

Mise en place de serveurs Galaxy dans le cadre du réseau CATI BBRIC

{[Sebastien.Carrere](mailto:Sebastien.Carrere@toulouse.inra.fr), [Ludovic.Legrand](mailto:Ludovic.Legrand@toulouse.inra.fr), [Jerome.Gouzy](mailto:Jerome.Gouzy@toulouse.inra.fr)}@toulouse.inra.fr
{[Fabrice.Legeai](mailto:Fabrice.Legeai@rennes.inra.fr), [Anthony.Bretaudeau](mailto:Anthony.Bretaudeau@rennes.inra.fr)}@rennes.inra.fr

CATI BBRIC



- ▶ 35 bioinformaticiens
- ▶ 15 unités
- ▶ 10 sites

Domaines d'applications

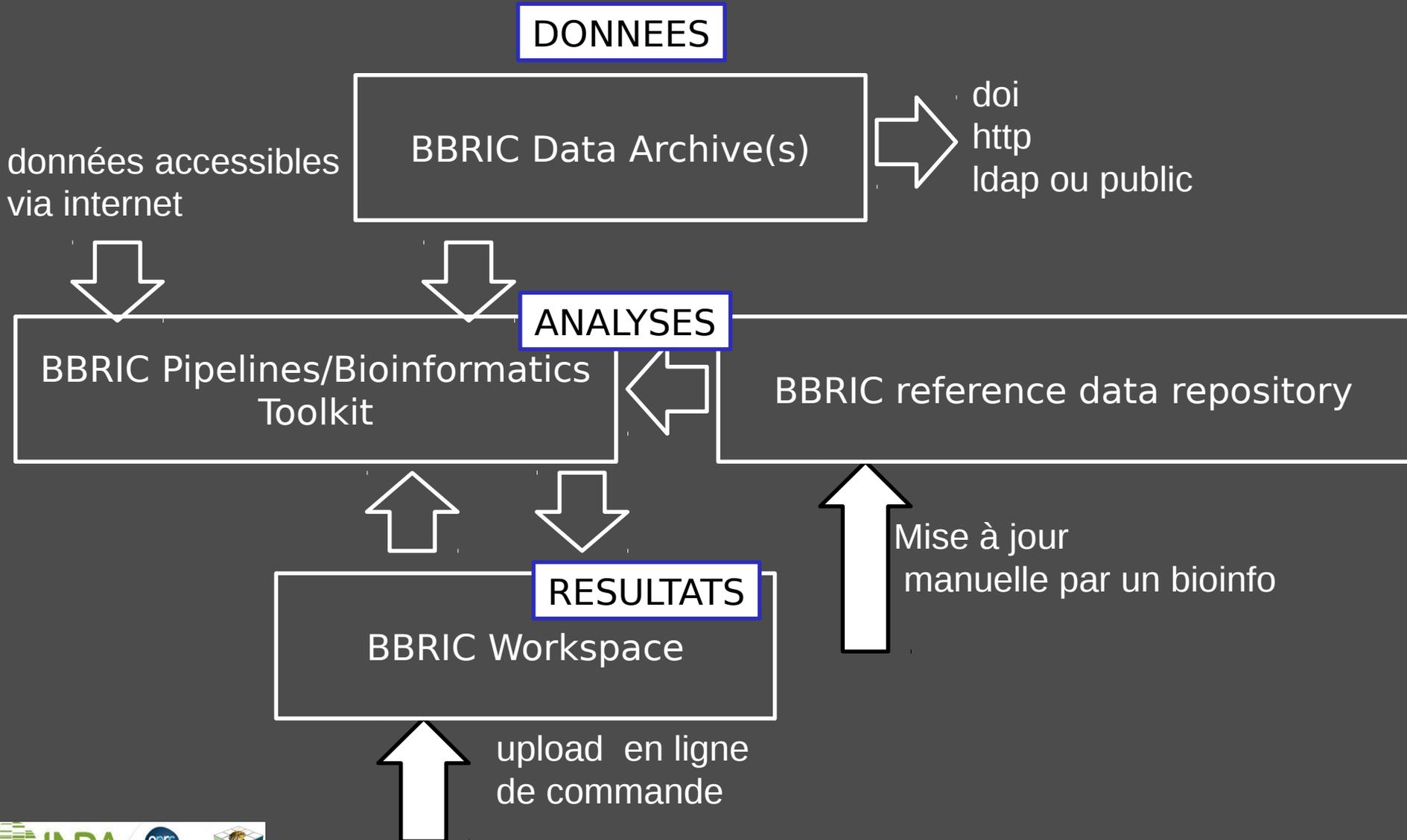
- ▶ Assemblage (genomes/transcriptomes)
- ▶ Annotation structurale des génomes (gènes codants pour des protéines / ncRNA)
- ▶ Annotation fonctionnelle (genomes/transcriptomes)
- ▶ Analyse de l'expression (RNAseq, puces)
- ▶ Détection et analyse du polymorphisme
- ▶ Métagénomique
- ▶ Epigénomique
- ▶ Modélisation des réseaux métaboliques et de régulation.
- ▶ Gestion de collections (bactéries, insectes, etc.)
- ▶ Systématique: identification des espèces de groupes d'espèces d'intérêt
- ▶ Phylogénie, évolution
- ▶ Génomique des populations

