

Lettre d'information

Cette lettre d'information est destinée aux membres des équipes utilisant la plate-forme Genotoul-Bioinfo. Elle a pour but de répondre aux questions et commentaires que vous nous avez fait remonter via le questionnaire de satisfaction annuel et de vous parler des évolutions de l'équipe, les nouveaux outils, services, conditions d'utilisation, projets et formations mis en place.



Publications

Lorsque vous utilisez l'infrastructure pour vos travaux de recherche, nous vous demandons, s'il vous plaît, de nous citer dans les remerciements des publications. La phrase type est sur notre FAQ dans la rubrique « [cite us](#) ». Il s'agit d'un des principaux indicateurs servant à justifier notre utilité auprès de nos tutelles. Merci par avance.



Retour sur les incidents

Vous avez pu constater le 19 septembre 2024, de 10h à 16h, que l'accès à l'espace `/work` était impossible. Nous avons subi une panne sur un port du switch Infiniband auquel étaient connectés les serveurs d'entrées/sorties du `/work`. Nous avons profité de la maintenance de janvier 2025 pour recâbler ces serveurs et assurer une meilleure tolérance de panne.

Du 1^{er} au 3 février dernier, une panne sur un switch REMIP (Réseau Rénater Régional), qui fait le lien entre le cœur de réseau INRAE et le DROCC (Datacenter Régional Occitanie), a coupé le cluster du reste du monde. Cet équipement est actuellement branché en simple attachement à notre réseau INRAE. Nous prévoyons de le basculer en double attachement au printemps 2025, afin, encore une fois, d'améliorer la disponibilité du cluster lors de pannes.

Évolutions de l'infrastructure

Nous vous remercions chaleureusement d'avoir rempli le questionnaire de satisfaction annuel. Nous mettons chaque fois en place des ajustements pour répondre à vos suggestions lorsque cela est possible. Cette lettre est dédiée aux réponses à vos remarques et questions.

Évolution du serveur GPU

Vous avez été nombreux à nous demander davantage de GPUs. Nous sommes en train d'acheter 2 nouveaux serveurs de calcul supplémentaires (avec chacun 64 cœurs 3.85GHz, 1.5TB de RAM, 20TB de disque local ultra rapide, et 4 cartes L40S). Nous avons, de plus, mis en place un script permettant de savoir si certaines partitions ou cartes sont libres : `sq_gpu`. Pour utiliser les GPUs, il est nécessaire de faire une demande de ressources via [ce formulaire](#) puis de suivre les instructions sur notre [FAQ](#) (« How to use GPU node ? »)

Espace de stockage

Nous avons racheté de l'espace work courant 2024, si vous avez besoin de plus d'espace que le 1 To de quota par défaut, il est possible de louer de l'espace supplémentaire en faisant une demande de ressources via [ce formulaire](#).

Vous rencontrez des problèmes pour vous loguer ?

N'hésitez pas à faire un ticket au [support](#) pour que nous investiguions. Nous avons remarqué que ce problème est fréquemment rencontré lorsque tout un laboratoire accède au cluster en utilisant la même IP et que celle-ci est bloquée par notre firewall suite à de multiples erreurs de mot de passe.

Création de compte et gestion des permissions

Nous ouvrons près de 300 comptes par an de manière automatisée. Pour être sûr des droits d'accès à donner sur les projets nous préférons donc agir en deux temps : nous créons le compte puis le responsable de l'espace projet nous demande de rajouter les accès souhaités.



Trucs et astuces

Vous nous avez demandé le retour de cette rubrique « Trucs et astuces ». À l'occasion de ce numéro 43 nous avons décidé de faire un focus sur Open On Demand (OOD) et sur les newsletters elles-mêmes dans le cadre de cette rubrique.

[OOD](#) permet de lancer via une interface web un [Rstudio](#), un [Jupyter Notebook](#) ou un [bureau Linux](#) sur le cluster, il permet aussi de naviguer dans les répertoires ou fichiers du cluster. **Nous avons récemment ouvert cette possibilité à l'ensemble de nos utilisateurs**, mais si cela engendre des problèmes nous pourrions le soumettre à demande exceptionnelle à nouveau.

