Rapport de stage au LBBE, mai juin 2020

Recherche de gènes humains dupliqués chez les chauves-souris à partir d'une approche bioinformatique

Antoine HEURTEL, dit Odd

Étudiant du master 1 Bioinformatique
Université Lyon 1 Claude Bernard

Les chauves-souris sont connues pour être réservoir de plusieurs virus qui peuvent-être mortels chez d'autres espèces de mammifères dont l'Homme. Leur particularité est d'être le plus souvent asymptomatique. Les chauves-souris ont une réponse inflammatoire très faible par rapport à d'autres mammifères.

Les chercheurs travaillent sur des hypothèses génétiques variées pour en expliquer la cause. Notamment, des duplications de gènes de l'immunité ou du métabolisme pourraient expliquer pourquoi la réponse immunitaire ou la réponse inflammatoire chez les chauves-souris font qu'elles ne sont généralement que des porteuses saines de virus. Afin de détecter les gènes dupliqués chez une espèce, le développement d'outils bioinformatiques est nécessaire pour traiter de manière rapide et efficace de grandes quantités de données.

Responsables du Master :
Vincent LACROIX & Céline BROCHIER-ARMANET

LBBE
LABORATOIRE DE BIOMÉTRIE ET BIOLOGIE ÉVOLUTIVE

Lyon 1

Sommaire

Abréviations utilisées

AA : acides aminés
RefSeq : code d'identification d'un élément dans la base NCBI

Index des tableaux

Index des figures

Liste des logiciels utilisés

- Python 3.7
- BLAST 2.7.1+

Stage réalisé au laboratoire de biologie, biométrie et évolution (LBBE) de l'université Lyon 1 Claude Bernard, dans l'équipe de Dominique PONTIER, sous l'encadrement de Dominique PONTIER et co-encadrement de Stéphanie JACQUET.

Introduction

Les changements climatiques et les destructions d'habitats de certaines niches écologiques font que certaines espèces se rapprochent de plus en plus des habitations humaines, augmentant ainsi les interactions entre espèces. Malheureusement, cette cohabitation forcée n'est pas sans conséquence sur la santé humaine comme le montre la pandémie mondiale actuelle de Covid-19. Actuellement, on identifie le pangolin (Cyranoski 2020) et une espèce de chauve-souris (Sen et al. 2020) comme étant les vecteurs et réservoirs d'un virus apparentés à l'actuel SRAS Cov 2, le virus responsable de l'actuelle pandémie mondiale chez l'Homme.

Les chauves-souris sont parmi les mammifères, qui hébergent le plus grand nombre de virus et parmi eux de zoonoses. D'après l'étude de Mollentze et Streicker, les chauves-souris porteraient 8 fois plus de virus que les lapins (Mollentze et Streicker 2020). Parmi ces virus, on compte les virus de la rage, du SRAS (Banerjee et al. 2019), du Nipah, d'Ébola, d'Hendra.

Il existe une très grande diversité virale chez les chiroptères (ordre des chauves-souris), qui en plus, généralement ne développent aucun symptôme, alors que certains de ces virus sont mortels pour l'Homme (Luis et al. 2013).

Pour expliquer que les chauves-souris restent asymptomatiques, plusieurs hypothèses ont été avancées. On a notamment évoqué leur métabolisme et des adaptations uniques de leur système immunitaire permettant aux chiroptères d'interagir avec les virus sans développer de signes cliniques de la maladie.

Les chauves-souris pratiquent le vol actif, ce qui implique que leur métabolisme est plus élevé que celui des primates par exemple (Sharma et al. 2018). Des adaptations génétiques liées au métabolisme des chiroptères permettraient de libérer l'énergie nécessaire au vol et de faire face aux effets néfastes liés au métabolisme élevé (libération d'oxydant), ce qui en retour, contribueraient également à diminuer leur réponse inflammatoire (O'Shea et al. 2014). De plus, leurs apports nutritionnels sont très différents de ceux d'autres mammifères, et reposent sur des régimes alimentaires très variés. Elles sont insectivores, frugivores, carnivores ou hématophages, ce qui pourrait avoir des conséquences sur les gènes du métabolisme, mais aussi participer à la diversité virale dont elles sont porteuses.

Une autre hypothèse voudrait que des gènes auraient subi plusieurs pressions de sélections comme des mutations entraînant des changements d'acides aminés (AA), ou des recombinaisons. Ceci générerait de nouvelles protéines qui conférerait aux chiroptères cette capacité de tolérance et de résistance aux virus.

Enfin l'hypothèse actuellement étudiée au sein de l'équipe du laboratoire met en avant une duplication de gènes responsable de l'immunité, comme l'ont montré de récentes recherches par exemple sur le gène APOBEC3 et son évolution au sein des chauves-souris (Hayward et al. 2018). Ces duplications permettraient à la chauve-souris de devenir tolérante et/ou résistante aux virus qui circuleraient dans leur organisme (Kacprzyk et al. 2017).

Étant donné que nous savons qu'il y a des duplications de certains gènes, nous sommes maintenant amenés à nous interroger sur la répartition de ces duplications au sein du génome. Sont-elles aléatoires, ou focalisées sur des familles de gènes spécifiques ?

Nous allons commencer notre analyse en nous intéressant aux gènes de l'immunité et du métabolisme. Pour cela, nous allons prendre comme référence des gènes humains. En effet, ces derniers sont relativement bien étudiés et donc documentés. Ceci facilitera le début de l'analyse, car nous pourrons commencer en utilisant une base humaine, fiable pour réaliser des comparaissions entre le nombre de duplications trouvées chez les humains et chez les chiroptères. Puis nous utiliserons une nouvelle base constituée de gènes de chiroptères pour réaliser une comparaison entre les gènes de différentes chauves-souris.

L'objectif de mon stage est de développer un pipeline. Nous avons besoin d'analyser plusieurs milliers de gènes de plusieurs génomes. Il existe actuellement des outils qui permettraient de répondre en partie à notre problématique (confère paragraphe ci-dessous). Malheureusement la plupart de ces outils ne permettraient pas de répondre correctement à nos questions, car ils sont très souvent trop spécifiques à un gène. De plus, les données d'entrées nécessaires pour notre étude ne sont pas toujours adaptées pour les pipelines existants comme *GenoDup* ou *DuplicationDetector* (Mao 2019; Djedatin et al. 2017). Ou encore, ces outils existants sont trop axés sur la phylogénie des espèces des chauves-souris en détectant des gènes orthologues ayant eu une sélection positive (Hawkins et al. 2019). Un développement bioinformatique s'avère nécessaire et un nouveau pipeline sera établi dans le cadre de ce projet.

