

Cette lettre d'information est destinée aux membres des équipes de recherche utilisant la plate-forme bio-informatique GenoToul. Elle a pour but de vous informer sur les évolutions de l'équipe, les nouveaux outils, services, conditions d'utilisation, projets et formations mis en place.

1/ Les prochains cycles d'apprentissage

Nous vous proposons 3 cycles d'apprentissage dans les mois qui viennent (2 sous Galaxy et un en ligne de commande) :

A) 2,5 jours de formation à l'analyse bioinformatique et biostatistique des données RNASeq sur génome de référence sous environnement Galaxy du 29 au 31 janvier 2017 (reste 4 places)

Ce cycle d'apprentissage vous est proposé par la plateforme Bioinfo Genotoul, l'équipe Sigeneae et l'unité MIAT.

Après un bref rappel sur l'environnement Galaxy nous vous présenterons les données RNAseq. Nous discuterons notamment des plans d'expérience et des biais des différentes technologies de séquençage de type NGS. Vous apprendrez à analyser la qualité des données RNAseq et à les nettoyer. Vous effectuerez ensuite l'alignement sur le génome de référence et vous serez en mesure de découvrir de nouveaux gènes ou de nouveaux transcrits. Vous effectuerez ensuite les comptages afin de générer le tableau qui vous servira à mettre en évidence les gènes différentiellement exprimés entre deux conditions. La dernière demi-journée sera consacrée à l'analyse statistique en utilisant les SARtools, c'est à dire l'exploration des données de comptage, la normalisation et l'analyse de l'expression différentielle.

Aucun pré-requis n'est nécessaire.

B) 2 jours de formation à l'alignement et à la détection de variants à partir de NGS sous Galaxy du 1^{er} au 2 février 2018 (reste 3 places)

Ce cycle d'apprentissage vous est proposé par l'équipe Sigeneae.

La première demi-journée sera consacrée à la présentation de l'environnement Galaxy, puis à la prise en main de l'instance toulousaine. Suivront ensuite 1,5 jours de formation vous permettant de comprendre les principaux formats de fichiers manipulés, d'apprendre à aligner les lectures sur un génome de référence et de détecter les SNP et petits indels via la suite GATK.

Aucun pré-requis n'est nécessaire.

Ces formations sont organisées sur le site INRA de Toulouse Auzeville.

Les tarifs sont disponibles à l'adresse suivante : <http://bioinfo.genotoul.fr/index.php/training-2/pricing/>.

Les inscriptions s'effectuent sur cette page : <http://bioinfo.genotoul.fr/index.php/training-2/training/>.

2/ Les logiciels installés

La liste des logiciels installés sur genotoul a été mise à jour sur le site web. Elle est accessible à la page : <http://bioinfo.genotoul.fr/index.php/resources-2/software/>.

Enfin, pour trouver un logiciel installé nous vous conseillons d'utiliser la commande suivante :

```
ls /usr/local/bioinfo/src|grep -i soft_name
```

De même pour connaître les différents modules configurés et les utiliser, les commandes : `module avail`, `module display NomDuModule` et `module load NomDuModule` s'avèrent particulièrement utiles.

3/ Les news de l'instance Galaxy Sigeneae / BioInfo Genotoul

De nouveaux outils sont disponibles sur votre instance Galaxy Sigeneae / BioInfo Genotoul (<http://sigeneae-workbench.toulouse.inra.fr/>):

- « Upload several files from Genotoul without impact too much your Galaxy quota » vous permet d'uploader un nombre important de fichiers en un seul clic, dans la limite de 1 To du work, sans trop consommer votre quota et sans préciser le chemin d'accès à chaque fichier, en ne donnant que le chemin d'accès à un répertoire qui contient l'ensemble des fichiers à charger dans Galaxy.
- « Tax4Fun » prédit les capacités métaboliques ou fonctionnelles des communautés bactériennes à partir des séquences 16S.
- « Detection viral sequences with VIP » : Pipeline d'analyse de données de métagénomique virale.

