# Hackathon

# Analyse de l’implication de la mutation au codon 625 du gène SF3B1 dans le mélanome uvéal

Théo Roncalli

Gustavo Magnaña López

Izem Mouhoubi

Université de Paris-Saclay

Master 1 Bioinformatique et biostatistiques (BIBS)

2020-2021

# Table des matières

# Introduction

Furney et al. (2013) et Harbour et al. (2013) ont remarqué une mutation récurrente au codon 625 de la sous-unité 1 du facteur d'épissage 3B (SF3B1) chez des individus atteints du mélanome uvéal de type 1. Leurs travaux ont consisté à étudier l'influence de cette mutation dans l'expression génique des individus atteints du type de cancer précédemment cité. La reproductibilité est un des principe fondamentaux de la méthode scientifique. Le présent travail consiste à réaliser un traitement et une analyse reproductible de données RNA-seq pour mettre en évidence une expression différentielle et un épissage alternatif entre les deux condition physio-pathologiques pour huit individus. Les données que nous utilisons reprennent celles utilisées par Furney et al. (2013) et Harbour et al. (2013).

Durant ce projet, nous vérifions que le travail soit reproductible par tous et non variable dans le temps. En effet, un travail de recherche doit respecter une rigueur scientifique, dont la reproductibilité en fait partie. La reproductibilité est souvent citée comme une des caractéristiques d'une « bonne science » et concerne tous les domaines scientifiques tels que la physique, la biologie, l'informatique, etc.

L'amélioration des connaissances scientifiques reposent sur la nature reproductible des résultats. La science acquise aujourd'hui s'appuie sur les découvertes d'hier. La reproductibilité du travail scientifique ne représente pas les objectifs finaux de la science, mais constitue un aspect essentiel de l'approche scientifique. La non-reproductibilité est un phénomène touchant tous les domaines (ML, médical, pharmaceutiques, etc.). Notre intérêt se porte précisément sur la reproductibilité computationnelle. Les expériences computationnelles manquent souvent de précisions dans les rapports et articles, et les résultats sont brièvement décris dans les figures et le script est souvent machine-subjectif.

Pour permettre la reproductibilité computationnelle des travaux de recherche, des outils ont été développés. Le logiciel Git s'occupe de la gestion des versions de code informatique. Docker permet quant à lui de pallier le problème de code machine-subjectif et améliore la portabilité du code en utilisant un système de conteneurisation, permettant ainsi de créer un environnement spécifique isolé contenant toutes les dépendances nécessaires à l'exécution d'un logiciel. Enfin, Nextflow est un système de gestion de workflow en bioinformatique permettant la portabilité, la gestion de logiciel multiples, un temps d'exécution rapide grâce à la parallélisation mais surtout la reproductibilité des pipelines d'analyse grâce à la gestion des versions des logiciels. Grâce aux trois outils énumérés, l'analyse des données RNA-Seq des travaux de Furney et al. (2013) et Harbour et al. (2013) a été réalisée.

Harbour el al. (2013) ont mis en évidence la présence de la mutation SF3B1 pour des données exomiques de 15 patients et ont montré, sur un autre jeu de données de 19 patients, que l'allèle qui porte la mutation SF3B1 est présent en proportion identique avec l'allèle sauvage. Les auteurs indiquent que la mutation SF3B1 est présente dans les tumeurs bénignes (classe 1) avec peu de métastase et peu de cellules épithélioïdes. Toutefois, les auteurs ne trouvent pas d'associations entre la mutation SF3B1 et un possible épissage alternatif.

Furney et al. (2013) réutilisent le même jeu de données (8 échantillons) que clui de Harbour et al. (2013) et découvrent une association entre des phénomènes d'épissages alternatifs et la mutation SF3B1. Les auteurs montrent que cette mutation est associée à 3 type d'épissage alternatif : rétention d'introns, épissage sur sites cryptiques et épissage en site 3' alternatif. Ces 3 épissages peuvent se produire sur des exons de gène tel que ABCC5 ou des gènes non codant tel que CRNDE. Puisque la mutation se produit sur un facteur d'épissage (composant du spliceosome), il est possible que ces épissages alternatifs soient impliqués dans le phénotype cancéreux des patients.

L'objectif est de reproduire l'analyse RNA-seq d'expression génique et d'épissage alternatif, pour trouver les gènes qui s'expriment différentiellement ou différemment pour les 4 échantillons portant la mutation contre les 4 échantillons sauvage ne portant pas la mutation au codon 625 sur le gène SF3B1.

# Organisation de l'équipe

La réalisation du projet s'est déroulée en étroite collaboration entre les différents membres de l'équipe, hormis Alex qui a été malheureusement très discret durant ce projet. Au début du projet, les tâches n'ont pas réellement été réparties, car le logiciel Nextflow était nouveau pour chacun d'entre nous et nous avions tous besoin de l'aide de chacun pour réaliser les premiers processus. Nous collaborions étroitement via Discord. Une fois que le logiciel a été pris en main par les différents membre de l'équipe (c'est-à-dire relativement tard), les tâches ont été davantage séparées et Github et son système de branches ont été très utilisés. Nous énumérons ci-dessous les tâches globalement effectués par chaque membre, mais en réalité chacun a été d'une grande aide pour les autres lors de certains bugs.

Alex a débugué un problème sur le logiciel Filezilla.

Izem s'est chargé de la recherche et la vérification des données à utiliser sur le NCBI. Il a également créé un script nextflow en dsl1 jusqu'à l'étape de création d'index. Il a fortement contribué à trouver les commandes bash nécessaires pour le projet, ce qui a permis aux autres membres de l'équipe de gagner beaucoup de temps. Ses connaissances en biologie ont permis d'éclairer les autres membres du groupe sur l'objectif du projet et les résultats des articles à lire. Izem s'est également intéressé à la qualité des reads.

Gustavo et Théo ont principalement contribué à la création du script Nextflow en dsl2 de l'étape initiale jusqu'à l'étape finale. C'est ce script que nous délivrons pour le projet. Ils ont tous deux aidé l'autre pour les débogages. Les connaissances de Gustavo en informatique ont été très appréciées. Théo s'est occupé de la gestion de la machine virtuelle.

