente d'introns (AT-AC) dont l'épissage est assuré par un spliceosome différent, **le spliceosome mineur**, constitué des snRNPs (« **small nuclear ribonucleoproteins ») U11, U12 and U6atac. Ces introns dits de **type U12**, sont rares et ne représentent que 0,05% des introns (Burset et al., 2001) soit environ 70 gènes chez l'humain contenant un ou des introns de type U12. Les interacteurs exoniques des deux types d'introns sont représentés **Figure 7B**. La proportion de sites non canoniques, différents de GT-AG, GC-AG ou AT-AC serait d'environ 0,02% des introns.

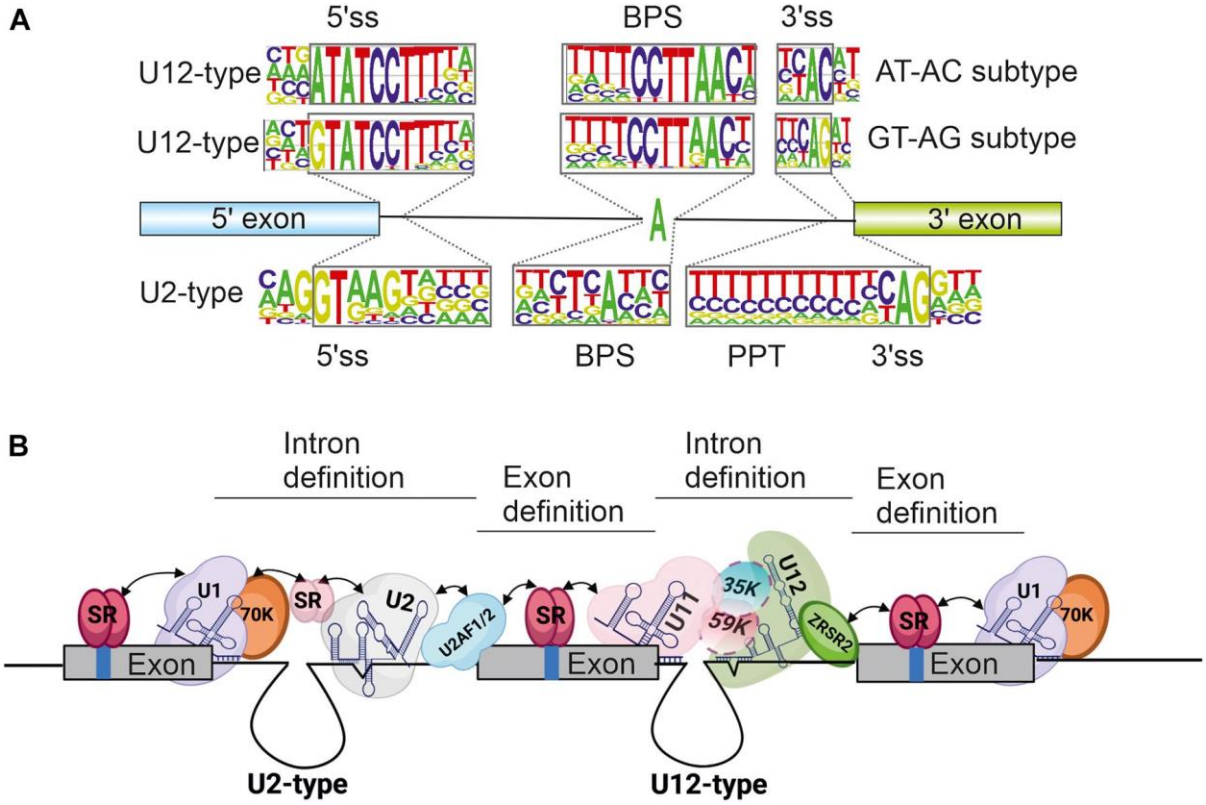

Figure 7 : A) Séquences consensuelles des sites d'épissage des introns mineurs et majeurs. Les introns mineurs ou de type U12 peuvent en outre être regroupés en sous-types AT-AC et GT-AG sur la base des premiers et derniers di-nucléotides. **B)** Schéma des interacteurs exonique ayant lieu dans un gène contenant un intron mineur flanqué d'introns de type U2. Les composants snRNA et protéiques sélectionnés sont indiqués. Les exons sont représentés par des boîtes pleines et les introns par des lignes. Les éléments de régulation à l'intérieur des exons sont représentés par des barres bleues (Akinyi and Frilander, 2021).

4.1.2 Les séquences cis-régulatrices auxiliaires

Les sites donneurs/accepteurs et de branchement ne sont pas les seuls éléments qui régulent l'épissage, d'autres séquences régulatrices (« splicing regulatory sequences ») interviennent dans l'épissage constitutif et alternatif (Goren et al., 2006). Ce sont de courtes séquences (en moyenne entre 4 et 18 nucléotides) qui recrutent différentes protéines de liaison à l'ARN (« RNA Binding Proteins, RBPs ») qui influencent la sélection des sites d'épissages, en favorisant le positionnement du spliceosome sur les sites d'épissage ou au contraire, en l'inhibant. Elles jouent un rôle particulièrement important dans la **régulation spatio-temporelle de l'épissage alternatif**.

Ces séquences sont de quatre types différents, classées selon leur position (exon ou intron) et l'effet de la RBP recrutée sur l'épissage (Activatrice ou inhibitrices) :

- Activateurs exoniques d'épissage ou « Exonic splicing enhancer » (ESE)
- Inhibiteurs exoniques d'épissage ou « Exonic splicing silencer » (ESS)
- Activateurs introniques d'épissage ou « Intronic splicing enhancer » (ISE)
- Inhibiteurs introniques d'épissage ou « Intronic splicing silencer » (ISS)

La présence de ces séquences régulatrices distribuées dans les exons et introns, à distance des sites, ont conduit à considérer que **n'importe quel variant exonique (ou intronique) peut altérer une séquence régulatrice et conduire à un défaut d'épissage**. La fréquence de ces variants d'épissage en pathologie humaine reste cependant très inférieure à celle des variants touchant les sites donneurs et accepteurs.

De plus, des variants nucléotidiques génomiques peuvent produire des transcrits anormaux par création/renforcement ou destruction/diminution de séquences signales.

4.2 Analyse de l'épissage

L'analyse de l'épissage à partir de données RNA-Seq en lectures courtes peut se schématiser en 3 étapes :

- Alignement (étape commune aux 3 types d'analyses**, Cf. 4. Alignement**)
- Détection
- Prédiction/priorisation

Chacune de ces étapes nécessite des outils bio-informatiques particuliers. Dans un contexte diagnostic, il est essentiel de tenir compte du fait que **les analyses seront réalisées sous forme de comparaison entre un échantillon (patient) et une cohorte** (soit un groupe contrôle, soit un groupe de patients, servant de contrôle pour chaque patient individuellement), plutôt que, selon la démarche plus courante en recherche, comparer une cohorte de contrôle à une autre cohorte de conditions différentes.

4.3 Détection

La majorité des outils de détection vont permettre d'identifier 4 types majeurs d'altérations ou partie de ces 4 types d'altérations :

- Site 3' alternatif (3A)
- Site 5' alternatif (5A)
- Rétention intronique (RI)
- Saut d'exon (qui s'exprime comme une combinaison des 2 premières altérations) (SE)

Le **Tableau 3** récapitule pour chaque outil considéré les types d'évènements détectables.

Tableau 3 : **Types d'évènements détectés par outil**

Outil	Type d'évènement détecté
ASGAL	SE, RI, 5A, 3A
ASpli	SE, RI, 5A, 3A
EventPointer	SE, RI, 5A, 3A, EME, SEM
IRFinder / IRFinder-S	RI (total)
MAJIQ	SE, RI, 5A, 3A
SGSeq	SE, RI, 5A, 3A, PEA, DEA, SEM (2 exons)
splAdder	SE, RI, 5A, 3A, EME, SEM
Whippet	SE, RI, 5A, 3A, PEA, DEA, initiation de transcription, site de polyadénylation
FRASER / FRASER2	SE, RI, 5A, 3A, SEM
SpliceLauncher	SE, 5A, 3A, RI partielles
LeafCutterMD	SE, RI, 5A, 3A, SEM

SE : Saut d'Exon, RI: Rétention Intronique, 5A: site 5' d'épissage Alternatif, 3A, site 3' d'épissage Alternatif, EME: Exon Mutuellement Exclusif, SEM: Saut d'Exon Multiple, PEA: Premier Exon Alternatif, DEA: Dernier Exon Alternatif. Adapté de (Fenn et al., 2023). Sont inclus les 8 outils testés dans l'étude de (Fenn et al., 2023), ainsi que 3 outils supplémentaires, FRASER, SpliceLauncher et LeafCutterMD.

4.3.1 IRFinder, IRFinder-S et autres algorithmes évalués dans l'étude Fenn et al., 2023

IRFinder et son successeur IRFinder-S sont des outils populaires (plus de 200 citations) qui permettent une **détection des rétentions introniques totales exclusivement**. La détection (IR ratio) se base sur le rapport des lectures introniques avec la totalité des lectures (introniques et exoniques) : plus ce rapport est élevé, plus la rétention intronique est importante. Les performances de l'outil sont directement liées à la qualité des annotations utilisées (identification des séquences exoniques). La publication descriptive de IRFinder-S (Lorenzi et al., 2021) montre que IRFinder-S est l'outil le plus performant pour la détection de ce type d'évènements (comparaison avec iREAD, Whippet et MAJIQ, 2 jeux de données différents, meilleur VPP, précision et taux de fausses découvertes (« False Discovery Rate », FDR)).

Fenn et collaborateurs ont **comparé les 8 premiers outils** du **Tableau 3** (Fenn et al., 2023). EventPointer, SGSeq (de novo version) et splAdder ont montré une sensibilité (TP/(TP+FN), capacité à détecter les évènements) faible sur un jeu de données simples (détection de moins de 10% des évènements pour ces 3 outils).

ASGAL demande beaucoup de temps de calcul. Sur ce jeu de données synthétiques de complexité croissante, les outils les plus pertinents étaient Whippet et SGSeq (version utilisant des annotations). Cependant, globalement, **tous les outils évalués dans cette étude ont montré une faible propension à détecter de nouveaux évènements** (non annotés), ce qui constitue une proportion importante des évènements à identifier en diagnostic. Les auteurs conseillent IRFinder pour l'identification des rétentions introniques *de novo*, et Whippet et/ou MAJIQ pour les autres types d'évènements. A noter enfin qu'IRFinder sera non-fonctionnel ou avec des performances très réduites si l'expérience de RNA-Seq est faite à partir de sondes de capture.

4.3.2 LeafCutter et LeafCutterMD

En 2020, un algorithme populaire (**LeafCutter**) et performant pour la comparaison de cohortes **a été adapté à la problématique du diagnostic** (un échantillon comparé à une cohorte, LeafCutterMD (Jenkinson et al., 2020)). Cette méthode combine une représentation sous forme de graphes des excisions d'introns (identifiés par les lectures divisées (« split reads ») ancrées par au moins 6pb dans chaque exon) et une approche statistique (construction d'une distribution multinomiale de Dirichlet) à partir d'une matrice de comptage introns/échantillon.

Cependant, l'approche statistique LeafCutterMD ne semble pas adaptée aux situations de comparaison d'un patient à une cohorte contrôle issue d'un ou de plusieurs runs de séquençage différents (Wang et al., 2023).

4.3.3 DROP, FRASER et FRASER2

Cette étude (Wang et al., 2023) considère que l'approche la plus pertinente est celle basée sur un auto-encodeur pour dé-bruiter le jeu de données, utilisée pour FRASER, et reconduite dans FRASER2 (Scheller et al., 2023). FRASER2 utilise comme seule métrique l'**indice intronique de Jaccard** (« Intron Jaccard Index »), efficace notamment pour diminuer l'effet des facteurs confondants (biologiques ou techniques) sur les résultats.

L'indice intronique de Jaccard (**Figure 8**) va capturer les lectures associées à un intron en rapport à toutes les autres lectures couvrant un des 2 sites accepteur ou donneur de l'intron considéré.

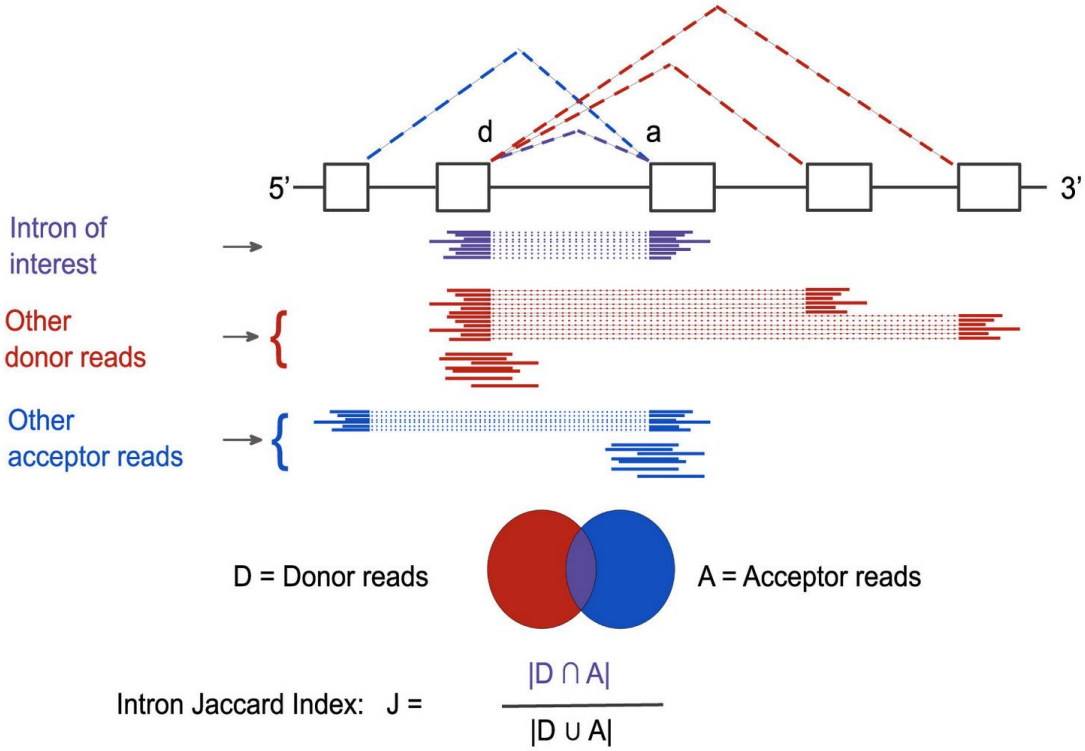

Figure 8 : Indice de Jaccard intronique (Scheller et al., 2023)

La **Figure 8** (Scheller et al., 2023) montre comment sera calculé l'indice intronique de Jaccard en fonction de différents types d'épissage aberrants. **Cet indice unique permet de recouvrir les 3 métriques utilisée dans la version 1 de FRASER** et permet un traitement plus rapide diminuant par 10 le nombre d'événements aberrants détectés sur la cohorte GTEx et en diminuant les biais liés à la profondeur de séquençage (Scheller et al., 2023).

