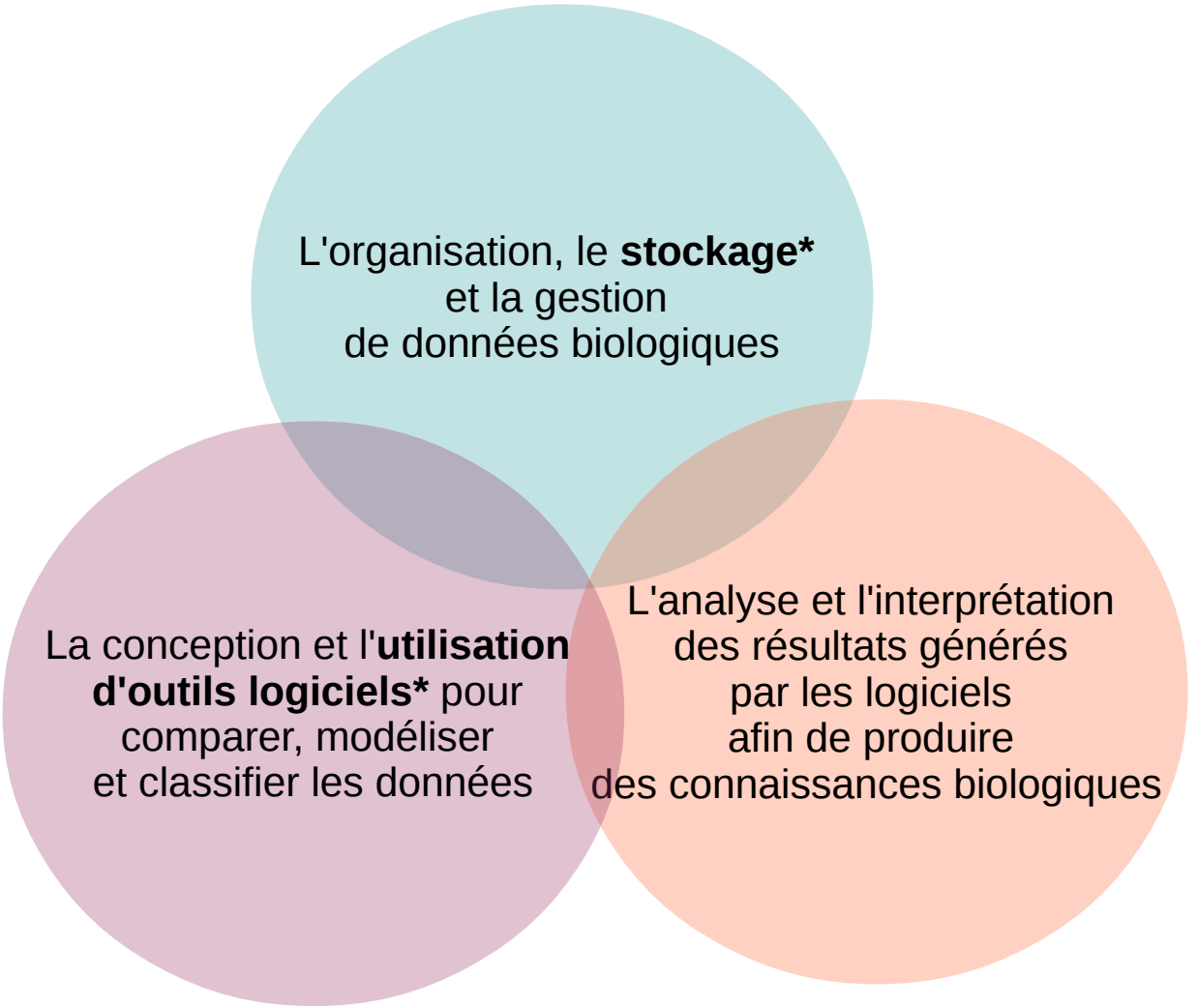
Analyses bioinformatiques pour le PCIM

Journée de rencontre des utilisateurs du Pôle de calcul intensif pour la mer
17 janvier 2014

La bioinfo, késaco ?

