



<http://get.genotoul.fr>



<http://bioinfo.genotoul.fr>

Gestion des données de l'HiSeq 2000

Plateformes GenoToul Bioinformatique et GeT-PlaGe
Ch. Klopp / G. Salin



07 octobre 2011 – gerald.salin@toulouse.inra.fr

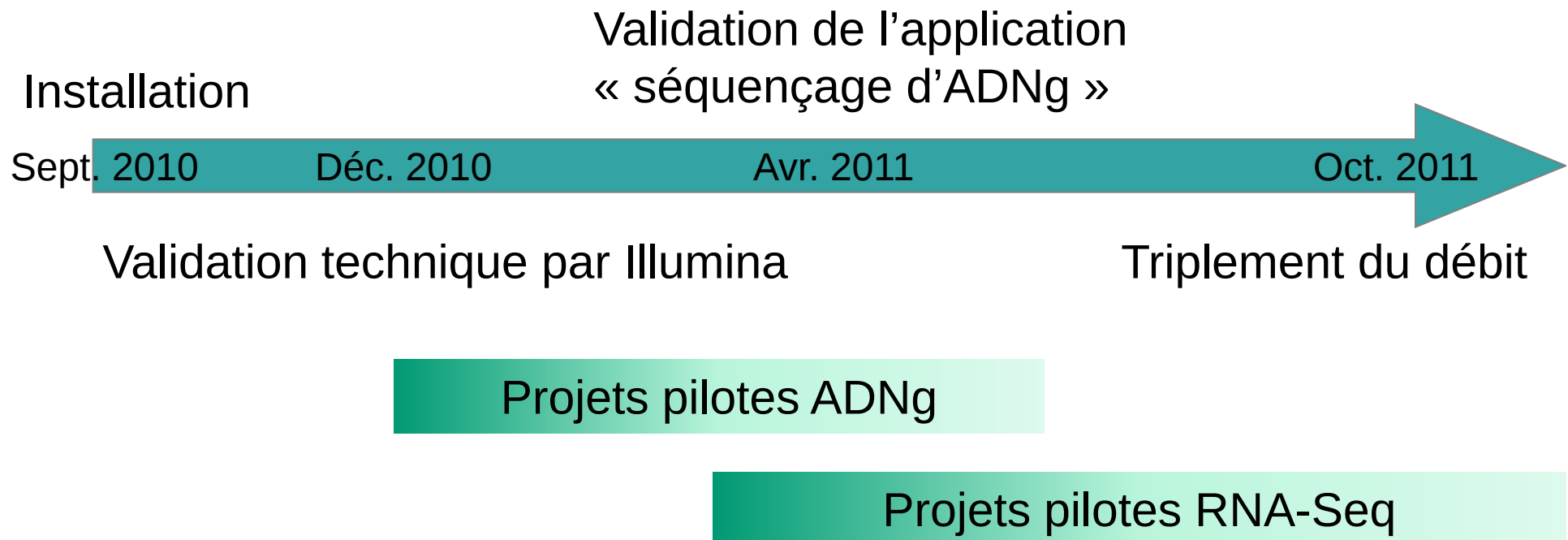


Contexte NGS

- Collaboration entre PF GeT-PlaGe et Bioinfo depuis 2007
 - PF GeT-PlaGe : production de données (454, HiSeq2000)
 - PF Bioinfo : moyens d'analyse et de mise à disposition
- CPER/IBiSA communs pour les investissements dans le NGS
- Acteurs du projet France Génomique (Investissements d'avenir)

L'HiSeq2000, côté PlaGe

Quelques dates



L'HiSeq2000, côté PlaGe

Quelques chiffres

36 flowcells en 2x101b depuis le 23/11/2010 + 8 HS

- 288 lanes - autant d'analyses nG6
 - 227 ADNg
 - 61 ARN
- 28 projets différents
- 366 librairies
- ~54To de données brutes transférées

19 flowcells planifiées

L'HiSeq2000, côté PlaGe

Notre expérience

- : Problèmes réactifs + machine (8 flowcells repassées)
- : Dépendance vis-à-vis de l'infrastructure informatique
- + : Beaucoup de problèmes logiciels corrigés par Illumina
- ++ : Un débit plus élevé qu'annoncé (+40%)
 - 130 M de reads / Lane attendus selon spécifications Illumina en Paired-ends
 - 185 M de reads / Lane en moyenne
- ++ : Retour positif des utilisateurs