Les newsletters sont un moyen de communication important pour nous, elles sont destinées à l'ensemble de nos utilisateurs (biologistes inclus) et contiennent des informations capitales sur les services que nous proposons. Nous en envoyons 2 à 3 par an maximum par mail. **Elles sont à tous moments disponibles sur notre site web sur [cette page](#).**



Prochains cycles d'apprentissage

Nous attirons votre attention sur le [calendrier](#) des formations et cycles d'apprentissage de l'IFB ainsi que sur ceux de la plateforme [GenoToul Biostat](#). De notre côté, nous proposons en partenariat avec Sigeneae et SaAB (MIAT) plusieurs cycles d'apprentissage : métagénomique du 5 au 7 mai, Python pour la bioinfo les 21 et 22 mai, Linux le 18 novembre, Cluster le 19 novembre, RNASeq (bioinfo et biostat) du 24 au 27 novembre, utiliser les workflows nf-core le 1^{er} décembre.

Toutes les informations et les formulaires d'inscription sont disponibles [ici](#).

La prochaine formation FROGS, analyse de données de metabarcoding, se tiendra les 12-13 et 19-20 mai en webinaire. Pour tout renseignement et inscription c'est [ici](#).

Pourquoi utiliser CodeServer plutôt que VSCode ?

De plus en plus d'utilisateurs utilisent Visual Studio Code (VSCode) sur leur poste personnel pour éditer leurs scripts et fichiers directement sur les nœuds de login à l'aide de l'extension SSH. Bien que cette pratique semble anodine, chaque utilisateur utilisant cette méthode lance en réalité une copie de VSCode sur les serveurs frontaux. Cela consomme énormément de ressources en mémoire, calcul et d'accès disques sur les serveurs de login jusqu'à les bloquer. Les processus lourds, comme VSCode, R, ou autres outils de bioinformatique, lancés sur les serveurs frontaux provoquent plusieurs crashes par an, interrompant les sessions et jobs interactifs des autres utilisateurs. Afin de limiter ces problèmes, nous vous proposons d'utiliser *code-server* sur les nœuds de calcul. Il s'agit de VSCode accessible dans votre navigateur, de la même façon que vous pourriez le faire avec un notebook jupyter.

Nous vous invitons à consulter le site web de [code-server](#) et [comment l'utiliser sur le cluster](#).

En cas d'utilisation (exceptionnelle) de VSCode, merci de respecter les bonnes pratiques suivantes :

- Ne pas exécuter de code au travers de VSCode, y compris vos notebooks jupyter, quarto, etc. Vous pouvez utiliser [Open On Demand \(OOD\)](#) ou *code-server* pour cela.
- Ne pas ouvrir tout votre *home*, *work*, *save* ou *projet*. Seulement le dossier qui contient le code que vous voulez éditer. Cela suppose que [votre projet est bien organisé](#) et que le code ne se trouve pas dans le même dossier que vos données.
- De configurer VSCode pour limiter au maximum le scan de fichiers, pour cela :
 - interdisez tous les dossiers contenant beaucoup de données, mais aussi les dossiers d'environnement (conda, snakemake, node_modules, etc.),
 - VScode comprend [le contenu des fichiers .gitignore, .ignore et .rignore](#) pour limiter le scan. Utilisez-les !
 - Vous pouvez également [éditer le fichier de configuration settings.json](#) de VSCode et définir des [motifs d'exclusion globaux](#). Configurez en particulier les clés "*files.exclude*", "*files.watcherExclude*" et "*search.exclude*".
 - Nous vous recommandons aussi d'interdire à VSCode de suivre les liens symboliques en ajoutant "*search.followSymlinks*": *false* au [fichier settings.json](#) et de n'installer/activer que les extensions nécessaires à votre projet.
- N'installer/activer que les extensions nécessaires à votre projet. Toute extension gourmande est à proscrire (analyse/génération de code local par exemple).