- Approche in silico de la biologie

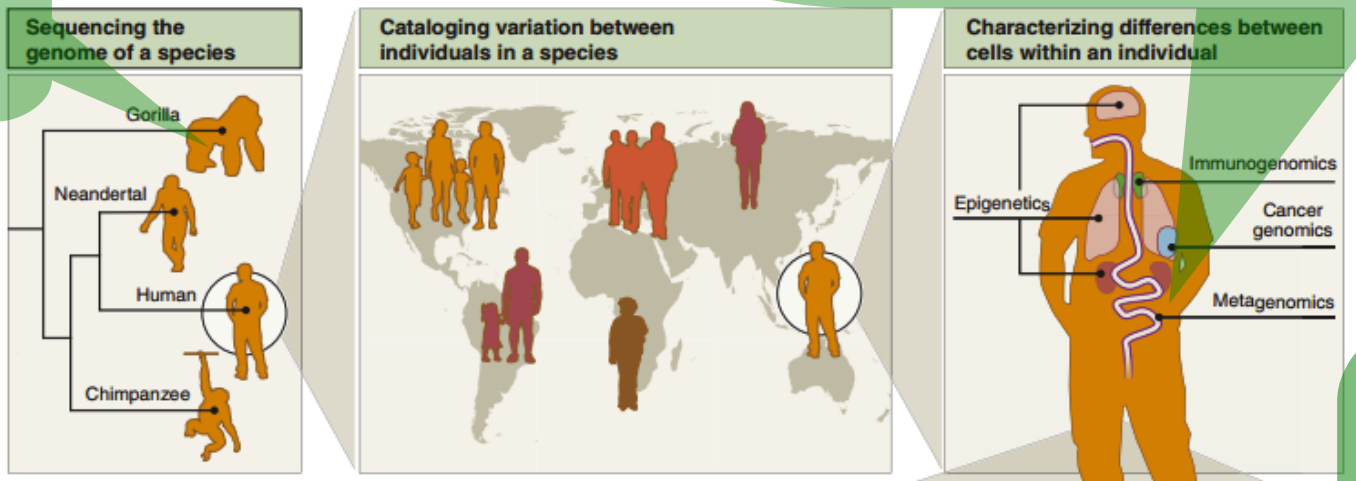


***** : là où intervient le PCIM

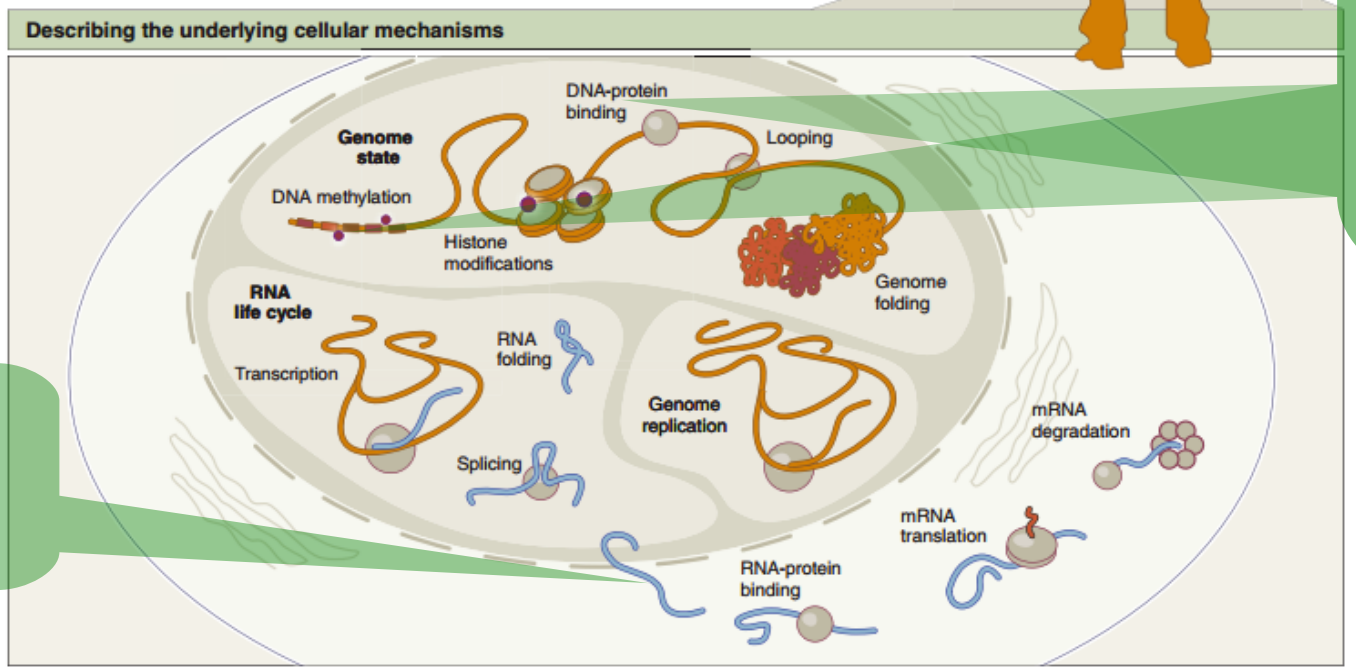
Les applications

DNA-seq (séquençage de novo nouvelles espèces)

Metagénomique (analyse génomique des micro-organismes appartenant à une communauté microbienne)