L'HiSeq2000, côté Bioinfo

Une infrastructure en conséquence

- Réseau à 1Gb/s avec connexion directe sur un NAS
- 14 To pour le transfert des données brutes
- 8 To pour la sauvegarde des données analysées
- Cluster avec 49 nœuds de calcul (> 400 coeurs)
- nG6, un système d'information dédié au NGS

L'HiSeq2000, côté Bioinfo

Historique de nG6

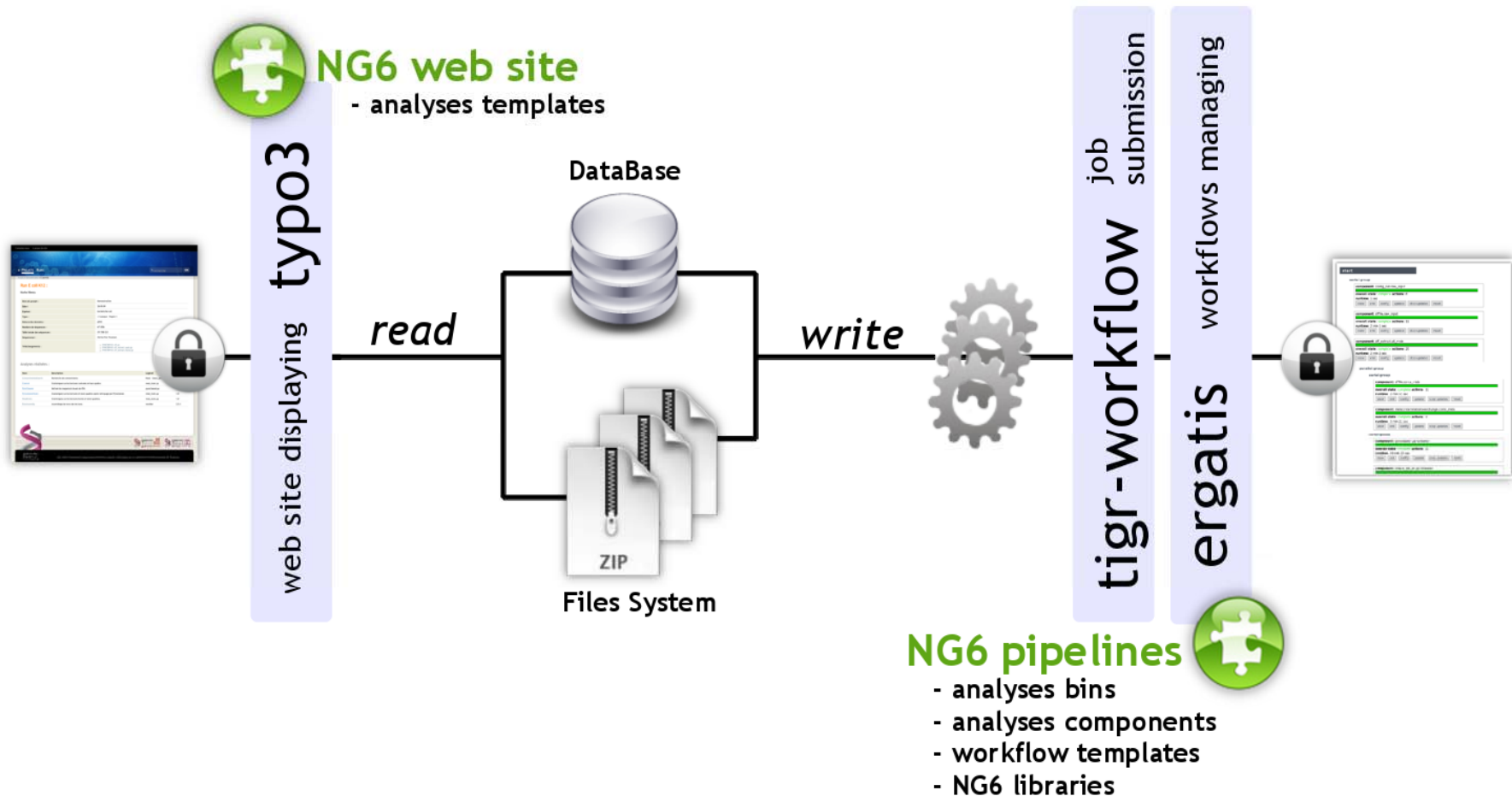
- Projet commun Bioinfo/PlaGe depuis 2008
- Pour la gestion des données Roche 454 :
 - Traçabilité des données et analyses associés
 - Portail sécurisé d'accès aux données pour les utilisateurs
 - Sauvegarde des données analysées
- Intégration des premières données HiSeq fin 2010

