Table des matières

[1 - Introduction 1](#_Toc512854511)

[2 - Matériels et Méthodes 2](#_Toc512854512)

[2.1 - Obtention de l’échantillon 2](#_Toc512854513)

[2.2 - Acquisition des données Illumina 2](#_Toc512854514)

[2.2.1 - Préparation de la banque 2](#_Toc512854515)

[2.2.2 - Séquençage 2](#_Toc512854516)

[2.2.3 - Démultiplexage et trimming des reads 3](#_Toc512854517)

[2.2.4 - Étape de préassemblage 3](#_Toc512854518)

[2.3 - Acquisition des données Nanopore 4](#_Toc512854519)

[2.3.1 - Préparation de la banque 4](#_Toc512854520)

[2.3.2 - Séquençage 4](#_Toc512854521)

[2.3.3 - Appel de bases 4](#_Toc512854522)

[2.3.4 - Trimming et filtration des reads 4](#_Toc512854523)

[2.3.5 - Correction des reads 5](#_Toc512854524)

[2.4 - Assemblage hybride des reads 5](#_Toc512854525)

[2.4.1 - SPAdes 5](#_Toc512854526)

[2.4.2 - MaSuRCA 5](#_Toc512854527)

[2.4.3 - Contrôle qualité de l’assemblage 5](#_Toc512854528)

[3 - Résultats 6](#_Toc512854529)

[3.1 - Séquençage Illumina 6](#_Toc512854530)

[3.2 - Séquençage Nanopore 6](#_Toc512854531)

[3.3 - Assemblage hybride 6](#_Toc512854532)

[4 - Discussion 7](#_Toc512854533)

[5 - Bibliographie 7](#_Toc512854534)

[6 - Figures 8](#_Toc512854535)

# Introduction

Le virus Nipah (NiV), découvert récemment en 1998 1, a depuis été associé à de sévère pathologies chez l’homme 2. Il présente une pathogénicité très élevée chez les espèces mammifères, à l’exception de la famille de chauves-souris Pteropus. En dépit d’être les hôtes de nombreux virus hautement pathogènes, dont le NiV, les membres de cette famille ne présente pas de symptômes cliniques d’infection (citation). Ce travail cherche à mettre en évidence des interactions moléculaires entre des individus du genre Pteropus et le virus NiV, qui pourraient aider à une meilleure compréhension concernant l’absence de pathogénicité chez Pteropus. Pour cela, l’étude se concentre sur l’espèce *Pteropus giganteus*, largement répandue en Inde et au Bangladesh où des foyers d’infection du NiV sont présents. Son génome n’étant pour le moment pas connu, il faut d’abord le séquencer *de novo*. La procédure expérimentale est décrite plus bas.

# Matériels et Méthodes

## Obtention de l’échantillon

Qualité :

Un ratio d’absorbance a d’abord été utilisé pour déterminer la pureté de l’ADN. Un ADN pur est caractérisé par un ratio 260/280 de 1,8 et 260/230 entre 1,8 et 2,2.

L’intégrité de l’échantillon a ensuite été déterminée par une électrophorèse sur gel d’agarose à 1%. Si les fragments d’ADN sont trop petits, cela indique une probable dégradation en cours au moment de la préparation de l’échantillon, ce qui peut biaiser les données.

Quantité :

Le kit utilisé (voir plus bas) nécessite une quantité minimum d’échantillon de 3µg. La quantification s’est effectuée avec l’appareil QuantiFluor dsDNA System (société Promega), qui utilise un agent fluorescent qui se lie avec l’ADN.

## Acquisition des données Illumina

### Préparation de la banque

La construction d’une banque s’est faite avec le kit NEXTflex PCR-Free DNA Library Prep, qui est adapté au séquençage sur Illumina (Figure 1). L’ADN est d’abord fragmenté en petits morceaux, puis une étape de nettoyage garde les brins d’ADN d’une longueur entre 300 et 400 pb approximativement (400-500 après fixation de l’adaptateur). Les extrémités sont ensuite réparées, et un adaptateur et un index (permettant d’identifier l’échantillon par la suite) viennent se fixer de chaque côté des brins.

Le Bioanalyzer (société Agilent), fonctionnant sur le principe d’une électrophorèse capillaire, a été utilisé pour contrôler la qualité de la banque. Une banque est de bonne qualité si la courbe représentant la densité de la taille des fragments d’ADN est une gaussienne avec un seul pic bien distinct autour de 450pb. Un résultat qui s’éloigne de la gaussienne signifie que la fragmentation ne s’est pas faîte aléatoirement, et pourrait présenter un biais.

