Mise en place d'un pipeline d'annotation des introns dans les génomes viraux à partir de données RNAseq : Application au virus de la grippe.

- Cahier des charges -

Equipe :
Barbier Jérémy
Ganovsky Jérémy
Koenig Natacha
Siekaniec Grégoire

Encadrants :
Benoit-Pilven Clara
Naffakh Nadia
Navratil Vincent
Lacroix Vincent

Master 2 Bioinformatique - 20 octobre 2017

Table des matières

1 Présentation du projet

1.1 Contexte scientifique

Les virus influenza A de la famille des Orthomyxoviridae est un pathogène majeur causant la grippe chez l'Homme, les mammifères et les oiseaux. Chaque année, 15–20% de la population mondiale est infectée [5], environ 25 000 décès sont dûs à une infection par un virus influenza aux Etats Unis [9] et entre 250 000 et 500 000 décès au niveau mondial. Il s'agit donc d'un réel problème de santé publique, et comprendre le fonctionnement intra-cellulaire du virus pourrait amener à proposer des traitements plus efficaces, voire même de meilleurs vaccins.

Le génome du virus comprend 8 segments simple-brin (ou monocaténaire) d'ARN de polarité négative représentant environ 14000 bp, codant pour 8 protéines majeures, et quelques protéines auxiliaires. Il possède une ARN polymérase hétérotrimérique, constituée de 3 sous unités (PB1, PB2 et PA). Cette ARN polymérase transcrit l'ARN antisens viral en ARN sens (ARNm), permettant la production, par la cellule hôte, de protéines virales. Le virus utilise également la machinerie cellulaire de l'hôte pour effectuer la maturation de ses ARN messagers (ARNm), notamment l'épissage de certains des ARNm qu'il produit. Deux protéines essentielles à la multiplication des cellules virales sont produites de cette manière : la protéine NS2/NEP (facteur d'exportation nucléaire) et la protéine M2 (canal ionique). Elles correspondent aux épissages respectifs des ARNm de NS1 et M1 (figure 1).

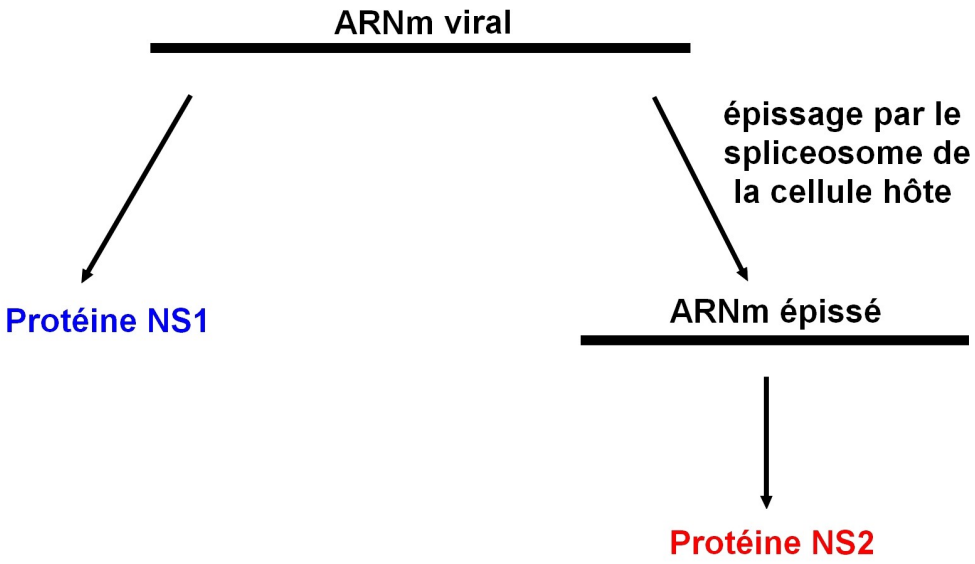

FIGURE 1 – **Lien entre NS1 et NS2.**

Deux composants du spliceosome (facteurs d'épissage) humains ont été identifiés comme nécessaires à la multiplication du virus : RED et SMU1 [1]. En effet, l'épissage des ARNm codant pour NS1 est diminué dans les cellules appauvries en RED ou