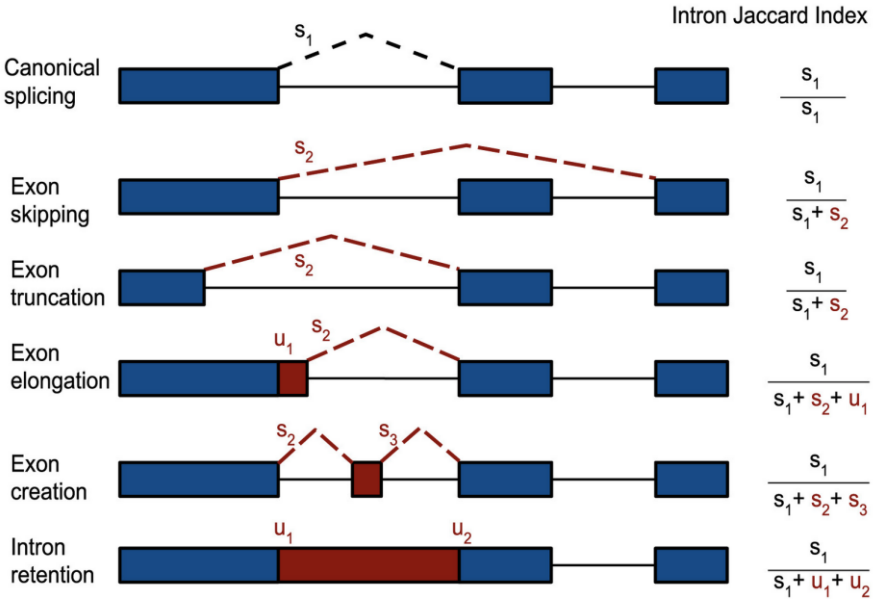

Figure 9 : Représentation de différents types d'événements d'épissage aberrants reconnus par l'indice intronique de Jaccard (Scheller et al., 2023)**.**

FRASER 2 semble représenter la méthode la plus récente et la plus utilisée pour la détection d'évènements d'épissage aberrants au sein d'une cohorte de patients (ou mieux patients VS contrôles) (**Figure 9**). Cependant, chaque nouvel échantillon à tester doit être comparé à la cohorte existante, ou à une cohorte contrôle fixe (chaque laboratoire doit donc avoir sa cohorte témoin). **La cohorte doit être issue de conditions d'analyses similaires** (même type de design, même type d'échantillon), même si FRASER 2 ne demande pas des échantillons issus d'un même run. FRASER et FRASER2 font partie du workflow DROP (Yépez et al., 2021).

DROP permet à la fois la détection d'évènements affectant l'épissage (par FRASER2), la détection des différentiels d'expression (OUTRIDER) et l'étude de l'expression mono-allélique. DROP inclut aussi un module de contrôle qualité (QC), d'identitovigilance en comparant les variants identifiés en RNA-Seq et DNA-Seq (VCFs DNA-Seq transmis en plus des BAMs RNA-Seq) et de priorisation des évènements en fonction du phénotype (utilisation des termes HPO).

DROP permet aussi une gestion intelligente des échantillons, à savoir qu'il peut considérer des groupes différents (tissus ou maladies par exemple), un échantillon pouvant appartenir à différents groupes. Chaque échantillon n'est analysé qu'une fois par chaque module.

OUTRIDER nécessite 50-60 échantillons et FRASER en demande une trentaine mais une cinquantaine d'échantillons permettra d'avoir de meilleurs résultats statistiques. Le dépôt github de DROP contient des matrices de comptage issues d'un certain nombre de protocoles et tissus, permettant de simuler une cohorte pour débuter (matrices de comptage de lectures orientées ou non, pour les assemblages hg19 et hg38, mais non adaptées pour RNA-Seq panel).

Pour utiliser DROP, les auteurs conseillent de fournir un BAM issu de STAR en mode 2-pass basic ainsi qu'un index de ce fichier BAM. Il faut aussi un VCF issu de séquençage génomique, un fichier FASTA du génome de référence et un fichier d'annotation, typiquement au format GTF issu de gencode.

4.3.4 SpliceLauncher

SpliceLauncher (Leman et al., 2020) présente certains avantages pratiques. On peut l'utiliser en sortie de séquenceur pour **analyser tous les échantillons du run** (contrairement à FRASER, SpliceLauncher est efficace sur un nombre restreint d'échantillons). L'outil est dimensionné pour du **RNA-**

Seq réalisé sur un panel de gène. Il répond ainsi à une demande relativement commune des laboratoires de diagnostic qui utilisent ce type de design pour leur facilité de mise en œuvre notamment (réutilisation de sondes de captures pré-existantes pour le séquençage génomique, possibilité d'utilisation de séquenceurs de moyenne capacité type MiSeq, volume de données moins important…). SpliceLauncher est distribué sous forme de conteneur, ce qui permet un **déploiement et une utilisation facilitée et standardisée**. SpliceLauncher est un outil capable de détecter les rétentions partielles d'introns notamment grâce à la détection de nouveaux sites accepteurs ou donneurs d'épissage dans un intron, mais ne permet pas la détection de rétentions complètes d'introns.

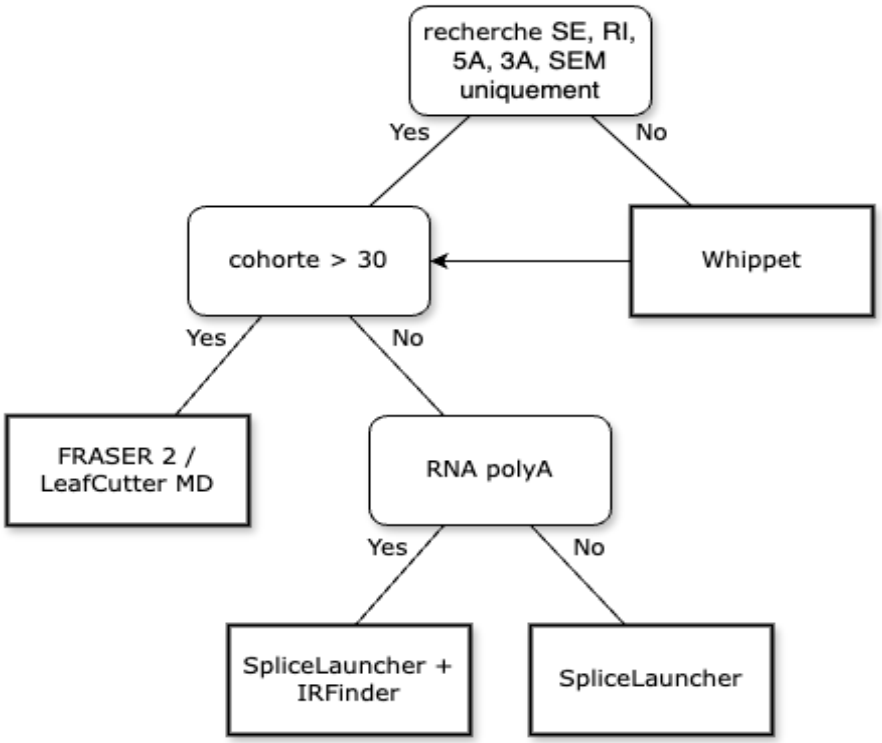

Figure 10 : Exemple d'arbre décisionnel de choix d'outils

La **Figure 10** propose un arbre décisionnel pour le choix des outils en fonction de la question et du design. D'autres approches comme SAMI pourraient intégrer cet arbre décisionnel.

4.3.5 SAMI

SAMI (Mareschal et al., 2025) est un outil qui propose de **détecter des évènements d'épissages aberrants et fusions de gènes à partir d'un alignement STAR** (Dobin et al., 2013) 2-pass en comparant les lectures de jonctions aux transcrits de l'annotation. Bien qu'un filtrage empirique sur la récurrence au sein du run soit toujours possible, cette approche basée sur l'annotation en fait un outil sensible sur des panels de gènes mais peu spécifique en transcriptome complet où les lacunes de l'annotation sont plus évidentes. SAMI propose une analyse intégrale en sortie de séquenceur (fichiers FASTQ bruts), une portabilité assurée par Nextflow et Singularity, une sélection de QC rassemblés par MultiQC (Ewels et al., 2016), une représentation graphique claire et complète de l'épissage observé dans chaque gène et une prise en charge des UMIs qui en font un outil complet.

4.3.6 Considérations autour du NMD

L'analyse des données RNA-Seq, quel que soit la méthode ou le design choisi, **nécessite une vigilance particulière afin de détecter l'absence de visibilité des défauts d'épissage** liés à la physiologie de la cellule ou du tissu étudié.

Un exemple est le NMD (« Nonsense Mediated mRNA Decay » ou système de dégradation des ARNm possédant un codon Stop). Wang et al., 2023 (Wang et al., 2023) cite l'exemple d'une étude (Mohammadi et al., 2019) qui réanalysait les échantillons de l'étude de Cummings (une cinquantaine de patients, (Cummings et al., 2017)). Ils ont résolu un cas en identifiant un variant dans le gène *DES* qui avait pour conséquence l'exonisation de 116 pb (inclusion d'un pseudo-exon) mais présent seulement dans 1% des lectures de la région du fait de l'activation du NMD sur ces transcrits. L'évènement était donc filtré dans l'étude originale. Les lectures longues proposent une solution élégante à ce problème en permettant une analyse allèle spécifique (par la détermination de la phase des variants environnants) (Glinos et al., 2022). Une autre possibilité est d'inhiber le NMD sur des cellules en culture (fibroblastes, lymphocytes, …). Enfin, une analyse quantitative performante peut permettre de pallier l'absence de visualisation directe de l'effet sur l'épissage en mettant en évidence la baisse du nombre de transcrits (voir partie sur l'expression).

4.4 Discussion

Plusieurs méthodes statistiques ont été récemment développées pour réaliser des analyses d'épissage différentiel dans le but de mettre en évidence des évènements « outliers » comparativement à des populations d'échantillons contrôles (LeafCutter MD ou FRASER2). Ces analyses sont très dépendantes de l'étape d'alignement préalable, et l'alignement peut par exemple être altéré par la présence de SNPs proches des régions 5' ou 3' d'épissage ce qui induit un bruit de fond non négligeable lors de l'analyse biologique. Afin d'augmenter la spécificité de l'analyse, autrement dit le fait qu'une anomalie d'épissage soit effectivement due à la présence d'un variant génomique, il est recommandé de croiser les données d'épissage aux données génomiques et de prioriser les évènements supportés par au moins un variant rare.

La validation d'un effet passe encore par une visualisation des évènements sur un navigateur de génome type IGV (« Integrative Genomics Viewer ») et il n'est pas rare que l'évènement appelé soit difficilement visualisable ou interprétable par ce biais. Cette visualisation est dépendante de la méthode RNA-Seq (poly-A ou capture), par exemple pour la visualisation des rétentions introniques, et nécessite une certaine habitude des opérateurs pour l'interprétation. Des formations à cette interprétation sont recommandées via les différents réseaux professionnels en bioinformatique et en génétique moléculaire.

5 Le séquençage de l'ARN pour l'analyse de la quantification des transcrits

Le séquençage des ARN pour l'analyse de l'expression génique est une technique permettant de **quantifier les niveaux d'expression des gènes**. La quantification des transcrits peut se diviser en 2 grandes sous-catégories selon la question posée : **l'analyse différentielle de l'expression**, pour comparer des cohortes selon différentes conditions, et l'**analyse d'une expression aberrante** pour détecter une valeur aberrante (« outlier », échantillon, gène, transcrit, voie de signalisation...) au sein d'une cohorte. Les méthodes et pipelines qui en découlent sont différents. La quantification des transcrits nécessite l'obtention d'un grand nombre de réplicats, sujets d'intérêt ou témoins, pour obtenir un résultat fiable et significatif. De manière générale, l'analyse de l'expression est particulièrement sensible aux effets « *batch* » (variabilité non désirée) et autres sources de variabilités techniques. Une particularité de cette analyse réside dans le fait qu'un gène peut donner différents transcrits, qu'il

conviendrait de quantifier séparément. Il est possible de s'affranchir de cette particularité en faisant une analyse par gène et non par transcrit afin d'éviter des biais liés au séquençage en lectures courtes.

5.1 Contrôles qualité spécifiques

Comme pour tout séquençage haut débit, un des points de contrôle à réaliser en priorité est d'observer la distribution du **nombre de lectures par patient à l'issue du run** (dépendant de la méthode utilisée, transcriptome complet ou ciblé, du nombre de patients, du séquenceur, de la chimie utilisée…). Plus cette distribution sera homogène d'un run à l'autre, plus les résultats pourront être comparables. Le pourcentage de lectures alignées pour chaque patient peut aussi servir de contrôle.

Il est intéressant de **vérifier la présence d'un effet « batch »** par rapport aux autres runs déjà réalisés, en utilisant par exemple une Analyse en Composantes Principales (ACP). Il s'agit notamment d'un point de contrôle essentiel avant de poursuivre les analyses secondaires. Après vérification de l'absence d'un effet « batch », on peut alors déterminer statistiquement la présence d'expressions aberrantes chez un individu par rapport à la cohorte notamment en utilisant des méthodes statistiques comme Outrider. Il est aussi possible de contrôler le **nombre d'appels (« calls ») par patient en sortie d'Outrider** : un nombre aberrant par rapport au reste de la cohorte permet de mettre en évidence certains biais ou erreur d'échantillons.

5.2 Obtention d'une matrice de comptage

Réaliser une matrice de comptage est une étape nécessaire et importante pour pouvoir ensuite analyser les différentiels d'expression. Elle est le résultat du comptage des lectures, attribués aux différents transcrits ou gènes. Pour cette étape, il faut choisir entre un comptage classique, ou un pseudo-comptage. Le pseudo-comptage sert notamment à lisser *a minima* les profondeurs de lecture pour les rendre comparables et exploitables dans une analyse différentielle. Il utilise une méthode plus rapide et également moins coûteuse en stockage.

5.2.1 Kallisto et Salmon

Kallisto (Bray et al., 2016) et Salmon (Patro et al., 2017) sont des **outils de pseudo-alignement rapide** pour l'analyse des données de séquençage, qui fonctionnent sans aligner chaque lecture sur un génome de référence. Kallisto utilise uniquement une collection de séquences transcriptomiques pour construire un graphe de De Bruijn (T-DBG) basé sur les k-mers, quand Salmon construit un index basé sur les k-mers du transcriptome et une structure de recherche associée. Les lectures sont décomposées en k-mers pour identifier leurs transcrits potentiels grâce à des classes d'équivalence, qui regroupent les lectures compatibles avec les mêmes transcrits. L'algorithme d'espérance-maximisation (EM) (Kallisto) ou un algorithme proche (Salmon) sont appliqués à ces classes pour quantifier les transcrits tout en maximisant la vraisemblance. Cette approche optimise la rapidité et l'efficacité, en évitant les étapes coûteuses d'alignement traditionnel.

5.2.2 RSEM

RSEM (RNA-Seq by Expectation-Maximization) (Li and Dewey, 2011) est un **outil de quantification de l'expression génique qui utilise une approche basée sur l'alignement des lectures sur un transcriptome**. Les lectures sont d'abord alignées, généralement à l'aide d'un outil comme Bowtie, pour déterminer leur correspondance précise avec les transcrits. RSEM calcule ensuite les probabilités qu'une lecture provienne de chaque transcrit possible, en tenant compte des erreurs de séquençage et des ambiguïtés dues aux transcrits similaires. Un algorithme d'espérance-maximisation (EM) est là encore utilisé pour estimer les abondances relatives des transcrits, en maximisant la vraisemblance

des données observées. Bien qu'il soit plus lent que les outils de pseudo-alignement, RSEM est réputé pour sa précision grâce à son approche rigoureuse et ses modèles statistiques robustes. RSEM reste cependant beaucoup plus coûteux en temps de calcul.

5.2.3 HTSeq-count

HTSeq-count (Putri et al., 2022) fonctionne en **comptant le nombre de lectures alignées** sur des régions spécifiques du génome ou du transcriptome, basées sur une annotation génomique fournie (fichier GTF ou GFF). Les lectures doivent d'abord être alignées à l'aide d'un outil comme Bowtie ou HISAT2, car HTSeq-count nécessite les positions exactes des lectures. Il associe chaque lecture à une entité génomique (gène ou exon) selon un mode choisi (par exemple, union, intersection stricte, ou intersection non vide pour gérer les chevauchements et les ambiguïtés dans l'attribution des lectures), et les regroupe pour obtenir un comptage final par gène. Bien que moins sophistiqué que d'autres outils pour la gestion des ambiguïtés ou des multi-alignements, HTSeq-count est apprécié pour sa simplicité et son adéquation à des analyses précises et ciblées.

5.2.4 SummarizeOverlaps

Il s'agit de l'**algorithme de comptage utilisé dans le package DROP**. Il propose différents modes de comptage, s'inspirant des modes de comptage de HTseq count (intersection stricte, intersection non vide et union). SummarizeOverlaps est une fonction de la bibliothèque Bioconductor GenomicAlignments, utilisée pour quantifier l'expression génique à partir de données RNA-Seq. Elle prend en entrée des fichiers d'alignement (au format BAM) et une annotation génomique (GRanges ou TxDb) pour compter le nombre de lectures qui se chevauchent avec des entités génomiques, comme des gènes ou des exons.

