Étude sur l'utilisation des données de séquençage des ressources génétiques pour l'alimentation et l'agriculture

Fondation pour la recherche sur la biodiversité

CONTRIBUTEURS

CITATION
Fondation pour la recherche sur la biodiversité (2019), *Rapport de l'étude sur l'utilisation des données de séquençage des ressources génétiques pour l'alimentation et l'agriculture*, Paris, France : FRB, 103p.

Étude réalisée dans le cadre du projet d'étude sur l'utilisation des données de séquençage de ressources génétiques pour l'agriculture et l'alimentation commandée par le ministère de l'agriculture et de l'alimentation en mai 2018.
Direction : Ministère de l'agriculture et de l'alimentation
Coordination : Robin Goffaux (FRB)
Réalisation de l'étude : Charlotte Navarro (FRB)
Rédaction : Charlotte Navarro (FRB), Robin Goffaux (FRB)

REMERCIEMENTS
La Fondation pour la recherche sur la biodiversité tient à remercier l'ensemble du Comité de pilotage pour leurs expertises et orientations, ainsi que toutes les personnes qui ont pris le temps de répondre au questionnaire.

Membres du Comité de pilotage :
Jean-François Agnese (IRD)
Catherine Aubertin (IRD)
Audrey Didier (GEVES)
Denis Duclos (MNHN)
Guillaume Faure (MTES)
Robin Goffaux (FRB)
Philippe Grandcolas (CNRS)
Florence Hervatin-Queney (MESRI)
Jean Lanotte (MAA)
Selim Louafi (CIRAD)
Sylvain Mahé (AllEnvi)
Claire Neirac (CIRAD)
Jean-Louis Pham (IRD)
Alexandrine Rey (CIRAD)
Jean-François Silvain (FRB)
Michèle Tixier-Boichard (INRA)

Relecteurs : Catherine Aubertin (IRD), Audrey Didier (Geves), Robin Goffaux (FRB), Jean Lanotte (MAA), Sélim Louafi (Cirad), Jean-Louis Pham (IRD), Alexandrine Rey (Cirad), Hélène Soubelet (FRB), Jean-François Silvain (FRB).

Table des matières

INTRODUCTION

Depuis la 2e moitié du XXe siècle, les technologies de séquençage ont révolutionné la recherche dans les sciences du vivant et de nombreux domaines des sciences du vivant en utilisant les données, de la recherche fondamentale à la recherche appliquée, par exemple la taxonomie, la biologie de l'évolution, l'amélioration des plantes et la biologie de synthèse (Laird *et al.*, 2018 in Karger, 2018). La rapidité des analyses et la baisse des coûts ont permis de séquencer l'ADN, l'ARN et les protéines de ressources génétiques pour les objectifs de la recherche et du développement (Lawson and Rourke, 2016, in Karger, 2018). Il existe des Pétaoctets[1] de données de séquençage dans différentes bases de données à travers le monde en accès libre pour la plupart, conçues et alimentées par les chercheurs dans un cadre de bénéfice mutuel. Il est aujourd'hui possible d'être utilisateurs de bases de données sans avoir à accéder aux ressources génétiques et à les séquencer (Marx, 2013 in Karger, 2018).

Les ressources génétiques pour l'alimentation et l'agriculture (RGAA) sont des objets de recherche dans les sciences du vivant. A ce titre, leur utilisation et leur partage sont questionnés à différents niveaux, international, régional ou local.

Leur importance et leur rôle particulier quant à la sécurité alimentaire et nutritionnelle ou encore l'adaptation et la mitigation des changements climatiques, sont ainsi soulignés par la Commission sur les ressources génétiques pour l'alimentation et l'agriculture (CRGAA).

Elles relèvent également de la convention de la diversité biologique qui porte sur trois objectifs principaux : la conservation de la diversité biologique, l'utilisation durable de ses composants, et le partage juste et équitable des avantages découlant de l'utilisation des ressources génétiques. Le Protocole de Nagoya est le cadre juridique pour la mise en œuvre du troisième objectif de la CDB. Il reconnaît explicitement l'importance des RGAA pour la sécurité alimentaire, la nature particulière de la biodiversité agricole et ses caractéristiques distinctives (cf. annexe 1). De même, il reconnaît l'importance de ces ressources pour le développement durable de l'agriculture dans le contexte de lutte contre la pauvreté et le changement climatique (FAO, 2016).

Le libre accès aux données de séquençage de ressources génétiques n'est pas sans susciter des questionnements de la part de certains États parties à la Convention sur la diversité biologique (CDB) et au protocole de Nagoya, qui interrogent l'équité des dispositifs associés. Les considérations autour des données de séquençage sont nombreuses et divergent. Nécessitant plus de compréhension et de clarté dans ce débat, le ministère en charge de l'agriculture a confié à la FRB la réalisation de la présente étude.

Ce travail avait pour objectif de réaliser un état des lieux sur l'utilisation des données de séquençage en France pour alimenter les réflexions au niveau international portées par les décideurs français voire européens.

La 1re partie de ce rapport revient sur le contexte et les enjeux de l'utilisation des données de séquençage des ressources génétiques pour l'alimentation et l'agriculture. Un dispositif méthodologique d'enquête reposant sur une analyse de la bibliographie et sur des entretiens semi-directifs ont permis, par des cas d'utilisation emblématiques, de révéler les différents usages et intérêts relatifs à l'accès aux données de séquençage. La population enquêtée englobait à la fois les acteurs de la recherche, principaux utilisateurs des données de séquençage, mais aussi les entreprises de l'agroalimentaire et les acteurs institutionnels qui doivent faire face à ces questionnements.

La 2e partie du rapport portera sur le concept de données de séquençage et ses implications pour les ressources génétiques pour l'alimentation et l'agriculture.

La 3e partie met en avant des cas d'utilisation des données de séquençage de RGAA pour illustrer les réalités et les pratiques derrière ces notions.

[1] Un Pétaoctet, représente 1024 Téraoctets, ou autrement dit 1 million de milliards d'octets. Soit dix fois la quantité d'informations contenues dans la Bibliothèque du Congrès aux États-Unis, si elle était entièrement numérisée. Cette dernière conserve près de 100 millions d'ouvrages.

CADRAGE DE L'ÉTUDE ET DISPOSITIF MÉTHODOLOGIQUE MIS EN ŒUVRE

Commande du ministère de l'agriculture et de l'alimentation :

Dans le cadre de la 14e Conférence des Parties relative à la Convention sur la diversité biologique (CDB) qui s'est tenue du 10 au 22 novembre 2018 en Égypte, le ministère de l'agriculture et de l'alimentation a délégué à la Fondation pour la recherche sur la biodiversité (FRB) une étude relative à l'état des lieux de l'utilisation de l'information de séquençage numérique sur les ressources génétiques.

Plus particulièrement, l'étude devait répondre aux questions suivantes :

- Qui sont, aujourd'hui, les utilisateurs de séquences numériques agricoles et agroalimentaires (au sens large) ?
- À quelles fins les utilisent-ils ?
- Selon quelles pratiques?

Cette étude avait précisément pour objectif de donner un éclairage sur ce qu'est l'« information numérique de données de séquençage » de ressources génétiques pour l'alimentation et l'agriculture, quels sont les utilisateurs et les pratiques. Il s'agissait donc de réaliser un état des lieux sur ce concept et les réalités qui l'entourent. Ces informations devant servir à éclairer les décideurs et appuyer la position de la France dans les échanges internationaux sur la question de la régulation des échanges de données de séquençages relatives aux ressources génétiques.

Le dispositif méthodologique :

Pour mener à bien la mission, le dispositif méthodologique d'enquête a été conduit autour d'une analyse de la bibliographie et sur des entretiens semi-directifs permettant de révéler les différents usages et intérêts relatifs à l'accès aux données de séquençage. Plus particulièrement, cette démarche avait pour buts :

- Le recueil de données sur l'utilisation des données de séquençage, les objets d'étude et les pratiques ;
- L'identification d'applications concrètes de la production et de l'utilisation des données de séquençage ;

Elle a permis de faire émerger les faits, les opinions et les aspirations futures des personnes interrogées autour de la question de l'inclusion des données de séquençage dans le champ d'application du protocole de Nagoya. Ceci n'était pas prévu dans la commande, les éléments essentiels recueillis sont présentés dans la partie II/D.

La population enquêtée a englobé à la fois les acteurs de la recherche, principaux utilisateurs des données de séquençage, mais aussi les entreprises de l'agroalimentaire et les acteurs institutionnels en lien à ces questionnements. Elle est ainsi constituée de :

- plus de quarante chercheurs français de quatre instituts de recherche
- quatre représentants d'instituts techniques différents,
- un représentant d'un institut de biologie,
- un représentant d'un Groupement d'intérêt public (GIP),
- un gestionnaire de collection,
- trois représentants de différents établissements à caractère industriel et commerciale (EPIC),
- un représentant d'un établissement à caractère administratif (EPCA),
- cinq représentants de différentes entreprises semencières dont une spécialisée dans la biotechnologie,
- un représentant d'une association,
- deux représentants de deux coopératives différentes,
- un représentant d'une interprofession,
- diverses personnes issues d'un réseau d'instituts techniques,

- trois représentants de trois sociétés privées de trois secteurs différents (biotechnologie, propriété intellectuelle et sélection génomique),
- et trois représentants de trois syndicats.

La liste précise de ces structures est disponible en annexe 2.

L'enquête a débuté début juin 2018 et s'est terminée fin novembre 2018. La méthodologie adoptée est participative, tenant compte de la diversité des acteurs et des contraintes de temps. La première phase de l'étude a consisté en une recherche bibliographique donnant lieu à une synthèse documentaire.

La seconde phase de l'étude a consisté à sélectionner un panel d'experts et constituer un comité de pilotage, avant de réaliser les entretiens avec les professionnels et les chercheurs. Ces entretiens ont permis de comprendre les différents points de vue et enjeux liés à l'utilisation des données de séquençage, le rôle des interviewés, les attitudes adoptées, leurs visions pour le futur. Les thèmes des entretiens ont traité, entre autres, de la production des données de séquençage, de leur utilisation, de la valorisation économique de la biodiversité, du stockage et de l'accessibilité aux données.

Une première restitution des résultats sous la forme de séminaire a eu lieu le lundi 8 octobre 2018 à la Maison des Océans à Paris. Il a réuni près de cinquante personnes de divers horizons (ministères, diplomates, chercheurs, industriels, journalistes). Un compte-rendu, ainsi que les présentations sont accessibles sur le site de la FRB.

I. LE CONTEXTE ET LES ENJEUX DE L'UTILISATION DES DONNÉES DE SÉQUENÇAGE DE RESSOURCES GÉNÉTIQUES POUR L'ALIMENTATION ET L'AGRICULTURE

A. LES PRINCIPALES INSTITUTIONS AU NIVEAU INTERNATIONAL ET LE PRINCIPE DE L'APA

Dans cette partie, après un rappel de l'origine de la CDB, nous ouvrons sur les enjeux actuels de l'utilisation croissante de données de séquençage de ressources génétiques. Le débat actuel pose la question de l'inclusion de ces données de séquençage dans le champ d'application du protocole de Nagoya. La loi française sur la reconquête de la biodiversité, de la nature et des paysages de 2016, inclue des dispositions pour l'accès et le partage des avantages liés à l'utilisation des ressources génétiques. Cependant des régimes spécifiques s'appliquent pour certains types de ressources génétiques.

1. LES RESSOURCES GÉNÉTIQUES COMME PATRIMOINE COMMUN : UNE VISION PORTÉE NOTAMMENT AU SEIN DE LA FAO

À la fin des années 1970, le débat sur le statut juridique des collections internationales de ressources végétales émerge. Les pays en développement, menés par le Mexique, s'inquiètent de la collecte, de la conservation et de l'appropriation par des droits de propriété intellectuelle (PI) des ressources phytogénétiques mondiales réalisées par les pays développés. Pour répondre à leur attente, l'Organisation des Nations unies pour l'alimentation et l'agriculture (FAO) crée en 1984 une commission chargée de considérer ces questions, qui défend l'idée de patrimoine commun des ressources biologiques végétales[2]. Devenue depuis la Commission des ressources génétiques pour l'alimentation et l'agriculture (CRGAA), cette instance s'occupe spécifiquement de tous les éléments de la biodiversité pour l'alimentation et l'agriculture.

La biodiversité pour l'alimentation et l'agriculture comprend la diversité biologique présente dans les systèmes de production agricoles, pastoraux, forestiers et aquatiques ou ayant une importance pour ces systèmes. Il englobe la variété et la variabilité des animaux, des plantes et des micro-organismes aux niveaux génétiques, des espèces et des écosystèmes qui soutiennent la structure, les fonctions et les processus des systèmes de production. Cette diversité est gérée ou influencée par les agriculteurs, les éleveurs, les habitants des forêts et les pêcheurs depuis des centaines de générations et reflète ainsi la diversité des activités humaines et des processus naturels. Cette commission intergouvernementale unique promeut l'utilisation durable et la conservation de la biodiversité pour l'alimentation et l'agriculture[3] pour l'ensemble des RGAA (les végétaux, les animaux, les ressources génétiques aquatiques, les forêts, les micro-organismes et les invertébrés).

Le Traité international sur les ressources phytogénétiques pour l'alimentation et l'agriculture (TIRPAA), quant à lui a pour finalité de favoriser les échanges de matériels génétiques entre les utilisateurs dans un objectif de partage dans l'intérêt général[4].

2. DES ÉTATS SOUVERAINS SUR LEURS RESSOURCES GÉNÉTIQUES : UNE VISION PORTÉE AU SEIN DE LA CDB

La CDB est signée en 1992 (cf. annexe 2). Elle défend une approche souverainiste sur les ressources génétiques qui s'éloigne du principe de patrimoine commun pour la gestion harmonieuse des domaines d'intérêt commun.

Dans le cadre de la CDB, les États ont ainsi le droit souverain d'exploiter leurs propres ressources selon leur politique environnementale. Les États ont aussi le devoir de faire en sorte que « les activités exercées dans les limites de leur juridiction ou sous leur contrôle ne causent pas de dommage à l'environnement dans

[2] Cette notion de patrimoine commun a émergé dans les années 1960 dans un contexte de guerre froide où les États-Unis et l'URSS décident de faire des espaces extra-atmosphériques et des grands fonds marins, dont les bénéfices d'une appropriation sont incertains, des espaces internationalisés. Cette notion est notamment utile pour les pays du Sud qui n'ont pas les moyens d'explorer ces espaces et y voient une manière de les conserver (Smouts, 2005).

[3] Au même moment, la France met en place le bureau des ressources génétiques, un GIS qui traitait de tous les types de ressources génétiques (animal, végétal, microbien) et était en charge d'apporter une expertise au gouvernement français dans le cadre des négociations au sein de la FAO. En 2008, il fusionna avec l'Institut français de biodiversité pour créer la Fondation pour la recherche sur la biodiversité (FRB).

[4] Le préambule de ce traité écarte la notion de patrimoine commun au profit de la notion de « préoccupation commune ».

d'autres États ou dans des régions ne relevant d'aucune juridiction nationale » (Article 3 de la Convention sur la diversité biologique). L'accès aux ressources est soumis aux législations nationales et matérialisé par des accords écrits.

3. LA DÉCLINAISON OPERATIONNELLE DE L'ACCÈS AUX RESSOURCES GÉNÉTIQUES AU NIVEAU INTERNATIONAL

Le protocole de Nagoya (2010, cf. annexe 2) précise les deux mécanismes selon lesquelles doivent s'effectuer l'accès et le partage des avantages. D'une part, l'accès aux ressources et aux connaissances traditionnelles associées est soumis au consentement préalable en connaissance de cause (*Prior Informed Consent* – PIC) du fournisseur, sauf décision contraire de sa part. Et d'autre part, le partage juste et équitable des avantages issus de l'utilisation des ressources, est établit sur la base d'un contrat définissant les conditions convenues d'un commun accord (*Mutually Agreed Terms* - MAT) (CDB, 2002).

Depuis 2004, l'accès aux ressources phytogénétiques utiles à l'alimentation et l'agriculture fait l'objet d'une procédure d'APA particulière dans le cadre du TIRPAA. Sont concernées les espèces et genres végétaux listés dans l'annexe I du Traité international, dès lors que leur utilisation vise la conservation et la recherche, la sélection et la formation pour l'alimentation et l'agriculture[5]. Les ressources qui ne répondent pas aux critères d'inclusion dans le TIRPAA sont soumises au régime de l'APA (CDB/Protocole de Nagoya).

Un système multilatéral d'APA s'applique aux 64 espèces de l'annexe I se trouvant dans le domaine public et étant directement gérées et contrôlées par les États parties au Traité, ainsi qu'à celles mises volontairement dans le système par les détenteurs publics ou privés de collections de ressources phytogénétiques. Le système consiste en un ensemble de ressources accessibles à tous. Il comprend également les collections *ex situ* détenues par des centres internationaux de recherche agricole, dont celle du consortium « Groupe consultatif pour la recherche agricole internationale » (Article 11.5. Traité international sur les ressources phytogénétiques).