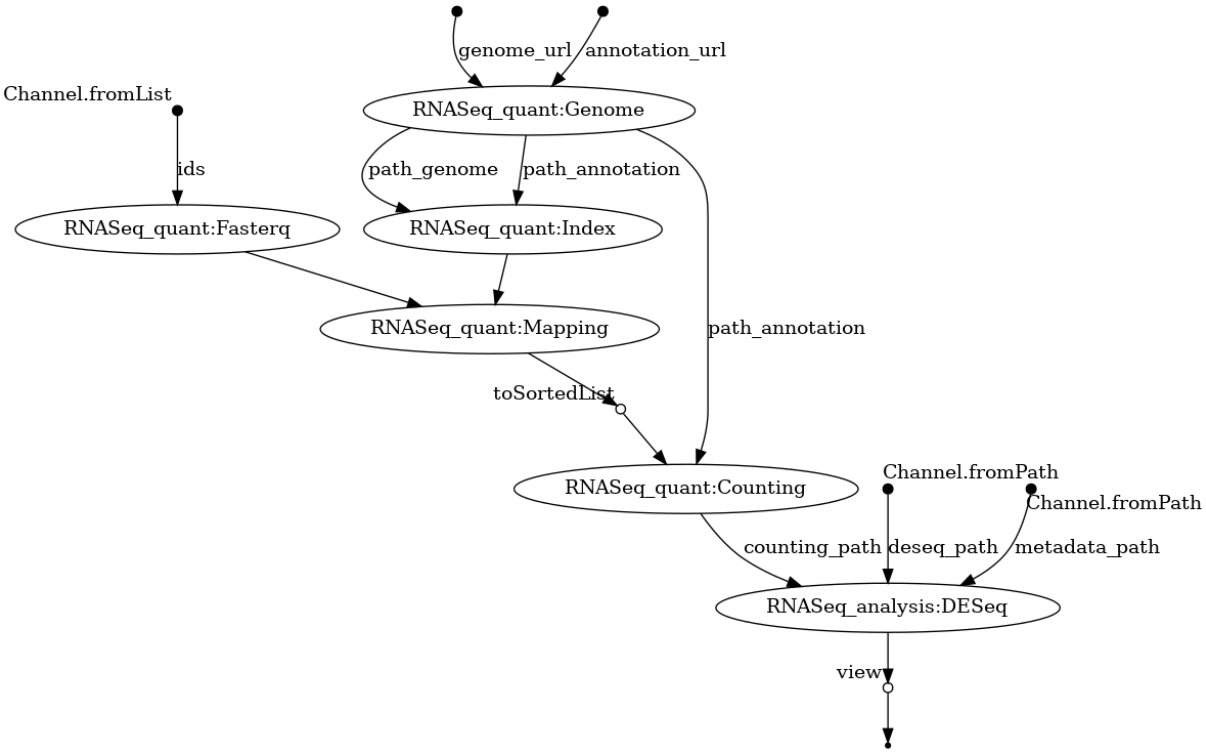
Izem et Théo ont généré un script DESeq permettant d'obtenir les figures pour les résultats biologiques et se sont intéressés à l'analyse de l'épissage alternatif en plus de l'expression différentielle génique. Ils ont également co-écrit le rapport.

Gustavo a beaucoup bénéficié à l'équipe concernant la gestion de Github (et surtout de son système de branches). Il s'est également occupé de créer la documentation sur Github ainsi que le fichier README.md.

# Résultats

## Workflow

Afin de veiller à ce que le projet soit reproductible dans le temps, un script Nextflow a été rédigé pour réaliser toutes les étapes du projet, allant de la collecte et la préparation des données à l'analyse de l'expression génique différentielle. Le script Nextflow permet de garantir la stabilité de nos résultats et leur reproductibilité grâce au système de gestion de versions d'environnement. De plus, tous les processus sont réalisés dans des conteneurs Docker qui sont isolés du système d'exploitation de la machine de l'utilisateur. L'exécution du workflow permet de réaliser l'analyse différentielle des expressions géniques des séquençages à haut débits obtenus via Illumina HiSeq 2000 et accessibles avec le numéro d'accès SRA062359. Les résultats, qui offrent de multiples figures d'analyse différentielle, sont contenues dans un même répertoire *Figures* dont le chemin d'accès est spécifié à la fin de l'exécution du workflow. Le code est portable et est exécutable en un temps relativement rapide (environ deux heures) grâce à la parallélisation. Le workflow bioinformatique utilisé exécute une suite de tâches organisées comme suit :




Deux sous-workflows sont utilisés : RNASeq_quant et RNASeq_analysis. Le premier permet de collecter et de préparer les données jusqu'à obtenir la table de comptage et le second permet de réaliser l'analyse statistique d'expression différentielle. On remarque de nombreux processus au sein du sous-workflow RNASeq_quant. Pour commencer, nous avons l'importation des données, avec le processus Fasterq pour récupérer les fichiers de reads au format fastq et le processus Genome permettant de récupérer le génome de référence au format fasta ainsi que le fichier d'annotation au format gtf. Ensuite, le processus Index est lancé pour indexer le génome de référence. Une fois ces étapes terminées, le processus Mapping est exécuté pour chaque fichier de reads au format fastq qui renvoie un fichier d'output. Lorsque chaque fichier BAM est généré, le processus Counting permet de créer la table de comptage pour chaque gène et pour chaque individu. Une fois ce processus terminé, la table de comptage généré est envoyée au sous-workflow RNASeq-analysis et plus précisément au processus DESeq pour réaliser l'analyse différentielle statistique via la librairie DESeq2 sur le logiciel R. Ce dernier processus renvoie toutes les figures générées pour l'analyse d'expression différentielle.

Pour utiliser le workflow, il faut avoir une machine d'au moins 32 Go RAM. Si ce n'est pas le cas, le script ne pourra pas fonctionner. Également, et par précaution, les paramètres de configuration de base sont proposés

pour une machine de 16 cœurs et 64 Go RAM (si vous souhaitez utiliser une machine plus petite de 8 cœurs ou 32 Go RAM, veuillez modifier les paramètres par défaut. Plus de détails sont fournis).