Lors du développement du pipeline, la difficulté principale est de détecter des gènes dupliqués et de ne pas relever un isoforme à la place. Ici un isoforme est un gène dont les exons sont agencés différemment. D'un point de vue biologique, il s'agit du même gène et non d'une duplication (i.e., des protéines issues de la traduction d'un même gène qui se différencient par l'ajout ou la perte d'une partie de leurs séquences en acides aminés). De plus, lors de la détection des gènes dupliqués il sera nécessaire de s'interroger sur son appartenance ou non à la même famille de gène. En effet, relever des gènes de la même famille pourrait biaiser nos résultats, alors qu'il s'agirait de gènes dupliqués. Mais ces gènes doivent être des duplications récentes, pour ne pas avoir d'ancêtre commun avec d'autres mammifères.

La réalisation de ce pipeline permettra de pouvoir répondre à nos hypothèses de travail concernant la présence de gènes dupliqués chez les chiroptères. L'objectif étant de guider les recherches biologiques pour comprendre la possible implication de ces duplications, dans la capacité des chiroptères à être asymptomatiques en présence de nombreux virus. De plus, il sera également possible de réaliser la même étude chez les primates afin d'obtenir une comparaison du niveau de duplication de leurs gènes à ceux du groupe des chiroptères.

Avec les avancées technologiques, de plus en plus de génomes sont disponibles permettant de faire de la génomique comparative. Chez les chiroptères il existe 43 génomes publiés sur le site du NCBI (à la date du 21 juin), dont plusieurs sont de bonne qualité ce qui permet d'investiguer les adaptations génétiques par rapport à l'évolution de leur génome et de répondre aux questions biologiques.

Pour répondre, nous nous intéresserons aux gènes de l'immunité, en récupérant ces gènes chez l'Homme. L'avantage des gènes de l'immunité humaine est que ces derniers sont relativement bien étudiés et donc documentés, ce qui facilitera le début de l'analyse; Nous pourrons commencer en utilisant une base humaine, fiable pour réaliser les comparaissions ; avant de passer sur les gènes de la chauve-souris.

Notons également que les gènes annotés comme intervenant dans la réponse inflammatoire seront traités séparément. Dans un second temps, nous nous intéressons aux gènes du métabolisme, en récupérant les gènes humains. Enfin, nous nous intéresserons aux gènes de la reproduction dans

un but de contrôle qualité de notre pipeline. En effet, pour cette dernière catégorie, nous n'attendons pas de duplications et donc pas de différences significatives entre le groupe des chiroptères et celui des primates.

Ces listes de gènes seront soumises à un `tblastn` sur des génomes annotés des chauves-souris. C'est à partir des résultats de ces `tblastn` que seront extraites les listes de gènes par espèces. Rajoutons également que lors du développement, de nombreuses étapes de contrôles seront mis en place afin de déceler tout éventuel bogue qui pourrait biaiser les résultats.

Ainsi, nous pourrons visualiser si les gènes impliqués dans la réponse immunitaire sont plus souvent dupliqués chez les chauves-souris que chez l'être humain (et primates) de manière générale. À terme, nous pourrons également envisager de réaliser une étude phylogénétique de certains gènes relevés comme étant dupliqués et donc intéressant pour de futures études. Cela permettra d'être plus précis en identifiant des familles de gènes, et de pouvoir identifier des gènes soumis à des pressions de sélections plus importantes. De plus, il serait également envisageable de pousser plus finement les analyses pour mieux comprendre, par exemple, les stratégies développées par les chauves-souris pour lutter contre les infections. De même, chez l'Homme, on pourrait mieux comprendre et améliorer les moyens de luttes contre les épidémies.

Ainsi s'il est attendu des duplications de gènes chez la chauve-souris par rapport à l'Homme. Il n'est pas censé en être de même avec les gènes de la reproduction qui sont des gènes conservés au cours du temps entre les espèces étudiées.

Matériels et méthodes

Constitution du jeu de données

Afin d'étudier comment les événements de duplications génétiques ont impacté le génome des chauves-souris, 3 catégories de gènes ont été retenues : l'immunité, le métabolisme et la reproduction. L'immunité et le métabolisme sont ici notre sujet d'étude. Les gènes de la reproduction nous servirons ici comme « groupe témoin ». En effet, les gènes impliqués dans la reproduction sont conservés chez les mammifères. Rajoutons simplement que, dans la catégorie de l'immunité, nous avons établi une liste de gènes responsables de l'inflammation.

La récupération des données s'est effectuée de manière manuelle depuis les bases d'*Uniprot*[1] (« UniProt: A Worldwide Hub of Protein Knowledge » 2019) et de *InnatDB*[2] (Breuer et al. 2013). Nous récupérons les séquences protéiques et pas génomiques, car les séquences protéiques sont plus conservées au cours du temps que les séquences génomiques qui sont davantage soumises aux mutations.

Nous choisissons *Uniprot* comme source principale de données. *Uniprot* se veut exhaustif en termes d'informations. De plus, nous utilisons la base *InnatDB* comme source complémentaire.

1 https://www.uniprot.org/
2 https://www.innatedb.com/

InnatDB est une base publique de gènes et de protéines (principalement axées sur l'immunité) dont les interactions et les voies de signalisations ont été vérifiées expérimentalement.

À partir de chaque base de données, les gènes ont été filtrés par taxon (celui de Homo Sapiens : 9606) et par `GO:therm`. Pour rappel, le *Gene Ontology*, est un projet ayant pour but de structurer et de contrôler le vocabulaire pour les annotations des gènes et des produits géniques dans une ontologie commune (« The Gene Ontology Project in 2008 » 2008).

La liste des `GO:therm` est disponible en ligne. Nous avons utilisé pour nos choix de gènes le site de l'*ebi*, QuickGO[3] (Huntley et al. 2015). Pour les gènes de l'immunité les `GO:therm` suivants ont été retenus : `GO:0002376` (pour l'ensemble du système immunitaire) et le `GO:0006954` (pour la réponse inflammatoire) ont été utilisés. Pour la catégorie métabolisme, le `GO:0008152` (pour le processus métabolique). Enfin pour la reproduction le go:`0022414` (pour le processus de reproduction).

Ces données ont été téléchargées en xml pour *Uniprot* et en tsv pour *InnatDB*. Ces fichiers constituent les entrées de notre pipeline d'analyse codé en Python. Le nombre de gènes récupérés depuis les bases de données en ligne est résumé dans le Tableau 1.

Lors de l'importation des données il peut arriver que des gènes soient présents dans les deux bases de données. Dans ce cas, le programme se chargera de rassembler les informations. En revanche, en cas d'informations contraires, ce sont les informations d'*Uniprot* qui seront considérées comme correctes par le programme et l'utilisateur sera notifié de ce problème. La base *Uniprot* a été arbitrairement choisie, car elle se veut exhaustive et rassemble un maximum d'informations.

Nous prêtons une attention particulière aux gènes de l'immunité en termes d'interactions virus/ hôte et aux gènes du métabolisme pour répondre à notre problématique.