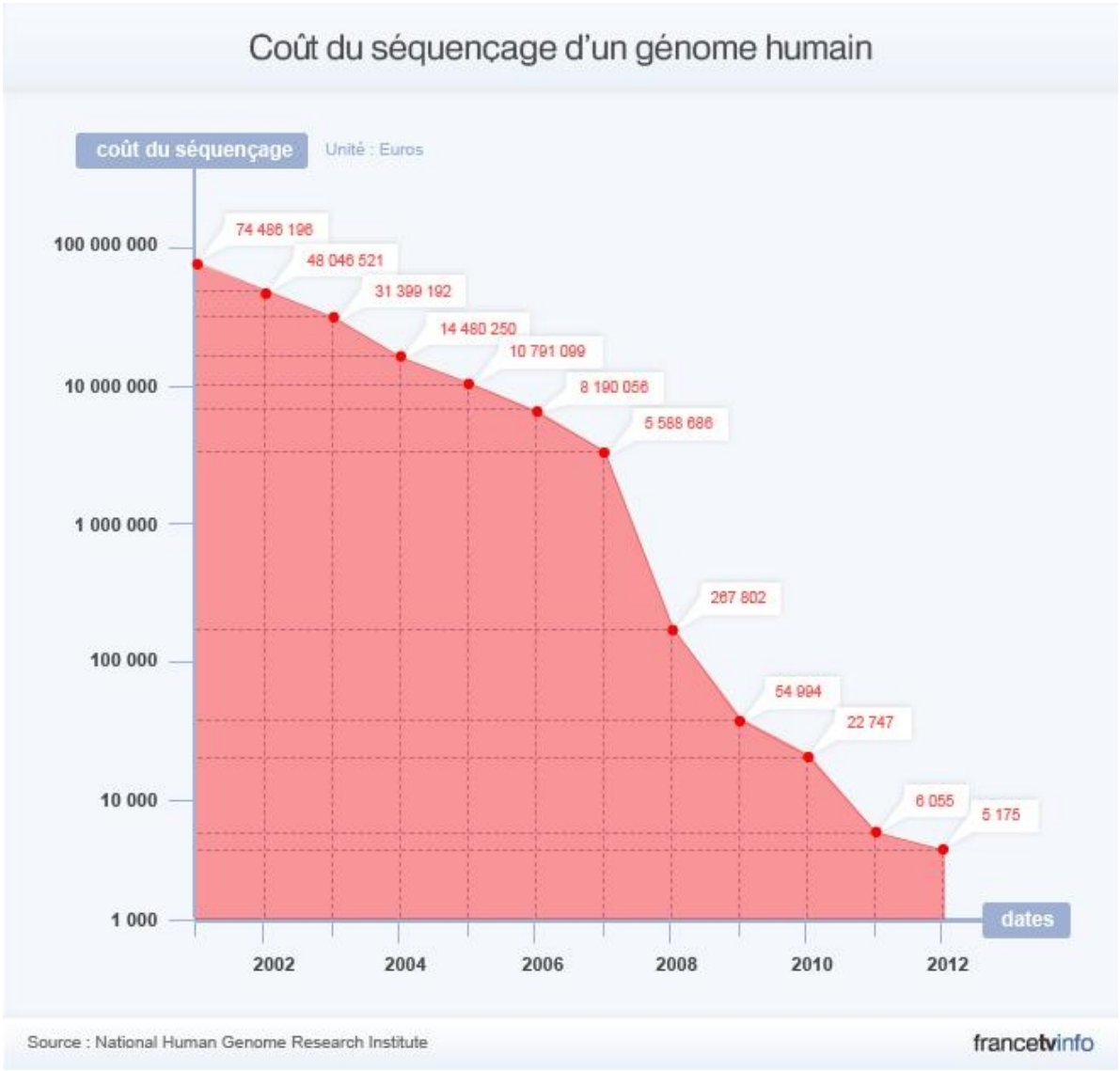


Epigenetique (étude régulation de gènes héréditaires impliquant des modifications de l'ADN) :
- *Méthylation de l'ADN*
- *Chip-Seq* (analyse de l'intéraction protéines/ADN)



RNA-seq (séquençage de transcriptome)