Pour utiliser le workflow, utiliser les commandes suivantes :

*cd Nextflow*

*nextflow run main.nf*

Il est possible que l'utilisateur rencontre des problèmes lors de son exécution. Si le problème est lié à une erreur réseau ou erreur système non spécifique aux caractéristiques de la machine – par exemple core dump (segmentation default) –, vous pouvez utiliser les points de contrôle continus afin de ne pas réexécuter les processus ayant déjà réussi avec succès. Pour cela, veuillez utiliser la commande suivante qui récupérera en caches les fichiers déjà compilés :

*nextflow run main.nf -resume*

Dans le cas où le problème vient d'un problème spécifique aux performances de la machine de l'utilisateur, il est possible de modifier les paramètres de gestion d'utilisation des ressources de la machine. Veuillez donc spécifier la valeur des paramètres que vous souhaitez modifier avec la commande suivante :

*nextflow run main.nf --<parameter 1> <value 1> … <parameter n> <value n>*

Attention, tous les paramètres que nous présenterons par la suite sont optionnels. Les différents paramètres d'allocation des ressources sont les suivants :

- index_cpus (nombre de cœurs alloués au processus d'indexation du génome étudié. Par défaut : 14)
- mapping_cpus (nombre de cœurs alloués au processus de mapping pour la création des fichiers BAM. Par défaut : 14)
- counting_cpus (nombre de cœurs alloués au processus de comptage. Par défaut : 7)
- mapping_memory (mémoire allouée au processus de mapping. Par défaut : '50GB')

La valeur par défaut de la mémoire allouée au mapping est très au-dessus de la valeur requise pour le cas présent. Une valeur fixée à '30GB' fonctionne avec les données RNA-Seq utilisée pour le projet.

D'autres paramètres sont ajustables. Par exemple, si vous disposez déjà de certaines données en locales et ne souhaitez par les recalculer, vous pouvez utiliser les paramètres suivants :

- reads (Répertoire contenant les fichiers fasterq. Par défaut : nul)
- genome (Répertoire contenant le fichier fasta du génome. Par défaut : nul)
- index (Répertoire contenant les fichiers d'index. Par défaut : nul)
- mapping (Répertoire contenant les fichiers BAM. Par défaut : nul)
- counting (Chemin d'accès entier au fichier de comptage – comprend le fichier lui-même. Par défaut : nul)
- metadata (Chemin d'accès entier au fichier de métadonnées – comprend le fichier lui-même. Par défaut : SraRunTable.txt)

Lorsque le chemin d'accès est nul, les données sont téléchargées ou calculées. Parmi les données à télécharger, nous avons les fichiers paired-end au format fastq, le génome de référence au format fasta et le fichier d'annotation du génome de référence au format gtf. Bien que le projet s'inscrive dans une démarche de reproductibilité computationnelle, nous pouvons utiliser les paramètres précédemment cités pour utiliser le workflow sur de nouvelles données ou un autre sujet dont la démarche est similaire. Dans ce contexte, il est donc également possible de modifier les données à télécharger via les paramètres suivants :

- ids (Liste des numéro d'accès SRR pour les fichiers paired-end. Par défaut : ['SRR628582', 'SRR628583', 'SRR628584', 'SRR628585', 'SRR628586', 'SRR628587', 'SRR628588', 'SRR628589'])

- genome_url (URL de téléchargement du génome de référence. Par défaut : ftp://ftp.ensembl.org/pub/release-101/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz)
- annotation_url (URL de téléchargement de l'annotation du génome de référence. Par défaut : ftp://ftp.ensembl.org/pub/release-101/gtf/homo_sapiens/Homo_sapiens.GRCh38.101.chr.gtf.gz)

Un dernier paramètre concerne le lancement de STAR :

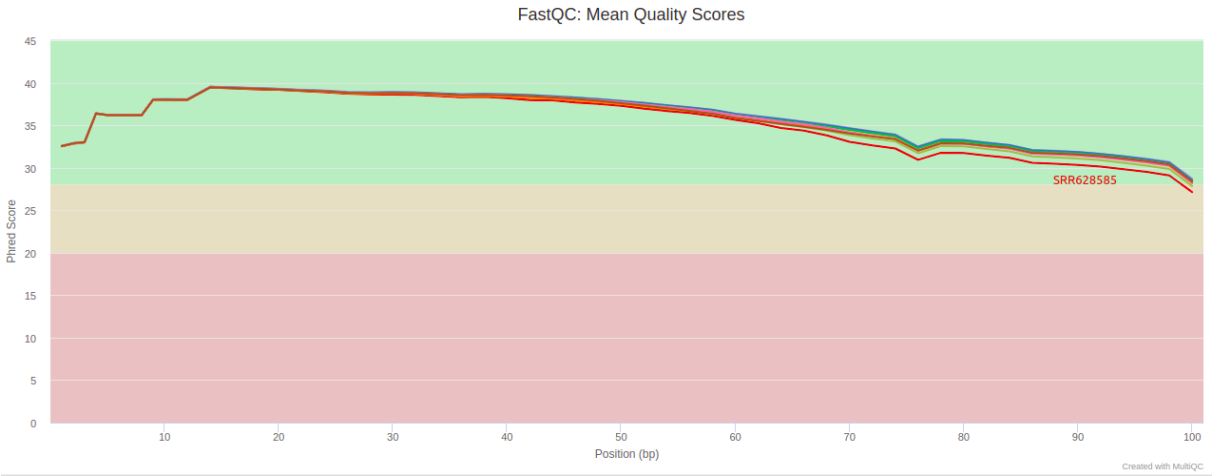
- sjdbOverhang (Par défaut : 99)

La pipeline propose donc une interface user-friendly avec de multiples paramètres afin de donner plus de souplesse à son utilisateur. De plus, si l'option *counting* est spécifiée, seul le sous-workflow RNASeq_analysis sera exécuté puisque l'utilisateur dispose déjà de la table de comptage. Dans ce cas-là, l'utilisateur devra toutefois veiller à ce que les individus soient dans le même ordre dans la table de comptage que dans le fichier de métadonnées utilisé pour la partie statistique.

