Projet Hackathon

Antoine DEGENNE, Louise DRY, Etienne PEREZ, Hélène PHILIPPE

Introduction

Pour reprendre Claude Bernard dans son *Introduction à l'étude de la médecine expérimentale* [1], la vérité en science se construit autour d'observations faites dans un cadre expérimental précis permettant de mettre à l'épreuve une théorie étudiée. Pour garantir la fiabilité d'un résultat, il faut être en capacité de reproduire l'expérience qui a mené à ce résultat le plus justement possible. Le problème de la reproductibilité s'applique dans toutes les sciences expérimentales, d'autant plus avec les outils que nous avons aujourd'hui à disposition.

La programmation informatique au service de la biologie permet d'automatiser des processus fastidieux, comme la lecture et l'analyse de données en grande quantité. Les données génétiques font partie de cette catégorie. L'analyse de ces données permet de mettre en évidence de potentiels marqueurs biologiques dans certaines conditions et maladies, dans notre cas pour le cancer.

L'idée est de pouvoir reproduire, à l'aide de données en libre service, des expériences et des programmes qui ont amené des chercheurs à tirer des conclusions, afin de voir si l'on peut arriver aux mêmes résultats et les interpréter de la même façon. Dans notre cas, nous disposions de deux articles : *SF3B1 mutations are associated with alternative splicing in uveal melanoma, Furney et al., 2013* [2] et *Recurrent mutations at codon 625 of the splicing factor SF3B1in uveal melanoma, Harbour et al., 2013* [3] ayant travaillé avec les mêmes données, la même technique, et n'étant pas parvenus aux mêmes conclusions.

Présentation des travaux reproduits

Ces deux articles se concentrent sur le cas du cancer uvéal, le cancer de l'œil le plus fréquent et causant souvent des métastases fatales.

1. Résumé de l'article de *Harbour et al.*

L'étude réalisée par *Harbour et al.* [2] présente deux étapes. Dans un premier temps, une recherche de mutations dans l'exome de tumeurs a été effectuée. Une mutation sur BAP1 (BRCA1 associated protein 1) ayant déjà été observée lors d'une étude réalisée au préalable [4], l'étude ici présente s'intéresse aux mutations autres que BAP1. Par séquençage et analyse d'échantillons de mélanomes uvéaux, une mutation sur la sous-unité 1 du facteur d'épissage 3B (SF3B1) a été identifiée dans 19% des tumeurs. Cette mutation semble être associée à une progression plus lente de la métastase principalement chez les patients jeunes ou ayant peu de cellules d'épithélium indifférenciées.

La deuxième partie de l'étude consiste donc à la recherche des effets de la mutation SF3B1 sur l'expression génétiques. Pour cela, la recherche d'une expression différentielle de l'ARN a été faite sur six mutants SF3B1 et six types sauvages avec la puce à ADN Illumina BeadArray Platform. Cette recherche n'a permis d'identifier uniquement les

gènes qui présentent de réelles différences d'expression, aucun d'eux ne donnant d'information sur les effets de la mutation SF3B1.

Ainsi, dans un second temps, une recherche d'épissage différentiel a été faite sur 3 mutants SF3B1 et 5 types sauvages à l'aide d'une analyse d'RNA-Seq. Cette analyse mène à la conclusion qu'il n'y a aucune réelle différence dans l'expression des gènes entre les types mutés et les types sauvages.

2. Résumé de l'article de *Furney et al.*

La méthodologie de l'étude de *Furney et al.* [3] se rapproche de celle mise en place par *Harbour et al.* [2] (recherche de mutations puis recherche des effets de la mutation SF3B1).

Tout d'abord, un séquençage du génome entier de 12 mélanomes uvéaux primaires a été réalisé. Par comparaison à l'ADN normal des patients, les mutations BAP1 et SF3B1 ont de nouveau été détectées (dans respectivement 7 et 3 des 12 tumeurs).

Ensuite, pour comprendre les effets de la mutation SF3B1, trois mutants SF3B1 et 3 types sauvages ont été hybridés avec la puce à ADN HTA2. Par ce procédé, 325 gènes ont été identifiés comme différentiellement exprimés (46 régulés à la hausse et 279 à la baisse par rapport au type sauvage de SF3B1). Au niveau de l'épissage, 130 gènes contenant au moins un exon exprimé différentiellement ont été détectés. Ensuite, les épissages alternatifs potentiels ont été identifiés par RNA-seq.