SMU1, ce qui conduit à une réduction de la production d'ARNm épissé codant pour NS2. Cela est dû au fait que pour effectuer l'épissage de ses ARNm, l'ARN polymérase virale recrute un complexe formé de RED et SMU1. Une cellule appauvrie en l'un des deux facteurs (ou les deux) effectue donc moins d'épissage, et la production d'ARNm de type NS2 (épissé) est donc logiquement réduite. L'apport protéique de NS2, qui se traduit par le ratio NS2/NS1 est donc réduit également. Ceci est intéressant dans le sens où la multiplication du virus serait très dépendante de l'expression de NS2 (cette protéine étant nécessaire à l'exportation des ribonucléoprotéines virales du noyau vers le cytoplasme de la cellule infectée). Ces résultats établissent donc les facteurs RED et SMU1 comme des régulateurs clés de l'expression des gènes du virus de la grippe.

Dans l'ensemble, les résultats dévoilent un nouveau mécanisme de subversion virale de la machinerie d'épissage cellulaire, en établissant que les facteurs d'épissage humain RED et SMU1 agissent conjointement en tant que régulateurs clés de l'expression des gènes du virus de la grippe. De plus, les données indiquent un rôle central de l'ARN polymérase virale en couplant la transcription et l'épissage alternatif des ARNm viraux.

Dans ce contexte, bloquer l'épissage du virus empêcherait sa multiplication et sa prolifération. Cela pourrait permettre de concevoir de nouvelles approches antivirales contre la grippe. Dans cette optique, nous chercherons à mieux comprendre les mécanismes d'épissage en en identifiant les sites. Cela permettra de mettre évidence de nouveaux introns.

1.2 Analyse préliminaire

Dans le cadre d'un projet collaboratif entre l'Institut Pasteur (Nadia Naffakh), le PRABI (Vincent Navratil) et le LBBE (Vincent Lacroix), le séquençage du transcriptome de lignées cellulaires humaines A549 a été fait dans les conditions suivantes : infection par un virus influenza A particulier : A/WSN/33 (H1N1) et/ou inactivation du facteur d'épissage RED/SMU1 par un siRNA. De plus, afin de vérifier que les siRNA (small interfering RNA ou petits ARN interférents) non spécifiques ne biaisent pas l'analyse des données, celle-ci a été aussi réalisée avec des siRNA contrôle n'ayant en théorie pas d'effet (tableau 1). Et en complément, un contrôle négatif a aussi été testé sans siRNA pour contrôler le séquençage de l'ARN.

	Non infecté par le virus	**Infecté par le virus**
siRNA RED-SMU1 (inactivation de RED-SMU1)	x 4	x 4
siRNA contrôle	x 4	x 4
contrôle négatif (sans siRNA)	x 4	x 4
TOTAL	24 jeux de données de séquençage	

TABLE 1 – **Plan expérimental des jeux de données de séquençage selon les conditions.**

Le séquençage des transcrits des lignées cellulaires humaines A549 infectées par le virus de la grippe (souche A/WSN/33) a été réalisé sur une plateforme Illumina HISeq 2500 en Paired-End et Multiplexé (2 échantillons par lane, 1 infecté et 1 non-infecté en proportions respectives de 70% et 30%). Suite à ce séquençage, un alignement avec le logiciel STAR [2] contre le génome humain et contre le génome viral a été réalisé afin de déterminer l'origine des lectures obtenues. Nos données contiennent celles provenant du virus. Une analyse différentielle de l'expression des gènes (DESeq2 [7]) et une analyse différentielle de l'épissage (DEXSeq [1]) ont été réalisées pour mettre en évidence les différences entre les 3 conditions.

Il est concrètement établi que RED/SMU1 favorise l'épissage de NS1/NS2 [3]. L'analyse préliminaire des données confirme ce résultat mais indique que d'autres isoformes de l'ARNm NS1 sont également présentes.

