

Partage des données relatives aux ressources génétiques et génomiques

Etat des lieux, analyse stratégique et besoins d'accompagnement

Participants au groupe de travail (affiliation INRA si non précisé) :

P. Bessi res, S. Carr re, S. Casaregola, G. Lassalle, F. Legeai, N. Mohellibi, F. Moreews, J-L. Noyer (CIRAD), H. Quesneville, L. Ranjard, S. Sidibe-Bocs (CIRAD), M. Tixier-Boichard (r daction)

Sommaire

L'�cosyst�me informationnel	2
Typologie des donn�es	2
Entrep�ts de donn�es	2
Propri�t�s des donn�es et conditions d'acc�s	3
Patrimoine � p�renniser et cycle de vie de la donn�e	4
Vision strat�gique : quelles donn�es partager avec qui et pourquoi	5
L'INRA dispose-t-il de donn�es originales	6
Projets pilotes dans le domaine	6
Attentes en termes d'offre INRA	8
Tableau 2 : liste des entrep�ts de donn�es	9

R sum 

On peut identifier 19 types de donn es (6 types de donn es brutes, 13 types de donn es  labor es). Les donn es g nomiques sont de plus en plus abondantes et disposent le plus souvent de standards d' change et de m ta-donn es. Elles sont stock es dans des entrep ts internationaux, mais le mod le centralis  de stockage semble atteindre sa limite et un syst me de stockage distribu  r pondra mieux aux besoins actuels et futurs des chercheurs. Les ressources humaines constituent le point sensible pour le bon fonctionnement et la p rennisation de ces entrep ts.

L'INRA devrait se positionner sur l'h bergement de certains entrep ts dans un syst me distribu  pour renforcer sa notori t  et s curiser les donn es patrimoniales qu'il produit. Cette s curisation est   r fl chir en relation avec l'infrastructure nationale de bioinformatique, IFB.

Une grande partie des donn es est obtenue en partenariat avec d'autres organismes de recherche ou avec des acteurs priv s. En r gle g n rale, les donn es ne sont pas partag es tant qu'elles n'ont pas fait l'objet d'une publication. La tra abilit  des producteurs de donn es est importante pour favoriser le partage. Une proposition de cycle de vie de la donn e est faite sur la base de l'abondance et de l'innovation technologique.

Il y a un int r t strat gique   partager les donn es de r f rence, mais il est important d' tre leader dans le consortium travaillant sur l'esp ce  tudi e, cela suppose un investissement massif mais procure plusieurs avantages : exclusivit  temporaire, vue int grative, capitalisation de toutes les connaissances.

Plusieurs projets pilotes sont pr sent s pour illustrer le partage et la gestion des donn es g nomiques dans le domaine animal, v g tal et microbien. Les besoins semblent concerner plus les outils informatiques (environnement collaboratif, plateforme de logiciels et d'acc s aux donn es) que les outils juridiques, la r alisation d'accords de consortium  tant une pratique courante. Il faut traiter le cas de 'l'abandon' de donn es.

a) L'écosystème informationnel

1- Typologie des données

On peut identifier 19 types de données selon l'élément biologique décrit (tableau 1).

On propose de distinguer la donnée brute obtenue directement en sortie d'un équipement de mesure, de la donnée élaborée issue d'une analyse de la donnée brute. La frontière entre les deux types de données n'est pas toujours facile à tracer. Lors de l'acquisition des données brutes, il existe déjà une analyse réalisée sur la plateforme transformant le signal lu (fluorescence, migration) en une donnée génomique, mais cette analyse n'implique généralement pas l'utilisateur final de la donnée, pour qui la donnée fournie par la plateforme est un point de départ. Sur les 19 types identifiés, 6 seulement sont des données brutes.

