Variance de l'expression des microARN et des ARN messagers dans le cancer

Christophe Le Priol

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **MBS - Modèles, méthodes et algorithmes en biologie, santé et environnement**

Arrêté ministériel : 25 mai 2016

Présentée par

Christophe LE PRIOL

Thèse dirigée par **Xavier GIDROL**, Directeur de recherche, Communauté Université Grenoble Alpes
et co-encadrée par **Chloé-Agathe AZENCOTT**, Chargée de recherche, MINES ParisTech

préparée au sein du **Laboratoire Biologie à Grande Echelle**
dans **l'École Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement**

Variance de l'expression des microARN et des ARN messagers dans le cancer

Thèse soutenue publiquement le **19/09/2019**,
devant le jury composé de :

Madame Nathalie VIALANEIX
Directrice de recherche, INRA Centre Toulouse Midi-Pyrénées, Rapporteure
Monsieur Pascal BARBRY
Directeur de recherche, CNRS Délégation Côte d'Azur, Rapporteur
Madame Adeline LECLERCQ-SAMSON
Professeur, Université Grenoble Alpes, Présidente
Monsieur Christophe BATTAIL
Cadre scientifique, CEA Grenoble, Examinateur

et des membres invités :

Monsieur Xavier GIDROL
Directeur de recherche, Communauté Université Grenoble Alpes, Directeur de thèse
Madame Chloé-Agathe AZENCOTT
Chargée de recherche, MINES ParisTech, Co-encadrante de thèse

Remerciements

Je tiens tout d'abord à remercier l'EDISCE de m'avoir attribué un financement du Ministère de l'Education Supérieure et de la Recherche me donnant la possibilité de mener mes travaux de thèse pendant trois ans.

Je remercie Madame Nathalie Vialaneix et Monsieur Pascal Barbry d'avoir accepté d'être rapporteurs dans mon jury de thèse. Je tiens à vous remercier pour vos retours qui ont permis d'améliorer mon manuscrit et de donner des pistes d'amélioration de mon travail. Je tiens aussi à remercier Madame Adeline Leclercq-Samson et Monsieur Christophe Battail d'avoir accepté de faire partie de mon jury et pour les échanges fructueux que nous avons pu avoir à l'occasion de mon comité de thèse.

Je remercie Xavier pour son aide tout au long de ces trois années et demi, pour les nombreux échanges scientifiques que l'on a pu avoir qui n'ont fait que confirmer mon envie de persévérer dans la recherche et pour les discussions informelles qui ont toujours été soit intéressantes, soit marrantes. Je tiens à te remercier pour la confiance que tu m'as accordée, en particulier pour la prolongation de financement de six mois qui fut indispensable pour mener à bien ma thèse. Enfin, je veux te remercier pour la qualité scientifique du sujet que tu m'as proposé pour cette thèse. La pertinence de ce projet m'a tout de suite convaincu.

Je remercie Chloé pour son aide précieuse durant ma thèse. Bien que la distance entre Grenoble et Paris n'a pas aidé à l'encadrement de ma thèse, ta réactivité et ton efficacité ont permis d'y remédier. Après le changement de direction au début de ma thèse, ton encadrement était littéralement indispensable à la réussite de ma thèse. Pour cela, je te suis très reconnaissant. Je tiens aussi à remercier tous les membres du CBIO pour leur accueil lors de mes passages à Paris.

Je tiens à remercier chaleureusement Sophie, Aleks et Julie pour tous les bons moments passés au quotidien durant ma thèse. Tous nos échanges, légers, passionnants, intéressants, amusants, voire même délirants, ont été autant de moments permettant de m'extirper de ma thèse et, au final, indispensables à sa réussite. Ces trois années et demi de thèse ont aussi été émaillées de moments difficiles. Vous m'avez aidé à les endurer. Merci.

Pour tous les bons moments passés avec vous, je tiens à remercier Maxime, Alejandro, Ville, Sean, Sergio, Bastien, Sylvain, Andrea, Axel, Farah, Lisa, Thibaud mon salaud, Paul, Mélissa, Tolgahan et Racha. I have special thanks to Ville for his support and having made me visit his wonderful country. In addition, these holidays were perfect to relax and get prepared for my last year of hard work.

Je tiens à remercier mes anciens collègues de bioMérieux, en particulier Jean-Baptiste, Maud, Pierre et Kevin pour m'avoir donné envie de découvrir la recherche académique en travaillant avec vous pendant plusieurs années et en partageant votre culture scientifique. Ce projet, un peu fou, de me lancer dans cette aventure de la thèse ne vient pas de nulle part.

Je tiens à remercier les colocs, Noëlle et Manu, de m'avoir supporté durant ma thèse. J'espère ne pas avoir ramené trop souvent tous mes tracas de thésard à l'appart.

Je souhaite remercier Oriane d'avoir partagé mon enthousiasme lorsque je lui faisais part de mon envie de me lancer dans une thèse alors que ce projet ne suscitait qu'incompréhension et peu d'adhésion chez la plupart des personnes à qui j'en parlais. Avec Carole, tu étais l'une des rares à autant me comprendre et me soutenir dans ce projet. Lorsqu'il s'est agi de véritablement se lancer dans cette aventure, ton soutien fut précieux.

Enfin, je remercie mes parents pour leur soutien indéfectible.

Contributions

Contributions en rapport avec les travaux de cette thèse

Le Priol C, Azencott CA et Gidrol X. « Large-scale RNA-seq datasets enable the detection of genes with a differential expression dispersion in cancer ». Poster présenté à :

— ISMB/ECCB (Intelligent Systems for Molecular Biology / European Conference on Computational Biology), 21-25 juillet 2019, Bâle ;
— JOBIM (Journées Ouvertes de Biologie Informatique & Mathématiques), 2-5 juillet 2019, Nantes.

Autres contributions

Jalbert R, Wong M et Le Priol C (2018). « Étude de trois outils à disposition des médecins généralistes isérois lors de la prise en charge de patients bénéficiant de soins palliatifs à domicile : les médecins généralistes isérois connaissent-ils leur existence ? », Revue internationale de soins palliatifs, vol. 33(HS), p. 51-51.

— analyse statistique de réponses à un questionnaire à choix multiples dans le cadre de la thèse de médecine de Robin Jalbert.

Le Priol C, Guyon L, Azencott CA et Gidrol X. « Analysis of microRNA sequences identifies conserved families of microRNAs ». Poster présenté à JOBIM (Journées Ouvertes de Biologie Informatique & Mathématiques), 28-30 juin 2016, Lyon.

— résultats obtenus à propos d'un premier sujet exploré au cours de cette thèse, abandonné au bout de 6 mois.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

ADN	**A**cide **D**ésoxyribo**N**ucléique
APL	Vraisemblance ajustée profilée (APL pour ***A**djusted **P**rofile **L**ikelihood*)
ARN	**A**cide **R**ibo**N**ucléique
ARNm	**A**cide **R**ibo**N**ucléique **m**essager
AUC	Aire sous la Courbe ROC (AUC pour **A**rea **U**nder the **C**urve)
BFGS	**B**royden-**F**letcher-**G**oldfarb-**S**hanno
BLCA	Carcinome de la vessie (BLCA pour ***BL**adder urothelial **CA**rcinoma*)
BRCA	Carcinome invasive du sein (BRCA pour ***BR**east invasive **CA**rcinoma*)
BH	**B**enjamini-**H**ochberg
BY	**B**enjamini-**Y**ekutieli
COAD	Adénocarcinome du colon (COAD pour ***CO**lon **AD**enocarcinoma*)
CPM	**C**omptes **P**ar **M**illion
CPU	***C**entral **P**rocessing **U**nit*
ESCA	Carcinome de l'œsophage (ESCA pour ***ES**ophageal **CA**rcinoma*)
FDR	Taux de Fausses Découvertes (FDR pour **F**alse **D**iscovery **R**ate)
FN	**F**aux **N**égatif
FP	**F**aux **P**ositif
FPKM	**F**ragments par **K**ilo-base et par **M**illion de *reads*
FPR	Taux de Faux Positifs (FPR pour **F**alse **P**ositive **R**ate)
GDC	***G**enomic **D**ata **C**ommons*
GLM	Modèles Linéaires Généralisés (GLM pour ***G**eneralized **L**inear **M**odels*)
GO	***G**ene **O**ntology*
HNSC	Carcinomes squameux de la tête et du cou (HNSC pour ***H**ead and **N**eck **S**quamous cell **C**arcinoma*)
miARN	**mi**cro-**A**cide **R**ibo**N**ucléique
KIRC	Carcinome à cellules claires du rein (KIRC pour ***KI**dney **R**enal **C**lear cell carcinoma*)
KIRP	Carcinome papillaire du rein (KIRP pour ***KI**dney **R**enal **P**apillary cell carcinoma*)
LIHC	Carcinome hépatique (LIHC pour ***LI**ver **H**epatocellular **C**arcinoma*)
LUAD	Adénocarcinome du poumon (LUAD pour ***LU**ng **AD**enocarcinoma*)
LUSC	Carcinome squameux du poumon (LUSC pour ***LU**ng **S**quamous cell **C**arcinoma*)
PCR	***P**olymerase **C**hain **R**eaction*
PRAD	Adénocarcinome de la prostate (PRAD pour ***PR**ostate **AD**enocarcinoma*)
QL	Quasi-vraisemblance (QL pour ***Q**uasi-**L**ikelihood*)
READ	Adénocarcinome du rectum (READ pour ***RE**ctum **AD**enocarcinoma*)
RNA-seq	Séquençage d'ARN (RNA-seq pour ***RNA-seq**uencing*)
ROC	***R**eceiver **O**perating **C**haracteristic*
STAD	Adénocarcinome de l'estomac (STAD pour ***ST**omach **AD**enocarcinoma*)
RPKM	***R**eads* par **K**ilo-base et par **M**illion de *reads*
TCGA	***T**he **C**ancer **G**enome **A**tlas*

THCA	Carcinome de la thyroïde (THCA pour ***TH****yroid* ***CA****rcinoma*)
TMM	Moyenne tronquée de M-valeurs (TMM pour ***T****rimmed* ***M****ean* ***M****-values*)
TN	Vrai Négatif (TN pour **T**rue **N**egative)
TNR	Taux de Vrai Négatifs (TNR pour **T**rue **N**egative **R**ate)
TP	Vrai Positif (TP pour **T**rue **P**ositive)
TPR	Taux de Vrai Positifs (TPR pour **T**rue **P**ositive **R**ate)
UCEC	Carcinome endométrial (UCEC pour ***U****terine* ***C****orpus* ***E****ndometrial* ***C****arcinoma*)
UQ	Troisième quartile (UQ pour ***U****pper* ***Q****uartile*)
UTR	Région transcrite non traduite (UTR pour ***U****ntranslated* ***T****ranscribed* ***R****egion*)
wqCML	Maximum de vraisemblance conditionnelle ajustée par quantiles pondérée (wqCML pour ***w****eighted* ***q****uantile-ajusted* ***C****onditional* ***M****aximum* ***L****ikelihood*)

Chapitre 1

Contexte biologique

1 microARN et cancer

1.1 Description des gènes de microARN

1.1.1 Localisation dans le génome

Les microARN (miARN) sont des petits ARN (environ 22 nucléotides) non codants, *i.e.* non traduits en protéines, qui régulent l'expression de gènes de manière post-transcriptionnelle. On retrouve les gènes de miARN dans les génomes d'animaux et de plantes à différentes localisations (figure 1.1). Les gènes de miARN sont loca-

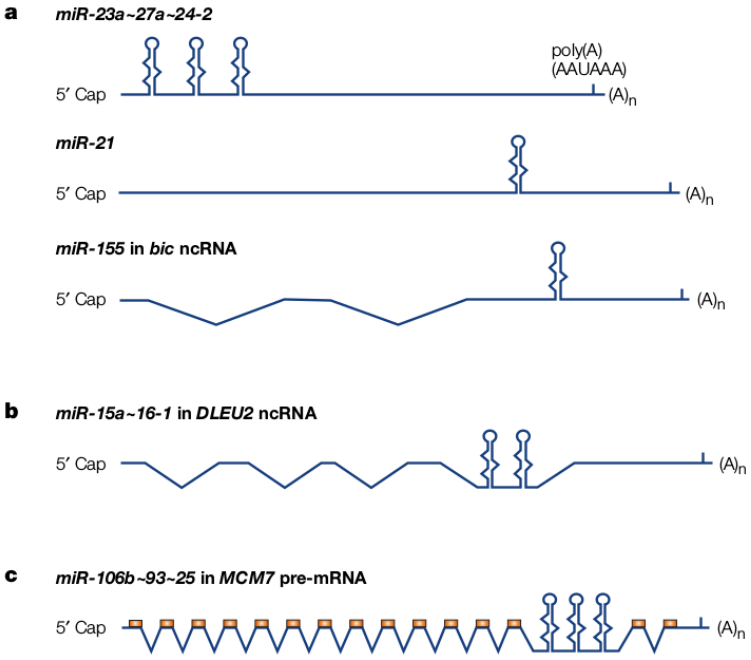

FIGURE 1.1 – Différentes localisations de miARN dans le génome (figure 1 de KIM, 2005) : dans des introns de gènes non codants (a), dans les exons de gènes non codants (b), dans les introns de gènes codant des protéines (c).

lisés à la fois dans des régions transcrites et des régions non transcrites du génome. Ils peuvent ainsi se retrouver dans les exons ou les introns d'ARN non codants, mais aussi dans les introns de gènes codant des protéines. La transcription de ces derniers se retrouve alors dépendante de celle du gène en ARN messager. Enfin, certains gènes de miARN, sont regroupés en une même région génomique, formant ce que l'on appelle un cluster de miARN. Ces miARN ainsi rassemblés partagent souvent les mêmes fonctions.

1.1.2 Synthèse

Les miARN sont transcrits par l'ARN poymérase II (Lee et al., 2004). Avant d'apparaître sous leur forme active, les miARN subissent plusieurs étapes de synthèse et de maturation. En fonction de leur localisation dans le génome, la synthèse des miARN peut présenter quelques spécificités, en particulier pour les miARN localisés dans les introns de gènes codant des protéines. Le principe général de la synthèse est illustré dans la figure 1.2. Ils sont tout d'abord synthétisés sous la forme de

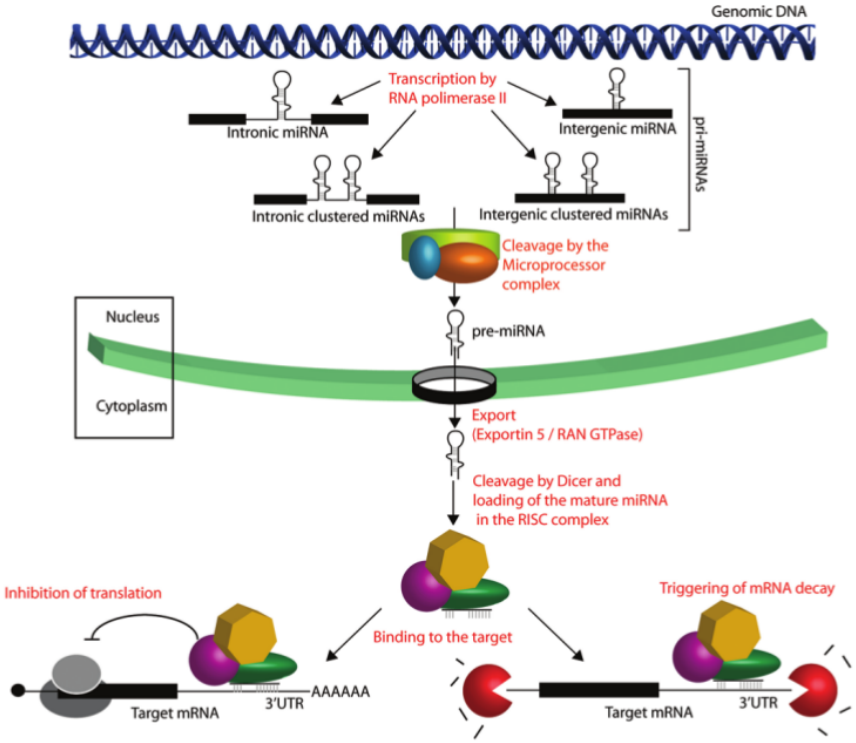

Figure 1.2 – Synthèse des miARN (figure 1 de Rosa et Brivanlou, 2009).

miARN primaires, plus longs que leur forme mature (plus de 1 kb). Ils sont ensuite clivés par l'endonucléase ARNase III Drosha au niveau de la base de la structure en tige-boucle (Lee et al., 2004). L'ARN ainsi formé d'environ 60 à 70 nucléotides est appelé précurseur de miARN et noté « pre-miARN ». Après avoir été exporté vers le cytoplasme par la protéine Exportin-5 (Lund et al., 2004), le pre-miARN subit un deuxième clivage. L'endonucléase ARNase III, Dicer, reconnaît la partie double brin du pre-miARN et le coupe à environ deux tours d'hélice à partir de la base de la structure en tige-boucle. La boucle est ainsi coupée, résultant en une structure en double brin (Hutvágner et al., 2001, Lee et al., 2004). Ce duplex est composé des deux brins du miARN : le brin qui constituera le miARN mature final et le brin opposé. Après leur dissociation par une hélicase, un seul des deux brins perdure et est incorporé dans un complexe ribonucléoprotéique, le complexe RISC (*RNA-Induced Silencing Complex*, Hutvágner et Zamore, 2002), qui constitue la forme active du miARN, permettant son action sur les ARNm messagers qu'il cible. L'autre brin, le brin opposé, est quant à lui dégradé. La stabilité des extrémités du duplex guide le choix du brin qui est incorporé dans RISC : il s'agit de celui dont l'extrémité 5' est la moins fortement appariée (Khvorova, Reynolds et Jayasena, 2003).

1.1.3 miRBase

La base de données miRBase (Kozomara et Griffiths-Jones, 2014) est la base de données de référence de séquences de miARN. En plus de référencer les séquences de miARN primaires et matures issus de différents organismes publiés dans la littérature scientifique, miRBase met à disposition un ensemble d'annotations telles que la localisation dans le génome ou une prédiction de structure en tige-boucle. Elle propose, en outre, une nomenclature et des identifiants uniques, les MIMAT ID, spécifiques aux miARN. Dans sa dernière version, miRBase répertorie 2 588 miARN matures humains.

1.2 Interactions miARN-ARNm

1.2.1 Reconnaissance de la cible

La liaison du complexe RISC aux ARN messagers ciblés est guidée par le miARN. Chez les animaux, il s'agit d'une partie de la séquence du miARN, une séquence de 7 nucléotides comprises entre les nucléotides 2 et 8 appelée séquence *seed*, qui spécifie l'interaction du complexe RISC dans la région 3'-UTR, *i.e.* la région non traduite en 3'(UTR pour *Untranslated Transcribed Region*), de l'ARN messager ciblé. Les miARN ayant la même séquence *seed* forment une famille de miARN et sont supposés avoir les mêmes cibles.

1.2.2 Actions

La reconnaissance d'un ARNm cible par un miARN induit soit l'inhibition de la traduction de l'ARNm ciblé soit sa dégradation (figure 1.3). Il est communément ad-

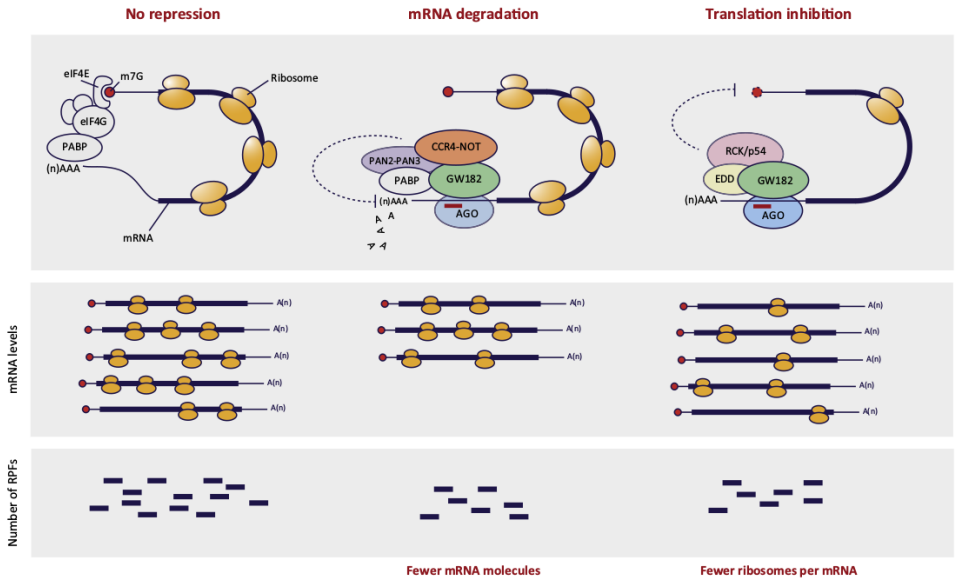

Figure 1.3 – Modes d'action des miARN sur leurs ARNm cibles et effets sur leur traduction (figure 1 de Vidigal et Ventura, 2015) : pas d'action des miARN (à gauche), le miARN induit la dégradation de son ARNm cible (au centre), le miARN induit l'inhibition de la traduction de son ARNm cible (à droite).

mis que les miARN contribuent à la répression de l'expression de leurs ARNm cibles mais les mécanismes moléculaires impliqués ne sont pas tous clairement établis (Fabian et Sonenberg, 2012). En particulier, le mode principal d'action, dégradation

de l'ARN messager cible ou inhibition de sa traduction, n'est pas complètement compris. Bien que leur rôle soit de réprimer l'expression d'ARNm, les miARN peuvent aussi plus rarement les activer (Catalanotto, Cogoni et Zardo, 2016).
Les interactions mises en place par les protéines constitutives du complexe RISC et d'autres cofacteurs semblent déterminer le mode d'action de celui-ci. La protéine GW182 peut ainsi interagir directement ou indirectement avec des complexes de désadénylation tels que CCR4-NOT et entraîner la dégradation de l'ARNm ciblé (Chen et al., 2009). Cette même protéine peut aussi interagir avec la protéine EDD qui, associée avec l'hélicase RCK/p54, permet le clivage de la coiffe de l'ARNm. Cette dernière étant nécessaire à la traduction, la traduction coiffe-dépendante de l'ARNm est alors réprimée (Su et al., 2011). Des études basées sur le séquençage d'ARNm et de fragments protégés par les ribosomes ont permis de montrer que le mode d'action principal des miARN chez l'humain est la dégradation de l'ARNm ciblé (Guo et al., 2010). Ainsi, l'action d'un miARN peut être estimée par une mesure de la diminution de l'expression de ses ARNm cibles.

1.2.3 Base de données d'interaction miARN-ARNm

Il existe de nombreuses bases de données et de logiciels de prédictions *in silico* d'interaction miARN-ARNm. Ces prédictions sont basées sur une multitude de caractéristiques différentes. Parmi les plus répandues, on peut citer :

- l'appariement de la séquence *seed*,
- la conservation inter-espèces des séquences *seed*,
- l'énergie libre,
- l'accessibilité du site.

L'appariement de la séquence *seed* du miARN avec l'ARNm cible doit résulter en un alignement parfait constitué d'appariements Watson-Crick, *i.e.* adénosine-uracil (A-U) et guanine-cytosine (G-C), sans insertion-délétion. La longueur de cet alignement peut varier entre 6 et 8 paires de bases en fonction des bases de prédictions.

Du fait de leur rôle crucial dans la reconnaissance des cibles, les séquences *seed* sont très conservées entre différentes espèces. Il en est de même pour les séquences en 3'-UTR des ARNm avec lesquelles les séquences *seed* interagissent. Dans des cas moins fréquents, des appariements ont lieu dans la région 3' du miARN avec sa cible pour compenser un mésalignement de la séquence *seed*, se traduisant également par la présence de ces séquences particulières dans différentes espèces. Ainsi, certains outils prennent en compte la conservation de séquences à ces différentes localisations sur plusieurs espèces pour prédire une interaction miARN-ARNm.

La stabilité de la liaison entre un miARN et sa cible est un indicateur que l'interaction prédite est une réelle interaction. Elle peut se mesurer par un changement d'énergie libre (ΔG), les régions impliquées dans des appariements ayant une faible énergie libre alors que les régions non appariées en ont une forte. Par exemple, pour une structure en tige-boucle, la région de la tige a une faible énergie libre alors que la région de la boucle a une énergie libre élevée. Dans le cadre de la prédiction d'interaction miARN-ARNm, l'énergie libre de l'appariement du miARN à sa cible peut être calculée et utilisée comme un indicateur de la véracité de cette interaction.

La structure secondaire des ARNm peut nuire à la liaison d'un miARN dans la région 3'-UTR. Cette liaison se fait généralement en deux étapes : le miARN s'hybride à sa

cible dans une courte région accessible avant de déclencher un changement de conformation de la structure de l'ARNm pour accomplir sa liaison complète. Ainsi, rendre un site de liaison accessible requiert de l'énergie qui peut être évaluée et utilisée pour estimer la vraisemblance d'une interaction miARN-ARNm.

Dans le cadre de cette thèse, trois bases de prédictions *in silico* d'interaction miARN-ARNm ont été utilisées : miRANDA (John et al., 2004), TargetScan (Agarwal et al., 2015) et PITA (Kertesz et al., 2007) dont les principales caractéristiques sur lesquelles sont basées les prédictions sont listées dans la table 1.1. Il existe bien

Bases de prédictions	***Seed***	**Conservation**	**Energie libre**	**Accessibilité**
miRANDA	✓	✓	✓	
PITA	✓	✓	✓	✓
TargetScan	✓	✓		

Table 1.1 – Principales caractéristiques des bases de prédictions d'interaction miARN-ARNm miRANDA, PITA et TargetScan.

d'autres bases de prédictions *in silico* d'interactions miARN-ARNm, chacune ayant ses propres spécificités, et ayant fait l'objet d'études comparatives (Riffo-Campos, Riquelme et Brebi-Mieville, 2016).

D'autres bases de données répertorient les interactions miARN-ARNm validées expérimentalement. La base de données miRTarBase (Chou et al., 2018) rassemble de manière automatisée les résultats d'études fonctionnelles mettant en évidence des interactions miARN-ARNm. Pour ce faire, les articles scientifiques d'étude d'interactions miARN-ARNm sont téléchargés depuis PubMed et sont analysés par des méthodes de fouille de textes pour extraire les miARN et les ARNm impliqués dans ces interactions. Les informations complémentaires, comme les techniques expérimentales ayant permis d'identifier les interactions miARN-ARNm, sont extraites manuellement. MirTarBase indique un degré de confiance pour chaque interaction en fonction de la technique expérimentale utilisée : fort pour les interactions identifiées par constructions reportrices (*reporter assays*), *western blot* ou qRT-PCR (*quantitative Reverse Transcription Polymerase Chain Reaction*) et faible pour celles révélées par des approches de séquençage et d'immunoprécipitation (CLIP-seq) ou de puces à ADN. Dans sa dernière version, miRTarBase contient 380 639 interactions miARN-ARNm dont 14 052 avec de fortes preuves expérimentales.

1.3 Rôle tampon des miARN

Les miARN peuvent cibler des dizaines d'ARNm différents et, à l'inverse, un ARNm peut être ciblé par plusieurs miARN différents. Les miARN se retrouvent ainsi impliqués dans de grands réseaux de régulation d'expression de gènes. Avec l'action conjointe de facteurs de transcription, les miARN peuvent former des boucles *feed-forward* dans lesquelles un miARN et son ARNm cible sont co-induits par un facteur de transcription (figure 1.4). Ce genre de boucle de régulation permet un contrôle précis de l'expression de gènes. Il a ainsi été montré que les miARN confèrent de la précision dans l'expression des gènes qu'ils ciblent (Schmiedel et al., 2015).

La robustesse est définie comme la capacité des systèmes biologiques à maintenir

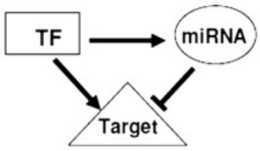

FIGURE 1.4 – Boucle *feed-forward* de régulation (portion de la figure 1 de LU et CLARK, 2012) : un miARN et son ARNm cible (*Target*) sont co-induits par un facteur de transcription (TF). L'activation de l'expression de l'ARNm par le facteur de transcription est contrebalancée par l'effet indirect du miARN. Dans cette situation, l'expression de l'ARNm peut être contrôlée finement.

des fonctions spécifiques lorsqu'ils sont exposés à des perturbations internes ou externes (KITANO, 2004). Les miARN sont ainsi suspectés de contribuer à la robustesse des systèmes biologiques (EBERT et SHARP, 2012). SICILIANO et al., 2013 ont effectivement apporté des preuves expérimentales, à l'aide de constructions reportrices, que les miARN contribuaient au contrôle de l'expression de leurs cibles en limitant leurs fluctuations, considérées comme du « bruit » dans leurs travaux.

L'action des miARN entraîne rarement l'extinction totale de l'expression de ses ARNm cibles mais le plus généralement seulement une diminution de celle-ci (SEGGERSON, TANG et MOSS, 2002). Les expériences d'extinction, ou *knockouts*, de miARN ne causent en effet généralement que de faibles changements phénotypiques (MISKA et al., 2007). Par exemple, chez *Caenorhabditis elegans*, moins de 10% des miARN sont individuellement nécessaires au développement normal ou à la viabilité de ce petit ver. Jusqu'à présent, seuls deux miARN, miR-17-92 et miR-96, semblent causer des défauts de développement chez l'Homme lorsqu'ils subissent une mutatation (MENCIA et al., 2009). Il semble donc que, plutôt que d'agir comme des interrupteurs génétiques majeurs, l'action des miARN soit semblable à celle d'un rhéostat, ajustant de façon synergique et fine l'expression de centaines de gènes codant des protéines (BARTEL et CHEN, 2004).

Ces différentes caractéristiques, fine régulation de l'expression des ARNm, grand nombre de cibles et contribution à la robustesse des système biologiques, confèrent aux miARN un rôle tampon. En modulant l'expression des miARN, la cellule peut s'adapter aux perturbations des conditions environnementales qu'elle subit. YANG et al., 2012 ont ainsi montré que l'expression des gènes ciblés par des miARN a tendance à être plus stable que celle de gènes qui ne sont pas ciblés par des miARN.

Le cancer peut être vu comme une perte de robustesse (BEN-DAYAN et al., 2015) ou, au contraire, comme un système robuste dans la mesure où les tissus cancéreux ne sont que peu affectés par les stimulis environnementaux (KITANO, 2004). Les miARN permettant de conférer de la robustesse aux systèmes biologiques dans des conditions non pathologiques, il apparaît pertinent d'analyser leur rôle dans la cancérogénèse.

1.4 Rôle dans le cancer

1.4.1 OncomiRs et miARN tumeurs suppresseurs

Du fait de leur grand nombre et du nombre important de gènes dont chaque miARN peut réguler l'expression, les miARN sont impliqués dans de nombreuses

fonctions cellulaires telles que la différenciation (Ivey et Srivastava, 2010) ou l'apoptose (Garofalo et al., 2010). Une dérégulation de leur expression peut ainsi avoir d'importantes conséquences et entraîner le développement de maladies telles que le cancer (Calin et Croce, 2006). Hanahan et Weinberg, 2000 ont décrit six caractéristiques majeures de la progression tumorale : l'auto-suffisance en facteurs de croissance, l'insensibilité aux signaux inhibant la croissance, l'échappement à l'apoptose, le potentiel de réplication sans limite, le soutien de l'angiogénèse et l'invasion tissulaire. En ciblant des gènes dont l'expression favorise l'une de ces fonctions moléculaires, *i.e.* des oncogènes, certains miARN ont une action qui prévient la survenue de cancers. On parle alors de miARN tumeurs suppresseurs. A l'inverse, d'autres miARN qui ciblent des gènes qui répriment ces fonctions moléculaires, ont une action favorable au développement de tumeurs et sont appelés oncomiRs (figure 1.5). Par

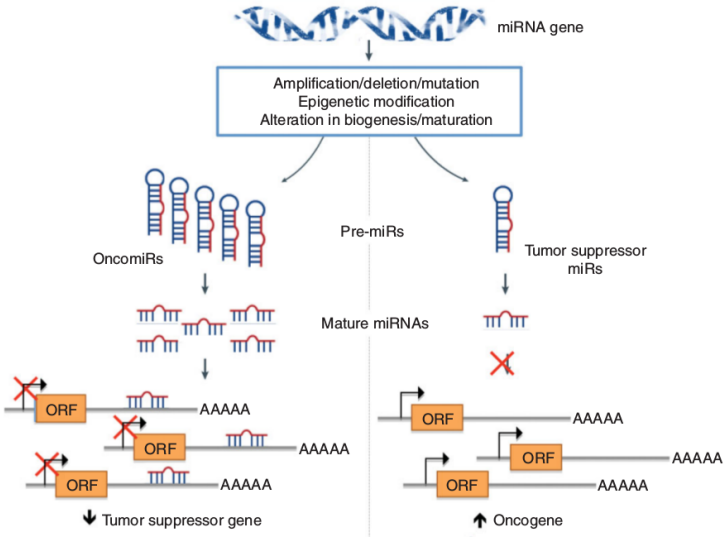

Figure 1.5 – Dérégulation de l'expression de miARN dans la cancérogénèse (portion de la figure 1 de Shah et Calin, 2014).

exemple, les miARN du cluster miR-17-92, dont la sur-expression dans les lymphomes et les leucémies favorise la prolifération et l'angiogénèse, ont été parmi les premiers oncomiRs identifiés (He et al., 2005). A l'inverse, miR-30-5p agit comme un tumeur suppresseur en limitant la prolifération. Sa répression est souvent observée dans les myélomes (Zhao et al., 2014a). De nombreux autres exemples de miARN impliqués dans le développement tumoral sont répertoriés dans la base de données miRCancer (Xie et al., 2013).

1.4.2 Outils de diagnostic

Excrétés par les cellules, les miARN sont présents dans de nombreux fluides tels que le plasma, le sérum, la salive ou le lait. Pour éviter d'être dégradés par des enzymes qui sont abondantes dans ces fluides, ces miARN sont associés dans des complexes avec des protéines Argonaute-2 (Turchinovich et al., 2011) ou incorporés dans des structures comme des exososomes, microvésicules ou complexes lipo-protéiques. Ces miARN non cellulaires, appelés miARN circulants, sont suspectés d'avoir un rôle dans la communication inter-cellulaire (Hrustincova, Votavova et Dostalova Merkerova, 2015). De la même manière que les miARN cellulaires, les miARN circulants pourraient refléter un état pathophysiologique et donc servir de biomarqueurs de diagnostic ou de prognostic. Par exemple, le miARN miR-26a dans le plasma a une valeur

de diagnostic dans le cas du cancer gastrique (Qiu et al., 2016) alors qu'une concentration plasmatique élevée de miR-19a a un facteur de prognostic du cancer du sein métastatique inflammatoire HER2+ (Anfossi et al., 2014). Ainsi, de nombreux essais sont en cours dans le but d'identifier de nouveaux miARN circulants biomarqueurs de différents cancers.

Leur capacité à refléter un état pathologique, leur grande stabilité et la facilité avec laquelle ils peuvent être détectés par des méthodes non invasives font des miARN circulants des outils potentiellement intéressants de diagnostic. Malheureusement, leur faible spécificité freine considérablement leur utilisation comme biomarqueurs. De nombreux miARN circulants sont en effet détectés dans différentes conditions physiopathologiques. Par exemple, le miARN circulant miR-141 a à la fois été détecté dans le plasma de femmes enceintes et d'hommes atteints par le cancer de la prostate (Mitchell et al., 2008).

1.4.3 Cibles thérapeutiques

Si les miARN peuvent être impliqués dans le développement tumoral à cause de leur action dans de nombreuses fonctions cellulaires (voir section 1.4.1), ils constituent autant de cibles thérapeutiques éventuelles. Différentes stratégies thérapeutiques ont ainsi été développées. L'approche la plus commune vise à inhiber l'action de miARN en employant des oligonuclétotides inhibiteurs (Krutzfeldt et al., 2005) ou des éponges à miARN composés de plusieurs sites de liaison spécifiques des miARN ciblés (Ebert, Neilson et Sharp, 2007). Ces structures ont la capacité de se lier au miARN ciblé, réduisant ainsi son action sur ses ARNm cibles. Le potentiel thérapeutique a ainsi pu être démontré dans différents cancers (Silber et al., 2008).

Les principales difficultés de l'application de thérapies basées sur les miARN résident dans la spécificité du tissu ciblé et la toxicité. L'utilisation de structures telles que les vecteurs viraux ou les nanoparticules permettent de surpasser cette première difficulté (Malik et Roy, 2008). La capacité des miARN à réguler l'expression d'une multitude de gènes peut avoir des effets indésirables, dont certains potentiellement toxiques, ce qui constitue une limitation importante de ce type d'approches thérapeutiques (Aagaard et Rossi, 2007).

2 Données d'expression de gènes

2.1 Du séquençage aux données d'expression

L'expression d'un gène peut être estimée par la quantité d'ARNm correspondant récupéré dans une condition biologique donnée. Au cours de la dernière décennie, le séquençage d'ARN ou RNA-seq (*RNA-sequencing*) a supplanté les puces à ADN comme technologie standard de mesure de l'expression de gènes. Cette section vise à présenter les différentes étapes qui mènent du séquençage de l'ARN récupéré dans un échantillon à la quantification de l'expression des gènes. La principale application du RNA-seq est le séquençage d'ARNm. Des adaptations ont été développées pour pouvoir séquencer les petits ARN, dont les miARN font partie, et seront également présentées dans cette section. Enfin, des exemples de technologies et de logiciels seront pris pour illustrer les différentes étapes. Ces exemples représentent les choix qui ont été faits pour la construction de la base de données d'expression GDC (*Genomic Data*

Commons), une base de données d'expression de gènes issues d'échantillons tumoraux et sains de référence.

2.1.1 RNA-seq

Séquençage Le RNAseq est une adaptation du processus classique de séquençage haut débit illustré dans la figure 1.6. Le séquençage haut débit consiste à fragmen-

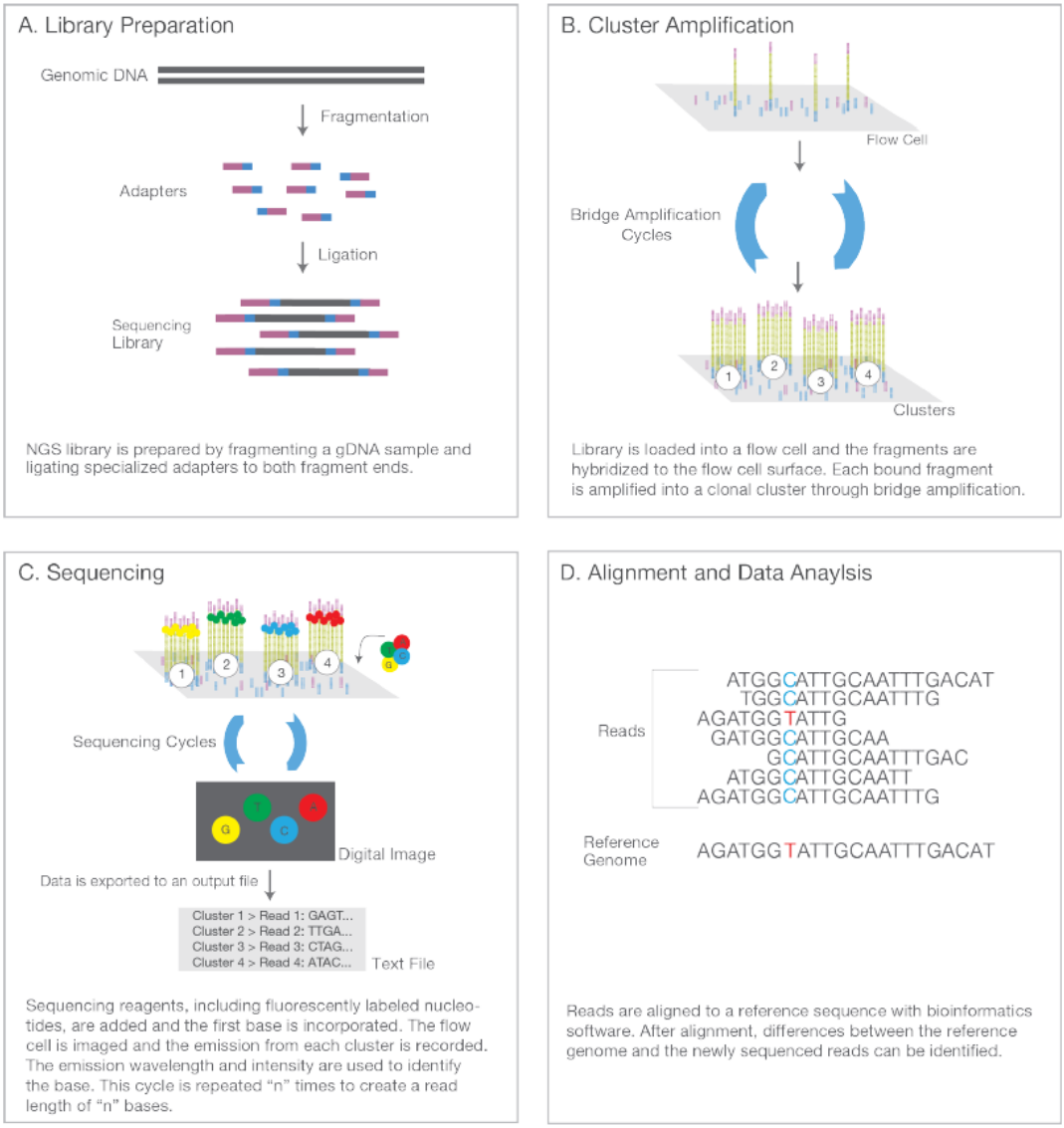

FIGURE 1.6 – Processus de séquençage haut débit suivi par les séquenceurs Illumina (ILLUMINA, 2017).

ter l'ADN présent dans un échantillon, fixer les fragments générés sur une puce, les amplifier de façon massive et parallèle avant de réaliser la réaction de séquençage à proprement parlé.
Le RNAseq vise à séquencer uniquement l'ensemble des transcrits (ARNm) présents dans un échantillon. La spécificité du RNAseq se manifeste essentiellement lors de l'étape de préparation de la librairie, c'est-à-dire l'ensemble des fragments d'ADN à séquencer. Des adaptateurs avec des oligonucléotides poly-T, complémentaires de l'extrêmité 3' poly-adénylée des ARNm, sont utilisés dans le but de ne conserver que les ARNm. Ceux-ci sont ensuite convertis en ADN complémentaire (ADNc) par une

réaction de transcription inverse. La librairie est ainsi constituée et la suite du processus de séquençage réalisée.
Les fragments d'ADN sont fixés sur la puce de séquençage puis amplifiés par PCR (Polymerase Chain Reaction). Le séquençage a ensuite lieu grâce à l'incorporation successive de nucléotides couplés à des agents fluorescents. La fluorescence émise permet de déterminer quel nucléotide a été incorporé à chaque cycle. La séquence des fragments d'ADN de la librairie est ainsi déterminée à la fin des réactions de séquençage. Cette séquence lue par le séquenceur est communément appelée *read*.
Dans le cadre du GDC, la plateforme de séquençage utilisée est l'Illumina HiSeq.

Alignement de *reads* Plusieurs traitements informatiques sont nécessaires pour pouvoir estimer des valeurs d'expression à partir des *reads* issus du séquençage. La suite classique de ces traitements commence par l'alignement des *reads* contre une séquence de référence avant de pouvoir estimer l'expression des gènes. Ces différentes étapes, ainsi que les logiciels utilisés par le GDC pour les accomplir, sont illustrés dans la figure 1.7.

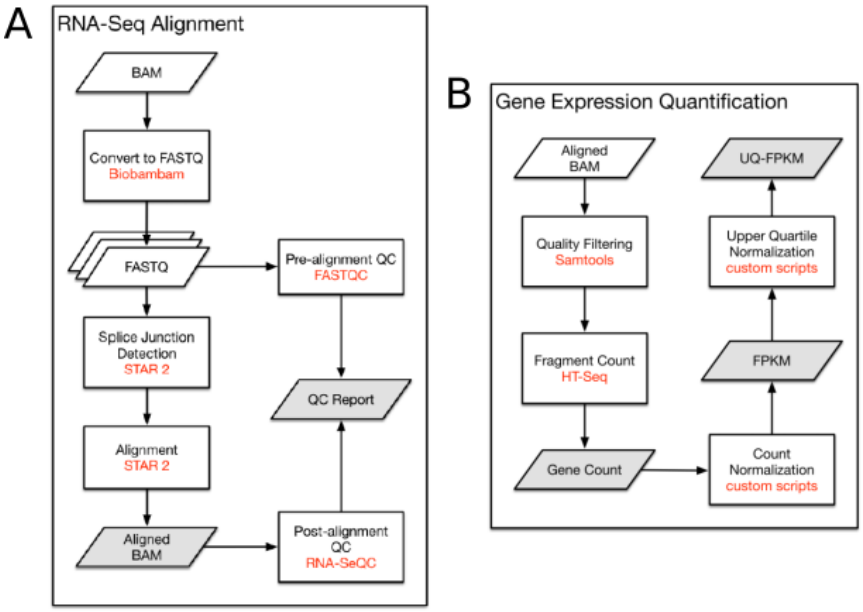

FIGURE 1.7 – *Workflows* d'alignement de *reads* (A) et de quantification d'expression (B) suivis par le GDC pour quantifier l'expression de gènes à partir du séquençage d'échantillons (`https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/`).

La suite d'opérations permettant d'aligner les *reads* contre une séquence de référence est illustré dans le diagramme de la partie A de la figure 1.7. Avant de procéder à l'alignement de *reads*, la qualité d'une expérience de séquençage d'un échantillon est évaluée à l'aide du logiciel FASTQC (ANDREWS, 2010). Les *reads* ayant passé le contrôle qualité sont ensuite alignés contre le génome de référence GRCh38 (GENOME REFERENCE CONSORTIUM, 2017) en utilisant la méthode « 2-pass » du logiciel STAR (*Spliced Transcripts Alignment to a Reference*, DOBIN et al., 2013). STAR aligne des *reads* contre une séquence de référence en recherchant les plus grandes sous-séquences contiguës alignables à l'aide de tableaux de suffixes, une structure de données très proche du FM-index utilisée par d'autres logiciels d'alignement de *reads* fréquemment utilisés tels que BWA (LI et DURBIN, 2009) ou BOWTIE2 (LANGMEAD et SALZBERG, 2012). La méthode « 2-pass » permet une détection plus fine de nouvelles jonctions

intron-exon issues de l'épissage alternatif. Pour ce faire, l'alignement de *reads* est réalisé en deux itérations : la première permet de détecter de nouvelles jonctions et la deuxième intègre les nouvelles jonctions en éditant le génome de référence utilisé pour l'alignement des *reads*. Les *reads* ainsi alignés, ainsi que ceux qui n'ont pas pu l'être, sont fournis dans un fichier au format BAM (Li et al., 2009), le format standard pour enregistrer des alignements de *reads*. La qualité de l'alignement des *reads* est ensuite évaluée à l'aide des logiciels RNA-SeQC (DeLuca et al., 2012) et Picard Tools (Broad Institute, 2015).

Quantification d'expression Les différentes étapes permettant la quantification de l'expression de gènes sont représentées dans le diagramme de la partie B de la figure 1.7. Elle peut être estimée à partir de l'alignement de *reads* précédemment effectué. Pour ce faire, les annotations présentes dans la version 22 de GENCODE (Harrow et al., 2012) sont utilisées. Ainsi, l'expression des gènes codant pour des protéines, mais aussi celle de longs ARN non codants et plus généralement celle de toutes les annotations présentes dans GENCODE, est estimée à partir du nombre de *reads* alignés dans les régions génomiques de chacune de ces annotations. La fonction *htseq-count* du logiciel HT-Seq (Anders, Pyl et Huber, 2015) est utilisée pour effectuer le comptage des *reads* par annotation. Cette fonction compte le nombre de *reads* alignés dans chaque intervalle, ou union d'intervalles, de positions définies par les annotations de GENCODE.
La plupart des *reads* ne sont alignés que contre une seule annotation mais il arrive que certains puissent être affectés à plusieurs annotations lorsque celles-ci sont très proches ou chevauchantes. Plusieurs exemples sont illustrés dans la figure 1.8. Pour at-

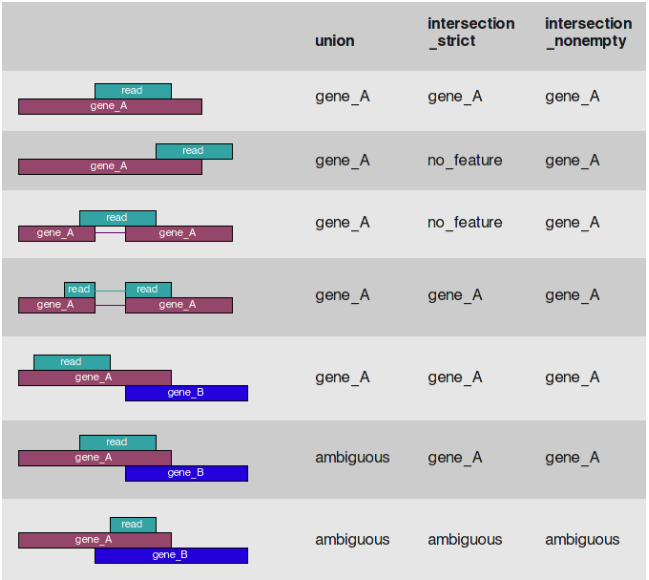

Figure 1.8 – Attribution de *read* à une ou plusieurs annotations. Trois stratégies peuvent être suivies pour les cas particuliers où un *read* est aligné contre une annotation, une région non annotée ou plusieurs annotations : « *union* », « *intersect_strict* » et « *intersect_nonempty* ». La stratégie suivie par le GDC est la stratégie « *intersect_nonempty* ».

tribuer un *read* à une annotation, les annotations couvrant chaque position d'un *read* sont déterminées. Différentes méthodes sont applicables pour attribuer un *read* à une annotation, plusieurs annotations (« *ambiguous* ») ou aucune (« *no_feature* »). Pour générer ses données d'expression, le GDC utilise l'option « *intersection_nonempty* ».

L'intersection des ensembles d'annotations non-vides pour chaque position d'un *read* est considérée pour attribuer un *read* à une annotation.
Le nombre brut de *reads* attribués à chaque annotation de GENCODE v22 est ainsi obtenu. En plus de ces valeurs brutes, des valeurs normalisées sont aussi fournies pour chaque annotation : les valeurs FPKM (*Fragments per Kilobase of transcript per Million mapped reads*) et FPKM-UQ (*Upper Quartile FPKM*). Les valeurs FPKM donnent le compte brut du nombre de *reads* alignés contre un gène divisé par sa longueur et le nombre total de *reads* alignés contre des gènes codant pour des protéines. Ces valeurs sont calculées selon la formule :

$$FPKM = \frac{\text{nb\_}reads \ \times 10^9}{\text{nb\_}reads_{\ gp} \times L},$$

où :

- nb_*reads* est le nombre de *reads* alignés contre un gène ;
- nb_*reads* $_{gp}$ est le nombre total de *reads* alignés contre des gènes codant pour des protéines ;
- L est la longueur du gène ;
- le facteur 10^9 permet d'obtenir une valeur par million de *reads* séquencés et par milliers de bases.

Les valeurs FPKM-UQ sont des valeurs modifiées des valeurs FPKM. Plutôt que d'utiliser le nombre total de *reads* alignés contre des gènes codant pour des protéines, le nombre brut des *reads* est ici divisé par le 75ème percentile des nombres de *reads* alignés contre des gènes codant pour des protéines. Les valeurs FPKM-UQ sont calculées selon la formule :

$$FPKM-UQ = \frac{\text{nb\_}reads \ \times 10^9}{\text{nb\_}reads_{\ gp75} \times L}$$

où :

- nb_*reads* est le nombre de *reads* alignés contre un gène ;
- nb_*reads* $_{gp75}$ est le 75ème percentile des nombres de *reads* alignés contre des gènes codant pour des protéines ;
- L est la longueur du gène ;
- le facteur 10^9 permet d'obtenir une valeur par million de *reads* séquencés et par milliers de bases.

Présence de valeurs aberrantes Il peut arriver que certains gènes présentent des valeurs d'expression aberrantes très élevées dans certains échantillons qui se distinguent très nettement du reste des valeurs d'expression observées pour les autres échantillons qui constituent le jeu de données considéré. Ces valeurs aberrantes sont généralement la conséquence d'un problème technique survenu au cours du séquençage de ces échantillons et ne revêtent donc pas de sens biologique. On parle alors d'« effets *batch* ». Par exemple, elles peuvent être due à un biais lors de l'étape d'amplification des fragments d'ADN complémentaire. De plus, à cause de leur côté exceptionnel, elles peuvent amener à des conclusions erronées lors des analyses statistiques ultérieures. Pour ces deux raisons, ces valeurs aberrantes, également appelées *outliers*, doivent être l'objet d'un effort particulier dans toute analyse de données RNA-seq pour les identifier au mieux et ne pas nuire aux analyses.

2.1.2 miRNA-seq

Un exemple de suite d'opérations permettant la quantification de l'expression de miARN à l'aide du séquençage haut débit est illustré dans la figure 1.9.

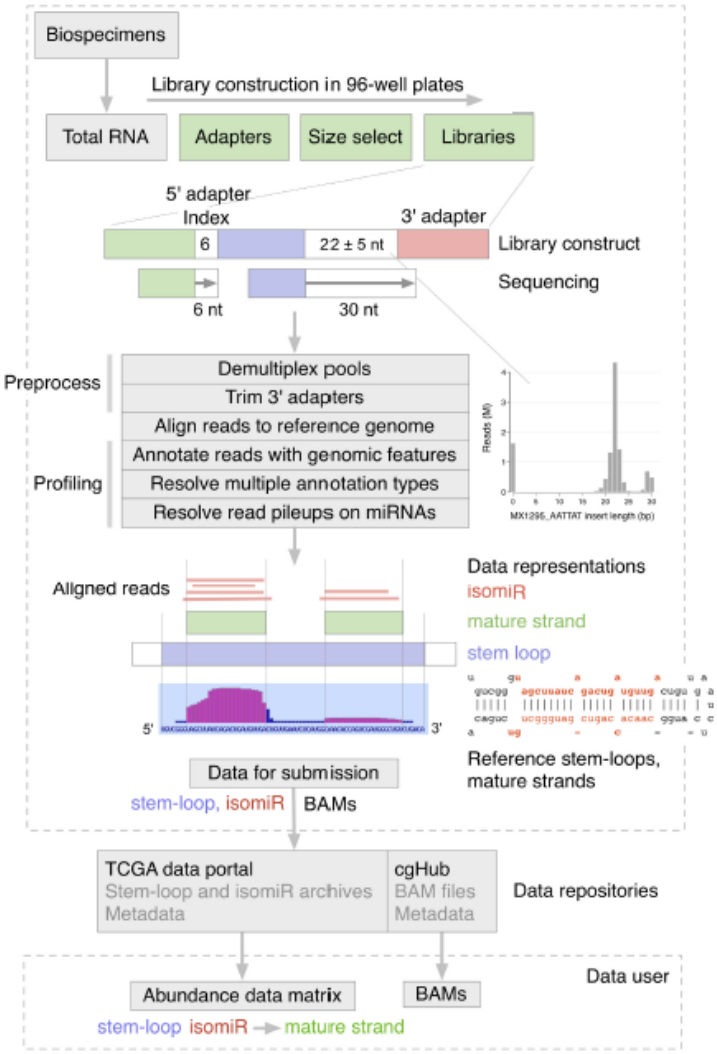

FIGURE 1.9 – *Workflows* de séquençage, d'alignement de *reads* et de quantification d'expression suivis par le GDC pour quantifier l'expression de miARN (CHU et al., 2016).

Séquençage La taille des *reads* de séquençage, plusieurs dizaines de bases (voire quelques centaines), étant plus grande que la taille moyenne des miARN, le workflow de séquençage haut débit nécessite quelques adaptations. Ces modifications concernent principalement l'étape de préparation de la librairie, c'est-à-dire l'ensemble des fragments d'ADN à séquencer. Plutôt que d'utiliser le contenu total d'un échantillon en ARN, les ARNm sont filtrés pour ne conserver que les petits ARN. Des adaptateurs, nécessaires à l'amplication des ARN, sont ensuite liés aux extrémités 5' et 3' à l'aide d'ARN ligases T4. L'adaptateur utilisé en 3' est spécifique des miARN, ce qui permettra de ne séquencer que ce type d'ARN parmi l'ensemble des petits ARN présents dans l'échantillon. La qualité des librairies préparées est évaluée à partir de la taille des fragments d'ADN obtenus. On s'attend en effet à obtenir une distribution

de la taille d'insert, *i.e.* le fragment d'ADN compris entre les adaptateurs, centrée autour de la taille moyenne d'un miARN, soit environ 22 bases.
Les librairies ayant une distribution de taille d'insert proche de celle attendue sont ensuite séquencées à l'aide de *reads* de 30 bases en suivant une méthode classique de séquençage haut débit : amplification de la librairie par PCR, multiplexage de différentes librairies issues de plusieurs échantillons grâce à l'utilisation de séquences d'index et séquençage à proprement parler de plusieurs librairies sur la même puce de séquençage.

Alignement de *reads* Les *reads* séquencés étant plus longs que les miARN, ils peuvent contenir en leur extrémité 3' des fragments d'adaptateur (voir figure 1.9). Cette partie des *reads*, qui n'est pas d'intérêt biologique, doit être supprimée lors d'une étape de *trimming* avant de procéder à l'alignement de *reads*. Les miARN n'étant pas sujet à l'épissage alternatif, l'alignement de *reads* ne requiert pas ici d'étape de détection de nouvelles jonctions intron-exon comme pour les ARNm. Le logiciel BWA-aln (Li et Durbin, 2009) est ainsi utilisé pour aligner les *reads* contre le génome de référence GRCh38.

Quantification d'expression Différents fichiers d'annotations sont utilisés pour déterminer quels types de petits ARN ont été séquencés et quantifier leur expression. Ainsi, les *reads* sont comparés aux annotations issues de miRBase v21 (Kozomara et Griffiths-Jones, 2014) pour quantifier l'expression de miARN, à celles de l'UCSC (*University of California Santa Cruz*) Genome Browser database (Rosenbloom et al., 2015) pour quantifier celle des petits ARN nucléolaires (snoARN). D'autres types d'ARN non codants sont aussi évalués tels que les ARN de transfert (ARNt), les ARN ribosomiques (ARNr) ou les petits ARN nucléaires (snARN). Un alignement parfait entre un *read* et une région du génome de référence est requis pour quantifier l'expression d'un miARN, *i.e.* aucun mésalignement n'est toléré.
De la même manière que pour les ARNm, un *read* peut être aligné contre une région génomique correspondant à plusieurs annotations. Une seule annotation est retenue en se basant sur une liste de priorités entre les différents types de petits ARN (voir table 1 de Chu et al., 2016). Dans cette liste de priorités, les miARN ont la priorité la plus élevée.
Un *read* peut aussi s'aligner de manière parfaite en différentes régions du génome. Cette situation est d'autant plus fréquente, par comparaison avec le séquençage d'ARNm, que les *reads* de miRNAseq sont particulièrement courts. En particulier, un *read* peut s'aligner contre différents miARN. S'il s'agit de régions génomiques différentes avec des séquences identiques, aboutissant donc au même miARN mature, alors le *read* est affecté à ce miARN mature et aléatoirement à l'une des régions génomiques. Si un *read* est aligné contre des miARN ayant des séquences différentes, alors ce *read* est considéré comme étant « *cross-mapped* ». Il est alors affecté à l'ensemble de ces miARN.
L'expression des miARN est ensuite calculée en faisant la somme des *reads* alignés contre chacune des annotations contenues dans miRBase. Une version normalisée du nombre brut de *reads* alignés contre chaque miARN par million de *reads* alignés contre des miARN est aussi fournie. Enfin, il est aussi indiqué si des *reads* « *cross-mapped* » figurent parmi les *reads* alignés contre chaque miARN. Ces données sont disponibles à la fois pour les gènes précurseurs de miARN mais aussi pour les miARN matures et leurs isoformes.

2.1.3 Type de données

Que ce soit pour les ARNm ou les miARN, les données d'expression de gènes issues du séquençage sont généralement représentées sous la forme d'une matrice où les lignes contiennent les nombres de *reads* associés à un gène et les colonnes contiennent les nombres de *reads* associés à un échantillon :

$$
\begin{array}{c}
 \\
\text{gene}_1 \\
\text{gene}_2 \\
\vdots \\
\text{gene}_a \\
\text{gene}_b \\
\vdots \\
\text{gene}_{m-1} \\
\text{gene}_G
\end{array}
\begin{array}{c}
\begin{array}{ccccccc}
\text{e}_1 & \text{e}_2 & \dots & \text{e}_k & e_{k+1} & \dots & \text{e}_n
\end{array} \\
\left(\begin{array}{ccccccc}
33 & 72 & \dots & 106 & 64 & \dots & 73 \\
11682 & 854 & \dots & 5740 & 1758 & \dots & 10450 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & \dots & 0 & 0 & \dots & 0 \\
0 & 0 & \dots & 0 & 54 & \dots & 89 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 16 & \dots & 4 & 2 & \dots & 0 \\
43 & 21 & \dots & 0 & 0 & \dots & 76
\end{array}\right)
\end{array}
$$

Ces matrices peuvent contenir jusqu'à quelques dizaines de milliers de lignes en fonction du type de séquençage. Le nombre de colonnes est en général toujours beaucoup plus petit que le nombre de lignes. En effet, les études d'expression basées sur le RNA-seq ne contiennent en général que quelques échantillons par condition d'intérêt. Cependant, avec la diminution continuelle du coût du séquençage, le développement de grandes bases de données de séquençage publiques et des techniques d'aggrégation de différentes études, ce nombre tend à augmenter et peut atteindre quelques dizaines, voire quelques centaines d'échantillons. Néanmoins, le nombre d'échantillons reste nettement inférieur à celui de gènes.

2.2 Le *Genomic Data Commons*

Le GDC (*Genomic Data Commons*, National Cancer Institute, 2017a) est un espace public de partage de données génomiques sur le cancer. Il vise à mettre à disposition des chercheurs des données issues de différents programmes de recherche du NCI (*National Cancer Institute*) *Center for Cancer Genomics* : TCGA (*The Cancer Genome Atlas*, National Cancer Institute, 2017b) et TARGET (*Therapeutically Applicable Research to Generate Effective Treatments*, National Cancer Institute, 2017c). TCGA vise à caractériser les changements génomiques au cours du développement tumoral pour les types et sous-types majoritaires de cancer. Cette base de données agrège des données issues de biopsies de tissus cancéreux et de tissus sains appariés provenant de plus de 11 000 patients et représentant 33 types et sous-types de cancer. TARGET est une source de données plus spécifique. En effet, elle s'intéresse en particulier aux cancers qui affectent les enfants et vise à découvrir des solutions thérapeutiques applicables au domaine clinique.