Choix des espèces étudiées

Le choix de ces génomes s'est fait sur les critères de qualité et de divergence phylogénétiques. En effet, nous avons fait le choix de garder les génomes des espèces ayant une bonne qualité de séquençage et d'annotation tout en assurant une bonne représentativité des familles. Nous avons ainsi pris en compte la longueur du génome, le nombre de contigs ainsi que la N50 des contigs et le nombre de scaffold. Cela nous permet d'avoir un maximum d'informations (annotations) sur le gène sans que l'on risque de le retrouver fragmenté entre plusieurs contigs.

Ainsi à partir de la base de données NCBI (NCBI Resource Coordinators 2018), nous avons récupéré les génomes des espèces de chauves-souris suivants : *Pteropus vampyrus*, *Rousettus aegyptiacus*, *Rhinolophus ferrumequinum*, *Hipposideros armiger*, *Miniopterus natalensis*, *Myotis lucifugus* et *Phyllostomus discolor*.

3 https://www.ebi.ac.uk/QuickGO/

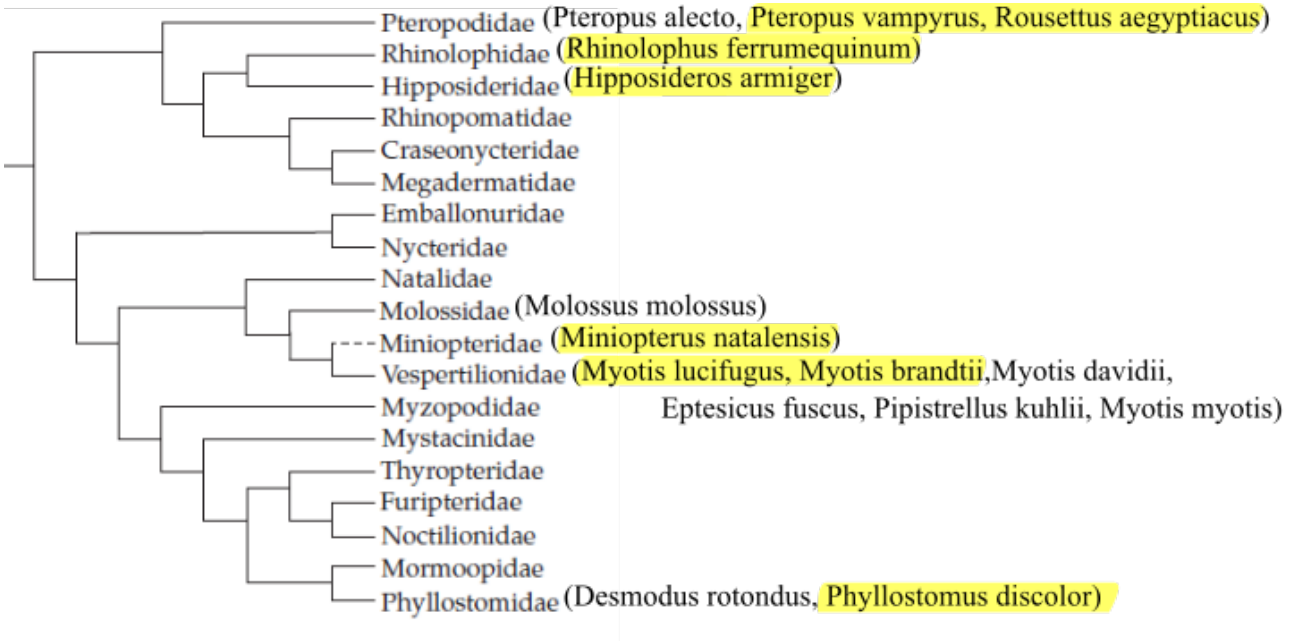

Figure 1 : Arbre phylogénétique des familles de l'ordre des chiroptères. Les espèces surlignées en jaune sont les espèces dont les génomes ont été récupérés et utilisés dans cette étude.

Source : Jacquet S., correspondance personnelle

Programme Python

Notre pipeline est constitué d'un programme codé en python contenant 2 sous-modules (`parser` et `blast`) selon les besoins de l'utilisateur Figure 2.

