Bioinformatique Biodiversité Représentation & Intégration des Connaissances	2eme Assemblée Générale « orientée biologistes » Formation analyse de données : génomes & transcriptomes 24 & 25 avril 2014 – Paris (réunion préparatoire le 23 avril à Paris)

Membres du CATI présents le 23 (16): Corinne Rancurel, Martine Da Rocha (Sophia), Martial Briand (Angers), Alexandre Dehne-Garcia, Bernhard Gschloessl, Franck Dorkeld, Marc Tauzin (Montpellier), Fabrice Legeai, Anthony Bretaudeau (Rennes), Joseph Tran, Adeline Simon (Versailles), Patrice Baa-Puyoulet (Lyon) Sébastien Carrere, Ludovic Legrand, Ludovic Cottret, Jérôme Gouzy (Toulouse).

Participants à la formation des 24/25 (16): PONTS Nadia (Bordeaux), LEPETIT Marc, TAUZIN Marc (Montpellier), PORQUIER Antoine, BARNY Marie-Anne (Versailles), DARRASSE Armelle (Angers), BARTOLI Claudia, MONTEIL Caroline (Avignon), GRAVOT Antoine, MONTARRY Josselin, EOCHE-BOSY Delphine, GRENIER Eric, DAVAL Stéphanie, GAZENGEL Kévin, JUBAULT Mélanie, (Rennes), COLELLA Stefano (Lyon).

Communication/inscriptions : Arnaud Ridel (dpt SPE)
Annonce : http://cati-bbric.toulouse.inra.fr/lib/exe/fetch.php/BBRIC-AG-Communaute-201304-24-25.pdf
Liste de diffusion : communaute-bbric@liste.inra.fr

Rédacteur : Jérôme Gouzy
Type du document : compte rendu et perspectives. Les documents de la formation sont disponibles sur https://listes.inra.fr/sympa/d_read/communaute-bbric/Formation-2014/

Planning des Assemblées Générales BBRIC

- 30 octobre 2012 : présentation des membres du CATI
- 15-16 avril 2013 : « orientée biologistes », présentation du CATI à notre « communauté servie »
- 28-30 octobre 2013 : « orientée technique informatique », présentation d'outils et de méthodes pour le développement. Travaux pratiques sur le déploiement de programmes via Galaxy.
- 23-25 avril 2014 : « orientée biologistes », présentation de l'architecture bioinformatique BBRIC, formation des utilisateurs de la communauté servie à l'utilisation de l'environnement BBRIC

- *23-24 octobre 2014 : « orientée technique (bio)informatique », présentation d'outils et de méthodes (bio)informatiques.*
    - *20-22 octobre : Couplée à l'AG, nous organisons avec l'aide de la FP de Toulouse, une formation aux technologies web : HTML5, CSS3, JAVASCRIPT (contact Sebastien.Carrere@toulouse.inra.fr)*
- *printemps 2015 : « orientée biologistes », présentation de l'architecture bioinformatique BBRIC, formation des utilisateurs de la communauté servie à l'utilisation de l'environnement BBRIC*

Objectifs des formations BBRIC

1. Illustrer à travers des exemples variés comment nous allons interagir avec les biologistes
   - à distance et sur un temps long
   - à travers quelques principes de fonctionnement
   - grâce aux outils que nous mettons à disposition (Portail bioinformatique https://bbric.toulouse.inra.fr, Archive, Workspace, etc.)
2. Rendre autonomes les utilisateurs sur les tâches d'analyse de données récurrentes et automatisées.
3. Illustrer un savoir faire bioinformatique pour la conception de pipeline d'analyses bioinformatiques
4. Produire un support de formation que nous pourrons réutiliser facilement
5. Essayer d'être complémentaire des autres formations, déjà disponibles ou en préparation.

Planning de la formation

Mercredi 23 : Répétition de la partie théorique de la formation entre membres du CATI

Jeudi 24

Module	Horaire	Durée
Introduction et présentation du portail BBRIC	10H00-11H15	1 h 15
Assemblage de petits génomes	11H15-13H15	2h
Repas – déjeuner (cantine INRA, 147 rue de l'univ.)	13H15-14H30	1 h 15
Assemblage de novo de transcriptomes avec mesure de l'expression par banque	14H30-16H30	2h
Pause		15 min
Prédiction et caractérisation de longs ARN non codant	16H45-18H45	2h

Vendredi 25

Module	Horaire	Durée
Annotation de génomes bactériens	9H00-10H45	1 h 45
Pause		15 min
Conversion de format pour soumission de génomes annotés aux banques publiques	11H00-12H00	1 h
Mesure de l'expression a partir de données RNAseq	12H00-13H15	1h15
Repas – déjeuner (cantine INRA, 147 rue de l'univ.)	13H15-14H30	1 h 15
Détection de transferts horizontaux	14H30-15H30	1h
Discussion	15H30-16H00	30min

Organisation de la formation

La formation a été organisée par Martine Da Rocha (Sophia) qui a suivi récemment la formation de formateur interne.