Les faits

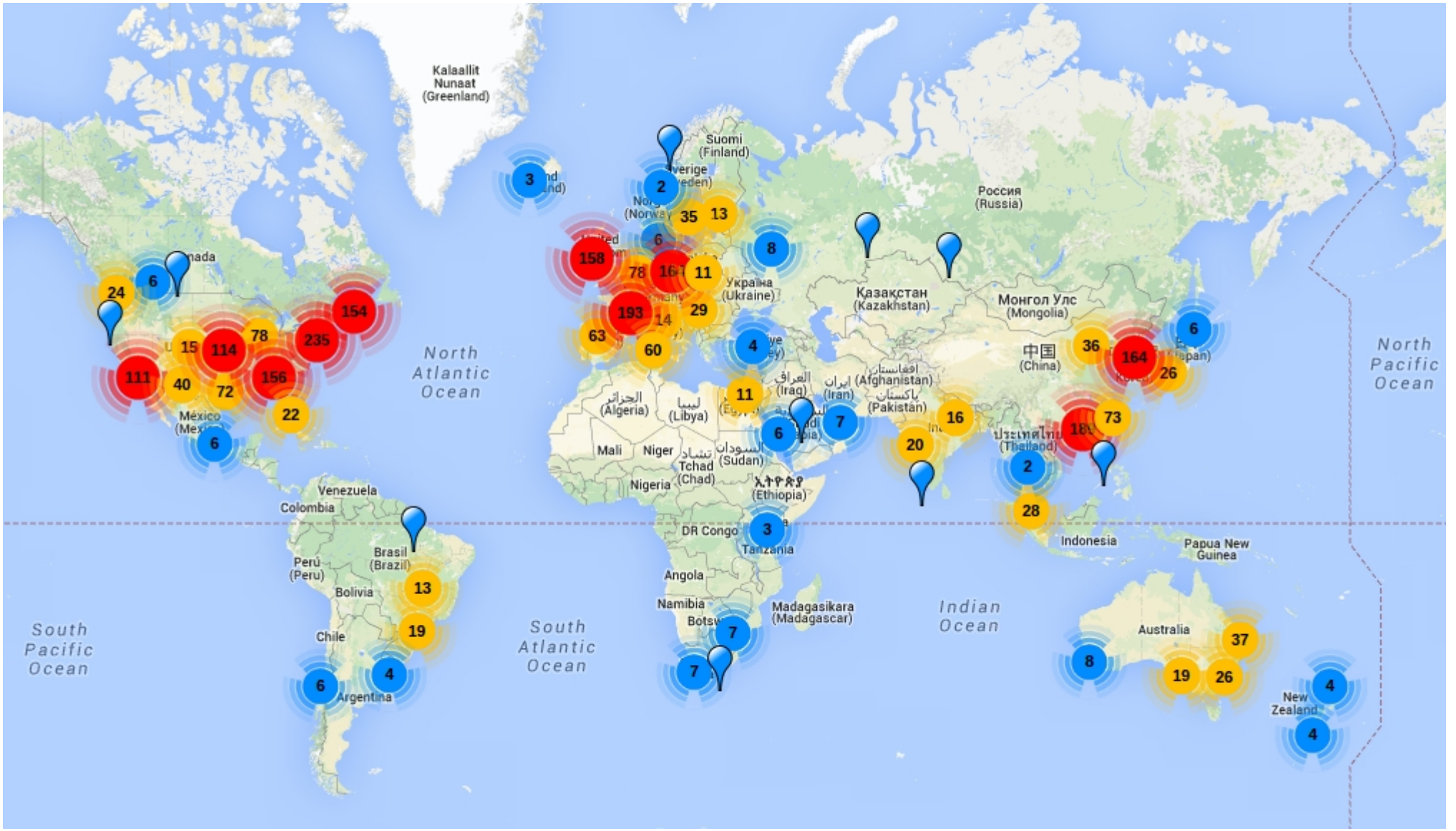


Rythme 300 fois plus rapide que celui qu'a suivi la chute des coûts dans l'informatique, décrit par la célèbre loi de Moore (une division des coûts par deux tous les deux ans à puissance égale)