Sites web

- Démo : <http://ng6.toulouse.inra.fr>
- Forge : <http://mulcyber.toulouse.inra.fr/projects/ng6/>

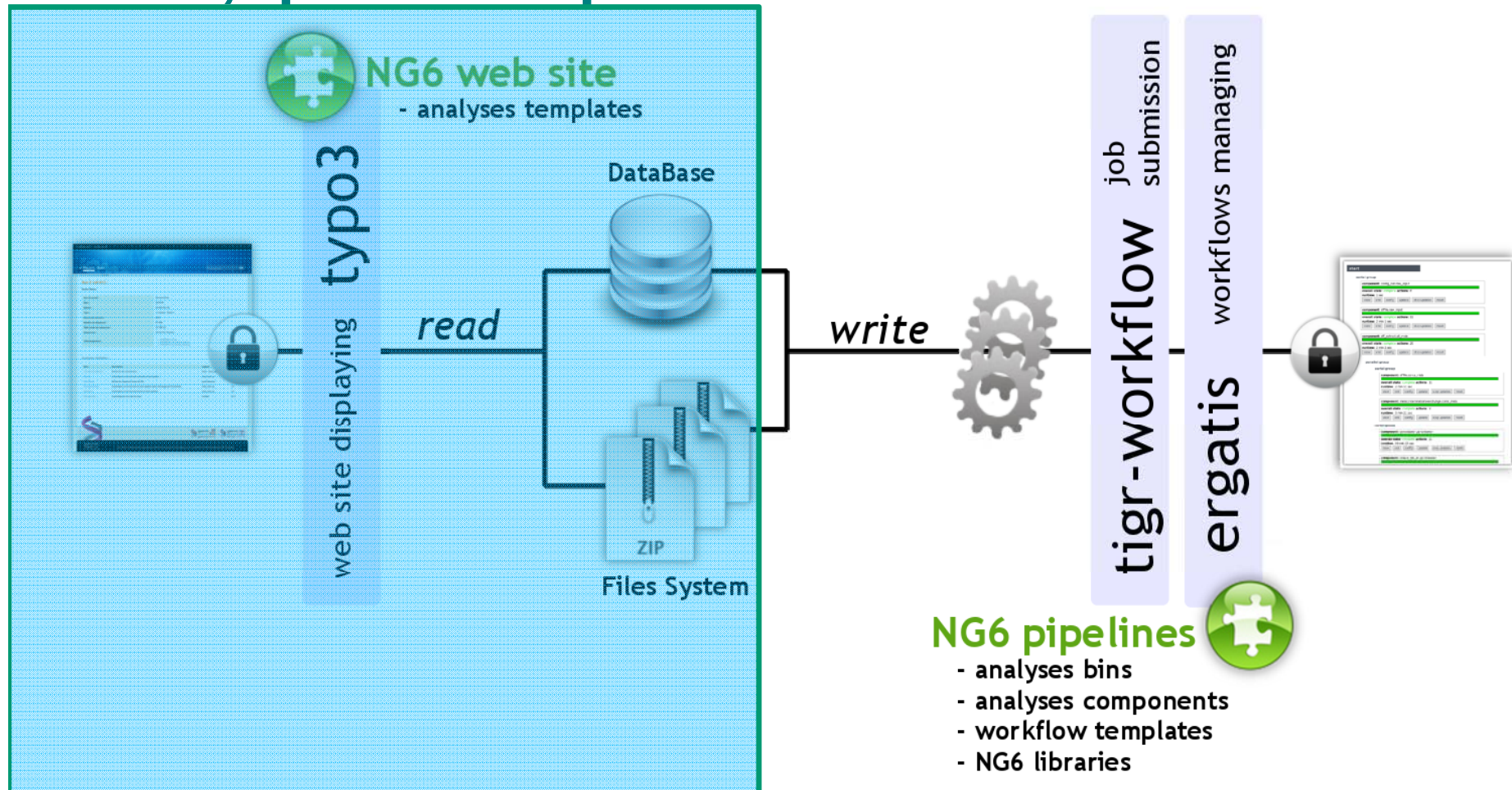
L'HiSeq2000, côté Bioinfo

nG6, qu'est-ce que c'est?



