Bioinformatique Biodiversité Représentation & Intégration des Connaissances	3eme Assemblée Générale « orientée biologistes » Formation: annotation fonctionnelle et analyse du polymorphisme 28 & 29 avril 2015 – Paris (UPMC) (réunion préparatoire le 27 avril à Paris)

Membres du CATI présents le 27 (15): Corinne Rancurel, Martine Da Rocha (Sophia), Martial Briand (Angers), Bernhard Gschloessl, Franck Dorkeld (Montpellier), Fabrice Legeai, Anthony Bretaudeau (Rennes), Joseph Tran, Adeline Simon (Versailles), Patrice Baa-Puyoulet (Lyon) Erika Sallet, Sébastien Carrere, Ludovic Legrand, Ludovic Cottret, Jérôme Gouzy (Toulouse).

Participants à la formation des 28/29 (16): PONTS Nadia (Bordeaux), NEGRE Nicolas, MONE Yves (Montpellier), JACQUIN-JOLY Emmanuelle (Paris), DARRASSE Armelle (Angers), NEGRE Sylvie, LE TRIONNAIRE Gael, RICHARD Gautier, MONTARRY Josselin, EOCHE-BOSY Delphine, DAVAL Stéphanie, GAZENGEL Kévin (Rennes), VINCENT –MONEGAT Carole (Lyon), PANABIERES Franck (Sophia), BOISSINOT Sylvaine, MULOT Michael (Colmar)
Les noms des personnes qui ont suivi la précédente formation BBRIC (2014) sont soulignés

Communication/inscriptions : Arnaud Ridel (dpt SPE)
Annonce : http://cati-bbric.toulouse.inra.fr/lib/exe/fetch.php/BBRIC-AG-Communaute-201504-28-29.pdf
Liste de diffusion : communaute-bbric@liste.inra.fr

Rédacteur : Jérôme Gouzy
Type du document : compte rendu, relevé de conclusion et perspectives. Les documents de la formation sont disponibles sur https://listes.inra.fr/sympa/d_read/communaute-bbric/Formation-2015/

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 à l'utilisation de l'environnement BBRIC. Thèmes : analyse de données génomes et transcriptomes
- 23-24 octobre 2014 : « orientée technique (bio)informatique », présentation d'outils et de méthodes (bio)informatiques.
    - 20-22 octobre : Couplée à l'AG, Sébastien a organisé avec l'aide de la FP de Toulouse, une formation aux technologies web : HTML5, CSS3, JAVASCRIPT.
- 27-29 avril 2015 : « orientée biologistes », formation des utilisateurs de la communauté servie à l'utilisation de l'environnement BBRIC. Thèmes : annotation fonctionnelle & détection polymorphismes
- *Automne 2015 : « orientée technique info/bioinfo», présentation d'outils et de méthodes (bio)informatiques.*
    - *Couplée (?) à l'AG, LudoC va organiser une formation au framework javascript de visualisation de données D3.js*
- *Printemps 2016 : « 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, Galaxy, 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.

Retours sur la Formation 2014

Au total 91 personnes ont suivi tout ou partie de la formation que nous avons préparé pour 2014 sur les thèmes « genomes & transcriptomes ».

- 16 personnes ont suivi la formation initiale à Paris en avril 2014
- 2 juin à Lyon (BF2I, Patrice), 5 personnes
- 17 et 20 juin à Montpellier (CBGP, Bernhard & Franck), 11 personnes
- 23 juin à Toulouse (LIPM, LudoL, LudoC, Sébastien, Erika), 12+6 personnes
- 23/24 et 25/26 sept 2014 à Sophia (ISA, Martine & Corinne), 12+10 personnes
- 27/28 octobre 2014 à Avignon (ISA, Martine & Corinne), 10 personnes
- 8 janvier 2015 à Angers (IRHS, Martial & Armelle Darasse), 9 personnes

Cela a été possible grâce a notre effort important de formalisation des scenarii et supports pédagogiques (250 diapositives) et surtout grâce a un effort important de celles et ceux qui ont reproduit la formation. Sur la base de cette réussite collective nous avons décidé d'appliquer la même méthode pour la mise en place de la formation 2015.