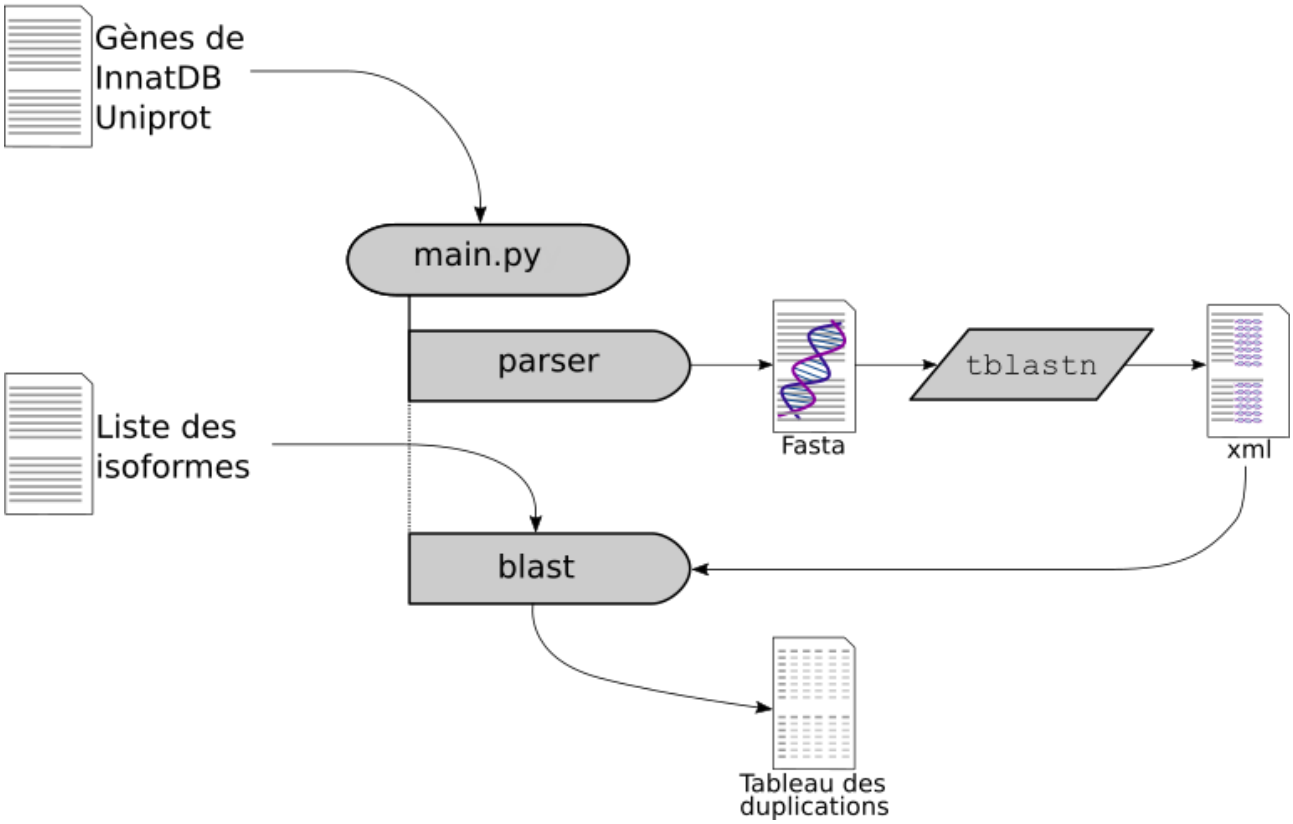

Figure 2 : logigramme du pipeline développé

Sous-module parser

Ce module permet à l'utilisateur de charger les données récupérées dans les bases de données en ligne. Afin d'accélérer les futurs chargements des données, j'ai implémenté une fonction d'exportation des données dans ce module.

Le parcours des fichiers en Python n'est pas une étape limitante dans le calcul et ne constitue pas une performance du programme (tout dépend de la puissance de l'ordinateur). En revanche, nos deux sources de données, *Uniprot* et *InnatDB* n'ont pas le même format d'écriture et donc pas la même distribution des données. Nous allons avoir besoin des séquences protéomiques pour la suite de notre analyse. Le fichier xml d'*Uniprot* est relativement exhaustif et contient les séquences d'acides aminées ; il n'en est pas de même pour le fichier tabulé de la base de données *InnatDB*, qui contient parmi ses informations le code *Ensembl* d'identification de la protéine, mais pas la séquence protéique.

Le module `parser` effectue 2 requêtes à *Uniprot* pour récupérer dans un premier temps le code accession du gène. Ce code permet ensuite d'obtenir la séquence protéique à l'aide de la seconde requête. Ces requêtes peuvent en fonction de la disponibilité des serveurs d'*Uniprot*, ne pas avoir le même temps de réponse. C'est pourquoi, j'ai implémenté une fonction d'exportation des données nouvellement chargées pour passer outre ces requêtes lors d'un futur appel au programme. Le fichier qui est alors écrit contient uniquement les informations nécessaires à notre analyse.

La commande `[In 1]` suivante permet de charger un fichier xml téléchargé depuis *Uniprot* et un fichier tabulé téléchargé depuis *InnatDB*. Enfin, l'option `-o` permet d'exporter les résultats dans un fichier `data.tsv`. Ce fichier nouvellement exporté pourra ensuite être directement chargé par le programme avec l'option `-l`.

```
[In 1]    parse -x fileUniprot.xml -i fileInnatdb.tsv -o data.tsv
```

Sous-module blast

Génération de fichiers fasta

Ce deuxième module propose à l'utilisateur de nombreuses fonctions pour manipuler des résultats issus d'un blast.

La première fonction de ce sous-module permet de générer des fichiers fasta à partir des séquences utilisées dans le sous-module `parser`. Cette fonction peut recevoir deux types de paramètres. Ce qui inclut soit un nombre de gènes défini à exporter, soit le tirage aléatoire d'un nombre défini de séquences protéiques à exporter en fasta parmi l'ensemble des gènes. Ainsi l'utilisateur pourra constituer ses propres jeux de données à partir des données initiales, qu'il pourra par la suite soumettre au blast.

Voici un exemple de commande pour générer un fichier fasta de 20 séquences protéiques à partir de données déjà traitées (`data.tsv` cf. commande `[In 1]`) :

```
[In 2]    blast -l inputData.tsv -fa -n 20
```

Traitement du fichier xml : résultat du tblastn

Avec les fichiers fasta générés précédemment, nous avons effectué un `tblastn` sur les génomes des chauves-souris. Le résultat de ce blast est téléchargé sous la forme d'un fichier xml qui sera ensuite parsé par le script.

Le `tbalstn` s'est effectué en local sur le serveur de l'université et du LBBE. Deux filtrages ont été appliqués : une E-Value supérieure à 10e-4 et une couverture des séquences à 60 %, en plus des paramètres par défaut.

La version actuelle de `tblastn` ne permet pas de filtrer le taux d'identité des hits, j'ai implémenté une fonction dans ce sous-module qui filtre le pourcentage d'identité des hits à conserver lors de l'importation du xml.

La commande `[In 3]` suivante parse le fichier de résultats `immune10.xml` en appliquant un filtre de positivité à 60 %.

```
[In 3]    main.py blast -b immune10.xml -f -o test -id 60
```

Notre objectif étant d'identifier les gènes dupliqués, il est nécessaire, lors du traitement du fichier de résultats de blast de ne pas retenir les protéines isoformes. Par définition un isoforme n'est qu'une version d'épissage possible connue de ce gène, ce n'est donc pas une duplication. Il est donc nécessaire pour chaque gène d'établir la liste des isoformes connus et de ne conserver que l'isoforme le plus long.

Pour chaque hit ayant été filtré, mon programme récupère les identifiants RefSeq des séquences pour la liste des isoformes connues associée à ce RefSeq (depuis les serveurs du NCBI). Cela permet de distinguer les duplications de gènes des isoformes. La liste des isoformes connus pour le gène renvoyé par `blast` est obtenu depuis les serveurs du NCBI. Seul l'isoforme ayant la séquence annotée la plus longue est retenu.

Effet du taux d'identité

Afin de correctement détecter les gènes d'intérêt, il va être important de correctement identifier les effets de la variation du taux d'identité.

Rappelons que ce taux d'identité correspond, dans notre cas, aux acides aminés (AA) qui partagent des caractéristiques physico-chimiques suffisamment équivalentes pour être considérées comme identiques. Une mutation d'un AA conduisant à un autre AA de même propriété sera donc pris en compte. En revanche, c'est le taux de positivité qui ne prend en compte que les correspondances strictes des AA.

Ainsi, le choix du taux d'identité ne doit pas être pris à la légère car, plus ce taux sera faible et plus nous aurons dans notre liste de gènes de séquences qui auront davantage divergées. Avec un taux faible, nous sommes plus à même d'obtenir des gènes appartenant à une même famille. Le fait que l'on ne retourne pas des gènes de la même famille peut paraître paradoxal ici. Mais notre but est la détection de gène dupliqué récemment.

Il s'agit donc ici d'une phase critique pour déterminer le meilleur pourcentage d'identité permettant à la fois de capter des éventuels paralogues tout en restreignant les hits de sortie au gène d'intérêt. Étant donné l'étendue de notre jeu de données, il est difficile de disposer pour chaque gène de son appartenance ou non à une famille de gènes. Si cette information avait été communiquée en amont, cela aurait permis de conforter la pertinence des résultats.

