Lettre d'information N°19

du 15 oct 2014 – Rédacteur : Claire Hoede

http://bioinfo.genotoul.fr/

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, projets et formations mis en place.

1/ Chronique de la nouvelle infrastructure de calcul et de stockage

En août dernier, l'accès au nouveau cluster a été ouvert. Il est composé de :

- deux nœuds de login (genotoul1&2) accessibles via la commande ssh username@genotoul.toulouse.inra.fr
- 68 nœuds de calcul, comportant chacun de 40 cœurs et 256 Go de RAM
- un nouvel espace /work de 158 To utiles (soit 240 disques SAS de 1,2 To chacun et 16 disques SSD de 400Go chacun). Toutes les métadonnées et les très petits fichiers sont stockés sur les disques SSD dans le but d'améliorer la performance
- une interconnexion Infiniband 40Gb/s.

Le 22 septembre : nous avons déménagé l'ancien cluster dans le datacenter de l'INRA de Toulouse.

L'architecture complète est donc composée de :

- l'ancien cluster AMD : 34 nœuds comportant chacun 48 cœurs et 384Go RAM
- le nouveau cluster INTEL : décrit ci-dessus
- BIGMEM : 2 nœuds comportant chacun 64 cœurs et 1To RAM
- SMP : 1 nœud de 240 cœurs et 3To RAM

Vous pouvez surveiller la charge des différents nœuds du cluster à partir de notre site web en sélectionnant « Resources » puis « Hardware » et « Monitoring genotoul ».

Depuis l'alias genotoul vous avez désormais accès à l'ensemble des nœuds via les queues suivantes :

- workq : adresse en priorité les nouvelles machines INTEL, puis les anciennes AMD (durée des jobs limitée à 48h)
- unlimitq : pas de limitation sur la durée des jobs
- hypermemq : BIGMEM (accessible sur demande exceptionnelle : http://bioinfo.genotoul.fr/index.php?id=82)
- smpq : SMP (accessible sur demande exceptionnelle : http://bioinfo.genotoul.fr/index.php?id=82)

Rappel des règles d'utilisation pour tous les utilisateurs (Galaxy compris) :

L'espace /work est un espace temporaire de travail. Les résultats d'analyse à conserver doivent être copiés sur le /save, les fichiers intermédiaires éliminés. Lorsque c'est nécessaire, la plate-forme efface les fichiers non accédés depuis plus de 120 jours.

Pour les académiques, les 100000 premières heures de calcul (par année civile) sont mises à disposition sans conditions particulières. Au-delà, nous demandons de remplir le formulaire de demande exceptionnelle sur notre site web : http://bioinfo.genotoul.fr/index.php?id=82. Si votre projet est un projet de bioinformatique de taille appropriée par rapport à notre infrastructure, nous réviserons ce quota. Pour le personnel des entreprises privés, au-delà de 500 h de calcul, nous facturerons les heures de calcul supplémentaires. Les tarifs sont disponibles sur demande.

L'espace /save est limité à 250Go par utilisateur (doublé avec la réplication). L'espace /work est limité à 1To par utilisateur. Si vous avez besoin de davantage d'espace, merci de nous contacter via le formulaire : http://bioinfo.genotoul.fr/index.php?id=82. L'espace /home, quant à lui, est réservé aux fichiers de configuration (ne doit contenir aucune donnée).

L'ancien /work appelé maintenant /work_old restera accessible sur genotoul en lecture seule jusqu'à la fin de l'année 2014.

Rappel des règles d'utilisation par groupe d'utilisateurs (Galaxy compris) :

Les utilisateurs du groupe « contributeurs » ont accès à 100 % des ressources de calcul, les utilisateurs situés dans la région ou employé par l'INRA ont accès à 75 % des ressources, tous les autres à 20 %.

2/ Les prochains cycles d'apprentissage

A/ Assemblage de données RNASeq *de novo* (01-03 décembre 2014) (reste 3 places).