Votre génome en 2015 pour la modique somme de 100$!!

Conséquence : N'importe quel laboratoire peut soumettre un projet de séquençage !!!

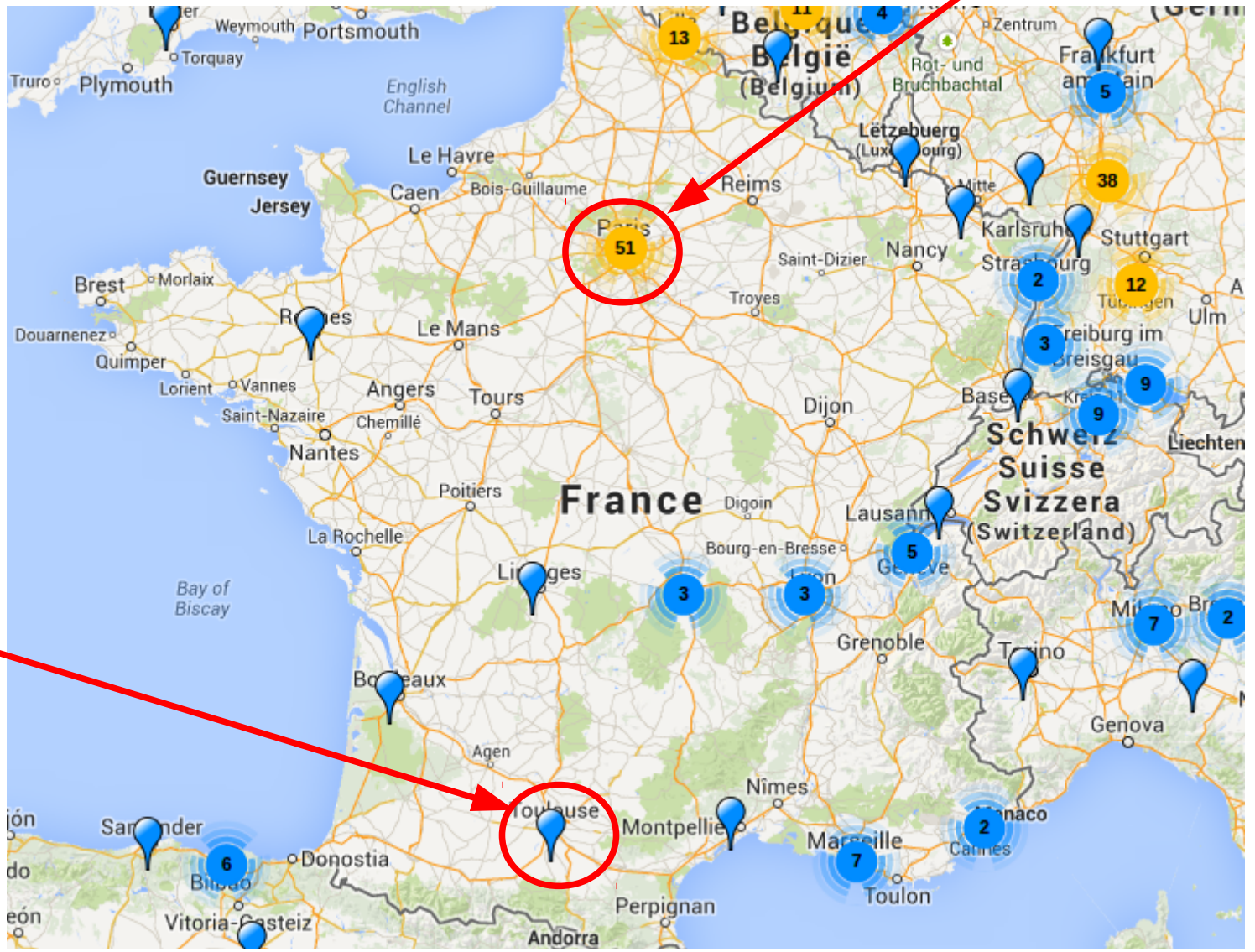
Les séquenceurs haut-débit dans le monde