Furney et al. ont aussi comparé leurs données à celles de l'étude de *Harbour et al.* [2] de manière à confirmer leur identification des épissages alternatifs.

Cette étude arrive donc à la conclusion qu'il existe bien des différences d'expression et d'épissage entre les types mutés et les types sauvages.

On a ainsi deux articles qui, partant d'un même jeu de données, arrivent à des conclusions divergentes. L'objectif de notre étude va donc être de réaliser une RNA-seq sur les données de l'article *Harbour et al.* et de comparer nos résultats aux conclusions de chacun des articles.

Matériel & Méthodes

1. Accès aux données

Les données RNA-seq ont été téléchargées sur le site de la NCBI en utilisant le lien fourni dans l'article *Furney et al* [3].

Les données de génome humain et d'annotations ont été téléchargées à partir du site emsebl.org (génome GRch38).

2. Construction d'un pipeline d'analyse de données

Nous avons utilisé le gestionnaire de workflow Snakemake pour développer un pipeline d'analyse de données. Snakemake propose une architecture en rules (étapes du processus d'analyse) qui prennent un fichier en entrée et retournent un autre fichier en sortie.

Pour exécuter chaque rule, Snakemake récupère une image Docker qui correspond à un environnement spécifique où les logiciels adéquats avec leurs packages associés sont installés et utilisables. Les codes des Dockerfiles ayant permis de générer les images Docker utilisées sont fournies avec ce rapport.

3. Outils d'analyse

Nous présentons ici la liste des outils utilisés dans le pipeline d'analyse.

3.1. Fastq-dump/Fasterq-dump/FastQC

Fastq-dump/Fasterq-dump [5] sont deux wrappers inclus dans le SRA-toolkit qui ont pour fonction de télécharger des fichiers disponibles sur le NCBI via leur numéro d'accession. En particulier nous nous en sommes servi pour télécharger les fichiers SRA au format fastq des données génomiques mises à disposition et utilisées dans les articles que nous avons vu précédemment.
L'image Docker utilisée pour accéder à ces outils est la suivante : https://hub.docker.com/r/pegi3s/sratoolkit/. Cette image donne accès à la version 2.9.6 de SRA Toolkit.
Les commande permettant de télécharger les fichiers au format SRA puis de les transformer au format FASTQ sont les suivantes :

```
prefetch -v {wildcards.list_sra} > sra_files/{wildcards.list_sra}.sra
fastq-dump -v --split-files {wildcards.list_sra} --outdir fastq_files/
```

L'option "*--split-files"* permet de générer deux fichiers FASTQ correspondant aux lectures d'ARN dans les 2 sens (3' → 5' et 5' → 3'). Avoir ces deux fichiers permet d'améliorer la précision du mapping que l'on réalise plus tard dans le pipeline.

Nous avons également utilisé l'outil FastQC (version 0.11.9) sur les fichiers FASTQ afin de contrôler la qualité des données obtenues à l'étape précédente. L'image Docker utilisée est l'image *biocontainers/fastqc/v0.11.9_cv8*.

3.2. STAR

STAR est un outil permettant d'aligner les séquences d'ARN [6]. Afin d'aligner les séquences, nous avons besoin de créer un index à partir du génome humain. Cet index a donc été créé au préalable en téléchargeant le génome humain puis une commande STAR. Une autre commande sert ensuite à aligner les séquences et sera appliquée à chacun de nos échantillons.

STAR présente les options générales suivantes :

- l'option -runMode permet d'indiquer la tâche que l'on veut que STAR réalise (création de l'index ou alignement des reads)
- l'option *-runThreandN* permet d'indiquer le nombre de "fils d'exécution" du processeurs alloués à cette tâche

Pour la tâche de création de l'index on a les options :

- l'option -genomeDir indique le chemin vers le répertoire où les fichiers d'index seront créés
- l'option -genomeFastaFiles indique le chemin vers les fichiers Fasta du génome humain

Pour la tâche de création de l'index on a les options :

- l'option -readFilesIn indique les fichiers d'entrée pairés qui vont être alignés
- l'option -genomeDir indique le chemin vers le fichier d'index préalablement créé
- l'option -outSAMtype BAM SortedByCoordinate permet d'obtenir en sortie des fichiers BAM triés par leurs coordonnées

Ainsi, la commande de création d'index est la suivante :

STAR --runMode genomeGenerate --genomeDir {emplacement du dossier d'index} --genomeFastaFiles {emplacement du fichier de génome} --runThreadN {threads}

Et la commande d'alignement des séquences :

