ENQUÊTE

SUR L'USAGE ET LES BESOINS EN RESSOURCES BIOINFORMATIQUES

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

RAPPORT ENQUETE 2020

ABSTRACT

At the end of 2019, IFB conducted a survey targeting the life sciences and bioinformatics communities. The targeted researchers belonged to our main funding bodies, as well as to the data-producing national infrastructures. In total, more than **400 answers were collected**, mostly from whole teams rather than individuals. The main goal of this survey was to orientate IFB actions towards the actual expectations of the user communities. It was intended to measure the evolution of the needs of the life sciences communities. This was achieved through a set of questions addressing several topics, of which:

- Exponential increase in storage and computing requirements
- Growing need in bioinformatics skills and competencies
- Use of specific and rapidly evolving software resources
- Access to specialized databases
- Need to control the entire processing chain to ensure the openness and reproducibility of results (Data Management Plan, workflows, software environments)

The main information gained is that both communities expect to see bioinformatics questions take on more and more importance. To illustrate that, we can see the huge expectations about recruitment, training and collaboration. **75% of participants wish to engage in training**.
The survey shows **the lack of knowledge about "Data Management Plan"** and raises the huge necessity of spreading knowledge and training about DMP. For that purpose, IFB established two courses on data management in bioinformatics following the survey.

The survey also highlights **the importance of training around bioinformatics competences (data analysis, workflows, biostatistics...)** and it's a point IFB improved significantly since 2019 with the creation of a University Diploma in Integrative Bioinformatics.

Finally, the survey shows **the growing need of computing & storage infrastructures due to the increase of projects generating a lot of data to store and the need of an increased compute capability to analyse them.** This raises questions like: how to deal with sensitive data (HDS)? How to answer the storage problem for old projects? And how to improve the computational and storage facilities? IFB will deal with these questions in its future projects.

RÉSUMÉ

Fin 2019, l'IFB a mené une enquête auprès des communautés des sciences de la vie et de la bioinformatique. Les chercheurs ciblés appartenaient à nos principales tutelles, ainsi qu'aux infrastructures nationales productrices de données. Au total, plus de **400 réponses ont été recueillies**, principalement auprès d'équipes entières plutôt que d'individus. L'objectif principal de cette enquête était d'orienter les actions de l'IFB vers les attentes réelles des communautés d'utilisateurs. Elle visait à mesurer l'évolution des besoins des communautés des sciences de la vie. Ceci a été réalisé grâce à une série de questions abordant plusieurs sujets, dont :

- Augmentation exponentielle des besoins de stockage et de calcul
- Besoin croissant d'aptitudes et de compétences en bioinformatique
- Utilisation de ressources logicielles spécifiques et évoluant rapidement
- Accès à des bases de données spécialisées
- Besoin de maîtriser l'ensemble de la chaîne de traitement pour garantir l'ouverture et la reproductibilité des résultats (Plan de Gestion des Données, workflows, environnements logiciels)

Cette enquête s'adressait en priorité aux équipes et unités de recherche ou de service, cependant les réponses individuelles ont été autorisées. L'écrasante majorité des réponses sont formulées au titre d'unités et d'équipes de recherche plutôt que d'individus. Elles couvrent aussi un éventail représentatif des communautés des sciences de la vie. Les résultats détaillés de la couverture de l'enquête sont fournis en annexe, en incluant une représentation (1) par instituts/département de chaque EPST ; (2) pondérée par le nombre d'ETP des unités/équipes. L'enquête a couvert des structures de tailles diverses (Annexes A10 et A11).

L'enquête montre d'énormes attentes en matière de recrutement, de formation et de collaboration. **75 % des participants souhaitent suivre une formation.**

L'enquête montre le **manque de connaissances sur le « Plan de gestion des données »**. A cet effet, l'IFB a mis en place deux cours sur la gestion des données en bio-informatique depuis l'enquête.

L'enquête souligne également l'**importance de la formation autour des compétences en bioinformatique (analyse de données, workflows, biostatistiques...)** et c'est un point sur lequel l'IFB s'est nettement amélioré depuis 2019 avec la création d'un Diplôme Universitaire en Bioinformatique Intégrative.

Enfin, l'enquête montre le **besoin croissant d'infrastructures de calcul et de stockage en raison de l'augmentation des projets générant beaucoup de données à stocker et le besoin d'une capacité de calcul accrue pour les analyser**. Cela soulève des questions comme : comment traiter les données sensibles (HDS)? Comment répondre au problème du stockage des anciens projets? Et comment améliorer les installations de calcul et de stockage? L'IFB traitera de ces questions dans ses futurs projets.

Une écrasante majorité des réponses (93%) est positive concernant les **besoins en compétences** et ce pour toutes les tutelles. Ce besoin est surtout marqué pour les postes d'ingénieurs en analyse de données, développement de workflows, d'algorithmes et de bases de données. Au-delà des recrutements nécessaires, on note la volonté des structures de s'engager dans des collaborations et d'améliorer leurs connaissances via la formation.

Du point de vue des formats de formations souhaités, ils sont partagés entre des formations en immersion d'une semaine et des formations courtes d'un jour ou deux. La formule permettant aux participants d'apporter leurs propres données est très appréciée.

Concernant la volumétrie, en 2019 et à 3 ans, soit 2022, la grande majorité des unités évoluent avec 1 à 100 To. Globalement, l'ensemble de répondants s'attend à une forte augmentation des données, autant pour les données en cours de traitement que pour les données archivées. En termes de capacité de calcul, l'utilisation des processeurs GPU se popularise, en accord avec l'augmentation des besoins en analyse d'image.

Un travail de sensibilisation est à mener auprès des utilisateurs qui privilégient des ordinateurs individuels équipés de forte capacité de RAM, plutôt que des **ressources mutualisées, à l'échelle régionale ou nationale**, qui pourtant s'avèrent avantageuses pour l'ensemble de la communauté sur plusieurs points : frais de maintenance, coût financier et environnemental, etc. L'IFB a dès lors une grande responsabilité dans l'évolution des pratiques de la communauté, cela s'illustre notamment par son engagement dans le projet MUDIS4LS.

Sur le plan des **Ressources logicielles**, on observe certaines spécificités d'usage des serveurs européens (EBI) et américains (NCBI) selon le type de données. La ressource la plus utilisée est la base de données bibliographiques du NCBI Pubmed, pour laquelle il n'existe pas d'équivalent européen. Pour les protéines, Uniprot est sans surprise la ressource la plus utilisée. Pour les séquences génomiques, les équipes françaises semblent utiliser à la fois les ressources européennes (Ensembl, EnsemblGenomes) et américaines (Genome). Pour les données NGS, on observe une plus forte adoption des bases de données américaines pour les séquences brutes (SRA versus ENA) et leur interprétation primaire (GEO versus ArrayExpress). On notera cependant une utilisation significative d'Expression Atlas, pour lequel il n'existe pas d'équivalent américain.

Après consultation de la littérature, on constate que les ressources les plus utilisées coïncident avec un taux de citation très élevé des publications associées (Fig. 16). La France produit des ressources logicielles qui bénéficient d'une forte reconnaissance internationale (indicateurs de citation) et nationale. Il faut noter que ces ressources sont produites en s'appuyant sur l'expertise d'équipes de recherche en bioinformatique auxquelles sont adossées des plateformes de l'IFB qui assurent le déploiement de services organisés autour de ces ressources, facilitant l'adoption par de larges communautés d'usagers.

Les **données génomiques** sont les plus utilisées, suivies par l'imagerie puis les autres omiques. Il est intéressant de noter que l'IFB s'est engagé dans plusieurs projets qui intègrent des données multi-omiques et d'imagerie (projets-pilotes InexMed et PhenoMeta, study cases de MUDIS4LS).

Les **méthodes d'IA** s'appliquent aux thématiques de près de 70% des répondants. De façon notable, les proportions s'inversent pour la question "collaborez-vous déjà avec des équipes spécialisées en IA pour l'analyse de vos données biologiques ?" Ceci suggère qu'il serait stratégique d'organiser des événements (ateliers, formations, hackathons) pour stimuler la rencontre entre les communautés des sciences de la vie et celles des spécialistes de l'IA. L'action "défis de la bioinformatique intégrative" de l'IFB vise précisément à fournir une réponse concrète à ces enjeux.

Les besoins exprimés concernent les principaux domaines d'application de la biologie (recherche fondamentale, santé, agriculture, environnement, biotechnologie).

A l'heure où les **plans de gestion de données** (PGD/DMP) deviennent obligatoires pour tous les projets scientifiques subventionnés au niveau national (ANR) ou international (EU), on note une très faible familiarité avec le concept. Fin 2019, seules 23% des équipes avaient recours à des PGD. La majorité des équipes (63%) expriment un besoin de formation en PGD. Pour ces trois questions (familiarité, utilisation, demande de formations), l'INRAE se démarque systématiquement par un plus grand taux de réponses positives. Ceci correspond à une politique active entreprise dès 2016 par l'établissement pour engager ses équipes dans une démarche de science ouverte (projet Datapartage).

Les résultats de cette enquête montrent de forts besoins et attentes des communautés françaises des sciences de la vie et de la bioinformatique dans chacune des thématiques abordées (compétences et formations recherchées, infrastructure de calcul et stockage, ressources logicielles, intelligence artificielle et gestion des données). Cette enquête soulève pour nos communautés des questionnements et domaines globaux : ceux dans lesquels nous sommes déjà très actifs (NGS, bioinformatique intégrative, techniques bioinfo pour les bioinfo, ...) et ceux sur lesquels nous sommes en progression (PGD/DMP, utilisation de l'IA...). Les problématiques qui ressortent de l'enquête sont d'une part l'arrivée de l'analyse des données d'imagerie et leur intégration avec d'autres données (omiques...) dans la bioinformatique, d'autre part la mise en évidence du besoin urgent d'infrastructure national HDS et enfin souligne l'importante montée en charge des données qui réclament de plus en plus de capacités de stockage et de calcul. Ce sont d'ailleurs des points abordés dans le cadre du projet d'infrastructure numérique MUDIS4LS. Pour faire face à cela et devant l'ampleur du travail et des coûts pour y répondre, la **mutualisation** semble être une réponse efficiente.

On peut citer comme exemples de projets qui profitent de la **complémentarité** des plateformes de l'IFB : le projet de surveillance des variants du SARS-CoV-2 (EMERGEN), celui de surveillance des antibiorésistances (ABRomics) ou encore la mise en place d'environnements logiciels modulaires (infrastructure environnée), taillés en fonction des besoins des équipes de recherche universitaires.

CONTEXTE ET OBJECTIFS DE L'ENQUÊTE

Ce rapport décrit les résultats d'une enquête sur les besoins des unités et équipes de recherche en matière de bioinformatique.

L'Institut Français de Bioinformatique (IFB) a lancé à l'automne 2019 une enquête sur les besoins des unités et équipes de recherche en matière de bioinformatique avec une large diffusion auprès des laboratoires de biologie et bioinformatique de France. Les résultats de cette enquête sont déterminants pour optimiser l'organisation de l'offre de services, et assurer une mutualisation des moyens nationaux dans le cadre de notre mission d'infrastructure nationale de support à la recherche. Cette enquête se donne pour objectif de caractériser les besoins afin d'organiser et d'adapter la réponse de l'infrastructure nationale, et ainsi anticiper les besoins à venir. Elle est aussi une occasion pour l'ensemble des tutelles de faire un état des lieux de ce qui se fait et de ce qui est attendu par leurs équipes respectives.

Cette enquête est cruciale pour répondre à l'indéniable évolution des sciences du vivant :

- la transition marquée de la biologie vers une science des données, qui se traduit non seulement par l'augmentation exponentielle de besoins de stockage et calcul, mais aussi par la nécessité de gérer la donnée tout au long de son cycle de vie (Data Management Plan)
- le besoin croissant en compétences bioinformatiques
- l'expression croissante de besoins de formation
- l'utilisation de ressources logicielles spécifiques et en évolution rapide
- l'accès à des banques de données spécialisées
- la nécessité de maîtriser l'ensemble de la chaîne de traitement pour assurer l'ouverture et la reproductibilité des résultats (workflows, environnements logiciels)
- l'intérêt exprimé par les biologistes pour les applications de l'intelligence artificielle

CIBLES

Cette enquête s'adressait en priorité aux équipes et unités de recherche ou de service, cependant les réponses individuelles ont été autorisées. Ainsi elle a été remplie à différents niveaux :

- Au **nom d'une unité,** en veillant à répercuter les besoins de l'ensemble des équipes
- Au **nom d'une équipe**, l'enquête s'applique aux unités de recherche et/ou de service
- A **titre individuel**, au cas où les deux situations précédentes ne seraient pas remplies (personnes ressentant des besoins non reconnus par l'équipe, ou dont l'équipe/unité a décidé de ne pas répondre à l'enquête)

Conformément aux recommandations du formulaire, la grande majorité des réponses sont formulées au titre d'unités et d'équipes de recherche plutôt que d'individus *(Fig.1).* Elles couvrent aussi un éventail représentatif des communautés des sciences de la vie *(Fig.2).* Les résultats détaillés de la couverture de l'enquête sont fournis en annexe, en incluant une représentation (1) par instituts/département de chaque EPST; (2) pondérée par le nombre d'ETP des unités/équipes. L'enquête a couvert des structures de tailles diverses (Annexes A10 et A11).

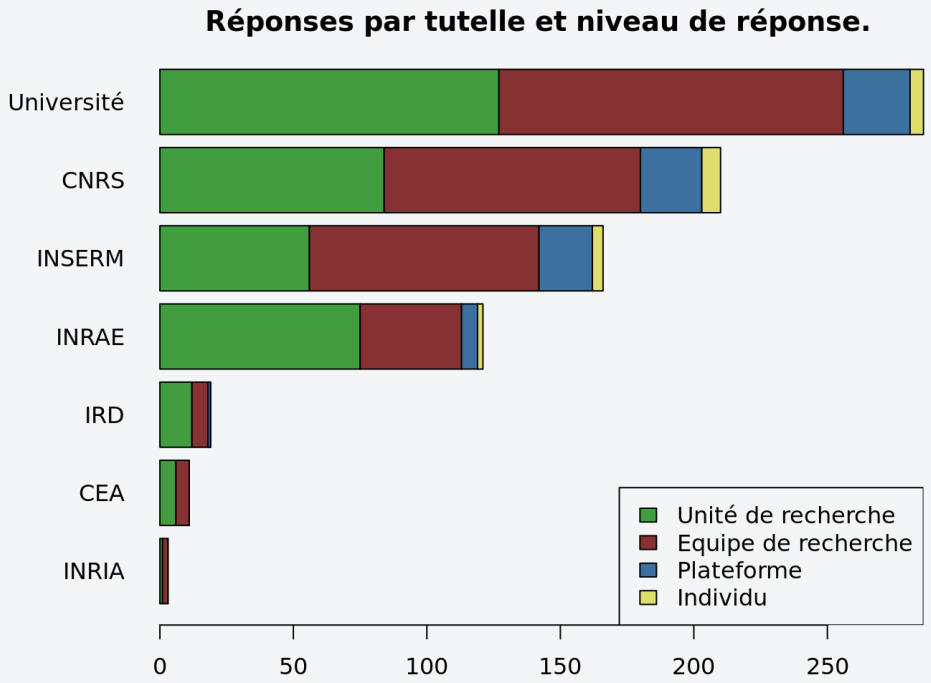


Figure 1. Nombre de réponses par établissement. Les couleurs indiquent le niveau de réponse (au titre d'une unité, équipe, plateforme, individuel)

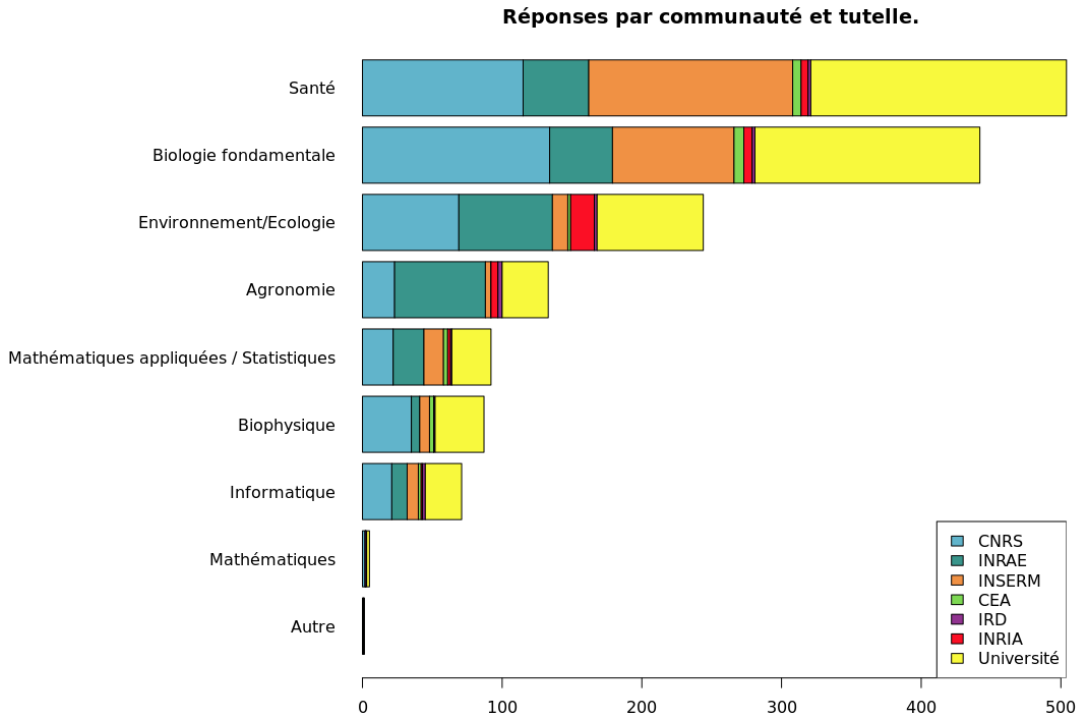


Figure 2. Communautés thématiques couvertes par l'enquête

MÉTHODOLOGIE

CONCEPTION DU QUESTIONNAIRE

Les questions ont été élaborées par le comité de pilotage et par les responsables d'actions de la feuille de route 2018-2021 de l'IFB. Un questionnaire a été élaboré, en reposant pour autant que possible sur des questions à choix multiples, tout en laissant systématiquement la possibilité de fournir une réponse spécifique en cochant "Autre" ou des réponses en texte libre.

Le questionnaire est fourni en annexe 1.

DIFFUSION DE L'ENQUÊTE

L'enquête a été largement diffusée via différents canaux:

- Unités de recherche et de services des différents organismes et instituts de recherche
- Liste des plateformes IFB en demandant de transmettre aux collègues biologistes
- Liste de la SFBI en demandant de transmettre aux collègues biologistes

La diffusion de l'enquête a donc couvert l'ensemble de nos tutelles, ainsi que plusieurs autres types d'infrastructure/organisme. Les organismes répondant sont donc nombreux et variés: Universités, CNRS, Inserm, INRAe, IRD, CEA, INRIA, CHU (Nantes, Limoges, Lille, APHP, APHM...), Cirad, Institut (Pasteur, Paoli Calmettes, INP, Gustave Roussy, INSAT, Ifremer, INTS, INP, Irstea), Grandes écoles (ENS, AgroParisTech, SupAgro...), autres institutions (Etablissement Français du Sang, ANSES, CGIAR, MNHN, Service de Santé des Armées, Généthon, Collège de France), et une start-up (Vect-Horus).

COMPÉTENCES ET FORMATIONS RECHERCHÉES

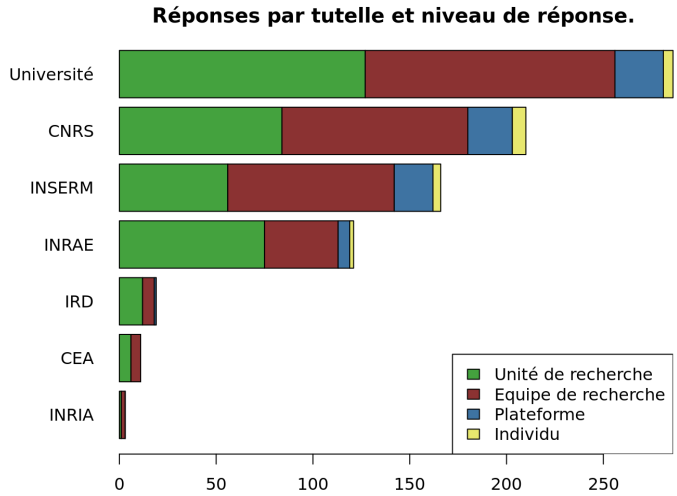


Figure 3. Filières professionnelles attendues pour couvrir les besoins de bioinformatique. Les données sont ventilées par tutelles

Une écrasante majorité des réponses (93%) est positive concernant les besoins en compétences et ce pour toutes les tutelles (annexe C1). Ce besoin est surtout marqué pour les postes d'ingénieurs *(Fig. 3)* en analyse de données, développement de workflows, d'algorithmes et de bases de données *(Fig. 4)*.

Au-delà des recrutements nécessaires, on note la volonté des structures de s'engager dans des collaborations et d'améliorer leurs connaissances via la formation *(Fig.5)*.

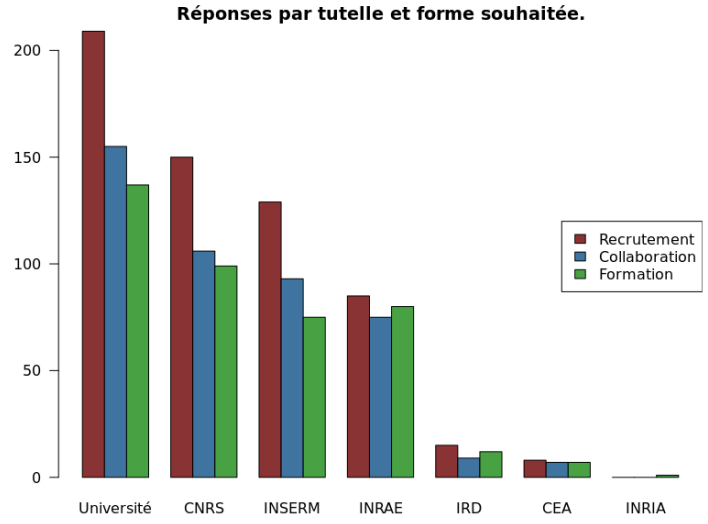


Figure 5. Stratégies envisagées pour répondre aux besoins en bioinformatique

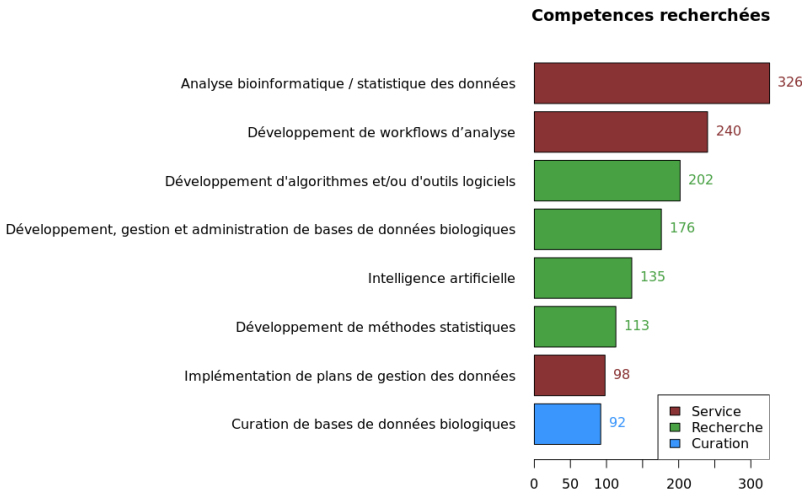


Figure 4. Compétences recherchées

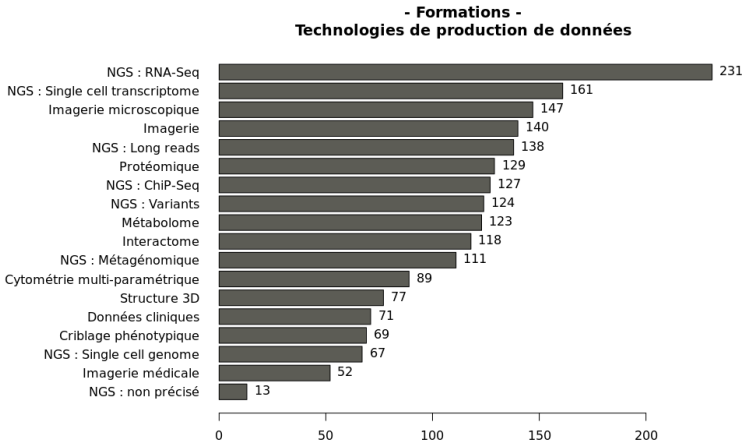


Figure 6. Demande de formations en technologies de production de données

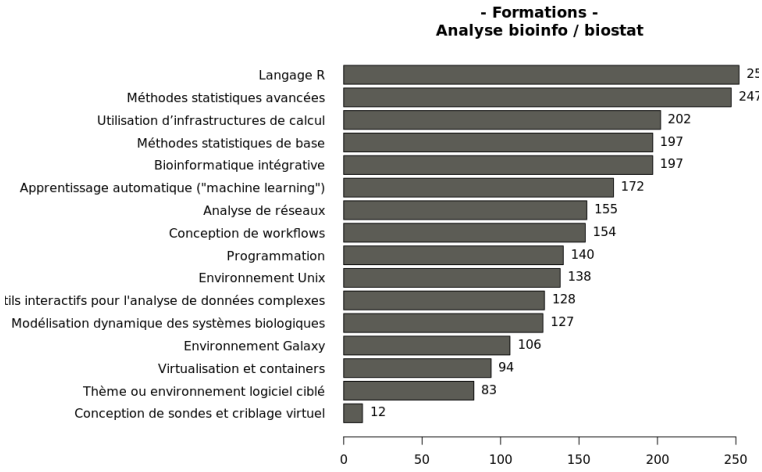


Figure 7. Besoins de formations à l'analyse bioinformatique/biostatistique

Le **RNA-seq** reste la technologie la plus demandée en matière de formation *(Fig 6)*. Ceci est vrai depuis plusieurs années. Le **single-cell transcriptomique** est aussi très demandé. Nous observons également une grosse demande en imagerie médicale et microscopique. Le besoin en formation concernant les autres omiques reste prégnant. Ceci correspond également aux tendances des projets récents en biologie intégrative. Ces résultats appellent notamment à mettre en place des formations en collaboration entre l'IFB et les deux infrastructures d'imagerie (FBI et FLI). Ceci pourra se faire dans le contexte de la feuille de route IFB (projet PIA3 MUDIS4LS, dans lequel nous avons une Implementation Study spécifiquement dédiée à l'intégration imagerie et multi-omiques). Sur l'analyse des données, les besoins en formation concernant les statistiques sont très en avant *(Fig. 7)*.

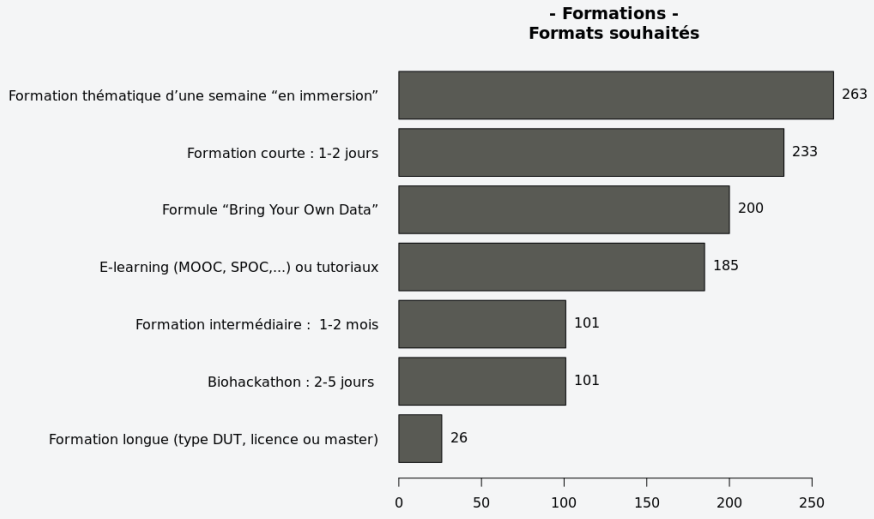


Figure 8. Formats souhaités pour les formations

Du point de vue des formats de formations souhaités, ils sont partagés entre des formations en immersion d'une semaine et des formations courtes d'un jour ou deux. La formule permettant aux participants d'apporter leurs propres données est très appréciée (*Fig. 8)*.

INFRASTRUCTURE DE CALCUL ET STOCKAGE

Concernant la volumétrie, en 2019 et à 3 ans, soit 2022, la grande majorité des unités évoluent avec 1 à 100 To. Globalement, l'ensemble de répondants s'attend à une forte augmentation des données, autant pour les données en cours de traitement que pour les données archivées (*Fig. 9, 10 et 11).*

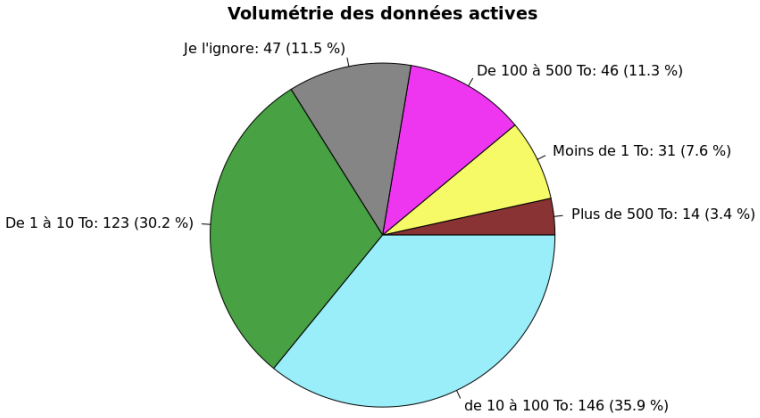


Figure 9. Volumétrie actuelle des données actives stockées par l'unité

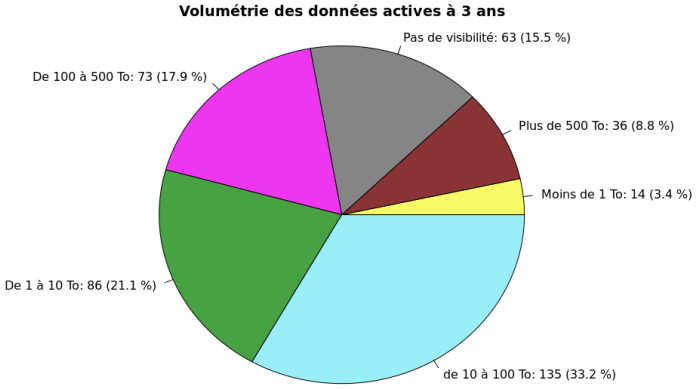


Figure 10. Visibilité sur les besoins en stockages de données actives à 3 ans

En termes de capacité de calcul, l'utilisation des processeurs GPU se popularise, en accord avec l'augmentation des besoins en analyse d'image.

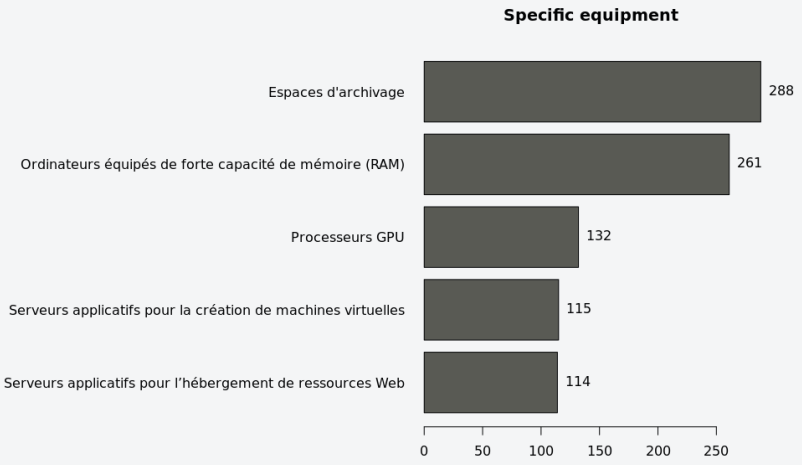


Figure 11. Anticipation des besoins spécifiques en termes de matériels

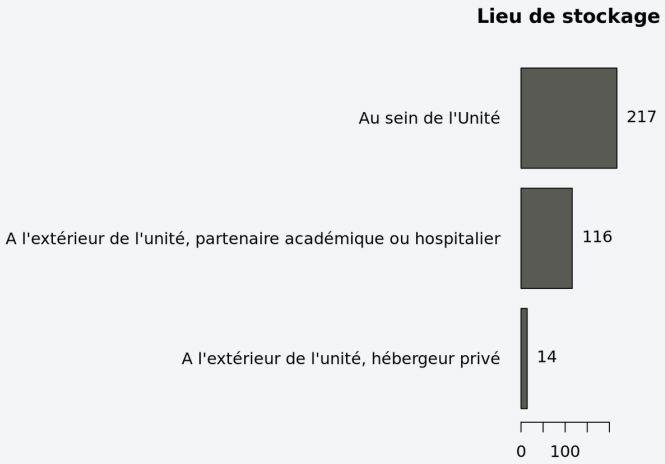


Figure 12. Lieu de stockage des données à caractères sensibles

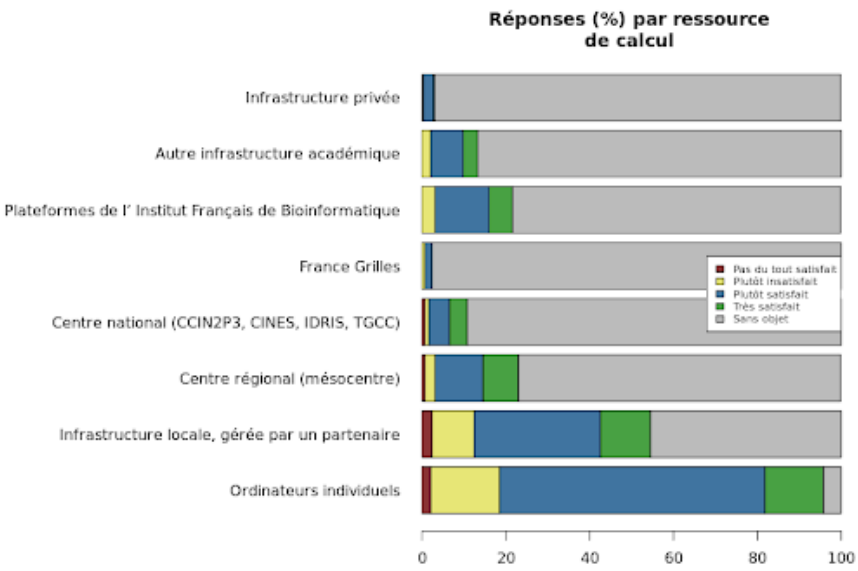


Figure 13. Ressources de calcul utilisées et niveau de satisfaction

Les résultats semblent refléter l'absence de structure de stockage HDS (hébergeur agréé de données de santé) au moment de l'enquête *(Fig. 12)* et encore aujourd'hui. Cette figure est construite uniquement à partir des réponses de structures qui ont des données sensibles à traiter. Dans une écrasante majorité des cas, les utilisateurs utilisent et sont satisfaits de leurs ressources de calcul locales (ordinateurs individuels ou infrastructures internes) *(Fig. 13)*.

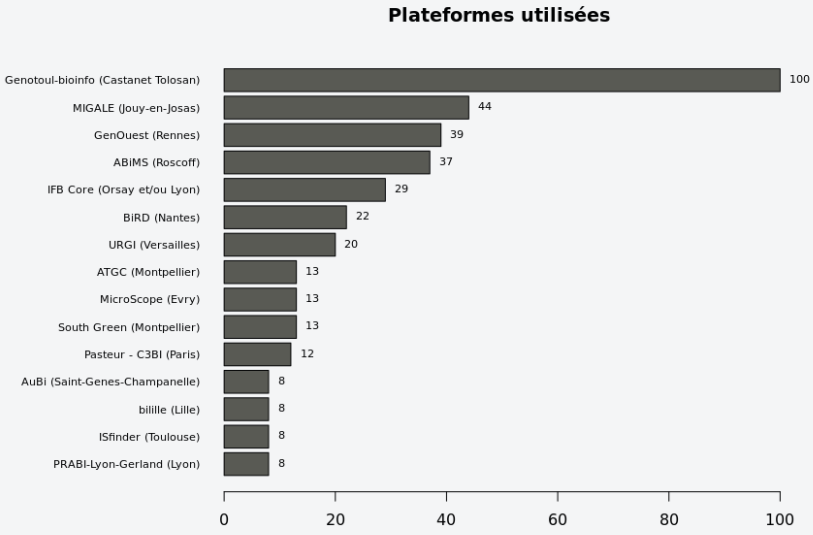


Figure 14. Plateformes bioinformatiques utilisées pour votre stockage et/ou calcul

Un travail de sensibilisation est à mener auprès des utilisateurs qui privilégient des ordinateurs individuels équipés de forte capacité de RAM, plutôt que des ressources mutualisées, à l'échelle régionale ou nationale, qui pourtant s'avèrent avantageuses pour l'ensemble de la communauté sur plusieurs points: frais de maintenance, coût financier et environnemental, etc.

L'IFB a dès lors une grande responsabilité dans l'évolution des pratiques de la communauté; cela s'illustre notamment par son engagement dans le projet MUDIS4LS.

RESSOURCES LOGICIELLES

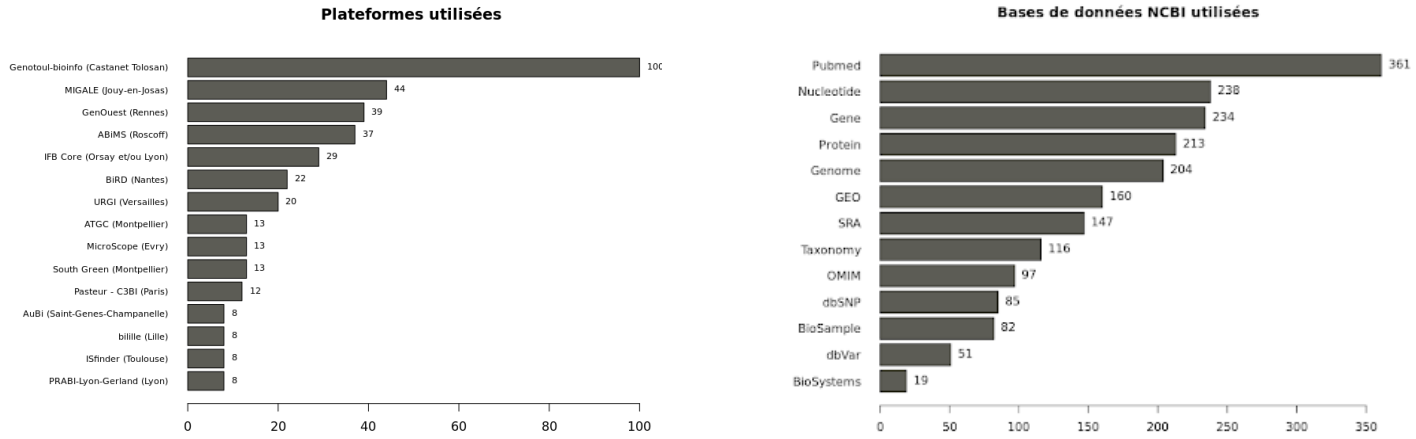


Figure 15. Utilisation par les équipes françaises des bases de données déployées par l'EBI (gauche) et le NCBI

On observe certaines spécificités d'usage des serveurs européens (EBI) et américains (NCBI) selon le type de données *(Fig. 15).* La ressource la plus utilisée est la base de données bibliographiques du NCBI Pubmed, pour laquelle il n'existe pas d'équivalent européen. Pour les protéines, Uniprot est sans surprise la ressource la plus utilisée. Pour les séquences génomiques, les équipes françaises semblent utiliser à la fois les ressources européennes (Ensembl, EnsemblGenomes) et américaines (Genome). Pour les données NGS, on observe une plus forte adoption des bases de données américaines pour les séquences brutes (SRA versus ENA) et leur interprétation primaire (GEO versus ArrayExpress). On notera cependant une utilisation significative d'Expression Atlas, pour lequel il n'existe pas d'équivalent américain.

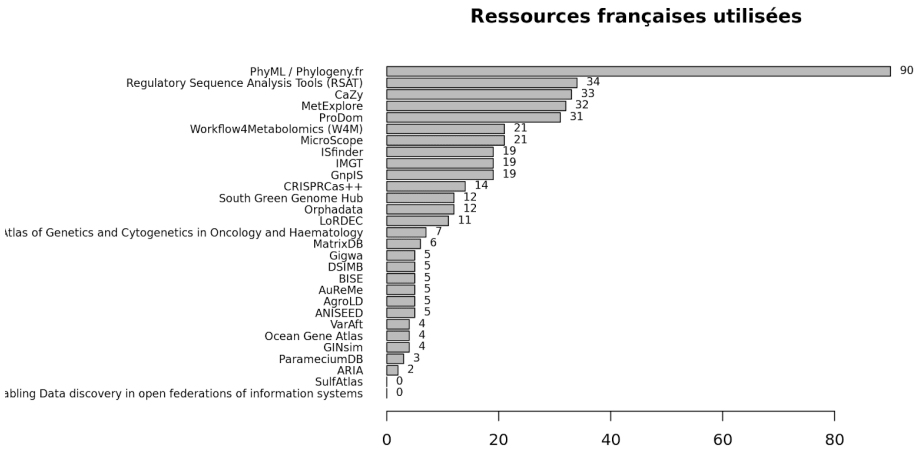


Figure 16. Utilisation des ressources bioinformatiques françaises

Après consultation de la littérature, on constate que les ressources les plus utilisées coïncident avec un taux de citation très élevé des publications associées *(Fig. 16).* La France produit des ressources logicielles qui bénéficient d'une forte reconnaissance internationale (indicateurs de citation) et nationale. Il faut noter que ces ressources sont produites en s'appuyant sur l'expertise d'équipes de recherche en bioinformatique auxquelles sont adossées des plateformes de l'IFB qui assurent le déploiement de services organisés autour de ces ressources, facilitant l'adoption par de larges communautés d'usagers.

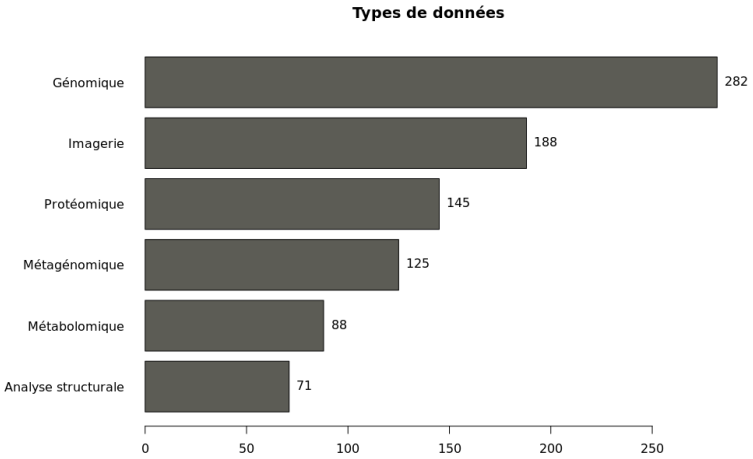


Figure 17. Types de données analysées

Les données génomiques sont les plus utilisées, suivies par l'imagerie puis les autres omiques *(Fig. 17)*. Il est intéressant de noter que l'IFB s'est engagé dans plusieurs projets qui intègrent des données multi-omiques et d'imagerie (projets-pilotes InexMed et PhenoMeta, study cases de MUDIS4LS).

Le traitement des données biologiques repose sur des enchaînements relativement complexes d'outils, formalisés dans des workflows. L'environnement le plus utilisé pour la gestion de ces workflows est Galaxy, dont l'interface Web a favorisé l'adoption rapide par les biologistes depuis l'avènement du NGS. On constate aussi une forte utilisation de deux solutions programmatiques plus récentes: Snakemake et NextFlow *(Fig. 18)*.

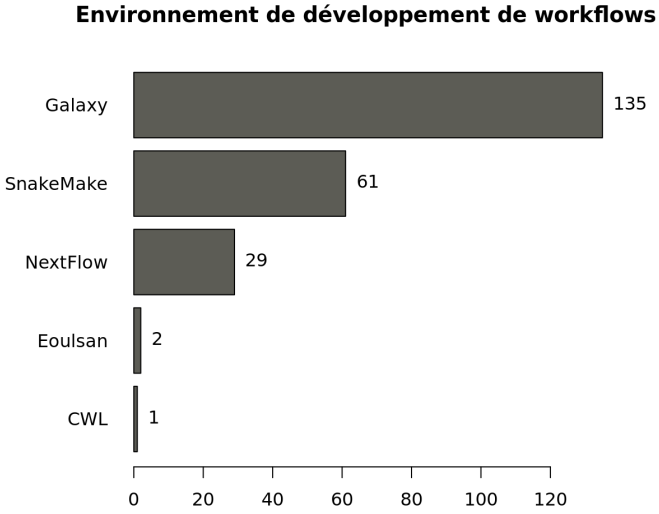


Figure 18. Environnement(s) de développement de workflows utilisés

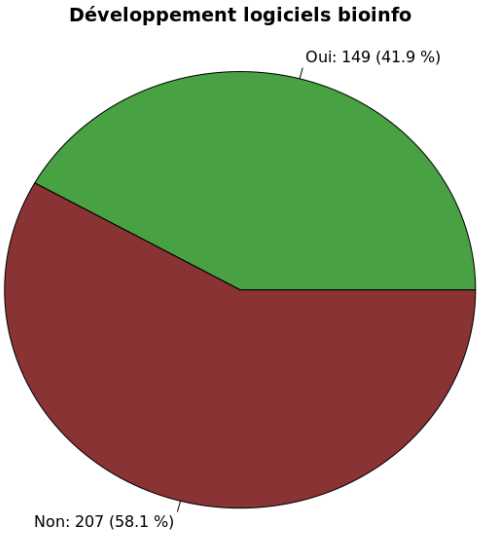


Figure 19. Développement des ressources logicielles pour la bioinformatique

42% des répondants indiquent que leur équipe développe des ressources logicielles pour la bioinformatique *(Fig. 19)*. Ceci pourrait refléter un biais de couverture de l'enquête en faveur d'équipes incluant des bioinformaticiens. Cependant, lors de futures enquêtes la notion de "ressource logicielle" mériterait d'être précisée, pour savoir quels types de ressources sont prises en considération (scripts R, workflows, méthodes, logiciels, bases de données), et pour savoir si ces ressources sont mises à disposition d'usagers externes, et sous quelle forme.

INTELLIGENCE ARTIFICIELLE

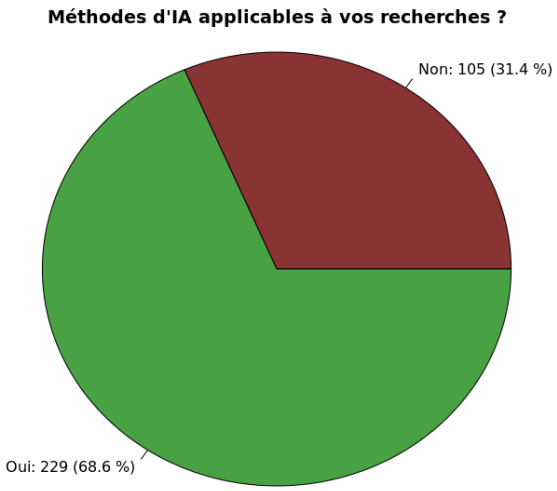


Figure 20. Les méthodes d'intelligence artificielle appliquées aux thématiques de recherche

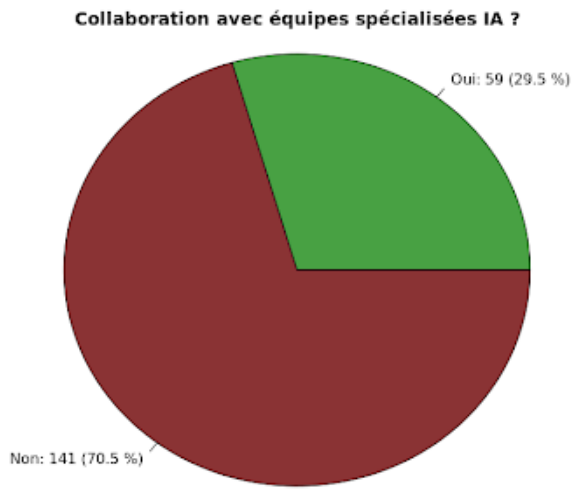


Figure 21. Collaboration

Les méthodes d'IA s'appliquent aux thématiques de près de 70% (229/334) des répondants *(Fig. 20).*De façon notable, les proportions s'inversent pour la question *"collaborez-vous déjà avec des équipes spécialisées en IA pour l'analyse de vos données biologiques ?" (Fig.21).* Ceci suggère qu'il serait stratégique d'organiser des événements (ateliers, formations, hackathons) pour stimuler la rencontre entre les communautés des sciences de la vie et celles des spécialistes de l'IA. L'action "défis de la bioinformatique intégrative" de l'IFB vise précisément à fournir une réponse concrète à ces enjeux.