SummarizeOverlaps gère également les lectures multi-alignées et les structures complexes grâce à une intégration fluide dans l'écosystème R. SummarizeOverlaps est particulièrement utile dans des flux de travail R pour des analyses flexibles et reproductibles d'expression génique.

Cet outil est gourmand en mémoire vive. La **Figure 11** représentent les différents modes et ce qu'ils sont capables de détecter.

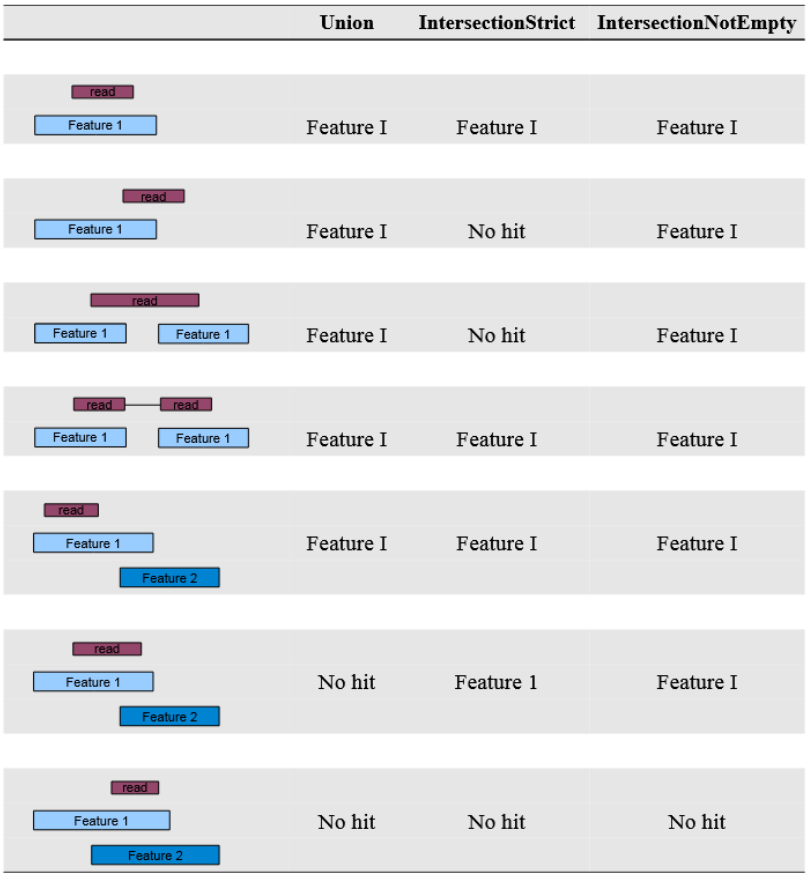

Figure 11 : Différents modes de SummarizeOverlaps, et types de détections associés. Source : Couting reads with SummarizeOverlaps de Valérie Obenchain (https://bioconductor.org/packages/release/bioc/vignettes/GenomicAlignments/inst/doc/summarizeOverlaps.pdf)

5.3 Détection d'une expression aberrante par rapport aux autres patients

5.3.1 OUTRIDER

OUTRIDER (Brechtmann et al., 2018) est un outil statistique **intégré au package DROP**, conçu pour **détecter des valeurs aberrantes d'expression génique** dans les données RNA-Seq. Il est recommandé d'analyser des **cohortes d'au moins 50 à 60 échantillons pour obtenir des résultats statistiquement significatifs**. Si la taille de la cohorte est insuffisante, des échantillons externes, comme ceux de GTEx, peuvent être ajoutés, à condition de minimiser les effets de « *batch* » (par exemple, via une robotisation des processus).

Avant d'utiliser OUTRIDER, il est nécessaire de procéder à l'alignement des lectures et à la génération de matrices de comptage. **Les matrices peuvent être enrichies progressivement en gérant soigneusement les effets « batch »**. OUTRIDER permet d'exclure certains gènes de la correction du FDR (« False Discovery Rate », méthode statistique utilisée pour corriger les tests multiples), comme des gènes OMIM ou ceux portant des variants d'épissage d'intérêt, ce qui renforce la détection des valeurs aberrantes sur ces gènes spécifiques.

L'outil propose des options pour exclure des échantillons problématiques via le « SampleExclusionMask » (réplicats techniques ou biologiques) et analyse les facteurs confondants (« batch », sexe, âge) à l'aide de méthodes similaires à DESeq2. Les gènes faiblement exprimés (« Fragments Per Kilobase Million », FPKM faibles) sont automatiquement retirés de l'analyse. Les résultats sont présentés sous forme de graphiques (volcanoplot, Qplot) et de tableaux, facilitant l'interprétation. OUTRIDER repose sur une combinaison de Snakemake et Conda pour simplifier son déploiement et son utilisation.

5.3.2 ABEILLE

ABEILLE (« ABerrant Expression Identification empLoying machine LEarning from sequencing data ») (Labory et al., 2022) est une **méthode basée sur un autoencodeur variationnel** (VAE) pour l'identification des gènes ayant une expression aberrante (AGE) à partir de l'analyse de données RNA-Seq sans besoin de réplicats ou groupe contrôle. ABEILLE combine l'utilisation d'un VAE et d'un arbre de décision pour classer les gènes comme AGE ou non AGE. La publication montre les performances d'ABEILLE sur un ensemble de données semi-synthétiques et expérimentales, avec des **résultats au moins comparables à OUTRIDER. ABEILLE présente l'intérêt de rester performant sur des petites cohortes.**

5.3.3 Conclusion et recommandations

Il est conseillé, lorsque le temps de calcul n'est pas limitant, de **privilégier un comptage classique** qui sera plus précis que le pseudo-comptage. Concernant OUTRIDER, dans le cas d'un nombre d'échantillons inférieur à 100 il est conseillé de filtrer sur la p-valeur plutôt que sur la p-valeur ajustée, sous peine d'une sensibilité très diminuée. En retour d'expérience du CHU de Nantes, avec un nombre d'échantillons atteignant les 200, l'usage de la p-valeur ajustée s'avère très spécifique et pour la majorité des évènements détectés, un variant génomique causal est identifié. La sensibilité devra cependant être évaluée dans de futurs tests. Le point fort de cette approche est l'identification de co-variations permettant de normaliser des biais aussi bien liés à la technique qu'à la biologie. Exemple : il est possible de détecter la surexpression des gènes qui échappent partiellement à l'inactivation de l'X d'un individu XXY, dans une cohorte mélangée d'individus XX et XY. La normalisation permet de « repérer » cet individu comme ayant un profil proche des « XY » et donc un profil anormal d'expression de certains gènes sur l'X alors même qu'il possède une expression similaire aux individus XX.

Une approche d'expression différentielle performante est la clé vers l'exploitation des données transcriptomiques après une analyse génomique négative. En effet, elle permet une détection à la fois des anomalies d'épissage prises en charge par le NMD, mais aussi de détecter des variants dans des régions régulatrices des gènes et ouvre également la voie à l'identification des conséquences secondaires ou « signatures transcriptomiques ».

5.4 Recherche d'une signature

L'analyse classique de l'expression différentielle de gènes en transcriptomique comprend 3 étapes : quantification (salmon, RSEM…), normalisation des données (DESEQ2, TMM) et l'identification des gènes différentiellement exprimés. **En diagnostic, il pourrait être possible de classer les patients selon leur signature transcriptomique** (atteints et non atteints) voire pour les pathologies avec différents degrés de sévérité d'avoir une approche pronostique en fonction de la signature. Il faut donc avoir un profil pour chaque degré de sévérité de la pathologie concernée (découpage des degrés de sévérité) ce qui peut permettre aussi de classer les variants et leurs conséquences phénotypiques. Des méthodes telles que le random forest ou SVM pourrait permettre de créer ces classifieurs. Cependant, pour obtenir une signature robuste, il faut une centaine d'échantillons. Il est possible de faire du 3' RNA-Seq pour réduire les coûts (séquençage uniquement de la partie c-terminale du transcrit, qui permet évaluer l'expression du transcrit, **cf paragraphe 2.3**).

5.5 Annotation des résultats

Il n'existe pas de méthode consensus pour l'annotation et l'interprétation d'évènements d'expression aberrante. Toutefois, l'utilisation de bases de données publiques peut améliorer la priorisation de ces événements. Par exemple, l'intégration des termes HPO (« Human Phenotype Ontology ») permet de relier les anomalies d'expression à des phénotypes cliniques. De plus, les fréquences

alléliques issues de gnomAD et les scores de contrainte tels que pLI aident à évaluer la probabilité qu'un gène soit intolérant à la perte de fonction. Pour les maladies rares, la présence d'un gène dans PanelApp peut renforcer sa pertinence, et la consultation des données OMIM ou des scores de dosage de ClinGen permet d'affiner l'interprétation en considérant les mécanismes de dosage génique.

5.6 Application en cancérologie : Identification des cancers primitifs inconnus

Le **siège primitif désigne l'emplacement initial où les cellules subissent leur première transformation maligne** et où le cancer prend naissance. Toutefois, ces cellules cancéreuses peuvent se détacher et se propager vers d'autres parties du corps, un phénomène connu sous le nom de métastase. **Le cancer primitif inconnu** est diagnostiqué lorsqu'un cancer est détecté dans l'organisme, mais que **son site d'origine reste indéterminé**. Il est également appelé cancer primitif occulte ou cancer d'origine primitive inconnue. **Déterminer l'origine d'un cancer est essentiel**, car le choix du traitement et son efficacité en dépendent directement. Une approche prometteuse pour cette identification repose sur l'analyse du profil transcriptomique, qui est propre à chaque type de cancer. En effet, **les signatures d'expression génique peuvent fournir des indices précieux sur l'origine tumorale**. Le principe consiste à établir une base de données transcriptomique regroupant différents types de cancers, puis à comparer le profil transcriptomique d'intérêt à cette base afin de déterminer de quel type de cancer il se rapproche le plus.

Cependant, cette méthode reste encore **peu répandue dans les laboratoires, et aucun consensus** n'existe actuellement sur les outils et protocoles à utiliser. À ce stade, les retours des laboratoires français ayant initié des expérimentations dans ce domaine permettent de mettre en évidence les points suivants :

- Concernant la méthodologie, l'outil Kallisto est privilégié pour la quantification des transcrits, offrant une performance comparable à Salmon ou HTSeq-count, tout en étant plus rapide. Pour l'analyse des données, l'utilisation de méthodes de clustering consensus est recommandée plutôt que les méthodes de réduction de dimension type UMAP ou tSNE, sauf pour certains cas spécifiques (comme les mésothéliomes par exemple).
- Concernant la gestion des données et leur qualité, une normalisation par quantile, ainsi que des contrôles par ACP pour détecter d'éventuelles valeurs aberrantes, doivent être implémentés lors de l'ajout de nouveaux échantillons dans la base de données. Les valeurs aberrantes doivent être identifiées et éliminées après consensus entre les décideurs. De plus, il est essentiel de valider les échantillons en s'assurant qu'ils contiennent un nombre suffisant de lectures. Il est important de noter que pour une analyse plus ciblée, les données doivent être corrigées afin d'éliminer les signatures liées aux types d'organes, permettant ainsi de se concentrer uniquement sur les signatures tumorales.
- Il est important d'éviter les modifications des sondes ou des versions génomiques afin de garantir la cohérence des analyses.

Enfin, il sera nécessaire pour assurer une implémentation rigoureuse et standardisée de cette méthode dans les laboratoires d'organiser la collaboration entre les équipes pour évaluer les performances des algorithmes et améliorer les performances des analyses. **Le partage des connaissances et des outils permettra ainsi d'optimiser les méthodologies employées** et d'accroître la robustesse des résultats.

6 Le séquençage du transcriptome pour la détection de transcrits de fusion

Les **transcrits de fusions, résultants de réarrangements chromosomiques** dans lesquels deux gènes distincts fusionnent pour former un gène hybride, constituent une classe importante des altérations somatiques des cancers (Mitelman et al., 2007). **Les transcrits de fusion sont des marqueurs clés en oncohématologie**, notamment avec la fusion *BCR-ABL1* dans la leucémie myéloïde chronique (LMC). Les inhibiteurs de tyrosine kinase (ITK) ciblant *BCR-ABL1* ont révolutionné le traitement de la LMC. D'autres fusions, telles que *PML-RARA* dans la leucémie aiguë promyélocytaire (LAP), sont également ciblées. **Les transcrits de fusion** ont non seulement un rôle diagnostique et pronostique, mais ils **guident également la thérapie**. Les transcrits de fusion jouent aussi un rôle croissant dans les tumeurs solides. Des fusions impliquant *ALK*, *ROS1*, *RET*, *NTRK*, *FGFR* et *MET* sont maintenant considérées comme des cibles thérapeutiques actionnables. Des inhibiteurs ciblant *FGFR2* sont également utilisés dans la prise en charge de certaines tumeurs présentant des fusions impliquant ce gène. Le ciblage de ces transcrits de fusion a permis de développer des approches de thérapie tumorale, où le traitement est guidé par les altérations génomiques plutôt que par le type de tissu (*RET*, *NTRK…)* (Lu et al., 2022; Subbiah et al., 2024).

Environ 40 % des sarcomes sont caractérisés par la présence de transcrits de fusion (Wachtel et al., 2024a). Ces protéines peuvent être des facteurs de transcription, des régulateurs épigénétiques ou des kinases. Les transcrits de fusions *SS18-SSX* dans le sarcome synovial, *PAX3-FOXO1* dans le rhabdomyosarcome alvéolaire, et *NAB2-STAT6* dans les tumeurs fibreuses solitaires sont des exemples clés. **Les sarcomes sont maintenant classifiés en fonction de leurs altérations moléculaires** et plus uniquement sur des critères histopathologiques. **Les transcrits de fusion agissent comme des oncogènes pilotes et leur inhibition représente une approche thérapeutique prometteuse**. La grande fréquence de fusions impliquant le gène *EWSR1* dans de nombreux sarcomes démontre que l'activité de cette protéine coopère avec différents partenaires de fusion (Wachtel et al., 2024b).

Les transcrits de fusion sont également importants dans les cancers pédiatriques. Le sarcome d'Ewing, par exemple, est caractérisé par des fusions impliquant *EWS*. Dans le neuroblastome, les fusions avec *ALK* sont un enjeu thérapeutique. D'autres fusions, notamment *ETV6-NTRK3*, *LMNA-NTRK1*, et *BRAF* sont également impliquées dans des sarcomes pédiatriques. La découverte de fusions récurrentes dans les cancers pédiatriques a conduit à des essais cliniques utilisant des inhibiteurs ciblés (Oliver et al., 2020).

La pertinence de l'analyse des transcrits de fusion dans le diagnostic de maladies rares héréditaires a été établie récemment. Des études ont montré que le RNA-Seq permet de détecter des fusions causales non identifiées par le séquençage de l'ADN (Oliver et al., 2019) avec par exemple *ATM-SLC35F2* et *SAMD12-EXT1*. Cependant, la présence de ces fusions, même pathogènes, à faible niveau dans des tissus normaux doit être prise en compte.

Comparé aux autres méthodes classiquement utilisées pour la détection des fusions (FISH, RT-PCR, ddPCR), le séquençage de 2ème génération est la **méthode la plus polyvalente, capable de détecter à la fois les transcrits de fusion connus et les nouveaux. Le séquençage de l'ARN est particulièrement adapté à la détection des transcrits de fusion et permet une analyse non biaisée du transcriptome.** Le séquençage ciblé de l'ADN (panel) est également utilisé pour détecter des transcrits de fusion, en particulier les fusions impliquant les kinases. Le séquençage du génome entier (« WGS, Whole Genome Sequencing ») peut potentiellement identifier tous les événements de fusion génomique, mais il est limité par son coût et sa complexité. Le séquençage de l'exome entier (« WES, Whole Exome Sequencing ») est plus économique, mais il est limité pour la détection des transcrits de fusion dont les points de cassure se situent dans les introns. **Dans tous les cas, il est important de déterminer si la fusion aboutit à un cadre de lecture ouvert (« ORF, Open Reading Frame »)**, ce qui est essentiel pour la fonctionnalité de la protéine de fusion.

