Écogénomique 2 : Synthèse du contrôle continue 3

Ce contrôle continue a été effectué en groupe de deux personnes dont Charlène BOULET et moi-même. Nous avons commencé ce contrôle le lundi 21 décembre 2020. Au départ, nous avons choisi de sélectionner l'article que j'ai présenté seule en UE Écogénomique 2, qui s'intitule « Comparison of DNA-, PMA-, and RNA-based 16S rRNA Illumina sequencing for detection of live bacteria in water » de Li. et al 2017. Nous avons décidé de sélectionner cet article, car son numéro d'accession était disponible alors que cela n'était pas le cas pour l'article que Charlène avait présenté. Ce dernier s'intitule « Prevalence of antibiotic resistance genes in drinking water and biofilms: The correlation with the microbial communities and opportunistics pathogens » de Jiping et al. 2020. Au cours de la recherche des données sur le site ENA, nous avons remarqué que toutes les données étaient dans un seul fichier .fastq.gz. Nous ne savions pas comment séparer les données et cela était encore plus difficile, car il y avait beaucoup de données. En effet, lorsque nous avons ouvert le fichier nous ne savions pas à quoi correspondait les différents caractères (amorces, l'origine des échantillons, et autres). Nous avons pris quelques jours pour prendre une décision et nous avons choisi de sélectionner un nouvel article qui s'intitule « Illumina MiSeq 16S amplicon sequence analysis of bovine respiratory disease associated bacteria in lung and mediastinal lymph node tissue » de Johnston et al. 2017. Nous avons choisi ce dernier car il était très intéressant et il correspondait à nos attentent, c'est-à-dire que les auteurs ont bien utilisé la technique d'illumina pour le séquençage. Nous avons pu observer également que les données étaient accessibles via le numéro d'accession. Nous avons pu voir que les données étaient classées dans différents fichiers .fastq.gz.

Après avoir importé les données, des bug ont subitement apparus. En premier lieu, au cours de l'importation des données ma VM était lente et ensuite elle a perdu tous les fichiers et le lien existant avec mon compte github. Malgré cela, j'ai reeffectué le lien avec mon compte github et reimporté les données, les packages et autres. Par la suite, j'ai pu importer les données de l'article. Ensuite, j'ai voulu me connecter à ma VM, mais la page chargeait dans le vide pendant des heures. Je ne pouvais pas ouvrir ma VM. J'ai essayé d'arrêter de me connecter pendant le soir du 24 décembre 2020 et j'ai essayé de me connecter tous les jours jusqu'au samedi 26 décembre 2020. Suite à cela, j'ai décidé de créer une nouvelle VM. J'ai pu installer les différents packages et importer les fichiers. J'ai pu reimporté les données de l'article, mais un problème est survenu dû à une panne au niveau du serveur de l'UBO. De ce fait, nous sommes restés quelques jours sans utiliser notre VM.

Après avoir essayé de se connecter tous les jours, j'ai pu réussir à ouvrir ma VM le dimanche 04 janvier 2021. Nous avons pu continuer le contrôle continue le lundi 05 janvier 2021, après notre stage. Nous avons pu commencer l'analyse des données de l'article via le tuto Dada2. Nous avons pu retracer les données dans notre fichier et observer le score de qualité des amorces Foward et Reverse. Après cela, nous avons pu filtrer nos données d'amorces. Cependant, un problème est survenu. En effet, nous avons affecté le modèle d'erreur à l'objet « errF », tel que la commande suivante : errF <- learnErrors(filtFs, multithread=TRUE). Après avoir effectué « RUN » de cette commande, j'ai perdu plusieurs fois l'environnement dans R. Suite à plusieurs tentative un message d'erreur s'affiche automatiquement. Nous ne pouvons pas sauter ou bâcler cette étape, car cela est nécessaire aux analyses antérieures. Nous pouvons émettre quelques hypothèses liées à ce problème.

Cela pourrait être dû à la quantité de données ou un problème au niveau de la VM. Nous essayons à mainte reprise de relancer les différentes commandes mais le message d'erreur apparait à chaque reprise.

Nous sommes déçus d'avoir ce problème au début du contrôle continue. On aurait voulu effectuer l'analyse des données, afin de les comparer avec les figures de l'article. En effet, il aurait été fort agréable et enrichissant d'effectuer par nos propres moyens l'analyse de données d'article via une plateforme de statistiques. Nous continuons mainte fois à essayer de régler ce problème en se référant à internet. Serait-il possible que ce soit une erreur de programmation ? ou une trop grande quantité de données à traiter ?

Après plusieurs essais, un second message d'erreur s'affiche : « Error in add(bin) : 'Calloc' could not allocate memory (100000512 of 1 bytes) ». En se référant à internet, il s'est avéré que le problème serait dû à la quantité de données. Comme on a une VM de 32 bits, elle serait limitée à environ 3 Go de mémoire, alors que dans notre cas nous dépassons la capacité de mémoire. Cela reste insuffisant pour charger nos données. La solution la plus adéquat serait de passer à un système de 64 bits.

Nous avons pensé à choisir un autre article. Cependant, en vue des dates de rendu qui s'approche et que nous sommes actuellement en stage, nous ne voulons pas bâcler un contrôle continu. Nous voulons essayer de trouver des solutions et comprendre réellement ce qui ne fonctionne pas. Nous préférons rendre un travail rendu avec de l'implication et de la compréhension qu'un travail bâclé.

Toutefois, nous voudrions revoir ce problème avec vous monsieur. Nous pouvons même attendre les vacances en fin juin, s'il le faut. Mais nous voudrions vraiment continuer ce CC3. Nous sommes très déçus de ne pas pouvoir le finir, car nous trouvons que c'est un très bon exercice pour l'utilisation de R. Nous aurions pu comparer ce que nous avons obtenus avec certaines figures de l'article.