Planning de la formation 2015

Lundi 27 : Répétition de la partie théorique de la formation entre membres du CATI

Mardi 28

Module	Horaire	Durée
Introduction et présentation du portail BBRIC	10H00-11H15	1 h 15
Annotation fonctionnelle automatique d'un proteome à partir de la banque de domaines protéiques InterPro	11H15-12H15	1h
Repas – déjeuner		1h30
Construction d'un protéome exhaustif, non redondant et annoté pour analyses protéomiques	13h45-14h45	1h
Transfert d'annotation par identification des relations d'orthologies entre deux espèces	14h45h-15h15	30mn
Identification des relations d'orthologies entre deux espèces pour accélérer l'identification des voies métaboliques	15h15-15h45	30mn
Pause		15 min
Analyse comparative de protéomes d'espèces apparentées (orthoMCL)	16h-17h	1h
Identification des voies métaboliques par intégration de méthodes d'assignation de codes enzymatiques et de classification selon la 'gene ontology'	17h-18h	1h

Mercredi 29

Module	Horaire	Durée
Analyse du polymorphisme AVEC génome de référence		
Détection des SNPs à partir de données de réséquencage (*GATK*) (09h00-10h00)	09h00-10h00	1h
Détection des SNPs (*samtools/VarScan*) pour analyse de leurs effets	10h00-10h30	30mn
Pause		15 min
Détection des SNPs (*samtools/VarScan*) pour calcul des fréquences alléliques	10h45-11h15	30mn
Analyse du polymorphisme SANS génome de référence		
Identification des SNP à partir de données de réséquencage	11h15-11h45	30mn
Analyse de données de type RAD-seq	11h45-12h45	1h
Repas – déjeuner		1 h 30
Présentation d'indicateurs qualité		
D'une expérience de RNAseq	14h55-15h30	35mn
D'un assemblage de transcriptome	14h30-14h55	25mn
Discussion	15h-16h30	1h30

Organisation de la formation

La formation a été organisée par Martine Da Rocha (Sophia) qui a suivi la formation de formateur interne et nous guide dans l'application de la méthode recommandée. Pour rappel la méthode est séquencée en trois parties « Découverte/Démonstration/Application », elle nous donne un cadre commun qui favorise à la fois l'homogénéité de notre support pédagogique et la reproductibilité de la formation.

Le résultat 2015 est un support de plus de 350 diapositives (100 de plus qu'en 2014) incluant les objectifs pédagogiques de chaque module, la partie théorique et les questions pour les travaux pratiques. Comme en 2014, ce support très complet sera réutilisé par ceux qui vont reproduire tout ou partie de la formation sur leurs sites respectifs (déjà prévu cet automne à Toulouse et à Sophia)

L'autre originalité de la méthode utilisée est 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 sait 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 27 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. Cette journée de préparation en conditions réelles nous permet d'affiner notre discours et de partager nos outils et méthodes entre bioinformaticiens du CATI. En termes d'animation interne, c'est un moment indispensable pour favoriser le développement de compétences partagées au sein de notre réseau.

Pour finir, l'organisation de cette formation a nécessité un travail important de la part de Martine pour nous trouver une salle parfaitement adaptée à notre formation et pour construire le support pédagogique dans les temps. Cela a aussi nécessité un travail très important de la part de Ludovic Legrand pour que tous les pipelines soient fonctionnels le jour J. Et comme l'an dernier, rien n'aurait été possible sans 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: Ludovic Legrand (LIPM/Toulouse), Jérôme Gouzy (LIPM/Toulouse) & Arnaud Ridel (SPE/Sophia)

Modules	Responsable et intervenant principal	Expert	Relecteur
Annotation fonctionnelle automatique d'un proteome à partir de la banque de domaines protéiques InterPro	Martial Briand (IRHS)	Sébastien Carrere (LIPM)	Adeline Simon (BIOGER)
Construction d'un protéome exhaustif, non redondant et annoté pour analyses protéomiques	Adeline Simon (BIOGER)	Martine Da Rocha & Corinne Rancurel (ISA)	Martial Briand (IRHS)
Analyse comparative de protéomes d'espèces apparentées (*orthoMCL*)	Corinne Rancurel (ISA)	Martial Briand (IRHS) Corinne Rancurel (ISA) Sébastien Carrere & Ludovic Cottret (LIPM)	Fabrice Legeai (IGEPP)
Transfert d'annotation par identification des relations d'orthologies entre deux espèces	Fabrice Legeai (IGEPP)	Ludovic Cottret (LIPM)	Corinne Rancurel (ISA)
Identification des relations d'orthologies entre deux espèces pour accélérer l'identification des voies métaboliques	Fabrice Legeai (IGEPP)	Ludovic Cottret (LIPM)	Ludovic Legrand (LIPM)
Identification des voies métaboliques par intégration de méthodes d'assignation de codes enzymatiques et de classification selon la 'gene ontology'	Ludovic Legrand (LIPM)	Patrice Baa-Puyoulet (BF2I)	Fabrice Legeai (IGEPP)
Détection des SNPs à partir de données de réséquencage (*GATK*)	Martine Da Rocha (ISA)	Fabrice Legeai (IGEPP)	Joseph Tran (IJPB)
Analyse de données de type RAD-seq	Bernhard Gschloessl (CBGP)	Anthony Bretaudeau (IGEPP)	Erika Sallet (LIPM)
Détection des SNPs (*samtools/VarScan*) pour analyse de leurs effets	Anthony Bretaudeau (IGEPP)	Ludovic Legrand (LIPM)	Martine Da Rocha (ISA)
Détection des SNPs (*samtools/VarScan*) pour calcul des fréquences alléliques	Joseph Tran (IJPB)	Ludovic Legrand (LIPM)	Anthony Bretaudeau (IGEPP)
Identification des SNP à partir de données de réséquencage (DiscoSNP) sans génome de référence	Erika Sallet (LIPM)	Fabrice Legeai (IGEPP)	Bernhard Gschloessl (CBGP)
Indicateurs qualité d'un assemblage de transcriptome	Franck Dorkeld (CBGP)	Bernhard Gschloessl (CBGP)	Anthony Bretaudeau (IGEPP)
Indicateurs qualité d'une expérience de RNAseq	Ludovic Cottret (LIPM)	Joseph Tran (IJPB) Adeline Simon (BIOGER)	Franck Dorkeld (CBGP)