Le développement de bases de données exhaustives est essentiel pour l'interprétation des transcrits de fusion. Les bases de données telles que Mitelman, COSMIC, ChimerDB, et ChiTaRS répertorient les fusions et fournissent des informations sur leur fréquence, leur association avec les cancers et les possibilités de thérapie ciblée. Cependant, l'intégration des données de différentes sources et la standardisation des analyses bioinformatiques sont nécessaires.

La priorisation des résultats se base sur l'utilisation de filtres dont le but comme souvent est de réduire les listes de faux-positifs sans affecter les taux de faux-négatifs.

- Les **fusions récurrentes** dans un type de cancer donné sont généralement plus pertinentes. La fréquence peut être déterminée à partir des bases de données existantes.
- L'identification de **fusions pathogènes** dans les tissus normaux souligne la nécessité d'un filtrage rigoureux des résultats de séquençage. Les **fusions polymorphes**, qui sont communes dans la population générale, peuvent être distinguées des fusions pathogènes par leur fréquence et leur niveau d'expression.
- **La fusion doit entraîner une altération fonctionnelle de la protéine**, comme une activation constitutive de kinases ou une modification de la régulation de la transcription. L'existence d'un ORF est un critère important.
- **L'analyse de la fonctionnalité des protéines de fusion est cruciale** pour évaluer leur pertinence clinique. Il faut s'assurer que la protéine de fusion a une activité oncogénique, que l'ORF est bien conservé, et qu'elle peut être ciblée par des médicaments spécifiques.
- **Les fusions doivent être interprétées dans le contexte clinique du patient**, en tenant compte de son type de cancer, de son stade et de ses antécédents.

Les gènes de fusion sont devenus des acteurs majeurs en oncologie, en hématologie, dans les sarcomes, les cancers pédiatriques et les maladies génétiques rares. Leur découverte et leur caractérisation ont considérablement amélioré le diagnostic, le pronostic et le traitement des cancers et de certaines maladies rares. Le couple séquençage de seconde génération et bio-informatique a permis de détecter de nouvelles fusions à l'échelle du génome. Le développement de bases de données exhaustives et l'analyse fonctionnelle des protéines de fusion sont essentiels pour une interprétation plus précise des données et le développement de thérapies ciblées toujours plus efficaces. En conclusion, **les transcrits de fusion représentent un domaine de recherche en pleine évolution, avec un potentiel considérable pour améliorer la prise en charge des patients**.

6.1 Analyse et contrôle qualité

La combinaison de plusieurs outils de détection de fusions est indispensable pour rendre une liste de fusions candidates. La diversité des algorithmes peut permettre dans un contexte diagnostique, en rattrapant d'éventuelles faiblesses entre les outils, d'améliorer l'exhaustivité de détection. La détection, avec plusieurs outils, d'une même fusion permet de lui donner du poids (ex : la confiance en une fusion détectée avec un faible nombre de lectures peut être augmentée lorsque celle-ci est détectée par plusieurs algorithmes différents). La préparation des librairies, le séquençage et l'analyse dans les mêmes conditions d'un groupe d'échantillons contrôles est utile pour valider les performances des outils sélectionnés mais également pour identifier les fusions faussement positives qui pourront être écartées lors des analyses qui suivront. Afin de valider une fusion ou son absence, **il est nécessaire de valider la librairie de séquences en amont et de s'assurer de la bonne couverture des gènes et exons portant l'altération**.

6.2 Détection des transcrits de fusion

Un certain nombre d'outils bio-informatiques ont été développés pour identifier les transcrits de fusion candidats à partir de données RNA-Seq. Ces méthodes se divisent en deux catégories conceptuelles :

- **Approches basées sur l'alignement** des lectures RNA-Seq sur le génome de référence pour identifier les lectures qui s'alignent de manière discordante, suggérant des réarrangements

chromosomiques (**Figure 12**). Les lectures discordantes peuvent inclure des lectures chimériques (« split reads ») qui chevauchent directement la jonction de fusion, ou des paires de lectures discordantes (« discordant paired ends reads » ou encore « spanning reads ») où chaque lecture s'aligne de part et d'autre de la jonction sans la chevaucher directement.

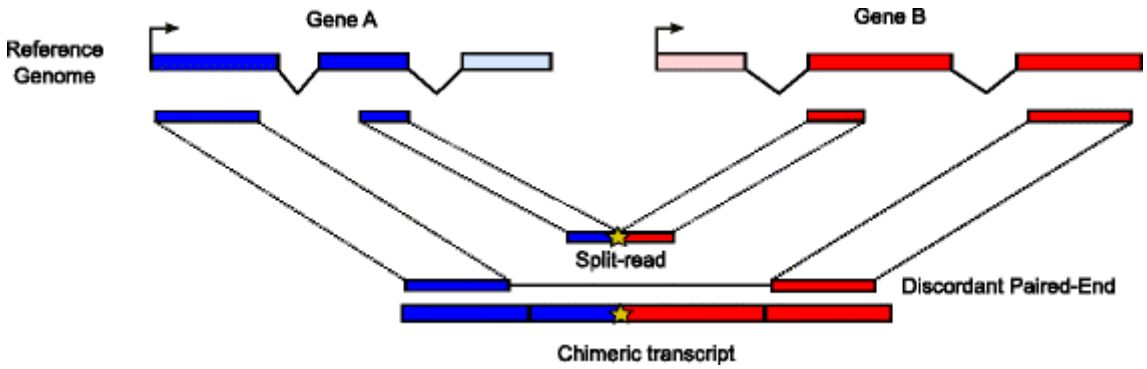

Figure 12 : Transcrit chimérique (en bas) composé d'exons de deux gènes, A et B (en bleu et rouge, en haut) soutenu par deux types de lectures : une lecture « split-read » et une lecture « discordante paired ends » ou « spanning read ». Ces lectures sont représentées à la fois alignées sur le génome (milieu-haut) et sur le transcriptome (milieu-bas). La position de la jonction chimérique (point de cassure) dans le transcriptome est marquée par une étoile jaune, visible dans la « split-read » et dans le transcrit chimérique (Rodríguez-Martín et al., 2017).

- **Approches basées sur l'assemblage** direct des lectures en séquences de transcrits plus longues, puis par identification des transcrits chimériques cohérents avec des réarrangements chromosomiques. Cette approche peut être particulièrement utile pour détecter des transcrits de fusion complexes ou inconnus.

Les preuves soutenant les fusions prédites sont généralement mesurées par le nombre de lectures chimériques et le nombre de paires de lectures discordantes (**Figure 13**) (Haas et al., 2019).

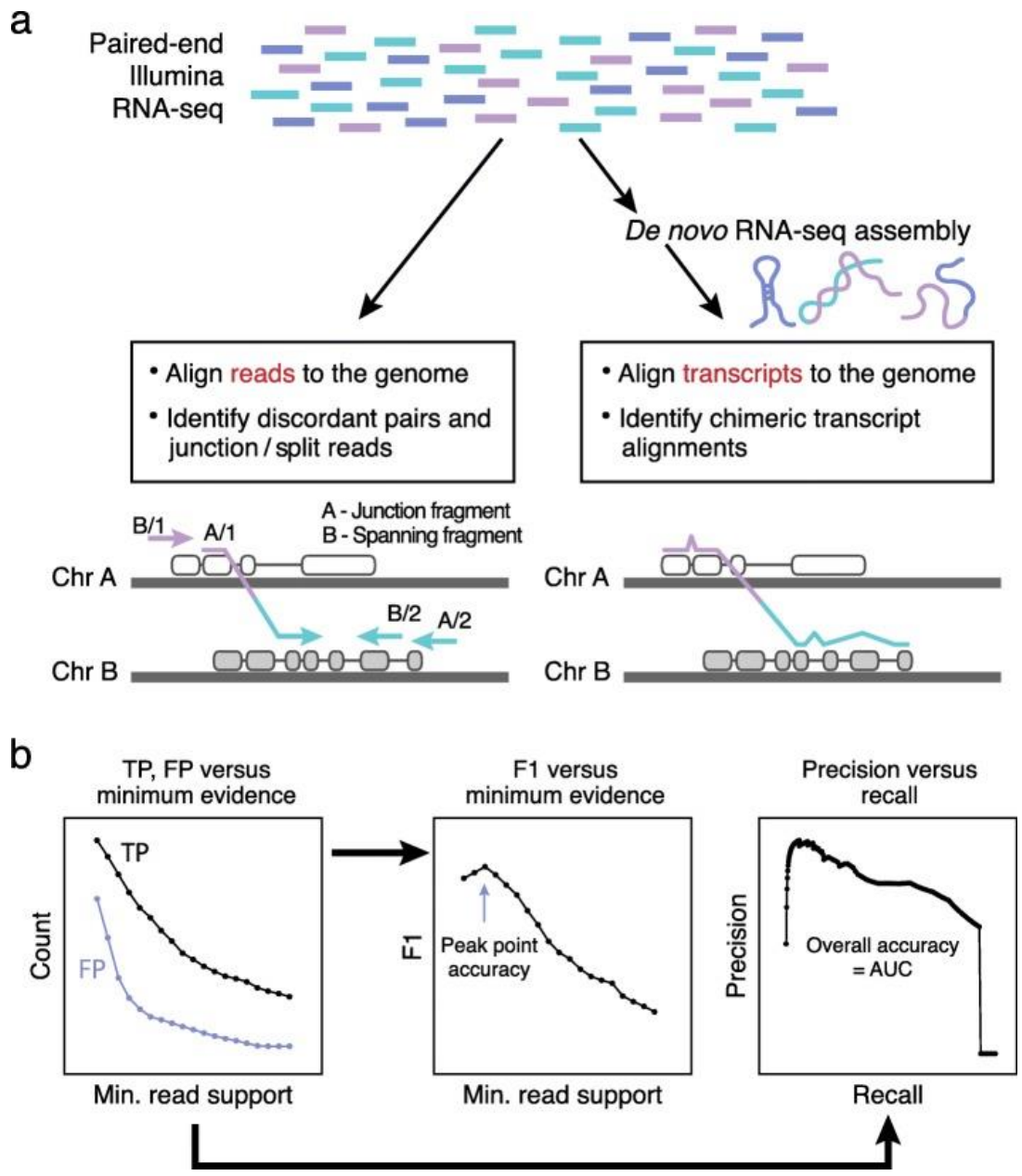

Figure 13 : Méthodes de prédiction de transcrits de fusion et d'évaluation de leur précision. A) Les deux modèles pour l'identification des transcrits de fusion comprennent (à gauche) l'alignement des lectures sur le génome et la capture des paires de lectures discordantes et des alignements de lectures chimériques et (à droite) la réalisation d'un assemblage de novo des transcrits suivi de l'identification des alignements de transcrits chimériques. **B)** Étant donné un ensemble bien défini de fusions réelles, les prédictions vraies et fausses positives sont comptabilisées en fonction du nombre minimum de lectures soutenant les fusions seuil). Le score F1 est calculé afin de déterminer le seuil qui produit une précision de prédiction maximale pour chaque méthode. De même, les valeurs de précision et de rappel sont calculées à chaque seuil, représentées sous la forme d'une courbe précision-rappel, et l'aire sous la courbe (AUC) est calculée comme mesure de la précision globale de la prédiction (Haas et al., 2019).

6.3 Outils de détection

6.3.1 Arriba

Il s'agit d'un outil développé pour la détection de transcrits de fusion dans le cadre de la recherche clinique, à partir de données RNA-Seq paired-ends ou single-ends (Uhrig et al., 2021). Arriba est aussi capable de détecter d'autres réarrangements structurels, tels que les duplications internes en tandem (ITD)**. Arriba utilise l'outil STAR pour aligner les séquences (Figure 14). Il se sert des alignements chimériques (« split reads et spanning reads ») pour établir une liste de fusions candidates**. Il utilise un modèle statistique pour filtrer cette liste de fusions candidates selon le nombre de lectures supportant la fusion par rapport au bruit de fond. L'outil applique ensuite des filtres positifs et filtres négatifs pour supprimer les artefacts et les faux positifs.

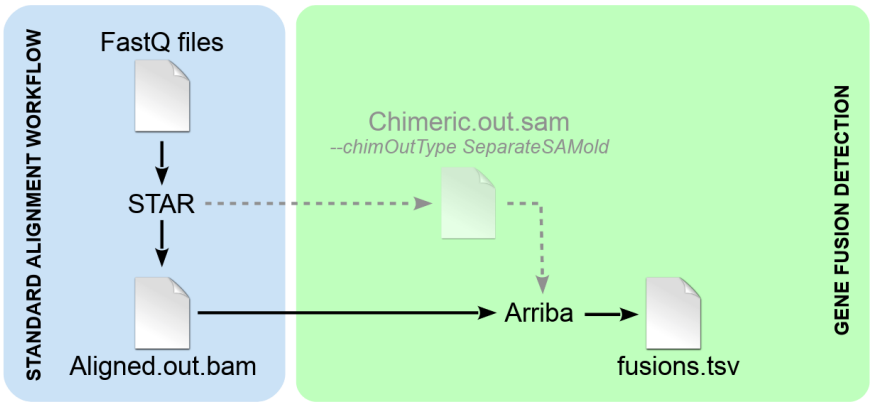

Figure 14 : Flux de travail de l'outil Arriba (Uhrig et al., 2021)

Arriba met également à disposition l'outil draw_fusions.R qui génère un graphique des fusions trouvées par échantillon. Pour chaque fusion sont représentés les transcrits impliqués, l'orientation de ces gènes, les exons inclus dans le transcrit de fusion, le point de cassure ainsi que la séquence environnante et les domaines protéiques retenus (**Figure 15)**.

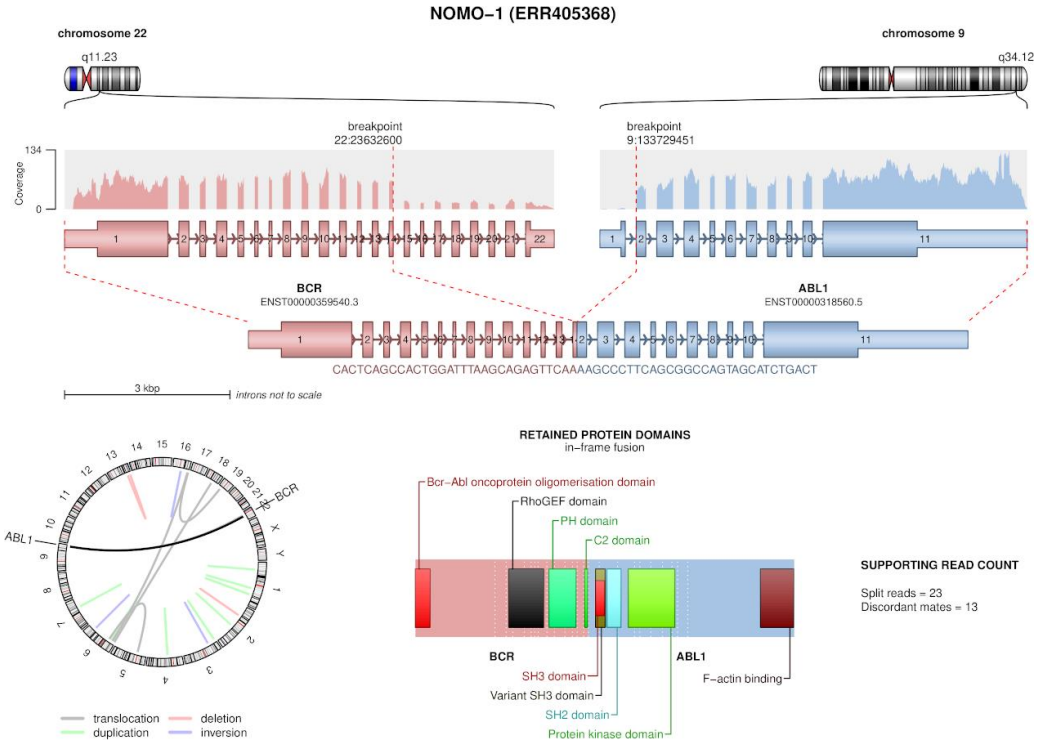

Figure 15 : Représentation de fusion de transcrits avec draw_fusions.R (Uhrig et al., 2021)

6.3.2 STAR-Fusion

Star-Fusion (Haas et al., 2019) **analyse les données RNA-Seq de préférence pairées**, pour identifier des fusions candidates à partir des alignements chimériques STAR. Les lectures discordantes et divisées sont alignées contre le transcrit de référence annoté. Les fusions candidates sont séparées des artefacts et faux positifs, puis un score leur est attribué selon le nombre de lectures qui les supportent (**Figure 16**).