Modèles d'intérêt

- ▶ Plantes
- ▶ Bactéries
- ▶ Insectes
- ▶ Nématodes
- ▶ Champignons
- ▶ Oomycetes
- ▶ Virus

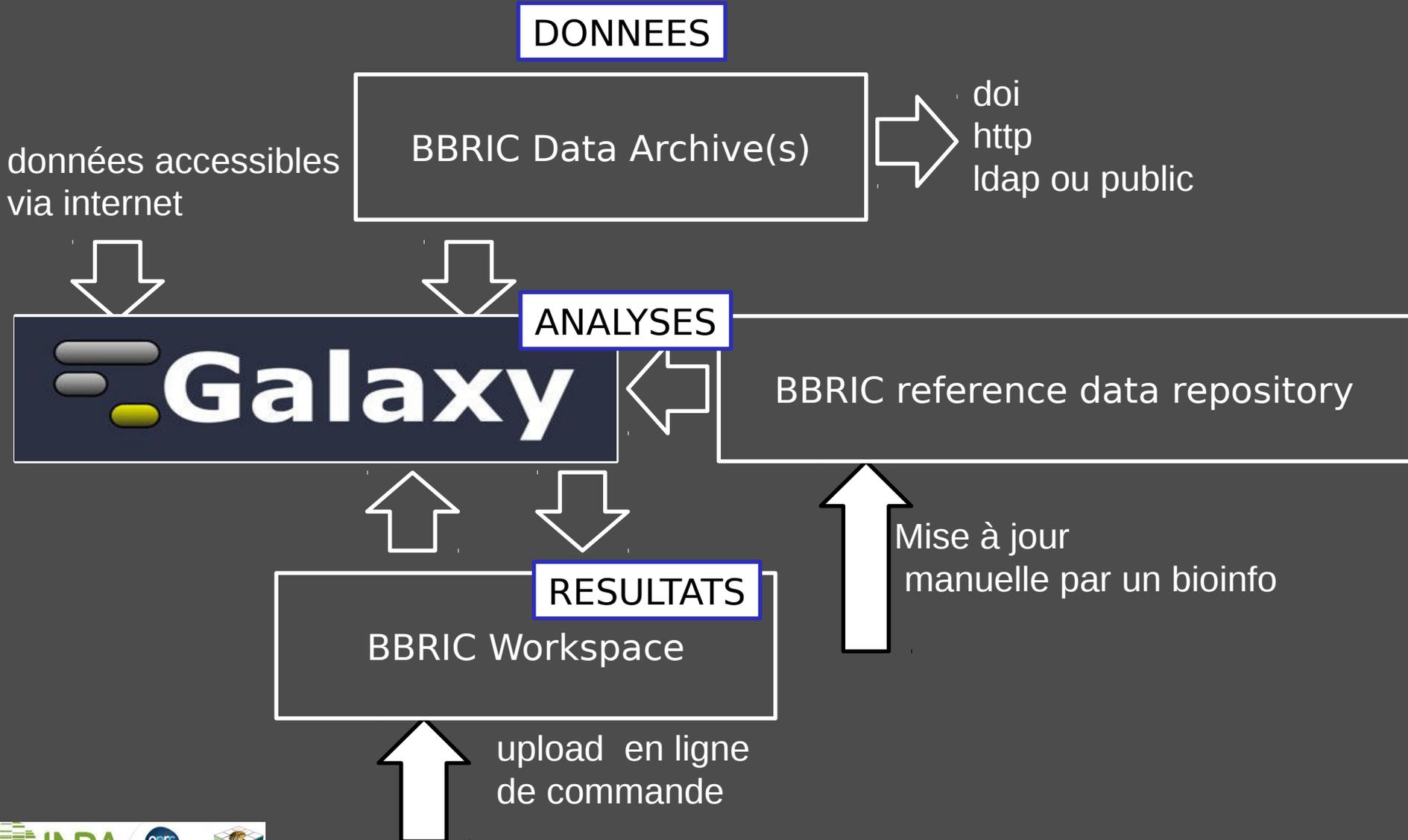
Architecture bioinformatique BBRIC

4 composants interoperables



Architecture bioinformatique BBRIC

4 composants interoperables



Galaxy @ BBRIC

S'adapter au réseau

- ▶ Des instances au plus proche des données
 - @Rennes
 - Accès contrôlé
 - ▶ annuaire LDAP
 - @Toulouse
 - Accès contrôlé
 - ▶ Fédération Education Recherche (Shibboleth)
 - @Ailleurs ?
- ▶ Un toolshed pour partager les outils
- ▶ Un système pour maintenir les données de référence

Galaxy @ BBRIC

Pré-requis avant un passage en production

- ▶ Utilisation de fichiers compressés
 - fastq.gz
- ▶ Interopérabilité avec les instances d'Archive
 - Les utilisateurs doivent pouvoir analyser leurs données
- ▶ Interopérabilité avec les instances Workspace
 - Les utilisateurs doivent pouvoir être autonomes dans le post-processing des résultats intermédiaires

Utilisation de fichiers compressés

- ▶ De nombreux outils le permettent, pourquoi s'en priver ?
- ▶ Pourquoi Galaxy décompresse automatiquement ?
- ▶ Extension des datatypes
 - `datatypes_conf.xml`
 - `lib/galaxy/datatypes/data.py`
- ▶ Modification du code de l'upload
 - `tools/data_source/upload.py`