Les données fournies par le GDC incluent des données de mutations, de variation de nombre de copies de gènes, de quantification d'expression de gènes et de modifications post-transcriptionnelles. La figure 1.10 représente le nombre de cas ('*Case*' dans les dénominations utilisées par le GDC), c'est-à-dire l'ensemble de toutes les données disponibles issues d'un patient pour un type ou sous-type de cancer, actuellement disponibles dans le GDC. Avec des données issues de centaines de patients pour les cancers les plus documentés, le GDC constitue une source de données génomiques sur le cancer volumineuse et extrêmement précieuse. L'ensemble de ces données a été

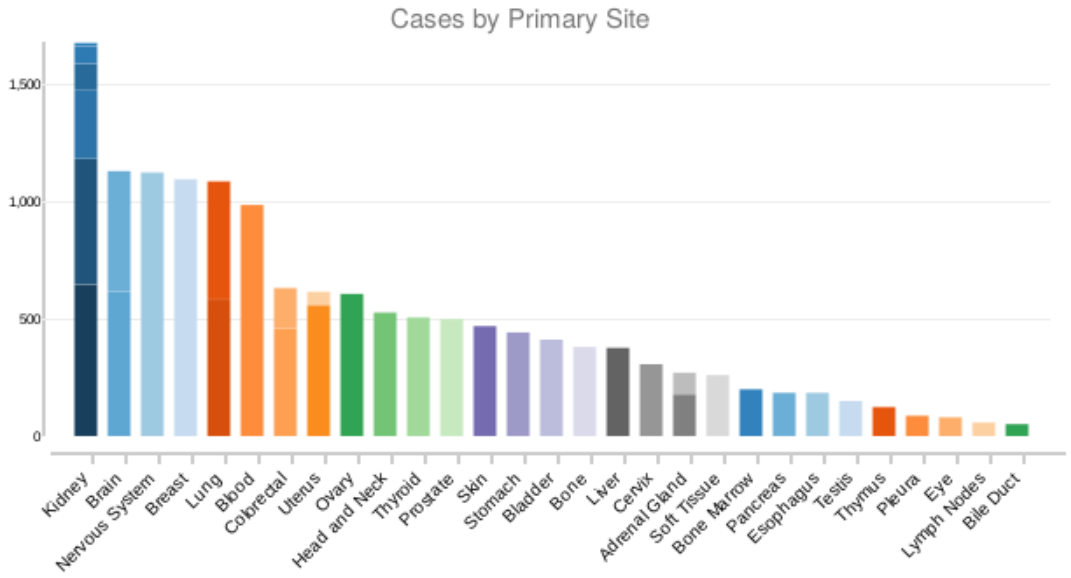

FIGURE 1.10 – Nombre de patients par type ou sous-type de cancer pour lesquels le GDC fournit des données incluant des données de mutations, de variation de copies de gènes, de quantification d'expression de gènes et de modifications post-transcriptionnelles (NATIONAL CANCER INSTITUTE, 2017a).

harmonisé par l'utilisation de technologies de séquençage et d'outils bioinformatiques communs dans le but de faciliter la comparaison de données issus de projets de recherche différents. Le GDC, par le volume et la standardisation des données qu'il fournit, constitue ainsi une source de données de haute qualité pour caractériser le développement cancéreux au niveau génomique.

En plus de données issus d'échantillons provenant de la partie cancéreuse du tissu, le GDC fournit aussi les mêmes types de données pour des échantillons de tissus sains adjacents à certains de ces échantillons de tissus cancéreux provenant du même donneur. Par la suite dans ce document, nous parlerons d'échantillons tumoraux ou sains. La figure 1.11 représente les nombres d'échantillons tumoraux et sains pour les tissus disposant des plus grands nombres d'échantillons sains. Pour la majorité des tissus, il y a dix fois moins d'échantillons sains que d'échantillons tumoraux. Bien que peu nombreux, ces échantillons issus de tissus sains ont une valeur essentielle car ils peuvent servir de contrôles et permettre de mieux caractériser le développement tumoral pour un tissu donné.

Les variations spatio-temporelles de l'expression du génome contrôlent le devenir cellulaire. La comparaison de données d'expression entre deux conditions permettent d'identifier les changements qui s'opèrent au niveau de l'expression génique. Dans le cas du cancer, il s'agit de comparer des données d'expression de gènes entre échantillons tumoraux et échantillons sains pour un tissu donné afin de déterminer les changement induits par le développement cancéreux.

2.3 Différence de moyenne et différence de variance d'expression

Lorsqu'il s'agit d'analyser l'expression de gènes dans différentes conditions, la plupart des études cherchent à comparer des moyennes d'expression d'un ou plusieurs

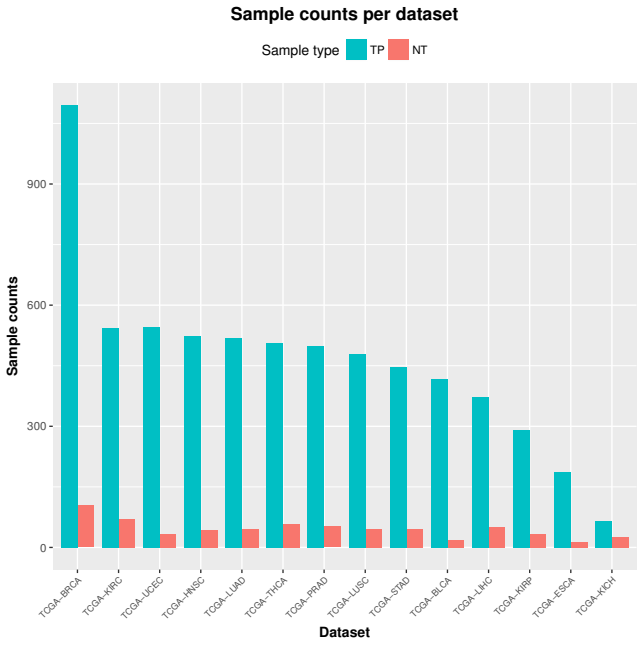

FIGURE 1.11 – Nombre d'échantillons tumoraux et normaux avec des données d'expression de miARN pour les tissus avec plus de 10 échantillons normaux (TP : *Tumor Primary*, NT : *Non Tumoral*, BLCA : *BLadder urothelial CArcinoma*, BRCA : *BReast invasive CArcinoma*, ESCA : *ESophageal CArcinoma*, HNSC : *Head and Neck Squamous cell Carcinoma*, KICH : *KIdney CHromophobe*, KIRC : *KIdney Renal Clear cell Carcinoma*, KIRP : *KIdney Renal Papillary cell carcinoma*, LIHC : *LIver Hepatocellular Carcinoma*, LUAD : *LUng ADenocarcinoma*, LUSC : *LUng Squamous cell Carcinoma*, PRAD : *PRostate ADenocarcinoma*, STAD : *STomach ADenocarcinoma*, THCA : *THyroid CArcinoma*, UCEC : *Uterine Corpus Endometrial Carcinoma*.

gènes dans deux conditions différentes (partie A de la figure 1.12). Le but de cette

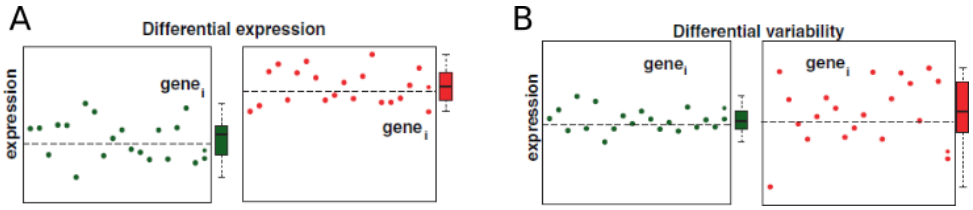

FIGURE 1.12 – Différence de moyenne d'expression (A) et de variabilité d'expression (B) d'un gène entre deux conditions (portion de la figure 3 de EMMERT-STREIB, TRIPATHI et MATOS SIMOES, 2012).

approche est de trouver un ou plusieurs gènes induits ou, au contraire, réprimés dans une condition par rapport à une autre. Cette différence d'expression pourrait caractériser une condition par rapport à une autre. Si on compare des échantillons issus de patients atteints d'une maladie à des échantillons issus de donneurs sains, les gènes différentiellement exprimés pourraient expliquer le développement pathologique. De manière générale, l'analyse de profils d'expression de gènes permet de lier le génotype au phénotype d'une population cellulaire ou même d'une cellule unique.

Une autre approche consiste à chercher des différences de variance d'expression d'un gène entre deux conditions. En effet, ces différences de variance pourraient refléter

des altérations de la régulation de l'expression des gènes, soit au niveau de leur synthèse, soit au niveau de leur dégradation. Dans le cadre de l'étude du cancer, une variance d'expression plus importante dans les échantillons tumoraux serait le reflet d'une régulation anarchique au cours du développement cancéreux. A l'inverse, un gène dont la gamme d'expression serait large dans les tissus sains et dont l'expression deviendrait beaucoup plus restreinte dans les tumeurs pourrait traduire le figement des cellules cancéreuses dans des voies de signalisation spécifiques.
De rares études ont été menées sur la variabilité de l'expression des gènes et ont permis d'identifier des gènes différentiellement variables impliqués dans le cancer (Ho et al., 2008, Ecker et al., 2015) ou des maladies neurologiques telles que la maladie de Parkinson ou la schizophrénie (Mar et al., 2011, Zhang et al., 2015). D'autres études se sont intéressées à identifier des gènes se caractérisant par leur variance d'expression au cours du développement. Mason et al., 2014 ont ainsi montré que des gènes, comme POU5F1, ont une très faible variance d'expression dans les cellules souches ayant la plus forte capacité d'auto-renouvellement. Hasegawa et al., 2015 ont identifié, à l'aide de données de séquençage de cellules uniques, des gènes présentant une très faible variance d'expression au cours de quatre différents stades de développement. Par ailleurs, ils ont mis en évidence un ensemble de gènes se distinguant par un profil de changement de variance d'expression au cours de ces stades. Parmi ces gènes, ils ont pu prouver expérimentalement que le gène HDDC2 joue un rôle dans le maintien de la pluripotence des cellules souches humaines.

3 But de ma thèse

L'objectif de ma thèse est double. Le premier est d'identifier les gènes codants et les miARN présentant une différence de variance d'expression entre les tissus cancéreux et les tissus sains. Nous espérons trouver dans cet espace peu investigué, la plupart des études se concentrant sur les gènes présentant une différence de moyenne d'expression entre deux conditions, de nouveaux biomarqueurs potentiels de diagnostic, de pronostic ou de théranostic. Le deuxième objectif est, à l'échelle du système, de mieux caractériser le rôle tampon des miARN en comparant leurs variances d'expression à celles de leurs ARNm cibles. Nous émettons l'hypothèse que les modifications de la variance d'expression entre deux conditions biologiques pourraient être le reflet du rôle tampon des miARN. Par exemple, il pourrait se matérialiser par le schéma suivant : perte de la souplesse de l'expression de miARN dans une condition pathologique, se traduisant par une diminution de leur variance d'expression par rapport à une condition saine, et perte de la précision de l'expression des ARNm cibles dans cette même condition pathologique, se traduisant par une augmentation de leur variance de l'expression.
Le cancer apparaît comme un domaine d'application pertinent de cette approche. L'hétérogénéité des profils d'expression observée dans les tumeurs et la robustesse de ces dernières leur permettant d'échapper à la plupart des traitements peuvent faire penser à une situation où les mécanismes de régulation de l'homéostasie cellulaire si efficaces en condition saine n'accomplissent plus leurs fonctions. Le rôle tampon des miARN pourrait alors se voir fortement affecté.
L'approche développée dans cette thèse se veut systémique. Le but n'est ainsi pas d'identifier quelques miARN et ARNm remarquables qui pourraient être les marqueurs d'une situation spécifique. L'objectif est d'identifier des traits biologiques les plus génériques possibles, marqueurs du développement tumoral. Dans cette approche, certains résultats d'une situation précise peuvent malgré tout être identifiés

et constituer le point de départ d'analyses complémentaires plus spécifiques. De manière générale, cette approche consistant à déterminer des modifications de la variance d'expression pourrait permettre d'identifier des gènes qui ne seraient pas détectés par l'approche classique de différence de moyenne d'expression. Cette approche pourrait ainsi aboutir à la découverte de nouveaux biomarqueurs du cancer ou de nouvelles cibles thérapeutiques.

Pour ce faire, des méthodes permettant de détecter des différences de variance d'expression de gènes entre deux conditions d'intérêt ont été identifiées, évaluées et appliquées aux données d'expression du TCGA. L'intégralité des méthodes utilisées dans ce manuscrit de thèse sont présentées dans le chapitre 2. L'application des tests de comparaison de variabilité et celle des méthodes permettant de détecter des différences de dispersion dans le cadre d'une modélisation des données par la distribution binomiale négative sont ainsi l'objet des chapitres 3 et 4 respectivement. Enfin, l'association de l'expression de miARN et d'ARNm cibles préalablement identifiés pour la modification de leur variance d'expression au cours de la cancérogénèse est évaluée dans le chapitre 5.

Chapitre 2

Méthodes

1 Prétraitement des données d'expression

Avant de pouvoir mener les analyses d'intérêt, *i.e.* l'identification de différence de variance ou de dispersion d'expression entre deux populations d'échantillons, les matrices de données d'expression issues du RNA-seq (voir section 2.1.3 du chapitre 1) doivent subir différentes étapes de pré-traitement. Elles ont notamment pour but de limiter la variabilité technique due au fait que les échantillons comparés ont été générés par différentes expériences de séquençage. Une bonne prise en compte de ces biais techniques permet une meilleure estimation de l'effet biologique recherché.

1.1 Filtrage des gènes faiblement exprimés

L'intérêt biologique d'un gène identifié dont le niveau moyen d'expression est très faible peut être mis en question. De plus, les très faibles valeurs de nombres de *reads* associées à un gène, avec la présence éventuelle de valeurs nulles pour de nombreux échantillons, peuvent perturber l'application de certaines méthodes statistiques. Enfin, retirer d'une analyse les gènes considérés comme trop faiblement exprimés permet de limiter la sévérité de la correction de tests multiples (voir section 4) et ainsi gagner en puissance statistique. Ainsi, tant sur le plan biologique que statistique, il y a tout intérêt à exclure de toute analyse les gènes faiblement exprimés.

Pour éviter tout biais pour les analyses ultérieures, les groupes d'échantillons d'intérêt ne sont pas pris en compte et un seuil sur la moyenne des nombres de *reads* est appliqué sur l'intégralité des échantillons qui composent le jeu de données (Bourgon, Gentleman et Huber, 2010). Les valeurs de seuil utilisées varient selon que les nombres de *reads* ont été transformés en $\log_2$ ou non. Dans le cas où ils n'ont pas été transformés, le filtrage est basé sur les valeurs CPM (*Counts Per Million*) pour prendre en compte les différences de profondeur de séquençage entre échantillons.

1.2 Normalisation

Dans le cadre du RNA-seq, l'expression d'un gène est estimée à partir du nombre de *reads* qui lui est affecté et dépend du nombre total de *reads* générés pour le séquençage de l'échantillon. Or, le rendement en nombre de *reads* peut beaucoup varier d'une expérience à une autre. De plus, au sein d'une même expérience multiplexée permettant le séquençage de plusieurs échantillons, la taille de librairie peut énormément varier d'un échantillon à l'autre. Ainsi, pour pouvoir comparer les valeurs d'expression issues de différents échantillons, les nombres de *reads* doivent être mis à la même échelle. Pour ce faire, différentes méthodes de normalisation des nombres de *reads* ont été évaluées par Dillies et al., 2013.

Ces méthodes ont toutes le même principe qui consiste à déterminer un facteur multiplicatif à appliquer à chaque échantillon. Des approches assez basiques consistent à diviser chaque nombre de *reads* par le nombre total de *reads* issus de l'échantillon ou à faire en sorte que l'ensemble des échantillons considérés aient les mêmes quantiles de nombres de *reads*. La méthode TMM (*Trimmed Mean of M-values*, Robinson et Oshlack, 2010) est une approche plus sophistiquée basée sur les M-valeurs et A-valeurs qui reflètent respectivement le *fold-change* de moyenne d'expression entre un échantillon j et un échantillon r pris comme référence et le niveau moyen d'expression :

$$M_i = \log_2\left(\frac{y_{ij}/m_j}{y_{ir}/m_r}\right)$$
$$A_{ij}^r = \frac{1}{2}\left(\log_2\left(\frac{y_{ij}}{m_j}\right) + \log_2\left(\frac{y_{ir}}{m_r}\right)\right)$$

où :

- y_{ij} est le nombre de *reads* pour le gène i dans l'échantillon j $(y_{ij} > 0)$;
- m_j est le nombre total de *reads* pour l'échantillon j ;
- r : échantillon de référence, *i.e.* l'échantillon dont le 3ème quartile est le plus proche du 3ème quartile moyen.

Pour chaque échantillon j, le facteur s_j est la moyenne pondérée et tronquée des M-valeurs après retrait des gènes faisant partie :

- des 30% des M-valeurs les plus petites ou les plus plus grandes ;
- des 5% des A-valeurs les plus petites ou les plus plus grandes.

Ces filtrages permettent de ne pas prendre en compte les gènes fortement différentiellement exprimés ou fortement exprimés qui pourraient fausser le facteur de normalisation. Pour n échantillons, les nombres de *reads* normalisés y_{ij}^{TMM} sont ainsi obtenus :

$$y_{ij}^{TMM} = \frac{y_{ij}}{m_j \times \mathrm{s}_j} \times \frac{\sum_{j=1}^{n} m_j \times \mathrm{s}_j}{n} \tag{2.1}$$

Au contraire de certaines méthodes, telle que RPKM, la méthode TMM conserve les valeurs de coefficients de variation présentes dans les nombres de *reads* bruts (Dillies et al., 2013). Dans le contexte de l'analyse de la variance d'expression, cette propriété est cruciale et explique pourquoi cette méthode a été retenue pour normaliser les nombres de *reads*.

Dans le cadre de l'application des tests de comparaison de variabilité (voir section 2), les nombres de reads sont normalisés à l'aide la formule 2.1. En revanche, lorsque les données d'expression sont modélisées par un modèle linéaire généralisé, les comptes de reads ne sont pas directement normalisés, le facteur s_j étant utilisé comme *offset* au sein du modèle (voir section 3.2.2).

1.3 Transformation $\log_2$

Les valeurs d'expression de gène déterminées à partir de données de séquençage se caractérisent par une distribution asymétrique du fait de la présence de valeurs extrêmes. La transformation des données d'expression par la fonction $\log_2$ permet de limiter la présence de ces valeurs extrêmes et de rapprocher la distribution des données d'une distribution normale, ce qui facilite les analyses statistiques ultérieures (Zwiener, Frisch et Binder, 2014).
Les valeurs d'expression comprises entre 0 et 2 deviennent négatives, nulles ou proches de 0 après la transformation $\log_2$. Cela peut poser problème dans les calculs ultérieurs, en particulier pour les calculs de coefficients de variation (voir section 2.2). Pour éviter

ces situations, les gènes dont l'expression moyenne après la transformation $\log_2$ est strictement inférieure à 2 parmi les échantillons d'intérêt sont retirés de l'étude.

1.4 Effets *batch*

Les effets *batch* (voir section 2.1.1 du chapitre 1) peuvent être traités de deux manières différentes. Les valeurs de nombres de *reads* identifiées comme étant affectées par ces effets peuvent être corrigées. D'autres approches basées sur des modèles linéaires généralisés consistent à ajouter au modèle une variable explicative contenant l'information des expériences ayant permis le séquençage des échantillons du jeu de données analysé.

1.4.1 Correction des données

La méthode ComBat (Johnson, Li et Rabinovic, 2007) a été appliquée pour corriger les effets *batch* dans les nombres de *reads*. Cette méthode a été initialement développée pour corriger les effets *batch* dans des données de puces à ADN et est considérée comme étant la meilleure (Chen et al., 2011). Elle peut tout de même être appliquée à des données RNA-seq après que celles-ci ont été log-transformées (Conesa et al., 2016).

1.4.2 Facteur bloquant dans un modèle linéaire généralisé

Dans le cadre des méthodes implémentant un modèle linéaire généralisé (voir section 3.2.2), les effets *batch* peuvent être contrôlés par l'intégration d'une variable explicative, ou facteur bloquant, au modèle (voir table 2.2). Cette variable explicative supplémentaire permet de capter la variabilité induite par les différentes expériences de séquençage ayant permis de générer les échantillons qui composent le jeu de données. L'intégration de cette variable au modèle permet à la comparaison d'intérêt biologique de ne pas être impactée par cette source de variabilité technique.

2 Tests de comparaison de variabilité

La variabilité de données peut être mesurée par différentes approches. La plus courante est la variance mais il en existe d'autres telles que le coefficient de variation ou l'entropie de Shannon. Le terme « variabilité », qui relève du langage courant mais qui n'est pas un terme statistique, est utilisé ici pour désigner l'ensemble de ces mesures statistiques. Pour chacune d'entre elles, des tests statistiques ont été développés pour comparer les données issues de deux populations d'échantillons. Cette section a pour but de présenter ces différentes mesures et les tests statistiques qui leur sont associés. De manière générale, l'ensemble de ces tests est dénommé sous l'appellation « tests de comparaison de variabilité ».

2.1 Comparaison de variances

La mesure la plus courante de la variabilité est la variance. Les tests présentés dans cette section permettent de tester l'égalité de variance de deux populations (ou plus de deux pour certains tests). Les hypothèses de ces tests sont donc :

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ et } H_1 : \sigma_1^2 \neq \sigma_2^2 \,,$$

où :

— σ_1^2 est la variance des données de la population 1 ;
— σ_2^2 est la variance des données de la population 2.

2.1.1 Test de Fisher

Le test le plus courant d'égalité de variance de deux populations est le test de Fisher. Il permet de comparer la variance de deux ensembles de données distribuées selon la loi normale. Sa statistique de test est la suivante :

$$F = \frac{\sigma_1^2}{\sigma_2^2},$$

où :
— σ_1^2 est la variance des données de la population 1 ;
— σ_2^2 est la variance des données de la population 2.

La statistique de test F suit la loi de Fisher avec (n-1) et (m-1) degrés de liberté sous l'hypothèse nulle, où n et m sont les nombres de valeurs observées pour les deux populations comparées.

Le test de Fisher est sensible à la non-normalité des données des populations comparées. La p-valeur calculée peut être affectée si la distribution des valeurs observées d'au moins l'une des populations comparées ne suit pas la loi normale. D'autres tests statistiques, tels que les tests de Bartlett, Brown-Forsythe et Levene, ont été développés pour pouvoir comparer la variance de variables aléatoires qui ne suivent pas une distribution normale.

2.1.2 Test de Bartlett

Le test de Bartlett (Bartlett, 1937) permet d'évaluer l'égalité de variance de données issues de k populations, avec k supérieur ou égal à 2. La statistique du test est la suivante :

$$B = \frac{(N-k)\,\ln(\sigma_p^2) - \sum_{i=1}^{k}(n_i-1)\,\ln(\sigma_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right) - \frac{1}{N-k}\right)},$$

où :
— k est le nombre de populations différentes auxquels les échantillons appartiennent ;
— N est le nombre total d'échantillons, toutes populations confondues ;
— n_i est le nombre d'échantillons appartenant à la i-ème population ;
— σ_p^2 est la variance de l'ensemble des données ;
— σ_i^2 est la variance de la i-ème population.

Sous l'hypothèse nulle d'égalité des variances, la statistique de test B suit la loi du χ^2 avec $(k-1)$ degrés de liberté.

Le test de Bartlett est connu pour être sensible à la non-normalité des données.

2.1.3 Tests de Levene et de Brown-Forsythe

Les tests de Levene (Levene, 1960) et de Brown-Forsythe (Brown et Forsythe, 1974) sont des tests statistiques très proches. Ils permettent de tester l'égalité de

variance de données de k populations, avec k supérieur ou égal à 2. La formule de leur statistique de test est la suivante :

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i \left(Z_{i.} - Z_{..}\right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i.})^2}, \tag{2.2}$$

où :

- k est le nombre de populations différentes auxquels les échantillons appartiennent ;
- N_i est le nombre d'échantillons appartenant à la i-ème population ;
- N est le nombre total d'échantillons, toutes populations confondues ;
- Y_{ij} est la valeur de la variable mesurée pour le j-ème échantillon de la i-ème population ;
- $Z_{ij} = \begin{cases} |Y_{ij} - \overline{Y}_{i.}|, \overline{Y}_{i.} \text{est la moyenne de la } i\text{-ème population (Levene)} ; \\ |Y_{ij} - \tilde{Y}_{i.}|, \tilde{Y}_{i.} \text{est la médiane de la } i\text{-ème population (Brown-Forsythe)} ; \end{cases}$
- $Z_{i.} = \frac{1}{N_i} \sum_{j=1}^{N_i} Z_{ij}$ est la moyenne des valeurs Z_{ij} pour la i-ème population ;
- $Z_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} Z_{ij}$ est la moyenne de toutes les valeurs Z_{ij}.

Les statistiques des tests de Levene et de Brown-Forsythe se différencient uniquement sur le terme Z_{ij}. Le test de Levene emploie la moyenne des valeurs observées dans son calcul tandis que le test de Brown-Forsythe utilise la médiane des valeurs observées. Le test de Brown-Forsythe est ainsi moins sensible aux valeurs extrêmes et donc à la non-normalité des données.

Pour les deux tests, la statistique de test W suit la loi de Fisher avec $(k-1)$ et $(N-k)$ degrés de liberté sous l'hypothèse nulle.

2.1.4 Test de Fligner-Killeen

Le test de Fligner-Killeen (Fligner et Killeen, 1976) permet la comparaison de la variance de k populations. Ce test est basé sur les rangs normalisés d'écarts à la médiane au sein de chaque population :

$$r_{ij}^{N} = \Phi^{-1}\left(\frac{1 + \frac{r_{ij}}{N+1}}{2}\right),$$

où :

- $\Phi(\cdot)$ est la fonction de répartition de la distribution normale ;
- r_{ij} est le rang de l'écart de l'observation i à la la médiane des observations de la population j à laquelle elle appartient parmi l'ensemble des observations des k populations ;
- N est le nombre total d'observations issues des k populations.

La statistique de ce test est définie par :

$$FK = \frac{\sum_{j=1}^{k} n_j (\bar{a}_j - \bar{a})^2}{s^2},$$

où :

- k est le nombre de populations ;
- n_j est le nombre d'observation pour la j-ème population ;
- $\bar{a}_j$ est la moyenne des rangs normalisés de la j-ème population ;

— $\bar{a}$ est la moyenne des rangs normalisés, toutes populations confondues ;
— s^2 est la variance de tous les rangs normalisés.

Sous l'hypothèse nulle, cette statistique de test suit la loi du χ^2 avec $(k-1)$ degrés de liberté. L'emploi de rangs rend ce test peu sensible aux données ne suivant pas une loi normale.

2.2 Comparaison de coefficients de variation

La variabilité de données issues de plusieurs populations peut être mesurée par le coefficient de variation (c_v) dont la formule est la suivante :

$$c_v = \frac{\sigma}{\mu},$$

où :
— σ est l'écart-type des valeurs observées d'une variable ;
— μ est la moyenne des valeurs observées d'une variable.

En divisant l'écart-type par la moyenne, le c_v permet d'avoir une valeur relative de la dispersion des données. Un des inconvénients de l'utilisation du c_v est sa sensibilité aux faibles valeurs. En effet, pour des valeurs dont la moyenne est proche de 0, le c_v peut tendre vers l'infini. De plus, on constate que le c_v est en général plus élevé pour les faibles valeurs que pour les valeurs élevées.

Il existe des tests statistiques de comparaison de variabilité de données de plusieurs populations basés sur le coefficient de variation. Les hypothèses des tests de cette section sont les suivantes :

$$H_0 : c_{v1} = c_{v2} \text{ et } H_1 : c_{v1} \neq c_{v2},$$

où :
— c_{v1} est le coefficient de variation des données de la population 1 ;
— c_{v2} est le coefficient de variation des données de la population 2.

Les tests de Feltz-Miller et de Krishnamoorthy-Lee sont des tests de comparaison de variabilité basés sur le coefficient de variation.

2.2.1 Test de Feltz-Miller

Le test de Feltz-Miller (Feltz et Miller, 1996) permet de comparer les coefficients de variation de k populations. Il est basé sur une estimation du coefficient de variation de l'ensemble des k populations :

$$\hat{\tau}_c = \frac{\left(\sum_{j=1}^{k} m_j \frac{s_j}{\bar{x}_j}\right)}{M},$$

où :
— s_j est l'écart-type de la j-ème population ;
— $\bar{x}_j$ est la moyenne observée de la j-ème population ;
— $m_j = n_j - 1$ avec n_j le nombre d'observations de la j-ème population ;
— $M = \sum_{j=1}^{k} m_j$.

La statistique de test reflète l'écart du coefficient de variation de chaque population à l'estimateur $\hat{\tau}_c$ du coefficient de variation de l'ensemble des k populations :

$$D'AD = \hat{\tau}_c^{-2} \left(0,5 + \hat{\tau}_c^2\right)^{-1} \left[\sum_{i=1}^{k} m_i \left(\frac{s_i}{\bar{x}_i}\right)^2 - \frac{\hat{\tau}_c^2}{M}\right] .$$

La statistiques $D'AD$ suit la loi du χ^2 avec $(k-1)$ degrés de liberté sous l'hypothèse nulle d'égalité des coefficients de variation.

2.2.2 Test de Krishnamoorthy-Lee

Le test de Krishnamoorthy-Lee (Krishnamoorthy et Lee, 2014) a été développé dans le but de comparer plusieurs populations suivant des lois normales. Les paramètres de ces distributions sont estimés par maximum de vraisemblance et le test de Krishnamoorthy-Lee consiste en une version modifiée du test de rapport de vraisemblance dans le but de comparer les coefficients de variation des différentes populations. Les développements nécessaires à l'obtention de la statistique de ce test sont présentés dans l'article de ce test (Krishnamoorthy et Lee, 2014). Les auteurs affirment que leur test a de meilleurs performances que celui de Feltz-Miller pour comparer les coefficients de variation de populations de tailles différentes.

2.3 Entropie de Shannon

L'entropie de Shannon est une mesure utilisée en théorie de l'information pour quantifier la quantité d'information contenue dans une source de données. Il s'agit d'une mesure de la diversité d'un ensemble de données. Sachant les valeurs et la fréquence à laquelle elles sont observées, l'entropie de Shannon quantifie l'incertitude liée à la prédiction d'une nouvelle valeur.

L'entropie de Shannon d'une suite de nombre $x_1, x_2, \ldots, x_n$ est calculée selon la formule :

$$SE = \frac{-\sum_{i=1}^{n} \frac{x_i}{x} \log_2 \left(\frac{x_i}{x}\right)}{\log_2 n} ,$$

où $x = \sum_{i=1}^{n} x_i$ est la somme des valeurs observées.

Par définition, l'entropie de Shannon prend des valeurs comprises dans l'intervalle $[0; 1]$. Pour un ensemble de données très variables où chaque valeur est observée à la même fréquence, l'entropie de Shannon tend vers 1. Au contraire, pour une population où une valeur est observée un très grand nombre de fois, l'entropie de Shannon tend vers 0.

Un test basé sur l'entropie de Shannon a déjà été utilisé pour identifier des gènes différentiellement variables entre deux conditions à partir de données de puces à ADN (Wang et al., 2015). Les p-valeurs sont déterminées à l'aide de permutations des données dans les deux populations considérées.

2.4 Test de normalité

Certains tests de comparaison de variabilité sont sensibles à la normalité des données comparées. Leur conclusion peut être faussée si les données d'au moins l'une des deux populations comparées ne sont pas distribuées selon la loi normale, ce qui est généralement le cas en présence de valeurs extrêmes.

Il existe des tests statistiques qui permettent de savoir si des données s'écartent de la loi normale. Les hypothèses sont :

$$H_0 : X = (x_1, \ldots, x_n) \text{ est un échantillon normalement distribué,}$$
$$H_a : X = (x_1, \ldots, x_n) \text{ n'est pas un échantillon normalement distribué.}$$

Le test de Shapiro-Wilk (Shapiro et Wilk, 1965) est l'un de ces tests. Sa statistique est la suivante :

$$SW = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

où :

- $x_{(i)}$ est la statistique d'ordre de rang i de X, *i.e.* la i-ème plus petite valeur de X ;
- $\bar{x}$ est la moyenne de l'échantillon X ;
- les constantes a_i sont données par la formule :

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}},$$

où :

- $m = (m_1, \ldots, m_n)^T$ sont les espérances des statistiques d'ordre de variables aléatoires indépendantes et identiquement distribuées issues d'une distribution normale ;
- V est la matrice de covariance des ces statistiques d'ordre.

La statistique SW est ensuite comparée à une table de valeurs limites à différents niveaux de risque α pour conclure le test (Shapiro et Wilk, 1965).

Une étude a montré que le test de Shapiro-Wilk est le plus puissant des tests de normalité (Yap et Sim, 2011).

2.5 Récapitulatif des tests statistiques évalués

Les caractéristiques de l'ensemble des tests considérés dans cette étude sont résumées dans le tableau de la figure 2.1.

	Brown-Forsythe	Levene	F-test	Bartlett	Fligner-Killeen	Feltz-Miller	Krishnamoorthy-Lee	Differential SE
Metric	variance	variance	variance	variance	variance	CV	CV	SE
Distribution	F-distribution	F-distribution	F-distribution	chi-squared	chi-squared	chi-squared	chi-squared	permutations
Number of groups	2 or more	2 or more	2	2 or more	2 or more	2 or more	2 or more	2
Non-normality sensitivity	⊕	⊕	⊕⊕⊕	⊕⊕⊕	⊕	⊕⊕	⊕⊕	⊘
Uneven sample groups	⊘	⊘	⊘	⊘	⊘	⊘	✓	⊘

Figure 2.1 – Tests statistiques évalués pour comparer la variabilité de deux ensembles de données. Symboles plus rouges : degré de sensibilité, panneau noir : pas d'*a priori*, symbole vert : test adapté.

2.6 Comparaison des différents tests de variabilité

Les tests statistiques de variabilité sont comparés entre eux en se basant sur les p-valeurs qu'ils génèrent. En fonction des tests, les p-valeurs peuvent beaucoup varier. Plutôt que de comparer les tests statistiques en se basant sur leurs p-valeurs ou en se basant sur un seuil unique que l'on appliquerait à tous les tests, il apparaît plus pertinent d'ordonner par ordre croissant les p-valeurs obtenues avec chaque test et de comparer les tests statistiques en se basant sur les rangs des p-valeurs.

Pour chaque comparaison de tests statistiques, les gènes les plus discordants entre les deux tests considérés, c'est-à-dire classés parmi les gènes les plus différentiellement variables (rang de p-valeur faible) selon l'un des deux tests et beaucoup moins bien classés par l'autre test (rang de p-valeur élevé), sont déterminés à l'aide de la formule :

$$\frac{\text{rang}(p_{t1}) - \text{rang}(p_{t2})}{\min(\text{rang}(p_{t1}),\, \text{rang}(p_{t2}))}\,, \tag{2.3}$$

où p_{t1} et p_{t2} sont les p-valeurs comparées issues de deux tests différents $t1$ et $t2$ et $\text{rang}(p_{t1})$ et $\text{rang}(p_{t2})$ leur rang respectif parmi l'ensemble des p-valeurs ordonnées par ordre croissant obtenues avec les tests $t1$ et $t2$.

3 Modèles basés sur la distribution binomiale négative

3.1 Loi de probabilité

Les nombres de *reads* bruts sont des entiers positifs (voir section 2.1.3 du chapitre 1), les lois de probabilités discrètes apparaissent ainsi comme des modélisations de choix pour représenter ce type de données. Une particularité des données RNA-seq est la présence de valeurs nulles en grand nombre.

La distribution binomiale négative Dans cette section, on note Y_{ij} la variable aléatoire quantitative décrivant le nombre de *reads* associés au gène i parmi le total de m_j *reads* issus de l'échantillon j. Les réalisations $y_{1j}, y_{2j}, \ldots, y_{Gj}$ de cette variable aléatoire sont les nombres de *reads* associés aux gènes 1 à G dans l'échantillon j. Chaque *read* peut être interprété comme la réalisation d'une variable aléatoire de Bernoulli X_r de paramètre p_{ij}, où p_{ij} est la probabilité que le *read* r soit issu du gène i :

$$X_r \sim \text{Bernouilli}(p_{ij})\,.$$

On note alors que Y_{ij} est la somme de m_j variables aléatoires de Bernoulli indépendantes et identiquement distribuées. Cette variable aléatoire suit une distribution binomiale de paramètres m_j et p_{ij} :

$$Y_{ij} = \sum_{r=1}^{m_j} X_r \sim \mathcal{B}(m_j, p_{ij})\,.$$

La probabilité que Y_{ij} prenne la valeur k est :

$$\mathbb{P}(Y_{ij} = k) = \binom{m_j}{k} p_{ij}^k (1 - p_{ij})^{m_j - k}\,.$$

Comme le nombre total de *reads* m_j est très grand et la probabilité p_{ij} qu'un *read* soit issu d'un gène i est généralement très faible, la distribution suivie par Y_{ij} peut

être approximée par une distribution de Poissson de paramètre λ_{ij}. On note :

$$Y_{ij} \sim \mathcal{P}(\lambda_{ij})\,.$$

La probabilité que Y_{ij} prenne la valeur k est alors :

$$\mathbb{P}(Y_{ij} = k) = \frac{\lambda_{ij}^k}{k!} e^{-\lambda_{ij}}\,.$$

Pour un ensemble d'échantillons, l'espérance et la variance de Y_{ij} sont alors égales à λ_{ij} :

$$\mathbb{E}(Y_{ij}) = \mathrm{Var}(Y_{ij}) = \lambda_{ij}\,.$$

Dans le cas de réplicats techniques, *i.e.* les échantillons sont les séquençages répétés d'un même échantillon biologique, cette modélisation capture bien la relation entre la moyenne et la variance des nombres de *reads* observés (Marioni et al., 2008). En revanche, en présence de réplicats biologiques, *i.e.* les échantillons séquencés sont issus d'échantillons biologiques différents, la variance des nombres de *reads* est plus importante et la distribution de Poisson ne parvient plus à bien la modéliser, en particulier pour les gènes fortement exprimés (figure 2.2). Les nombres de *reads* sont dits « sur-dispersés ». La distribution binomiale négative apparaît alors comme une

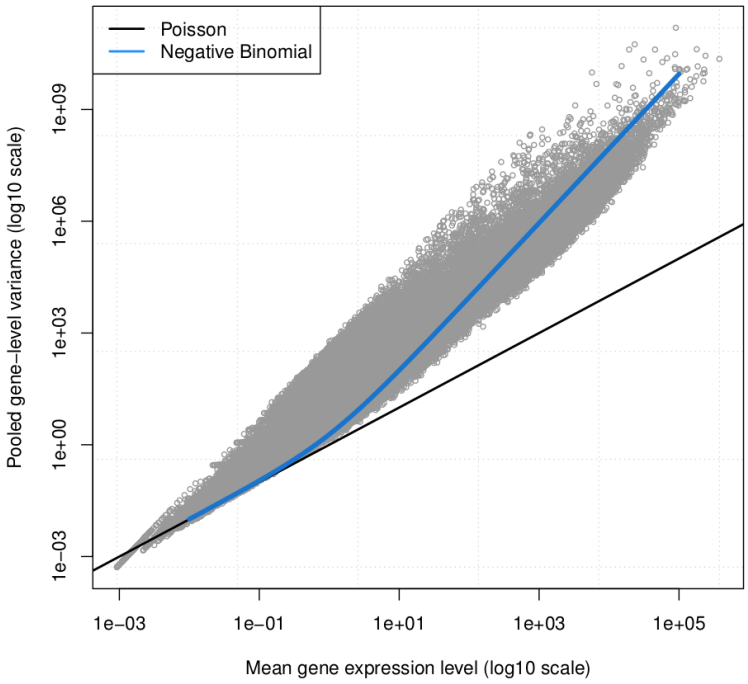

Figure 2.2 – Moyennes et variances de nombres de *reads* et ajustements d'une distribution de Poisson et d'une distribution binomiale négative.

distribution plus adaptée pour modéliser ce type de données. La variable Y_{ij} décrit alors le nombre de succès d'expériences de Bernoulli de probabilité de succès p_{ij} avant d'observer un nombre d'échecs r_{ij}. Dans le cas de données RNA-seq, un succès est l'observation d'un *read* issu du gène i et un échec est l'observation d'un *read* non issu du gène i. On note :

$$Y_{ij} \sim \mathcal{NB}(r_{ij}, p_{ij})\,, \tag{2.4}$$

où :

— r_{ij} est le nombre pré-défini d'échecs à observer ;

— p_{ij} est la probabilité de succès de chaque expérience de Bernoulli.
La fonction de masse, l'espérance et la variance sont alors :

$$\begin{aligned} &\mathbb{P}(Y_{ij}=k) = \binom{k+r_{ij}-1}{k} p_{ij}^k (1-p_{ij})^{r_{ij}}\,, \\ &\mathbb{E}(Y_{ij}) = \tfrac{p_{ij}\, r_{ij}}{1-p_{ij}}\,, \\ &\mathrm{Var}(Y_{ij}) = \tfrac{p_{ij}\, r_{ij}}{(1-p_{ij})^2}\,. \end{aligned} \tag{2.5}$$

La distribution binomiale négative est en fait une distribution de Poisson de paramètre $\lambda_{ij}\Theta_{ij}$, où Θ_{ij} suit une distribution Gamma de paramètres de forme et d'échelle prenant la même valeur α_{ij} (de sorte que $\mathbb{E}(\Theta_{ij}) = 1$ et donc $\mathbb{E}(Y_{ij}) = \lambda_{ij}$), permettant la modélisation de la sur-dispersion des données. On note :

$$Y_{ij} \,|\, \Theta_{ij} \sim \mathcal{P}(\lambda_{ij}\Theta_{ij}) \text{ avec } \Theta_{ij} \sim \mathrm{Gamma}(\alpha_{ij}, \alpha_{ij})\,.$$

La probabilité que Y_{ij} prenne la valeur k s'écrit alors :

$$\begin{aligned} \mathbb{P}(Y_{ij}=k) &= \int_0^{+\infty} \frac{\lambda^k e^{-k}}{k!} \frac{\alpha_{ij}^{\alpha_{ij}}}{\Gamma(\alpha_{ij})} \lambda^{\alpha_{ij}-1} e^{-\alpha_{ij}\lambda} d\lambda \\ &= \frac{\Gamma(\alpha_{ij}+k)}{k!\,\Gamma(\alpha_{ij})} \left(\frac{1}{1+\alpha_{ij}}\right)^k \left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right)^{\alpha_{ij}}, \end{aligned}$$

où $\Gamma(\cdot)$ est la fonction gamma.
En identifiant $\alpha_{ij} = r_{ij}$ et $p_{ij} = \frac{1}{1+\alpha_{ij}}$, on a :

$$\mathbb{P}(Y_{ij}=k) = \frac{\Gamma(r_{ij}+k)}{k!\,\Gamma(r_{ij})}\, p_{ij}^k\, (1-p_{ij})^{r_{ij}}.$$

On retrouve la fonction de masse définie dans la formule 2.5 étendue aux cas où le paramètre r_{ij} prend des valeurs réelles. En effet, quand r_{ij} prend des valeurs entières,

$$\binom{k+r_{ij}-1}{k} = \frac{\Gamma(r_{ij}+k)}{k!\,\Gamma(r_{ij})}\,.$$

Il s'agit ainsi de la distribution binomiale négative, ou distribution de Pólya, de paramères r_{ij} et p_{ij}. L'espérance et la variance s'écrivent alors :

$$\begin{aligned} \mathbb{E}(Y_{ij}) &= \frac{p_{ij} r_{ij}}{1-p_{ij}} = \lambda_{ij} \text{ et} \\ \mathrm{Var}(Y_{ij}) &= \frac{p_{ij} r_{ij}}{(1-p_{ij})^2} = \lambda_{ij} + \phi_{ij}\, \lambda_{ij}^2, \end{aligned}$$

où ϕ_{ij} est le paramètre de dispersion, $\phi_{ij} = \frac{1}{\alpha_{ij}}$.
C'est la dispersion ϕ_{ij} qui permet de prendre en compte le surplus de variance observée dans les données RNA-seq issues de réplicats biologiques (figure 2.2) et lorsque ϕ_i tend vers 0, alors la distribution binomiale négative revient à la distribution de Poisson (Cameron et Trivedi, 1998). Ainsi, cette loi est plus appropriée pour modéliser les nombres de *reads* issus de réplicats biologiques et s'est imposée comme la distribution la plus utilisée par les méthodes analysant des données RNA-seq.

Notation en fonction de la moyenne et de la dispersion Dans les modèles de régression, la distribution binomiale négative est communément présentée en fonction de sa moyenne et de sa variance plutôt qu'en fonction des paramètres r_{ij} et p_{ij} telle

qu'indiquée dans la notation 2.4. En effet, comme la distribution binomiale négative est issue d'un mélange des distributions de Poisson et Gamma où le paramètre λ_{ij} de la loi de Poisson, qui est aussi la moyenne, est distribué selon une loi Gamma de paramètres $(\alpha_{ij}, \alpha_{ij})$, on peut écrire :

$$\begin{aligned} &Y_{ij} \sim \mathcal{NB}(\lambda_{ij}, \phi_{ij})\,, \\ &\mathbb{E}(Y_{ij}) = \lambda_{ij}\,, \\ &\mathrm{Var}(Y_{ij}) = \lambda_{ij} + \phi_{ij}\,\lambda_{ij}^2\,. \end{aligned} \tag{2.6}$$

où $\phi_{ij} = \frac{1}{\alpha_{ij}}$.
La fonction de masse s'écrit alors :

$$f(k;\,\lambda_{ij},\,\phi_{ij}) = \mathbb{P}(Y_{ij} = k) = \frac{\Gamma(k + \phi_{ij}^{-1})}{\Gamma(\phi_{ij}^{-1})\Gamma(k+1)} \left(\frac{1}{1+\lambda_{ij}\,\phi_{ij}}\right)^{\phi_{ij}^{-1}} \left(\frac{\lambda_{ij}}{\phi_{ij}^{-1} + \lambda_{ij}}\right)^{k}. \tag{2.7}$$

Les paramètres classiques r_{ij} et p_{ij} d'une distribution binomiale négative se retrouvent à l'aide des formules :

$$r_{ij} = \frac{1}{\phi_{ij}} \text{ et } p_{ij} = \frac{1}{1+\lambda_{ij}\,\phi_{ij}}\,.$$

3.2 Analyse de différence de moyenne d'expression

Une des principales applications du RNA-seq est d'identifier des gènes différentiellement exprimés (DE), *i.e.* des gènes présentant des différences de moyenne d'expression entre deux conditions d'intérêt. Pour ce faire, de nombreuses méthodes ont été développées. La plupart de ces méthodes sont basées sur la distribution binomiale négative et, parmi celles-ci, les plus utilisées sont celles implémentées dans les *packages* R edgeR (Robinson, McCarthy et Smyth, 2010) et DESeq (Anders et Huber, 2010, Love, Huber et Anders, 2014).
Dans un premier temps, des tests exacts ont été développés dans le but d'identifier des gènes différentiellement exprimés entre deux populations d'échantillons d'intérêt (section 3.2.1). Dans un second temps, des schémas expérimentaux plus complexes constitués de plusieurs variables d'intérêt et pouvant définir plus que deux populations d'échantillons ont été pris en compte par des méthodes plus élaborées basées sur des modèles linéaires généralisés (section 3.2.2).

3.2.1 Comparaison de deux groupes d'échantillons

Dans cette section, on suppose que les nombres de *reads* issus de différents échantillons sont indépendants. De plus, on suppose qu'après normalisation, les nombres de *reads* sont identiquement distribués. Les variables aléatoires Y_{ij} représentant les nombres de *reads* sont ainsi considérées comme étant indépendantes et suivant une distribution binomiale négative telle que définie dans la formule 2.6 de paramètres $\mu_{ij} = s_j\,\lambda_{ij}$ et ϕ_{ij} où s_j est un facteur de normalisation de la taille de la librairie de l'échantillon j (voir section 1.2), λ_{ij} est la proportion de la librairie issue du gène i et ϕ_{ij} est la dispersion de l'expression du gène i. Le test de différence de moyenne d'expression entre deux populations d'échantillons peut ainsi s'écrire :

$$H_0 : \lambda_{i1} = \lambda_{i2}\ ,\ H_a : \lambda_{i1} \neq \lambda_{i2}\ .$$

Les paramètres λ_{ij} et ϕ_{ij} sont inconnus et doivent être estimés. Pour pouvoir réaliser des tests sur la moyenne des nombres de *reads*, la dispersion ϕ_{ij} doit avoir été estimée

au préalable. Les *packages* edgeR et DESeq diffèrent par leurs méthodes d'estimation de la dispersion qui sont l'objet de la section 3.3. En particulier, la méthode d'estimation de la dispersion employée par edgeR est détaillée dans la section 3.3.3.
Dans cette section, on considère qu'un estimateur de ϕ_{ij} est disponible pour chaque gène i et chaque condition. Les *packages* edgeR et DESeq proposent chacun un test exact, inspiré du test exact de Fisher, la distribution binomiale négative étant utilisée pour le calcul des probabilités des nombres de *reads* observés. Les sommes des nombres de *reads* pour le gène i sont calculées pour les deux conditions et sont utilisées comme statistiques de test :

$$z_{i1} = \sum_{j \in S_1} y_{ij}\,,\ z_{i2} = \sum_{j \in S_2} y_{ij} \ \text{ et } z_i = z_{i1} + z_{i2}\,,$$

où S_1 et S_2 sont les ensembles d'échantillons appartenant aux conditions 1 et 2 respectivement.
Les statistiques z_{i1} et z_{i2} sont supposées suivre une distribution binomiale négative et les probabilités d'observer les comptes z_{i1} et z_{i2} peuvent être calculées. Comme pour le test exact de Fisher, la p-valeur de ce test exact est la somme de toutes les probabilités inférieures ou égales à celle observée étant donnée la somme totale des nombres de *reads* z_i.
Sous l'hypothèse nulle, $\lambda_{i1} = \lambda_{i2} = \lambda_i$, on estime donc λ_i à partir des nombres de *reads* issus des deux conditions. En prenant en compte les facteurs de normalisation s_j, les tailles de librairie sont considérées comme étant équivalentes. Les nombres de *reads* issus de différents gènes sont supposés indépendants. Bien que cette dernière supposition soit discutable biologiquement, en particulier pour les gènes partageant le même promoteur, elle présente l'avantage de considérer les variables aléatoires Y_{ij} comme étant indépendantes et identiquement distribuées, permettant ainsi d'estimer λ_i par maximum de vraisemblance. L'estimateur $\hat{\lambda}_i$ est alors la moyenne des comptes observés :

$$\hat{\lambda}_i = \frac{1}{|S_1| + |S_2|} \sum_{j \in \{S_1; S_2\}} \frac{y_{ij}}{s_j}\,.$$

L'estimateur de μ_i pour les deux populations d'échantillons est ensuite calculé en appliquant les facteurs de normalisation s_j propres à chaque échantillon :

$$\hat{\mu}_{i1} = \sum_{j \in S_1} s_j\,\hat{\lambda}_i \text{ et } \hat{\mu}_{i2} = \sum_{j \in S_2} s_j\,\hat{\lambda}_i\,.$$

Les estimateurs $\hat{\mu}_i$ et $\hat{\phi}_i$ étant connus pour chaque condition, les probabilités $\mathbb{P}(z_{i1} = k_1)$ et $\mathbb{P}(z_{i2} = k_2)$ peuvent être calculées à l'aide de la fonction de masse définie dans la formule 2.7 pour toutes paires de valeurs (k_1, k_2) telles que $k_1 + k_2 = z_i$. Soit $\mathbb{P}(k_1, k_2)$ cette probabilité. Sous l'hypothèse nulle, on suppose que les nombres de *reads* issus de différents échantillons sont indépendants, donc $\mathbb{P}(k_1, k_2) = \mathbb{P}(z_{i1} = k_1)\,\mathbb{P}(z_{i2} = k_2)$. La p-valeur p_i pour une paire de valeurs observée est alors la somme des probabilités $\mathbb{P}(k_1, k_2)$ inférieures ou égales à $\mathbb{P}(z_{i1}, z_{i2})$ étant donné z_i :

$$p_i = \sum_{k_1 + k_2 = z_i} \mathbb{P}(k_1, k_2)\,,\ \text{avec } \mathbb{P}(k_1, k_2) \leq \mathbb{P}(z_{i1}, z_{i2})\,.$$

Les tests exposés dans cette section ne permettent la comparaison que de deux groupes d'échantillons. Des schémas expérimentaux plus complexes, permettant la comparaison de plus de deux groupes d'échantillons ou l'intégration d'autres effets que la comparaison d'intérêt, ne peuvent être pris en compte par ces tests et nécessitent

l'emploi d'autres méthodes basées sur les modèles linéaires généralisés.

3.2.2 Les modèles linéaires généralisés

Dans le cadre de l'analyse de différence de moyenne d'expression à partir de données RNA-seq, les modèles linéaires généralisés (GLM, pour *Generalized Linear Models*) permettent la prise en compte de schémas expérimentaux plus complexes que la simple comparaison de deux groupes d'échantillons, habituellement un groupe contrôle et un groupe d'intérêt (voir section 3.2.1). Plusieurs traitements peuvent être pris en compte et évalués à l'aide de variables explicatives X_i dans ces modèles. L'une d'entre elles représente généralement les comparaisons d'intérêt principal définissant les groupes d'échantillons à comparer. Dans le cas le plus simple, cette variable peut prendre deux valeurs différentes correspondant aux deux populations d'échantillons à comparer. D'autres mesures que l'on soupçonne pouvoir avoir une influence sur l'expression de gènes peuvent aussi être intégrées dans ce genre de modèle, que ce soit des facteurs biologiques tels que l'âge ou le sexe des patients dont sont issus les échantillons analysés ou des facteurs techniques tels que les expériences de séquençage ayant généré les échantillons. On parle de covariables lorsque plusieurs variables explicatives sont intégrées au modèle. Leur intégration permet de dissocier les effets induits par ces covariables de l'effet principal à observer qui constitue l'intérêt majeur de l'étude.

Définition Les modèles linéairés généralisés sont une extension des modèles linéaires classiques à des variables aléatoires non-gaussiennes (Nelder et Wedderburn, 1972, McCullagh et Nelder, 1989). Ils permettent de caractériser les relations entre une variable Y et k variables explicatives $X_1, X_2, \ldots, X_k$ et sont composés de trois éléments :

1. une variable aléatoire Y suivant une distribution faisant partie de la famille exponentielle ;
2. une combinaison linéaire de variables explicatives $X_1, X_2, \ldots, X_k$ suivant une loi arbitraire :
$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \,,$$
où les coefficients β_j sont des coefficients de régression associés aux variables X_j pour j compris entre 1 et k, et β_0 est une valeur d'intercept ;
3. une fonction de lien linéaire et bijective g qui modélise l'espérance de la variable aléatoire Y à partir de la combinaison linéaire des variables explicatives X_j :
$$\mathbb{E}(Y) = g^{-1}(\eta) = g^{-1}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \,.$$
La fonction de lien peut être toute fonction linéaire bijective. Les fonctions identité, log, inverse ou logit peuvent ainsi être utilisées comme fonction de lien dans le cadre d'un GLM.

Dans ce cadre, la variable aléatoire Y peut suivre n'importe quelle loi de probabilité appartenant à la famille exponentielle. Les lois de probabilité appartenant à la famille exponentielle sont toutes les distributions dont la densité peut s'écrire sous la forme :

$$f(x|\theta) = h(x) \exp(\eta(\theta)\, T(x) - A(\theta) + B(x)) \,.$$

La fonction de masse de loi binomiale négative telle que définie dans la formule 2.5 peut s'écrire sous la forme :

$$f(x|p,r) = \binom{x+r-1}{x} p^x (1-p)^r = \binom{x+r-1}{x} \exp(x\log(p) + r\log(1-p))\,.$$

A la condition que r, et donc ϕ, soit fixé, la loi binomiale négative fait partie de la famille des distributions exponentielles avec :

- $\theta = p$;
- $\eta : \theta \mapsto \log(\theta)$;
- $T : x \mapsto x$;
- $A : \theta \mapsto -r\log(1-\theta)$;
- $h : x \mapsto \binom{x+r-1}{x}$.

La relation entre l'espérance et la variance de Y doit donc être connue pour pouvoir appliquer un GLM à cette variable aléatoire. La variance d'une loi de probabilité appartenant à la famille exponentielle est une fonction de la moyenne et, éventuellement, d'un paramètre de dispersion. Dans le cas de la loi binomiale négative, la variance est fonction de la moyenne et de la dispersion telle que définie dans la formule 2.6, ce qui permet donc l'application d'un GLM à Y_{ij}.

Application aux données RNA-seq Dans le cas de la modélisation de nombres de *reads* RNA-seq, la variable aléatoire Y_{ij} représente les nombres de *reads* issus du gène i et de l'échantillon j. On suppose que Y_{ij} suit une loi binomiale négative de paramètres de moyenne μ_{ij} et de dispersion ϕ_i telle que définie dans la formule 2.6. Les méthodes classiques d'estimation de paramètres pour les GLM de la famille de lois exponentielles peuvent ainsi être appliquées à la condition que le paramètre de dispersion ϕ_i soit fixé. Dans cette section, on suppose qu'un estimateur $\hat{\phi}_i$ est disponible pour chaque gène i préalablement à l'application d'un GLM grâce aux méthodes exposées dans les sections 3.3.4 et 3.3.5 pour les *packages* edgeR et DESeq2 respectivement. Cette condition est donc bien vérifiée. Si cela n'est pas le cas, la dispersion ainsi que les coefficients de régression peuvent tout de même être estimés par d'autres approches de maximum de vraisemblance.
L'espérance de la variable aléatoire Y_{ij} étant strictement positive, la fonction log apparaît comme un choix naturel de fonction de lien. Le GLM que l'on applique à Y_{ij} s'écrit donc :

$$\begin{aligned} Y_{ij} &\sim \mathcal{NB}(\mu_{ij},\, \hat{\phi}_i)\,, \\ \log(\mu_{ij}) &= \textstyle\sum_k x_{jk}\,\beta_{ik} + \log(s_j)\,, \end{aligned} \tag{2.8}$$

où :

- x_{jk} est la valeur de la variable explicative X_k concernant l'échantillon j ;
- β_{ik} sont les coefficients de régression applicables aux variables explicatives X_k pour le gène i ;
- s_j sont les facteurs de normalisation (voir section 1.2) capturant la dépendance des nombres de *reads* à la profondeur de séquençage et servent d'*offset*.

Matrice de *design* Dans le cadre d'analyses de données RNA-seq, les GLM sont souvent représentés sous forme vectorielle :

$$\log(\mu_{ij}) = x_j^T\,\beta_i\,,$$

où :

— x_j^T est un vecteur contenant les valeurs des variables explicatives à appliquer à l'échantillon j ;
— β_i est un vecteur contenant les coefficients de régression applicables aux variables explicatives pour le gène i.

Le vecteur x_j^T est issu d'une matrice X, appelée matrice de *design*, contenant l'ensemble des valeurs des variables explicatives pour les n échantillons. Pour un GLM constitué de K variables explicatives pouvant prendre au total L valeurs différentes, cette matrice est constituée de $L - K + 1$ colonnes et de n lignes correspondant aux n échantillons. Par exemple, pour la situation simple de comparaison de deux groupes de trois échantillons tels que définis dans la table 2.1, le GLM s'écrit :

$$\log(\mu_{ij}) = \beta_{i0} + \beta_{i1}\, x_{j1}\,,$$

où :

— β_{i0} représente le niveau d'expression du gène i dans l'un des deux groupes pris comme référence ;
— β_{i1} représente la différence d'expression du gène i entre les deux groupes d'échantillons ;
— la variable explicative d'intérêt x_{j1} représente l'appartenance de l'échantillon j à l'un des deux groupes d'échantillons.

La matrice de *design* de ce GLM s'écrit :

Echantillon	Groupe
1	A
2	A
3	A
4	B
5	B
6	B

$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$$

TABLE 2.1 – Exemple de comparaison de deux groupes d'échantillons et matrice de *design* correspondante. Dans cet exemple, le groupe A est pris comme groupe de référence.

La première colonne de la matrice de *design*, appelée intercept, représente le niveau d'expression des gènes dans le groupe d'échantillons pris comme référence. Dans cet exemple, il s'agit des échantillons du groupe A. Les autres colonnes de la matrice représentent les différences entre le groupe de référence et les autres groupes définis par les variables explicatives. Ici, le GLM n'intégrant qu'une seule variable explicative ne pouvant prendre que deux valeurs différentes, la matrice de *design* n'est constituée que de deux colonnes. La deuxième colonne représente les différences entre les échantillons du groupe B et le groupe de référence, *i.e.* le groupe A dans cet exemple. Elle est ainsi constituée de '0' pour les échantillons appartenant au groupe de référence et de '1' pour les échantillons du groupe B.
Les lignes uniques de la matrice de *design* représentent toutes les combinaisons possibles des coefficients de régression pour modéliser l'expression des gènes dans les groupes d'échantillons tels que définis par le GLM. Dans cet exemple, la moyenne de l'expression pour les groupes d'échantillons A et B s'écrivent ainsi :

$$\log\left(\mu_i^A\right) = x_A^T\,\beta_i = (1,0)^T\,(\beta_{i0},\beta_{i1}) = 1\times\beta_{i0} + 0\times\beta_{i1} = \beta_{i0}\,,$$

$$\log\left(\mu_i^B\right) = x_B^T\,\beta_i = (1,1)^T\,(\beta_{i0},\beta_{i1}) = 1\times\beta_{i0} + 1\times\beta_{i1} = \beta_{i0} + \beta_{i1}\,.$$

Le vecteur de contraste $(0,1)^T$ permet ainsi la comparaison de l'expression des gènes entre les groupes d'échantillons B et A. Grâce à l'utilisation d'un intercept et de la fonction log comme fonction de lien du GLM, identifier des *fold-changes* de moyenne d'expression entre deux groupes d'échantillons définis par la variable explicative X_k revient à estimer le coefficient de régression β_{ik} correspondant.