- Les besoins exprimés concernent les principaux domaines d'application de la biologie (recherche fondamentale, santé, agriculture, environnement, biotechnologie) (*Fig. 22).*

- Parmi ceux-ci, 122 équipes utilisent déjà l'IA au sens large. On notera également 55 réponses positives pour le deep learning *(Fig. 23a)*, qui ne sont pas liées uniquement à l'analyse d'images microscopiques ou macroscopiques (*Fig. 24a, 24b, 23b et 23c).*

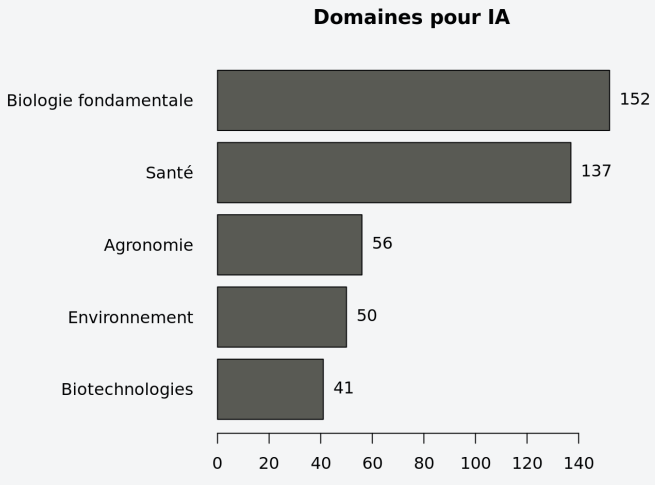


Figure 22. Domaines d'application pour l'IA

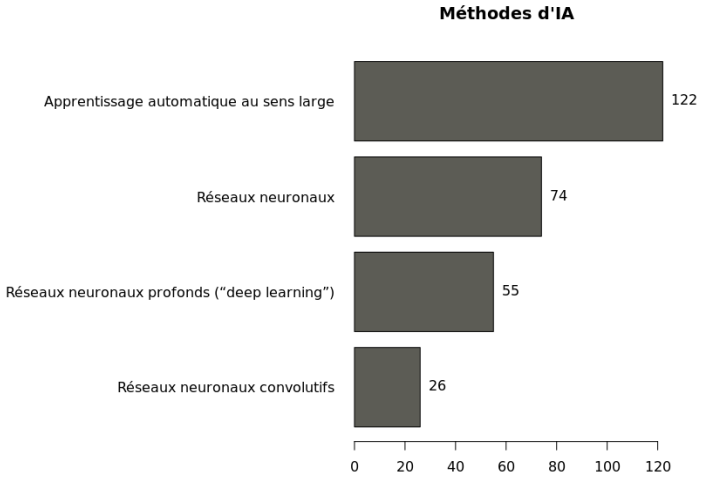


Figure 23a. Méthodes d'IA déjà utilisées par les équipes

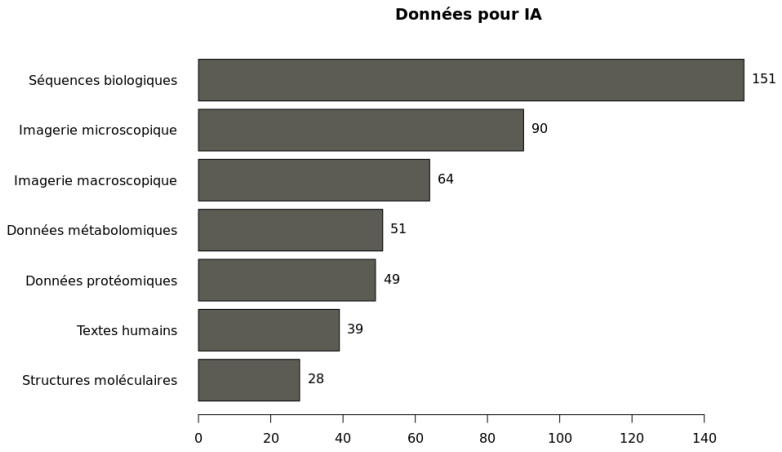


Figure 24a. Types de données pour lesquelles les méthodes d'IA sont utlisées

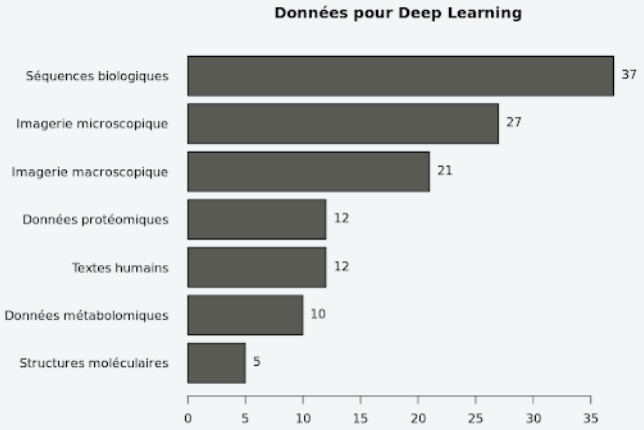


Figure 24b. Types de données pour lesquelles le Deep Learning est utilisé.

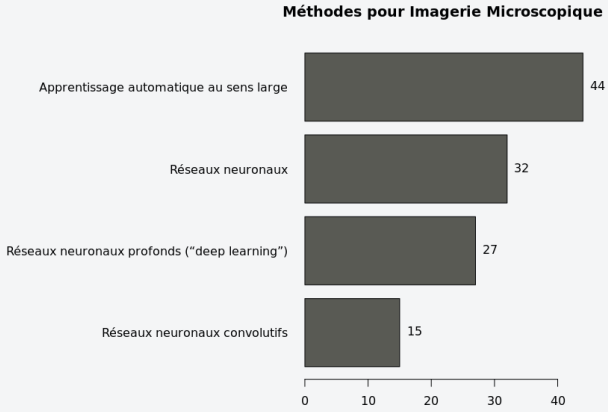


Figure 23b. Méthodes d’IA utilisées pour l’imagerie microscopique

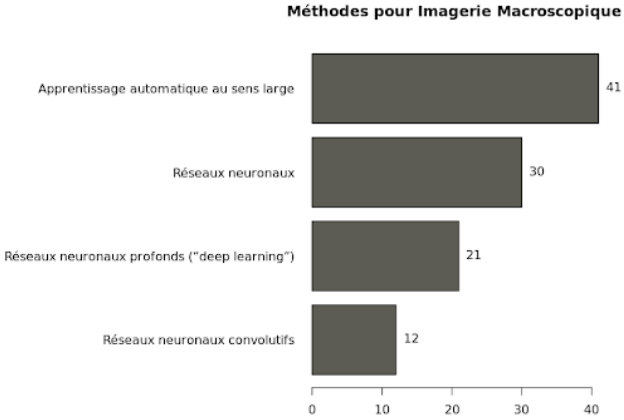


Figure 23c. Méthodes d’IA utilisées pour l’imagerie macroscopique

GESTION DES DONNÉES

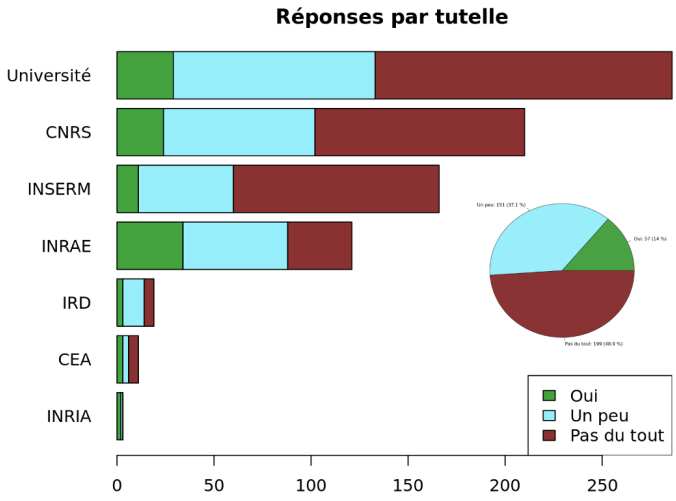


Figure 25. Familiarité avec le concept de plan de gestion de données (DMP, Data Management Plan)

A l'heure où les plans de gestion de données (PGD/DMP) deviennent obligatoires pour tous les projets scientifiques subventionnés au niveau national (ANR) ou international (EU), on note une très faible familiarité avec le concept (14% "oui", 37% "un peu" et 49% "non", cf (*Fig. 25)*. Fin 2019, seules 23% des équipes avaient recours à des PGD *(Fig. 26)*. La majorité des équipes (63%) expriment un besoin de formation en PGD *(Fig. 27*).

Pour ces trois questions (familiarité, utilisation, demande de formations), l'INRAE se démarque systématiquement par un plus grand taux de réponses positives. Ceci correspond à une politique active entreprise dès 2016 par l'établissement pour engager ses équipes dans une démarche de science ouverte (projet Datapartage).

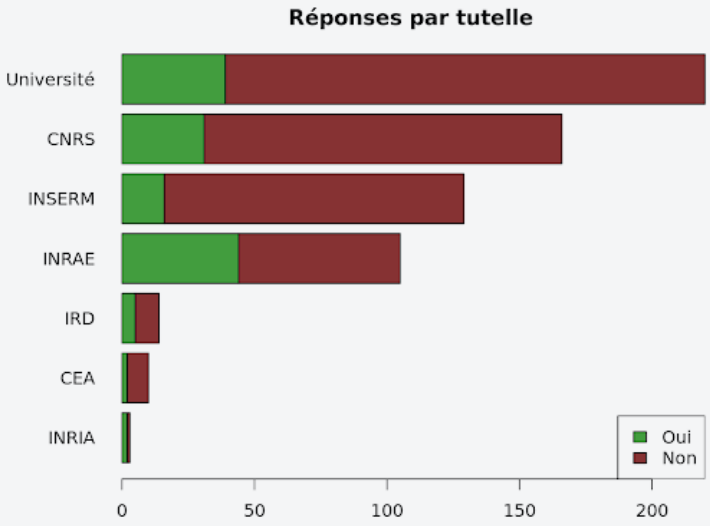


Figure 26. Equipe ayant recours à des PDG pour ses projets

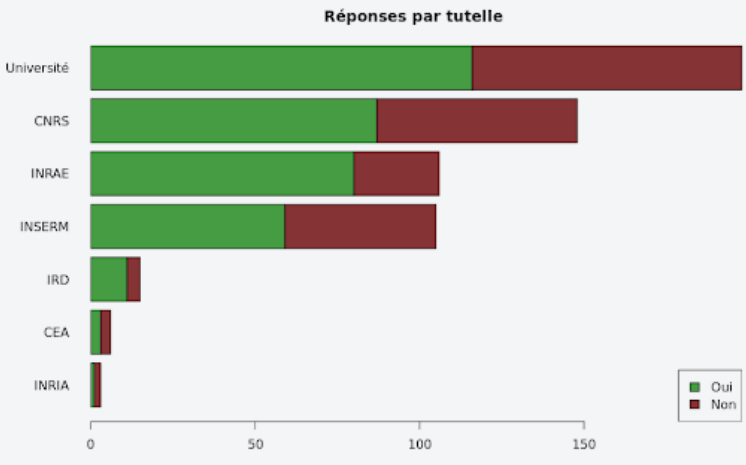


Figure 27. Besoin de formation à la gestion des PGD

AFFRONTER LES BESOINS

Les résultats de cette enquête montrent de forts besoins et attentes des communautés françaises des sciences de la vie et de la bioinformatique dans chacune des thématiques abordées:

- Compétences et formations recherchées
- Infrastructure de calcul et stockage
- Ressources logicielles
- Intelligence artificielle
- Gestion des données

Cette enquête soulève pour nos communautés des questionnements et domaines globaux: ceux dans lesquels nous sommes déjà très actifs (NGS, bioinformatique intégrative, techniques bioinformatiques pour les bioinformaticiens, ...) et ceux sur lesquels nous sommes en progression (PGD/DMP, utilisation de l'IA...).

Les problématiques qui ressortent de l'enquête sont d'une part l'arrivée de l'analyse des données d'imagerie et leur intégration avec d'autres données -omiques, d'autre part la mise en évidence du besoin urgent d'infrastructure national HDS (Hébergeur de Données de Santé), et enfin l'importante montée en charge des données qui réclament de plus en plus de capacités de stockage et de calcul.

Ce sont d'ailleurs des points abordés dans le cadre du projet d'infrastructure numérique MUDIS4LS. Pour faire face à cela et devant l'ampleur du travail et des coûts pour y répondre, la **mutualisation** semble être une réponse efficiente.

On peut citer comme exemples de projets qui profitent de la **complémentarité** des plateformes de l'IFB :

- Le projet de surveillance des variants du SARS-CoV-2: EMERGEN,
- Celui surveillance des antibiorésistances: ABRomics,
- Ou encore la mise en place d'environnements logiciels modulaires (infrastructure environnée), dimensionnés en fonction des besoins des équipes de recherche universitaires.

ANNEXE 1: QUESTIONNAIRE

Télécharger le questionnaire en PDF :

https://www.france-bioinformatique.fr/wp-content/uploads/form_enq_besoins-bioinfo_2020.pdf

ANNEXE 2: ANALYSE DÉTAILLÉE DES RÉPONSES

PARTIE A. IDENTITÉ DU RÉPONDANT

• A1. A QUEL TITRE RÉPONDEZ-VOUS (UNITÉ/ÉQUIPE/INDIVIDU)?

La grande **majorité des réponses** sont formulées **au titre d'unités et d'équipes de recherche** plutôt que d'individus.

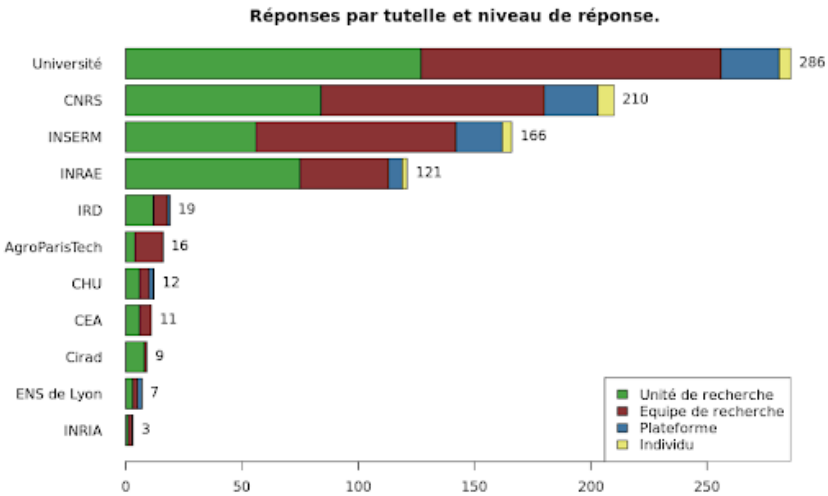
Ceci correspond au ciblage de l'enquête :

- Nous avions demandé de répondre si possible au nom d'équipes/unités, tout en laissant la possibilité de répondre à titre individuel.

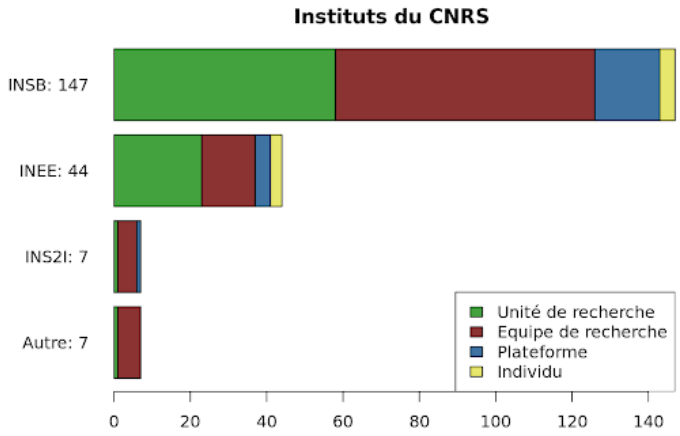
- Nous avons demandé aux tutelles et instituts de diffuser largement l'annonce auprès des unités/laboratoires de recherche, tout en laissant la possibilité aux plateformes de services de répondre.

Conformément à la demande exprimée sur le formulaire, la grande majorité des réponses sont des groupes: équipes, plateformes, unités.

La tutelle **"Université" est majoritaire**, mais n'est généralement pas unique, car la plupart proviennent d'unités et équipes affiliées à plusieurs tutelles (Unités Mixtes de Recherche, voir heatmap). Chaque répondant pouvait cocher plusieurs tutelles. La somme des réponses par tutelle est donc systématiquement supérieure au nombre total de réponses.



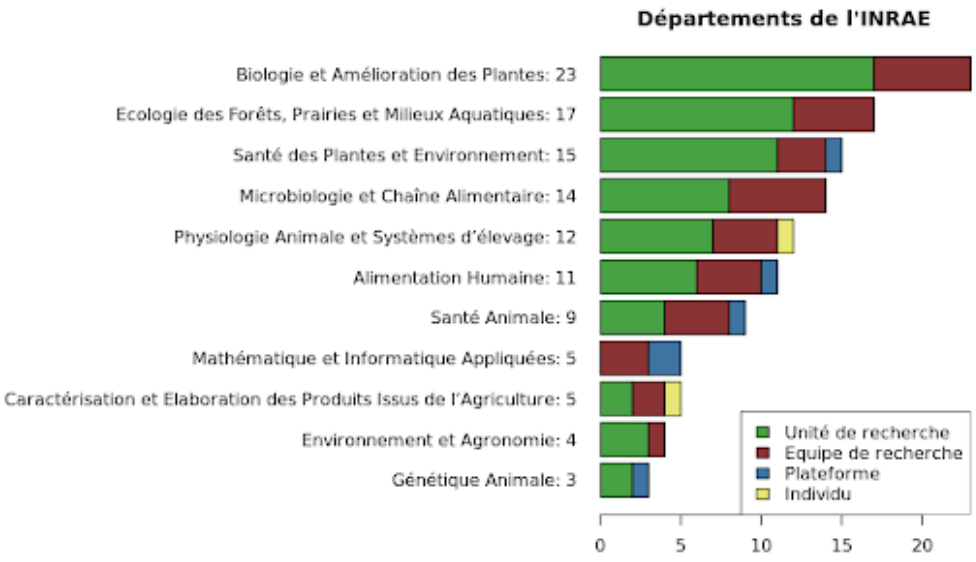
• A5. INSTITUTS DU CNRS



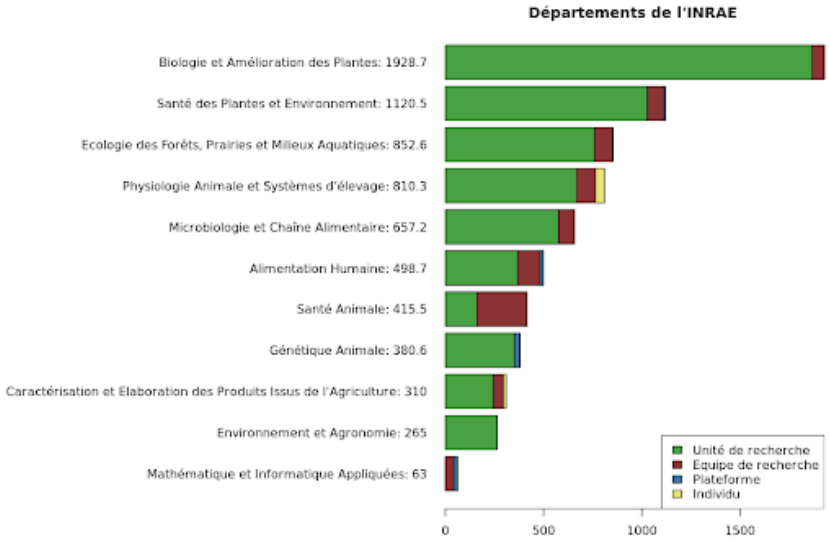
L'institut majoritaire en termes de réponses est l'INSB, mais nous avons également une quarantaine de réponses de l'INEE, dont une partie des thématiques inclut les sciences de la vie.

- Pour le CNRS, la couverture par institut est conforme à ce qu'on s'attendait à trouver.
- Les proportions unités / équipes / plateformes / individus sont similaires entre INSB et INEE.
- Les plateformes INEE et INSB incluent sans doute des services non seulement en bioinfo (PF IFB) mais aussi en biologie (ex: séquençage).
- A noter pour l'INS2I, 6 des 7 réponses sont des équipes (5) / unité (1), on a donc ici les réponses concernant la demande de services IFB pour des équipes INS2I. Le 7ème réponse provient d'une plateforme.

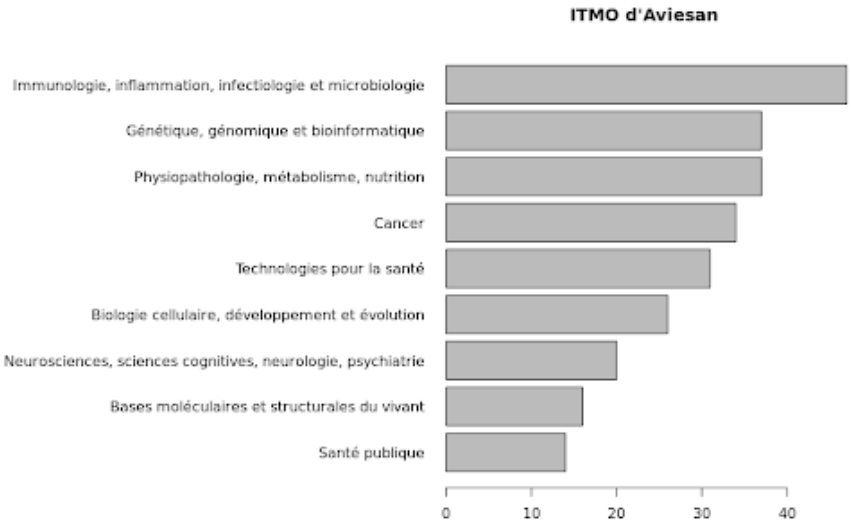
• A6. DÉPARTEMENT DE RECHERCHE INRAE



Le nombre de réponses pourrait jouer sur la réponse, car en physiologie animale et élevage il y a deux très grosses unités.



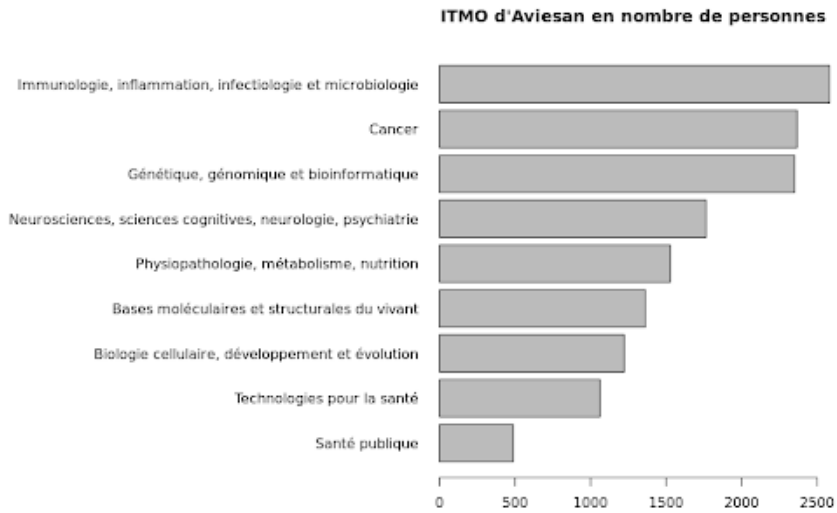
• A7. INSTITUT(S) THÉMATHIQUE(S) MULTI-ORGANISME(S) D'AVIESAN



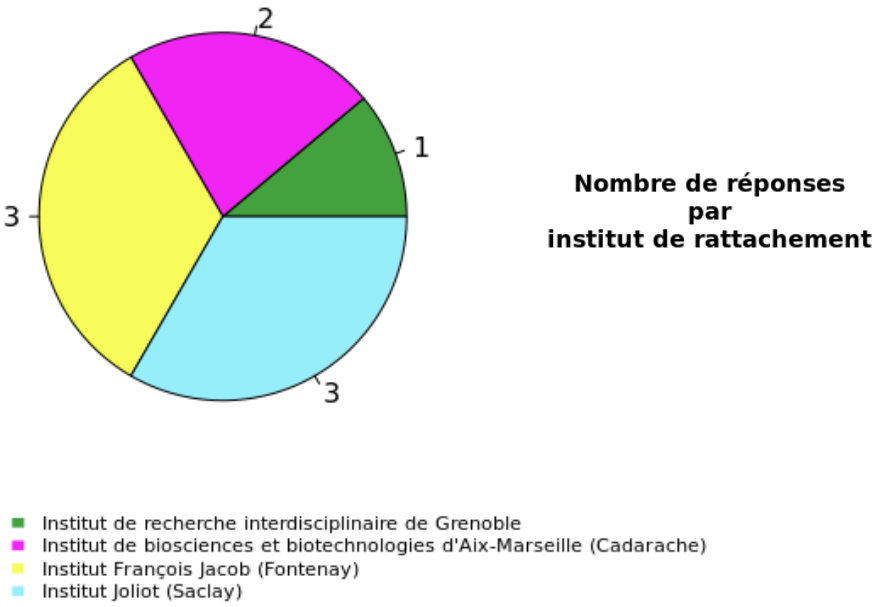
L'enquête a réellement touché tous les ITMO:

- À relativiser par rapport aux tailles des équipes /unités
- Très intéressant de constater la forte réponse de l'I3

- En nombre de personnes (réponses * tot ETP), la thématique du cancer passe en seconde position, car des unités de grande taille y travaillent.

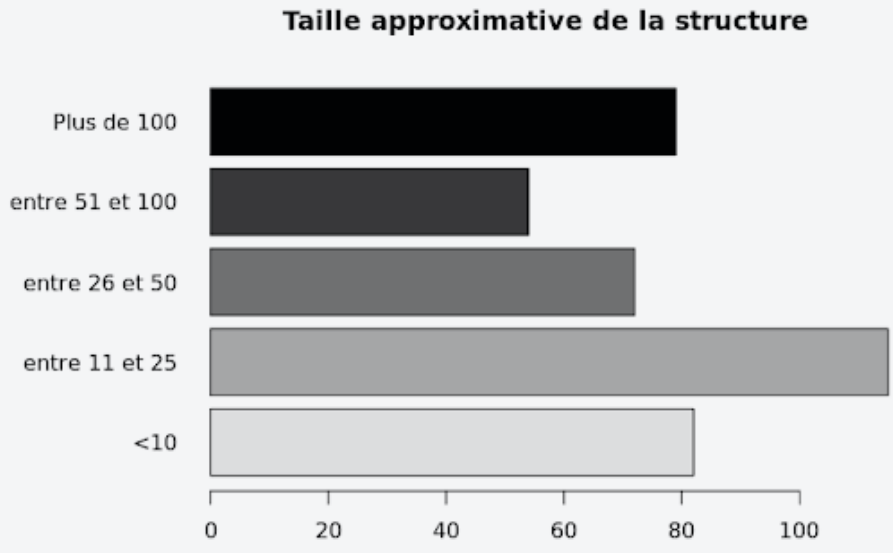


• A8. INSTITUT DE RATTACHEMENT

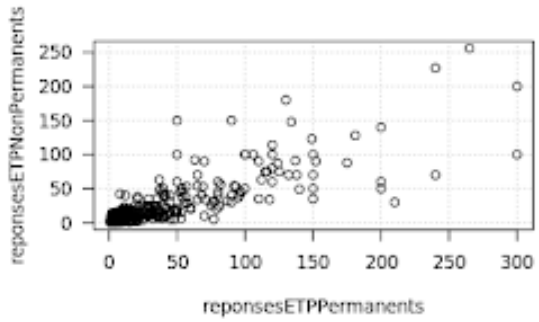


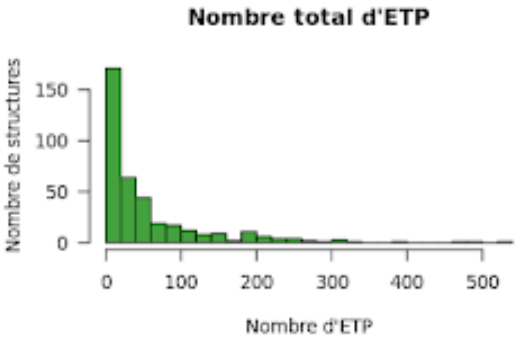
• A10. TAILLE APPROXIMATIVE DE LA STRUCTURE

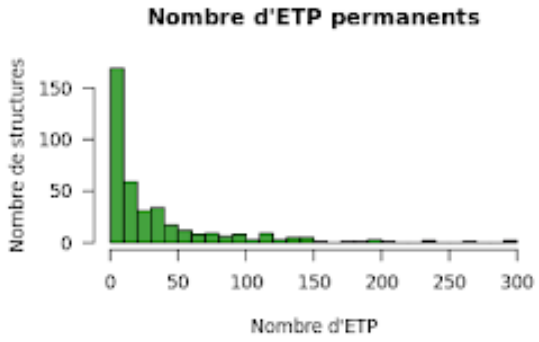
L'enquête a couvert des structures de tailles diverses

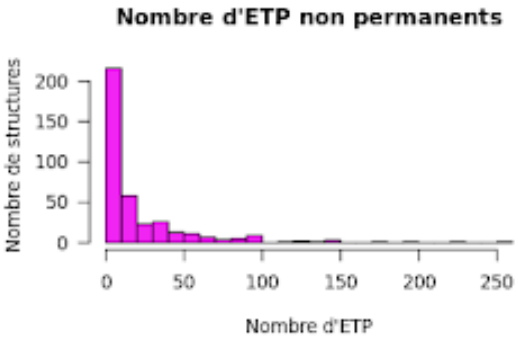


• A11. NOMBRE D'ETP

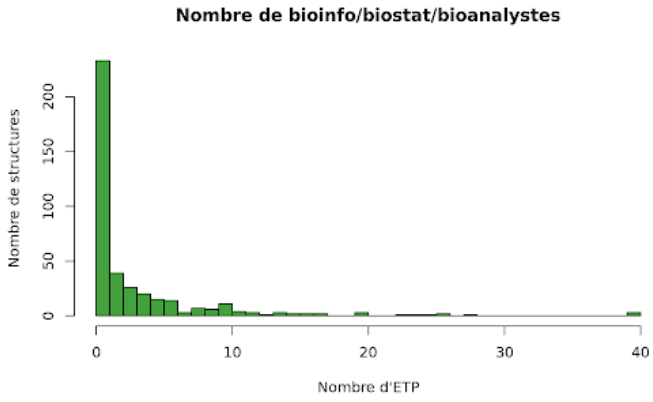


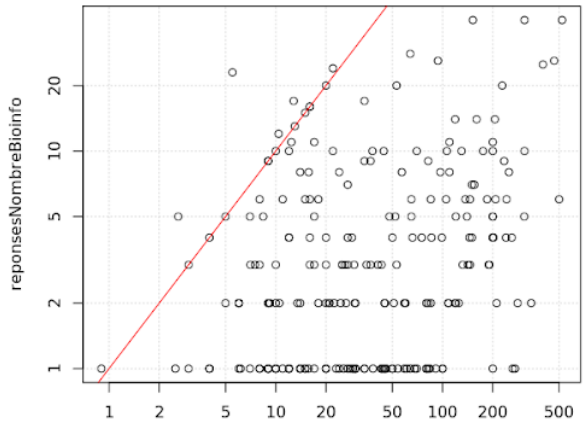




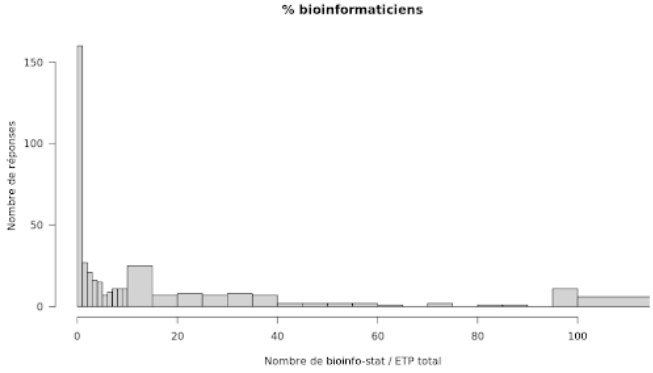


• A12. NOMBRE DE BIOINFORMATICIENS/BIOISTATISTICIENS/BIOANALYSTES DE VOTRE STRUCTURE



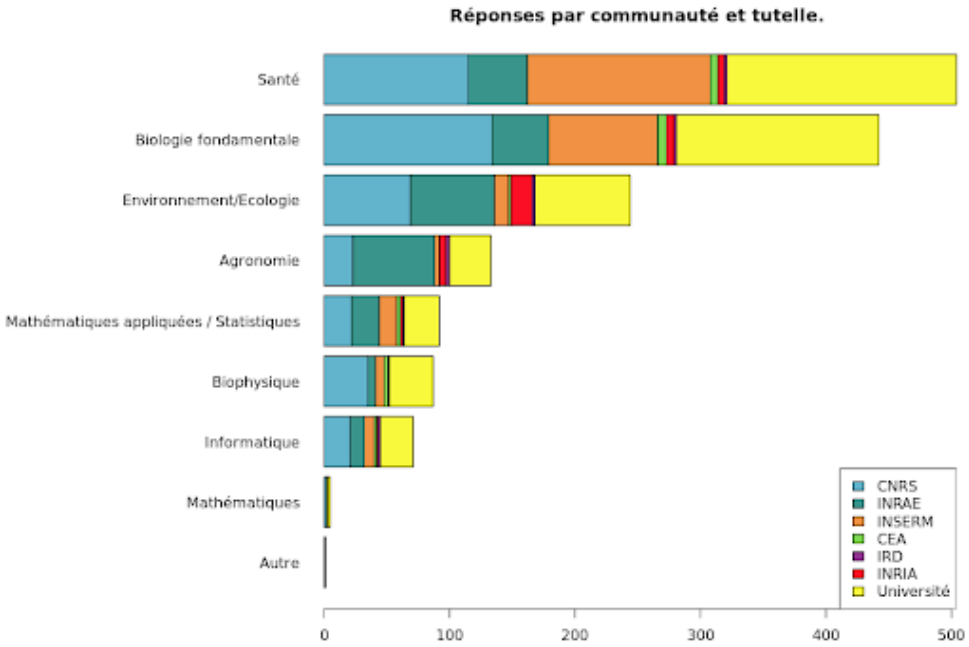


Dans quelques cas, le nombre de bioinformaticiens dépasse le nombre total d'ETP de la structure. Une interprétation possible serait que des membres d'équipes associées à ces structures y collaborent une partie de leur temps.



PARTIE B. THÉMATIQUE(S) DE RECHERCHE ET/OU DE SERVICE DE L'UNITÉ/ÉQUIPE

• B3. A QUELLE COMMUNAUTÉ SE RATTACHE VOTRE UNITÉ/ÉQUIPE?



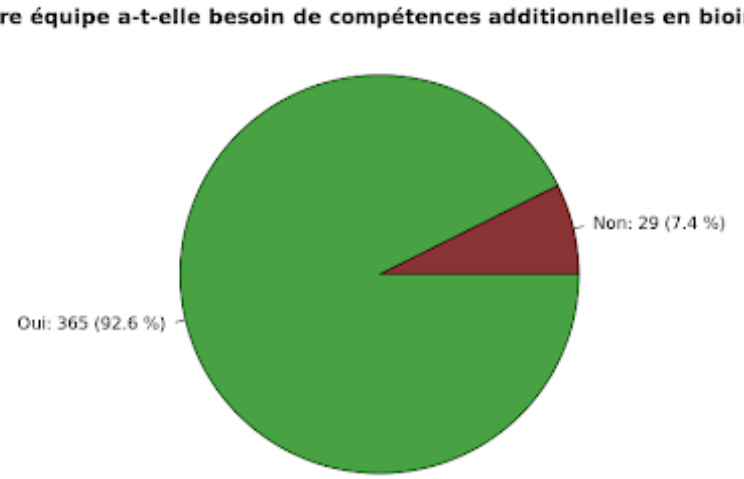
Les **communautés** les plus représentées dans les réponses sont celles de **Santé, de Biologie fondamentale et d'Environnement/Ecologie.**

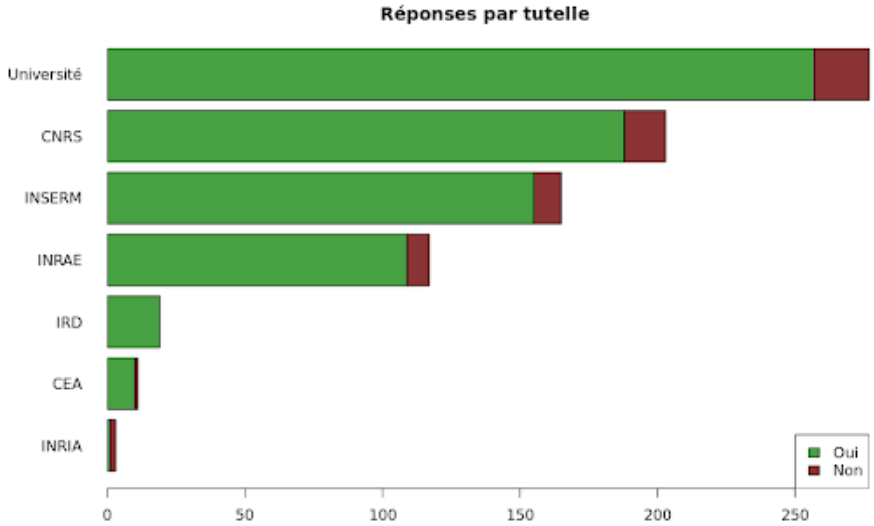
PARTIE C. BESOINS DE L'UNITÉ/ÉQUIPE EN COMPÉTENCES BIOINFORMATIQUES/BIOSTATISTIQUES/BIOANALYSES

Les besoins en compétences pourront être traités par différentes actions:

- Formation
- Recrutement
- Prestation de service des plateformes
- Collaborations avec des équipes de bioinformatique

• C1. VOTRE UNITÉ/ÉQUIPE ÉPROUVE-T-ELLE LE BESOIN DE COMPÉTENCES ADDITIONNELLES EN BIOINFORMATIQUE?

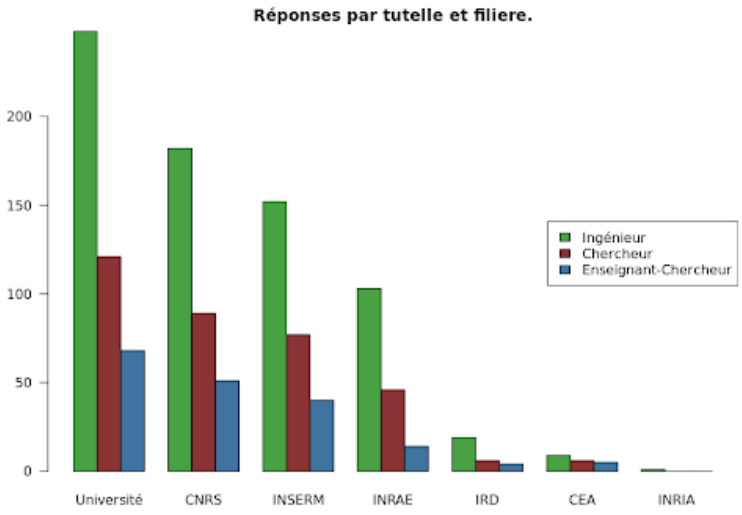




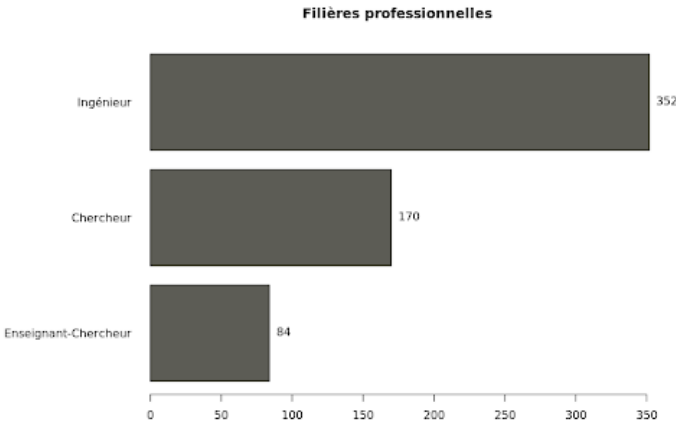
- Oui pour toutes les tutelles. Une écrasante majorité des réponses (93%) est positive concernant les besoins en compétences.

Clé de compréhension: les réponses “non” proviennent de 3 plateformes, 16 équipes, 9 unités et 1 individu.

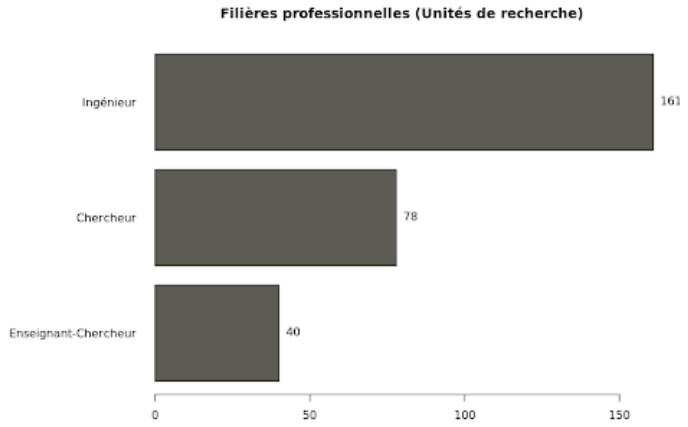
• C2. SI VOUS AVEZ RÉPONDU POSITIVEMENT À LA QUESTION PRÉCÉDENTE, PRÉCISEZ LE(S) FILIÈRES(S) PROFESSIONELLE(S) ATTENDUE(S)



- **Dans toutes les tutelles, le premier besoin est en ingénieurs.**
- De façon inattendue, les rapports entre ces 3 types de métiers sont les mêmes pour toutes les tutelles (on va écarter la réponse unique d'une équipe INRIA, qui n'est pas représentative de l'ensemble de l'EPST).
- Si on regroupe chercheurs et enseignants-chercheurs (qui font de la recherche) on constate que **les unités ne limitent pas la bioinformatique à l'aspect service/support.**

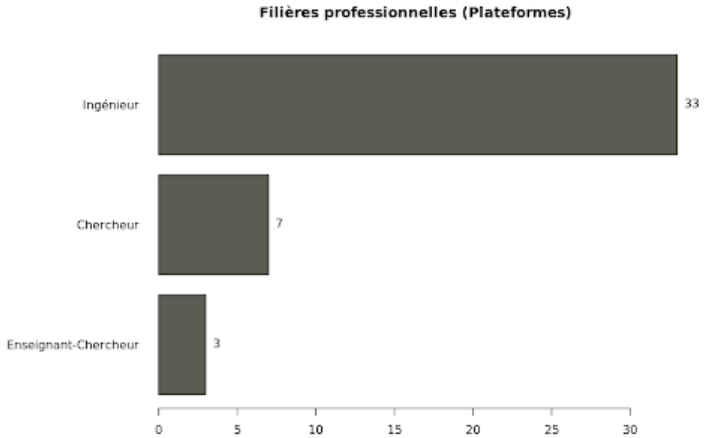


Pour toutes les réponses

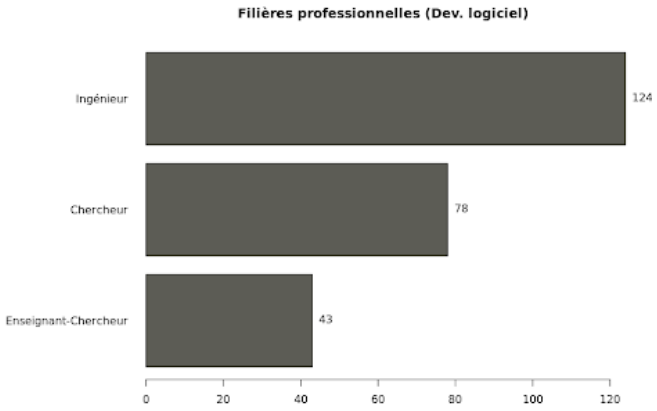


Pour les unités seulement

Pour les équipes seulement



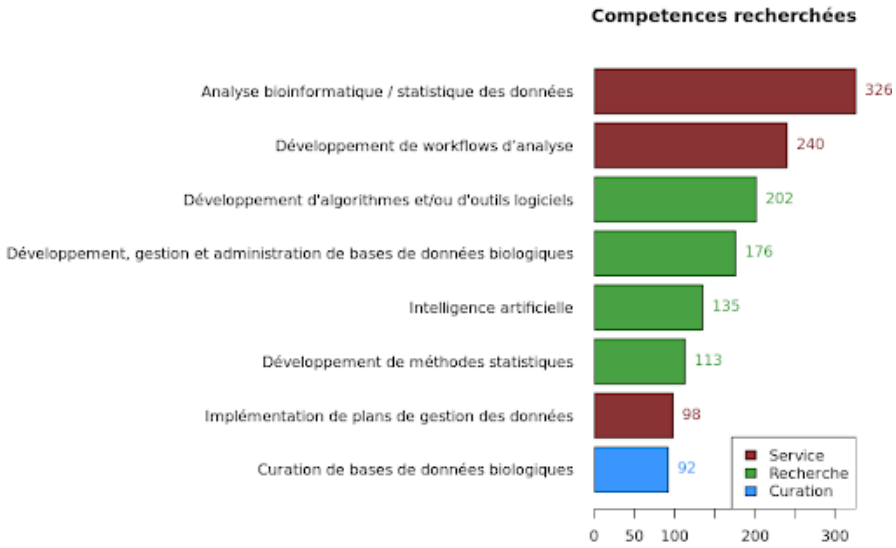
Pour les plateformes seulement

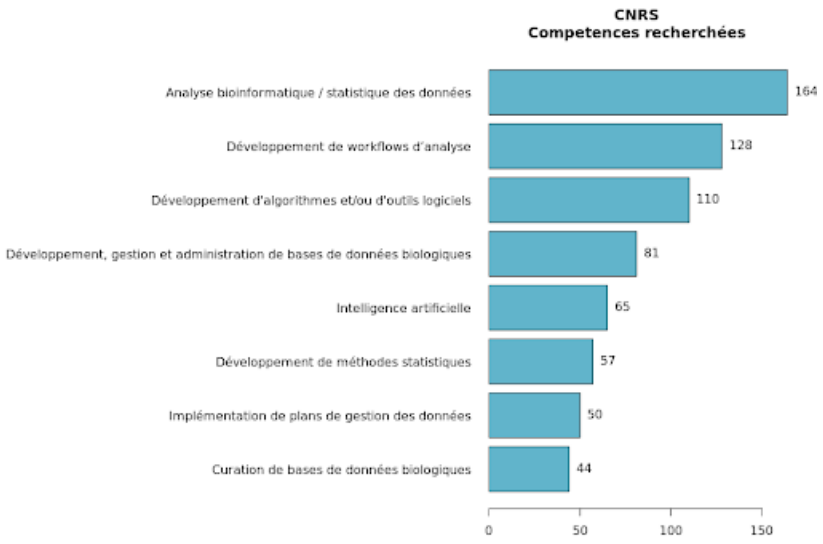


Structures qui développent des logiciels

Structures qui ne développent pas des logiciels

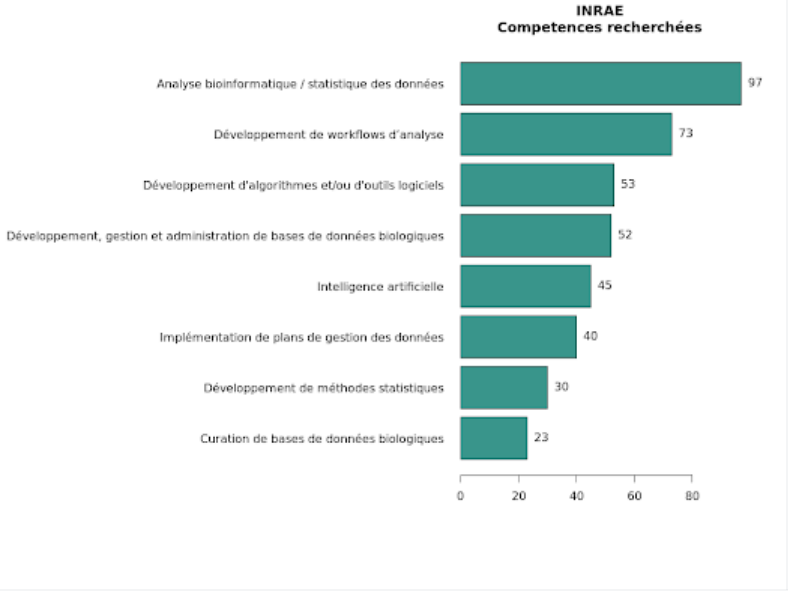
• C3. QUELLES SERAIENT LES COMPÉTENCES RECHERCHÉES ?

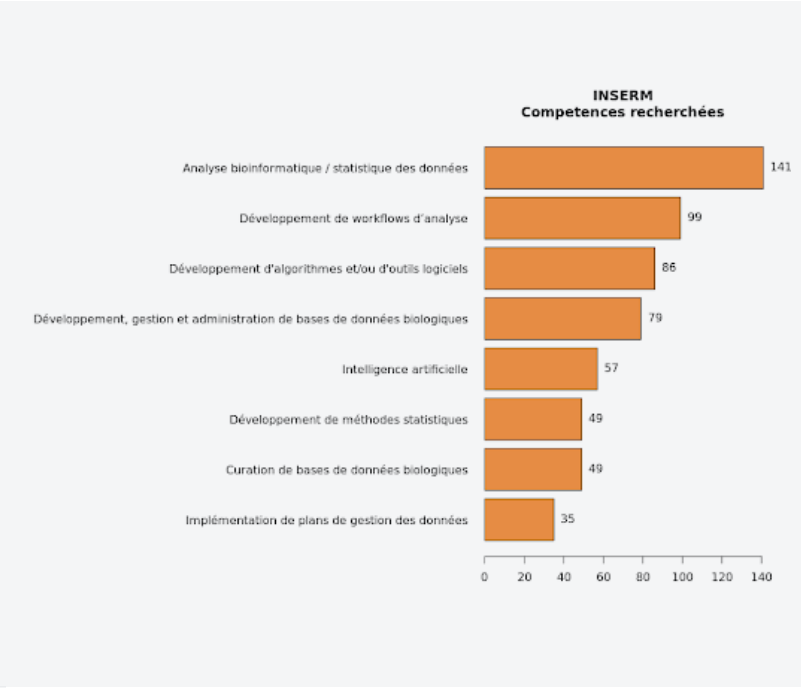


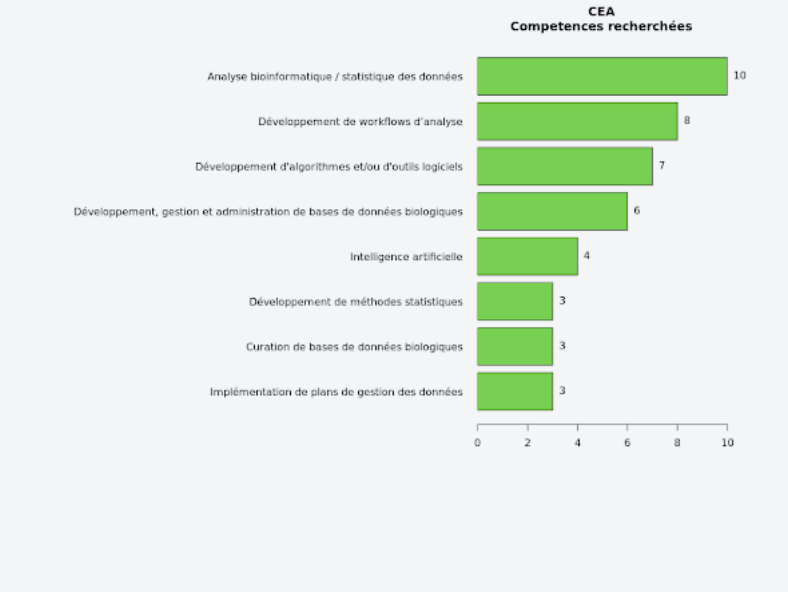


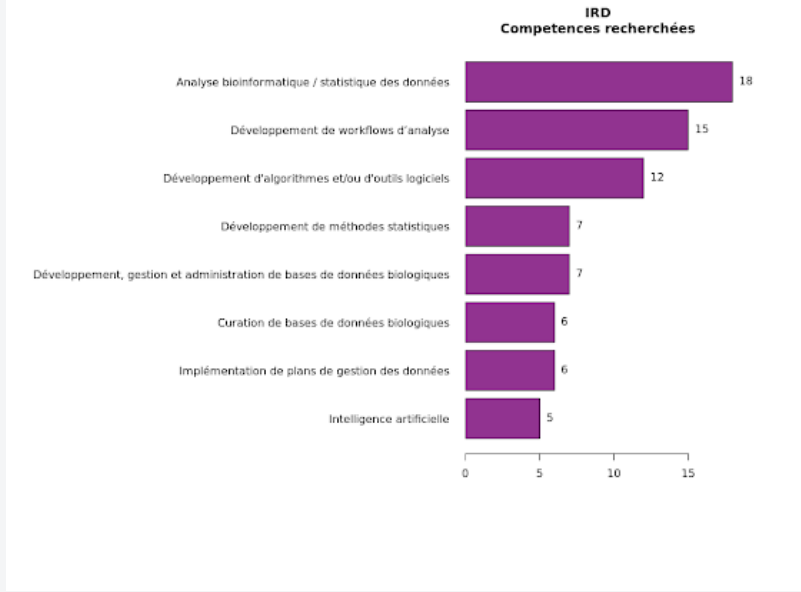
- **Premier besoin des données = analyse bioinformatique / statistique des données**, sachant que la rubrique “workflows” correspond en grande partie aux mêmes objectifs (workflows d’analyse de données).