- ▶ Modification des fiches des outils pour ajout du nouveau type accepté

Interopérabilité avec Archive

Qu'est-ce que l'Archive ?

- ▶ Un système de conservation des données brutes
 - générique
 - sans limite de taille de fichier
- ▶ Un système ouvert
 - accès programmatique (API, client)
 - conversion vers les formats standards (SRA, GEO pour les séquences)
 - publication des données/métadonnées (DOI, lien permanent)
- ▶ Un accès sécurisé et authentifié
 - droits d'accès par fichier
 - authentification par la fédération Education Recherche
- ▶ Un outil accessible aux producteurs de données (les biologistes)
 - interface web (consultation, gestion)
 - moteur de recherche

Interopérabilité avec Archive

Galaxy / BBRIC

Analyze Data Workflow Shared Data Visualization Admin Help User

Tools

search tools

LIPM RNAseq LIPM RNAseq
eggnip Annotation on bacterial genome with RNAseq
Get Data

BBRIC Workspace server
BBRIC Archive server
Upload File from your computer
UCSC Main table browser
UCSC Archaea table browser
BX table browser
EBI SRA ENA SRA
BioMart Central server
GrameneMart Central server
Flymine server
modENCODE fly server
modENCODE modMine server
MouseMine server
Ratmine server
YeastMine server
modENCODE worm server
WormBase server
EuPathDB server
EncodeDB at NHGRI
EpiGRAPH server
GenomeSpace import from file browser