Lorsque les groupes d'échantillons d'intérêt sont issus de différentes expériences de séquençage, la variabilité due uniquement à ces différentes expériences peut être isolée et ne pas interférer avec les effets que l'on cherche à observer entre les groupes d'échantillons d'intérêt. Il s'agit alors d'ajouter une variable explicative au GLM, contenant les expériences ayant généré les échantillons. On parle d'effet bloquant. Par exemple, pour la comparaison de trois groupes d'échantillons d'intérêt ayant été séquencés par deux expériences différentes tels que définis dans la table 2.2, le GLM s'écrit :

$$\log(\mu_{ij}) = \beta_{i0} + \beta_{i1}\, x_{j1} + \beta_{i2}\, x_{j2} + \beta_{i3}\, x_{j3}\,.$$

Et la matrice de *design* s'écrit :

Echantillon	Groupe	Expérience
1	A	1
2	A	2
3	B	1
4	B	1
5	B	2
6	C	1
7	C	2
8	C	2

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

TABLE 2.2 – Exemple de comparaison de trois groupes d'échantillons issus de deux expériences de séquençage et matrice de *design* correspondante. Dans cet exemple, l'échantillon du groupe A issu de l'expérience 1 est pris comme référence.

Les deuxième et troisième colonnes permettent de représenter les différences entre les trois groupes d'échantillons définis par la variable explicative d'intérêt. La quatrième colonne permet de prendre en compte les différences induites par les deux expériences ayant permis de séquencer les échantillons.
Dans cet exemple, la comparaison des échantillons issus des groupes A et C en prenant en compte la variabililité introduite par les différentes expériences de séquençage se fait à l'aide du vecteur de contraste $(0,0,1,0)^T$. Les différences d'expression entre ces deux groupes sont ainsi modélisées par les coefficients β_{i2}. La prise en compte du coefficient de régression β_{i3} dans la modélisation de l'expression des gènes permet de capter les variations d'expression dues aux différentes expériences de séquençage. Le coefficient de régression β_{i2} représente alors les différences d'expression entre les deux groupes d'échantillons d'intérêt, non perturbées par les effets des différentes expériences.

Malgré l'utilisation d'un intercept, des comparaisons ne faisant pas intervenir le groupe d'échantillons de référence sont aussi possibles par l'emploi de vecteurs de contraste spécifiques.

Ajustement du modèle linéaire généralisé Les paramètres de régression β_{ik} sont estimés de manière itérative pour chaque gène i par des méthodes classiques de la théorie des GLM. Brièvement, le processus consiste en la régression par les moindres carrés pondérés d'une variable auxiliaire Z issue de la dérivée de la log-vraisemblance relativement aux coefficients de régression β_{ik} sur les variables explicatives X_k. Chaque itération produit généralement une augmentatation de la fonction de vraisemblance du GLM. Le processus est initialisé avec des valeurs de départ et la régression est répétee jusqu'à ce que les valeurs des coefficients de régression convergent vers les estimateurs du maximum de vraisemblance.
La convergence n'est pas toujours atteinte, en particulier lors de l'ajustement de GLM composés de nombreuses variables explicatives ou pour les petits jeux de données. Pour y remédier, McCarthy, Chen et Smyth, 2012 ont amélioré l'algorithme classique d'estimation des coefficients de régression avec une méthode d'optimisation de recherche linéaire. Conceptuellement, cette optimisation consiste en l'utilisation d'un écart suffisamment petit entre deux itérations pour que la fonction de vraisemblance augmente (Smyth, 1998).

Inférence statistique Le but de l'analyse de différentiel d'expression classique est de trouver des différences de moyenne d'expression entre deux conditions. On parle de *fold-change* (FC) d'expression :

$$FC_\mu = \frac{\mu_{i1}}{\mu_{i2}}\,.$$

Dans le cadre du GLM défini dans la formule 2.8, on utilise le *log-fold-change* pour écrire les hypothèses de test. L'hypothèse nulle H_0 représentant l'absence de différence de moyenne s'écrit donc $H_0 : \log(FC_\mu) = 0$ et l'hypothèse alternative s'écrit $H_a : \log(FC_\mu) \neq 0$. Dans le cadre des GLM basés sur la distribution binomiale négative et avec la fonction log comme fonction de lien tels que définis dans la formule 2.8, les coefficients de régression β_{ik} peuvent être interprétés comme une mesure du *log-fold-change* de moyenne d'expression entre les groupes d'échantillons définis par les niveaux de la variable explicative X_k. Le test porte alors sur le coefficient de régression β_{ik} correspondant à la variable explicative X_k testée et les hypothèses s'écrivent :

$$H_0 : \beta_{ik} = 0 \text{ et } H_a : \beta_{ik} \neq 0\,.$$

Le test revient à comparer le modèle complet à un modèle privé de la variable explicative X_k à évaluer, *i.e.* le modèle avec $\beta_{ik} = 0$. Pour ce faire, des tests dont la distribution asymptotique sous H_0 est connue sont utilisés. Le test du rapport de vraisemblance entre le modèle complet et le modèle contraint avec $\beta_{ik} = 0$ dont la satistique suit une distribution du χ^2 est implémenté dans les *packages* edgeR et DESeq2. Par défaut, DESeq2 utilise des tests de Wald dont la statistique de test est :

$$W = \frac{\hat{\beta}_{ik} - \beta_{ik}^{H_0}}{se(\hat{\beta}_{ik})} \sim \mathcal{N}(0,1)\,,$$

où :
- $\beta_{ik}^{H_0}$ est la valeur de β_{ik} sous H_0, ici $\beta_{ik}^{H_0} = 0$;
- $se(\hat{\beta}_{ik})$ est l'écart-type de l'estimateur $\hat{\beta}_{ik}$ issu de la matrice de variance-covariance des coefficients de régression.

A noter qu'une combinaison linéaire de coefficients de régression peut être testée simultanément.

3.3 La dispersion dans les modèles de différence de moyenne d'expression

Pour pouvoir réaliser des tests statistiques sur la moyenne, les méthodes exposées dans la section 3.2, tests exacts et GLM, requièrent que le second paramètre de la distribution binomiale négative, la dispersion, ait été estimé préalablement. L'estimation du paramètre de dispersion d'une distribution binomiale négative est une tâche plus complexe que l'estimation de la moyenne. En effet, elle nécessite beaucoup plus de données pour pouvoir être réalisée de manière fiable. Or, les jeux de données RNA-seq sont en général constitués de seulement quelques échantillons, rendant très difficile l'estimation de la dispersion pour chaque gène de manière indépendante. Dans le but d'augmenter le volume de données utilisées pour estimer la dispersion, plusieurs approches ont été développées, en estimant la dispersion :

- pour chaque gène de manière indépendante à partir des nombres de *reads* issus de tous les échantillons qui constituent le jeu de données, quelle que soit la population à laquelle ils appartiennent ;
- pour un ensemble de gènes, cet ensemble pouvant être l'intégralité des gènes du jeu de données ;
- par un compromis entre les deux premières approches.

Selon qu'elles s'insèrent dans le cadre d'un test exact ou d'un GLM, les méthodes d'estimation de la dispersion subiront quelques modifications mais suivront les mêmes principes.

3.3.1 Estimateur du maximum de vraisemblance

La moyenne et la dispersion d'une distribution binomiale négative telle que définie dans la formule 2.7 peuvent être estimées par l'approche du maximum de vraisemblance. Les fonctions de vraisemblance et log-vraisemblance sont ainsi définies (Lawless, 1987) :

$$\begin{aligned} L(\mu,\,\phi|y_1,\,y_2,\,\ldots,\,y_n) &= \textstyle\prod_{i=1}^{n} \frac{\Gamma(\phi^{-1})+y_i}{\Gamma(\phi^{-1})y_i!} \left(\frac{1}{1+\mu\,\phi}\right)^{\phi_{ij}^{-1}} \left(\frac{\mu\,\phi}{\mu\,\phi+1}\right)^{y_i}\,, \\ \ln(L(\mu,\,\phi)) = l(\mu,\,\phi) &= \textstyle\sum_{i=1}^{n} \Bigg(\sum_{j=0}^{y_i-1} \ln(j+\phi^{-1}) - (y_i+\phi^{-1})\ln(1+\mu\,\phi) \\ &\qquad + y_i \ln(\mu) - \ln(y_i!) \Bigg)\,. \end{aligned} \tag{2.9}$$

où $\sum_{j=0}^{y_i-1} \ln(j+\phi^{-1}) = 0$ si $y_i < 1$, $j = 0,\,1,\,\ldots,\,y_i - 1$.
Les estimateurs $\hat{\mu}$ et $\hat{\phi}$ du maximum de vraisemblance de la moyenne et de la dispersion respectivement s'obtiennent en maximisant la log-vraisemblance, *i.e* en trouvant en quelles valeurs les dérivées partielles de la log-vraisemblance en fonction de μ et de ϕ s'annulent :

$$\frac{\partial\, l}{\partial\, \mu} = \sum_{i=1}^{n} \left(\frac{y_i}{\mu} - \frac{y_i+\phi^{-1}}{\mu+\phi^{-1}} \right) = 0\,,$$

$$\frac{\partial\, l}{\partial\, \phi} = \sum_{i=1}^{n} \left(\sum_{j=0}^{y_i-1} \frac{j}{1+\alpha\, j} + \phi^{-2}\ln(1+\phi\,\mu) - \frac{\mu(y_i+\phi^{-1})}{1+\phi\,\mu} \right) = 0\,.$$

Résoudre la première équation révèle que l'estimateur $\hat{\mu}$ est la moyenne des valeurs observées. L'estimateur $\hat{\phi}$ est obtenu en résolvant la seconde équation avec $\mu = \hat{\mu}$. Pour y parvenir, une méthode d'optimisation approchée telle que la méthode de Newton-Raphson doit être utilisée (Park et Lord, 2008).

L'estimateur du maximum de vraisemblance de la dispersion est biaisé et tend à sous-estimer celle-ci car il échoue à prendre en compte le fait que la moyenne est estimée à partir des mêmes données (Robinson et Smyth, 2008). De plus, un grand nombre d'échantillons est requis pour une bonne estimation de la dispersion par maximum de vraisemblance. Cette contrainte ne constitue pas un problème dans certains domaines d'études. Lord a ainsi estimé la fiabilité de cet estimateur en fonction du nombre d'échantillons disponibles à l'aide d'une étude de simulation dans le cadre de ses travaux de modélisation d'accidents de la route sur les autoroutes (Lord, 2006). Pour des nombres d'échantillons jugés faibles, *i.e.* 50 et 100 dans son étude, l'estimateur du maximum de vraisemblance est peu fiable pour les faibles valeurs de dispersion (0,5 dans son étude). En revanche, pour une grande population de 1 000 échantillons, l'estimateur du maximum de vraisemblance fournit une estimation précise, quelle que soit la valeur de dispersion à estimer.
Dans le contexte de l'analyse de données RNA-seq, de telles tailles de population d'échantillons paraissent irréalistes, en particulier à l'époque de la publication des premières méthodes visant à les analyser où les jeux de données étaient en général constitués de 3 à 5 échantillons par condition d'intérêt. Ainsi, d'autres stratégies ont dû être déployées pour pouvoir estimer la dispersion de l'expression de gènes.

3.3.2 Quasi-vraisemblance

Les fonctions de quasi-vraisemblance sont des fonctions qui ont les mêmes propriétés que les fonctions de log-vraisemblance mais ne sont pas la fonction de log-vraisemblance d'une quelconque loi de probabilité (Wedderburn, 1974). Elles permettent, par exemple, l'estimation d'un paramètre de sur-dispersion dans le cadre d'un modèle basé sur la loi de Poisson. On parle alors de modèle quasi-Poisson. Ici, dans le cadre d'une modélisation par la loi binomiale négative, la dispersion faisant partie des paramètres de cette loi de probabilités, le terme de quasi-vraisemblance fait référence à l'emploi d'une fonction différente de la log-vraisemblance de la binomiale négative pour la dispersion telle que définie dans la formule 2.9.

La méthode de quasi-vraisemblance (QL, pour *Quasi-Likelihood*) introduite par Robinson et Smyth, 2008 consiste à estimer la moyenne μ_{ij} par maximum de vraisemblance et la dispersion ϕ_{ij} par maximum de quasi-vraisemblance :

— l'estimateur du maximum de vraisemblance de la moyenne, $\hat{\mu}_{ij}$, en maximisant la log-vraisemblance de la distribution binomiale négative, étant donné les nombres de *reads* observés y_{ij} et l'estimateur de la dispersion $\hat{\phi}_i$;
— l'estimateur de la dispersion, $\hat{\phi}_i$, est calculé à l'aide d'un modèle de quasi-vraisemblance et de l'estimateur $\hat{\mu}_{ij}$ de la moyenne déterminé à l'étape précédente en résolvant l'équation :

$$2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\mu}_{ij}}\right) - (y_i + \phi_i^{-1}) \log\left(\frac{y_i + \phi^{-1}}{\hat{\mu}_{ij} + \phi^{-1}}\right)\right] = n - 1\,.$$

Un estimateur de la dispersion est ainsi obtenu de manière indépendante pour chaque gène, c'est-à-dire que seul les nombres de *reads* associés au gène i sont utilisés pour déterminer $\hat{\phi}_i$. Lorsque le nombre d'échantillons est faible, et le nombre de valeurs de dispersion à estimer étant beaucoup plus important, cette méthode est sous-optimale car elle ne tire pas profit de toute l'information disponible dans le jeu de données dans

son ensemble. Ainsi, d'autres méthodes, partageant l'information issue de plusieurs gènes ont été développées.

3.3.3 Vraisemblance conditionnelle ajustée par quantiles pondérée

Dans la première version de leur méthode edgeR, Robinson et Smyth, 2008 font le constat que le très faible nombre d'échantillons par condition rend impossible une estimation fiable de la dispersion ϕ_i pour chaque gène i de manière indépendante. En effet, à l'époque de la publication de leur méthode, les jeux de données RNA-seq n'étaient très souvent composés que de trois échantillons par condition. De plus, ils notent qu'il n'y a aucune preuve à cette époque que des gènes puissent avoir des dispersions d'expression très différentes. Ils décident donc d'estimer une seule valeur de dispersion d'expression ϕ commune à l'ensemble des gènes. Leur approche consiste à estimer la dispersion à partir des nombres de *reads* issus des échantillons de l'ensemble des conditions. A l'aide d'une étude de simulation, ils montrent que cette méthode est la plus fiable comparée aux approches classiques de vraisemblance ou de quasi-vraisemblance. Dans cette approche, les échantillons issus des différentes conditions sont donc supposés avoir la même valeur de dispersion.

Dans le cas où les librairies ont des tailles équivalentes, la somme $Z_i = Y_{i1} + Y_{i2} + \cdots + Y_{in}$ suit aussi une distribution binomiale négative de paramètres $n\,m\,\lambda_i$ et $\phi\,n^{-1}$, où m est la moyenne géométrique des tailles de librairie des échantillons. La somme des nombres de *reads* peut être utilisée pour estimer λ_i. Une fonction de vraisemblance conditionnelle exacte pour la dispersion indépendante du paramètre λ_i peut alors être définie :

$$l_{(Y_i|Z_i=z)}(\phi_i) = \left[\sum_{i=1}^{n} \log\Gamma(y_i + \phi_i^{-1})\right] + \log\Gamma(n\,\phi_i^{-1}) - \log\Gamma(z + n\phi_i^{-1}) - n\log\Gamma(\phi_i^{-1})\,. \tag{2.10}$$

Un estimateur de la dispersion commune $\hat{\phi}$ à l'ensemble des gènes G est obtenu en maximisant la vraisemblance conditionnelle de l'ensemble des gènes :

$$l_c(\phi) = \sum_{i=1}^{G} l_{(Y_i|Z_i=z)}(\phi_i)\,. \tag{2.11}$$

Dans le cas où les librairies ne sont pas de tailles équivalentes, les données sont alors ajustées de telle sorte qu'elles soient issues d'une même distribution $\mathcal{NB}(m^*\,\lambda_i, \phi)$ où m^* est la moyenne géométrique des tailles de librairie : $m^* = (\prod_{i=1}^{n} m_i)^{\frac{1}{n}}$. Les données sont transformées de telle sorte que les distributions des nombres de *reads* des librairies aient les mêmes quantiles. Il s'agit en fait d'une normalisation des nombres de *reads* par quantiles (voir section 1.2). La dispersion commune à l'ensemble des gènes peut alors être estimée en maximisant la vraisemblance conditionnelle exacte définie dans la formule 2.11.

La méthode d'estimation d'une valeur de dispersion ϕ commune à l'ensemble des gènes permet d'avoir une estimation fiable de la dispersion en s'appuyant sur les nombres de *reads* de l'intégralité du jeu de données. Cependant, elle repose sur l'hypothèse que tous les gènes ont la même valeur de dispersion d'expression. Cette hypothèse apparaît comme biologiquement peu réaliste. Permettre l'estimation de différentes valeurs de dispersion pour le même jeu de données requiert le développement de nouvelles méthodes. Robinson et Smyth, 2007 proposent ainsi une nouvelle méthode où la

dispersion ϕ_i du gène i est déterminée comme un compromis entre une valeur estimée à partir des seuls nombres de *reads* de ce gène et la valeur de dispersion commune ϕ déterminée à l'aide de la vraisemblance conditionnelle. La dispersion ϕ_i de chaque gène est alors estimée en maximisant la vraisemblance pondérée :

$$WL(\phi_i) = l_i(\phi_i) + \alpha\, l_c(\phi_i)\,, \tag{2.12}$$

où :

- $l_i(\phi_i)$ est la log-vraisemblance conditionnelle pour les pseudo-comptes associés à chaque gène i définie dans la formule 2.10 ;
- $l_c(\phi_i)$ est la log-vraisemblance commune définie dans la formule 2.11 ;
- α est un coefficient permettant de pondérer l'importance relative de la log-vraisemblance commune $l_c(\phi_i)$ et de la log-vraisemblance individuelle $l_i(\phi_i)$.

L'estimation de la dispersion à partir des seuls nombres de *reads* du gène i est ainsi pondérée par la valeur de dispersion commune à l'aide du facteur α. Plus α est grand, plus la vraisemblance tend vers la vraisemblance commune et la dispersion de chaque gène i se rapproche de la valeur commune ϕ. En pratique, α est estimé à l'aide d'un modèle bayésien empirique dans lequel la dispersion ϕ_i pour chaque gène i est supposée suivre une loi normale de moyenne ϕ_0 et de variance τ_0^2 inconnues. Ces deux hyperparamètres peuvent être estimés à partir de la distribution marginale des estimateurs $\hat{\phi}_i$ de la dispersion de chaque gène i et ainsi former la règle de décision de cette approche bayésienne empirique. Le paramètre α est ensuite choisi de telle sorte que la log-vraisemblance conditionnelle pondérée coïncide avec cette règle.

Cette approche est implémentée dans le *package* R edgeR (Robinson, McCarthy et Smyth, 2010) et peut s'utiliser avec deux fonctions :

- *estimateCommonDisp()* : estime une valeur commune de dispersion $\hat{\phi}$ pour l'ensemble des gènes par l'approche de vraisemblance conditionnelle ajustée par quantiles définie dans la formule 2.11 ;
- *estimateDisp()* : estime une valeur de dispersion $\hat{\phi}_i$ pour chaque gène par l'approche de vraisemblance conditionnelle ajustée par quantiles pondérée définie dans la formule 2.12.

3.3.4 Vraisemblance profilée par ajustement de Cox-Reid

Les fonctions de vraisemblance profilée s'appliquent dans les cas où la fonction de vraisemblance dépend de plusieurs paramètres et que l'intérêt se porte principalement sur l'un d'entre eux. Les autres paramètres, dits de nuisance, sont exprimés en fonction du paramètre d'intérêt et remplacés dans la fonction de vraisemblance. Ici, le paramètre d'intérêt est la dispersion ϕ_i et la moyenne μ_i est considérée comme un paramètre de nuisance. L'ajustement de Cox-Reid (Cox et Reid, 1987) vise à corriger le biais introduit par l'estimateur du maximum de vraisemblance en pénalisant la log-vraisemblance relative à la dispersion par un terme contenant l'information observée pour la moyenne et est défini par :

$$l_{CR}(\phi) = l(\hat{\mu}_i, \phi_i) - \frac{1}{2}\log\left|\mathcal{I}_{\mu_i\mu_i}(\phi, \hat{\mu}_i)\right|\,, \tag{2.13}$$

où :

- $\hat{\mu}_i$ est l'estimateur du maximum de vraisemblance de la moyenne ;
- $\mathcal{I}_{\mu_i\mu_i}(\phi, \hat{\mu}_i)$ est l'information observée de Fisher pour la moyenne.

La vraisemblance conditionnelle définie dans la formule 2.10 suit une approche similaire dans la mesure où considérer la somme des nombres de *reads* comme une statistique pour la moyenne permet d'écrire une fonction de vraisemblance pour la dispersion qui ne dépend pas de la moyenne. L'avantage de la vraisemblance profilée par ajustement Cox-Reid est qu'elle peut être appliquée dans le cadre d'un GLM, au contraire de la vraisemblance conditionnelle qui ne peut s'appliquer qu'à la comparaison d'échantillons selon un unique facteur. Dans le cadre d'un GLM, la moyenne est exprimée dans la fonction de vraisemblance par la matrice d'information de Fisher. L'estimation de la dispersion ϕ_i pour chaque gène revient alors à maximiser la vraisemblance profilée ajustée :

$$APL_i(\phi_i) = l(\phi_i; y_i, \hat{\beta}_i) - \frac{1}{2} \log |\mathcal{I}_i| \,, \tag{2.14}$$

où :

- y_i est le vecteur de nombres de *reads* du gène i ;
- $\hat{\beta}_i$ est le vecteur de paramètres de régression estimés dans le cadre du GLM défini dans la formule 2.8 en l'absence de *fold-change* de moyenne à appliquer au gène i ;
- l est la fonction de log-vraisemblance relative à la dispersion calculée à partir de la fonction de masse de la distribution binomiale négative définie dans la formule 2.7 et la valeur estimée $\hat{\mu}_i^0$ de la moyenne à partir de $\hat{\beta}_i$:

$$l(\phi_i; y_i, \hat{\mu}_i^0) = \sum_j \log f(y_{ij}; \hat{\mu}_{ij}^0, \phi_i) \ ;$$

- $\log |\mathcal{I}_i|$ est le déterminant de la matrice d'information de Fisher, obtenu par la décomposition de Cholesky (Stewart, 1973).

Comme dans la section précédente, la dispersion ϕ_i n'est pas estimée pour chaque gène i de manière indépendante mais à partir d'un ensemble de gènes. Elle est obtenue grâce à un compromis entre l'estimation d'une dispersion commune à un ensemble de gènes incluant le gène i et une estimation indépendante pour chaque gène i. La manière la plus simple et la plus fiable est de considérer l'ensemble des gènes du jeu de données et d'estimer une dispersion commune $\hat{\phi}$ à l'ensemble des gènes. Elle est obtenue en maximisant la fonction de vraisemblance partagée :

$$APL_S(\phi) = \frac{1}{G} \sum_{i=1}^{G} APL_i(\phi) \,, \tag{2.15}$$

où :

- APL_i est la vraisemblance profilée ajustée du gène i définie dans la formule 2.14 ;
- G est le nombre total de gènes.

Cette maximisation peut être obtenue numériquement de différentes manières. Les auteurs d'edgeR ont opté pour la méthode de Newton-Raphson (Brent, 1973).

Une approche plus fine consiste à considérer un ensemble plus restreint de gènes et à exprimer la dispersion comme une fonction de la moyenne d'expression. Des sous-ensembles de gènes sont ainsi constitués selon l'expression moyenne des gènes et une dispersion commune est estimée pour chaque sous-ensemble. Une courbe de régression est ensuite obtenue à travers les estimations de la dispersion par régression locale (loess ou spline). La dispersion ϕ_i est alors estimée par la moyenne pondérée des APL

du gène i et de gènes dont les nombres de *reads* moyens sont proches de celui du gène i.

Enfin, une dernière approche consiste à estimer la dispersion ϕ_i de chaque gène i à l'aide d'un compromis entre l'estimation de la dispersion commune à un ensemble de gènes incluant le gène i et une estimation indépendante pour chaque gène i en maximisant la vraisemblance partagée :

$$APL_i(\phi_i) + G_0\, APL_{S_i}(\phi_i)\,, \tag{2.16}$$

où :

— $APL_{Si}(\phi_i)$ est la log-vraisemblance locale partagée d'un ensemble de gènes S_i ;

— G_0 est le poids donné à la log-vraisemblance locale partagée.

De manière similaire à la vraisemblance pondérée définie dans la formule 2.12, cette approche peut être vue comme une approche bayésienne où la vraisemblance partagée $APL_{Si}(\phi_i)$ est la distribution *a priori* de ϕ_i, la vraisemblance pondérée comme la distribution *a posteriori* et G_0 est le poids donné à la distribution *a priori*. McCarthy, Chen et Smyth, 2012 recommendent de prendre une valeur G_0 petite lorsque l'on suspecte que la dispersion varie beaucoup au sein du jeu de données. Ils recommendent $G_0 = \frac{20}{df}$ où df est le nombre de degrés de liberté résiduels pour estimer la dispersion (*i.e.* le nombre d'échantillons moins le nombre de populations d'échantillons). Une caractéristique de cette modélisation est le fait que l'estimation de ϕ_i tend d'autant plus vers l'estimation partagée que l'estimation indépendante de ϕ_i à partir des seuls comptes y_i est incertaine. Par exemple, les gènes faiblement exprimés verront leur estimateur de la dispersion tendre fortement vers l'estimateur partagé.

Les différentes approches basées sur la vraisemblance profilée ajustée sont implémentées dans le *package* R edgeR (Robinson, McCarthy et Smyth, 2010) par les fonctions suivantes :

— *estimateGLMCommonDisp()* : estime une valeur commune de dispersion $\hat{\phi}$ pour l'ensemble des gènes à l'aide de la vraisemblance profilée ajustée partagée définie dans la formule 2.15 ;

— *estimateGLMTagwiseDisp()* : estime une valeur de dispersion $\hat{\phi}_i$ pour chaque gène à l'aide de la vraisemblance profilée ajustée pondérée définie dans la formule 2.16 ;

— *estimateGLMTrendedDisp()* : estime une valeur de dispersion $\hat{\phi}_i$ pour chaque gène à partir de la tendance observée entre la dispersion déterminée à l'aide de la vraisemblance profilée ajustée partagée par sous-ensembles de gènes et la moyenne d'expression.

3.3.5 DESeq2

Dans la première version de leur méthode, DESeq, Anders et Huber, 2010 n'estiment pas un paramètre de dispersion pour quantifier la variance des données. Pour représenter les comptes *reads*, ils utilisent une paramétrisation de la distribution binomiale négative différente de celle définie dans la formule 2.6 et estiment directement la variance.

Pour la deuxième version de leur méthode, DESeq2, (Love, Huber et Anders, 2014) s'inscrivent dans le cadre d'un GLM tel que défini dans la formule 2.8 et estiment le paramètre de dispersion. Leur approche suppose que les gènes qui ont des

moyennes d'expression très proches ont aussi des valeurs de dispersion similaires et est très proche de l'approche d'estimation de la dispersion d'edgeR par maximisation de la vraisemblance profilée ajustée pondérée (voir section 3.3.4). Elle procède en 3 étapes :

1. estimation de la dispersion ϕ_i pour chaque gène i de manière indépendante par maximum de vraisemblance profilée par ajustement de Cox-Reid définie dans la formule 2.14 ;
2. une courbe de tendance entre la dispersion et la moyenne est obtenue par régression des ϕ_i sur la moyenne des nombres de *reads* normalisés ;
3. l'estimateur final de la dispersion est déterminée par une approche bayésienne empirique dont la distribution *a priori* est une distribution log-normale centrée sur la courbe de régression.

L'approche d'estimation de la dispersion développée par DESeq2 est illustrée par la figure 2.3.

Etant donnée la dépendance décroissante de la dispersion à la moyenne communément observée dans les données RNA-seq, Love, Huber et Anders, 2014 ont opté pour la paramétrisation suivante de la courbe de tendance :

$$\phi_{tr}(\bar{\mu}_i) = \frac{a_1}{\bar{\mu}} + \alpha_0 \,, \tag{2.17}$$

où :

— $\bar{\mu}$ est la moyenne des nombres de *reads* observée ;
— a_1 et α_0 sont deux hyperparamètres à estimer.

La distribution des estimateurs pouvant être asymétrique autour de la vraie valeur ϕ_i, une régression par GLM de la famille gamma est appliquée plutôt qu'une régression classique par les moindres carrés. Les hyperparamètres a_1 et α_0 sont obtenus lors de l'ajustement itératif du GLM de la famille gamma. La courbe de tendance est ensuite utilisée dans le cadre d'une approche bayésienne empirique pour paramétrer la distribution *a priori* telle que :

$$\log \phi_i \sim \mathcal{N}(\log \phi_{tr}(\bar{\mu}_i), \sigma_d^2) \,,$$

où σ_d^2 représente la largeur de la distribution *a priori* et décrit à quel point les vraies dispersions ϕ_i sont éparpillées autour de la courbe de tendance.
A l'inverse d'edgeR, l'hyperparamètre de variance σ_d^2 est estimé à partir des données. Il dépend du nombre de degrés de liberté du modèle qui est égal à la différence entre le nombre d'échantillons et le nombre de coefficients du GLM. Plus le nombre d'échantillons est grand, plus l'estimateur final tend vers l'estimateur du maximum de vraisemblance profilée du gène i.
Un seuil minimal est utilisé pour éviter que l'estimateur final de la dispersion $\hat{\phi}_i$ tende totalement vers la valeur estimée par la courbe de tendance. A l'inverse, les gènes dont l'estimateur de la dispersion par maximum de vraisemblance profilée est très éloigné de la courbe de tendance ne voient pas leur estimateur final de la dispersion affecté par la courbe de tendance.
Le maximum de la distribution *a posteriori* (MAP) construite à partir de la vraisemblance profilée par ajustement de Cox-Reid et de la distribution *a priori* est utilisé comme estimateur final de la dispersion :

$$\hat{\phi}_i^{MAP} = \text{argmax}_\phi \left(l_{CR}(\phi_i) + \frac{-(\log \phi - \log \phi_{tr}(\bar{\mu}_i))^2}{2\,\sigma_d^2} \right) , \tag{2.18}$$

où :

— $l_{CR}(\phi_i)$ est la vraisemblance profilée par ajustement de Cox-Reid ;

— ϕ est une constante additive.

La propension de l'estimateur final à tendre vers la courbe de tendance dépend de la proximité de la vraie valeur de la dispersion avec la courbe (figure 2.3).

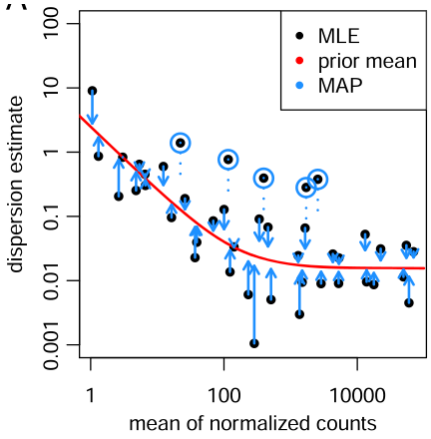

FIGURE 2.3 – Illustration de l'approche d'estimation de la dispersion développée dans le *package* DESeq2 (LOVE, HUBER et ANDERS, 2014). Les points représentent les estimateurs du maximum de vraisemblance de la dispersion obtenus à l'aide des nombres de *reads* de chaque gène respectivement. La courbe rouge est la courbe d'ajustement des estimateurs pour capturer la tendance entre la moyenne et la dispersion. Elle est utilisée comme *a priori* d'une seconde estimation de la dispersion consistant à ajuster les estimateurs de la dispersion propres à chaque gène à la tendance globale. Cet ajustement est représenté par les flèches bleues. Certains estimateurs sont considérés comme des valeurs aberrantes et ne sont pas modifiés par la tendance globale (cercles bleus).

3.3.6 Impact de l'estimation de la dispersion sur la détection de gènes différentiellement exprimés

Dans le cadre des méthodes de détection de différence de moyenne d'expression basées sur la distribution binomiale négative, l'estimation de la dispersion peut avoir des conséquences sur la détection de gènes différentiellement exprimés. En effet, une surestimation de celle-ci peut empêcher la détection de différence de moyenne d'expression et, à l'inverse, une sous-estimation peut amener à l'identification à tort de gènes différentiellement exprimés.

LANDAU et LIU, 2013 ont évalué la précision de différentes méthodes d'estimation de la dispersion et leur impact sur les performances des méthodes de détection de différence de moyenne d'expression basées sur les tests exacts présentés dans la section 3.2.1. Cette étude est basée sur des jeux de données simulées à partir des jeux de données RNA-seq réels de PICKRELL et al., 2010 et de HAMMER et al., 2010. Les caractéristiques des ces jeux de données, grand nombre d'échantillons pour PICKRELL et al., 2010 et grande profondeur de séquençage pour HAMMER et al., 2010, permettent d'avoir une estimation fiable de la dispersion par quasi-vraisemblance (section 3.3.2). Les valeurs d'estimation ainsi obtenues sont ensuite utilisées lors de la génération des jeux de données simulées.

Les méthodes d'estimation de la dispersion évaluées dans cette étude sont :

— la quasi-vraisemblance (QL, voir section 3.3.2) ;

— la vraisemblance conditionnelle ajustée par quantiles pondérée (wqCML pour *Weighted Quantile-adjusted Conditional Maximum Likelihood*, voir section 3.3.3) ;
— la vraisemblance profilée par ajustement de Cox-Reid (APL pour *Cox-Reid Adjusted Profile Likelihood*, voir section 3.3.4) ;
— la méthode *Dispersion Shrink for Sequencing* (DSS), non présentée dans ce document, qui est une approche très proche de la méthode d'estimation de la dispersion employée par DESeq2 (voir section 3.3.5).

La méthode d'estimation de la variance de DESeq est aussi incluse dans leurs comparaison.
Les tests de différence de moyenne d'expression évalués dans cette étude sont les tests exacts d'edgeR et de DESeq, ainsi que trois variantes d'un test issues du *package* QuasiSeq (Lund et al., 2012). Les méthodes d'estimation de la dispersion s'inscrivant dans le cadre d'un GLM ne font donc pas partie du périmètre de cette étude.
Les méthodes permettant une pondération modérée de l'estimateur de la dispersion propre à chaque gène par une tendance générale (*i.e.* DSS, wqCML et APL avec leur option d'estimation pour chaque gène et DESeq avec l'option « maximum »), donnent les estimations les plus précises de la dispersion en comparaison avec les méthodes sans correction (*i.e.* QL et DESeq avec l'option « non ») et les méthodes avec forte correction vers la tendance globale (*i.e.* les estimations d'une valeur commune de dispersion et les versions d'estimation de DESeq et d'APL ajustées à une tendance globale).
De manière assez logique, ce sont ces mêmes méthodes d'estimation de la dispersion avec une pondération modérée de l'estimateur de la dispersion propre à chaque gène par une tendance générale qui, utilisées avec les tests exacts d'edgeR et de DESeq, permettent d'avoir les meilleurs performances d'identification de gènes présentant une différence de moyenne d'expression entre deux conditions.

D'autres études ont été menées pour évaluer l'impact de l'estimation de la dispersion dans le cadre des tests de différence de moyenne d'expression (Wu, Wang et Wu, 2013, Yu, Huber et Vitek, 2013). Les conclusions de Landau et Liu, 2013 confirment les principales conclusions de ces études avec les avantages d'une simulation réaliste de jeux de données et d'un plus large ensemble de méthodes évaluées.

3.4 Méthodes de détection de différence de dispersion

Les méthodes d'analyse de données d'expression issues du RNA-seq se sont concentrées sur la détection de différence de moyenne entre populations d'échantillons d'intérêt. A cause du faible nombre d'échantillons qui composent les jeux de données RNA-seq, la dispersion est estimée à partir de l'ensemble des échantillons. Cela permet d'avoir plus d'échantillons pour estimer ce paramètre et ainsi avoir une meilleure précision. Par contre, le fait qu'un gène puisse avoir une dispersion d'expression différente entre les conditions d'intérêt n'est pas considéré, ce qui est biologiquement assez peu réaliste.
La baisse du coût du séquençage ainsi que le développement de grandes bases de données publiques ont permis d'envisager d'étudier la dispersion dans les données d'expression de gène. Ainsi, deux méthodes, MDSeq (Ran et Daye, 2017) et DiPhiSeq (Li et Lamere, 2018), permettant la détection de différence de moyenne et de dispersion au sein du même modèle, ont été récemment développées.
La présence de valeurs aberrantes dans les nombres de *reads* peut nuire à l'identification de différence de moyenne. Leur effet peut être encore plus important dans le

cadre de l'estimation de différence de dispersion. Ces deux méthodes portent ainsi une attention particulière à leur détection.

3.4.1 MDSeq

Reparamétrisation de la loi binomiale négative Dans le but de pouvoir construire un GLM à la fois pour la moyenne et la dispersion, Ran et Daye, 2017 proposent dans leur méthode MDSeq une paramétrisation de la distribution binomiale négative en fonction de la moyenne et la dispersion différente de celle exposée dans la formule 2.6 :

$$\begin{aligned} &Y_{ij} \sim \mathcal{NB}(\mu_{ij}, \phi_{ij})\,, \\ &\mathbb{E}(Y_{ij}) = \mu_{ij}\,, \\ &\mathrm{Var}(Y_{ij}) = \phi_{ij}\,\mu_{ij} \text{ pour } \phi_{ij} > 1\,. \end{aligned} \tag{2.19}$$

La fonction de masse s'écrit alors :

$$\mathbb{P}(Y_i = k|\mu_{ij}, \phi_{ij}) = \frac{\Gamma(k+\theta_{ij})}{\Gamma(k+1)\Gamma(\theta_{ij})} \left(\frac{1}{\phi_{ij}}\right)^{\theta_{ij}} \left(1-\frac{1}{\phi_{ij}}\right)^{k},$$

où $\theta_{ij} = \theta(\mu_{ij}, \phi_{ij}) = \frac{\mu_{ij}}{\phi_{ij}-1}$.

Cette nouvelle paramétrisation de la distribution binomiale négative permet d'appliquer un GLM reliant à la fois la moyenne et la dispersion à un ensemble de covariables à l'aide d'une relation log-linéaire :

$$\begin{aligned} Y_{ij} &\sim \mathcal{NB}(\mu_{ij}, \phi_{ij})\,, \\ \log(\mu_{ij}) &= \beta_0 + \textstyle\sum_p x_{jp}\beta_{ip}\,, \\ \log(\phi_{ij}) &= \gamma_0 + \textstyle\sum_q u_{jq}\gamma_{iq}\,. \end{aligned} \tag{2.20}$$

où :

- x_{jp} et u_{jq} sont des éléments des matrices de *design* pour la moyenne et la dispersion respectivement ;
- β_{ip} et γ_{iq} sont les coefficients de régression relatifs à x_{jp} et u_{jq} pour la moyenne et la dispersion respectivement à appliquer à l'échantillon j.

Ce modèle linéaire généralisé ne s'inscrit pas dans la définition des GLM de la famille des lois de probabilités exponentielles puisque la dispersion n'est pas fixée (voir section 3.2.2). Les méthodes classiques d'estimation des paramètres des GLM de cette famille de distributions ne peuvent donc pas être appliquées. Les auteurs de MDSeq optent donc pour d'autres méthodes issues de la théorie du maximum de vraisemblance. Les coefficients de régression β_{ip} et γ_{iq} pour la moyenne et la dispersion respectivement sont ainsi estimés en maximisant la log-vraisemblance du modèle :

$$\begin{aligned} l_{MD}(\beta_i, \gamma_i; y_i) = \textstyle\sum_{j=1}^{n} &\log\Gamma(y_{ij}, \theta_{ij}) - \log\Gamma(\theta_{ij}) - \log\Gamma(y_{ij}+1) - \theta_{ij}\,\log\phi_{ij} \\ &+ y_{ij}\log\left(1-\tfrac{1}{\phi_{ij}}\right), \end{aligned} \tag{2.21}$$

où :

- $y_i = (y_{i1}, y_{i2}, \ldots, y_{in})^T$ est le vecteur contenant les n nombres de *reads* du gène i ;
- $\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{iP})^T$ est le vecteur contenant les P coefficients de régression pour la moyenne ;
- $\gamma_i = (\gamma i1, \gamma_{i2}, \ldots, \gamma_{iQ})^T$ est le vecteur contenant les Q coefficients de régression pour la dispersion ;

— $\theta_{ij} = \frac{\mu_{ij}}{\phi_{ij}-1}$;
— $\Gamma(\cdot)$ est la fonction gamma ;
— $\phi_{ij} > 1$ implique que $\theta_{ij} > 0$.

Des techniques d'optimisation contraintes, un algortithme d'adaptive barrier (Lange, 2010) et la méthode de Broyden-Fletcher-Goldfarb-Shanno (BFGS, Nocedal, 2006), sont utilisées pour déterminer les estimateurs de maximum de vraisemblance des coefficients de régression β_{ip} et γ_{iq} pour la moyenne et la dispersion respectivement.

Les auteurs de MDSeq soulignent qu'ajuster un GLM pour à la fois la moyenne et la dispersion est un problème numérique plus complexe que le cas classique d'un GLM pour la moyenne avec la dispersion fixée. En particulier, la méthode développée ne converge pas dans tous les cas vers un maximum pour à la fois la moyenne et la dispersion. En effet, la log-vraisemblance n'est pas tout le temps concave pour les coefficients de la moyenne et de la dispersion (Daye, Chen et Li, 2012).

Tests sur les coefficients du modèle linéaire généralisé Comme indiqué dans la section 3.2.2, les coefficients β_{ip} et γ_{iq} sont évalués en testant le modèle complet avec le modèle dépourvu du coefficient à tester, *i.e.* avec $\beta_{ip} = 0$ ou $\gamma_{iq} = 0$. Les hypothèses s'écrivent donc :

$$H_0 : \beta_{ip} = 0 \text{ vs } H_a : \beta_{ip} \neq 0\,,$$

$$H_0 : \gamma_{iq} = 0 \text{ vs } H_a : \gamma_{iq} \neq 0\,.$$

MDSeq applique des tests de Wald pour évaluer ces hypothèses avec les statistiques suivantes :

$$W_{\beta_{ip}} = \frac{\hat{\beta}_{ip}^2}{\mathrm{Var}(\hat{\beta}_{ip})} \text{ et } W_{\gamma_{iq}} = \frac{\hat{\gamma}_{iq}^2}{\mathrm{Var}(\hat{\gamma}_{iq})}\,,$$

où :

— $\hat{\beta}_{ip}$ et $\hat{\gamma}_{iq}$ sont les estimateurs des coefficients de régression pour la moyenne et la dispersion respectivement obtenus par la méthode BFGS ;
— $\mathrm{Var}(\hat{\beta}_{ip})$ et $\mathrm{Var}(\hat{\gamma}_{iq})$ sont l'inverse des informations de Fisher observées.

Les statistiques $W_{\beta_{ip}}$ et $W_{\gamma_{iq}}$ suivent la loi du χ_1^2 sous H_0.

Tests de différence de moyenne et de dispersion Les différences de moyenne et de dispersion sont évaluées sous la forme de *log-fold-changes* entre deux ensembles d'échantillons définis par deux valeurs d'une variable explicative. On note :

$$\log\left(\frac{\hat{\mu}_{il_1}}{\hat{\mu}_{il_0}}\right) = \sum_{p=1}^{L-1}(c_p^{l_1} - c_p^{l_0})\hat{\beta}_{ip} \text{ et } \log\left(\frac{\hat{\phi}_{il_1}}{\hat{\phi}_{il_0}}\right) = \sum_{q=1}^{L-1}(c_q^{l_1} - c_q^{l_0})\hat{\gamma}_{iq}\,,$$

où :

— l_0 et l_1 sont deux valeurs d'une variable explicative et L est le nombre total de valeurs que peut prendre cette variable explicative ;
— $\log\left(\frac{\hat{\mu}_{il_1}}{\hat{\mu}_{il_0}}\right)$ et $\log\left(\frac{\hat{\phi}_{il_1}}{\hat{\phi}_{il_0}}\right)$ sont les *log-fold-changes* de moyenne et de dispersion respectivement ;
— $c^l = (c_1^l, c_2^l, \ldots, c_{L-1}^l)^T$ est un vecteur de contraste définissant les échantillons ayant la variable explicative prenant la valeur l (voir section 3.2.2).

Les hypothèses des tests de différence de moyenne et de dispersion s'écrivent donc ainsi :

$$H_0^{\mu_i} : \log\left(\frac{\mu_{il_1}}{\mu_{il_0}}\right) = 0 \text{ vs } H_a^{\mu_i} : \log\left(\frac{\mu_{il_1}}{\mu_{il_0}}\right) \neq 0\,,$$

$$H_0^{\phi_i} : \log\left(\frac{\phi_{il_1}}{\phi_{il_0}}\right) = 0 \text{ vs } H_a^{\phi_i} : \log\left(\frac{\phi_{il_1}}{\phi_{il_0}}\right) \neq 0\,.$$

Des tests de Wald sont utilisés pour évaluer ces hypothèses. Les statistiques sont les suivantes :

$$W_{\mu_i} = \frac{\left[\log\left(\frac{\hat{\mu}_{il_1}}{\hat{\mu}_{il_0}}\right)\right]^2}{Var\left[\log\left(\frac{\hat{\mu}_{il_1}}{\hat{\mu}_{il_0}}\right)\right]} \text{ et } W_{\phi_i} = \frac{\left[\log\left(\frac{\hat{\phi}_{il_1}}{\hat{\phi}_{il_0}}\right)\right]^2}{Var\left[\log\left(\frac{\hat{\phi}_{il_1}}{\hat{\phi}_{il_0}}\right)\right]}\,,$$

où $\text{Var}\left[\log\left(\frac{\hat{\mu}_{il_1}}{\hat{\mu}_{il_0}}\right)\right] = (c_p^{l_1} - c_p^{l_0})^T \text{Var}\left(\hat{\beta}_{i1}, \hat{\beta}_{i2}, \ldots, \hat{\beta}_{iL-1}\right)(c_p^{l_1} - c_p^{l_0})$ et $\text{Var}\left[\log\left(\frac{\hat{\phi}_{il_1}}{\hat{\phi}_{il_0}}\right)\right] = (c_q^{l_1} - c_q^{l_0})^T \text{Var}\left(\hat{\gamma}_{i1}, \hat{\gamma}_{i2}, \ldots, \hat{\gamma}_{iL-1}\right)(c_q^{l_1} - c_q^{l_0})$ sont les variances des *log-fold-changes* de la moyenne et de la dispersion respectivement.
Les statistiques W_{μ_i} et W_{ϕ_i} suivent la loi du χ_1^2 et les p-valeurs de ces tests sont déduites de cette distribution.

Ces tests étant appliqués à l'ensemble des gènes qui composent le jeu de données à analyser, une correction de tests multiples doit être appliquée pour contrôler le taux de faux positifs dans cette situation de tests multiples. La méthode de Benjamini-Yekutieli est employée ici pour sa capacité à être appliquée à des tests impliquant des grandeurs dépendantes (Benjamini et Yekutieli, 2001).

Tests de différence de moyenne et de dispersion au-delà d'un seuil de *log-fold-change* Dans le cas où un grand nombre d'échantillons est disponible, les tests standards visant à détecter la moindre différence de moyenne d'expression peut amener à l'obtention d'un très grand nombre de gènes différentiellement exprimés alors que la différence de moyenne d'expression peut être faible. MDSeq propose donc la possibilité d'utiliser des seuils de *log-fold-change* τ pour la détection de différence de moyenne et de dispersion en évaluant les hypothèses suivantes :

$$H_0^{\tau,\mu_i} : \left|\log\left(\frac{\mu_{il_1}}{\mu_{il_0}}\right)\right| = \tau \text{ vs } H_a^{\tau,\mu_i} : \left|\log\left(\frac{\mu_{il_1}}{\mu_{il_0}}\right)\right| > \tau\,,$$

$$H_0^{\tau,\phi_i} : \left|\log\left(\frac{\phi_{il_1}}{\phi_{il_0}}\right)\right| = \tau \text{ vs } H_a^{\tau,\phi_i} : \left|\log\left(\frac{\phi_{il_1}}{\phi_{il_0}}\right)\right| > \tau\,.$$

Ces tests consistent en la réalisation de deux tests unilatéraux qui, dans le cas de la moyenne, sont définis par les hypothèses alternatives :

$$H_{a-}^{\tau,\mu_i} : \log\left(\frac{\mu_{il_1}}{\mu_{il_0}}\right) < -\tau \text{ et } H_{a+}^{\tau,\mu_i} : \log\left(\frac{\mu_{il_1}}{\mu_{il_0}}\right) > \tau\,.$$

Les statistiques de Wald dans ces espaces de paramètres restreints sont ainsi définies :

$$W_{\mu_i}^{+} = \begin{cases} \frac{\left[\log\left(\frac{\hat{\mu}_{il_1}}{\hat{\mu}_{il_0}}\right) - \tau\right]^2}{\text{Var}\left[\log\left(\frac{\hat{\mu}_{il_1}}{\hat{\mu}_{il_0}}\right)\right]} \text{ si } \log\left(\frac{\hat{\mu}_{il_1}}{\hat{\mu}_{il_0}}\right) > \tau\,, \\ 0 \text{ sinon } ; \end{cases}$$

$$W^-_{\mu_i} = \begin{cases} \frac{\left[\log\left(\frac{\hat{\mu}_{il_1}}{\hat{\mu}_{il_0}}\right)+\tau\right]^2}{\mathrm{Var}\left[\log\left(\frac{\hat{\mu}_{il_1}}{\hat{\mu}_{il_0}}\right)\right]} & \text{si } \log\left(\frac{\hat{\mu}_{il_1}}{\hat{\mu}_{il_0}}\right) < \tau\,, \\ 0 \text{ sinon}\,. \end{cases}$$

Les p-valeurs de ces deux tests sont obtenues à l'aide de mélanges de distributions du χ^2 et le minimum de ces deux p-valeurs est utilisé comme p-valeur du test de l'hypothèse alternative H_a^{τ,μ_i} selon de principe d'union-intersection.

La même procédure est appliquée pour les tests de différence de dispersion utilisant un seuil de *log-fold-change*.

Détection de valeurs aberrantes La présence de valeurs aberrantes peut avoir un fort impact sur les tests de différence de moyenne et de dispersion d'expression. MDSeq propose une fonction qui vise à retirer ces nombres de *reads* extrêmes pour ne pas perturber l'inférence de différence de moyenne et de dispersion d'expression. Pour ce faire, chaque nombre de *reads* y_{ij} est évalué individuellement à l'aide de deux tests de rapport de vraisemblance entre la vraisemblance telle que définie dans la formule 2.21 avec les coefficients β_{ip} et γ_{iq} d'intérêt fixé à 0 et celle du modèle complet en présence et en l'absence de y_{ij}. L'écart entre les deux statistiques de test calculées permet de mesurer l'influence du compte y_{ij} sur les tests de différence de moyenne et de dispersion basés sur les coefficients β_{ip} et γ_{iq}. Ainsi, une grande différence de statistique suggère que y_{ij} est un *outlier* et ne doit pas être pris en compte pour l'identification de gènes différentiellement exprimés ou dispersés.

Cette approche nécessite le calcul de log-vraisemblances pour chaque nombre de *reads*, ce qui peut engendrer des temps de calculs très longs pour les grands jeux de données. Pour y remédier, les auteurs de MDSeq ont implémenté une procédure d'optimisation permettant le calcul des différences de statistique de test de rapport de vraisemblance pour tous les nombres de *reads* d'un même gène en une seule étape (voir la section 1.7 des méthodes supplémentaires de l'article de MDSeq (Ran et Daye, 2017)).

3.4.2 DiPhiSeq

Récemment, une nouvelle méthode, DiPhiSeq, a été introduite pour permettre la détection de différence de moyenne et de dispersion d'expression dans des données RNA-seq (Li et Lamere, 2018). Les auteurs de cette méthode modélisent les nombres de *reads* de la façon suivante :

$$Y_{ij} \sim \mathcal{NB}\left(\sum_{k=1;2} d_j \exp(\beta_{ik}) I_{j\in S_k}, \sum_{k=1;2} \phi_{ik} I_{j\in S_k}\right), \tag{2.22}$$

où :

- d_j est la profondeur de séquençage de l'échantillon j ;
- S_k est l'ensemble des échantillons appartenant à la condition k ;
- $\exp(\beta_{ik})$ est l'espérance de l'expression du gène i dans S_k ;
- ϕ_{ik} est la dispersion de l'expression du gène i dans S_k ;
- I est la fonction indicatrice qui vaut 1 si la condition est remplie, 0 sinon.

Cette modélisation permet ainsi d'estimer la moyenne et la dispersion ϕ_{ik} pour chaque condition séparément. Une différence de moyenne d'expression (DE) et de dispersion (DD) d'expression entre deux conditions (1 et 2) peuvent ainsi être évaluées à l'aide

des hypothèses de test :

$$H_0^{DE} : \beta_{i1} = \beta_{i2} \text{ et } H_a^{DE} : \beta_{i1} \neq \beta_{i2}\,,$$
$$H_0^{DD} : \phi_{i1} = \phi_{i2} \text{ et } H_a^{DD} : \phi_{i1} \neq \phi_{i2}\,.$$

L'espérance est modélisée à l'aide d'un GLM basé sur une fonction de lien log :

$$\log(\mu_{ij}) = \log\, d_j + \beta_{ik}\,.$$

Contrairement aux GLM classiques définis dans la section 3.2.2, le GLM utilisé ne peut prendre en compte qu'une seule variable explicative, celle spécifiant l'appartenance des échantillons à leur condition. De plus, cette variable explicative est limitée à deux valeurs distinctes : les deux conditions d'intérêt.

Dans le cadre du modèle defini dans la formule 2.22, Li et Lamere, 2018 soulignent la sensibilité du test de rapport de vraisemblance à la présence d'*outliers* pour identifier des différences de dispersion entre les deux conditions. Pour y remédier, ils proposent d'utiliser un M-estimateur robuste (Cantoni et Ronchetti, 2001, Aeberhard, Cantoni et Heritier, 2014) pour estimer la moyenne et la dispersion dans les deux conditions. Les M-estimateurs sont une généralisation de l'estimation du maximum de vraisemblance et consistent en la minimisation d'une fonction ρ sur un ensemble de données x_i (Huber, 1964) :

$$\hat{\theta} = \text{argmin}_\theta \left(\sum_{i=1}^{n} \rho(x_i, \theta) \right), \tag{2.23}$$

où θ est le paramètre à estimer et $\hat{\theta}$ son M-estimateur.
Les M-estimateurs s'obtiennent lorsque la dérivée partielle ψ de la fonction ρ en fonction de θ s'annule :

$$\psi(x_i) = \frac{\partial}{\partial\theta} \left(\sum_{i=1}^{n} \rho(x_i, \theta) \right) = 0\,.$$

De nombreuses fonctions ρ peuvent être utilisées, notamment la fonction $\rho(x) = x^2$ qui est la fonction utilisée par la méthode des moindres carrés. Dans le cadre du modèle implémenté par DiPhiSeq, les M-estimateurs de β_{ik} et de ϕ_{ik} sont obtenus en résolvant les équations de quasi-vraisemblance robustes :

$$U_\beta(\beta_{ik}, \phi_{ik}) = \sum_{j \in S_k} \left[\psi_{c_\beta}(r_{ijk}) V_{ijk}^{-\frac{1}{2}} \mu_{ijk} - a_{ijk} \right] = 0\,,$$

$$U_\phi(\beta_{ik}, \phi_{ik}) = \sum_{j \in S_k} \left[\frac{\psi_{c_\phi}(r_{ijk})}{r_{ijk}} \Psi_{ijk} - b_{ijk} \right] = 0\,,$$

où :

- μ_{ijk} est l'espérance de Y_{ij} dans la condition k ;
- $V_{ijk} = \mu_{ijk} + \phi_{ik}\mu_{ijk}^2$ est la variance de Y_{ij} dans la condition k ;
- $r_{ijk} = \frac{(y_{ijk} - \mu_{ijk})}{\sqrt{V_{ijk}}}$ est le résidu de Pearson ;
- $\psi_c(r) = \begin{cases} ((r/c)^2 - 1)^2\, r \text{ si } |r| \leq c \\ 0 \text{ si } |r| > c \end{cases}$ est la dérivée de la fonction de Tukey à 2 poids et c est une constante prédéfinie pour β (c_β) et ϕ (c_ϕ) qui détermine

la robustesse des estimateurs ;

— $\Psi_{ijk} = (-1/\phi_{ik}^2)\left[(y_{ijk} + 1/\phi_{ik}) - F(1/\phi_{ik}) - \log(\phi_{ik}\mu_{ijk} + 1) - \frac{\phi_{ik}(y_{ijk}-\mu_{ijk})}{\phi_{ik}\mu_{ijk}+1}\right]$ est la dérivée partielle de la fonction de log-vraisemblance de ϕ_{ik}, où $F(x) = \partial \log \Gamma / \partial x$ est la fonction digamma ;

— $a_{ijk} = E\left[\psi_{c_\beta}(r_{ijk})\right] V_{ijk}^{-1/2}$ et $b_{ijk} = E\left[\frac{\psi_{c_\phi}(r_{ijk})}{r_{ijk}}\Psi_{ijk}\right]$ sont deux constantes assurant la consistance de Fisher des estimateurs.

Si $\psi(r) = r$, alors les deux équations reviennent aux équations de quasi-vraisemblance classiques pour β_{ik} et ϕ_{ik}. Ici, la fonction de Tukey à deux poids est la fonction ρ utilisée dans le cadre de la M-estimation des paramètres β_{ik} et ϕ_{ik}. La robustesse de ce modèle réside dans l'application d'un poids $\psi(r_{ijk})/r_{ijk}$ pour chaque nombre de *reads* y_{ijk} à l'aide de la fonction de Tukey à 2 poids qui donne la valeur de 0 aux *outliers*, *i.e.* les nombres de *reads* dont le résidu de Pearson est supérieur à une valeur c, les excluant ainsi de la fonction de quasi-vraisemblance. Le paramètre c détermine ainsi le niveau de robustesse : plus c est petit, plus l'estimateur est robuste et, à l'inverse, plus c est grand, moins la fonction de Tukey a d'influence et le modèle tend vers la modélisation classique de quasi-vraisemblance. La fonction de Tukey à deux poids est considérée par les auteurs de DiPhiSeq comme étant plus robuste aux *outliers* que la fonction de Huber définie par : $\psi_{c,Huber}(r) = \max(-c, \min(c, r))$. En effet, avec cette fonction de M-estimation, les *outliers* prennent la valeur $\psi_{c,Huber}(r) = c$ et sont donc pris en compte dans la quasi-vraisemblance du modèle.

L'estimation de β_{ik} et de ϕ_{ik} se fait en résolvant les deux équations itérativement, *i.e.* en obtenant ϕ_{ik} à partir de l'équation $U_\phi(\beta_{ik}, \phi_{ik}) = 0$ en fixant la valeur de β_{ik}, puis en obtenant β_{ik} à partir de l'équation $U_\beta(\beta_{ik}, \phi_{ik}) = 0$ en fixant la valeur de ϕ_{ik} avec la valeur précédemment obtenue. Les paramètres β_{ik} et de ϕ_{ik} sont initialisés et plusieurs cycles d'itérations sont réalisés tant que la différence des estimations de β_{ik} et ϕ_{ik} avec celles de l'itération précédente sont supérieures à des seuils Δ ($\Delta_\beta = 0{,}01$ et $\Delta_\phi = 0{,}005$).

Aeberhard, Cantoni et Heritier, 2014 ont montré que les estimateurs suivent asymptotiquement une loi gaussienne :

$$\sqrt{|S_k|}\begin{pmatrix}\hat{\beta}_{ik} - \beta_{ik} \\ \hat{\phi}_{ik} - \phi_{ik}\end{pmatrix} \sim \mathcal{N}\left(0, M_{ik}^{-1} Q_{ik} M_{ik}^{-T}\right),$$

où M_{ik} et Q_{ik} sont des matrices 2×2 dont les éléments sont précisés dans les données supplémentaires de l'article de DiPhiSeq (Li et Lamere, 2018).
Les différences de moyenne et de dispersion peuvent ensuite être testées à l'aide des statistiques de test :

$$\frac{\hat{\beta}_{i2} - \hat{\beta}_{i1}}{\sqrt{A_{i2,11} + A_{i1,11}}} \sim \mathcal{N}(0,1) \ , \ \frac{\hat{\phi}_{i2} - \hat{\phi}_{i1}}{\sqrt{A_{i2,22} + A_{i2,22}}} \sim \mathcal{N}(0,1),$$

où $A_{ik} = \frac{1}{|S_k|} M_{ik}^{-1} Q_{ik} M_{ik}^{-T}$ telle que $\mathrm{Var}(\hat{\beta}_{ik}) = A_{ik,11}$ et $\mathrm{Var}(\hat{\phi}_{ik}) = A_{ik,22}$.
Sous l'hypothèse nulle, ces deux statistiques suivent une loi normale centrée et réduite. Les p-valeurs de ces tests sont ainsi déduites et corrigées selon la méthode de Benjamini-Hochberg pour contrôler le taux de faux positifs dans le cadre de tests multiples (Benjamini et Hochberg, 1995).

4 Correction de tests multiples

Les tests de différence de moyenne et de dispersion d'expression sont appliqués de manière simultanée à chaque gène du jeu de données considéré. Pour chacun de ces tests, la probabilité de rejeter à tort l'hypothèse nulle d'égalité de moyenne, ou de dispersion, d'expression est contrôlée à l'aide du risque de première espèce α, communément fixé à 5%. A l'échelle du jeu de donnée tout entier, la probabilité d'effectuer des erreurs de type I, *i.e.* rejeter à tort l'hypothèse nulle, n'est pas de 5% mais égale à $1-(1-0{,}05)^n$. Ainsi, la probabilité d'observer un rejet de l'hypothèse nulle à tort augmente très rapidement avec le nombre de tests statistiques réalisés. Par exemple, elle est égale à 40% pour 10 tests. Les jeux de données d'expression pouvant être composés de milliers de gènes, un nombre important de résultats faux figurera parmi les résultats positifs. Il est donc nécessaire de contrôler les erreurs de type I. Pour ce faire, de nombreuses approches existent. Les deux principales sont les suivantes :

— le FWER (*Family-Wise Error Rate*) qui vise à maintenir la probabilité d'observer au moins un faux positif au niveau α ;
— le FDR (*False Discovery Rate*) qui vise à maintenir le nombre de faux positifs parmi l'ensemble des résultats positifs au niveau α.