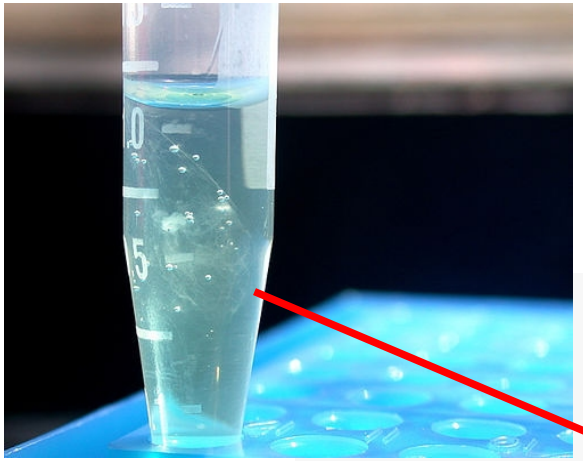
Les séquenceurs haut-débit en France

Génoscope, Evry

Plate-forme Genotoul, Toulouse



Données générées



Fichiers plats format fastq (reads, scores qualité)

Enjeux ?

- Manipulation de données haut-débit
  - Stockage :
    - Jeu brut : ~ 150Go
    - Jeu traité : facilement 1To

    } 1 seul projet !
  - Puissance de calcul :
    - Parallélisation possible de certains outils
    - Chargement en mémoire des données de référence

Prérequis

- Débit réseau (récupération, partage, soumission des données)
- Stockage (données brutes et traitées)
- Linux (décompresser et visualiser les données brutes)
- Savoir compiler/configurer des outils d'analyse bioinfo en ligne de commande
- Faire le tri parmi la montagne d'outils bioinfo permettant de faire plus ou moins là même chose...
- Maîtriser les bases d'utilisation d'une machine de calcul
- ...

Prérequis

- Débit réseau (récupération, partage, soumission des données)
- Stockage (données brutes et traitées)
- Linux (décompresser et visualiser les données brutes)
- Savoir compiler/configurer des outils d'analyse bioinfo en ligne de commande
- Savoir choisir pa bioinfo permettant de faire plus ou moi
- Maîtriser les bas chine de calcul
- ...

Solutions ?

- Tout sous-traiter (limites...)

- Faciliter l'accès aux outils complexes par l’intermédiaire d'outils moins complexes
- Formation

Solutions ?

- Tout sous-traiter (limites...)
- Faciliter l'
- Formation

ngKlast

- Développé par Korilog (Muzillac)
- Outil de recherche d'homologies de séquences basé sur les algorithmes :
  - Blast (NCBI)
  - Klast (Korilog et INRIA Rennes)
- Visualisations graphiques
- Pré-annotation des résultats

ngKlast architecture

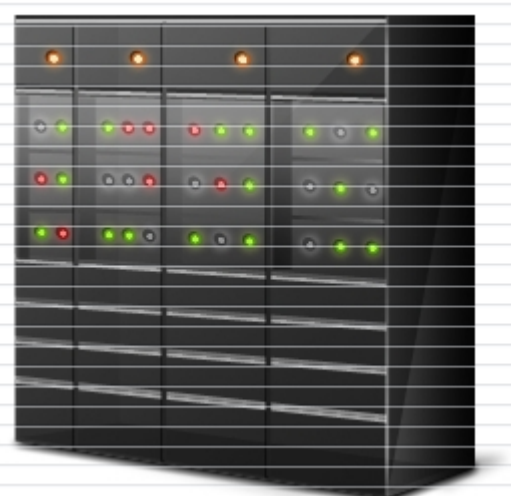
vkoriblast2
Windows serveur
2008
Application ngKlast