## Contrôle qualité des données

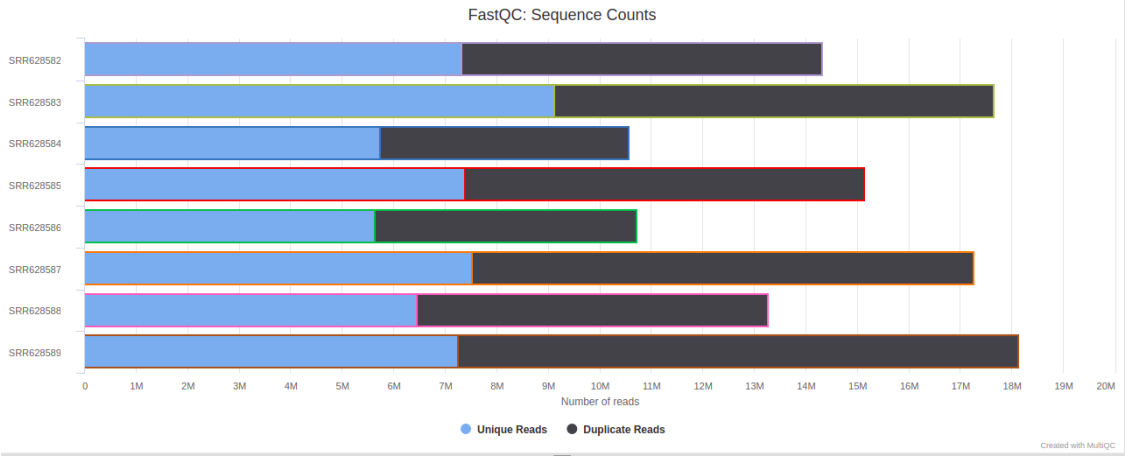
Nos résultats concernent principalement l'analyse différentielle des gènes et, dans une moindre mesure, l'épissage alternatif. La qualité des reads paired-end de taille 100 issus des huit échantillons ont été soumis à un contrôle qualité. Figure 1 fournit les scores Phred, métrique permettant de mesurer la qualité de séquençage. Il est proportionnel à la probabilité qu'une base soit correcte. On remarque qu'au début et à la fin du séquençage, le score Phred est relativement moins bon. Le Q score moyen est de 35, ce qui signifie que nous avons une précision d'identification d'environ 99.9% pour une base (1 chance sur 1000 que la base soit incorrecte). A la fin de séquençage, le Q score moyen descend à environ 27. Dans l'ensemble, les données de séquençage sont de très bonne qualité. Il est possible d'utiliser l'outil de trimming pour enlever les bases de faible qualité sur les extrémités 3' et 5' de nos reads. Furney et al. (2013) ont procédé au trimming en enlevant la dernière base de chaque read.

Figure 1 : Analyse de la qualité des reads




La quantité de séquences récupérées varie selon les échantillons. En moyenne, nous comptons environ 14.35 millions de reads par échantillon. On observe également un taux de séquences dupliqués correct qui varie entre 50% et 56% pour les échantillons. L'échantillon SRR628589 possède un taux de séquences dupliquées de 60% (Figure 2), ce qui pourrait être dû à quelques séquences qui seraient surreprésentées dans cet échantillon. Cela pourrait résulter de la taille des librairies avant le séquençage.

Figure 2 : Analyse de la quantité de séquences pour chaque échantillon




Concernant l'analyse des séquences, on observe un taux de GC équilibré (50% environ) pour tous les échantillons hormis le SRR628689 pour lequel ce taux s'élève à 60%. Le taux GC renseigne sur la densité génique dans l'ADN.

## Traitement des données

Le traitement des données consiste à créer l'index, le mapping puis le comptage de reads par gène. L'index a été généré à partir du génome humain GRCh38 proposé dans la Section *Méthodes*. Ce processus prend une trentaine de minutes à s'exécuter pour 14 cœurs et 30 Go de mémoire vive. Harbour et al. (2013) utilisent le génome hg19 et créent l'index avec le logiciel Bowtie2. Le Mapping consiste à créer huit fichiers BAM par reads triés selon les coordonnées chromosomiques du génome humain. Le temps estimé pour faire le mapping est d'environ dix minutes par tâche, soit un total d'une heure vingt en moyenne. La taille de chaque fichier BAM est d'environ 3 Go. Pour l'alignement, Furney et al. (2013) utilisent le logiciel TopHat et Harbour et al. (2013) utilisent le logiciel Bowtie2.

Concernant l'étape de comptage, nous avons testé deux solutions. Depuis la version 2.5, le logiciel STAR propose un mode de comptage de gène pendant le mapping. STAR compte si le read chevauche un et un seul gène. Le logiciel vérifie cette condition pour les deux fichiers paired-end au format fastq. Nous avons également utilisé FeatureCounts qui a le même critère de comptage. Toutefois, nous avons remarquer des différences significatives dans les comptages entre ces deux logiciels. Le nombre de reads mappés est beaucoup plus important avec le fonction FeatureCounts. Presque deux fois plus de reads sont assignés avec FeatureCounts contrairement à son homologue STAR (voir Figure 3 et 4). Remarquons également que le mode de comptage STAR a une proportion de reads non alignés (no feature) plus faible que FeatureCounts. En moyenne, six pourcents des résultats provenant de STAR sont des reads non alignés, contre quinze pourcents en moyenne avec FeatureCounts. STAR assigne environ neuf pourcents des reads comme multi-mapping, c'est-à-dire qu'il y a des reads qui s'alignent sur plus d'un gène dans le génome. Puisque nous alignons sur le génome complet, il est compréhensible d'avoir dans cet ADN des gènes présents en copies multiples. FeatureCounts n'assigne aucune séquence dans la catégorie multi-mapping car nous n'avons pas spécifié l'option qui le permet (option -O). Dans la documentation du package Subread, il est recommandé de ne pas utiliser l'option multi-mapping pour les expériences RNA-seq, car en théorie, un fragment d'ADNc (ARN) devrait être originaire d'un seul gène[1]. FeatureCounts désigne les séquences "Unassigned Ambiguity" comme étant des séquences qui chevauchent 2 exons ou plus. Ce logiciel classe environ 17% des reads dans la catégorie "Unassigned Ambiguity" contre 9 % des reads comme étant "Ambiguous" pour STAR. Comme le nombre de reads mappés est plus important avec FeatureCounts, nous avons retenu la table de comptage générée par cette dernière pour l'analyse statistique d'expression différentielle. Furney et al. (2013) ont quant à eux utilisé le logiciel Htseq pour générer la table de comptage.

[1] Voir page 36 de la documentation.