L'accès à l'ensemble de ces ressources se fait au travers d'un accord-type de transfert de matériel (ATTM) entre le gestionnaire de la collection et l'utilisateur (FRB, 2017).

Le débat entre les approches « patrimoine commun » et « souverainiste » est toujours d'actualité. Face à l'augmentation du nombre de brevet sur le vivant à partir des années 1980, les tensions se sont cristallisées. Aujourd'hui, le débat sur l'équité de l'utilisation des ressources génétiques s'étend aux données de séquençage. Le débat en cours est une opportunité pour éclairer les réalités qui entourent ce concept, ses caractéristiques et les acteurs impliqués dans son utilisation.

B. LES RÉGLEMENTATION EUROPÉENNE ET FRANÇAISE

Le règlement européen n° 511/2014 du 16 avril 2014 transpose le troisième volet du protocole de Nagoya, il ne concerne que le contrôle de conformité des procédures, il n'a pas vocation à réglementer l'accès aux ressources génétiques, qui reste de la compétence des États (annexe 2).

L'article 46 de la loi du 8 août 2016 via un décret signé le 21 novembre 2016 a opéré la ratification, par la France, du protocole de Nagoya et l'a donc rendu légalement contraignant en introduisant des dispositions d'accès aux ressources génétiques, aux connaissances traditionnelles associées et au partage des avantages découlant de leur utilisation (APA)[6].

La loi française et ses textes d'application définissent ainsi un régime général pour l'APA et plusieurs régimes spécifiques applicables à des catégories de ressources génétiques identifiées et notamment :

- Les ressources génétiques issues d'espèces cultivées ou domestiquées, c'est-à-dire « dont le processus d'évolution a été influencé par l'homme pour répondre à ses besoins » ;

- Les ressources génétiques des espèces sauvages apparentées, c'est-à-dire « toute espèce animale ayant la capacité de se reproduire par voie sexuée avec des espèces domestiquées, ainsi que toute

[5] Et à condition qu'ils ne soient pas destinés à des utilisations chimiques ou pharmaceutiques, ni à d'autres utilisations industrielles non alimentaires et non fourragères (cosmétique, bioénergie, fibres, etc.).

[6] Décret n° 2016-1615 du 21 novembre 2016 portant publication du protocole de Nagoya du 29 octobre 2010 sur l'accès aux ressources génétiques et le partage juste et équitable des avantages découlant de leur utilisation, relatif à la convention sur la diversité biologique, signé par la France le 20 septembre 2011 à New York.

espèce végétale utilisée en croisement avec une espèce cultivée dans le cadre de la sélection variétale » ;

- Les ressources génétiques objets de sylviculture ;
- Les ressources génétiques collectées par les laboratoires dans le cadre de la prévention, la surveillance et de la lutte contre les dangers sanitaires concernant les animaux, les végétaux et la sécurité sanitaire des aliments ;
- Les ressources génétiques collectées par les laboratoires au titre de la prévention et de la maîtrise des risques graves pour la santé humaine.

Le ministère chargé de l'agriculture ayant décidé de ne pas réglementer les ressources génétiques agricoles et agro-alimentaires, ces dernières ne sont pas soumises au régime général et ne font pas l'objet de dispositions particulières. Aucune démarche n'est donc à entreprendre dans le cadre de leur utilisation.

Enfin, la loi française ne s'applique pas aux ressources génétiques couvertes par des traités internationaux particuliers distincts de la CDB mais qui n'y sont pas contraires. C'est le cas du TIRPAA qui prévoit un système d'APA multilatéral pour 64 espèces listées dans son annexe I via un accord de « transferts de matériel ».

C. LES DONNÉES DE SÉQUENÇAGE : ENJEUX ET ACTUALITÉS

Dans un contexte de progrès rapides du génie génétique et des biotechnologies avec des développements difficilement contrôlables, de mondialisation des échanges avec une diffusion généralisée des droits de propriété intellectuelle, en particulier des brevets dans le domaine du vivant jusqu'alors relativement protégé, à la fin des années 1980, les ressources génétiques deviennent un enjeu économique et sont considérées comme des matières premières soumises aux lois de l'offre et de la demande.

La relation entre génétique et numérique est de plus en plus étroite. Elle est utilisée dans deux cas, le traitement de l'information et sa transmission.

L'utilisation de l'informatique pour **le traitement** de l'information génétique a permis des progrès majeurs (notamment la compréhension de la structure et l'expression des gènes) et a démultiplié les capacités d'analyse et donc la valeur de l'information contenue dans la ressource génétique.

La **transmission** des informations de séquençage est absolument essentielle dans un contexte d'augmentation du volume des bases de données bio-informatiques et le développement de la biologie de synthèse. La création de la première cellule vivante dotée d'un génome synthétique en 2008 par Craig Venter, le coauteur du premier séquençage humain en (2003), renforce l'idée de l'autonomisation des données de séquençage par rapport à la ressource génétique. L'édition de ce génome synthétique est alors une copie d'un génome existant ensuite transplanté dans une bactérie (*Mycoplasma capricolum*) (*Gibson et al.*, Science, 2010). Bien que ces découvertes ouvrent la voie à de nombreuses applications environnementales ou énergétiques, certains avis mettent en garde contre ce champ d'activité à haut risque, comme celui, par exemple, de reconstituer le génome d'un virus à partir de sa séquence.

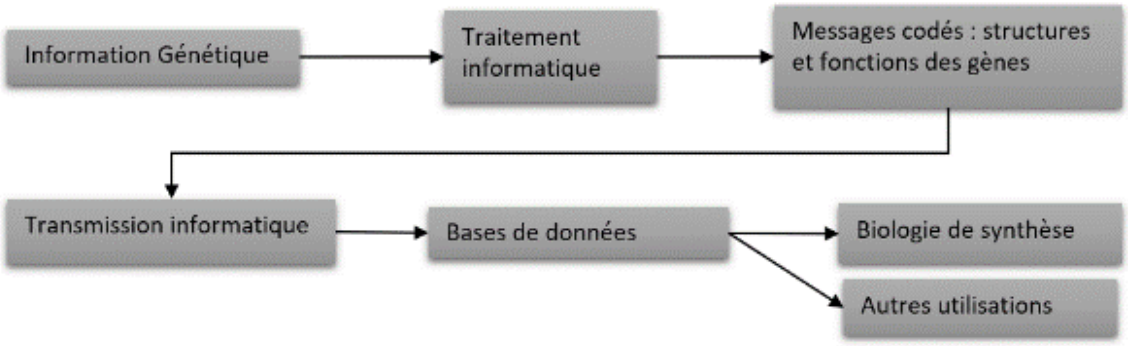


Figure 1 : Une relation étroite entre génétique et numérique (Rey, 2017)

Le développement des outils moléculaires, considéré par certains comme une révolution dans le monde de la biologie (Ledford, 2015 in Ducos, 2017), a permis la production d'espèces génétiquement modifiées et posent

également de nombreuses questions d'ordre scientifique, éthique, réglementaire, économique et stratégique (Ducos et al., 2017). Après la transgénèse, qui consistait en un transfert de matériel génétique d'un individu ou d'une espèce à une autre, la technique Crispr-Cas9[7], une technique d'édition de gènes permettant de modifier directement le génome sans transfert de matériel exogène, simple en termes de conception et d'utilisation et moins coûteuse, est de plus en plus accessible (cf. annexe 3). Dans ce cas, l'accès à l'information génétique de données de séquençage *in silico*[8] remplace l'accès au matériel génétique. Les progrès dans l'édition de génome ont accru les tensions et le débat sur les données de séquençage. Des inquiétudes liées à son utilisation pour le domaine de l'agriculture et de l'alimentation ont émergé : contournements de résistance[9], gestion de la diversité des populations, nature juridique des innovations, diffusion cachée de mutations, problèmes éthiques, etc. En juillet 2018, la Cour de justice européenne a rendu son arrêt sur les organismes obtenus par mutagénèse dirigée qu'elle considère comme des organismes génétiquement modifiés[10].

Le passage de l'utilisation de ressources naturelles aux ressources informatiques et synthétiques pose alors la question du statut de ces dernières : les bases de données et la bio-informatique sont-elles comprises dans la notion de « ressource génétique » (Cirad, 2011) ? En d'autres termes, sont-elles des ressources génétiques à part entière ?

C'est le cœur du débat qui anime les discussions des Parties signataires entre autres, de la CDB, du TIRPAA et de la Convention sur le droit de la mer.

Les conclusions des dernières Conférences des Parties (COP) à la CDB :

Lors de la COP13 et de la COP-MOP2 à Mexico en décembre 2016, la question des données de séquençage (« *Digital sequence information* », DSI) a été soulevée. Plusieurs visions ont été portées, allant d'une exclusion des données de séquençage des réglementations portant sur les ressources génétiques à une volonté d'inclure ces données dans les questions de partage des avantages.

Les Parties ont mis en place un processus de collecte d'informations et de points de vue sur la question en préparation des négociations futures sur le statut à attribuer aux données de séquençage. Un groupe d'experts international (AHTEG) a été mobilisé pour examiner les implications potentielles de l'utilisation des données de séquençage pour les objectifs de la CDB et du protocole de Nagoya. En février 2018, l'AHTEG a rendu compte de ses conclusions à l'organe subsidiaire chargé de fournir des avis scientifiques, techniques et technologiques (SBSTTA) qui s'est réuni en juillet 2018.

Lors de cette réunion, l'AHTEG a proposé un cadre de recommandations concernant les données de séquençage de ressources génétiques pour la préparation de la prochaine Conférence des parties de novembre 2018. Il faut noter que des questions restent en suspens et que la majorité des décisions font toujours l'objet de négociations (cf. annexe 4).

Le mandat du second groupe d'experts comprend les éléments suivants :

- Prendre en considération la compilation et la synthèse des points de vue et des informations
- Examiner la compilation des points de vue et des informations ;
- Clarifier le concept d'information numérique de séquençage et désigner un terme fonctionnel.

À ce stade, les parties s'accordent sur deux aspects, d'une part l'importance de la DSI pour les objectifs de la CDB et d'autre part la nécessité de clarifier les termes, d'approfondir les connaissances et d'étudier les effets d'une éventuelle réglementation avant son adoption. L'accès libre est considéré par certaines parties comme une forme de partage des avantages alors que d'autres y voient au contraire un régime qui ne résout en rien les problèmes d'équité et qui de plus pourrait constituer une forme possible de contournement des règles de partage établie pour les ressources génétiques.

[7] Voir le glossaire.

[8] Voir le glossaire.

[9] Des contournements de résistance ont été régulièrement décrits dans le cas des plantes génétiquement modifiées (résistance de populations d'insectes ravageurs aux toxines sécrétées par les variétés de maïs Bt notamment ; voir par exemple Tabashnik *et al.*, 2013) (Ducos, 2017).

[10] Cour de justice de l'Union européenne, le 25 juillet 2018, ECLI:EU:C:2018:583.

Au niveau international, la France s'exprime à travers la présidence tournante du Conseil de l'Union européenne, actuellement la Roumanie, et la Commission européenne. Lors de la COP en Égypte, la position de la France rejoint celle de l'Europe : le champ d'application du protocole de Nagoya ne couvre pas les DSI et, étant donné qu'il n'y a pas de définition consensuelle au niveau international, il apparait difficile de discuter d'un élargissement du protocole de Nagoya.

À l'issu de la COP14, le projet de décision relatif aux données de séquençage sur les ressources génétiques décide de créer un nouveau groupe d'experts techniques en vue de compiler les points de vue et informations pour préciser le concept, la terminologie et le champ d'application et comment les mesures nationales sur l'accès et le partage des avantages tiennent compte des DSI. Ce groupe a aussi pour objectif de réaliser une étude fondée sur la science et évaluée par des pairs, sur :

- le concept et le champ d'application ainsi que son utilisation ;
- l'évolution en cours dans le domaine de la traçabilité, notamment dans les bases de données ;
- les bases de données publiques et privées, notamment sur les conditions d'accès, la champ d'application biologique et la taille des bases de données, le nombre d'accès et leur origine, les politiques de gouvernance et les fournisseurs et utilisateurs des DSI ;
- les mesures nationales qui traitent du partage des avantages découlant de l'utilisation des données de séquençage à des fins commerciales ou non.

II. LES DONNÉES DE SÉQUENÇAGE DE RESSOURCES GÉNÉTIQUES POUR L'ALIMENTATION ET L'AGRICULTURE

Les ressources génétiques pour l'alimentation et l'agriculture sont définies au niveau international mais cette définition n'inclue pas les données de séquençage. L'évolution des recherches dans le domaine des sciences du vivant nous amène à comprendre quelle réalité recouvre le terme de données de séquençage.

A. LES RESSOURCES GÉNÉTIQUES POUR L'ALIMENTATION ET L'AGRICULTURE ET LES DÉBATS SUR LES DÉFINITIONS DE LA CDB ET DU PROTOCOLE DE NAGOYA

Dans cette partie, nous revenons sur plusieurs définitions de la FAO et de la CDB et leur lien avec les données de séquençage.

1. LES RESSOURCES GÉNÉTIQUES POUR L'ALIMENTATION ET L'AGRICULTURE

L'Article 2 de la CDB définit **les ressources biologiques** comme : « les ressources génétiques, les organismes ou éléments de ceux-ci, les populations, ou autres éléments biotiques des écosystèmes ayant une utilisation ou une valeur effective ou potentielle pour l'humanité ».

Les ressources génétiques[11] font référence au matériel génétique ayant une valeur effective ou potentielle. Elles contiennent le matériel génétique, d'origine végétale, animale, microbienne ou autre, contentant des unités fonctionnelles de l'hérédité. Les ressources génétiques peuvent être *in situ* (dans le milieu) ou *ex situ* (en collection). A l'époque où cette définition a été créée, les ressources *in silico* n'ont pas été prises en compte.

La notion de « **matériel génétique**[12] » comprend l'ensemble des caractères héréditaires d'un organisme vivant, exprimés en termes d'information génétique, et leur support matériel.

La CRGAA, l'organe intergouvernemental chargé de la biodiversité pour l'alimentation et l'agriculture, donne une définition pour chaque type de ressources génétiques pour l'alimentation et l'agriculture[13] (cf. tableau 1). Ces définitions sont très hétérogènes, en fonction du type de ressources on se réfère à du matériel, des espèces ou des semences.

[11] Définition des termes dans l'Article 2 de la Convention sur la diversité biologique.

[12] Définition des termes dans l'Article 2 de la Convention sur la diversité biologique.

[13] En annexe 4, le tableau recensant les caractéristiques propres de la FAO et de la CRGAA.

Tableau 1 : Les Ressources génétiques pour l'alimentation et l'agriculture considérées pour la présente étude

Types de Ressources génétiques pour l'alimentation et l'agriculture	Définitions de la CRGAA de la FAO[14]
Ressources phytogénétiques	Semences et matériel génétique de variétés traditionnelles et de cultivars[15] modernes, de plantes sauvages apparentées et d'autres espèces de plantes sauvages.
Ressources zoogénétiques	Espèces de mammifères et aviaires domestiquées.
Ressources génétiques aquatiques	Espèces aquatiques d'élevage et espèces sauvages apparentées, dans les juridictions nationales.
Ressources génétiques forestières	Matériel héréditaire contenu dans les arbres et autres espèces ligneuses.
Ressources génétiques de microorganismes et d'invertébrés	Groupe d'espèces le plus nombreux sur Terre. Les invertébrés sont des animaux sans colonne vertébrale (95% de tous les animaux). Les microorganismes sont trop petits pour être vus par l'œil humain.

Un des points de débat lors des négociations entre les parties au sujet des données de séquençage de ressources génétiques est la terminologie employée.

2. DÉBATS SUR LES DÉFINITIONS AU SEIN DE LA CDB ET DU PROTOCOLE DE NAGOYA

Les considérations sur le statut juridique des informations de données de séquençage révèlent plusieurs manques en termes de définitions au sein de la CDB et du protocole de Nagoya.

Unité fonctionnelle de l'hérédité : En raison de l'évolution rapide des connaissances sur le gène, la notion d'unité fonctionnelle de l'hérédité n'a pas été scientifiquement définie. Les parties ne sont pas d'accord sur l'idée que le concept de « séquence » inclut ou non « unités fonctionnelles de l'hérédité ».

Les notions de « **valeur effective ou potentielle** » associées aux ressources génétiques interrogent quant à la valeur intrinsèque des ressources génétiques. Il semblerait que ce sont les Etats qui assignent ces valeurs en fonction des travaux des sélectionneurs ou des investisseurs liés à l'utilisation des biotechnologies (Rey, 2017). Cette attribution de valeurs aux ressources génétiques devrait être liée au partage des avantages dans le cadre de la CDB. Au-delà du potentiel commercial des ressources génétiques, la notion de valeur englobe également leur rôle dans les cultures traditionnelles.

L'« **utilisation des ressources génétiques** » est comprise comme « l'ensemble des activités de recherche et de développement sur la composition génétique et/ou biochimique de ressources génétiques, notamment par l'application de la biotechnologie », conformément à l'article 2 de la Convention.

