Lettre d'information N°18

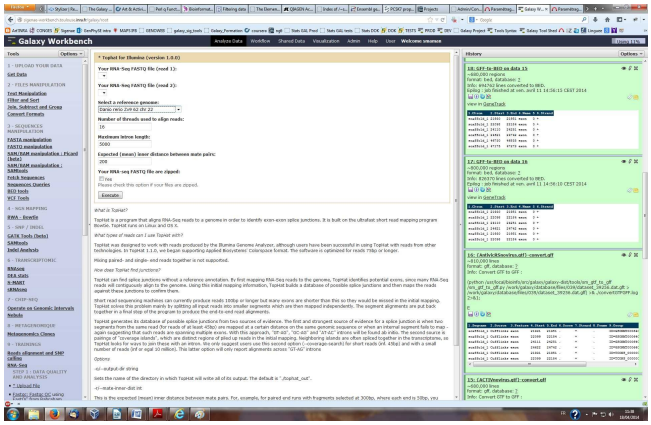
du 19 mai 2014 – Rédacteur : Claire Hoede

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/ Qu'est-ce que Galaxy et comment se connecter à l'instance Toulousaine ?

Galaxy est une interface web vous permettant de lancer des traitements (bio)informatiques sans connaître linux ou la ligne de commande. Cet outil, accessible depuis Internet (sans aucune installation préalable), est connecté à l'infrastructure de calcul locale Genotoul. Sigenae (et Genotoul Bioinfo) mettent ce service à votre disposition depuis l'URL http://sigenae-workbench.toulouse.inra.fr (protégé par un compte Genotoul http://bioinfo.genotoul.fr/index.php?id=81). Des sessions de formation à l'analyse des données sous Galaxy sont proposées à Toulouse par les plate-formes BioInfo Genotoul et Sigenae. D'autre part, un environnement d'auto-formation en ligne est accessible depuis la plate-forme « sig-learning » : http://sig-learning.toulouse.inra.fr (login et mot de passe Genotoul)



Pour toute demande (renseignement, installation, ajout d'outil, bug, etc.), veuillez envoyer un mail à sigenae-support@listes.inra.fr.

2/ Les prochains cycles d'apprentissage

A/ L'analyse métagénomique des données 16S sous Galaxy (26 mai 2014).

En collaboration avec l'équipe Sigenae (http://www.sigenae.org/), nous organisons le 26 mai 2014 un cycle d'apprentissage 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. Par contre il n'y a aucune nécessité de connaître la ligne de commandes Linux/Unix.

B/ Assemblage de données RNASeq de nov (22-24 septembre 2014).

La plate-forme bioinfo genotoul et l'équipe Sigenae propose du 22 septembre 2014 à 14 heure au 24 septembre 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.

C/ Phylogénie et évolution de séquences (25/26 septembre et 29 septembre 2014).

Un nouveau parcours d'apprentissage sur le thème de l'évolution de 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 différents. Les inscriptions sont donc indépendantes (sauf les deux demi-journées qui sont couplées du point de vue de l'inscription).

Nom du module	Public visé et pré-requis	Date (durée)
1) Initiation à l'alignement de séquences et à la phylogénie	Biologistes et bio-informaticiens souhaitant s'initier à l'analyse phylogénétique	25 septembre (1 journée)
2) Présentation et mise en œuvre de différentes méthodes de construction d'arbres phylogénétiques	Biologistes avertis en ligne de commandes et bio-informaticiens souhaitant découvrir l'analyse phylogénétique. Pré-requis : Savoir générer et interpréter un alignement de séquences.	26 septembre (1 journée)
3) Présentation et mise en œuvre de méthodes de phylogénomique	Biologistes avertis en ligne de commandes et bio-informaticiens souhaitant découvrir l'analyse phylogénétique. Pré-requis : Savoir générer et interpréter un alignement de séquences, connaître les modèles évolutifs et les méthodes de construction d'arbres phylogénétiques.	29 septembre matin (½ journée)
4) Présentation et mise en œuvre de méthodes de détection de pressions de sélection	Biologistes avertis en ligne de commandes et bio-informaticiens souhaitant découvrir l'analyse phylogénétique. Pré-requis : Savoir générer et interpréter un alignement de séquences, connaître les modèles évolutifs et les méthodes de construction d'arbres phylogénétiques.	29 septembre après-midi (½ journée)

D/ Quatre jours de formation aux traitements de données issues des séquenceurs haut débit sous environnement Galaxy

Cette formation est organisée du 2 au 6 novembre 2014 en collaboration avec l'équipe Sigenae. Toute inscription à cette formation se fera obligatoirement pour le cycle complet de 4 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 dans l'environnement Galaxy. Ce cycle ne nécessite aucune connaissance préalable en ligne de commande.