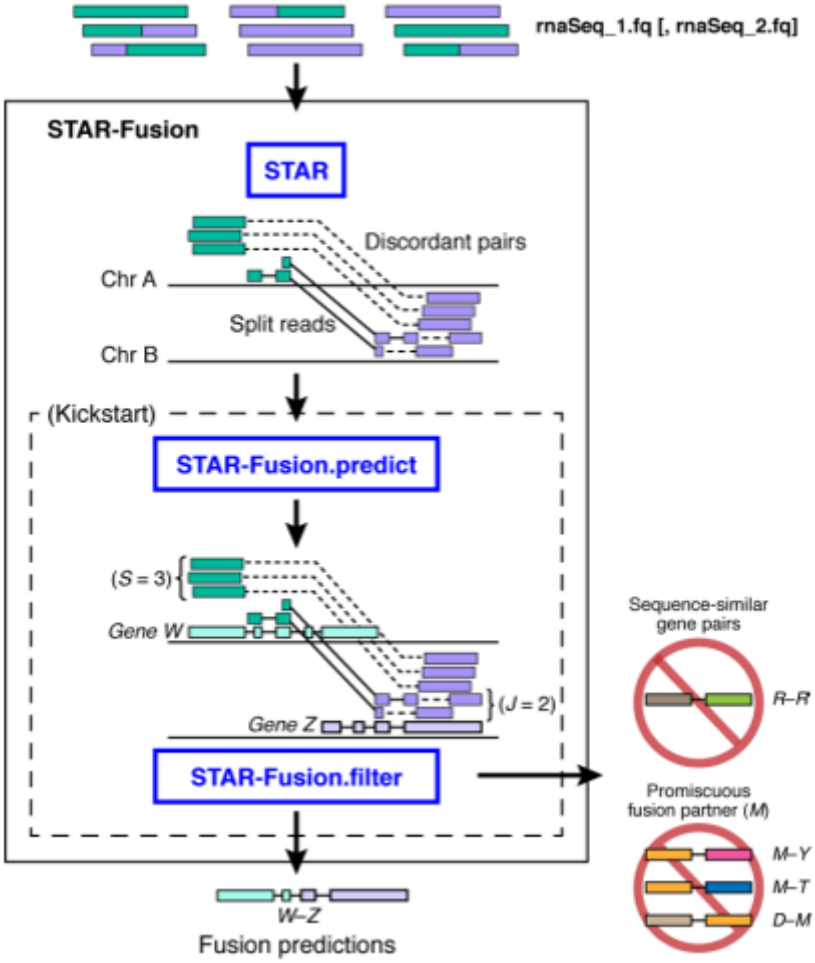

Figure 16 : Vue d'ensemble du pipeline STAR-Fusion. Les lectures courtes Illumina sont alignées sur le génome de référence par STAR. Les alignements discordants et chimériques sont identifiés, alignés, comparés à la structure du transcriptome de référence, filtrés pour enlever les artefacts, et notés selon l'abondance des lectures soutenant la fusion. Les candidats à la fusion contenant des paires de gènes similaires en termes de séquence ou des partenaires de fusion multiples au sein d'un même échantillon sont exclus car il s'agit probablement de faux positifs.

6.3.3 Fusioncatcher

FusionCatcher (Nicorici et al., 2014) **permet d'identifier les fusions somatiques dans les données RNA-Seq pairées**. Il effectue un prétraitement des données en filtrant les lectures de faible qualité, en supprimant les lectures s'alignant sur l'ARN ribosomal ou de transfert, l'ADN mitochondrial, les gènes HLA, ou les génomes connus de virus, phages ou bactéries. Les lectures sont ensuite alignées

sur le transcriptome référence « Ensembl » par l'outil d'alignement Bowtie **(Figure 17**). Une liste préliminaire de transcrits de fusions candidats est générée en recherchant des paires de gènes pour lesquels une lecture s'aligne sur les transcrits du gène A et sa lecture appariée sur ceux du gène B.

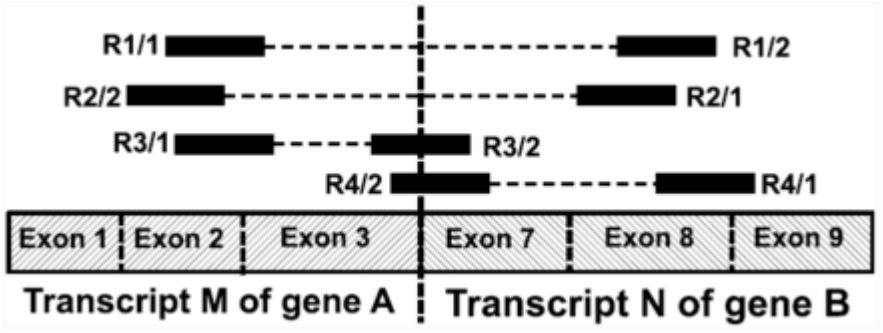

Figure 17 : Alignement Bowtie des lectures et de leurs partenaires qui confirment la fusion entre les gènes A et B avec une jonction de fusion entre l'exon 3 du gène A et l'exon 7 du gène B. Les paires de lectures (R1/1, R1/2) et (R2/1, R2/2) soutiennent la fusion et les lectures R3/2 et R4/2 soutiennent la jonction exon-exon de la fusion (Nicorici et al., 2014)

Les lectures non alignées sont ensuite réalignées contre une base de données construite avec toutes les combinaisons exon/exon pour chaque couple de gène candidat, à l'aide de Bowtie. Le but étant de réduire le nombre de lectures avant d'aligner contre une base de séquences gène/gène à l'aide des aligneurs STAR, BLAT et Bowtie2 pour la détection du point de cassure (**Figure 18**).

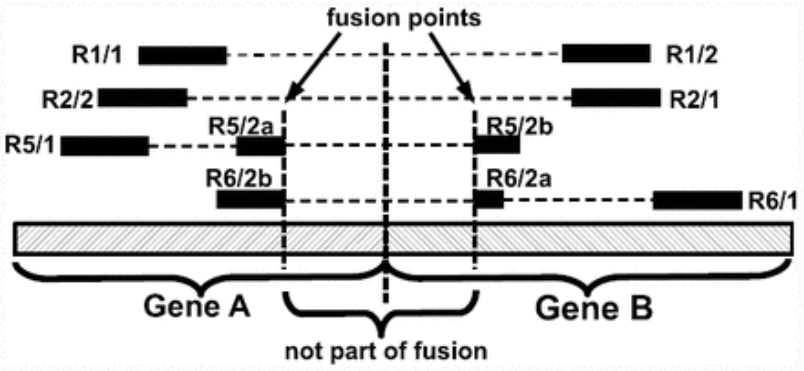

Figure 18 : Alignement des lectures et des partenaires qui soutiennent la fusion entre les gènes A et B. Les paires de lectures (R1/1, R1/2) et (R2/1, R2/2) soutiennent la fusion entre les gènes A et B et les lectures R5/2 et R6/2 soutiennent la jonction de la fusion. Les algorithmes BLAT, STAR et Bowtie2 divisent la lecture R5/2 en deux parties R5/2a et R5/2b et la lecture R6/2 en R6/2a et R6/2b. Les lectures R1/1, R1/2, R2/1, R2/2, R5/1 et R6/1 sont alignées par Bowtie sur le transcriptome et les lectures R5/2 et R6/2 sont alignées par BLAT, STAR et Bowtie2 sur les séquences des gènes (Nicorici et al., 2014).

Des critères biologiques et des bases de données connues sont utilisés pour éliminer les faux positifs potentiels. Cette approche vise à améliorer la précision de la détection des fusions tout en réduisant le nombre de faux positifs. La méthode par étapes permet de réduire la complexité computationnelle et d'augmenter la précision des résultats en excluant progressivement les lectures non pertinentes.

6.3.4 Autres outils

On pourra citer **Pizzly** (**pmelsted/pizzly: Fast fusion detection using kallisto**), qui utilise les pseudo-alignements générés par kallisto pour identifier des gènes de fusions, **Manta** (Chen et al., 2016), **FusionScan** (Kim et al., 2019b) ou encore **deFuse** (McPherson et al., 2011). L'algorithme de **Cicero** (Tian et al., 2020) est basé sur un alignement local avec STAR, et permet la détection des transcrits de fusion ainsi que les ITD (« Internal Tandem Duplication »). La liste des fusions candidates est établie à partir de l'analyse des lectures portant des bases « soft-clippées » (partie minoritaire d'une lecture qui ne s'aligne pas correctement à la localisation génomique de la partie majoritaire). Les contigs de fusion sont ensuite assemblés et cartographiés, en utilisant BLAT pour identifier le point de cassure de la fusion.

6.4 Performances

Les outils de détection de fusions varient en termes d'outils d'alignement utilisés, du génome de référence, des critères de validation d'une fusion candidate et des filtres mis en place pour limiter les faux positifs. Le besoin, avec l'adoption du RNA-Seq en médecine de précision et en diagnostic clinique, est de trouver les méthodes avec de bonnes performances de prédiction et un temps d'exécution raisonnable. **Une étude récente évalue 41 outils de détection** de transcrits de fusions, testés sur des données transcriptomiques à travers 10 évaluations couvrant 38 jeux de données, en termes de temps d'exécution et de besoin en ressources de calcul. La performance de ces outils en sensibilité et précision varie considérablement en fonction du type d'échantillon (réel ou simulé) et de la longueur des lectures (**Figure 19**). **Cette étude montre clairement la nécessité de combiner plusieurs algorithmes** pour aboutir à un résultat fiable (Ostrowska and Gambin, 2025).

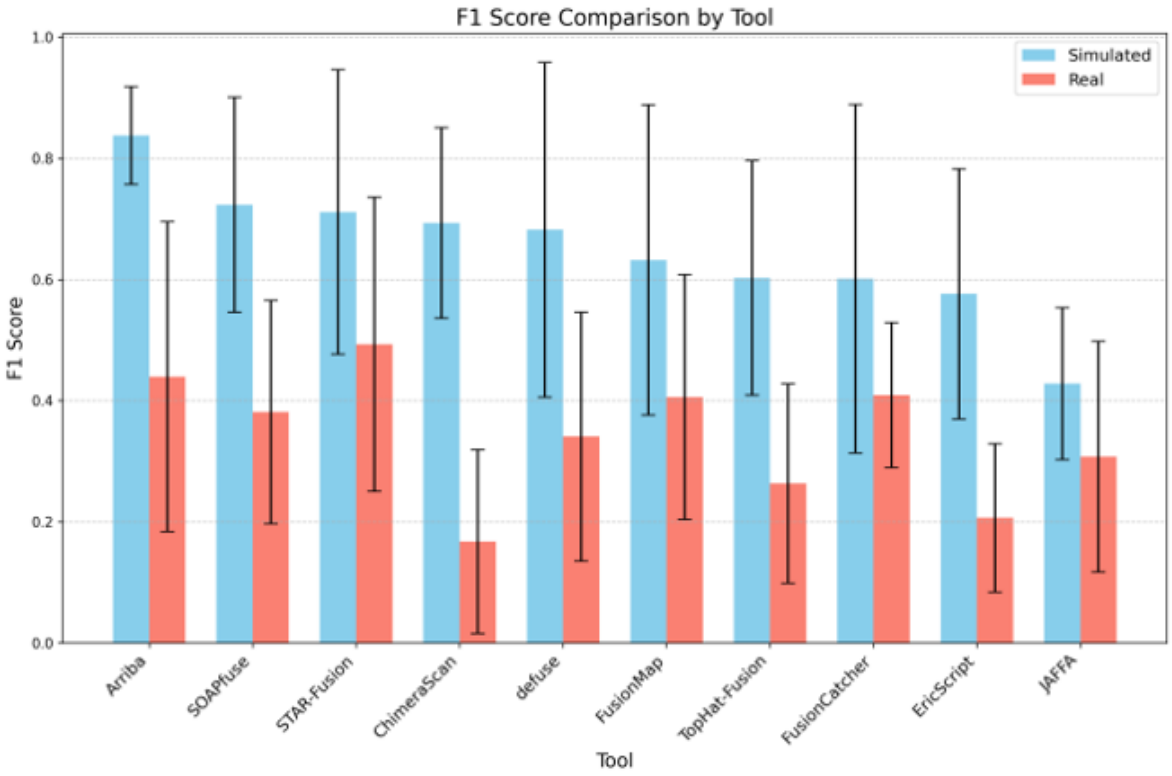

Figure 19 : Comparaison des F1-scores avec écarts-types (barres d'erreurs) pour les données réelles (rouge) et les données simulées (bleu) pour 10 outils de détection de fusions de gènes.

Tableau 4 : Performances comparées de 8 algorithmes de détection des fusions de gènes selon 2 benchmarks. Ref. : [A] : (Vicente-Garcés et al., 2023) ; [B] (Creason et al., 2021)

Outils	Codes	Type de données d'entrées	Metrics (benchmark [A] estimation "Global")	Metrics benchmark [B]	Remarques
Arriba	Releases suhrig/arriba · GitHub	Alignements issus de l'outil STAR (SAM, BAM ou CRAM)	F1 score : 0,89 Sensitivity : 82,5 Precision: 97,1	F1 = 0.73	
Star-Fusion	Releases STAR-Fusion/STAR-Fusion GitHub	Fastq ou alignement de STAR	F1 score : 0,81 Sensitivity : 85 Precision: 77,3	F1 = 0.70	FusionInspector inclus dans le package génère des graphes
FusionCatcher	Release Version 1.33 · ndaniel/fusioncatcher GitHub	Fastq	F1 score : 0,88 Sensitivity : 92,5 Precision: 84,1	F1 = 0.58	Dernière version en 2021
Pizzly	Release Bug fixes and license switch pmelsted/pizzly GitHub	Dépend de sortie de kallisto	/	/	Pas de version stable depuis 2017
manta	Release manta-v1.6.0 Illumina/manta GitHub	Prend des BAM et CRAM	/	/	Plus maintenu
FusionScan	Site inaccessible fusionscan.ewha.ac.kr	Fastq	/		Dernière version en 2015
Cicero	Release v1.9.6 stjude/CICERO GitHub	Fastq ou BAM de STAR avec retrait de duplicat	F1 score : 0,85 Sensitivity : 75 Precision: 96,8	/	Détecte des ITD avec une précision supérieure à Arriba et STAR-fusion. Sorties graphiques disponibles avec FusionEditor.
deFuse	http://compbio.bccrc.ca	Fastq	F1 score : 0,76 Sensitivity : 70 Precision: 82,4	/	Site inaccessible à la date du 07/05/2024

Comme on le constate dans le **Tableau 4** et la **Figure 19,** le choix du jeu de données de test va influencer grandement les performances. Néanmoins, **STAR-fusion et Arriba sont les outils les plus performants**.

Afin de compléter l'analyse de performance, l'article de Zhong et collaborateurs apporte des références de données de test utiles pour des évaluations complémentaires (Zhong et al., 2023). Par rapport au développement de nouvelles approches, l'outil SFyNCS (Zhong et al., 2023) semble, selon ses propres tests, plus performant que les outils STAR-fusion et Arriba. Enfin, l'apprentissage profond (« deep learning ») commence aussi à être utilisé pour la détection de fusion comme le décrit une étude récente (Zhong et al., 2023) utilisant FusionAI (Kim et al., 2022).

Les variations de performance des outils utilisés sur un ensemble de données de test identique requiert une étape de décision soit par la majorité (Apostolides et al., 2023) (Vicente-Garcés et al., 2023), soit en utilisant une approche de type apprentissage automatique (« machine learning ») (Hafstað et al., 2023).

6.5 SoFuR et autres workflows intégratifs

Plusieurs équipes développent l'intégration de divers outils pour finalement sélectionner des fusions candidates.