Figure 3 : statistiques sur les comptages avec FeatureCounts

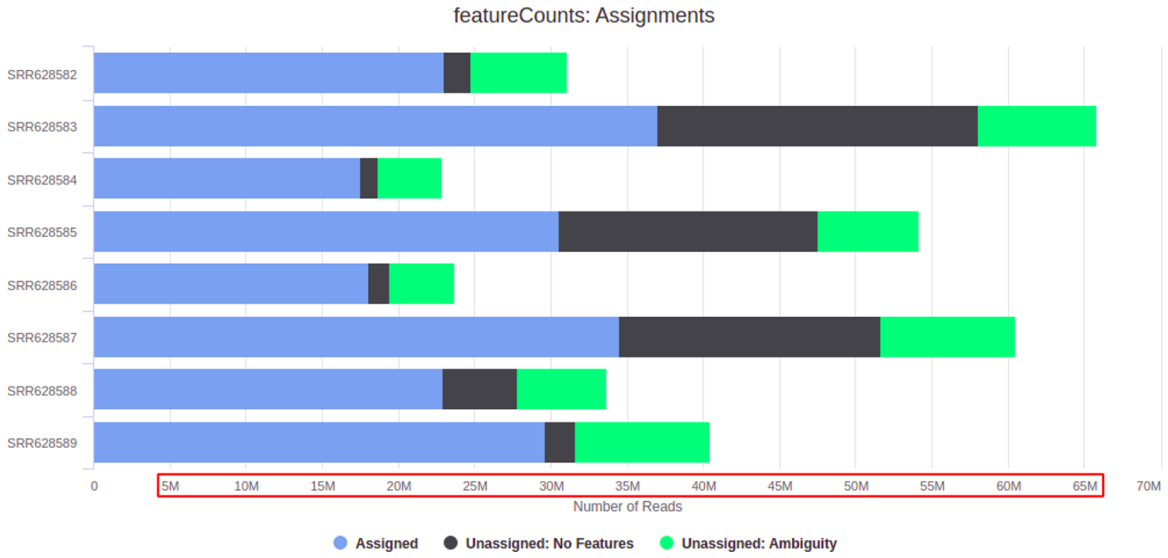
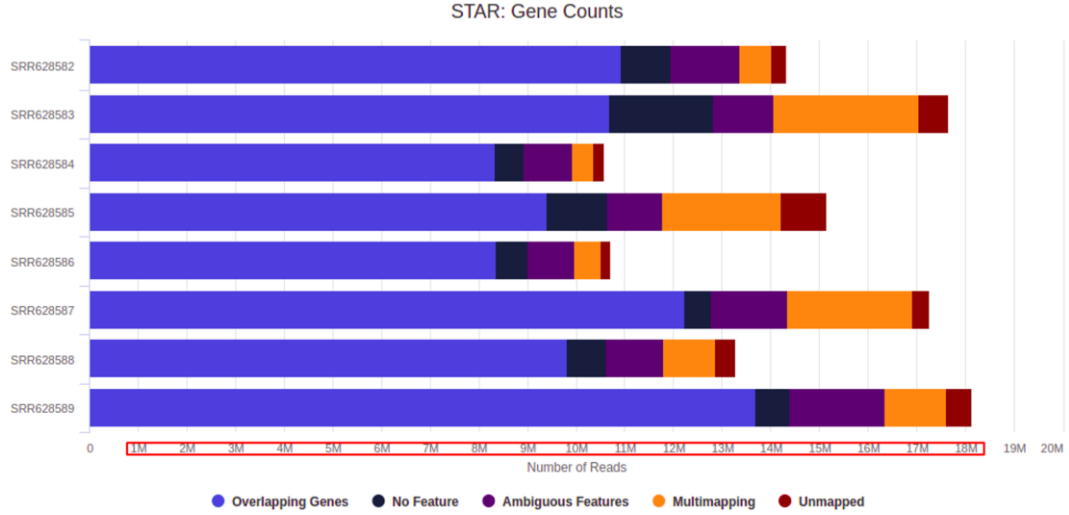



Figure 4 : Statistiques sur les comptages avec STAR



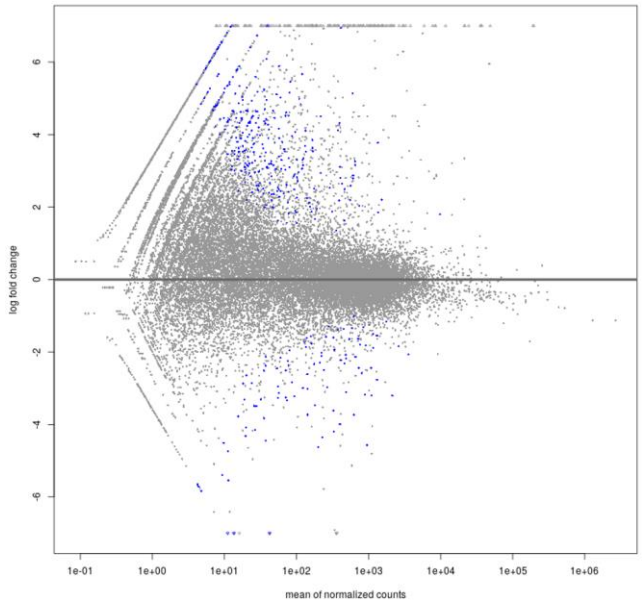

## Analyse Statistique et résultats biologiques

Nous avons utilisé les comptages produits par FeatureCounts pour la suite de cette analyse. La librairie DESeq2 sur R utilise un modèle linéaire généralisé pour modéliser les données de comptage RNA-seq. On suppose que les reads suivent une distribution de loi binomiale négative (méthode paramétrique). Furney et al. (2013) utilisent DEXseq pour l'analyse de l'expression différentielle des gènes. Cette librairie issue de BioConductor utilise aussi un modèle linéaire généralisé et suppose une distribution selon une loi binomiale négative pour les reads. Harbour et al. (2013) utilisent la librairie SAMseq qui utilisent une méthode non paramétrique.

La pipeline DESeq2 (Love et al. ,2014) implémente la méthode de normalisation appelée *méthode des ratios médians*. Elle trouve le rapport entre chaque nombre de reads et la moyenne géométrique de tous les nombres de reads pour ce gène pour chaque échantillon (Anders et Huber, 2010) afin d'effacer le biais technique dû au séquençage et de remettre à l'échelle les gènes non différentiellement exprimés. Nous séparons nos conditions entre les individus ayant subi une mutation du gène SF3B1 et ceux n'en ayant pas subi. Figure 5 fournit le MA-plot et permet donc de montrer les sous-expressions et sur-expressions des gènes entre les deux conditions en fonction du comptage moyen normalisé. S'il n'y avait pas d'expression différentielle, nous aurions des points uniformément répartis autour de l'axe horizontal. Or nous avons ici une forme de poisson, ce qui signifie que les gènes peu exprimés sont très fortement sous-exprimés ou sur-exprimés par rapport aux patients n'ayant pas de