Ces méthodes visent à ajuster les p-valeurs de l'ensemble des tests statistiques réalisés. Cet ajustement consiste à augmenter les p-valeurs de telle sorte que le critrère de contrôle des faux positifs soit vérifié.

Les approches contrôlant le FWER, comme par exemple la correction de Bonferroni (Dunn, 1961), considèrent chaque test de manière indépendante et ont pour but de prévenir tout faux positif. Ce sont des méthodes de correction très conservatives et sont donc à privilégier lorsqu'aucune erreur ne peut être tolérée. Dans le contexte d'une analyse exploratoire, où la présence de quelques faux positifs en faible nombre est acceptable, les approches moins contraignantes visant à contrôler le FDR sont préférées du fait de leur meilleure puissance statistique, *i.e.* meilleure capacité à détecter de vrais positifs. Ainsi, en génomique, la méthode de contrôle du FDR de Benjamini-Hochberg (Benjamini et Hochberg, 1995) est la méthode de correction de tests multiples la plus largement utilisée. Elle consiste en la procédure suivante :

1. ordonner les n p-valeurs de manière croissante ;
2. multiplier chaque p-valeur par un facteur défini par $\frac{n}{i}$ où i est le rang de la p-valeur dans la liste ordonnée de p-valeurs ;
3. si l'ordre initial des p-valeurs est modifié, alors pour chaque paire de p-valeur dont les rangs ont été modifiés, la p-valeur la plus élevée est abaissée à la valeur de la plus basse ;
4. ramener à 1 toute p-valeur ajustée supérieure à 1.

La correction des p-valeurs par la méthode de Benjamini-Hochberg est ainsi d'autant plus forte que le nombre total de tests statistiques réalisés est grand et que le nombre de vrais positifs est faible. Elle suppose que les tests statistiques sont indépendants ou dépendants selon une régression positive.
La méthode de Benjamini-Yekutieli (Benjamini et Yekutieli, 2001) est une amélioration de la procédure de Benjamini-Hochberg dont le but est de pouvoir prendre en compte une éventuelle relation de dépendance entre les tests statistiques réalisés. Elle se distingue de la méthode de Benjamini-Hochberg uniquement par le facteur

multiplicatif f_i apporté aux p-valeurs :

$$f_i = \frac{n \sum_{j=1}^{n} \frac{1}{j}}{i}$$

Cette méthode de correction est plus forte que la méthode de Benjamini-Hochberg, ce qui la rend plus conservative et moins puissante que cette dernière.

Le choix de la méthode de correction de tests multiples peut se faire en fonction de critères connus avant l'étude : relation de dépendance entre les tests statistiques, proportion attendue de résultats positifs, nombre total de tests (Hwang, Chu et Ou, 2011). La forme de l'histogramme des p-valeurs ajustées peut aussi indiquer si la méthode de correction des p-valeurs est trop forte ou pas assez.

5 Simulation de jeux de données RNA-seq

Les méthodes de détection de différence de dispersion entre deux groupes d'échantillons RNA-seq, MDSeq et DiPhiSeq, ont été évaluées à l'aide de jeux de données simulées. De manière générale, les données simulées, dont les principales caractéristiques sont connues, permettent d'estimer les performances d'algorithmes en mesurant leur capacité à détecter les effets pour lesquels ils ont été développés et, à l'inverse, à ne pas les détecter lorsqu'ils sont absents. Ici, les moyennes et dispersions des nombres de *reads* simulés sont connues et il s'agit de déterminer les performances de MDSeq et de DiPhiSeq pour l'identification de gènes dont l'expression présente une différence de dispersion entre deux populations d'échantillons. Bien que ces deux méthodes permettent aussi la détection de différence de moyenne, leur capacité à détecter des différences de dispersion constituera le principal objectif de cette étude de simulation.

5.1 *Packages* de simulation de nombres de *reads*

Le *package* R compcodeR (Soneson, 2014) est l'un des rares *packages* de simulation de nombres de *reads* RNA-seq à donner la possibilité d'avoir des *fold-changes* de dispersion entre les deux populations d'échantillons simulées pour un gène donné. En effet, les autres *packages* de simulation de nombres de *reads*, PoiClaClu (Witten, 2011), PROPER (Wu, Wang et Wu, 2015), ssizeRNA (Bi et Liu, 2016), qui s'inscrivent dans le cadre classique de l'analyse de différence de moyenne, imposent une même valeur de dispersion dans les deux conditions pour un gène donné et ne permettent donc pas de répondre aux besoins des simulations considérées dans cette étude. Le *package* simSeq (Benidt et Nettleton, 2015) simule des nombres de *reads* pour chaque population d'échantillons de manière non paramétrique à l'aide de valeurs de moyenne et de dispersion estimées à partir de jeux de données RNA-seq réels. Les populations d'échantillons générées avec ce *package* peuvent donc présenter des *fold-changes* de moyenne et de dispersion mais le manque de contrôle à l'aide de paramètres dans leur introduction ne permet pas de répondre aux besoins de cette étude de simulation.

En plus de permettre d'introduire des *fold-changes* de dispersion de manière paramétrique, le *package* compcodeR présente l'avantage de générer des nombres de *reads* à partir de valeurs de moyenne et de dispersion observées par paires dans deux jeux de données réelles (Cheung et al., 2010 et Pickrell et al., 2010). Cette caractéristique renforce le réalisme des données simulées en évitant de générer des nombres de *reads*

issus d'une distribution binomiale négative de manière totalement aléatoire, *e.g* en associant de manière irréaliste une valeur de moyenne et une valeur de dispersion.

Le *package* compcodeR a donc été utilisé pour simuler des nombres de *reads* réalistes pouvant présenter des différences de dispersion entre deux populations d'échantillons.

5.2 Simulation de nombres de *reads* à partir de la distribution binomiale négative

Des nombres de *reads* ont été simulés pour deux populations d'échantillons, ou conditions, à partir de la distribution binomiale négative telle que définie dans la formule 2.6. Soit Y_{ij} la variable aléatoire décrivant le nombre de *reads* pour le gène i dans l'échantillon j. On note :

$$Y_{ij} \sim \mathcal{NB}(\mu_i^k, \phi_i^k)\,,$$

où :

— k est la condition à laquelle appartient l'échantillon j, $k = \{1; 2\}$;
— μ_i^k et ϕ_i^k sont la moyenne et la dispersion du gène i dans la condition k.

Pour chaque gène i, la moyenne μ_i^1 et la dispersion ϕ_i^1 pour les échantillons de la condition 1 sont estimées par paires à partir de jeux de données réelles de Cheung et al., 2010 et de Pickrell et al., 2010. La précision avec laquelle ces valeurs de paramètres ont été estimées est ainsi cruciale pour pouvoir garantir le réalisme des valeurs d'expression simulées.

5.2.1 Estimation des valeurs de moyenne et de dispersion à partir de jeux de données réelles

Le jeu de données de Pickrell et al., 2010 est constitué de 69 échantillons issus de lignées cellulaires de lymphoblastoïdes provenant de 69 individus différents. Un sous-ensemble de 44 échantillons, dont le nombre total de *reads* associés à un transcrit est compris entre 10 et 16 millions, a été sélectionné. Les transcrits ayant en moyenne moins d'un *read* ont été retirés, aboutissant à un total de 46 446 transcrits. Les nombres de *reads* ont été ensuite normalisés de telle que sorte que la taille des librairies soient égales à la plus petite du jeu de données.

Le jeu de données de Cheung et al., 2010 est constitué de 41 échantillons issus de cellules-B immortalisées provenant d'individus différents. La taille moyenne des librairies séquencées est de 41 millions de *reads*. Aucune indication à propos du filtrage des échantillons, des transcrits et de la normalisation des nombres de *reads* n'est précisée par les auteurs du *package* compcodeR.

Pour ces deux jeux de données, les échantillons étant issus d'individus différents, les nombres de *reads* qui les composent sont supposés être « sur-dispersés », tels qu'habituellement observés dans les jeux de données RNA-seq composés de réplicats biologiques. Les estimateurs $\hat{\mu}_i$ et $\hat{\phi}_i$ de la moyenne et de la dispersion ont été déterminés par maximum de vraisemblance tel qu'indiqué dans la section 3.3.2 pour chaque transcrit i. Du fait de la grande taille des jeux de données et de l'importante profondeur de séquençage, on suppose que l'estimation indépendante de la dispersion à partir des nombres de *reads* de chaque transcrit est suffisamment fiable. Les auteurs précisent toutefois que le but n'est pas d'avoir une estimation précise de la moyenne et de la dispersion pour chaque transcrit mais d'avoir une distribution de valeurs de

ces deux estimateurs décrivant une tendance à l'échelle du transcriptome (figure 2.4).

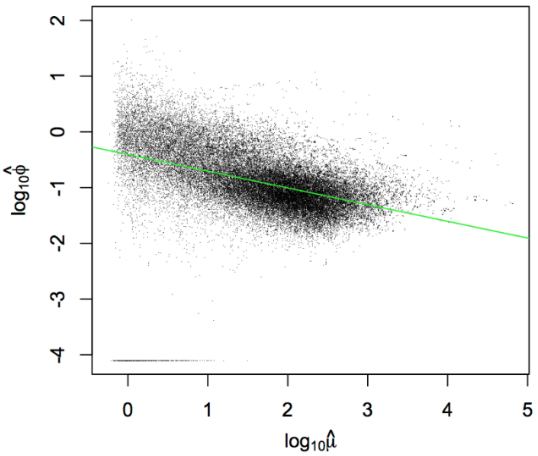

FIGURE 2.4 – Estimateurs $\hat{\mu}_i$ de la moyenne et $\hat{\phi}_i$ de la dispersion par maximum de vraisemblance à partir du jeu de données de PICKRELL et al., 2010 (figure issue de PICKRELL et al., 2010).

5.2.2 *Fold-changes* de moyenne et de dispersion

Les nombres de *reads* des échantillons appartenant à la seconde condition sont générés en fonction de ceux des échantillons de la première condition. Des facteurs, ou *fold-changes*, sont appliqués à la moyenne μ_i^1 et à la dispersion ϕ_i^1 pour générer les comptes des gènes différentiellement exprimés et/ou dispersés dans la seconde condition :

$$\mu_i^2 = \begin{cases} f_i^{\mu}\,\mu_i^1 \text{ pour les gènes différentiellement exprimés,} \\ \mu_i^1 \text{ pour les gènes non différentiellement exprimés ;} \end{cases},$$

$$\phi_i^2 = \begin{cases} f_i^{\phi}\,\phi_i^1 \text{ pour les gènes différentiellement dispersés,} \\ \phi_i^1 \text{ pour les gènes non différentiellement dispersés.} \end{cases}$$

où f_i^{μ} et f_i^{ϕ} sont les *fold-changes* appliqués à la moyenne et à la dispersion respectivement pour générer les comptes des gènes différentiellement exprimés et/ou dispersés dans la condition 2. Que ce soit pour la moyenne ou pour la dispersion, le *fold-change* f_i est généré selon la formule suivante :

$$f_i = b_i + c_i\,, \tag{2.24}$$

où :

- b_i est la valeur minimale de *fold-change* ;
- c_i est un facteur supplémentaire pour générer de plus grandes valeurs de *fold-change*, c_i est tiré aléatoirement à partir d'une distribution exponentielle de paramètre λ.

5.3 Paramètres

La simulation de nombres de *reads* de RNA-seq avec le *package* compcodeR (SONESON, 2014) peut prendre en compte les paramètres suivants :

— dimension de la matrice de nombres de *reads* ;
— profondeur de séquençage ;
— gènes différentiellement exprimés et/ou dispersés ;
— gènes avec des comptes sans sur-dispersion ;
— gènes avec des valeurs de nombres de *reads* aberrantes.

Deux paramètres permettent de contrôler la dimension de la matrice de nombres de *reads* simulés. Le premier détermine le nombre de gènes et le second le nombre d'échantillons par condition. Par défaut, compcodeR simule des nombres de *reads* pour deux populations d'échantillons de taille identique. La fonction qui simule les nombres de *reads* (la fonction *generateSyntheticData()*) a été modifiée pour permettre la génération de populations de tailles différentes.

La profondeur de séquençage, *i.e.* la somme des nombres de *reads* de chaque échantillon, est contrôlée par trois paramètres : une valeur de base, le minimum et le maximum d'une distribution uniforme utilisée pour obtenir un facteur à appliquer à la valeur de base dans le but d'introduire de la variabilité pour ce paramètre. L'objectif de cette étude de simulation n'étant pas d'évaluer les méthodes de normalisation, les échantillons ont été simulés de telle sorte qu'ils aient exactement la même profondeur de séquençage.

Les valeurs de moyenne et de dispersion de chaque gène pour les deux conditions peuvent être générées comme indiqué dans la section 5.2.

Le *package* compcodeR peut générer des nombres de *reads* ne présentant pas de « sur-dispersion » à l'aide d'une distribution de Poisson. Le but de cette étude de simulation étant d'évaluer des méthodes de détection de différence de dispersion sur des données issues de réplicats biologiques, cette option n'est pas utilisée et tous les nombres de *reads* sont générés à partir de distributions binomiales négatives.

Enfin, le *package* compcodeR donne la possibilité d'introduire des valeurs aberrantes, ou *outliers*, dans les nombres de *reads* générés. Ces valeurs aberrantes sont des valeurs anormalement très élevées ou très basses. Deux méthodes sont proposées pour introduire ces valeurs aberrantes. La méthode « *single* » sélectionne un échantillon dont le nombre de *reads* est multiplié ou divisé par un facteur compris entre 5 et 10 pour un pourcentage donné de transcrits. La méthode « *random* » consiste à considérer tous les nombres de *reads* de manière indépendante et de les multiplier ou de les diviser par un facteur compris entre 5 et 10 selon une probabilité donnée. Soneson et Delorenzi, 2013 ont montré que les *outliers* avec des comptes très bas n'avaient que très peu d'effets sur les méthodes de détection de différence de moyenne d'expression et que seules les valeurs aberrantes de nombres de *reads* très élevées impactaient leurs performances. Ainsi, seules des valeurs aberrantes anormalement élevées sont prises en compte dans cette étude de simulation.

5.4 Valeurs de paramètres testées

5.4.1 *Fold-changes* de dispersion

Les *fold-changes* de dispersion ont été introduits selon la formule 2.24. Les valeurs par défaut de b_i (1,5) et de λ (1) ont été conservées. Il existe ainsi une différence de dispersion d'au moins 50% entre les deux populations d'échantillons pour les gènes dont l'expression est différentiellement dispersée. L'opposé du *fold-change* généré a été appliqué à la moitié des gènes présentant une différence de dispersion dans le

but d'avoir autant de gènes présentant une augmentation qu'une diminution de leur dispersion d'expression dans la seconde condition.

5.4.2 *Fold-changes* de moyenne

Les *fold-changes* de moyenne ont été introduits selon deux scénarios. Le premier consiste à introduire des différences de moyenne modérées entre les deux populations d'échantillons dans le but d'évaluer les performances de détection de dispersion d'expression pour des gènes très peu à faiblement différentiellement exprimés. Pour ce faire, les *fold-changes* de moyenne ont été générés aléatoirement selon une distribution uniforme :

$$f_i^{\mu} \sim \mathcal{U}([1; f_{max}^{\mu}]) \,, \tag{2.25}$$

où f_{max}^{μ} est le *fold-change* maximum toléré.
Les valeurs f_{max}^{μ} allant de 1,1 à 1,5 par pas de 0,1 ont été évaluées.

Le second scénario consiste à introduire des *fold-changes* de moyenne selon la formule 2.24 en utilisant différentes valeurs de b_i comprises entre 1,1 et 1,5 et des valeurs de λ comprises entre 0,85 et 1, de telle sorte que les valeurs maximales de *fold-change* observées soient similaires pour chaque paire de paramètres (b_i, λ) (figure 2.5). Les

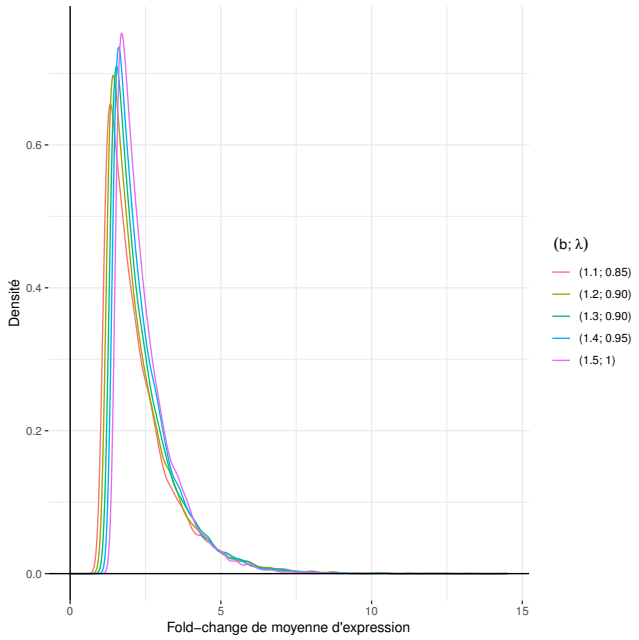

FIGURE 2.5 – Courbes de densité de 10 000 valeurs de *fold-changes* obtenues avec différentes paires de paramètres b et λ.

gènes considérés comme non différentiellement exprimés n'ont pas été générés de telle sorte qu'ils aient exactement la même moyenne d'expression dans les deux conditions, *i.e.* $f_i^{\mu} = 1$. En effet, le but principal de cette étude de simulation n'étant pas l'évaluation de la performance des méthodes pour la détection de différence de moyenne d'expression, une distinction aussi nette, et assez peu réaliste, entre les gènes différentiellement exprimés et non différentiellement exprimés n'est pas requise. C'est la raison pour laquelle des *fold-changes* de moyenne modérés ont été appliqués aux gènes considérés comme non différentiellement exprimés selon la formule 2.25 utilisée dans le premier scénario, le *fold-change* maximum toléré étant la valeur b_i utilisée pour les gènes différentiellement exprimés.

De manière similaire aux *fold-changes* de dispersion, l'opposé des *fold-changes* de moyenne a été appliqué à la moitié des gènes différentiellement exprimés.

5.4.3 Répartition des *fold-changes* de moyenne et de dispersion

Des jeux de données ont été simulés avec 50% de gènes différentiellement exprimés et 50% de gènes différentiellement dispersés. Les *fold-changes* de moyenne et de dispersion ont été introduits de telle sorte que toutes les catégories de gènes relatives à ces *fold-changes* soient également représentées. Par exemple, parmi les gènes différentiellement exprimés, la moitié a un *fold-change* de dispersion et l'autre moitié n'en a pas. La même répartition selon le *fold-change* de dispersion est respectée pour les gènes non différentiellement exprimés. De plus, la même répartition uniforme est suivie concernant le signe des *fold-changes* : parmi les gènes différentiellement exprimés ayant une augmentation de moyenne d'expression dans la seconde condition, 25% présentent une différence de dispersion positive, 25 % présentent une différence de dispersion négative et 50% ne sont pas différentiellement dispersés.

Ce seuil de 50% a été choisi pour l'avantage qu'il présente de générer de manière égale les différentes catégories de gènes relatives à leurs *fold-changes* et aux signes de ceux-ci. Ainsi, les capacités des méthodes à détecter ces *fold-changes* lorsqu'ils existent et à ne pas les détecter lorsqu'ils n'existent pas pourront être évaluées de manière égale. Cependant, dans la plupart des études d'évaluation de performances de méthodes de détection de différence de moyenne d'expression, cette valeur de seuil à 50% est jugée trop élevée. Des valeurs comprises entre 10 et 40% sont considérées comme étant plus réalistes (Landau et Liu, 2013, Bonafede et al., 2016). D'autres jeux de données ont donc été simulés avec des seuils de *fold-change* de moyenne et de dispersion égaux à 30%.

5.4.4 Taille des populations d'échantillons

Des jeux de données contenant des populations d'échantillons de taille égale ont été générés. Les tailles évaluées sont {5 ; 10 ; 20 ; 30 ; 40 ; 50 ; 100}. Des populations de tailles différentes sont aussi prises en compte dans le but de s'approcher des déséquilibres que l'on peut observer dans les tailles de population d'échantillons réels, comme c'est le cas pour les échantillons tumoraux et non tumoraux disponibles dans TCGA. Un facteur multiplicatif a été appliqué au nombre d'échantillons de la première condition pour déterminer le nombre d'échantillons dans la seconde condition. Les valeurs suivantes de facteur multiplicatif ont été évaluées : {1,5 ; 2 ; 3 ; 5 ; 10}. Ainsi, pour un nombre d'échantillons dans la première condition égal à 10, des jeux de données simulées ont été générés avec 15, 20, 30, 50 et 100 échantillons dans la seconde condition. Cet ensemble de jeux de données a été créé de telle sorte que la seule différence entre eux soit le nombre d'échantillons dans la seconde condition. Les valeurs de moyenne, de dispersion, de *fold-changes* pour la moyenne et la dispersion sont identiques pour l'ensemble des gènes de ces jeux de données. Cette modélisation permet ainsi d'évaluer l'impact de l'augmentation du nombre d'échantillons dans la seconde condition sur les performances des méthodes évaluées.

5.4.5 Présence d'*outliers*

La présence d'*outliers* est considérée comme étant plus réaliste que leur absence des nombres de *reads* simulés (Li et Lamere, 2018). Ainsi, seuls des jeux de données contenant des *outliers* ont été considérés dans cette étude de simulation. Les *outliers*

ayant des valeurs anormalement basses n'ayant pas d'impact sur les performances des méthodes de détection de différence de moyenne (Soneson et Delorenzi, 2013), ils n'ont pas été considérés non plus. Enfin, de manière analogue à l'étude de simulation menée par Bonafede et al., 2016, seule l'option « *single* » appliquée à 10% des gènes a été évaluée.

5.4.6 Réplicats

Des jeux de données ont été simulés pour chaque ensemble de paramètres en dix réplicats dans le but d'évaluer la variabilité des performances de MDSeq et de DiPhiSeq pour la détection de différence de dispersion.

5.4.7 Récapitulatif des paramètres de simulation évalués

- Taille des populations d'échantillons :
  - première condition : {5 ; 10 ; 20 ; 30 ; 40 ; 50 ; 100} ;
  - facteur multiplicatif pour la deuxième condition : {1 ; 1,5 ; 2 ; 3 ; 5 ; 10} ;
- nombre de gènes : 10 000 ;
- profondeur de séquençage : {5} ;
- gènes différentiellement exprimés :
  - pourcentage : {30 ; 50} ;
  - seuil de *fold-change* : {1,1 ; 1,2 ; 1,3 ; 1,4 ; 1,5} ;
  - paramètre de la loi exponentielle : {0,85 ; 0,9 ; 0,9 ; 0,95 ; 1} en correspondance avec le seuil de *fold-change* ;
  - scénarios : *fold-changes* modérés et réalistes ;
- gènes différentiellement dispersés :
  - pourcentage : {30 ; 50} ;
  - seuil de *fold-change* : {1,5} ;
  - paramètre de la loi exponentielle : 1 ;
- présence d'outliers :
  - option de compcodeR : « *single* » ;
  - pourcentage de gènes avec outliers : 10.

5.5 Evaluation de performance

5.5.1 Tables de contingence

Simuler des jeux de données et appliquer des méthodes sur ces jeux permet d'évaluer les performances de ces dernières. En effet, connaissant les conditions dans lesquelles ces jeux de données ont été simulés, le résultat attendu des méthodes évaluées est connu. On peut donc évaluer leurs performances en confrontant les résultats qu'elles prédisent aux résultats attendus. Dans le cas où les résultats peuvent se ranger en deux classes distinctes : résultats positifs et résultats négatifs, les classes d'une table de contingence peuvent être définies (table 2.3). Un résultat attendu positif est un vrai positif (TP pour *True Positive*) s'il est correctement prédit comme étant un résultat positif ou, à l'inverse, un faux négatif (FN pour *False Negative*) s'il est prédit comme étant un résultat négatif. De même, un résultat négatif est un vrai négatif (TN pour *True Negative*) s'il est correctement prédit ou un faux positif (FP pour *False Positive*) s'il ne l'est pas. Les FP constituent donc les erreurs de première espèce alors que les FN constituent les erreurs de deuxième espèce.
Dans le cadre de cette étude de simulation, les résultats attendus positifs pour la moyenne sont les gènes différentiellement exprimés et les résultats attendus négatifs

		Résultats attendus	
		P	N
Résultats prédits	P	TP	FP
	N	FN	TN

TABLE 2.3 – Table de contingence des classes de résultats pour une méthode classant les résultats en deux catégories : résultats positifs (P) et résultats négatifs (N). Les résultats se rangent en quatre catégories : vrai positif (TP), faux positif (FP), faux négatif (FN), vrai négatif (TN).

sont les gènes non différentiellement exprimés. De même, concernant la dispersion, les résultats attendus positifs sont les gènes dont les nombres de *reads* ont été simulés de telle sorte qu'ils présentent une différence de dispersion entre les deux populations d'échantillons. Les résultats attendus négatifs sont les gènes ayant la même valeur de dispersion pour les deux populations d'échantillons.
Les classes prédites sont définies par la p-valeur des tests de différence de moyenne et de dispersion. Après correction de tests multiples, les résultats prédits positifs sont ainsi les gènes ayant une p-valeur inférieure à 0,05 et les résultats prédits négatifs sont ceux ayant une p-valeur supérieure à ce seuil.

5.5.2 Indicateurs de performance

A partir de la table de contingence définie précédemment, différents indicateurs peuvent être calculés, captant chacun une caractéristique de la performance des méthodes évaluées.

$$TPR = \frac{TP}{TP + FN}$$

La sensibilité ou taux de vrais positifs (TPR pour *True Positive Rate*) mesure la capacité d'un classifieur à détecter un résultat positif lorsqu'il existe. Une méthode est dite sensible lorsqu'elle détecte comme positif la plupart des résultats positifs attendus.

$$TNR = \frac{TN}{TN + FP}$$

La spécificité ou taux de vrais négatifs (TNR pour *True Negative Rate*) mesure la capacité d'une méthode à ne pas considérer comme positifs des résultats qui ne le sont effectivement pas. Une méthode est dite spécifique lorsqu'elle détecte négatif la plupart des résultats qui le sont vraiment, ce qui par conséquent signifie qu'elle identifie très peu de résultats positifs parmi les résultats négatifs.

$$FDR = \frac{FP}{FP + TP}$$

Une mesure très couramment utilisée est le taux de fausses découvertes (FDR pour *False Discovery Rate*). Elle mesure la proportion de faux positifs parmi l'ensemble des résultats détectés comme positifs par une méthode. La précision est une mesure semblable et est la proportion de vrais positifs parmi l'ensemble des résultats détectés comme positifs par une méthode (la précision est égale à $1 - FDR$). On dit qu'une méthode est précise lorsque la part des faux positifs parmi les résultats positifs qu'elle prédit est faible.

5.5.3 Courbes ROC et aire sous la courbe

Une mesure couramment utilisée pour évaluer la performance d'une méthode est l'aire sous la courbe ROC (*Receiver Operating Characteristic*). Celle-ci représente la sensibilité en fonction de la spécificité en utilisant une gamme de valeurs de seuil séparant les résultats positifs des négatifs. Un mauvais algorithme prédictif a une aire sous la courbe ROC (AUC pour *Area Under the Curve*) proche de 0,5. Cette valeur signifie que la méthode n'est pas meilleure que le hasard pour identifier les deux classes. Au contraire, un excellent algorithme prédictif a une AUC proche de 1, cette valeur signifiant que la méthode ne se trompe jamais.

5.6 Réalisme des jeux de données simulées

La qualité de l'évaluation des performances de méthodes à partir de données simulées dépend essentiellement du réalisme avec lequel les données simulées ont été générées. Dans cette étude, les données simulées ont été générées à l'aide du *package* R compcodeR (Soneson, 2014) qui utilisent les valeurs de moyenne et de dispersion estimées par paires à partir de jeux de données RNA-seq réels pour générer les nombres de *reads* pour l'une des deux populations d'échantillons considérés. Le grand nombre d'échantillons et l'importante profondeur de séquençage de ces deux jeux de données réels (Pickrell et al., 2010 et Cheung et al., 2010) garantissent une bonne estimation de ces paramètres. De plus, l'application des valeurs estimées de moyenne et de dispersion par paire permet de respecter la tendance généralement observée entre ces deux paramètres dans les jeux de données RNA-seq, renforçant ainsi le réalisme des données simulées.
Les nombres de *reads* de la seconde population d'échantillons sont obtenus en appliquant éventuellement des *fold-changes* à la moyenne et/ou à la dispersion. L'application de ces *fold-changes* peut aboutir à l'association peu réaliste de valeurs de moyenne et de dispersion. En effet, l'application d'une valeur importante de *fold-change* à une valeur de moyenne ou de dispersion déjà élevée peut aboutir à une paire de valeurs de moyenne et de dispersion s'écartant nettement de la tendance réaliste. Les valeurs extrêmes étant assez peu nombreuses parmi les *fold-changes* introduits (figure 2.5) et le nombre de gènes pour lesquels des nombres de *reads* sont générés étant important (10 000 par jeu de données), le nombre de paires de valeurs de moyenne et de dispersion irréalistes utilisées pour générer les nombres de *reads* pour la seconde population d'échantillon est supposé être assez faible relativement au nombre total de gènes simulés par jeu de données.

6 Enrichissement de termes *Gene Ontology*

L'analyse classique de différence de moyenne d'expression à partir de données RNA-seq aboutit à l'obtention de listes de gènes différentiellement exprimés. Ces listes en tant que telles sont difficilement interprétables. Une analyse couramment réalisée consiste à identifier des fonctions biologiques dans lesquelles les gènes différentiellement exprimés sont sur-représentés, permettant ainsi une interprétation biologique des différences observées entre les conditions d'intérêt. De manière générale, cette approche peut s'appliquer à toute liste de gènes d'intérêt. Dans le cadre de l'analyse de différence d'expression, il s'agit de l'ensemble des gènes différentiellement exprimés. Ici, cette approche est appliquée aux gènes différentiellement dispersés.
Pour pouvoir mener ce type d'analyse, une base de connaissances reliant les gènes à des fonctions biologiques est nécessaire. *Gene Ontology* (Ashburner et al., 2000,

The Gene Ontology Consortium, 2017) est l'une d'entre elles et celle utilisée dans le cadre de ces travaux de thèse.

6.1 *Gene Ontology*

Une ontologie est la représentation de connaissances issues d'un domaine sous la forme de concepts et de relations les reliant dans une organisation hiérarchique. L'objectif de *Gene Ontology* est de fournir une base de connaissances sur la fonction des gènes dans différents organismes. Ces connaissances sont réparties en trois domaines : processus biologiques, fonctions moléculaires et localisations cellulaires. Pour chacun de ces domaines, *Gene Ontology* fournit une ontologie (figure 2.6) et un ensemble d'annotations reliant les gènes à des termes de l'ontologie. L'ontologie est

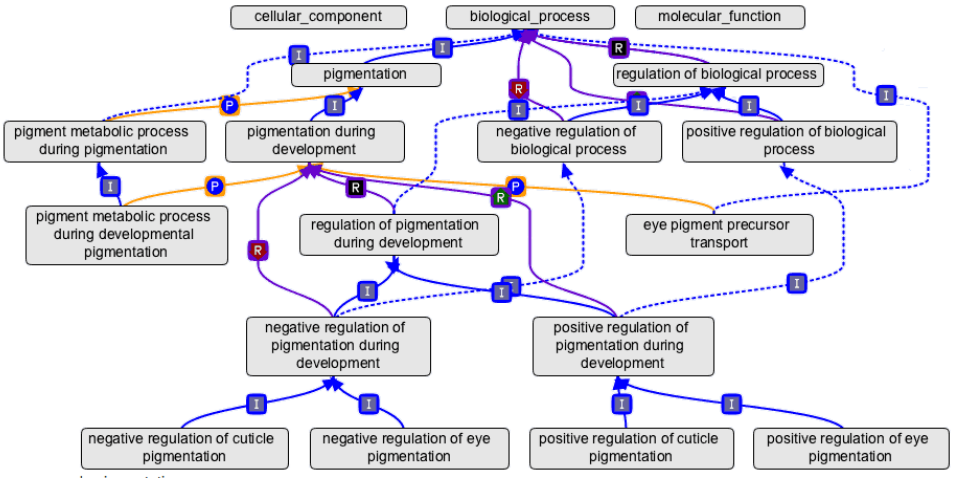

Figure 2.6 – Exemples de concepts et de relations dans le domaine des processus biologiques de *Gene Ontology* (`http://geneontology.org/page/ontology-structure` Ashburner et al., 2000 The Gene Ontology Consortium, 2017).

un ensemble de classes, ou termes, reliées les unes aux autres par différents types de relations. Elle peut être représentée sous la forme d'un graphe direct acyclique. La principale relation est la relation « *is_a* » qui marque un lien hiérarchique entre un terme générique et un terme plus spécifique, *e.g.* « pigmentation durant le développement » est plus spécifique que « pigmentation » sur la figure 2.6. Les autres types de relation sont les relations de partition (« *part_of* ») et de régulation (« regulates »). Les annotations, reliant les gènes à des classes de l'ontologie, sont basées sur des preuves expérimentales issues de plus de 150 000 articles.

6.2 Enrichissement de termes *Gene Ontology*

L'enrichissement de termes *Gene Ontology* , ou termes GO, consiste à identifier les termes GO sur-représentés dans une liste de gènes d'intérêt. Cette analyse peut être représentée sous la forme d'un diagramme de Venn (figure 2.7) où l'enrichissement d'un terme GO est représenté par le recouvrement du cercle représentant des gènes d'intérêt (cercle bleu « *in list* ») et l'ensemble des gènes dans *Gene Ontology* annoté avec ce terme ontologique (cercle vert « *with annotation* »). La significativité de l'enrichissement d'un terme GO dépend ainsi du nombre total de gènes d'intérêt, du nombre total de gènes annotés avec ce terme dans *Gene Ontology*, le nombre de gènes annotés avec le terme GO parmi les gènes d'intérêt et l'ensemble des gènes présents dans le jeu de données (cercle gris « *tested* »).

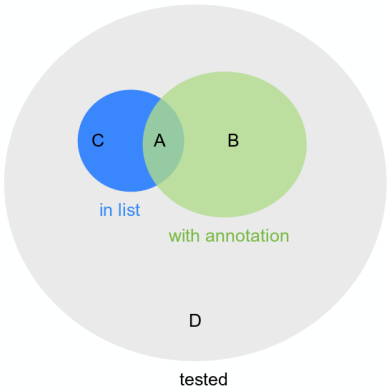

FIGURE 2.7 – Représentation schématique du test d'enrichissement d'un terme *Gene Ontology* représenté par le cercle vert (« *with annotation* ») parmi un ensemble de gènes d'intérêt représenté par le cercle bleu (« *in list* »). Les gènes d'intérêt ont été identifiés parmi un ensemble de gènes testés représenté par le cercle gris (« *tested* »).

La fonction *enrichGO()*, qui implémente un test hypergéométrique, du *package* R clusterProfiler (YU et al., 2012) est utilisée pour réaliser les tests d'enrichissement de termes *Gene Ontology*.

6.3 Réduction de redondance

Les listes de termes GO enrichis peuvent être très longues dans ce type d'approche, en particulier lorsque l'ensemble des gènes considéré est important. De plus, du fait de la structure hiérarchique et inclusive des termes au sein d'une ontologie, les listes de termes GO significatifs sont aussi souvent redondantes. Pour simplifier leur interprétation, les termes GO similaires peuvent être rassemblés à l'aide de mesures de similarité sémantique.

6.3.1 Mesures de similarité sémantique

Ces mesures font appel à la notion de contenu d'information mesurable par la fréquence d'occurrence d'un terme au sein d'un corpus :

$$p(t) = \frac{n_{t'}}{N} | t' \,,$$

où :

- t' est un terme GO appartenant à l'ensemble défini par le terme GO t et ses fils ;
- $n_{t'}$ est le nombre d'occurrences du terme t' au sein d'un corpus de N annotations.

Le contenu d'information est alors défini par :

$$IC(t) = -\log(p(t)) \,.$$

Ainsi, un terme GO est d'autant plus informatif qu'il est peu utilisé.
La similarité sémantique de deux terms GO t_1 et t_2 peut alors être calculée à partir du contenu informatif de leur plus proche ancêtre commun (MICA pour *Most Informative Common Ancestor*). RESNIK, 1999 propose d'utiliser l'information contenue par MICA :

$$\mathrm{sim}_{Resnik}(t_1, t_2) = IC(MICA) \,.$$

Lin, 1998 normalise l'information contenue par MICA par celles des deux termes GO respectivement :

$$\mathrm{sim}_{Lin}(t_1, t_2) = \frac{2\,IC(MICA)}{IC(t_1) + IC(t_2)}\,.$$

La méthode *Relevance* (Schlicker et al., 2006) combine les méthodes de Resnik et de Lin :

$$sim_{Rel}(t_1, t_2) = \frac{2\,IC(MICA)(1 - p(MICA))}{IC(t_1) + IC(t_2)}\,.$$

Le *package* R GoSemSim a été utilisé pour calculer ces mesures de similarité (Yu et al., 2010).

6.3.2 Simplification de listes de termes GO

Simplification d'une liste de termes GO Les listes de termes GO enrichis ont été simplifiées en rassemblant les termes GO similaires. Pour ce faire, les valeurs de similarité selon la méthode *Relevance* proposée par Schlicker et al., 2006 ont été calculées pour chaque paire de termes GO. Les termes GO ont ensuite été regroupés par *clustering* hiérarchique en utilisant ces valeurs de similarité comme mesure de distance et des groupes de termes GO similaires ont été formés selon une valeur de seuil. Pour chaque groupe de termes GO similaires ainsi formés, le terme GO avec la p-valeur la plus faible a été retenu comme terme représentant du groupe. Les termes GO du groupe de similarité sont alors remplacés par l'unique terme GO représentant du groupe.

Comparaison de plusieurs listes de termes GO Dans le but de faciliter la comparaison de plusieurs listes de termes GO, le même procédé de simplification par regroupement de termes GO similaires a été appliqué aux listes à comparer. Les termes GO issus de l'ensemble des listes à comparer ont été rassemblés par *clustering* hiérarchique basé sur la mesure de similarité *Relevance* et des groupes ont été formés en fonction de seuils. Les termes GO présents au sein d'un groupe de similarité pouvant ne pas être présents dans toutes les listes comparées, le terme GO ancêtre le plus proche à l'ensemble des termes du groupe a été sélectionné pour représenter ce groupe de termes GO similaires, à la condition qu'il figure parmi les listes de termes GO enrichis. Dans le cas contraire, aucun terme GO n'est utilisé comme représentant du groupe et les termes GO du groupe de similarité ne sont pas remplacés.

7 Association entre expression de miARN et d'ARNm

La plupart des études visant à associer l'expression de miARN à celles de leurs cibles sont basées sur des liens de corrélation (Bossel Ben-Moshe et al., 2012). Ces approches supposent une répression induite par les miARN sur l'expression de leurs ARNm cibles, se matiéralisant par des anti-corrélations entre les profils d'expression. Ces méthodes s'inscrivent parfaitement dans le cadre de l'analyse classique de différence de moyenne d'expression entre deux conditions d'intérêt, les liens d'interaction étant trouvés entre miARN et ARNm dont les expressions différentielles évoluent de manière opposée, *e.g.* entre un miARN sur-exprimé et un ARNm sous-exprimé. En revanche, ces approches d'anti-corrélation entre expressions de miARN et d'ARNm ne sont pas adaptées à l'approche menée dans le cadre de cette thèse où l'intérêt ne se concentre pas principalement sur l'identification de miARN et d'ARNm se caractérisant par des différences de moyenne d'expression entre deux conditions d'intérêt.

Les approches suivies par Iterson et al., 2013 et Rauschenberger et al., 2016 proposent d'identifier une association entre l'expression d'ARNm et un ensemble de covariables à l'aide d'un test global. Iterson et al., 2013 se sont intéressés à identifier des associations entre miARN et ARNm à partir de données de puces à ADN dans le but d'améliorer les prédictions d'interaction miARN-ARNm. Rauschenberger et al., 2016 ont plus tard étendu leur approche aux données RNA-seq et à la prise en compte de différents types de données génomiques comme covariables tels que les SNPs (*Single Nucleotide Polymorphisms*) ou les CNVs (*Copy-Number Variations*). Dans le cadre de cette thèse, les fonctions développées par Rauschenberger et al., 2016 ont été utilisées en suivant l'approche menée par Iterson et al., 2013.
Ces méthodes présentent l'avantage de pouvoir considérer l'action collective des miARN ciblant un ARNm plutôt que de limiter les interactions à une paire miARN-ARNm tel que c'est le cas avec les méthodes d'anti-corrélation. Cette caractéristique permet ainsi de mieux capter la réalité biologique de l'action conjointe de plusieurs miARN sur un même ARNm.

7.1 Test global d'association

L'association entre miARN et ARNm est estimée à l'aide d'un GLM tel que défini dans la section 3.2.2 :

$$\mathbb{E}(Y_i) = h^{-1}\left(\alpha_j + \sum_{k=1}^{p} X_{jk}\beta_{ik}\right),$$

où :

- la variable réponse Y_i est l'expression d'un ARNm i à travers l'ensemble des n échantillons du jeu de données et y_{ij} sa valeur d'expression dans l'échantillon j ;
- h^{-1} est l'inverse de la fonction de lien du GLM ;
- X est la matrice de covariables ;
- β_k sont les coefficients de régression du GLM ;
- α_j est une valeur d'intercept pour l'échantillon j.

L'approche consiste à estimer l'association entre l'ensemble des covariables et se traduit par les hypothèses de test suivantes :

$$H_0 : \beta_{i1} = \ldots = \beta_{ip} = 0\,,\; H_a : \beta_{i1} \neq 0 \cup \ldots \cup \beta_{ip} \neq 0\,.$$

Cette situation peut être difficile à résoudre si le nombre de covariables est important. En particulier, les méthodes classiques de résolution de GLM ne peuvent être appliquées si le nombre de covariables est supérieur au nombre d'échantillons, *i.e.* $p > n$. Pour pouvoir tout de même estimer la significativité de l'ensemble des paramètres de régression, Goeman et al., 2004 proposent de considérer que ces paramètres sont aléatoires avec une espérance nulle, $\mathbb{E}(\beta_i) = 0$ et une matrice de variance-covariance $\mathrm{Var}(\beta_i) = \tau_i^2 I$ où I est la matrice d'identité de taille $p \times p$ et $\tau_i^2 > 0$. Un modèle à effets aléatoires est alors défini :

$$\begin{aligned} \mathbb{E}(y_{ij}|r_{ij}) &= h^{-1}(\alpha_j + r_{ij})\,, \\ r_{ij} &= \textstyle\sum_{k=1}^{p} X_{jk}\beta_{ik}\,. \end{aligned} \tag{2.26}$$

L'hypothèse de non-association des covariables à la variable réponse peut alors s'écrire :

$$H_0 : \tau_i^2 = 0\,.$$

Un test de score est alors appliqué avec l'hypothèse alternative $H_a : \tau^2 > 0$.
Le test global a initialement été développé par Goeman et al., 2004 pour les distributions de la famille exponentielle. Ici, la variable réponse y_i est représentée par une distribution binomiale négative de paramètres μ_{ij} et ϕ_i. La distribution binomiale négative ne faisant pas partie de cette famille de distributions lorsque le paramètre de dispersion est inconnu (voir section 3.2.2), Rauschenberger et al., 2016 ont étendu l'application de ce test à ce type de variable aléatoire. Le test de score est alors basé sur la dérivée partielle de la log-vraisemblance marginale par rapport τ_i^2. Conceptuellement, toute modification de la vraisemblance marginale à un changement de τ_i^2 proche de 0 est une indication pour le rejet de l'hypothèse nulle. La fonction de score est calculée à l'aide des résultats de Le Cessie et Houwelingen, 1995 et la statistique de test est définie :

$$u = \sum_{j=1}^{n}\sum_{l=1}^{n} \frac{R_{jl}}{2} \frac{(y_{ij} - \hat{\mu}_{ij})(y_{il} - \hat{\mu}_{il})}{(1 + \hat{\phi}_i\hat{\mu}_{il})(1 + \hat{\phi}_i\hat{\mu}_{ij})} - \sum_{i=1}^{n} \frac{R_{jj}}{2} \frac{(\hat{\mu}_{ij} + y_{ij}\hat{\phi}_i\hat{\mu}_{ij})}{(1 + \hat{\phi}_i\hat{\mu}_{ij})^2},$$

où :

- R_{jl} est la valeur à la j-ème ligne et l-ème colonne de la matrice $n \times n$ $R = (1/p)XX^T$;
- $\hat{\mu}_{ij}$ et $\hat{\phi}_i$ sont les estimateurs du maximum de vraisemblance de la moyenne et de la dispersion de y_{ij} respectivement sous l'hypothèse nulle.

La distribution de la statistique de test u sous l'hypothèse nulle est inconnue. La p-valeur du test d'association est alors calculée par permutation des valeurs de la variable réponse $y = (y_1, \ldots, y_n)^T$ et de la moyenne estimée $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)^T$. Si la statistique du test est supérieure à la majorité de celles obtenues après permutations, alors l'hypothèse nulle peut être rejetée.

De plus, la statistique de test peut être décomposée en la contribution indépendante de chacune des covariables. La statistique de test est la moyenne des contributions des différentes covariables. Cette décomposition permet ainsi de déterminer quelles covariables contribuent le plus à l'association avec la variable réponse observée à l'échelle de l'ensemble des covariables.

7.2 Prédiction de paires miARN-ARNm

Dans le but de limiter le nombre de covariables utilisées dans le GLM pour chaque ARNm, les miARN considérés ont été limités à ceux identifiés comme interagissant avec l'ARNm en question selon différentes bases de données. Les bases de données d'interaction miARN-ARNm utilisées sont les bases de prédictions *in silico* TargetScan, PITA et microCosm et la base d'interactions validées expérimentalement miRTarBase (voir section 1.2.3 du chapitre 1).

Les bases de prédictions *in silico* ont été choisies pour leur complémentarité. Ces bases de données sont connues pour contenir un grand nombre de faux positifs (Pinzon et al., 2017). Une pratique courante consiste alors de ne retenir une interaction que si elle est prédite par plusieurs bases de prédictions *in silico* différentes. Le nombre de prédictions différentes requis pour conserver une interaction miARN-ARNm dépendra du nombre de bases de prédictions considérées et du niveau de confiance souhaité. Le *package* R miRNAmRNA (Iterson et al., 2013) a été utilisé pour identifier les interactions miARN-ARNm concordantes entres différentes bases de prédictions. Ce

package ne se base que sur les bases de prédictions *in silico* TargetScan, PITA et microCosm pour déterminer des interactions miARN-ARNm. Pour donner plus de poids à ces prédictions, la base d'interactions validées expérimentalement miRTarBase a été ajoutée à l'ensemble des bases de prédictions prises en compte. Pour ce faire, certaines fonctions du *package* miRNAmRNA ont été modifiées et des nouvelles fonctions ont été ajoutées. Parmi ces dernières, une option permet de prendre en compte miRTarBase dans son intégralité ou uniquement les interactions qui ne sont pas considérées comme faibles par cette base de données (voir section 1.2.3 du chapitre 1).

Chapitre 3

Identification de gènes différentiellement variants

Les méthodes exposées dans la section 2 du chapitre 2 ont été appliquées aux données d'expression d'ARNm et de miARN dans le but d'identifier des gènes ayant une différence de variance d'expression entre échantillons sains et échantillons tumoraux. Les jeux de données analysés sont ceux pour lesquels au moins 10 échantillons sains sont disponibles (table 3.1). Le nombre d'échantillons sains disponibles pour les

Jeux de données	**miARN**		**ARNm**	
	Echantillons sains	Echantillons tumoraux	Echantillons sains	Echantillons tumoraux
TCGA-BLCA	19	417	19	414
TCGA-BRCA	104	1096	113	1102
TCGA-COAD	8	455	41	478
TCGA-ESCA	13	186	11	161
TCGA-HNSC	44	523	44	500
TCGA-KICH	25	66	24	65
TCGA-KIRC	71	544	72	538
TCGA-KIRP	34	291	32	288
TCGA-LIHC	50	372	50	371
TCGA-LUAD	46	519	59	533
TCGA-LUSC	45	478	49	502
TCGA-PRAD	52	498	52	498
TCGA-READ	3	161	10	166
TCGA-STAD	45	446	32	375
TCGA-THCA	59	506	58	502
TCGA-UCEC	33	545	35	551

TABLE 3.1 – Nombres d'échantillons sains et tumoraux des jeux de données d'expression de miARN et d'ARNm du TCGA analysés dans le cadre de ce chapitre 3. BLCA : *BLadder urothelial CArcinoma*, BRCA : *BReast invasive CArcinoma*, COAD : *COlon ADenocarcinoma*, ESCA : *ESophageal CArcinoma*, HNSC : *Head and Neck Squamous cell Carcinoma*, KIRC : *KIdney Renal Clear cell carcinoma*, KIRP : *KIdney Renal Papillary cell carcinoma*, LIHC : *LIver Hepatocellular Carcinoma*, LUAD : *LUng ADenocarcinoma*, LUSC : *LUng Squamous cell Carcinoma*, PRAD : *PRostate ADenocarcinoma*, READ : *REctum ADenocarcinoma*, STAD : *STomach ADenocarcinoma*, THCA : *THyroid CArcinoma*, UCEC : *Uterine Corpus Endometrial Carcinoma*. Les jeux de données d'expression de miARN pour les cancers du colon (COAD) et du rectum (READ) ne sont pas pris en compte à cause du trop faible nombre d'échantillons sains.

autres jeux de données sont considérés comme étant trop petit pour permettre une estimation de différence de variance d'expression fiable.

1 Résultats

1.1 Sélection de test statistique

Les p-valeurs obtenues par les tests de différence de variabilité ont été comparées en vue de sélectionner le test statistique le plus approprié aux données analysées. En particulier, la sensibilité à la présence de valeurs extrêmes et au déséquilibre de taille des populations d'échantillons comparées ont été évaluées. Pour ce faire, les jeux de données d'expression d'ARNm et de miARN des cancers du sein (TCGA-BRCA), du poumon (TCGA-LUAD) et de la prostate (TCGA-PRAD) ont été analysés. Les résultats obtenus pour tous ces jeux de données aboutissent aux mêmes conclusions. Ainsi, seuls les résultats concernant le cancer de la prostate sont exposés dans cette section.
Les valeurs normalisées d'expression d'ARNm et de miARN selon la méthode TMM (voir section 1.2 du chapitre 2) ont été transformées en $\log_2$. Les ARNm et les miARN dont l'expression moyenne est supérieure ou égale à 2 parmi les échantillons tumoraux et sains respectivement ont été conservés.

1.1.1 Comparaisons des tests entre eux

Les différents tests statistiques de comparaison de variabilité (voir section 2 du chapitre 2) ont été comparés les uns avec les autres. Pour cela, les p-valeurs de chaque test ont été obtenues et ordonnées. Les rangs des p-valeurs ont ensuite été comparés pour chaque paire de tests statistiques. La figure 3.1 représente les nuages des points et les corrélations de Pearson des rangs de p-valeur pour chaque comparaison de tests statistiques. Les corrélations de Pearson révèlent des groupes de tests statistiques proches dans la classification des miARN différentiellement variables entre les échantillons tumoraux et les échantillons sains. Comme attendu à cause de leurs statistiques de test très proches, les tests de Levene et de Brown-Forsythe engendrent des rangs de p-valeur très corrélés (corrélation de Pearson de 0,99). Le test de Bartlett et le test de Fisher, tous deux très sensibles aux données qui ne suivent pas la loi normale, ordonnent les p-valeurs de manière similaire (corrélation de Pearson de 1). De manière générale, les tests utilisant la même mesure de variabilité (variance, coefficient de variation) ordonnent les p-valeurs de manière corrélée.
Enfin, les tests de Fligner-Killeen et de différence d'entropie de Shannon génèrent des résultats discordants par rapport à l'ensemble des autres tests statistiques. En particulier, pour l'ensemble des ARN considérés, l'entropie de Shannon est très proche de 1. Les différences d'entropie de Shannon entre échantillons tumoraux et sains sont ainsi systématiquement très proches de 0, ce qui donne peu de confiance dans les p-valeurs obtenues par permutations et explique pourquoi ce test est si différent de l'ensemble des autres tests statistiques. En plus des temps de calculs beaucoup plus longs que pour les autres tests statistiques, ce test basé sur l'entropie de Shannon ne semble donc pas adapté aux données comparées dans cette étude.

1.1.2 Normalité des données

Les écarts à la distribution normale des données comparés dans le cadre des tests statistiques évalués peuvent fausser les résultats obtenus (voir section 2.4 du chapitre 2). La figure 3.2 représente les p-valeurs obtenues par le test de Shapiro-Wilk pour les données normalisées d'expression des miARN pour le cancer de la prostate. Sur les 217 miARN analysés, 165 ont une p-valeur du test de Shapiro-Wilk inférieure à 0,05 pour les échantillons tumoraux, les échantillons sains ou les deux groupes d'échantillons.

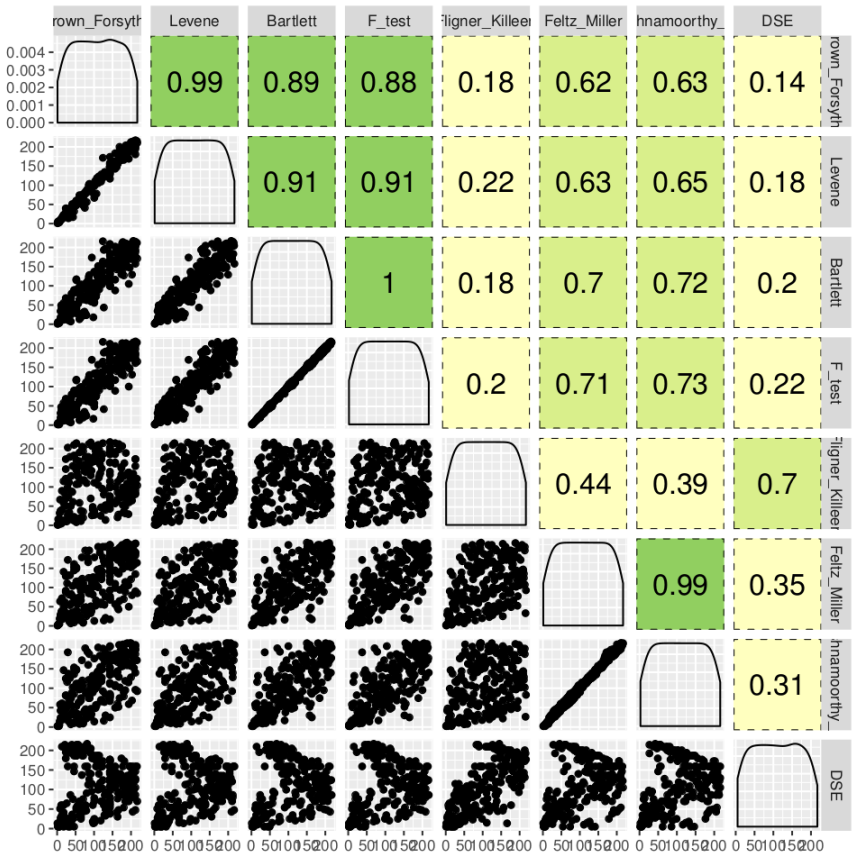

FIGURE 3.1 – Nuage de points des rangs de p-valeur obtenues pour chaque paire de tests statistiques (partie inférieure de la matrice) pour la comparaison des valeurs normalisées d'expression de miARN des échantillons tumoraux et des échantillons sains pour le cancer de la prostate. Corrélation de Pearson des rangs de p-valeur pour chaque paire de tests statistiques (partie supérieure de la matrice).

Les données normalisées d'expression de miARN, ainsi que celles d'ARNm (données non représentées) s'écartent donc de la distribution normale dans la majorité des cas. Les tests statistiques de comparaison de variance peu sensibles aux données ne suivant pas la distribution normale sont donc conseillés pour comparer la variabilité d'expression de miARN et d'ARNm entre échantillons tumoraux et échantillons sains.

La figure 3.3 illustre la sensibilité du test de Fisher aux données dont la distribution ne suit pas la loi normale. Le nuage de points des rangs des p-valeurs obtenues avec les tests de Levene et de Fisher pour les données d'expression normalisées de miARN pour le cancer de la prostate est représenté en figure 3.3.A. Une attention particulière est apportée aux miARN dont les valeurs d'expression ne suivent pas la loi normale pour au moins l'un des deux groupes d'échantillons considérés (points verts sur la figure 3.3.A). L'expression de certains de ces miARN est illustrée en figure 3.3.B. On observe que certains de ces miARN ont des p-valeurs significatives selon le test de Fisher et pas (ou beaucoup moins significatives) avec le test de Levene, se traduisant par des rangs particulièrement faibles pour le test de Fisher, alors qu'ils ne sont pas différentiellement variables entre les échantillons tumoraux et les échantillons sains. Par exemple, hsa-mir-1-1 et hsa-mir-505, dont les *boxplots* des valeurs d'expression ne se différencient que par la présence de quelques valeurs extrêmes, sont bien classés (respectivement 25ème et 9ème) selon le test de Fisher alors qu'ils ne le

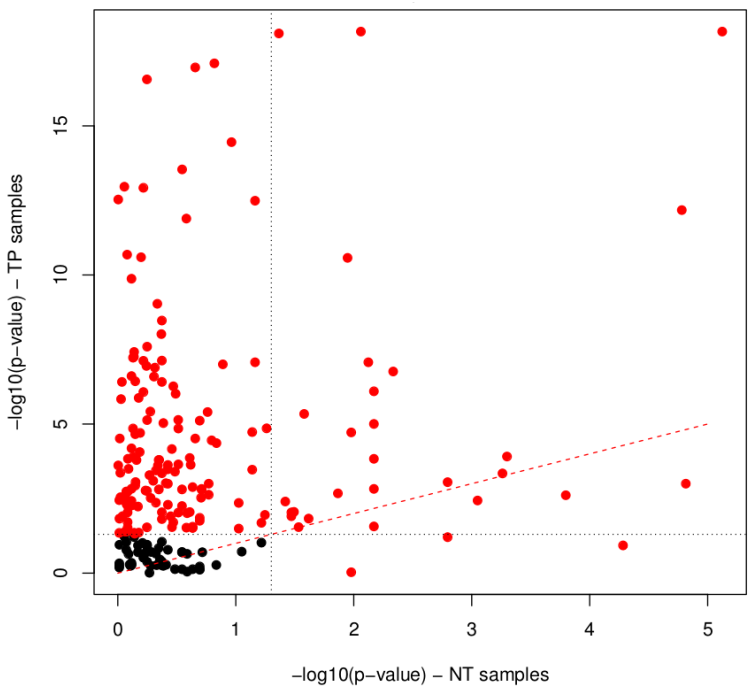

FIGURE 3.2 – P-valeurs obtenues avec le test de Shapiro-Wilk après ajustement par la procédure de Benjamini-Hochberg pour les données d'expression de miARN des échantillons tumoraux et sains du cancer de la prostate. Les valeurs normalisées d'expression ont été transformées en $\log_2$. Seuls les 217 miARN dont l'expression moyenne est supérieure ou égale à 2 parmi les deux groupes d'échantillons ont été conservés. Les valeurs $-\log_{10}$ des p-valeurs sont représentées. Les seuils à 0,05 pour les deux groupes et la droite d'identité ont été tracés en pointillés. Les points rouges représentent des miARN dont la p-valeur est inférieure à 0,05 pour au moins l'un des deux groupes d'échantillons. TP : *Tumor Primary*, NT : *Non Tumoral*.

sont pas avec le test de Levene. Cette conclusion erronée du test de Fisher est due à la présence de valeurs extrêmes parmi les échantillons tumoraux (pour hsa-mir-1-1) et parmi les échantillons sains (pour hsa-mir-505), qui écartent la distribution des valeurs d'expression de la loi normale.
A l'inverse, certains miARN (comme hsa-mir-486-1 et hsa-mir-181b-2), dont les *boxplots* des valeurs d'expression ont des écarts interquartiles différents, ont des rangs de p-valeur plus élevés selon le test de Fisher (42ème et 31ème respectivement) que selon le test de Levene (14ème et 9ème). Ceci s'explique par la présence de valeurs extrêmes parmi le groupe d'échantillons de plus faible variabilité (les échantillons tumoraux pour ces deux miARN).

Il apparaît donc que la sensibilité d'un test statistique à des données dont la distribution ne suit pas la loi normale doit être prise en compte dans le choix du ou des tests les plus adaptés pour la détermination d'ARN différentiellement variables.