Tableau 1 : types de données et standards correspondants

obtention	Nature	Format standard pour l'échange
brute	séquences lues ADN-ARN	fastq, sff
brute	génotypes SNP	Illumina (matrice marker*organisme) + métadonnées en en-tête, VCF
brute	génotypes SSR	Csv – excel (GeneMapper)
brute	données d'expression (arrays, qPCR)	MIAME
brute	données métabolome	EC number, SBML
brute	profils protéiques (quantitatifs)	?
élaborée	séquences protéines	fasta, asn, embl
élaborée	séquences alignées/assemblées	fasta, bam, sam
élaborée	données d'expression RNAseq	bam, sam, sff, bed, wig
élaborée	polymorphismes SNP	VCF, fasta, flatfile
élaborée	polymorphismes SSR	gff3
élaborée	variants structuraux	VCF (V4.1+)
élaborée	patrons de méthylation	bed ?
élaborée	annotations des gènes	gff3, asn, embl
élaborée	orthologues, paralogues, familles de gènes	Tables xml
élaborée	cartes (génétiques, QTLs, physiques)	acp, text formatted
élaborée	données passeport populations/souches	Voir bases FAO
élaborée	données passeport croisements temporaires	?
élaborée	données passeport banques génomiques	?

Il existe le plus souvent (16 types sur 19) un format standard pour échanger les données, au moins au sein des domaines animaux, végétaux ou microbiens.

Il ne semble pas qu'il y ait de format standard pour les données de populations mais la FAO en propose probablement.

Les données brutes sont généralement plus volumineuses que les données élaborées.

2- Entrepôts de données

Il existe déjà des entrepôts de données bien identifiés sur le plan national et international (tableau 2, en fin de document). Les entrepôts de données génomiques ne sont généralement pas spécialisés sur un type de données et donne un accès assez complet aux données de génomique structurale et fonctionnelle. Les données expressionnelles disposent d'un entrepôt spécifique (GEO) de même que les données sur les protéines et les données sur le métabolome. Toutefois, les bases de données

sur les plantes tropicales rassemblent souvent plusieurs types de données (ADN, ARN, protéine, métabolome).

Le volume des données peut devenir un facteur limitant leur échange et limitant leur stockage. Ainsi, on constate de plus en plus un encombrement des BD internationales, avec un temps d'attente important au dépôt, ce n'est pas un frein au partage mais une source de lourdeurs, les entrepôts sont victimes de leur succès. L'INRA dispose de quelques entrepôts spécifiques (tableau 2).

Si le stockage coûte plus cher que la production de la donnée, on peut évoluer vers une situation où la donnée est 'ré-obtenue' en fonction des besoins (cas du séquençage). Pour cela, il serait souhaitable de connecter les conservatoires de ressources génétiques et génomiques aux entrepôts de données (exemple : GenoSol pour les ressources métagénomiques des microbes des sols).

Il faut distinguer le stockage du traitement de la donnée, il s'agit de métiers différents. La donnée brute doit être proche du calculateur. L'accès aux données est facilité par la standardisation des métadonnées : l'EBI propose le kit ISA pour la gestion des métadonnées, ISA est intégré à la plateforme Galaxy, toutefois la convivialité de ISA est discutée. Ainsi, le LIPM gère directement le format de ses données pour les soumettre dans les bases publiques. Dans le cas de *Bacillus subtilis*, l'entrepôt des données internationales est centralisé en Suisse (ETH) avec un contrôle du format d'entrée et l'utilisation de fichiers excel avec vocabulaire contrôlé.

Les entrepôts centralisés étant de plus en plus encombrés et peu réactifs, l'avenir réside certainement dans un système distribué où chacun stocke ses données. C'est la solution retenue par le projet européen TransPlant coordonné par l'EBI, les données sont réparties sur plusieurs nœuds, si les template et les ontologies sont identiques, les outils développés par chaque nœud sont interopérables. L'ensemble des nœuds est accessible via un seul annuaire ou un méta-annuaire si plusieurs annuaires sont liés à des partenaires.

L'INRA devrait se positionner sur l'hébergement de certains entrepôts dans un tel système : un institut responsable d'un nœud gère les métadonnées et données, développe des pipelines d'analyse qui deviennent eux-mêmes des données. L'INRA peut délivrer des DOI pour tracer les jeux de données qu'il gère. L'hébergement d'un entrepôt renforce la notoriété d'un institut et permet de sécuriser les données patrimoniales produites par l'institut. Cette sécurisation est à réfléchir en relation avec l'infrastructure nationale de bioinformatique, IFB.