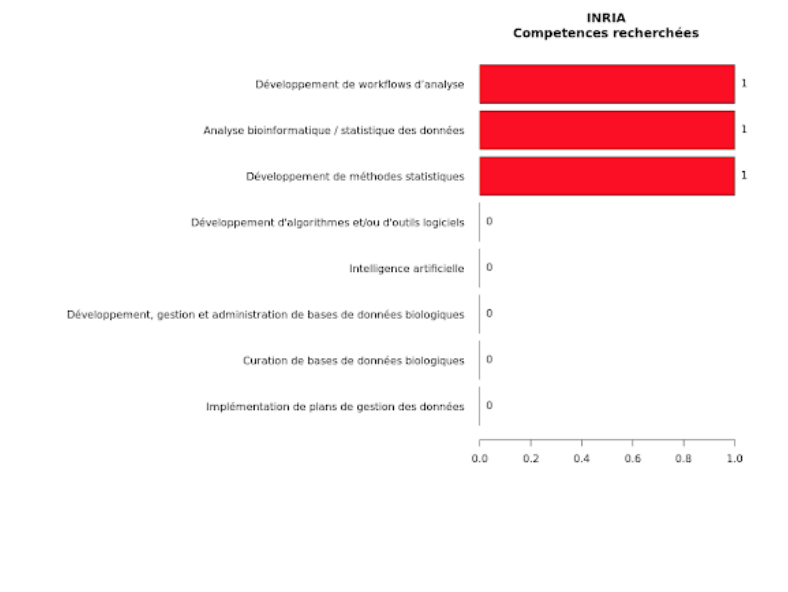
- La demande en développement d’algorithmes et/ou outils logiciels couvre la moitié des réponses (202/407).

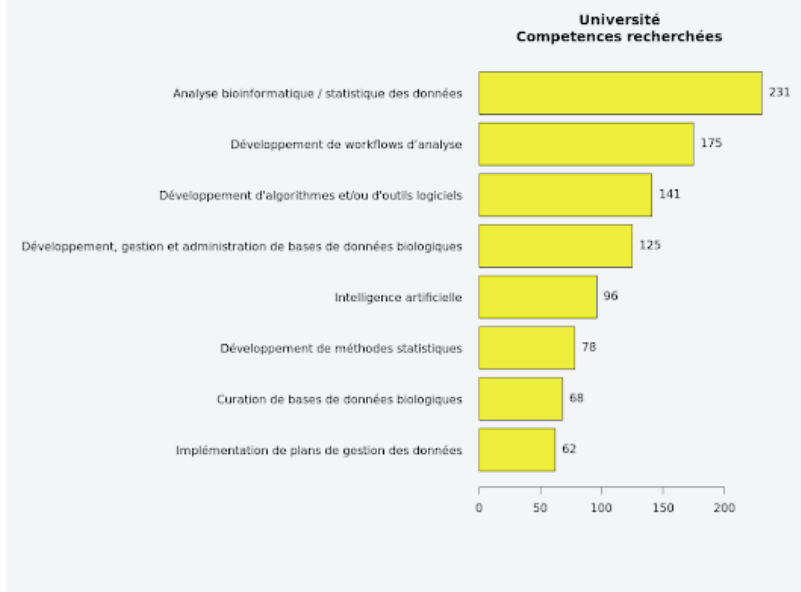

INRAE
Competences recherchées
Analyse bioinformatique / statistique des données
97
Développement de workflows d'analyse
73
Développement d'algorithmes et/ou d'outils logiciels
53
Développement, gestion et administration de bases de données biologiques
52
Intelligence artificielle
45
Implémentation de plans de gestion des données
40
Développement de méthodes statistiques
30
Curation de bases de données biologiques
23
0
20
40
60
80


INSERM
Competences recherchées
Analyse bioinformatique / statistique des données
141
Développement de workflows d'analyse
99
Développement d'algorithmes et/ou d'outils logiciels
86
Développement, gestion et administration de bases de données biologiques
79
Intelligence artificielle
57
Développement de méthodes statistiques
49
Curation de bases de données biologiques
49
Implémentation de plans de gestion des données
35
0
20
40
60
80
100
120
140

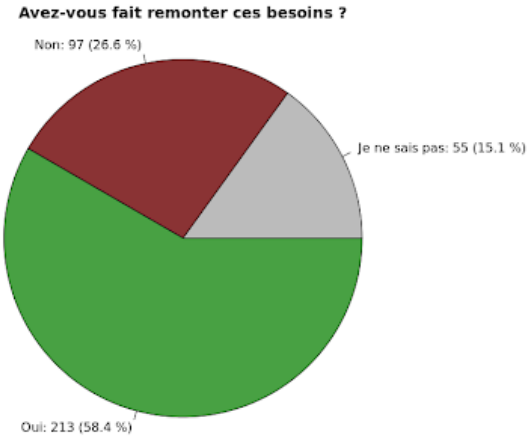

CEA
Competences recherchées
Analyse bioinformatique / statistique des données
10
Développement de workflows d'analyse
8
Développement d'algorithmes et/ou d'outils logiciels
7
Développement, gestion et administration de bases de données biologiques
6
Intelligence artificielle
4
Développement de méthodes statistiques
3
Curation de bases de données biologiques
3
Implémentation de plans de gestion des données
3
0
2
4
6
8
10

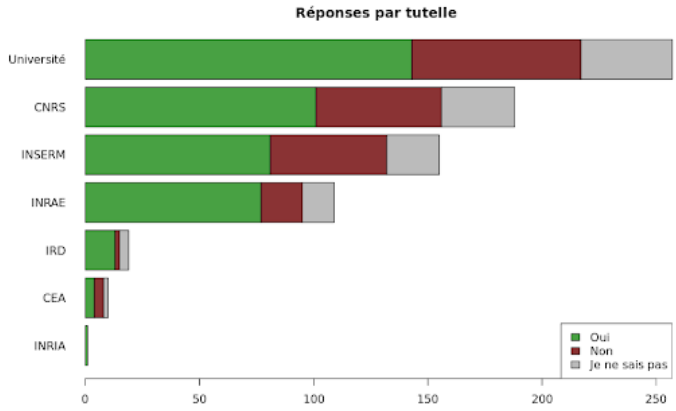

IRD
Competences recherchées
Analyse bioinformatique / statistique des données
18
Développement de workflows d'analyse
15
Développement d'algorithmes et/ou d'outils logiciels
12
Développement de méthodes statistiques
7
Développement, gestion et administration de bases de données biologiques
7
Curation de bases de données biologiques
6
Implémentation de plans de gestion des données
6
Intelligence artificielle
5
0
5
10
15


INRIA
Competences recherchées
Développement de workflows d'analyse
1
Analyse bioinformatique / statistique des données
1
Développement de méthodes statistiques
1
Développement d'algorithmes et/ou d'outils logiciels
0
Intelligence artificielle
0
Développement, gestion et administration de bases de données biologiques
0
Curation de bases de données biologiques
0
Implémentation de plans de gestion des données
0
0.0
0.2
0.4
0.6
0.8
1.0


Université
Competences recherchées
Analyse bioinformatique / statistique des données
231
Développement de workflows d'analyse
175
Développement d'algorithmes et/ou d'outils logiciels
141
Développement, gestion et administration de bases de données biologiques
125
Intelligence artificielle
96
Développement de méthodes statistiques
78
Curation de bases de données biologiques
68
Implémentation de plans de gestion des données
62
0
50
100
150
200

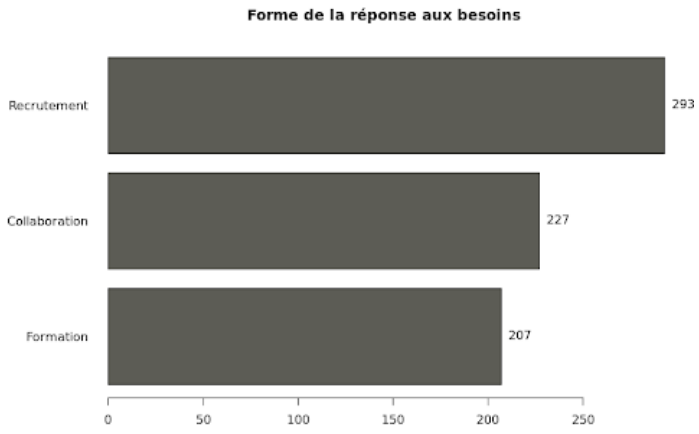
• C4. VOTRE UNITÉ/ÉQUIPE A-T-ELLE DÉJÀ FAIT REMONTER CES BESOINS?





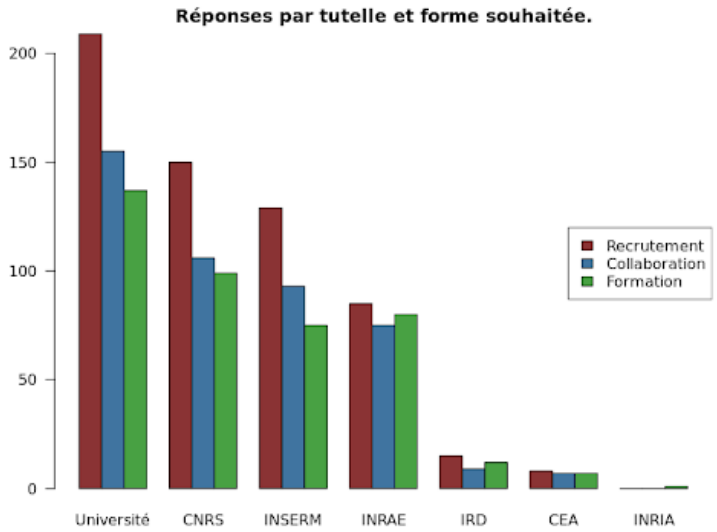
- **Les besoins dépassent les remontées (qui dépassent sans doute largement les recrutements).**
- Les proportions de remontées sont assez semblables entre tutelles.

• C5. SOUS QUELLE FORME VOTRE UNITÉ PENSE-T-ELLE POUVOIR RÉPONDRE À CES BESOINS?



- **Les stratégies sont équilibrées entre recrutement, collaboration et formation.**

- Ceci reflète peut-être une prise de conscience que le recrutement de bioinformaticien au sein de l'équipe n'est pas forcément la panacée, car on a besoin de compétences de plus en plus pointues, qu'il vaut mieux trouver dans d'autres équipes que de recruter dans chaque équipe.

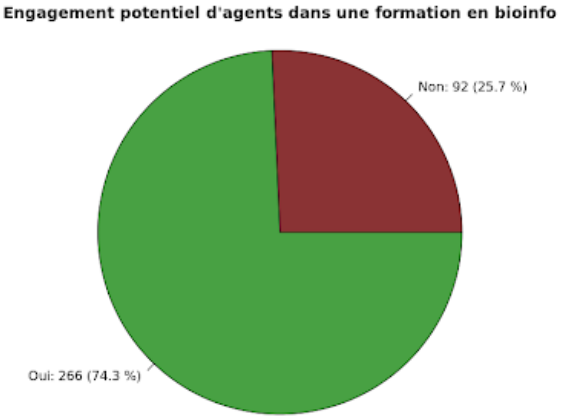


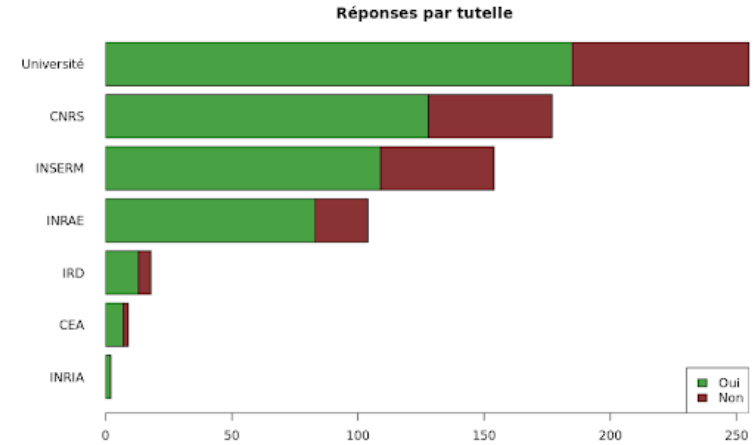
- On observe qu'**à l'INRAe la formation est encore plus mise en avant que dans les autres tutelles.** Ceci reflète sans doute la politique délibérée de l'INRAe. C'est un message assez fort que l'INRAe a fait passer, on c**onstate une évolution de la mentalité (ou en tout cas de la perception des stratégies raisonnables).**

- Les réponses "collaborations" mettent en lumière la volonté des répondant de travailler avec des unités des autres tutelles /instituts ou éventuellement des sociétés externes.

• C6. VOTRE UNITÉ/ÉQUIPE INCLUT-ELLE DES AGENTS SUSCEPTIBLES DE S'ENGAGER DANS UNE FORMATION EN VUE D'ACQUISITION DE COMPÉTENCES COMPLÉMENTAIRES EN BIOINFORMATIQUE?

75% de oui, indépendamment des tutelles

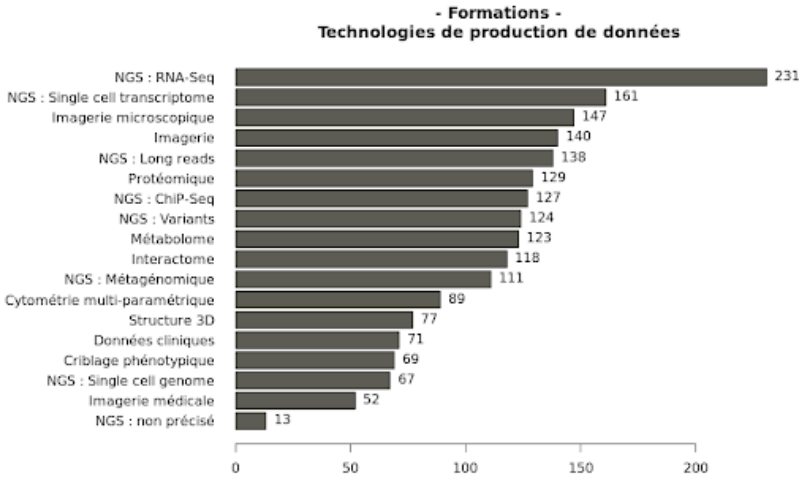


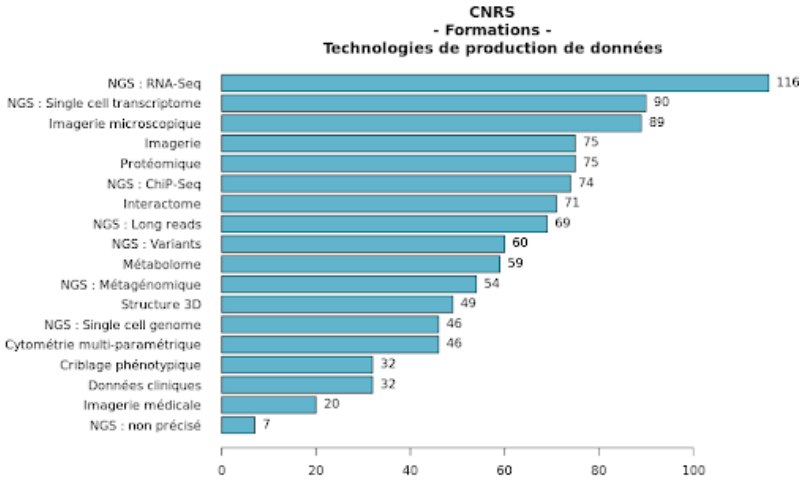


Les universités sont tout aussi demandeuses de formation continue que les organismes de recherche.

PARTIE D. BESOINS EN FORMATION

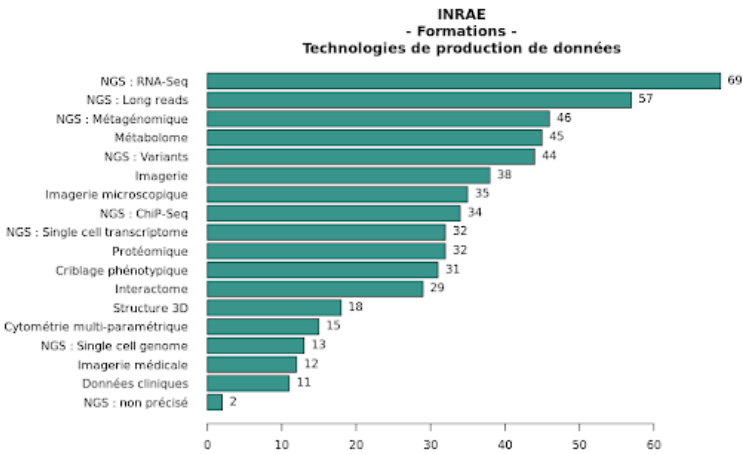
• D1. EN MATIÈRE DE FORMATIONS AUX TECHNOLOGIES DE PRODUCTION ET TYPES DE DONNÉES, QUELS SONT VOS BESOINS?

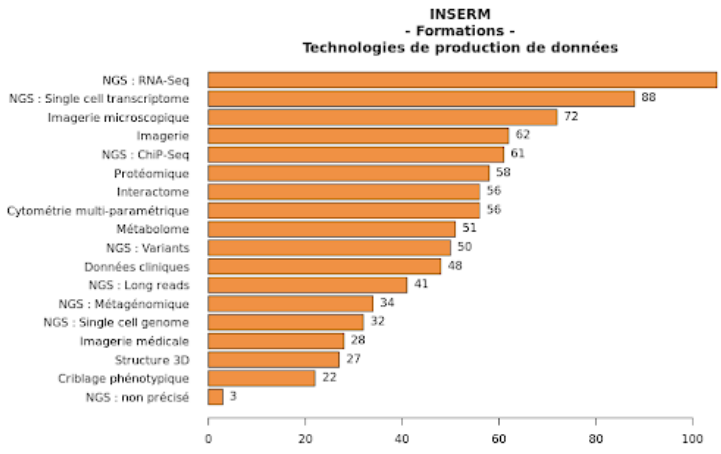


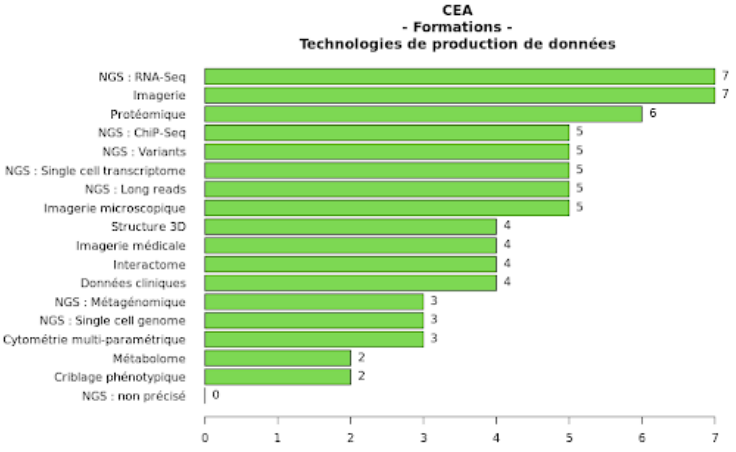


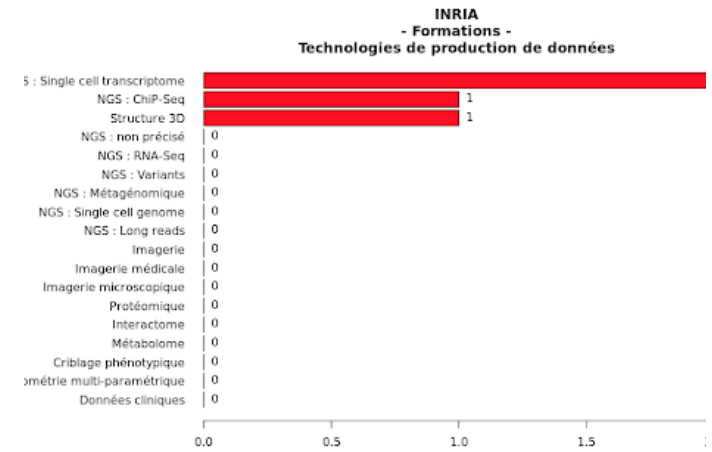
- **Le RNA-seq reste la technologie la plus demandée en matière de formation.** Ceci est vrai depuis plusieurs années. **Le single cell-transcriptomique est aussi très demandé.**

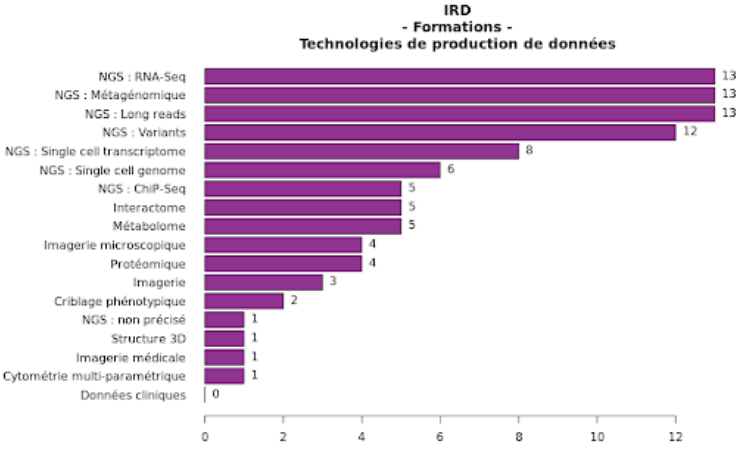
- **Grosse demande en imagerie médicale et microscopique**. Ceci correspond également aux tendances des projets récents en biologie intégrative. Il faudrait mettre en place des formations en collaboration entre IFB et les deux infra d'imagerie. Ceci pourra se faire notamment dans le contexte de la feuille de route IFB (projets -pilotes et PIA3, où nous avons une Implementation Study spécifiquement à l'intégration imagerie multi-omique).

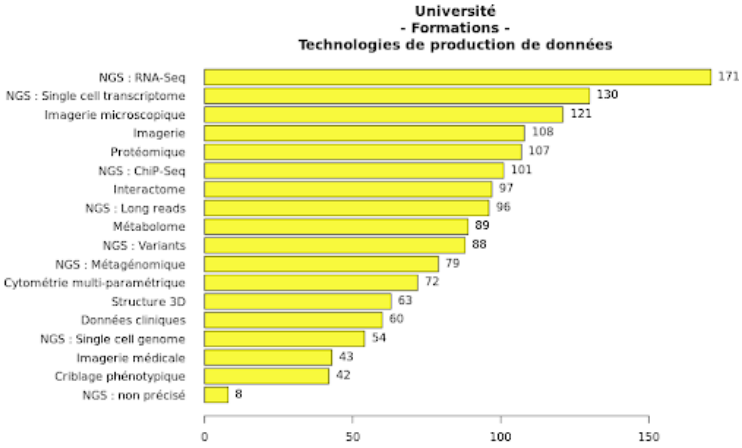




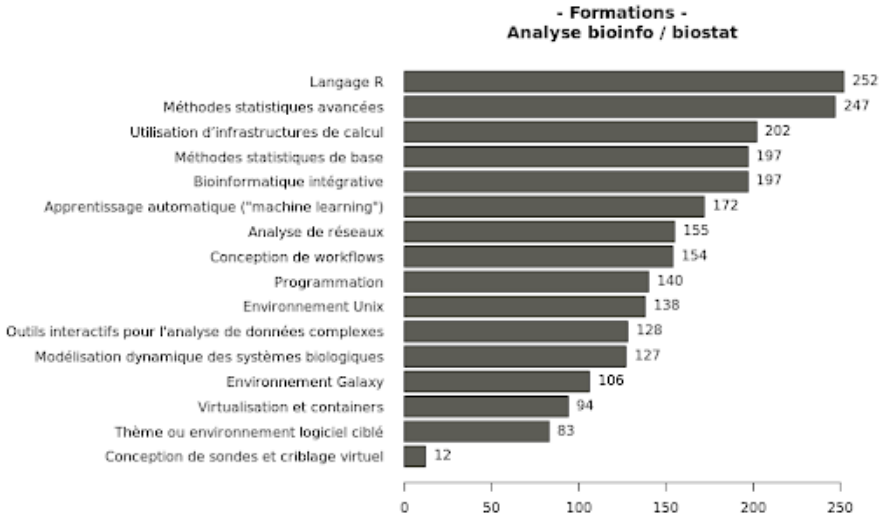






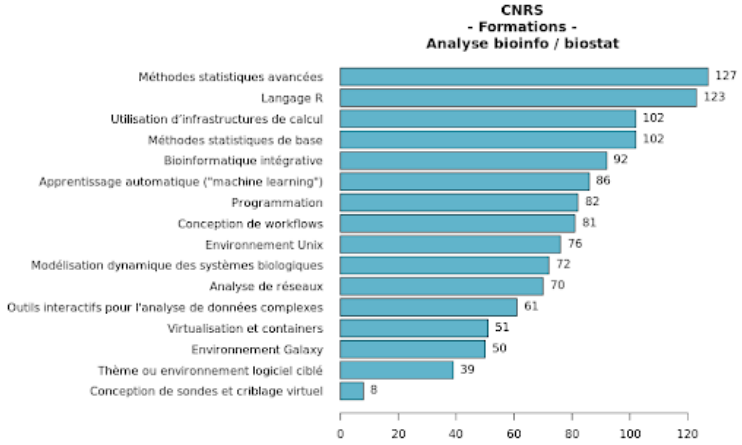


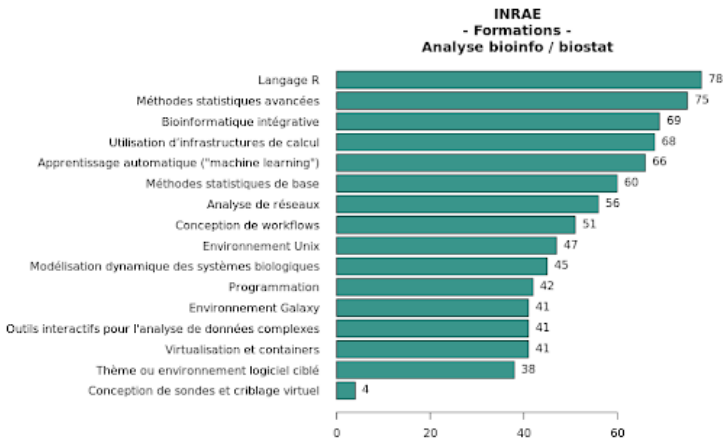
- **D2. EN MATIÈRE DE FORMATIONS À L'ANALYSE BIOINFORMATIQUE/BIOSTATISTIQUE, QUELS SONT VOS BESOINS?**

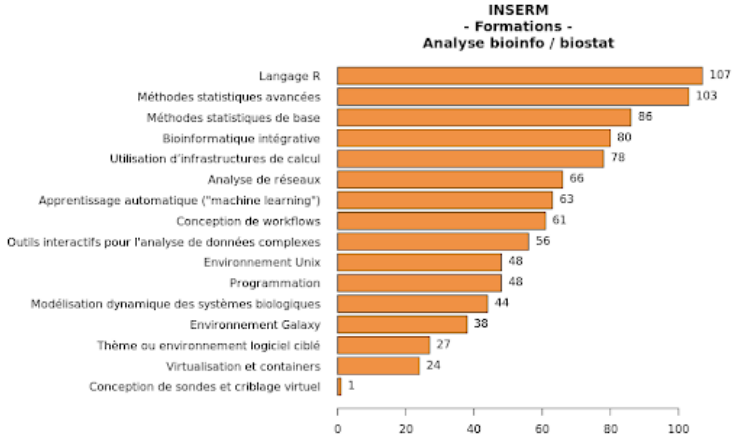


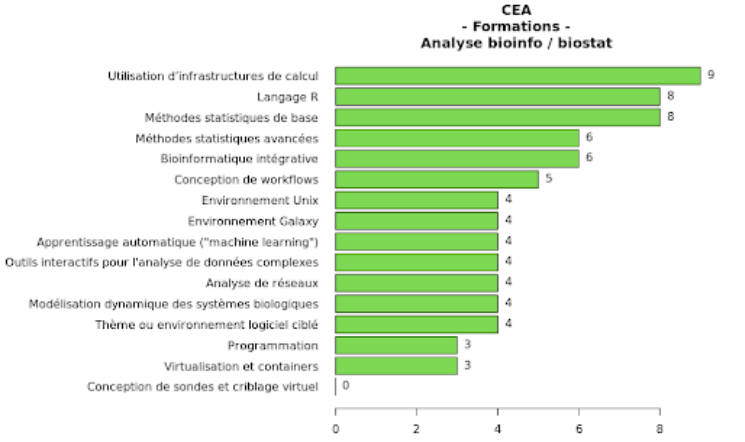
- Les réponses correspondent à notre enquête préliminaire auprès de l'Inserm.
- **Très forte demande R, méthodes statistiques de base et avancées, bioinformatique intégrative.**
- On note également une forte demande pour l'utilisation d'infrastructures de calcul.
- Les formations EBAII et DUBii apportent des réponses à la plupart de ces besoins de formation en statistique, analyse des réseaux, bioinformatique intégrative, etc.

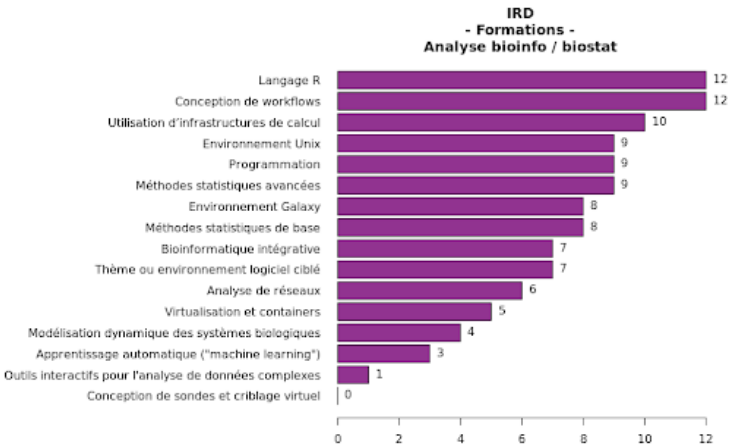
- Plusieurs formations organisées par les plateformes de l'IFB utilisent l'environnement Galaxy pour former les biologistes.
- Il subsistait un manque en formation concernant l'apprentissage automatique, mais une formation sur ce thème est à présent mise en place ("Introduction to Machine Learning using R").

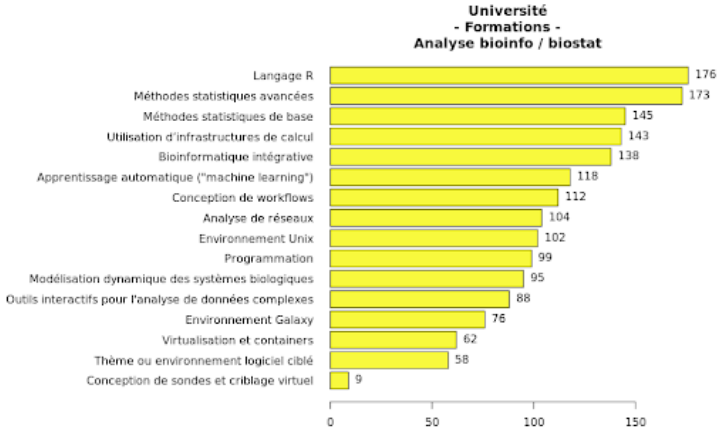


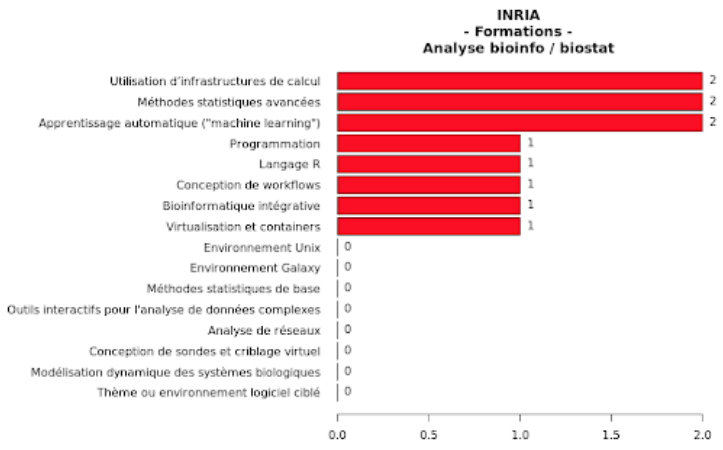




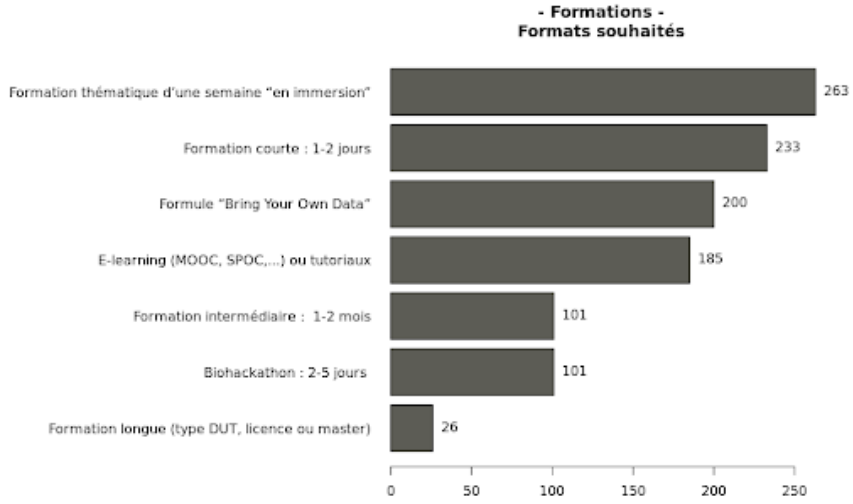




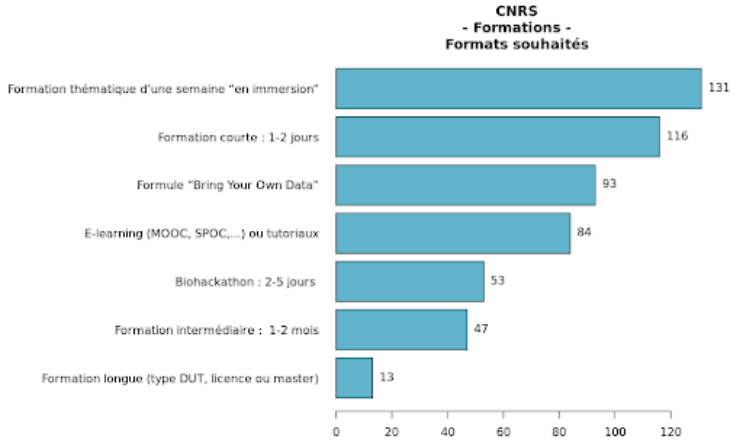


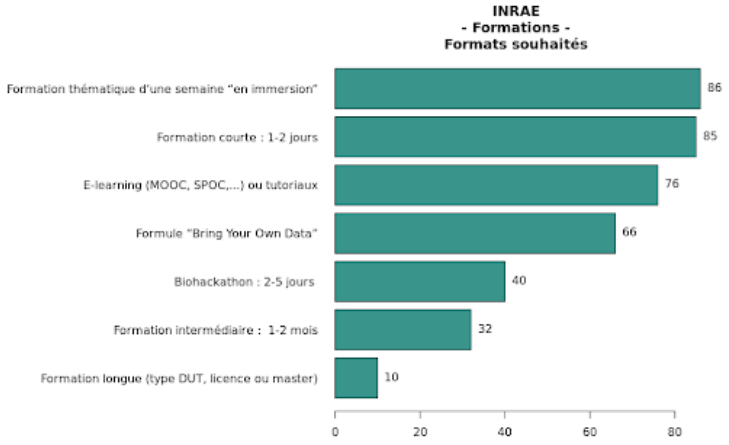


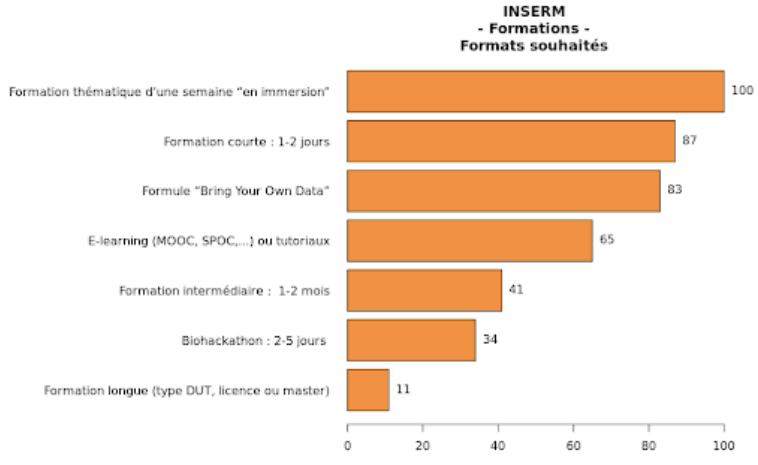
• D3. QUELS FORMATS SOUHAITERIEZ-VOUS POUR LES MEMBRES CONCERNÉS DE VOTRE ÉQUIPE/UNITÉ?

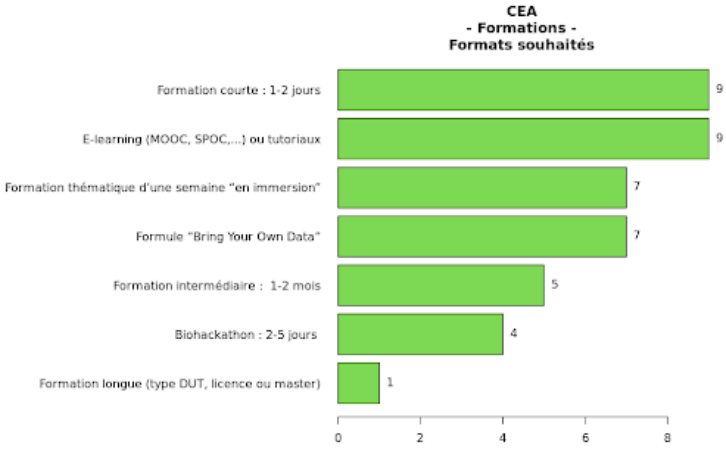


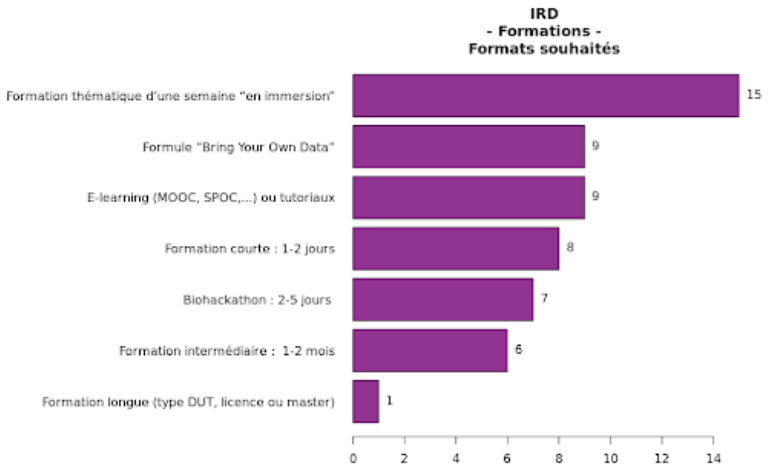
- **Forte demande pour la formation en immersion**
- **Fort intérêt pour le format "Bring Your Own Data"**
- **Forte demande de e-learning**

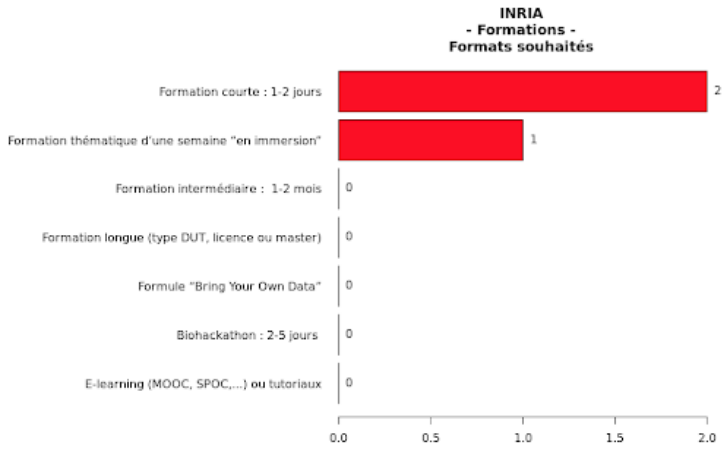


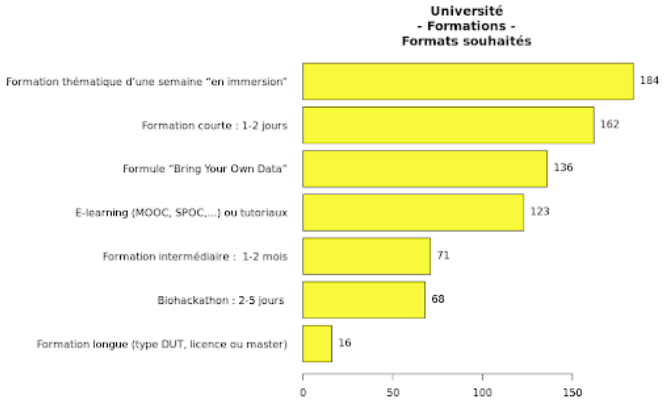




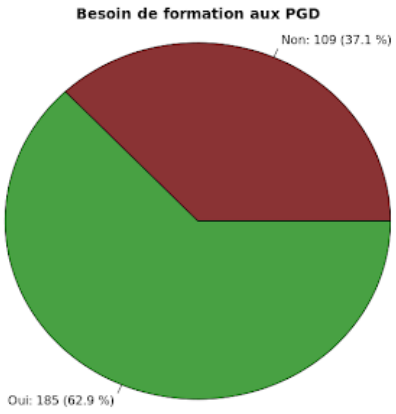


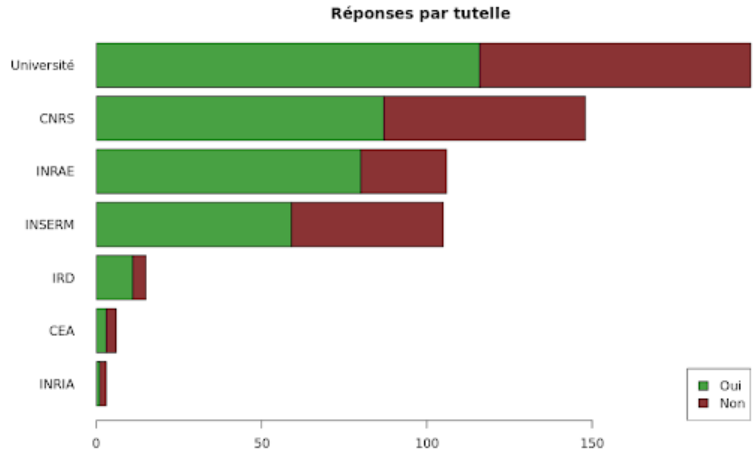






• D4. EPROUVEZ-VOUS DES BESOINS DE FORMATION À LA GESTION DES PLANS DE DONNÉES (DMP)?



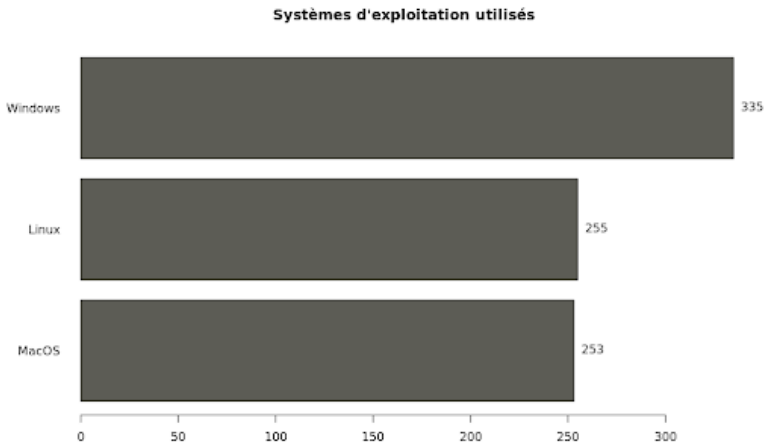


- **Demande importante en formations pour les Plan de Gestion de Données (PGD)**. L'IFB a précisément mis en place une offre nationale. ELIXIR développe également des formations (ELIXIR-CONVERGE).
- **Les instituts comme l'INRAE qui ont déjà adopté les PGD** (voir questions G.1 et G.2) **ont la plus forte demande de formation.** Mais globalement la demande est forte partout.

- Les réponses négatives concernent peut-être des équipes / unités qui :
  a) soit sont déjà formées à la rédaction de PGD
  b) soit n'ont pas encore assez d'information des enjeux autour de la **Science Ouverte**
  c) soit on une attitudes de résistance à cette nouvelle pratique

PARTIE E. RESSOURCES LOGICIELLES (OUTILS ET BASES DE DONNÉES) UTILISÉES PAR VOTRE UNITÉ/ÉQUIPE

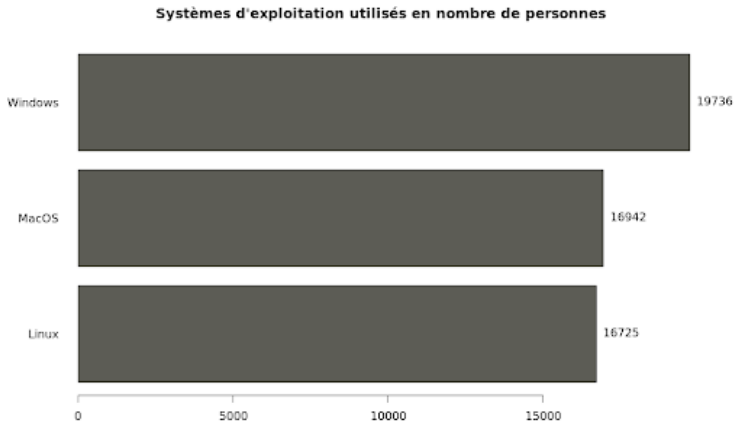
• E1. QUELS SONT LES SYSTÈMES D'EXPLOITATION UTILISÉS?

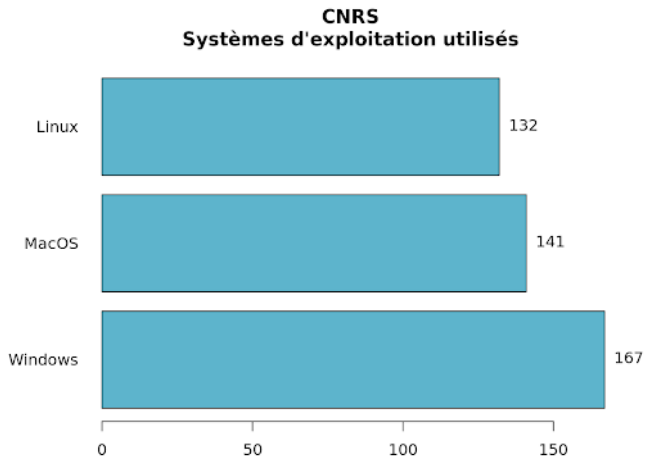


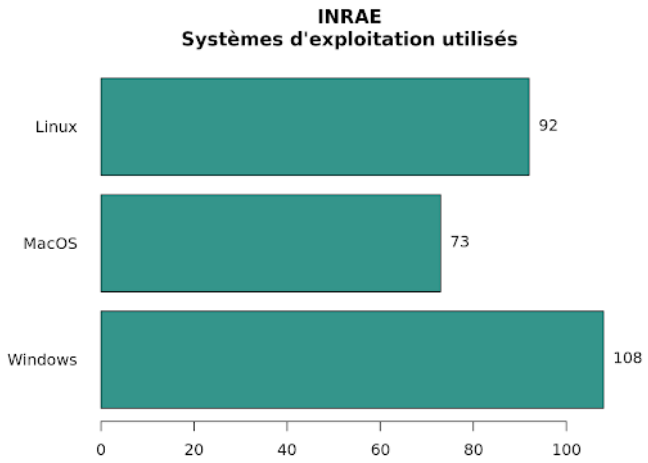
- Si on fait **la somme Linux et Mac OS, on constate qu'ils sont plus utilisés que Windows dans notre communauté.**

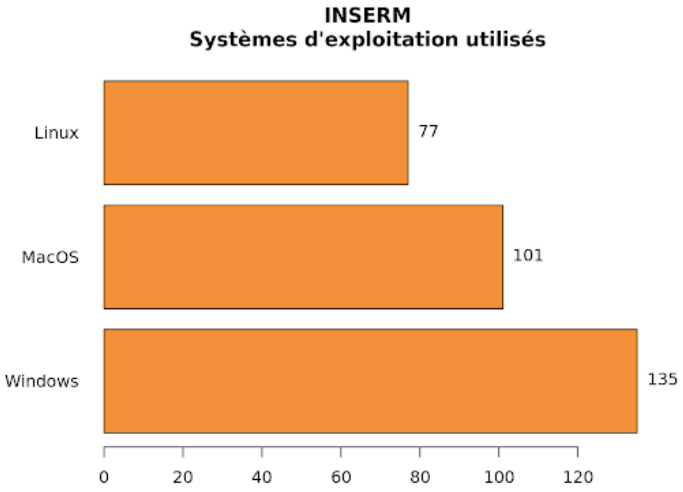
- A nuancer que tous les utilisateurs de Mac n'utilisent pas forcément Linux.