La « **biotechnologie** » est définie par l'Article 2 du Protocole de Nagoya comme « toute application qui utilise des systèmes biologiques, des organismes vivants, ou des dérivés de ceux-ci, pour réaliser ou modifier des produits ou des procédés à usage spécifique », conformément à la définition fournie dans l'article 2 du Protocole de Nagoya.

Le terme « **dérivé** » n'est pas définit par la CDB mais le protocole de Nagoya propose de le définir comme tout composé biochimique qui existe à l'état naturel résultant de l'expression génétique ou du métabolisme de ressources biologiques ou génétiques, même s'il ne contient pas d'unités fonctionnelles de l'hérédité.

[14] La CRGAA s'efforce d'obtenir un consensus international sur des politiques en faveur de la conservation et de l'utilisation durable des ressources génétiques pour l'alimentation et l'agriculture, ainsi que d'un partage juste et équitable des avantages découlant de leur utilisation.

[15] Terme internationalement reconnu représentant une variété de plantes cultivées. Cette dernière doit se différencier des autres variétés par des caractéristiques données qu'elle doit conserver quand elle est reproduite dans des conditions bien déterminées.

La question des dérivés a pris de plus en plus d'importance au fur et à mesure des avancées et pratiques de la recherche sur le vivant. Il existe donc une « course poursuite » entre les avancées et pratiques de la recherche d'une part et l'extension du champ de l'APA d'autre part (Aubertin, 2018).

Ces définitions confirment que l'utilisation de ressources génétiques ne peut pas être restreinte au matériel génétique contenant des unités fonctionnelles de l'hérédité mais inclut aussi l'ARN, les protéines, les composés biochimiques existant à l'état naturel. En élargissant le champ d'application de l'APA aux dérivés ne contenant pas d'« unités fonctionnelles de l'hérédité », la ressource génétique semble être qualifiée plus par l'utilisation des ressources biologiques et moins par la présence d'information génétique (Rey, 2017).

B. HISTOIRE DE LA BIOLOGIE MOLÉCULAIRE À LA GÉNOMIQUE

Cette partie vise à décrire les évolutions dans le domaine des sciences du vivant liées aux progrès technologiques. Les données de séquençage deviennent des éléments d'analyse essentiels et la communauté scientifique s'organise pour permettre le partage de ces données à travers des bases de données en accès libre.

1. L'ESSOR DE L'OBSERVATION DU GÉNOME

Une cellule vivante contient un ensemble d'instructions qu'on appelle le génome où chaque instruction est un gène. Une instruction est codée sous forme chimique et s'organise en une molécule constituée de quatre éléments appelés bases nucléiques (Adénine, Cytosine, Guanine, Thymine pour l'ADN ou Adénine, Cytosine, Guanine, Uracile pour l'ARN[16]). L'enchaînement de ces quatre bases, appelé « séquence », permet de coder ces instructions, comme la succession d'octet code des informations pour des programmes informatiques. Ici, le codage est chimique. Une séquence est donc une représentation de l'enchaînement des constituants élémentaires des molécules d'ADN (bases) symbolisés par les lettres ACGT (Weissenbach, 2000).

Le séquençage est l'ensemble des manipulations permettant de déterminer la séquence d'une molécule d'ADN, d'ARN, ou d'une protéine. Les bio-informaticiens manipulent trois alphabets : ADN, ARN et protéines. L'étude du génome peut être comparée à celle de la lecture d'un grand texte où les biologistes cherchent des mots, des fonctions, pour comprendre les mécanismes du vivant (figure 2). La taille des génomes varie fortement en fonction des espèces considérées.

La biologie moléculaire est au cœur des activités scientifiques d'une grande partie des chercheurs qui étudient l'expression de l'information génétique et ses régulations (cf. annexe 6). Elle a pour objet l'étude des macromolécules biologiques comme les acides nucléiques, dont l'ADN, et les protéines. Les techniques rattachées à la biologie moléculaire sont de l'ordre de l'exploration du vivant et de l'informatique (Gallezot, 2002).

Encadré 1 : La biologie moléculaire

[16] ADN : acide désoxyribonucléique. Molécules de taille importante contenant les instructions (gènes) et qui constituent les chromosomes (J. Weissenbach, 2000). ARN : acide ribonucléique. Molécule constituée d'un enchainement de nucléotides, considéré comme support intermédiaire des gènes pour synthétiser les protéines, et qui possède d'autres fonctions. A pour adénine, C pour cytosine, G pour guanine, T pour thymine et U pour uracile.

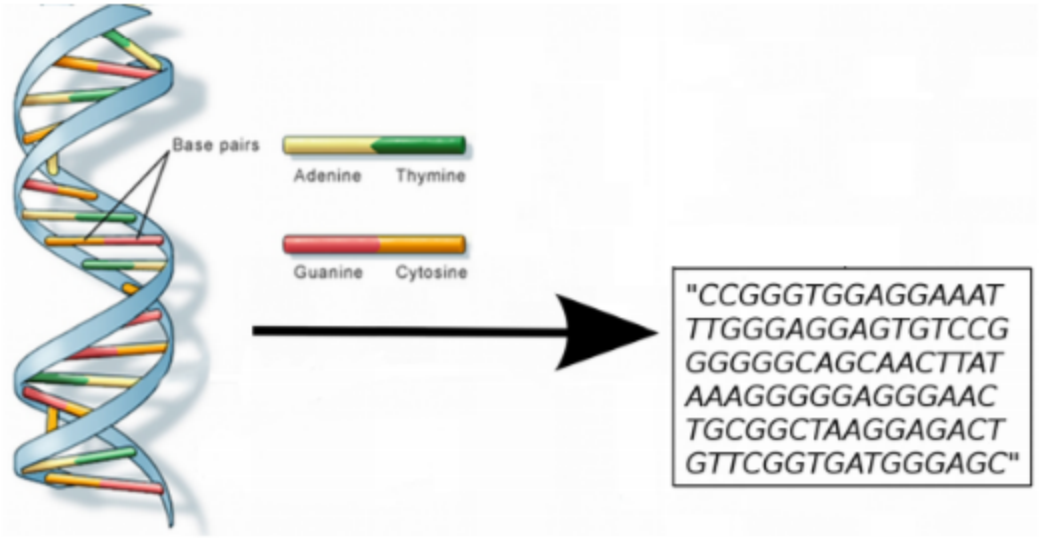


Figure 2 : Aspect biologique et aspect informatique de l'ADN (Leleux, 2014)

La biologie moléculaire est une discipline en soi, qui produit entre autres, des outils permettant de décoder et manipuler le génome et utilisés dans d'autres disciplines. Par exemple, la génétique évolutive est une discipline qui utilise des outils de biologie moléculaire (tableau 2) afin d'étudier et comparer les programmes génétiques développementaux d'organismes modèles.

Tableau 2 : Domaines et outils de la biologie moléculaire (adapté de Gallezot, 2002)

Domaines techniques liées à la Biologie moléculaire	**Outils**
Manipulation du vivant	Clonage, électrophorèse, hybridation, PCR (*Polymerase chain reaction*), séquençage, transcriptomique, protéomique, transgénèse, édition de génome etc.
Informatique	Ordinateurs, langages informatiques et systèmes d'exploitation, Internet, Systèmes de gestion de bases de données (SGBD),etc.

La génomique est l'étude de l'ensemble de gènes qui caractérisent les différentes espèces et définissent leur génome. Elle a subi des évolutions marquantes ces dernières années (Gaspin, 2015), passant d'une phase descriptive à une phase d'expérimentation fonctionnelle. L'analyse du génome est essentielle dans l'étude du vivant.

Encadré 2 : La génomique. Gaspin Christine (2015). « Les données de la recherche dans le domaine des sciences du vivant : évolution et perspectives à la lumière des nouvelles technologies du numérique et d'exploration du vivant », Présentation à Toulouse.

S'intéresser aux activités des chercheurs permet de comprendre les évolutions de l'utilisation de l'information de données de séquençage. Leurs activités comprennent : **la collecte d'informations** qui peut être réalisée à partir de différents supports (banques de données, sites internet, expériences en laboratoires, collecte sur le terrain); **le traitement de cette information** qui est généralement effectué par des bioinformaticiens ou des scientifiques avec des connaissances en bioinformatique ; **la diffusion des connaissances** qui correspond à des opérations standardisées (Gallezot, 2002). Ces activités forment un cycle marqué par l'acquisition d'informations dans les bases de données et le traitement informatique par les nouvelles techniques de séquençage (cf. figure 2).

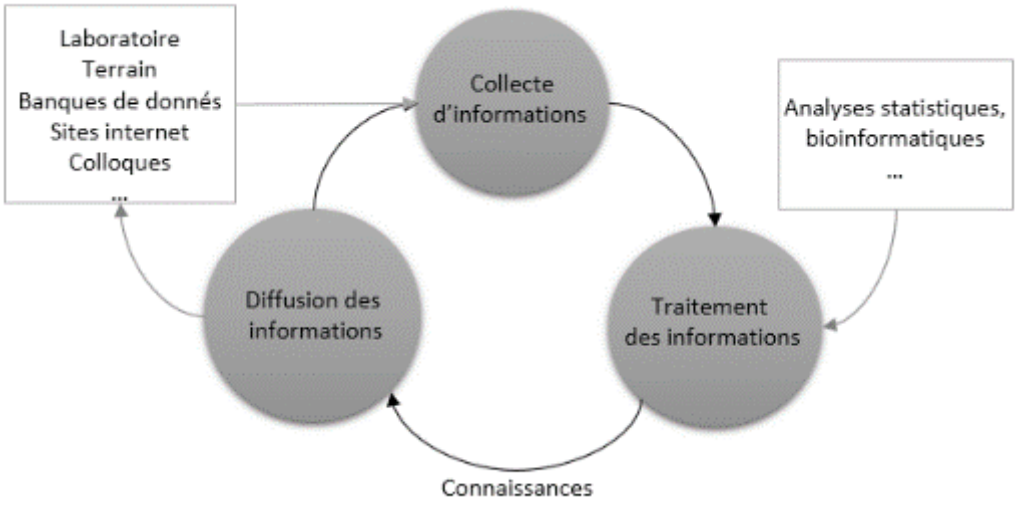


Figure 3 : La cycle de l'information scientifique et technique (Gallezot, 2002)

La frise chronologique ci-après résume des grandes étapes historiques de l'évolution de la génomique et notamment les évolutions des techniques de séquençage (Gaspin, 2015) :

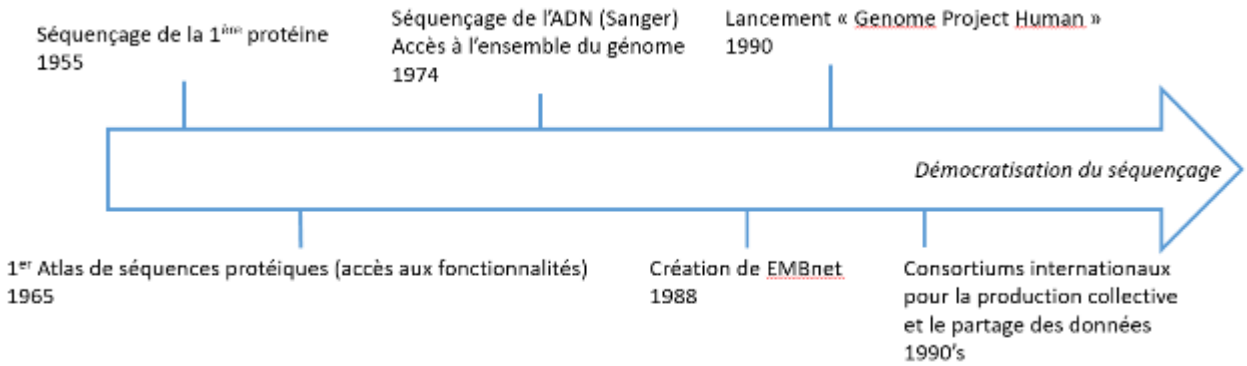


Tableau 3 : Frise chronologique : Les grandes étapes historiques de l'évolution de la génomique et notamment les évolutions des techniques de séquençage[17]

Les techniques de séquençage ont beaucoup évolué, de la méthode Sanger développée par la société Thermo dans les années 1970, aux nouvelles générations de séquençage développées par Illumina puis Oxford Nanopore Technologies en 2015 (cf. annexe 7).

Les dernières techniques de séquençage permettent la lecture de dizaine, voire de centaine de milliers de paires de bases. Leur faible coût les rend très attractives. Cependant certaines techniques engendrent des taux d'erreurs importants, de l'ordre de 10%. L'enjeu pour les bioinformaticiens est de travailler pour minimiser ces taux d'erreurs. De plus, ces séquençages requièrent de l'ADN de haut poids moléculaire, c'est-à-dire de bonne qualité et en quantité suffisante (mesurée en paires de base).

Actuellement on peut classer les domaines d'expertise liés aux nouvelles technologies de séquençage de l'ADN en quatre catégories[18] selon l'objectif poursuivi:

[17] Gaspin Christine (2015). « Les données de la recherche dans le domaine des sciences du vivant : évolution et perspectives à la lumière des nouvelles technologies du numérique et d'exploration du vivant », Présentation à Toulouse.

[18] Domaines d'expertise in France génomique, *france-genomque.org*, novembre 2017 [consulté le 7 juillet 2018]. https://www.france-genomique.org/spip/spip.php?article63&lang=fr

- Le séquençage *de novo* qui consiste à obtenir la séquence d'un organisme entier (assemblage de données de séquence d'un génome inconnu),
- le reséquençage du génome complet pour répertorier les variations nucléotidiques et structurales[19],
- le reséquençage ciblé d'ADN (ou « capture ») par amplicon ou hybridation[20],
- le génotypage qui consiste à l'extraction de l'ADN et à son analyse sur puce[21].

Des techniques existent pour le séquençage de l'ARN et permettent l'étude des transcrits (ou transcriptome), elle permet par exemple de quantifier l'expression des gènes d'un organisme. D'autres techniques encore permettent l'étude de la régulation de l'expression des gènes (épigénétique).

Les types de séquences générés peuvent être de différentes natures : mitochondriale, ribosomique, nucléaire, génome partiel, complet, etc. La démocratisation et la baisse des coûts du séquençage génère un accroissement du nombre de données produites (cf. annexe 8) et posent des défis de gestion et de stockage.

2. GESTION DES DONNÉES DE SÉQUENÇAGE

Depuis les années 1970 et les techniques de génie génétique comme le clonage, les capacités de traitement informatique et l'avancée de la génétique se sont développées conjointement.

La bioinformatique est le traitement automatique de l'information biologique. Elle applique les approches de l'informatique à la génomique : acquisition, organisation, analyse, visualisation, modélisation de l'information. Aujourd'hui, les équipes de recherche sont constituées d'informaticiens, de mathématiciens, de biologistes et des chercheurs en science de l'information.

Encadré 3 : La bioinformatique

Les banques internationales voient le nombre de séquences augmenter rapidement depuis le séquençage réalisé avec des séquenceurs informatisés dans le cadre du Projet de séquençage du génome humain *(Human Genome Project*). Composé de 3 milliards de bases, ce séquençage a couté 300 millions de dollars et 10 ans de travail. Aujourd'hui il peut être réalisé en quelques jours et pour quelques milliers de dollars seulement. Ce projet a poussé la recherche dans l'ère de la génomique. Les résultats de ce projet ouvrent la voie à une nouvelle génération de programmes de recherche qui vise à décrypter la fonction des gènes nouvellement détectés et ce chez toutes les espèces du vivant. La génétique fonctionnelle regroupe alors les approches biochimiques et physiologiques à des analyses de génomes entiers.

Pour gérer les données produites, les systèmes de gestion de bases de données (SGBD) se développent. Pour chaque base de données, il existe un format de soumission propre et des outils de recherche adaptés au type d'accès prévu. La condition pour publier dans certaines revues du domaine est le dépôt des séquences dans une base de données ouverte avant publication, notamment lorsque la recherche est financée sur fonds publics (Gallezot, 2002). Ces conditions s'inscrivent dans un cadre déontologique propre aux utilisateurs de données de séquençage qui prône le partage pour la vérification des résultats et la réutilisation des données. Ce dépôt s'accompagne de l'attribution d'un identifiant référencé, le « DOI », pour identifier les séquences et en assurer une traçabilité.

Les séquences générées sont donc stockées et référencées dans des bases de données internationales. L'accès libre permet la réutilisation des données dans un objectif de valorisation. Dès les années 1990, les accords de consortium internationaux ont défini un cadre de production collective et de partage des ressources génétiques et des données de séquençage en vue de leur exploitation. Le partage de la donnée dans le domaine de la biologie n'est donc pas récent.

[19] Les séquences issues de reséquençage sont alignées sur la séquence de référence pour identifier les variations génomiques de type SNP, CNV, indels (insertion ou délétion) et les réarrangements chromosomiques (déplacement, suppression, duplication de parties de séquence).

[20] Séquençage des fragments d'intérêt amplifiés par PCR (*Polymerase Chain Reaction*) ou par hybridation sur des sondes spécifiques.

[21] Une fois l'ADN préparé et sur puce, il est scanné pour permettre la lecture des marqueurs génétiques.