sebastien.carrere@toulouse.inra.fr
scarrere@inra.fr

Your Archive Dev Portal - Collection

Sequences Account Admin Logout

Your Archive Dev Portal

Interopérabilité avec Archive

Galaxy / BBRIC

Analyze Data Workflow Shared Data Visualization Admin Help User

Tools

search tools

LIPM RNAseq LIPM RNAseq
eggnip Annotation on bacterial genome with RNAseq
Get Data
BBRIC Workspace server
BBRIC Archive server
Upload File from your computer
UCSC Main table browser
UCSC Archaea table browser
BX table browser
EBI SRA ENA SRA
BioMart Central server
GrameneMart Central server
Flymine server
modENCODE fly server
modENCODE modMine server
MouseMine server
Ratmine server
YeastMine server
modENCODE worm server
WormBase server
EuPathDB server
EncodeDB at NHGRI
EpiGRAPH server
GenomeSpace import from file browser

sebastien.carrere@toulouse.inra.fr
scarrere@inra.fr

Your Archive Dev Portal - Sequence Collection

Sequences Account Admin Logout
Submit Search Browse Manage

Quick search Envoyeur (use * for partial search, ex:'phospho*' will match with for 'phosphorylase', 'phosphokinase')

Contributor	Date	Title	Species	Molecule	Project	Metadata
BBRIC Demo	20131024	GSM1185104	Arabidopsis thaliana	total_RNA	INRA	
<input type="button" value="Send to galaxy"/>		111.94 Mo	bbbic.demo@gmail.com, LIPM			

All files are public All files are shared or belong to you Some files are shared All files are private

Interopérabilité avec Archive

The screenshot displays the Galaxy / BBRIC web interface. At the top, the navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The main header reads 'Your Archive Dev Portal - Sequence Collection'. A search bar for tools is visible on the left. A large green notification box in the center states: 'The following job has been successfully added to the queue: 2: SRR932114.fastq.gz'. Below this, instructions are provided on how to check job status and view data in the 'History' pane. A legend at the bottom of the notification explains the status changes: 'running' to 'finished' for successful jobs, and 'error' for failed ones. On the right, a 'History' panel shows two entries: 'Unnamed history' (111.3 MB) and '2: SRR932114.fastq.gz'. Below the notification, a table shows a file 'Send to galaxy' (111.94 Mo) shared with 'bbric.demo@gmail.com, LIPM'. A legend below the table defines the sharing status: 'All files are public' (blue), 'All files are shared or belong to you' (green), 'Some files are shared' (yellow), and 'All files are private' (red). The left sidebar contains a list of external data sources and servers, including UCSC Main table browser, UCSC Archaea table browser, BX table browser, EBI SRA ENA SRA, BioMart Central server, GrameneMart Central server, Flymine server, modENCODE fly server, modENCODE modMine server, MouseMine server, Ratmine server, YeastMine server, modENCODE worm server, WormBase server, EuPathDB server, EncodeDB at NHGRI, EpiGRAPH server, and GenomeSpace import from file browser.

Interopérabilité avec Archive

Galaxy / BBRIC Analyze Data Workflow Shared Data Visualization Admin Help User

Tools Your Archive Dev Portal - Sequence Collection

search tools

LIPM **egnp** **geno** **Get I** **BB** **BB** **Up** **cor** **UCSC Main tabl** **UCSC Archaea t** **BX table browse** **EBI SRA ENA SR** **BioMart Central** **GrameneMart C** **Flymine server** **modENCODE fly** **modENCODE md** **MouseMine serv** **Ratmine server** **YeastMine serv** **modENCODE wd** **WormBase serv** **EuPathDB serv** **EncodeDB at NHGRI** **EpiGRAPH server** **GenomeSpace import from file browser**