Le coin des débutants

Suite à vos retours nous nous sommes lancés en 2024 dans une nouvelle modalité de formation, les captures vidéos courtes (30 minutes à 1 heure maximum). Ce format semble vous plaire, nous en proposerons donc de nouvelles en 2025.

Pour le moment nous avons déjà enregistré quelques vidéos dont une vidéo traitant de la connexion au cluster et le lancement d'un blast en srin : <https://nextcloud.inrae.fr/s/Z4XRmciKw5364sr>, et trois vidéos sur Open On demand (OOD) pour utiliser Rstudio : <https://nextcloud.inrae.fr/s/ePpb3oDdjYT9kmG>, Linux Desktop : <https://nextcloud.inrae.fr/s/rNWYqTSFa8YjWqp>, et Jupyter Notebook : <https://nextcloud.inrae.fr/s/5w2jB8tEJ2Koj2j>. L'utilisation d'OOD n'est plus soumise à demande exceptionnelle, si vous êtes trop nombreux ou si vous demandez plus de ressources que disponibles, vous resterez donc en file d'attente.

Nous proposons également un service d'accompagnement des projets de recherche aux biologistes, si vous cherchez un partenaire bioinformatique pour monter un projet par exemple ou si vous avez besoin d'aide pour analyser des données, vous pouvez nous contacter via le [formulaire dédié](#) sur notre site web. Nous étudierons votre demande et reviendrons rapidement vers vous.



Logiciels

Sur le cluster plus de 800 logiciels ont été installés sur demande de nos utilisateurs via [ce formulaire](#), la quasi totalité n'a pas été développée par nous, par conséquent, nous ne pouvons pas maîtriser l'utilisation et la documentation de l'ensemble de ces logiciels, même si nous faisons de notre mieux pour vous aider.

Nous avons mis en place des scripts, comme celui pour trouver les packages Python : `search_Python_package NomDuPackage`. Par exemple : `search_Python_package biopython` trouvera les 5 versions actuellement installées. Vous trouverez beaucoup d'information pour vous aider à trouver ou utiliser les logiciels sur notre FAQ rubrique [« software »](#). Par exemple un [tutoriel](#) expliquant comment utiliser R en ligne de commande sur le cluster, comment installer des packages R ou encore compiler un fichier Rmarkdown. Vous êtes plusieurs à nous avoir demandé un tutoriel pour vous guider dans la création et l'utilisation d'environnement conda, celui-ci est en cours d'écriture, il devrait être très bientôt disponible sur notre FAQ. A ce propos, pour mieux contrôler vos environnements, si vous rencontrez des incompatibilités entre différents modules chargés par exemple, la création sur mesure de votre environnement conda peut être une solution.



Banques

Il existe plusieurs façons de trouver les banques disponibles sur notre cluster. Celles-ci sont pointées sur notre FAQ rubrique [« databanks »](#). En particulier, vous pouvez vous connecter à [Biomaj Watcher](#) avec votre login et mot de passe du cluster. Vous y trouverez la liste des banques puis en cliquant, leur localisation sur le cluster et les formats mis à disposition. Si vous préférez la ligne de commande nous avons mis en place le script `bank_info.sh`.

Pour obtenir la liste des banques : `bank_info.sh -a`

Pour obtenir des informations sur une banque (formats, url, date de mise à jour...): `bank_info.sh -b <nomDeLaBanque>` ;

Autre solution : explorer le répertoire `/work/bank2`. Vous y trouverez l'ensemble des banques (même celles non gérées via biomaj), les indexs blast sont dans le répertoire `blastdb`, les banques kraken dans `kraken`, les banques ensembl dans `ebi/ensembl`, etc.

Nous prévoyons d'enregistrer une capsule vidéo pour mieux vous guider dans l'utilisation des banques dans les prochains mois. Nous espérons qu'elle répondra à vos attentes.