### Séquençage

Le séquençage s’est effectué sur un appareil Illumina Nextseq 500 utilisant une Flow cell à high output, et une lecture des fragments 2 x 150 Nt (paired-end). Les fragments sont lus en paired-end, sur une longueur de nucléotides choisie (75, 150, etc…). Selon la taille des fragments et les applications concernées, la longueur des reads à choisir est importante car elle permet d’obtenir l’information souhaitée (Figure 1). Les fichiers de sortie du séquençage sont au format .bcl et contiennent les bases et leur score de qualité, stockés en format binaire.

Pour le type de flow cell utilisé, les spécifications du fournisseur sont les suivantes :

* Le nombre de clusters analysés doit être d’environ 400 millions (x2 en paired-end).
* Le taux d’erreurs doit être inférieur à 2% en moyenne pour chaque read
* La quantité de reads obtenus doit être supérieure à 130 millions.

### Démultiplexage et trimming des reads

Le démultiplexage consiste à transformer les fichiers BCL issus du séquençage en fichiers FASTQ, qui sont des fichiers textes contenant des séquences nucléotidiques (dans notre cas) et leur score de qualité associé. Lorsque plusieurs échantillons sont séquencés en même temps, il faut aussi lors de cette étape associer chaque read au bon échantillon. Dans ce but, un index spécifique à l’échantillon, généralement de 6pb, est rajouté aux fragments. Lors du démultiplexage, un seuil de misalignement est spécifié au niveau de l’index, qui peut être 0 (pas d’erreur acceptée lors de la lecture de l’index) ou 1 (une erreur acceptée). Par exemple, si pour un échantillon E l’index est ACTAGT, et le seuil spécifié est 1, la lecture d’un index ACTCGT sera bien attribué à l’échantillon E. Il est important de vérifier que la proportion du nombre de read associé à chaque échantillon est similaire aux proportions de départ (par exemple, une proportion de 50/50 est attendue dans le cas où deux échantillons ont été séquencés). Le logiciel utilisé pour le démultiplexage est bcl2fastq (v2.17.1.14).

Ce contrôle s’effectue avec le logiciel FastQC 3. FastQC analyse l’ensemble des reads pour en tirer des statistiques permettant d’estimer la qualité des reads. Les résultats sont présentés sous la forme de graphiques au format html.

Le trimming consiste à éliminer les reads ou parties de reads de mauvaise qualité. Généralement, au fur et à mesure de la lecture d’un read, la précision diminue, et donc le score de qualité du read. C’est pourquoi il est préférable de couper la fin des reads, afin d’obtenir des reads d’une longueur désirée approprié à l’étude et de bonne qualité. Les reads présentant également un score de qualité trop bas sont éliminés. Le logiciel Cutadapt (v1.3) 4 a ici été utilisé pour le trimming.

Un nouveau contrôle est effectué avec FastQC pour s’assurer que, si ce n’était pas le cas après le démultiplexage, les reads ont cette fois-ci une qualité suffisante pour les analyses ultérieures.

Un autre logiciel, FastQ Screen 5, est utilisé pour vérifier la pureté des échantillons concernant la contamination. FastQ Screen fait appel à un outil d’alignement externe. Ici bowtie2 a été utilisé pour ce contrôle. FastQ Screen prend un échantillon de reads (généralement 100 000) et aligne ces reads grâce à bowtie2 contre des génomes et des ensembles de génomes. Ici, l’échantillon a été comparé aux génomes d’Homo sapiens, de la Souris, de la Levure, d’E.coli, à deux génomes d’espèces proches (P. alecto et P. vampyrus) ainsi qu’à des bases de données génomique pour les plasmides et vecteurs (UNIVec), les virus et les bactéries.

### Étape de préassemblage

La taille du génome ainsi que la taille idéale des k-mers ont été estimées par le logiciel KmerGenie 6. Ce logiciel utilise l’histogramme d’abondance des k-mers distincts. Pour différentes valeurs de k, il calcule par échantillonnage aléatoire combien de k-mers distincts contient le set de reads. La valeur optimale pour k est alors celle qui donne le plus de k-mers distincts.