1.3 Objectifs

L'objectif du projet peut être décomposé en trois points. Le premier objectif est, en utilisant uniquement les données issues de la condition contrôle négatif – infecté, de faire un pipeline permettant l'annotation exhaustive des introns présents sur les ARNm viraux, en tirant parti de la très grande profondeur de séquençage des ARN issus de cellules infectées. Le but est de détecter des évènements d'épissage n'ayant pas été détecté auparavant, pour chaque intron détecté (selon un seuil à discuter en fonction du nombre ou du pourcentage de lectures chevauchantes).
Deuxièmement on reportera le nombre de lectures supportant l'excision de ces introns. Nous pourrons également calculer le taux de G-C, la longueur des introns ainsi que leurs coordonnées sur le segment. La force des sites d'épissage (MaxEntScan [4]) viendra compléter l'analyse.
Enfin ces étapes seront à nouveau réalisées pour les conditions : siRNA contrôle - infecté et siRNA RED-SMU1 - infecté. Une fois cela réalisé, il sera possible de comparer les résultats obtenus dans les différentes conditions afin de déterminer quels évènements d'épissage identifiés sont régulés par RED-SMU1.

1.4 Pour aller plus loin

- Le pipeline pourra être appliqué à d'autres jeux de données disponibles publiquement pour évaluer la prévalence des nouveaux introns découverts.
- La séquence de la protéine correspondant à chaque ARNm épissé pourra également être prédite.
- Les informations sur les domaines de la protéine pourront être étudiés via la base de données Uniprot [10] et/ou Interpro [6].
- Certains introns prédits pourront être validés expérimentalement par RT-qPCR.
- Créer une interface graphique facilitant l'utilisation du pipeline par des non-bioinformaticiens pourrait également être intéressant.

1.5 Description de l'existant

Sjcount [8] : sjcount est un outil du package IPSA (Integrative Pipeline for Splicing Analyses) permettant la quantification des jonctions introns/exons permettant de trouver les sites d'épissage en calculant des métriques d'épissage exon-centriques et intron-centriques. Cet outil pourra être utilisé pour déterminer et quantifier les sites d'épissages à partir des fichier .bam.

MaxEntScan [4] : MaxEntScan est un outil utilisant une méthode basée sur le principe de l'entropie maximale permettant de quantifier la force des sites contenant un signal permettant l'épissage d'ARN. Cet outil nous permettra de calculer la force des sites d'épissages lors de nos analyses.

2 Expression des besoins

2.1 Besoins fonctionnels

Le pipeline (figure 2) sera réalisé en bash (script sh) en utilisant différentes "briques" pouvant être à la fois des scripts "maison" en python et/ou des programmes déjà existants. Le pipeline prendra en entrée le fichier d'alignement des lectures sur les segments viraux au format .bam et la sortie sera la liste des introns trouvés ainsi que leurs positions sur le segment, le nombre de lectures supportant l'excision de l'intron, la longueur et le taux de G-C de l'intron et de l'ARN sans cet intron. Un autre fichier contiendra la force des sites d'épissages les plus importants calculée avec MaxEntScan. Pour réaliser l'analyse, le package Integrative Pipeline for Splicing Analyses (IPSA) et en particulier sjcount [8] seront testés. Si le test du package IPSA n'est pas concluant le fichier d'alignement au format .bam sera transformé au format .sam à l'aide de samtools et l'identification des introns ainsi que la comparaison entre les différentes conditions seront réalisées grâce à des scripts "maison" et/ou des programmes ou scripts existants. Une analyse différentielle des comptages pourra être réalisée mais pour cela les données devront être normalisées (possiblement par la profondeur de séquençage).