La collaboration internationale des bases de données sur les séquences de nucléotides (INSDC) est une initiative majeure entre le centre national pour l'information relative à la biotechnologie (NCBI) aux États-Unis, le laboratoire européen de biologie moléculaire (EMBL) et la banque de données ADN du Japon (DDBJ) (cf. annexe 9). Cette collaboration couvre le spectre des lectures de données brutes, des annotations fonctionnelles et des informations contextuelles relatives aux échantillons et aux configurations expérimentales (tableau 4).

Tableau 4 : Les types de données proposés par la collaboration internationale des bases de données sur les séquences de nucléotides (INSDC) (tableau issu du site officiel du NCBI)

Type de données	Base de données DDBJ	Base de données EMBL-EBI	Base de données NCBI
Données de séquençage brutes	Sequence Read Archive	ENA (European Nucleotide Archive)	Sequence Read Archive
Lecture capillaire	Trace Archive		Trace archive
Séquences annotées	DDBJ		GenBank
Echantillon	BioSample		BioSample
Etudes	BioProject		BioProject

Dans ces bases de données on retrouve les données de séquençage brutes qui sont les données obtenues en sortie d'un équipement de mesure (cf. figure 4) ; les lectures de séquences issues d'électrophorèse capillaire qui sont les chromatogrammes[22] de séquence d'ADN (cf. figure 5) ; les séquences annotées qui sont les régions fonctionnelles identifiées, souvent les gènes codant les protéines (cf. figure 6). Enfin, les bases de données BioSample contiennent des descriptions de matériels biologiques pour des essais expérimentaux (identifiant, organisme, titre, description, lien, etc.) et la collection BioProject est une collection de données biologiques relié à un projet (champ, méthodologie, objectifs).

```
>gnl|ti|2344447466 name:C07-S82_F_Control-A_Verifications3-ForwardSeq_059 mate_pairs:2344447470,2344447471,2344447472,2344447473
AACTTCNTTCNGCATGAAATGTAGCTCATTACCTGGTAGGCTTTAACAATTTCTCGTTTGCNTGGCGTTC
TTTTTCACACCGAGTATTTTGTAATAATCACGTTTCTTTGCTTGTTTTTGAAGTTTCTGAACTTTTTGCA
TTCCTTCCTTTGCTCTATTGAATGATTCTTCGATTTCTAATGCCTAAAGATAAGTTGCTTGATGATTAAT
CCAAGACATCAATATATCAATATGTTAATTAATAATTTACCTTAGCATAATCTTGTTGAGCCTCGTCGAG
CATATCAAGAGCAATGTAATCTTCAGCTCTGTCGCACAATATCCTCGGCTCTGTGTGCCGAGCTAAAGCT
GAAGAGCAGAACTCNAGACTTTCCGAGTGTTCACCTAGTTGAAGATGGCAATGNCAAAGTTTATCTTCAG
CCAAAAATAATAATTCTGATACGGTGGGATCCACTTTCAGGACCTATATAAATGGTTACATCGTTATTTA
TTGTGGAATTTAATAATTCAAAGTAAAATACTTTTTTAGCTGAAGTGATGCATCCCTTGGTAANNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Figure 4 : Données de séquençage brute de l'espèce Daphnia pulex (Sequence Read Archive, 2018)

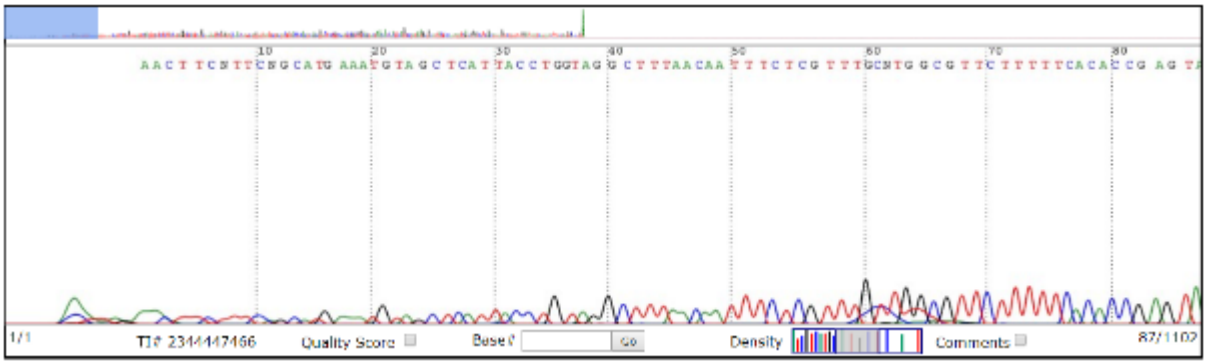


Figure 5 : Les lectures capillaires ou chromatogrammes de séquence d'ADN de l'espèce Daphnia pulex (Trace archive NCBI, 2018)

[22] Diagramme résultant d'une chromatographie, technique permettant de séparer les composants chimiques.

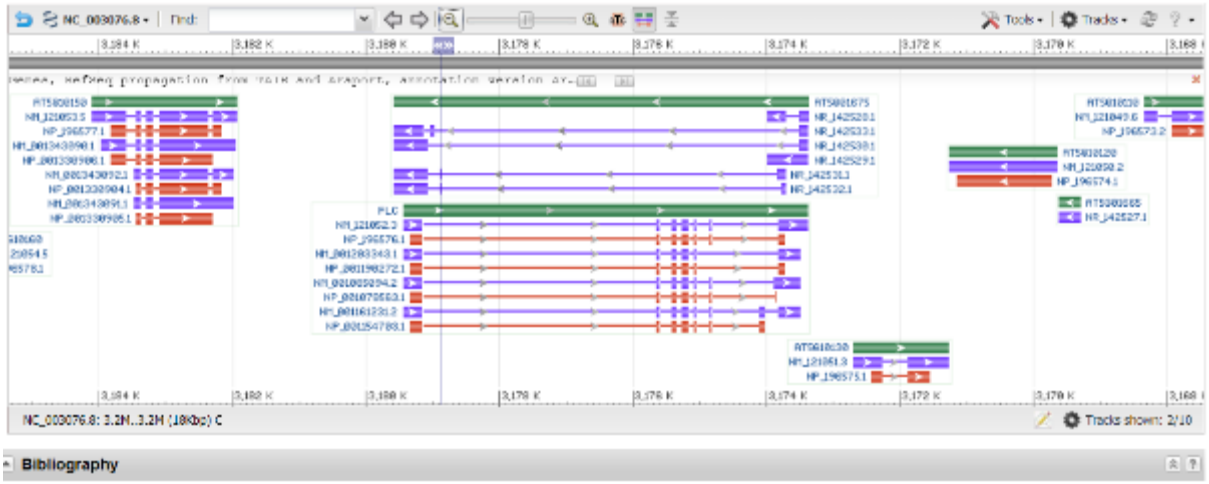


Figure 6 : Les séquences annotées du gène Flowering Locus C codant pour la protéine MADS-box (floraison) de l'espèce Arabidopsis thaliana (GenBank, 2018)

3. LE RÉSEAU DE PLATEFORMES EN FRANCE

Face à la croissance exponentielle des banques de données permise par la vitesse accrue du séquençage, en France est créé à la fin des années 1990 le réseau national de pôles de génomique pour mettre en place des plateformes dans les régions afin de mutualiser des équipements coûteux pour l'acquisition de données sur le vivant (Gaspin, 2015).

Les programmes d'investissement d'avenir (PIA) de l'État français sont initiés en 2010 et sont toujours en cours en 2018. Les domaines concernés sont, entre autres, l'enseignement supérieur et la formation professionnelle, le développement durable et les secteurs d'avenir (numérique, biotechnologies), la recherche et l'innovation. Ces programmes permettent de financer de grands projets innovants tels que France Génomique.

Encadré 4 : les programmes d'investissement d'avenir (PIA)

France Génomique est une infrastructure créée grâce à un financement « Investissement d'Avenir » dans le but de coordonner et renforcer les capacités françaises dans le domaine de la génomique à haut débit et de la bioinformatique associée en termes d'équipements, de compétences et d'innovation. Elle constitue le cadre de référence pour comprendre l'organisation des activités de génomique sur le territoire français[23]. Elle réunit la majorité des plateformes de séquençage et de bioinformatique. En termes de capacités, France Génomique intègre les analyses génomique (séquençage/génotypage) et le traitement bioinformatique. Les services proposés par l'infrastructure sont, entre autres :

- L'accès au réseau de plateformes de séquençage et/ou bioinformatique opérationnelles ;
- Des appels d'offres annuels pour des projets à forte visibilité ;
- L'accès et la dissémination des expertises et des innovations liées aux nouvelles méthodologies et technologies de séquençage et de bioinformatique ;

[23] Les partenaires du Consortium sont le CEA (coordinateur), l'Inra, le CNRS, l'INSERM, l'Institut Pasteur, la Fondation Paris Sciences & Lettres et le CERBM-GIE de Strasbourg.

- L'accès au Très grand centre de calcul (TGCC) du Commissariat à l'énergie atomique et aux énergies alternatives (CEA).

Dans le domaine de l'agriculture et de l'alimentation, il existe plus de vingt pôles de génomique et de bioinformatique regroupant des équipements et des compétences complémentaires pour les projets de recherche (cf. annexe 10). Les principaux sont situés en Ile de France (Genoscope d'Evry), à Toulouse (GenoToul Bioinfo, Get-PlaGe), à Montpellier (MGX) et à Rennes (GenOuest) (cf. annexe 11). Le Genoscope d'Evry représente 2400 emplois à lui seul. Ces différents pôles regroupent chacun des dizaines de plateformes inter organismes. Chaque organisme est responsable d'au moins une plateforme au sein des pôles.

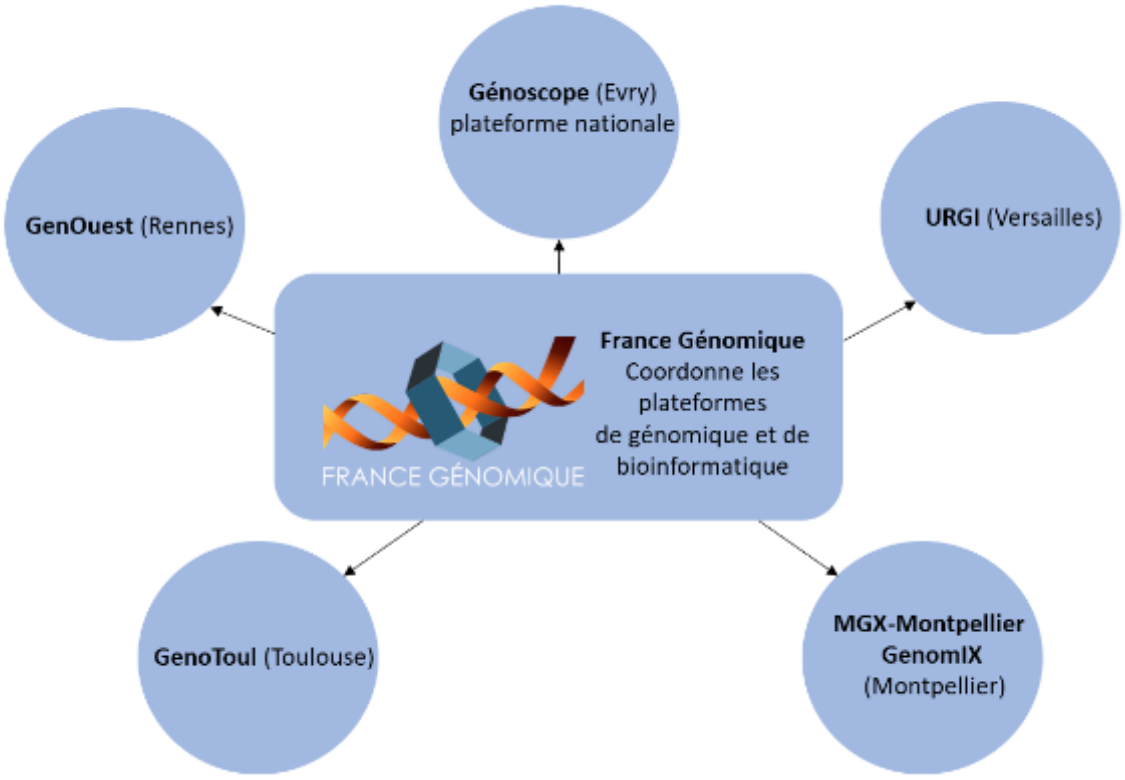


Figure 7 : Les principaux pôles de génomique pour l'alimentation et l'agriculture en France dans le cadre de France Génomique

Un des enjeux liés à la production des données de séquençage est la capacité de stockage. Aujourd'hui, les capacités de production de données sont supérieures aux capacités de stockage des données.

Les plateformes de bioinformatique sont en charge de stocker les données et éventuellement de réaliser leurs traitements. Aussi un contrôle qualité des données est réalisé par ces dernières à la sortie des séquenceurs.

En décembre 2011, Le NCBI a alerté la communauté scientifique sur le problème posé par la gestion des données de séquençage en indiquant qu'elle ne prendrait en charge que des données de séquençage d'espèces modèles[24] (arabette des dames, riz, tomate, maïs, etc.). De même, si l'institut européen de bioinformatique (EBI) continue d'accepter les dépôts de de toute nature, l'énorme volume généré devient un facteur limitant leur partage et limitant le stockage des données. Ainsi, les bases de données internationales

[24] Genome Biol. 2011; 12(3): 402. Published online 2011 Mar 22. doi: 10.1186/gb-2011-12-3-402 PMCID: PMC3129670 PMID: 21418618 Closure of the NCBI SRA and implications for the long-term future of genomics data storage. Une espèce modèle est fait l'objet d'études approfondies pour comprendre certains phénomènes biologiques notables, afin de mieux comprendre le fonctionnement d'autres organismes.

sont encombrées et le dépôt de données nécessite un temps d'attente. « Les entrepôts sont victimes de leur succès » (Inra, 2014).

Encadré 5 : Les enjeux de stockage des données

C. QUE RECOUVRE LE TERME DE « DONNÉES DE SEQUENÇAG » OU « INFORMATION NUMÉRIQUE DE DONNÉES DE SÉQUENÇAGE » ?

Un des principaux résultats issus des enquêtes menées dans le cadre de cette étude est que le concept de « données de séquençage » ou d'« information numérique de données de séquençage » ne fait pas consensus tant au niveau de la communauté internationale qu'au niveau national. Des typologies de données de séquençage sont proposées afin de mieux appréhender ce concept.

« Données de séquençage » ou « information numérique de séquençage » :

La notion de « données de séquençage » ou d'« information de données de séquençage numérique » n'a pas été définie ni officiellement, ni scientifiquement. La CDB utilise le terme « *digital sequence information* » mais de nombreux termes sont employés dans la communauté scientifique pour s'y référer[25] : données *in sillico*, données de séquence, information de séquence génétique, données de séquence numérique, information génétique, ressources génétiques dématérialisées, information de séquence d'acides nucléiques, etc. Dans cette étude, les experts se sont penchés sur la terminologie à adopter et concluent qu'il est préférable d'utiliser le terme « **données numériques de séquence de ressources génétiques** » à la place « d'information de séquençage numérique » utilisé dans les instances internationales (cf. annexe 12).

Les principaux éléments de divergences concernent le caractère matériel ou non des données de séquençage, son caractère autonome ou dépendant à la ressource génétique associée. Plusieurs experts ont insisté pour dire que les données de séquençage sont un produit de la recherche, et une représentation intellectuelle de la ressource génétique.

Le caractère polysémique du terme « information numérique de séquençage » a été rappelé lors de l'enquête. En effet, selon les spécialités des utilisateurs, ce terme revêt des sens différents. Par exemple, pour un informaticien, une séquence de lettre A, T, G, C constitue une information, tandis que pour un biologiste, cette même séquence n'est pas une information, tant qu'elle n'a pas été analysée.

Certains experts ont admis que le concept de « données de séquençage » est dynamique et qu'il peut s'étendre à de nouvelles données (métabolomiques, épigénétiques) en fonction des évolutions des techniques de séquençage.

Certains estiment que ce défaut de consensus sur une définition est une stratégie visant à ne pas légiférer sur le statut. La communauté scientifique a proposé de définir un périmètre et caractériser les différents types d'information de données de séquençage.

L'étude a permis de faire avancer les réflexions autour du terme à adopter pour se référer aux DSI en proposant des typologies auxquelles il serait possible de se référer.

1. LA NATURE DES DONNÉES

Les données issues des travaux de génomique, c'est-à-dire de la discipline réunissant les différentes techniques visant l'étude de l'information génétique, sont de différentes natures (Gallezot, 2002). Les données factuelles se matérialisent sous la forme de suites de nucléotides (A, T, C, G, U) auxquelles sont associées des annotations renseignées par les dépositaires des séquences. Les données textuelles, elles, sont associées aux publications scientifiques.

1) **Les données factuelles**, ou représentations de séquences nucléotidiques, sont issues des expérimentations (« paillasse ») ou des banques de séquences internationales. Les principales banques publiques de données sont GenBank (maintenue par le NCBI), EMBL ou DDBJ pour les séquences d'ADN et participent à une collaboration internationale de base de données de séquences de nucléotides (INSDC). Dans

[25] Rapport d'étude de l'AHTEG sur l'information de séquence numérique de ressources génétiques dans le contexte de la CDB et du Protocole de Nagoya, 12 janvier 2018 (CBD/DSI/AHTEG/2018/1/3).

ces banques, les fichiers sont codés en ASCII (*American Standard Code for Information Interchange*) et sont dits « à plat » (*flat file*) pour leur caractère brut fournis sans outil d'organisation. Leur nomenclature de description suit une notice précise standardisée. Les dépositaires doivent donc spécifier des champs pour l'enregistrement de leurs données :

- L'identité biologique, sorte d'état civil : nom, type de molécule, affiliation biologique, date de son entrée (champ LOCUS), numéro d'accès (champ ACCESSION) comme identificateur de l'enregistrement dans la banque, définitions et mots-clefs, origine (SOURCE), etc. (cf. figure 8) ;

- Les références bibliographiques relative à la production de la séquence ;

- Les propriétés de la séquence (champ FEATURES) : annotations qui décrivent la séquence, soit les fonctions des sous-séquences ainsi que leur position et attributs spécifiques (cf. figure 9) ;

- Le texte de la séquence d'ADN (champ ORIGINE) : représentation de la séquence nucléotidique (symboles ATGC) (cf. figure 10).

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
```

Figure 8 : Exemple d'enregistrement des champs LOCUS, ACCESSION et REFERENCE d'un échantillon annoté en format flat file dans la base de données GenBank

```
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
                     VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
```

Figure 9 : Exemple d'enregistrement du champ FEATURE d'un échantillon annoté en format flat file dans la base de données GenBank

```
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
      601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
      661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
      721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
      781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
      841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
      901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct tctgacttac
      961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg gactctgccg
```

Figure 10 : Exemple d'enregistrement du champ ORIGIN d'un échantillon annoté en format flat file dans la base de données GenBank

Cette notice évolue en fonction des connaissances mais garantit une utilisation pérenne des documents. Il existe un système d'harmonisation du champ FEATURES entre les trois banques en cas de doublon. Il existe cependant des erreurs liées à la saisie des informations et des redondances qui altèrent la qualité des données.