La pérennisation et l'actualisation du contenu d'un entrepôt de données dépend du modèle économique retenu. Dans le cas de l'EBI, son financement bénéficie d'une cotisation des états européens à l'infrastructure que représente l'EMBL. Pour la France, ce financement représente environ 2,5 M€ par an. C'est un exemple de financement mutualisé en amont puisque c'est le MESR qui cotise directement.

Dans le cas de la base de données sur la plante modèle *Arabidopsis* (<http://www.arabidopsis.org>), le modèle de financement va changer pour assurer la pérennité de cette ressource stratégique. Dès Avril 2014, les chercheurs académiques devront payer pour accéder à ces informations. Le montant demandé pour cet accès à un institut comme l'INRA est de 7500 \$/an ce qui semble très raisonnable pour pérenniser de façon institutionnelle l'accès à ces données, sachant que l'INRA est le 7ème utilisateur mondial (17 centres impliqués et + de 22 000 visites/an). La stratégie retenue est de placer cette cotisation au niveau des départements concernés qui doivent se concerter.

D'une manière générale, les ressources humaines constituent le point sensible pour la pérennisation.

3- Propriété des données et conditions d'accès

Une grande partie des données est obtenue en partenariat avec d'autres organismes de recherche ou avec des acteurs privés.

Aucun frein au partage des données n'est signalé dans le domaine de la génomique microbienne ni de la métagenomique microbienne du sol et de l'environnement.

En règle générale, les données ne sont pas partagées tant qu'elles n'ont pas fait l'objet d'une publication. Selon les règles internationales, les organismes publics producteurs de données tel le

Genoscope doivent déposer les séquences brutes 6 mois après l'obtention. Toutefois, aucune analyse globale ne peut être effectuée avant publication. De nombreuses revues exigent la mise à disposition des données pour accepter une publication. La revue Gigasciences du BGI propose de publier avec doi et uri (universal resource identifier) et met les données à disposition sur un serveur Galaxy afin que les relecteurs puissent réexécuter l'analyse et vérifier la validité de la publication.

Le changement de statut des données de 'privé' à 'public' nécessite de pérenniser le stockage et entraîne un travail lourd de diffusion/publication qui a été rarement prévu au début d'un projet. Cela peut constituer un frein technique au partage des données.

Lorsqu'il subsiste des freins au partage, ils s'expliquent par :

- *concurrence entre l'INRA et un consortium impliquant des instituts étrangers,

- *concurrence entre équipes INRA travaillant sur la même espèce

- *habitudes de certains chercheurs qui ont du mal à diffuser leurs données

- *applications économiques par des entreprises concurrentes de nos partenaires (en sélection génomique notamment), même une fois la publication réalisée, l'accès aux données brutes, comme aux données élaborées (fichier vcf), peut fournir une information stratégique à un concurrent.

Il peut être intéressant de rendre la donnée anonyme dans le cadre d'élaboration de référentiels. Dans le cas de MicroSol sur la diversité microbienne de sols, la donnée peut être anonymisée et faire l'objet d'une convention d'échange entre partenaires publics ou entre partenaire public et partenaire privé, pour pouvoir l'utiliser en modélisation (ou autre approche) afin de construire et d'alimenter la base de données de ces référentiels.

b) **patrimoine à pérenniser** (cf approche patrimoine numérique du CIRAD)

La réflexion sur les données de génomique à l'INRA ne s'est pas orientée jusqu'à maintenant vers une approche patrimoniale. Il s'agit d'évaluer l'intérêt d'une donnée ancienne. On peut au moins envisager 2 pistes de réflexion :

- la donnée est relative à une ressource biologique qui est encore disponible, peut encore être étudiée, caractérisée avec des techniques plus récentes ; la donnée nouvelle pourrait alors remplacer la donnée ancienne ou donner des clés pour ré-analyser la donnée ancienne ;
- si la ressource biologique n'existe plus ou a fortement évolué (ex des échantillons de sols), la donnée prend alors une valeur patrimoniale.