History Refresh Settings

Unnamed history Share Comment

111.3 MB

History Refresh Settings

Unnamed history Share Comment

222.7 MB

2: SRR932114.fastq.gz View Edit Delete

done
format: fastq.gz, database: 2

1: SRR932114.fastq.gz View Edit Delete

Tool: BBRIC Archive

Name: SRR932114.fastq.gz
Created: Wed Nov 27 15:58:49 2013 (UTC)
Filesize: 111.3 MB
Dbkey: ?
Format: fastq.gz
Galaxy Tool Version: 1.0.0
Tool Version:
Tool Standard Output: [stdout](#)
Tool Standard Error: [stderr](#)
Tool Exit Code: 0
API ID: 22c94bd07149de56
Full Path: /www-galaxy/database/files/000/dataset_214.dat
Job Command-Line: python /www-galaxy/shed_tools/toolshed.genouest.org/repos/abretaud/bbric_archive/e4b9c92213f0/bbric_archive/bbric_archive.py /www-galaxy/database/files/000/dataset_214.dat Sebastien.Carrere@toulouse.inra.fr 2 0

Input Parameter	Value	Note for rerun
user_email	not used (parameter was added after this job was run)	
user_id	not used (parameter was added after this job was run)	
GALAXY_URL	not used (parameter was added after this job was run)	

Inheritance Chain

SRR932114.fastq.gz

Interopérabilité avec Workspace

Galaxy / BBRIC Analyze Data Workflow Shared Data Visualization Admin Help User

Workspace DEV

Tools

search tools

[LIPM RNAseq](#) [LIPM RNAseq](#)
[eggnog Annotation on bacterial genome with RNAseq](#)

Get Data

BBRIC Workspace server

[BBRIC Archive server](#)

[Upload File from your computer](#)

[UCSC Main table browser](#)

[UCSC Archaea table browser](#)

[BX table browser](#)

[EBI SRA ENA SRA](#)

[BioMart Central server](#)

[GramenaMart Central server](#)

[Flymine server](#)

[modENCODE fly server](#)

[modENCODE modMine server](#)

[MouseMine server](#)

[Ratmine server](#)

[YeastMine server](#)

[modENCODE worm server](#)

[WormBase server](#)

[EuPathDB server](#)

[EncodeDB at NHGRI](#)

[EpiGRAPH server](#)

[GenomeSpace import from file browser](#)

sebastien.carrere@toulouse.inra.fr
scarrere@inra.fr

Analyses **Search** **Manage** **Admin** **Logout**

Last 10 Analyses All Analyses

Quick search Envoyer

List of the 10 last analyses

	Date	Title	Description	Context
	20131106	BRADI_data	Donnees FATAL	service

Interopérabilité avec Workspace

Galaxy / BBRIC Analyze Data Workflow Shared Data Visualization Admin Help User

Workspace DEV

Tools

search tools

[LIPM RNAseq](#) [LIPM RNAseq](#)
[eggnip](#) [Annotation on bacterial genome with RNAseq](#)

Get Data

BBRIC Workspace server

[BBRIC Archive server](#)

[Upload File from your computer](#)

[UCSC Main table browser](#)

[UCSC Archaea table browser](#)

[BX table browser](#)

[EBI SRA](#) [ENA SRA](#)

[BioMart](#) [Central server](#)

[GramenaMart](#) [Central server](#)

[Flymine](#) [server](#)

[modENCODE](#) [fly server](#)

[modENCODE](#) [modMine server](#)

[MouseMine](#) [server](#)

[Ratmine](#) [server](#)

[YeastMine](#) [server](#)

[modENCODE](#) [worm server](#)

[WormBase](#) [server](#)

[EuPathDB](#) [server](#)

[EncodeDB](#) [at NHGRI](#)

[EpiGRAPH](#) [server](#)

[GenomeSpace](#) [import from file browser](#)

sebastien.carrere@toulouse.inra.fr
scarrere@inra.fr

Analyses **Search** **Manage** **Admin** **Logout**

Last 10 Analyses **All Analyses**

Quick search Envoyer

List of the 10 last analyses

	Date	Title	Description	Context
	20131106	BRADI_data	Donnees FATAL	service

Interopérabilité avec Workspace

Galaxy / BBRIC

Analyze Data Workflow Shared Data Visualization Admin Help User

Workspace DEV

Tools

search tools

LIPM RNAseq LIPM RNAseq
Annotation on bacterial genome with RNAseq

Get Data

- BBRIC Workspace server
- BBRIC Archive server
- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- BX table browser
- EBI SRA ENA SRA
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- MouseMine server
- Ratmine server
- YeastMine server
- modENCODE worm server
- WormBase server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- GenomeSpace import from file browser

sebastien.carrere@toulouse.inra.fr
scarrere@inra.fr

Bioinformatique
Biodiversité
Représentation
Intégration
des
Connaissances

Analyses Search Manage Admin Logout

Last 10 Analyses All Analyses

BRADI_data - 20131106

Summary

Owner	sebastien.carrere@toulouse.inra.fr
Title	BRADI_data
Date	20131106
Context	service
Description	Donnees FATAL
Provider	sebastien.carrere@toulouse.inra.fr
Share	