J'ai donc testé l'effet du seuil de pourcentage d'identité et les éventuels biais associés en choisissant une dizaine de gènes, qui sont membres de différentes familles de gènes et dont le taux de divergence est variable entre homologues. J'ai plus particulièrement testé différents seuils d'identités. Puis j'ai comparé les résultats obtenus à ceux obtenus de manière manuelle. En effet, il est important de passer par une méthode manuelle afin de pouvoir correctement identifier si les résultats obtenus sont ceux du gène d'intérêt (potentiellement dupliqué) ou ceux d'un autre gène de la même famille.

J'ai pu établir une première liste de gènes pour cette phase critique. Il s'agit des gènes de la famille IFIT (1, 2 et 3). Le cas de IFIT est intéressant, car nous savons qu'il s'agit d'une même famille qui s'est dupliquée il y a plusieurs millions d'années chez plusieurs mammifères (Liu et al. 2013).

Pour la détermination d'un taux d'identité convenable, nous avons retenu des gènes conservés chez plusieurs espèces comme `ITK`, `TAP` et `IFITMS`, et également `NLRC5` dans le but de visualiser l'impact du taux d'identité sur une protéine peu conservée (les séquences protéiques utilisées pour ce test sont disponibles dans le supplément d'informations du document).

Nous avons lancé le programme python pour les pourcentages d'identités suivants : 40, 50, 55, 60, 65, 70, 80 et 100. Pour chaque résultat, nous avons relevé le nombre de résultats du *tblastn* avant et après effet du paramètre de filtrage du taux d'identité. Puis pour calculer les sensibilités (Se) et spécificités (Sp), nous utilisons les nombres de vrais positifs (VP, nombre de duplications obtenues par le programme et confirmées par une analyse manuelle), les nombres de vrais négatifs (VN, nombre de résultats du *tblastn* n'étant pas des duplications trouvées par le programme et confirmées manuellement), de faux positifs (FP, nombre de gènes trouvés comme étant dupliqués par le programme à tort) et de faux négatifs (FN, nombre de gènes trouvés comme étant dupliqués par la méthode manuelle et non détectés par le programme python). Ainsi, nous obtenons les formules suivantes :

$$Se = \frac{VP}{VP+FP} \quad \text{et} \quad Sp = \frac{VN}{VN+FP}$$

Résultats

Gènes récupérés depuis les bases de données

La récupération des données depuis *Uniprot* et *InnateDB* nous permet d'obtenir un total de 14831 gènes, répartis dans plusieurs différentes catégories comme l'illustre le Tableau 1.

Nous constatons qu'*Uniprot* possède plus de gènes qu'*InnatDB* pour toutes les catégories de gènes envisagées : immunité, métabolisme et reproduction. En revanche, *InnatDB* possède un nombre de gènes de l'inflammation non négligeable par rapport à la base d'*Uniprot*. En effet, cela est bien en accord avec la caractéristique de la base d'*InnatDB* qui se veut spécialisée dans l'immunité. Le fait de rassembler les données de ces deux bases de données est donc un choix pertinent.

Tableau 1 : Synthèse des données téléchargées depuis les bases de données en ligne Uniprot et InnateDB

Base de données	Processus biologique (GoAccession)	Nombre de gènes (date)
Uniprot	Immunité GO:0002376	2774 (4 juin 2020)
	Inflammation GO:0006954	450 (4 juin 2020)
	Métabolisme GO:0008152	8306 (4 juin 2020)
	Reproduction GO:0022414	1419 (4 juin 2020)
InnatDB	Immunité GO:0002376	367 (4 juin 2020)
	Inflammation GO:0006954	315 (4 juin 2020)
	Métabolisme GO:0008152	1195 (4 juin 2020)
	Reproduction GO:0022414	5 (4 juin 2020)

Parsing du `tblastn`

À l'issue de cette étape, nous avons obtenu un total de 1112 résultats à partir d'un pool de 10 gènes dont la répartition des hits est donnée par le Tableau 2 suivant.

Tableau 2 : Répartition des hits (résultats) du tblastn effectué sur notre liste de 10 gènes.

Gènes	Nombre de hits
IFIT1	38
IFIT2	38
IFIT3	38
IFIT5	38
ITK	363
TAP1	106
TAP2	269
NLRC3	128
NLRC4	41
NLRC5	6

Nous constatons que la répartition des hits à l'issue du `tblastn` est très hétérogène.

Effet de la variation du taux d'identité

Pour la détermination du taux d'identité nous avons utilisé les gènes décris dans la partie Effet du taux d'identité. Après avoir lancé notre script, nous avons obtenu les résultats résumés dans le Tableau 3 et le Tableau 4.

À la vue des résultats, nous retiendrons qu'un pourcentage choisi entre 60 et 65 %, permet de récupérer les paralogues et les orthologues.

Nous remarquons que le nombre de résultats totaux obtenus par le `tblastn` reste constant au cours du temps, démontrant que nous n'avons aucune perte de résultats.

Le nombre de résultats après l'application du filtre décroît de manière progressive, montrant ainsi la bonne application de ce paramètre de filtrage.

Les nombres de faux positifs et de faux négatifs évoluent de manières cohérentes sur ce type de test. En effet, plus le taux est élevé et moins le programme trouvera des gènes orthologues et paralogues, faisant ainsi diminuer les faux positifs au profit des faux négatifs.

Tableau 3 : Résultats obtenus sur une liste de 10 gènes pour mesurer l'impact du taux d'identité

Taux d'identité	Nombre de résultats tblastn	Nombre de résultats tblatn après filtrage	Nombre de Vrais Positifs	Nombre de Vrais Négatifs	Nombre de Faux Positifs	Nombre de Faux Négatifs
20	1065	556	75	517	473	0
40	1065	310	75	837	153	0
50	1065	120	72	943	47	3
55	1065	95	72	967	23	3
60	1065	72	71	989	1	4
65	1065	64	64	990	0	11
70	1065	50	50	990	0	25
80	1065	33	32	990	0	43
90	1065	14	14	990	0	61
100	1065	0	0	990	0	75

Comme attendu, nous obtenons une baisse des vrais positifs. En effet, plus notre taux d'identité augmente et plus nous obtenons des séquences qui ont faiblement divergé et donc des duplications. Rappelons que nous cherchons les duplications les plus récentes chez les chauves-souris. Notons rapidement qu'avec un taux de 100 % nous n'obtenons aucun résultat de gène, puisque cela impliquerait de retrouver exactement le même gène qu'en entrée du `tblastn`.

Les valeurs calculées à partir du Tableau 3 et à l'aide des formules précédentes, permettent à nouveau de montrer une cohérence dans les résultats obtenus. Les taux de sensibilité et de spécificité évoluent de manière inverses l'une par rapport à l'autre.