SoFur (SOmatic FUsions from Rna), est un pipeline de détection, d'annotation et de filtration des fusions qui traite les données « stranded ». Il combine Arriba, Manta et STAR-Fusion et produit une visualisation graphique des fusions détectées ainsi qu'un rapport qualité (https://github.com/bialimed/SoFuR).

On peut aussi citer InPipe (Vicente-Garcés et al., 2023), METAFusion (Apostolides et al., 2023) et le pipeline rna-fusion de nf-core (https://nf-co.re/rnafusion).

7 Le séquençage de l'ARN en lectures longues pour l'étude des transcrits de fusion

Le séquençage en lectures longues (« long reads », séquençage de 3ème génération) d'ARN par Pacific Biosciences (PacBio) ou Oxford Nanopore Technologies (ONT) **permet en théorie de capturer des transcrits pleine longueur** et lire les régions répétées. Cette approche **facilite la résolution de la diversité et la quantité des isoformes**, aussi bien en condition normale que pathologique. Malgré une couverture plus faible, la reconstruction des transcrits par lectures longues est meilleure que celle par lectures courtes. Cependant les technologies en lectures longues induisent des erreurs et des biais qui peuvent altérer l'alignement et la reconstruction des transcrits, menant à une identification erronée.

7.1 Alignement

7.1.1 Minimap2

Une étude de 2018 (Križanovic et al., 2018), a comparé plusieurs outils d'alignement, y compris ceux faisant référence pour les lectures courtes. Les résultats, basés sur des jeux de données réels et synthétiques issus de technologies PacBio et ONT, ont montré que GMAP était l'aligneur le plus performant (**Tableau 5**).

Tableau 5 : Pourcentage des lectures alignées selon les jeux de données et les algorithmes d'alignement. Les jeux de données A et 5-8 sont de type réel, issus de *D. melanogaster,* alors que les jeux de données 1-4 sont synthétiques. Les technologies de séquençage diffèrent aussi selon les jeux. D'après (Križanovic et al., 2018).

Jeu de données	# lectures	Tophat2 (%)	Hisat2 (%)	STAR (%)	BBMap (%)	GMAP (%)
A - Illumina	4M	85,2	94,8	96,8	**97,6**	96,7
1 - PacBio	185K	0,7	6,77	48,9	**91,4**	89,2
2 - PacBio	412K	0	0	33,3	84,5	**92,0**
3 - PacBio	84K	0	0	32,3	64,3	**88,3**
4 - ONT R9	342K	0	0	5,5	43,0	**98,8**
5 - PacBio	192K	0	0	46,1	74,5	**85,4**
6 – PacBio - avec correction d'erreur	192K	0	0,4	67,2	82,8	**88,5**
7 - PacBio	243K	0	0	0,1	72,8	**89,7**
8 - ONT R9	40K	0	0	16,7	88,0	**98,3**

En 2018, **minimap2** a été publié avec une comparaison approfondie sur les lectures ONT. Cette étude a démontré que minimap2 surpasse nettement GMAP et STAR en termes de **temps de calcul, de consommation de ressources, et de performances d'alignement**. Ces performances incluent un **nombre plus élevé de lectures alignées** et une **meilleure identification des jonctions d'introns** (voir **Tableau 6**, (Li, 2018)).

Les auteurs notent aussi que STAR est moins performant en présence de lectures bruitées, on peut donc s'attendre avec l'amélioration des performances ONT à une amélioration parallèle des performances de STAR sur ces mêmes données (cependant, en 2018 la différence de performance entre minimap2 et STAR était très importante). De plus, **la plupart des outils d'analyse ARN lectures longues sont utilisés en aval d'un alignement réalisé par minimap2** (**cf paragraphe 7.7**).

Tableau 6 : Évaluation de la précision de la détection de jonctions sur lectures ONT. Lectures issues de chimie R9.4, alignés sur génome murin GRCm38. D'après, (Li, 2018).

	GMAP	minimap2	SpAln	STAR (long)
Temps de traitement (CPU min)	631	**15.9**	2076	33.9
RAM max (GByte)	8.9	14.5	**3.2**	29.2
#lectures alignées	103 669	**104 199**	103 711	26 479
#alignements chimériques	**1904**	1488	0	0
#alignements non épissés	15 854	14 798	**17 033**	10 545
#introns alignés	692 275	**693 553**	692 945	78 603
#nouveaux introns	**11 239**	3113	8550	1214
% introns exacts	83.8	**94.0**	87.9	55.2
% introns (approx)	91.8	**96.9**	92.5	82.4

On citera pour les techniques d'alignement de transcriptome en vue d'assemblage une méthode récente et performante n'utilisant pas de génome de référence : RNABloom2 (Nip et al., 2023).

7.2 Détection

Ce que l'on constate dans la littérature des outils dédiés aux technologies de séquençage de 3ème génération, c'est la **prééminence d'approches pipelines multi-tâches qui ne se retrouve pas forcément dans les outils dédiés aux lectures courtes**, qui s'attachent à étudier les différents aspects qualitatifs et quantitatifs du transcriptome.

7.2.1 JAFFAL

JAFFAL (Davidson et al., 2022) est un **outil de détection des fusions de gènes** (par extension de JAFFA pour le RNA-Seq). Ce pipeline aligne les lectures sur un transcriptome de référence. Les lectures couvrant deux gènes différents sont identifiées comme lectures candidates et réalignées sur le génome de référence (minimap2) pour confirmation. Les lectures sont ensuite classifiées selon les positions des points de cassure, et ceux-ci sont classés selon un niveau de confiance relatif au nombre de lectures supportant la cassure et leur localisation ou non au bord d'un exon (**Figure 20**). Il pourrait donc dévaloriser (à tort) les jonctions *de novo* n'impliquant pas d'exon connu. La sortie est rapportée dans un tableau ainsi qu'un fichier de séquence des jonctions.

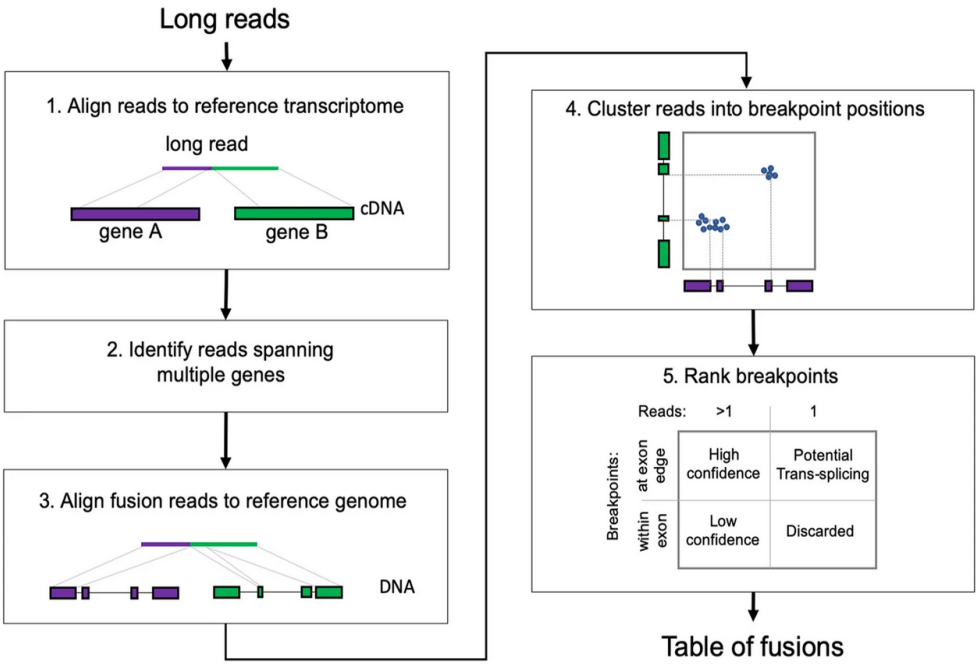

Figure 20 : **Description du pipeline JAFFAL pour la détection des fusions** (Davidson et al., 2022)

7.3 SOSTAR

SOSTAR (iSofOrmS annoTAtoR) (Aucouturier et al., 2024) est un **pipeline polyvalent en snakemake pour assembler et décrire les isoformes issues du séquençage « long read RNA-Seq » ciblé**. Un premier module **(Figure 21)** procède à un alignement primaire et secondaire (minimap2), assemble les isoformes et calcule leur expression (StringTie). Le second module annote chaque isoforme assemblée issue d'un épissage alternatif. En théorie, **il peut quantifier les isoformes et en découvrir de nouvelles (sans détection de fusions)**. La sortie est un fichier tableur avec une table de comptage. Le protocole de capture permet ici d'atteindre une profondeur suffisante pour la caractérisation des isoformes.

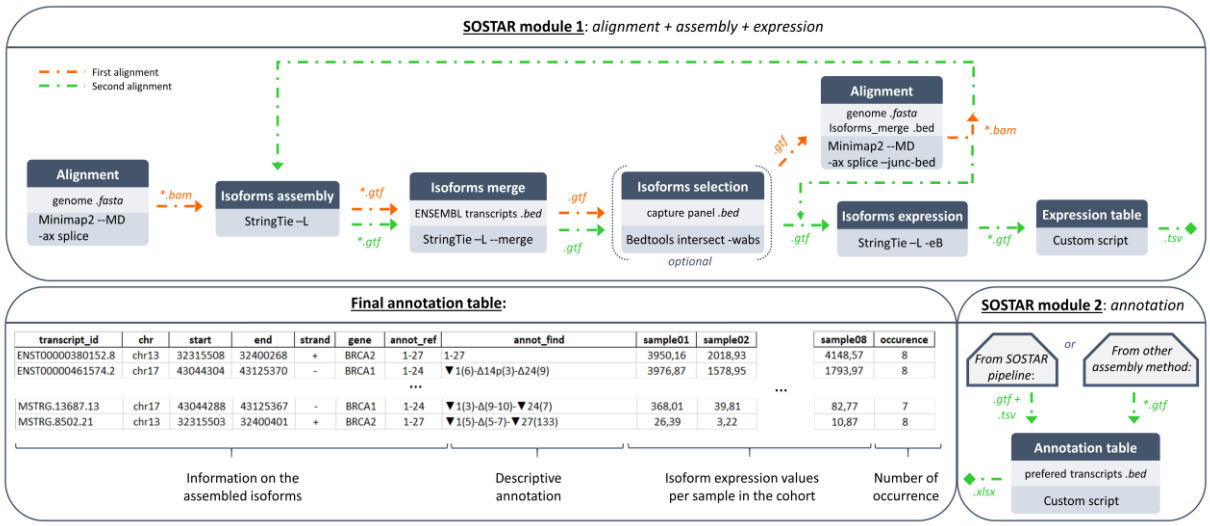

Figure 21 : Description du pipeline SOSTAR (Aucouturier et al., 2024)

7.4 TAGET: Toolkit for Analyzing full-length GEne Transcripts.

Dédié au séquençage PacBio Iso-seq dont le taux d'erreur est faible, TAGET (Xia et al., 2023) est aussi capable de traiter les lectures ONT pour **l'alignement, l'annotation, l'expression et la fusion**. La publication est basée sur le jeu de données GM12878 et une chimie R9.4 plus bruitée que la plus récente R10. L'approche de TAGET est de mixer un alignement lectures longues (minimap2, **Figure 22**) avec un alignement lectures courtes (STAR ou HISAT2), les lectures courtes étant obtenues par découpage en fenêtres glissantes des lectures longues. Puis TAGET assemble les 2 types d'alignement pour obtenir une localisation plus fine des jonctions. Sur données ONT, les pourcentages de lectures alignées sont similaires entre TAGET, minimap2 et GMAP, mais les jonctions d'épissage de TAGET sont plus précises.

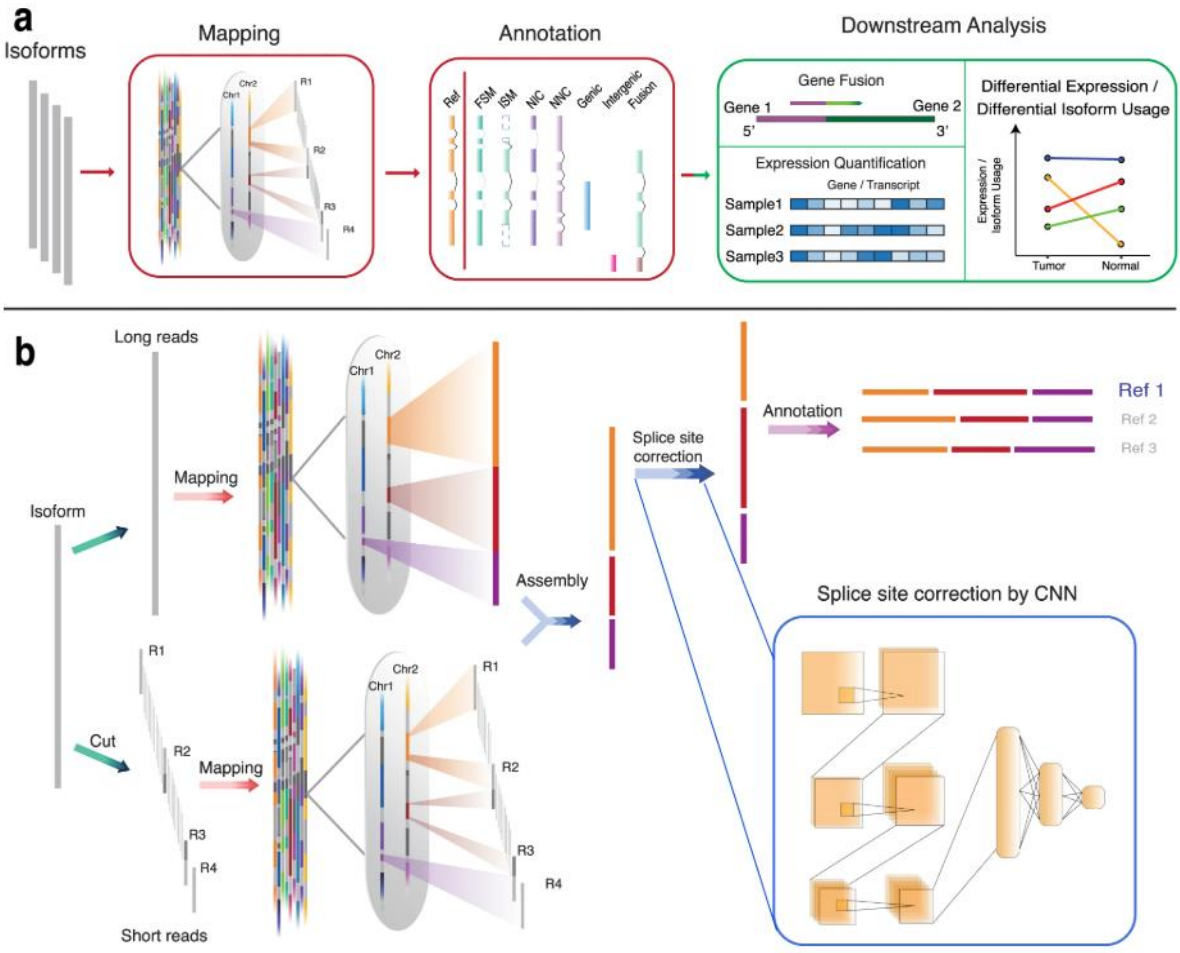

Figure 22 : Description du pipeline TAGET (Xia et al., 2023)

En ce qui concerne la détection de gènes de fusions, TAGET a une F-mesure supérieure à SQANTI et JAFFAL sur le jeu de données testé dans leur publication (**Figure 23**).