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 par nos collègues biologistes. Les réponses au questionnaire se trouvent à l'url http://cati-bbric.toulouse.inra.fr/lib/exe/fetch.php/bbric-ag-communaute-201504-28-29-resultats_questionnaire_evaluation_biologistes.pdf

La critique principale est le manque de temps pour analyser les résultats des TP. L'an passé nous avions moins de modules et prévu trop de temps pour certains modules. Cette année nous avions plus de modules et pour certains pas assez de temps.

Pour le reste il s'agit plus de commentaires que de critiques et le curseur est nettement vers les commentaires positifs. Nos collègues ont eu la sensation d'apprendre, de développer leur culture en génomique, leur connaissances des possibles dans l'analyse de leur données et se sentent capables d'utiliser nos outils ou de venir nous trouver si besoin.

Il était très intéressant que 6 personnes aient suivi les formations 2014 et 2015 car elles ont pu comparer notre fonctionnement et nous avons également pu comparer l'évolution de leur perception.

La remarque la plus agréable est que nous avons tenu compte de leurs critiques/demandes de l'an passé. Aussi bien pour trouver une salle de formation plus agréable que dans les thèmes abordés. L'aspect des conditions de travail est loin d'être négligeable, la salle trouvée par Martine a l'UPMC et le service/support proposé par l'UPMC étaient parfaits et très professionnels. Le fait que la formation se déroule pendant les vacances scolaires a sans doute joué pour rendre les pauses (dans les couloirs vides) vraiment efficaces.

De notre côté, nous avons pu observer une différence de perception sur les formats bioinformatiques standards. La première année les « débutants » en génomique critiquent le fait qu'il n'y a pas assez de temps pour les analyser et les comprendre, la deuxième année cette critique tend à disparaitre.

Autoanalyse collective et pistes d'améliorations pour l'an prochain

La structure de la formation est très bonne, nous n'allons pas la changer mais l'ajuster pour la rendre encore plus efficace. Cette partie est une synthèse des (meilleures) idées que nous avons échangées lors du débriefing.

Grace au soutien financier du SPE nous avons pu payer les 1230€ pour louer la salle de formation de l'UPMC, les conditions étaient vraiment excellentes, nous essayerons de revenir l'an prochain.

La question des formats bioinformatiques standards (fasta, gff3, vcf) est centrale. Pour que les débutants se sentent plus vite à l'aise **nous rajouterons à notre introduction** (a la suite de la présentation de BBRIC, de l'Archive et de Galaxy) **un module de description des formats standards et de visualisation de ces formats dans un genome browser** (1h). En faire un module nous permettra de le jouer ou pas en fonction du public de notre formation et devrait nous permettre de gagner du temps sur les autres présentations et de limiter les redondances sur ce sujet. Dans le support du module il faudra penser à inclure des diapositives de type « pense bête » indépendantes du support.

Concernant le temps alloué aux modules, le compromis est difficile à trouver mais la première année nous avions prévu plutôt 2h, cette année plutôt 1h, l'an prochain cela sera **1h30 a 2h par module**.