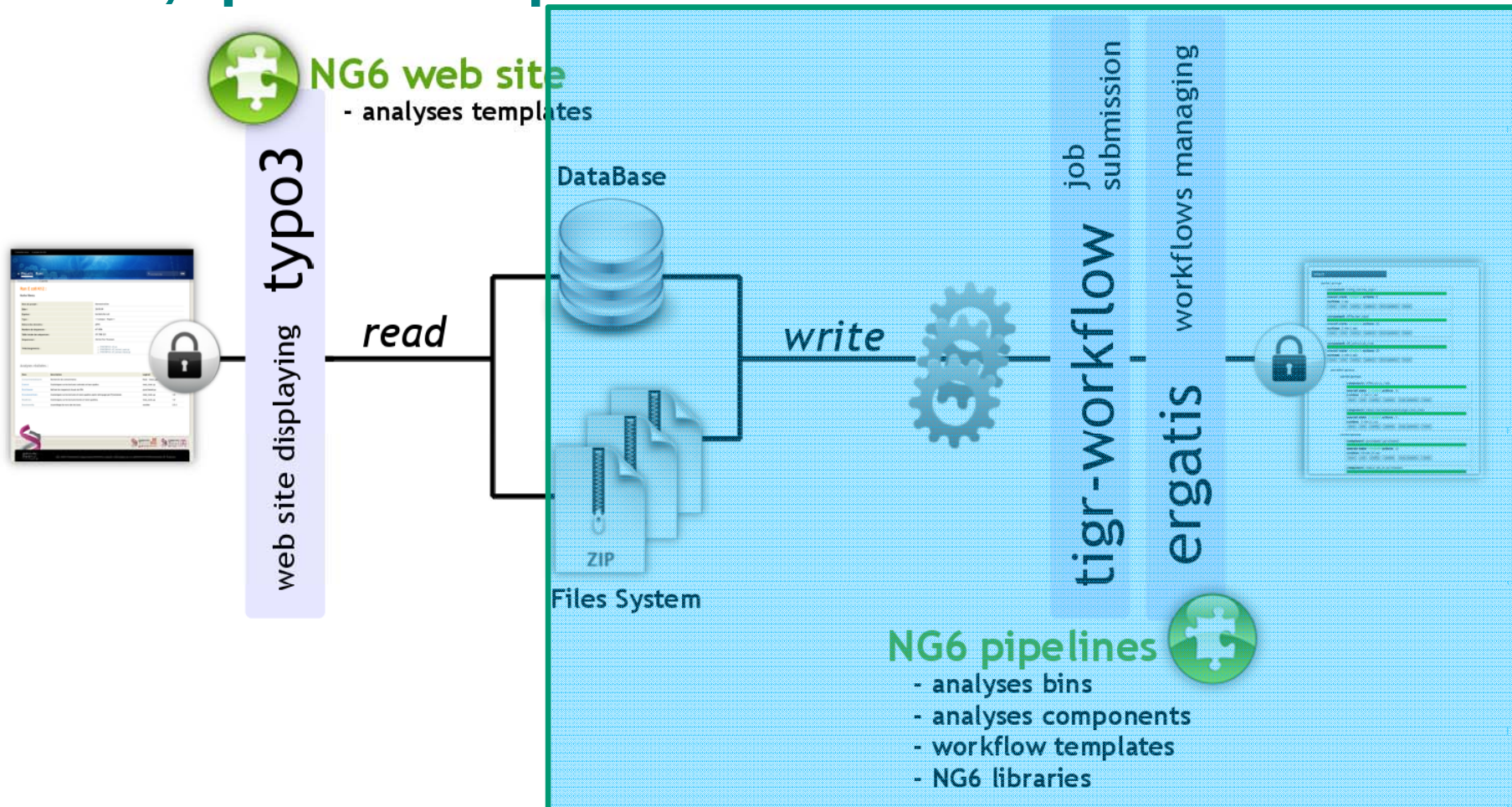
L'HiSeq2000, côté Bioinfo

nG6, qu'est-ce que c'est?