L'application par tous de la même méthode séquencée en trois parties « Découverte/Démonstration/Application » a donné un cadre commun à tous les responsables de modules. Il en va de même de la rédaction du scénario pédagogique en amont de la conception du support de présentation et de TP.

Le résultat est un support de plus de 250 pages incluant les objectifs pédagogiques de chaque module, la partie théorique et les questions pour les travaux pratiques. Ce support très complet sera réutilisé par ceux qui vont reproduire tout ou partie de la formation sur leurs sites respectifs (déjà prévu à Sophia, Toulouse, Montpellier, Angers).

L'autre originalité de la méthode utilisée a été que ce n'est pas l'expert de l'outil (celui/celle qui a implémenté l'outil) qui a rédigé les supports de cours et de TP. Cela a nécessité un investissement important en amont aussi bien de la part de l'expert pour transférer ce qu'il savait que du responsable du module (et du relecteur) pour acquérir les informations et le recul nécessaire pour faire fonctionner l'outil, connaitre ses limites et être capable de le présenter. Le bénéfice a été que pour chaque module, trois ou quatre personnes étaient disponibles et capables de guider nos collègues biologistes. Cela a permis un déroulement impeccable des TP de tous les modules. Pour la partie théorique, le fait que nous l'ayons répétée tous ensemble la journée du 23 nous a permis d'échanger sur nos stratégies d'analyse et d'affuter nos arguments aussi bien sur les points forts que les limitations des outils que nous avons présentés. Au delà des nombreuses visioconférences, cette journée de préparation en conditions réelles a contribué a évacuer le stress et à affiner notre discours pour les deux jours suivants.

Pour finir, l'organisation de cette formation a nécessité un travail important d'animation (bravo Martine) pour que tout soit bien fait et fait dans les temps. Ainsi qu'une implication importante des membres du CATI qui ont conçu et mis en œuvre cette formation pour nos collègues biologistes, merci à vous tous. Le tableau ci-après récapitule le rôle de chacun.

Responsable: Martine Da Rocha (ISA/Sophia)
Assistants: Jérôme Gouzy (LIPM/Toulouse) & Arnaud Ridel (SPE/Sophia)

Modules	Responsable et intervenant principal	Expert	Relecteur
Introduction et présentation du portail BBRIC	Jérôme Gouzy, Ludovic Legrand LIPM/Toulouse ; Anthony Bretaudeau IGEPP/Rennes		
Assemblage de petits génomes	Martine Da Rocha ISA/Sophia	Jérôme Gouzy LIPM/Toulouse	Ludovic Cottret LIPM/Toulouse
Assemblage de novo de transcriptomes avec mesure de l'expression par banque	Joseph Tran IJBP/Versailles	Anthony Bretaudeau IGEPP/Rennes	Franck Dorkheld CBGP/Montpellier
Prédiction et caractérisation de longs ARN non codant	Corinne Rancurel ISA/Sophia	Fabrice Legeai IGEPP/Rennes	Martine Da Rocha ISA/Sophia
Annotation de génomes bactériens	Bernhard Gschloessl CBGP/Montpellier	Jérôme Gouzy LIPM/Toulouse	Martial Briand IRHS/Angers
Conversion de format pour soumission de génomes annotés aux banques publiques	Martial Briand IRHS/Angers	Sébastien Carrere LIPM/Toulouse	Corinne Rancurel ISA/Sophia
Mesure de l'expression a partir de données RNAseq	Adeline Simon BIOGER/Versailles	Ludovic Legrand LIPM/Toulouse	Joseph Tran IJBP/Versailles
Détection de transferts horizontaux	Ludovic Cottret LIPM/Toulouse	Corinne Rancurel ISA/Sophia	Sébastien Carrere LIPM/Toulouse ; Fabrice Legeai IGEPP/Rennes

Feedback des personnes formées

Afin de recueillir les premiers retours des personnes ayant assisté à la formation, nous nous appuyons sur la discussion de fin de formation ainsi que sur le petit questionnaire rempli. Les réponses au questionnaire se trouvent à l'url http://cati-bbric.toulouse.inra.fr/lib/exe/fetch.php/bbric-ag-communaute-201304-24-25-resultats_questionnaire_evaluation_biologistes.pdf.