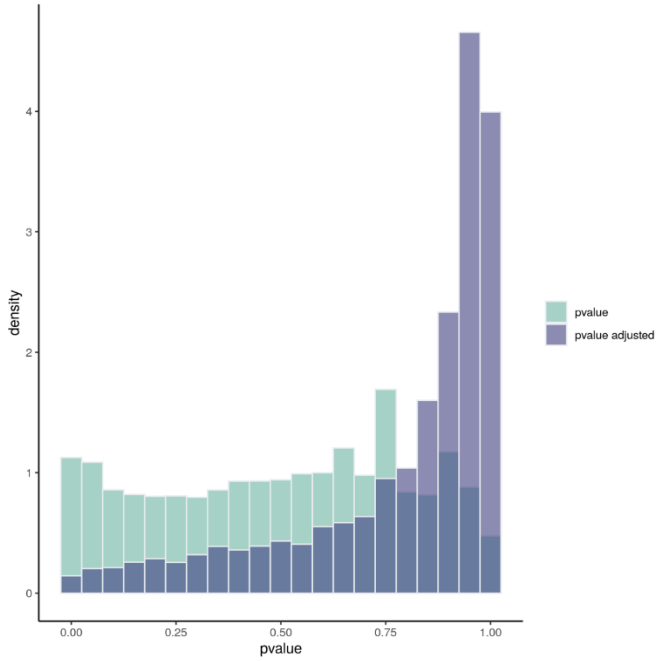
mutation du gène SF3B1. Nous observons donc que cette mutation du gène SF3B1 semble fortement impliqué dans les expressions différentielles des gènes chez les individus atteints du mélanome uvéal.

Figure 5 : MA-plot




Afin d'éviter de considérer des gènes comme différentiellement exprimés par hasard, nous appliquons la correction de Benjamini-Hochberg. Les évènements avec une p-valeur ajustée inférieure à 5% sont considérés comme significatifs. Figure 6 fournit les histogrammes des p-valeurs standards et ajustées. On remarque que l'histogramme pour les p-valeurs standards ressemble à une densité de loi uniforme. La correction permet donc d'éviter la découverte de faux positifs. Au total, nous retrouvons 240 gènes significativement différentiellement exprimés entre les patients ayant une mutation du gène SF3B1 et les autres.

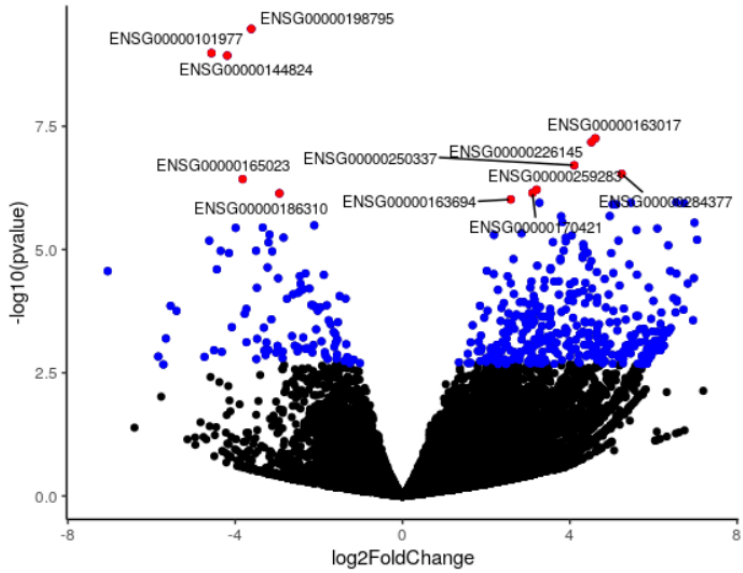
Figure 6 : histogramme des p-valeurs




La métrique couramment utilisée pour apprécier l'expression différentielle entre les gènes est le Log2FoldChange (LFC). Le LFC traduit le ratio en log2 entre 2 moyenne de reads. Ainsi, une valeur de 0 signifie qu'il n'y a pas de différence d'expression génique, une valeur de 1 signifie une expression deux fois plus importante d'un gène

sous une condition donnée et une valeur de -1 signifie une expression deux fois moins importante d'un gène sous une condition donnée. Figure 7 fournit le volcano plot, figure permettant de mettre en évidence la relation entre les p-valeurs et la différence d'expression génique relative entre deux conditions données. Toutes les valeurs en bleu correspondent aux gènes associés à une p-valeur ajustée inférieure à 1%. Toutes les valeurs en rouge correspondent aux gènes ayant une p-valeur standard inférieure à 0.0001%, soit une p-valeur ajustée inférieure à presque 0.01%. Parmi les gènes les plus différentiellement exprimés, nous observons le gène p53 et le gène RNA binding motif protein 47 (RBM47). Dans la littérature, Radine et al. (2020) décrivent que la sur-expression du gène RMB47 conduit à l'amélioration de l'expression de p53. Ces 2 protéines sont impliquées dans certains cas dans la cancérogénèse des tissus épithéliales.

Figure 6 : Volcano plot




Jusqu'à présent, nous avons étudié l'expression différentielle des gènes. Toutefois, nous pouvons également nous intéresser aux épissages alternatifs. Furney et al. (2013) ont montré des différences d'épissage alternatif pour les gènes CRNDE, ABCC5 et UQCC lorsque les patients atteints de mélanome uvéal ont une mutation du gène SF3B1. Nous retrouvons des résultats très similaires en utilisant l'outil IGV. Figure 7 fournit l'expression de chaque paire de bases sur le gène ABCC5 et met en évidence une rétention de l'intron 5 pour ce gène. On observe une couverture des reads pour la région intronique très élevée pour les individus n'ayant pas de mutation du gène SF3B1. En revanche, ceux qui portent l'allèle SF3B1 muté ont un épissage de l'intron 5 car la présence de reads dans la région intronique est quasi inexistante. De même, la jonction d'épissage en rouge et l'épaisseur de l'arc montrent que cet évènement d'épissage est majeur. Les arcs en rouge montrent la jonction d'épissage pour le brin + et les arcs en bleu montrent la jonction d'épissage pour le brin –. L'épaisseur de l'arc est proportionnelle à la profondeur de la couverture des reads.