L'HiSeq2000 et nG6

nG6, qu'est-ce que c'est?



L'HiSeq2000, côté Bioinfo

Frontend, site web utilisateur

- Accès aux résultats d'analyse par projet ou par run + module de recherche



Liste des projets auxquels vous avez accès :

Vous avez accès à 80 projets.

L'ensemble des données brutes et des résultats d'analyses occupent **4.65 Tb** d'espace disque pour l'ensemble des projets.

Afficher éléments par page

Rechercher :

☐	Nom du projet ▲	Description ▼
☐	FLUSEQH6	Analyse des quasi-especes de virus influenza aviaires faiblement pathogenes (VIAFP) infectant le canard versus le poulet ou la dinde
☐	FosFine	sequencage d'inserts fosmidiques grands fragments issus de criblages de clones metagenomiques
☐	FRUITPALM	Ce projet vise l'identification, le clonage et la caracterisation de l'expression des genes controlant les differences d'activite lipase chez E. guineensis pour preparer la selection assistee par marqueurs d'un materiel a faible activite lipase.
☐	FRUITSET	Analyse transcriptomique chez la Tomate
☐	GENSURHY	Étude métabolisme de dégradation
☐	GLOMITO	Variabilite des genomes mitochondriaux de Glomus sp.
☐	COBIO	Séquençage de cDNA pour identification de séquences codantes et en particulier séquences codantes du système de détoxification.
☐	HEMOGENE	Caracterisation du locus BoLA class I en relation avec la survenue de la pancytopenie neonatale bovine (PNB)
☐	HISEQOC	Recherche de polymorphisme chez le cheval
☐	IMMUNREP	Séquençage d'un pool de produits PCR pour étude du répertoire immunitaire.

Pour la selection :

Affiche les éléments de 31 à 40 sur un total de 80 enregistrements

<<
 <

>
 >>

Projet CARTOSEQ :

Resequencage de taureaux

35 run(s) et 0 analyse(s) ont été réalisées sur le projet CARTOSEQ.

L'ensemble des données brutes et des résultats d'analyses occupent **587.36 Gb** d'espace disque pour l'ensemble du projet.

Afficher 10

☐ Nom

☐ FLUSI

☐ FosFI

☐ FRUIT

☐ FRUIT

☐ GENS

☐ GLOM

☐ COBI

☐ HEMC

☐ HISEC

☐ IMML

Pour la selecti

Affiche les élér

Runs réalisés :

Afficher 10 éléments par page

Rechercher :

☐	Nom du run	Nom du projet	Date	Espèce	Nature des données	Type	Nombre de séquences	Taille totale des séquences	Description	Séquenceur
☐	FR2290038601-FATAL	CARTOSEQ	12-09-11	Bovin	ADNg	1/8 FlowCell B - lane 8	184 076 390	18 591 715 390	Insert : 406 pour ech FR2290038601-FATAL. [C] 8 pM	HiSeq 2000
☐	FR2296001756-MERDRIGNAC	CARTOSEQ	12-09-11	Bovin	ADNg	1/8 FlowCell B - lane 7	179 068 682	18 085 936 882	Insert : 360 pour ech FR2296001756-MERDRIGNAC. [C] 7 pM	HiSeq 2000
☐	FR4486041658-BESNEBUCK	CARTOSEQ	23-08-11	Bovin	ADNg	1/8 Flowcell A - lane 1	159 262 244	16 085 486 644	Insert : 313 pour ech FR4486041658-BESNEBUCK. [C] 7 pM	HiSeq 2000
☐	FR4486041658-BESNEBUCK	CARTOSEQ	23-08-11	Bovin	ADNg	1/8 Flowcell A - lane 2	159 674 806	16 127 155 406	Insert : 313 pour ech FR4486041658-BESNEBUCK. [C] 7 pM	HiSeq 2000
☐	FR4486041658-BESNEBUCK	CARTOSEQ	23-08-11	Bovin	ADNg	1/8 Flowcell A - lane 3	163 355 506	16 498 906 106	Insert : 313 pour ech FR4486041658-BESNEBUCK. [C] 7	HiSeq 2000



Run FR2290038601-FATAL :

Insert : 406 pour ech FR2290038601-FATAL. [C] 8 pM

3 analyses ont été réalisées sur le run FR2290038601-FATAL.