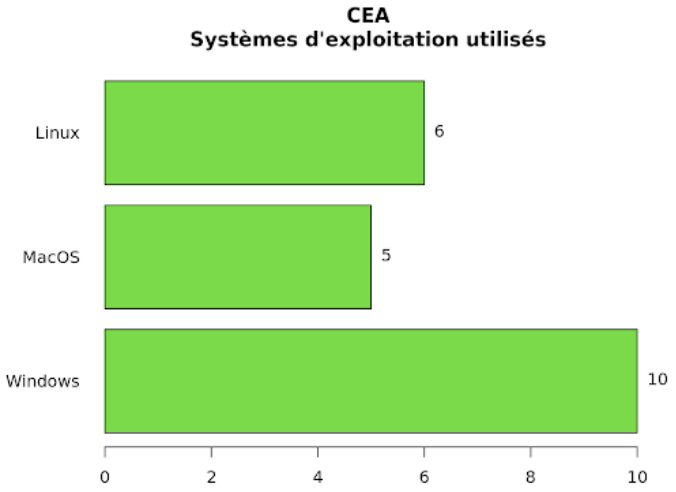
- En nombre de personnes: ça ne bouleverse pas drastiquement le profil général

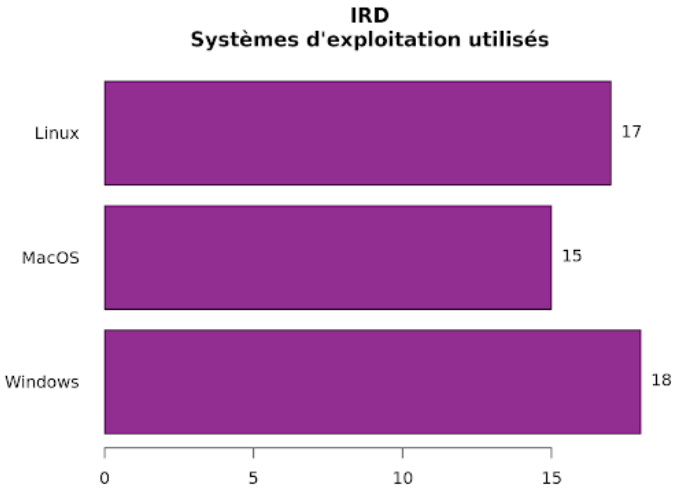


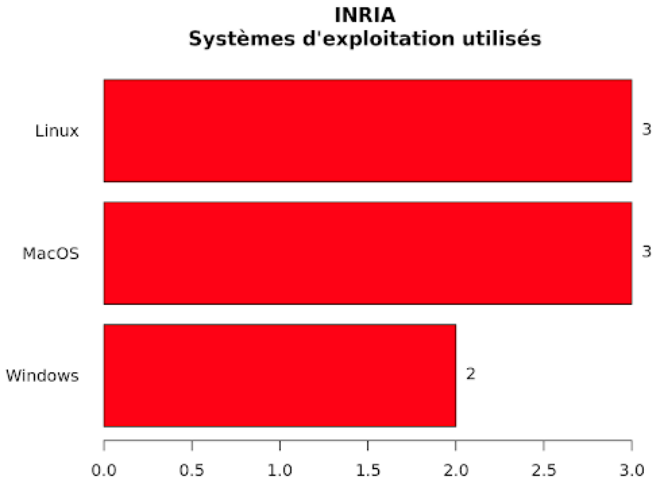


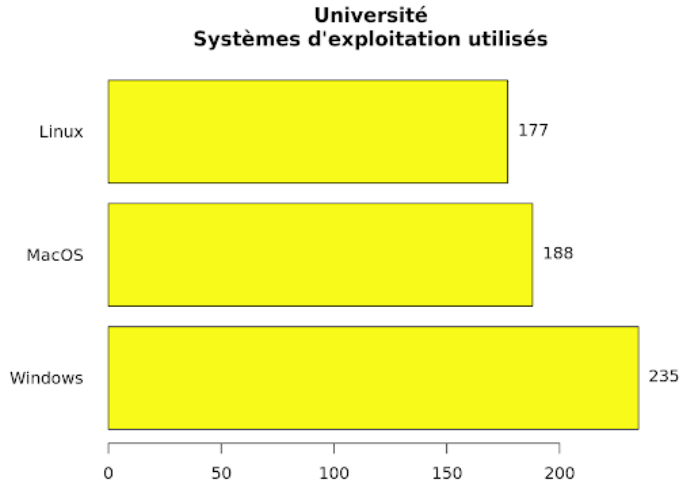


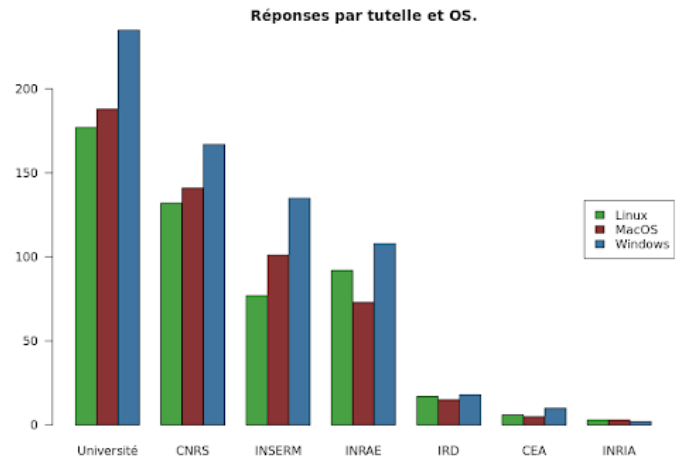

INSERM
Systèmes d'exploitation utilisés
Linux
MacOS
Windows
77
101
135
0
20
40
60
80
100
120


CEA
Systèmes d'exploitation utilisés
Linux
MacOS
Windows
6
5
10
0
2
4
6
8
10

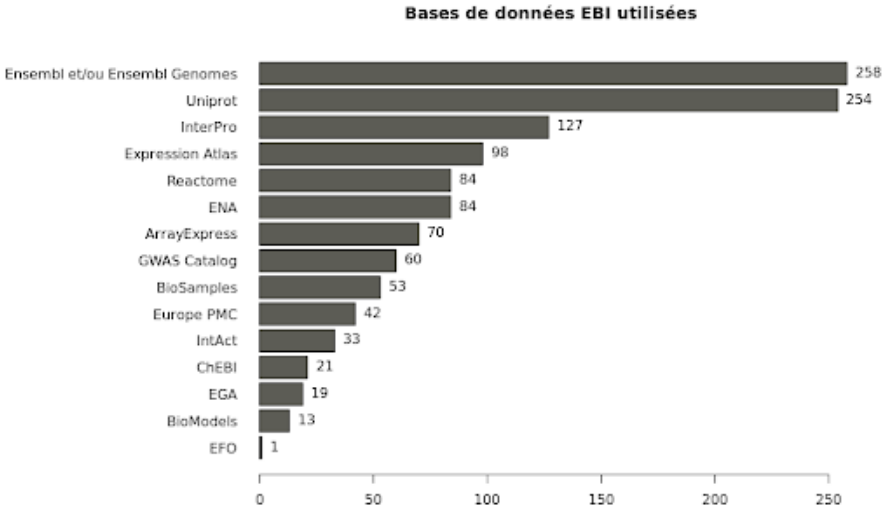

IRD
Systèmes d'exploitation utilisés
Linux
MacOS
Windows
17
15
18
0
5
10
15


INRIA
Systèmes d'exploitation utilisés
Linux
MacOS
Windows
3
3
2
0.0
0.5
1.0
1.5
2.0
2.5
3.0

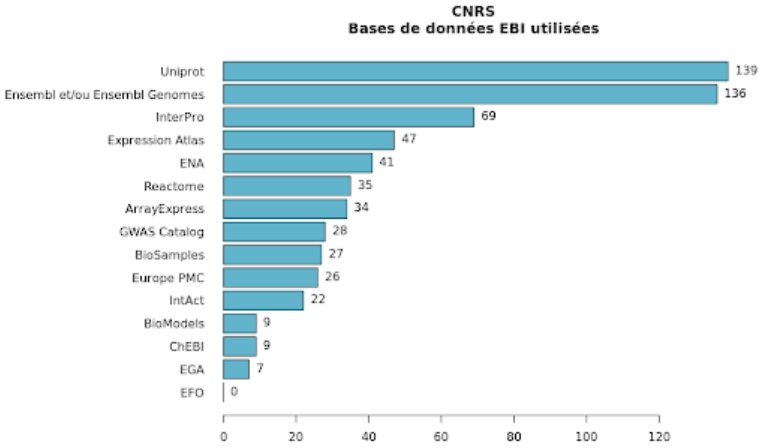

Université
Systèmes d'exploitation utilisés
Linux
MacOS
Windows
177
188
235
0
50
100
150
200

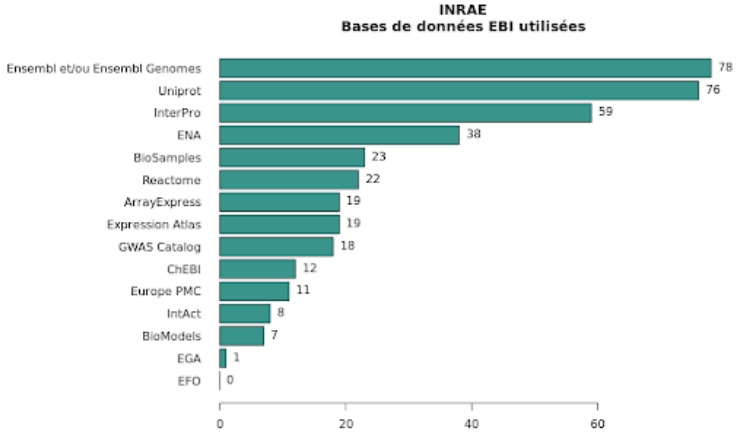

Réponses par tutelle et OS.
200
150
100
50
0
Linux
MacOS
Windows
Université
CNRS
INSERM
INRAE
IRD
CEA
INRIA

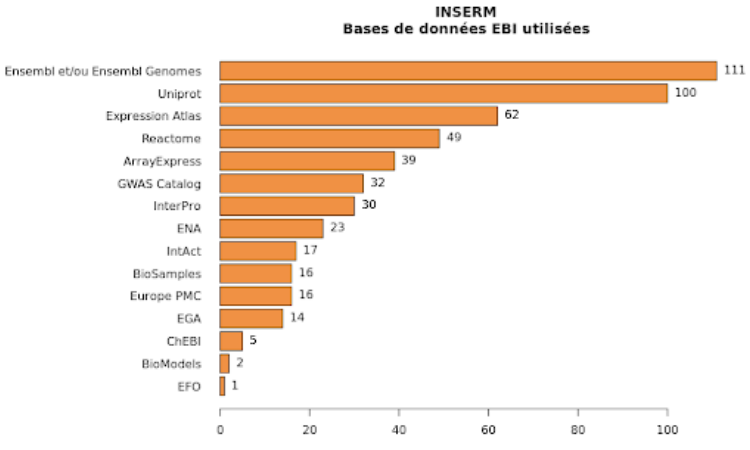
• E2. UTILISEZ-VOUS DES BASES DE DONNÉES DE L'EBI?

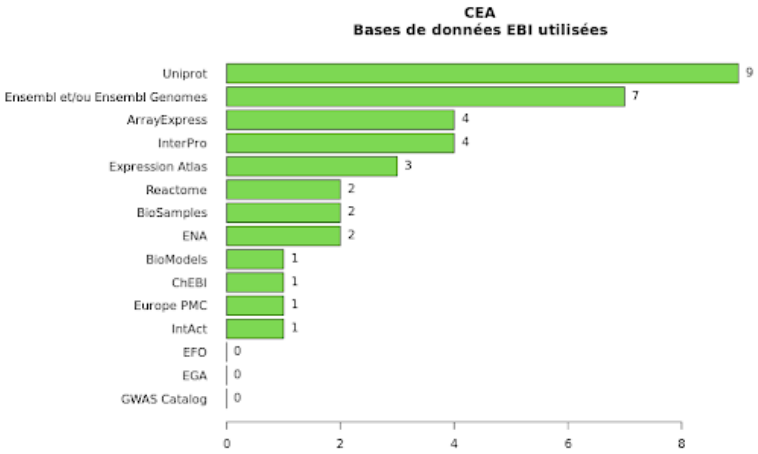


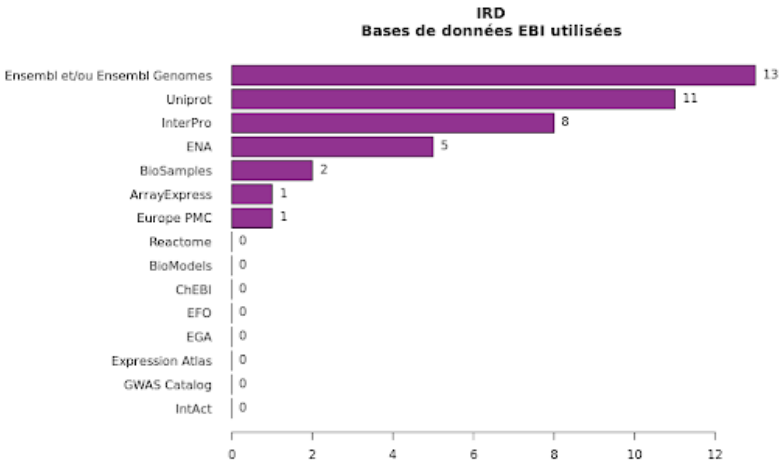
Forte utilisation des ressources de l'EBI d'une manière générale, en particulier UniProt, Ensembl et Interpro. En revanche, apparente méconnaissance de la ressource bibliographique Europe PMC.

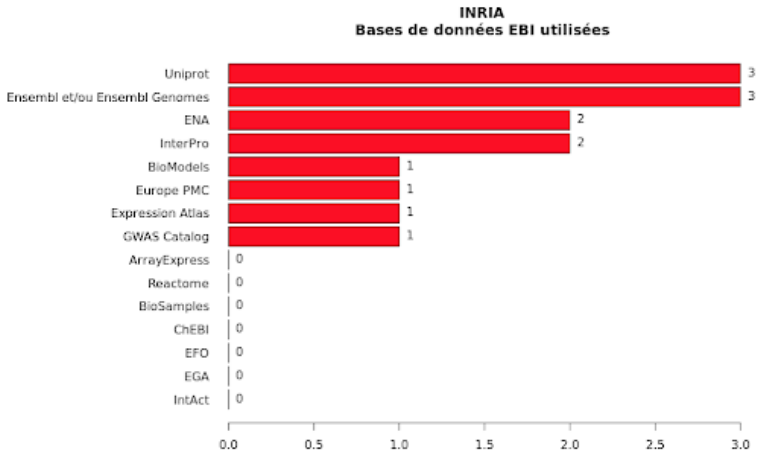


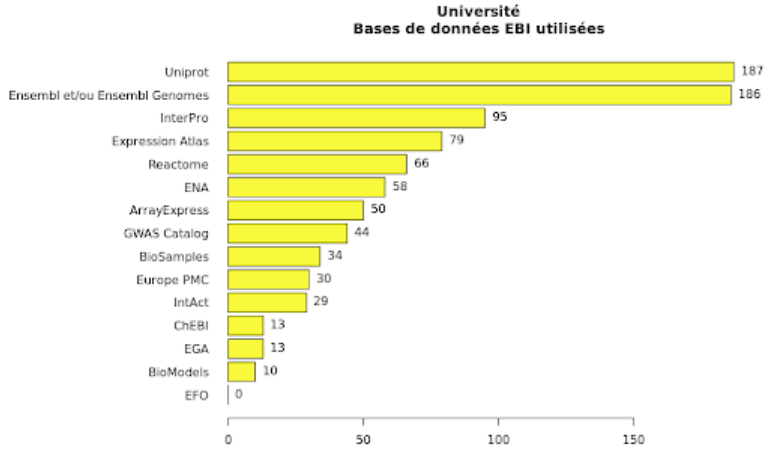




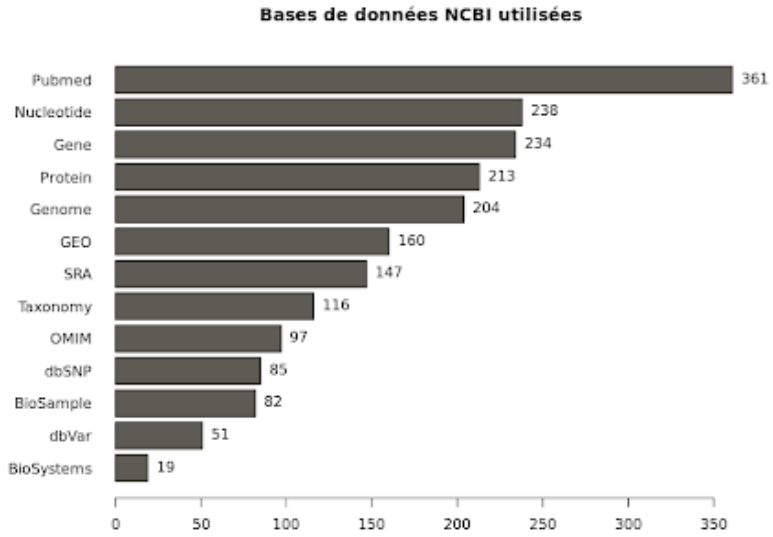






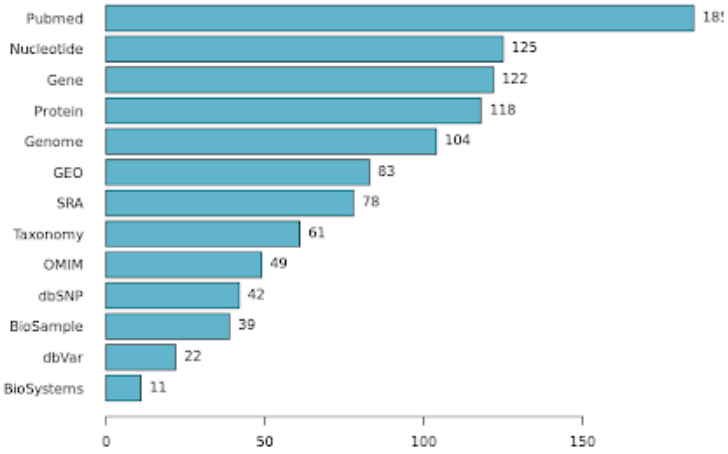


- **E3. UTILISEZ-VOUS DES BASES DE DONNÉES DU NCBI?**

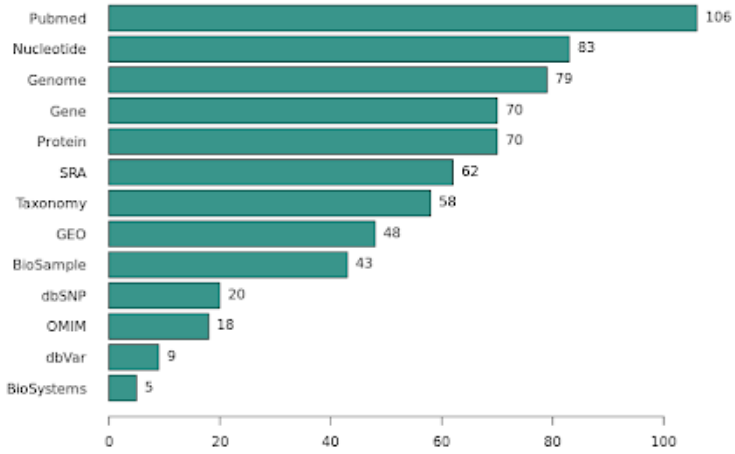


On note l'importance de Pubmed dans les ressources NCBI

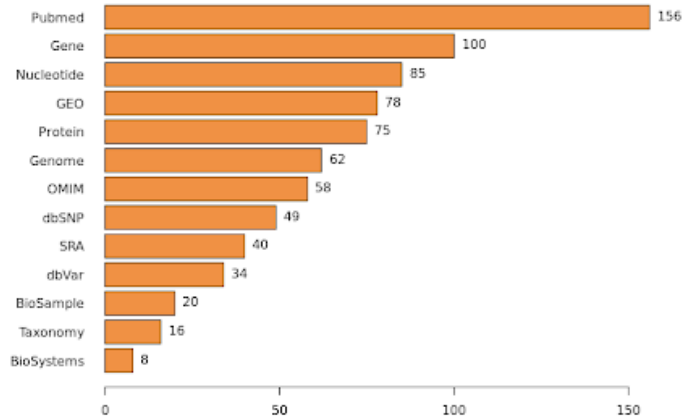
CNRS
Bases de données NCBI utilisées



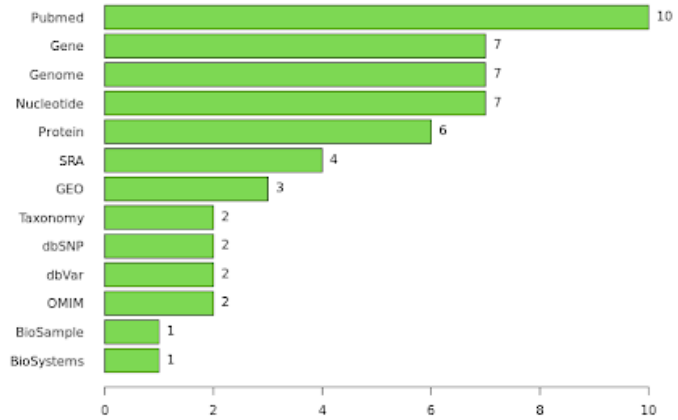
INRAE
Bases de données NCBI utilisées



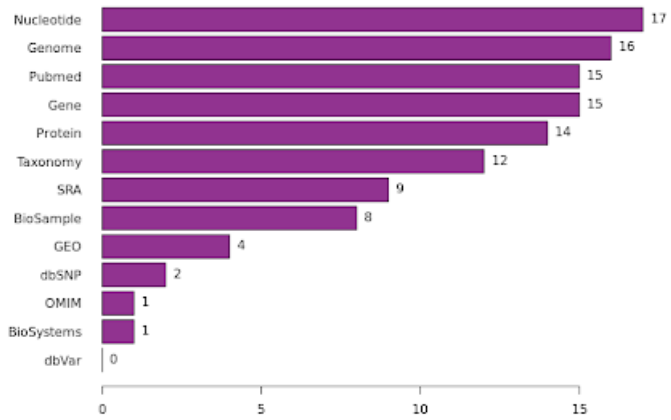
INSERM
Bases de données NCBI utilisées



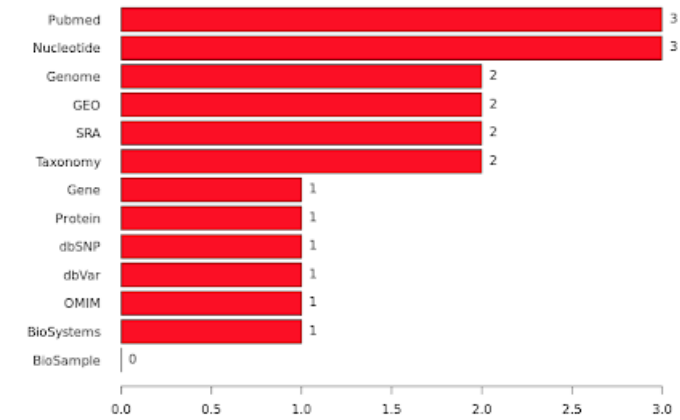
CEA
Bases de données NCBI utilisées



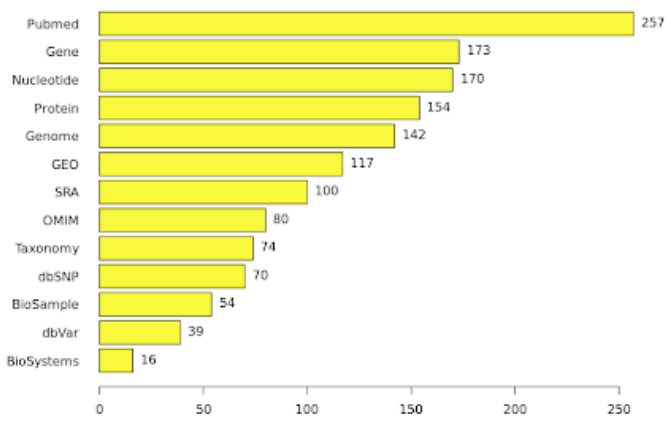
IRD
Bases de données NCBI utilisées



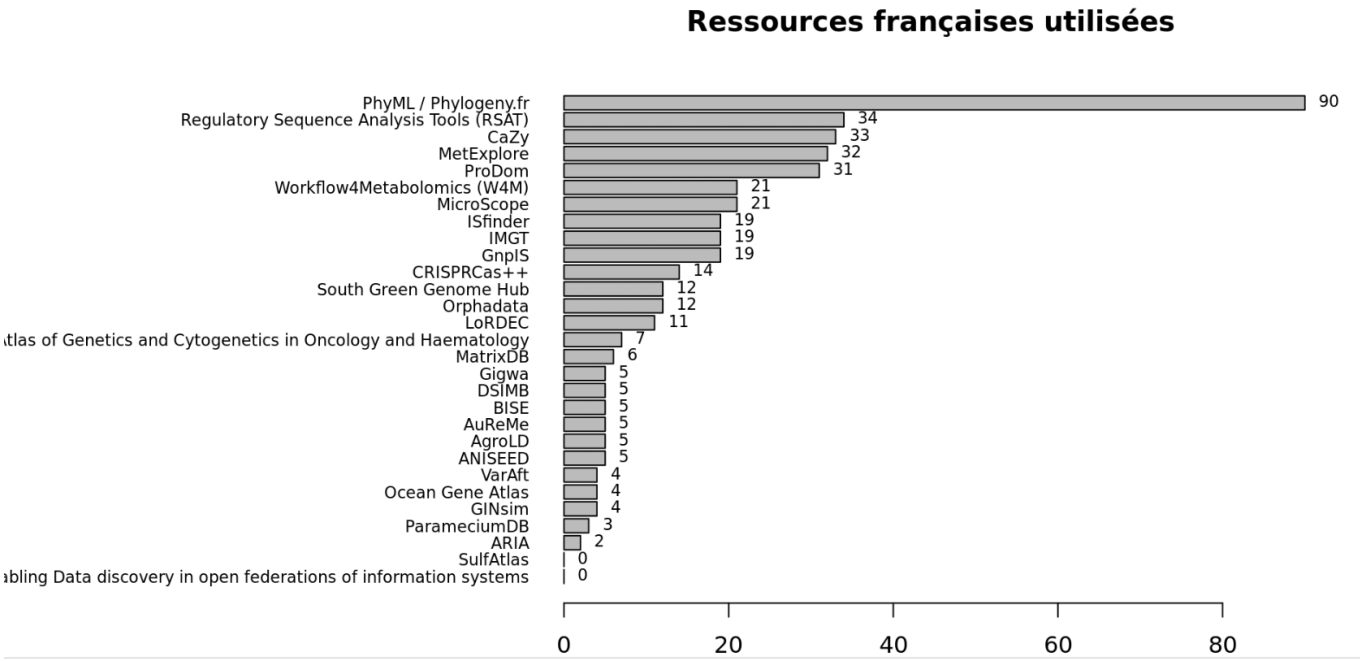
INRIA
Bases de données NCBI utilisées



Université
Bases de données NCBI utilisées

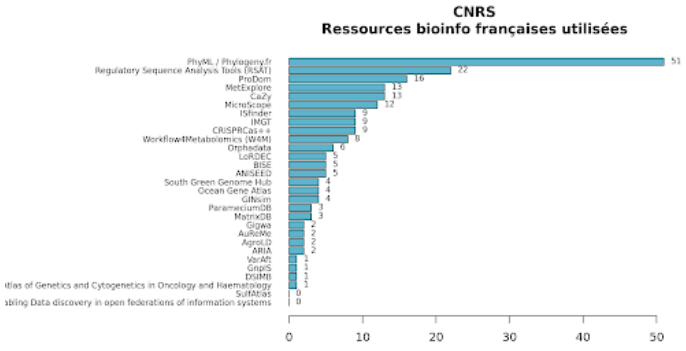


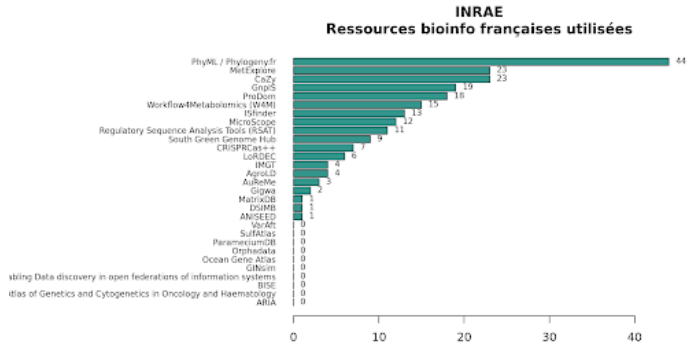
• E4. UTILISEZ-VOUS DES RESSOURCES BIOINFORMATIQUES FRANÇAISES?

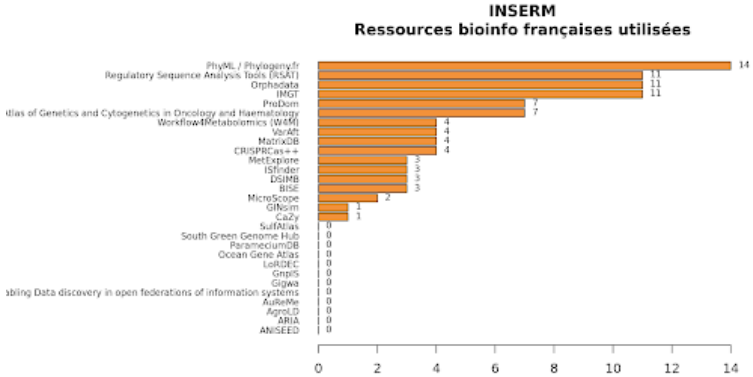


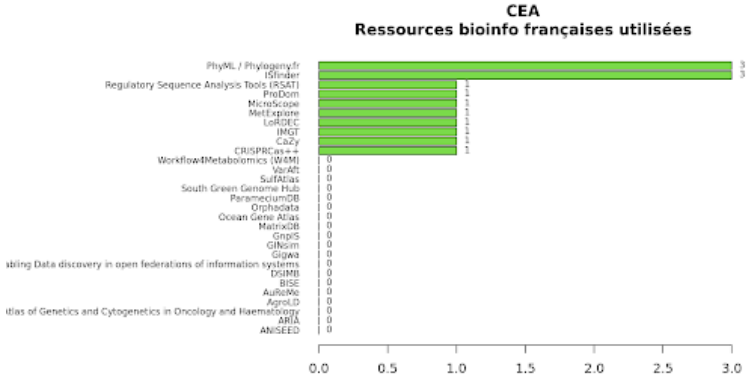
- **Résultat marquant** : outil généraliste **phylogeny.fr le plus utilisé.**

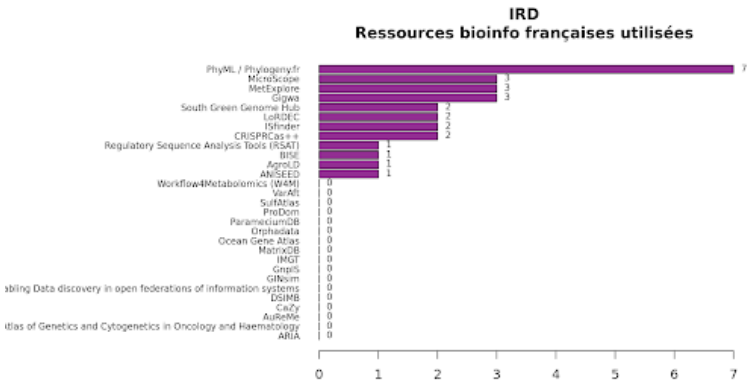
- **Ensuite des ressources plus spécialisées par communauté** (CAZY, Microscope, IMGT, GNPis…), **montre l'effort de communication pour mieux faire connaître les ressources.**

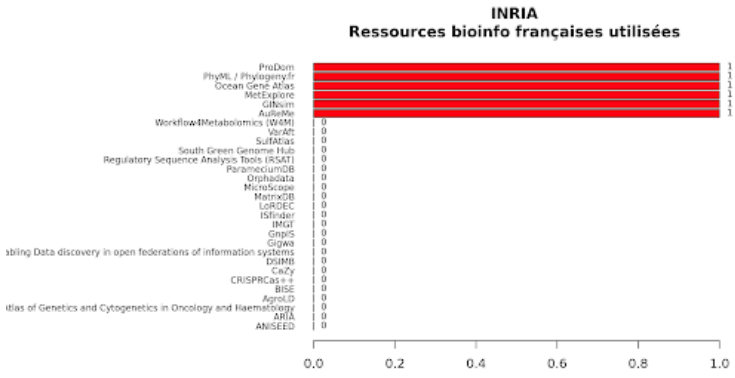


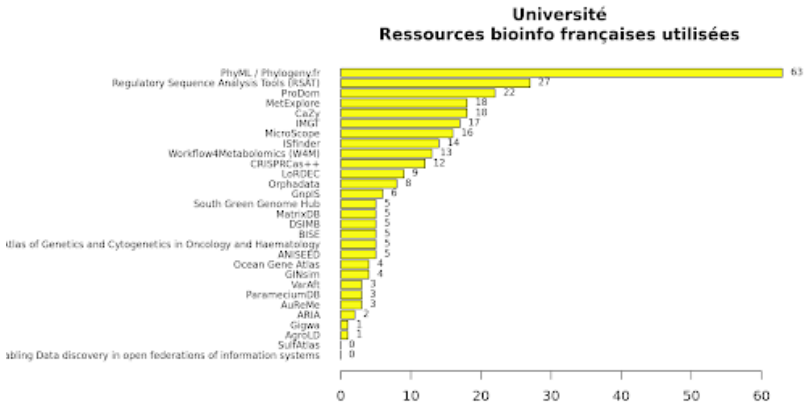




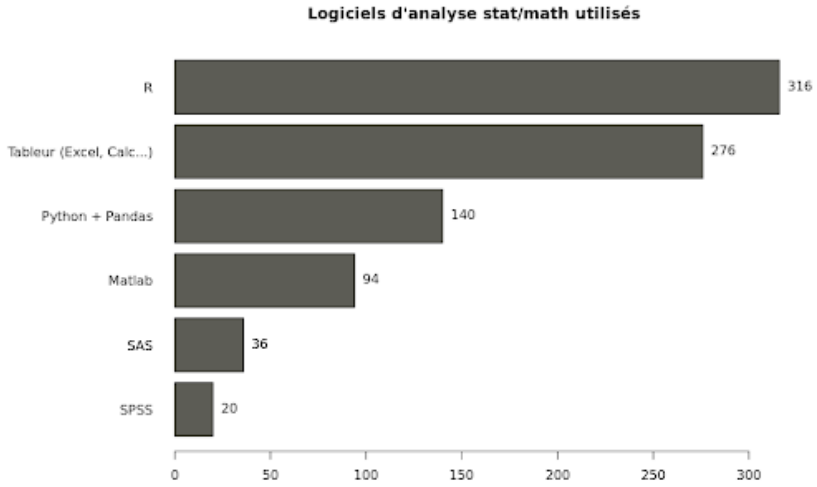




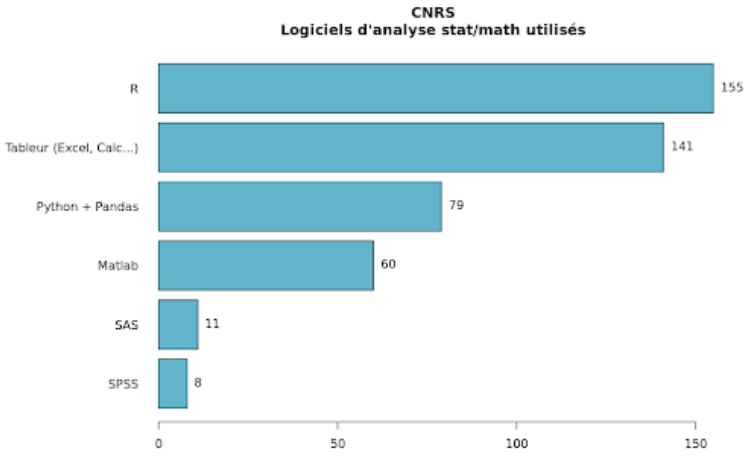


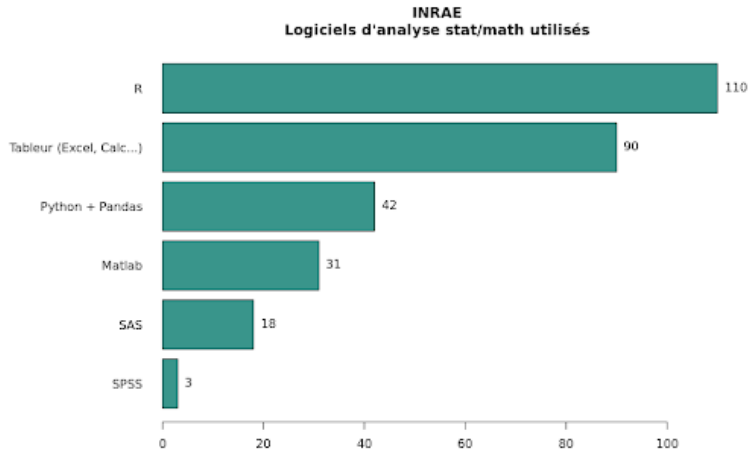


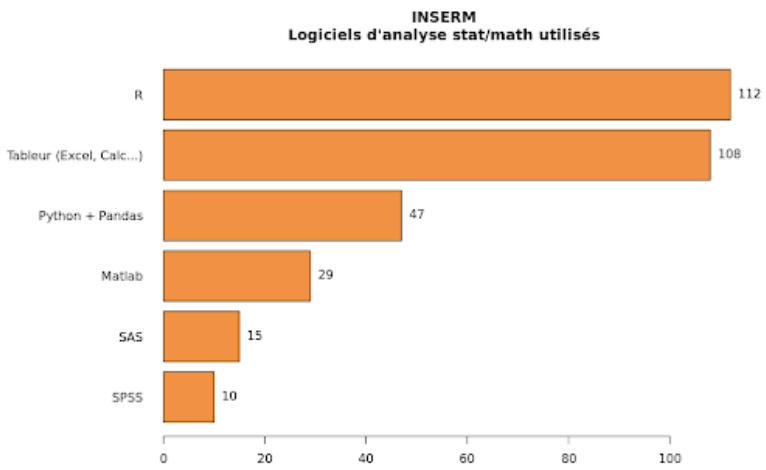
• E6. QUEL(S) LOGICIEL(S) UTILISEZ-VOUS POUR L'ANALYSE STATISTIQUE ET MATHÉMATIQUE?

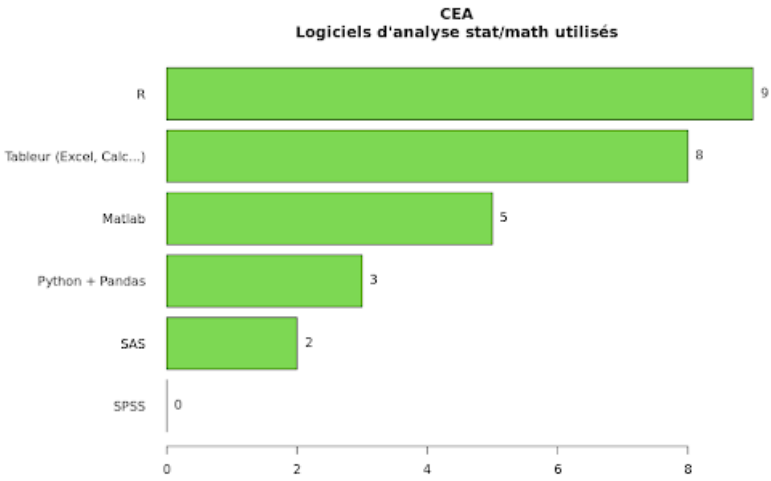


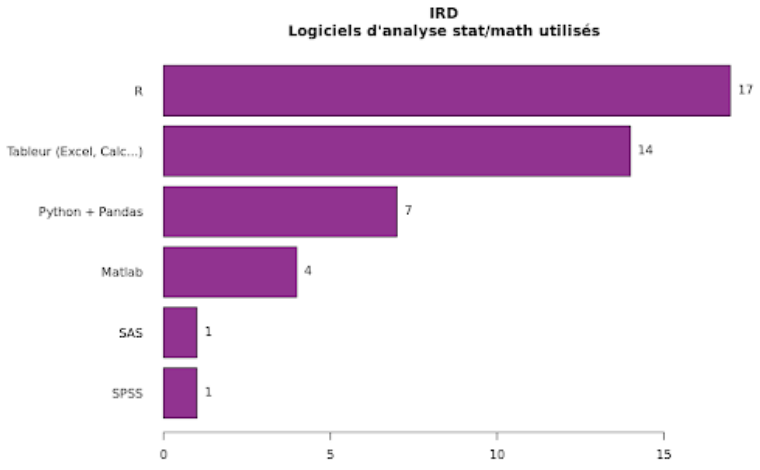
- **Montre la forte utilisation de R, très majoritaire.**
- Importante utilisation malgré tout des tableurs.
- **Ceci reflète éventuellement une dichotomie entre biologistes et bioinformaticiens. Cependant, de plus en plus de biologistes utilisent R.**
- On note aussi le succès de Python, qui reflète vraisemblablement la forte adoption de ce langage par les bioinformaticiens, et un intérêt plus récent de la part de biologistes expérimentaux.

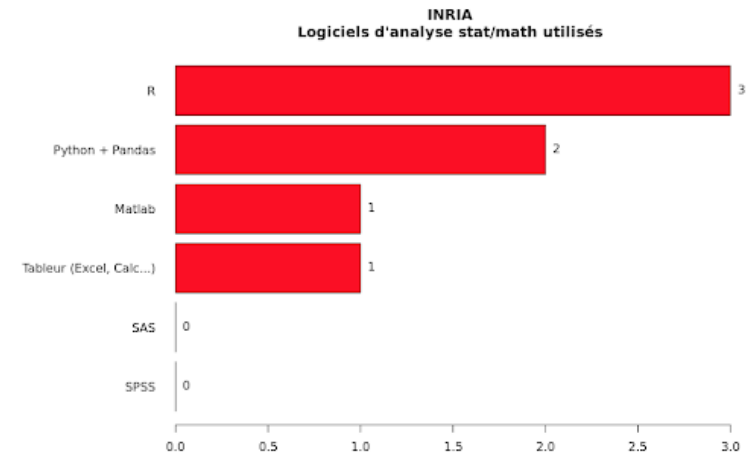

CNRS
Logiciels d'analyse stat/math utilisés
R
Tableur (Excel, Calc...)
Python + Pandas
Matlab
SAS
SPSS
155
141
79
60
11
8
0
50
100
150

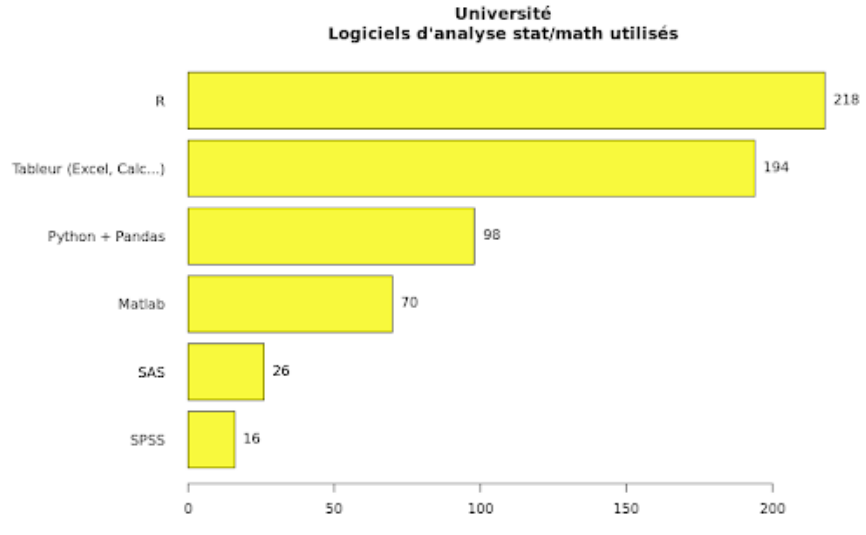

INRAE
Logiciels d'analyse stat/math utilisés
R
Tableur (Excel, Calc...)
Python + Pandas
Matlab
SAS
SPSS
110
90
42
31
18
3
0
20
40
60
80
100


INSERM
Logiciels d'analyse stat/math utilisés
R
Tableur (Excel, Calc...)
Python + Pandas
Matlab
SAS
SPSS
112
108
47
29
15
10
0
20
40
60
80
100

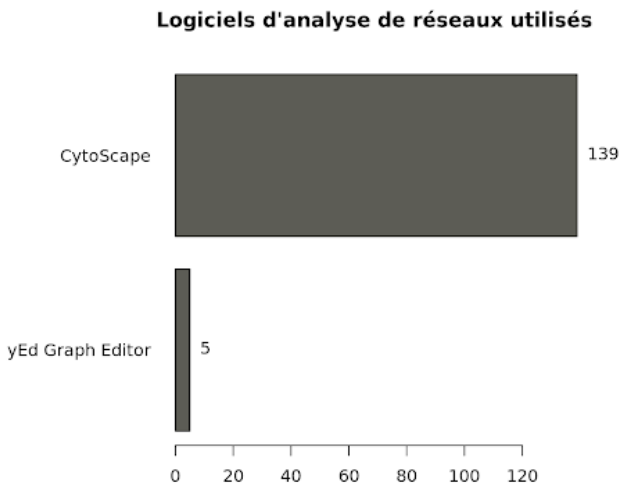

CEA
Logiciels d'analyse stat/math utilisés
R
Tableur (Excel, Calc...)
Matlab
Python + Pandas
SAS
SPSS
9
8
5
3
2
0
2
4
6
8


IRD
Logiciels d'analyse stat/math utilisés
R
Tableur (Excel, Calc...)
Python + Pandas
Matlab
SAS
SPSS
17
14
7
4
1
1
0
5
10
15

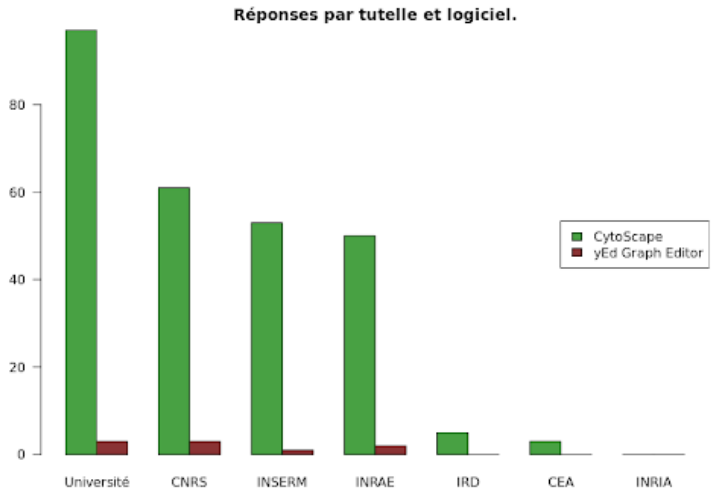

INRIA
Logiciels d'analyse stat/math utilisés
R
Python + Pandas
Matlab
Tableur (Excel, Calc...)
SAS
SPSS
3
2
1
1
0
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

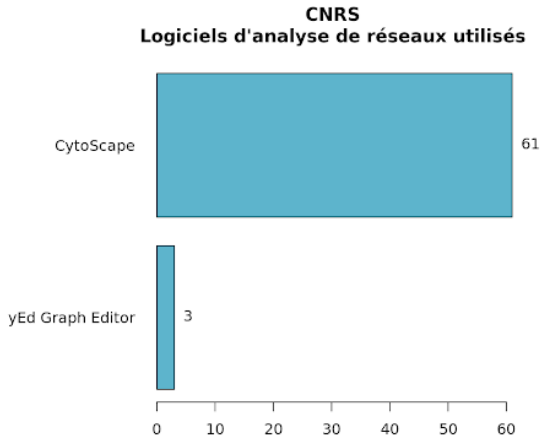

Université
Logiciels d'analyse stat/math utilisés
R
Tableur (Excel, Calc...)
Python + Pandas
Matlab
SAS
SPSS
218
194
98
70
26
16
0
50
100
150
200

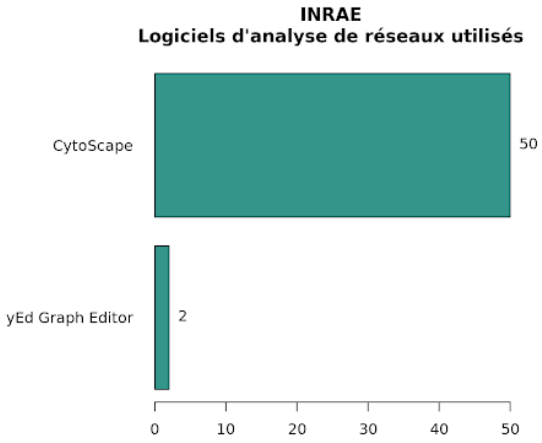
• E7. QUEL LOGICIEL UTILISEZ-VOUS POUR L'ANALYSE DE RÉSEAUX?

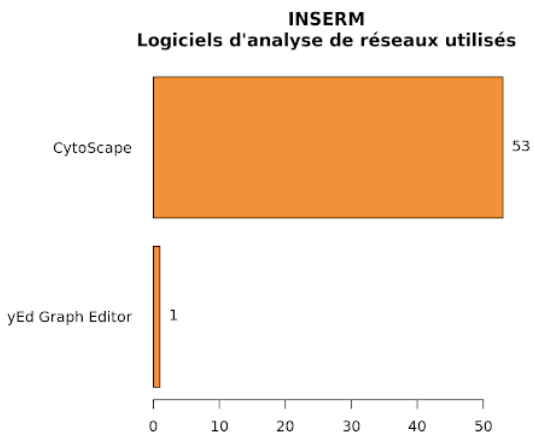


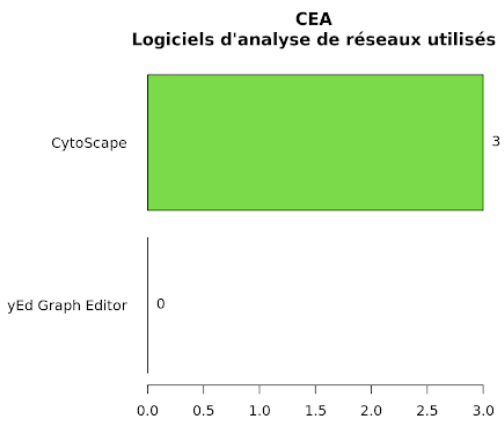
Cytoscape domine fortement le paysage
et il n'y a pas d'équivalent sur le marché.