Tableau 4 : Évolution des valeurs de sensibilité et spécificité

Taux d'identité	Sensibilité	Spécificité	Précision
20	1	0,36	0,29
40	1	0,54	0,5
50	0,96	0,78	0,73
55	0,96	0,86	0,86
60	0,95	0,9	0,99
65	0,89	0,9	1
70	0,72	0,9	0,9
80	0,46	0,9	0,6
90	0,2	0,9	0,2
100	0	0,9	0

À partir de ces données, nous pouvons tracer les courbes d'évolution de ces valeurs afin de déterminer un seuil d'identité acceptable pour la suite de notre étude.

L'évolution du taux de précision donne une courbe de type gaussienne centrée sur une valeur proche de 62 % (cf Figure 4), d'où le fait que nous retenons 60 % comme étant un taux d'identité acceptable pour toute la suite de cette analyse. Précisons, qu'il s'agit de mesures moyennes effectuer sur les 10 gènes.

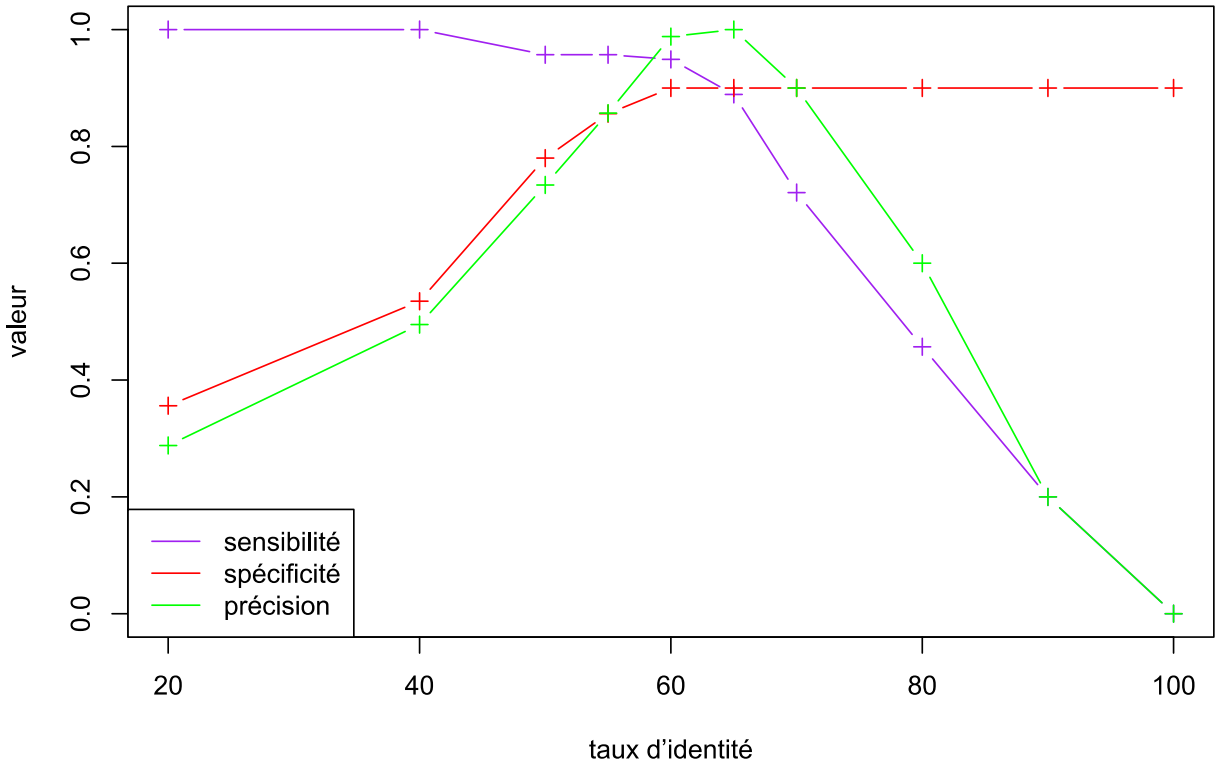

Figure 3 : Évolution des valeurs de sensibilité, spécificité et de la précision en fonction du taux d'identité utilisé dans le script python comme filtrage des résultats du tblastn. Chaque point est une moyenne calculée sur 10 valeurs, les variations de la moyenne ne sont pas représentées sur ce graphique pour des questions de visibilités

Première liste de gènes dupliqués

Après avoir appliqué le taux d'identité de 60 % sur les résultats du `tblastn` et sur notre liste de 10 gènes nous détectons plusieurs duplications de gènes.

Tableau 5 : nombre de duplications à partir d'une liste de 10 gènes chez 7 espèces de chauves-souris. Les valeurs dans le tableau indiquent le nombre de gène retenu par le programme.

gène	Pteropus vampyrus	Myotis lucifugus	Rhinolophus ferrumequinum	Hipposideros armiger	Miniopterus natalensis	Rousettus aegyptiacus	Phyllostomus discolor
IFIT1	1	2	2	2	1	1	3
IFIT2	1	1	1	1	1	1	1
IFIT3	1	1	1	1	1	1	1
IFIT5	1	1	1	1	1	1	1
ITK	1	2	1	1	1	1	1
TAP1	1	1	1	1	1	1	1
TAP2	0	1	1	1	0	1	1
NLRC3	0	1	1	1	1	1	1
NLRC4	1	1	1	1	1	1	1
NLRC5	1	0	1	1	1	1	1

Notons que le gène IFIT1 est dupliqué chez l'ensemble des espèces étudiées sauf chez *Miniopterus natalensis* et *Rousettus aegyptiacus*. Et que ce gène est même triplé chez *Phyllostomus discolor*. Ce gène est un inhibiteur des ARN viraux (Pichlmair et al. 2011).

Nous avons également le gène ITK qui est dupliqué chez *Myotis lucifugus*. Ce gène est indispensable dans la différentiation et la prolifération des lymphocytes T (Grasis, Browne, et Tsoukas 2003). La vérification des duplications s'est faite à l'aide d'arbre phylogénétique disponible en ligne comme *phylomedb* (Huerta-Cepas et al. 2014).

Discussion et Conclusion

L'objectif de ce stage était l'implémentation d'un pipeline dans le but d'évaluer si les duplications de certains gènes immunitaires ou métaboliques avaient un impact sur la capacité des chauves-souris à être asymptomatique, malgré le fait alors qu'elles sont le réservoir de plusieurs virus.

Le pipeline développé au cours de ce stage s'est montré performant dans la mesure où des gènes humains dupliqués chez la chauve-souris ont pu être mis en évidence. Ce pipeline a su répondre aux objectifs biologiques principaux. En effet, à l'issue de l'exécution de ce pipeline, nous obtenons 2 gènes de l'immunité qui présentent au moins une duplication parmi nos espèces de chauves-souris étudiés. Il s'agit des gènes IFIT1 et ITK (qui ne sont pas dupliqués chez l'Homme).