L'ensemble des données brutes et des résultats d'analyses occupent **14.23 Gb** d'espace disque.

Nom du projet	Date	Espèce	Type	Nature des données	Nombre de séquences	Taille totale des séquences	Séquenceur
CARTOSEQ	12-09-11	Bovin	1/8 FlowCell B - lane 8	ADNg	184 076 390	18 591 715 390	HiSeq 2000

Téléchargements :

[s_8_1_qseq_UMD3_1_sorted.bam](#)

Analyses réalisées :

Afficher 10 éléments par page

Rechercher :

☐	Nom	Description	Logiciel	Version
☐	ContaminationSearch	Recherche de contaminants.	bwa	0.5.9-r16
☐	ReadsStats	Statistiques sur les lectures et leurs qualites.	fastqc	v0.9.1
☐	AlignmentStats	Statistiques alignement contre genome de reference.	samtools flagstat	0.1.13

Pour la selection :

cacher

décacher

supprimer

Affiche les éléments de 1 à 3 sur un total de 3 enregistrements

<< < 1 > >>

L'HiSeq2000, côté Bioinfo

Frontend, site web utilisateur

- Analyses proposées en routine pour des données 454 :
 - Statistiques de base (longueur, qualité, N...)
 - Recherche de contaminants (blastall 2.2.25)
 - Nettoyage sur les biais connus du 454 (Pyrocleaner 1.2)
 - Assemblage (newbler 2.6)
- Analyses proposées en routine pour des données HiSeq2000:
 - Statistiques de base (FastQC 0.9.1)
 - Recherche de contaminants (bwa 0.5.9)
 - Alignement contre génome de ref (bwa 0.5.9)

L'HiSeq2000, côté Bioinfo

Frontend, site web utilisateur

- Possibilité de télécharger les données « brutes» (sff, bam, fastq.gz) après archivage ou de créer un lien symbolique sur serveur Bioinfo pour analyse sur le cluster

L'HiSeq2000, côté Bioinfo

nG6, q

★ PROJETS RUNS **DOWNLOAD** recherche OK



Téléchargement de vos données

Sélectionner les données que vous souhaitez télécharger, sélectionner la façon dont vous voulez récupérer vos données sur la droite, puis cliquer sur le bouton de téléchargement.

Liste des fichiers à télécharger :

- Run FR2290038601-FATAL (12-09-11) :
Données brutes
- Run FR2296001756-MERDRIGNAC (12-09-11) :
Données brutes

Format :

☐ .tar.gz
☒ liens sur snp.toulouse.inra.fr

 **Télécharger**

Création de liens symboliques

Cette fonctionnalité est uniquement valable si vous disposez d'un compte utilisateur sur le serveur snp.toulouse.inra.fr.