Le temps supplémentaire sera alloué aux TP. De plus, comme nous l'avions déjà constaté l'an passé, le rôle du relecteur est trop marginal, il faut le renforcer. Pour cela il faudra :
- finir plus tôt la mise en place des workflows
    - ➔ LudoL rédigera un petit guide des bonnes pratiques pour faire en sorte que l'intégration des protocoles dans galaxy soit plus facile/rapide (ex : éliminer des formulaires les paramètres que l'on ne change jamais ou qui n'ont aucun sens pour personne si ce n'est l'auteur de l'outil)
    - ➔ seuls les pipelines en production au 31 janvier feront l'objet de la formation
- finir plus tôt les supports de cours
    - ➔ un coordinateur par thème sera nommé afin de faire respecter un calendrier plus strict et afin de supprimer au maximum les redondances dans le support (au prix de plus de visio de préparation à la demande expresse de Fabrice)
- mieux équilibrer les contributions de chacun.

A ces conditions **le « relecteur » pourra avoir la charge de préparer et de présenter les exercices**. L'idée sera de faire un TP en deux parties. La première correspondra « au cas qui marche » et la deuxième correspondra « au cas qui ne marche pas » ou qui pose un problème d'interprétation.

Prévision pour la formation de l'an prochain

Comme en 2013 et 2014, 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 formation de l'an prochain. Voici les commentaires que l'on peut faire et les priorités que l'on peut définir pour l'an prochain.

Thèmes	Commentaires
Metagénomique (7 votes)	**C'est un thème déjà listé l'an passé et que nous avions redirigé vers le metaprogramme MEM. Il s'avère que plusieurs d'entre nous sommes sollicités pour assurer le support de projets de métagénomique. Cela devient donc prioritaire pour nous de maitriser les outils et méthodes et d'en profiter pour offrir une formation sur ce sujet l'an prochain.**

Epigénétique/smallRNA (8 votes)	**C'est un sujet qui intéresse de plus en plus de laboratoires. Nous avions prévu d'offrir des outils et formations pour 2016, on le fera.**
Phylogénie avancée	On peut faire des formations d'initiation mais pour des formations avancées il vaut 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 à participer mais pour une formation il vaut encore mieux s'adresser directement aux formateurs plutôt qu'aux élèves que nous étions.
Traitement d'autres technologies	Il est important de mettre à jour nos pipelines existant pour fournir des solutions pour analyses d'autres types de données. C'est très vrai pour PacBio qui se développe de plus en plus.
Visualisation	**L'importance de la problématique de visualisation/représentation était apparue avec force lors de l'AG « orientée biologiste » de 2013 et est récurrente depuis. Nous avons commencé à l'aborder cette année et nous allons continuer tout 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 2016 seront donc les outils pour la métagénomique, l'epigénomique et les solutions de visualisation.

Pour Phylogénie et Génétique, ce sont les mêmes items (peu de personnes) et donc les mêmes réponses que l'an passé. Il n'est pas question de transférer toutes les annonces de formations sur la liste communaute-bbric mais par contre il faut que nous le fassions lorsqu'il y a une annonce sur des questions que nous ne pensons pas traiter (phylogénie/genet des pop/GWA).

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

Depuis la fin 2014, le SPE soutien financièrement et significativement les projets et le fonctionnement du CATI. En 2014, le projet BBRIC Archive a ainsi bénéficié d'un financement de 32k€ (LIPM : 20k€ ; CBGP : 6.5k€ ; IGEPP : 6k€)

Pour 2015, le budget du CATI est de 46,7k€ composé de 5,4k€ de métabolisme de base (4,6k€ SPE+ 0,8k€ BAP) et de 41,3 k€ de soutien (cf. budget prévisionnel). Le budget est géré par le LIPM et les actions réalisées à ce jour sont les suivantes :

Aide en fonction des contributions aux projets du CATI en 2014	
LIPM Toulouse	1370
ISA Sophia	900
CBGP Montpellier	600
IRHS Angers	300
BF2I Lyon	200
IJPB Versailles	400
Location de la salle de formation à l'UPMC	1230
Contribution aux missions du CATI pour 2015	
LIPM Toulouse	<3000
IGEPP Rennes	2000
ISA Sophia	2000

BIOGER Versailles	600
IRHS Angers	600
BF2I Lyon	800
SPE « contributeur » de la PF bioinformatique Genotoul (période mai 2015-avril 2016)	5000

La prochaine action (d'ici la fin de l'été) est d'acheter un serveur puissant sur lequel nous pourrons créer des machines virtuelles pour nos projets spécifiques ou partagés, partager des logiciels sous licence (on va acheter blast2go), exécuter le galaxy du CATI et surtout partager encore plus facilement nos outils d'analyse. L'autre principale action est d'organiser à l'automne une formation technique sur le framework D3.js.