Il n'existe pas de restriction quant à l'utilisation ou à la distribution des données GenBank. Cependant, certains auteurs peuvent revendiquer un brevet, un droit d'auteur ou d'autres droits de propriété intellectuelle sur les données qu'ils ont soumis.

2) **Les données textuelles** renvoient à la littérature au sens large du terme (articles de revue, ouvrages scientifiques, actes de colloque, etc.) qui se base sur l'utilisation de ces données, et où ces dernières sont analysées, interprétées et discutées. Leur description suit une notice dite catalographique (auteurs, titre, résumé, revue, date, etc.) permettant leur référencement et facilitant leur diffusion. Le développement en bioinformatique utilise cette nomenclature pour extraire des informations de façon automatique depuis les

bases de données textuelles. En effet, les connaissances biologiques y sont souvent mieux décrites dans les articles scientifiques que dans les banques. Ceci renforce l'enjeu d'exploitation des informations présentes dans ces bases de données.

2. LES TYPOLOGIES PROPOSÉES

a. Entretien avec Pierre Mournet, responsable du plateau de génotypage de Montpellier[26]

La biologie moléculaire identifie des types de séquence qui peuvent être classées en fonction des technologies mobilisées :

- le séquençage Sanger bas-débit (1re génération) de moins en moins utilisé ;
- le séquençage nouvelle génération très haut-débit (NGS) :
    - le séquençage NGS de 2e génération (développé par la société Illumina) de petits fragments d'une centaine de paires de base[27].
    - depuis deux à trois ans développement de méthode du type de séquençage longs fragments avec de nouvelles sociétés qui ont commercialisé des nouvelles machines (PacBio et Nano pore, 3e génération). Elles permettent de récupérer le séquençage complet des ressources.

Elles peuvent aussi être identifiées par les types de séquences utilisés en biologie : des séquences de génomes entièrement ou partiellement séquencés, séquences d'ADN, d'ARN, de protéines.

Enfin, ces séquences peuvent être classées en fonction de leur degré de traitement, des séquences brutes, obtenues en sortie du séquenceur, aux séquences entièrement « annotées » (Varchney *et al.*, 2014 in Krager, 2018). Ces annotations peuvent fournir différents types d'informations, telles que la fonction des gènes, les mutations, la relation entre les génotypes et les phénotypes etc. Les bases de données incluent des informations liées à l'origine de l'échantillon (date et lieu de collecte). Une des difficultés est que toutes les données publiées et rendues publiques ne sont pas forcément accompagnées de ces « informations sur les données », également appelées « métadonnées ».

b. Entretien avec Arnaud Lemainque, Chef du laboratoire de séquençage du Genoscope[28]

On peut retenir une typologie plus précise des données de séquençage. Premièrement, les données brutes sont les données issues des séquenceurs mais ne sont pas conservées. Deuxièmement, les fichiers texte en format fast.q communément appelés « séquence », sont les données « propres » ou « nettoyées » par des procédés informatiques[29]. Enfin, l'assemblage[30] permet l'obtention de nouveaux fichiers texte. Cette typologie suit le chainage des programmes qui constitue le protocole bioinformatique (***pipeline***) pour le traitement qualité et l'analyse des données issue d'un séquençage haut débit. Les types de fichiers générés au cours des étapes du protocole sont différents en termes de volume et d'utilité pour l'utilisateur[31].

---

[26] Entretien avec Pierre Mournet, responsable du plateau de génotypage, 18 juin 2018.

[27] Appariement de deux bases nucléiques sur deux brins complémentaires d'ADN et d'ARN. C'est l'unité de référence utilisée par les scientifiques pour décrire la longueur des fragments séquencés (pb, kb, Gb, Mg, Tb).

[28] Entretien avec Arnaud Lemainque, Chef du laboratoire de Séquençage du Genoscope, août 2018.

[29] Le fichier FASTQ (extension. fastq ou .fq) est un fichier texte standard pour l'échange de données de séquence et de qualité utilisé pour tout type de séquenceur y compris Sanger. Il contient les noms des séquences, les séquences et la valeur de qualité des nucléotides. Ce fichier texte doit être compressé pour le stockage.

[30] L'assemblage est l'étape d'alignement ou de fusion des fragments de séquence permettant de reconstruire la séquence originale. Il peut être comparé à la reconstruction du texte d'un livre à partir de plusieurs copies de celui-ci, préalablement déchiquetées en petits morceaux (Rayan Chikhi, 2012, in Leleux, 2014).

[31] Groupe de travail « Cahier des charges informatique, bio-analyse/bioinformatique, bases de données mutations » dans le cadre du Réseau NGS Diagnostic, Recommandations générales pour la gestion informatique des données et des analyses de séquençage à haut débit pour les laboratoires de diagnostic moléculaire de maladies génétiques, mai 2016.

Figure 11 : Pipeline ou protocole bioinformatique pour le traitement des données issues du séquenceur

```
1  > 1801005_brs1541 -- unclipped
2  GGGCTGGCTCTTGTTACGACTTCACCCCAGTCATCTGTCCTGCCTTAGACGGCTCCCTCC
3  CGAAGGTTAGGCCACCGGCTTTGGGCATTACAAACTCCCATGGTGTGACGGGCGGTGTGT
4  ACAAGACCCGGGAACGTATTCACCGCGGCGTGCTGATCCGCGATTACTAGCGATTCCGAC
5  TTCGTGCAGTCGAGTTGCAGACTGCAGTCCGAACTGAGACGTACTTTAAGAGATTAGCTC
6  ACCTTCGCAGGTTGGCAACTCGTTGTATACGCCATTGTAGCACGTGTGTAGCCCAGGTCA
7  TAAGGGGCATGATGATCTGACGTCGTCCCCGCCTTCCTCCGGTTTGTCACCGGCAGTCTC
8  GCTAGAGTGCCCAACTGAATGCTGGCAACTAACAATAAGGGTTGCGCTCGTTGCGGGACT
9  TAACCCAACATCTCACGACACGAGCTGACGACGACCATGCACCACCTGTCACTTTGTCTC
```

Figure 12 : Visualisation d'une séquence de données brutes de ressources génétiques

```
>1801005_brs1541 -- 11..1024 of sequence
TTGTTACGACTTCACCCCAGTCATCTGTCCTGCCTTAGACGGCTCCCTCCCGAAGGTTAG
GCCACCGGCTTTGGGCATTACAAACTCCCATGGTGTGACGGGCGGTGTGTACAAGACCCG
GGAACGTATTCACCGCGGCGTGCTGATCCGCGATTACTAGCGATTCCGACTTCGTGCAGT
CGAGTTGCAGACTGCAGTCCGAACTGAGACGTACTTTAAGAGATTAGCTCACCTTCGCAG
GTTGGCAACTCGTTGTATACGCCATTGTAGCACGTGTGTAGCCCAGGTCATAAGGGGCAT
GATGATCTGACGTCGTCCCCGCCTTCCTCCGGTTTGTCACCGGCAGTCTCGCTAGAGTGC
CCAACTGAATGCTGGCAACTAACAATAAGGGTTGCGCTCGTTGCGGGACTTAACCCAACA
```

Figure 13 : Visualisation d'une séquence de données nettoyées de ressources génétiques

```
>Chr1 Chr1:3631..5899 (+ strand) class=mRNA length=2269
AAATTATTAGATATACCAAACCAGAGAAAACAAATACATAATCGGAGAAATACAGATTACAGAG
AGCGAGAGAGATCGACGGCGAAGCTCTTTACCCGGAAACCATTGAAATCGGACGGTTTAGTGAA
AATGGAGGATCAAGTTGGGTTTGGGTTCCGTCCGAACGACGAGGAGCTCGTTGGTCACTATCTC
CGTAACAAAATCGAAGGAAACACTAGCCGCGACGTTGAAGTAGCCATCAGCGAGGTCAACATCT
GTAGCTACGATCCTTGGAACTTGCGCTGTAAGTTCCGAATTTTCTGAATTTCATTTGCAAGTAA
```

Figure 14 : Visualisation d'une séquence de données annotées de ressources génétiques

Dans ce cas, trois catégories de formats sont distinguées, les données brutes (directement issues du séquenceur), les données primaires (séquences nettoyées) et les données secondaires (analysées).

c. Les typologies de données de l'Inra :

Dans le cadre du séminaire sur les données de la recherche, méthodes et outils pour l'open data, l'Inra propose le schéma suivant pour caractériser l'état des données. Elle identifie les données brutes qui après traitement deviennent les données dites « curées », puis les données que l'on retrouve « dans les publications », analysées pour produire de la connaissance. Cette typologie simple est proche de celle de A. Lemainque.

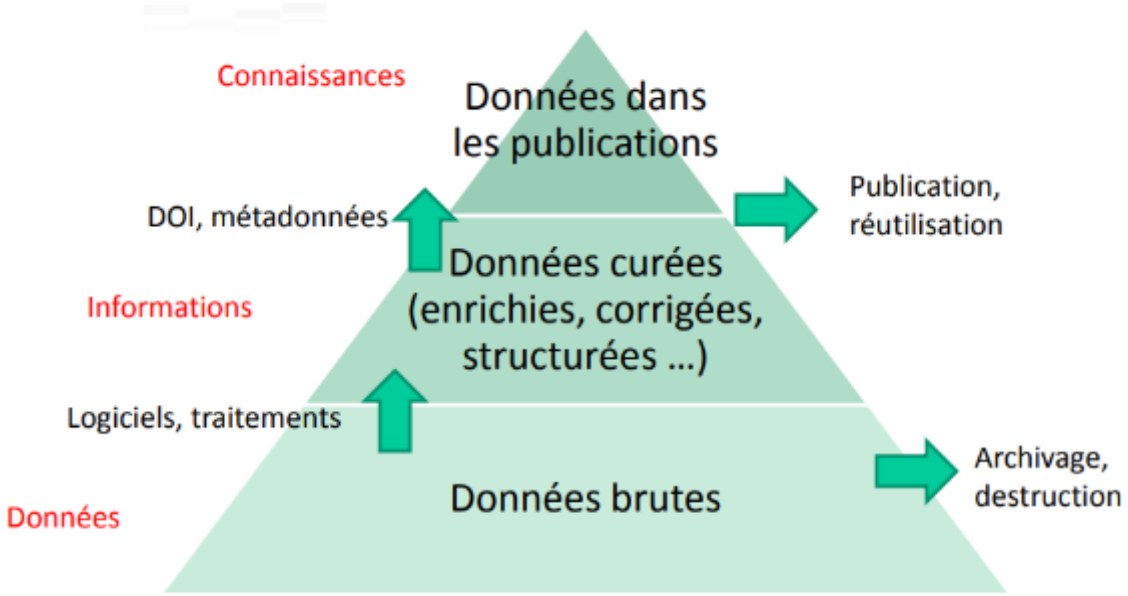


Figure 15 : Les états de la donnée au sens large
(schéma issu de la présentation du séminaire méthodes et outils pour l'Open data)

D'une façon plus détaillée, l'Inra propose aussi une typologie plus complexe composée de 19 types (6 types de données brutes, 13 types de données élaborées)[32]. Elle définit la donnée brute comme la donnée obtenue directement en sortie d'un équipement, tandis que la donnée élaborée est issue d'une analyse de la donnée brute. Les types sont définis selon l'élément biologique séquencé (nature) (cf. annexe 13).

Cette typologie recouvre la diversité des types de données directement issues du séquenceur. Il peut s'agir de données issues d'ADN, d'ARN, de protéines, de métabolome[33], de variation génétique (SNP, SSR), d'ARNm. Une fois analysées, ces données conduisent à davantage de types de données dites « élaborées » (séquences de protéines, données de polymorphismes[34], à des données passeport[35], etc.). Il est donc possible en fonction des outils et analyses possibles d'obtenir plusieurs données élaborées à partir d'un même type de donnée brute.

La production croissante des données de séquençage permise par Les nouvelles pose le problème de leur traitement. La communauté scientifique s'organise au niveau européen et français pour harmoniser les bases de données entre elles et faciliter l'accès et l'utilisation des données.

D. L'ACCÈS AUX DONNÉES : LES ENJEUX D'INTEROPÉRABILITÉ

1. L'ORGANISATION DE L'INTEROPÉRABILITÉ DES BASES DE DONNÉES

La multiplication des bases de données de séquençage de ressources génétiques pose la question de leur interopérabilité et celle de la standardisation des systèmes d'information. L'interopérabilité des bases de données, nécessaire pour la circulation des informations d'un système à l'autre, dépend de plusieurs aspects : les formats de renseignement des données et des métadonnées, ceux d'échange entre systèmes d'information, les référentiels utilisés, les formats d'identifiant, les langues, les versions des fichiers, ou encore les droits associés.

[32] Le groupe de travail « Partage des données relatives aux ressources génétiques et génomiques : états des lieux, analyse stratégique et besoins d'accompagnement » de l'INRA. Ce groupe de travail a été mis en place pour répondre aux limites du modèle de stockage centralisé des données (INSDC) et propose un système de stockage distribué.

[33] Le métabolome est l'ensemble des métabolites retrouvés dans un échantillon biologique. Ces métabolites renferment une diversité de molécules : nucléotide, vitamine, hormone, énergétique.

[34] C'est-à-dire de n'importe quelle différence entre les individus.

[35] Origines, généalogies.

Ces questions sont traitées dans le projet européen ELIXIR et le GnpIS en France, ou encore par la plateforme Southgreen à une échelle locale.

L'Institut français de bio-informatique (IFB) est l'infrastructure nationale de service en bio-informatique. Elle a été créée dans le cadre du programme national des « Investissements d'avenir » (voir encadré 4) pour mutualiser, soutenir et coordonner le développement des ressources et activités dépendant d'organismes publics de recherche, universités, Instituts Pasteur et Curie. L'IFB est le nœud français du programme européen ELIXIR[36]. Ses missions sont de soutenir la recherche publique et privée par l'accès à des services, l'accompagnement, la possibilité de participer à des projets ambitieux nationaux et internationaux. Parmi les services proposés, l'IFB coordonne le développement de bases de données thématiques et des bases de données publiques de référence pour veiller à l'interopérabilité de leurs systèmes d'informations.

À titre d'exemple, l'Unité de Recherche génomique info (URGI) de l'Inra a été créée en 2001 pour venir en support au GIS Genoplante[37] (aujourd'hui le GIS biotechnologies vertes) et, gérer les données des séquenceurs. L'URGI assure la gestion du système d'information et d'intégration des données ainsi que l'analyse des génomes. Le système d'information est mis en œuvre grâce à un ensemble de serveurs de l'Inra Versailles-Grignon où sont stockées les données : le GnpIS qui vise l'interopérabilité des bases de données pour une meilleure utilisation. Le GnpIS assure l'accès aux données génétiques[38] et génomiques[39] pour des espèces d'intérêt agronomique (cf. annexe 14). Il gère d'une part les données issues des travaux scientifiques de l'Inra et d'autre part, constitue un système ouvert sur l'extérieur à travers des projets collaboratifs de très grande envergure. Il est associé à des projets nationaux « Investissements d'avenir » pour les espèces blé, maïs, pois, betterave, miscanthus, sorgho, colza et pour le phénotypage haut-débit (Infrastructure Phenome). À l'international, le GnpIS et les équipes qui l'animent sont engagés dans les consortiums « blé » (*International Wheat Genome Sequencing Consortium, Wheat Initiative*) et « vigne » (*International Grape Genome Program*) et participent au groupe de travail réunissant l'Union européenne et les Etats-Unis sur la recherche en biotechnologie dans le domaine de la bioinformatique végétale (*EC-US Plant Biotechnology task force*) (Inra, 2015). Le volume de données géré par le GnpIS est estimé à 20 Téraoctets.