Nous avons commencé notre analyse en utilisant comme protéines de départs, des protéines humaines, ce qui est un point faible dans notre pipeline. En effet, BLAST est très efficace pour détecter les homologies. Si nous voulons détecter avec une plus grande précision les gènes dupliqués, il sera nécessaire de refaire cette analyse, mais cette fois avec une protéine de chauve-souris. Cela pourra être fait par exemple avec les deux gènes que nous avons détecté avec le pool de 10 gènes de départs.

Le choix du `tblastn` a été fait en utilisant les propriétés de BLAST. En effet, BLAST permet de réaliser des alignements locaux sur des génomes entiers ce qui permet de détecter correctement l'homologie des séquences.

Sur notre liste de 10 gènes, notre pipeline a pu relever 2 gènes immunitaires humains présentant des duplications chez les chiroptères. Ces 2 gènes ont un rôle dans l'immunité, le gène IFIT1 permet d'inhiber les ARN viraux (Pichlmair et al. 2011) et le gène ITK permet de réguler la prolifération et la différenciation des lymphocytes T (Grasis, Browne, et Tsoukas 2003). Ces 2 gènes ont donc un rôle majeur dans la réponse immunitaire contre les pathogènes.

Ces résultats suggèrent que la duplication de IFIT1 chez les chiroptères pourrait leur conférer la capacité à diminuer plus efficacement la charge virale des virus à ARN circulant dans leur organisme. D'autre part, la duplication du gène ITK, leur permet d'avoir une réponse immunitaire plus importante grâce à un meilleur contrôle et a une meilleure prolifération des lymphocytes T, qui rappelons-le, joue un rôle fondamental dans la réponse immunitaire. Ces adaptations génétiques majeures pourraient ainsi contribuer à expliquer pourquoi les chiroptères développent aussi peu de symptômes face aux infections virales.

Ces résultats sont en revanche à nuancer avec la conception de ce pipeline. Je pense notamment à la diminution du nombre de faux-négatifs. Nous pourrions améliorer la façon de détection et de tri des isoformes. Actuellement nous nous basons sur une seule liste d'isoforme. Nous pourrions par exemple étudier d'autres listes d'isoformes et ne pas forcément se focaliser sur l'isoforme le plus long mais par exemple sur un des isoformes les plus probables. Aussi, nous pourrions envisager de prendre en considération d'autres facteurs pour trier ces gènes, comme des paramètres phylogénétiques.

Le pipeline développé au cours de ce stage prend des séquences protéiques comme entrée pour les analyses subséquentes. En particulier, nous avons utilisé les séquences de protéines humaines, ce qui présente un point faible dans notre pipeline. En effet, BLAST est très efficace pour détecter les homologies. Pour détecter avec une plus grande précision les orthologues et les gènes dupliqués chez les chiroptères. Par la suite, il serait plus judicieux d'utiliser une protéine de chauve-souris.

Le choix du `tblastn` a été fait en utilisant les propriétés de BLAST. En effet, BLAST permet de réaliser des alignements locaux sur des génomes entiers ce qui permet de détecter correctement l'homologie des séquences. Un simple `blast` pourrait détecter les orthologues même si la divergence entre l'Homme et la chauve-souris peut parfois être élevée.

Le temps d'évolution des séquences doit également être pris en compte pour mieux choisir une liste de gène de références. Il est préférable de choisir un temps court. En effet, moins les séquences auront divergées entre celles de références et celles du génome d'étude et plus le blast sera capable de les aligner correctement et ainsi favoriser la détection des duplications. Si j'avais eu plus de temps de développement, j'aurais pu intégrer un sous module supplémentaire à ce pipeline qui prendrait en compte ce type d'événements, voire qui intégrerait des outils phylogénétiques.

Même avec ce pipeline, la vérification des gènes dupliqués passe par un contrôle manuel. Malgré cela, mon pipeline apporte la possibilité de détecter des gènes orthologues et paralogues. Son utilisation peut donc être exploitée dans d'autres études scientifiques qui exploiteraient des listes de gènes orthologues. De plus, ce pipeline n'est pas très exigeant sur les données d'entrées. Il se contente de fichiers tabulés et de résultats d'un `tblastn`. Cela offre donc un champ d'application non limité à une espèce ou à une famille de gène comme il en est question dans certains pipelines. D'autre part, il peut être facilement accessible au plus grand nombre de chercheurs.

Malheureusement, mon pipeline se limite à la prédiction de gènes en se basant sur les résultats obtenus après un `tblastn`. Ces résultats sont comme nous l'avons démontré lors de l'établissement d'un taux d'identité, très dépendants de ces paramètres. Ici nous nous sommes intéressées à

l'influence de ce taux. Mais il serait également nécessaire de mesurer les influences sur les résultats des autres paramètres du BLAST, tels que le nombre de gap, ou la taille du mot initial qui permet de réaliser les alignements locaux.

Ce projet mériterait une étude plus approfondie des résultats. Je pense notamment à une comparaison statistique gène par gène entre les espèces, qui permettrait d'obtenir une liste de gènes beaucoup plus pertinente. Ce test statistique pourrait déterminer s'il existe une différence significative du nombre de duplication chez un groupe par rapport à l'autre. À l'heure de la rédaction de ce rapport, je n'ai pas pu lancer le pipeline sur les autres catégories de gènes. On ne peut donc pas ici présenter une liste de gènes plus exhaustive que celle présentée plus haut.

Enfin une dernière perspective d'utilisation de mon pipeline serait des études évolutives. Comme la détermination de gènes dupliqués qui seraient soumis à des sélections positives, par exemple en combinant mon pipeline avec DGINN. Il s'agit d'un autre pipeline conçu par le laboratoire pour déterminer les régions géniques soumises à sélections positives (Picard et al. 2020).

Enfin, on pourrait peut-être envisager de changer de groupes d'études comme les rongeurs ou les oiseaux qui sont également des réservoirs pathogènes.

Remerciements

Je remercie Dominique Pontier et Stéphanie Jacquet pour leur accueil et leur sujet de stage qui a été très intéressant, et au cœur de l'actualité. Durant toute la période de mon stage, elles ont été disponibles malgré la période difficile. Je les remercie également pour les nombreux retours et commentaires sur la rédaction de ce rapport de stage. Merci à Dominique pour son partage de connaissances sur les chauves-souris et bien d'autres sujets scientifiques. Merci à Stéphanie pour ses nombreux conseils et retours d'expériences qui m'ont permis de progresser tout au long de mon stage.

Je remercie aussi mon camarade de promotion, Théophile Tesseraud, *dit Thoto*, pour les bons moments de camaraderies malgré le confinement, et sa collaboration durant les premières semaines de notre stage où le travail collaboratif, même distancé par plusieurs dizaines de kilomètres, a toujours été efficace.

Je remercie également le service informatique du LBBE. Je pense notamment à Bruno Spataro et à Adil El Filali, qui m'ont permis d'avoir accès aux ressources informatiques du laboratoire et m'ont donné des conseils techniques.