D'une manière générale, la pérennisation d'un patrimoine suppose de pérenniser des moyens, notamment en personnel, pour actualiser les données ou au moins les outils donnant accès à des données anciennes. Les données plus ou moins compressées sont plus ou moins longues à récupérer, cela impose 2 critères à remplir pour le data center stockant ces données : prévoir une fonctionnalité d'indexation et intégrer un outil d'extraction /reformatage des données.

La démarche paraît assez semblable à celle des CRB : collecter/caractériser/sécuriser/distribuer.

Le cycle de vie d'une donnée génomique peut être réfléchi en fonction de la nouveauté de la technique, de l'abondance des données du même type, de la standardisation des données qui facilite leur comparaison, et du caractère patrimonial de la donnée. La durée de valorisation d'une donnée est de 3 ans en moyenne, mais pour les espèces modèles, cette durée peut être bien plus longue. Les données mises à disposition par le portail d'accès de l'urgi sont maintenues depuis 10 ans.

Une donnée élaborée est répliquable si le pipeline d'analyse est stocké et donc réutilisable pour reproduire la donnée à partir des données brutes. Les données brutes sont alors utiles à garder si elles sont bien décrites.

On peut récapituler le cycle de vie par une figure assez simple, avec 4 cas typiques :

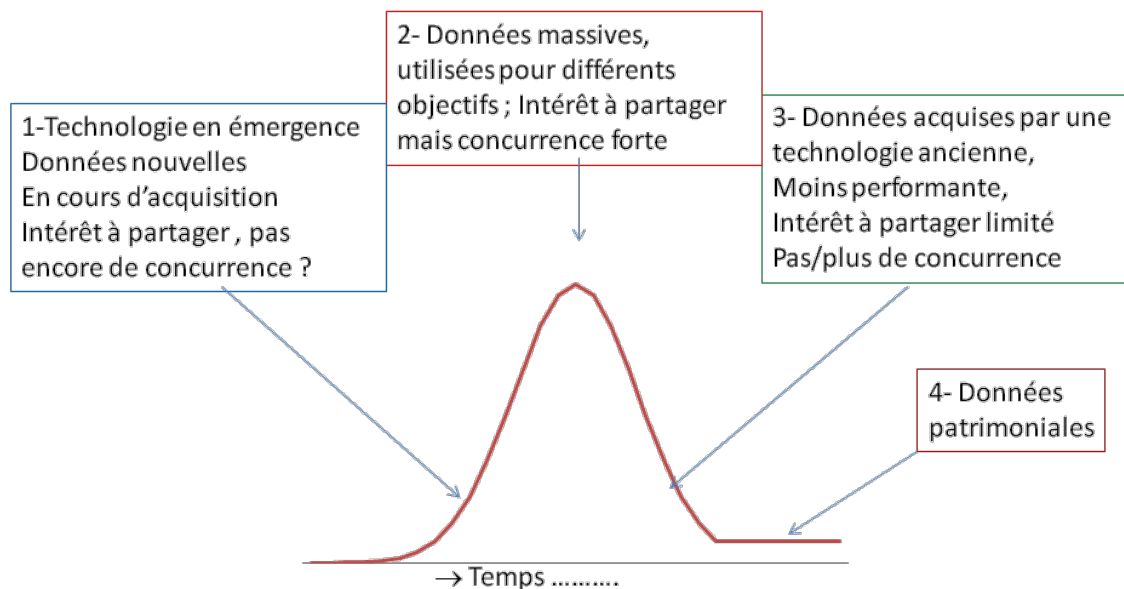
Cas 1 : exemple = patrons de méthylation,

Cas 2 : exemple = données SNP, données de séquence, données du transcriptome

Cas 3 : exemple = génotypes microsatellites , RFLP

Cas 4 : exemple données métagénomiques d'un échantillon daté, disparu ou modifié depuis l'acquisition de la donnée.