Un autre système d'information coordonné par l'IFB est MOSAIC, proposé par la plateforme bioinformatique MIGALE, dédiée aux génomes de bactéries.

La plateforme de bioinformatique South Green est un réseau local de scientifiques rassemblant des compétences en bioinformatique. Elle est située sur le campus de l'association Agropolis[40]. Celle-ci héberge des instituts de recherche tels que le Cirad, l'IRD, l'Inra, SupAgro, Bioversity international ainsi que le CGIAR. Basé sur cette forte communauté locale dans le domaine de l'agriculture, de l'alimentation, de la biodiversité et de l'environnement, ce réseau développe des applications et des ressources bioinformatiques dédiées à la génétique et à la génomique des plantes tropicales et méditerranéennes. South Green s'appuie sur les plates-formes techniques informatiques de ses instituts partenaires.

Enfin, le projet RGscope, volet « Ressources génétiques » de l'infrastructure nationale ECOSCOPE, coordonnée de 2012 à 2017 par la FRB, visait à structurer et renforcer les dispositifs d'observation sur les ressources génétiques. Dans ce cadre, un panorama des dispositifs dédiés aux ressources génétiques végétales et animales et des relations entre les systèmes d'information a été dressé en 2013[41]. Il décrit les efforts réalisés pour favoriser l'harmonisation entre ces outils et à l'international, mais soulève les manques d'homogénéité et de standardisation entre et à l'intérieur de ces systèmes (ignorance de l'existence de standards, besoins spécifiques, différences de notation entre observateurs, etc.). La multiplicité des acteurs

---

[36] Le projet ELIXIR est un programme H2020, https://www.elixir-europe.org/about-us/who-we-are/nodes/france

[37] Genoplante - programme fédérateur en génomique végétale - a été créé en 1999. Il associait la recherche publique (Inra, CNRS, Cirad, IRD) et des acteurs privés de l'amélioration des plantes (Biogemma, Sofiprotéol, Arvalis Institut du Végétal). Les partenaires privés Bayer CropScience et Bioplante ont également contribué aux phases 1 et 2 du programme. Après les phases 1 et 2 (1999 - 2004) soutenue par les Ministères de la Recherche et de l'Agriculture, Genoplante a bénéficié en 2005 du soutien de l'Agence Nationale de la Recherche (ANR).

[38] Les informations génétiques sont des cartes génétiques, locus quantitatifs, gènes d'association, marqueurs, polymorphismes, germoplasmes, phénotypes et génotypes.

[39] Les données génomiques sont les séquences génomiques et données d'expression.

[40] Agropolis rassemble une communauté scientifique dans les domaines agriculture, alimentation, biodiversité, environnement de 2 700 chercheurs et enseignants.

[41] Le projet RGscope constitue le volet « biodiversité domestique et sauvage apparentée » du projet ECOSCOPE, réseau des observatoires de recherche sur la biodiversité. Plus d'informations, http://ecoscope.fondationbiodiversite.fr/images/presentations/2013_RGscope_CIAg_VF.pd.

et des réseaux rend notamment le travail d'harmonisation des pratiques plus complexe. Cette infrastructure a évolué en Pôle national de données de biodiversité (PNDB). Elle interagit avec l'infrastructure Ressources Agronomiques pour la recherche (RARe) qui regroupe des centres de ressources biologiques (CRB) conservant les ressources génétiques/biologiques pour la recherche sur les animaux domestiques, les plantes modèles/cultivées, les espèces apparentées aux domestiques, les micro-organismes d'intérêt agronomique/agro-alimentaire, les micro-organismes/organismes de l'environnement.

2. LES ENJEUX DES DROITS SUR LES DONNÉES : DIFFÉRENTS DEGRÉS D'OUVERTURE DES DONNÉES

Différents cadres juridiques s'appliquent sur l'accès aux données en France. Dans cette partie nous ne ferons pas l'inventaire des cadres existants mais nous montrerons que le libre accès aux données est conditionné par les projets, comme l'illustre l'exemple de l'accord de consortium d'un projet du GIS Biotechnologies vertes.

a. Aperçu de la réglementation

Pour toute publication d'un article faisant référence à une ou plusieurs séquences de nucléotides dans une revue scientifique, il est obligatoire de déposer préalablement la ou les séquences dans une des bases de l'INSDC lorsque la recherche est financée sur fonds publics. En échange du dépôt dans une base de données, le chercheur reçoit un numéro d'identification unique qu'il intégrera dans sa publication, aujourd'hui le « *Digital Object Identifier* » (DOI).

La Commission européenne est très active en matière d'ouverture des données. « **Le programme H2020 [42] comporte notamment l'obligation d'assurer le libre accès aux publications** issues des recherches qu'il aura contribuées à financer, sous peine de sanctions financières »[43]. Les scientifiques sont encouragés à mettre en place un **Plan de gestion des données** en amont des projets décrivant notamment les modalités d'accès (restreint, libre, gestion, documentation, ...). En France, le Plan national pour la science ouverte, annoncé le 4 juillet 2018[44], rend obligatoire l'accès ouvert pour les publications et pour les données issues de recherches financées sur projets.

L'ouverture de données se met donc en place sous les impulsions concomitantes des États, des agences de financement et des organismes et communautés de recherche.

> « Un Data Management Plan (DMP) ou plan de gestion de données est un document formalisé explicitant la manière dont seront obtenues, documentées, analysées, disséminées et archivées les données produites au cours et à l'issue d'un processus ou d'un projet de recherche. Ce guide est conçu pour répondre aux exigences des financeurs, mais il est aussi un outil pour gérer les données tout au long du projet en intégrant la notion de cycle de vie. Le DMP s'étend donc de la production (ou la collecte) des données à leur diffusion et/ou leur archivage, en passant par leur stockage, leur traitement/curation, leur analyse et leur description. Le DMP s'appuie sur le cycle de vie des données/documents et décrit les choix réalisés en termes de métadonnées, formats des bases de données, méthodes et sécurité d'accès, durées d'archivage, ainsi que les coûts associés à la gestion des données. Une mention particulière doit être apportée aux données venant en appui des publications et qui doivent à ce titre rester disponibles et accessibles au plus grand nombre. L'établissement d'un data management plan est de plus en plus demandé dans les appels à projets financés sur fonds publics, notamment européens. [...] L'objectif est ainsi de documenter la manière dont les données seront produites ou collectées au cours et à l'issue d'un processus de recherche, en s'attachant notamment à définir comment elles seront décrites, partagées, protégées puis conservées. La gestion des données n'est pas une fin en soi, mais le moyen de conduire à la découverte de connaissances et d'innovations par l'intégration et la réutilisation des connaissances produites. »

[42] Horizon 2020 est un programme de financement de la recherche et de l'innovation de l'Union européenne pour la période 2014-2020.

[43] Site officiel du programme Horizon H2020 (http://www.horizon2020.gouv.fr/cid82025/le-libre-acces-aux-publications-aux-donnees-recherche.html).

[44] http://cache.media.enseignementsup-recherche.gouv.fr/file/Actus/67/2/PLAN_NATIONAL_SCIENCE_OUVERTE_978672.pdf

Source : Nathalie Reymonet, Magalie Moysan, Aurore Cartier, Renaud Délémontez. Réaliser un plan de gestion de données "FAIR " : modèle. 2018. ffsic_01690547v2f

b. Les avantages du libre accès aux données

Il existe de nombreux avantages liés à l'ouverture des données de la recherche[45] :

- Les scientifiques peuvent rapatrier des données tierces pour les exploiter dans le cadre de leurs recherches.
- Le partage permet une vision globale des objets étudiés, car la production de données concerne différentes disciplines comme la génomique mais aussi la protéomique ou le phénotypage[46].
- La publication et la mise en bases permettent une meilleure visibilité des travaux, les données référencées sont plus citées et donc davantage reconnues.
- L'accès aux données permet la reproductibilité des analyses et l'amélioration des méthodes dans un souci d'efficacité par la réutilisation des données.

Ces avantages sont synthétisés dans les principes FAIR (Trouvable, Accessible, Interopérable, Réutilisable).
L'accès libre aux données implique des rôles et responsabilités pour les différents acteurs. « Entrer dans le mouvement de l'*open science* nécessite pour les organismes de travailler sur les classes de données, la propriété intellectuelle, les aspects juridiques, les méthodes et outils, les compétences, les offres de service. Des contraintes et des compétences nouvelles pour le scientifique apparaissent. Il faut penser le cycle de vie des données et savoir construire un plan de gestion des données (formulaires) » (Christine Gaspin, directrice de recherche à l'UMR MIAT[47]).

c. Les inconvénients du libre accès aux données

Ces questions renvoient, entre autres, au débat sur le financement de la recherche et la valeur économique des résultats. Alors que la recherche publique doit publier les données, les entreprises privées ne sont pas soumises à cette obligation bien qu'elles bénéficient de ces données de séquençage. Qui plus est, les programmes de recherche en génomique sont généralement financés conjointement par les secteurs public et privé. Le retour sur investissement attendu par les secteurs publics et privés se traduit par un embargo sur les données et une réserve sur la diffusion des résultats d'analyse pour une exploitation optimale des résultats et leur publication dans des journaux scientifiques.

En effet, selon le groupe de l'Inra sur le partage des données, il subsiste des freins au partage qui s'expliquent par :

- la concurrence notamment avec les instituts étrangers,
- la concurrence entre équipes d'une même structure travaillant sur la même espèce,
- les habitudes de certains chercheurs qui ont du mal à diffuser leurs données
- même une fois la publication réalisée, l'accès aux données brutes, comme aux données élaborées, peut fournir une information stratégique aux entreprises concurrentes des partenaires du chercheur (en sélection génomique notamment) (Inra, 2014).

La FRB a réalisé une enquête[48] sur les données produites dans le cadre des observatoires de recherche sur la biodiversité relevant les freins évoqués par ces acteurs à l'ouverture de leurs données. On y retrouve ces arguments portant sur la propriété intellectuelle rattachée à ces données, mais d'autres types de freins sont

[45] Odile Hologne, Données de la recherche : données ouvertes ? Inra, opendatagropolis, 28 mars 2017.

[46] La protéomique est l'étude de l'ensemble des protéines (identification, fonction, structure). Le phénotypage consiste à observer des organismes et les caractères qu'ils expriment dans leur environnement.

[47] Unité mixte de recherche en mathématiques et informatique appliquées de Toulouse.

[48] Fondation pour la Recherche sur la Biodiversité (2016), État des lieux et analyse du paysage national des observatoires de recherche sur la biodiversité, une étude de l'infrastructure ECOSCOPE. Série FRB, Expertise et synthèse. Ed. Aurélie Delavaud et Robin Goffaux, 72 pp.

également cités : i) le manque de standards d'échanges ; ii) le manqué de contrôle sur leurs données et le risque mauvaises interprétations de leurs données; mais également très fréquemment iii) le temps nécessaire à consacrer.

d. Des règles d'accès variables au sein des partenariats dans les Programmes d'investissement d'avenir

Il existe une communauté publique-privée structurée autour des projets d'Investissement d'avenir réunissant entres autres des laboratoires de recherche publique et des entreprises semencières. La société Génoplante-Valor vient en appui aux activités du GIB Biotechnologies vertes (GIS BV) et est en charge de la propriété intellectuelle et de la valorisation. L'accord de consortium type, établi dans le cadre de Programme d'investissements d'avenir conduits par la communauté de recherche représentée au sein du GIS BV, différencie trois catégories de « Lots de travaux » présentés dans le tableau ci-après. Un lot de travaux est un ensemble cohérent de travaux exécuté par plusieurs parties en collaboration et répondant aux mêmes objectifs. Par exemple, pour le projet SUNRISE (cf. III. 2, *infra*), un lot de travaux répond à l'objectif de découverte de gènes et autres facteurs moléculaires impliqués dans les mécanismes de rendement en huile des tournesols hybrides sous contraintes hydrique.

Tableau 5 : Types de lots définis selon la part des financements des partenaires dans le cadre d'un projet du GIS BV, Informations issues de l'Accord de consortium SUNRISE.

Types de Lots	Part des financements des partenaires privés hors aide publique	Nombre de lots du projet
Lots de travaux **précompétitifs**	Financement <15% du coût total du Lot de travaux	1 lot
Lots de travaux **appliqués** avec financement substantiel	15% ≥ financement > 50% du coût total du Lot de travaux	4 Lots
Lots de travaux **appliqués** avec financement majoritaire	Financement ≥ 50% du coût total du Lot de travaux	1 lot + Exceptions pour le matériel végétal dérivé du matériel Inra
		1 lot qui exploite les résultats générés dans un autre lot appliqué substantiel (dépendant de 3 autres lots de cette catégorie).

La classification est définie par le Comité stratégique du GIS Biotechnologies Vertes et peut se retrouver dans d'autres projets. **Ce type d'accord illustre les conditions générales qui peuvent être établies pour l'ouverture des données.**

Selon le lot considéré, les règles relatives à la publication et la communication des résultats diffèrent. Un Comité de revue et de propriété intellectuelle du projet (CRPI) a été mis en place au sein du GIS BV. C'est à ce comité que revient la responsabilité de donner l'autorisation de publication et de communication immédiate si les travaux et les résultats ne sont pas susceptibles de faire l'objet d'un titre de propriété intellectuelle (PI) ou ne portent pas préjudice à une valorisation commerciale. Dans le cas contraire le CRPI s'assure que les résultats des travaux restent privés.

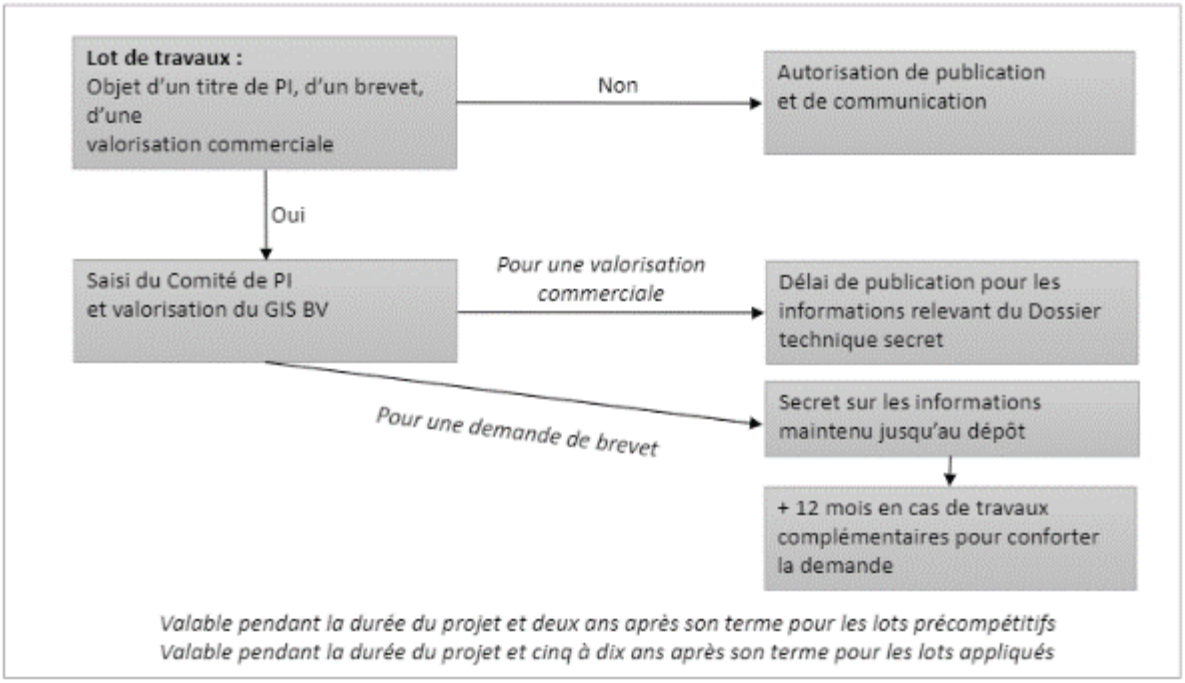


Figure 16 : Règlementation relative à la publication et la communication des résultats dans le cadre d'un accord de consortium défini par Genoplante-Valor

Les droits de propriété et d'accès diffèrent également selon le type de lots de travaux, résultats ou des matériels considérés (cf. tableau 5).