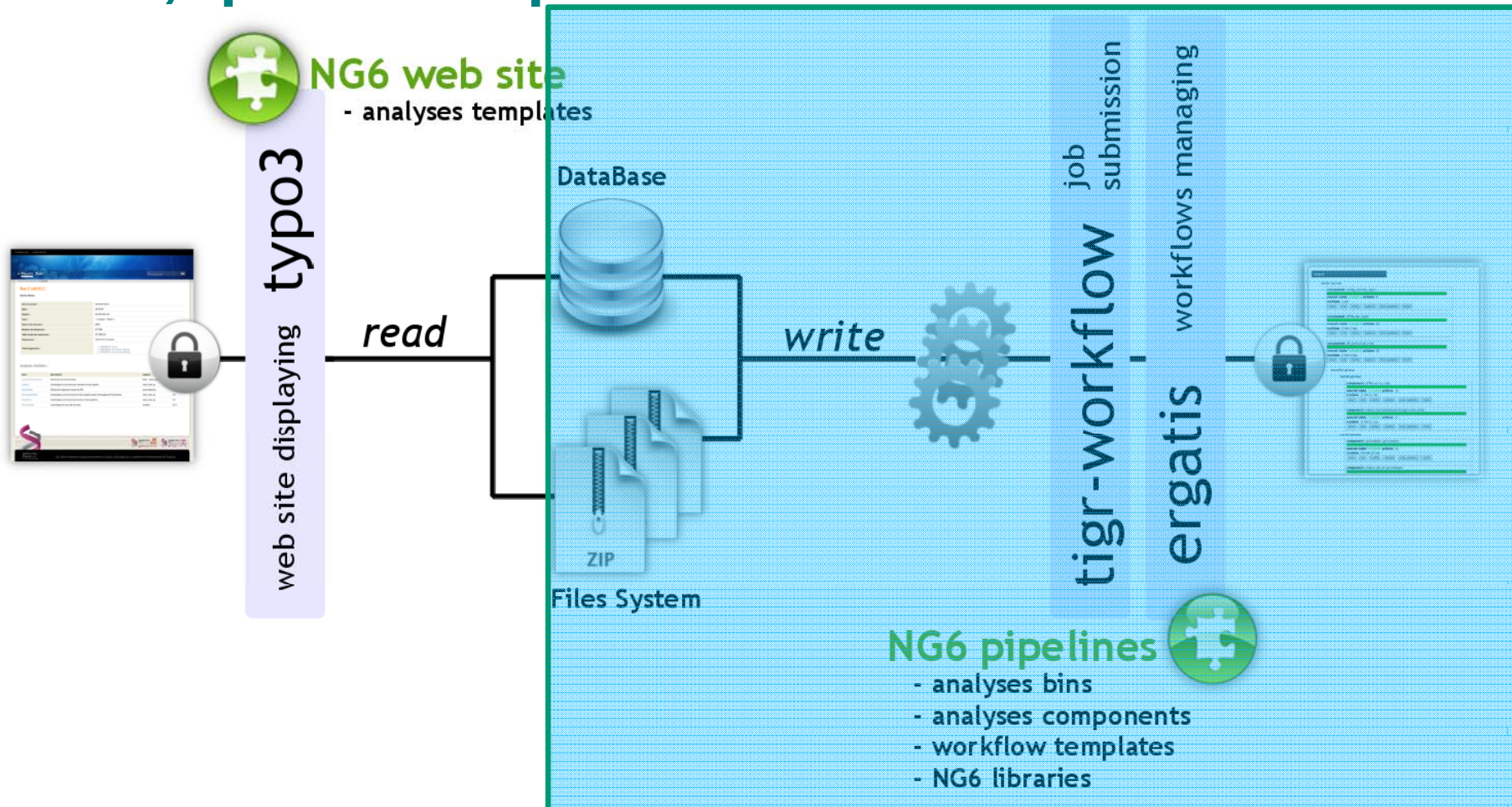
Nom d'utilisateur :
 Mot de passe :
 Répertoire :

Créer **Annuler**

- ☐ Projet ACTIV
- ☐ Projet ALEXA
- ☐ Projet ALOM
- ☐ Projet APACH
- ☐ Projet ARACE
- ☐ Projet ARCAD
- ☐ Projet ASEDM
- ☐ Projet BALAN
- ☐ Projet BIOFO
- ☐ Projet BIOME
- ☐ Projet BOVRN
- ☐ Projet BRASE
- ☐ Projet CAPRISNP
- ☐ Projet CAPTSEQ
- ☒ **Projet CARTOSEQ**
 - ☒ Run FR2290038601-FATAL (Bovin) - réalisé le 12-09-11, 184076390 lectures
 - ☒ Données brutes
 - ☐ Analyse AlignmentStats
 - ☐ Analyse ContaminationSearch
 - ☐ Analyse ReadsStats
 - ☒ Run FR2296001756-MERDRIGNAC (Bovin) - réalisé le 12-09-11, 179068682 lectures
 - ☒ Données brutes
 - ☐ Analyse AlignmentStats
 - ☐ Analyse ContaminationSearch
 - ☐ Analyse ReadsStats
 - ☐ Run FR4486041658-BESNEBUCK (Bovin) - réalisé le 23-08-11, 159262244 lectures
 - ☐ Run FR4486041658-BESNEBUCK (Bovin) - réalisé le 23-08-11, 159674806 lectures
 - ☐ Run FR4486041658-BESNEBUCK (Bovin) - réalisé le 23-08-11, 163355506 lectures

L'HiSeq2000 et nG6

nG6, qu'est-ce que c'est?



