

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/ Le séminaire du 15 mars

La plate-forme vous invite à la présentation des nouveautés liées au nouveau cluster le jeudi 15 mars de 10h à 12h au plus tard en salle de conférences Marc Ridet (INRA, Auzeville). Le nouveau cluster sous SLURM apporte de nouvelles performances, de nouvelles fonctionnalités mais il faudra se former aux nouveaux usages (la façon de soumettre un job ou d'utiliser un logiciel changeant par exemple). Un lien skype sera fourni une heure avant pour ceux qui seraient dans l'incapacité de se déplacer.

### 2/ 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 jours de formation à l'alignement et à la détection de variants à partir de NGS sous Galaxy du 14 au 15 mai 2018 (reste 8 places)**

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

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 formations vous permettant de comprendre les principaux formats de fichiers manipulés, d'apprendre à aligner les lectures sur un génome de référence et de détecter SNP et petits indels via la suite GATK.

**Aucun pré-requis n'est nécessaire.**

#### **B) 4 jours de formation à la métagénomique à partir d'amplicons sous Galaxy seront organisés du 25 au 28 juin 2018 (reste 2 places)**

Vous avez, ou allez obtenir, des données de métagénomique 16S ou 18S, Miseq ou 454? Vous souhaitez les analyser efficacement sans utiliser la ligne de commande? Les équipes Sigenae, NED (GenPhySE), TWB et Bioinfo Genotoul vous proposent une formation de 4 jours indivisibles sous environnement Galaxy. La première demi-journée sera consacrée, après une initiation à l'environnement Galaxy à la prise en main de l'instance Galaxy Toulousaine. Pendant deux jours vous apprendrez à utiliser le pipeline FROGS dont la description est disponible à l'url : <http://frogs.toulouse.inra.fr/>. Il est constitué des étapes principales suivantes: nettoyage, clustering, affiliation taxonomique, filtre et statistiques descriptives que nous vous détaillerons. Enfin la dernière journée et demie vous permettra d'interpréter les résultats grâce aux outils statistiques proposés.

**Aucun pré-requis n'est nécessaire.**

#### **C) 1 journée de formation à l'assemblage de novo de données RNASeq : le 5 juin 2018 (reste 9 places)**

L'équipe Sigenae vous propose une journée de formation à l'assemblage de transcriptome sans génome de référence à partir de données de séquençage de type Illumina Hiseq. Vous utiliserez le pipeline développé par Sigenae appelé DRAP. Il vous permet d'assembler les données RNASeq de vos différents échantillons, de faire le méta-assemblage et d'évaluer la qualité des différents assemblages produits.

**Pré-requis nécessaire** : Connaissance de la ligne de commande Unix/Linux 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/>.

### 3/ Quelques consignes de bonnes pratiques

#### **A) La partition /work est réservée pour les fichiers temporaires de calcul**

Le quota utilisateur de cet espace est de 1To, nous vous rappelons que ce volume n'est pas répliqué (contrairement à la partition /save) et peut être purgé si besoin (fichiers inutilisés depuis plus de 120 jours). Il est de la responsabilité de chacun de vous de gérer vos données en terme d'organisation, de volumétrie, pertinence et d'ancienneté. Les résultats finaux d'analyse doivent être déplacés dans le /save pour en assurer la réplication.

#### **B) Les traitements de données sur les serveurs frontaux sont interdits**

Pour que l'ensemble des utilisateurs de genotoul puissent travailler dans de bonnes conditions le traitement des données sur le serveur frontal est interdit. Les serveurs genotoul sont réservés exclusivement à la connexion, au transfert de données, à la compilation, au test très rapide de la ligne de commande et à la soumission de jobs sur le cluster de calcul. Tout traitement de données lancé directement (sans «qsub», «qarray», «qrsh» ou «qlogin» pour SGE) sera systématiquement interrompu sans préavis par les administrateurs système. Ces quatre commandes sont explicitées dans notre FAQ, rubrique «Job Submission» ([http://bioinfo.genotoul.fr/index.php/faq/job\\_submission\\_faq/](http://bioinfo.genotoul.fr/index.php/faq/job_submission_faq/)) à la question: «Which commands can I use to submit my job»? De plus, lorsque vous êtes en qlogin et en qrsh, si vous avez fini vos traitements, nous vous conseillons de fermer votre session car cela occupe un nœud et compte dans votre quota CPU.

## 4/ Nouveautés et rappels concernant les logiciels

### A) Installation des logiciels et mise à jour (<http://bioinfo.genotoul.fr/index.php/ask-for/install-soft/>)

A partir du 16 mars prochain, lorsque vous demanderez l'installation d'un logiciel ou sa mise à jour, celle-ci sera effectuée sur le nouveau cluster. L'installation ou la mise à jour du logiciel sur l'ancien cluster sera encore possible si vous en justifiez le besoin.