Les résultats issus de lots de travaux précompétitifs appartiennent aux partenaires publics. Les résultats de travaux de lots appliqués à financement majoritaires appartiennent aux partenaires privés. Dans le cas de lots de travaux appliqués à financement substantiel, c'est la société Genoplante-Valor qui est désignée propriétaire. Genoplante-Valor est une société commune à des organismes publics de recherche et des acteurs privés (Inra, Biogemma, etc.) associés. Elle vient à l'appui du GIS BV. C'est elle qui aide notamment à la rédaction et au suivi juridique de l'accord consortium du projet SUNRISE (cf. III. 2, *infra*), sur les aspects de diffusion des résultats entre autres. Genoplante-Valor détient et gère la propriété intellectuelle de projets appliqués dits « substantiels ».

Des exceptions sont prévues pour certains résultats, notamment pour le matériel végétal produit. Pour l'accès, les résultats sont rendus disponibles à tous les partenaires, gratuitement et de manière non-exclusive, à des fins de réalisation du projet. D'une manière générale, les résultats sont disponibles pour les partenaires, pour d'autres recherches interne ou en collaboration, selon des pas de temps tenant compte de leur implication dans le projet SUNRISE. S'il s'agit de finalités commerciales, l'accès aux résultats est payant et comprend des droits de sous-licence pour les affiliés des partenaires concernés.

Les négociations sur ce sujet visent à établir un équilibre entre promotion du libre accès aux données, la réglementation sur l'accès aux données et les droits de propriété intellectuelle et de brevetabilité (cf. annexe 15).

Le libre accès aux données promu par les réglementations françaises et européennes, notamment dans le cadre des projets H2020, peut présenter des situations conflictuelles vis à vis les droits de propriété intellectuelle et de brevetabilité qui existent.

III. LES UTILISATIONS DES DONNÉES DE SÉQUENÇAGE DE RESSOURCES GÉNÉTIQUES POUR L'ALIMENTATION ET L'AGRICULTURE : QUELLES RÉALITÉS, QUELLES PRATIQUES ?

Dans cette partie, après un propos liminaire, nous présenterons les différents acteurs et pratiques liés à l'utilisation de données de séquençage de ressources génétiques pour l'alimentation et l'agriculture. Pour chaque type de ressources génétiques considéré (animales, végétales, forestières, aquatiques, de microorganismes et d'invertébrés), des cas d'utilisations emblématiques seront présentés. Pour chaque illustration, seront indiqués les objectifs, les partenaires, les techniques mobilisées, les résultats et les modalités d'accès aux données.

A. LES TYPES D'UTILISATION DE DONNÉES DE SÉQUENÇAGE DE RESSOURCES GÉNÉTIQUES POUR L'ALIMENTATION ET L'AGRICULTURE

Les projets incluant du séquençage de microorganismes sont les plus nombreux en raison de la petite taille des génomes de microorganismes, inférieure à celle des végétaux ou des animaux et donc plus rapide à séquencer et à analyser. Néanmoins, avec les progrès réalisés dans les technologies de séquençage, les ressources génétiques d'animaux ou de végétaux au génome complexe font l'objet de nouveaux projets toujours plus nombreux.

Au travers des cas de figures rencontrés lors de ce travail plusieurs finalités ont motivé la production de données de séquençage de ressources génétiques : de la connaissance fondamentale de la composition et du fonctionnement du génome, de l'identification et l'exploration de la diversité génétique, à l'amélioration génétique par différentes techniques de sélection reposant sur l'étude de DSI.

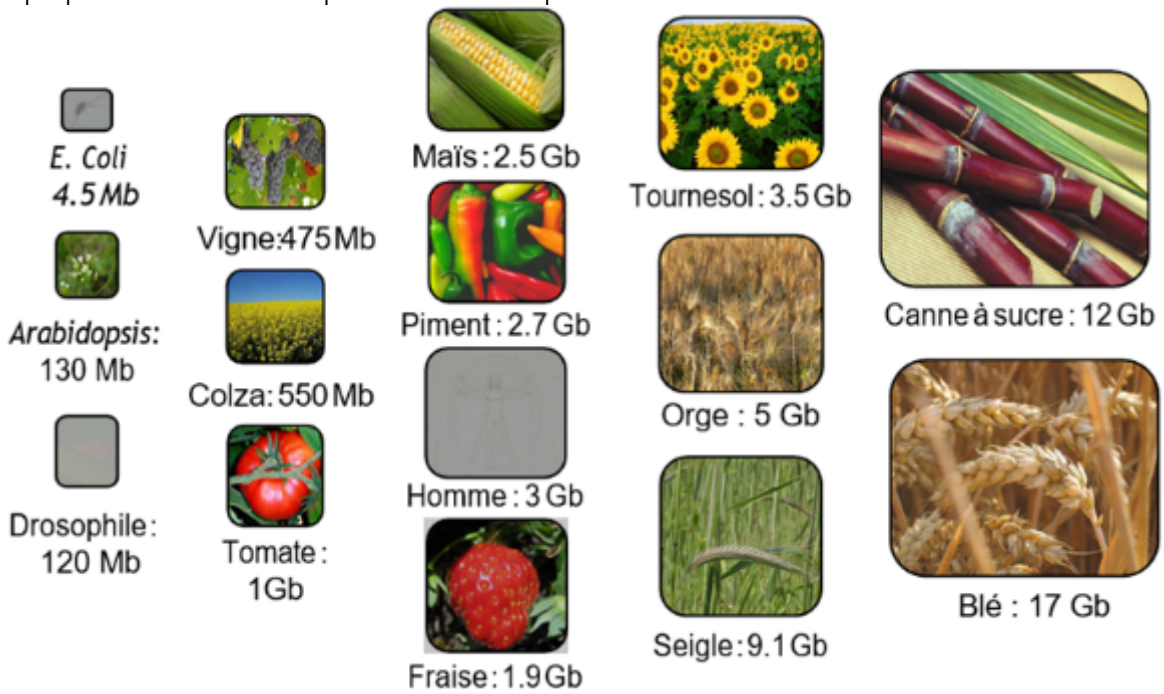


Figure 17 : La taille des génomes, Centre national de ressources génétiques végétales (CNRGV)

L'utilisation des données de séquençage de ressources génétiques pour l'alimentation et l'agriculture est fonction des applications commerciales. Notamment, pour ce qui concerne les ressources génétiques animales, les bovins ont traditionnellement bénéficié de programmes de sélection génétique. Par exemple, certains de ces programmes consistaient à définir des marqueurs génétiques responsables de maladie et d'effectuer une contre sélection des mâles grâce au séquençage du génome des animaux.

Ces programmes de sélection génétique s'étendent peu à peu à l'ensemble des RGAA. La découverte de marqueurs d'intérêt (sexe, résistance à un parasite, etc.) permet une sélection précoce des descendants. Par

exemple pour l'esturgeon, comprendre le déterminant majeur du sexe permet une sélection plus précoce des femelles produisant du caviar.

De manière générale, quel que soit le domaine, pour les espèces présentant des intérêts économiques moindres, des études de diversité génétique sont plus fréquentes. Elles mesurent le degré de variétés des gènes au sein d'une même espèce. La cartographie consiste à situer, le long des chromosomes, des séquences d'ADN connues. Elle permet une représentation d'un génome sous forme de balises. La sélection assistée par marqueurs permet le suivi des gènes et le tri précoce des ressources considérées après croisement naturel. Plus récemment, l'édition de génome est une utilisation récente des données de séquençage et permet la modification ciblée du génome. On la retrouve en grande majorité pour les ressources génétiques de microorganismes (bactéries) et en faible proportion chez les plantes. Cela s'explique notamment par la taille des génomes. Plus un génome est complexe, plus il est difficile de comprendre son fonctionnement.

Tableau 6 : Utilisations principales des données de séquençage de génomes de RG, typologie extraite des entretiens de la présente étude

Utilisations principales des données de séquençage de génomes de ressources génétiques :

- Analyse de la diversité allélique
- Mise au point des outils (puces à ADN) capables de suivre l'activation de gènes selon certaines conditions
- Caractérisation des ressources génétiques
- Détermination et Identification des gènes présents dans les zones du génome (cartographie génétique)
- Recherche des allèles les plus efficaces/intéressants
- Sélection assistée par marqueurs : choix des géniteurs, tri des descendants obtenus
- Clonage plus facile des gènes pour des possibilités de transformations génétiques mieux ciblées, plus efficaces utilisant des gènes spécifiques ou cherchant à éteindre ou désactiver l'expression de certains gènes indésirables.

Les entretiens menés au cours de l'enquête ont permis de rendre compte de diverses utilisations de données de séquençage de RGAA. Un tableau récapitulatif de ces exemples est disponible en annexe (cf. annexe 16).

B. DES EXEMPLES EMBLÉMATIQUES DE CAS D'UTILISATION DE DONNÉES DE SÉQUENÇAGE

1. L'UTILISATION DE DONNÉES DE SÉQUENÇAGE DE RESSOURCES ZOOGÉNÉTIQUES : CARACTÉRISATION GÉNÉTIQUE POUR LA CONSERVATION DES RACES LOCALES

Des acteurs incontournables :

En France, les différentes filières agroalimentaires sont organisées de façon à développer les activités de recherche et développement dans le domaine de la génomique. Les entreprises sont en général rattachées à une interprofession et un centre technique.

Pour la filière viande, l'Institut de l'élevage (Idele) a pour vocation d'améliorer la compétitivité des élevages d'herbivores et de leurs filières. Il assure une expertise dans plusieurs domaines, dont la génétique à travers des partenariats avec des unités mixtes technologiques (UMT) et des réseaux mixtes technologiques (RMT) créés par le ministère de l'agriculture pour renforcer les liens entre les acteurs des instituts techniques et de la recherche publique, l'Inra notamment.

Le Syndicat des sélectionneurs avicoles et aquacoles français (Sysaaf[49]) apporte un appui technique à une quarantaine d'entreprises de sélection et de gestion génétique concernant aujourd'hui plus d'une

[49] Le Syndicat des sélectionneurs avicoles et aquacoles français (SYSAAF) regroupe des entreprises de sélection développant des programmes de gestion et/ou d'amélioration génétique des espèces avicoles et aquacoles. Leurs actions visent à maîtriser la diversité génétique des populations en sélection tout en contribuant à préserver la qualité et la diversité de la gastronomie nationale issue de ces productions : huître creuse, filet de truite, truite fumée, turbot, bar, daurade, esturgeon sibérien, caviar d'esturgeon ou de truite, maigre, ombrine & crevette bleue dans les territoires ultramarins, perche commune, foie gras de canard ou d'oie, poulet, chapon, œuf de poule ou de caille, ovo produits, dinde fermière, pigeon, oie à rôtir, magret de canard ou d'oie, pintade, gibiers (faisan, perdrix, canard colvert).

trentaine d'espèces avicoles et aquacoles, dans le cadre d'une mission officielle qui s'inscrit dans le Programme génétique animale du Programme national de développement agricole et rural. En mutualisant des compétences et des moyens, il s'implique dans de nombreux programmes de recherche en collaborations avec des équipes de recherche de l'Inra, l'Ifremer, l'Anses, le CNRS et de nombreux laboratoires français et étrangers.

a. Le Projet BioDivA : caractérisation génétique pour la conservation de races locales avicoles

Figure 18 : Poule grise du Vercors (association Ouantia Grise du Vercors)

Objectif :

Les faibles effectifs observés chez les races traditionnelles locales et/ou anciennes françaises de poule présentent une menace pour leur diversité génétique et donc, à terme, pour leur existence. Afin de répondre ce problème, le projet BioDivA a pour ambition de caractériser la diversité génétique des races locales françaises de poules et de contribuer à favoriser la mise en place des programmes de conservation adaptés.

Les partenaires :

Le projet BioDivA, financé le ministère de l'agriculture et de l'alimentation et la Région Rhône-Alpes, a débuté en 2013 pour une durée de trois ans. Les partenaires de ce projet sont, entre autres, l'unité mixte de recherche Gabi (Génétique animale et biologie intégrative) de l'Inra, le Sysaaf, l'organisme de recherche appliquée indépendant et reconnu par les pouvoirs publics (Itavi), le Centre de sélection de Béchanne et le laboratoire d'analyse génétique animale de l'Inra, Labogena.

Aspects techniques :

Les analyses moléculaires se révèlent être l'outil idéal pour la caractérisation de la diversité génétique. Entre 2013 et 2016, 1 517 animaux ont été génotypés avec la puce Illumina Infinium 60K spécifique du poulet par la plateforme Labogena DNA, correspondant à 26 populations et révélant une grande diversité génétique au sein des races locales françaises (Restoux *et al.*, 2017). Cette étude de diversité constitue une étape avant la mise en place de programmes de préservation *in vivo* et/ou *in vitro* (Cryobanque nationale).

Résultats :

La caractérisation génétique des races menacées et la remontée des pedigrees en bases de données ont permis la mise en place de programmes de préservation adaptés (Chiron *et al.*, 2018). Le développement d'outils de gestion génétique adaptés aux races à petits effectifs permet entre autres le choix approprié des candidats reproducteurs au regard d'objectifs prédéfinis et d'établir les plans d'accouplement pour les races locales. Par exemple, le développement de logiciels permet au Sysaaf de proposer des listes de reproducteurs mâles et femelles pour la conservation de la diversité et la création de progrès génétique tout en maîtrisant la consanguinité. Les graphiques suivants montrent l'évolution génétique pour trois caractères clés (poids vif, poids des œufs et ponte) de la poule grise du Vercors (cf. figure 18A) et l'évolution de la consanguinité moyenne pour trois races locales (Bresse Blanche, Géline de Touraine, Noire de Berry) (cf. figure 18B). Le ralentissement de la progression du taux moyen de consanguinité pour la race Bresse Blanche révèle la prise de conscience de l'importance de la variabilité génétique pour une gestion durable (Chiron *et al.,* 2018). La race Bresse Blanche bénéficie de programmes de sélection depuis le début des années 1990, ce qui explique que la pente d'augmentation est moins importante que pour les races Géline de Touraine et Noire de Berry.

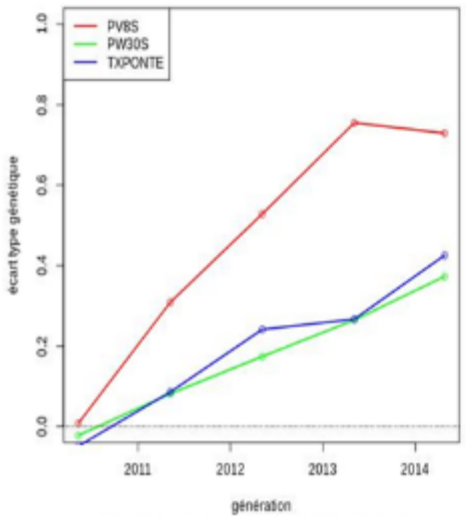


Figure A : Evolution génétique du poids vif (PV8S), du poids des œufs (PW30s) et du taux de ponte (Txponte) pour la race Poule Grise du Vercors

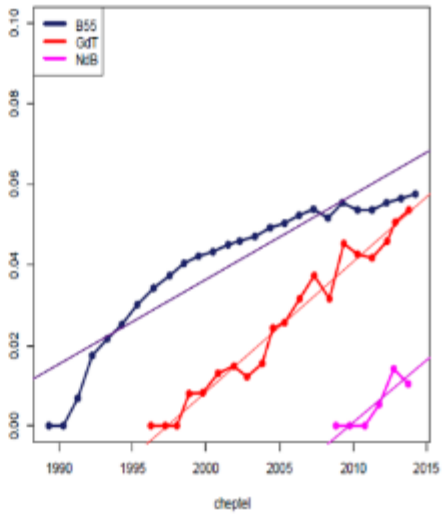


Figure B : Evolution de la consanguinité pour 3 races locales : Bresse Blanche (B55), Géline de Touraine (Gdt) et Noire du Berry (NdB)

Figure 19 : Évolution génétique pour la race Poule Grise du Vercors et évolution de la consanguinité pour trois races locales (Chinon et al*., 2018)*

Les données de sélection sont désormais toutes transférées, contrôlées, analysées au fur et à mesure permettant l'évaluation des lignées des adhérents au Sysaaf pour une pérennisation des noyaux de sélection sur le long terme et une variabilité au sein des cheptels.

Sur le transfert des données :

Un accord de transfert de données et établi entre l'Inra et l'entreprise de génétique Hendrix[50], ainsi que pour les associations propriétaires de races locales, pour l'acquisition de données issues des bases de données de ces derniers pour le projet. Ce contrat pose le cadre de l'utilisation des données transférées, ici pour la recherche et l'expérimentation. L'utilisateur n'a aucun droit de brevet, ou d'utilisation pour des finalités commerciales. Si le cas se présente, cela devra faire l'objet d'un contrat spécifique. Par ailleurs, toute publication utilisant la base de données partagées devra faire apparaître les deux parties. Le nom de l'entreprise ou de l'association sera cité comme source de données pour toute publication utilisant les données. Ce partage des données est limité dans le temps (trois ans). Au-delà, l'acquéreur devra émettre une demande de licence ou bien effacer les données acquises du fournisseur.

b. Le Projet ADAPTmap : caractérisation de la diversité génétique caprine mondiale

Figure 20 : Chèvres créoles. © Inra, NIORE Jacqueline

[50] Entreprise dont les filiales sont spécialisées dans la sélection génétique des différentes espèces domestiques : poules pondeuses, poulets à croissance lente, truites, dindes, porcs, volailles, truites, saumons et insectes.