Le logiciel Lighter 7 a été utilisé pour la correction des erreurs. Il utilise le même principe que KmerGenie, l’abondance des k-mers distincts. Lighter a l’avantage d’utiliser des Bloom Filters, qui est une structure de données probabiliste, dont l’utilisation peut entraîner des faux positifs (mais pas de faux négatifs). Pour cette raison Lighter est plus rapide que d’autres logiciels de correction d’erreurs grâce à ces Bloom Filters, qui diminuent énormément la demande en temps de calcul et en mémoire sans sacrifier la précision ni la sensibilité.

## Acquisition des données Nanopore

### Préparation de la banque

### Séquençage

Le séquençage s’est fait sur un appareil MinION de Oxford Nanopore Technologies (ONT). Le principe est basé sur la mesure du profil électrique d’un nanopore, par lequel passe un courant ionisé. Lorsque qu’une molécule traverse le pore, une perturbation du courant électrique est enregistrée. La flowcell utilisée, r9.5, contient au minimum 800 puits, avec quatre pores au fond de chaque puits. Le séquençage génère des fichiers au format fast5, contenant les profils enregistrés des courants électriques.

### Appel de bases

Pour pouvoir identifier les reads et leurs bases à partir de ces profils, il faut les traduire en séquences interprétables. Il n’est en fait pas possible d’identifier correctement la ou les bases simplement en observant une perturbation dans le profil de courant, ou même une portion du profil. En effet, le profil du courant dépend de l’état du pore, de ce qui précède la base, de ce qui suit, et de l’encombrement général du milieu. Pour régler ce problème, Albacore utilise le principe des réseaux neuronaux : il s’agit d’un logiciel composé de plusieurs couches de fonctions. Lorsqu’on lui fournit des données à déchiffrer dont on connaît la solution, le logiciel va définir des suites de fonctions qui vont aboutir à un résultat correct. Une phase d’apprentissage sur un grand volume de données est donc nécessaire pour que le logiciel puisse ensuite interpréter correctement de nouvelles informations.

Nanopore fournit un logiciel d’appel de bases appelé Albacore, dont plusieurs versions existent. Nous avons utilisé la version la plus récente au moment de l’étude, la 2.1.10, la plus performante considérant plusieurs paramètres 8.

La qualité est estimée par Albacore, mais nous avons aussi choisi d’utiliser notre propre outil pour savoir si l’estimation fournie était correcte. L’outil Filtlong propose sa propre mesure de la qualité, en alignant les long reads issus du MinION avec des short reads issus de la technologie Illumina. Malheureusement, nous n’avons pas trouvé de corrélation significative entre les deux qualités estimées, c’est pourquoi nous avons choisi de ne pas tenir compte de la qualité estimée par Nanopore, pour ne prendre en compte que celle utilisé par Filtlong.

### Trimming et filtration des reads

Le trimming a été réalisé par le logiciel porechop. Ce logiciel scanne d’abord les données pour déterminer les adaptateurs issus du séquençage présents dans les reads, puis va couper les portions des reads contenant ces adaptateurs. Il détecte aussi les reads chimères, c’est-à-dire un read unique issu de deux fragments d’ADN distinct. De tels reads sont coupés par porechop.

Les reads trimmés ont ensuite été filtrés avec le logiciel Filtlong, qui a pour but, à partir d’un ensemble de données, de produire un second ensemble plus petit mais de meilleure qualité. Pour cela il prend en compte un ensemble de critère pour établir un score de qualité. Il est possible de spécifier quels critères utiliser, ainsi que le poids à leur donner. Nous avons choisi d’utiliser comme unique critère de qualité l’alignement des long reads sur les short reads d’Illumina.

### Correction des reads

Les reads ont été corrigés selon deux méthodes différentes, en utilisant Jabba et en utilisant HG-CoLoR. Les deux pipelines suivent le même principe : d’abord, la construction d’un graphe De Bruijn à partir de short reads corrigés issus de la technologie Illumina, puis l’alignement des long reads issus de la technologie Nanopore afin de corriger ces derniers. En revanche, la manière de construire le graphe, ainsi que celle d’aligner les long reads, ne sont pas les même pour Jabba et pour HG-CoLoR.

## Assemblage hybride des reads

Afin de réaliser un assemblage *de novo* avec des reads issus de deux technologies différentes, le choix de l’assembleur est une étape importante. Il existe deux assembleurs majeurs permettant de réaliser un assemblage hybride *de novo*: *SPAdes* 9 et MaSuRCA 10.