Files

Send to galaxy

e566972ed31b8ae67b1a96e32147e7db_20131106.BRADI_data.zip

- BRADI_data
 - BRADI.clusters
 - BRADI.cfg.mask
 - tair
 - BRADI.peptides
 - BRADI.reads
 - BRADI.pep.ipr...
 - BRADI.report
 - BRADI.FrameD

Expand this branche
Display info
galaxy Link
Save as...

Interopérabilité avec Workspace

Galaxy / BBRIC

Analyze Data Workflow Shared Data Visualization Admin Help User

Workspace DEV

Tools

search tools

LIPM RNAseq LIPM RNAseq
 eqpp Annotation on bacterial genome with RNAseq
 Get Data
 BBRIC Workspace server

rrere@toulouse.inra.fr
 ra.fr

Bi-informatique
 Biodiversité
 Représentation
 Intégration
 des
 Connaissances

Analyses Search Manage Admin Logout

Last 10 Analyses All Analyses

Analyze Data Workflow Shared Data Visualization Admin Help User

Using 224.2 MB

Brad000371_132_566_3	044996DE9E52498F	144	HMMPfam	PF00234	Tryp_alpha_amyl	62	139	0.0010999999999999999
Brad000371_132_566_3	044996DE9E52498F	144	superfamily	SSF47699	Bifunc_inhib/LTP/seed_store	64	143	1.700002955900538E-2
Brad000371_132_566_3	044996DE9E52498F	144	HMMSmart	SM00499	AAI	62	143	6.400000387884858E-
Brad000371_132_566_3	044996DE9E52498F	144	Gene3D	G3DSA:1.10.110.10	LPT_helical	59	143	1.3999989204987804E-2
Brad000001_211_1497_1	A82E34902D6AF5CA	428	HMMPfam	PF00004	AAA	154	295	6.6e-1
Brad000004_76_852_1	09578AC44C4B89C2	258	HMMPfam	PF00504	Chloroa_b-bind	58	224	1.4e-5
Brad000005_1_1274_3	0CCC384FD028407E	423	HMMPfam	PF00274	Glycolytic	79	423	2.3e-16
Brad000006_91_1104_1	9203040545794FB7	337	HMMPfam	PF02800	Gp_dh_C	159	316	5.6e-7
Brad000006_91_1104_1	9203040545794FB7	337	HMMPfam	PF00044	Gp_dh_N	4	154	3.3e-5
Brad000007_630_1309_3	20098A0D4B7AB60E	226	HMMPfam	PF01536	SAM_decarbox	12	226	1e-8
Brad000008_1_1262_3	145AD88C77835C39	419	HMMPfam	PF00240	ubiquitin	44	112	1.2e-3
Brad000008_1_1262_3	145AD88C77835C39	419	HMMPfam	PF00240	ubiquitin	120	188	1.2e-3
Brad000008_1_1262_3	145AD88C77835C39	419	HMMPfam	PF00240	ubiquitin	196	264	1.2e-3
Brad000008_1_1262_3	145AD88C77835C39	419	HMMPfam	PF00240	ubiquitin	272	340	1.2e-3
Brad000008_1_1262_3	145AD88C77835C39	419	HMMPfam	PF00240	ubiquitin	348	416	1.2e-3
Brad000009_109_1599_1	E8AED979303DE70D	496	HMMPfam	PF00199	Catalase	18	399	2.3e-17

History

Unnamed history
222.8 MB

4: BRADI.pep.iprscan

done
format: tabular, database: ?