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/ Sur genotoul : utiliser qsub, qarray, qrsh ou qlogin

Nous vous rappelons que le serveur genotoul est uniquement réservé à la connexion, au développement et au transferts de fichiers. **En aucun cas, il ne doit être utilisé pour traiter des données au risque de le surcharger.**
Il est nécessaire d'utiliser le cluster de calcul pour traiter des données :

- **qrsh** si vous voulez soumettre un job en interactif (adresse l'ensemble des nœuds du cluster).
- **qlogin** si vous souhaitez soumettre un job en interactif avec le déport de l'interface graphique (adresse les deux premiers nœuds du cluster).
- **qsub/qarray** si vous souhaitez soumettre un job en batch (adresse l'ensemble des nœuds du cluster).

4/ Désactivation de l'option -M dans les scripts SGE

Nous avons désactivé la fonction SGE qui permet de changer le mail dans le script en raison des erreurs de frappe qui génèrent un grand nombre de retours de mails infructueux (et qui saturent le serveur de mail du centre INRA). Désormais c'est l'@ email de l'utilisateur qui compte (enregistré avec la création de compte). En clair, l'option $ -m bea est toujours valable mais l'option $ -M my_email@toulouse.inra.fr est inhibée.

5/ Déménagement des serveurs de virtualisation prévu prochainement

Nous prévoyons de déménager dans le Datacenter INRA l'infrastructure de serveurs de virtualisation courant septembre. Cela occasionnera des coupures de vos machines virtuelles hébergées sur la plate-forme.

6/ Acquisition d'une nouvelle machine SMP (Shared Memory Processor)

Nous venons d'acquérir et de mettre en place une nouvelle machine de calcul (genosmp) dans le datacenter. Cette machine dispose de 120 cœurs et de 3To de RAM. Elle est accessible via le gestionnaire OGE (Open grid Engine) en utilisant la file « smpq » (qrsh ou qlogin).
Son utilisation est soumise à une demande exceptionnelle : http://bioinfo.genotoul.fr/index.php?id=82.
Étant donné son éloignement avec le cluster de calcul actuel, elle ne peut pas bénéficier du réseau de disques à haut débit (/work) en place sur genotoul. C'est pourquoi elle dispose de son propre espace temporaire de calcul /scratch de 22To en attendant la mise en commun des ressources de calcul dans le datacenter.

7/ Comment obtenir du support informatique ?

La FAQ : http://bioinfo.genotoul.fr/index.php?id=11
Les formations : http://bioinfo.genotoul.fr/index.php?id=10.

- Nous mettons à votre dispositions tous nos supports de formations (en cliquant sur la description de chacune des formations à partir de l'URL précédente).
- De plus, le e-learning est disponible à cette adresse : http://sig-learning.toulouse.inra.fr.

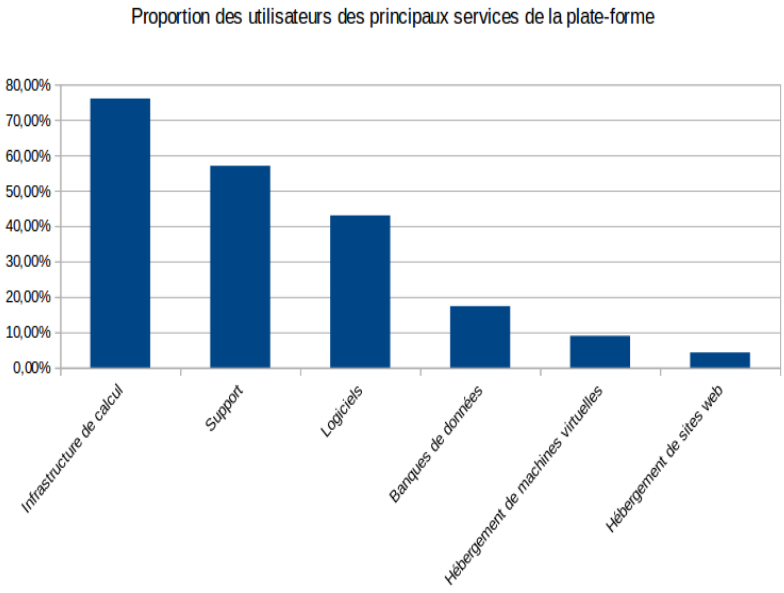
Le mail : support.genopole@toulouse.inra.fr
Les demandes exceptionnelles : http://bioinfo.genotoul.fr/index.php?id=82

8/ Bilan de l'enquête de satisfaction 2013

Nous tenons à remercier tous les utilisateurs qui ont répondu à notre questionnaire de satisfaction annuelle.
Il portait sur l'ensemble des activités liées à l'infrastructure :

– le calcul
– les logiciels
– les banques de données
– l'hébergement de machines virtuelles
– l'hébergement de sites web
– et transversalement à tous les points précédents : le support

La satisfaction liée à nos activités d'accompagnement de projets et de formation est mesurée par ailleurs.
Parmi les 46 utilisateurs qui ont répondu à l'enquête, le graphique ci-dessous présente la proportion d'utilisateurs des différents services que nous proposons.