### SPAdes

SPAdes construit un graphe De Bruijn avec les short reads Illumina, puis cherche à résoudre ce graphe, en s’aidant des long reads afin d’obtenir un assemblage de meilleure qualité et moins fracturé. Il a déjà été utilisé avec succès pour assembler des génomes eucaryotes 11,12.

Nous avons fait tourner SPAdes avec différentes tailles de k-mer afin de retenir la taille produisant la meilleure qualité d’assemblage, puis l’assemblage final a été fait sur la taille de kmer la plus appropriée.

### MaSuRCA

L’approche de MaSuRCA est de transformer les short reads initiaux en super-reads, qui permettent (entre autres) un meilleur alignement avec les long reads sans perdre d’information. Ces super-reads sont pré-assemblés à l’aide des long reads selon un algorithme de ‘fusion et juxtaposition’, afin d’obtenir des mega-reads. Ces mega-reads sont finalement assemblés à l’aide d’une approche OLC (Overlap-Layout-Consensus).

### Contrôle qualité de l’assemblage

Afin d’évaluer l’assemblage crée par SPAdes, nous avons cherché à aligner les shorts reads d’Illumina dessus, en utilisant BWA-MEM 13, suivi d’une étape de réalignement des indels avec l’outil GATK (Genome Analysis Toolkit) InDelRealigner. Cet alignement a été évalué par Quast, qui permet d’obtenir de nombreuses statistiques pour établir une qualité.

# Résultats

## Séquençage Illumina

9,3µg d’échantillon ont été obtenus, ce qui est suffisant pour l’utilisation du kit, et l’analyse de la qualité de l’ADN a montré une bonne intégrité de l’échantillon (Figure 2) ainsi que la pureté de l’ADN, ce qui a permis de valider l’échantillon pour la construction de la banque.

La taille moyenne des fragments obtenus était de 462pb après préparation de la banque, ce qui est conforme pour le séquençage (Figure 3). En revanche, la courbe obtenue n’était pas une gaussienne comme attendue ce qui indique un possible biais de fragmentation.

Après séquençage, le taux d’erreur était bien inférieur à 2% pour chaque read (Tableau 1), et la quantité de reads excédait 130 millions, donc les résultats sont bien rentrés dans les spécifications du fournisseur.

L’analyse de FastQC après démultiplexage indiquait une bonne qualité des bases et des séquences. Cependant, certains reads présentaient des polyG à leur extrémité, ce qui a pu venir d’un biais de séquençage. FastQ Screen n’a pas montré de contamination avec d’autres organismes (Figure 4).

Les résultats avaient ensuite été calculés pour différentes options de démultiplexage et de trimming (Tableau 2). Comme on pourrait s’y attendre, un trimming plus élevé donne moins de reads et une moindre couverture. La qualité était très similaire pour des seuils choisis entre 115 et 140. Les reads présentant une chaîne polyG ont été trimmés, mais le résultat était sensiblement identique, ce qui a montré que la profondeur de séquençage n’est pas affectée par cet enrichissement.

La taille du génome a été estimée à 1 994 539 846pb. L’outil KmerGenie a permis d’estimer la taille idéale de k-mer, qui était 65 pour un seuil de trimming à 115 et 71 pour un seuil à 140 (Figure 5). La taille de l’assemblage prédite varie légèrement pour les deux seuils (1 977 606 166 et 1 923 969 911, respectivement).

Après correction des reads par Lighter, les tailles optimales de k-mers étaient 81 pour un trimming à 115 et 65 pour un trimming à 140 (Figure 5). Les tailles d’assemblages estimées sont légèrement plus élevées que sans la correction des reads (1 995 433 827 et 2 029 521 422, respectivement).

## Séquençage Nanopore

Quatre banques ont été préparées pour le séquençage, numérotées de 2 à 5. Le nombre total de bases pour chaque banque obtenu après le basecalling, le trimming et le filtre est visible sur la Table XX. Le séquençage par le MinION a permis d’obtenir 1 898 200 reads comptabilisant 12 921 689 433bp. Après trimming par Porechop, 12 132 994 458 ont étés retenues pour un nombre de reads sensiblement similaire (1 873 102). 1 756 444 reads ont été séléctionnés par Filtlong, pour un total de 12 119 501 988 bp. La taille du génome ayant précédemment été estimée à 2Gb, la couverture obtenue pour les long reads est 6X.