1.1.3 Populations de tailles différentes

Les échantillons tumoraux sont beaucoup plus nombreux que les échantillons sains parmi les données fournies par le GDC. Pour la majorité des jeux de données, il y a environ dix fois plus d'échantillons tumoraux que d'échantillons sains (figure 1.11 et table 3.1). Cette différence de taille d'effectif entre les deux groupes considérés dans les tests de comparaison de variabilité peut influencer leurs conclusions. Pour évaluer

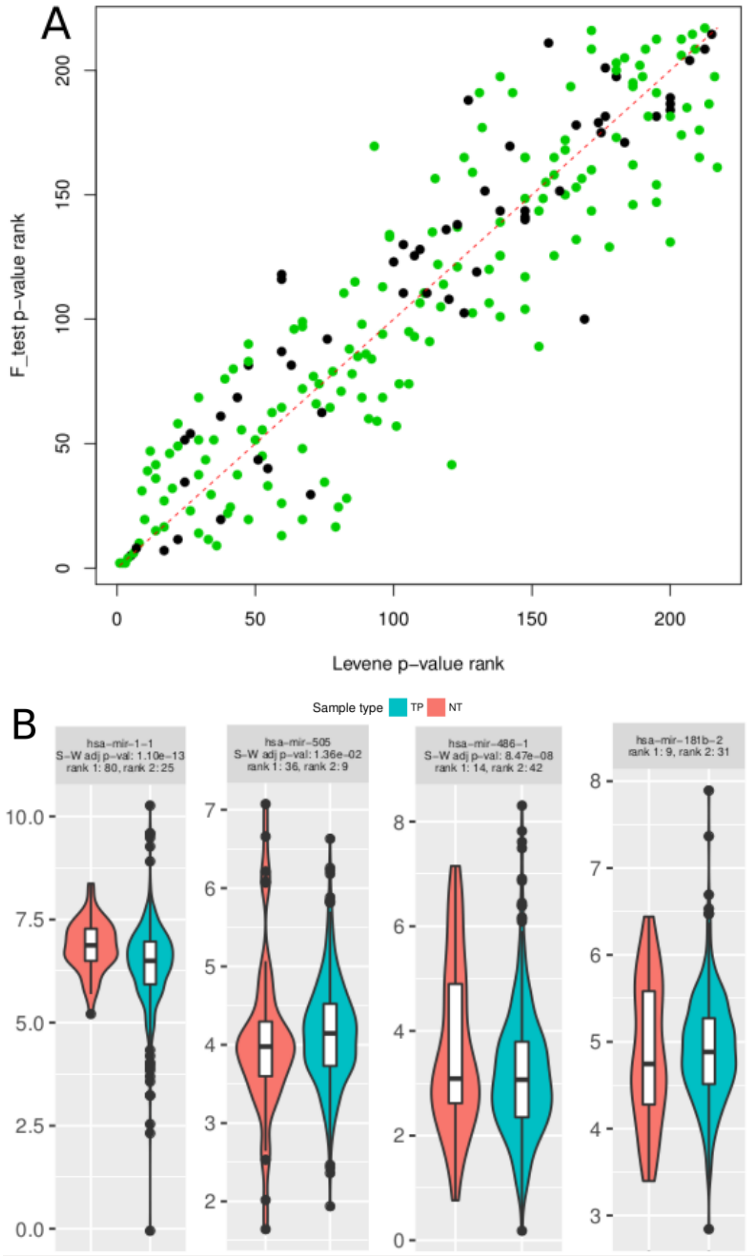

FIGURE 3.3 – A : Nuage de points des rangs de p-valeur obtenues avec le test de Levene et le test de Fisher à partir des données d'expression normalisées des miARN des échantillons tumoraux et sains du cancer de la prostate. Les points verts représentent des miARN dont la p-valeur du test de Shapiro-Wilk est inférieure à 0,05 pour au moins l'un des deux groupes d'échantillons. B : *Boxplots* des valeurs d'expression de quatre miARN dont l'expression ne suit pas la loi normale pour au moins l'un des deux groupes d'échantillons et dont les rangs de p-valeur sont discordants selon le test de Levene et le test de Fisher. Rang 1 : rang de p-valeurs selon le test de Levene, rang 2 : rang de p-valeurs selon le test de Fisher.

si cette différence importante d'effectif entre les deux groupes de données influe sur les conclusions des tests statistiques, les échantillons tumoraux ont été échantillonnés aléatoirement au même nombre que le nombre d'échantillons sains. Pour chaque jeu de données, mille échantillonnages ont ainsi été réalisés. Les tests statistiques ont ensuite été appliqués à ces données sous-échantillonnées et les résultats ont été comparés à

ceux obtenus avec l'intégralité des échantillons disponibles. La figure 3.4 représente les rangs obtenus par les 20 miARN ayant les rangs moyens les plus faibles sur l'ensemble des échantillonages pour les tests de Feltz-Miller, Levene et de Fligner-Killeen. Le test de Feltz-Miller est le test moins sensible à la différence de taille de population. En effet, les distributions de rangs de p-valeur obtenue pour les mille tirages aléatoires sont les plus étroites et les plus proches des rangs obtenus en utilisant l'ensemble des échantillons tumoraux. Au contraire, le test de Fligner-Killeen montre le plus de variabilité dans les rangs de p-valeur obtenue après échantillonnages. Le test de Levene, quant à lui, a des performances proches de celles du test de Feltz-Miller et semble donc assez peu sensible aux tailles de groupe différentes. Les tests de Levene et de Feltz-Miller sont donc les plus appropriés pour comparer la variabilité d'expression des ARN entre les populations d'échantillons tumoraux et sains fournies par le GDC.

1.1.4 Sélection de tests statistiques

Les tests statistiques évalués dans ce chapitre ont des sensibilités différentes aux données ne suivant pas la loi normale et aux tailles d'ensemble d'échantillons différentes. Dans le cadre de la détermination de gènes différentiellement variables entre échantillons tumoraux et échantillons sains, un test statistique de comparaison de variance peu sensible aux écarts à la distribution normale (voir section 1.1.2) et à des populations de tailles différentes (voir section 1.1.3) est requis. Les tests de Levene, de Brown-Forsythe et de Feltz-Miller répondent à ces critères. La figure 3.5 illustre la comparaison des rangs de p-valeur obtenues avec les tests de Levene et de Feltz-Miller à partir des données d'expression de miARN du cancer de la prostate. Les rangs de p-valeur obtenue avec ces deux tests sont assez peu corrélés (corrélation de Pearson égale à 0,63, voir figures 3.1 et 3.5.A). Parmi les miARN dont les rangs sont les plus discordants pour ces deux tests, il apparaît que le test de Feltz-Miller fait des erreurs manifestes. En effet, parmi les miARN les plus discordants, certains miARN dont les distributions des valeurs d'expression présentent des variabilités différentes entre échantillons tumoraux et échantillons sains, comme hsa-mir-486-1 et hsa-mir-20a, qui sont mal classés par le test de Feltz-Miller (rangs 132 et 64 respectivement) alors qu'ils figurent parmi les miARN les plus différentiellement variables selon le test de Levene (rangs 14 et 8 respectivement). A l'inverse, certains miARN, comme hsa-mir-183 et hsa-mir-141, qui font partie des miARN les mieux classés selon le test de Feltz-Miller (rangs 8 et 14 respectivement) ne le sont pas selon le test de Levene (rangs 70 et 121) alors qu'ils ne semblent pas présenter de différence de variance d'expression (figure 3.5.B). Le test de Feltz-Miller n'apparaît donc pas comme étant le plus approprié pour identifier des gènes différentiellement variables à partir des données d'expression issues du TCGA.

Les tests de Levene et de Brown-Forsythe sont des tests statistiques très proches. Leurs statistiques de test ne se différencient que par un seul terme (voir section 2.1 et formules 2.2 du chapitre 2). S'appuyant sur la médiane plutôt que sur la moyenne, le test de Brown-Forsythe est moins sensible aux valeurs extrêmes que le test de Levene. On observe ainsi une tendance parmi les miARN les plus discordants entre ces deux tests : ils ont tous des rangs de p-valeur plus élevés avec le test de Brown-Forsythe qu'avec le test de Levene (figure 3.6.A). Il s'agit de miARN dont certains échantillons sains ont une valeur d'expression nettement différente du reste des échantillons sains (voir figure 3.6.B, notammment hsa-mir-148b). Ces valeurs extrêmes ont d'autant plus de poids que les échantillons sains sont beaucoup moins nombreux que les échantillons tumoraux et que la moyenne des valeurs observées est utilisée par le test de Levene.

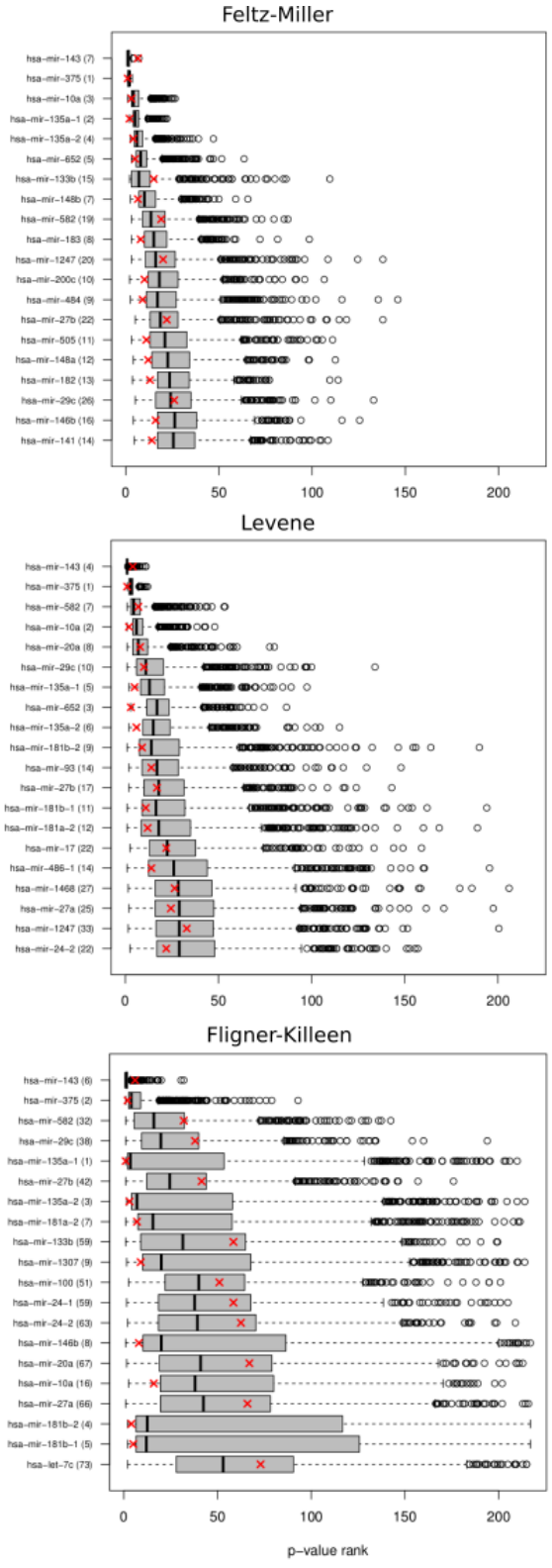

FIGURE 3.4 – Distribution des rangs de p-valeur obtenues pour mille tirages aléatoires des échantillons tumoraux au nombre des échantillons sains. Les 20 miARN ayant les rangs de p-valeur moyens les plus faibles sur l'ensemble des échantillonages sont représentés pour chacun des tests suivants : Feltz-Miller, Levene et Fligner-Killeen. Les croix rouges indiquent les rangs de p-valeur obtenue en considérant l'ensemble des échantillons disponibles.

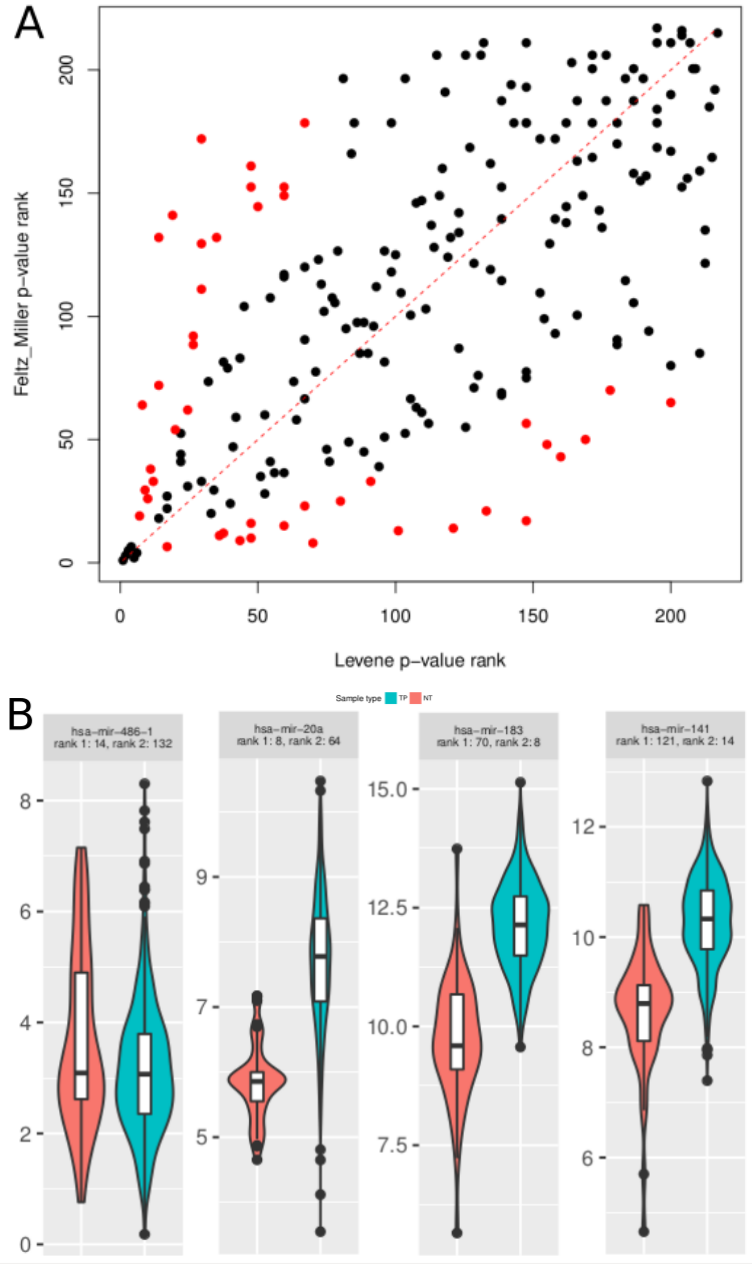

FIGURE 3.5 – A : Comparaison des rangs des p-valeurs obtenues avec les tests de Levene et de Feltz-Miller à partir des données d'expression normalisées de miARN du cancer de la prostate. Les points rouges représentent les miARN les plus discordants selon la formule 2.3. B : *Boxplots* des valeurs d'expression parmi les échantillons tumoraux (TP) et sains (NT) de quelques miARN discordants. Rang 1 : rang de p-valeur selon le test de Levene, rang 2 : rang de p-valeur selon le test de Feltz-Miller.

Ce constat est encore plus marquant lorsque l'on considère les données d'expression d'ARNm (voir figure 3.7). Les ARNm les plus discordants ont, pour quasiment l'intégralité d'entre eux, des rangs de p-valeur plus élevés selon le test de Brown-Forsythe que selon le test de Levene. On observe aussi la présence de valeurs extrêmes parmi les échantillons sains encore plus marquée que pour les miARN. Ces gènes, présentant des valeurs extrêmes parmi les échantillons sains et qu'on ne retrouve pas parmi les échantillons tumoraux, sont d'intérêt biologique dans le cadre de notre étude. Le test

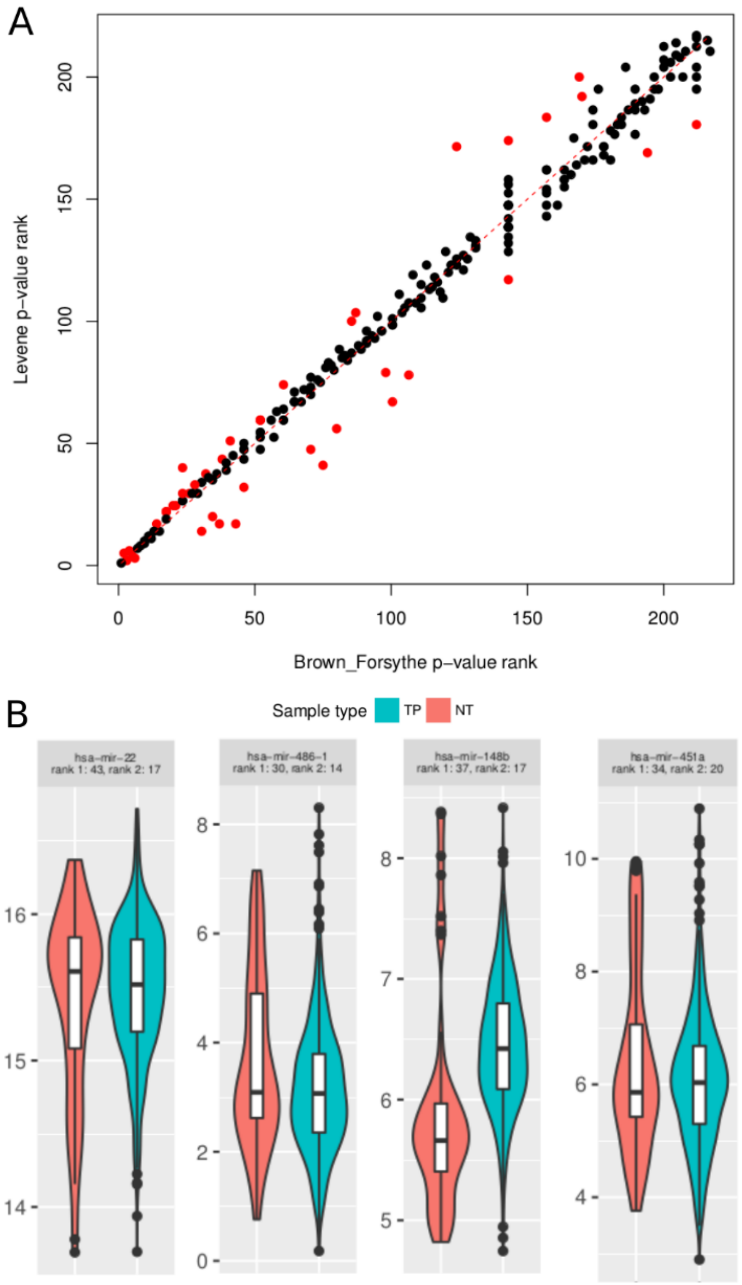

FIGURE 3.6 – A : Comparaison des rangs des p-valeurs obtenues avec les tests de Brown-Forsythe et de Levene à partir des données d'expression normalisées de miARN du cancer de la prostate. Les points rouges représentent les miARN les plus discordants selon la formule 2.3. B : *Boxplots* des valeurs d'expression parmi les échantillons tumoraux (TP) et sains (NT) de quelques miARN discordants. Rang 1 : rang de p-valeur selon le test de Brown-Forsythe, rang 2 : rang de p-valeur selon le test de Levene.

de Levene est donc préféré au test de Brown-Forsythe pour détecter des différences de variabilité et est donc retenu comme le test à appliquer pour caractériser des miARN et des ARNm différentiellement variables entre échantillons tumoraux et échantillons sains issus du TCGA.

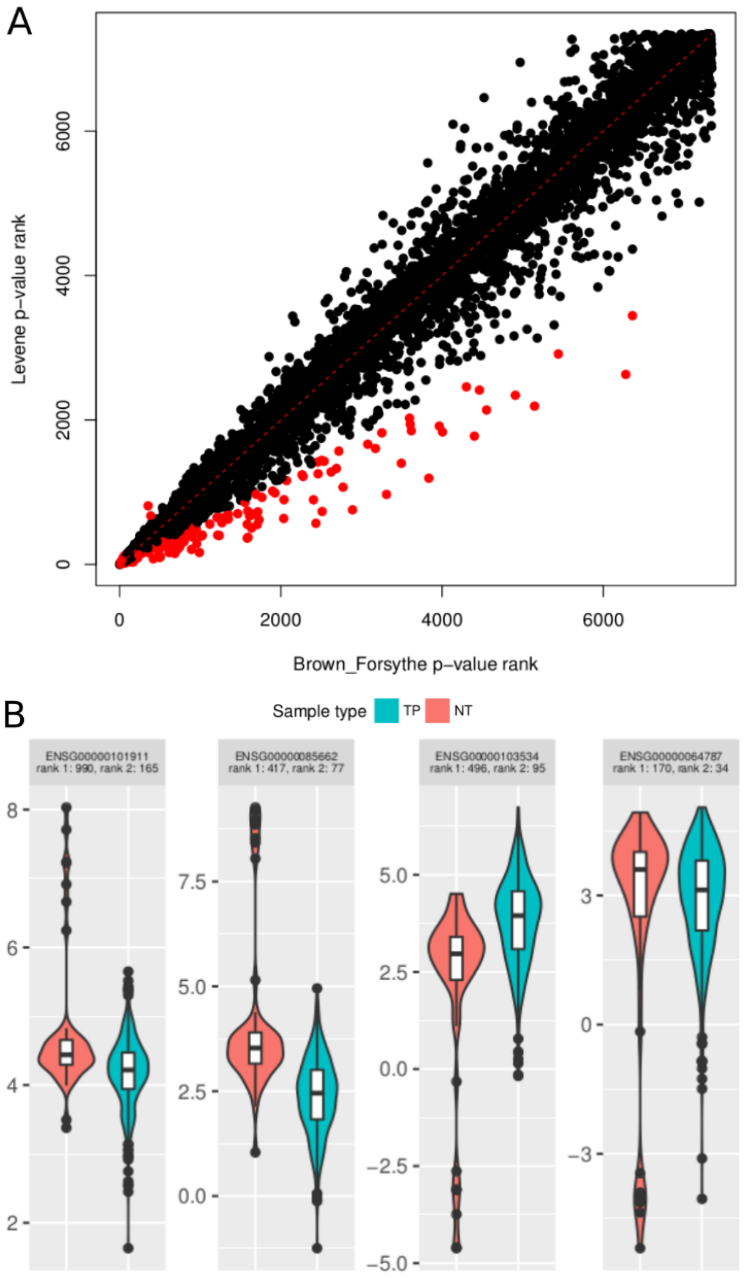

FIGURE 3.7 – A : Comparaison des rangs des p-valeurs obtenues avec les tests de Brown-Forsythe et de Levene à partir des données d'expression normalisées d'ARNm du cancer de la prostate. Les points rouges représentent les ARNm les plus discordants selon la formule 2.3. B : *Boxplots* des valeurs d'expression parmi les échantillons tumoraux (TP) et sains (NT) des quatre ARNm les plus discordants. Rang 1 : rang de p-valeur selon le test de Brown-Forsythe, rang 2 : rang de p-valeur selon le test de Levene (B).

1.2 Application aux données TCGA

Suite à la comparaison des différents tests statistiques visant à détecter des différences de variabilité (section 1.1), le test retenu, le test de Levene, a été appliqué à l'ensemble des jeux de données TCGA listés dans la table 3.1.

1.2.1 Pré-traitement

Avant d'identifier des différences de variance dans l'expression de gènes entre échantillons sains et tumoraux, les données d'expression ont subi différentes étapes de pré-traitement. Les nombres de *reads* ont été normalisés par la méthode TMM puis transformés en $\log_2$ pour permettre la correction d'effets *batch* et l'application de tests de comparaison de variance.

Normalisation des nombres de *reads* Les différentes méthodes de normalisation évaluées par Dillies et al., 2013 ont été appliquées aux données TCGA. Les résultats obtenus confirment ceux de Dillies et al., 2013. La méthode TMM (voir section 1.2 du chapitre 2) permet effectivement de mettre les nombres de *reads* issus de l'ensemble des échantillons à la même échelle. De plus, cette méthode présente l'avantage de laisser une part de variabilité entre les échantillons, au contraire de méthodes telles que les normalisations selon la médiane ou le troisième quartile. De plus, elle conserve intacte la distribution des coefficients de variabilité, ce qui est un critère essentiel pour pouvoir détecter des différences de variance par la suite (données non montrées). Ainsi, corroborant les constatations et les conclusions de Dillies et al., 2013, les nombres de *reads* ont été normalisés selon la méthode TMM.

Log-transformation Après normalisation, les nombres de *reads* ont été transformés en $\log_2$ dans le but de pouvoir appliquer la méthode de correction d'effets *batch* et le test de Levene par la suite (voir section 1.3 du chapitre 2).

Filtrage des gènes faiblement exprimés Un filtre sur les comptes moyens à travers l'ensemble des échantillons du jeu de données a été appliqué pour filtrer les gènes faiblement exprimés sans introduire de biais (voir section 1.1 du chapitre 2). Après $\log_2$-transformation, les gènes dont l'expression moyenne est inférieure à 2 ont été retirés de l'analyse.

Correction d'effets *batch* La présence de valeurs extrêmes peut fortement influencer les conclusions de tests de comparaison de variabilité (voir section 1.1.2). Bien que le test de Levene y soit assez peu sensible, il convient d'y porter attention, d'autant plus si elles sont le reflet de biais techniques et non le reflet d'un véritable effet biologique.

TCGA a fait un effort d'harmonisation des différents protocoles utilisés pour constituer leurs jeux de données. Ainsi, les mêmes séquenceurs et les mêmes programmes informatiques de conversion des données brutes en nombres de *reads* ont été utilisés dans le but de limiter les sources de biais techniques. Les effets *batch*, qui sont l'une des sources principales de ces biais techniques, sont malgré tout possibles au sein des jeux de données TCGA. Ils se matérialisent par la présence de valeurs d'expression aberrantes, très hautes ou très basses, pour un gène provenant uniquement d'une ou de quelques rares expériences de séquençage ayant permis de constituer le jeu de données. Ces effets peuvent être facilement détectés à l'aide d'Analyses en Composantes Principales (ACP). Lorsqu'ils sont présents, ces effets captent en effet la variabilité représentée par les premières composantes. Dans l'ensemble, les jeux de données TCGA sont assez peu impactés par les effets *batch*. Cependant, quelques jeux de données comme celui de l'expression des ARNm pour le cancer du colon présentent manifestement des effets *batch* (figure 3.8). Globalement, les échantillons se rassemblent bien en fonction de la catégorie biologique, échantillons sains et échantillons tumoraux,

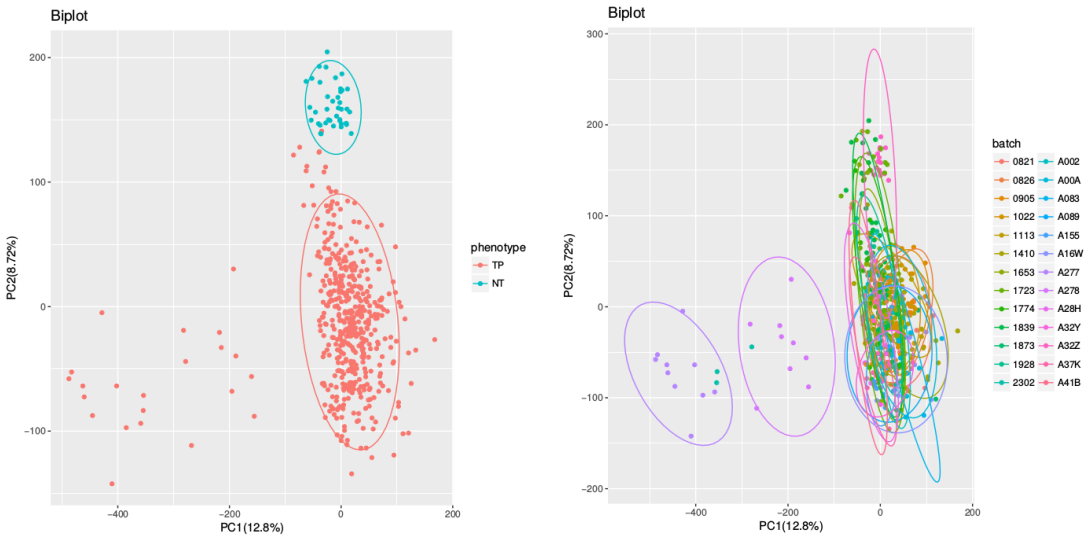

FIGURE 3.8 – ACP des données d'expression du cancer du colon (TCGA-COAD) avant l'application de la méthode de correction d'effets *batch* ComBat. Graphique de gauche : les échantillons sont colorés en fonction de leur appartenance aux catégories biologiques d'intérêt (échantillons sains en bleu, échantillons tumoraux en rouge). Graphique de droite : les échantillons sont colorés en fonction des expériences de séquençage les ayant générés. Des ellipses ont été ajoutées pour chaque groupe d'échantillons ainsi définis pour pouvoir mieux les distinguer.

selon les deux premières composantes. Cependant, parmi les échantillons tumoraux, certains se démarquent très nettement selon la première composante. Le fait que les expériences de séquençage A277 et A278 n'aient séquencé que des échantillons présentant des valeurs extrêmes captées par la première composante est le reflet d'effets *batch*.

Ainsi, les jeux de données analysés dans ce chapitre ont été corrigés en fonction de la présence d'effets *batch* en appliquant la méthode ComBat (voir section 1.4.1 du chapitre 2). Dans le cas du jeu de données d'expression des ARNm pour le cancer du colon, l'application permet effectivement de corriger ces effets (figure 3.9). En effet, la présence d'échantillons issus de quelques expériences de séquençage précises se démarquant très fortemement selon l'une des deux premières composantes, qui captent le plus de variabilité, n'est plus observée alors que la distinction entre les catégories biologiques est conservée.

1.2.2 Identification de gènes différentiellement variants

Le test de Levene a été appliqué à l'ensemble des jeux de données d'expression de miARN et d'ARNm listés dans la table 3.1 après avoir subi les différentes étapes de pré-traitement énumérées dans la section 1.2.1. Une p-valeur inférieure à 0,05 après correction selon la méthode de Benjamini-Hochberg (voir section 4 du chapitre 2) est utilisée comme critère pour identifier des gènes, miARN ou ARNm, différentiellement variants.
La figure 3.10 représente les proportions de gènes différentiellement variants par rapport à l'ensemble des gènes analysés ainsi que la proportion de gènes dont la variance augmente dans les tumeurs par rapport à l'ensembles des gènes différentiellement variants. Les miARN et les ARNm présentent des tendances similaires selon leur chan-

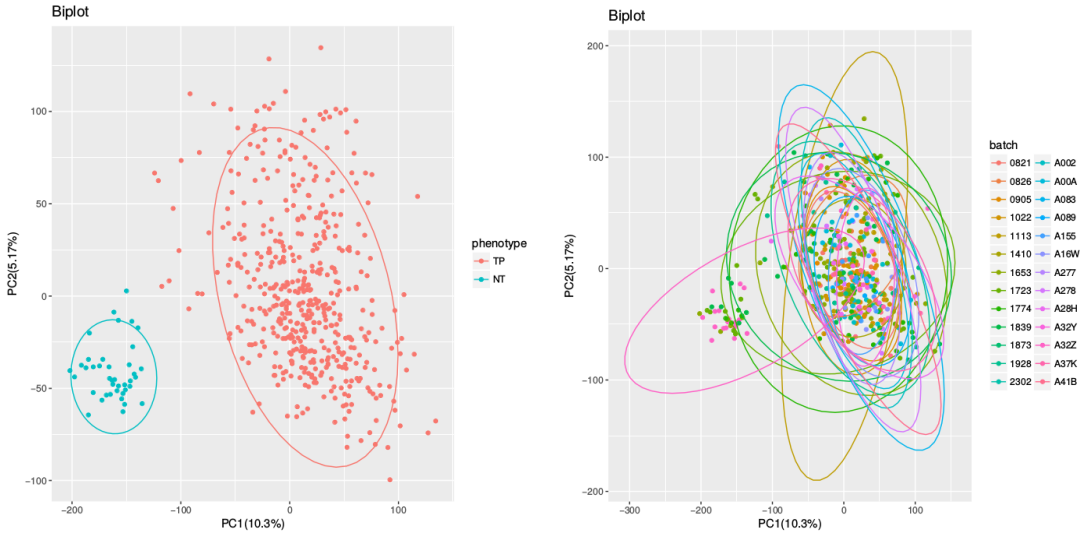

FIGURE 3.9 – ACP des données d'expression du cancer du colon (TCGA-COAD) après l'application de la méthode de correction d'effets *batch* ComBat. Les mêmes groupes d'échantillons sont définis sur ces deux graphiques que pour la figure 3.8.

gement de variance d'expression entre échantillons tumoraux et échantillons sains. Pour de nombreux différents cancers, une part très importante (près de la moitié ou plus) de gènes sont identifiés comme différentiellement variants. Seuls les cancers de l'œsophage, de la prostate et de la vessie se distinguent par un nombre très faible de gènes différentiellement variants. De plus, parmi les gènes différentiellement variants, l'écrasante majorité des gènes présentent une augmentation de variance d'expression dans les tumeurs. Cette part frôle même les 100% pour certains cancers pour lesquels un grand nombre de gènes différentiellement variants ont été détectés (foie, poumon, rein, sein).

2 Discussion

Les tendances extrêmement marquées observées dans les résultats de la détection de gènes différentiellement variants semblent indiquer que les tests de comparaison de variabilité sont peu pertinents dans le cadre de la comparaison d'échantillons sains et tumoraux. En effet, pour un nombre important de cancers différents, un très grand nombre de gènes sont identifiés comme différentiellement variants et, parmi ceux-ci, la quasi totalité présente une augmentation de la variance d'expression dans les tumeurs. Ces très fortes tendances ne permettent pas d'évaluer l'hypothèse émise dans le cadre de cette thèse de caractérisation de la perturbation du rôle tampon des miARN lors de la cancérogénèse à travers la modification de leur variance d'expression et de celle de leurs ARNm cibles (voir section 3 dans le chapitre 1).

Pour pouvoir appliquer le test de Levene, les données ont dû être lourdement transformées. En particulier, la transformation en $\log_2$, indispensable à l'application de méthodes de correction d'effets *batch* et de tests de différence de variabilité, modifie énormément la variance des données. Cela aboutit à considérer plus variants des gènes dont l'expression très faible s'étale sur plusieurs valeurs log par rapport à des gènes très fortement exprimés dont les valeurs sont comprises entre deux valeurs log. Ainsi, certains auteurs déconseillent très fortement de transformer des valeurs de comptage

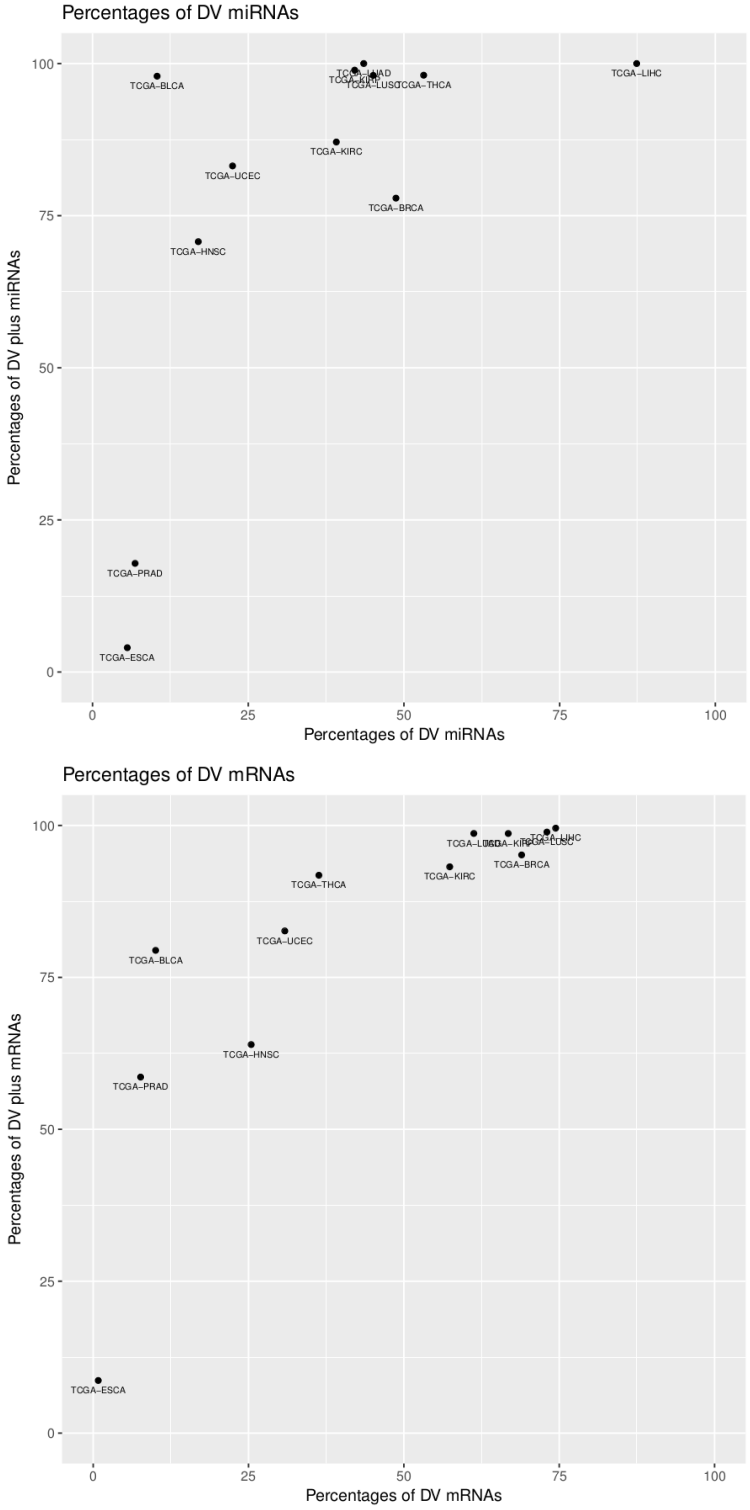

FIGURE 3.10 – Pourcentages de gènes, miARN (en haut) et ARNm (en bas), identifiés comme étant différentiellement variants (DV) après l'application du test de Levene dans le cadre de la comparaison des échantillons tumoraux et sains du TCGA. L'axe des ordonnées sur ces graphiques représente le pourcentage de gènes présentant une augmentation de leur variance d'expression (« DV plus ») dans les tumeurs parmi l'ensemble des gènes différentiellement variants.

sur une échelle logarithmique (O'Hara et Kotze, 2010). Par cette approche, il est ainsi difficile d'analyser une différence de variance indépendamment d'une différence de moyenne. De plus, les méthodes de correction d'effets *batch* modifient les valeurs des nombres de *reads* jugées aberrantes. En plus de les transformer, ces méthodes peuvent introduire des biais vers l'observation de l'effet biologique recherché. Nygaard, Rodland et Hovig, 2016 ont en effet montré que la correction d'effets *batch* tendait à renforcer les différences entre les groupes biologiques dont la comparaison constitue l'intérêt principal de l'étude.

Dans le cadre de l'analyse classique de détection de différence de moyenne d'expression, les nombres de *reads* issus du RNA-seq sont représentés par des variables aléatoires suivant des distributions binomiales négatives. Cette distribution représente mieux les caractéristiques de ces données que l'approximation de distribution normale après $\log_2$-transformation faite dans le cadre de ce chapitre. En effet, les nombres de *reads* prenant des valeurs entières, une loi de probabilité de valeurs discrètes se prête mieux à la représentation de ce type de données. Elle permet, en outre, de représenter l'augmentation de la variance des nombres de *reads* généralement observée pour les gènes fortement exprimés grâce à l'emploi d'un paramètre de dispersion. Ce paramètre supplémentaire permet ainsi de caractériser une modification de la variance de manière indépendante de la moyenne. Enfin, ces méthodes emploient généralement un GLM qui permet de prendre en compte, entre autres, les effets *batch* par une variable explicative (voir section 1.4.2 du chapitre 2) sans modifier les valeurs de nombre de *reads*. Les données sont ainsi beaucoup moins transformées en ne subissant que l'étape de normalisation, réduisant ainsi l'introduction de biais avant même l'analyse de différence de variance d'expression. Ces modèles basés sur la distribution binomiale négative apparaissent plus appropriés pour analyser des données RNA-seq en général et pour analyser leur variance en particulier. Toutefois, ces méthodes n'ont permis pendant longtemps que la détection de différence de moyenne d'expression. La variance, définie par le paramètre de dispersion, n'est considérée que comme un paramètre de nuisance qu'il s'agit d'estimer au mieux de manière préalable à l'analyse d'intérêt de différence de moyenne. Mais, très récemment, la publication au cours de ma thèse des méthodes MDSeq (Ran et Daye, 2017) en mai 2017 et DiPhiSeq (Li et Lamere, 2018) en novembre 2018 étend l'emploi de ces modèles basés sur la distribution binomiale négative à la détection de différence de variance. Ces nouvelles méthodes ouvrent ainsi de nouvelles perspectives pour l'identification de gènes présentant des différences de variance d'expression entre populations d'échantillons. Ran et Daye, 2017 ont comparé leur méthode, MDSeq, à certains tests de différence de variabilité évalués dans ce chapitre (les test de Bartlett, Levene et Fligner-Killeen) et ont montré, à l'aide de données d'expression simulées ne présentant pas de différence de moyenne entre deux populations d'échantillons, qu'elle permettait un meilleur contrôle des erreurs de type I (figure 3.11) et une meilleure sensibilité (figure 3.12) pour l'identification de gènes différentiellement variants que ces tests. En particulier, MDSeq présente l'avantage d'être plus puissante que le test de Levene, tout en présentant un nombre comparable d'erreurs de type I. De plus, cette méthode permet une meilleure prise en compte d'une caractéristique des données RNA-seq et qui peut influer sur les tests statistiques : la présence en grand nombre de valeurs nulles.

Ces deux nouvelles méthodes semblent ainsi beaucoup plus adaptées pour répondre à la problématique de l'identification de gènes présentant des différences de variance d'expression entre deux populations d'échantillons que l'adaptation de tests classiques de différence de variabilité après transformation des données. L'emploi de ces tests de

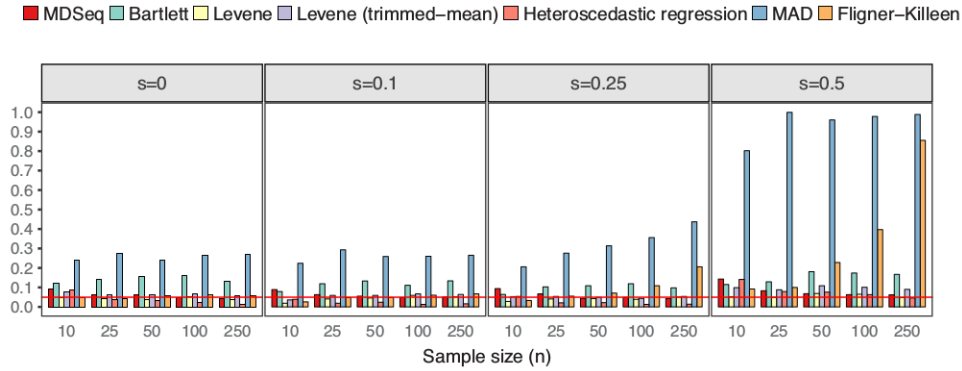

FIGURE 3.11 – Proportions d'erreurs de type I de MDSeq et différents tests de différence de variabilité obtenues à l'aide de données simulées ne présentant pas de différence de variance d'expression entre deux populations d'échantillons de tailles égales (figure 1 de RAN et DAYE, 2017). Différentes tailles de population (*sample size*) et différentes proportions de valeurs nulles (s) dans les valeurs d'expression de chaque gène ont été évaluées.

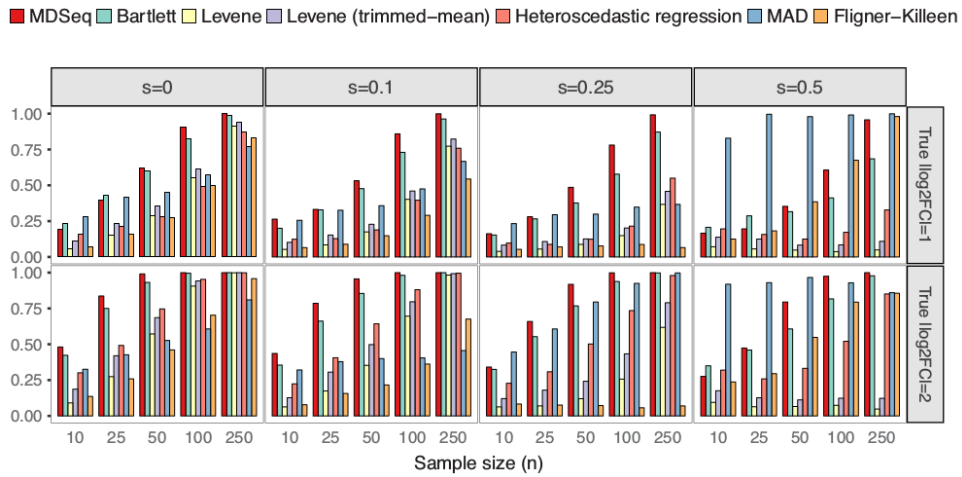

FIGURE 3.12 – Puissances de MDSeq et différents tests de différence de variabilité obtenues à l'aide de données simulées présentant des différences de variance d'expression (*log2FC*) entre deux populations d'échantillons de tailles égales (figure 2 de RAN et DAYE, 2017). Différentes tailles de populations (*sample size*) et différentes proportions de valeurs nulles (s) dans les valeurs d'expression de chaque gène ont été évaluées.

différence de variabilité n'a donc pas plus été approfondi et l'évaluation de MDSeq et de DiPhiSeq et leur application aux mêmes jeux de données sont les objets du chapitre suivant.

Chapitre 4

Identification de gènes différentiellement dispersés

L'application des deux méthodes basées sur la distribution binomiale négative permettant la détection de différence de dispersion entre deux populations d'échantillons de RNA-seq, MDSeq (Ran et Daye, 2017) et DiPhiSeq (Li et Lamere, 2018), est l'objet de ce chapitre 4. Leurs performances pour la détection de différence de dispersion ont été évaluées à l'aide de jeux de données simulées (section 1). Les résultats de cette étude de simulation a guidé ensuite leur application aux jeux de données d'expression TCGA (section 2).

1 Données simulées

1.1 Résultats

1.1.1 Pré-traitement

Avant d'être analysés par les méthodes MDSeq et DiPhiSeq, les jeux de données simulées ont subi des étapes classiques de pré-traitement. Bien que les échantillons des jeux de données aient été simulés de telle sorte qu'ils aient tous la même taille de librairie (voir section 5.3 du chapitre 2), les nombres de *reads* ont été normalisés selon la méthode TMM (Robinson et Oshlack, 2010) pour prendre en compte la présence éventuelle de gènes très fortement exprimés dans certains échantillons (voir section 1.2 du chapitre 2). Après normalisation, les gènes faiblement exprimés ont été filtrés en utilisant un seuil minimal d'expression moyenne de 1 CPM sur l'ensemble des échantillons des deux conditions (voir section 1.1 du chapitre 2).

1.1.2 Méthodes de correction de tests multiples

Par défaut, les auteurs de MDSeq et de DiPhiSeq ont opté pour deux méthodes différentes de correction de tests multiples : la méthode de Benjamini-Yekutieli (BY) pour MDSeq et la méthode de Benjamini-Hochberg (BH) pour DiPhiSeq. Les histogrammes de p-valeurs non corrigées et corrigées selon les méthodes BH et BY permettent de visualiser l'impact de ces méthodes de correction (figures 4.1 et 4.2). Ici, un jeu de données simulées pour 10 000 gènes est pris en exemple. Un test étant réalisé pour chaque gène et le nombre total de gènes pris en compte étant élevé, on s'attend à ce que les méthodes de correction aient un effet conséquent sur les p-valeurs non corrigées.

La méthode BH a un effet assez limité sur les p-valeurs obtenues avec MDSeq (figures 4.1) et ne parvient pas à maintenir le FDR inférieur à 0,05. En effet, il est égal à 0,13 avec les p-valeurs non corrigées et n'est abaissé qu'à 0,08 après correction. De manière

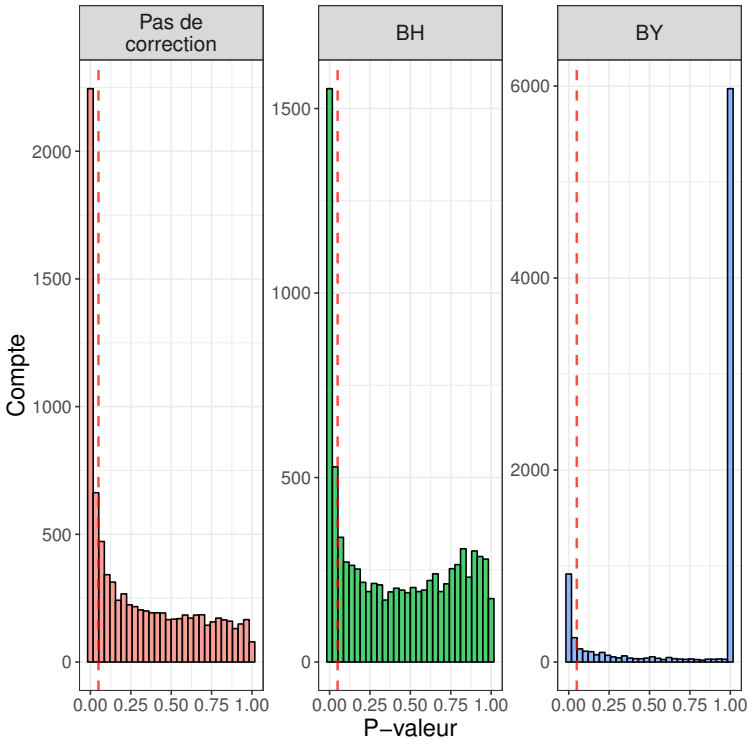

FIGURE 4.1 – Histogrammes des p-valeurs non corrigées et corrigées selon les méthodes de Benjamini-Hochberg (BH) et de Benjamini-Yekutieli (BY) obtenues avec MDSeq pour la détection de différence de dispersion sur un jeu de données simulées composé de deux populations de 50 échantillons. Paramètres de simulations : 10 000 gènes, 50% de gènes différentiellement dispersés avec une différence de dispersion d'au moins 50%, *fold-changes* de moyenne contenus entre 1 et 1,3, introduction d'un *outlier* pour 10% des gènes. Les nombres de *reads* ont été normalisés selon la méthode TMM et les gènes dont la moyenne d'expression est inférieure à 1 CPM ont été filtrés, la fonction de retrait d'*outliers* de MDSeq a été appliquée. Les gènes différentiellement dispersés ont été identifiés à l'aide d'un seuil de *fold-change* égal à 1. La ligne rouge en pointillés représente le seuil de p-valeur à 0,05.

générale, la distribution est assez peu impactée par cette méthode de correction, ce qui suggère que cette méthode de correction de p-valeurs n'est pas assez forte. En revanche, l'histogramme des p-valeurs corrigées selon la méthode BY est plus conforme aux effets attendus d'une correction de p-valeurs. Les p-valeurs significatives, *i.e.* inférieures à 0,05 ont nettement diminué tout en restant assez nombreuses et la forme de l'histogramme a radicalement changé avec l'apparition d'un pic de p-valeurs corrigées à 1. Le FDR est ainsi contrôlé avec cette méthode, prenant la valeur de 0,02. Ainsi, comme attendu, la méthode BY est une méthode très conservatrice mais, appliquée aux p-valeurs obtenues par MDSeq, elle permet tout de même de conserver de nombreuses très faibles p-valeurs après correction.

Les méthodes de correction BH et BY ont des effets différents sur les p-valeurs obtenues avec DiPhiSeq sur le même jeu de données. La méthode de correction BH a en effet bien plus d'impact avec une nette diminution du nombre de très faibles p-valeurs et parvient à contrôler le FDR en l'abaissant de 0,07 à 0,01. En revanche, en augmentant les p-valeurs de telle sorte qu'il ne reste que quelques p-valeurs inférieures à 0,05, la correction par la méthode BY apparaît comme étant beaucoup trop conservatrice.

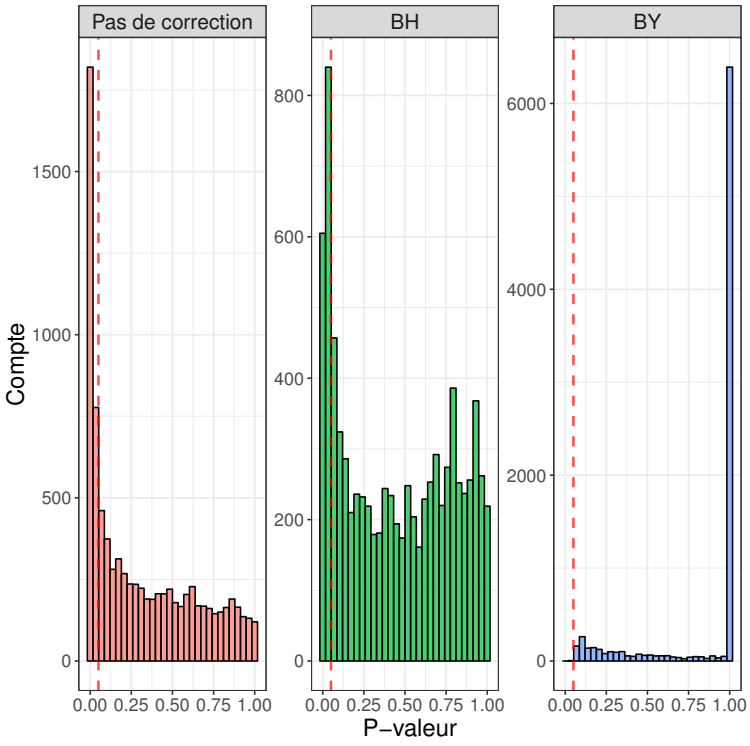

FIGURE 4.2 – Histogrammes des p-valeurs non corrigées et corrigées selon les méthodes de Benjamini-Hochberg (BH) et de Benjamini-Yekutieli (BY) obtenues avec DiPhiSeq pour la détection de différence de dispersion sur le même jeu de données simulées que pour la figure 4.1. Les nombres de *reads* ont été normalisés selon la méthode TMM et les gènes dont la moyenne d'expression est inférieure à 1 CPM ont été filtrés. La ligne rouge en pointillés représente le seuil de p-valeur à 0,05.

Ces tendances sont observées pour l'ensemble des jeux de données considérés et confirment donc le choix de méthode correction des p-valeurs fait par les auteurs respectifs de MDSeq et de DiPhiSeq. Dans la mesure où elles semblent être adaptées respectivement à ces deux méthodes, cette différence de méthode de correction de p-valeurs n'empêche pas de comparer les résultats obtenus par MDSeq et DiPhiSeq mais sera à prendre en considération pour certaines interprétations.

1.1.3 Influence de la présence d'une différence de moyenne

Des jeux de données avec des populations de tailles égales et des gènes dont le *fold-change* de moyenne d'expression n'est pas limité par un maximum ont été utilisés dans un premier temps. Les performances de MDSeq pour la détection de gènes différentiellement dispersés sont fortement impactées par la présence de forts *fold-changes* de moyenne (figure 4.3). En effet, pour des populations de 50 échantillons, l'AUC est de 0,78 pour les gènes ayant un faible *fold-change* de moyenne, *i.e.* les gènes ayant un *fold-change* de moyenne inférieur à 1,3 et appelés « non DE » sur la figure 4.3, alors qu'elle chute à 0,58 pour les gènes ayant un *fold-change* de moyenne supérieur à 1,3. Les performances de DiPhiSeq, en revanche, ne varient pas en fonction de la présence de forts *fold-changes* de moyenne d'expression.

De manière générale, l'AUC obtenue avec DiPhiSeq est plus élevée que celle obtenue avec MDSeq, y compris pour les gènes ayant un faible *fold-change* de moyenne

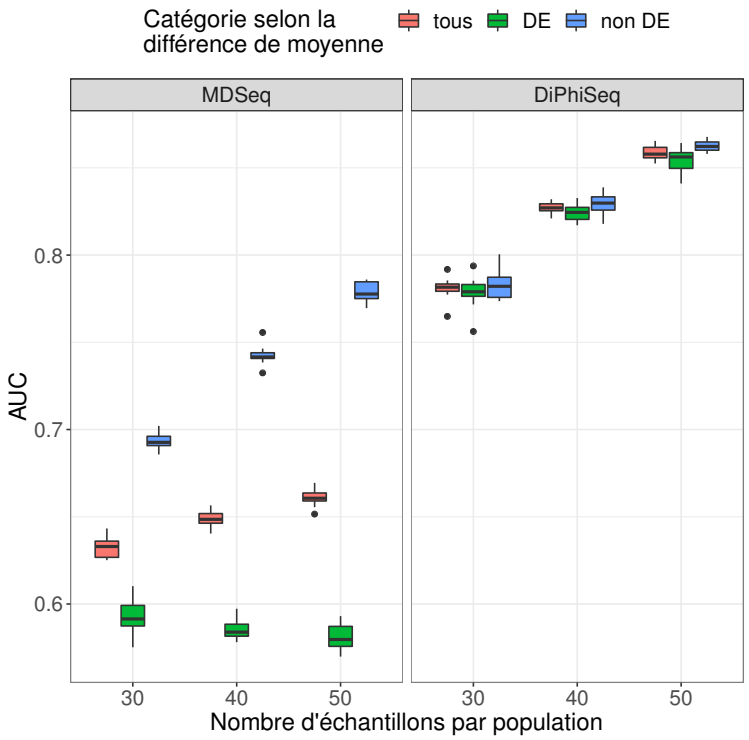

FIGURE 4.3 – AUC obtenues avec MDSeq et DiPhiSeq pour la détection de différence de dispersion avec des jeux de données simulées composés de populations d'échantillons de tailles égales. Paramètres de simulation : 10 000 gènes, 50% de gènes différentiellement dispersés avec une différence de dispersion d'au moins 50%, les gènes non différentiellement dispersés ont la même valeur de dispersion pour les deux populations, une valeur de *fold-change* de moyenne de 1,3 définit la séparation entre gènes différentiellement exprimés (DE) et non différentiellement exprimés (non DE), présence d'un *outlier* pour 10% des gènes. Les performances ont été mesurées à l'aide de 10 réplicats pour chaque jeu de paramètres. La fonction de retrait d'*outliers* de MDSeq a été appliquée. Les gènes différentiellement dispersés ont été identifiés à l'aide d'un seuil de *fold-change* égal à 1 pour les deux méthodes.

d'expression. La forte correction des p-valeurs obtenues avec MDSeq par la méthode de Benjamini-Yekutieli qui se traduit par un très grand nombre de p-valeurs égales à 1 (figure 4.1) peut expliquer les plus faibles valeurs d'AUC obtenues avec MDSeq. Les p-valeurs de DiPhiSeq étant corrigées par une méthode beaucoup moins conservatrice, la méthode de Benjamini-Hochberg, les performances de ces deux méthodes pour l'identification de gènes différentiellement dispersés ne peuvent être évaluées uniquement à l'aide de cet indicateur.

Bien que les *fold-changes* de dispersion et de moyenne aient été introduits de manière indépendante lors de la simulation des jeux de données, les *fold-changes* de dispersion et de moyenne estimés par MDSeq sont fortement corrélés (figure 4.4). En effet, les prédictions correctes de différence de dispersion se concentrent pour les gènes dont la moyenne et la dispersion varient de manière similaire, *i.e.* augmentation ou diminution pour les deux valeurs, entre les deux populations d'échantillons. A l'inverse, il y a très peu de vrais positifs pour les gènes dont la moyenne et la dispersion d'expression varient de manière opposée, *i.e.* augmentation pour l'une et diminution pour l'autre. Enfin, les faux positifs se concentrent autour de la droite de corrélation $y = x$. Ils se trouvent donc majoritairement parmi les gènes dont les *fold-changes*

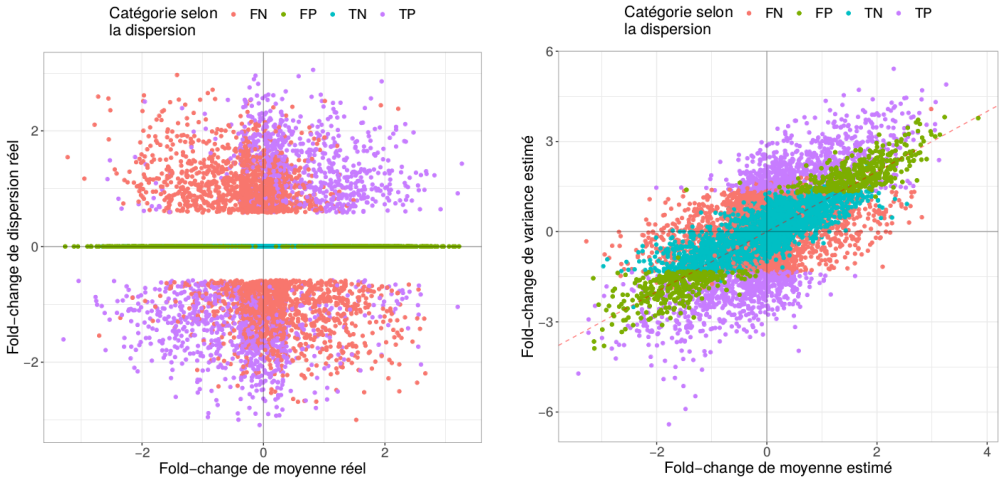

FIGURE 4.4 – Valeurs réelles de *fold-changes* de moyenne et de dispersion d'un jeu de données simulées composé de 50 échantillons par population (à gauche). Valeurs de *fold-changes* de moyenne et de variance estimées par MDSeq sur ce même jeu de données simulées. Les couleurs des points correspondent aux résultats de test de différence de variance réalisés par MDSeq avec un *fold-change* de 1 (à droite). La droite rouge en pointillés représente la diagonale $y = x$.

estimés de moyenne et de variance ont des valeurs très proches. Cette corrélation est due au fait que le GLM proposé par MDSeq (voir formule 2.20) ne permet, en réalité, pas de modéliser la dispersion de la variable aléatoire Y_{ij} mais sa variance. La corrélation observée entre les *fold-changes* de moyenne et de variance s'explique alors par la reparamétrisation de la distribution binomiale négative où $Var(Y_{ij}) = \phi_{ij}\,\mu_{ij}$ (voir formule 2.20). Ce que les auteurs de MDSeq appellent « dispersion » ou « variabilité » est donc en fait la variance de la variable aléatoire. Or, des variables aléatoires suivant des distributions binomiales négatives peuvent avoir des variances différentes tout en ayant la même valeur de dispersion, la différence de variance étant uniquement due à une différence de moyenne. Ce sont donc des gènes différentiellement exprimés et non différentiellement dispersés qui constituent les faux positifs observés autour de la droite de corrélation sur la figure 4.4.

Le paramètre de dispersion est le paramètre qui permet de comparer les variances de variables aléatoires suivant des distributions binomiales négatives indépendamment de leurs moyennes respectives. Bien que MDSeq modélise la variance de la variable aléatoire Y_{ij}, la dispersion peut malgré tout être déduite à partir des *fold-changes* de moyenne et de variance estimés. En effet, comme indiqué par les auteurs de MDSeq, la dispersion peut être estimée grâce à la reparamétrisation de la distribution binomiale négative utilisée :

$$\log\left[Var(Y_{ij})\right] = \log(\phi_{ij}) + \log(\mu_{ij}).$$

Appliquée à la comparaison de deux variables aléatoires, le *fold-change* de dispersion peut ainsi être calculé :

$$\log(FC_\phi) = \log(FC_{Var}) - \log(FC_\mu)\,,$$

où $\log(FC_\phi)$, $\log(FC_{Var})$ et $\log(FC_\mu)$ sont les *fold-changes* de dispersion, de variance et de moyenne respectivement de l'expression d'un gène entre deux populations d'échantillons.