Gestion des jobs

Serveur web
(apache/tomcat)
Kserver

Soumission des
requêtes
Blast/Klast

Caparmor



Connexion bureau à
distance sur serveur
vkoriblast2

Montage nfs

Caparmor-nas

/home12/caparmor/bioinfo/
Kserver
Kdms (banques de
références)
Klast
Blast

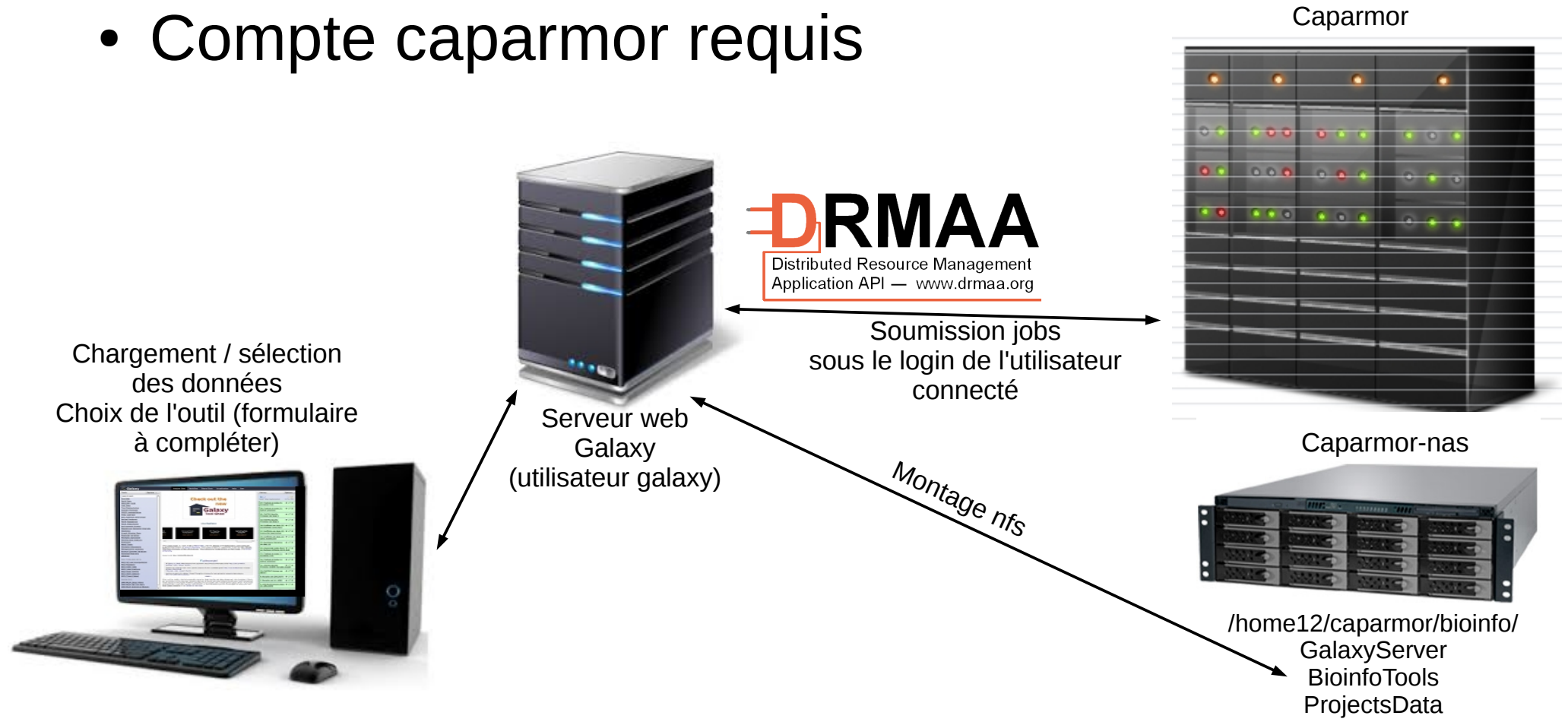
Démo ngKlast

Galaxy

- Plate-forme web développée par le centre de génomique comparative et de bioinformatique de l'université de Penn State (Pennsylvanie)
- Idée : Donner accès à une multitude d'outils de bioinformatique dans un unique support sans passer par la ligne de commande
- Solution open source, Code python
- Communauté très active (Galaxy Community Conference chaque année)
- Utilisé pour :
  - Analyse qualité des données haut-débit
  - Analyses RNA-Seq
  - Métagénomique
  - Phylogénie

