Lettre d'information N°17

du 23 décembre 2013 – 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/ Les prochains cycles d'apprentissage :

A - Une semaine de formation aux traitements de données issues des séquenceurs haut débit en environnement Unix

Cette formation est organisée du 7 au 10 avril 2014 en collaboration avec l'équipe Sigenae (http://www.sigenae.org/).

Titre du module	Date
Premiers contacts avec la plateforme bioinformatique de la génopole, apprendre à utiliser un environnement Linux.	07/04/2014
Utilisation du cluster de calcul de la plateforme bioinformatique de la génopole.	08/04/2014
Alignement de séquences issues des NGS et à la recherche de polymorphismes.	09/04/2014
Analyse d'expression des gènes codant pour des protéines en utilisant des séquences de type RNA-Seq.	10/04/2014

Savoir utiliser unix / Linux est un pré-requis pour être en capacité de suivre les autres modules. Les modules des jours 2, 3, 4 sont indépendants les uns des autres.

B - Trois jours de formation aux traitements de données issues des séquenceurs haut débit sous environnement Galaxy

Cette formation est organisée du 12 au 14 mai 2014 en collaboration avec l'équipe Sigenae. Toute inscription à cette formation se fera obligatoirement pour le cycle complet de 3 jours. Cette formation permettra de vous initier à l'environnement Galaxy, à l'utilisation d'outils permettant l'alignement de séquences et la recherche de polymorphismes, ainsi qu'à l'analyse de données RNA-Seq et sRNAseq dans l'environnement Galaxy. Ce cycle ne nécessite aucune connaissance préalable en ligne de commande. Vous pouvez vous connecter à l'environnement Galaxy que nous proposons à l'url : http://galaxy-workbench.toulouse.inra.fr (avec le login et le mot de passe LDAP Genotoul).

C - Un nouveau cycle d'apprentissage est proposé : L'analyse métagénomique des données 16S sous Galaxy.

En collaboration avec l'équipe Sigenae ((http://www.sigenae.org/), nous organisons le 5 février 2014 une journée de formation à l'analyse métagénomique de données d'ADN 16S (produite par un séquenceur 454 ou Illumina Solexa).
Après une petite introduction à l'instance Galaxy de Toulouse, vous apprendrez comment démultiplexer les reads, les nettoyer, les aligner contre une banque de données d'ARN 16S de référence, faire l'assignement taxonomique et construire les OTUs (Operationnal Taxonomic Unit) et enfin faire les analyses de diversité.
Savoir utiliser un environnement Galaxy est un pré-requis pour ce module.

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 sont aussi disponibles sur la plateforme d'e-learning sig-learning à l'adresse suivante : http://sig-learning.toulouse.inra.fr.

2/ Sur genotoul : utiliser qsub, qarray, qrsh ou qlogin pour les gros traitements :

Pour lancer vos gros traitements (plus de 2 heures), merci d'utiliser une de ces 4 commandes : "qsub", "qarray", "qrsh" ou "qlogin".

Le qsub permet de lancer des jobs en batch sur le cluster, le qarray des jobs array. Les jobs array permettent de lancer un ensemble de tâches en parallèle (et non pas en séquence). Le job id sera le même pour toutes les tâches, il lui sera rajouté un numéro de tâche.

Le qrsh et le qlogin permettent de soumettre des jobs en interactif sur le cluster de calcul. Les différences :

- le qlogin permet un déport de l'interface graphique alors que le qrsh non.
- le qlogin n'adresse qu'une seule machine (ceri002) alors que le qrsh adresse l'ensemble du cluster.

Donc, si vous souhaitez soumettre des jobs en interactif qui ne nécessitent pas l'interface graphique, le mieux est d'utiliser qrsh !

3/ /home /save /work : quelles différences ?

La partition /home est réservée aux fichiers de configuration (quota utilisateur 100Mo).
La partition /save doit être utilisée pour les données sauvegardées (quota utilisateur 200Go), la durée de rétention des sauvegardes est de 1 mois.
La partition /work est réservée aux fichiers temporaires de calcul (quota utilisateur 1To).

Attention : la partition /work n'est pas sauvegardée. Elle est automatiquement purgée selon les modalités suivantes :

– le premier du mois, un mail d'information sera envoyé avec la liste des fichiers non accédés depuis plus de 120 jours.

– sans action de votre part, le 15 du mois les fichiers de la précédente liste seront effacés.

Pour connaître les fichiers qui vont être purgés dans votre /work directory :
#find /work/username -type f -atime +120

Il est de votre responsabilité de gérer vos données (organisation, volumétrie, pertinence, ancienneté).