La répartition des *fold-changes* estimés de dispersion et de moyenne est alors plus conforme à ceux qui ont été réellement introduits lors de la simulation du jeu de données et aucune corrélation entre ces deux types de *fold-change* n'est observée (figure 4.5). Les gènes présentant une différence de dispersion d'expression entre les deux

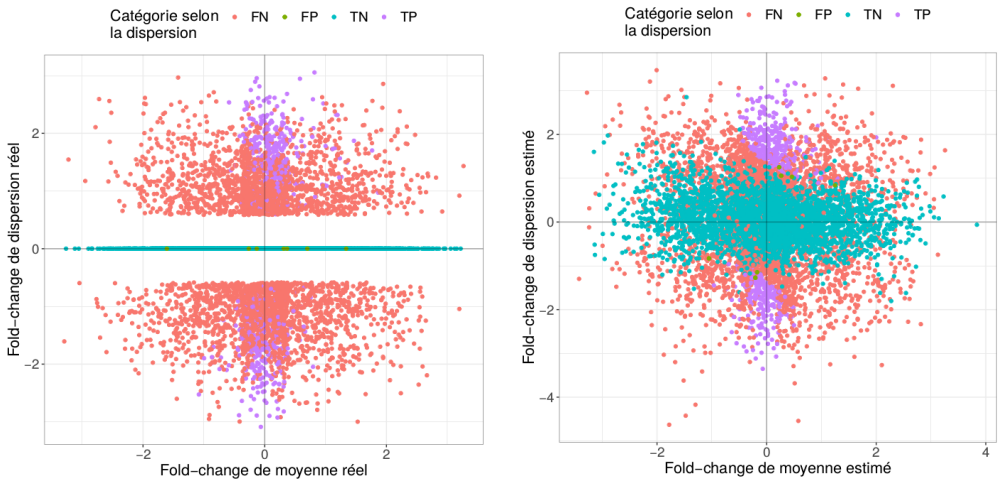

FIGURE 4.5 – Valeurs réelles de *fold-changes* de moyenne et de dispersion d'un jeu de données simulées composé de 50 échantillons par population (le même jeu de données que pour la figure 4.4, à gauche). Valeurs de *fold-changes* de moyenne et de dispersion estimées par MDSeq sur ce même jeu de données simulées. Les couleurs des points correspondent aux résultats de test de différence de variance réalisés par MDSeq avec un *fold-change* de 1 (à droite).

populations d'échantillons sont alors déterminés à l'aide des p-valeurs des tests de différence de moyenne et de variance développés par MDSeq. Une différence de variance pouvant potentiellement être uniquement due à une différence de moyenne, une différence de dispersion ne peut être trouvée que parmi les gènes non différentiellement exprimés. Les résultats positifs pour la détection de différence de dispersion sont ainsi définis par une p-valeur non significative pour le test de différence de moyenne et une p-valeur significative pour le test de différence de variance. En appliquant cette pratique, les vrais positifs se retrouvent uniquement parmi les gènes faiblement différentiellement exprimés et très peu de faux positifs sont obtenus (figure 4.5).

MDSeq ne peut donc être utilisé pour trouver des différences de dispersion que parmi des gènes non différentiellement exprimés, *i.e.* des gènes présentant une faible différence de moyenne d'expression. Les performances de détection de différence de dispersion de MDSeq ont donc été évaluées, dans un premier temps, avec des jeux de données simulées constitués de gènes dont les *fold-changes* de moyenne ont été limités par différentes valeurs maximales comprises entre 1,1 et 1,5 (voir section 5.3 du chapitre 2). Pour ces jeux de données où les *fold-changes* de moyenne sont limités, les gènes différentiellement dispersés ont été identifiés avec MDSeq uniquement à l'aide de la p-valeur du test de différence de variance. Les performances de détection de différence de dispersion de MDSeq ainsi obtenues sont comparées à celles obtenues avec DiPhiSeq sur les mêmes jeux de données.
De manière générale, les valeurs d'AUC obtenues par DiPhiSeq sont meilleures que celles obtenues par MDSeq (figure 4.6). Comme pour la section précédente, les différentes méthodes de correction de p-valeurs expliquent certainement en partie cet écart de valeurs d'AUC et ne permet donc pas d'utiliser cet indicateur comme critère

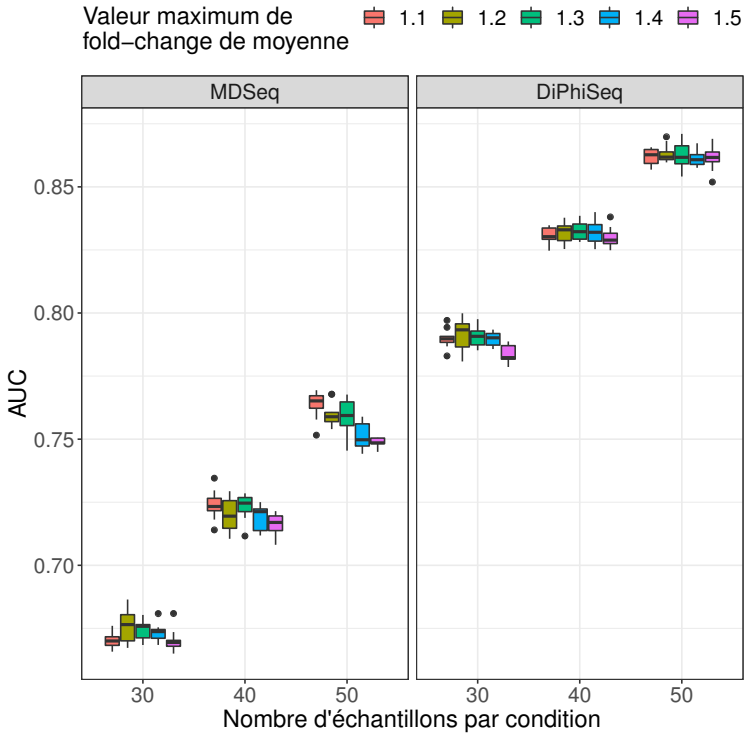

FIGURE 4.6 – AUC obtenues avec MDSeq et DiPhiSeq pour la détection de différence de dispersion avec des jeux de données simulées composés de populations d'échantillons de tailles égales. Paramètres de simulation : 10 000 gènes, 50% de gènes différentiellement dispersés avec une différence de dispersion d'au moins 50%, les gènes non différentiellement dispersés ont la même valeur de dispersion pour les deux populations, les *fold-changes* de moyenne sont limités par différentes valeurs maximales comprises entre 1,1 et 1,5, présence d'un *outlier* pour 10% des gènes. Les performances ont été mesurées à l'aide de 10 réplicats pour chaque jeu de paramètres. La fonction de retrait d'*outliers* de MDSeq a été appliquée. Les gènes différentiellement dispersés ont été identifiés à l'aide d'un seuil de *fold-change* égal à 1 pour les deux méthodes.

de comparaison des deux méthodes. Toutefois, on peut noter que l'augmentation de la taille des populations d'échantillons permet une augmentation de l'AUC pour les deux méthodes.

Le but des méthodes de correction de p-valeurs est de maintenir le risque de première espèce α à 5% de chaque test individuel à l'échelle du jeu de données entier. Les deux méthodes parviennent à contrôler le FDR inférieur à 0,05 (figure 4.7). Cependant, des différences notables entre les deux méthodes existent. DiPhiSeq parvient à maintenir le FDR à des valeurs très basses, inférieures ou égales à 0,02, quelles que soient les tailles des populations d'échantillons considérées et les valeurs maximales de *fold-change* de moyenne d'expression tolérées. Par contre, l'augmentation des valeurs maximales de *fold-change* de moyenne d'expression tolérées entraîne une augmentation du FDR avec MDSeq, atteignant 0,05 pour quelques réplicats lorsque les *fold-changes* de moyenne peuvent atteindre 1,5. Cela suggère une perte de contrôle du FDR et confirme l'influence de l'existence d'un *fold-change* de moyenne sur les p-valeurs des tests de différence de variance.

Le contrôle du FDR à une valeur désirée est le critère principal d'évaluation d'une

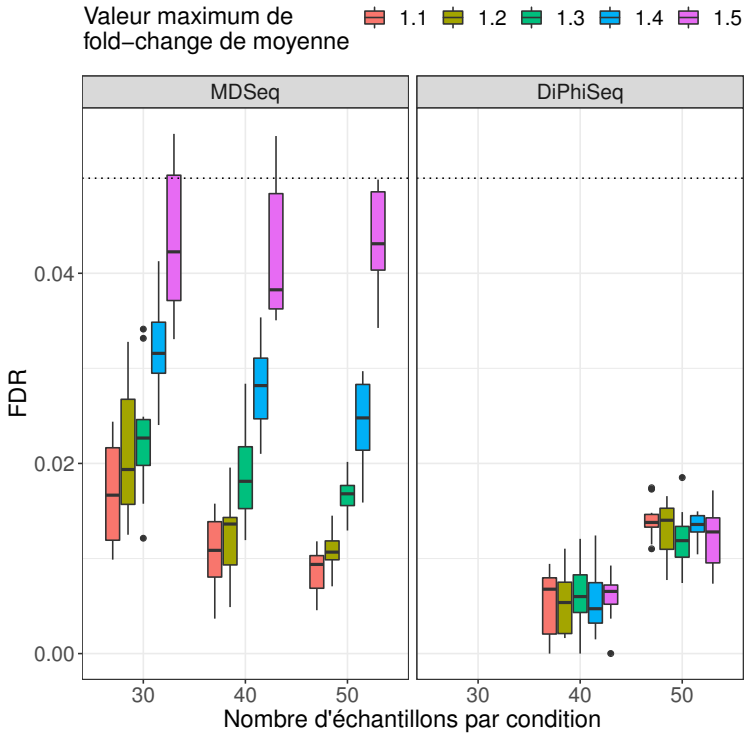

FIGURE 4.7 – FDR obtenus avec MDSeq et DiPhiSeq pour la détection de différence de dispersion avec des jeux de données simulées composés de populations d'échantillons de tailles égales. Les mêmes jeux de données simulées que pour la figure 4.6 ont été utilisés.

méthode. La capacité d'une méthode à détecter un effet lorsqu'il est présent, *i.e.* sa sensibilité, est aussi un critère important. MDSeq et DiPhiSeq sont des méthodes assez peu sensibles pour détecter des différences de dispersion d'expression entre deux populations d'échantillons (figure 4.8). En effet, quels que soient les paramètres considérés ici, la sensibilité est toujours inférieure à 0,35. MDSeq est plus sensible que DiPhiSeq pour des populations composées jusqu'à 40 échantillons. Pour des populations de 30 échantillons, la sensibilité de DiPhiSeq est nulle. Pour cette taille de population d'échantillons, DiPhiSeq n'est en fait pas en mesure de prédire le moindre résultat positif, le FDR étant aussi nul (figure 4.7). En revanche, pour des populations de 50 échantillons, DiPhiSeq a une meilleur sensibilité : environ 0,32 contre 0,27 pour MDSeq. Enfin, la sensibilité de MDSeq augmente légèrement avec les valeurs maximales de *fold-change* de moyenne tolérées alors que DiPhiSeq, de manière similaire avec le FDR, y est insensible.

L'augmentation du FDR et de la sensibilité de MDSeq sont les conséquences de la prise en compte d'un *fold-change* de moyenne dans l'obtention d'une p-valeur significative pour la différence de variance. Elle est bénéfique lorsque la dispersion et la moyenne varient de manière similaire (figure 4.4) en permettant de détecter de faibles valeurs de différence de dispersion. A l'inverse, elle est néfaste lorsqu'un *fold-change* de moyenne permet à lui seul d'obtenir une p-valeur de différence de variance significative.

Les temps de calculs pour l'analyse de ces jeux de données par les deux méthodes

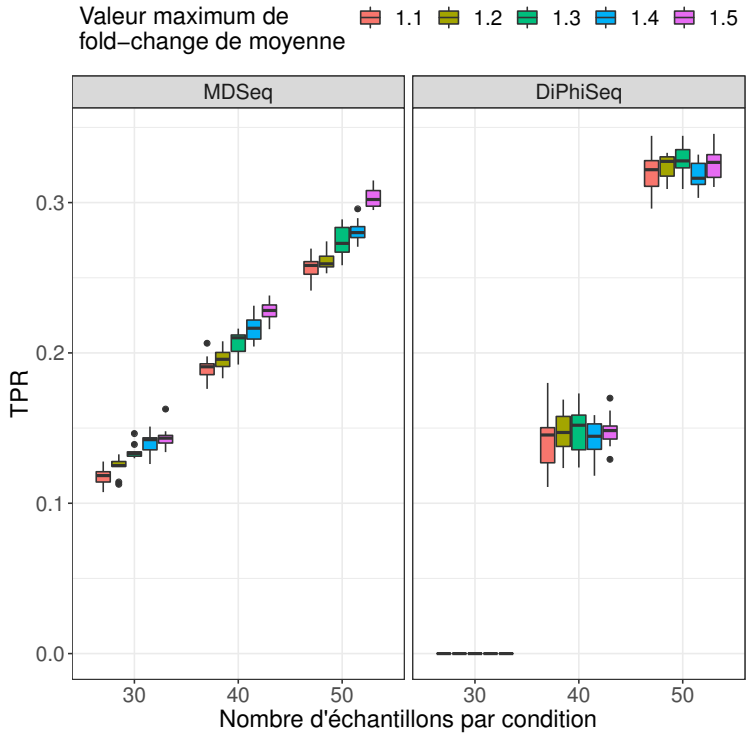

FIGURE 4.8 – Sensibilités, ou TPR, obtenues avec MDSeq et DiPhiSeq pour la détection de différence de dispersion avec des jeux de données simulées composés de populations d'échantillons de tailles égales. Les mêmes jeux de données simulées que pour la figure 4.6 ont été utilisés.

sont précisés dans la table 4.1. Dans leurs implémentations respectives, MDSeq permet l'utilisation de plusieurs CPU pour analyser un jeu de données alors que DiPhiSeq ne le permet pas. Les analyses ont été réalisées à l'aide d'un ordinateur de bureau équipé de 4 CPU Intel(R) Xeon(R) E3-1220 v5 @ 3.00 GHz et de 16 Go de RAM. Les analyses avec la méthode MDSeq ont été réalisées en utilisant 3 CPU. MDSeq

Taille des populations	**MDSeq**	**DiPhiSeq**
30	2	340
40	2,5	450
50	3	560

TABLE 4.1 – Temps de calculs moyens en minutes par jeu de données obtenus avec un ordinateur équipé de 4 CPU Intel(R) Xeon(R) E3-1220 v5 @ 3.00 GHz et de 16 Go de RAM.

est extrêmement plus rapide que DiPhiSeq pour analyser les différences de dispersion entre populations d'échantillons. DiPhiSeq met entre 100 et 500 fois plus de temps. L'implémentation de MDSeq permettant l'utilisation de plusieurs CPU explique en partie cette très grande différence de temps de calculs. La différence de temps de calculs entre les deux méthodes est telle que tous les paramètres n'ont pu être évalués avec DiPhiSeq. Les performances de MDSeq ont ainsi été évaluées avec beaucoup plus de paramétrages des jeux de données simulés que celles de DiPhiSeq.

1.1.4 Approfondissement de l'analyse des performances de MDSeq

Tailles des populations d'échantillons L'augmentation du nombre d'échantillons par condition, de 20 à 100, permet une amélioration des performances de MDSeq (figure 4.9). Pour les jeux de données simulées dont les gènes ont un *fold-change* de

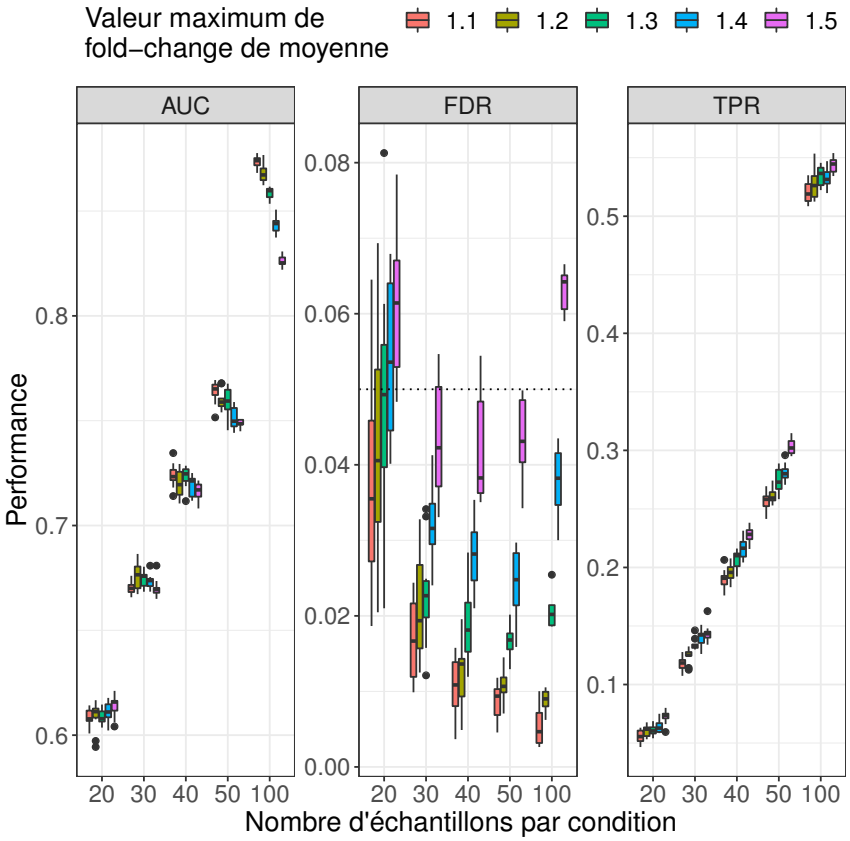

FIGURE 4.9 – Performances obtenues avec MDSeq pour la détection de différence de dispersion avec des jeux de données simulées composés de populations d'échantillons de tailles égales. Paramètres de simulation : 10 000 gènes, 50% de gènes différentiellement dispersés avec une différence de dispersion d'au moins 50%, les gènes non différentiellement dispersés ont la même valeur de dispersion pour les deux populations, les *fold-changes* de moyenne sont limités par différentes valeurs maximales comprises entre 1,1 et 1,5, présence d'un *outlier* pour 10% des gènes. Les performances ont été mesurées à l'aide de 10 réplicats pour chaque jeu de paramètres. La fonction de retrait d'*outliers* de MDSeq a été appliquée. Les gènes différentiellement dispersés ont été identifiés à l'aide d'un seuil de *fold-change* égal à 1.

moyenne inférieur ou égal 1,3, l'AUC augmente ainsi de 0,61 à 0,86, la sensibilité augmente de 0,06 à 0,54 et le FDR diminue de 0,05 à 0,02. Les mêmes tendances d'augmentation du FDR et de la sensibilité en fonction des valeurs maximales de *fold-change* de moyenne qu'observées précédemment sont retrouvées et sont d'autant plus marquées que les populations d'échantillons sont grandes. La perte de contrôle du FDR dépasse le seuil de 0,05 pour des populations de 100 échantillons et avec des gènes dont le *fold-change* de moyenne peut atteindre 1,5. Dans ce cas, il est nécessaire de filtrer ces gènes ayant de trop grandes valeurs de *fold-change* de moyenne pour pouvoir maintenir le FDR de la détection de différence de dispersion inférieur à 0,05.

Populations d'échantillons de tailles inégales L'impact de la différence de taille de population d'échantillons entre les deux conditions sur les performances de la détection de différence de dispersion d'expression a été évaluée avec MDSeq. Pour une taille de population donnée pour la première condition, différents jeux de données ont été simulés avec 1,5 à 10 fois plus d'échantillons pour la seconde condition (voir section 5.3 du chapitre 2). Quelle que soit la taille de la population d'échantillons de la première condition, l'AUC augmente avec l'augmentation de la taille de la population de la seconde condition lorsque celle-ci est limitée à un facteur multiplicatif de 3 (figure 4.10). Des populations encore plus grandes pour la seconde condition n'ap-

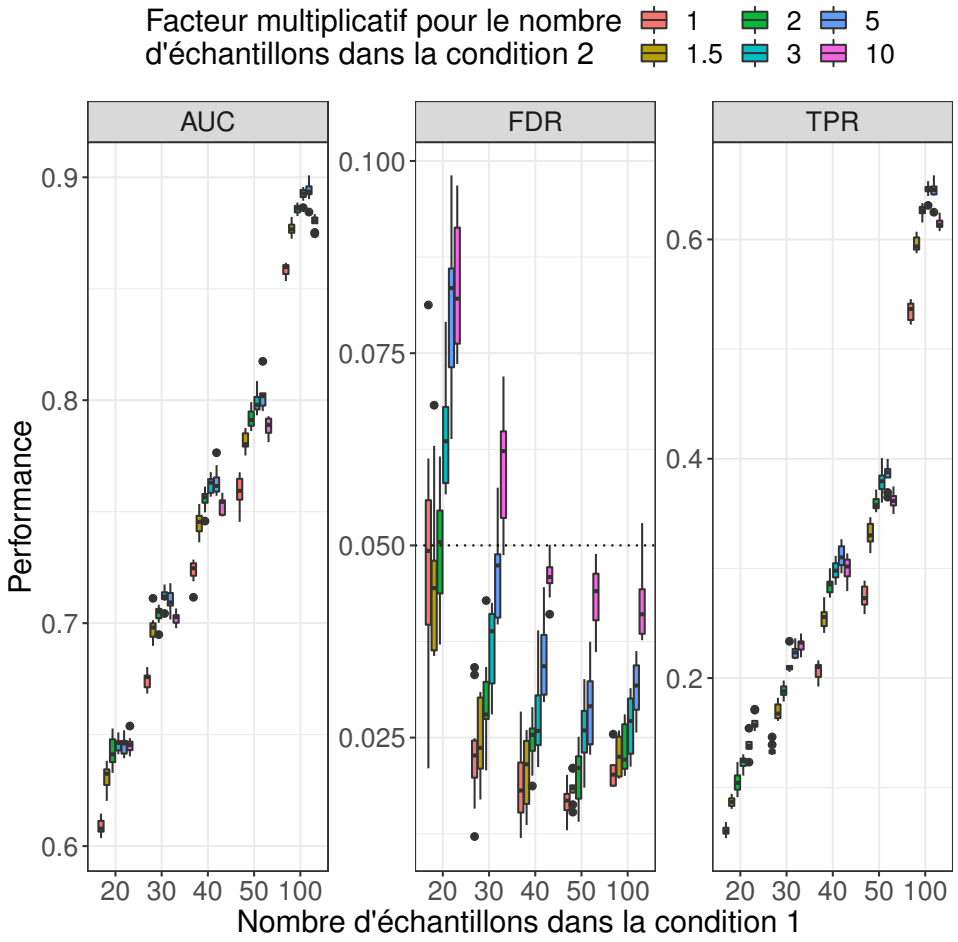

FIGURE 4.10 – Performances obtenues avec MDSeq pour la détection de différence de dispersion avec des jeux de données simulées composés de populations d'échantillons de tailles inégales. La taille de la population de la seconde condition peut prendre différentes valeurs, de 1,5 à 10 fois plus grande que la population de la première condition. Paramètres de simulation : 10 000 gènes, 50% de gènes différentiellement dispersés avec une différence de dispersion d'au moins 50%, les gènes non différentiellement dispersés ont la même valeur de dispersion pour les deux populations, les *fold-changes* de moyenne sont limités par une valeur maximale de 1,3, présence d'un *outlier* pour 10% des gènes. Les performances ont été mesurées à l'aide de 10 réplicats pour chaque jeu de paramètres. La fonction de retrait d'*outliers* de MDSeq a été appliquée. Les gènes différentiellement dispersés ont été identifiés à l'aide d'un seuil de *fold-change* égal à 1.

portent aucune amélioration de l'AUC. Elle régresse même pour les jeux de données constitués d'une population 10 fois plus grande que l'autre population, atteignant des

valeurs comparables à celles obtenues avec des populations pour la seconde condition 1,5 à 2 fois plus grandes que celles pour la première condition.
La sensibilité suit une tendance similaire à celle de l'AUC. Augmenter la taille de la population de la seconde condition permet une augmentation de la sensibilité pour les jeux de données ayant jusqu'à 5 fois plus d'échantillons pour cette condition. Au-delà de cette valeur, l'augmentation de la taille de la population pour la seconde condition n'apporte aucun gain de sensibilité, voire même est néfaste pour les jeux de données dont la population de la première condition est composée d'au moins 40 échantillons. Enfin, l'augmentation de la taille de la population pour la seconde condition entraîne une augmentation du FDR, quelle que soit la taille de la population de la première condition et le facteur multiplicatif d'augmentation de la taille de la population pour la seconde condition. De plus, cette augmentation du FDR est d'autant plus marquée que la population de la première condition est petite. Les valeurs de FDR obtenues avec des populations de tailles égales d'au moins 30 échantillons étant très faibles (autour de 0,02), cette augmentation du FDR n'est pas rédhibitoire dans la mesure où il ne dépasse le seuil de 0,05 uniquement pour quelques cas, *e.g.* les jeux de données composés de populations de 30 et 300 échantillons et quelques jeux de données composés de populations de 100 et 1000 échantillons.
Les mêmes tendances en fonction de l'augmentation de la taille de la population d'échantillons de la deuxième condition sont observées pour les autres valeurs maximales tolérées de *fold-change* de moyenne d'expression comprises entre 1,1 et 1,5 (figures A.1 à A.4 en annexes).

L'augmentation de la taille de la population d'échantillons pour la seconde condition permet une meilleure détection des gènes différentiellement dispersés dont la dispersion d'expression est plus grande dans la plus grande des deux populations d'échantillons (figure 4.11). Le jeu de données simulées composé de populations de 50 et de 500 échantillons dont les *fold-changes* réels et estimés de dispersion sont représentés sur la figure 4.11 est issu du même réplicat que le jeu de données composé de populations de tailles égales à 50 échantillons dont les mêmes *fold-changes* sont représentés sur la figure 4.5. Ainsi, les *fold-changes* réels de moyenne et de dispersion sont les mêmes pour ces deux jeux de données (voir section 5.4.4 du chapitre 2). En revanche, une très nette tendance à sur-estimer les *fold-changes* de dispersion reflétant une augmentation de la dispersion dans la population de 500 échantillons peut être observée (*fold-changes* de dispersion positifs sur la figure 4.11). L'augmentation de la taille de la population de la seconde condition introduit donc un biais dans la détection de différence de dispersion. Les gènes différentiellement dispersés dont la dispersion est plus grande dans la plus grande des populations ont ainsi tendance à être plus facilement détectés que ceux dont la dispersion diminue dans la plus grande des populations.

Les performances de détection de différence de dispersion d'expression par MDSeq sont donc influencées par l'amplitude des *fold-changes* de moyenne présents ainsi que par la différence de taille des populations d'échantillons considérées. Parmi les indicateurs de performances calculés, le FDR est d'intérêt principal puisqu'il reflète le taux d'erreurs parmi les résultats positifs prédits par une méthode. Il est ainsi en général préférable de contrôler le FDR à une valeur basse au détriment d'une bonne valeur de sensibilité plutôt que l'inverse. Dans cette mesure, la figure 4.12 récapitule les valeurs maximales de *fold-change* de moyenne permettant d'obtenir un FDR inférieur à 0,05 dans l'ensemble des réplicats de simulation en fonction de l'ensemble des tailles de population d'échantillons considérées.

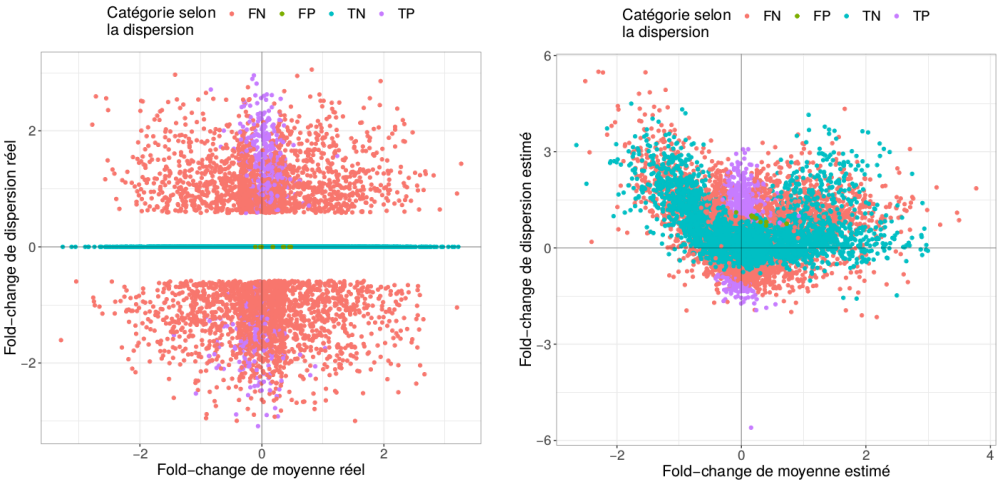

FIGURE 4.11 – Valeurs réelles de *fold-changes* de moyenne et de dispersion d'un jeu de données simulées composé de populations de 50 échantillons pour la première condition et de 500 échantillons pour la seconde condition (les mêmes valeurs que pour le jeu de données de la figure 4.5, à gauche). Valeurs de *fold-changes* de moyenne et de dispersion estimées par MDSeq sur ce même jeu de données simulées (à droite). Les couleurs des points correspondent aux résultats de test de différence de variance réalisés par MDSeq avec un *fold-change* de 1.

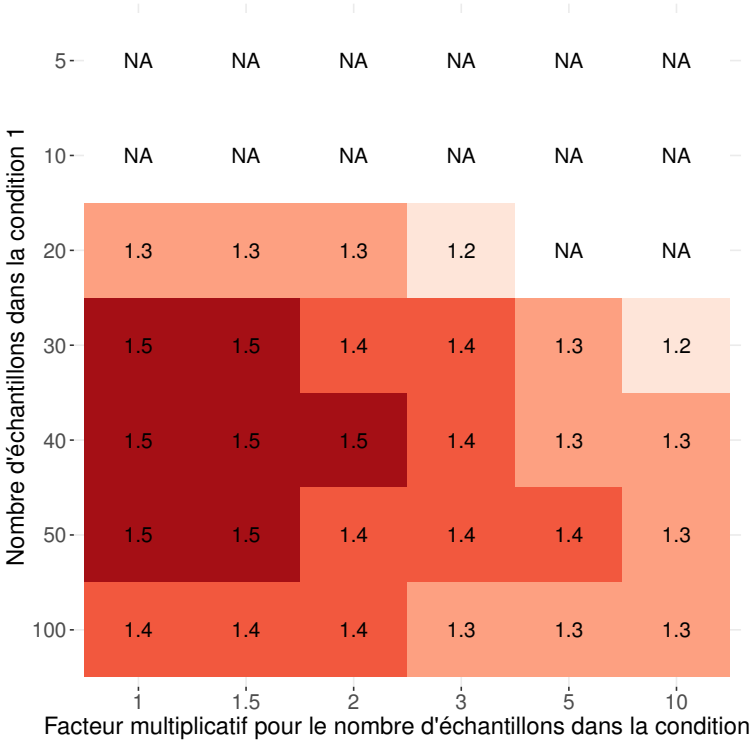

FIGURE 4.12 – Valeurs maximales de *fold-change* de moyenne permettant d'obtenir un FDR pour la détection de différence de dispersion inférieur à 0,05 dans l'ensemble des réplicats de simulation en fonction de l'ensemble des tailles de population d'échantillons considérées. Une valeur « NA » indique qu'il n'est pas possible d'obtenir un FDR inférieur à 0,05 avec les tailles de population d'échantillons correspondantes, quelle que soit la valeur maximale de *fold-change* de moyenne.

Filtrage des gènes différentiellement exprimés Pour éviter l'augmentation du FDR due à la présence d'un *fold-change* de moyenne, les gènes différentiellement exprimés identifiés à l'aide des tests de différence de moyenne peuvent être retirés de l'analyse de détection de différence de dispersion. Pour ce faire, des tests portant sur la moyenne avec une valeur de *fold-change* égale à 1 (voir section 3.4.1 du chapitre 2) ont été réalisés. Les gènes différentiellement dispersés sont alors identifiés par une p-valeur non significative pour le test de différence de moyenne et une p-valeur significative pour le test de différence de variance. Retirer les gènes différentiellement exprimés permet effectivement de contrôler le FDR (figure 4.13), aucune augmentation du FDR n'est observée en fonction des valeurs de maximales de *fold-change* de moyenne telle qu'observée précédemment (figure 4.9). En revanche, le filtrage des

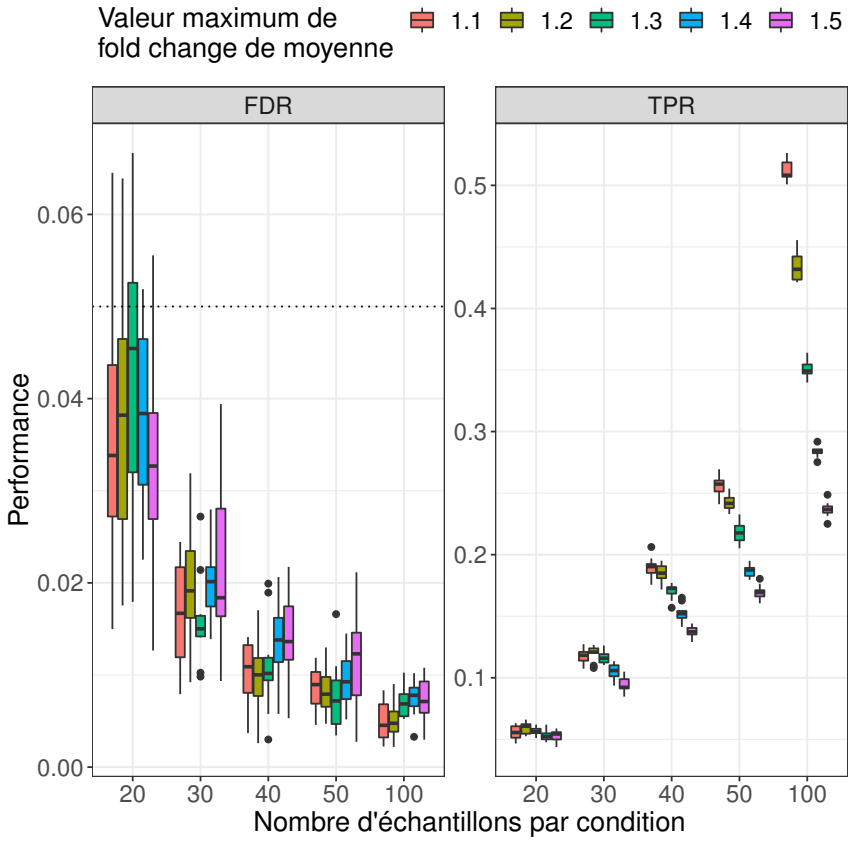

FIGURE 4.13 – Performances obtenues avec MDSeq pour la détection de différence de dispersion avec des jeux de données simulées composés de populations d'échantillons de tailles égales. Paramètres de simulation : 10 000 gènes, 50% de gènes différentiellement dispersés avec une différence de dispersion d'au moins 50%, les gènes non différentiellement dispersés ont la même valeur de dispersion pour les deux populations, les *fold-changes* de moyenne sont limités par différentes valeurs maximales comprises entre 1,1 et 1,5, présence d'un *outlier* pour 10% des gènes. Les performances ont été mesurées à l'aide de 10 réplicats pour chaque jeu de paramètres. La fonction de retrait d'*outliers* de MDSeq a été appliquée. Les gènes différentiellement dispersés ont été identifiés par une p-valeur supérieure à 0,05 pour le test de différence de moyenne et une p-valeur inférieure à 0,05 pour le test de différence de variance avec des valeurs de seuil de *fold-change* égales à 1.

gènes présentant une différence de moyenne d'expression a pour conséquence de faire fortement chuter la sensibilité pour les jeux de données composés d'un grand nombre d'échantillons et où les *fold-changes* de moyenne peuvent prendre des valeurs jusqu'à

1,5. Cela s'explique par le fait que des grandes populations d'échantillons permettent de détecter des gènes ayant un faible *fold-change* de moyenne comme différentiellement exprimés à l'aide de tests visant à détecter toute différence de moyenne, *i.e.* le *fold-change* de moyenne testé est de 1. Pour éviter cela, MDSeq offre la possibilité d'utiliser des valeurs plus élevées de *fold-change* de moyenne pour identifier des gènes différentiellement exprimés (voir section 3.4.1 du chapitre 2).

Le filtrage des gènes différentiellement exprimés permet d'évaluer les performances de MDSeq sur des jeux de données simulées plus réalistes contenant des gènes pouvant présenter de grandes différences de moyenne d'expression entre les deux populations d'échantillons. Les jeux de données simulées utilisés sont ainsi ceux où aucune contrainte n'a été mise sur les *fold-changes* de moyennes en suivant le second scénario de simulation de jeux de données exposé dans la section 5.4.2 du chapitre 2. Précédemment, les valeurs maximales de *fold-change* de moyenne permettant de contrôler le FDR pour la détection de différence de dispersion inférieur à 0,05 à l'aide du seul test de différence de variance ont été déterminées en fonction du nombre d'échantillons dans les deux populations considérées (figure 4.12). Ici, les jeux données simulées utilisés sont ceux dont les valeurs de *fold-change* séparant les gènes différentiellement exprimés et les gènes considérés comme non différentiellement exprimés correspondent à ces valeurs pour chaque paire de tailles de population d'échantillons considérée. Un ensemble de valeurs de seuils de *fold-change* de moyenne, comprises entre 1 et ces valeurs, a été utilisé pour filtrer les gènes différentiellement exprimés. La sensibilité est ainsi nettement améliorée en utilisant des valeurs élevées de *fold-change* pour les tests de détection de différence de moyenne pour filtrer les gènes différentiellement exprimés (figure 4.14). Pour des populations de 50 échantillons, pour lesquelles le FDR pour la détection de différence de dispersion peut être maintenu inférieur à 0,05 avec des gènes dont le *fold-change* de moyenne peut atteindre 1,5, faire varier le seuil de *fold-change* de 1 à 1,5 pour le filtrage des gènes différentiellement exprimés permet d'augmenter la sensibilité de 0,26 à 0,33. En revanche, cette augmentation du seuil a aussi pour conséquence une augmentation du FDR au-delà de 0,05. Dans ce cas, un seuil de 1,3 doit être utilisé pour maintenir le FDR inférieur de 0,05 tout en permettant une augmentation de la sensibilité à 0,31, correspondant à un doublement du nombre de vrais positifs (passant de 390 à 790 en moyenne). La valeur de seuil de *fold-change* à utiliser pour filtrer les gènes différentiellement exprimés permettant une augmentation de la sensibilité tout en maintenant le FDR inférieur à 0,05 a ainsi été déterminée pour l'ensemble des tailles de population considérées (figure 4.15).

1.2 Discussion

Cette étude de simulation a permis d'établir de manière rigoureuse les conditions dans lesquelles des différences de dispersion peuvent être détectées dans des données d'expression issues du RNA-seq. En contrôlant différents paramètres lors de la simulation de jeux de données, les principales caractéristiques des méthodes MDSeq et DiPhiSeq, les seules à ce jour permettant d'identifier des gènes différentiellement dispersés à l'aide de données RNA-seq, ont pu être établies.

1.2.1 Qualité des jeux de données simulés

En plus du réalisme des données simulées abordé dans la section 5.6 du chapitre 2, la qualité d'une étude de simulation dépend également de la fidélité avec laquelle elle

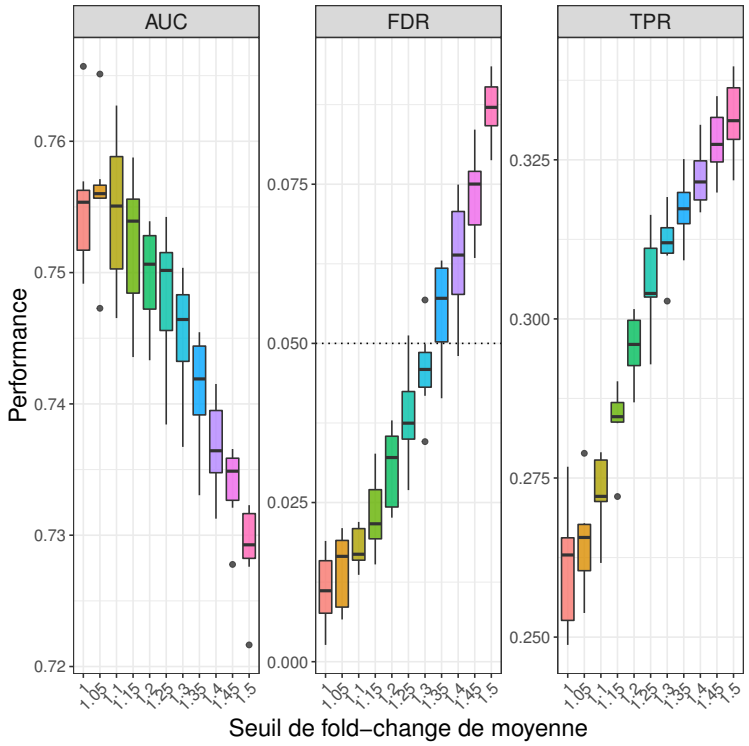

FIGURE 4.14 – Performances obtenues avec MDSeq pour la détection de différence de dispersion avec des jeux de données simulées composés de deux populations de 50 échantillons. Paramètres de simulation : 10 000 gènes, 50% de gènes différentiellement dispersés avec une différence de dispersion d'au moins 50%, les gènes non différentiellement dispersés ont la même valeur de dispersion pour les deux populations, une valeur de *fold-change* de moyenne de 1,5 définit la séparation entre gènes différentiellement exprimés et non différentiellement exprimés, présence d'un *outlier* pour 10% des gènes. Les performances ont été mesurées à l'aide de 10 réplicats pour chaque jeu de paramètres. La fonction de retrait d'*outliers* de MDSeq a été appliquée. Les gènes différentiellement exprimés ont été filtrés selon différentes valeurs de seuil de *fold-change* de moyenne. Les gènes différentiellement dispersés ont été identifiés à l'aide d'un seuil de *fold-change* égal à 1.

reflète les principales caractéristiques des jeux de données réels sur lesquels les méthodes évaluées devront être appliquées. Dans le cadre de cette étude de simulation, les *fold-changes* de moyenne et de dispersion ont été introduits de manière à ce que l'expression de 50% des gènes présente au moins l'un des deux types de *fold-change*. La plupart des études visant à évaluer les performances de méthodes de détection de gènes différentiellement exprimés à l'aide de données simulées fixent ce pourcentage entre 10 et 40%. Introduire un *fold-change* de moyenne d'expression à 50% des gènes peut donc paraître peu réaliste. Cependant, le seuil définissant la limite entre gènes différentiellement exprimés et gènes non différentiellement exprimés peut prendre des valeurs relativement basses, comprises entre 1,1 et 1,5. Par conséquent, des gènes considérés comme différentiellement exprimés peuvent avoir de faibles *fold-changes* de moyenne, *e.g.* 1,2 pour des jeux de données simulées avec une valeur de seuil de *fold-change* de moyenne fixée à 1,1. Ces faibles valeurs de *fold-change* de moyenne parmi les gènes différentiellement exprimés contrebalancent ainsi la proportion de gènes différentiellement exprimés dans les jeux de données simulées.

La plupart des études visant à évaluer les performances de méthodes de détection

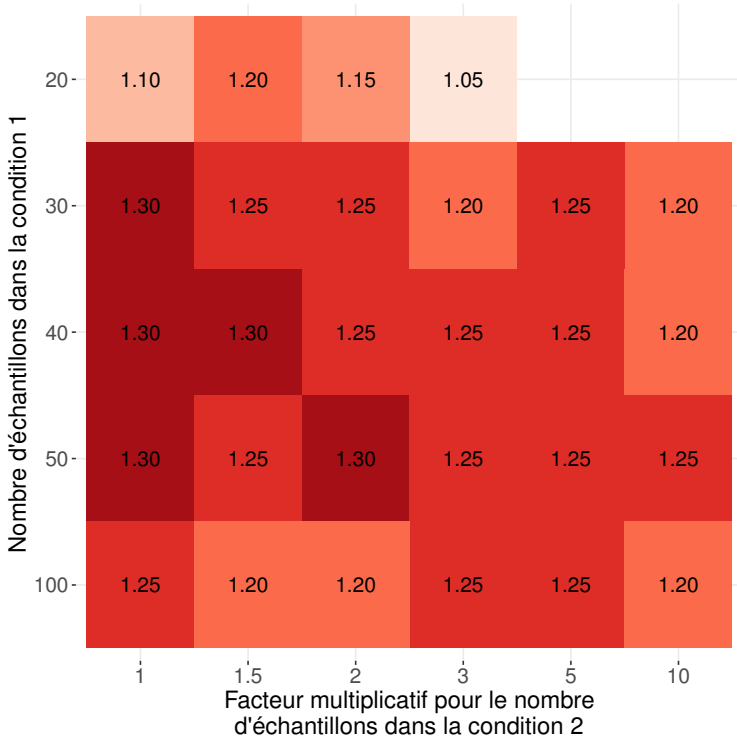

FIGURE 4.15 – Valeurs de seuil de *fold-change* de moyenne à utiliser pour filtrer les gènes différentiellement exprimés permettant d'obtenir un FDR pour la détection de différence de dispersion inférieur à 0,05 en fonction de l'ensemble des tailles de population d'échantillons considérées. Les cases vides sont dues au fait qu'il n'est pas possible d'obtenir un FDR pour la détection de différence de dispersion inférieur à 0,05 avec les tailles de population d'échantillons correspondantes, quelle que soit la valeur de seuil de *fold-change* de moyenne (voir figure 4.12).

de gènes différentiellement exprimés ne prennent pas en compte la possibilité que l'expression des gènes puissent avoir des différences de dispersion entre les populations d'échantillons comparées. Seule l'étude de YU, FERNANDEZ et BROCK, 2017 prend en compte cette possibilité en envisageant deux scénarios de simulation : avec ou sans différence de dispersion entre les deux conditions. Dans le scénario envisageant une différence de dispersion, les nombres de *reads* de l'intégralité des gènes présentent un *fold-change* de dispersion égal à 1,5 entre les deux conditions. Ainsi, peu de repères existent pour guider l'introduction de *fold-changes* de dispersion dans le cadre d'une étude de simulation. Ici, le principal objectif est d'appliquer les méthodes de détection de différence de dispersion à des jeux de données d'expression permettant la comparaison entre des échantillons tumoraux et des échantillons sains. Le développement cancéreux se caractérise par la dérégulation d'un grand nombre de gènes. Dans ce contexte, l'introduction de *fold-changes* de dispersion à 50% des gènes dans les jeux de données simulées apparaît pertinente.

Pour se conformer aux pourcentages de gènes différentiellement exprimés utilisés par la plupart des études de simulation et abaisser le nombre de gènes différentiellement dispersés, des jeux de données ont été simulés avec 30% de gènes présentant un *fold-change* de moyenne ou de dispersion. Les gènes non différentiellement dispersés étant plus nombreux et MDSeq étant assez peu sensible pour détecter des différences de dispersion, en particulier avec de petites populations d'échantillons, les faux positifs

sont plus nombreux et ont une plus grande influence sur le FDR. Ainsi, les valeurs maximales de *fold-change* de moyenne permettant de maintenir le FDR de la détection de différence de dispersion (figure 4.16) sont plus basses que celles obtenues avec les jeux de données simulées avec 50% de gènes différentiellement dispersés (figure 4.12). Les valeurs maximales de seuil pour le filtrage des gènes différentiellement exprimés à

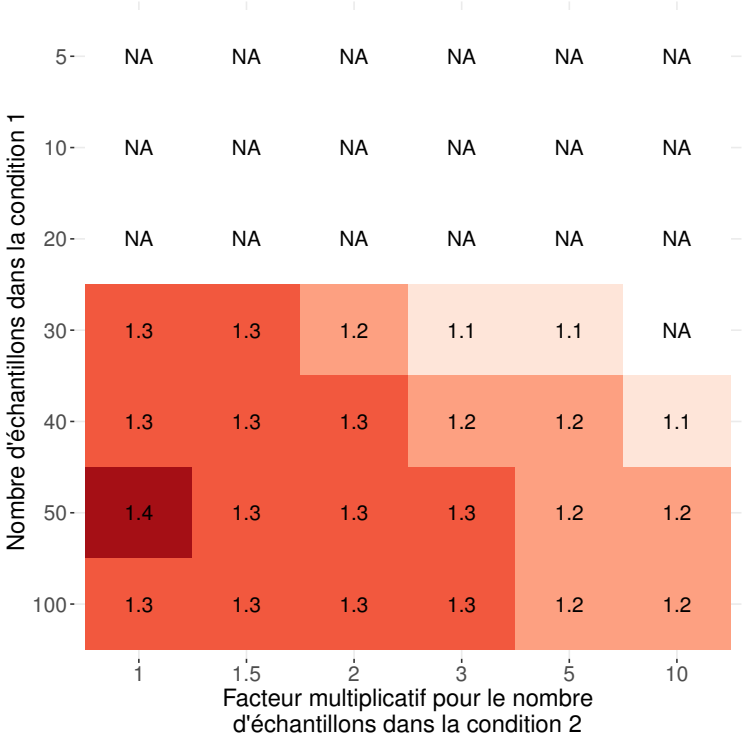

FIGURE 4.16 – Valeurs maximales de *fold-change* de moyenne permettant d'obtenir un FDR pour la détection de différence de dispersion inférieur à 0,05 dans l'ensemble des réplicats de simulation en fonction de l'ensemble des tailles de population d'échantillons considérées. Les jeux de données sont constitués de 30% de gènes différentiellement dispersés. Une valeur « NA » indique qu'il n'est pas possible d'obtenir un FDR inférieur à 0,05 avec les tailles de population d'échantillons correspondantes, quelle que soit la valeur maximale de *fold-change* de moyenne.

utiliser pour maintenir le FDR pour la détection de différence de dispersion inférieur à 0,05 suivent logiquement la même tendance (données non montrées). Ce sont ces valeurs de seuil de *fold-change* de moyenne qui devront être utilisées avec MDSeq pour la détection de différence de dispersion lorsque les gènes différentiellement dispersés sont supposés être présents en faible nombre dans le cadre de la comparaison de deux populations d'échantillons d'intérêt.

1.2.2 Caractéristiques de MDSeq et DiPhiSeq

Principaux résultats L'étude de simulation a permis d'identifier les caractéristiques des méthodes MDSeq et DiPhiSeq pour la détection de différence de dispersion dans des données RNA-seq. Dans le cadre de modèles basés sur la distribution binomiale négative, ce problème s'avère être bien plus complexe que la détection de différence de moyenne. Pour y parvenir, beaucoup plus d'échantillons sont en effet nécessaires. Des populations d'au moins 20 échantillons sont un minimum pour pouvoir détecter des différences de dispersion en maintenant le FDR inférieur à 0,05 avec MDSeq. DiPhiSeq est une méthode moins sensible et requiert au moins 40 échantillons

par population pour pouvoir détecter des gènes différentiellement dispersés selon les tailles de population évaluées dans cette étude. Pour des populations plus petites, cette méthode est en effet incapable de détecter le moindre résultat positif. Le jeu de données utilisé par les auteurs de DiPhiSeq pour appliquer leur méthode dans leur article (Li et Lamere, 2018) est constitué de deux populations de 35 échantillons. En accord avec les résultats de cette étude de simulation, cette valeur semble être le nombre minimal d'échantillons par population permettant à DiPhiSeq de détecter des gènes différentiellement dispersés.

En ce qui concerne les gènes faiblement différentiellement exprimés, les deux méthodes ont des performances globalement similaires. Des populations composées d'au moins 30 échantillons sont ainsi nécessaires pour pouvoir détecter des différences de dispersion. Cependant, la sensibilité à détecter des différences de dispersion d'au moins 50%, telle qu'elle a été introduite dans les jeux de données simulées dans le cadre de cette étude (voir section 5.4.1 du chapitre 2), est très basse (autour de 0,15) avec ces tailles de population. L'amélioration de la sensibilité requiert de plus grandes populations d'échantillons composées d'au moins 40 ou 50 échantillons (figure 4.9). Toutefois, les méthodes MDSeq et DiPhiSeq disposent chacune de caractéristiques propres. MDSeq est une méthode plus sensible mais aussi moins spécifique que DiPhiSeq. En effet, plus de vrais positifs sont détectés par MDSeq mais au prix d'une présence plus importante de faux positifs sous l'influence de la présence de *fold-changes* de moyenne d'expression, y compris à des valeurs assez faibles. DiPhiSeq se distingue par sa capacité à contrôler le FDR à des taux très bas. En revanche, sa faible sensibilité, en particulier avec de petites populations d'échantillons, et des temps de calculs énormément plus longs que MDSeq ne permettent pas à DiPhiSeq d'être la méthode de choix pour l'identification de différence de dispersion d'expression.

L'évaluation des performances de MDSeq a donc été approfondie en testant de nombreuses autres tailles de population d'échantillons, en particulier des jeux de données composés de populations d'échantillons de tailles inégales. Les valeurs maximales de *fold-change* moyenne permettant de maintenir le FDR pour la détection de différence de dispersion inférieur à 0,05 ont été déterminées en fonction de la taille des populations d'échantillons. De la même manière, les valeurs de seuil à utiliser pour filtrer les gènes différentiellement exprimés en accord avec ces valeurs maximales de *fold-change* de moyenne ont été identifiées. Enfin, les conséquences de l'augmentation de la taille de l'une des deux populations sur les performances d'identification des gènes différentiellement dispersés ont été évaluées. Celui-ci n'apporte qu'un gain limité de sensibilité en maintenant le FDR inférieur à 0,05. De plus, un biais en faveur de l'identification de gènes différentiellement dispersés dont la dispersion est plus grande dans la plus grande des deux populations d'échantillons est introduit.

Identification de différence de dispersion Les méthodes MDSeq et DiPhiSeq sont les deux seules méthodes basées sur la distribution binomiale négative permettant d'identifier des différences de dispersion dans les données d'expression issues du RNA-seq. Pour y parvenir, elles emploient des approches différentes.

La méthode MDSeq étend les GLM traditionnellement utilisés pour modéliser l'espérance d'une variable aléatoire en fonction de variables explicatives à la variance de celle-ci. Les tests implémentés par cette méthode ne portent donc pas directement sur le paramètre de dispersion mais sur l'espérance, qui est égale à la moyenne, et la variance de la variable aléatoire. Cependant, des différences de dispersion peuvent

tout de même être identifiées parmi les gènes non différentiellement exprimés. En effet, une différence de variance dans l'expression de ces gènes est nécessairement due à une différence de dispersion. De plus, un *fold-change* de dispersion peut être calculé à partir des *fold-changes* de moyenne et de variance estimés grâce la reparamétrisation de la distribution binomiale négative (voir section 3.4.1 du chapitre 2). Cette valeur de *fold-change* de dispersion calculée pourrait être utilisée pour identifier des différences de dispersion d'expression parmi les gènes différentiellement exprimés en la comparant à un seuil. La valeur minimale de *fold-change* de dispersion parmi les gènes différentiellement variants et non différentiellement exprimés et la moyenne des 50 valeurs les plus faibles de ces *fold-changes* de dispersion observés parmi cet ensemble de gènes ont été utilisés pour identifier des différences de dispersion parmi les gènes différentiellement exprimés (figure 4.17). L'utilisation d'une valeur suffisamment élevée de *fold-change* de dispersion semble permettre de détecter des gènes différentiellement dispersés parmi les gènes différentiellement exprimés en contrôlant le FDR. En effet, l'extension de la détection de différence de dispersion à l'ensemble du jeu de données permet d'augmenter le nombre de vrais positifs au prix d'une faible augmentation du nombre de faux positifs. En revanche, l'utilisation d'une valeur trop faible de *fold-change* de dispersion engendre un trop grand nombre de faux positifs.
Les mêmes valeurs de seuil de *fold-change* de dispersion ont été appliquées à un jeu de données composé de populations de 50 et 500 échantillons. Avec l'augmentation de la taille de la population de la seconde condition, le biais introduit vers la détection de gènes dont la dispersion est plus grande dans la plus grande des deux populations (voir section 1.1.4) nécessite des valeurs plus élevées de *fold-change* de dispersion pour maintenir le nombre de faux positifs au même niveau que celui obtenu avec des populations de tailles égales.
L'utilisation d'un seuil de *fold-change* de dispersion pour détecter des gènes différentiellement dispersés parmi les gènes différentiellement exprimés semble être une piste intéressante mais nécessite des développements méthodologiques pour identifier les valeurs de seuil à utiliser pour contrôler le FDR en fonction des tailles de population qui constituent le jeu de données. En l'état, l'utilisation de MDSeq pour l'identification de gènes différentiellement dispersés doit être limitée aux gènes non différentiellement exprimés pour pouvoir bénéficier du cadre rigoureux des tests statistiques portant sur la moyenne et la variance. L'intérêt principal de la détection de gènes différentiellement dispersés est d'identifier des gènes qui ne sont pas détectés par l'approche classique de différence de moyenne ou qui ne figurent pas parmi les gènes présentant les plus grands *fold-changes* de moyenne. Cette limite de l'emploi de MDSeq aux gènes non différentiellement exprimés ne constitue donc pas un problème majeur.

Au contraire de MDSeq, le test implémenté par DiPhiSeq porte directement sur le paramètre de dispersion. Les gènes présentant des différences de dispersion dans leur expression peuvent ainsi être identifiés à l'aide de cette méthode quelle que soit la différence de moyenne d'expression qu'ils présentent par ailleurs. Au-delà de l'aspect rigoureux lié au fait que les tests statistiques développés par DiPhiSeq portent directement sur le paramètre d'intérêt, DiPhiSeq présente l'avantage de pouvoir identifier des différences de dispersion dans l'expression de l'ensemble des gènes du jeu de données, qu'ils soient différentiellement exprimés ou non. Bien que l'intérêt principal de l'approche consiste à identifier des gènes différentiellement dispersés parmi les gènes non différentiellement exprimés, il peut aussi être intéressant biologiquement d'identifier des changements de dispersion dans l'expression de gènes présentant un *fold-change* de moyenne.

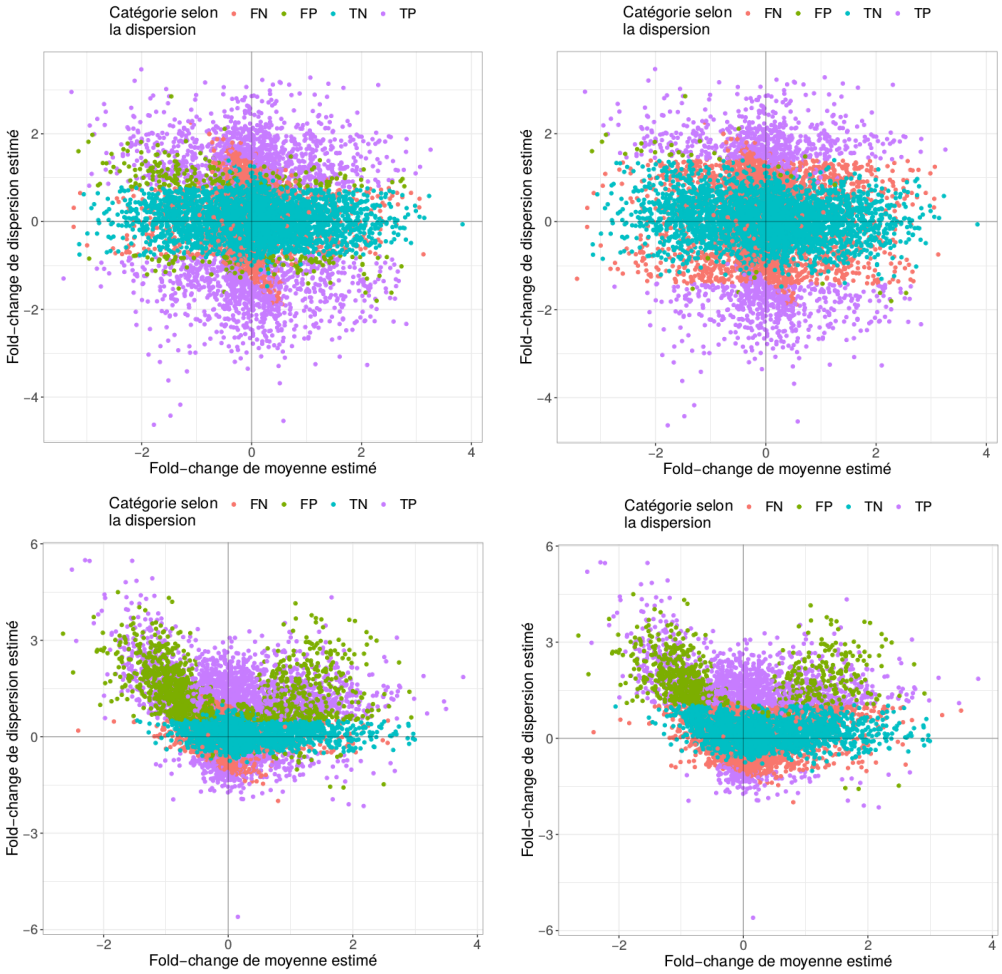

FIGURE 4.17 – Valeurs de *fold-changes* de moyenne et de dispersion estimées par MDSeq pour un jeu de données simulées composé de 50 échantillons par population, le même que pour la figure 4.5 (en haut). Les couleurs des points correspondent aux résultats de tests de différence de moyenne et de variance réalisés par MDSeq avec un seuil de *fold-change* de 1 pour les gènes non différentiellement exprimés. Pour les gènes différentiellement exprimés, les gènes différentiellement dispersés sont identifiés à l'aide d'un seuil de *fold-change* de dispersion fixé à la valeur minimale de *fold-change* de dispersion parmi les gènes différentiellement variants et non différentiellement exprimés (en haut à gauche) ou de la moyenne des 50 plus faibles valeurs de *fold-change* de dispersion parmi les gènes différentiellement variants et non différentiellement exprimés (en haut à droite). Graphiques équivalents avec des populations de 50 et 500 échantillons au bas de la figure, utilisation du même jeu de données simulées que pour la figure 4.11 (en bas).

Temps de calculs Les méthodes d'optimisation et d'implémentation employées par MDSeq lui permettent d'avoir des temps de calculs beaucoup plus rapides que DiPhiSeq. Avec l'établissement de bases d'échantillons RNA-seq de plus en plus importantes, le développement de méthodes d'intégration d'échantillons RNA-seq issus de sources hétérogènes et l'obtention de profondeur de séquençage de plus en plus importantes, le temps de calculs est un aspect important d'évaluation de méthodes d'analyse de ce type de données. Les temps extrêmement longs de calculs de DiPhiSeq observés avec les jeux de données simulées (voire table 4.1) constituent un frein important à son application. A minima, des méthodes d'optimisation de l'implémentation, comme la possibilité d'utiliser plusieurs processeurs, devront être apportées à

DiPhiSeq pour lui permettre d'être une méthode facilement applicable à de grands jeux de données. En revanche, des méthodes d'optimisation statistiques, au prix d'une perte de robustesse, qui est son principal avantage, ne devraient pas être apportées.