* Stats sur Filtlong
* Correction des reads

## Assemblage hybride

# Discussion

# Bibliographie

1. Chua, K. B. Nipah Virus: A Recently Emergent Deadly Paramyxovirus. *Science* **288,** 1432–1435 (2000).

2. Marsh, G. A. & Wang, L.-F. Hendra and Nipah viruses: why are they so deadly? *Curr. Opin. Virol.* **2,** 242–247 (2012).

3. Andrews, S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 30th April 2018)

4. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).

5. Wingett, S. Babraham Bioinformatics - FastQ Screen. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastq\_screen/. (Accessed: 30th April 2018)

6. Chikhi, R. & Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30,** 31–37 (2014).

7. Song, L., Florea, L. & Langmead, B. Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biol.* **15,** (2014).

8. Wick, R., Judd, L. M. & Holt, K. E. *Comparison Of Oxford Nanopore Basecalling Tools*. (Zenodo, 2018). doi:10.5281/zenodo.1188469

9. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).

10. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29,** 2669–2677 (2013).

11. Cao, M. D. *et al.* Scaffolding and completing genome assemblies in real-time with nanopore sequencing. *Nat. Commun.* **8,** 14515 (2017).

12. Faino, L. *et al.* Single-Molecule Real-Time Sequencing Combined with Optical Mapping Yields Completely Finished Fungal Genome. *mBio* **6,** e00936-15 (2015).

13. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).

# Figures

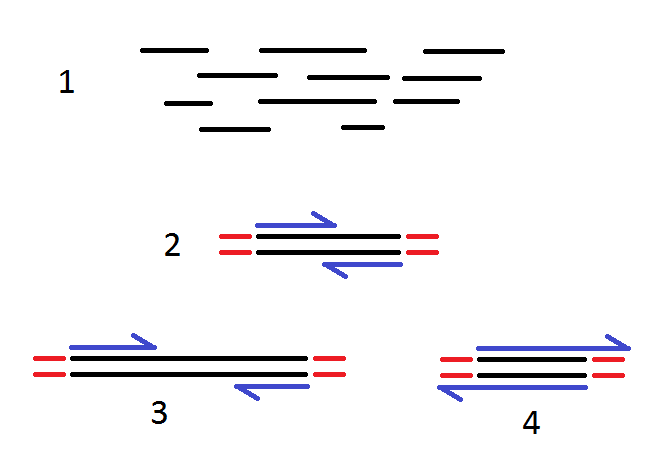


Figure 1: 1) Fragmentation de l'ADN. 2) Séquençage des reads en paired-end habituel. 3) Cas où les fragments sont très grands par rapport à la taille de reads choisie. 4) Cas où les fragments sont très petits par rapport à la taille de reads choisie.

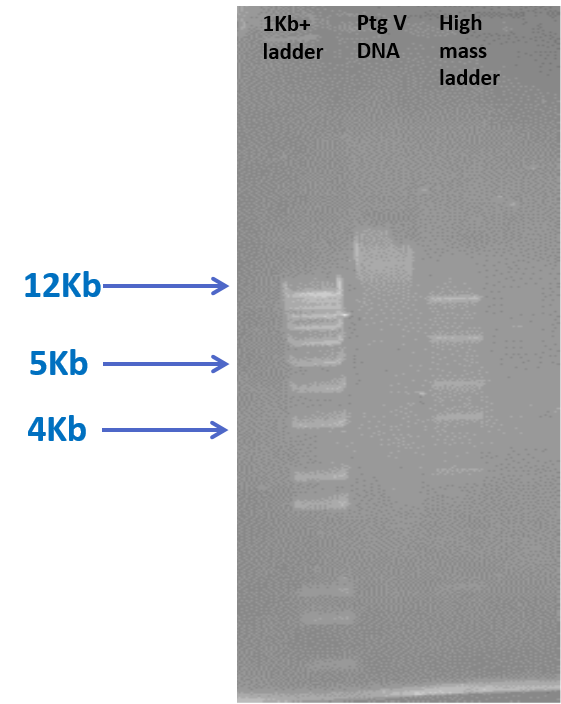


Figure 2: Contrôle qualité de l'échantillon sur un gel d'agarose à 1%.