Pour connaître votre consommation d'espace disque, utilisez la commande suivante :
du -csh */DIR_NAME/USER_NAME/**

Si vous avez besoin de davantage de ressources, merci d'utiliser le formulaire de demandes exceptionnelles à l'adresse suivante : http://bioinfo.genotoul.fr/index.php?id=82

4/ Des quotas de slots de calcul ont été mis en place sur le cluster :

Différents quotas sont appliqués suivant votre appartenance aux 3 différents groupes que nous avons définis :

- les contributeurs de la plate-forme
- les utilisateurs qui proviennent de l'INRA et/ou de la région,
- les autres

Le pourcentage maximum de ressources accessible par chacun des groupes varie de la manière suivante :

- contributeurs = 100 % des ressources,
- INRA et/ou REGION = 80 % des ressources,
- autres = 23 % des ressources

De plus il existe une limite maximum de slots par utilisateurs de 256 slots sur le workq et de 32 sur la unlimitq.

Par exemple, si vous êtes de l'INRA, vous êtes comme tout utilisateur limité à 256 slots sur la workq et à 32 slots sur la unlimitq (utilisés simultanément). Mais vous partagez aussi avec tous les autres membres du groupe INRA et/ou région les 80 % des ressources autorisées à ce groupe.

5/ Fin de l'infrastructure SNP :

A ce jour, tous les utilisateurs de l'ancienne infrastructure "snp" ont désormais migré sur l'environnement "genotoul". Les ressources de calcul "snp" seront recyclées pour d'autres usages (test de cloud par exemple).
Les seules exceptions sont les machines "bigmem" et "hypermem" qui seront intégrées dans le cluster genotoul lors du déménagement dans le datacenter.

6/ Les évolutions de l'infrastructure en 2014 :

Les serveurs, cluster de calcul et baies de stockages GENOTOUL seront déménagés dans le nouveau datacenter de l'INRA de Toulouse au cours du premier trimestre 2014.
Il y aura donc des coupures de services et une dizaine de jours d'indisponibilité à prévoir.

Par la suite (vers le mois de juin 2014), la plate-forme fera évoluer ses moyens de calcul et de stockage de manière significative. Veuillez nous faire part de vos besoins éventuels afin qu'ils soient pris en compte dans les acquisitions (nœuds à grosse quantité de mémoire, nœuds GPU, nœuds de visualisation graphique, espaces disques nécessaires...). Pour cela vous pouvez utiliser l'adresse mail : support.genopole@toulouse.inra.fr.

7/ Où trouver les banques de données ?

La liste des banques de données disponibles sur la plateforme ainsi que la date de leur mise à jour ou le numéro de version lorsque celui-ci est disponible sont ici : http://genoweb.toulouse.inra.fr/BmajWatcher/

Dans le tableau ci-dessous vous trouverez les répertoires dans lesquels sont accessibles les banques indexées pour différents outils.

Outil	Répertoire dans lequel se situe les index
NCBI blast	/bank/blastdb
bwa	/bank/bwadb
bowtie	/bank/bowtiedb
bowtie2	/bank/bowtie2db

Lorsqu'on utilise un index, il ne faut pas ajouter l'extension du fichier.

Les index STAR seront très bientôt disponibles pour les génomes Ensembl déjà disponibles sur la plate-forme.

8/ L'école de bioinformatique de Roscoff :

La seconde édition de l'école de bioinformatique organisée par l'ITMO Génétique, génomique et bioinformatique d'AVIESAN s'est déroulée à la station biologique de Roscoff la semaine du 17 au 23 novembre 2013. Cette école est organisée sous forme d'ateliers thématiques exclusivement réalisés sous Galaxy, via des serveurs et des instances hébergées par Roscoff et Sigenae Toulouse. Après une présentation de Galaxy et des principales notions bioinformatiques (technologies de séquençage, types de librairies, formats de fichiers, logiciels, et quelques notions de statistiques), les participants se regroupent en ateliers en fonction de leurs besoins et de leurs données : RNAseq, ChIPseq, miRNA, SNP et/ou CNV. La semaine est ponctuée par une journée d'analyses spécifiques des données personnelles des participants, encadrés par des tuteurs spécialisés.

L'équipe Sigenae et la plate-forme BioInfo Genotoul s'est mobilisée pour participer, en terme d'intervention, d'encadrement, et de co-organisation, à cette l'école.

La plate-forme vous souhaite à tous de très bonnes fêtes de fin d'années !

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