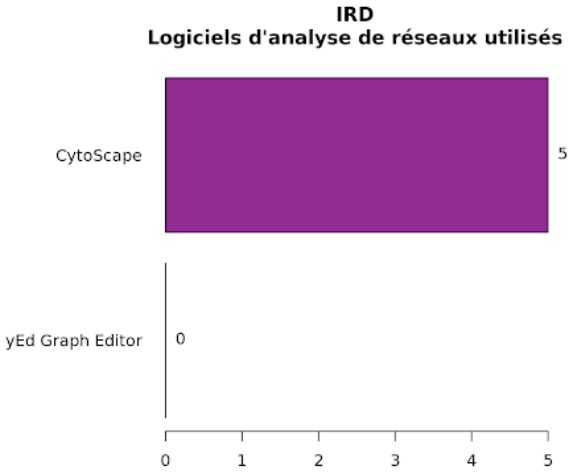


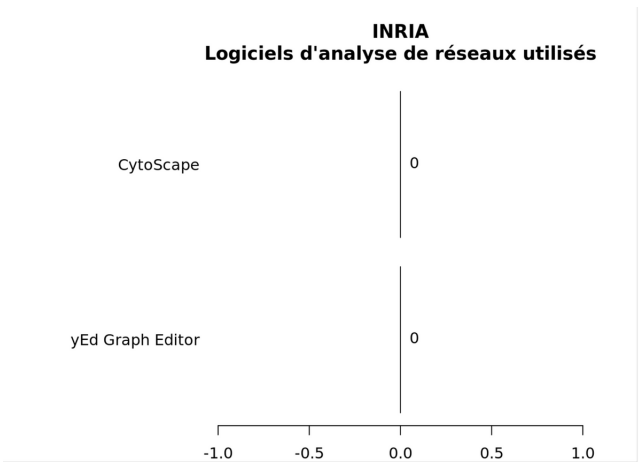


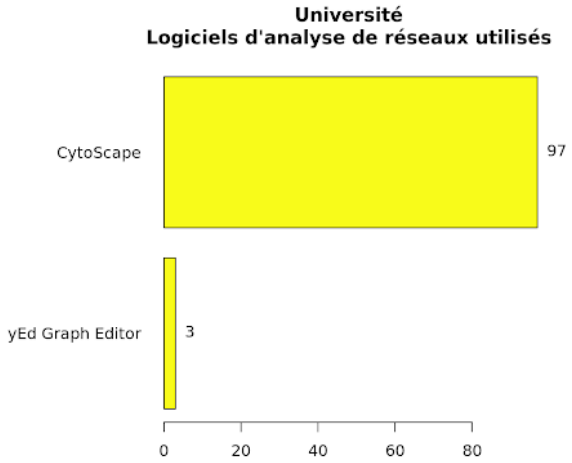




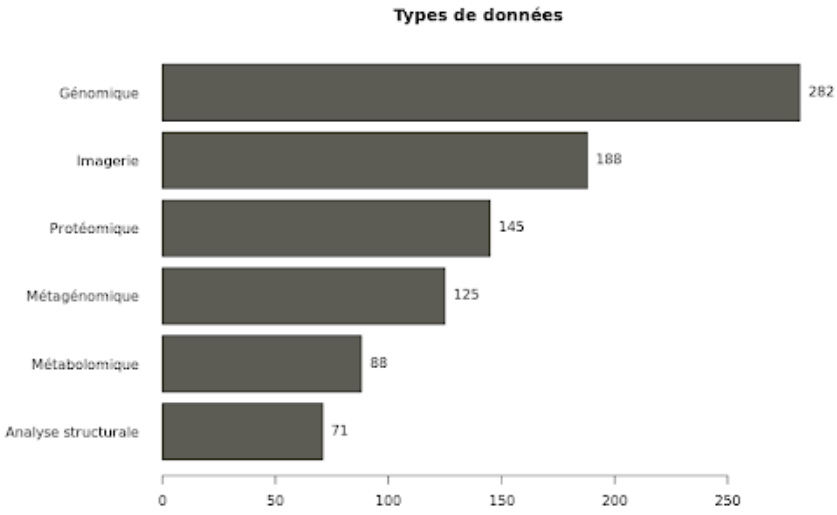



IRD
Logiciels d'analyse de réseaux utilisés
CytoScape
5
yEd Graph Editor
0
0
1
2
3
4
5

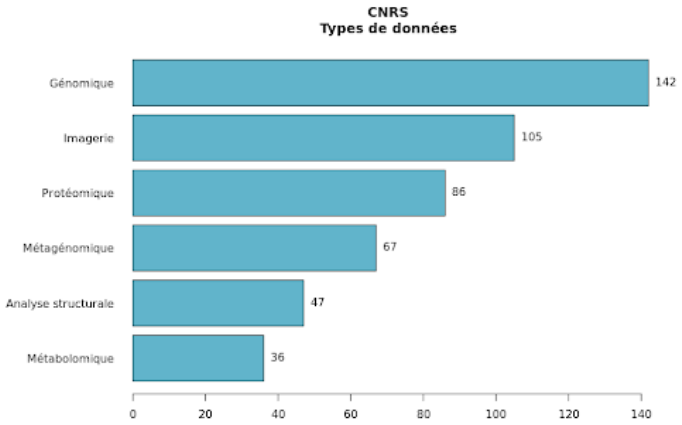

INRIA
Logiciels d'analyse de réseaux utilisés
CytoScape
0
yEd Graph Editor
0
-1.0
-0.5
0.0
0.5
1.0

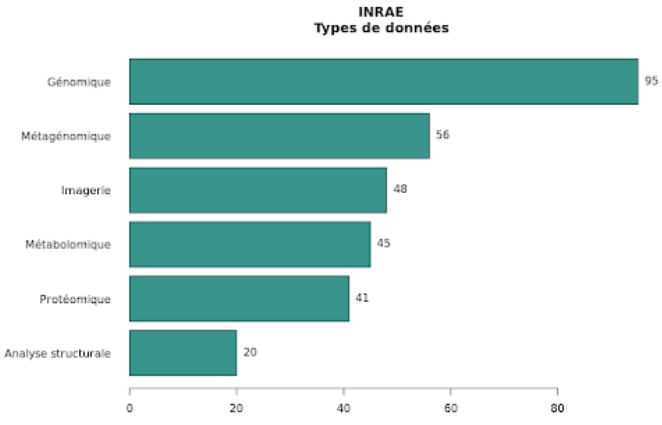

Université
Logiciels d'analyse de réseaux utilisés
CytoScape
97
yEd Graph Editor
3
0
20
40
60
80

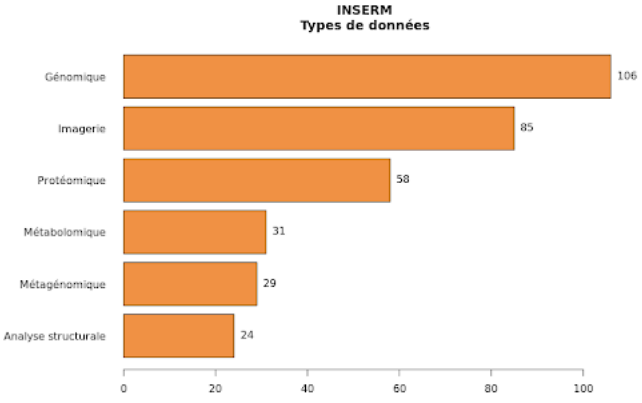
• E8. POUR QUELS TYPES DE DONNÉES UTILISEZ-VOUS DES OUTILS LOGICIELS?

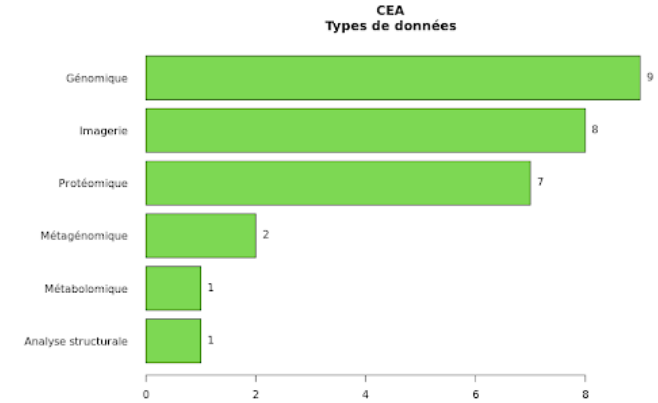


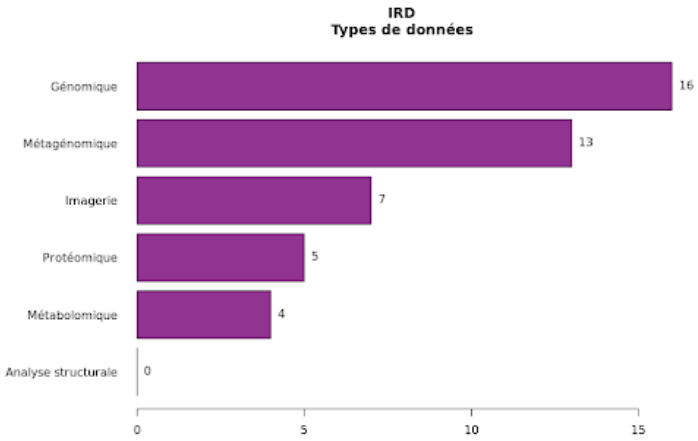
- **L'imagerie semble prendre une grande ampleur et entre dans le champ de la bioinformatique comme le souligne sa 2ème position juste derrière la génomique.**
- L'ampleur des réponses confirme l'intérêt du développement et de la mise à disposition de logiciels.
- Pour nuancer, l'imagerie peut recouvrir beaucoup de domaines différents mais le besoin reste présent pour chaque tutelle.

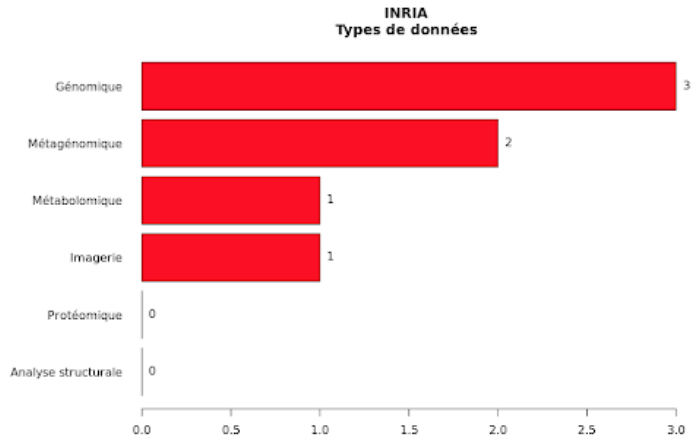


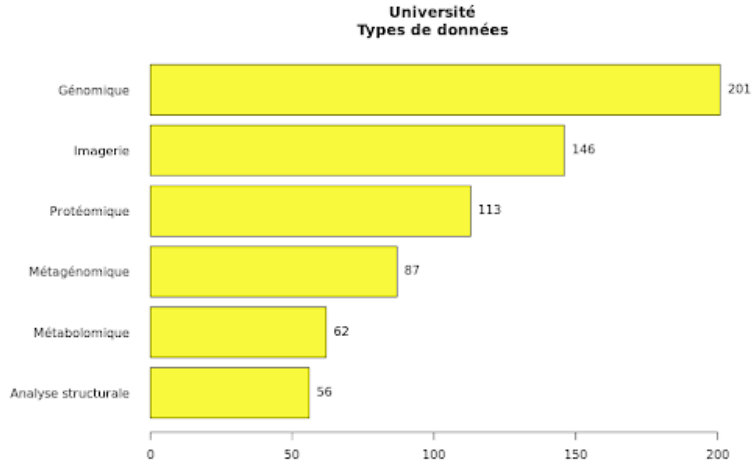







IRD
Types de données
Génomique
Métagénomique
Imagerie
Protéomique
Métabolomique
Analyse structurale
16
13
7
5
4
0
0
5
10
15


INRIA
Types de données
Génomique
Métagénomique
Métabolomique
Imagerie
Protéomique
Analyse structurale
3
2
1
1
0
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

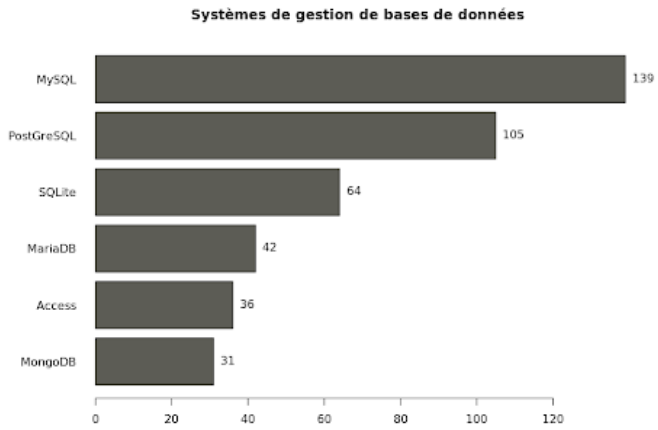

Université
Types de données
Génomique
Imagerie
Protéomique
Métagénomique
Métabolomique
Analyse structurale
201
146
113
87
62
56
0
50
100
150
200

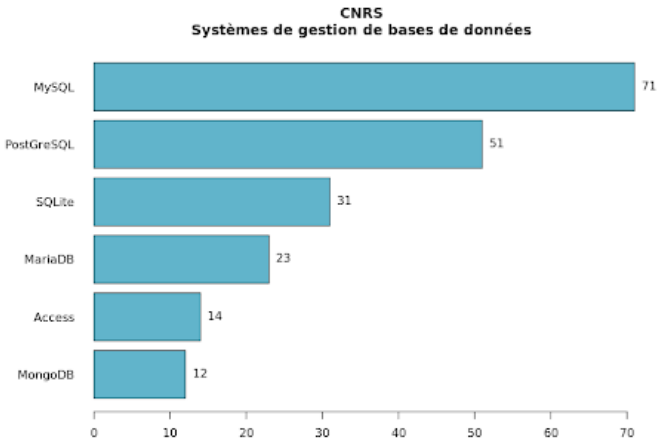
• E9. QUEL(S) SYSTÈME(S) DE GESTION DE BASES DE DONNÉES UTILISEZ-VOUS?

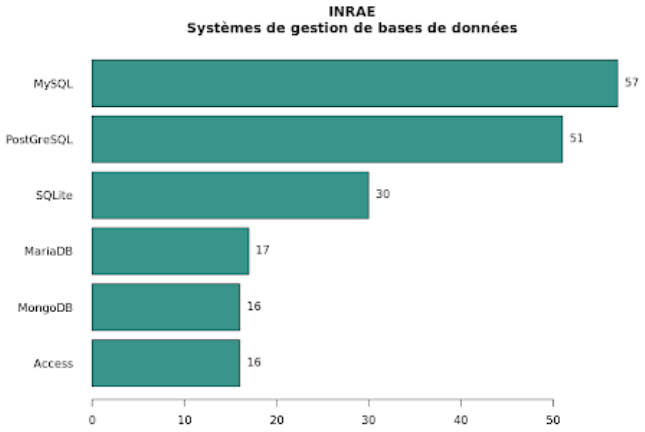
- **En terme de Base de Données, SQL est très majoritaire. C'est un langage robuste et éprouvé. MySQL est le Système de Gestion de Base de Données dominant devant PostgreSQL, SQLite et MariaDB (version open source de MySQL, racheté par Oracle).**

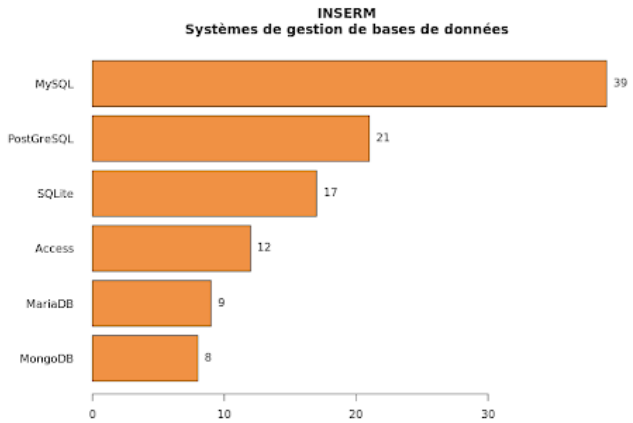
- MongoDB : technologie noSQL qui indique sûrement des initiatives intéressantes autour du big data.

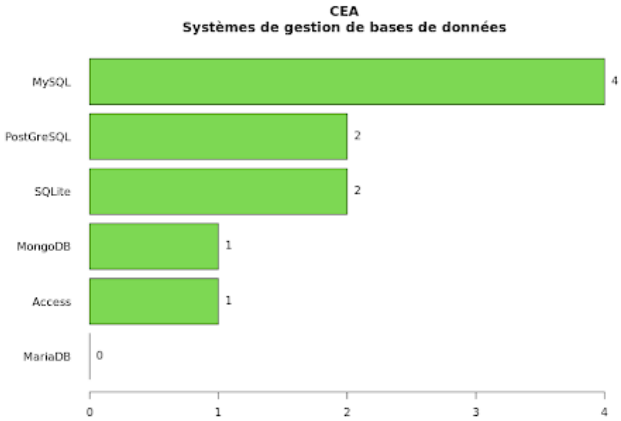
- L'utilisation d'Access reste possible dans le pack Office, cependant on peut se demander quelles données cela concerne encore. Intéressant de réfléchir à proposer une migration vers les infrastructures IFB.

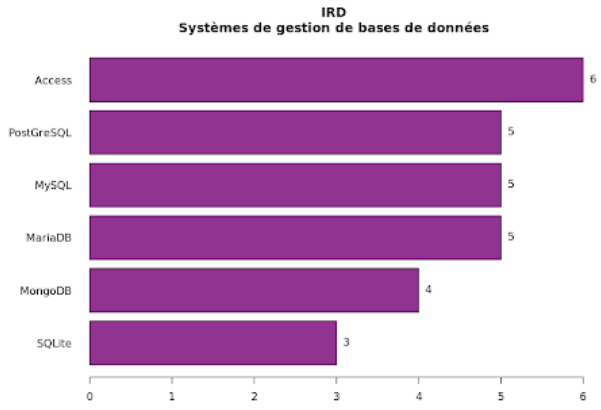


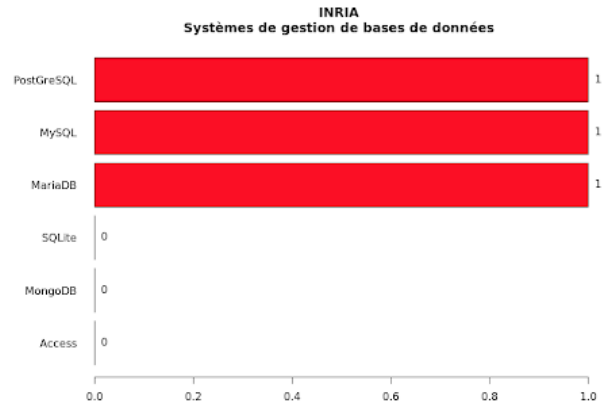


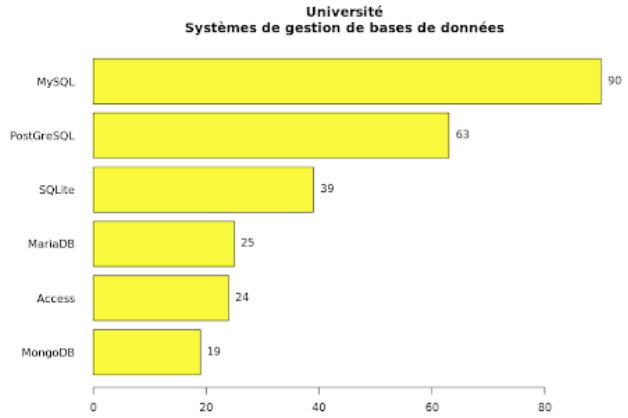




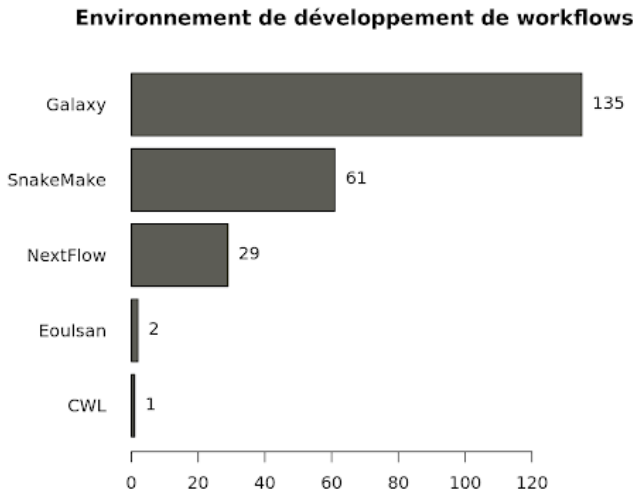



IRD
Systèmes de gestion de bases de données
Access
PostGreSQL
MySQL
MariaDB
MongoDB
SQLite
6
5
5
5
4
3
0
1
2
3
4
5
6

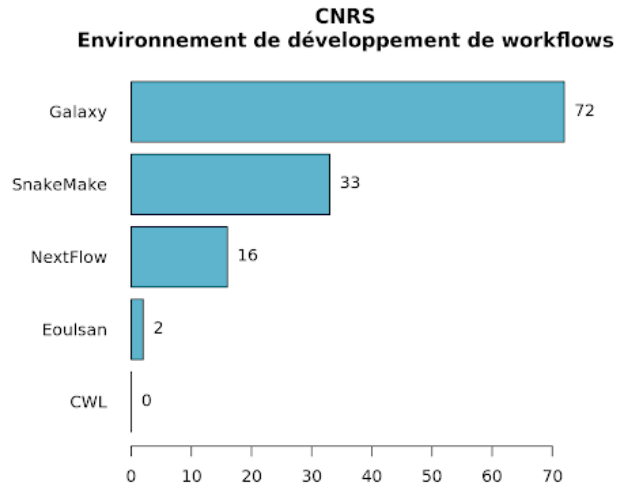

INRIA
Systèmes de gestion de bases de données
PostGreSQL
MySQL
MariaDB
SQLite
MongoDB
Access
1
1
1
0
0
0
0.0
0.2
0.4
0.6
0.8
1.0

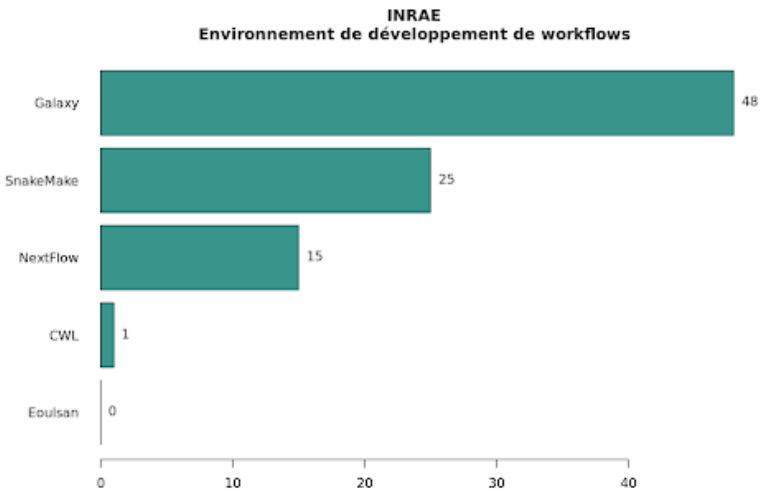

Université
Systèmes de gestion de bases de données
MySQL
PostGreSQL
SQLite
MariaDB
Access
MongoDB
90
63
39
25
24
19
0
20
40
60
80

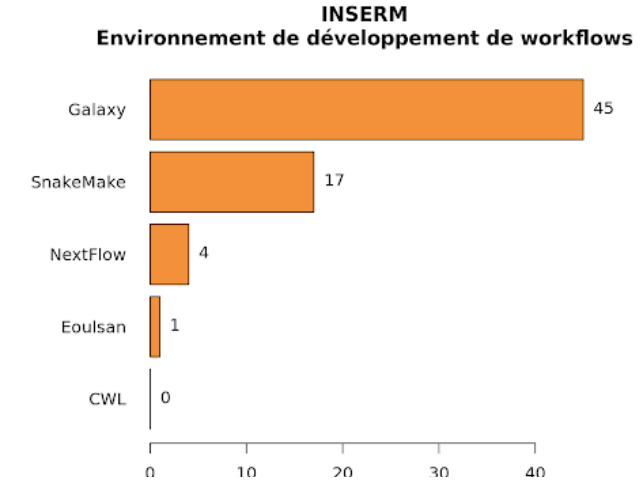
• E10. QUEL(S) ENVIRONNEMENT(S) DE DÉVELOPPEMENT DE WORKFLOWS UTILISEZ-VOUS?

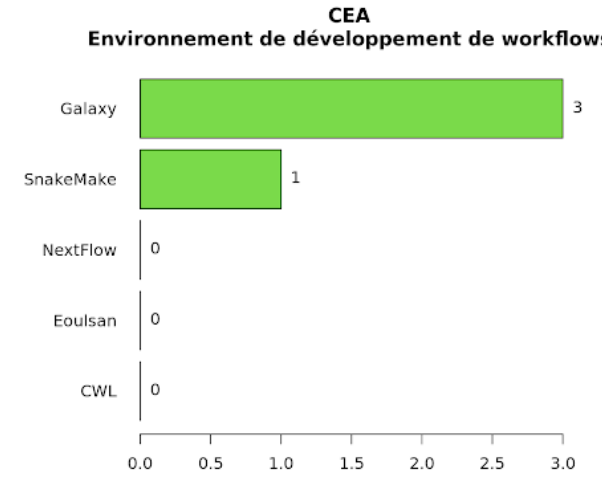


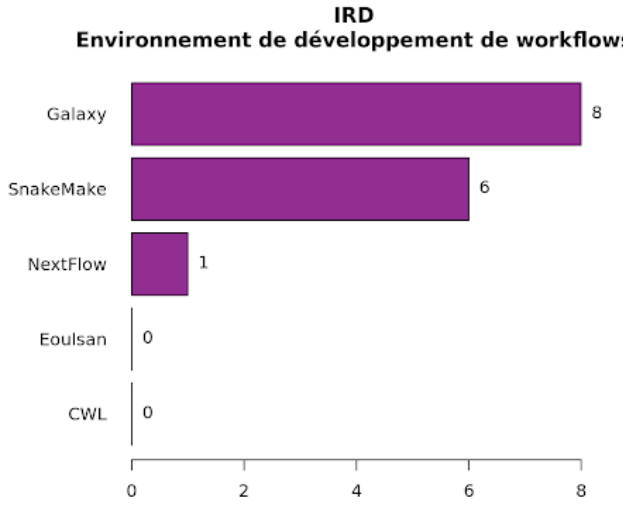
- **Galaxy très fortement majoritaire.**
- Snakemake majoritaire par rapport à Nextflow, sans doute parce que ce dernier est plus récent.
- **Attention, il ne faut pas comparer Galaxy avec Snakemake/Nextflow puisqu'ils n'ont pas le même type d'approche (interface web / programmation).**

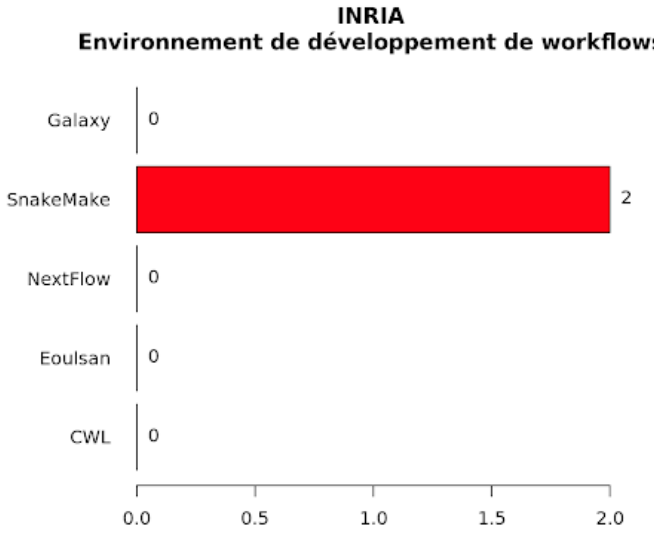


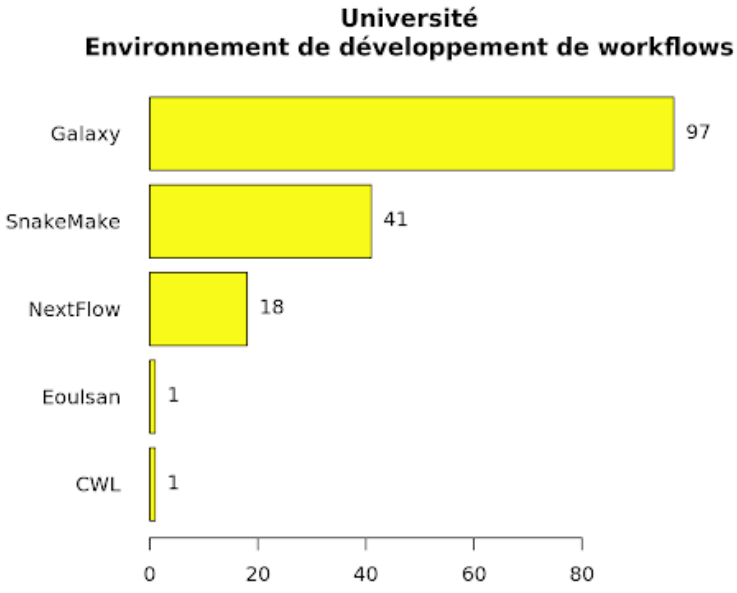




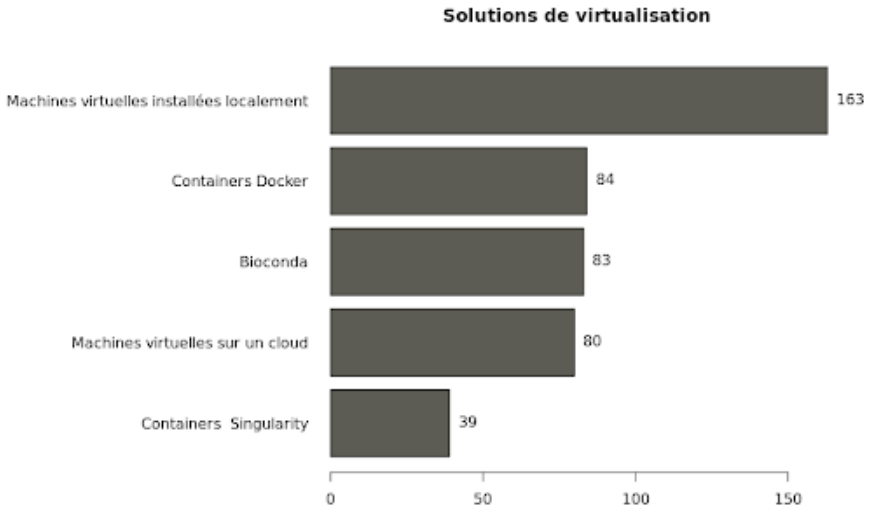



IRD
Environnement de développement de workflows
Galaxy
SnakeMake
NextFlow
Eoulsan
CWL
8
6
1
0
0
0
2
4
6
8

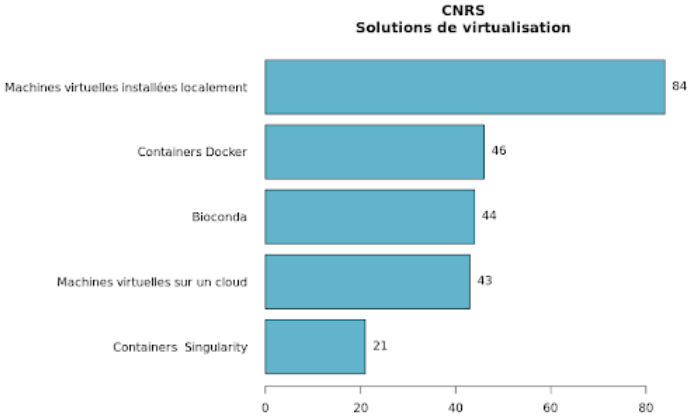

INRIA
Environnement de développement de workflows
Galaxy
SnakeMake
NextFlow
Eoulsan
CWL
0
2
0
0
0
0.0
0.5
1.0
1.5
2.0

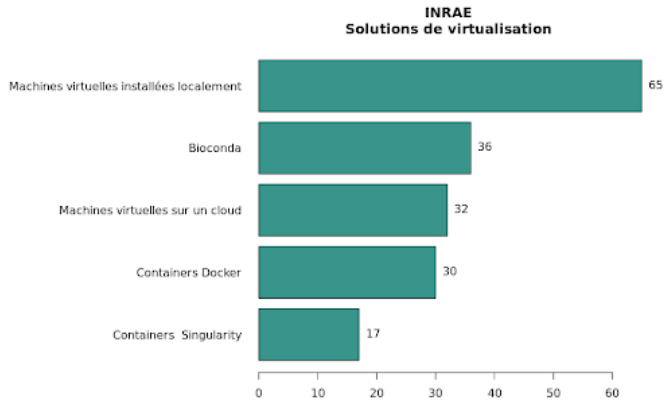

Université
Environnement de développement de workflows
Galaxy
SnakeMake
NextFlow
Eoulsan
CWL
97
41
18
1
1
0
20
40
60
80

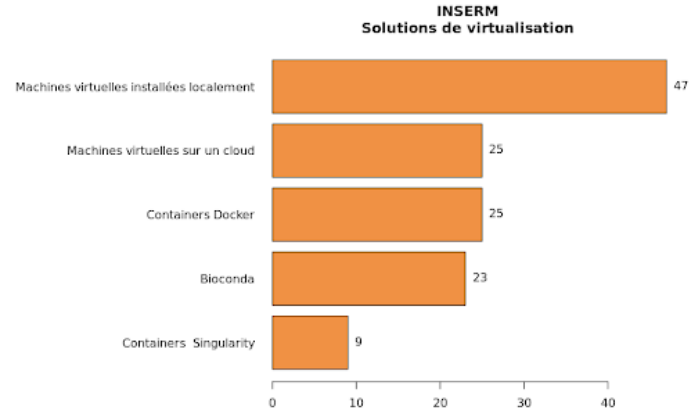
- **E11. UTILISEZ-VOUS LES SOLUTIONS DE VIRTUALISATION ET DE DÉPLOIEMENT SUIVANTES ?**

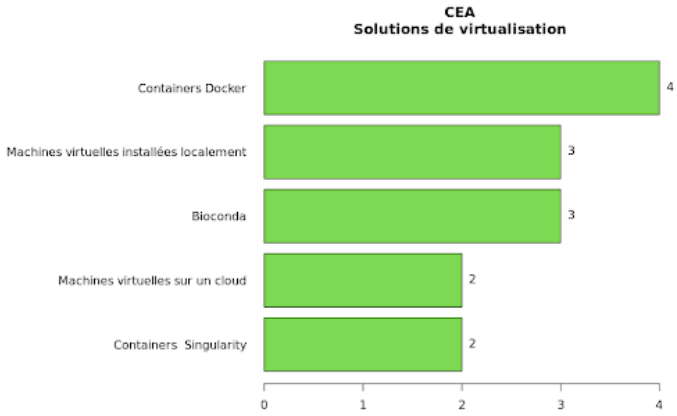


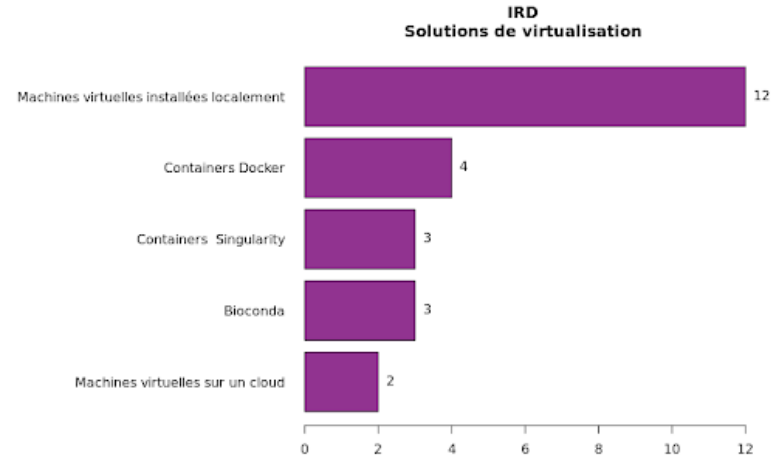
- **Technologie de virtualisation largement adoptée.**
- **Préférence apparente pour l'utilisation de ressources locales.**

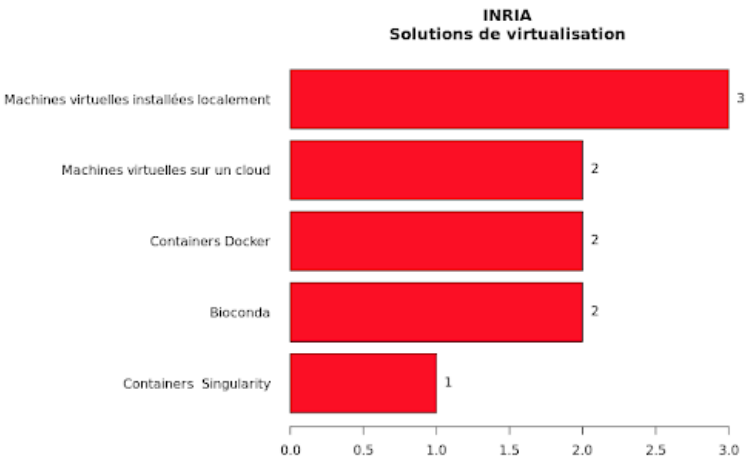


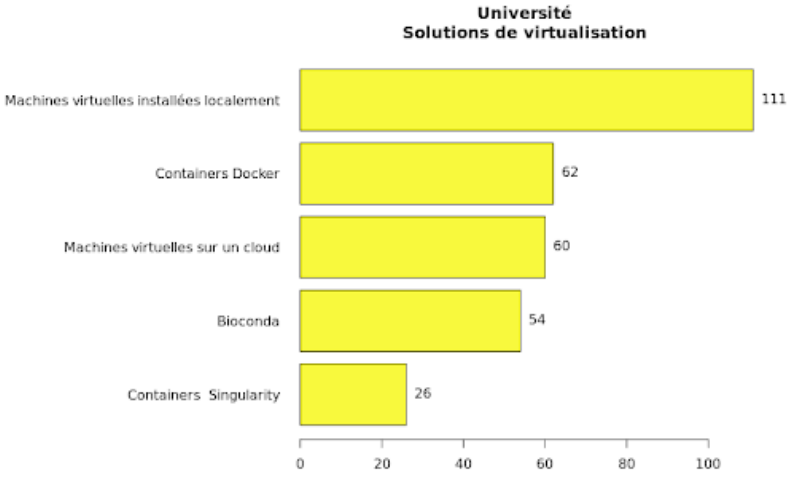






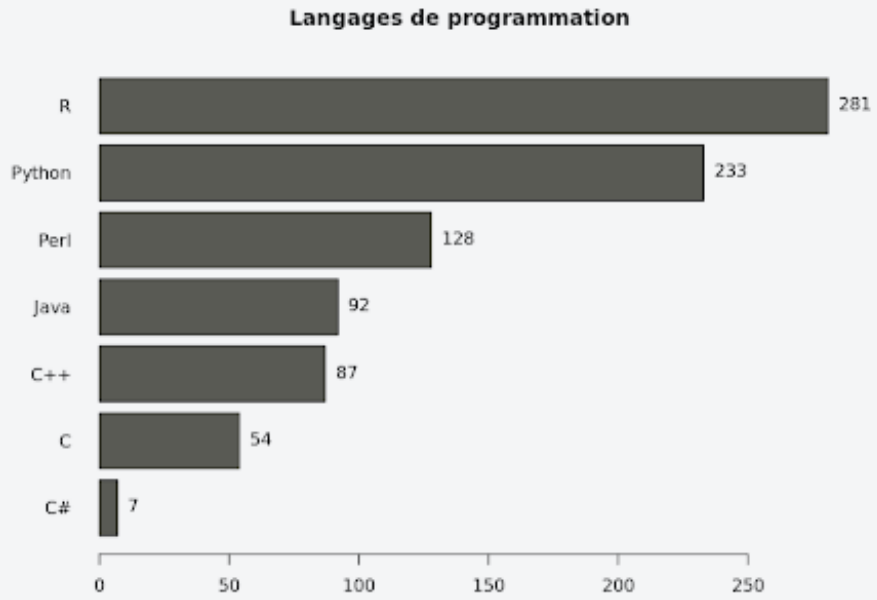

IRD
Solutions de virtualisation
Machines virtuelles installées localement
Containers Docker
Containers Singularity
Bioconda
Machines virtuelles sur un cloud
12
4
3
3
2
0
2
4
6
8
10
12

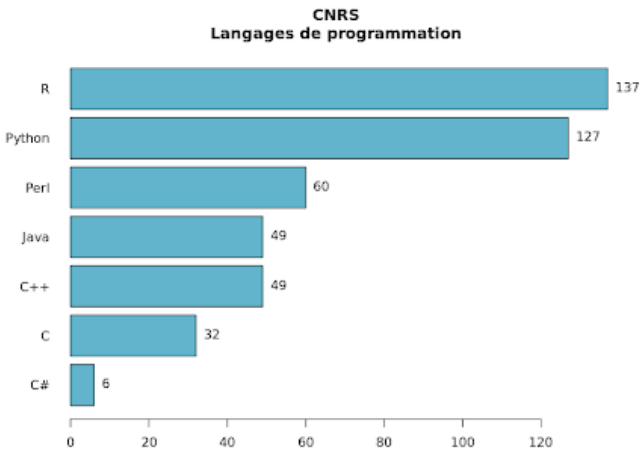

INRIA
Solutions de virtualisation
Machines virtuelles installées localement
Machines virtuelles sur un cloud
Containers Docker
Bioconda
Containers Singularity
3
2
2
2
1
0.0
0.5
1.0
1.5
2.0
2.5
3.0

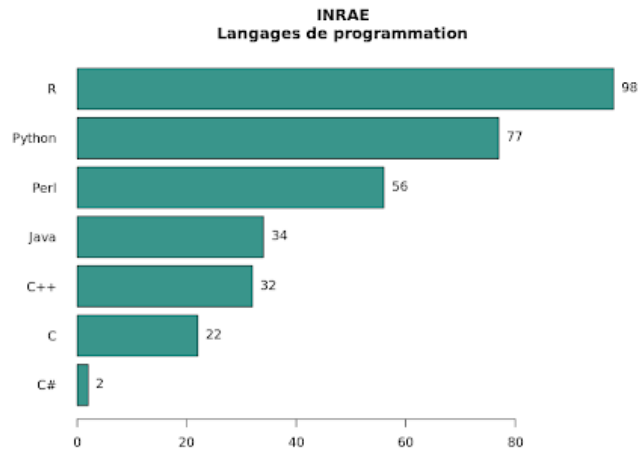

Université
Solutions de virtualisation
Machines virtuelles installées localement
Containers Docker
Machines virtuelles sur un cloud
Bioconda
Containers Singularity
111
62
60
54
26
0
20
40
60
80
100

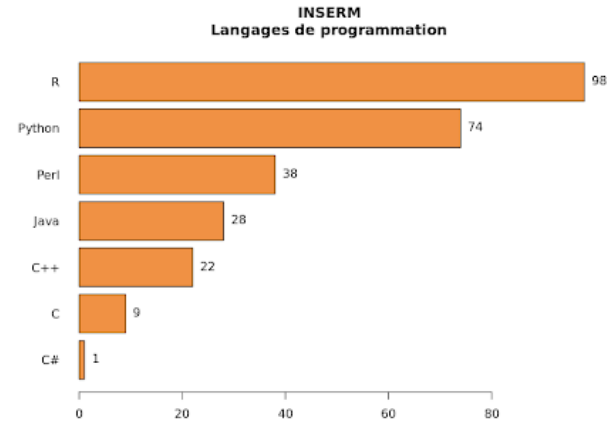
• E12. QUEL(S) LANGAGE(S) DE PROGRAMMATION UTILISEZ-VOUS?

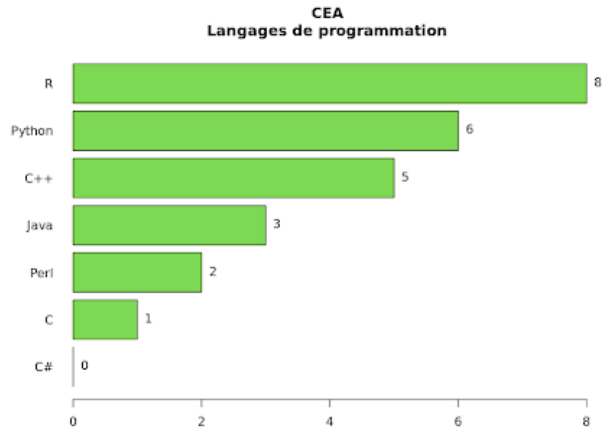
- **Forte dominance de R et Python**

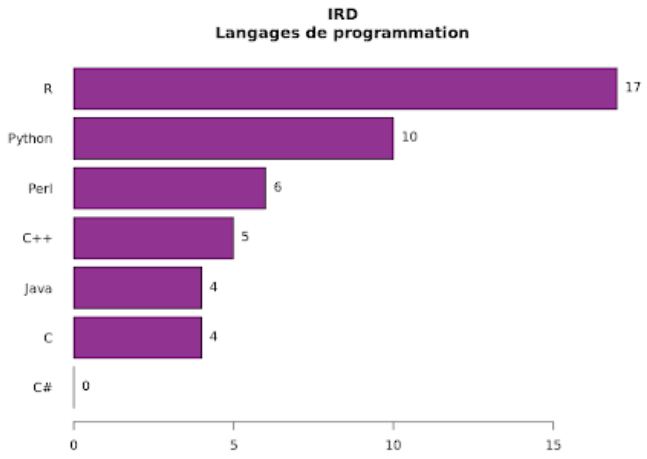


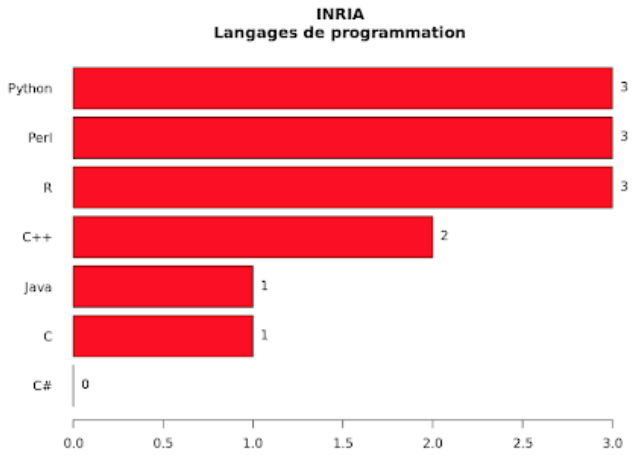


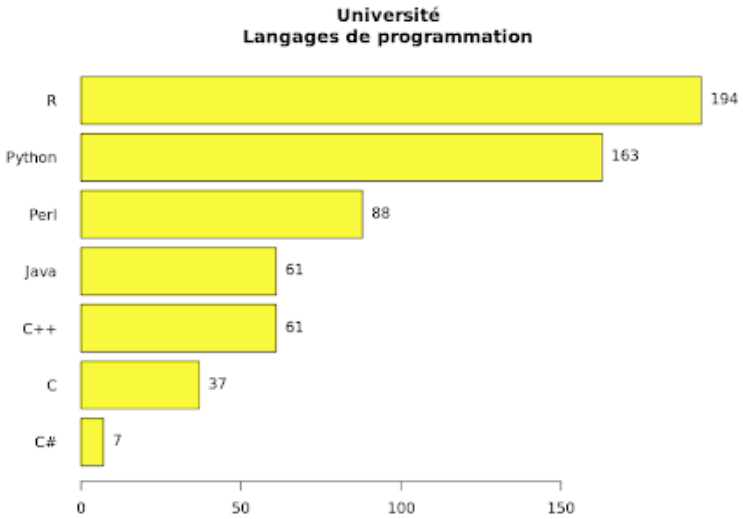








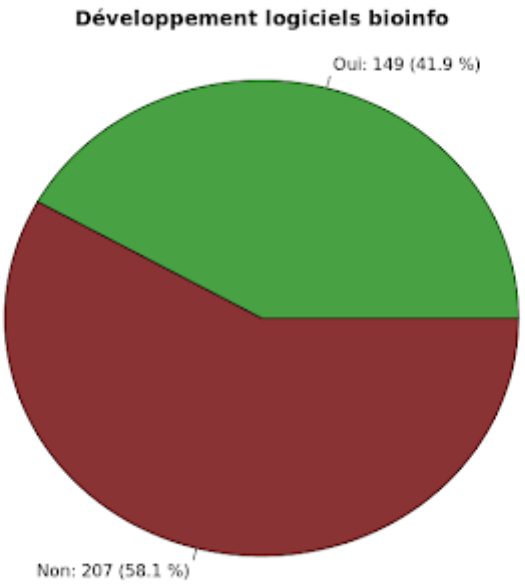




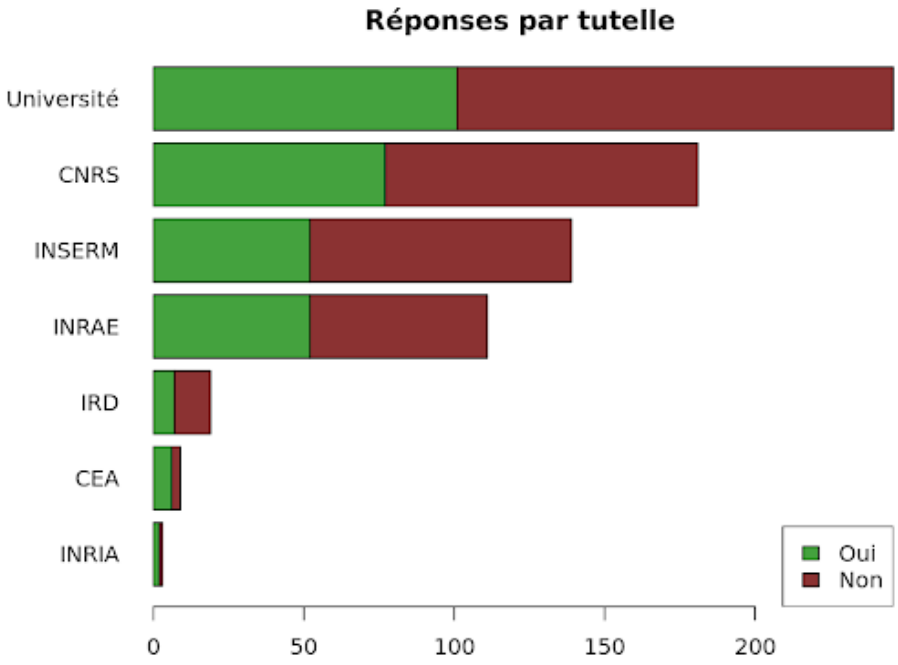
AUTRES LANGUAGUES UTILISÉS

- 4D
- Assembleur ARM
- Awk
- Bash
- Cuda
- Delphi
- Elm
- FIJI
- Fortran
- Go
- Groovy
- HTML
- Igor Pro (WaveMetrics)
- JavaScript et dérivés
- Julia
- LaTeX
- Labview
- lisp
- Mathematica
- Matlab
- MySQL
- Octave
- OjectifC
- Perl
- PHP
- Rev
- Ruby
- sed
- SQL
- Scala
- Scilab
- Shell
- Tcl
- VirtualBasic
- VisualBasic

- **E13. VOTRE UNITÉ/ÉQUIPE DÉVELOPPE-T-ELLE DES RESSOURCES LOGICIELLES POUR LA BIOINFORMATIQUE?**

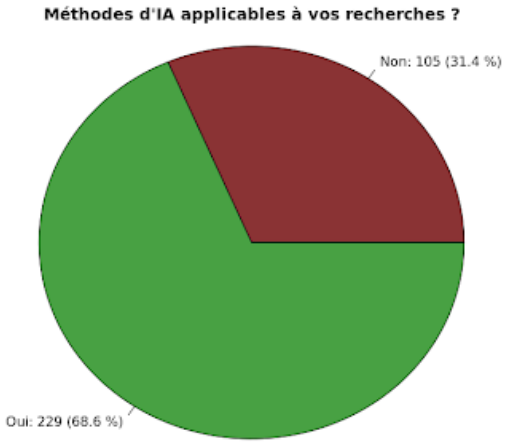


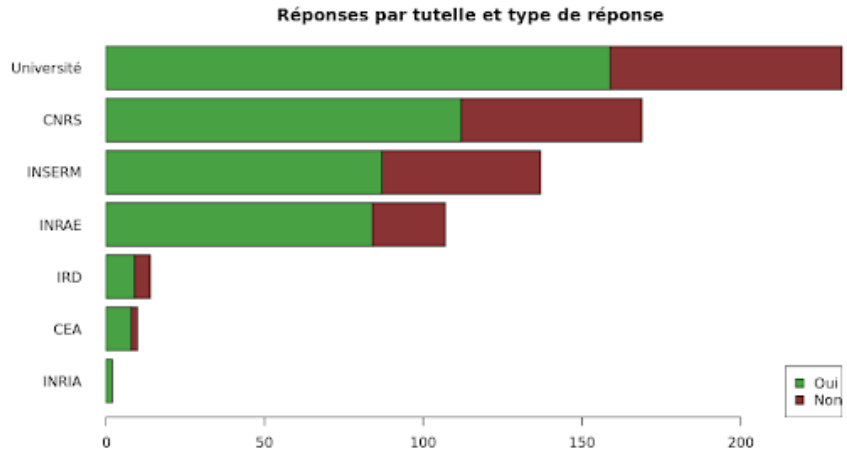
Près de la moitié des labos/ équipes (42%) qui ont répondu font du développement.