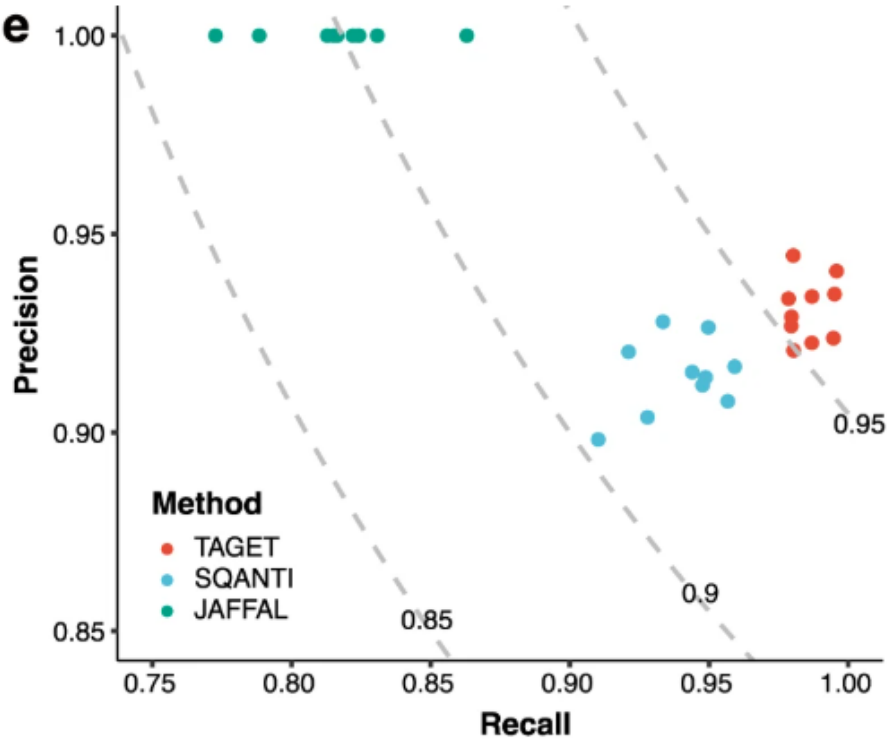

Figure 23 : capacité de TAGET, SQANTI et JAFFAL à détecter correctement des fusions de gènes sur un jeu de données simulées réalisé pour la publication de TAGET (Xia et al., 2023)

7.5 SQANTI3

Le Consortium LRGASP (« Long-read RNA-Seq Genome Annotation Assessment Project », https://www.gencodegenes.org/pages/LRGASP/) a pour objectif d'évaluer l'efficacité des approches lectures longues pour les analyses de transcriptomique, en termes de détection/quantification d'isoformes, transcrits *de novo*. Pour cela, plusieurs transcriptomes de différentes espèces avec différents protocoles et plateformes ont été séquencés.

Utilisé dans le cadre du LRGASP, **SQANTI3** (Pardo-Palacios et al., 2024) est un **outil d'évaluation (caractérisation structurale, contrôle qualité) des transcriptomes générés par lectures longues**, composé de 3 modules : « Quality control, Filter, Rescue » **(Figure 24**). Il génère un rapport (html/pdf) assez conséquent (113 pages dans l'exemple donné SQANTI3/example/SQANTI3_QC_output/UHR_chr22_SQANTI3_report.pdf at master · ConesaLab/SQANTI3 · GitHub).

Le principal module « quality control » combine des scripts Python et R pour annoter et classifier le transcriptome en catégories incluant **les fusions, les nouvelles isoformes, le niveau d'expression** dans un fichier de classification au format texte. Il peut accepter des données orthogonales comme du CAGE Peak, polyA motif ou des lectures courtes pour aider à l'identification et la quantification (grâce à STAR et Kallisto). Le module « filter » repose sur de l'apprentissage automatique par random forest pour calculer la probabilité qu'un transcrit soit une isoforme ou un artefact. Cette étape requiert des jeux de données d'entraînement. Les règles de ce module sont paramétrables par un fichier au format JSON.

Le module « rescue » est dédié à la récupération des transcrits classés artefact à tort. L'annotation est réalisée par tappAS (de la Fuente et al., 2020) qui génère un fichier au format gff3 et IsoAnnotLite (IsoAnnot Lite – IsoAnnot).

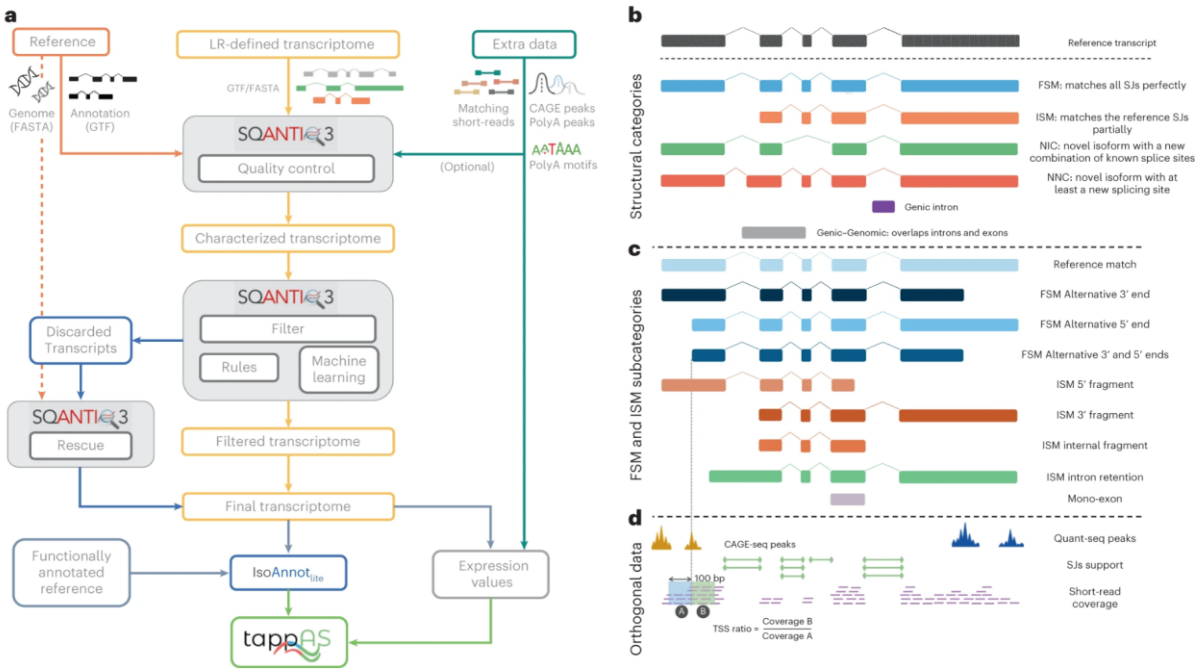

Figure 24 : vue d'ensemble de SQANTI3. **A)** workflow SQANTI. **B)** Principales catégories structurales de SQANTI3 pour les modèles de transcrits de gènes connus. **C)** sous-catégories de SQANTI3 pour les transcrits FSM (« Full Splice Match ») et ISM (« Incomplete Splice Match »). **D)** données orthogonales traitées par le module Quality control de SQANTI3. LR : long read, SJ, Splice junction (de la Fuente et al., 2020). NIC: « Novel In Catalog » ; NNC: « Novel Not In Catalog » ; SJ : « Splice junction » ; LR: « long read »

7.6 Approches combinées en lectures courtes et longues

Dans un contexte de laboratoire de diagnostic, l'utilisation combinée des techniques lectures courtes et longues reste peu probable en raison du surcoût, mais elle offrirait l'avantage de maximiser les performances. Plusieurs publications mettent en évidence les bénéfices de cette approche combinée : (Almeida et al., 2024; Gong et al., 2024; Han et al., 2024; Shumate et al., 2022).

7.7 Workflows multitâches

Ces workflows intègrent plusieurs programmes complémentaires, pour répondre à l'ensemble des questions transcriptomiques (fusion, épissage, expression). Ces deux workflows utilisent minimap2 pour l'alignement et JAFFAL pour la détection de fusions.

7.7.1 NANOSEQ (NextFlow) par nf-core

Nanoseq (https://nf-co.re/nanoseq/) permet la détection des fusions mais aussi une analyse de l'expression (**Figure 25**).

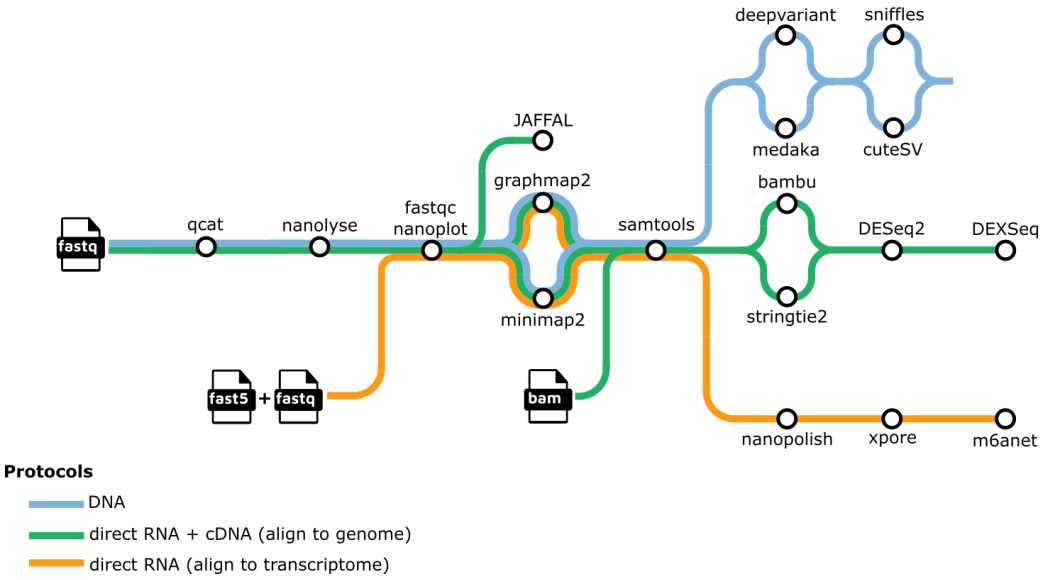

Figure 25 : Schéma des différents chemins disponibles dans nanoseq à suivre selon l'entrée et la sortie voulue.

7.7.2 EPI2ME (NextFlow) par ONT

Proposé par ONT, il existe une **version en ligne de commande ou interface graphique** (https://github.com/epi2me-labs/wf-transcriptomes). Ce workflow exécute un alignement (minimap2), un assemblage des transcrits (StringTie), l'annotation, le comptage et essentiellement une analyse d'expression différentielle par gène (edgeR) et par transcrit (DEXSeq).

7.8 Jeux de données

7.8.1 SG-NEx

Le SG-NEx project est le fruit d'une collaboration internationale initiée au Genome Institute of Singapore afin de fournir les transcriptomes de référence des principales lignées cellulaires cancéreuses utilisées, basés sur les données « long read RNA-Seq » de Nanopore (**Figure 26**).

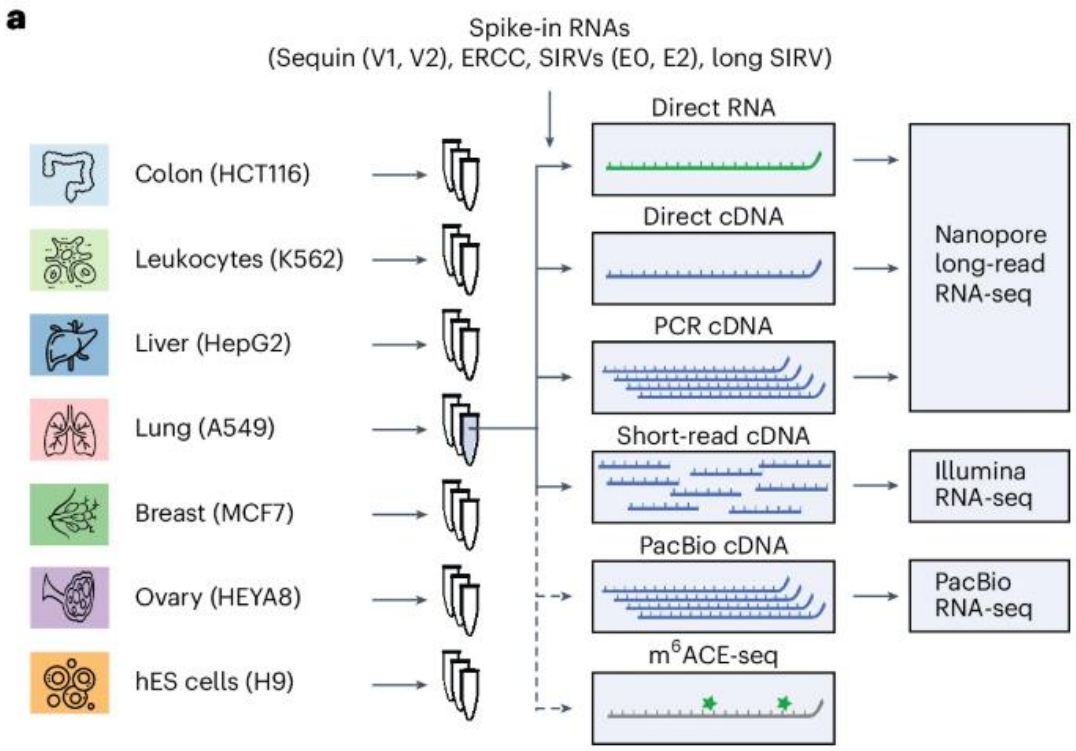

Figure 26 : Vue d'ensemble du pipeline SG-NEx : 7 lignées cellulaires issues de cancers ont été séquencées en de multiples réplicats selon des protocoles RNA-Seq différents (Chen et al., 2025).

Les transcriptomes sont réalisés par différentes stratégies de séquençage (**Figure 26**) :

- Séquençage de produits de PCR-ADNc,
- Direct cDNA: séquençage d'ADNc sans amplification,
- Direct RNA: séquençage d'ARN natif,
- RNA-Seq en lectures courtes ou PacBio
- Analyse de la N6-méthyladénosine (modification de l'ARN la plus abondante, ayant un effet sur l'expression génique)

Trois réplicats *a minima* de bonne qualité ont été réalisés. Des contrôles internes (spike-in, SIRV) ont été inclus. Les données brutes, alignées et traitées sont disponibles (https://github.com/GoekeLab/sg-nex-data (Chen et al., 2025).

7.8.2 GM12878 / NA12878

Les données issues de la lignée cellulaire GM12878 (Coriell, lignée Lymphoblastoide) séquencées par MinION ONT au Mortazavi Lab sont disponibles dans le Gene Expression Omnibus (GEO) avec le numéro d'accession GSE132766. Le même échantillon NA12878 a été séquencé en « Direct RNA et cDNA » avec le 1D ligation kit (SQK-LSK108) sur chimie R9.4 (FLO-MIN106) sur MinION et GridION, utilisant pour le basecalling Guppy 4.2.2 flip flop (high accuracy) model (https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md) (Workman et al., 2019)).

7.8.3 GTEx Portal

Le portail GTEx donne aussi accès à des données en lecture longues (protocole « cDNA-PCR ») générées à partir de 88 tissus et lignées GTEx, alignées en hg38 et transcrits quantifiés. (https://www.gtexportal.org/home/downloads/adult-gtex/long_read_data) (Glinos et al., 2022).

7.8.4 Resources diverses

Le site « long-read-tools » recense un nombre impressionnant de ressources : https://long-read-tools.org/index.html.

On trouvera aussi des détails sur le séquençage ARN direct par ONT sur le site de Nanopore : https://nanoporetech.com/products/prepare/rna-library-preparation.

8 Conclusion

Le séquençage de l'ARN est actuellement en train de transformer le diagnostic en génétique moléculaire, offrant de nouvelles perspectives d'étude du transcriptome et d'évaluation de son implication dans les pathologies humaines. Ce document vise à passer en revue les principales applications du RNA-Seq en diagnostic, allant de la préparation d'un échantillon au séquençage en lectures longues en passant par l'identification des transcrits de fusions, des variants d'épissage et l'analyse de l'expression différentielle de transcrits. Nous avons détaillé les outils bioinformatiques pour chaque étape, du contrôle qualité et de l'alignement à la détection, l'annotation et l'interprétation des résultats. L'émergence de technologies de séquençage haut débit, couplée au développement constant d'algorithmes bioinformatiques, permet d'analyser le transcriptome avec une précision et une rapidité croissante.