STAR --runMode alignReads --genomeDir {emplacement du dossier d'index} --outSAMtype BAM SortedByCoordinate --readFilesIn {.fastq} --runThreadN {threads}

3.3. FeatureCounts

FeatureCounts fait partie du package Subread [7]. C'est un outil qui permet de compter les reads des séquences d'ARN précédemment alignées.
L'intérêt ici est d'observer si certains reads reviennent particulièrement dans certains gènes chez les wildtypes, les mutants ou bien les deux.
La commande est la suivante :

featureCounts -T {threads} -t gene -g gene_id -p -s 0 -a {input.gtf} -o {output.counts} {input.bam}

L'option "*-T*" permet d'indiquer le nombre de cœurs utilisés pour cette tâche. Dans notre cas nous en avions 16.
L'option "*-t*" précise le type d'élément qui a été utilisé pour faire le fichier d'annotation gtf, "*-g*" spécifie le type d'attribut du fichier gtf, "*-p"* indique que nous avons des séquences pairées (une séquence lue dans chacun des 2 sens), "*-s 0*" permet de préciser que le comptage est effectué sur les branches d'ARN sans porter d'attention à l'orientation du brin, "*-a*" permet d'indiquer le fichier d'annotation à utiliser, et "*-o*" permet de préciser où vont les fichiers en sortie.

3.4. DESeq2

DESeq2 est un package disponible sur R qui permet de faire de l'analyse différentielle [8]. C'est un outil qui s'utilise en 2 étapes.

Étape 1: Création d'un objet via DESeqDataSet.
L'objet créé contient le fichier avec les comptes de reads (issu de FeatureCounts),
un fichier de métadonnées qui contient les noms des samples et leur sens de lecture (1 ou 2) et s'ils correspondent à un contrôle ou un muté. Cette dernière information est cruciale car l'analyse consiste à comparer l'expression pour le groupe contrôle à l'expression pour le groupe muté.

Étape 2: Application de la fonction DESeq() sur l'objet créé au préalable.
La fonction DESeq() donne accès à de nombreuses informations sur la dispersion des gènes, leur taille et les résultats de tests statistiques. [9]
Deux indicateurs calculés sont particulièrement intéressants :

- Le logarithme de base 2 du rapport d'expression (log2FoldChange): il est positif si le gène est surexprimé chez le groupe mutant (comparé au sauvage), et négatif si le gène est sous-exprimé chez le groupe mutant. On considère cette différence comme significative si la valeur absolue du log2FoldChange est supérieure à 1.5.
- La p-value du test de Wald.

L'interprétation de ces deux indicateurs permet de conclure pour chaque gène s'il existe une expression différentielle entre le groupe sauvage et mutant.

Figure 1 : Schéma du pipeline

Résultats

1. Analyse différentielle sur tous les gènes

Dans un premier temps, nous avons exécuté la fonction DESeq() sur tous les gènes associés aux fichiers SRA à notre disposition. Cela nous a permis de récupérer les comptages de read normalisés, ainsi que l'expression différentielle de ces gènes.