Furney et al. (2013) décrivent, pour les échantillons présentant l'allèle SF3B1 muté, une couverture très faible des reads pour l'exon 8 et l'exon terminal et une expression relativement plus importante dans la condition non mutée pour l'exon 8 et l'exon terminal concernant le gène UQCC. Nous retrouvons bien cela avec les fichiers BAM que nous avons généré. Figure 8 fournit l'expression de chaque paire de bases sur le gène UQCC. Les rectangles en rouge montrent une expression relativement plus importante pour la condition non mutée et une faible expression de ses 2 exons pour la condition non mutée.

Le dernier gène étudié par Furney et al. (2013) dans le cadre de l'épissage alternatif concerne le gène non-codant CRNDE. Figure 9 fournit l'expression des introns et exons pour le gène CRNDE. Furney et al. (2013) décrivent un site accepteur alternatif pour l'exon 4 dans la condition mutée, ce qui revient à prendre uniquement une partie de l'exon 4 dans la condition SF3B1 mutée pour le gène CRNDE. Nous observons bien un phénomène d'épissage alternatif pour les échantillons avec l'allèle SF3B1 muté. Nous remarquons que les individus ayant une mutation

SF3B1 exprime une partie de l'exon 4 (rectangle rouge) car il y a présence d'un site d'épissage alternatif qui produit un isoforme du transcrit original CRNDE (avec l'exon 4 complet). Les échantillons qui portent le gène SF3B1 non muté exprime la totalité de l'exon 4 tel que décrit par Furney et al. (2013).

Figure 7 : Rétention d'intron 5 pour le gène ABCC5

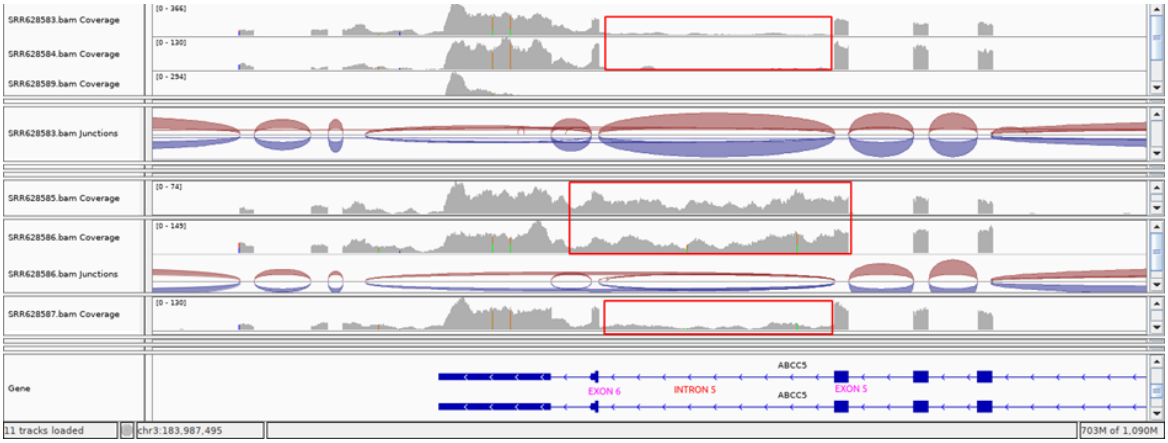



Figure 8 : Exon terminal alternatif pour le gène UQCC

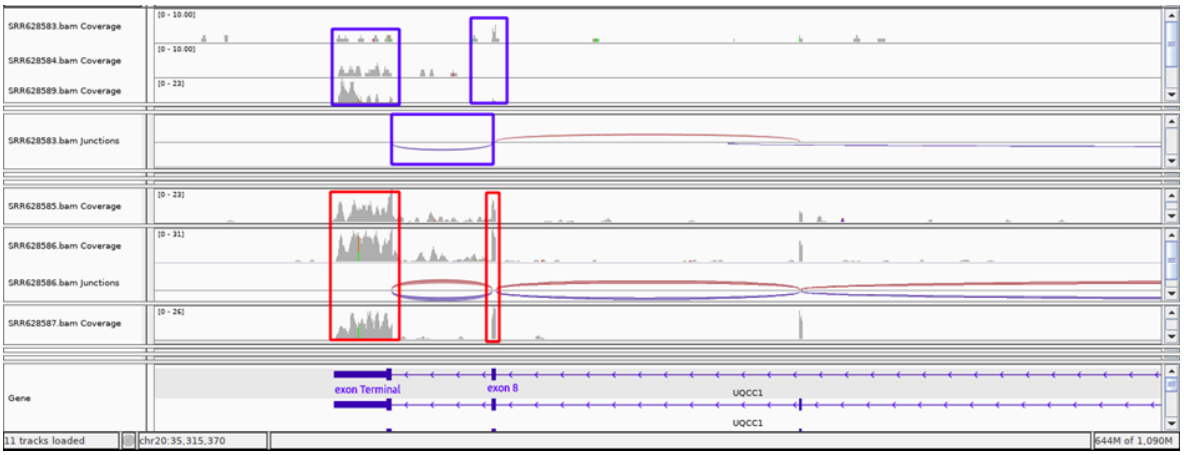
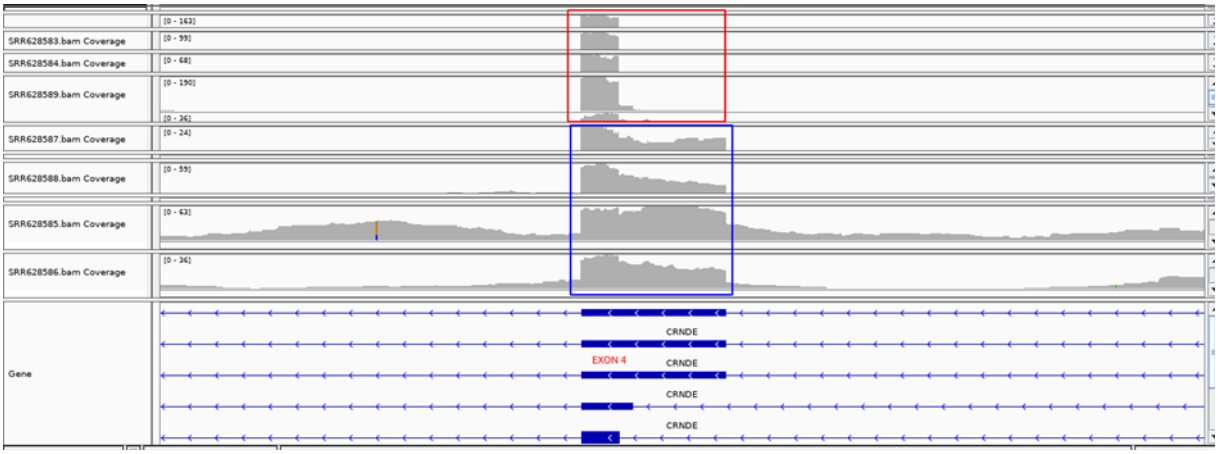