La figure ci-dessous illustre ces 4 cas.



Proposition pour décrire le cycle d'une vie d'une donnée 'omique'

c) vision stratégique : quelles données partager avec qui et pourquoi ?

On peut identifier différents cas selon que l'utilisation des données est plus ou moins générique:

Cas 1 : données de référence nécessaires à l'interprétation de quasiment toutes les autres données,

ex : génome de référence, généralement produit par un consortium, donc c'est une donnée partagée dès le début, actualisée par le consortium, mise à disposition de toute la communauté étudiant l'espèce concernée, avec référence vis-à-vis du consortium (publication de base ou site web).

Même s'il y a un intérêt stratégique à partager les données de référence, il est important d'être leader dans le consortium travaillant sur l'espèce étudiée, cela suppose un investissement massif mais procure plusieurs avantages : exclusivité temporaire, vue intégrative, capitalisation de toutes les connaissances. Il y a un intérêt stratégique à héberger un entrepôt de données pour une espèce modèle.

Cas 2 : données nécessaires à la comparaison de méthodes d'analyse, sous-ensemble de données dont l'origine est connue, référencée

ex : données de RNAseq servant à comparer des méthodes d'analyse statistique, il faut une concertation entre producteurs et analystes des données

Cas 3 : données complémentaires facilitant ou confortant l'interprétation des données d'un projet délimité

ex : accès aux données de séquence d'animaux non porteurs d'un phénotype particulier, objet du projet, intérêt à partager les données de réséquençage pour augmenter la fiabilité de l'association entre phénotype et génotype, le bénéfice apporté par l'accès à un plus grand nombre de données compense le risque de concurrence ;

D'une manière générale, la traçabilité des producteurs de données est importante pour favoriser le partage.

Le partage des données dans un cadre collaboratif permet de réaliser des méta-analyses, une approche de plus en plus importante pour l'intégration des connaissances.

Le partage des données suppose d'avoir confiance dans la **qualité des données** proposées, c'est en général la plateforme technologique produisant la donnée qui garantit sa qualité pour les technologies en 'milieu de vie' (typiquement qualité des séquences, des microarrays). Pour des technologies en début de vie, la validation de la qualité des données relève encore d'une démarche de recherche. L'existence de standards pour produire et échanger les données est déterminante pour permettre une analyse comparative entre données.

d) L'INRA dispose-t-il de données originales ?

Il faut distinguer l'originalité liée au matériel biologique étudié, de l'originalité liée au type de donnée ou à la technologie ayant produit la donnée.

Il paraît assez facile de trouver à l'INRA des exemples de données liées à un matériel original (génotype particulier, conditions expérimentales spécifiques).

En ce qui concerne les technologies, il faudrait contacter les plateformes. On pourrait penser à des données de transcriptome issues de types cellulaires spécifiques (ressource+ technique originales).

e) projets pilotes dans le domaine

Les exemples de partage de données 'omiques' ne manquent pas.

*Exemple en génomique animale : programme '1000 génomes bovins'

objectif : construire un jeu de données de référence

pour améliorer la prédiction de la séquence à partir de données de génotypage (car de nombreux animaux génotypes et quelques uns seulement séquencés, mutualiser les données des animaux séquencés pour mieux tirer parti des données de génotypage des nombreux candidats à la sélection)
pour trouver les mutations causales de phénotypes variés et analyser les effets biologiques
pour améliorer l'efficacité de la sélection

initiative : quelques chercheurs emmenés par un australien

étapes :

rédiger un accord de consortium explicitant les règles d'utilisation des résultats, dans le cas présent, grande liberté d'utilisation, mais pas de propriété intellectuelle exclusive, toute prise de brevet est associée à une licence gratuite pour tous les signataires et leurs financeurs, publics ou privés

définir les conditions techniques minimales pour entrer dans un 'club' mutualisant les données : quantité et qualité des données à fournir

répartir les rôles : producteurs de séquences envoient leurs fichiers bam au responsable de l'analyse qui centralise tous les fichiers bam, en extraient tous les polymorphismes et renvoient les fichiers vcf à tous les producteurs de séquences (qui ont alors accès aux polymorphismes issus de toutes les séquences produites), le responsable de l'analyse est le seul à conserver tous les fichiers bam.