Nous vous rappelons que vous pouvez demander l'installation ou la mise à jour d'un logiciel ou d'une banque de données via notre adresse mail de support : support.genopole@toulouse.inra.fr.

9/ Politique tarifaire de la plate-forme

A/ Stockage / sauvegarde des données.

Rappel : l'ouverture d'un compte utilisateur sur la plate-forme GenoToul Bioinfo donne droit à la mise à disposition gratuite d'un espace de stockage sauvegardé de 250 Go.Au-delà de ce volume de 250 Go, l'espace de stockage/sauvegarde est facturé. L'unité de facturation est le To non sauvegardé (tarifs disponibles sur demande). A partir du 1er mai 2014, le montant minimum facturé passe à 450€, ce qui correspond, pour les académiques en région et personnels Inra (national), à soit 3 ans de stockage d'un volume de 1 To, soit à 3 To de stockage sur un an. La sauvegarde s'effectuant par réplication sur un site distant, le tarif demandé pour un volume de stockage sauvegardé est doublé.

B/ Calcul.

Pour les académiques en région et hors région
Pour assurer à l'ensemble de ses utilisateurs un accès de meilleure qualité à l'infrastructure bio-informatique mise à disposition de la communauté du Vivant, la plate-forme GenoToul Bioinfo met en place une nouvelle règle de fonctionnement. Un quota de temps de calcul de 100.000h par an (année civile) sera attribué par utilisateur. Au-delà de ces 100.000h, le dépôt d'un projet scientifique sera nécessaire pour évaluer les besoins réels de l'environnement bio-informatique mis à disposition. En fonction des résultats de cette évaluation mais aussi de l'origine géographique et institutionnelle, les utilisateurs pourront alors :

i. soit poursuivre leurs traitements ;
ii. soit être sollicités pour contribuer financièrement au fonctionnement de l'infrastructure (tarifs disponibles sur demande) ;
iii. soit être ré-orientés vers des mésocentres de calcul régionaux ou nationaux.

Pour les entreprises privées
Un quota de 500 h de calcul par an (année civile) sera attribué afin de tester l'adéquation de l'infrastructure bio-informatique aux besoins de l'entreprise. Au-delà de ce quota, l'heure de calcul sera facturée au tarif en vigueur (disponible sur demande). Le nombre d'heures cumulées attribuées aux entreprises privées restera limité à 5 % du nombre total d'heures disponibles.

Pour l'année 2014, exceptionnellement, la période de validité de cette règle et des tarifs en vigueur va du 1er juin 2014 au 31 décembre 2014. A partir du 1er janvier 2015, ce quota sera calculé par année civile complète.

Nous avons mis en place un script vous permettant de connaître le nombre d'heure de calcul utilisé, et le nombre d'heure de calcul restant avant de dépasser votre quota, il s'utilise de la façon suivante :

qquota_cpu votreNomDeUserGenotoul

Vous obtiendrez les temps de calcul en seconde et en heure, comme sur l'exemple suivant :

```
Quota du user "votreNomDeUserGenotoul"
######################

Temps de calcul autorisé : 360000000 secondes <-> 100000 heures
Temps de calcul utilisé: 174878 secondes <-> 48 heures
Temps de calcul restant: 359825122 secondes <-> 99951 heures
```

10/ RNAbrowse : un environnement de visualisation des résultats RNAseq *de novo*

L'analyse du transcriptome basée sur un assemblage *de novo* de données RNASeq est maintenant réalisée fréquemment dans de nombreux laboratoires. Les résultats obtenus tels que les séquences des transcrits, la quantification, l'annotation fonctionnelle des contigs et les sorties de la détection de variants nécessite un environnement de visualisation pour rendre exploitable ces résultats par des biologistes.
Nous avons développé, en collaboration avec l'équipe SIGENAE, et publié récemment un environnement de stockage et de visualisation de ces résultats qui fournit des graphiques présentant les données de manière globale et permet également d'explorer l'ensemble des résultats détaillés. RNAbrowse est basé sur biomart et propose une procédure d'installation simplifiée. Il permet un chargement aisé des données locales. Il est disponible à l'adresse suivante : http://bioinfo.genotoul.fr/RNAbrowse.
Un site de démonstration a été mis en place : http://ngspipelines.toulouse.inra.fr:9012/.
Voici les références du papier : **Mariette J**, **Noirot C**, Nabihoudine I, Bardou P, **Hoede C**, Djari A, Cabau C, **Klopp C**. (2014) RNAbrowse: RNA-Seq De Novo Assembly Results Browser. PLoS ONE 9(5): e96821. doi: 10.1371/journal.pone.0096821.

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