- Nextflow -

Initiation sur genobioinfo

- TP -

Objectifs :

Cette formation a pour objectif de vous familiariser à l'utilisation des lignes de commande Nextflow.

Vous découvrirez notamment comment :

- Traiter des fichiers
- Comprendre les paramétrages
- Interpréter les résultats

Pour réaliser l'ensemble de ces exercices, vous avez besoin :

- De vous connecter au cluster de calculs en utilisant les login et mot de passe de votre compte «genobioinfo», que je vais vous donner en cours.
- Attention, ces comptes de formation ne sont valables qu'un mois.
- Merci de me rendre vos exercices dans un seul document PDF envoyé par mail : sarah.maman@inrae.fr

Exercice n°1 : Connexion à Genologin, création d'un répertoire de travail, téléchargement de fichiers à traiter.

Connexion à Genologin

Vous pouvez accéder au cluster en précisant votre login et mot de passe «genobioinfo».

Préparer son espace de travail

Créer un répertoire de travail « NEXTFLOW » dans son /work/user/NomDeFleur/ et télécharger les fichiers d'entrée du TP
http://web-genobioinfo.toulouse.inrae.fr/~sigenae/sarah/UPS/DATA/TP_TOMATES/

Merci de penser à organiser son répertoires en sous-répertoires significatifs

Exercice n°2: Préparer son fichier bash de lancement Nextflow

Préparer le fichier de lancement et lancer le job sur les données tomates

- Utiliser le pipeline Nextflow nf-core/rnaseq pour un traitement RNAseq des données tomates. Possibilité de choisir un autre pipeline nf-core si vous le souhaitez.
- Paramétrer au mieux votre run et expliquant le choix de vos paramètres. N'oubliez pas de vous référer à la documentation en ligne de Nextflow pour le choix et l'écriture des paramètres.
- Insérer votre ligne de Nextflow dans un fichier bash (.sh)
- Le nom du job sur le cluster doit être sous la forme de votre PrenomNom
- Lancer le job sur la workq
- La durée du job doit être de 1 jour maximum, affiner si possible.
- La mémoire du job doit être paramétrée au maximum de 6 G
- Lancer Nextflow avec le profil genobioinfo: Exemple : nextflow run nf-core/rnaseq -profile genotoul ……..
- La révision à utiliser est la version 3.4 ou 3.0 avec bioinfo/nfcore-Nextflow-v21.04.1
- Et enfin, lancer le job sur le cluster avec sbatch

Pour vous donner une idée du temps de traitement :
-[nf-core/rnaseq] Pipeline completed successfully-
Completed at: 30-sept.-2022 10:22:32
Duration : 47m 29s
CPU hours : 1.6
Succeeded : 48

Suivre le job avec seff, utiliser resume si nécessaire.

- Expliquer la sortie du seff.
- Expliquer l'intérêt du resume (qu'il soit utilisé ou pas).

Exercice n°3: Interpréter le report MultiQC ainsi que les principaux fichiers résultats obtenus.

Interprétation des principaux résultats

- Expliquer les principaux répertoires et fichiers de sortie.

Interprétation du report MultiQC

- Sources bibliographiques :
  https://github.com/nf-core/rnaseq/blob/master/docs/output.md#quality-control

  https://github.com/hbctraining/Intro-to-rnaseq-hpc-salmon/blob/master/lessons/qc_fastqc_assessment.md
  https://youtu.be/qPbIlO_KWN0
  https://multiqc.info/docs/#using-multiqc

Exercice n°3: Lancer ce pipeline sur des données NCBI.

Choisir deux échantillons Gadus m. sur le site du NCBI avec la réference fasta associées ; puis lancer ce pipeline RNAseq sur ces samples.
Je vous propose de choisir les échantillons Gadus_morhua : SRR2045415 , SRR2045416, SRR2045417.

Expliquer votre démarches et les résultats obtenus.