L'HiSeq2000, côté Bioinfo

Backend, Ergatis

Pipelines préconfigurés suivant

- le séquenceur
 - 454
 - HiSeq2000, GA
- le type d'échantillon
 - ADNg / ARN
 - Multiplexé ou non
 - Métagénomique
- Le format d'entrée des données
 - sff
 - fasta
 - fastq.gz
 - qseq

component: **config_run_casava (raw_input)**

configuration

parallel group

component: **ng6runstat (all_mids)**

configuration

component: **fastqc (all_mids)**

configuration

serial group

component: **bwa_index (ref)**

configuration

component: **bwa_aln_sorted_bam (default)**

configuration

component: **samtools_flagstat (default)**

configuration

add component add pipeline add serial group add parallel group

component: **bwa_contamination_search (default)**

configuration

add component add pipeline add serial group add parallel group

parallel group

component: **archive2ng6 (ref)**

configuration

add component add pipeline add serial group add parallel group

component: **ng6synchronization (default)**

configuration

add component add pipeline add serial group add parallel group

add component add pipeline add serial group add parallel group

2000, côté Bioinfo

GeT

http://ng6.toulouse.inra.fr/ergatis/cgi/build_pipeline.cgi?

component: **config_run_casava (raw_input)**

configuration

parallel group

component: **ng6runstat (all_mids)**

configuration

component: **fastqc (all_mids)**

configuration

serial group

component: **bwa_index (ref)**

configuration

component: **bwaalnsortedbam (default)**

configuration

component: **samtools_flagstat (default)**

configuration

add component add pipeline add serial group

component: **bwacontaminationsearch (default)**

configuration

add component add pipeline add serial group

parallel group

component: **archive2ng6 (ref)**

configuration

add component add pipeline add serial group

component: **ng6synchronization (default)**

configuration

add component add pipeline add serial group

component: **config_run_casava (raw_input)**

configuration

parameters

run name

The run name.

date

The run date in dd/mm/yyyy format.

species

The species used for this run.

data nature

The nature of the data.

run type

The type of run done.

run description

Run description.

sequencer

The sequencer name used.

lane number

The number of the lane (to grab the right qseq files)

project name

The project name the run belongs to. If the project already exists, the description and front end group is not required.

project description

The project description the run belongs to.

project fe group

The project fe group name allowed for this project. If it doesn't exist it will be created.

project fe group description

The project fe group description.

project fe group parent name

The project fe group parent name.

input

The raw directory or the demultiplexed directory if it's a multiplexed run

casava directory

info

L'HiSeq2000, les Hommes

nG6, qui sont les petites mains?

PF Bioinfo :

- C. Klopp : Responsable technique
- J. Mariette : l'expert, développeur principal
- D. Laborie : administrateur système

PF GeT-* :

- J. Lluch, E. Robe, S. Ruzafa, E. Lhuillier, N. Marsaud, D. Esquerre : équipes 454 et HiSeq, exécutent les analyses qualité sur ergatis et rédigent les comptes-rendus
- O. Bouchez : responsable NGS GeT-PlaGe, valide les analyses qualité
- F. Escudié et moi : gestion des données, suivi des analyses qualité et développement, interface PF Bioinfo

L'HiSeq2000, les Hommes

nG6, qui sont les petites mains?

Groupes de travail sur différentes thématiques visant à créer des pipelines à intégrer à nG6 (RNASeq, variations, miRNA, ChIP-SEQ, méthylation...)

<https://mulcyber.toulouse.inra.fr/projects/ngspipelines> >

Onglet MediaWiki

L'HiSeq2000, futur PlaGe

Et dans l'année à venir...

PF GeT-PlaGe :

- Triplement du débit de l'HiSeq2000 dès maintenant
- Doublement du débit annoncé courant d'année prochaine
- 1 voire 2 HiSeq supplémentaires (130To de données à sauvegarder)

L'HiSeq2000, futur Bioinfo

Et dans l'année à venir...

PF Bioinfo :

- Passage en production des pipelines CASAVA1.8
- Continuer le travail d'optimisation de l'espace dans la gestion des données
- Nouveaux pipelines d'analyse + nouvelles fonctionnalités nG6
- Formations NGS (polymorphisme SNP / RNA-Seq)
- Mise en place nouvelle infrastructure en janvier 2012 (CPER)
 - Calcul + + +
 - Stockage +

Remerciements / Questions



- Toute l'équipe PF Bioinfo
- L'EIC Toulouse



- Les équipes de recherche pour leurs retours
- Toute l'équipe PlaGe...et plus particulièrement les biologistes moléculaires qui mettent la main au clavier avec le sourire...ou pas