Les points négatifs sont relativement limités :
- une salle trop petite pour 16+3 personnes. C'est tout à fait vrai mais pour cette 1ere session nous ne connaissions pas la salle et nous avons fait le choix de prendre le maximum de candidats.
- trop de modules « orientés procaryotes » pour ceux qui travaillent sur les eucaryotes. C'est un choix délibéré du LIPM d'avoir plutôt présenté ses outils « orientés procaryotes » car il y a des dizaines de génomes sur de nombreuses espèces séquencées dans le département SPE. La priorité est de rendre autonome les chercheurs car on dispose de l'ensemble de la panoplie d'outils pour que les chercheurs puissent eux même aller de la séquence génomique à la soumission aux bases de données. Rennes et Sophia ont présenté des outils « orientés eucaryotes » et au final l'équilibre était respecté.
- certains auraient préféré plus de détails sur les pipelines. L'objectif principal était que les gens se sentent capables de faire des analyses avec un peu de recul sur le contrôle du pipeline mais pas de disséquer les pipelines. Ce choix était délibéré car il est vraisemblable que le contenu des pipelines (le « comment ») change au fil du temps et décrire en profondeur la partie bioinformatique reviendrait à former à l'acquisition de connaissances non pérennes, ce que nous souhaitons éviter.

Les points positifs sont largement majoritaires. De nombreuses personnes ont répondu que la formation correspondait à leurs attentes et qu'elles se sentaient capables de faire elles même les analyses, ce qui était

l'objectif premier de cette formation. Les commentaires sur la qualité des présentations sont aussi très encourageants pour ceux et celles qui ont passé beaucoup de temps à préparer. Les participants ont beaucoup apprécié les présentations en trinômes et les interactions bio/bioinfo lors des pauses et repas car cela leur a permis d'identifier et de discuter avec un maximum d'interlocuteurs bioinformaticiens.

Questions des personnes formées

Nous avons été questionnés sur deux sujets qui méritent que l'on s'y attarde un peu.

Citations: La question s'est posée de la reconnaissance liée aux pipelines d'analyses que nous mettons à disposition. Sauf indication contraire sur le formulaire du workflow, les pipelines mis à disposition sur le portail BBRIC peuvent *a priori* être utilisés sans qu'il soit nécessaire de formaliser une collaboration avec celui qui l'a mis en place. Par contre, il serait souhaitable que les utilisateurs nous envoient les références des publications qui ont exploité les résultats produits par nos outils d'analyse. Dans le cas où ce que nous mettons à disposition à titre gracieux ne suffit pas (Stockage, Calcul, fonctionnalités des outils) il faut contacter l'expert (ou le responsable du CATI) pour voir s'il est possible de trouver une autre solution.

Moyens de calcul et de stockage: L'architecture BBRIC présentée lors de la formation repose sur les moyens informatiques propres du LIPM et de l'IGEPP/Genouest, ceci aussi bien pour la partie « Archive » (collecte et structuration des données/metadonnées de séquence) que la partie « Protocoles d'analyse » (mise à disposition de pipelines d'analyse à travers l'interface Galaxy). Nous avons donc clairement une démarche proactive pour répondre aux besoins de la communauté servie par le CATI. Mais vu nos moyens limités nous ne pouvons vraiment pas fournir des moyens/outils aux chercheurs qui ne sont pas dans la communauté servie par BBRIC (il y a 7 CATI orientés bioinformatique à l'INRA, les chercheurs qui ne sont pas dans notre communauté doivent prendre contact avec les CATI dont ils dépendent). Il est aussi évident que nous aurons besoin d'un soutien financier si nos outils deviennent largement utilisés en dehors des deux laboratoires qui fournissent actuellement les moyens.

Prévision pour la formation de l'an prochain

Comme l'an dernier, nous avons profité de la rencontre avec des biologistes qui ne sont pas de nos laboratoires pour affiner notre perception des thèmes sur lesquels nous pourrions centrer notre session de formations de l'an prochain. Selon de même principe de ne pas se lancer dans ce que d'autres font déjà, voici les commentaires que l'on peut faire et les priorités que l'on peut définir pour l'an prochain.