PARTIE F. RECOURS À L'INTELLIGENCE ARTIFICIELLE

• F1. LES MÉTHODES D'INTELLIGENCE ARTIFICIELLE S'APPLIQUENT-ELLES À VOS THÉMATIQUES DE RECHERCHE?

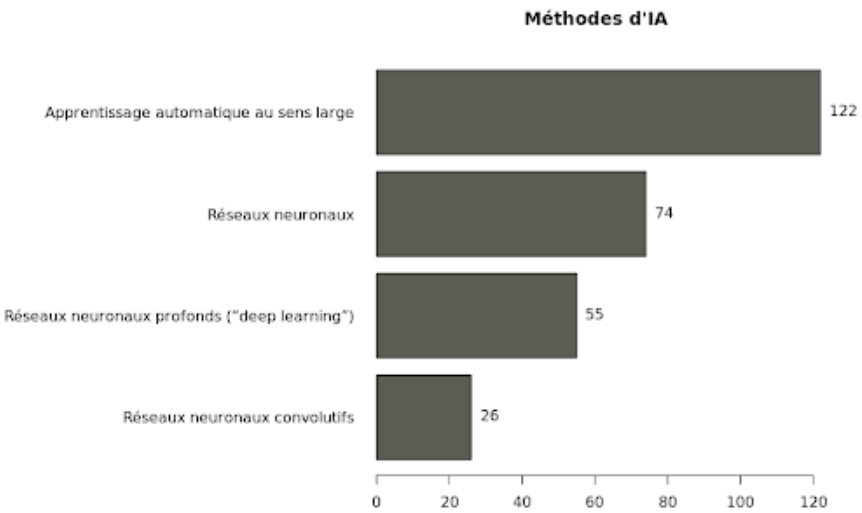




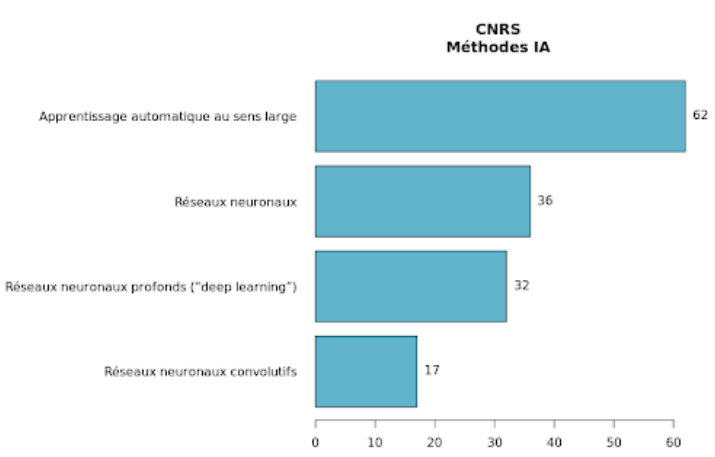
Fort intérêt pour cette thématique dans tous les instituts.

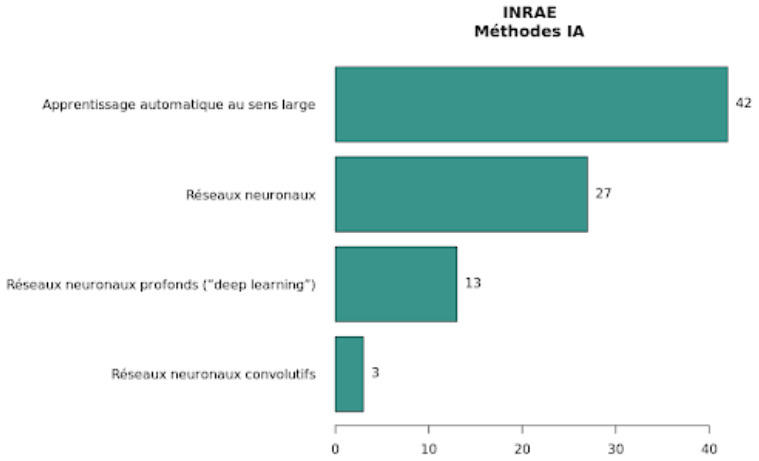
• F2. UTILISEZ-VOUS DÉJÀ L'UNE OU PLUSIEURS DES MÉTHODES D'IA SUIVANTES ?

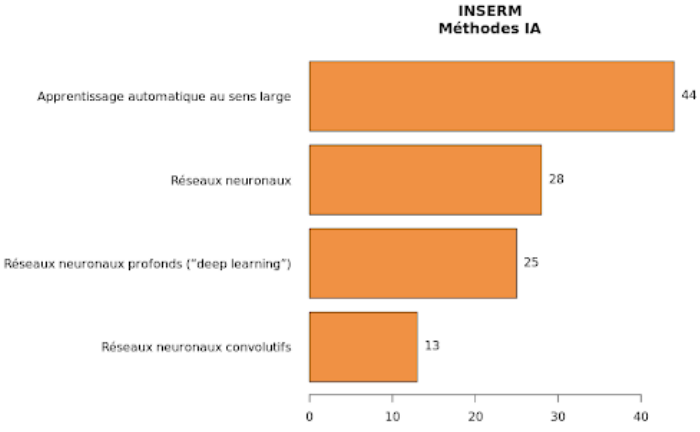
(cette question ne concerne que les personnes ayant répondu "oui" à la question F1)

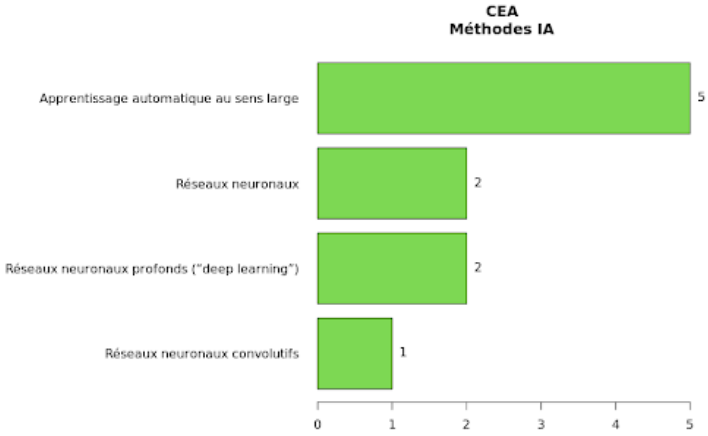


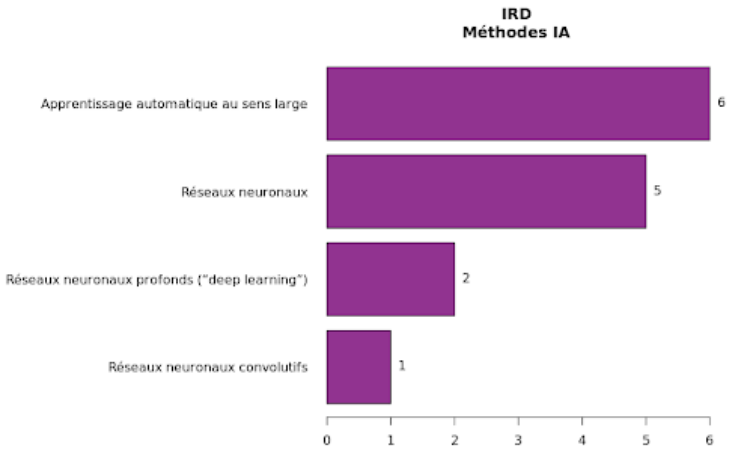
- **Il y a une utilisation non négligeable.**
- 55 équipes/unités sur 407 **(13,5%)** déclarent utiliser déjà des méthodes de deep learning.

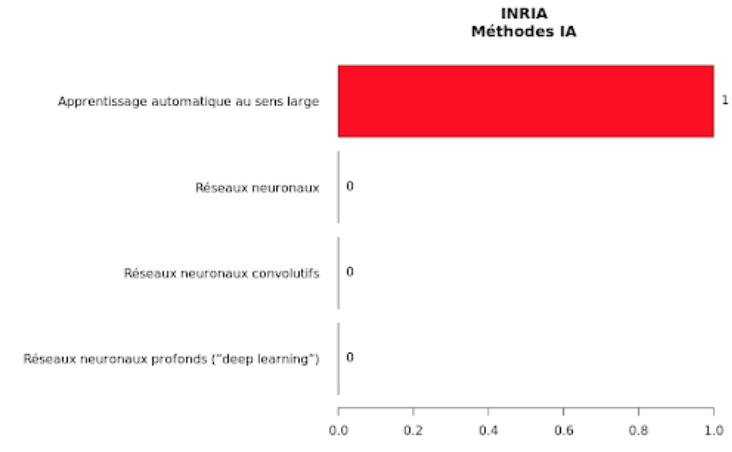

CNRS
Méthodes IA
Apprentissage automatique au sens large
62
Réseaux neuronaux
36
Réseaux neuronaux profonds ("deep learning")
32
Réseaux neuronaux convolutifs
17
0
10
20
30
40
50
60

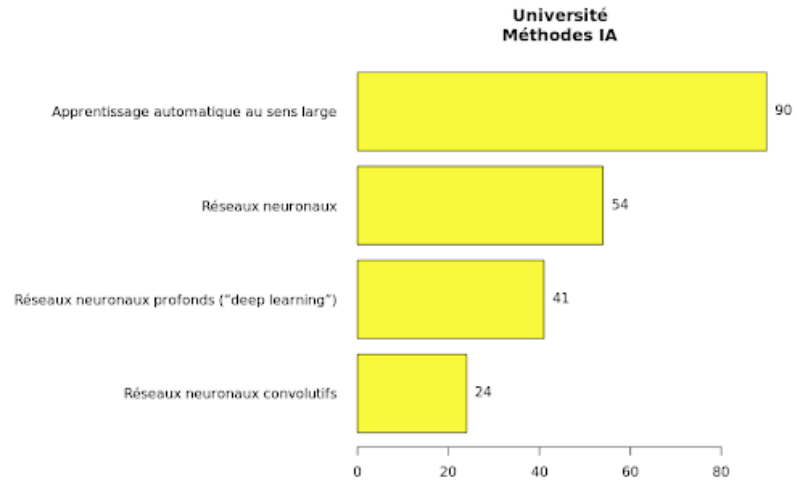

INRAE
Méthodes IA
Apprentissage automatique au sens large
42
Réseaux neuronaux
27
Réseaux neuronaux profonds ("deep learning")
13
Réseaux neuronaux convolutifs
3
0
10
20
30
40


INSERM
Méthodes IA
Apprentissage automatique au sens large
44
Réseaux neuronaux
28
Réseaux neuronaux profonds ("deep learning")
25
Réseaux neuronaux convolutifs
13
0
10
20
30
40


CEA
Méthodes IA
Apprentissage automatique au sens large
5
Réseaux neuronaux
2
Réseaux neuronaux profonds ("deep learning")
2
Réseaux neuronaux convolutifs
1
0
1
2
3
4
5

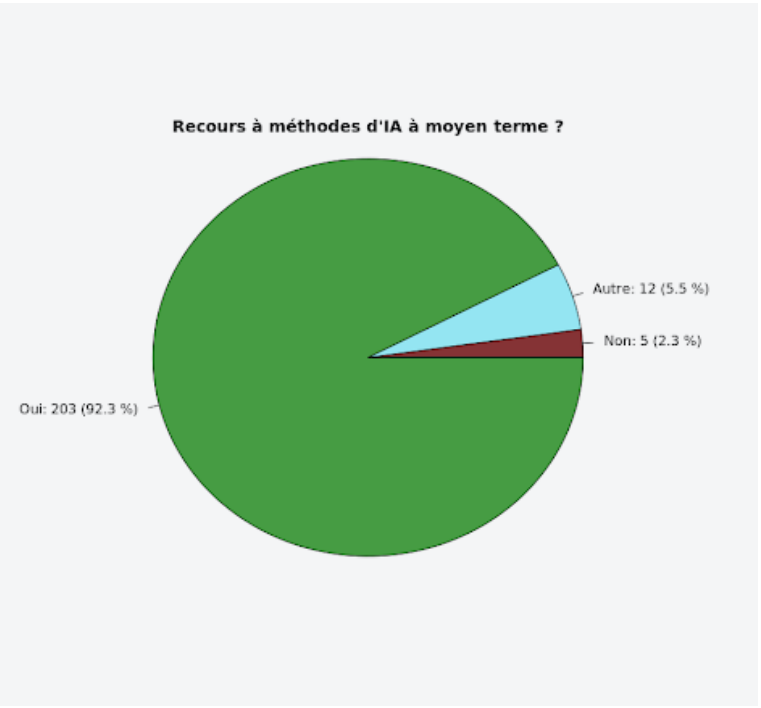

IRD
Méthodes IA
Apprentissage automatique au sens large
6
Réseaux neuronaux
5
Réseaux neuronaux profonds ("deep learning")
2
Réseaux neuronaux convolutifs
1
0
1
2
3
4
5
6

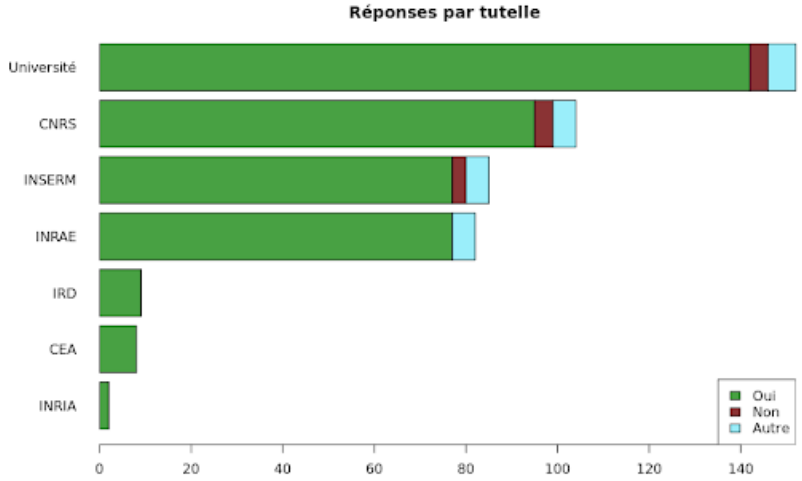

INRIA
Méthodes IA
Apprentissage automatique au sens large
1
Réseaux neuronaux
0
Réseaux neuronaux convolutifs
0
Réseaux neuronaux profonds ("deep learning")
0
0.0
0.2
0.4
0.6
0.8
1.0


Université
Méthodes IA
Apprentissage automatique au sens large
90
Réseaux neuronaux
54
Réseaux neuronaux profonds ("deep learning")
41
Réseaux neuronaux convolutifs
24
0
20
40
60
80

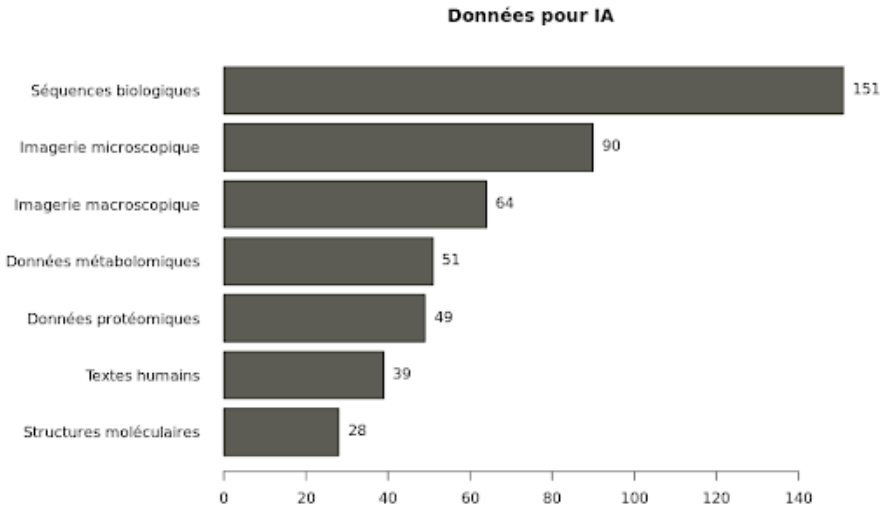
- **F3. A MOYEN TERME, COMPTEZ-VOUS RECOURIR À DES MÉTHODES D'IA POUR VOS PROJETS DE RECHERCHE?**

(cette question ne concerne que les personnes ayant répondu "oui" à la question F1)





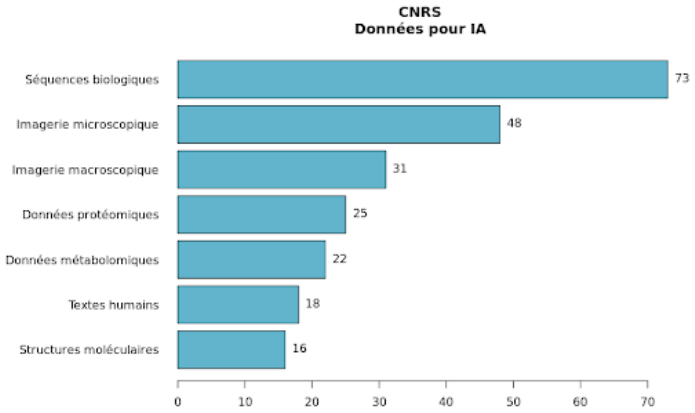
- **F5. POUR QUEL(S) TYPE(S) DE DONNÉES?**

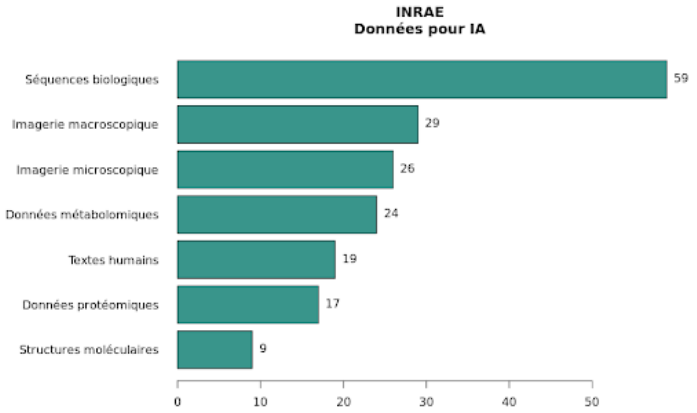


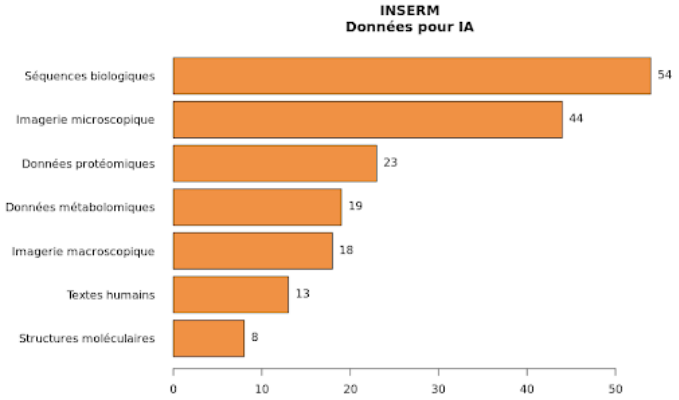
- En sommant les deux catégories de l'imagerie (micro et macro), celles-ci atteignent la même importance que les données des séquences biologiques.

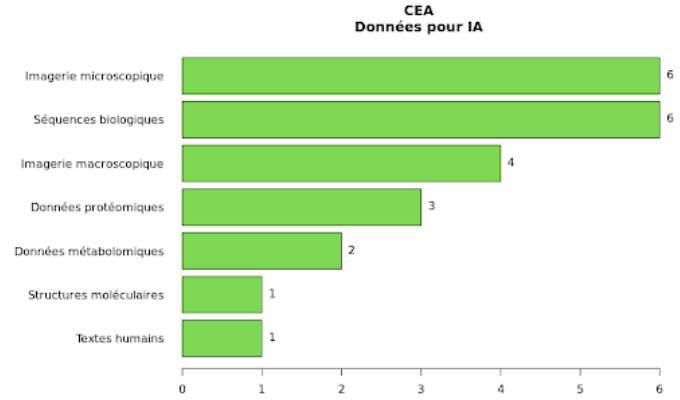
- Tradition d'apprentissage pour les motifs biologiques et les bases de connaissances en bioinformatique.

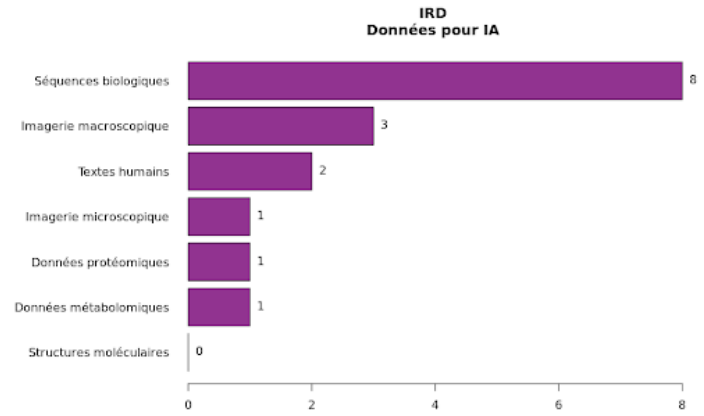
- Réponses cohérentes avec les domaines des tutelles.

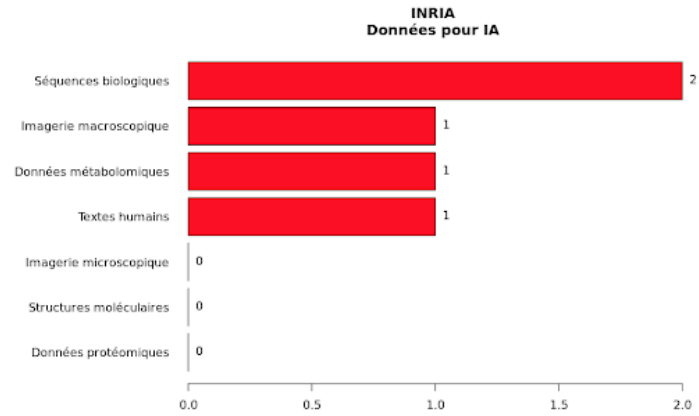

CNRS
Données pour IA
Séquences biologiques
Imagerie microscopique
Imagerie macroscopique
Données protéomiques
Données métabolomiques
Textes humains
Structures moléculaires
73
48
31
25
22
18
16
0
10
20
30
40
50
60
70

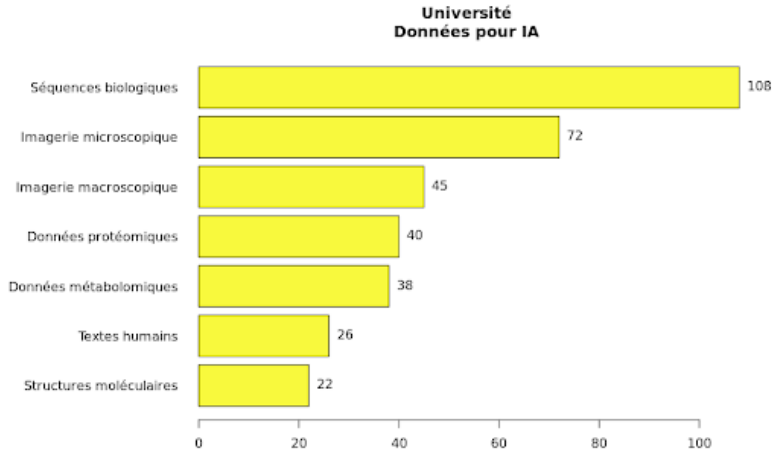

INRAE
Données pour IA
Séquences biologiques
Imagerie macroscopique
Imagerie microscopique
Données métabolomiques
Textes humains
Données protéomiques
Structures moléculaires
59
29
26
24
19
17
9
0
10
20
30
40
50


INSERM
Données pour IA
Séquences biologiques
Imagerie microscopique
Données protéomiques
Données métabolomiques
Imagerie macroscopique
Textes humains
Structures moléculaires
54
44
23
19
18
13
8
0
10
20
30
40
50

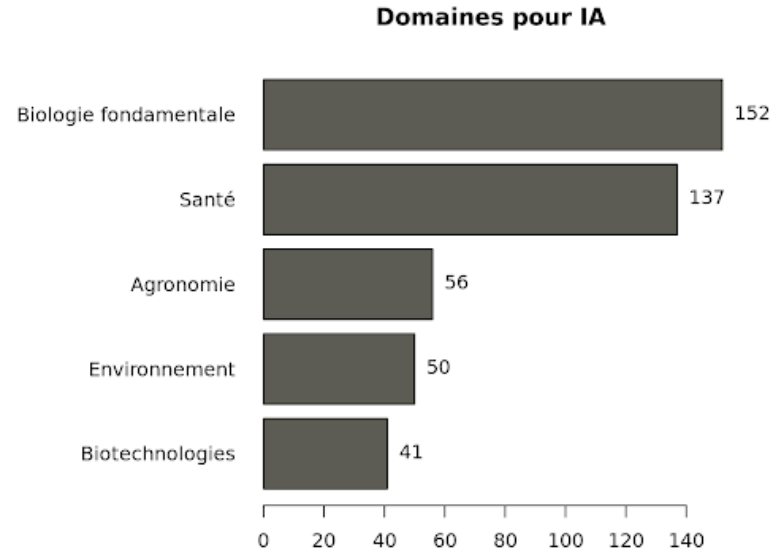

CEA
Données pour IA
Imagerie microscopique
Séquences biologiques
Imagerie macroscopique
Données protéomiques
Données métabolomiques
Structures moléculaires
Textes humains
6
6
4
3
2
1
1
0
1
2
3
4
5
6


IRD
Données pour IA
Séquences biologiques
Imagerie macroscopique
Textes humains
Imagerie microscopique
Données protéomiques
Données métabolomiques
Structures moléculaires
8
3
2
1
1
1
0
0
2
4
6
8

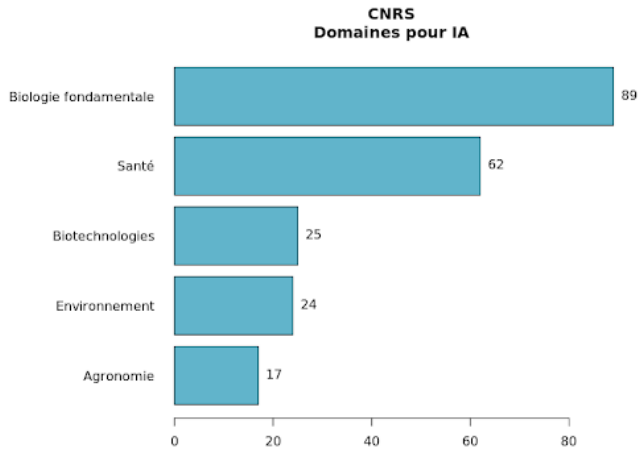

INRIA
Données pour IA
Séquences biologiques
Imagerie macroscopique
Données métabolomiques
Textes humains
Imagerie microscopique
Structures moléculaires
Données protéomiques
2
1
1
1
0
0
0
0.0
0.5
1.0
1.5
2.0

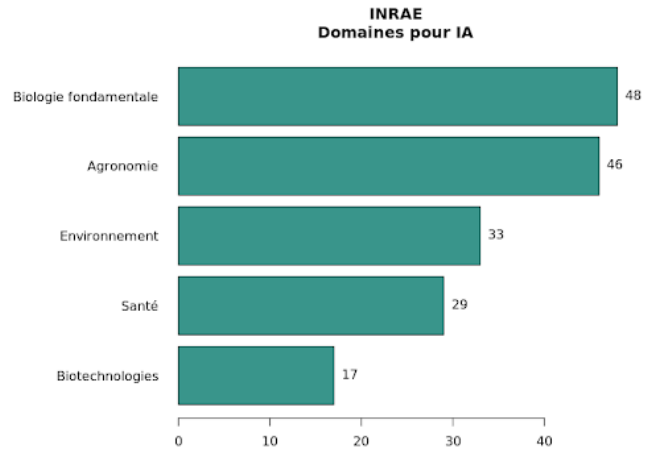

Université
Données pour IA
Séquences biologiques
Imagerie microscopique
Imagerie macroscopique
Données protéomiques
Données métabolomiques
Textes humains
Structures moléculaires
108
72
45
40
38
26
22
0
20
40
60
80
100

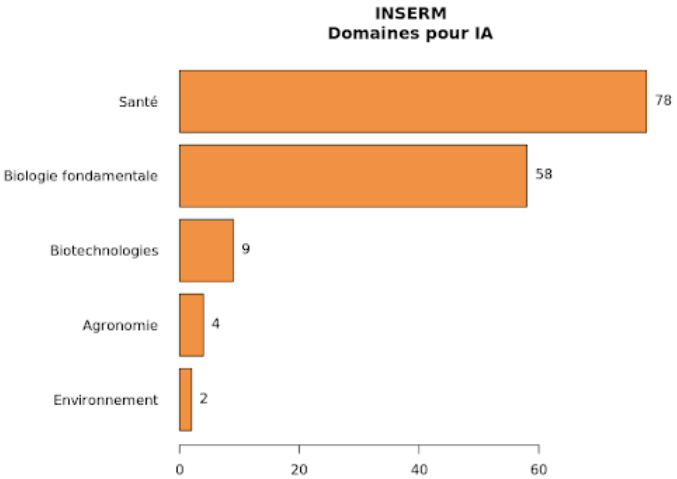
• F6. DANS QUELS DOMAINES D'APPLICATION?

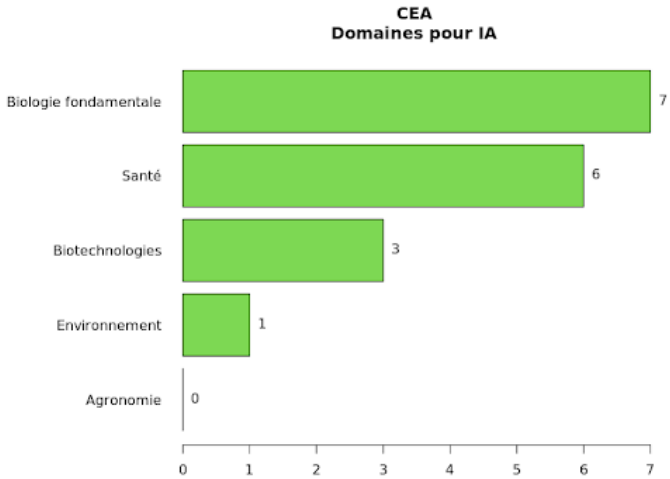


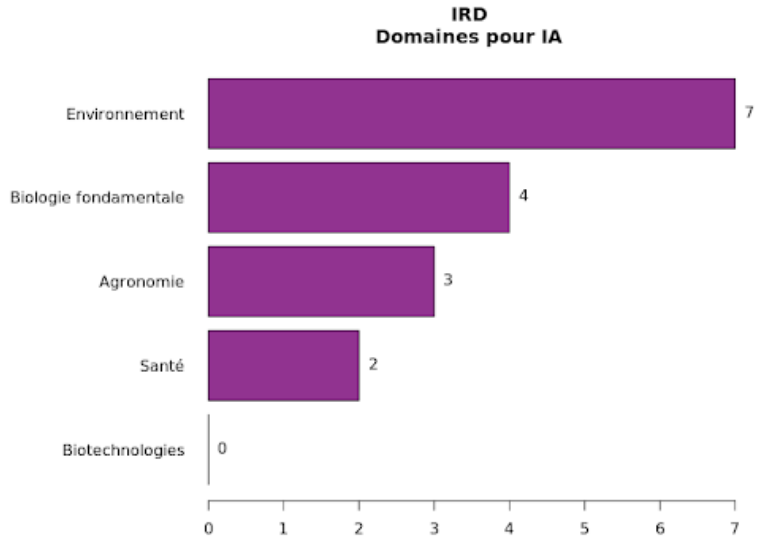
Les domaines d'application prioritaires sont la **biologie fondamentale, la santé et l'environnement.**

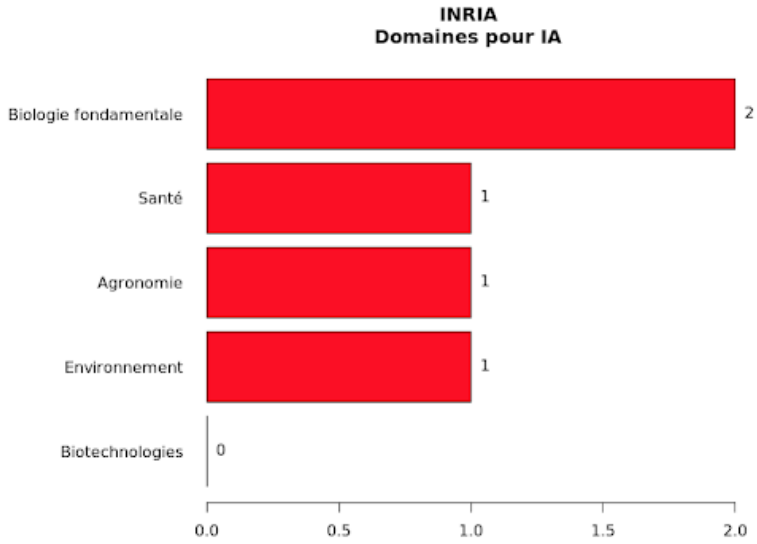


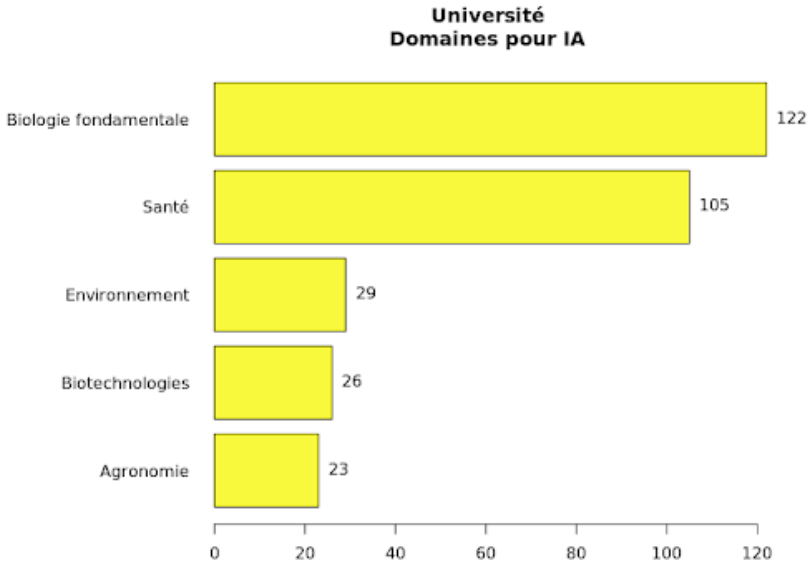




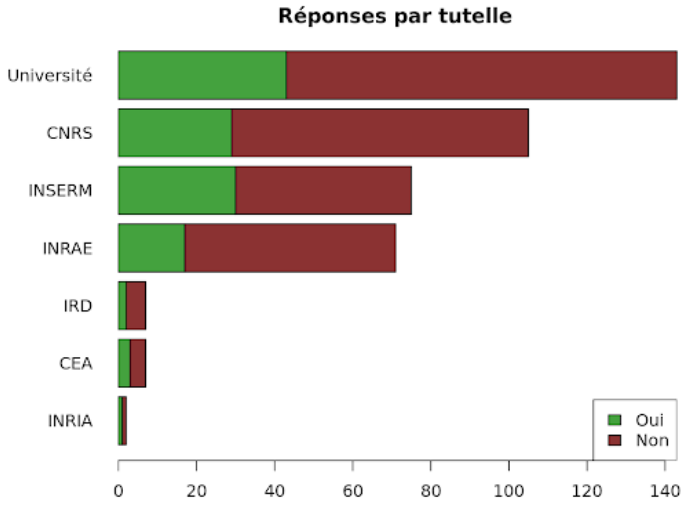


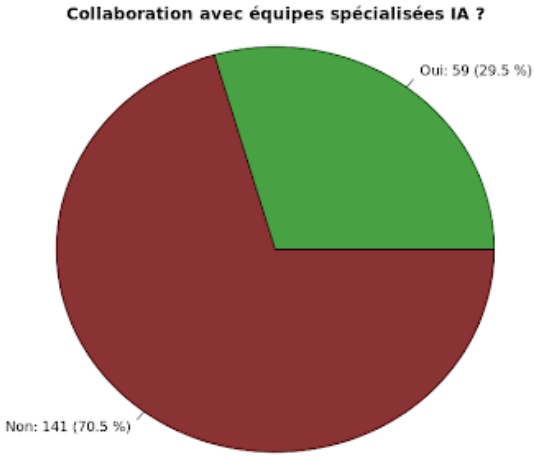

IRD
Domaines pour IA
Environnement
Biologie fondamentale
Agronomie
Santé
Biotechnologies
7
4
3
2
0
0
1
2
3
4
5
6
7


INRIA
Domaines pour IA
Biologie fondamentale
Santé
Agronomie
Environnement
Biotechnologies
2
1
1
1
0
0.0
0.5
1.0
1.5
2.0


Université
Domaines pour IA
Biologie fondamentale
Santé
Environnement
Biotechnologies
Agronomie
122
105
29
26
23
0
20
40
60
80
100
120

- **F7. COLLABOREZ-VOUS AVEC DES ÉQUIPES SPÉCIALISÉES EN IA POUR L'ANALYSE DE VOS DONNÉES BIOLOGIQUES?**

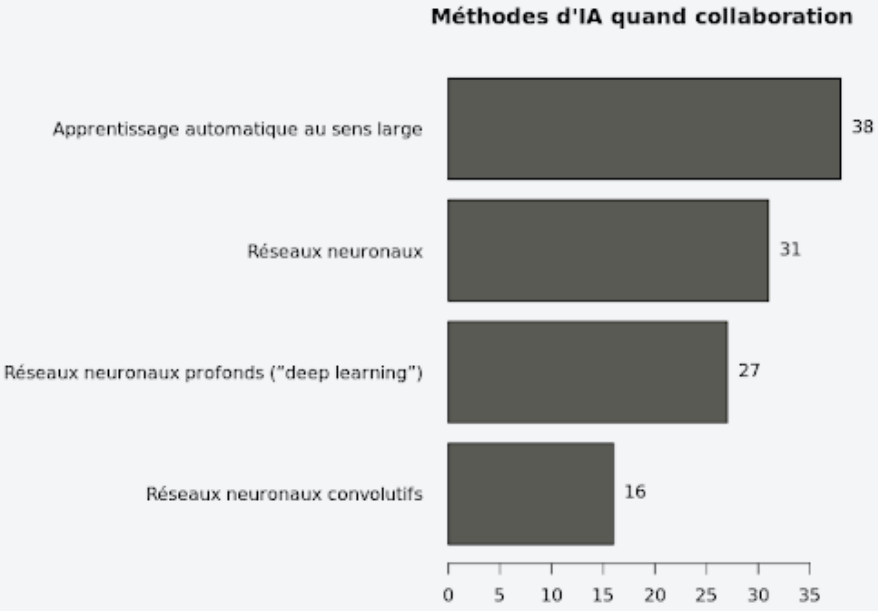




Le **faible nombre de "oui" (~ 30%)** pourrait indiquer que des outils "clés en main" sont disponibles (i.e. pas besoin de collaborer avec des experts de l'IA pour développer de nouvelles méthodes).

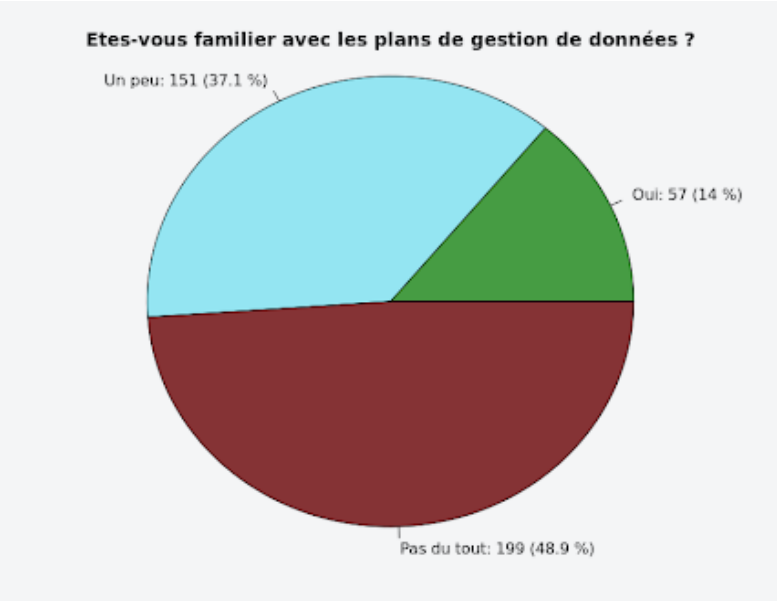
En croisant cette réponse "oui" avec celles de la question F2 (méthodes IA utilisées), on fait ressortir les méthodes qui incitent le plus à des collaborations:

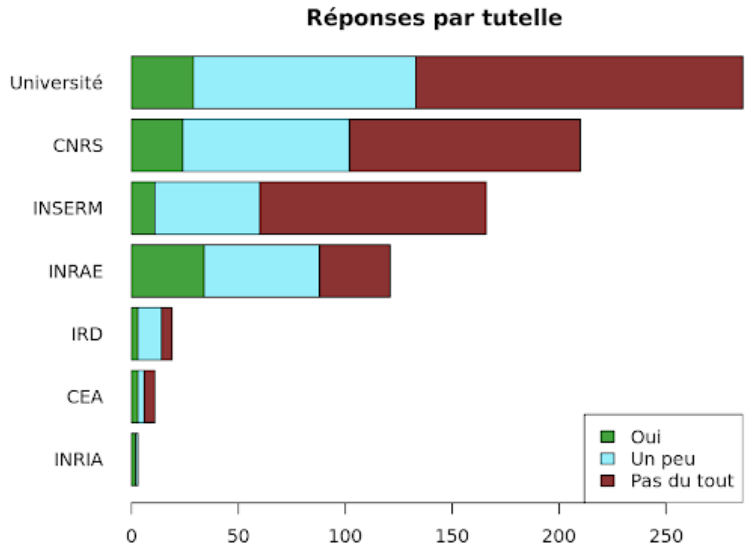
- Apprentissage automatique au sens large 38/122 (31,1%)
- Réseaux neuronaux 31/74 (41,9%)
- Réseaux neuronaux profonds 27/55 (49%)
- Réseaux neuronaux convolutifs 16/26 (61,5%)



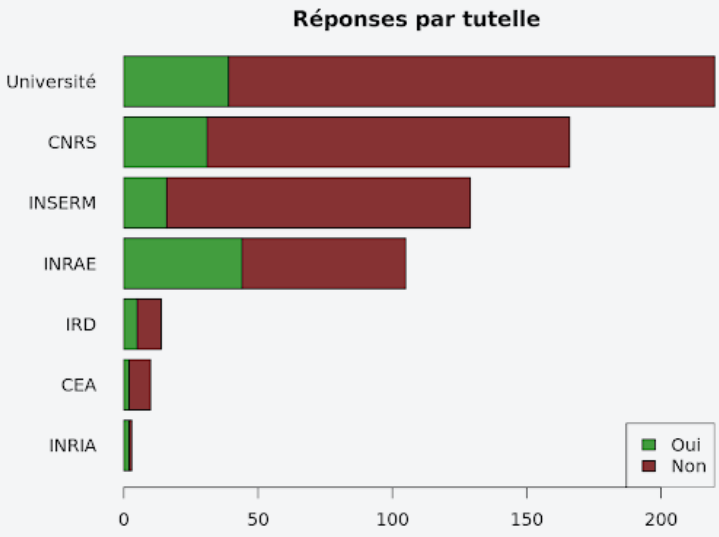
PARTIE G. PLAN DE GESTION DES DONNÉES (PGD) / DATA MANAGEMENT PLAN (DMP)

- **G1. ÊTES-VOUS FAMILIER AVEC LE CONCEPT DE DATA MANAGEMENT PLAN (DMP)?**

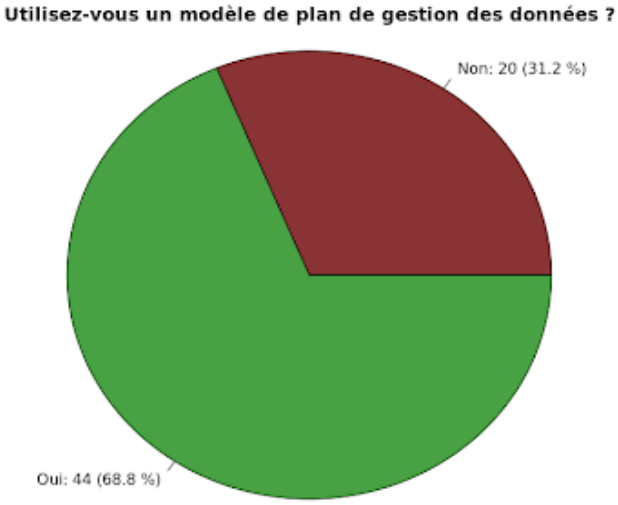


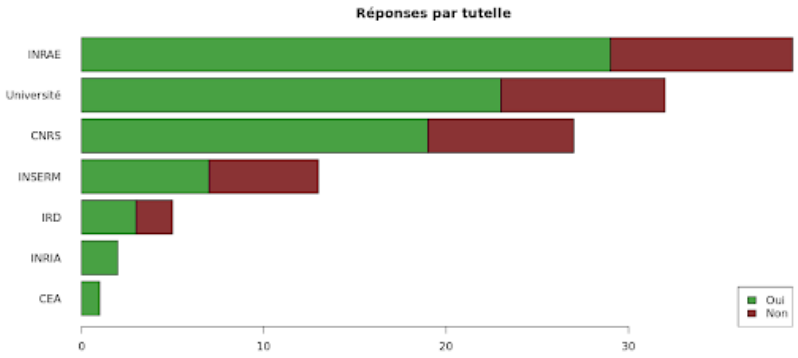


- **G2. VOTRE ÉQUIPE RECOURT-ELLE À DES DMP POUR SES PROJETS?**

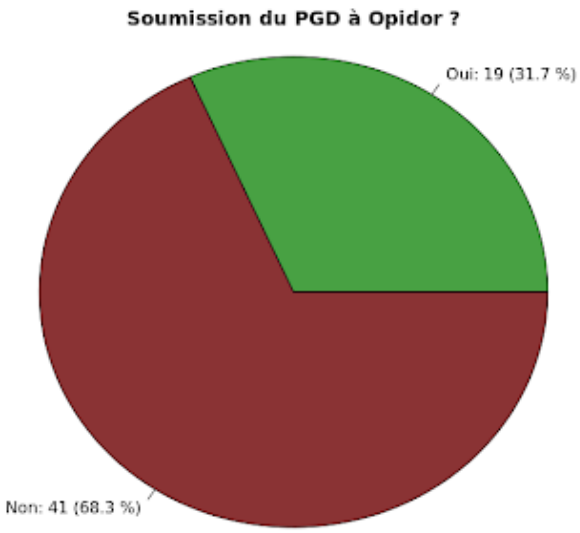


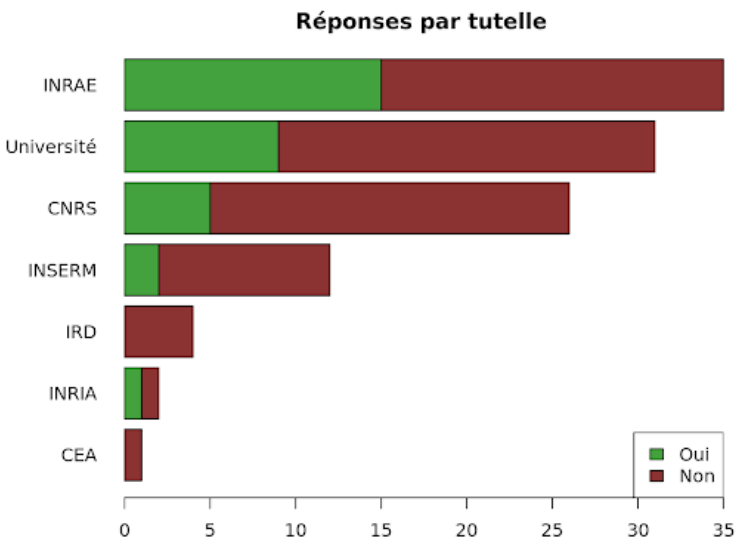
• G3. UTILISEZ-VOUS UN MODÈLE DE DMP ?