Brad000371_132_566_3	044996DE9E52498F
Brad000001_211_1497_1	A82E34902D6AF5CA
Brad000004_76_852_1	09578AC44C4B89C2

YeastMine server
 modENCODE worm server
 WormBase server
 EuPathDB server
 EncodeDB at NHGRI
 EpiGRAPH server
 GenomeSpace import from file browser

- BRADI.clusters
- BRADI.cfg.mask
- tair
- BRADI.peptides
- BRADI.reads
- BRADI.pep.iprscan
- BRADI.report
- BRADI.FrameD

Expand this branche
 Display info
 galaxy Link
 Save as...

Atelier 29-30/10/2013

- ▶ Quinzaine de bioinformaticiens du CATI BBRIC
- ▶ 1/3 connaissait les aspects techniques de Galaxy
- ▶ Etape 1: déployer un outil simple dans Galaxy from scratch
 - Démystifier Galaxy
- ▶ Etape 2: utiliser Appli.pm
 - Se faciliter la vie
- ▶ Etape 3: Déployer un outil complexe (pipeline)
 - Se rendre utile

Atelier 29-30/10/2013

- ▶ Etape 1: déployer un outil simple dans Galaxy from scratch
 - Predotar, tmhmm, signalpHMM
 - Bilan des difficultés rencontrées
 - Gestion dynamique des formats de sortie
 - ▶ Paramètres conditionnels
 - ▶ Itérateurs / collection
 - ▶ Sélection des *datatypes*
 - ▶ Gestion des paths
 - ▶ Débogage

Mais tout le monde peut y arriver
et tout le monde y est arrivé !

Atelier 29-30/10/2013

► Etape 2: Appli.pm

► Qu'est-ce ?

- Un module Perl de la « lipmutils » permettant de décrire un programme.
- Développé pour déployer des services Mobyly & BioMOBY étendu à Galaxy

► Pourquoi faire?

- décrire une et une seule fois son programme
- générer autant d'usage que d'outils (CLI, Mobyly, Galaxy, JSON, **next ?**)

► Sous quelles contraintes?

- Perl
- Connaitre les datatypes des interfaces (mobyly ≠ biomoby ≠ galaxy ≠ GetOptLong ≠ EDAM)

Atelier 29-30/10/2013

► Etape 2: Appli.pm

► Quel cas d'utilisation ?

- un programme à ajouter dans Galaxy
- ce programme a une interface en ligne de commande
- j'écris un wrapper Perl utilisant Appli pour générer son usage Galaxy

► Résultat

- Permet d'obtenir rapidement un outil fonctionnel
- Permet de générer facilement une trame du XML Galaxy
 - *≧ Gère les collections, mais pas les paramètres conditionnels*

Atelier 29-30/10/2013

- ▶ Etape3: Déployer un outil complexe (pipeline)
- ▶ EuGène-P
 - Pipeline d'annotation de génomes procaryotes
 - ▶ Intègre des données de similarité (Blastx)
 - ▶ Intègre des données de transcriptomique (RNA-seq)
 - ▶ Intègre des résultats de prédicteurs *ab-initio* (Prodigal, tRNAScan-SE,...)
 - Le wrapper devait
 - ▶ prendre en entrée X banques de RNA-seq, un génome nu
 - ▶ générer l'environnement (fichiers de conf, arborescence pour l'exécution du pipeline)
 - ▶ produire un fichier d'annotation au format GFF3

OK pour la majorité des binômes, quelque soit le langage et la méthode utilisée

Bilan de l'atelier

- + Montée en compétence commune
- + Prise en main de l'outil
- Crash des *handlers*
 - En cause: l'écriture de caractères spéciaux sur stderr !
- Difficultés logs
 - Comment trouver les commandes exécutées ?
- ~ Les types de données, les ontologies
- ~ Nécessité de redémarrer régulièrement
 - Problèmes de cache

Bonnes pratiques

- ▶ Toolshed
 - Partager les outils
- ▶ Appli.pm (ou autre ?)
 - Si demain, l'outil à la mode n'est plus Galaxy
- ▶ Déployer des outils de haut niveau
 - Utilisables
 - Utiles
- ▶ Fichiers compressés
 - Obligatoire dans un environnement de production

Merci