### B) Liste et utilisation des logiciels

Sur genotoul, près de 900 logiciels (sans compter les différentes version du même logiciel) sont installés. La liste de ces logiciels a été mise à jour sur notre site web (<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`. Dans le répertoire d'installation du logiciel, s'il y a un fichier « How\_to\_use », il est très important de le lire car il fournit des informations indispensables à son utilisation.

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.

## 5/ D-GENIES : un nouveau logiciel de comparaison de génomes

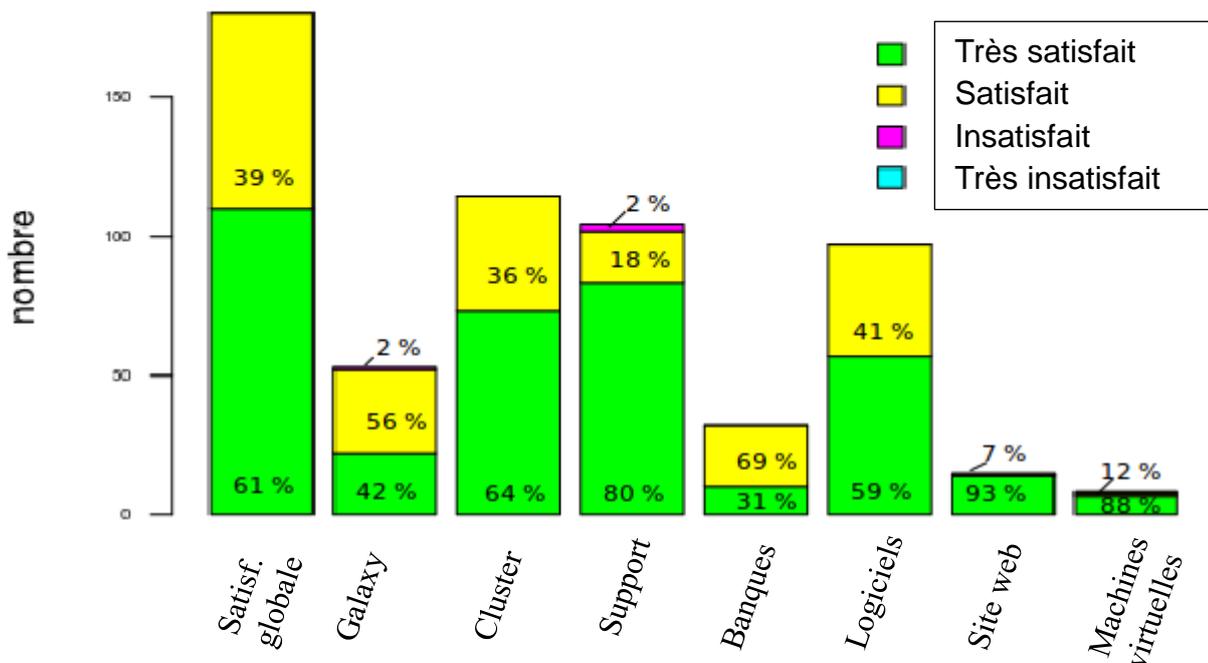
La plate-forme Genotoul Bioinfo met à disposition de ses utilisateurs un nouveau logiciel de comparaison de génomes : D-GENIES (Dot plot large GENomes in an Interactive, Efficient and Simple way).

Les dot plots sont largement utilisés pour comparer deux génomes car ils permettent de mettre en évidence de manière synthétique les répétitions, les cassures et les inversions. Les outils actuels générant des dots plots à partir d'alignement sont souvent limités par la taille des séquences en entrée. D-genies est une application (standalone ou web) qui réalise l'alignement de grands génomes en utilisant minimap2 et génère un dot plot interactif. Il permet aux utilisateurs de trier les séquences par rapport à la référence, zoomer sur le graphique et de télécharger les images, les alignements ou les fichiers de séquences. D-genies est un logiciel facile à installer, sous licence GPL, développé en Python et en JavaScript. Le code source est téléchargeable ici : <https://github.com/genotoul-bioinfo/dgenies> et une version de démonstration est disponible : <http://dgenies.toulouse.inra.fr/>. L'article soumis peut être consulté à l'adresse suivante : <https://peerj.com/preprints/26567/>.

## 6/ Bilan du questionnaire de satisfaction 2017

180 personnes ont répondu à notre dernière enquête de satisfaction annuelle, soit 26 % de réponses par rapport au nombre d'utilisateurs actifs en 2017 (15 % par rapport au nombre total d'utilisateurs déclarés). Ce qui est un très bon taux de retour et nous vous en remercions. La synthèse des réponses relatives à la satisfaction de chacun des types de services proposés par la plateforme est représenté sur le graphique ci-dessous.