• G4. SOUMETTEZ-VOUS VOS DMP AU DÉPÔT OPIDOR?

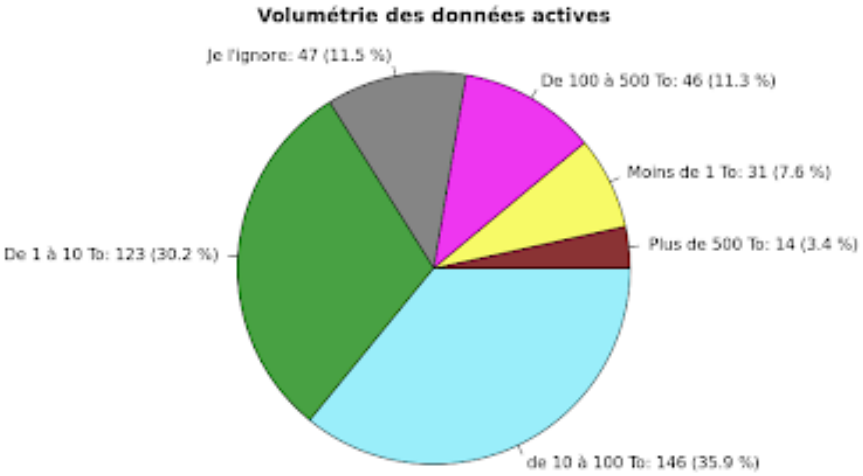


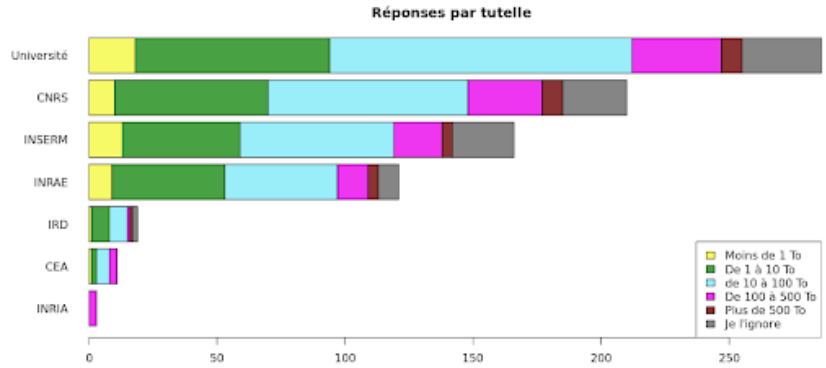


L'utilisation des DMP reste un point de progression majeur pour toutes les tutelles. L'IFB participe à cet effort en proposant régulièrement des formations dédiées. Le nombre important de demandes de participation à ces formations montre l'intérêt croissant pour cette notion.

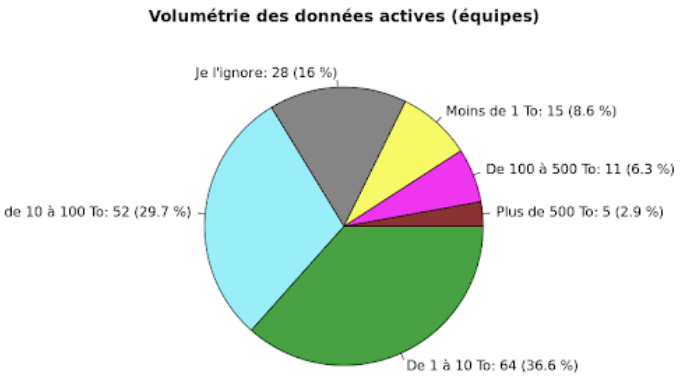
PARTIE H. INFRASTRUCTURE DE CALCUL ET DE STOCKAGE

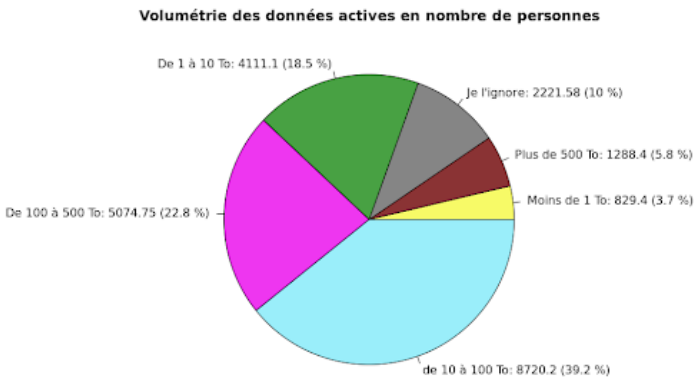
- **H1. QUELLE EST LA VOLUMÉTRIE ACTUELLE DES DONNÉES ACTIVES STOCKÉES PAR L'UNITÉ?**

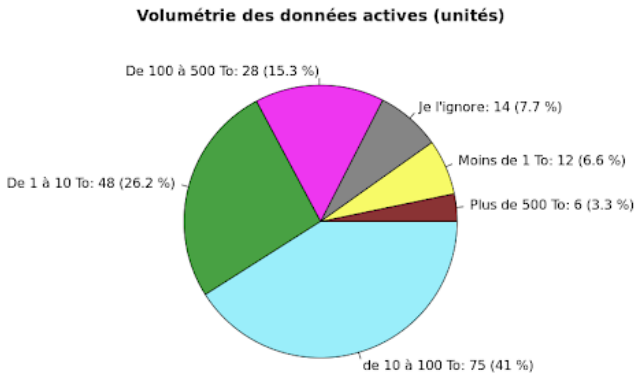




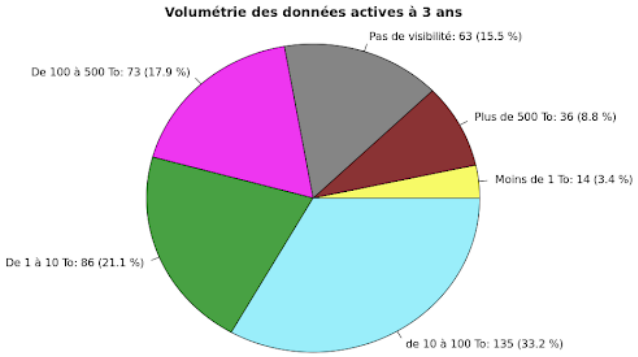
Globalement s'on somme les résultats on arrive à plusieurs dizaines de Pétaoctets de données. Cela confirme l'**énorme besoin en volumétrie**. L'IFB ne pourra pas résoudre seul ce problème de stockage.

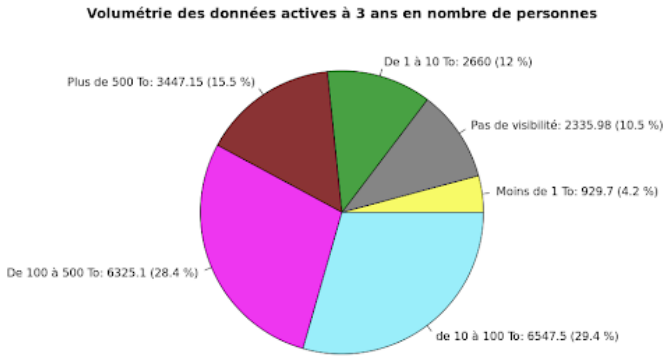


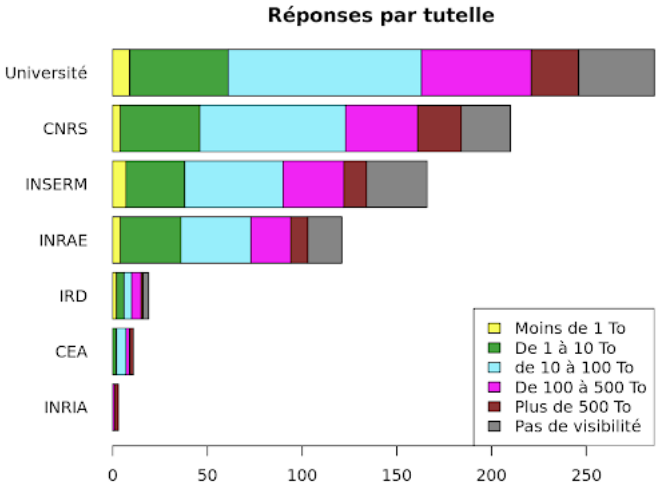




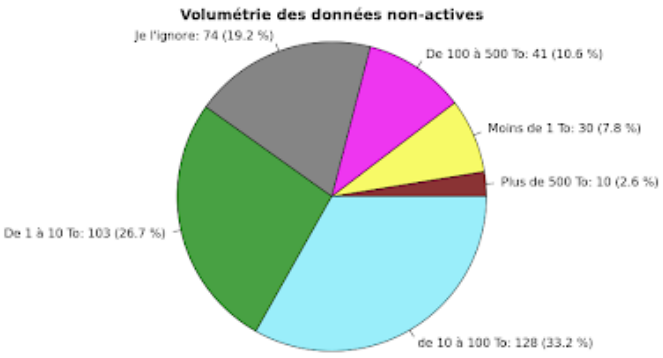
- **H2. QUELLE EST VOTRE VISIBILITÉ SUR VOS BESOINS EN STOCKAGE DE DONNÉES ACTIVES À 3 ANS?**

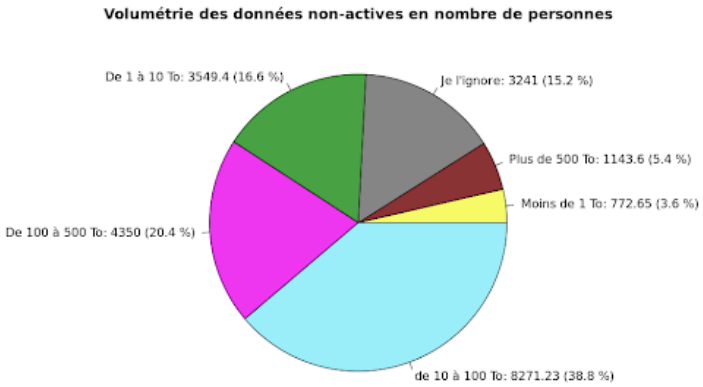


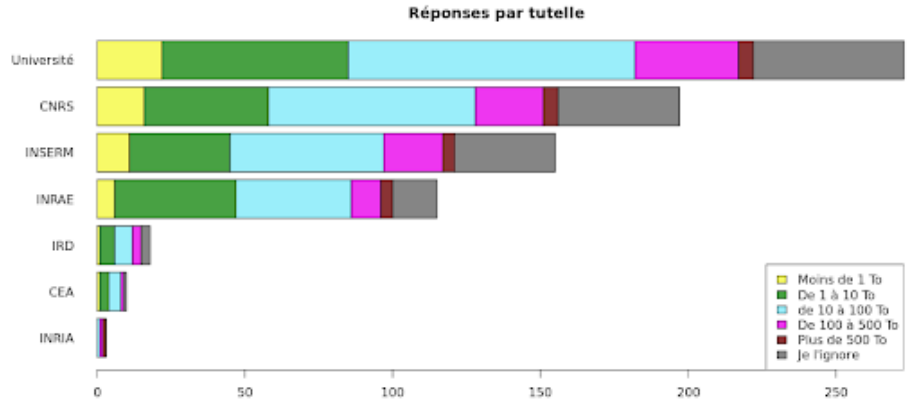




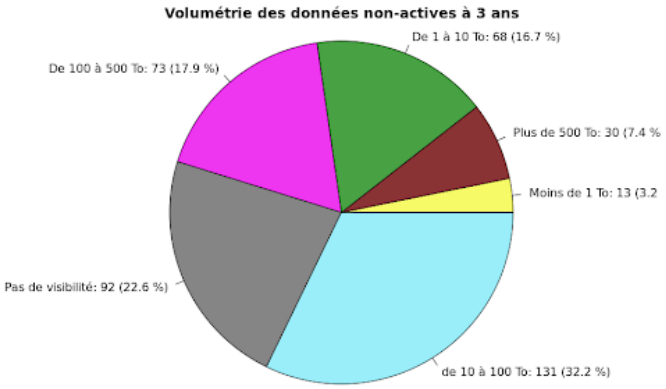
- **H3. QUELLE EST LA VOLUMÉTRIE ACTUELLE DES DONNÉES NON-ACTIVES STOCKÉES PAR L'UNITÉ?**

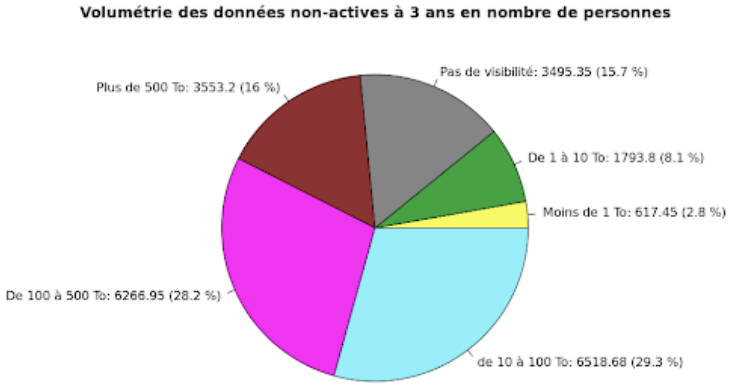


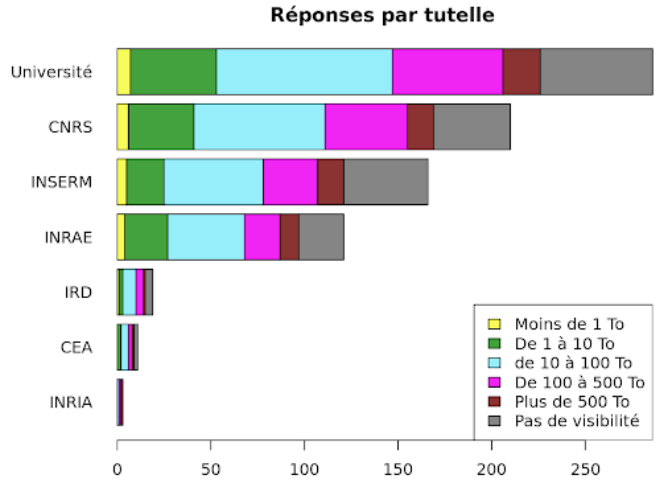




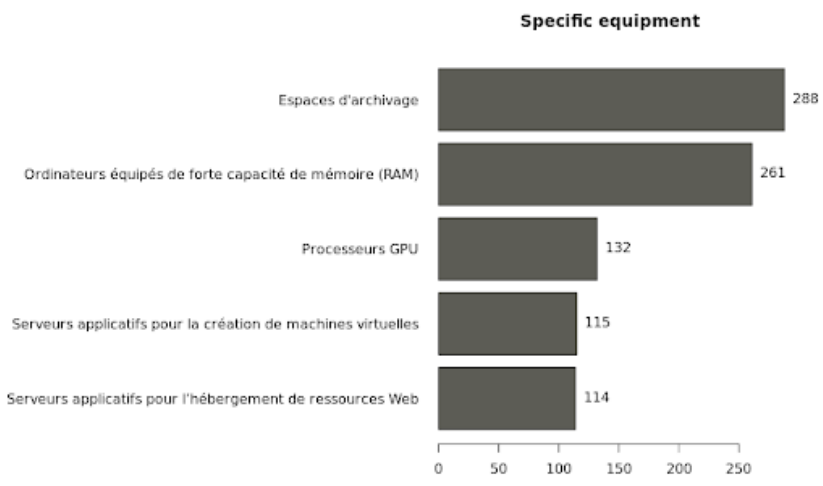
- **H4. QUELLE EST VOTRE VISIBILITÉ SUR VOS BESOINS EN STOCKAGE DE DONNÉES NON-ACTIVES À 3 ANS?**



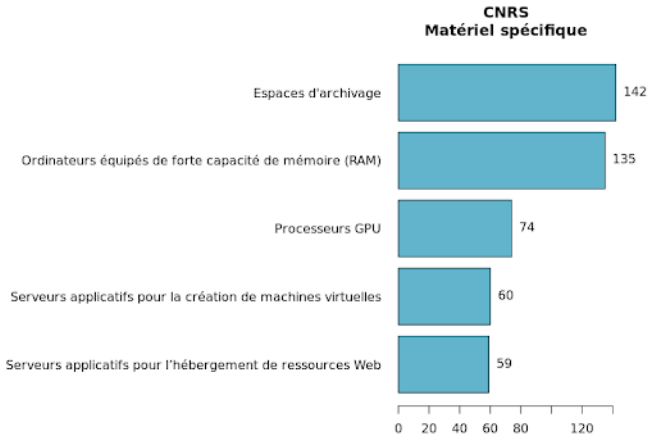


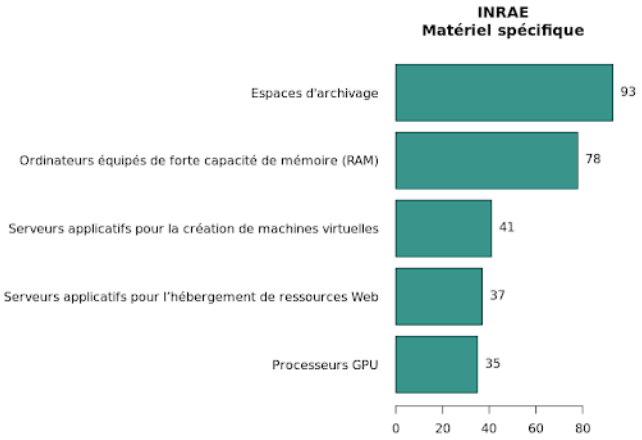


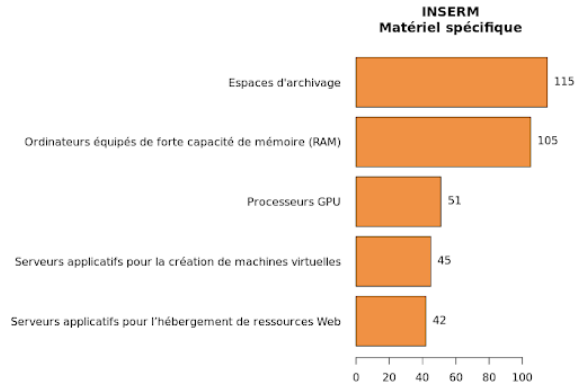
- **H5. ANTICIPEZ-VOUS DES BESOINS SPÉCIFIQUES EN TERME DE MATÉRIEL?**

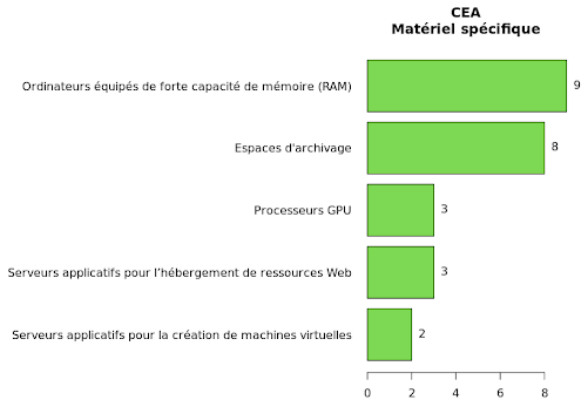


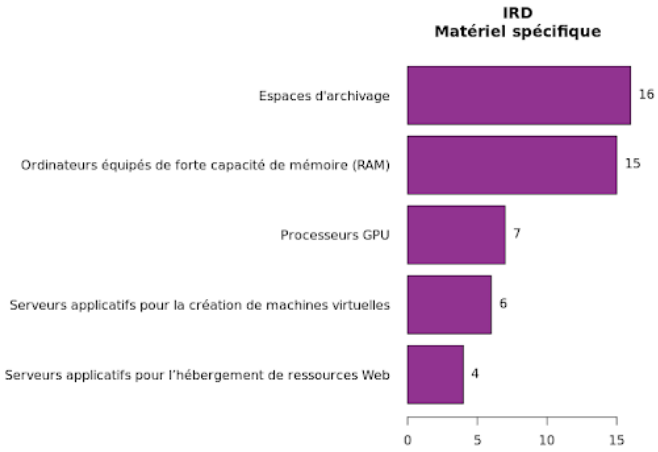
- Les résultats sont cohérents suivant les tutelles.
- Importante demande d'ordinateurs avec forte capacité de RAM.
- A noter que les besoins exprimés reflètent des cas très diversifiés (calcul, stockage, machine individuelle...).

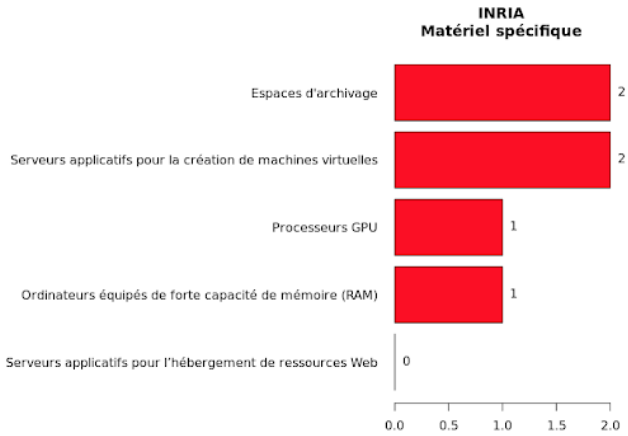

CNRS
Matériel spécifique
Espaces d'archivage
142
Ordinateurs équipés de forte capacité de mémoire (RAM)
135
Processeurs GPU
74
Serveurs applicatifs pour la création de machines virtuelles
60
Serveurs applicatifs pour l'hébergement de ressources Web
59
0 20 40 60 80 120

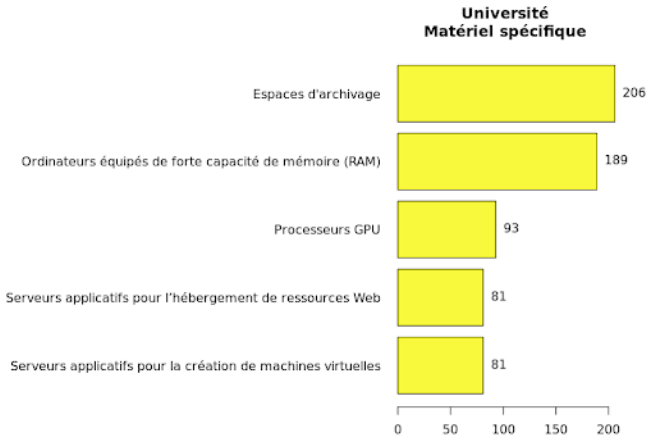

INRAE
Matériel spécifique
Espaces d'archivage
93
Ordinateurs équipés de forte capacité de mémoire (RAM)
78
Serveurs applicatifs pour la création de machines virtuelles
41
Serveurs applicatifs pour l'hébergement de ressources Web
37
Processeurs GPU
35
0 20 40 60 80


INSERM
Matériel spécifique
Espaces d'archivage
115
Ordinateurs équipés de forte capacité de mémoire (RAM)
105
Processeurs GPU
51
Serveurs applicatifs pour la création de machines virtuelles
45
Serveurs applicatifs pour l'hébergement de ressources Web
42
0 20 40 60 80 100

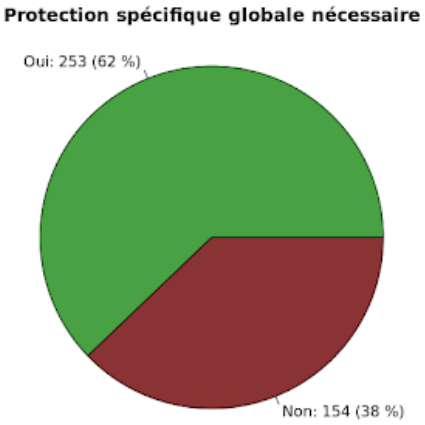

CEA
Matériel spécifique
Ordinateurs équipés de forte capacité de mémoire (RAM)
9
Espaces d'archivage
8
Processeurs GPU
3
Serveurs applicatifs pour l'hébergement de ressources Web
3
Serveurs applicatifs pour la création de machines virtuelles
2
0 2 4 6 8

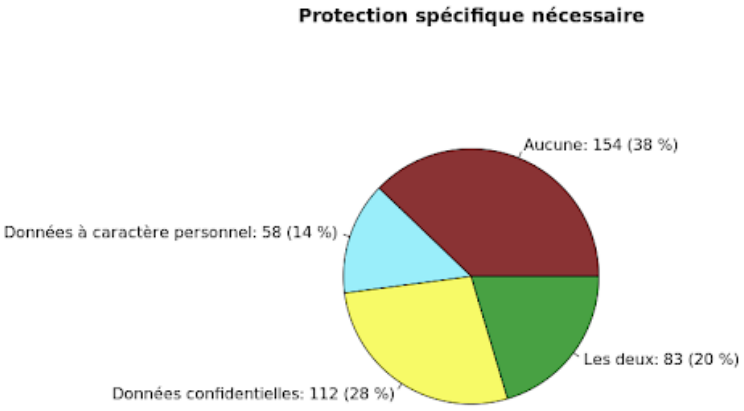

IRD
Matériel spécifique
Espaces d'archivage
16
Ordinateurs équipés de forte capacité de mémoire (RAM)
15
Processeurs GPU
7
Serveurs applicatifs pour la création de machines virtuelles
6
Serveurs applicatifs pour l'hébergement de ressources Web
4
0 5 10 15


INRIA
Matériel spécifique
Espaces d'archivage
2
Serveurs applicatifs pour la création de machines virtuelles
2
Processeurs GPU
1
Ordinateurs équipés de forte capacité de mémoire (RAM)
1
Serveurs applicatifs pour l'hébergement de ressources Web
0
0.0 0.5 1.0 1.5 2.0


Université
Matériel spécifique
Espaces d'archivage
206
Ordinateurs équipés de forte capacité de mémoire (RAM)
189
Processeurs GPU
93
Serveurs applicatifs pour l'hébergement de ressources Web
81
Serveurs applicatifs pour la création de machines virtuelles
81
0 50 100 150 200

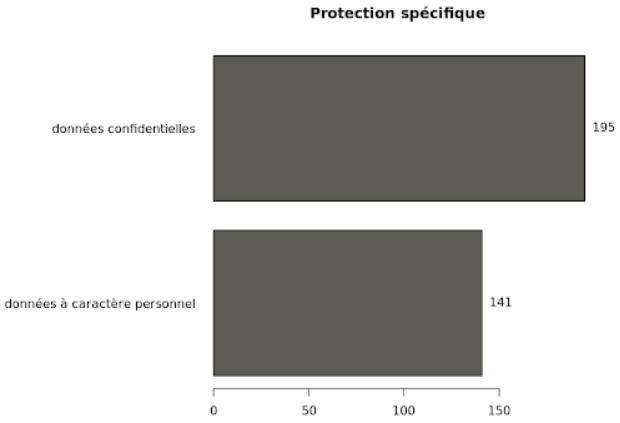
• H6. MANIPULEZ-VOUS DES DONNÉES NÉCESSITANT UNE PROTECTION SPÉCIFIQUE?

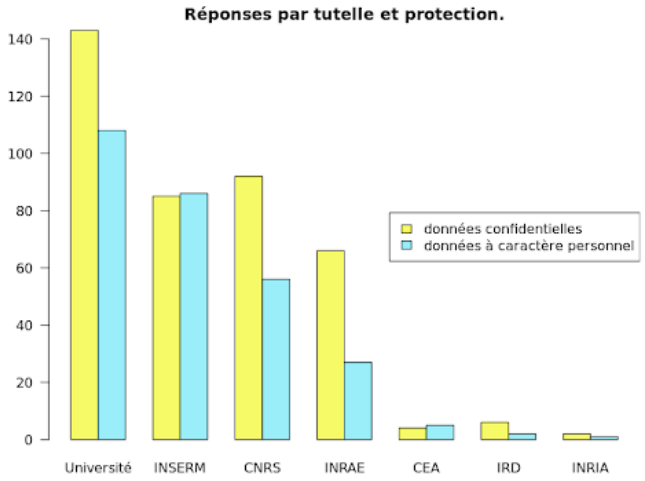




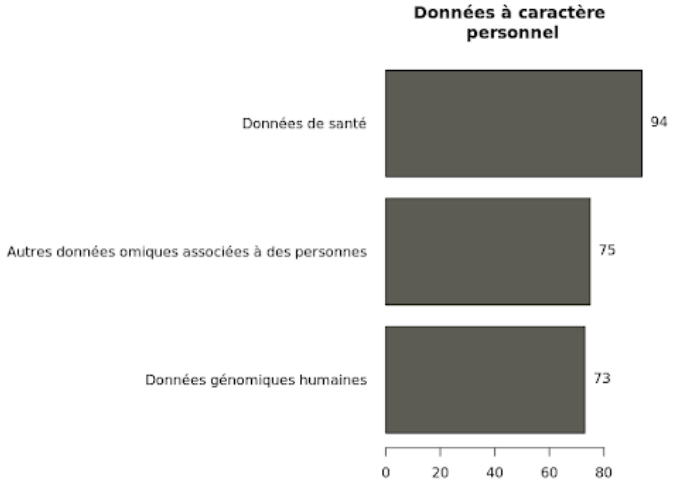
- Que ce soit pour des données confidentielles et/ou à caractère personnel, un grand nombre d'équipes sont concernées.

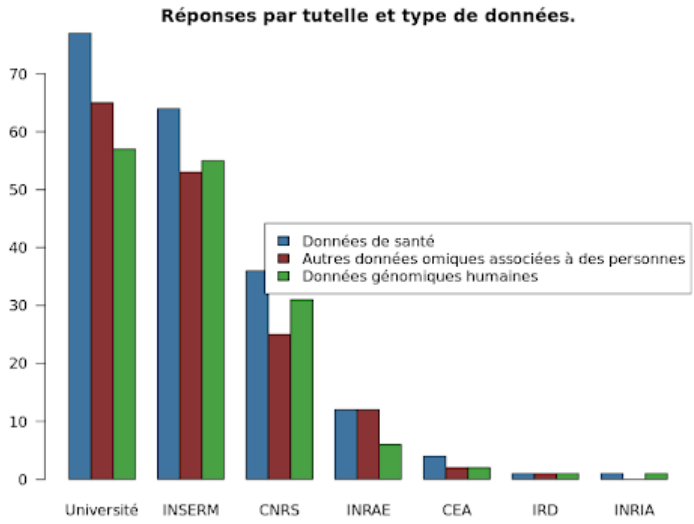
- A noter, les réponses sur les protections spécifiques ne sont pas exclusives.

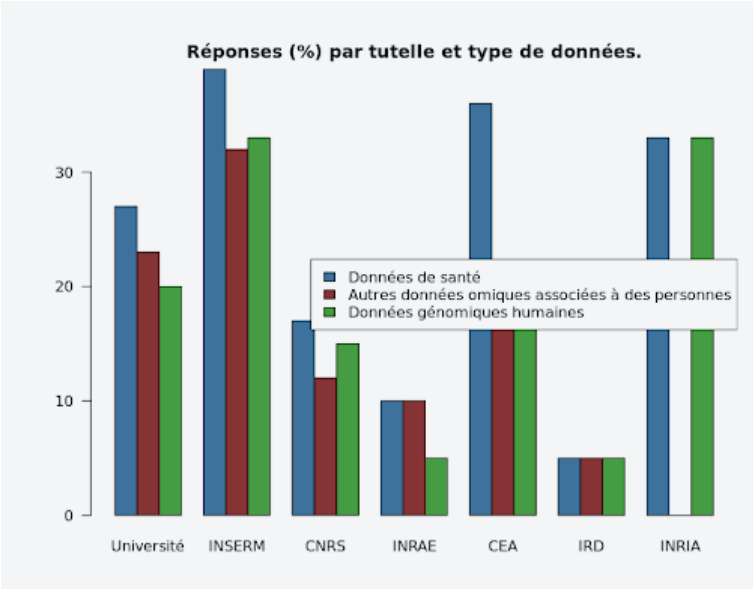




• H7. PRÉCISION DES "DONNÉES À CARACTÈRE PERSONNEL"





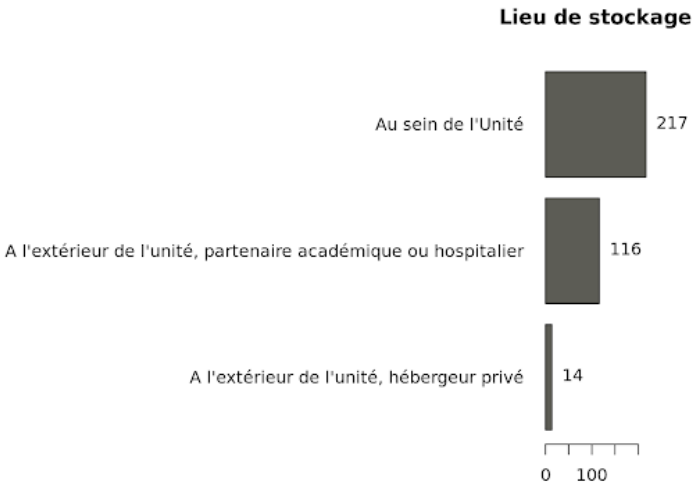


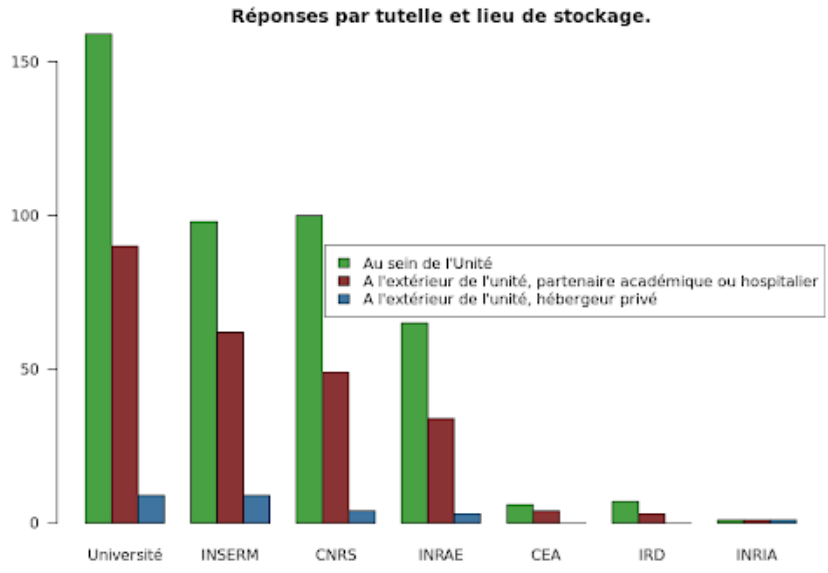
• H8. CES DONNÉES SONT-ELLES STOCKÉES… ?

(attention cette question ne concerne que les réponses "Oui" à la question H6)

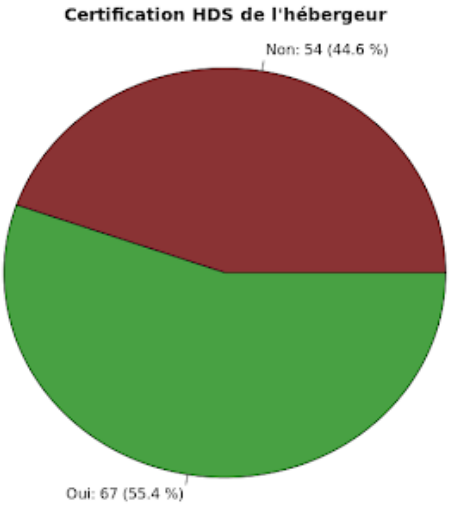
- **Résultat marquant :** la moitié des unités/équipes qui ont répondu (217/407) gèrent en interne leurs données.

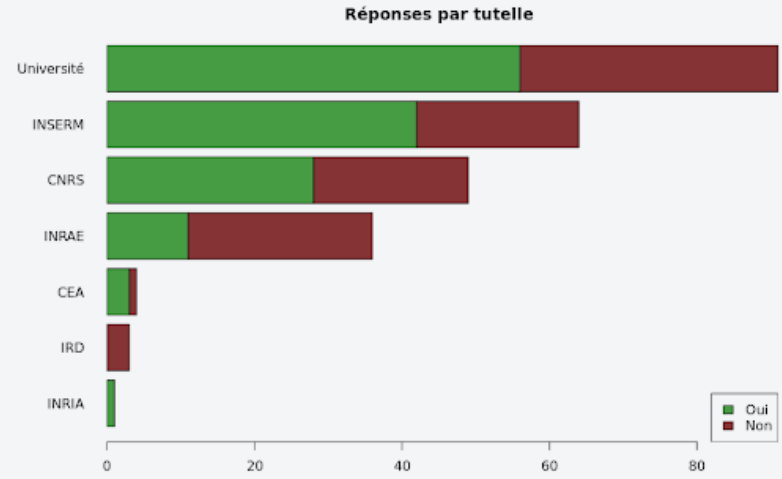
- Une explication possible pourrait être le manque d'infrastructure en capacité de garantir la sécurité de ce type de données.

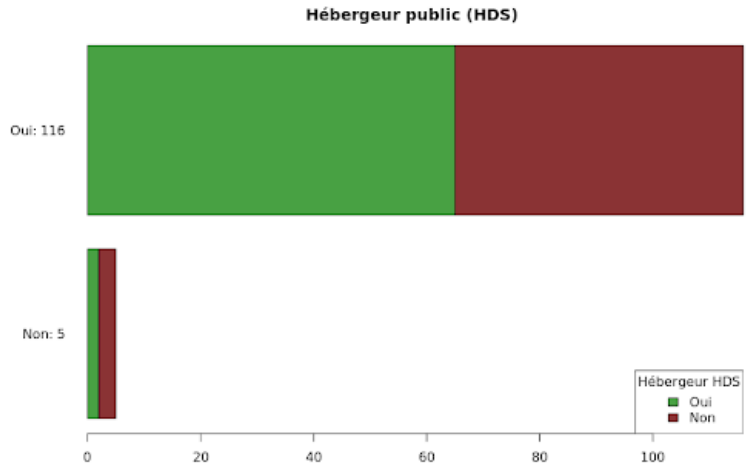


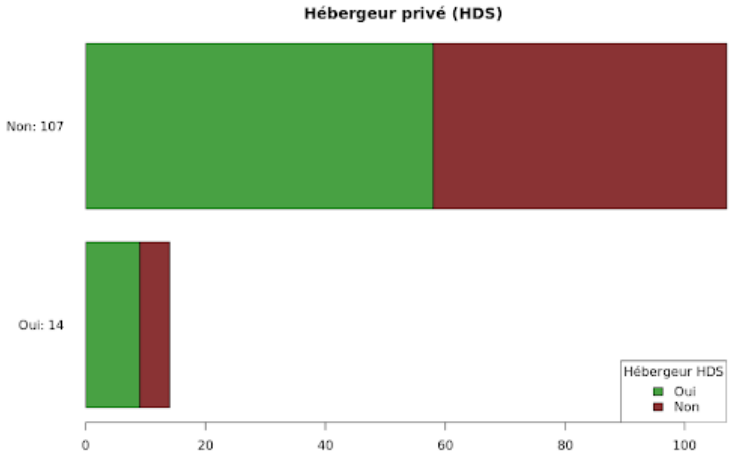


- **H9. PRÉCISION SUR L'HÉBERGEUR EXTÉRIEUR : EST-IL CERTIFIÉ "HÉBERGEUR DE DONNÉES DE SANTÉ (HDS)?**

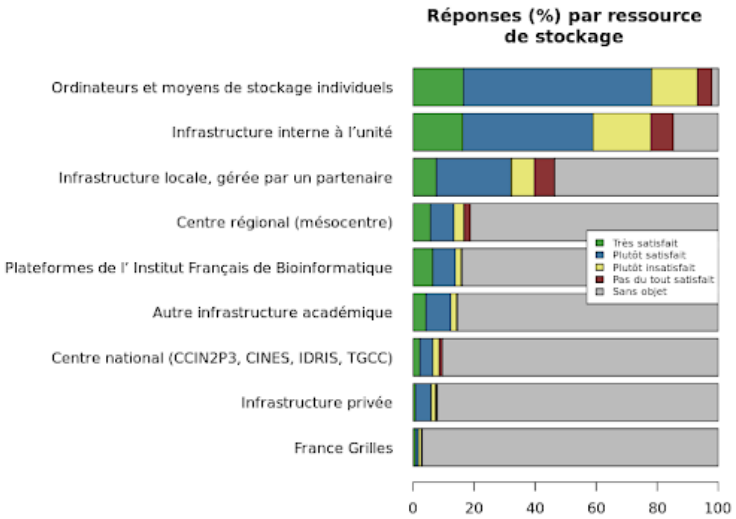


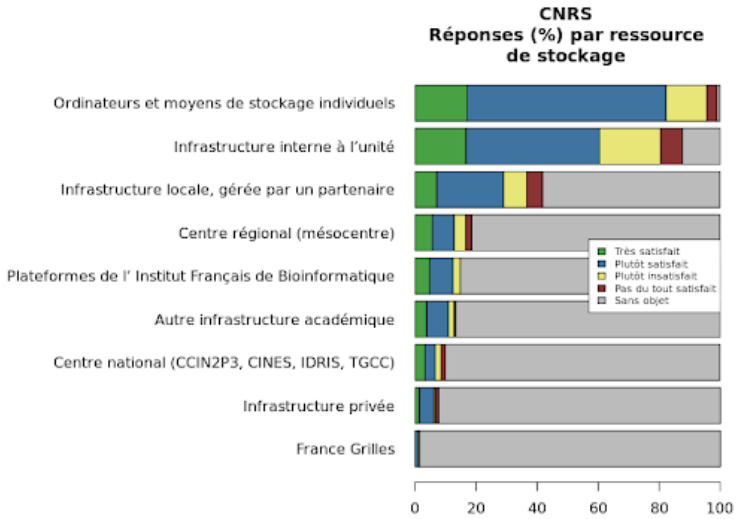




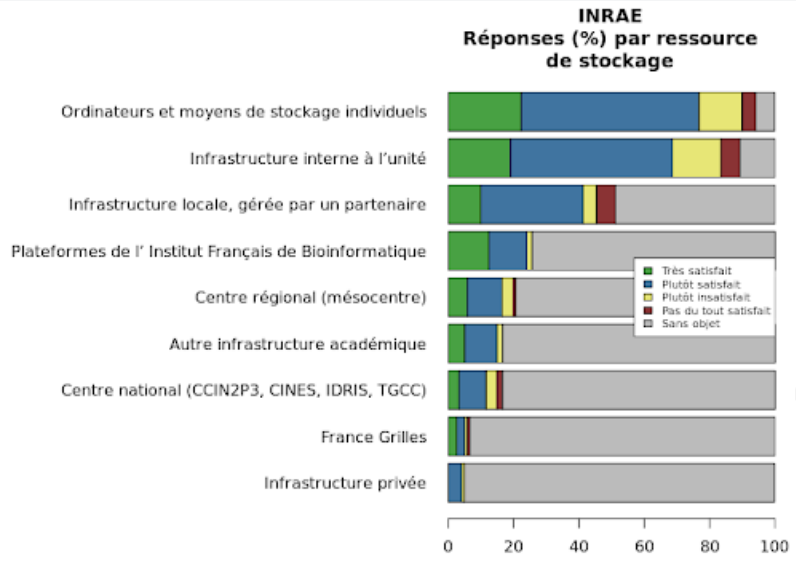


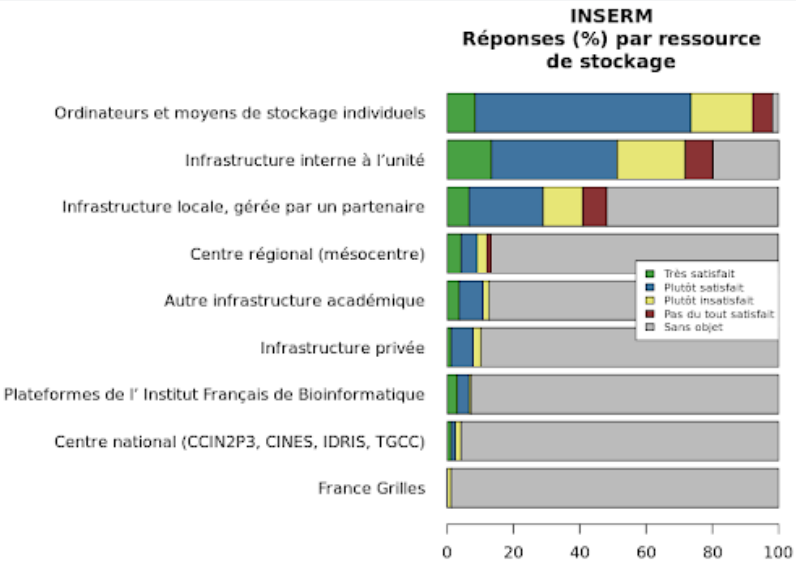
- **H10. INDIQUEZ LES RESSOURCES DE STOCKAGE UTILISÉES AU SEIN DE L'UNITÉ/ÉQUIPE ET UNE ESTIMATION DU TAUX DE SATISFACTION POUR CHACUNE**

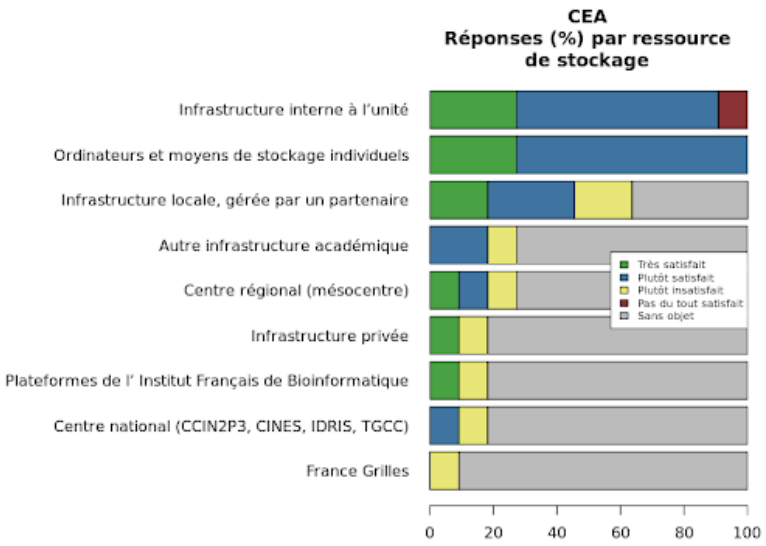


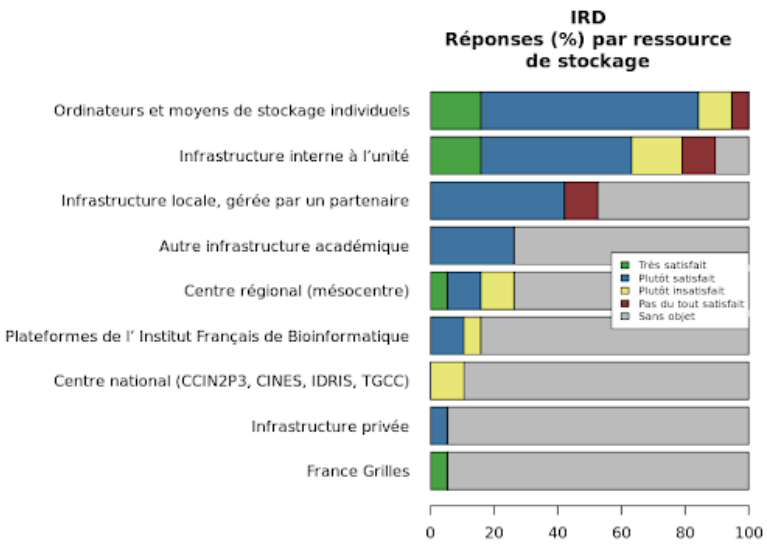


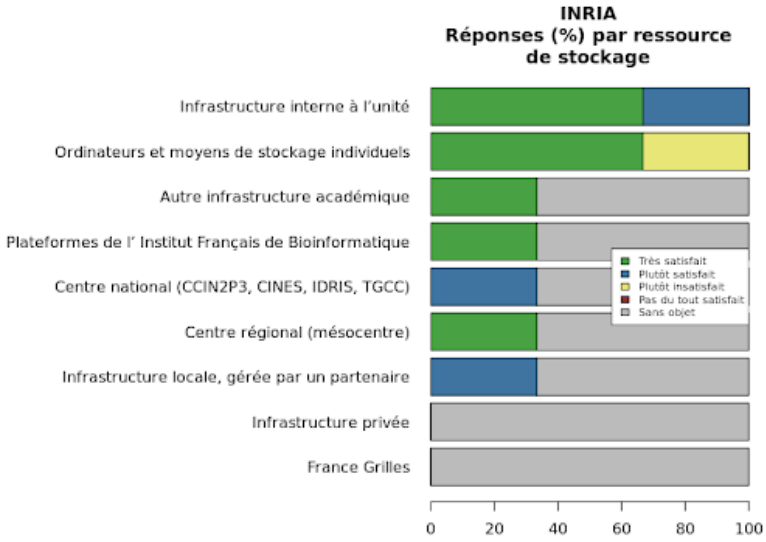
- Ordinateurs et moyens de stockage individuels: quasi tout le monde a répondu et est plutôt satisfait ou très satisfait (malgré tout, quasiment un quart ne sont pas satisfait).
- Les réponses montrent une domination de l'utilisation des ressources locales de l'unité pour le stockage.
- Plateformes de l'IFB : 16% des réponses mais avec un bon niveau de satisfaction.