mise en œuvre :

4 institutions publiques de 4 pays européens s'entendent avec le partenaire australien, elles apportent des données co-financées par des partenaires privés (sélectionneurs), conditions pour le premier consortium : fournir > 100X de séquences, > 10 animaux, >4X par animal, puis extension au Canada, puis Belgique, Finlande, Italie, Suisse (privé), 2014 : renforcement des conditions : fournir > 250X de séquences, > 25 animaux, >6X par animal, pour tous les membres + extension à plusieurs universités américaines, le Royaume-Uni, une université suisse. Un papier 'collectif' en révision pour Nature genetics.

QUESTION : l'INRA ne pourrait-il pas avoir l'ambition de jouer le rôle de centralisation et d'analyse pour d'autres espèces animales qui est joué par le partenaire australien pour l'espèce bovine ? Qu'en est-il d'une stratégie 'mille génomes' pour le poulet, le porc, le cheval ??

*Exemple en génomique végétale : Wheat IS

La *Wheat Initiative (International Research Initiative for Wheat Improvement)* est un projet international dédié à une meilleure connaissance de la génétique du blé afin d'accroître la résistance aux maladies et les rendements pour répondre à l'augmentation de la demande alimentaire mondiale. La création d'un système d'information global regroupant les données actuellement éparpillées est un des leviers clés qui permettrait de relever un tel challenge.

Le projet *Wheat IS*, coordonné par Hadi Quesneville, pour l'Inra, a pour perspective de construire un système d'information intégré sur le blé et de fournir ainsi à la communauté scientifique, d'une part, un accès facile aux données de la génétique et de la génomique de cette plante et d'autre part des outils bio-informatiques performants.

Ce projet repose sur un réseau international d'experts rassemblés autour de cet objectif ambitieux. Il réunit d'ores et déjà 16 partenaires académiques de 8 pays et le centre international d'amélioration du maïs et du blé (Cimmyt, MX) auxquels se sont associés des acteurs privés de la génétique et de la génomique du blé comme la compagnie *Dow seeds*.

A terme, le projet *Wheat IS* proposera un portail unique où les scientifiques du monde entier pourront disposer de données intégrées et consolidées, à même d'accélérer leurs recherches sur le blé.

*Exemple 1 en génomique microbienne : MicroSol.

MicroSol database (c) est le système d'information de la plateforme GenoSol de l'UMR Agroécologie de l'INRA de Dijon. Cette plateforme assure la centralisation et conservation des ressources génétiques microbiennes des sols des principaux sites expérimentaux et réseaux de surveillance nationaux. Afin d'appréhender l'écologie microbienne des sols, elle réalise la caractérisation moléculaire de ces échantillons de sol. Dans ce contexte, MicroSol database (c) garantit la traçabilité, le stockage, la sécurité et la mise à disposition des données métagénomiques des communautés microbiennes des sols ; il est constitué d'une base de données en PostgreSQL et d'une interface web (PHP, Zend) protégées par l'INRA (IDDN.FR.001.370007.000.R.A.2011.000.10300). Les données de diversité sur les microorganismes sont obtenues par séquençage massif du métagénome microbien (ADNr-16S et 18S des bactéries et champignons, respectivement). En sortie de séquenceur, le pipeline d'analyse GnS-PIPE spécialement développé permet la transformation des données brutes en composantes de la diversité microbienne : indices de diversité et inventaires taxonomiques. Ce sont ces données traitées qui sont sauvegardées dans MicroSol database (c). A ce jour, le SI contient les informations métagénomiques de 3 100 échantillons de sol, ce qui correspond à 53 276 761 séquences. Ces données sont mises à la disposition des partenaires de la plateforme via l'interface web qui assure la sécurité des données par identifiant et mot de passe individuel. Actuellement, le système n'est donc ouvert que pour des partenaires dans le cadre et le temps d'un projet bien défini.