La plate-forme bioinfo genotoul et l'équipe Sigenae propose du 01 décembre 2014 à 14 heure au 03 décembre 17 heure une formation sur 2,5 jours traitant de l'assemblage des données de transcriptomes *de novo* obtenues grâce aux nouvelles technologies de séquençage. Vous apprendrez comment vérifier la qualité des données, et pré-traiter les lectures en conséquence, comment fonctionne un assembleur et comment l'utiliser sur ce type de données. Enfin, vous apprendrez comment appréhender la qualité d'un assemblage dans le but de choisir le meilleur.
Savoir utiliser la ligne de commandes Linux/Unix est un pré-requis pour ce module.

B/ Phylogénie et évolution de séquences (08-10 décembre 2014)

Un nouveau parcours d'apprentissage sur le thème de l'évolution des séquences est proposé par le CATI Bios4Biol dont notre plate-forme fait partie. Il est constitué de 4 modules s'adressant potentiellement à des publics variés. Les inscriptions sont donc indépendantes (sauf les deux demi-journées qui sont couplées du point de vue de l'inscription).

Nom du module	Pré-requis	Date (durée)
1) Initiation à l'alignement de séquences et à la phylogénie	Aucune nécessité de connaître la ligne de commande Linux	08 décembre (1 journée)
2) Présentation et mise en œuvre de différentes méthodes de construction d'arbres phylogénétiques	Savoir générer et interpréter un alignement de séquences. Connaître la ligne de commande Linux.	09 décembre (1 journée)
3) Présentation et mise en œuvre de méthodes de phylogénomique	Savoir générer et interpréter un alignement de séquences, connaître les modèles évolutifs et les méthodes de construction d'arbres phylogénétiques. Connaître la ligne de commande Linux.	10 décembre matin (½ journée)
4) Présentation et mise en œuvre de méthodes de détection de pressions de sélection	Savoir générer et interpréter un alignement de séquences, connaître les modèles évolutifs et les méthodes de construction d'arbres phylogénétiques. Connaître la ligne de commande Linux.	10 décembre après-midi (½ journée)

C/ 2 jours de formation Linux/Cluster seront organisés les 23 et 24 mars 2015

Vous souhaitez utiliser notre infrastructure de calcul via la ligne de commande. Mais vous avez besoin d'aide pour démarrer : connaître les commandes de base, mieux comprendre notre infrastructure et découvrir comment s'en servir efficacement. Cette formation a été conçue dans ce but.
Lors de la première journée vous apprendrez comment vous connecter, comment copier des fichiers, les éditer et comment en sécuriser l'accès. Vous apprendrez aussi comment gérer votre espace disque, compresser et décompresser des fichiers.
La formation au cluster de calcul vous permettra de lancer vos premiers « jobs » sur le cluster. Nous aborderons aussi les réservations de ressources (CPU, mémoire…) et nous apprendrons comment suivre les différents « jobs » lancés. Lors des travaux pratiques vous découvrirez, entre autre, où sont situés les logiciels et les banques de données que nous mettons à votre disposition.
Les deux journées sont indépendantes mais savoir utiliser Linux/Unix est un pré-requis pour la formation au cluster du calcul.

Pour tous nos cycles d'apprentissage :
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?id=115.
Les inscriptions s'effectuent sur cette page : http://bioinfo.genotoul.fr/index.php?id=10.

La plupart des formations que nous dispensons sont aussi disponibles sur la plate-forme d'e-learning sig-learning à l'adresse suivante : http://sig-learning.toulouse.inra.fr.

3/ Nouvelle charte d'utilisation des infrastructures de la plateforme

Nous avons rédigé une nouvelle charte disponible à partir du premier item de notre FAQ sur le site web de la plateforme. Elle est téléchargeable directement à partir de ce lien : http://bioinfo.genotoul.fr/fileadmin/Documents/ChartPFBioinfoGenoToul.pdf.
Cette charte définit nos engagements et les conditions d'utilisation des serveurs, clusters de calcul et espace disques de la plateforme.
Lors de la demande d'un nouveau compte, nous demandons d'en accepter les termes.