***Outliers* et modèles linéaires généralisés** Dans le cadre d'une modélisation des nombres de *reads* par une distribution binomiale négative, les modèles linéaires généralisés permettent la prise en compte de schémas expérimentaux élaborés incluant, en plus de l'effet à observer, d'autres effets d'intérêt qu'ils soient biologiques ou techniques. Les effets *batch* sont un biais technique bien connu dans les données RNA-seq, se caractérisant par la présence de valeurs aberrantes ou *outliers*. Ils doivent être pris en compte pour ne pas interférer avec les effets biologiques que l'on souhaite observer. Il est ainsi fortement recommandé de les inclure comme un facteur bloquant dans les matrices de design des GLM (voir section 3.2.2 du chapitre 2). C'est ainsi que les principales méthodes de détection de *fold-change* de moyenne ont opté pour ces modèles en remplacement de tests exacts plus basiques (Robinson, McCarthy et Smyth, 2010, Love, Huber et Anders, 2014). Dans ce contexte, il est appréciable que les auteurs de MDSeq aient aussi opté pour cette modélisation pour la détection de différence à la fois de moyenne et de variance d'expression, au prix d'une reparamétrisation de la distribution binomiale négative (voir section 3.4.1 du chapitre 2). DiPhiSeq, de son côté, n'emploie pas ce type de modèle. Il prend en compte tout de même la présence d'*outliers* à l'aide d'une fonction de Tukey à deux poids (voir section 3.4.2 du chapitre 2). En appliquant un poids de 0 à ces *outliers*, cette caractéristique permet à DiPhiSeq d'annihiler les effets des *outliers* et ainsi d'atteindre un niveau élevé de spécificité. Cette caractéristique peut aussi expliquer sa faible sensibilité. De plus, aucune distinction ne peut être faite entre un *outlier* dû à un effet *batch* et un *outlier* d'origine biologique. La prise en compte des *outliers* par DiPhiSeq ne semble donc pas optimale et l'impossibilité de prendre en compte d'autres variables que l'effet biologique d'intérêt est préjudiciable.

Dans cette étude de simulation, les *outliers* ont été introduits de telle sorte qu'un seul échantillon était affecté par gène. Dans le cas de populations d'échantillons assez petites, comme c'est le cas dans l'étude de Bonafede et al., 2016 où des populations de 3 à 10 échantillons ont été simulées, l'impact de l'introduction d'*outliers* est significatif sur les performances des méthodes classiques de différence de moyenne d'expression. Ici, les populations d'échantillons sont beaucoup plus grandes que celles considérées par la plupart des études utilisant des données simulées. L'influence de ces *outliers* y est ainsi supposée moindre. Il n'est toutefois pas négligeable pour la détection de différence de dispersion comme l'ont montré les auteurs de DiPhiSeq (voir figure 1 de l'article de DiPhiSeq (Li et Lamere, 2018)). Des jeux de données ont été simulés avec l'introduction aléatoire d'*outliers* en utilisant l'option « *random* » et une probabilité de 0,0005 (voir section 5.3 du chapitre 2). Au contraire de l'option « *single* », les *outliers* sont introduits de manière proportionnelle à la taille des populations d'échantillons simulées. Ainsi, l'influence de la présence des *outliers* est la même quelle que soit la taille des populations des jeux de données. Des performances similaires à celles obtenues avec l'utilisation de l'option « *single* » (figure 4.12) ont été observées avec l'usage de cette option, indiquant que l'introduction d'*outliers* pour les plus grandes populations d'échantillons n'impactent pas significativement les performances de MDSeq pour la détection de différence de dispersion d'expression (données non montrées).

2 Application aux données TCGA

Les données RNA-seq TCGA (voir section 2.2 du chapitre 1) ont été utilisées pour identifier des gènes différentiellement dispersés entre échantillons tumoraux et échantillons sains.

2.1 Résultats

2.1.1 Sélection des jeux de données

L'étude de simulation a permis de mettre en évidence que des populations d'au moins 20 échantillons étaient nécessaires pour l'identification de différence de dispersion d'expression à l'aide de MDSeq. Pour permettre un meilleur contrôle du FDR à une valeur inférieure à 0,05 ainsi qu'une meilleure sensibilité, des populations plus grandes, d'au moins 30 échantillons, sont nécessaires. TCGA fournit des données pour beaucoup plus d'échantillons tumoraux que d'échantillons sains pour l'ensemble des cancers. Ainsi, seuls les jeux de données pour lesquels des données RNA-seq pour au moins 30 échantillons sains sont disponibles ont été retenus pour cette analyse et sont listés dans la table 4.2.

Jeux de données	**miARN**		**ARNm**	
	Echantillons sains	Echantillons tumoraux	Echantillons sains	Echantillons tumoraux
TCGA-BRCA	104	1096	113	1102
TCGA-COAD	8	455	41	478
TCGA-HNSC	44	523	44	500
TCGA-KIRC	71	544	72	538
TCGA-KIRP	34	291	32	288
TCGA-LIHC	50	372	50	371
TCGA-LUAD	46	519	59	533
TCGA-LUSC	45	478	49	502
TCGA-PRAD	52	498	52	498
TCGA-STAD	45	446	32	375
TCGA-THCA	59	506	58	502
TCGA-UCEC	33	545	35	551

TABLE 4.2 – Nombres d'échantillons sains et tumoraux des jeux de données d'expression de miARN et d'ARNm du TCGA pour lesquels au moins 30 échantillons sains sont disponibles. BRCA : *BReast invasive CArcinoma*, COAD : *COlon ADenocarcinoma*, HNSC : *Head and Neck Squamous cell Carcinoma*, KIRC : *KIdney Renal Clear cell carcinoma*, KIRP : *KIdney Renal Papillary cell carcinoma*, LIHC : *LIver Hepatocellular Carcinoma*, LUAD : *Lung Adenocarcinoma*, LUSC : *Lung Squamous cell Carcinoma*, PRAD : *PRostate Adenocarcinoma*, STAD : *STomach Adenocarcinoma*, THCA : *THyroid CArcinoma*, UCEC : *Uterine Corpus Endometrial Carcinoma*.

Le jeu de données d'expression de miARN du cancer du colon n'a pas été analysé à cause du trop faible nombre d'échantillons sains disponibles.

2.1.2 Pré-taitement

Les nombres de *reads* ont été normalisés selon la méthode TMM (voir section 1.2 du chapitre 2) et les gènes faiblement exprimés dont l'expression moyenne sur l'ensemble du jeu de données est inférieure à 1 CPM ont été filtrés (voir section 1.1 du chapitre 2).

2.1.3 Prise en compte des effets *batch*

Les effets *batch* ont été pris en compte sous la forme d'une variable explicative dans le cadre du GLM implémenté par MDSeq (voir section 1.4.2 du chapitre 2). La figure 4.18 représente les valeurs d'expression du gène HMBS (*Hydroxymethylbilane Synthase*, ENSG00000256269) dans les échantillons tumoraux et sains du jeu de données du cancer de la prostate en fonction des expériences de séquençage qui ont généré ces données. MDSeq a été utilisé avec et sans l'intégration d'une variable explicative

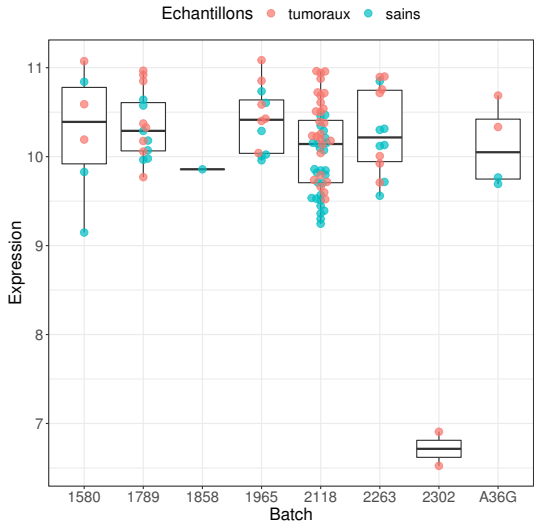

FIGURE 4.18 – Valeurs d'expression du gène HMBS (*Hydroxymethylbilane Synthase*, ENSG00000256269) pour les patients atteints du cancer de la prostate ayant fourni un échantillon tumoral et un échantillon sain en fonction des expériences de séquençage, ou *batch*, ayant généré ces données. P-valeurs du test de différence de dispersion sans intégration des effets *batch* par un facteur bloquant dans le GLM : $7.50\,10^{-3}$, avec intégration des effets *batch* par un facteur bloquant dans le GLM : $2.96\,10^{-1}$.

dans le GLM décrivant l'expérience de séquençage, ou *batch*, de chaque échantillon comme facteur bloquant. Pour le gène HMBS, le test de différence de dispersion sans l'intégration du facteur bloquant fournit une p-valeur significative de 7.50×10^{-3} alors que le même test avec le facteur bloquant fournit une p-valeur non significative de 2.96×10^{-1}. A l'exception du *batch* 2302, le niveau moyen d'expression étant assez peu variable entre les différents *batches* pour à la fois les échantillons tumoraux et sains, cette différence est très certainement due aux valeurs aberrantes fournies par ce *batch* n'ayant permis le séquençage que d'échantillons tumoraux. Ainsi, la dispersion de l'expression de ce gène est fortement augmentée par la seule influence de ces deux échantillons, reflétant parfaitement une situation d'effets *batch*. Le $\log_2$-*fold-change* de dispersion estimé par MDSeq sans l'intégration du facteur est en effet de 1,65 alors qu'il n'est que de 1,22 avec l'intégration du facteur bloquant. L'intégration d'une variable explicative dans le GLM décrivant l'expérience de séquençage d'origine des échantillons permet ainsi effectivement de prendre en compte la variabilité introduite par des valeurs aberrantes provenant de certaines expériences de séquençage dans le cadre de la comparaison de dispersion d'expression entre deux populations d'échantillons d'intérêt.

2.1.4 Identification de gènes différentiellement dispersés

Les différences de dispersion dans l'expression de miARN et d'ARNm entre échantillons tumoraux et sains ont été identifiées à l'aide de MDSeq avec l'intégration des différentes expériences de séquençage ayant généré les données sous la forme d'un facteur bloquant dans le GLM. Les analyses ont d'abord été réalisées en ne prenant en compte que les patients pour lesquels un échantillon tumoral et un échantillon sain sont disponibles avant d'inclure l'intégralité des échantillons tumoraux disponibles.

Patients pour lesquels un échantillon tumoral et un échantillon sain sont disponibles MDSeq a été appliqué pour identifer les gènes présentant une différence de dispersion d'expression entre échantillons tumoraux et échantillons sains parmi les gènes non différentiellement exprimés. Pour ce faire, les valeurs de seuil de *fold-change* de moyenne déterminés en fonction de la taille des populations d'échantillons dans le cadre de l'étude de simulation (figure 4.15) ont été utilisées pour filtrer les gènes différentiellement exprimés et une valeur de seuil de *fold-change* de dispersion égale à 1 a été utilisée pour identifier des différences de dispersion (figure 4.19). Que ce soit pour les miARN ou les ARNm, les seuils de *fold-change* de moyenne utilisés permettent de filtrer un nombre substantiel de gènes différentiellement exprimés (de 2710 à 7389 gènes pour les ARNm). Les gènes conservés pour l'analyse de différence de dispersion ne figurent ainsi pas parmi les ensembles de gènes retenus par l'analyse classique de différence d'expression basée sur la moyenne, ou du moins pas dans les tout premiers rangs, *i.e.* les gènes les plus différentiellement exprimés. L'effectif conservé pour l'analyse de différence de dispersion est toutefois de taille conséquente, avec entre 10 000 et 14 000 ARNm et entre 250 et 400 miARN.

Les miARN et les ARNm suivent des tendances similaires par rapport à la différence de dispersion d'expression entre échantillons sains et échantillons tumoraux. Deux groupes de cancers peuvent être formés en fonction du nombre de gènes différentiellement dispersés : les cancers pour lesquels un grand nombre de gènes sont différentiellement dispersés (sein, colon, rein, foie, poumon et thyroïde) et les cancers pour lesquels relativement peu de gènes ont un changement de dispersion d'expression (tête et cou, prostate, estomac et utérus). De manière générale, la très grande majorité des gènes, miARN ou ARNm, différentiellement dispersés voient leur dispersion d'expression augmenter dans les tumeurs par rapport aux échantillons sains. A l'exception de quelques cancers (adénocarcinome du poumon et prostate pour les miARN, tête et cou, prostate, estomac et utérus pour les ARNm), la proportion des gènes dont la dispersion d'expression diminue dans les tumeurs est extrêmement faible au sein de l'effectif des gènes différentiellement dispersés. Enfin, s'agissant des cancers dont l'expression des gènes est fortement différentiellement dispersée, en proportion, plus d'ARNm que de miARN sont identifiés comme différentiellement dispersés.

Prise en compte de l'intégralité des échantillons tumoraux Pour pouvoir bénéficier de plus de données d'expression, et ainsi pouvoir mieux estimer la dispersion d'expression, l'intégralité des échantillons tumoraux présents dans TCGA pour les cancers analysés dans cette étude a été utilisée et les gènes différentiellement dispersés ont été identifiés suivant le même principe que précédemment (figure 4.20). De manière générale, l'importante augmentation de taille de la population d'échantillons tumoraux permet une plus grande puissance statistique se traduisant par la détection de plus grands nombres de gènes différentiellement exprimés et de gènes différentiellement dispersés. En fonction des jeux de données, cette augmentation peut être très

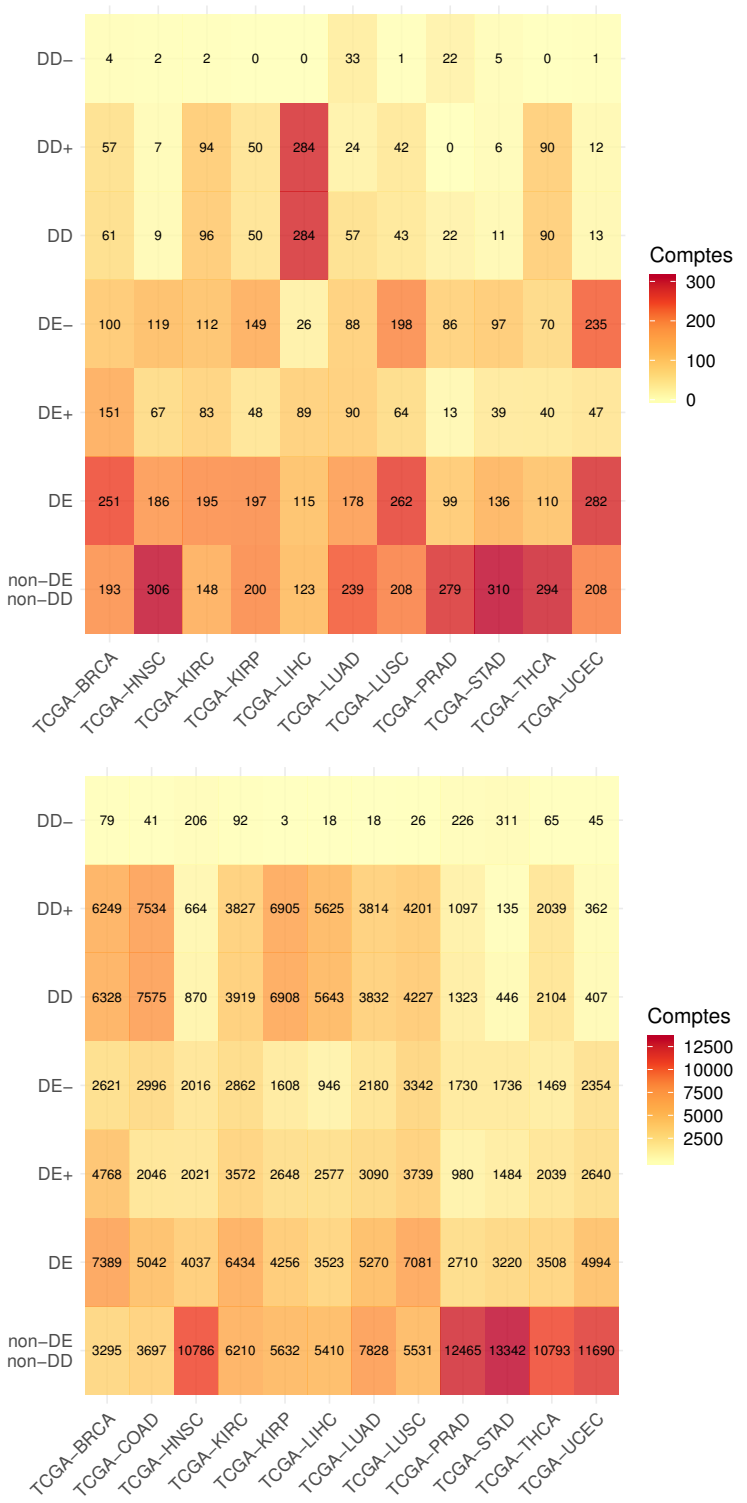

FIGURE 4.19 – Nombres de gènes, miARN (en haut) ou ARNm (en bas), différentiellement exprimés (DE), différentiellement dispersés (DD) et non différentiellement exprimés et non différentiellement dispersés (non-DE non-DD) dans le cadre de la comparaison des échantillons tumoraux et sains du TCGA fournis par paires. Le signe « + » indique une augmentation de la moyenne ou de la dispersion dans les échantillons tumoraux et le signe « - » indique une diminution.

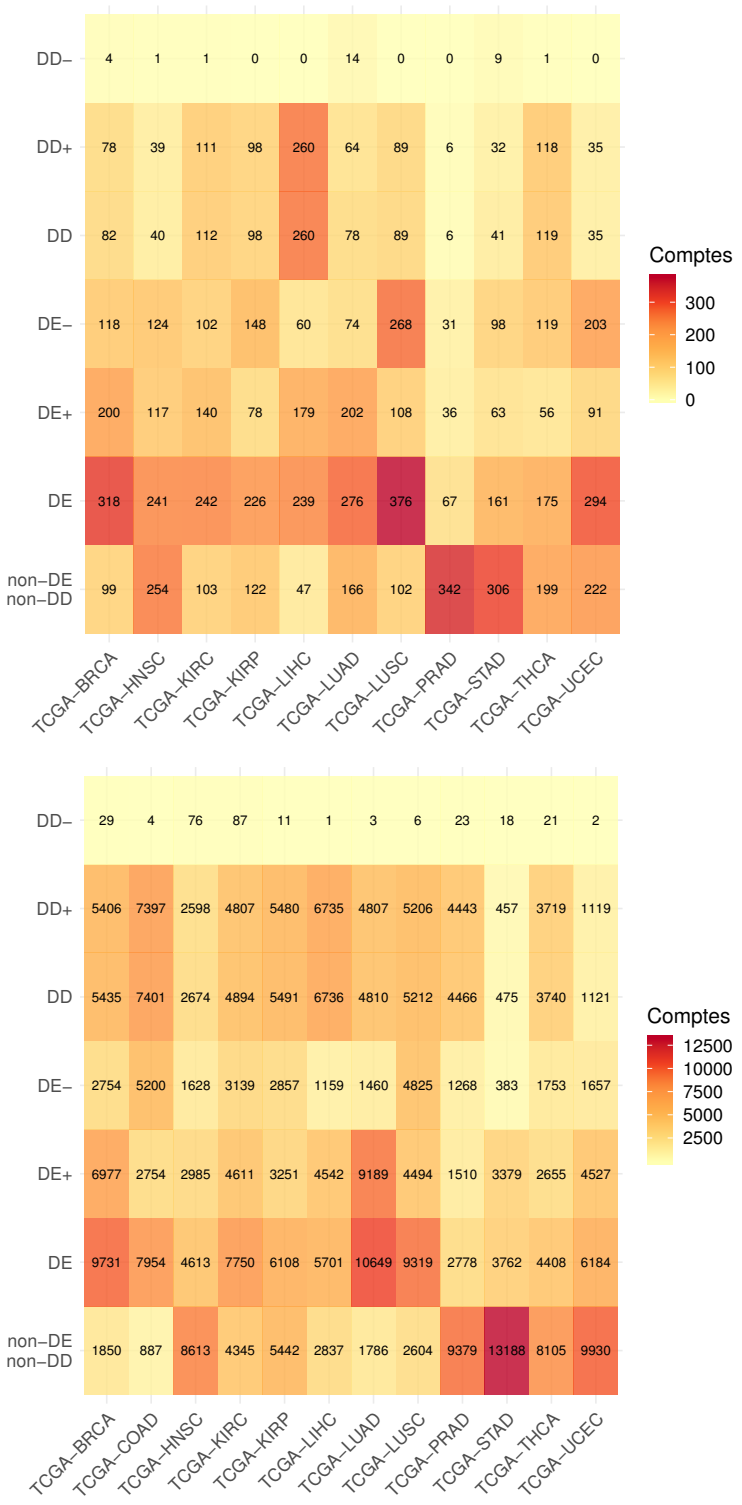

FIGURE 4.20 – Nombres de gènes, miARN (en haut) ou ARNm (en bas), différentiellement exprimés (DE), différentiellement dispersés (DD) et non différentiellement exprimés et non différentiellement dispersés (non-DE non-DD) dans le cadre de la comparaison de l'intégralité des échantillons tumoraux et sains du TCGA. Le signe « + » indique une augmentation de la moyenne ou de la dispersion dans les échantillons tumoraux et le signe « - » indique une diminution.

importante, comme pour les ARNm différentiellement dispersés dans le cancer de la prostate, passant de 1 323 à 4 466.

Malgré l'augmentation quasi générale des nombres de gènes différentiellement exprimés et de gènes différentiellement dispersés, les tendances observées avec les populations d'échantillons tumoraux et sains de tailles égales se confirment. Toutefois, parmi les gènes différentiellement dispersés, la proportion de gènes différentiellement dispersés avec une augmentation de la dispersion dans les échantillons tumoraux augmente fortement pour certains cancers. A l'inverse, le nombre de gènes dont la dispersion d'expression diminue dans les tumeurs décroit de manière généralisée pour la quasi totalité des jeux de données, atteignant des valeurs extrêmement faibles.

L'intégralité des gènes identifiés comme différentiellement exprimés ou différentiellement dispersés pour l'ensemble des jeux de données analysés sera publiée dans le cadre d'un article en cours de rédaction au moment de l'écriture de cette thèse.

2.1.5 Enrichissement de termes *Gene Ontology*

Les termes *Gene Ontology* (GO) significativement sur-représentés parmi les ARNm différentiellement dispersés ont été identifiés pour chaque jeu de données séparément à l'aide du *package* R clusterProfiler (voir section 6.2 du chapitre 2). Parmi ces termes GO enrichis, les termes similaires ont été rassemblés à l'aide de la mesure de similiraité proposée par Resnik, 1999 et un seuil de similarité de 0,8 (voir section 6.3 du chapitre 2). Par exemple, pour le jeu de données des cancers de la tête et du cou, les termes `GO:0031647` (« *regulation of protein stability* ») et `GO:0050821` (« *protein stabilization* ») forment un groupe de termes GO similaires (figure 4.21, groupe de termes GO sous le seuil de similarité le plus à gauche). Ces deux termes ont une similarité de 0,959. Cette forte similarité s'explique par le fait que ces deux termes sont directement reliés par une relation « *is_a* », le terme `GO:0050821` (« *protein stabilization* ») étant plus spécifique que `GO:0031647` (« *regulation of protein stability* »). Pour l'ensemble des groupes de termes GO similaires, seul le terme avec la p-valeur la plus faible a été conservé. Dans le cadre de l'exemple donné dans cette section, il s'agit du terme `GO:0031647`.

Les listes de termes GO enrichis simplifiées ainsi obtenues pour chaque jeu de données d'expression d'ARNm ont ensuite été comparées entre elles. Dans le but de faciliter cette comparaison, le même procédé de simplification par regroupement de termes GO similaires a été appliqué. Cette fois-ci, les termes GO étant issus de différents jeux de données, les groupes de termes GO similaires sont représentés par le terme GO qui leur est le plus proche dans l'ontologie et non par celui qui a la p-valeur la plus faible. Un nombre assez important de termes GO se retrouvent enrichis pour la plupart des jeux de données analysés (figure 4.22). Sur cette figure, les termes GO enrichis sont ordonnés par ordre décroissant du nombre de jeu de données pour lesquels la p-valeur du test d'enrichissement parmi les gènes différentiellement dispersés avec une augmentation de la dispersion dans les tumeurs est inférieure à 0,05, puis par ordre croissant de moyenne de p-valeur sur l'ensemble des jeux de données. La figure 4.22 répertorie les termes GO enrichis parmi au moins 6 jeux de données.
Les plus fortes p-valeurs se retrouvent parmi les termes GO les plus répandus à travers les cancers analysés. En effet, les p-valeurs d'enrichissement des termes GO spécifiques à seulement quelques cancers (visibles sur la figure A.5 en annexes) sont en général moins faibles que celles obtenues pour certains cancers parmi les termes GO enrichis

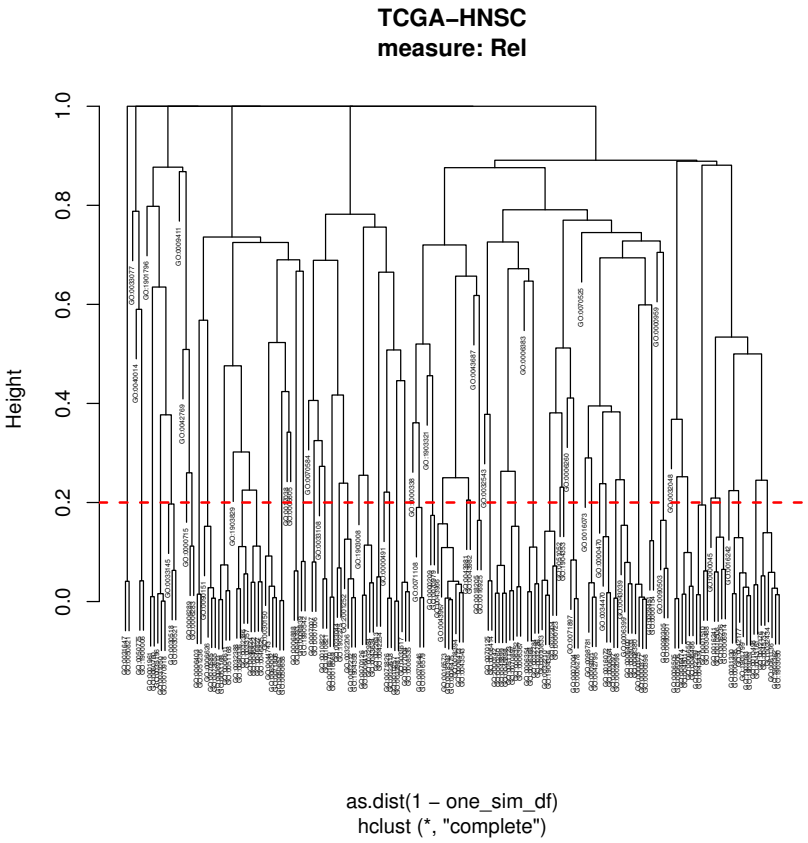

FIGURE 4.21 – Dendrogramme obtenu par *clustering* hiérarchique des termes *Gene Ontology* enrichis parmi les gènes différentiellement dispersés dont la dispersion augmente dans les tumeurs des cancers de la tête et du cou basé sur la mesure de similarité de Resnik. La droite rouge en tirets représente le seuil de similarité, fixé à 1 - 0,8, en dessous duquel les termes *Gene Ontology* sont rassemblés.

les plus répandus tels que « *ribonucleoprotein complex biogenesis* », « *RNA splicing* » ou « *translation initiation* ». Ces termes GO enrichis retrouvés dans un grand nombre de cancers différents suggèrent l'existence de processus biologiques communs dont l'expression des gènes impliqués ne se caractérise pas ou peu par une modification de la moyenne mais par une augmentation significative de la dispersion dans les tumeurs. Parmi ces processus communs au développement tumoral, de nombreux sont liés au catabolisme, à la synthèse de protéines et aux mécanismes de transport. Ces résultats font l'objet de plus amples discussions et interprétations biologiques dans la section 2.2.

Enfin, les jeux de données des cancers de l'estomac (TCGA-STAD) et de l'utérus (TCGA-UCEC) se distinguent par l'absence quasi totale de termes GO enrichis parmi les gènes présentant une augmentation de leur dispersion d'expression dans les tumeurs. En effet, un seul terme GO enrichi est trouvé pour le cancer de l'estomac, visible au bas de la figure A.5 en annexes, et aucun pour le cancer de l'utérus, ce qui explique son absence de la figure 4.22. Ces résultats peuvent s'expliquer par le faible nombre de gènes différentiellement dispersés détectés pour ces deux jeux de données en comparaison avec les autres jeux de données (voir figure 4.20), limitant de manière importante la découverte de termes *Gene Ontology* enrichis parmi ces ensembles de gènes.

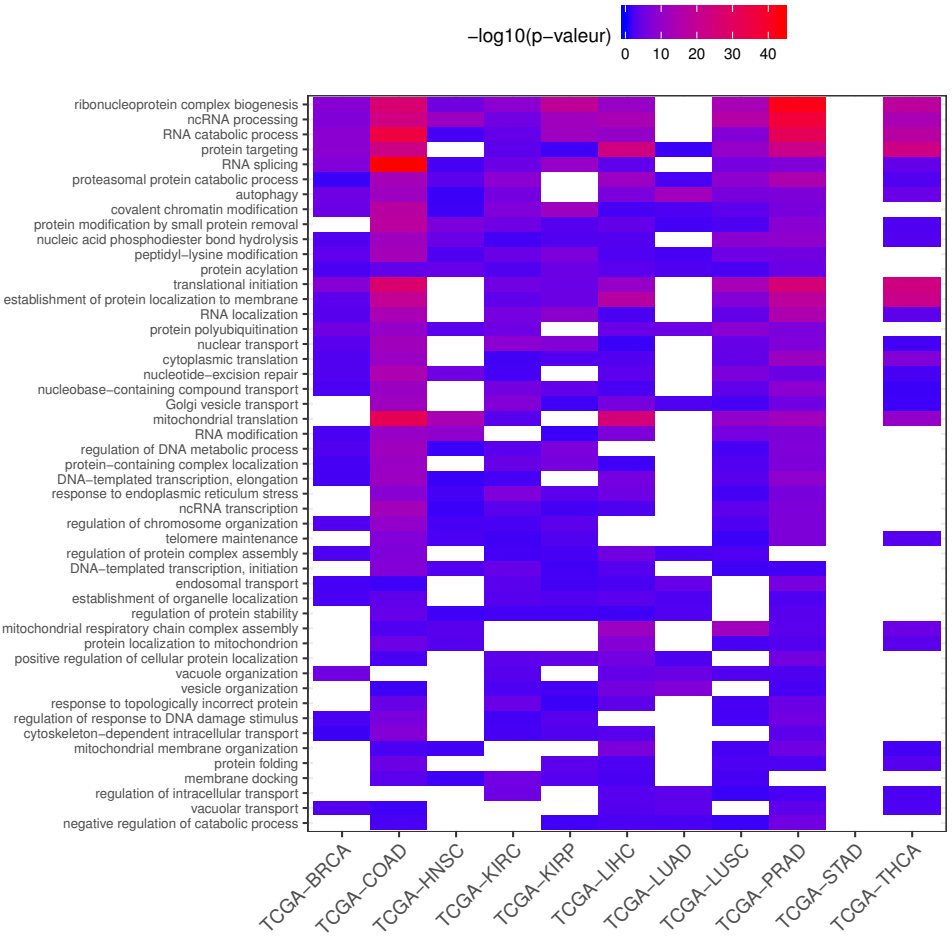

FIGURE 4.22 – Termes GO enrichis parmi les gènes différentiellement dispersés avec une augmentation de la dispersion d'expression dans les échantillons tumoraux de certains jeux de données TCGA. Pour chaque jeu de données, les listes de termes GO ont été simplifiées en utilisant la mesure de similarité sémantique proposée par RESNIK, 1999 et un seuil de 0,8. Les termes GO enrichis similaires issus de différents jeux de données ont été rassemblés pour faciliter la comparaison. Les termes GO sont ordonnés selon le nombre de jeux de données pour lesquels ils sont enrichis et la moyenne des p-valeurs des tests d'enrichissement. Seuls les termes *Gene Ontology* enrichis pour au moins 6 jeux de données différents sont visibles sur cette figure. L'intégralité des termes GO enrichis est visible sur la figure A.5 en annexes.

2.2 Discussion

2.2.1 Prise en compte de l'intégralité des échantillons tumoraux

La prise en compte de l'intégralité des échantillons tumoraux disponibles au TCGA entraîne une augmentation du nombre de gènes différentiellement dispersés avec une augmentation de la dispersion dans les échantillons tumoraux. A l'inverse, le nombre de gènes dont la dispersion d'expression diminue dans les tumeurs diminue de manière généralisée. Ces changements peuvent être dus à un véritable effet biologique suite à l'intégration de nombreux nouveaux échantillons tumoraux faisant augmenter la dispersion d'expression de nombreux gènes dans cette population d'échantillons. Ils peuvent aussi s'expliquer par le biais introduit par le fort déséquilibre de taille de population d'échantillons, un rapport de 1 à 10 pour la plupart des jeux de données TCGA (voir table 4.2), mis en évidence par l'étude de simulation (voir section 1.1.4).

Etant donné qu'un grand nombre d'échantillons est nécessaire pour pouvoir estimer le paramètre de dispersion et dans la mesure où l'étude de simulation a permis de mettre en évidence que l'augmentation de la taille de l'une des deux populations d'échantillons permet un gain de sensibilité tout en maintenant le FDR inférieur à 0,05, il serait dommageable de se priver de l'ajout de ces nombreux échantillons tumoraux. De plus, le biais introduit ne fait qu'accentuer une forte tendance observée avec des populations d'échantillons tumoraux et sains de tailles égales. Enfin, il est intéressant de noter les deux contre-exemples que constituent les jeux de données d'expression des ARNm pour le cancer du rein (TCGA-KIRC et TCGA-KIRP). Dans le cas du jeu de données TCGA-KIRC, le nombre d'ARNm dont la dispersion d'expression diminue dans les tumeurs ne diminue que très peu. S'agissant du jeu de données TCGA-KIRP, ce nombre, bien que très faible, augmente alors que celui des gènes dont la dispersion d'expression augmente dans les tumeurs diminue, allant à l'encontre de la tendance générale. Ainsi, au moins pour ces jeux de données, les différences de résultats observées suite à l'intégration de l'ensemble des échantillons tumoraux dans les jeux de données ne peuvent être entièrement imputées au biais introduit par le déséquilibre de taille des populations d'échantillons comparées.

2.2.2 Variance d'expression et robustesse

L'application de MDSeq aux jeux de données TCGA a permis de mettre en évidence une forte tendance à l'augmentation de la variance de l'expression de très nombreux gènes dans les tumeurs par rapport à des cellules saines. Cette tendance très marquée a déjà été documentée par le passé. Par exemple, Han et al., 2016 ont montré l'augmentation du coefficient de variation de l'expression des gènes dans les tumeurs de cancers du sein, du colon, du poumon et du foie et sa corrélation avec une faible expression de p53, un facteur de transcription suppresseur de tumeurs, et une faible réponse immunitaire ainsi qu'avec les stades tumoraux tardifs. L'approche proposée dans le cadre de cette thèse confirme et étend ces résultats à d'autres cancers. Les méthodes employées, basées sur la distribution binomiale négative qui reflète bien la nature des données RNA-seq, permettent d'identifier des changements de variance dus à une réelle modification biologique de la variabilité d'expression, matérialisée par un changement de dispersion et non simplement à un changement de moyenne. De plus, la validation rigoureuse de ces méthodes, à l'aide d'une étude de simulation, notamment en vue de contrôler le FDR, donne d'autant plus de poids à ces résultats.

Cette augmentation de la variance d'expression reflète certainement la diversité des perturbations génétiques à l'origine du développement tumoral et le caractère polyclonal des tumeurs. Cette grande variance d'expression peut aussi être vue comme le reflet de la grande robustesse de la pathologie cancéreuse. En effet, comme le définit Kitano, 2004, la robustesse d'un système biologique est sa capacité à maintenir ses fonctions malgré des perturbations externes ou internes. L'augmentation de la dispersion de l'expression de centaines de gènes dans les tumeurs pourrait ainsi leur permettre de s'adapter rapidement et efficacement à toute nouvelle perturbation. C'est un élément important à prendre en compte dans la prise en charge des patients et qui pourrait expliquer la résistance au traitement très fréquemment observée. En particulier, cette augmentation de dispersion d'expression pourrait expliquer l'échappement thérapeutique de certains traitements, pourtant efficaces dans leurs premières années d'application. Ces gènes dont le niveau moyen d'expression ne varie pas ou peu dans les cancers mais dont la dispersion d'expression augmente fortement dans

les tumeurs constituent *de facto* un nouvel espace où caractériser des biomarqueurs potentiels de diagnostic et/ou de pronostic des cancers.

2.2.3 Processus cataboliques

Les processus biologiques les plus significativement enrichis en gènes différentiellement dispersés dont la dispersion augmente dans les tumeurs sont aussi ceux que l'on trouve dans le plus grand nombre de cancers analysés. Ce résultat frappant suggère des caractéristiques communes de développement et/ou de progression du cancer quel que soit le tissu concerné se concentrant autour de quelques processus biologiques clés.

Parmi les processus biologiques les plus représentés dont de nombreux gènes ont une augmentation de leur dispersion d'expression dans les tumeurs, de nombreux sont en lien avec le catabolisme. En effet, parmi les dix processus biologiques significativement enrichis dans les gènes différentiellement dispersés les plus répandus dans les cancers analysés, sept correspondent à des processus de dégradation des ARN ou des protéines : « *ncRNA processing* », « *RNA catabolic process* », « *protein targeting* », « *proteosomal protein catabolic process* », « *autophagy* », « *protein modification by small protein removal* », « *nucleic acid phosphodiester bond hydrolysis* ». Le terme « *ribonucleotide complexe biogenese* », bien que contenant le terme « *biogenese* », pourrait rallonger cette liste. En effet, certains de ces complexes, comme le complexe RISC (*RNA Induced Silencing Complex*), visent aussi à dégrader les ARNm. Ce résultat suggère qu'au cours de la progression tumorale la dispersion de l'expression des gènes impliqués dans des processus cataboliques serait beaucoup plus impactée que celle des gènes impliqués dans les processus anaboliques. Cette augmentation de la dispersion d'expression de ces gènes pourrait être le reflet d'une régulation moins efficace de ces processus. Déjà mis en évidence par Han et al., 2016, l'augmentation de la dispersion de l'expression des gènes impliqués dans des processus cataboliques constitue une piste prometteuse d'investigation à mener pour mieux comprendre les mécanismes du développement tumoral.

Le protéasome et notamment le système d'ubiquitination, processus majeurs de dégradation des protéines et de recyclage des acides aminés, figurent aussi parmi les processus biologiques dont les gènes sont les plus différentiellement dispersés dans le cancer. En effet, en plus des termes « *protein targeting* », « *proteosomal protein catabolic process* » et « *protein modification by small protein removal* » déjà évoqués, le terme « *protein polyubiquitination* » figure en 17ème position sur la figure 4.22 et est retrouvé significativement enrichi pour les cancers du sein, du colon, de la tête et du cou, du rein, du foie, du poumon et de la prostate. Le système UPS (*Ubiquitin-Proteasome System*) est fortement impliqué dans les cancers (Shen et al., 2013). Il n'est donc pas surprenant de retrouver les gènes impliqués dans ces processus dans le cadre d'analyses de données issues du TCGA. En revanche, il est intéressant de noter que, par l'approche développée ici, l'expression de ces gènes est principalement marquée par une augmentation de la dispersion plutôt qu'un changement de moyenne.

2.2.4 Autophagie

Les gènes de l'autophagie sont aussi nombreux à être différentiellement dispersés dans les cancers. Ce processus est en effet retrouvé comme significativement enrichi parmi les gènes différentiellement dispersés avec une augmentation de la dispersion dans les tumeurs pour l'intégralité des cancers analysés hormis le cancer de l'estomac,

qui fait figure d'exception de manière générale, et le cancer du rein. En plus du protéasome précédemment évoqué, l'autophagie est un autre système de recyclage des molécules biologiques qui permet aux cellules de survivre à des situations critiques de leur environnement externe, comme la privation de nutriments, aussi bien qu'interne en se débarassant de composants moléculaires potentiellement dangereux tels que des organites endommagés, des pathogènes ou des infections. Il existe trois types d'autophagie : la macroautophagie, la microautophagie et l'autophagie à médiation par chaperon. Ces différentes formes d'autophagie ont pour point commun la dégradation et le recyclage de composants cellulaires dans un lysosome. La variante d'autophagie la mieux caractérisée, la macroautophagie, consiste en la séquestration progressive du matériel cytoplasmique par des organites, appelés autophagosomes, qui fusionnent avec les lysosomes pour déclencher la dégradation hydrolytique de leur charge. Les produits de dégradation autophagique, incluant potentiellement les sucres, les acides aminés, les acides gras et les nucléotides, sont généralement transportés vers le cytoplasme pour alimenter le métabolisme cellulaire et les mécanismes de réparation.

L'oncogénèse commence par la transformation d'une cellule pré-cancéreuse saine en un précurseur néoplasique qui n'est pas éliminé par les mécanismes de contrôle endogène, processus qui coïncide souvent avec l'inactivation d'un gène suppresseur de tumeur et/ou l'activation d'un oncogène. L'autophagie dans les cellules précancéreuses favorise fortement la préservation de l'homéostasie physiologique de multiples fonctions (figure 4.23). En particulier, l'élimination d'entités potentiellement mutagènes,

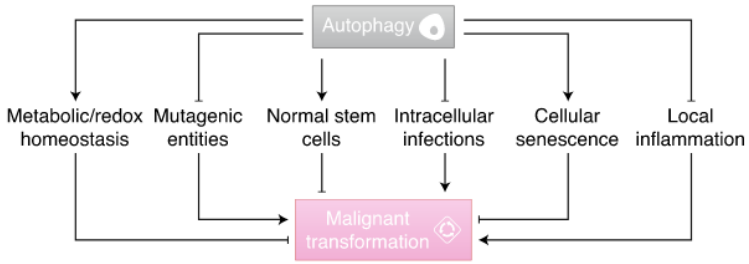

FIGURE 4.23 – Réponses autophagiques dans les cellules précancéreuses (figure 2.a de RYBSTEIN et al., 2018).

l'éradication d'infection intracellulaires et la restriction de l'inflammation locale sont autant de moyens pour enrailler le développement tumoral. Dans leur ensemble, le maintien des multiples fonctions faisant appel à des mécanismes autophagiques permettent de conserver la transformation maligne pré-cancéreuse sous contrôle.

Les cellules cancéreuses nouvellement transformées qui échappent aux mécanismes de contrôle endogènes et exogènes, y compris la sénescence cellulaire, la mort cellulaire régulée et l'immunosurveillance, se multiplient intensivement lorsqu'elles envahissent l'hôte. Ce processus, communément appelé progression tumorale, est associé à la l'apparition de manifestations cliniques de la maladie et, généralement, au début de certaines formes de traitement. Dans les cellules malignes, l'autophagie affecte la progression tumorale et la réponse au traitement de manière complexe en faisant intervenir des processus cellulaires intrinsèques et extrinsèques (figure 4.24). D'une part, l'autophagie favorise la progression tumorale et l'échappement thérapeutique par divers moyens. De manière similaire à son rôle dans le maintien des fonctions physiologiques des cellules non cancéreuses, l'autophagie soutient l'homéostasie des cellules malignes exposées à des conditions microenvironnementales difficiles. De plus,

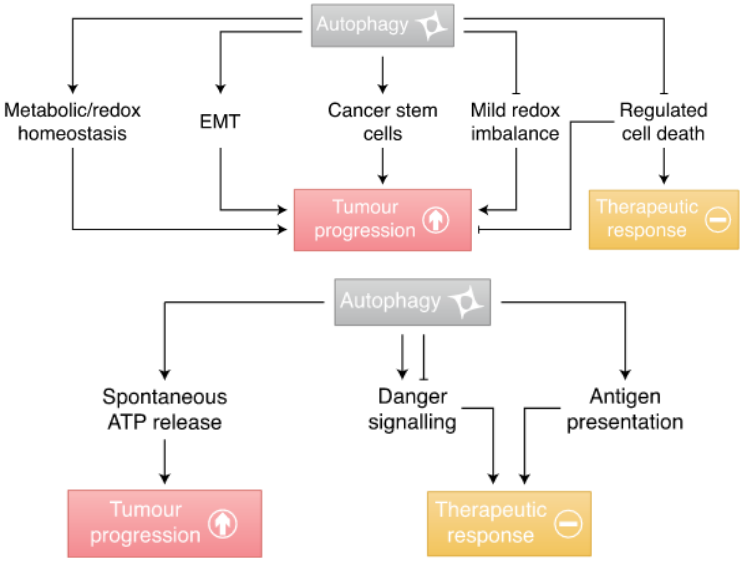

FIGURE 4.24 – Effets contradictoires de l'autophagie dans la progression tumorale : la favorisant (en haut) et provoquant une réponse immunitaire ciblée contre la tumeur (en bas, figures 2.b et 2.c de RYBSTEIN et al., 2018).

elle désensibilise les cellules néoplasiques à la mort cellulaire régulée par chimiothérapie, radiothérapie et certaines formes de l'immunothérapie. D'autre part, les réponses autophagiques dans les cellules malignes sont à l'origine de la libération spontanée d'ATP, qui favorise la progression tumorale, mais aussi de signaux de danger en réponse au déclenchement de la mort cellulaire immunogénique et peuvent déclencher des réponses immunitaires par le biais de la présentation d'antigène en favorisant la capture de cellules cancéreuses mourantes par des cellules dendritiques. Ainsi, l'effet net des réponses autophagiques au sein des cellules malignes sur la progression tumorale et la réponse au traitement dépend de nombreux paramètres et présente donc un degré élevé de dépendance au contexte. L'augmentation de la dispersion de l'expression des gènes impliqués dans ces processus observée dans les jeux de données TCGA illustrent alors la complexité de ces processus dans la progression tumorale. Cela peut amener à s'interroger sur l'opportunité de cibler ces processus dans le cadre du traitement du cancer, d'autant plus que certains traitements visent à les stimuler alors que d'autres visent à les inhiber (LEVY, TOWERS et THORBURN, 2017). Une bonne connaissance du contexte apparaît alors indispensable à l'efficacité du traitement. Les gènes impliqués dans ces processus étant pas ou peu différentiellement exprimés, une estimation de la variance de leur expression apparaît alors comme un élément pertinent pour définir le contexte.

Les processus cataboliques sont ainsi les mécanismes biologiques les plus représentés parmi les processus les plus enrichis en gènes dont la dispersion d'expression augmente dans les tumeurs. Les gènes impliqués dans d'autres processus tels que la traduction et les mécanismes de transport ont leur dispersion d'expression significativement augmentée dans les tumeurs. Ces processus, et de manière générale ceux spécifiques à seulement quelques cancers et qui n'apparaissent donc pas sur la figure 4.22, peuvent aussi constituer des pistes intéressantes pour une meilleure compréhension de leur implication dans le développement tumoral.

Chapitre 5

Association de l'expression de miARN et d'ARNm

Parmi les termes *Gene Ontology* (GO) précédemment identifiés correspondant aux gènes différentiellement dispersés présentant une augmentation de la dispersion d'expression dans les tumeurs, le terme « *ncRNA processing* » figure parmi les plus répandus et les plus significativement enrichis à travers les jeux de données analysés. Ce terme générique regroupe, entre autres, les termes plus spécifiques « *pre-miRNA processing* » et « *primary miRNA processing* » qui rassemblent des gènes impliqués dans le contrôle de la synthèse et l'action des miARN. Ces termes plus spécifiques ne présentent en général pas d'enrichissement parmi les jeux de données analysés. Ceci peut probablement s'expliquer par l'effectif important des gènes différentiellement dispersés et le faible nombre de gènes annotés par ces termes. Ces deux éléments imposent qu'une part importante des gènes annotés par ces termes GO soit retrouvé dans le groupe de gènes d'intérêt pour aboutir à une p-valeur de test d'enrichissement significative. Certains gènes différentiellement dispersés sont tout de même retrouvés pour ces termes, incitant à la recherche d'association entre l'expression de miARN et d'ARNm.

1 Résultats

Des associations ont été recherchées en appliquant le test global (voir section 7 du chapitre 2) entre les données d'expression des miARN et des ARNm issues du TCGA.

1.1 Pré-traitement

Pour permettre l'identification d'association entre valeurs d'expression de miARN et d'ARNm, il est impératif de disposer de ces deux types de données pour le même échantillon. Les échantillons pour lesquels ne sont disponibles que l'un des deux types de données ont donc été exclus de cette analyse. TCGA fournit, pour la grande majorité des échantillons, ces deux types de données. La seule différence notable concerne le jeu de données d'expression du cancer du colon (TCGA-COAD) pour lequel les données d'expression des miARN ne sont disponibles que pour un très faible nombre d'échantillons sains (voire table 4.2). Ce jeu de données tout entier a ainsi été exclu de cette analyse.

Après le retrait de ces échantillons pour lesquels l'un des deux types de données d'expression n'est pas disponible, les matrices de nombres de *reads* pour les miARN et les ARNm ont chacune subi les mêmes étapes de pré-traitement que pour l'analyse de différence de dispersion (voir section 2.1.2 du chapitre 4), à savoir : normalisation

par la méthode TMM et filtrage des gènes faiblement exprimés selon un seuil d'1 CPM appliqué sur l'ensemble du jeu de données.

1.2 Bases de données d'interaction

Les bases de données TargetScan, PITA, microCosm et miRTarBase ont été utilisées pour prédire des interactions miARN-ARNm (voir section 1.2.3 du chapitre 1). L'intégralité des interactions de miRTarBase, y compris celles classées comme faibles, a été prise en compte. Les bases de données de prédiction d'interaction *in silico* étant différentes et donc complémentaires, une prédiction par deux de ces quatre bases de prédiction d'interaction a été retenue pour valider une interaction. Ainsi, pour une interaction non validée expérimentalement, *i.e.* non prédite par miRTarBase, deux prédictions parmi les trois bases de prédiction *in silico* sont nécessaires pour la prendre compte. A l'inverse, une interaction validée expérimentalement ne sera pas retenue si elle n'est pas aussi prédite par au moins une des bases de prédiction *in silico*.

1.3 miARN associés aux ARNm différentiellement dispersés dans les tumeurs

Des associations ont été recherchées parmi les interactions miARN-ARNm identifiées telles que définies dans la section 1.2. Pour l'ensemble des associations trouvées, les catégories des miARN et des ARNm relatives à leurs différences de moyenne et de dispersion entre les échantillons tumoraux et les échantillons sains détérminées précédemment (voir section 2.1.4 du chapitre 4) ont été récupérées. La répartition de ces différentes catégories parmi ces associations significatives, avec un intérêt particulier pour celles impliquant un miARN et/ou un ARNm dont la dispersion d'expression est significativement plus grande dans les tumeurs, est représentée dans la figure 5.1. Pour les associations impliquant un ARNm dont la dispersion d'expression est augmentée dans les tumeurs, la répartition des miARN en fonction des catégories de différences de moyenne et de dispersion suit la tendance globale observée avec l'ensemble des miARN exprimés (voir figure 4.20). De manière générale, sur l'ensemble des jeux de données analysés, les miARN non différentiellement exprimés et non différentiellement dispersés sont les plus représentés parmi les associations avec des ARNm dont la dispersion d'expression augmente dans les tumeurs. Les miARN différentiellement exprimés, que leur expression augmente ou diminue dans les tumeurs, sont la deuxième catégorie la plus représentée dans ces associations. Enfin, pour certains cancers, un nombre conséquent d'associations impliquent un miARN non différentiellement exprimé et dont la dispersion d'expression augmente dans les tumeurs (foie, sein, rein, thyroïde).

2 Discussion

Le rôle tampon des miARN pour la régulation de l'expression du génome pourrait se matérialiser par des miARN dont la dispersion d'expression diminuerait entre deux conditions d'intérêt et par des ARNm cibles qui, à l'inverse, verraient leur dispersion d'expression augmentée dans une situation pathologique telle que le cancer. La mise en évidence d'un tel rôle miARN de manière systémique ne semble pas évidente avec ces jeux de données d'expression issus de différents cancers. En effet, très peu de miARN non différentiellement exprimés et dont la dispersion d'expression diminue

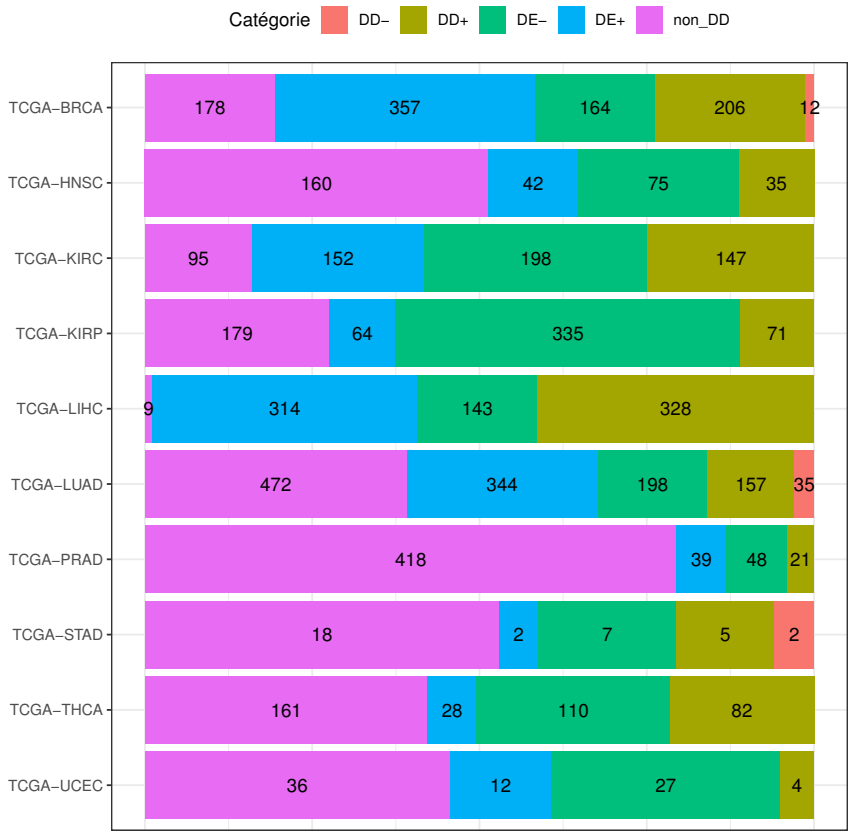

FIGURE 5.1 – Catégories selon les différences de moyenne et de dispersion d'expression entre échantillons sains et tumoraux des miARN associés à l'expression d'ARNm différentiellement dispersés ayant une augmentation de leur dispersion d'expression dans les tumeurs. Les associations sont identifiées parmi les interactions miARN-ARNm prédites par au moins deux des quatre bases d'interactions utilisées (microCosm, PITA, TargetScan et miRTarBase).

dans les tumeurs ont été identifiés (voir figure 4.20). Ainsi, très peu d'association impliquant ces miARN ont été identifiées dans ce chapitre.

On pourrait aussi imaginer qu'une dérégulation de l'expression de miARN, qui ne se traduirait pas forcément par une différence de moyenne d'expression entre une condition saine et une condition pathologique mais par une différence de dispersion d'expression, impliquerait aussi les mêmes modifications de l'expression des ARNm cibles. Ce type d'association a pu être mis en évidence en nombre assez important pour certains cancers (foie, sein, rein, thyroïde). Ces associations sont particulièrement intéressantes dans la mesure où elles impliquent des miARN et des ARNm qui ne sont pas ou peu différentiellement exprimés entre des échantillons sains et tumoraux. Ces gènes ne seraient donc pas détectés par l'analyse classique de différence de moyenne d'expression et constituent ainsi des pistes intéressantes pour une meilleure compréhension du développement tumoral.

Les ARNm non différentiellement exprimés dont la dispersion d'expression augmente dans les tumeurs ont constitué l'intérêt principal des résultats de l'approche développée dans ce chapitre. Les associations impliquant des ARNm appartenant aux autres catégories relatives aux différences de moyenne et de dispersion d'expression

ont aussi été analysées. Par exemple, les associations impliquant des ARNm différentiellement exprimés avec augmentation de la moyenne d'expression dans les tumeurs et des ARNm non différentiellement exprimés et non différentiellement dispersés sont représentés dans les figures 5.2 et 5.3 respectivement. De manière générale, la même

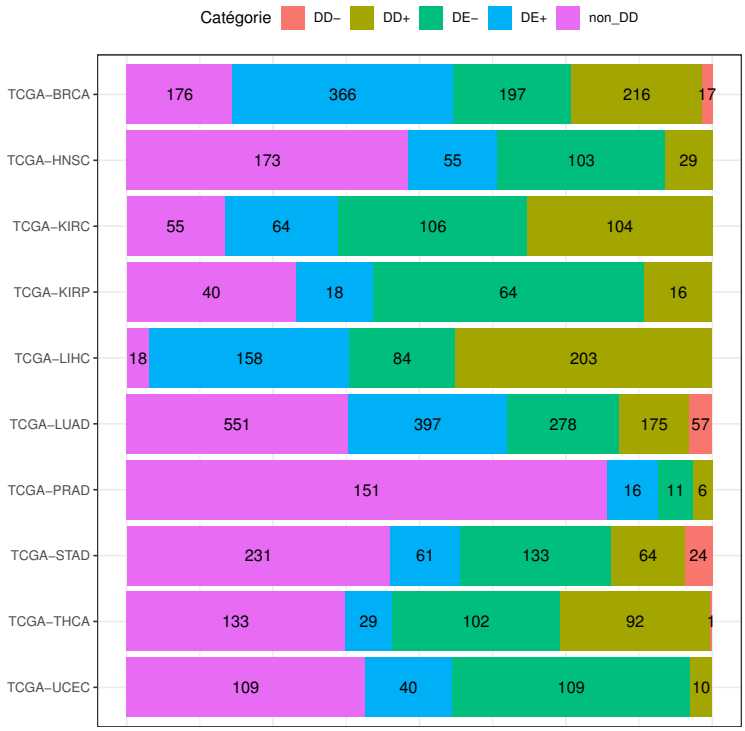

FIGURE 5.2 – Catégories selon les différences de moyenne et de dispersion d'expression entre échantillons sains et tumoraux des miARN associés à l'expression d'ARNm différentiellement exprimés ayant une augmentation de leur moyenne d'expression dans les tumeurs. Les associations sont identifiées parmi les interactions miARN-ARNm prédites par au moins deux des quatre bases d'interactions utilisées (microCosm, PITA, TargetScan et miRTarBase).

répartition des miARN selon les catégories définies par les différences de moyenne et de dispersion d'expression est observée quel que soit le profil d'expression des ARNm ciblés (différentiellement exprimés, différentiellement dispersés, ou ne présentant aucun changement d'expression entre échantillons sains et tumoraux). Ainsi, aucune tendance entre les profils d'expression de miARN et de leurs cibles prédites ne peut être observée relativement à leurs modifications de moyenne et de dispersion d'expression. En d'autres termes, par exemple, les miARN différentiellement dispersés présentant une augmentation de la dispersion d'expression dans les tumeurs n'ont pas plus tendance à être associé à des ARNm différentiellement exprimés qu'à des ARNm différentiellement dispersés ou à quelqu'autre catégorie d'ARNm que ce soit. Cette absence de tendance dans les résultats peut s'expliquer par la méthode utilisée pour identifier des associations entre l'expression de miARN et d'ARNm et son application dans le cadre de la comparaison d'échantillons tumoraux et d'échantillons sains. Bien que présentant les avantages de considérer l'association entre l'expression d'un ARNm et de l'ensemble des miARN prédits pour le cibler et de ne pas se limiter à des relations d'anti-corrélation entre expressions de miARN et d'ARNm, le test global pourrait ne pas être tout à fait adapté à l'identification d'association entre miARN et ARNm différentiellement dispersés. En effet, de manière similaire

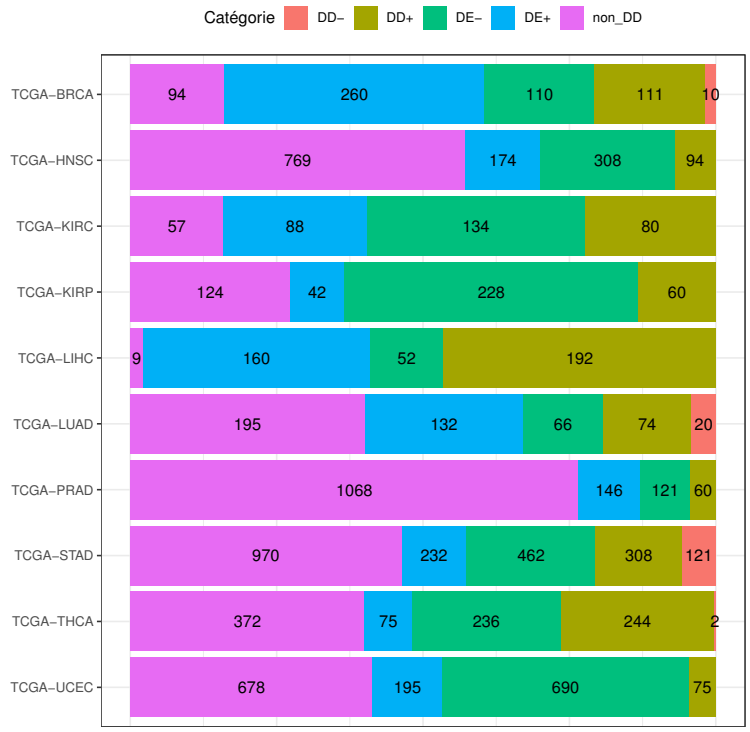

FIGURE 5.3 – Catégories selon les différences de moyenne et de dispersion d'expression entre échantillons sains et tumoraux des miARN associés à l'expression d'ARNm non différentiellement exprimés et non différentiellement dispersés. Les associations sont identifiées parmi les interactions miARN-ARNm prédites par au moins deux des quatre bases d'interactions utilisées (microCosm, PITA, TargetScan et miRTarBase).

aux méthodes classiques de détection de différence de moyenne d'expression basées sur la distribution binomiale négative, le test global ne permet pas la modélisation de différence de dispersion entre les deux populations d'échantillons d'intérêt (voir section 7.1 du chapitre 2). Cette limite peut expliquer l'absence de tendance relative aux différences de dispersion entre échantillons tumoraux et sains dans les associations identifiées entre les expressions de miARN et d'ARNm et ainsi justifier de futurs développements méthologiques pour y remédier.