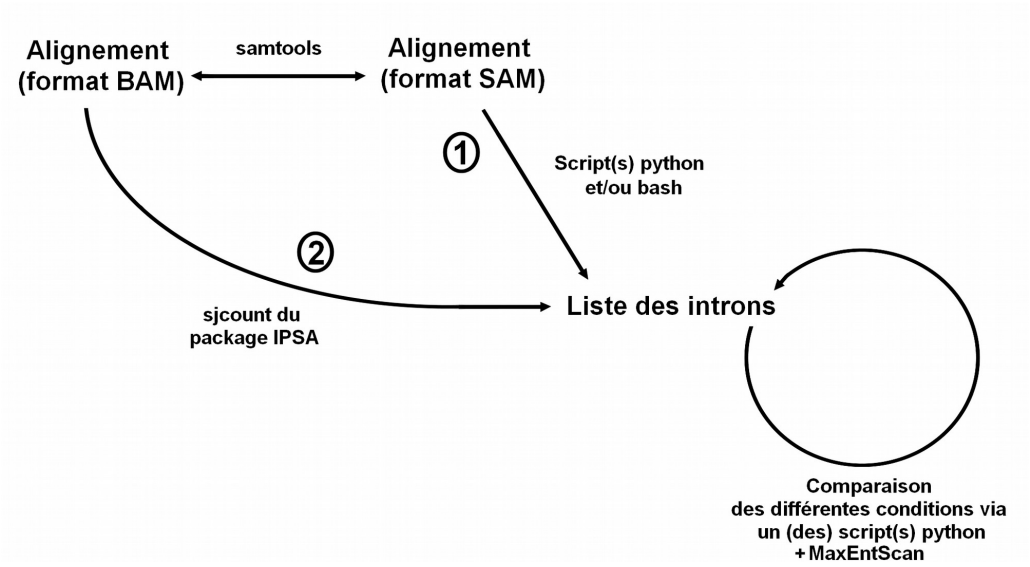

FIGURE 2 – **Schéma du pipeline d'analyse.** Le numéro 1 représente le pipeline si sjcount ne permet pas de réaliser l'analyse et inversement pour le numéro 2.

2.2 Besoins non fonctionnels

Le pipeline devra être réutilisable facilement ce qui implique l'utilisation de programmes robustes et peu gourmands en ressources.

3 Contraintes

3.1 Délais

Les délais du projet correspondent à ceux de la figure 3 ci-dessous.

FIGURE 3 – **Échéances intermédiaires et date de livraison du projet.**

4 Déroulement du projet

4.1 Planification

La récupération des introns devra être réalisée avant la fin du mois de novembre pour avoir le temps de comparer les résultats pour les différentes conditions et analyser les différents sites d'épissages afin de déterminer les plus intéressants et calculer leur force.

4.2 Documentation

Les scripts fournis en tant que délivrable seront accompagnés d'une documentation explicite au format texte et/ou pdf permettant une utilisation facile du pipeline afin que celui-ci puisse être réutilisé et adapté à des analyses ultérieures.

Bibliographie

[1] Anders S, Reyes A and Huber W. (2012) "Detecting differential usage of exons from RNA-seq data." Genome Research, 22, pp. 4025. doi : 10.1101/gr.133744.111.

[2] Dobin A, Davis CA, Schlesinger F, et al. STAR : ultrafast universal RNA-seq aligner. Bioinformatics. 2013 ;29(1) :15-21. doi :10.1093/bioinformatics/bts635.

[3] Fournier G, Chiang C, Munier S, Tomoiu A, Demeret C, Vidalain PO, Jacob Y, Naffakh N. Recruitment of RED-SMU1 complex by Influenza A Virus RNA polymerase to control Viral mRNA splicing. PLoS pathogens. 2014 ;10(6) :e1004164. Epub 2014/06/20. doi : 10.1371/journal.ppat.1004164. PubMed PMID : 24945353 ; PubMed Central PMCID :PMC4055741.

[4] Gene Yeo and Christopher B. Burge. Journal of Computational Biology. July 2004, 11(2-3) : 377-394. https ://doi.org/10.1089/1066527041410418.

[5] Jiang T, Nogales A, Baker SF, Martinez-Sobrido L, Turner DH. (2016) Mutations Designed by Ensemble Defect to Misfold Conserved RNA Structures of Influenza A Segments 7 and 8 Affect Splicing and Attenuate Viral Replication in Cell Culture. PLOS ONE 11(6) : e0156906.

[6] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter (2014). InterProScan 5 : genome-scale protein function classification. Bioinformatics, Jan 2014 ; doi :10.1093/bioinformatics/btu031.

[7] Love MI, Huber W and Anders S. (2014) "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, pp. 550. doi : 10.1186/s13059-014-0550-8.

[8] Dmitri D. Pervouchine, David G. Knowles, Roderic Guigó. Intron-centric estimation of alternative splicing from RNA-seq data, Bioinformatics, Volume 29, Issue 2, 15 January 2013, Pages 273–274, https ://doi.org/10.1093/bioinformatics/bts678.

[9] Priore SF, Kierzek E, Kierzek R, Baman JR, Moss WN, et al. (2013) Secondary Structure of a Conserved Domain in the Intron of Influenza A NS1 mRNA. PLoS ONE 8(9) : e70615. doi :10.1371/journal.pone.0070615.

[10] The UniProt Consortium. Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D158–D169, https ://doi.org/10.1093/nar/gkw1099.