Galaxy au PCIM

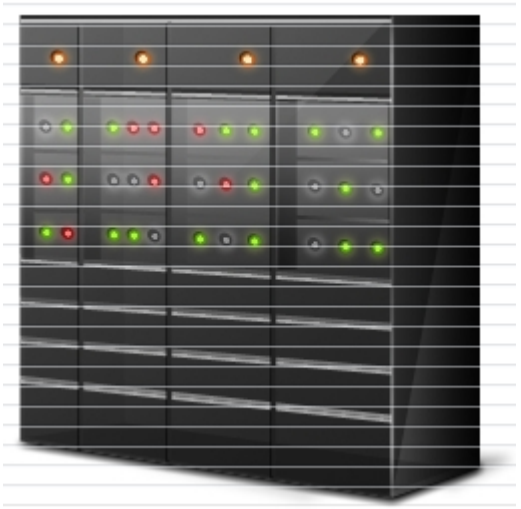
- w3.ifremer.fr/galaxy (intranet)
- Authentification CAS
- Compte caparmor requis



Galaxy au PCIM

- Ce que ça implique :

Caparmor



Installation / Compilation des outils de bioinformatique

Serveur web Galaxy

- Définition des outils

- Création des Wrappers :
  - 1 formulaire par outil : fichier XML
  - [1 script de génération de la ligne de commande par outil]

Démo Galaxy

Critiques de l'outil

- Utilisateurs :
  - Accès aux outils facilité
  - Rapidité d'éxécution des traitements
  - Possibilité de créer des worflows pour ré-exécuter des tâches récurrentes
  - Etre rigoureux dans le renommage des fichiers
  - Certains outils/worflows restent difficiles à prendre en main
  - Visualisation limitée pour certain types de données

- Administrateurs :
  - Rien de plus simple pour permettre l'accès à des outils unix
  - Limites dans la configuration des outils (trop d'options peuvent perdre les utilisateurs)
  - Stockage des données (dépendant de l'utilisateur, peut devenir rapidement gourmand)
  - Système complexe (nombreuses dépendances)
  - Plate-forme largement utilisée
  - Le bio-informaticien préférera toujours la ligne de commande

Perspectives autour de Galaxy

- Projets 2014 :
  - Ouvrir une instance internet :
  - Agrémenter la boîte à outils pour la phylogénie
  - Simplifier la prise en main de la suite QIIME 1.7 et agrémenter la boîte à outils métagénomiques (nombreux projets sur ce thème à venir dans les labos Ifremer)
  - Améliorer l'accès aux données générées pour une meilleure gestion par les utilisateurs (quotas – utilisation du caparmor home)

- Présentation à gen2bio :
  - Saint Malo (3 avril 2014)
    - Galaxy pour l'étude de données de métagénomique (Microflore des bacs d'élevage des huîtres C. Gigas, J.L. Nicolas PFOM-PI)

- Collaboration avec la plate-forme Abims (Station biologique de Roscoff)
  - Sept 2014 : « Marine Day »
    - Pertinence de la création d'un site chapeau, donnant accès aux plate-formes thématiques galaxy Brest/Roscoff + goodies (e-learning, bases de données espèces marines...) ?

- Plate-forme pouvant servir à n'importe quelle thématique

La cellule bioinfo (IDM/RIC)

Fanny Marquer
(Ingénieur Bioinfo)

Laure Quintric
(Ingénieur Bioinfo)

Pierrick Lucas
(Contrat Pro alternance Master 2 bioinfo
université de Nantes) : Sept 2013 à Sept 2015

- Pour contacter la cellule bioinfo : bioinfo@ifremer.fr