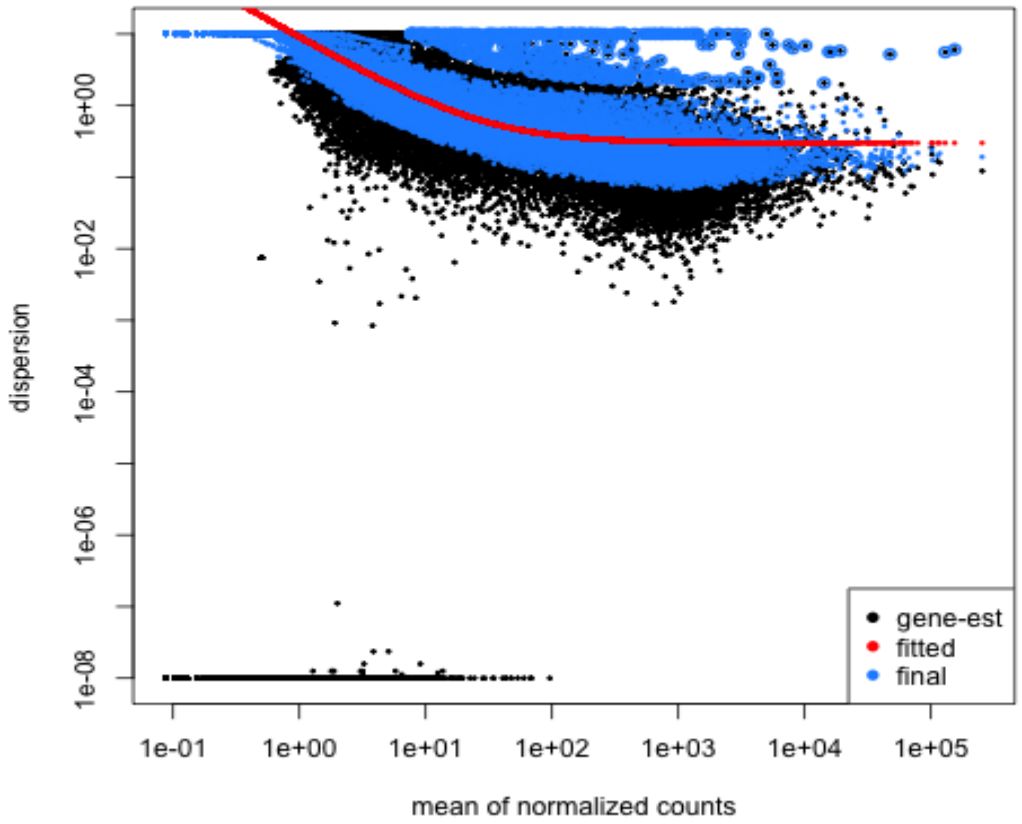

***Figure 2** : Dispersion des gènes. Un point noir = dispersion pour un gène. Plus un gène a un faible nombre de read, plus sa dispersion est élevée.*

On peut remarquer sur la *Figure 2* que de nombreux gènes ont des dispersions très faibles, donc des nombres de reads élevés. Cependant les gènes à forte dispersion ne sont pas très bien fittés, ce qui prédit une analyse relativement peu fiable sur l'ensemble des gènes.

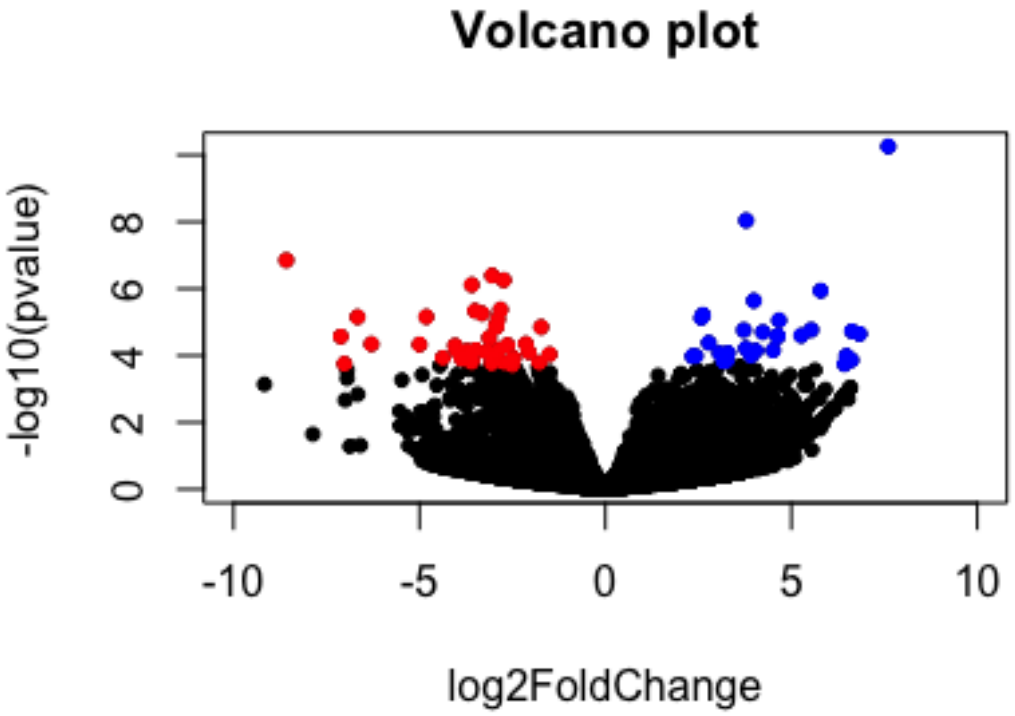

Figure 3 *: Volcano plot de l'ensemble des gènes étudiés.*

- *Gènes dont le L2FC est inférieur à -1.5 et la p-value ajustée < 0.05*
- *Gènes dont le L2FC est supérieur à 1.5 et la p-value ajustée < 0.05*

Le volcano plot permet d'observer quels gènes sont surexprimés ou sous-exprimés dans certains échantillons. Les points bleus et rouges correspondent aux gènes significativement différentiellement exprimés selon les deux critères (p-value et log2FoldChange).
Ici on observe que la majorité de nos gènes ne sont pas significatifs.