### réponses des utilisateurs par service en 2017



Globalement vous semblez apprécier nos modes de communications car 92% des répondants déclarent trouver l'information qu'ils recherchent sur notre site web et 74% lisent régulièrement la newsletter. Parmi les nombreux commentaires utiles que vous avez laissés lors de cette enquête : certains d'entre vous nous ont fait part de leur besoin de davantage de tutoriels : à ce sujet, nous nous permettons de vous rappeler les cycles d'apprentissage que nous proposons dont les supports sont disponibles à la fin de la page de description de chacun d'entre eux sur notre site web. Dans la plupart des cas les travaux pratiques sont également disponibles. De plus la FAQ sur notre site web (<http://bioinfo.genotoul.fr/index.php/faq/>) recense de nombreuses informations comme les différentes queues (Job submission / Which are the available queues?), ou encore les ressources

disponibles sur le cluster (default resources / Cluster available resources). Pour connaître la charge actuelle d'un nœud ou d'un serveur, le monitoring est disponible via notre site web : <http://bioinfo.genotoul.fr/index.php/resources-2/monitoring/>.

Certains d'entre vous ont remarqué également que les temps d'attente sur le cluster SGE ont augmentés, nous nous excusons de ces désagréments. En effet, depuis le début de la transition vers le nouveau cluster, nous avons enlevé les nœuds obsolètes. Le détail des nombres de slots restants pour chacun des groupes d'utilisateurs était présenté dans le tableau I de la newsletter numéro 27 (<http://bioinfo.genotoul.fr/index.php/news/newsletter-27/>). Le nouveau cluster sera plus performant et devrait donc permettre de diminuer le temps moyen d'attente. Pour apprendre à vous en servir, nous vous conseillons de venir au séminaire dédié le 15 mars prochain.

Vous nous avez aussi posé des questions sur les banques de données : pour avoir les informations les concernant nous mettons à disposition l'outil Biomaj Watcher (<http://genoweb.toulouse.inra.fr/BmajWatcher/> : accessible depuis la page databanks de notre site web avec votre login et mot de passe genotoul). Grâce à cette interface graphique vous avez accès à l'URL de téléchargement, la date à laquelle la version a été téléchargée et le chemin d'accès sur genotoul en double-cliquant sur la banque qui vous intéresse.

**Plus spécifiquement au sujet de l'instance Galaxy** : parmi les 260 utilisateurs de l'instance Galaxy Sigeneae en 2017, 53 ont répondu à l'enquête. Parmi eux 98% sont très satisfaits ou satisfaits de l'instance Galaxy Sigeneae / BioInfo Genotoul. Une majorité (58%) a suivi une session de formation de la plateforme BioInfo Genotoul, et 40% n'ont pas besoin de plus de support sur Galaxy. Pour 80% des utilisateurs, Galaxy leur permet de traiter leurs données bioinformatiques et pour 72% d'entre eux, les résultats obtenus ont été utiles dans leur projet. 9% des répondants déclarent avoir publiés ces résultats. A ce sujet nous vous rappelons que le nombre de publications remerciant Sigeneae et la plateforme bioinfo genotoul est un indicateur de l'utilité des services que nous proposons. Nous nous en servons lors des demandes de financement par exemple. Si vous avez utilisé l'instance Galaxy Sigeneae pour votre analyse, une phrase type de remerciements est disponible sur la page d'accueil : <http://sigeneae-workbench.toulouse.inra.fr/galaxy/>.

Les principaux types d'outils utilisés sont ceux servant à Frogs (51%), l'alignement de séquences (43%), la manipulation de fichiers (37%), la manipulation de séquences (32% des répondants), et le traitement de données RNASeq (24%).

Les actualités spécifiques à Galaxy (Galaxy News) sont archivées et accessibles via la page d'accueil de l'instance Galaxy Sigeneae.

## [7/ Cycles d'apprentissage proposés par la plateforme biostat](#)

Vous êtes plusieurs à nous avoir fait part de vos besoins en formations sur le logiciel R pour la biostatistique via le dernier questionnaire de satisfaction annuelle. La plateforme biostat du GIS genotoul propose régulièrement des formations aux biostatistiques ou à l'utilisation de R en statistique et biostatistiques. En juin prochain sera organisé pendant deux jours une formation « Initiation à la programmation en R » et en septembre une formation « statistique niveau 1 » de deux jours également. Pour plus d'information, n'hésitez pas à consulter leur page d'actualité : <https://perso.math.univ-toulouse.fr/biostat/>.