Toutefois, la démocratisation de l'application du RNA-Seq au diagnostic est confrontée à de nombreux défis. A la différence de la recherche, où l'on compare des cohortes, le diagnostic se concentre sur la comparaison d'un patient à une cohorte contrôle. Cela nécessite des approches statistiques particulières, comme celles employées par FRASER2, LeafCutterMD et OUTRIDER. Filtrer efficacement les faux positifs, notamment lors de la recherche de transcrits de fusions avec des outils comme Arriba et STAR-Fusion, est nécessaire pour un diagnostic fiable. L'interprétation des données RNA-Seq requiert une expertise combinée en bioinformatique et en génétique médicale, afin de relier les variations transcriptomiques aux phénotypes cliniques et d'identifier des cibles thérapeutiques.

➔ **Quels défis restent à surmonter ?**

La standardisation des pipelines bioinformatiques est aujourd'hui primordiale pour la reproductibilité et la comparabilité inter-laboratoires.

Des bases de données de référence plus complètes, intégrant des données de populations diverses et de tissus sains et pathologiques, sont nécessaires pour améliorer l'interprétation des variants.

L'intégration du RNA-Seq avec d'autres données omiques (génomiques, protéomiques, métabolomiques) permettra une compréhension plus holistique des maladies.

Enfin, les aspects éthique et réglementaire de l'utilisation du RNA-Seq en diagnostic ne doivent pas être négligés : confidentialité des données, consentement éclairé et validation des tests. Une collaboration étroite entre bioinformaticiens, généticiens, cliniciens et éthiciens est indispensable pour établir des pratiques robustes, fiables et éthiquement responsables. L'objectif final est d'exploiter la puissance du RNA-Seq pour améliorer le diagnostic, le pronostic et le traitement des maladies, ouvrant la voie à une médecine personnalisée plus efficace. L'exploration de nouvelles pistes, comme l'identification des cancers primitifs inconnus et l'utilisation des technologies en lectures longues, représente un futur prometteur pour le RNA-Seq en diagnostic et nous laisse entrevoir un futur où le RNA-Seq jouera un rôle central dans la médecine de précision.

Bibliographie

Akinyi, M. V and Frilander, M. J. (2021). At the Intersection of Major and Minor Spliceosomes: Crosstalk Mechanisms and Their Impact on Gene Expression. *Front Genet* **12**, 700744.

Almeida, I., Lu, X., Edwards, S. L., French, J. D. and Bitar, M. (2024). HyDRA: a pipeline for integrating long- and short-read RNAseq data for custom transcriptome assembly.

Apostolides, M., Li, M., Arnoldo, A., Ku, M., Husić, M., Ramani, A. K., Brudno, M., Turinsky, A., Hawkins, C. and Siddaway, R. (2023). Clinical Implementation of MetaFusion for Accurate Cancer-Driving Fusion Detection from RNA Sequencing. *J Mol Diagn* **25**, 921–931.

Aucouturier, C., Soirat, N., Castéra, L., Bertrand, D., Atkinson, A., Lavolé, T., Goardon, N., Quesnelle, C., Levilly, J., Barbachou, S., et al. (2024). Fine mapping of RNA isoform diversity using an innovative targeted long-read RNA sequencing protocol with novel dedicated bioinformatics pipeline. *BMC Genomics* **25**, 909.

Bieler, J., Kubik, S., Macheret, M., Pozzorini, C., Willig, A. and Xu, Z. (2023). Benefits of applying molecular barcoding systems are not uniform across different genomic applications. *J Transl Med* **21**, 305.

Bray, N. L., Pimentel, H., Melsted, P. and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525–7.

Brechtmann, F., Mertes, C., Matusevičiūtė, A., Yépez, V. A., Avsec, Ž., Herzog, M., Bader, D. M., Prokisch, H. and Gagneur, J. (2018). OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *Am J Hum Genet* **103**, 907–917.

Burset, M., Seledtsov, I. A. and Solovyev, V. V (2001). SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res* **29**, 255–9.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S. and Saunders, C. T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–2.

Chen, Y., Davidson, N. M., Wan, Y. K., Yao, F., Su, Y., Gamaarachchi, H., Sim, A., Patel, H., Low, H. M., Hendra, C., et al. (2025). A systematic benchmark of Nanopore long-read RNA sequencing for transcript-level analysis in human cell lines. *Nat Methods* **22**, 801–812.

Creason, A., Haan, D., Dang, K., Chiotti, K. E., Inkman, M., Lamb, A., Yu, T., Hu, Y., Norman, T. C., Buchanan, A., et al. (2021). A community challenge to evaluate RNA-seq, fusion detection, and isoform quantification methods for cancer discovery. *Cell Syst* **12**, 827-838.e5.

Cummings, B. B., Marshall, J. L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A. R., Bolduc, V., Waddell, L. B., Sandaradura, S. A., O'Grady, G. L., et al. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med* **9**,.

Davidson, N. M., Chen, Y., Sadras, T., Ryland, G. L., Blombery, P., Ekert, P. G., Göke, J. and Oshlack, A. (2022). JAFFAL: detecting fusion genes with long-read transcriptome sequencing. *Genome Biol* **23**, 10.

de la Fuente, L., Arzalluz-Luque, Á., Tardáguila, M., Del Risco, H., Martí, C., Tarazona, S., Salguero, P., Scott, R., Lerma, A., Alastrue-Agudo, A., et al. (2020). tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biol* **21**, 119.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21.

Ewels, P., Magnusson, M., Lundin, S. and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–8.

Fenn, A., Tsoy, O., Faro, T., Rößler, F. L. M., Dietrich, A., Kersting, J., Louadi, Z., Lio, C. T., Völker, U., Baumbach, J., et al. (2023). Alternative splicing analysis benchmark with DICAST. *NAR Genom Bioinform* **5**, lqad044.

Glinos, D. A., Garborcauskas, G., Hoffman, P., Ehsan, N., Jiang, L., Gokden, A., Dai, X., Aguet, F., Brown, K. L., Garimella, K., et al. (2022). Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**, 353–359.

Gong, B., Li, D., Łabaj, P. P., Pan, B., Novoradovskaya, N., Thierry-Mieg, D., Thierry-Mieg, J., Chen, G., Bergstrom Lucas, A., LoCoco, J. S., et al. (2024). Targeted DNA-seq and RNA-seq of Reference Samples with Short-read and Long-read Sequencing. *Sci Data* **11**, 892.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T. and Ast, G. (2006). Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. *Mol Cell* **22**, 769–781.

Haas, B. J., Dobin, A., Li, B., Stransky, N., Pochet, N. and Regev, A. (2019). Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol* **20**, 213.

Hafstað, V., Häkkinen, J., Larsson, M., Staaf, J., Vallon-Christersson, J. and Persson, H. (2023). Improved detection of clinically relevant fusion transcripts in cancer by machine learning classification. *BMC Genomics* **24**, 783.

Han, S. W., Jewell, S., Thomas-Tikhonenko, A. and Barash, Y. (2024). Contrasting and combining transcriptome complexity captured by short and long RNA sequencing reads. *Genome Res* **34**, 1624–1635.

Jenkinson, G., Li, Y. I., Basu, S., Cousin, M. A., Oliver, G. R. and Klee, E. W. (2020). LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. *Bioinformatics* **36**, 4609–4615.

Kim, D., Paggi, J. M., Park, C., Bennett, C. and Salzberg, S. L. (2019a). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915.

Kim, P., Jang, Y. E. and Lee, S. (2019b). FusionScan: accurate prediction of fusion genes from RNA-Seq data. *Genomics Inform* **17**, e26.

Kim, P., Tan, H., Liu, J., Kumar, H. and Zhou, X. (2022). FusionAI, a DNA-sequence-based deep learning protocol reduces the false positives of human fusion gene prediction. *STAR Protoc* **3**, 101185.

Križanovic, K., Echchiki, A., Roux, J. and Šikic, M. (2018). Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* **34**, 748–754.

Labory, J., Le Bideau, G., Pratella, D., Yao, J.-E., Ait-El-Mkadem Saadi, S., Bannwarth, S., El-Hami, L., Paquis-Fluckinger, V. and Bottini, S. (2022). ABEILLE: a novel method for ABerrant Expression Identification empLoying machine LEarning from RNA-sequencing data. *Bioinformatics* **38**, 4754–4761.

Leman, R., Harter, V., Atkinson, A., Davy, G., Rousselin, A., Muller, E., Castéra, L., Lemoine, F., de la Grange, P., Guillaud-Bataille, M., et al. (2020). SpliceLauncher: a tool for detection, annotation and relative quantification of alternative junctions from RNAseq data. *Bioinformatics* **36**, 1634–1636.

Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., Gnirke, A. and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**, 709–15.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323.

Lorenzi, C., Barriere, S., Arnold, K., Luco, R. F., Oldfield, A. J. and Ritchie, W. (2021). IRFinder-S: a comprehensive suite to discover and explore intron retention. *Genome Biol* **22**, 307.

Lu, S., Cheng, Y., Huang, D., Sun, Y., Wu, L., Zhou, C., Guo, Y., Shao, J., Zhang, W. and Zhou, J. (2022). Efficacy and safety of selpercatinib in Chinese patients with advanced RET fusion-positive non-small-cell lung cancer: a phase II clinical trial (LIBRETTO-321). *Ther Adv Med Oncol* **14**, 17588359221105020.

Ma, F., Fuqua, B. K., Hasin, Y., Yukhtman, C., Vulpe, C. D., Lusis, A. J. and Pellegrini, M. (2019). A comparison between whole transcript and 3' RNA sequencing methods using Kapa and Lexogen library preparation methods. *BMC Genomics* **20**, 9.

Mareschal, S., Wucher, V., Huet, S., Léonce, C., Chabane, K., Hayette, S., Bringuier, P.-P., Pinson, S., Barritault, M. and Bardel, C. (2025). Detecting aberrant splicing events in short-read RNA-seq with SAMI, an UMI-aware Nextflow pipeline.

McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M. G. F., Griffith, M., Heravi Moussavi, A., Senz, J., Melnyk, N., et al. (2011). deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol* **7**, e1001138.

Mitelman, F., Johansson, B. and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* **7**, 233–45.

Mohammadi, P., Castel, S. E., Cummings, B. B., Einson, J., Sousa, C., Hoffman, P., Donkervoort, S., Jiang, Z., Mohassel, P., Foley, A. R., et al. (2019). Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science* **366**, 351–356.

Nicorici, D., Şatalan, M., Edgren, H., Kangaspeska, S., Murumägi, A., Kallioniemi, O., Virtanen, S. and Kilkku, O. (2014). FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data.

Nip, K. M., Hafezqorani, S., Gagalova, K. K., Chiu, R., Yang, C., Warren, R. L. and Birol, I. (2023). Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2. *Nat Commun* **14**, 2940.

Oliver, G. R., Tang, X., Schultz-Rogers, L. E., Vidal-Folch, N., Jenkinson, W. G., Schwab, T. L., Gaonkar, K., Cousin, M. A., Nair, A., Basu, S., et al. (2019). A tailored approach to fusion transcript identification increases diagnosis of rare inherited disease. *PLoS One* **14**, e0223337.

Oliver, G. R., Jenkinson, G. and Klee, E. W. (2020). Computational Detection of Known Pathogenic Gene Fusions in a Normal Tissue Database and Implications for Genetic Disease Research. *Front Genet* **11**, 173.

Ostrowska, I. and Gambin, T. (2025). Meta-analysis of RNA-seq gene fusion detection tools: performance and variability across benchmarks.

Pardo-Palacios, F. J., Arzalluz-Luque, A., Kondratova, L., Salguero, P., Mestre-Tomás, J., Amorín, R., Estevan-Morió, E., Liu, T., Nanni, A., McIntyre, L., et al. (2024). SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat Methods* **21**, 793–797.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. and Hellmann, I. (2018). zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**,.

Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E. and Zanini, F. (2022). Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics* **38**, 2943–2945.

Rodríguez-Martín, B., Palumbo, E., Marco-Sola, S., Griebel, T., Ribeca, P., Alonso, G., Rastrojo, A., Aguado, B., Guigó, R. and Djebali, S. (2017). ChimPipe: accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data. *BMC Genomics* **18**, 7.

Roloff, G. W., Lai, C., Hourigan, C. S. and Dillon, L. W. (2017). Technical Advances in the Measurement of Residual Disease in Acute Myeloid Leukemia. *J Clin Med* **6**,.

Scheller, I. F., Lutz, K., Mertes, C., Yépez, V. A. and Gagneur, J. (2023). Improved detection of aberrant splicing with FRASER 2.0 and the intron Jaccard index. *Am J Hum Genet* **110**, 2056–2067.

Sharp, P. A. (1994). Split genes and RNA splicing. *Cell* **77**, 805–15.

Shumate, A., Wong, B., Pertea, G. and Pertea, M. (2022). Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS Comput Biol* **18**, e1009730.

Smith, T., Heger, A. and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491–499.

Srivastava, A., Malik, L., Smith, T., Sudbery, I. and Patro, R. (2019). Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. *Genome Biol* **20**, 65.

Subbiah, V., Gouda, M. A., Ryll, B., Burris, H. A. and Kurzrock, R. (2024). The evolving landscape of tissue-agnostic therapies in precision oncology. *CA Cancer J Clin* **74**, 433–452.

Tian, L., Li, Y., Edmonson, M. N., Zhou, X., Newman, S., McLeod, C., Thrasher, A., Liu, Y., Tang, B., Rusch, M. C., et al. (2020). CICERO: a versatile method for detecting complex and diverse driver fusions using cancer RNA sequencing data. *Genome Biol* **21**, 126.

Uhrig, S., Ellermann, J., Walther, T., Burkhardt, P., Fröhlich, M., Hutter, B., Toprak, U. H., Neumann, O., Stenzinger, A., Scholl, C., et al. (2021). Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res* **31**, 448–460.

Ura, H., Togi, S. and Niida, Y. (2022). Poly(A) capture full length cDNA sequencing improves the accuracy and detection ability of transcript quantification and alternative splicing events. *Sci Rep* **12**, 10599.

Vicente-Garcés, C., Maynou, J., Fernández, G., Esperanza-Cebollada, E., Torrebadell, M., Català, A., Rives, S., Camós, M. and Vega-García, N. (2023). Fusion InPipe, an integrative pipeline for gene fusion detection from RNA-seq data in acute pediatric leukemia. *Front Mol Biosci* **10**,.

Wachtel, M., Surdez, D., Grünewald, T. G. P. and Schäfer, B. W. (2024a). Functional Classification of Fusion Proteins in Sarcoma. *Cancers (Basel)* **16**,.

Wachtel, M., Surdez, D., Grünewald, T. G. P. and Schäfer, B. W. (2024b). Functional Classification of Fusion Proteins in Sarcoma. *Cancers (Basel)* **16**,.

Wang, R., Helbig, I., Edmondson, A. C., Lin, L. and Xing, Y. (2023). Splicing defects in rare diseases: transcriptomics and machine learning strategies towards genetic diagnosis. *Brief Bioinform* **24**,.

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., Zuzarte, P. C., Gilpatrick, T., Payne, A., Quick, J., et al. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* **16**, 1297–1305.

Xia, Y., Jin, Z., Zhang, C., Ouyang, L., Dong, Y., Li, J., Guo, L., Jing, B., Shi, Y., Miao, S., et al. (2023). TAGET: a toolkit for analyzing full-length transcripts from long-read sequencing. *Nat Commun* **14**, 5935.

Yépez, V. A., Mertes, C., Müller, M. F., Klaproth-Andrade, D., Wachutka, L., Frésard, L., Gusic, M., Scheller, I. F., Goldberg, P. F., Prokisch, H., et al. (2021). Detection of aberrant gene expression events in RNA sequencing data. *Nat Protoc* **16**, 1276–1296.

Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., von Schack, D. and Zhang, B. (2015). Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics* **16**, 675.

Zhao, S., Zhang, Y., Gamini, R., Zhang, B. and von Schack, D. (2018). Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep* **8**, 4781.

Zhong, X., Luan, J., Yu, A., Lee-Hassett, A., Miao, Y. and Yang, L. (2023). SFyNCS detects oncogenic fusions involving non-coding sequences in cancer. *Nucleic Acids Res* **51**, e96.

Zhou, Q., Su, X., Jing, G., Chen, S. and Ning, K. (2018). RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. *BMC Genomics* **19**, 144.