On peut également observer l'Analyse en Composantes Principales (*Figure 3*) pour voir si certains groupes se forment en fonction de si les individus ont des mutations de SF3B1 ou non.

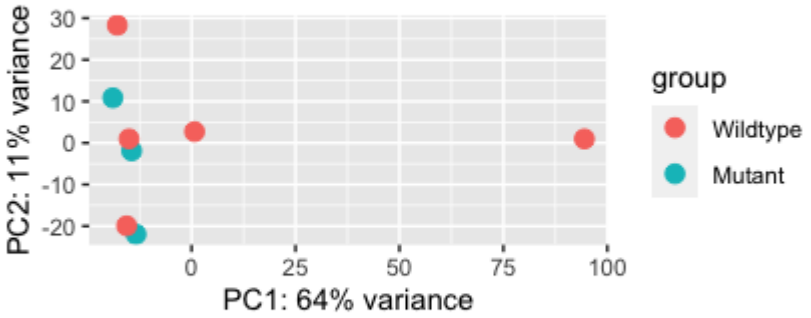

Figure 4 *: ACP sur l'ensemble des gènes.*

On observe que 64% de la variance est expliquée par le premier axe, cependant un seul groupe est isolé du reste. Il n'y a pas de groupe spécifique qui se détache selon les conditions (Wildtype et Mutant).

La liste des 10 gènes les plus différenciés est trouvable dans le *Tableau 2.* Nous avons trouvé au total 69 gènes différenciés (L2FC < -1.5 ou L2FC > 1.5 et pvalue ajustée < 0.05).

Gène	abs(L2FC)	Sens de différenciation	pvalue ajustée	Nature de la protéine produite
HLA-DQA2	8.581	down	8.427e-4	Antigénique
HLA-DQB2	7.115	down	1.67e-2	Antigénique
TPBGL	6.664	down	8.877e-3	"Trophoblast Glycoprotein Like", signalisation
LMO7DN	4.999	down	2.462e-2	Gène ARN
PDGFRA	4.818	down	8.877e-3	Code un récepteur pour facteur de croissance
KCNJ2	4.373	down	3.843e-2	Protéine membranaire
MCF2	4.0412	down	2.533e-2	Protéine oncogénique (favorise l'apparition de cancers)
PTCHD4	4.004	down	3.033e-2	Protéine membranaire
C1R	3.867	down	3.930e-2	Peptidase
GRIK3	3.764	down	3.969e-2	Récepteur glutamate

Tableau 1*: Liste des 10 gènes dont l'expression est la plus différenciée selon notre analyse, classés par valeur de abs(L2FC).*

Nous pouvons remarquer que tous ces gènes sont en sous-expression (sens de différenciation = down). Le premier gène qui est sur-exprimé chez les mutants et le 41e par ordre dans le tableau avec une valeur absolue de L2FC égale à 2.34 et une pvalue ajustée de 1.053e-4. Au total, il y a 40 gènes en sous-expression et 29 en sur-expression selon nos analyses, en prenant une limite de L2FC à -1.5 (respectivement 1.5).

Cependant, les articles étudiés se sont concentrés sur une liste de gènes spécifiques (*Tableau 2*), que l'on a pu récupérer grâce à leur code (ENSXXX) sur la page officielle du NCBI. Cela nous a permis de regarder les résultats de ces gènes particuliers plus en détail.

2. Analyse différentielle des gènes mentionnés par l'article

Code	Gène correspondant	baseMean	L2FC	pvalue	padj
ENSG00000115524	SF3B1	6244.429	0.00364	0.992	0.999
ENSG00000101019	UQCC1	1420.551	0.925	0.003	0.167
ENSG00000088256	GNA11	1354.525	-0.143	0.727	0.983
ENSG00000148848	ADAM12	1137.352	0.0226	0.967	0.998
ENSG00000245694	CRNDE	660.164	0.140	0.807	0.990
ENSG00000114770	ABCC5	5898.646	0.612	0.217	0.855
ENSG00000156052	GNAQ	1958.503	0.098	0.800	0.990
ENSG00000163930	BAP1	2649.031	-0.101	0.797	0.990
ENSG00000131503	ANKHD1	0	NA	NA	NA

***Tableau 2** : Gènes d'intérêts, leur code associé et les résultats exploités de l'analyse différentielle*

D'après les valeurs de pvalue, seul le gène UQCC1 (ubiquinol-cytochrome c reductase complex assembly factor 1) est exprimé différentiellement de façon significative. Ce gène encode une protéine transmembranaire. L'épissage alternatif de ce gène produit de très nombreux variants. On remarque également que le gène ANKHD1 n'a pas eu suffisamment de reads pour passer l'analyse différentielle.