Figure 9 : Epissage alternatif de l'exon 4 pour le gène CRNDE



# Méthodes

## Importation et nettoyage des données RNA-seq

L'ARN a été récolté via le séquençage Illumina HiSeq 2000 à partir d'un échantillon de huit individus ayant un mélanome uvéal et a fourni des reads paired-end de 100 paires de base. Les données RNA-seq sont issues de l'étude de Habour et al. (2009), qui sont téléchargeables depuis le site hébergé par NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra?term=SRA062359). Plus d'informations sont accessibles directement depuis le SRA Run Selector du NCBI[2] avec le numéro de projet PRJNA182344. La qualité des reads a été étudiée via le logiciel fastq. Parmi les huit individus étudiés, quatre ont subi une mutation du gène SF3B1 dont trois sont une mutation R625C et une est une mutation R625H. Les fichiers FASTQ issus de ces reads ont été alignés contre le génome humain de référence GRCh38[3]. Pour réaliser l'alignement des reads, l'indexation et le mapping du génome ont été réalisés avec STAR[4] issu du conteneur evolbioinfo/star. Une fois ces étapes terminées, la matrice de comptage a été réalisée via featureCounts issu du conteneur evolbioinfo/subread en utilisant les fichiers BAM obtenus lors du mapping.

## Analyse différentielle des expressions géniques

Afin de mettre en évidence l'implication d'une mutation du gène SF3B1 dans le cancer du mélanome uvéal, l'analyse statistique différentielle d'expression génique s'est appuyée sur la librairie DESeq2 gérée par Bioconductor pour le logiciel R. Bien que la matrice de comptage ait été calculée lors de l'étape d'importation des données RNA-seq, les tests de qualité produits jusqu'ici ne concernaient que les reads eux-mêmes. D'autres biais persistent malgré tout, et les comptages des gènes ont été recalculés afin d'éviter des problèmes de taille de librairie et en prenant en compte la dispersion des comptages pour chaque gène. Les p-valeurs ajustée suivent la correction de Benjamini-Hochberg et permettent d'éviter la découverte de faux positifs. Les évènements avec une p-valeur ajustée inférieure à 5% sont considérés comme significatifs, ce qui correspond à 240 gènes significativement différentiellement exprimés entre les individus ayant une mutation du gène SF3B1 et les autres.

## Analyse d'épissages alternatifs

Afin d'analyser l'implication d'une mutation du gène SF3B1 dans les épissages alternatifs, l'outil Integrative Genomics Viewer (IGV) a été utilisé sur les fichiers BAM produits par le script nextflow et les fichiers BAI produits avec samtools index. Cet outil permet de visualiser le nombre de fois que chaque paire de base est exprimé avec les reads d'un échantillon. Il diffère de l'analyse différentielle des expressions géniques puisqu'il permet de s'intéresser précisément à l'expression partielle ou complète des exons.

---

[2] Le SRA Run Selector est accessible à l'adresse suivante : https://www-ncbi-nlm-nih-gov.ezproxy.u-pec.fr/Traces/study/

[3] Le fichier du génome humain est téléchargeable à l'adresse suivante : ftp://ftp.ensembl.org/pub/release-101/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz

[4] Le fichier d'annotation utilisé pour l'indexation est téléchargeable à l'adresse suivante : ftp://ftp.ensembl.org/pub/release-101/gtf/homo_sapiens/Homo_sapiens.GRCh38.101.chr.gtf.gz

# Conclusion

A l'aide des outils Nextflow et Docker, nous avons pu créer un projet reproductible qui permet l'analyse des données RNA-seq. Ce workflow est aussi portatif : il n'est pas centré sur les données de Harbour et al. (2013) mais peut aussi être utilisé pour d'autres données RNA-seq en ajustant quelques paramètres. La pipeline construite parallélise les étapes d'analyse pour optimiser le temps d'exécution. La pipeline est adaptable et peut être lancée sur des plateformes différentes (local à condition d'avoir un ordinateur ayant 32 Go RAM minimum, cloud ou cluster de calcul).

Concernant les résultats obtenus à la fin de cette analyse, nous retrouvons quelques gènes différentiellement exprimés (240 sur 30988). Nos résultats sont différents des 2 articles sur lesquels nous nous somme appuyés. Plusieurs raisons peuvent expliquer ces différences. L'utilisation des outils n'est pas la même pour chaque étape de l'analyse RNA-seq. Également, les auteurs des 2 articles utilisent des outils différents, justifiant ainsi la diversité dans les résultats. Cependant pour l'analyse des phénomènes d'épissage avec IGV, nous retrouvons bien la rétention d'intron pour le gène ABCC5 , l'épissage alternatif de l'exon 4 pour le gène CRNDE et l'exon terminal alternatif pour le gène UQCC.

En raison de quelques contraintes de temps, nous n'avons pas pu explorer l'analyse des sites donneurs et accepteurs d'épissage avec un logiciel d'alignement. Il serait intéressant d'effectuer cette analyse avec les outils utilisés lors de ce projet pour voir l'inclusion différentielle des exons et pour comparer nos résultats et ainsi voir les problématiques de reproductibilité qui peuvent être soulevées pour ce genre de workflow.

# Références

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, 1-1.

Furney, S. J., Pedersen, M., Gentien, D., Dumont, A. G., Rapinat, A., Desjardins, L., ... & Marais, R. (2013). SF3B1 mutations are associated with alternative splicing in uveal melanoma. *Cancer discovery*, *3*(10), 1122-1129.

Harbour, J. W., Roberson, E. D., Anbunathan, H., Onken, M. D., Worley, L. A., & Bowcock, A. M. (2013). Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma. *Nature genetics*, *45*(2), 133-135.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 1-21.

Radine, C., Peters, D., Reese, A., Neuwahl, J., Budach, W., Jänicke, R. U., & Sohn, D. (2020). The RNA-binding protein RBM47 is a novel regulator of cell fate decisions by transcriptionally controlling the p53-p21-axis. *Cell Death & Differentiation*, *27*(4), 1274-1285.