4/ L'utilisation des logiciels sur le cluster de calcul

Les logiciels mis à disposition par la plate-forme sur notre infrastructure de calcul (genotoul.toulouse.inra.fr), et leur documentation associée, se trouvent dans le répertoire /usr/local/bioinfo/src/. Pour chaque logiciel, selon la demande, vous pourrez trouver une ou plusieurs versions. La version correspondant au lien « current » est celle dont les exécutables sont inclus dans l'environnement par défaut de l'utilisateur (/usr/local/bioinfo/bin). La mise à jour des liens « current » vers les dernières versions installées est effectuée périodiquement, après avertissement par mail à l'ensemble des utilisateurs. La liste des logiciels mis à disposition par la plate-forme est également consultable sur notre site web: http://bioinfo.genotoul.fr/index.php?id=5.

5/ Nouvelle version de Galaxy : quels changements ?

Profitant de l'arrêt de service lors de la migration de l'infrastructure GenoToul dans le datacenter de l'INRA de Toulouse, Galaxy a fait peau neuve. Depuis début octobre 2014, l'instance Sigenae de Galaxy (http://galaxy-workbench.toulouse.inra.fr ou http://sigenae-workbench.toulouse.inra.fr) a été mise à jour afin que vos jobs puissent s'exécuter à votre nom, avec plus d'équité, de sécurité, de traçabilité et de rapidité. Plusieurs banques, outils (R, SNPs, RNAseq) ont été ajoutés. Vos « datasets » peuvent être transférés de l'ancienne version à la nouvelle facilement, après tri de votre part. L'infrastructure web est maintenant plus robuste et récente. Pour toute assistance, n'hésitez pas à nous contacter : sigenae-support@listes.inra.fr

Chaque utilisateur Galaxy, est donc maintenant soumis aux mêmes quotas et conditions d'utilisations que l'ensemble des utilisateurs de l'infrastructure.

6/ Jvenn : un utilitaire de construction de diagramme de Venn

Les diagrammes de Venn sont couramment utilisés pour comparer des listes. En biologie, par exemple, ils sont très largement utilisés afin de comparer des listes de gènes différentiellement exprimés. Ils permettent la comparaison entre différentes conditions expérimentales ou entre différentes méthodes. Lorsque le nombre de listes d'entrée est supérieure à quatre, le diagramme devient rapidement illisible. Cependant, un affichage alternatif et dynamique peuvent améliorer son utilisation et sa lisibilité. Dans ce but, nous avons développé, en collaboration avec l'équipe Sigenae, une librairie JavaScript (jvenn) capable de gérer jusqu'à six listes et présentant les résultats sous la forme d'un diagramme de Venn classique ou sous la forme d'un diagramme d'Edwards. Jvenn est livré avec une documentation complète et un exemple disponible à l'adresse suivante : http://bioinfo.genotoul.fr/jvenn.

Voici la référence du papier : Philippe Bardou, **Jérôme Mariette, Frédéric Escudié, Christophe Djemiel and Christophe Klopp**. (2014) jvenn: an interactive Venn diagram viewer. BMC bioinformatics, 15:293 doi:10.1186/1471-2105-15-293

7/ Publications 2014

Vous trouverez sur notre site web (http://bioinfo.genotoul.fr/index.php?id=54), la liste des publications ayant utilisé l'infrastructure bioinfo Genotoul. Si vous avez publié en utilisant nos ressources et que votre article n'y apparaît pas, merci de nous en communiquer les références par mail à anim.bioinfo@toulouse.inra.fr.

Pour toute demande d'information ou de travaux, veuillez envoyer un mail à support.genopole@toulouse.inra.fr en précisant vos noms et coordonnées.