Thèmes	Commentaires
Metagénomique	C'est certainement un sujet dont s'est saisi le métaprogramme MEM, il vaut mieux que ceux qui ont ce besoin contactent les responsables du métaprogramme pour être redirigés vers les bons interlocuteurs.
Polymorphisme/Estimation des fréquences alléliques	**C'est clairement un besoin largement partagé. De nombreux outils sont déjà implémentés en ligne de commande, le portage dans l'interface galaxy est tout à fait possible dans l'année.**
Epigénétique	C'est un sujet qui intéresse de nombreux laboratoires mais pour le moment on a l'impression que ce sont surtout dans des laboratoires ou des bioinformaticiens sont présents pour mettre en œuvre et l'urgence ne semble donc pas immédiate. On fera quelque chose mais plutôt en 2016.
Phylogénie avancée	On peut faire des formations d'initiation mais pour des formations avancées il faut mieux voir avec les chercheurs experts du domaine qui organisent certainement des formations (➔ Lyon,Montpellier)
Génétique/Génomique des populations	En mars 2014, il y a eu l' « école chercheurs » organisée en partie par les départements SPE/EFPA. Nous étions 4 du CATI a participer mais pour une formation il vaut encore mieux s'adresser directement aux formateurs plutôt qu'aux élèves que nous étions.

Data mining/Annotation fonctionnelle/GO	**C'est la suite logique de la formation de cette année et nous avons déjà des outils prêts à être diffusés.**
Analyse des réseaux	Il y a des compétences au sein du CATI mais il vaut mieux prendre directement contact avec Ludovic.Cottret@toulouse.inra.fr pour voir si le besoin peut être traité au sein du projet MetExplore qui organise aussi des sessions de formations.
Visualisation	**L'importance de la problématique de visualisation/représentation était apparue avec force lors de l'AG « orientée biologiste » de 2013. Nous avions fait l'impasse cette année car nous ne pouvions pas tout faire. L'an prochain nous ferons un effort sur ce sujet important tout en essayant d'être complémentaire des outils commerciaux comme CLCbio qui sont de plus en plus présents dans nos laboratoires.**

Les trois priorités pour la session de formation de 2015 seront donc les outils pour l'analyse du polymorphisme, l'annotation fonctionnelle et les solutions de visualisation.

Bien sûr cela n'exclut pas la possibilité de mettre à disposition n'importe quel protocole d'analyse sur le portail BBRIC car tout protocole est éligible à partir du moment où il a déjà été utilisé pour de l'analyse de données dans le cadre d'un projet scientifique.

Budget de fonctionnement du CATI

Un petit point sur la répartition du financement du CATI pour formaliser ce que nous avons discuté en février. Le principe général est d'essayer d'aider un peu ceux qui contribuent le plus. En 2013 c'est le LIPM qui a essentiellement contribué à l'animation et aux projets, la totalité des financements venant des départements (4600€ SPE+800 BAP+200 EFPA) ont servi à payer une partie des frais de déplacement des agents du LIPM à Paris. En 2014, en plus du LIPM, Fabrice/Antony (BBRIC Référence, 2 Pipelines & Formation) et Martine/Corinne (Formation & 1 Pipeline) et a un degré moindre le CBGP (Archive&Formation) ont principalement apporté leur contribution, d'autres ont aussi contribué mais moins (ou ils sont sur Paris !). Sur la somme reçue (4600€ du SPE) le LIPM a transféré 800€ à Rennes et Sophia et 400€ au CBGP pour contribuer aux frais de déplacement (il restera 2600€ au LIPM).

Cout d'organisation de la formation 2014 :

Les frais engagés par les membres du CATI pour cette session de formation s'élèvent à 5672€ (détail ci-dessous). Le soutien actuel ne couvre pas cette somme et en outre dans notre CATI nous organisons 2 assemblées générales par an. Que ce soit pour les moyens informatiques ou les formations, nous faisons d'abord et demandons de l'aide ensuite. C'est une démarche qui permet aux financeurs potentiels de juger de la pertinence réelle et non pas théorique d'une action, mais il est évident qu'arrive le moment ou il n'est plus possible que les laboratoires des membres du CATI financent une structure ou des actions destinées a un collectif qui les dépasse (même si on fait tout pour qu'il leur serve aussi). Ainsi, pour la session de formation de l'an prochain, un soutien financier supplémentaire serait nécessaire et sera demandé aux départements et/ou à la formation permanente.

*LIPM/Toulouse (4 personnes/3j) : 2100€ ; ISA/Sophia (2 personnes/3j) : 1030€ ; CBGP/Montpellier (3 personnes/(1j + 2*3j)) : 1217€ ; IRHS/Angers (1 personne/3j) : 397€ ; IGEPP/Rennes (2 personnes/2j) : 768€ ; BIOGER/Versailles (1 personne/3j) : 80€ ; IJBP/Versailles (1 personne/3j) : 80€ (estimation)*