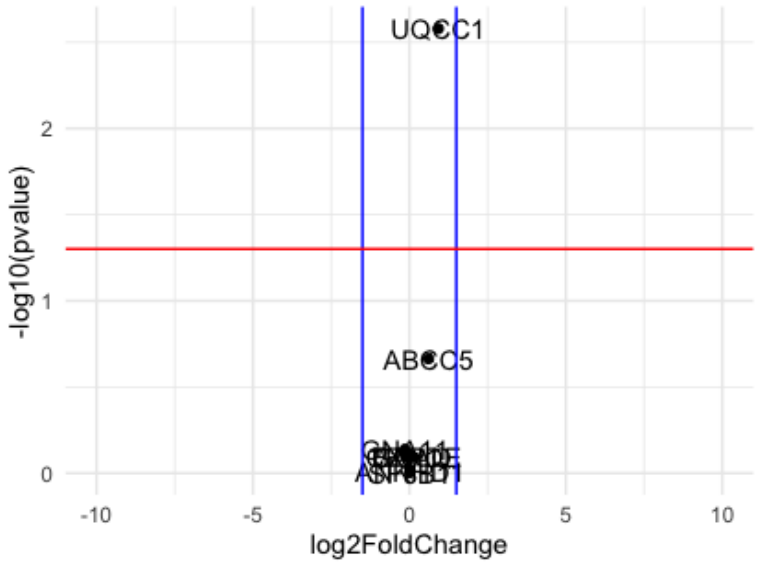

***Figure 5** : Volcano plot des gènes mentionnés dans les articles.*

Les gènes sous la ligne rouge ont une pvalue ajustée < 0.05.
Les gènes à l'intérieur des lignes bleues sont des gènes dont le L2FC est compris entre -1.5 et 1.5.

On peut observer que le seul gène estimé significatif n'est pas exprimé différentiellement. Globalement, il n'y a pas de résultat particulier à extraire de l'analyse différentielle des gènes observés dans les articles.

Conclusion

L'article *Harbour et al.* a mis en évidence qu'il n'y avait ni d'expression différentielle des gènes intervenant dans le mélanome uvéal, ni d'épissage différentiel de ces gènes chez les individus mutants. L'article *Furney et al.* a mis en évidence l'existence d'une expression différentielle de 325 gènes, et d'un épissage différentiel de 130 gènes. Dans notre analyse, nous avons trouvé une expression différentielle de 69 gènes.

Notre analyse a montré que la reproductibilité des articles *Furney et al.* et *Harbour et al.* n'est pas optimale dans la mesure où nous avons reproduit une analyse RNA-seq sur les mêmes données et nous n'avons pas obtenu les mêmes résultats. Parmi les divergences possibles de nos démarches, les paramètres utilisés dans l'alignement des séquences sur le génome humain peuvent expliquer la différence de résultats obtenus.

Bibliographie

[1] Introduction à l'étude de la médecine expérimentale, Claude Bernard, 1865

[2] J William Harbour et al. "Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma". In: Nature genetics 45.2 (2013), pp. 133–135.

[3] Simon J Furney et al. "SF3B1 mutations are associated with alternative splicing in uveal melanoma". In: Cancer discovery 3.10 (2013), pp. 1122–1129.

[4] J William Harbour et al. "Frequent mutation of BAP1 in metastasizing uveal melanomas". In: Science 330.6009 (2010), pp. 1410–1413.

[5] Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010

[6] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013

[7] Yang Liao, Gordon K. Smyth, Wei Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, *Bioinformatics*, Volume 30, Issue 7, 1 April 2014, Pages 923–930

[8] Anders, S., Huber, W. Differential expression analysis for sequence count data. *Genome Biol* 11, R106 (2010). https://doi.org/10.1186/gb-2010-11-10-r106

[9] McDermaid, A., Monier, B., Interpretation of differential gene expression results of RNA-seq data : review and integration, Briefings in Bioinformatics, 20(6), 2019, 2044–2054, doi: 10.1093/bib/bby067