Chapitre 6

Conclusion et perspectives

La détection de différence de variance dans les données d'expression issues du RNA-seq est une tâche complexe. L'application de tests classiques de différence de variance ou de coefficient de variation semble peu adaptée à ce type de données. Pour pouvoir être appliqués, ces tests nécessitent que les données aient subi d'importantes transformations, $\log_2$ transformation et correction d'effets *batch*, pouvant modifier grandement leur variance. La distribution binomiale négative apparaît comme étant plus appropriée pour modéliser des nombres de *reads* . Dans le cadre de cette modélisation, la dispersion, qui définit la variance de manière indépendante de la moyenne, est le paramètre à estimer pour identifier des gènes présentant une différence de variance d'expression entre deux populations d'échantillons. Ce paramètre nécessite plus de données pour pouvoir être estimé avec précision que la moyenne. Les premières méthodes basées sur cette distribution se sont ainsi concentrées sur la détection de différence de moyenne d'expression entre deux populations d'échantillons en accord avec la petite taille des jeux de données RNA-seq à l'époque de leur publication. Dans le cadre de ces approches, la dispersion doit tout de même être estimée avant de pouvoir identifier une différence de moyenne. Cette estimation est une tâche complexe et différentes stratégies de partage d'information entre gènes et échantillons de différentes populations d'intérêt ont été déployées pour augmenter le nombre de données disponibles et ainsi pallier la petite taille des jeux de données. Bien que ces approximations aient un impact limité pour la détection de différence de moyenne d'expression, ces approches sont assez peu réalistes biologiquement dans la mesure où elles ne permettent pas d'envisager qu'un gène puisse avoir une différence de dispersion d'expression entre les populations d'échantillons d'intérêt.

La baisse du coût du séquençage a permis le développement de bases de données d'expression publiques volumineuses. Des jeux de données d'expression composés de populations d'échantillons de grandes tailles sont ainsi de plus en plus disponibles, ouvrant la perspective de pouvoir estimer la dispersion sans utiliser de stratégies d'optimisation. Ainsi, très récemment, deux méthodes, MDSeq (Ran et Daye, 2017) et DiPhiSeq (Li et Lamere, 2018), permettant la détection à la fois de différences de moyenne et de dispersion ont été publiées. Les simulations menées au cours de cette thèse ont permis de déterminer les performances de ces deux méthodes pour la détection de différence de dispersion de manière rigoureuse. Elles ont permis notamment de confirmer qu'un grand volume de données était nécessaire pour détecter des différences de dispersion d'expression entre populations d'échantillons. Des populations d'au moins 30 échantillons représentent en effet la taille minimale pour y parvenir. En outre, elles ont permis de mettre en évidence les principales caractéristiques de ces deux méthodes. MDSeq ne permet en réalité pas de tester des différences de dispersion mais des différences de variance. Ses performances pour la détection de différence de dispersion sont ainsi affectées par la présence de différence de moyenne,

limitant son application à des gènes pas ou faiblement différentiellement exprimés. A l'inverse, DiPhiSeq permet effectivement de tester des différences de dispersion, ce qui permet d'envisager de détecter des différences de dispersion pour l'intégralité des gènes qui composent le jeu de données d'expression. Malheureusement, cette différence de paramètre testé se traduit par des temps de calculs énormément plus longs que pour MDSeq. De plus, l'impossibilité de prendre en compte les effets *batch* par un effet bloquant dans le cadre d'un modèle linéaire généralisé nuit à l'application de DiPhiSeq à des données RNA-seq dont on sait qu'elles peuvent être affectées par ce genre de biais techniques. C'est pour ces raisons que MDSeq est la méthode qui a été retenue pour analyser les données du TCGA en restreignant son application aux gènes faiblement différentiellement exprimés tels que définis lors de l'étude de simulation. Cette limite n'apparaît pas comme un obstacle majeur dans la mesure où l'intérêt biologique de cette démarche est justement de pouvoir identifier des gènes qui ne seraient pas détectés par l'approche classique de différence de moyenne. Toutefois, à l'avenir, il sera intéressant de pouvoir aussi identifier des différences de dispersion pour des gènes différentiellement exprimés à l'aide de développement de nouvelles méthodes ou d'amélioration des méthodes existantes.

Conformément au premier but de cette thèse, l'application de MDSeq aux jeux de données du TCGA a permis de mettre en évidence qu'un nombre important de miARN et d'ARNm faiblement différentiellement exprimés présentent en revanche une augmentation de leur dispersion d'expression dans les tumeurs par rapport à des tissus sains. La comparaison des termes *Gene Ontology* obtenus pour l'ensemble des cancers disponibles, facilitée par l'utilisation de mesures de similarité sémantique, a révélé l'existence de fonctions biologiques communes dont l'expression des gènes impliqués se caractérise plus par une augmentation de la dispersion de leur expression dans les tumeurs que par un changement de moyenne. Parmi celles-ci, de nombreuses fonctions sont en lien avec le catabolisme et la synthèse protéique. L'exemple de l'autophagie montre comment l'analyse de différence de dispersion d'expression peut permettre une meilleure compréhension des processus biologiques du développement tumoral et leur meilleure prise en compte dans le cadre de traitements. En effet, cette fonction cellulaire est la cible de différents traitements anti-cancéreux dont l'efficacité dépend grandement du contexte. De nombreuses situations d'échappement thérapeutique sont ainsi observées. L'importante dispersion d'expression des gènes impliqués dans cette fonction cellulaire ciblée par ces stratégies de traitement peut refléter l'adaptation des tumeurs aux perturbations de leur environnement engendrées par ces traitements et expliquer ces situations d'échappement thérapeutique.

Le second but de cette thèse était de mieux caractériser le rôle tampon des miARN en étudiant la variance de leur expression ainsi que celles de leurs cibles. L'étude de jeux de données d'expression dans une condition pathologique telle que le cancer n'est peut-être pas la situation la plus appropriée pour mettre en évidence ce rôle. En effet, une tendance générale très marquée à l'augmentation de la dispersion d'expression des miARN et des ARNm est observée dans les tumeurs, caractérisant un état de dérégulation de l'expression du génome quasi généralisé. Il serait peut-être plus pertinent d'appliquer cette approche à une situation biologique plus contrôlée, telle que la différenciation de cellules souches au cours du développement, par exemple. Il est en effet possible d'émettre l'hypothèse que l'éventail des possibles en terme d'expression du génome ne fait que se réduire au cours du développement alors que les tissus se spécialisent de plus en plus. Le PCBC (*Progenitor Cell Biology Consortium*,

Salomonis et al., 2016, Daily et al., 2017) a fourni un effort d'harmonisation similaire à celui fourni par TCGA pour la caractérisation de différentes lignées de cellules souches. Différents types de données de séquençage, comprenant notamment des données d'expression de miARN et d'ARNm, sont ainsi disponibles pour différents stades de différenciation. Bien que le nombre d'échantillons disponibles soit beaucoup moins grand que pour TCGA, il permet tout même d'envisager la détection de gènes présentant une différence de dispersion d'expression entre certains stades de différenciation. Il serait alors intéressant de comparer les résultats obtenus à partir de ces jeux de données avec ceux obtenus avec TCGA et d'identifier les miARN et les ARNm qui voient leurs profils d'expression les plus bouleversés entre le développement tumoral et la différenciation de cellules souches.

De manière générale, les jeux de données d'expression issus du RNA-seq de plus en plus grands disponibles dans des bases de données telles que GEO (*Gene Expression Omnibus*) ou GTEx (*Genotype-Tissue Expression*, Battle et al., 2017) donnent l'opportunité d'analyser la dispersion d'expression dans une grande variété de tissus et de conditions biologiques. Le développement d'autres bases de données dédiées à des pathologies particulières permettent d'envisager de pouvoir identifier des gènes présentant des modifications de leur dispersion d'expression dans le développement de ces pathologies. Par exemple, la base de données PD_NGSAtlas (Zhao et al., 2014b) vise à rassembler des profils d'expression et de méthylation obtenus à partir de données RNA-seq dans le cadre de l'étude de troubles psychiatriques tels que la schizophrénie ou les troubles bipolaires. Cette base de données offre la perspective de poursuivre et d'approfondir les efforts déjà entrepris dans l'identification de gènes se caractérisant par une modification de leur variance d'expression dans le développement de ces pathologies. Enfin, le développement de méthodes d'aggrégation de jeux de données d'expression obtenues de manière hétéroclite, avec la prise en compte de biais techniques sous-jacents, est une autre voie possible vers l'obtention de grands jeux de données d'expression dans une variété de conditions biologiques différentes (Sanders et al., 2017). Dans cette optique, notons l'effort mené par Wang et al., 2018 visant à augmenter le nombre d'échantillons sains dans les jeux de données du TCGA en intégrant des échantillons provenant de GTEx, avec une intention particulière portée sur le contrôle de l'introduction d'éventuels effets *batch*. Dans le cadre des travaux menés au cours de cette thèse, cette approche est particulièrement intéressante dans la mesure où elle pourrait permettre d'analyser de nouveaux jeux de données issus d'autres types de cancer et d'augmenter la puissance statistique des analyses déjà réalisées.

Ces dernières années, l'émergence du séquençage de cellules uniques a ouvert de nouvelles perspectives dans l'analyse de l'expression des gènes. Comme son nom l'indique, il permet d'atteindre une résolution de l'expression des gènes au niveau de la cellule. Par opposition, le séquençage *bulk*, *i.e.* la technique qui a généré les données analysées dans le cadre de cette thèse, reflète l'expression des gènes au niveau de populations de plusieurs milliers de cellules. Le séquençage de cellules uniques offre ainsi la possibilité d'observer une hétérogénéité dans les profils d'expression alors que seule une expression moyenne peut être captée par séquençage *bulk* (Teschendorff et Enver, 2017). Cette caractéristique est particulièrement intéressante pour étudier les tumeurs dont on sait que le caractère polyclonal peut engendrer une grande hétérogénéité d'expression. Le séquençage de cellules uniques est aussi très pertinent pour étudier la différenciation de cellules souches dans la mesure où le changement de profil d'expression d'une seule cellule souche peut entrainer toute une population dans une

voie de différenciation ou un lignage particulier (Kester et Oudenaarden, 2018). Etudier la dispersion d'expression de gènes dans ce type de données apparaît ainsi pertinente. Ces jeux de données, où un échantillon est une cellule, sont beaucoup plus grands en nombre d'échantillons que pour le séquençage *bulk*. Il n'est en effet pas rare que des jeux de données soient constitués de centaines, voire de milliers de cellules. En revanche, la profondeur de séquençage peut être très faible, se caractérisant par la présence de valeurs nulles en nombre encore plus important que pour le séquençage *bulk*. Il y a donc un choix à faire entre le nombre de cellules séquencées et la profondeur de séquençage avant toute expérience de séquençage de cellules uniques en fonction du but de l'étude (Haque et al., 2017). Une profondeur de séquençage très faible pouvant très probablement nuire à l'identification de différence de dispersion et dans la mesure où les plus petits jeux de données sont tout de même composés de quelques dizaines de cellules, la profondeur de séquençage serait à privilégier dans la perspective d'identifier des différences de dispersion dans des données issues du séquençage de cellules uniques. Dans une telle situation, les valeurs nulles demeurent toujours présentes en grands nombres et doivent être prises en compte dans toute analyse statistique. L'éventuelle application de MDSeq à des jeux de données de séquençage de cellules uniques a été prise en compte lors de son développement avec la modélisation de valeurs nulles en excès. Cette fonctionnalité vise à estimer si les valeurs nulles ont une valeur biologique, *i.e.* l'absence de l'expression d'un gène, ou si elles sont le résultat d'une trop faible profondeur de séquençage. Cette caractéristique apporte un avantage supplémentaire à l'utilisation de cette méthode dans le but de détecter des différences de dispersion d'expression entre deux populations d'échantillons. L'écrasante majorité des expériences de séquençage de cellules uniques a pour but de séquencer les gènes codants des protéines. Il existe en effet très peu de jeu de données d'expression de gènes non codants en cellules uniques. A ce jour, il n'existe qu'un seul jeu de données mettant à disposition à la fois des données d'expression d'ARNm et de petits ARN, incluant les miARN, pour les même cellules (Faridani et al., 2016) ouvrant la perspective de poursuivre la recherche d'associations dans l'expression de ces différents types d'ARN au niveau de la cellule.

La dispersion est un paramètre à prendre en compte dans l'analyse de jeux de données d'expression de grandes tailles, soit pour une meilleure estimation de différence de moyenne, soit pour la volonté explicite d'analyser des changements de variance d'expression. Dans le cadre de l'analyse classique de différence de moyenne, l'hypothèse qu'un gène ait la même valeur de dispersion pour l'ensemble des populations d'échantillons d'intérêt ne devrait plus être émise. La prise en compte d'une différence de dispersion d'expression pourrait ainsi améliorer les performances de détection de différence de moyenne. Pour des approches dont le but premier de l'analyse porte sur la variance d'expression, l'émergence de nouvelles méthodes dédiées, le développement de bases de données d'expression toujours plus grandes et l'expansion du séquençage de cellules uniques offrent des perspectives enthousiasmantes de meilleure compréhension des processus biologiques et de découverte de nouveaux biomarqueurs et de nouvelles cibles thérapeutiques. La publication très récente d'un article de revue au sujet de la variabilité de l'expression des gènes ne fait que confirmer l'intérêt grandissant autour de cette « *autre dimension de l'analyse du transcriptome* »(De Jong, Moshkin et Guryev, 2019).

Annexe A

Modèles basés sur la distribution binomiale négative

1 Données simulées

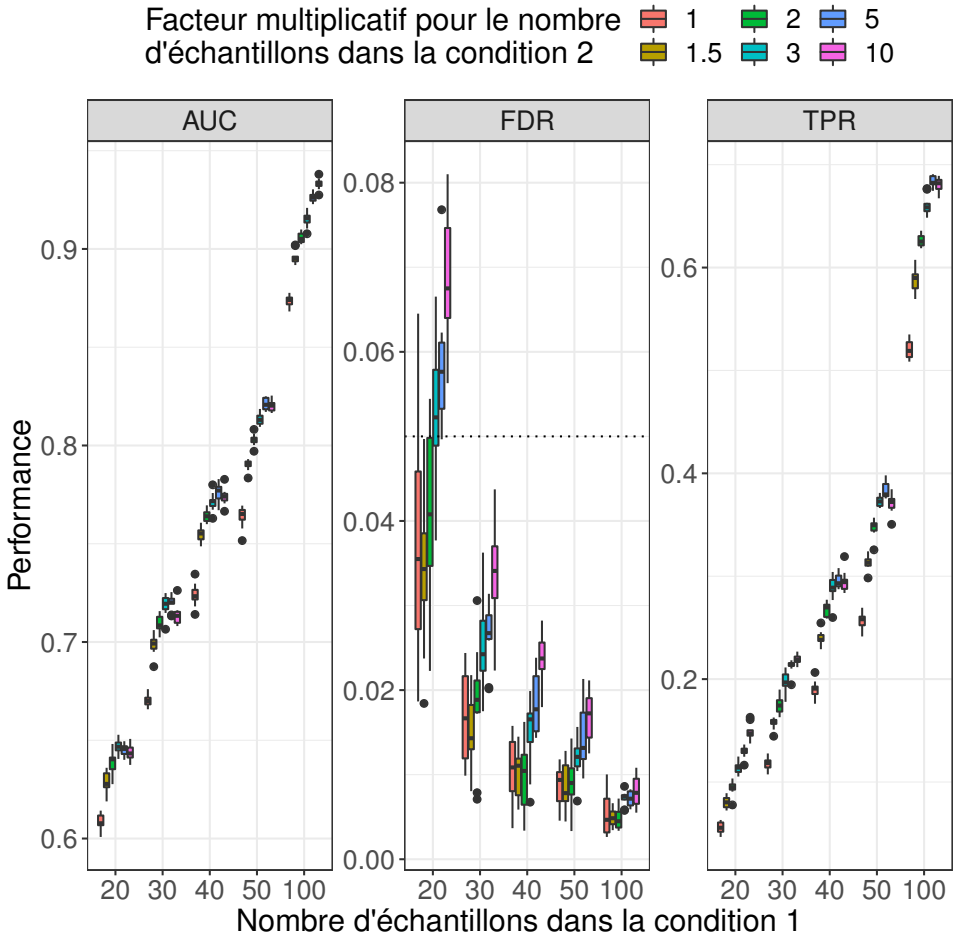

FIGURE A.1 – Performances obtenues avec MDSeq pour la détection de différence de dispersion avec des jeux de données simulées composés de populations d'échantillons de tailles inégales. La taille de la population de la seconde condition peut prendre différentes valeurs, de 1,5 à 10 fois plus grande que la population de la première condition. Paramètres de simulation : 10 000 gènes, 50% de gènes différentiellement dispersés avec une différence de dispersion d'au moins 50%, les gènes non différentiellement dispersés ont la même valeur de dispersion pour les deux populations, les *fold-changes* de moyenne sont limités par une valeur maximale de 1,1, présence d'un *outlier* pour 10% des gènes. Les performances ont été mesurées à l'aide de 10 réplicats pour chaque jeu de paramètres. La fonction de retrait d'*outliers* de MDSeq a été appliquée. Les gènes différentiellement dispersés ont été identifiés à l'aide d'un seuil de *fold-change* égal à 1.

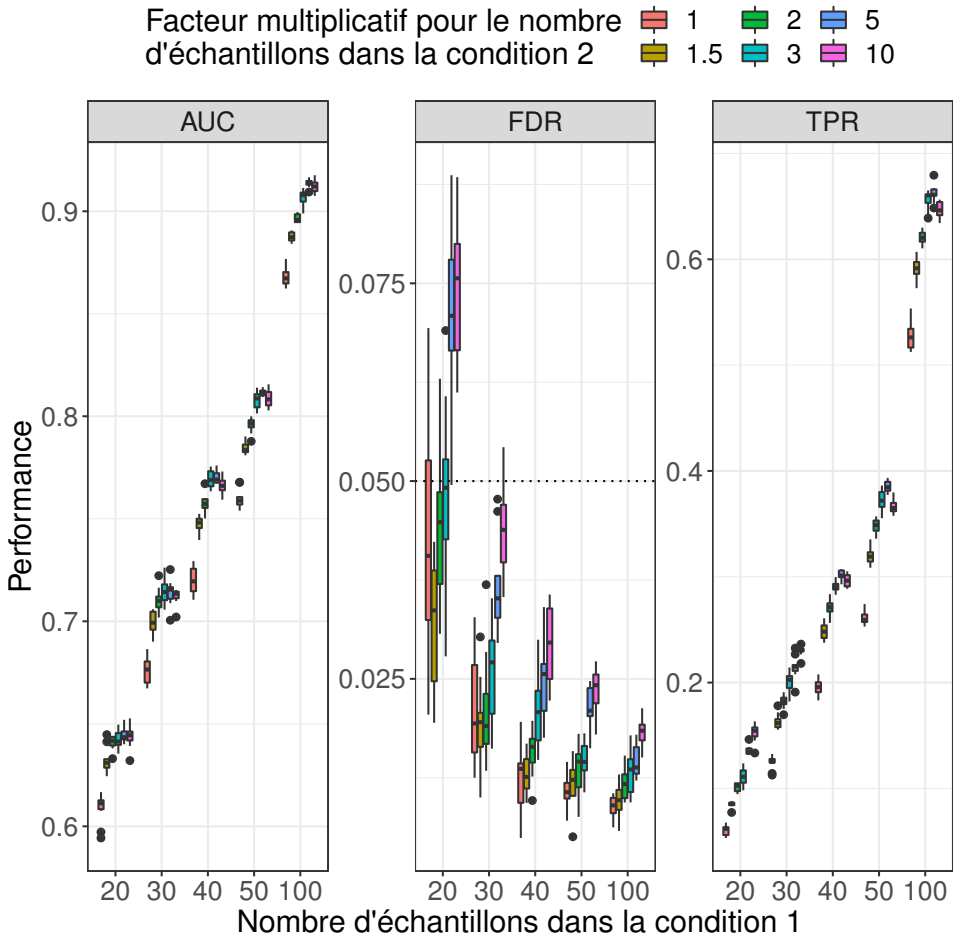

FIGURE A.2 – Performances obtenues avec MDSeq pour la détection de différence de dispersion avec des jeux de données simulées composés de populations d'échantillons de tailles inégales. La taille de la population de la seconde condition peut prendre différentes valeurs, de 1,5 à 10 fois plus grande que la population de la première condition. Paramètres de simulation : 10 000 gènes, 50% de gènes différentiellement dispersés avec une différence de dispersion d'au moins 50%, les gènes non différentiellement dispersés ont la même valeur de dispersion pour les deux populations, les *fold-changes* de moyenne sont limités par une valeur maximale de 1,2, présence d'un *outlier* pour 10% des gènes. Les performances ont été mesurées à l'aide de 10 réplicats pour chaque jeu de paramètres. La fonction de retrait d'*outliers* de MDSeq a été appliquée. Les gènes différentiellement dispersés ont été identifiés à l'aide d'un seuil de *fold-change* égal à 1.

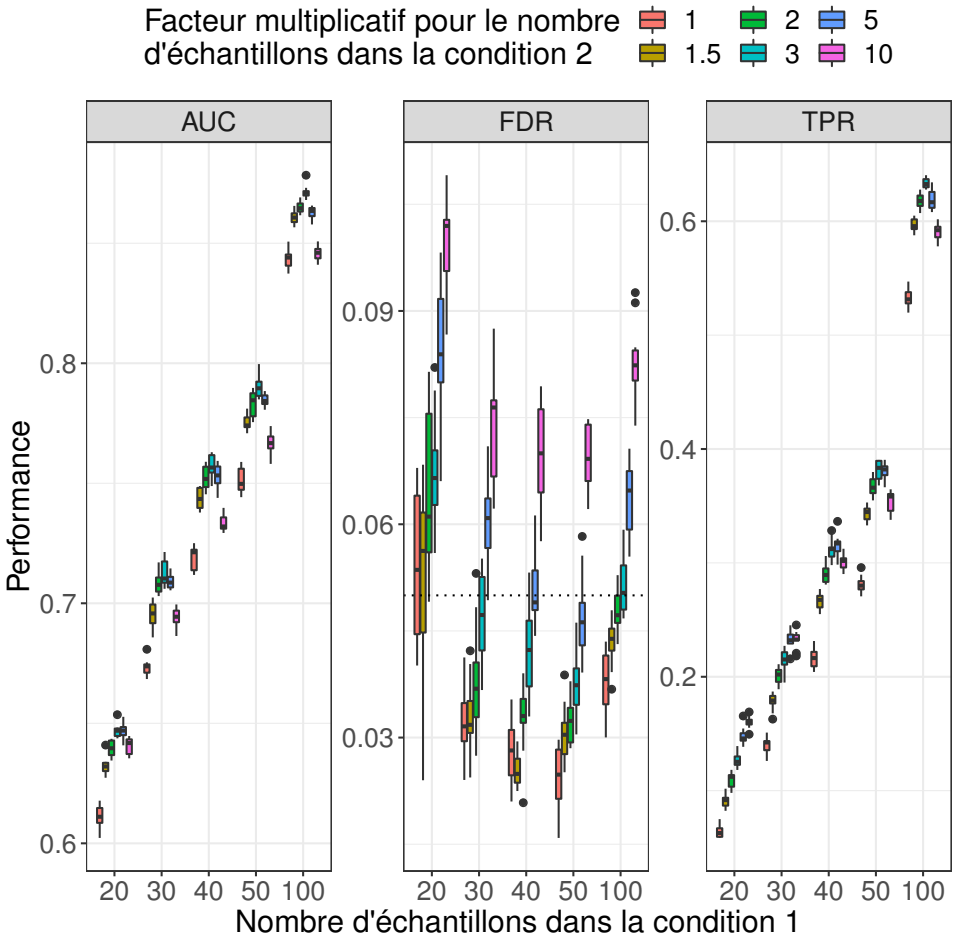

FIGURE A.3 – Performances obtenues avec MDSeq pour la détection de différence de dispersion avec des jeux de données simulées composés de populations d'échantillons de tailles inégales. La taille de la population de la seconde condition peut prendre différentes valeurs, de 1,5 à 10 fois plus grande que la population de la première condition. Paramètres de simulation : 10 000 gènes, 50% de gènes différentiellement dispersés avec une différence de dispersion d'au moins 50%, les gènes non différentiellement dispersés ont la même valeur de dispersion pour les deux populations, les *fold-changes* de moyenne sont limités par une valeur maximale de 1,4, présence d'un *outlier* pour 10% des gènes. Les performances ont été mesurées à l'aide de 10 réplicats pour chaque jeu de paramètres. La fonction de retrait d'*outliers* de MDSeq a été appliquée. Les gènes différentiellement dispersés ont été identifiés à l'aide d'un seuil de *fold-change* égal à 1.

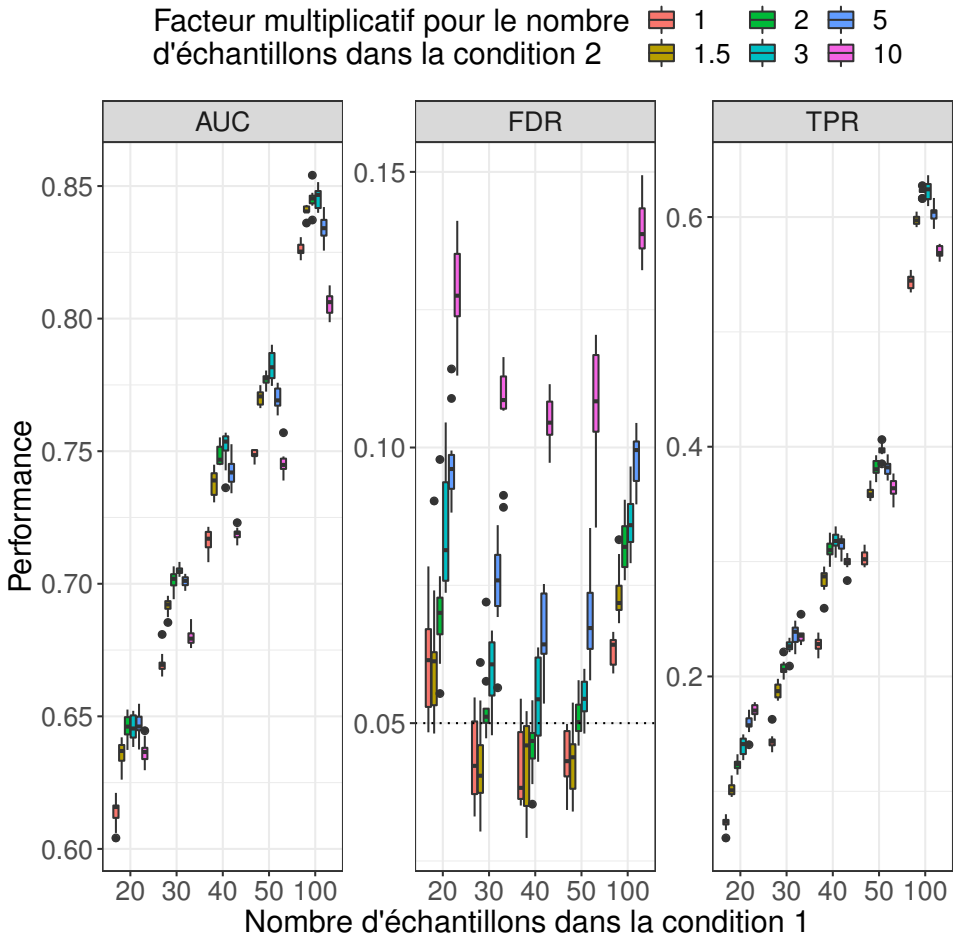

FIGURE A.4 – Performances obtenues avec MDSeq pour la détection de différence de dispersion avec des jeux de données simulées composés de populations d'échantillons de tailles inégales. La taille de la population de la seconde condition peut prendre différentes valeurs, de 1,5 à 10 fois plus grande que la population de la première condition. Paramètres de simulation : 10 000 gènes, 50% de gènes différentiellement dispersés avec une différence de dispersion d'au moins 50%, les gènes non différentiellement dispersés ont la même valeur de dispersion pour les deux populations, les *fold-changes* de moyenne sont limités par une valeur maximale de 1,5, présence d'un *outlier* pour 10% des gènes. Les performances ont été mesurées à l'aide de 10 réplicats pour chaque jeu de paramètres. La fonction de retrait d'*outliers* de MDSeq a été appliquée. Les gènes différentiellement dispersés ont été identifiés à l'aide d'un seuil de *fold-change* égal à 1.

2 Données réelles

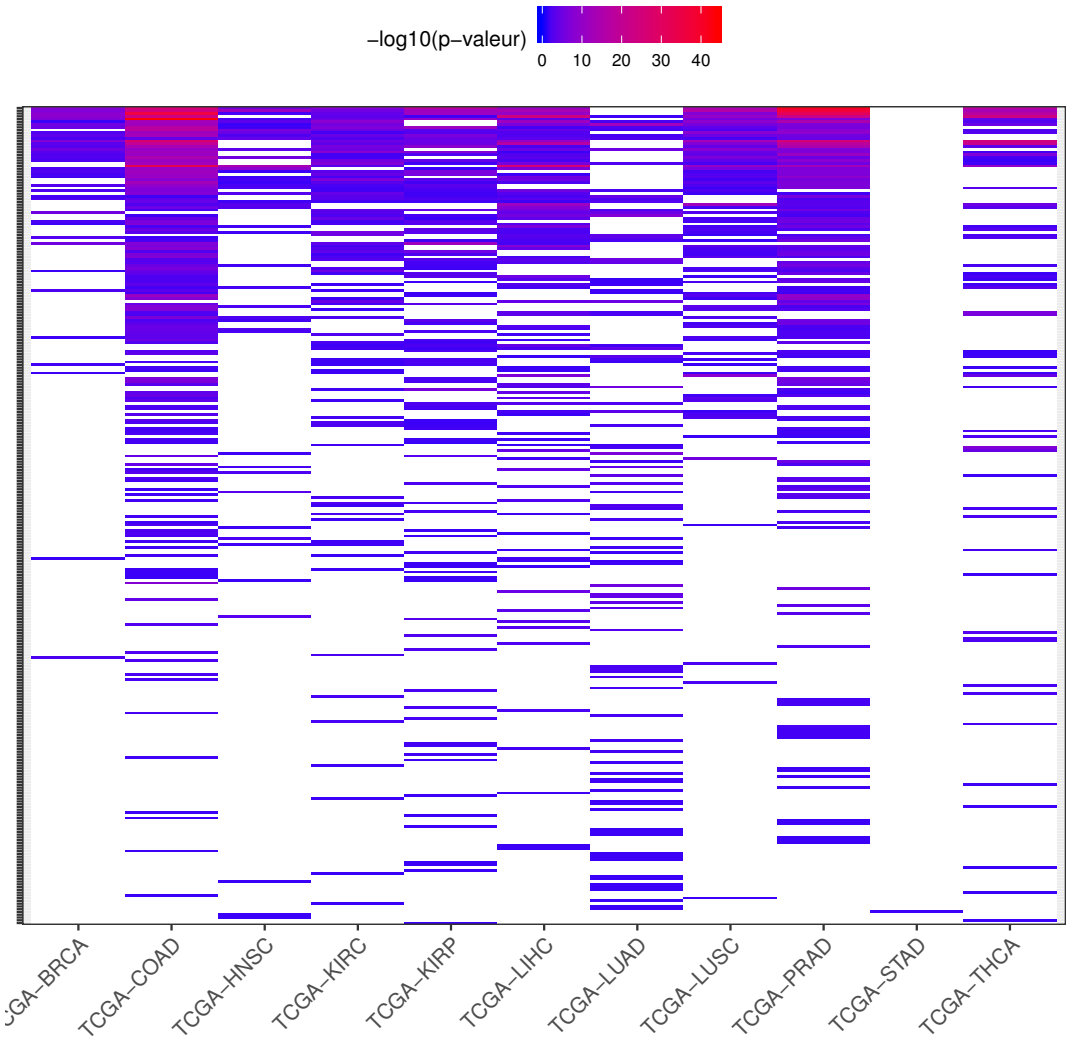

FIGURE A.5 – Intégralité des termes GO enrichis parmi les gènes différentiellement dispersés avec une augmentation de la dispersion d'expression dans les échantillons tumoraux de certains jeux de données TCGA. Les termes GO les plus répandus (en haut de la figure) sont visibles sur la figure 4.22. Pour chaque jeu de données, les listes de termes GO ont été simplifiées en utilisant la mesure de similarité sémantique proposée par Resnik et un seuil de 0,8. Les termes GO enrichis similaires issus de différents jeux de données ont été rassemblés pour faciliter la comparaison. Les termes GO sont ordonnés selon le nombre de jeux de données pour lesquels ils sont enrichis et la moyenne des p-valeurs des tests d'enrichissement.

Références

Aagaard, L. et J. J. Rossi (2007). « RNAi therapeutics : principles, prospects and challenges ». In : *Adv. Drug Deliv. Rev.* 59.2-3, p. 75–86.

Aeberhard, W. H., E. Cantoni et S. Heritier (2014). « Robust inference in the negative binomial regression model with an application to falls data ». In : *Biometrics* 70.4, p. 920–931.

Agarwal, V. et al. (2015). « Predicting effective microRNA target sites in mammalian mRNAs ». In : *Elife* 4.

Anders, S. et W. Huber (2010). « Differential expression analysis for sequence count data ». In : *Genome Biol.* 11.10, R106.

Anders, S., P. T. Pyl et W. Huber (2015). « HTSeq–a Python framework to work with high-throughput sequencing data ». In : *Bioinformatics* 31.2, p. 166–169.

Andrews, Simon (2010). *FASTQC : A quality control tool for high throughput sequence data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Anfossi, S. et al. (2014). « High serum miR-19a levels are associated with inflammatory breast cancer and are predictive of favorable clinical outcome in patients with metastatic HER2+ inflammatory breast cancer ». In : *PLoS ONE* 9.1, e83113.

Ashburner, M. et al. (2000). « Gene ontology : tool for the unification of biology. The Gene Ontology Consortium ». In : *Nat. Genet.* 25.1, p. 25–29.

Bartel, D. P. et C. Z. Chen (2004). « Micromanagers of gene expression : the potentially widespread influence of metazoan microRNAs ». In : *Nat. Rev. Genet.* 5.5, p. 396–400.

Bartlett, M. S. (1937). « Properties of Sufficiency and Statistical Tests ». In : *Proceedings of the Royal Society of London A : Mathematical, Physical and Engineering Sciences* 160.901, p. 268–282. issn : 0080-4630.

Battle, A. et al. (2017). « Genetic effects on gene expression across human tissues ». In : *Nature* 550.7675, p. 204–213.

Ben-Dayan, M. M. et al. (2015). « Cancer as the Disintegration of Robustness : Population-Level Variance in Gene Expression Identifies Key Differences Between Tobacco- and HPV-Associated Oropharyngeal Carcinogenesis ». In : *Arch. Pathol. Lab. Med.* 139.11, p. 1362–1372.

Benidt, S. et D. Nettleton (2015). « SimSeq : a nonparametric approach to simulation of RNA-sequence datasets ». In : *Bioinformatics* 31.13, p. 2131–2140.

Benjamini, Yoav et Yosef Hochberg (1995). « Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, p. 289–300.

Benjamini, Yoav et Daniel Yekutieli (2001). « The control of the false discovery rate in multiple testing under dependency ». In : *Ann. Statist.* 29.4, p. 1165–1188.

Bi, R. et P. Liu (2016). « Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments ». In : *BMC Bioinformatics* 17, p. 146.

Bonafede, E. et al. (2016). « Modeling overdispersion heterogeneity in differential expression analysis using mixtures ». In : *Biometrics* 72.3, p. 804–814.

Bossel Ben-Moshe, N. et al. (2012). « Context-specific microRNA analysis : identification of functional microRNAs and their mRNA targets ». In : *Nucleic Acids Res.* 40.21, p. 10614–10627.

Bourgon, R., R. Gentleman et W. Huber (2010). « Independent filtering increases detection power for high-throughput experiments ». In : *Proc. Natl. Acad. Sci. U.S.A.* 107.21, p. 9546–9551.

Brent, R. P. (Richard Peirce) (1973). *Algorithms for minimization without derivatives*. Englewood Cliffs, N.J. : Prentice-Hall. isbn : 0130223352.

Broad Institute (2015). *Picard tools*. http://broadinstitute.github.io/picard/.

Brown, Morton B. et Alan B. Forsythe (1974). « Robust Tests for the Equality of Variances ». In : *Journal of the American Statistical Association* 69.346, p. 364–367.

Calin, G. A. et C. M. Croce (2006). « MicroRNA signatures in human cancers ». In : *Nat. Rev. Cancer* 6.11, p. 857–866.

Cameron, A. Colin et Pravin K. Trivedi (1998). *Regression Analysis of Count Data*. Econometric Society Monographs. Cambridge University Press. isbn : 0521635675.

Cantoni, Eva et Elvezio Ronchetti (2001). « Robust Inference for Generalized Linear Models ». In : *Journal of the American Statistical Association* 96.455, p. 1022–1030.

Catalanotto, C., C. Cogoni et G. Zardo (2016). « MicroRNA in Control of Gene Expression : An Overview of Nuclear Functions ». In : *Int J Mol Sci* 17.10.

Chen, C. et al. (2011). « Removing batch effects in analysis of expression microarray data : an evaluation of six batch adjustment methods ». In : *PLoS ONE* 6.2, e17238.

Chen, C. Y. et al. (2009). « Ago-TNRC6 triggers microRNA-mediated decay by promoting two deadenylation steps ». In : *Nat. Struct. Mol. Biol.* 16.11, p. 1160–1166.

Cheung, V. G. et al. (2010). « Polymorphic cis- and trans-regulation of human gene expression ». In : *PLoS Biol.* 8.9.

Chou, C. H. et al. (2018). « miRTarBase update 2018 : a resource for experimentally validated microRNA-target interactions ». In : *Nucleic Acids Res.* 46.D1, p. D296–D302.

Chu, Andy et al. (2016). « Large-scale profiling of microRNAs for The Cancer Genome Atlas ». In : *Nucleic Acids Research* 44.1, e3.

Conesa, A. et al. (2016). « A survey of best practices for RNA-seq data analysis ». In : *Genome Biol.* 17, p. 13.

Cox, D. R. et N. Reid (1987). « Parameter Orthogonality and Approximate Conditional Inference ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 49.1, p. 1–39.

Daily, K. et al. (2017). « Molecular, phenotypic, and sample-associated data to describe pluripotent stem cell lines and derivatives ». In : *Sci Data* 4, p. 170030.

Daye, Z. J., J. Chen et H. Li (2012). « High-Dimensional Heteroscedastic Regression with an Application to eQTL Data Analysis ». In : *Biometrics* 68.1, p. 316–326.

De Jong, T. V., Y. M. Moshkin et V. Guryev (2019). « Gene expression variability : the other dimension in transcriptome analysis ». In : *Physiol. Genomics* 51.5, p. 145–158.

DeLuca, D. S. et al. (2012). « RNA-SeQC : RNA-seq metrics for quality control and process optimization ». In : *Bioinformatics* 28.11, p. 1530–1532.

Dillies, M. A. et al. (2013). « A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis ». In : *Brief. Bioinformatics* 14.6, p. 671–683.

Dobin, A. et al. (2013). « STAR : ultrafast universal RNA-seq aligner ». In : *Bioinformatics* 29.1, p. 15–21.

Dunn, O. J. (1961). « Multiple Comparisons among Means ». In : *Journal of the American Statistical Association* 56.293, p. 52–64.

Ebert, M. S., J. R. Neilson et P. A. Sharp (2007). « MicroRNA sponges : competitive inhibitors of small RNAs in mammalian cells ». In : *Nat. Methods* 4.9, p. 721–726.

Ebert, M. S. et P. A. Sharp (2012). « Roles for microRNAs in conferring robustness to biological processes ». In : *Cell* 149.3, p. 515–524.

Ecker, S. et al. (2015). « Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia ». In : *Genome Med* 7.1, p. 8.

Emmert-Streib, F., S. Tripathi et R. de Matos Simoes (2012). « Harnessing the complexity of gene expression data from cancer : from single gene to structural pathway methods ». In : *Biol. Direct* 7, p. 44.

Fabian, M. R. et N. Sonenberg (2012). « The mechanics of miRNA-mediated gene silencing : a look under the hood of miRISC ». In : *Nat. Struct. Mol. Biol.* 19.6, p. 586–593.

Faridani, O. R. et al. (2016). « Single-cell sequencing of the small-RNA transcriptome ». In : *Nat. Biotechnol.* 34.12, p. 1264–1266.

Feltz, C. J. et G. E. Miller (1996). « An asymptotic test for the equality of coefficients of variation from k populations ». In : *Statistics in Medicine* 15.6, p. 647–658.

Fligner, M.A. et T. J. Killeen (1976). « Distribution-Free Two-Sample Tests for Scale ». In : *J. Amer. Statist. Assoc.* 71, p. 210–213.

Garofalo, M. et al. (2010). « MicroRNAs as regulators of death receptors signaling ». In : *Cell Death Differ.* 17.2, p. 200–208.

Genome Reference Consortium (2017). *Human Genome Assembly GRCh38.p11*. https://www.ncbi.nlm.nih.gov/grc/human/data.

Goeman, J. J. et al. (2004). « A global test for groups of genes : testing association with a clinical outcome ». In : *Bioinformatics* 20.1, p. 93–99.

Guo, H. et al. (2010). « Mammalian microRNAs predominantly act to decrease target mRNA levels ». In : *Nature* 466.7308, p. 835–840.

Hammer, P. et al. (2010). « mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain ». In : *Genome Res.* 20.6, p. 847–860.

Han, R. et al. (2016). « Increased gene expression noise in human cancers is correlated with low p53 and immune activities as well as late stage cancer ». In : *Oncotarget* 7.44, p. 72011–72020.

Hanahan, D. et R. A. Weinberg (2000). « The hallmarks of cancer ». In : *Cell* 100.1, p. 57–70.

Haque, A. et al. (2017). « A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications ». In : *Genome Med* 9.1, p. 75.

Harrow, J. et al. (2012). « GENCODE : the reference human genome annotation for The ENCODE Project ». In : *Genome Res.* 22.9, p. 1760–1774.

Hasegawa, Y. et al. (2015). « Variability of Gene Expression Identifies Transcriptional Regulators of Early Human Embryonic Development ». In : *PLoS Genet.* 11.8, e1005428.

He, L. et al. (2005). « A microRNA polycistron as a potential human oncogene ». In : *Nature* 435.7043, p. 828–833.

Ho, J. W. et al. (2008). « Differential variability analysis of gene expression and its application to human diseases ». In : *Bioinformatics* 24.13, p. i390–398.

Hrustincova, A., H. Votavova et M. Dostalova Merkerova (2015). « Circulating MicroRNAs : Methodological Aspects in Detection of These Biomarkers ». In : *Folia Biol. (Praha)* 61.6, p. 203–218.

Huber, Peter J. (1964). « Robust Estimation of a Location Parameter ». In : *Ann. Math. Statist.* 35.1, p. 73–101.

Hutvágner, G. et P. D. Zamore (2002). « A microRNA in a multiple-turnover RNAi enzyme complex ». In : *Science* 297.5589, p. 2056–2060.

Hutvágner, György et al. (2001). « A Cellular Function for the RNA-Interference Enzyme Dicer in the Maturation of the let-7 Small Temporal RNA ». In : *Science* 293.5531, p. 834–838.

Hwang, Yi-Ting, Shih-Kai Chu et Shyh-Tyan Ou (2011). « Evaluations of FDR-controlling procedures in multiple hypothesis testing ». In : *Statistics and Computing* 21.4, p. 569–583.

Illumina (2017). *An Introduction to Next-Generation Sequencing Technology*. https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf.

Iterson, M. van et al. (2013). « Integrated analysis of microRNA and mRNA expression : adding biological significance to microRNA target predictions ». In : *Nucleic Acids Res.* 41.15, e146.

Ivey, K. N. et D. Srivastava (2010). « MicroRNAs as regulators of differentiation and cell fate decisions ». In : *Cell Stem Cell* 7.1, p. 36–41.

John, B. et al. (2004). « Human MicroRNA targets ». In : *PLoS Biol.* 2.11, e363.

Johnson, W. E., C. Li et A. Rabinovic (2007). « Adjusting batch effects in microarray expression data using empirical Bayes methods ». In : *Biostatistics* 8.1, p. 118–127.

Kertesz, M. et al. (2007). « The role of site accessibility in microRNA target recognition ». In : *Nat. Genet.* 39.10, p. 1278–1284.

Kester, L. et A. van Oudenaarden (2018). « Single-Cell Transcriptomics Meets Lineage Tracing ». In : *Cell Stem Cell* 23.2, p. 166–179.

Khvorova, A., A. Reynolds et S. D. Jayasena (2003). « Functional siRNAs and miRNAs exhibit strand bias ». In : *Cell* 115.2, p. 209–216.

Kim, V. N. (2005). « MicroRNA biogenesis : coordinated cropping and dicing ». In : *Nat. Rev. Mol. Cell Biol.* 6.5, p. 376–385.

Kitano, H. (2004). « Biological robustness ». In : *Nat. Rev. Genet.* 5.11, p. 826–837.

Kozomara, A. et S. Griffiths-Jones (2014). « miRBase : annotating high confidence microRNAs using deep sequencing data ». In : *Nucleic Acids Res.* 42.Database issue, p. 68–73.

Krishnamoorthy, K. et Meesook Lee (2014). « Improved Tests for the Equality of Normal Coefficients of Variation ». In : *Comput. Stat.* 29.1-2, p. 215–232.

Krutzfeldt, J. et al. (2005). « Silencing of microRNAs in vivo with 'antagomirs' ». In : *Nature* 438.7068, p. 685–689.

Landau, W. M. et P. Liu (2013). « Dispersion estimation and its effect on test performance in RNA-seq data analysis : a simulation-based comparison of methods ». In : *PLoS ONE* 8.12, e81415.

Lange, Kenneth (2010). *Numerical analysis for Statisticians*. New York : Springer. isbn : 9781441959454.

Langmead, B. et S. L. Salzberg (2012). « Fast gapped-read alignment with Bowtie 2 ». In : *Nat. Methods* 9.4, p. 357–359.

Lawless, Jerald F (1987). « Negative binomial and mixed Poisson regression ». In : *Canadian Journal of Statistics* 15.3, p. 209–225.

Le Cessie, Saskia et Johannes (Hans van Houwelingen (1995). « Testing the Fit of a Regression Model Via Score Tests in Random Effects Models ». In : *Biometrics* 51, p. 600–14.

Lee, Y. et al. (2004). « MicroRNA genes are transcribed by RNA polymerase II ». In : *EMBO J.* 23.20, p. 4051–4060.

Levene, Howard (1960). « Robust Tests for Equality of Variances ». In : *Contributions to Probability and Statistics : Essays in Honor of Harold Hotelling*. Sous la dir. d'Ingram Olkin et al. Stanford, Calif : Stanford University Press, p. 278–292.

Levy, J. M. M., C. G. Towers et A. Thorburn (2017). « Targeting autophagy in cancer ». In : *Nat. Rev. Cancer* 17.9, p. 528–542.

Li, H. et R. Durbin (2009). « Fast and accurate short read alignment with Burrows-Wheeler transform ». In : *Bioinformatics* 25.14, p. 1754–1760.

Li, H. et al. (2009). « The Sequence Alignment/Map format and SAMtools ». In : *Bioinformatics* 25.16, p. 2078–2079.

Li, J. et A. T. Lamere (2018). « DiPhiSeq : Robust comparison of expression levels on RNA-Seq data with large sample sizes ». In : *Bioinformatics*.

Lin, Dekang (1998). « An Information-Theoretic Definition of Similarity ». In : *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML '98. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., p. 296–304.

Lord, D. (2006). « Modeling motor vehicle crashes using Poisson-gamma models : examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter ». In : *Accid Anal Prev* 38.4, p. 751–766.

Love, M. I., W. Huber et S. Anders (2014). « Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 ». In : *Genome Biol.* 15.12, p. 550.

Lu, J. et A. G. Clark (2012). « Impact of microRNA regulation on variation in human gene expression ». In : *Genome Res.* 22.7, p. 1243–1254.

Lund, E. et al. (2004). « Nuclear export of microRNA precursors ». In : *Science* 303.5654, p. 95–98.

Lund, S. P. et al. (2012). « Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates ». In : *Stat Appl Genet Mol Biol* 11.5.

Malik, R. et I. Roy (2008). « Design and development of antisense drugs ». In : *Expert Opin Drug Discov* 3.10, p. 1189–1207.

Mar, J. C. et al. (2011). « Variance of gene expression identifies altered network constraints in neurological disease ». In : *PLoS Genet.* 7.8, e1002207.

Marioni, J. C. et al. (2008). « RNA-seq : an assessment of technical reproducibility and comparison with gene expression arrays ». In : *Genome Res.* 18.9, p. 1509–1517.

Mason, E. A. et al. (2014). « Gene expression variability as a unifying element of the pluripotency network ». In : *Stem Cell Reports* 3.2, p. 365–377.

McCarthy, D. J., Y. Chen et G. K. Smyth (2012). « Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation ». In : *Nucleic Acids Res.* 40.10, p. 4288–4297.

McCullagh, P. et J.A. Nelder (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. isbn : 9780412317606.

Mencia, A. et al. (2009). « Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss ». In : *Nat. Genet.* 41.5, p. 609–613.

Miska, E. A. et al. (2007). « Most Caenorhabditis elegans microRNAs are individually not essential for development or viability ». In : *PLoS Genet.* 3.12, e215.

Mitchell, P. S. et al. (2008). « Circulating microRNAs as stable blood-based markers for cancer detection ». In : *Proc. Natl. Acad. Sci. U.S.A.* 105.30, p. 10513–10518.

National Cancer Institute (2017a). *Genomic Data Commons Data Portal*. https://portal.gdc.cancer.gov/.

— (2017b). *The Cancer Genome Atlas*. https://cancergenome.nih.gov/.

— (2017c). *Therapeutically Applicable Research To Generate Effective Treatments*. https://ocg.cancer.gov/programs/target.

Nelder, J. A. et R. W. M. Wedderburn (1972). « Generalized Linear Models ». In : *Journal of the Royal Statistical Society. Series A (General)* 135.3, p. 370–384.

Nocedal, Jorge (2006). *Numerical Optimization*. New York : Springer. isbn : 9780387400655.

Nygaard, V., E. A. Rodland et E. Hovig (2016). « Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses ». In : *Biostatistics* 17.1, p. 29–39.

O'Hara, Robert B. et D. Johan Kotze (2010). « Do not log-transform count data ». In : *Methods in Ecology and Evolution* 1.2, p. 118–122.

Park, Byung-Jung et Dominique Lord (2008). « Adjustment for maximum likelihood estimate of negative binomial dispersion parameter ». In : *Transportation Research Record* 2061.1, p. 9–19.

Pickrell, J. K. et al. (2010). « Understanding mechanisms underlying human gene expression variation with RNA sequencing ». In : *Nature* 464.7289, p. 768–772.

Pinzon, N. et al. (2017). « microRNA target prediction programs predict many false positives ». In : *Genome Res.* 27.2, p. 234–245.

Qiu, X. et al. (2016). « Circulating MicroRNA-26a in Plasma and Its Potential Diagnostic Value in Gastric Cancer ». In : *PLoS ONE* 11.3, e0151345.

Ran, D. et Z. J. Daye (2017). « Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq ». In : *Nucleic Acids Res.* 45.13, e127.

Rauschenberger, A. et al. (2016). « Testing for association between RNA-Seq and high-dimensional data ». In : *BMC Bioinformatics* 17, p. 118.

Resnik, Philip (1999). « Semantic Similarity in a Taxonomy : An Information-based Measure and Its Application to Problems of Ambiguity in Natural Language ». In : *J. Artif. Int. Res.* 11.1, p. 95–130.

Riffo-Campos, A. L., I. Riquelme et P. Brebi-Mieville (2016). « Tools for Sequence-Based miRNA Target Prediction : What to Choose ? » In : *Int J Mol Sci* 17.12.

Robinson, M. D., D. J. McCarthy et G. K. Smyth (2010). « edgeR : a Bioconductor package for differential expression analysis of digital gene expression data ». In : *Bioinformatics* 26.1, p. 139–140.

Robinson, M. D. et A. Oshlack (2010). « A scaling normalization method for differential expression analysis of RNA-seq data ». In : *Genome Biol.* 11.3, R25.

Robinson, M. D. et G. K. Smyth (2007). « Moderated statistical tests for assessing differences in tag abundance ». In : *Bioinformatics* 23.21, p. 2881–2887.

— (2008). « Small-sample estimation of negative binomial dispersion, with applications to SAGE data ». In : *Biostatistics* 9.2, p. 321–332.

Rosa, A. et A. H. Brivanlou (2009). « MicroRNAs in early vertebrate development ». In : *Cell Cycle* 8.21, p. 3513–3520.

Rosenbloom, K. R. et al. (2015). « The UCSC Genome Browser database : 2015 update ». In : *Nucleic Acids Res.* 43.Database issue, p. D670–681.

Rybstein, M. D. et al. (2018). « The autophagic network and cancer ». In : *Nat. Cell Biol.* 20.3, p. 243–251.

Salomonis, N. et al. (2016). « Integrated Genomic Analysis of Diverse Induced Pluripotent Stem Cells from the Progenitor Cell Biology Consortium ». In : *Stem Cell Reports* 7.1, p. 110–125.

Sanders, A. R. et al. (2017). « Transcriptome sequencing study implicates immune-related genes differentially expressed in schizophrenia : new data and a meta-analysis ». In : *Transl Psychiatry* 7.4, e1093.

Schlicker, A. et al. (2006). « A new measure for functional similarity of gene products based on Gene Ontology ». In : *BMC Bioinformatics* 7, p. 302.

Schmiedel, J. M. et al. (2015). « Gene expression. MicroRNA control of protein expression noise ». In : *Science* 348.6230, p. 128–132.

Seggerson, K., L. Tang et E. G. Moss (2002). « Two genetic circuits repress the Caenorhabditis elegans heterochronic gene lin-28 after translation initiation ». In : *Dev. Biol.* 243.2, p. 215–225.

Shah, M. Y. et G. A. Calin (2014). « MicroRNAs as therapeutic targets in human cancers ». In : *Wiley Interdiscip Rev RNA* 5.4, p. 537–548.

Shapiro, Samuel Sanford et Martin B Wilk (1965). « An Analysis of Variance Test for Normality (Complete Samples) ». In : *Biometrika* 52.3/4, p. 591–611.

Shen, M. et al. (2013). « Targeting the ubiquitin-proteasome system for cancer therapy ». In : *Expert Opin. Ther. Targets* 17.9, p. 1091–1108.

Siciliano, V. et al. (2013). « MiRNAs confer phenotypic robustness to gene networks by suppressing biological noise ». In : *Nat Commun* 4, p. 2364.

Silber, J. et al. (2008). « miR-124 and miR-137 inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells ». In : *BMC Med* 6, p. 14.

Smyth, Gordon K (1998). « Optimization and Nonlinear Equations ». In : *Encyclopedia of Biostatistics*, p. 3174–3180.

Soneson, C. (2014). « compcodeR–an R package for benchmarking differential expression methods for RNA-seq data ». In : *Bioinformatics* 30.17, p. 2517–2518.

Soneson, C. et M. Delorenzi (2013). « A comparison of methods for differential expression analysis of RNA-seq data ». In : *BMC Bioinformatics* 14, p. 91.

Stewart, G. W. (1973). *Introduction to matrix computations*. New York : Academic Press. isbn : 9780126703504.

Su, H. et al. (2011). « Mammalian hyperplastic discs homolog EDD regulates miRNA-mediated gene silencing ». In : *Mol. Cell* 43.1, p. 97–109.

Teschendorff, A. E. et T. Enver (2017). « Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome ». In : *Nat Commun* 8, p. 15599.

The Gene Ontology Consortium (2017). « Expansion of the Gene Ontology knowledgebase and resources ». In : *Nucleic Acids Res.* 45.D1, p. D331–D338.

Turchinovich, A. et al. (2011). « Characterization of extracellular circulating microRNA ». In : *Nucleic Acids Res.* 39.16, p. 7223–7233.

Vidigal, J. A. et A. Ventura (2015). « The biological functions of miRNAs : lessons from in vivo studies ». In : *Trends Cell Biol.* 25.3, p. 137–147.

Wang, K. et al. (2015). « EntropyExplorer : an R package for computing and comparing differential Shannon entropy, differential coefficient of variation and differential expression ». In : *BMC Res Notes* 8, p. 832.

Wang, Q. et al. (2018). « Unifying cancer and normal RNA sequencing data from different sources ». In : *Sci Data* 5, p. 180061.

Wedderburn, R. W. M. (1974). « Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method ». In : *Biometrika* 61.3, p. 439–447.

Witten, Daniela M. (2011). « Classification and clustering of sequencing data using a Poisson model ». In : *Ann. Appl. Stat.* 5.4, p. 2493–2518.

Wu, H., C. Wang et Z. Wu (2013). « A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data ». In : *Biostatistics* 14.2, p. 232–243.

— (2015). « PROPER : comprehensive power evaluation for differential expression using RNA-seq ». In : *Bioinformatics* 31.2, p. 233–241.

Xie, B. et al. (2013). « miRCancer : a microRNA-cancer association database constructed by text mining on literature ». In : *Bioinformatics* 29.5, p. 638–644.

Yang, Z. et al. (2012). « Preferential regulation of stably expressed genes in the human genome suggests a widespread expression buffering role of microRNAs ». In : *BMC Genomics* 13 Suppl 7, S14.

Yap, B. W. et C. H. Sim (2011). « Comparisons of various types of normality tests ». In : *Journal of Statistical Computation and Simulation* 81.12, p. 2141–2155.

Yu, D., W. Huber et O. Vitek (2013). « Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size ». In : *Bioinformatics* 29.10, p. 1275–1282.

Yu, G. et al. (2010). « GOSemSim : an R package for measuring semantic similarity among GO terms and gene products ». In : *Bioinformatics* 26.7, p. 976–978.

Yu, G. et al. (2012). « clusterProfiler : an R package for comparing biological themes among gene clusters ». In : *OMICS* 16.5, p. 284–287.

Yu, L., S. Fernandez et G. Brock (2017). « Power analysis for RNA-Seq differential expression studies ». In : *BMC Bioinformatics* 18.1, p. 234.

Zhang, F. et al. (2015). « Increased Variability of Genomic Transcription in Schizophrenia ». In : *Sci Rep* 5, p. 17995.

Zhao, J. J. et al. (2014a). « miR-30-5p functions as a tumor suppressor and novel therapeutic tool by targeting the oncogenic Wnt/B-catenin/BCL9 pathway ». In : *Cancer Res.* 74.6, p. 1801–1813.

Zhao, Z. et al. (2014b). « PD_NGSAtlas : a reference database combining next-generation sequencing epigenomic and transcriptomic data for psychiatric disorders ». In : *BMC Med Genomics* 7, p. 71.

Zwiener, I., B. Frisch et H. Binder (2014). « Transforming RNA-Seq data to improve the performance of prognostic gene signatures ». In : *PLoS ONE* 9.1, e85150.

Résumé

La majorité des études sur l'expression des gènes cherchent à identifier des gènes présentant des différences de moyenne d'expression entre plusieurs populations d'échantillons. Dans ce cadre, la variance est considérée comme un paramètre à contrôler. Cependant, à l'instar d'une différence de moyenne, une différence de variance d'expression de gènes entre populations d'échantillons peut avoir un sens biologique et physiologique.
Les microARN (miARN) sont d'importants régulateurs de l'expression des gènes. Le nombre important de leurs cibles et leur mode d'action confèrent aux miARN un rôle tampon. L'objectif de ma thèse est d'étudier la variance d'expression des miARN et des ARN messagers (ARNm), en particulier ceux ciblés par des miARN, durant la cancérogénèse. Nous espérons que cette approche permettra d'identifier des gènes qui ne peuvent pas être détectés par l'analyse classique de différence de moyenne d'expression et qui pourraient servir de biomarqueurs potentiels ou de cibles thérapeutiques. En outre, en combinant l'expression de miARN et d'ARNm et en analysant leur variance à une échelle systémique, nous espérons pouvoir mieux caractériser le rôle tampon des miARN.
Plusieur méthodes incluant des tests statistiques d'égalité de variance et des modèles basés sur la distribution binomiale négative ont été évaluées. Les performances de ces méthodes ont été étudiées en détails à l'aide de jeux de données simulées. Par la suite, elles ont été appliquées aux jeux de données *The Cancer Genome Atlas* dans le but d'identifier des gènes ayant une différence de variance d'expression entre échantillons sains et tumoraux. De nombreux miARN et ARNm présentant une augmentation de leur variance d'expression dans les tumeurs ont été détectés. Pour la plupart des cancers, certaines fonctions biologiques importantes telles que le catabolisme ou l'autophagie sont sur-représentées parmi ces ARNm. Ainsi, analyser des gènes différentiellement variants semble être une approche pertinente pour avoir une meilleure compréhension de la progression tumorale et devrait être prise en compte dans le cadre de la recherche de nouveaux biomarqueurs et cibles thérapeutiques potentiels.

Mots-clés : bioinformatique, cancer, séquençage, expression, variance, dispersion, microARN

Abstract

The majority of gene expression studies focus on looking for genes whose mean expression is different when comparing two or more populations of samples. In this context, the variance is treated as a parameter to be controlled. However, similarly to a difference of mean, a difference of variance in gene expression between sample populations may also be biologically and physiologically relevant.
MicroRNAs (miRNAs) are key gene expression regulators. The large number of their targets and the fine tuning of their regulation confer to miRNAs a buffering role. The objective of my thesis is to study the variance in expression of miRNAs and messenger RNAs (mRNAs), especially those targeted by miRNAs, in particular during cancerogenesis. We hope that this approach can identify genes which cannot be identified by the tradional differential expression analysis and yet serve as potential biomarkers or therapeutic targets. Furthermore, by combining both miRNA and mRNA expression and analyzing their variance at a system level, we aim at better characterize the buffering role of miRNAs.
Several methods including statistical tests of equality of variance and models based on the negative binomial distribution were evaluated. The performances of these methods were thoroughly tested on simulated datasets. Then, they were applied to The Cancer Genome Atlas datasets in order to identify genes with a differential expression variance when comparing normal and tumor samples. Many miRNAs and mRNAs with an increase of expression variance in tumors were detected. Interestingly, among these mRNAs, some key biological functions such as catabolism or autophagy are over-represented in most cancers. Thus, analyzing genes having a differential expression variance is relevant to gain knowledge in tumor progression and opens a new space for the discovery of new potential biomarkers and therapeutic avenues.

Keywords : bioinformatics, cancer, sequencing, expression, variance, dispersion, microRNA