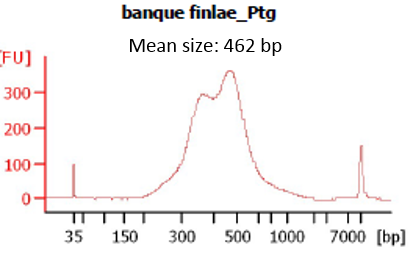


Figure 3: Taille moyenne des fragments de la banque. Elle doit se trouver entre 400 et 500pb.Les deux pics à chaque extrémité du graphique correspondent à des marqueurs et ne font pas partis de la banque.

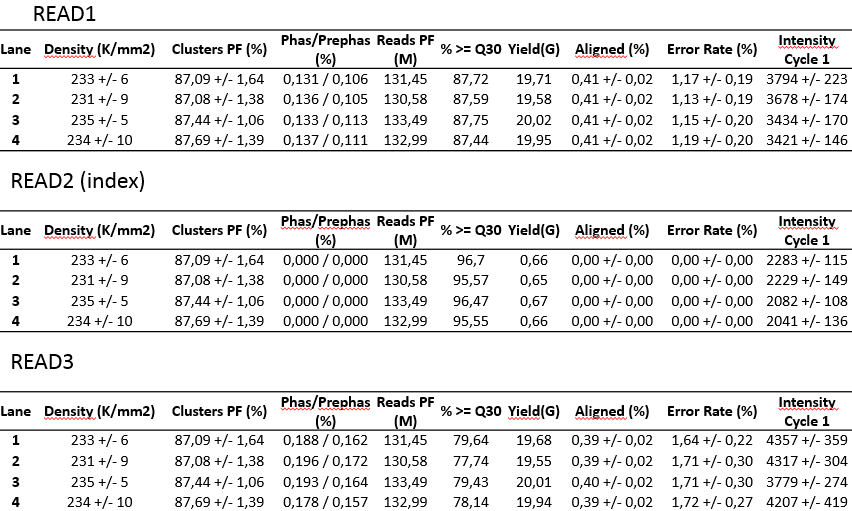


Tableau 3 : Résultat du séquençage.

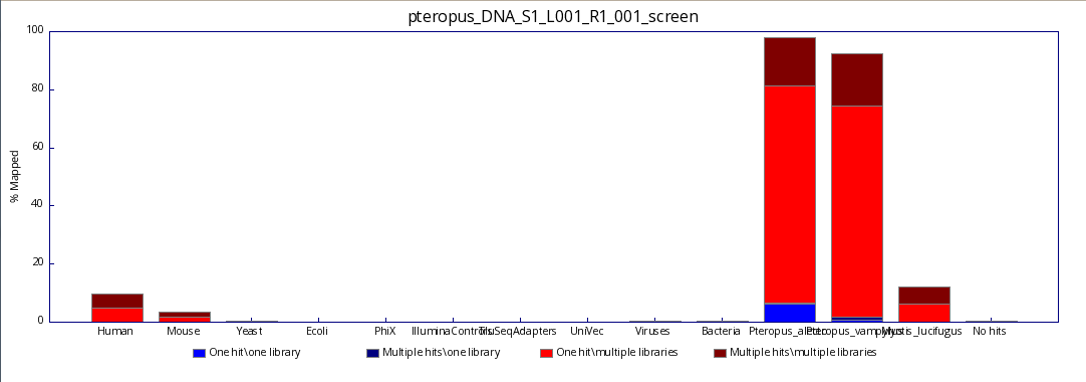


Figure 9 :Figure 4: Résultat de FasQ Screen. On peut voir qu'il n'y a pas de contamination d'autres organismes. Il est normal d'obtenir des hits pour les deux espèces très proches de chauve-souris, ainsi que pour les autres mammifères dans une moindre mesure.

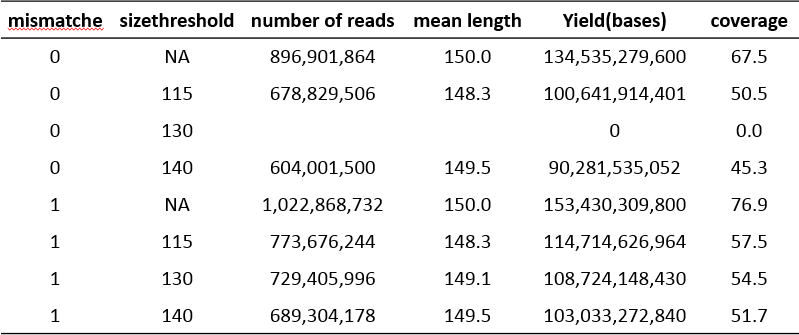


Tableau 4: Résultats selon différent paramètres de misalignement et de seuil de trimming, avant suppression des polyG.