QUESTION : le point faible réside au niveau des ressources humaines, l'IE informatique est en CDD depuis trois ans, sans pérennisation de cette compétence, la plateforme risque de régresser dans les mois et années à venir.

*Exemple 2 en génomique microbienne : Genomique comparée des levures du clade Lachancea

Dans le cadre du projet ANR GB-3G porté à l'INRA par C. Neuvéglise (Equipe BimLip, Unite Micalis), les génomes de 10 espèces de levures ascomycètes du clade Lachancea ont été séquencés et (ré)annotés dans un but de génomique comparative. Un ensemble d'outils a été développé à cet effet, soit directement dans l'équipe, soit en collaboration avec la société ISoft sur la base de leur suite logicielle Amadea (outil de transfert automatique d'annotation). En parallèle à cette étude, 6 génomes de levures ascomycètes du clade Yarrowia ont été séquencés et annotés avec une curation des modèles de gènes par données RNAseq obtenues pour différentes conditions de croissance. Pour l'analyse bio-informatique et la mise à disposition des données de séquence et annotation, l'équipe a opté pour l'acquisition en interne de 3 serveurs étant donné la saturation de la plateforme de bioinformatique la plus proche : serveurs pour les calculs (Dell R710 12 cores/24 threads, 96 Go

RAM), le web (Dell R420 12 cores/24 threads, 48 Go RAM) et les bases de données (Dell R420 6 cores/12 threads, 32 Go RAM), assortis d'une baie de stockage de 24 To. Une telle capacité de stockage est nécessaire au traitement des données haut-débit (assemblage de novo de génomes, analyse comparative de données RNAseq). Le projet a été rendu possible grâce aux compétences d'un post-doc en matière d'administration de serveur, puis grâce à un ingénieur bio-informatique INRA récemment recruté dans l'équipe. Ce dernier a réalisé le développement d'une base de données et d'un site web dédié pour héberger et diffuser les données. Le site porte le nom de domaine GRYC.inra.fr pour Genome Resources for Yeast Chromosomes. Ce site va être intégré dans une plateforme plus vaste destinée à héberger les données de génomiques d'un large panel de levures, sous l'égide du GDR-I CNRS iGénolevures nouvellement créé (2014-2018).

A terme plusieurs centaines de génomes de levures ascomycètes et basidiomycètes seront hébergées. Ces données proviendront de séquençages INRA, mais également de tout institut pour peu que les qualités de séquençage et annotations soient suffisantes.

***exemple d'un environnement collaboratif de recherche développé sur Genouest**

Genouest offre l'accès à un environnement intégrant partage de documents, gestion de projets, d'analyses de données, permettant la traçabilité des échanges et offrant des pointeurs vers les entrepôts de données et méta données. La gestion de méta-données (<http://emme.genouest.org/>) est basée sur ISA-tools (<http://isa-tools.org/>). Ces composants sont compatibles avec différentes architectures (répertoires de méta données uniques ou multiples, stockage des données distribué ou centralisé), et facilite la diffusion d'identifiants de type DOI/URI suivant une gestion fine des droits d'accès.

f) attentes en termes d'offre INRA : offre de services

→ boîte à outils pour partager : outils juridiques ; informatiques ? communication ?

Leçon du projet 1000 génomes bovins : l'outil 'accord de consortium' existe, il faut un 'centralisateur' de données ayant une forte capacité informatique + compétences d'analyse, la politique de publications n'a pas posé de difficultés jusqu'à maintenant, un ou quelques articles fondateurs co-signés de tous, et des articles ciblés sur tel ou tel point mais toujours avec toutes les données ;

Outils juridiques :

-S'appuyer sur les outils déjà disponibles pour définir les accords de consortium.

-Identifier les particularités liées aux traditions de partenariat selon les secteurs, les particularités liées au matériel biologique ;

-Prévoir le cas de l'abandon de données, comme celui de l'abandon de l'échantillon, le gestionnaire des données devient alors le propriétaire de données.