- « PEAR v0.9.6 » : ultrafast pair-end read merger.
- « Salmon » : Quantification de l'expression des transcrits à partir des données RNAseq en utilisant le concept de quasi-mapping.
- « multiQC » : MultiQC compile les fichiers de log et de résultats de nombreux outils (fastQC, STAR etc...) pour en faire une synthèse graphique sous la forme d'une page web.
- « BEAST » : Analyse évolutive de séquences et d'arbres phylogénétiques basée sur les principes bayésiens.

Des procédures détaillées, news et workflows ready-to-use sont disponibles sur la page d'accueil de Galaxy Sigeneae / BioInfo Genotoul.

Pour finir, il est maintenant possible de lancer un outil avec un nombre important de fichiers entrants grâce à la "dataset collection". En effet, si un outil doit être lancé de la même manière sur plusieurs fichiers, il n'est plus nécessaire de paramétrer et lancer l'outil autant de fois que vous avez de fichiers concernés. Il vous suffit de regrouper l'ensemble des fichiers à traiter par l'outil dans une "dataset collection", puis de paramétrer/lancer une seule fois l'outil sur cette "dataset collection".

Galaxy traite alors chaque fichier séparément sans avoir eu besoin de créer un pipeline pour chacun de ces fichiers ni avoir à lancer un même workflow pour chacun de ces fichiers.

4/ Attention : modification des adresses mails

Les adresses mails pour nous contacter ont changé :

- si vous avez besoin d'accompagnement bioinformatique pour un projet en construction ou déjà commencé : anim.bioinfo-Occitanie-Toulouse@inra.fr
- pour contacter le support de l'infrastructure : pour faciliter le traitement de votre demande nous vous conseillons de privilégier au maximum les formulaires disponibles dans la rubrique « Ask for » du site web. Si vous souhaitez tout de même contacter le support par mail, l'adresse est : support.bioinfo.genotoul@inra.fr.

5/ Evolution du cluster de calcul : démarrage de la phase 2

Le nouveau cluster de calcul est installé. Nous démarrons la deuxième phase. Comme nous l'avions annoncé dans la dernière lettre d'information de juin 2017, cette phase consiste à faire cohabiter l'ancien système (CentOS-6, SGE) avec le nouveau (CentOS-7, SLURM). La migration progressive des applications de bioinformatique va également commencer.

Vous serez progressivement invités à utiliser le nouveau cluster mais il faudra préalablement se former au nouveau système d'ordonnancement de travaux SLURM (<https://slurm.schedmd.com/>). Une journée d'information sera organisée courant 2018 sur les pratiques mises en place avec le nouveau cluster.

Pendant cette 2ème phase, la disponibilité du support informatique sera impactée. Nous vous remercions donc par avance d'anticiper au minimum d'une semaine vos demandes (calcul, mémoire, installation de logiciels, ouverture de comptes...).

Les principales caractéristiques techniques de ce nouveau cluster sont :

- deux nouveaux serveurs frontaux pour accéder à ce nouveau cluster (l'équivalent de genotoul).
- 48 nouveaux nœuds de calcul disposant chacun de 32 cœurs (64 threads) et 256Go de RAM.
- une nouvelle machine SMP de 48 cœurs (96 threads) et 1536Go de RAM (qui remplacera bigmem).
- un nœud de visualisation graphique avec une carte GPU.
- le stockage /work renforcé (en performance et en capacité).

6/ Article : « Intégration de données hétérogènes »

Le 25 octobre dernier est paru l'article suivant : Jérôme Mariette, Nathalie Villa-Vialaneix; Unsupervised multiple kernel learning for heterogeneous data integration dans Bioinformatics. Dans ce papier, les auteurs proposent une nouvelle méthode basée sur les noyaux multiples pour intégrer et analyser des jeux de données de nature hétérogène. Elle a été appliquée à deux jeux de données multi-omiques : des données de métagénomique de l'expédition TARA et des données concernant le cancer du sein issues du « Cancer Genome Atlas ». Les méthodes proposées sont implémentées dans un package R : mixKernel compatible avec le package mixOmics.

7/ Retrouvez-nous sur twitter !

Nous nous sommes dotés d'un compte twitter : [@BioInfoGenotoul](https://twitter.com/BioInfoGenotoul). Vous y retrouvez toutes nos actualités !

Pour toute demande d'information ou de travaux, veuillez remplir le formulaire adéquat sur la page : <http://bioinfo.genotoul.fr/index.php/ask-for/support/>.