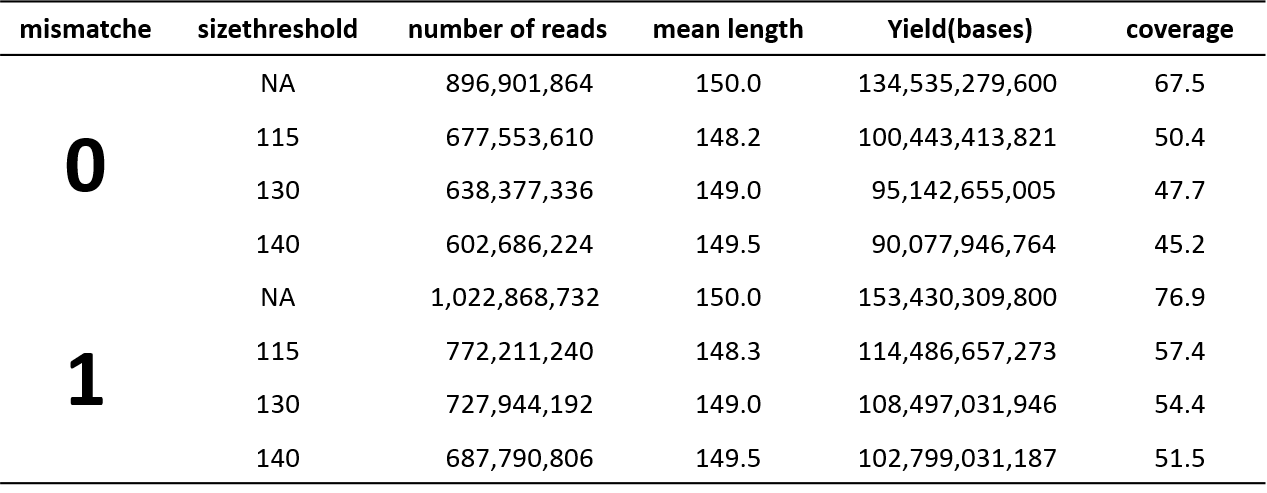


Tableau 5: Résultats selon différent paramètres de misalignement et de seuil de trimming, après suppression des polyG.

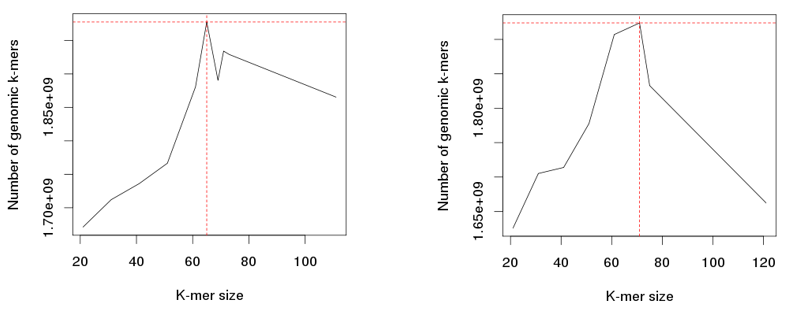


Figure 9 : Résultat de KmerGenie pour un trimming à 115 (gauche) et à 140 (droite).

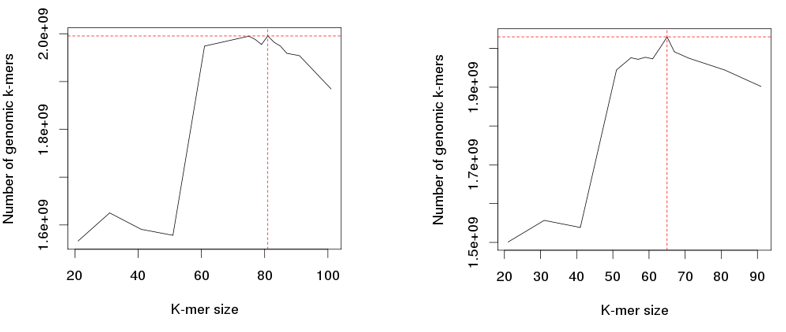


Figure 9 : Résultats de KmerGenie après correction des reads par Lighter, pour un trimming à 115 (gauche) et à 140 (droite).