Lorsque les données ont été recueillies sans accord préalable, est-il plus difficile d'établir un accord a posteriori ?

Outils informatiques :

-développer les environnements collaboratifs pour faciliter l'accès aux données au sein de l'INRA,

-former les chercheurs et ingénieurs à la définition, le renseignement et l'utilisation des métadonnées, en commençant par ceux qui travaillent avec des plateformes d'acquisition de données

-porter quelques entrepôts de données de nature stratégique et/ou patrimoniale (espèce modèle, espèce d'intérêt économique majeur) et connecter ces choix aux projets de data center.

Tableau 2 : inventaire des entrepôts de données

type	nature	Entrepôt de données	INRA en accès ouvert	INRA accès contrôlé
brute	séquences lues ADN-ARN	ncbi/sra ; genbank ; MIPS-DB ;	ng6-GENOTOUL BIPAA-Archive, BBRIC-Archive, SUNRISE-Archive, GnpIS (module GnpSeqNGS)	Microsol, ng6
brute	génotypes SNP			
brute	génotypes SSR			
brute	données d'expression	GEO	GnpIS (module GnpArray)	
brute	données métabolome	Pathway Tools, GNPAnnot, ESTtik, Banana Genome Hub ; MetExplore		
brute	profils protéiques (quantitatifs)		BIPAA,	PAPSSO-ProticDB
élaborée	séquences protéines	Swissprot, uniprot, MIPS-DB, genbank, ENA, Ensembl Plant, CoGe, GNPAnnot, GreenPhyl, Banana Genome Hub, OryGenesDB, Phytozome	Portails LIPM BIPAA	
élaborée	séquences alignées/assemblées	Ensembl, ncbi, ucsc, Banana Genome Hub, CocoaGenDB, GreenPhylDB, OryGenesDB,	GnpIS (module GnpGenome)	
élaborée	données d'expression RNAseq	GEO, BIOS	GnpIS (en cours de développement)	
élaborée	polymorphismes SNP	Ensembl/dbSNP ; TropGeneDB, SNIPlay, Banana Genome Hub, CocoaGenDB, GenDiversity, OryGenesDB	GnpIS (module polymorphisme - GnpSNP)	CTIG (animal)
élaborée	polymorphismes SSR	dbSNP/Gramene, TropGeneDB, SNIPlay, Banana Genome Hub, CocoaGenDB, GenDiversity, GNPAnnotDB, OryGenesDB	GnpIS (module polymorphisme - GnpSNP)	CTIG (animal)
élaborée	variants structuraux	Ensembl	GnpIS (module polymorphisme - GnpSNP)	
élaborée	patrons de méthylation	GEO		
élaborée	annotations des gènes	HGNC, Ensembl-Plant, CoGe, GNPAnnot, Banana Genome Hub, CocoaGenDB, OryGenesDB ; genbank, ENA, Genolevures	Portails LIPM BIPAA, GRYC GnpIS (module GnpGenome)	

type	nature	Entrepôt de données	INRA en accès ouvert	INRA accès réservé
elaborée	orthologues, paralogues, familles de gènes	KEGG, PhylomeDB, GreenPhyl, Banana Genome Hub, COG, HomoloGene,	Narcisse	
elaborée	cartes (génétiques, QTLs, physiques)	QTLdb (animal) ; /Gramene/MIPSDB/ MaizeGDB/ GrainGene ; Banana Genome Hub, CocoaGenDB, TropGeneDB	GnpIS (module GnpMap)	
elaborée	données passeport populations/souches	EFABIS, DAD-IS, Eurisco/GBIF/Genesys ; CocoaGenDB, Oryza Tag Line, GenDiversity, TropGeneDB ; StrainInfo	CIRM (3 databases http://www.inra.fr/cirm), GnpIS (module Siregal)	CFBP (http://www-intranet.angers.inra.fr/cfbnet/p/)
elaborée	données passeport croisements temporaires	GenDiversity, TropGeneDB,		bases de données UE INRA
elaborée	données passeport banques génomiques	TropGeneDB, CNRGV, CRB-GADIE		