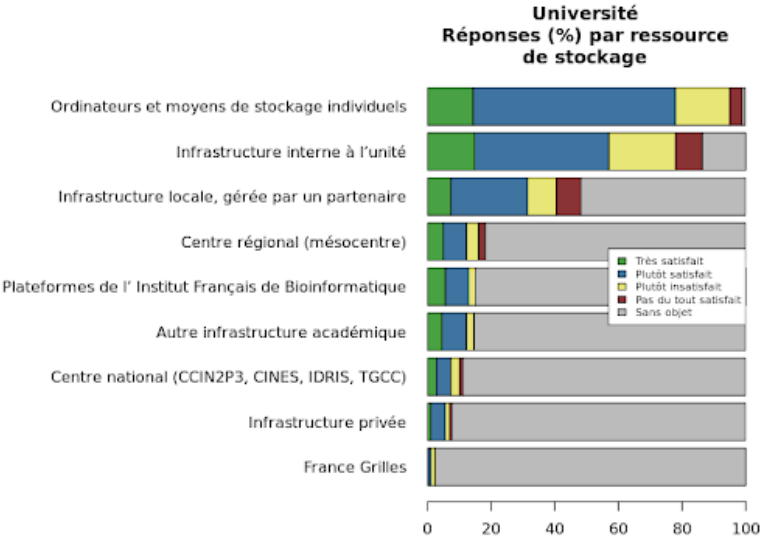












Taux de satisfaction pour les ressources de stockage

STOCKAGES	Ordinateurs et moyens de stockage individuels	Infra. interne à l'unité	Infra. locale, gérée par un partenaire	Centre régional	Centre national	France Grille	Plateformes de l'IFB	Autre infra. académique	Infra. privée
Taux d'utilisation (/réponses)	97,8% (398/407)	85,3% (347/407)	46,4% (189/407)	18,7% (76/407)	9,6% (39/407)	2,9% (12/407)	16,0% (65/407)	14,5% (59/407)	7,9% (32/407)
Très satisfait (/utilisation)	16,8%	19,0%	16,4%	30,3%	23,1%	25,0%	40,0%	28,8%	9,4%
Plutôt satisfait (/utilisation)	63,1%	50,1%	52,9%	40,8%	43,6%	33,3%	46,2%	55,9%	65,6%
Plutôt insatisfait (/utilisation)	15,3%	22,2%	16,4%	18,4%	23,1%	33,3%	12,3%	13,6%	18,8%
Pas du tout satisfait (/utilisation)	4,8%	8,6%	14,3%	10,5%	10,3%	8,3%	1,5%	1,7%	6,3%

H11. STOCKAGE : MOTIFS D'INSATISFACTION ET AUTRES COMMENTAIRES

- "Pas assez d'espace disponible et pas de modalités de sauvegarde standardisée/automatisée."
- "Au niveau de l'infrastructure locale : pas de vision sur la limite de stockage possible."
- "Quelle que soit le type d'infrastructure, nous sommes extrêmement limités... Au prix du To, c'est vraiment dommage."
- "Serveurs vieillissants et manquant d'espaces disque + manque d'administrateur système."
- "Le partage de gros fichier est un problème vu que nous n'avons pas accès aux services type dropbox et que mycore a montré quelques faiblesses de fonctionnement. Le stockage et le back up se fait de manière manuelle et individuelle pour chaque équipe et n'est pas automatisé comme on pourrait le souhaiter."
- "Incapacité de l'établissement hébergeur à mettre en place un système de partage de données + absence de plan de gestion de données de la recherche à l'échelon national."
- "Moyen de stockage au sein de l'équipe entraîne un coût financier. Les centres de calculs régionaux sont payants. Les centres nationaux sont très biens mais parfois surchargés."
- "Les modes de stockage individuels sont redondants et leur stabilité/sécurité n'est pas assurée."
- "En fait, nous n'accédons pas à une vraie offre de stockage: nous sommes obligés de bricoler au sein de l'équipe (NAS et disques)."
- "Pour le cloud: la solution proposée par les tutelles (OwnCloud CNRS) n'est pas fonctionnelle (volume et fiabilité), les utilisateurs se tournent vers des solutions privées et personnelles (DropBox, Google Drive...)."
- "Accessibilité hors site."
- "Aucun accès au mésocentre Nouvelle-Aquitaine. Absence de gestion de serveurs propres au stockage par la DSI de l'université. Capacité trop restreinte à nos besoins."
- "Pas de solution pour stocker depuis qu'on nous a enlevé la possibilité d'avoir un NAS."
- "Le volume de données augmente très fortement, nos systèmes tendent à saturation."
- "C'est vraiment dommage de ne pas avoir accès à dropbox, google drive etc. pour collaborer sur les documents etc."
- "Coût et impact écologique."
- "Stockage sur des supports différents (université, APHP)."
- "Stockage sur disque dur externe nécessite le branchement/débranchement, et est accessible au vol (et détérioration). Le stockage individuel est limité en espace. Le stockage en infrastructure interne est en cours de réaménagement, et surtout n'est pas dédié à du stockage scientifique (architecture Windows, place limitée). L'infrastructure locale est tout juste mise en place, on ne sait donc pas encore si cela nous satisfera sur le plus long terme."
- "Nos serveurs internes ne sont pas accessibles depuis l'extérieur (politique de l'établissement d'accueil), ce qui pose régulièrement des problèmes et nécessite de trouver des solutions alternatives."
- "Manque de ressources institutionnelles adaptées pour le stockage et la réplication/sécurisation des données"
- "Malcommode et onéreux"

- "pas assez de stockage, doit acheter sur mes fonds propres de recherche des espaces de stockage pour toute l'Unité"
- "Problèmes de rationalisation des moyens de stockage face à l'augmentation constante des besoins. Le coût du stockage devient trop lourd pour le laboratoire (80€/To/an au Datacenter AMU). Manque d'une politique de gestion du cycle de vie de la donnée."
- "Conditions d'accès (certificats personnels et limitation de ressources)"
- "Nous avons essayé d'utiliser le datacenter de Toulouse INRA; les problèmes: 1. problème de bandes passantes pour transférer nos données depuis l'unité vers le data center de Toulouse; 2. si on y hébergeait nos serveurs, on serait obligé de se rendre sur place en cas de panne physique."
- "Difficultés à obtenir des ressources financières et humaines pour opérer un stockage pérenne en interne"
- "données non sécurisées, pas de sauvegarde automatisées"
- "manque d'information sur les opportunités"
- "Les infrastructures de proximité ne veulent pas prendre en charge la sauvegarde de données"
- "Manque de solution intégrée ou de visibilité sur la disponibilité des data."
- "On manque d'infrastructures internes pour le stockage"
- "manque les compétences et les ressources humaines pour gérer un serveur de stockage"
- "Pas de solution unique"
- "Difficulté à pérenniser"
- "- Gestion limitée des droits sur les isques du data-center (MIGALE). Stockage sur center national : Interruption de service trop importante"
- "pas de sécurité pour les moyens individuels; Pas de support DSI pour les infrastructures d'Unité"
- "Coût de l'hébergeur privé et manque de souplesse (nombre d'utilisateurs prédéfini, exportations de résultats limitées, etc.). Lenteur de la connexion à distance"
- "Manque de stockage en infrastructure interne à l'unité. Manque d'information sur les possibilités de stockage national."
- "pas de stratégie commune, multiplication des disques durs externes"
- "la gestion des sauvegardes est inexistantes au sein de l'IBPS et le réseau d'une qualité pauvre"
- "Fiabilité hardware des systèmes. Prix. Temps nécessaire à la maintenance"
- "Aucune solution de stockage/sauvegarde"
- "- data lost on Galaxy servers"
- "Actuellement nous n'avons pas de personnel pour développer nos besoin en bioinformatique. C'est grâce à une unité voisine que nous avons de l'espace d'archivage."
- "La sauvegarde sur les PC ou disque dur en local permet un accès rapide mais n'assure pas bien la sécurité des données sur le long terme. L'utilisation de serveurs locaux permet un accès rapide, une sauvegarde sur le long terme mais nécessite un informaticien sur site. L'utilisation de serveurs distants permet une très bonne sauvegarde sur le long terme avec duplication des données mais l'accès n'est pas assez rapide pour de gros fichiers ce qui oblige à les dupliquer en local, ce qui n'est pas satisfaisant."
- "Aucune aide apportée par Gustave Roussy"
- "ordinateurs obsolètes"

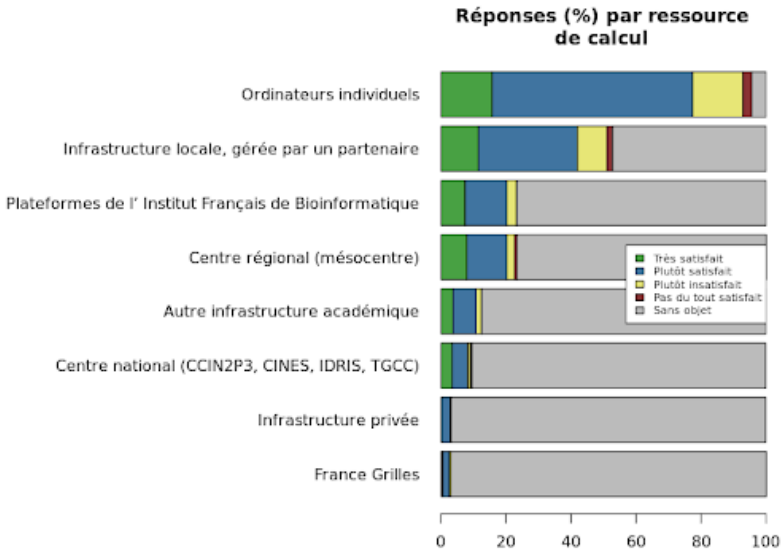
- "Pas de stockage au sein du mésocentre régional, hors les calculs qui y sont faits. (Quid de l'archivage, qui devient une obligation "")"
- "Absence infrastructure de stockage de coûts raisonnables."
- "Nous avons commencé à investir dans des serveurs de stockage mais à cause du manque de personnel informatique au sein de l'unité et du bâtiment, ces systèmes ne sont pas installés. Ce manque de personnel spécialisé nous empêche de bien évaluer nos besoins et les solutions à apporter."
- "La notion de long-terme; Peu de mode projet avec des espaces évolutifs sur un pas de temps long."
- "Le NAS de l'unité fonctionne très bien mais risque de destruction si évènement local (ex : feu)."
- "Pas assez d'espace"
- "L'espace alloué par le partenaire APHM est insuffisant"
- "pas de solution d'archivage mais en cours de mise au place au niveau université (infra locale)"
- "trop peu d'accès à des serveurs de stockage facilement utilisables et entretenus"
- "pas de système de sauvegarde centralisé, besoin d'un serveur pour partage samba"
- "L'organisation de la sauvegarde des ordinateurs individuels n'est pas suffisante"
- "NAS sous-dimensionné pour toutes les applications en imagerie"
- "Si le NAS peut répondre à des besoins personnels et administratifs, il ne répond pas aux besoins de stockage de données issues de la recherche. Définir un modèle économique et des modalités d'accès (protocoles) uniques entre Mésocentres et Plateformes de l'Institut Français de Bio informatique."
- "espace de stockage payant, limité et peu facile de gestion"
- "Pas de solution de stockage pour l'instant à l'échelle du MNHN"
- "Pas de sauvegarde automatique à partir des PC. Les sauvegardes à partir des Mac ne sont pas vérifiées régulièrement"
- "Les solutions centralisées (mésocentre, ccIN2P3) sont récentes. Elles devraient à terme remplacer le stockage individuel."
- "pas informaticien pour gérer cette problématique"
- "Volumes sauvegardés depuis les PC individuels par les offres institutionnelles bien trop restreints par rapport à la capacité des disques locaux; 2. données brutes archivées mais pas toutes les données actives avec en particulier un manque d'espace de projet partagé permettant un peu de volumétrie (actuellement limité de l'ordre du TO sur une solution institutionnelle). Va poser pb en particulier pour la mise en œuvre des DMP réclamés par les financeurs de projets (ANR, Europe...); 3. Pb de vitesse réseau en sortie de site pour la manipulation de gros volumes (besoin ponctuel mais pouvant nécessiter plusieurs jours de transfert pour quelques dizaines de TO sous forme de nombreux petits fichiers -formats propriétaires de données-)"
- "Manque d'espace de stockage"
- "une semaine d'arrêt du data centre IdF est trop long"
- "Pas d'espace pour installer une unité de stockage propre à l'unité. Doit externaliser, soit au sein de l'IBPC, soit au niveau des infrastructures nationales."

- "Infrastructure interne à l'unité sous dimensionnée, non adaptée du point de vue capacité et sécurité."
- "Nous avons à disposition une baie de stockage de 250 To au PSMN (ENS de Lyon) mais elle n'est accessible que par ssh et elle n'est pas sauvegardée"
- "Nécessité de disposer d'un dispositif de stockage/sauvegarde hors bâtiment de l'unité."
- "manque de compréhension"
- "j'aimerais être informé sur les plateformes de stockage régionales et nationales"
- "pas de protection / pas de sauvegarde au sein de l'IRD"
- "prix trop élevés, vitesse d'accès trop lente, pas assez de place"
- "Pas de stockage sécurisé (serveur local). Volume de stockage faible. Pas d'accès aux infrastructures académiques de notre tutelle."
- "Le stockage via le CNRS (myCore) n'offre pas une capacité suffisante pour sauvegarder l'ensemble d'un disque dur d'ordinateur contenant des données de taille importante (données de génomique), ce qui ne permet pas d'utiliser ce service pour des sauvegardes utiles."
- "Difficulté d'accès (espaces serveurs université et CHU non reliés ; discussion entre DSI en cours pour résolution problème). Autre infrastructure académique utilisée = GenOuest (Rennes) - Initialement utilisation de GenoToul mais saturé donc utilisation autre structure"
- "Pas suffisamment d'espace de stockage pour nos besoins"
- "espace de stockage / mise en partage des données de séquences et microarrays"
- "inopérant sous environnement Linux selon le PRI INRA\courte durée de vie des données sur les centres nationaux, CEPH du mésocentre est peu fonctionnel"
- "Moyens insuffisants et fragiles"
- "Nécessité d 'organisation. Un informaticien vient d 'être recruté pour le C2VN. Durant l' année 2020, une infrastructure va être mise en place pour assurer la protection des données, le stockage et la gestion des logiciels."
- "Etant en Guyane, le débit de connexion à internet est trop faible pour le transfert des grosses masses de données avec la métropole"
- "Ordinateurs de bureau plutôt anciens donc sans grande capacités de stockage. Disques durs de stockage à la charge de l'équipe, rapidement saturés. Quelques incidents de perte de données par panne de disques durs externes par le passé, qui ont rendu nécessaire une redondance des sauvegardes."
- "La sécurité des données stockées en interne est à perfectionner."
- "Manque de soutien par les tutelles."
- "lourd de gérer ses sauvegardes individuellement, surtout avec l'accumulation de données (X-ray, cryoEM)"
- "Nous venons mettre en place un moyen de stockage au sein de l'université"
- "Nous sommes dans une situation privilégiée avec la présence d'un informaticien ressource dédié. Attention à la situation des ressources réseaux niveau centre INRA qui doivent permettre des vitesses de transfert adaptées à la manipulation de grands volumes de données."

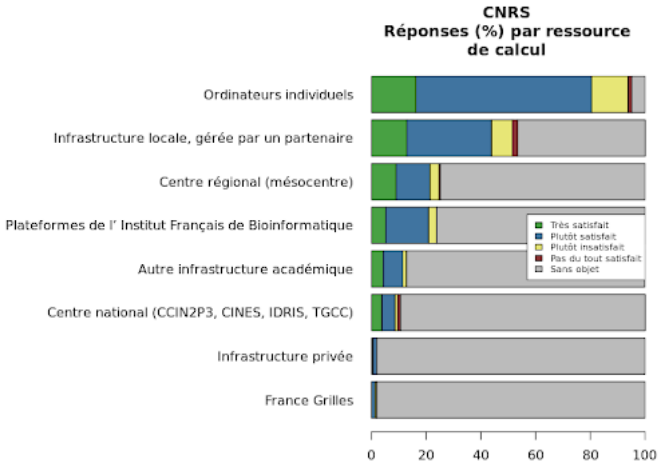
- "mise en place d'une solution propre à l'unité (NAS) comme pas de solution académique au sein de l'université de rennes 1. Toutefois nous gérons et maintenons nous même le NAS, or le NAS nous a laché. Nous avons donc un problème de fiabilité (malgré RED) / coût. Cette solution demanderait une solution mutualisée avec personnel dédié."
- "Alternative à l'existant trop limité, personnes ressources dédiées à la maintenance et l'évolution des infrastructures quasi inexistantes sur le centre"
- "harmoniser, mutualiser et sécuriser les solutions de stockage"
- "Nous avons de grandes quantités de données (actuellement 30 To) stockées actuellement sur des disques durs externes et nous aurions besoin de solutions de stockage pérennes et fiables en interne, le réseau est lent et les transferts prennent beaucoup de temps. Nous n'avons pas de sauvegarde automatique des données du laboratoire."
- "Manque de personnel informaticien au sein de la Fédération"
- "Difficulté d'accès aux données. Pas de procédure de dépôts pour des gros volumes. Limitation accès réseau. Gestion des sauvegardes et archives (volumétrie, sécurité).Question de la Pérennité des accès extérieurs"
- "utilisation de disque durs externes, peu rassurants"
- "Un service de stockage fourni par l'IFB pourrait être un bon complément des ressources disponibles localement, pour le stockage des données issues des ressources de calcul (centres nationaux et régionaux)."
- "Certains fichiers perdus peut-être lié à un problème de noms incluant des caractères inappropriés (avant de le savoir) ou chemin trop long à causes des sous répertoires nombreux (pas de moyens de contrôle)"
- "Pièce informatique de taille très limitée ne permettant pas actuellement l'implantation des nouveaux serveurs et baies de stockage indispensables suite à l'arrivée en 2019 de 2 nouvelles équipes ayant une forte activité de bioinformatique en génomique et en imagerie. Il est envisagé pour une partie de ces matériels de les implanter dans la salle informatique de l'INSERM sur le site de Saint Louis mais l'exploitation en sera rendue difficile en raison d'un débit de transfert de données insuffisant. Nous travaillons avec CATIBiomed mais ceci ne répond pas à tous les besoins de stockage de l'unité"
- "dans la liste il manque le data center institutionnel, tels que ceux de l'INRAe" "Difficulté d'identification des infrastructures et d'accès à celles-ci"
- "Nous n'avons pas de moyen de stockage pour nos données de omics autre que des moyen individuel de type disques durs ou ordinateurs."
- "Remonté par une équipe : manque d'espace sur infrastructure INRAe partagée locale."
- "Stockage géré en interne par la plateforme"
- "Aucun personnels pour la gestion des ressources informatiques et toute l'infra-structures informatiques de l'unité"
- "Aucune personnel pour la gestion des ressources informatiques"
- "Manque de visibilité/réactivité sur stockage. Coût de stockage des données NGS augmente chaque année et les projets qui les sont générées n'ont plus de budget. Difficultés pour organiser l'archivage des données"

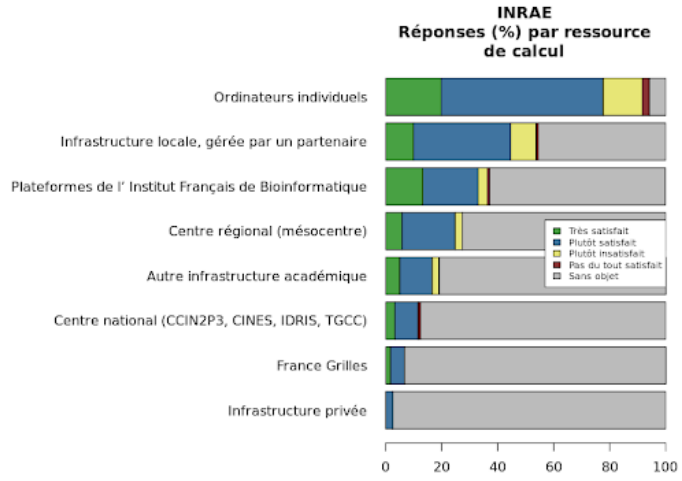
- "Autres commentaires :- difficultés du choix de la nature des données à stocker - données mises sur bandes par CEA et restaurées par le CEA dans le cadre de notre convention. Hors cadre de cette convention, n'existant pas de solution abordable financièrement, nous ne pourrions pas stocker les données à long terme comme proposé actuellement à nos partenaires de projet."
- "manque de moyens de sauvegarde"
- "Espaces insuffisants car besoins de + en + croissants"
- "Aucune sécurité en cas de problème. Les données sont éparpillées sur plusieurs ordinateurs."
- "le stockage sur ordinateur personnel n'est pas bien fait. La plupart des collègues ne savent pas gérer leur données correctement."
- "- La plupart des agents rapportent ne pas connaître suffisamment “France Grilles” et “Les Plateformes de l’Institut Français de Bioinformatique”.
- - Plusieurs agents rapportent la nécessité d’un système de sauvegarde avec synchronisation automatique, similaire à celui offert avec des solutions commerciales."
- "manque d'information et de tutoriel pour utilisation mésocentre du fait du manque de personnel dédié à la bioinfo dans l'équipe ou l'UMR"
- "Nos données sont principalement à caractère personnel de santé, c'est pourquoi elles sont toutes stockées en interne. Nos données sont archivées dans un site INRAe distant (sur le centre de Jouy-en-Josas). Nous prévoyons une certification ISO 27001 pour la fin 2020."
- "Serveur souvent lent et parfois inaccessibles, pas prévu pour le stockage de données non-actives"
- "Stockage IFB faible voire inexistant."
- "Besoin d'une solution institutionnelle de stockage et d'archivage des données"
  "Pas d'accès pour notre unité"
- "sur le mésocentre de Clermont-Ferrand : manque de personnel et du coup réponse longue pour un besoin identifié"
- "Manque d'une infrastructure locale de stockage et d'archivage"

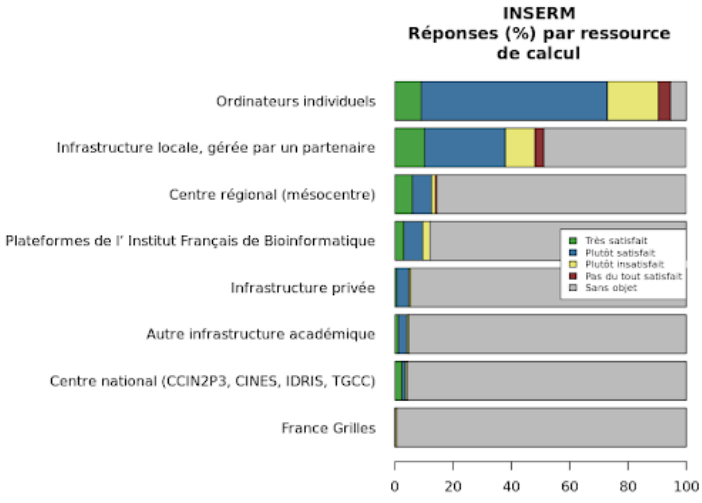
• H12. INDIQUEZ LES RESSOURCES DE CALCUL UTILISÉES AU SEIN DE L'UNITÉ/ ÉQUIPE ET VOTRE NIVEAU DE SATISFACTION

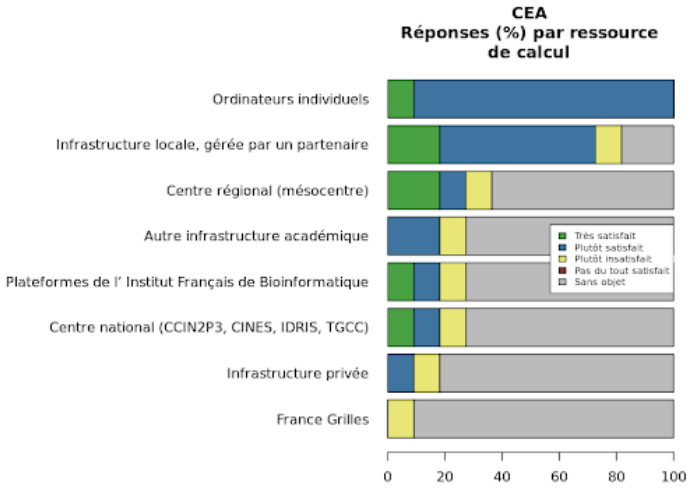


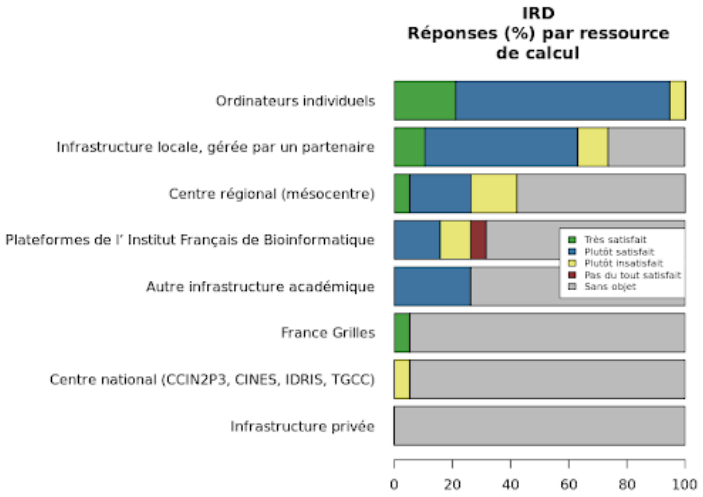
- Ordinateurs individuels: quasi tout le monde a répondu et est plutôt satisfait ou très satisfait (malgré tout, quasiment un quart ne sont pas satisfait).
- Plateformes de l'IFB : 23% des réponses mais avec un bon niveau de satisfaction.

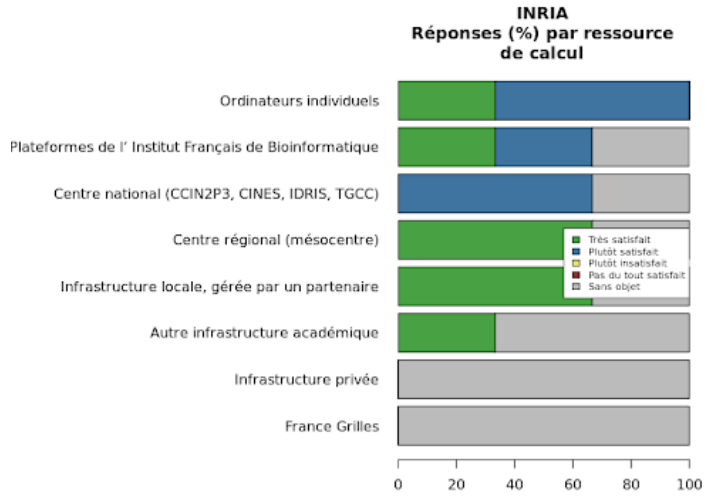


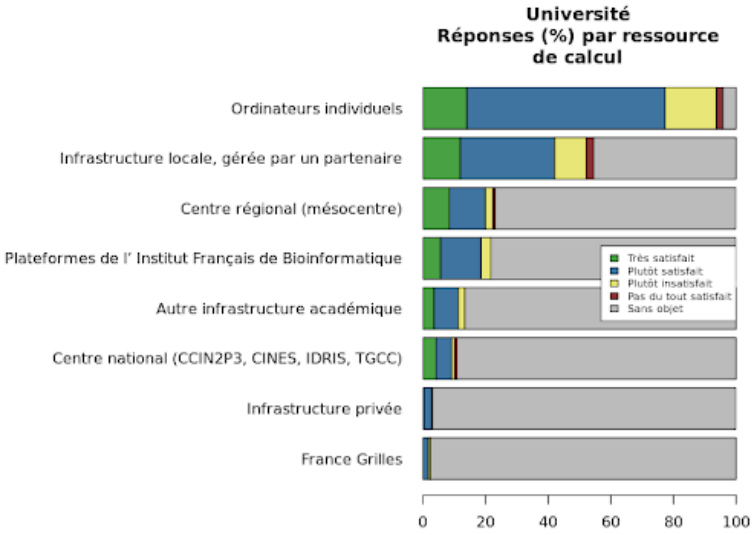












Taux de satisfaction pour les ressources de calcul

CALCULS	Ordinateurs et moyens de stockage individuels	Infra. locale, gérée par un partenaire	Centre régional	Centre national	France Grille	Plateformes de l'IFB	Autre infra. académique	Infra. privée
Taux d'utilisation (/réponses)	95,6% (389/407)	52,8% (215/407)	23,3% (95/407)	9,6% (39/407)	2,9% (12/407)	23,3% (95/407)	12,5% (51/407)	3,2% (13/407)
Très satisfait (/utilisation)	16,5%	21,9%	33,7%	35,9%	16,7%	31,6%	29,4%	7,7%
Plutôt satisfait (/utilisation)	64,5%	57,7%	52,6%	51,3%	66,7%	54,7%	56,9%	84,6%
Plutôt insatisfait (/utilisation)	16,2%	17,2%	10,5%	7,7%	16,7%	12,6%	13,7%	7,7%
Pas du tout satisfait (/utilisation)	2,8%	3,3%	3,2%	5,1%	0,0%	1,1%	0,0%	0,0%

H13. CALCULS : MOTIFS D'INSATISFACTION ET AUTRES COMMENTAIRES

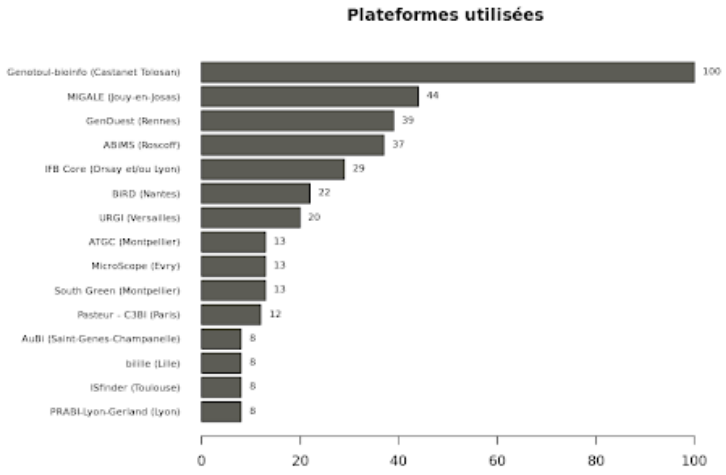
- "Infrastructure de calcul gérée par l'unité"
- "Meilleur accès à l'INRIA, ENS ..."
- "Serveurs vieillissants + manque d'administrateur système"
- "Nous faisons nos calculs en collaboration avec d'autres instituts (math et informatique)"
- "Nous possédons notre propre cluster de calcul."
- "Notre unité ne fournit aucun moyen de calcul. Nous utilisons des moyens internes à l'équipe (serveur) et le cloud du mésocentre lillois. C'est mieux que rien, mais loin d'être suffisant. Nous avons récemment testé le cloud de l'IFB: nous manquons encore de recul à ce sujet et n'avons pas de visibilité sur l'évolution de ce service."
- "Nous nous reposons principalement sur l'infrastructure de calcul de notre unité (et nous en sommes satisfaits). Problèmes avec les solutions externes: transfert des données, capacité de stockage limitante. Nous avons eu de mauvaises expériences avec EBIO (problèmes récurrents d'accès et d'ouvertures de compte, pannes, etc), et IFB Core (capacités très inférieures aux moyens internes du laboratoire)."
- "difficulté à maintenir des machines virtuelles distantes fonctionnelles"
- "Pas d'accès au mésocentre Nouvelle-Aquitaine"
- "Nous sommes obligés de payer un prestataire pour nos calculs"
- "L' infrastructure locale est une infrastructure hébergée et gérée à l'IBMP que nous partageons avec 2 autres unités."
- "Il nous manque la rubrique “Infrastructure interne à l’unité"pour laquelle nous aurions répondu plutôt satisfait."
- "Nos équipes de recherche sont contraintes en terme d'espace de calcul et de stockage"
- "Temps de calcul très longs"
- "Là encore, l'infrastructure locale est tout juste mise en place, on ne sait donc pas encore si cela nous satisfera sur le plus long terme. Calcul sur station de travail jusqu'à maintenant, mais puissance limitée et surtout sensible aux coupures de courant !!! (notre site est un peu vétuste). Pas eu de besoin d'utiliser les infrastructures nationales pour l'instant (génomes bactériens peu lourds)"
- "Manque d'ordinateurs puissants pour faire des analyses"
- "Pour l'instant, nous arrivons à fonctionner avec les partenariats sur projet pour avoir accès à des ressources de calcul sans frais. Nous avons également une infrastructure locale en accès libre. Mais jusqu'à quand ?"
- "Problème d'hébergement sec : incertitude sur le devenir de notre data center local; pas de visibilité sur les possibilités d'hébergement de serveurs dans ce DC"
- "Nous utilisons nos propres serveurs de calcul. Nous avons essayé d'utiliser le datacenter de Toulouse INRA; les problèmes: pas de possibilité de louer des machines virtuelles suffisamment performantes pour le calcul; problème de bandes passantes pour transférer nos données depuis l'unité vers le data center de Toulouse; si on y hébergeait nos serveurs, on serait obligé de se rendre sur place en cas de panne physique; pas assez de réactivité."
- "L'approche cloud de l'IFB est très intéressante mais elle doit être complétée d'un cluster de calcul"

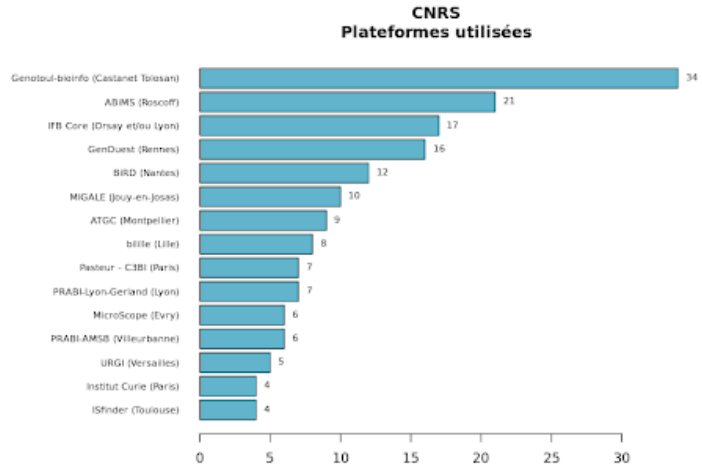
- "Très mauvaise optimisation des ressources informatiques par manque de compétences pour le choix des machines lors de l'achat et manque totale de maintenance."
- "en cours d'ouverture"
- "Certaines analyses sont difficiles à réaliser sur les ordinateurs individuels (notamment Mac)"
- "Manque d'homogénéité entre les différentes plateformes bioinformatiques (modèles de fonctionnement et économique), environement logiciels et de stockage"
- "Saturation et manque de contrôle sur la priorisation des projets."
- "Des demandes de temps de calcul sont prévues"
- "Difficulté à pérenniser"
- "Les infrastructures cloud académiques ne sont pas du tout compétitives par rapport au cloud commerciaux (Google, Amazon, Azur) en termes d'efficacité comme de facilité d'utilisation et d'intégration avec d'autres services. Il y a très peu de chance qu'elles le soient jamais et une réflexion sérieuse et honnête devrait être conduite sur ce point: en situation de compétition internationale, les chercheurs finiront toujours pas utiliser les ressources les plus performantes."
- "Manque d'information sur les possibilités en ressource de calcul externe à l'unité."
- "1-ordinateur personnel: la capacité de renouvellement est malheureusement assez faible, certains ordinateurs sont surannés, ce qui empêche l'utilisation de certains logiciels.
- 2- mésocentre: pour l'analyse de données NGS, les mésocentres sont moins efficaces en termes de puissance de calcul, par rapport aux plateformes.
- 3- les plateformes sont, pour certaines, saturées en nombre d'utilisateurs"
- "faible souplesse d'utilisation centre de calcul unité voisine"
- "Manque de ressource propre. Une grosse dizaine de serveurs pour l'équipe. Les calculs nécessitants des espaces temporaires conséquents (200 Gb à plus de 2TB) limitant ou empêchant l'utilisation de ressources partagées (cluster universitaire ou nationaux) et sont plutôt très longs. ex: dynamique moléculaire typique: 30 jours sur 2 CPU 24 cores réels et 48 GB RAM réservé. GPU: pas assez performant pour ce type de calcul."
- "Pas d'accès facile et de personnel formé pour nous aider à les utiliser"
- "Les serveurs locaux permettent d'avoir accès sans fil d'attente à un nombre de CPU assez important ce qui n'est pas toujours le cas sur des serveurs distants même s'ils sont très puissants du fait du nombre d'utilisateurs. Un intérêt des serveurs distants est de pouvoir bénéficier des conseils de bioinformaticiens travaillant sur des sujets similaires (...). Un point délicat est le financement de l'accès aux plateformes qui est un financement annuel difficilement justifiable sur des projets alors que l'achat d'un serveur ou de mémoire est justifiable."
- "Difficultés à établir du benchmarking (mesure du temps de calcul) de nos méthodes sur des clusters de calcul à nœuds partagés et hyperthreading."
- "Le point d'insatisfaction porte sur "la flexibilité" pour accéder à une capacité de calcul importante de façon ponctuelle."
- "Voir commentaire précédent"
- "Utilisation principalement du cluster de calcul de la plateforme Migale."
- "Nos ressources sont limitées et il nous arrive souvent de se trouver dans des situations où nous devions attendre la fin du calcul du voisin avant de lancer son calcul."
- "Ordis individuels utilisés pour calculs légers. Modalités d'accès aux calculs nécessitant une formation, peu accessible aux personnes non initiées."

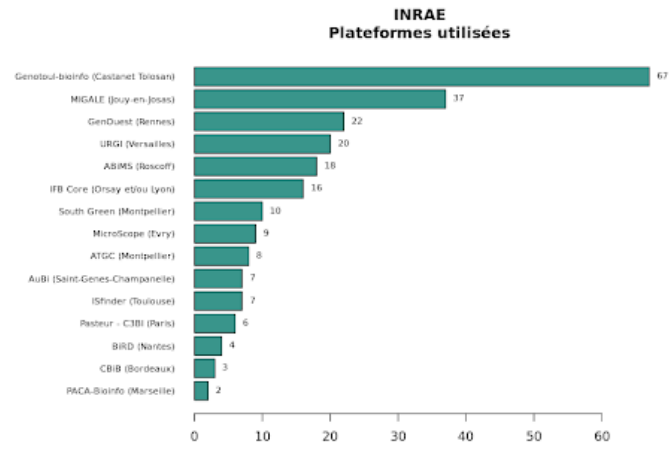
- "trop peu d'accès à des serveurs pour réaliser les approches de phylogénie et annotation"
- "utilisation IFB dans un avenir proche, objectif d'avoir un serveur de calcul"
- "manque de puissance de calcul pour certaines applications : protéomique, modélisation multiphysique"
- "c'est l'équipe qui est responsable de l'achat de tout matériel informatique, sur ses budgets propres. Aucune aide, ni ressource autre accessible, extérieures"
- "L'École normale de Lyon maintient un centre de calcul, le Pôle scientifique de modélisation numérique (PSMN) qui remplit parfaitement nos besoins en matière de calcul. Par la collaboration entre le PSMN et le ccIN2P3, nous avons également accès à du stockage de gros volumes à long terme, qui remplit là encore nos besoins."
- "j'aimerais être informé sur les ressources de calcul régionales et nationales"
- "Quota trop faible sans engagement financier pour les ressources externes (Génotoul). Pannes fréquentes (IFB). Les machines virtuelles ne sont pas stables (IFB)"
- "besoins de mieux connaître les ressources disponibles à l'INRA pour utiliser des clusters de calcul. Comment les utiliser? pour quelle utilisation?"
- "prix trop élevés, prise en main difficile, problème de disponibilité des logiciels, faibles performances"
- "Utilisation des ordinateurs personnels pas satisfaisante pour l'analyse d'images (pas assez de capacité)."
- "Autre infrastructure académique utilisée = GenOuest (Rennes)"
- "Nous avons un cluster de calcul propre à l'UMR, et administrons un 2e cluster de calcul ouvert à la communauté "Eco/Evo"de Montpellier; évaluation: très satisfait."
- "Ces ressources dépendent des projets."
- "Matériel individuel vieillissant"
- "idem stockage"
- "Nécessité de gérer les licences sur des postes dont l'utilisateur n'est pas administrateur."
- "Étant en Guyane, le débit de connexion à internet est trop faible pour le transfert des grosses masses de données avec la métropole"
- "Ordinateurs de bureau plutôt anciens donc capacités de calculs plutôt limités."
- "Manque de soutien par les tutelles"
- "Plutôt satisfait de notre infrastructure interne à l'unité (serveur de calcul, nombre de coeurs, ram...)"
- "les analyses de données sont réalisées sur ordinateur individuel, à l'exception des analyses des données massives sur la plateforme Genouest a à Rennes (données génomiques) et de la plateforme Biosit sur le campus médecine de l'université de Rennes 1 (imagerie multiparamétrique)"
- "contraintes d'utilisation, temps de transfert de données, espace de stockage"
- "Mise ne place de nouvelles actions de recherche sur la modélisation moléculaire -> ressources insuffisantes mais trop tôt pour dresser un bilan"
- "Outre nos ordinateurs individuels (Mac et PC sous linux), nous avons accès à un cluster de calculs du laboratoire de biochimie théorique voisin. Cet accès est satisfaisant la plupart du temps, mais nous ne sommes pas toujours informés des changements et de leurs impacts."
- "Remarque: il manque une ligne: "serveurs de calculs, NAS, ...” qui aurait été cochée ici avec mention "plutôt satisfait"(cf tableau précédent)"
- "Nécessité d'upgrader nos systèmes fréquemment. Coût de l'achat et du maintien de telles infrastructures. Vétusté rapide des moyens informatiques au vue des accroissements en calcul et gestion de gros volumes de données"

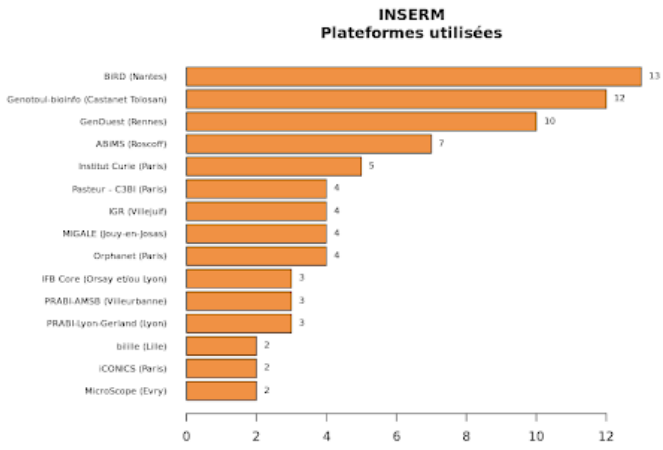
- "passage par une plateforme d'analyses des données qui est très occupée"
- "Pour le moment, je ne fais que des calculs simples sur Excel. Avec le développement de la métaprotéomique et des analyses statistiques qui vont avec, ça risque d'être autre chose"
- "Difficulté d'achat de machines de calcul personnalisées par les tutelles de l'unité"
- "Le renouvellement du parc d'ordinateurs individuels est insuffisant. Le renouvellement récent du centre de calcul de l'Université de La Réunion s'accompagne d'une gouvernance et gestion encore assez floue rendant son utilisation peu pratique; Nous espérons une amélioration du service à court terme. Le "centre régional"utilisé est la plateforme southGreen"
- "Nous disposons de moyens de calcul assez importants au sein de l'unité (cluster condor). Pour le reste, nous faisons appel aux plate-formes bioinformatiques INRAe"
- "Même remarque que pour le stockage"
- "les calculs sur plateforme à distance restent souvent limitant à cause du système de "queueing."
- "Calcul géré en interne par la plateforme"
- "Aucun personnels pour la gestion des ressources informatiques et toute l'infra-structures informatiques de l'unité"
- "Utilisation de Plateformes de l' INRA : plutôt satisfait"
- "Coût processeur GPU performant"
- "Ressources de calcul utilisées via les plateformes bioinformatique INRAe Genotoul ou MIGALE"
- "Insatisfaction : limitation sur l'espace alloué (Genotoul). Autres commentaires : cf H5."
- "Seule ressource externe étant l'IFB, cela implique une grosse dépendance"
- "Projet de création d'une salle de bioinformatique dans notre institut porté depuis 5 ans... toujours pas mis en œuvre. Objectif: mutualiser un serveur de calcul, et juxtaposer les bioinformaticiens des différentes équipes de biologie en un lieu unique sur site."
- "- La plupart des agents rapportent ne pas connaître suffisamment “France Grilles” et “Les Plateformes de l’Institut Français de Bioinformatique”. Plusieurs agents ont fait remonter le besoin d’une salle de visualisation au niveau local."
- "manque d'information et de tutoriel pour utilisation mésocentre du fait du manque de personnel dédié à la bioinfo dans l'équipe ou l'UMR"
- "MGP dispose de sa propre infrastructure de calcul. Les calculs sur les plateformes de l'IFB sont relativement ponctuels du fait de la volumétrie. Nous explorons la possibilité de faire du débordement de calcul dans des infrastructures cloud commercial. Azure a été testé en 2018-2019. Pour des raisons de sécurité et de besoin d'agrément HDS, nous allons tester OVHCloud."
- "Infrastructure inexistante au niveau de l'unité et du campus."
- "Pas d'accès pour notre UR"
- "Suite au départ de l'agent en charge de l'administration du serveur de calcul, la maintenance et la mise à jour n'est pas assurée ou de façon sporadique."

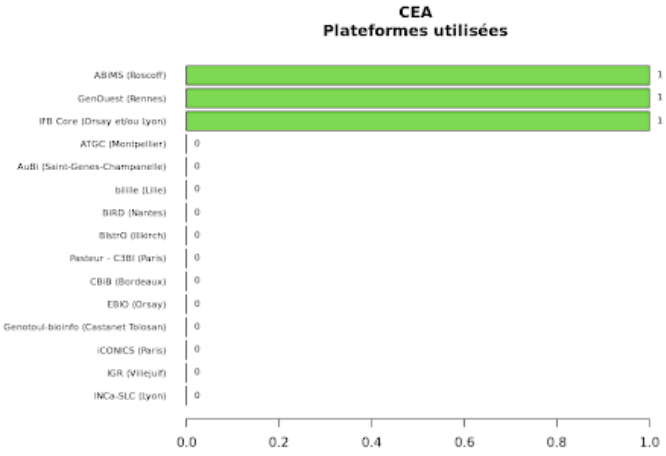
- **H14. PLATEFORMES BIOINFORMATIQUES UTILISÉES POUR VOTRE STOCKAGE ET/OU CALCUL. PRÉCISEZ SI VOTRE ÉQUIPE/UNITÉ RECOURS À DES PLATEFORMES SPÉCIALISÉES EN BIOINFORMATIQUE POUR LE STOCKAGE ET CALCUL**

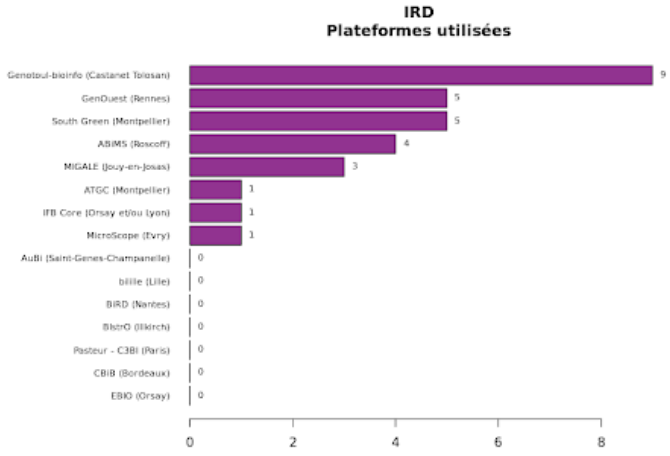


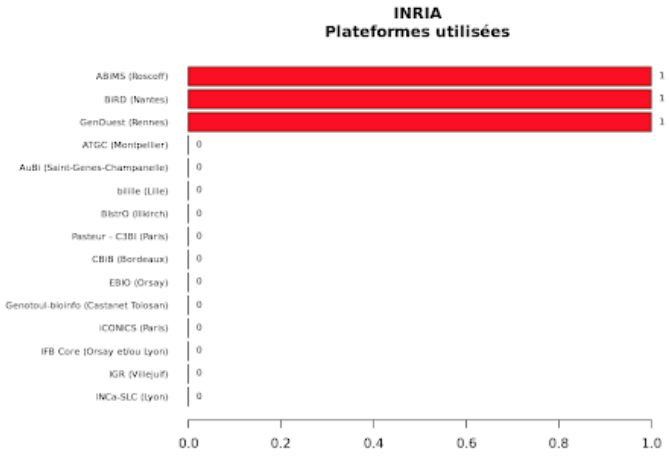


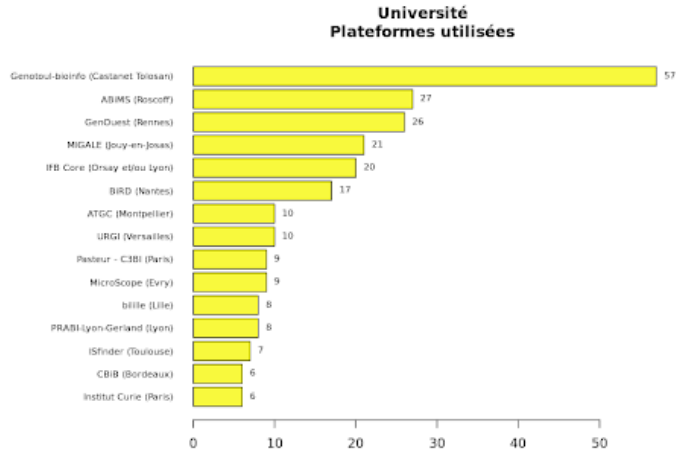









INRIA
Plateformes utilisées
ABIMS (Roscoff)
BiRD (Nantes)
GenDuest (Rennes)
ATGC (Montpellier)
AuBI (Saint-Genes-Champanelle)
bilille (Lille)
BistrO (Illkirch)
Pasteur - C3BI (Paris)
CBiB (Bordeaux)
EBIO (Orsay)
Genotoul-bioinfo (Castanet Tolosan)
ICONICS (Paris)
IFB Core (Orsay et/ou Lyon)
IGR (Villejuif)
INCa-SLC (Lyon)
1
1
1
0
0
0
0
0
0
0
0
0
0
0
0
0.0
0.2
0.4
0.6
0.8
1.0


Université
Plateformes utilisées
Genotoul-bioinfo (Castanet Tolosan)
ABIMS (Roscoff)
GenDuest (Rennes)
MIGALE (Jouy-en-Josas)
IFB Core (Orsay et/ou Lyon)
BiRD (Nantes)
ATGC (Montpellier)
URGI (Versailles)
Pasteur - C3BI (Paris)
MicroScope (Evry)
bilille (Lille)
PRABI-Lyon-Gerland (Lyon)
ISfinder (Toulouse)
CBiB (Bordeaux)
Institut Curie (Paris)
57
27
26
21
20
17
10
10
9
9
8
8
7
6
6
0
10
20
30
40
50

CRÉDITS & REMERCIEMENTS

Réalisé par Suzanne Lauriou, Claudine Médigue, Sylvain Milanesi, Hamid Ouahioune, Angela Saenz, Olivier Sand, & Jacques van Helden in November 2021.

Un grand merci à tous ceux qui ont contribué à ce rapport en répondant au questionnaire ou en aidant à l'analyse des résultats

Pour plus d'information:
contact@france-bioinformatique.fr

L'IFB/ELIXIR-FR est financé par le Programme d'Investissement d'Avenir PIA, subvention de l'Agence Nationale de la Recherche, numéro ANR-11-INBS-0013