Enfin, je remercie également mes amis de promotion vers qui j'ai pu trouver et partager, malgré la distance, des conseils et un soutien constant durant ces deux derniers mois.

À vous tous, bonne continuation…

Codes sources

L'intégralité du code Python ayant servi pour cette analyse est disponible sur la plateforme *GitHub* à l'adresse suivante : https://github.com/AntoineHeurtel/FDGBM

Sur ce même dépôt se trouvent les séquences fasta qui ont servi à établir le taux d'identité ainsi que les résultats détaillés qui ont servi à établir le graphique montrant l'évolution de la sensibilité, spécificité et précision.

Bibliographie

Banerjee, Arinjay, Kirsten Kulcsar, Vikram Misra, Matthew Frieman, et Karen Mossman. 2019. « Bats and Coronaviruses ». *Viruses* 11 (1). https://doi.org/10.3390/v11010041.

Breuer, Karin, Amir K. Foroushani, Matthew R. Laird, Carol Chen, Anastasia Sribnaia, Raymond Lo, Geoffrey L. Winsor, Robert E. W. Hancock, Fiona S. L. Brinkman, et David J. Lynn. 2013. « InnateDB: Systems Biology of Innate Immunity and beyond—Recent Updates and Continuing Curation ». *Nucleic Acids Research* 41 (D1): D1228-33. https://doi.org/10.1093/nar/gks1147.

Cyranoski, David. 2020. « Mystery Deepens over Animal Source of Coronavirus ». *Nature* 579 (7797): 18-19. https://doi.org/10.1038/d41586-020-00548-w.

Djedatin, Gustave, Cécile Monat, Stefan Engelen, et Francois Sabot. 2017. « DuplicationDetector, a Light Weight Tool for Duplication Detection Using NGS Data ». *Current Plant Biology*, Special issue on Plant Development, 9-10 (juin): 23-28. https://doi.org/10.1016/j.cpb.2017.07.001.

Grasis, Juris A., Cecille D. Browne, et Constantine D. Tsoukas. 2003. « Inducible T Cell Tyrosine Kinase Regulates Actin-Dependent Cytoskeletal Events Induced by the T Cell Antigen Receptor ». *The Journal of Immunology* 170 (8): 3971-76. https://doi.org/10.4049/jimmunol.170.8.3971.

Hawkins, John A., Maria E. Kaczmarek, Marcel A. Müller, Christian Drosten, William H. Press, et Sara L. Sawyer. 2019. « A Metaanalysis of Bat Phylogenetics and Positive Selection Based on Genomes and Transcriptomes from 18 Species ». *Proceedings of the National Academy of Sciences* 116 (23): 11351-60. https://doi.org/10.1073/pnas.1814995116.

Hayward, Joshua A., Mary Tachedjian, Jie Cui, Adam Z. Cheng, Adam Johnson, Michelle L. Baker, Reuben S. Harris, Lin-Fa Wang, et Gilda Tachedjian. 2018. « Differential Evolution of Antiretroviral Restriction Factors in Pteropid Bats as Revealed by APOBEC3 Gene Complexity ». *Molecular Biology and Evolution* 35 (7): 1626-37. https://doi.org/10.1093/molbev/msy048.

Huerta-Cepas, Jaime, Salvador Capella-Gutiérrez, Leszek P. Pryszcz, Marina Marcet-Houben, et Toni Gabaldón. 2014. « PhylomeDB v4: Zooming into the Plurality of Evolutionary Histories of a Genome ». *Nucleic Acids Research* 42 (Database issue): D897-902. https://doi.org/10.1093/nar/gkt1177.

Huntley, Rachael P., Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J. Martin, et Claire O'Donovan. 2015. « The GOA Database: Gene Ontology Annotation Updates for 2015 ». *Nucleic Acids Research* 43 (D1): D1057-63. https://doi.org/10.1093/nar/gku1113.

Kacprzyk, Joanna, Graham M. Hughes, Eva M. Palsson-McDermott, Susan R. Quinn, Sébastien J. Puechmaille, Luke A. J. O'Neill, et Emma C. Teeling. 2017. « A Potent Anti-Inflammatory Response in Bat Macrophages May Be Linked to Extended Longevity and Viral Tolerance ». *Acta Chiropterologica* 19 (2): 219-28. https://doi.org/10.3161/15081109ACC2017.19.2.001.

Liu, Ying, Yi-Bing Zhang, Ting-Kai Liu, et Jian-Fang Gui. 2013. « Lineage-Specific Expansion of IFIT Gene Family: An Insight into Coevolution with IFN Gene Family ». *PloS One* 8 (6): e66859. https://doi.org/10.1371/journal.pone.0066859.

Luis, Angela D., David T. S. Hayman, Thomas J. O'Shea, Paul M. Cryan, Amy T. Gilbert, Juliet R. C. Pulliam, James N. Mills, et al. 2013. « A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? » *Proceedings of the Royal Society B: Biological Sciences* 280 (1756). https://doi.org/10.1098/rspb.2012.2753.

Mao, Yafei. 2019. « GenoDup Pipeline: a tool to detect genome duplication using the dS-based method ». *PeerJ* 7 (janvier). https://doi.org/10.7717/peerj.6303.

Mollentze, Nardus, et Daniel G. Streicker. 2020. « Viral Zoonotic Risk Is Homogenous among Taxonomic Orders of Mammalian and Avian Reservoir Hosts ». *Proceedings of the National Academy of Sciences* 117 (17): 9423-30. https://doi.org/10.1073/pnas.1919176117.

NCBI Resource Coordinators. 2018. « Database Resources of the National Center for Biotechnology Information ». *Nucleic Acids Research* 46 (D1): D8-13. https://doi.org/10.1093/nar/gkx1095.

Pichlmair, Andreas, Caroline Lassnig, Carol-Ann Eberle, Maria W Górna, Christoph L Baumann, Thomas R Burkard, Tilmann Bürckstümmer, et al. 2011. « IFIT1 Is an Antiviral Protein That Recognizes 5′-Triphosphate RNA ». *Nature Immunology* 12 (7): 624-30. https://doi.org/10.1038/ni.2048.

Sen, Sourav, Kavita Bala Anand, Santosh Karade, et R. M. Gupta. 2020. « Coronaviruses: Origin and Evolution ». *Medical Journal, Armed Forces India*, avril. https://doi.org/10.1016/j.mjafi.2020.04.008.

« The Gene Ontology Project in 2008 ». 2008. *Nucleic Acids Research* 36 (suppl_1): D440-44. https://doi.org/10.1093/nar/gkm883.

« UniProt: A Worldwide Hub of Protein Knowledge ». 2019. *Nucleic Acids Research* 47 (D1): D506-15. https://doi.org/10.1093/nar/gky1049.