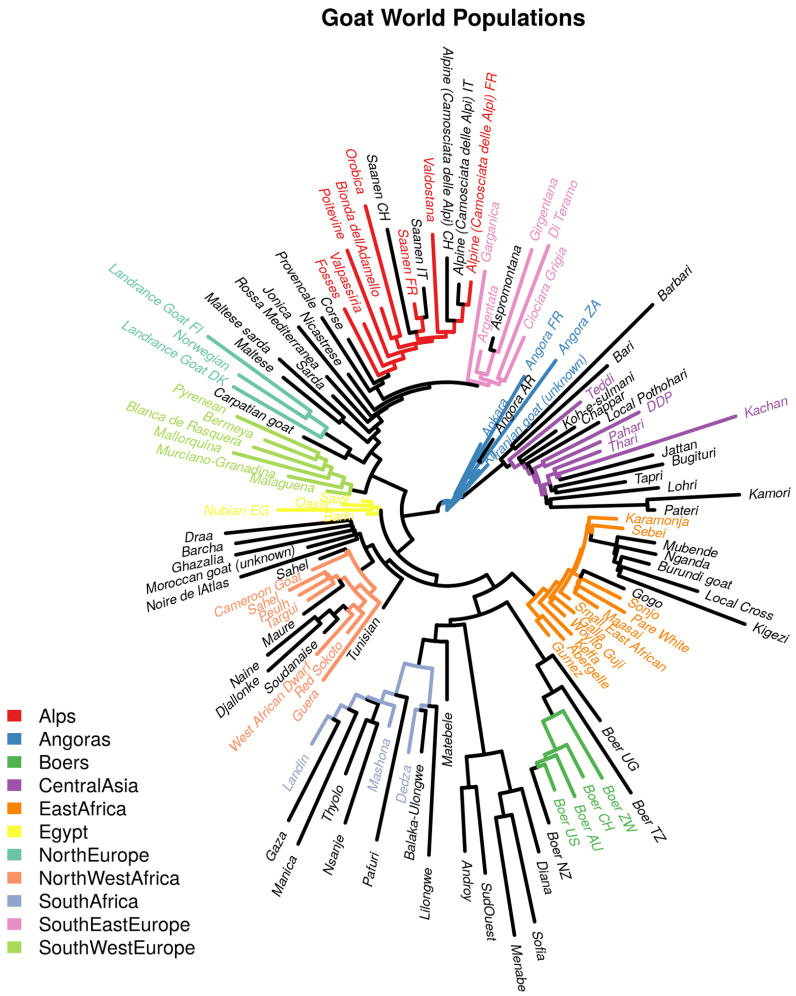


Figure 21 : Phylogénie des races caprines mondiales (tiré de Bertolini et al.*, 2018)*

Contexte :

Les chèvres sont élevées dans le monde entier dans une grande variété d'environnements géographiques, climatiques et de systèmes d'élevages. Certaines races locales permettent de valoriser des zones difficiles, tant dans les pays développés que dans ceux en développement. Domestiquées il y a environ 10 000 ans dans une seule région, le Croissant fertile, elles ont été soumises à une hybridation limitée entre les races. Elles représentent donc l'une des meilleures espèces pour l'étude de la diversité génétique et de l'adaptation.

Objectif :

L'arrivée des technologies de séquençage haut-débit a permis, en moins de 10 ans de produire de grands jeux de données permettant à la fois de développer des outils de génomique caprine, puis d'étudier les impacts des phénomènes adaptatifs. La compréhension des conséquences fonctionnelles des processus adaptatifs est un enjeu important pour la caractérisation des processus biologiques influençant la fitness (ou la valeur sélective[51]), la compréhension des potentiels adaptatifs des populations et contribuer à retracer les phénomènes historiques aboutissant à la diversité actuelle des populations.

Ainsi, au début des années 2010, de premières données de séquence ont permis de concevoir une puce SNP de moyenne densité (Tosser-Klopp *et al.*, 2014), qui a été utilisée largement au niveau international. Le projet ADAPTmap a démarré en 2014 dans le but de coordonner des projets internationaux variés ayant généré des données de génotypage et de phénotypage chez la chèvre (Stella *et al.*, 2018).

Les partenaires :

ADAPTmap a été soutenu par le consortium IGGC (*International Goat Genome Consortium*), coordonné par l'Inra, le réseau AGIN (*African Goat Improvement Network*) et les projets européens 3SR (*Sustainable Solutions for Small Ruminants*) et NextGen.

[51] La valeur sélective, ou *fitness*, d'un individu est mesuré par le nombre de descendants qui atteignent la maturité sexuelle.

Aspects techniques :

La puce Illumina GoatSNP50 (Tosser-Klopp *et al.*, 2014) a été génotypée sur des plate-formes différentes. Une base de données a permis de mettre à disposition du Consortium ADAPTmap les 4653 génotypages (148 populations, 35 pays) ainsi générés avec des informations de race, localisation géographique, phénotypes.

Résultats :

L'analyse des données caprines ADAPTmap a permis de déterminer la structuration génétique des chèvres dans le monde. Globalement, les populations étudiées se structurent selon leur origine géographique (Colli *et al.*, 2018, LIU *et al.*, 2018) : trois principaux groupes génétiques correspondent aux chèvres d'Europe, d'Afrique et d'Asie occidentale. Au sein de ces trois bassins, les modèles de variation sont compatibles avec les mesures de « distances » géographiques, l'histoire humaine et les pratiques d'élevage courantes.

Par exemple, une plus grande variabilité génétique est observée chez les populations proches du centre de domestication présumé, avec une différenciation relativement faible entre les races de ces régions. On note aussi que la structure génétique des populations reflète les principales voies de migration post-domestication. En outre, on observe des flux de gènes important dans des zones spécifiques (par exemple, l'Europe du Sud, le Maroc et le Mali-Burkina Faso-Nigéria), tandis que l'isolement géographique (mer, montagne) ou les pratiques d'élevage ont réduit ce flux dans d'autres zones du globe (Bertolini *et al.*, 2018a).

Ainsi, certaines populations iliennes (Canaries, Baléares, Islande) présentent des zones d'homozygotie spécifiques (Cardoso *et al.*, 2018).

Des signatures de sélection liées aux paramètres géographiques et climatiques ont également été identifiées, la plus forte d'entre elles étant associée à la température annuelle moyenne (Bertolini *et al.*, 2018b).

Enfin, un panel de SNP utilisable au niveau international a été mis au point pour permettre des études de filiation et l'assignation de parenté (Talenti *et al.*, 2018).

Perspectives :

ADAPTmap a maintenant terminé sa première phase. A la faveur de l'acquisition de nouvelles données de génotypage 50K, une deuxième phase est envisagée pour enrichir les résultats avec de nouvelles populations. Il s'agit d'une part de développer de nouvelles méthodes d'analyse visant à exploiter des données environnementales ou agronomiques sur les populations ou d'exploiter de nouveaux jeux de données issues d'expérience de paléogénomique fournissant des données temporelles riches en information. D'autre part, le projet France Génomique 1000-génomes caprins VarGoats, coordonné par l'Inra, s'appuie sur les résultats d'ADAPTmap pour son échantillonnage. Il permettra de revisiter et d'affiner ces études, à l'aide de données de séquence tout génome. En effet, les données SNP qui en découlent sont nettement plus denses que celles de la puce 50K. Par ailleurs, les variants de structure (et pas seulement les variations ponctuelles de type SNP) seront utilisés.

Gestion des données :

Le jeu de données du projet ADAPTmap est maintenant public (Stella *et al.*, 2018). Ce sera également le cas de toutes les séquences du projet VarGoats. Dans l'intervalle, le stockage de ces données et l'espace disque nécessaire aux analyses est un véritable enjeu.

2. LES UTILISATIONS DE DONNÉES DE SÉQUENÇAGE DE RESSOURCES GÉNÉTIQUES AQUATIQUES (RGA) : POUR UN SUIVI GÉNÉTIQUE DES ESPÈCES AQUACOLES

a. Le projet FishBoost : le développement d'outils de génotypage et l'utilisation de la génomique pour la sélection du bar et de la daurade

Contexte :

L'augmentation de la demande en protéines animales est corrélative à l'augmentation de la population mondiale. Les programmes de sélection permettent de contrôler et d'améliorer la production aquacole qui répond à cette demande. Deux entreprises françaises conduisent des programmes de sélection chez le bar et la daurade en France. Leur génétique étant performante, ces PME exportent également dans de nombreux pays situés sur le pourtour méditerranéen, mais aussi ailleurs dans le monde. Plusieurs programmes de

recherche successifs (Re-Sist (FUI[52]), FishBoost (FP7[53]), GeneSea (FEAMP[54]) et PerformFish (H2020) ont permis de développement d'outils de génotypage et l'utilisation de la génomique pour la sélection de ces deux espèces.

Objectif :

Le programme européen FishBoost a démarré en 2013 et a pour objectif d'améliorer l'efficacité et la rentabilité de l'aquaculture européenne pour six espèces de poissons, dont la dorade royale et le bar, grâce à une recherche en collaboration avec l'industrie. En particulier, un des objectifs est de découvrir l'architecture génétique de la résistance aux principales maladies pour lesquelles il n'existe pas de méthodes de prévention qui nécessite l'utilisation de traitement médicamenteux et occasionnent souvent d'importantes mortalités. Les programmes Re-Sist et PerformFish y contribuent également par la mise au point de nouvelles méthodes de phénotypage pour de nouveaux caractères. Plus récemment, le programme GeneSea, porté par l'Ifremer, utilise les résultats de ces précédents programmes pour le développement de la sélection génomique pour ces espèces.

Partenaires :

Le projet FishBoost réuni 14 partenaires européens, publics et privés dont pour les français l'Inra et l'Ifremer, des partenaires des secteurs recherche et développement privés (Labogena DNA du groupe Evolution, Sysaaf) et/ou professionnels avec les entreprises de sélection Ecloserie Marine de Gravelines et Ferme Marine du Douhet, toutes deux, adhérentes du Sysaaf. On retrouve peu ou prou les mêmes acteurs français dans les programmes Re-Sist, PerformFish et GeneSea ; auxquels il convient d'adjoindre l'Anses et le CNRS.

Techniques :

Deux axes ont été développés, la production de données, par le séquençage notamment, et la création d'outils de génotypage, de sélection assistée par marqueurs et de sélection génomique. Les puces de génotypages HD (Haute densité) destinées à la réalisation de la sélection génomique sont développées dans les programmes GeneSea et PerformFish.

Résultats :

Le développement d'une puce de génotypage mixte au bar et à la daurade est envisagé dans le programme PerformFish. Néanmoins, une puce de génotypage pour le bar a été créée dans le programme GeneSea et les premières utilisations pour de la sélection génomique sont en cours. Parallèlement, une puce de génotypage est en cours de création pour la daurade dans ce même programme GeneSea et devrait être opérationnelle début 2019.

Les programmes Re-Sist, FishBoost, PerformFish contribuent parallèlement à la mise au point de nouvelles méthodes de phénotypage pour de nouveaux caractères, permettant d'estimer l'héritabilité de ces caractères et ainsi d'apprécier s'ils sont sélectionnables ou pas. Dans la prévision d'une utilisation ultérieure pour constituer une population de référence pour laquelle nous avons l'information phénotypique et génomique, des échantillons d'ADN ont été stockés et seront analysés avec ces puces de génotypage.

La sélection génomique permet également de sélectionner les candidats plus précocement et potentiellement sur des caractères liés au sexe, comme la ponte chez la femelle, ou la qualité du sperme chez le mâle.

b. Le projet Vivaldi : contrôle de maladies affectant la filière conchylicole par le suivi épidémiologique des espèces

Contexte :

La conchyliculture européenne occupe une place privilégiée à l'échelle mondiale. La production européenne de coquillages repose principalement sur les moules, huîtres et palourdes. Ces dernières années, la filière a été fragilisée par des phénomènes de mortalités, associés à divers virus (ex. OsHV-1), bactéries (ex. *Vibrio aestuarianus*) et parasites (ex. *Marteilia cochillia*), qui entrainent de lourdes pertes économiques.

[52] Fonds unique interministériel à destination des pôles de compétitivité pour des projets de 5 ans.

[53] Septième programme-cadre de recherche et de développement technologique (2007-2013) de l'Union européenne.

[54] Fonds européen pour les affaires maritimes et la pêche.

Objectif :

Elevage d'huîtres ou ostréiculture

Le projet européen VIVALDI vise à augmenter la durabilité et la compétitivité du secteur conchylicole européen, qui regroupe les différentes cultures de coquillage, en développant des outils et approches pour mieux prévenir et contrôler les maladies d'une classe de mollusques marins : les bivalves[55]. Le projet a pour objet d'étude les ressources génétiques de mollusques et leurs pathogènes, extraites en Europe, Israël et Norvège. Pour répondre à ces besoins, VIVALDI doit apporter non seulement de nouvelles connaissances sur les interactions complexes entre coquillages, environnement et organismes pathogènes mais s'attachera aussi au développement d'outils et d'approches pratiques afin de mieux prévenir et contrôler les maladies affectant les bivalves marins. Comme les maladies ne connaissant pas de frontière, un réseau international rassemblant des experts des principaux pays producteurs de coquillages comme la Chine, le Japon, la Corée, l'Australie, la Nouvelle Zélande, les États-Unis et le Canada sera mis en place. Au cœur de ce réseau, VIVALDI contribuera ainsi à partager l'information et les expériences de chacun sur les mortalités de coquillages pour un meilleur contrôle des maladies associées.

Les partenaires :

Le projet VIVALDI est un projet européen d'Horizon2020 qui a démarré en 2016 pour 4 ans. Il réunit 21 partenaires publics et privés de dix pays. Pour répondre à certains objectifs, les partenaires du projet peuvent être amenés à explorer les ressources génétiques de mollusques et leurs pathogènes dans les pays concernés.

Les techniques :

Le génome complet des bivalves a été séquencé.

Les résultats :

De nombreux prélèvements ont déjà été effectués. Ces échantillons sont en cours d'analyse pour l'étude de la diversité des pathogènes affectant les mollusques bivalves. Un résultat novateur a montré qu'il est possible de détecter de l'ADN de virus dans les parcs à huîtres en utilisant des bandes de plastiques immergées qui jouent le rôle de capteurs (Ifremer, 2017).

Le cadre d'échanges des données :

Il existe un *Data Transfer Agreement* (DTA) pour l'échange de jeu de données entre les parties au projet. Par exemple, un DTA a été établi entre l'Institut français pour l'exploitation de la mer (Ifremer) et l'université de Galway, qui fixe les stipulations suivantes :

- Le DTA concerne les séquences d'ARN séquencées par l'Ifremer et partagées avec l'université irlandaise,
- Les types d'utilisation autorisées dans le cadre du projet sont l'étude de la variabilité génétique et des marqueurs d'intérêt.
- L'utilisation commerciale des données nécessite le consentement préalable de l'Ifremer. Les termes de la négociation entre les deux parties doivent faire l'objet d'un nouvel accord.
- Les résultats des recherches doivent être obligatoirement communiqués à l'Ifremer. En cas d'invention, les deux parties se réunissent pour décider d'un brevet ou d'une application.
- Le récepteur des données est prié de demander l'accord du fournisseur pour toute divulgation à une troisième partie.
- Lors d'une publication, le fournisseur de données doit être cité.

[55] Les bivalves marins sont une classe de mollusques d'eau de mer, nommée également *Pelecypoda* (les pélécypodes) ou *Lamellibranchia* (les lamellibranches). Cette classe comprend notamment les palourdes, les huîtres, les moules, les pétoncles et de nombreuses autres familles de coquillages.

3. LES UTILISATIONS DE DONNÉES DE SÉQUENÇAGE DE RESSOURCES PHYTOGÉNÉTIQUES (RPG) : DU SÉQUENÇAGE COMPLET DES GÉNOMES AU PROGRAMME D'AMÉLIORATION VARIÉTALE

Dans cette partie l'utilisation des données de séquençage de ressources phytogénétiques est illustré par plusieurs exemples (riz et tournesol notamment). L'intérêt pour ces ressources n'est pas récent et répond, entre autres, aux enjeux de sécurité alimentaire et de changement climatique.

Une espèce emblématique : l'arabette des dames

Dans le domaine végétal, les utilisations de données de séquençage remontent aux années 1990. L'arabette des dames (*Arabidopsis thaliana*) est la plante modèle choisie dans les premiers programmes de recherche en génomique végétale, car elle présente un génome de petite taille qui facilite les manipulations en laboratoires (son génome est 20 fois plus petit que celui du maïs). Son génome est séquencé entièrement en 2000 (cf. annexe 17). L'objectif de cette étude est la compréhension de la biologie des plantes à fleurs, la caractérisation de la structure, de la fonction et de la régulation des gènes qui le composent. Ces études sont menées afin d'identifier des gènes présentant des caractéristiques d'intérêt agronomique, comme la résistance au froid ou à la sécheresse, qui pourront être ensuite transférés à l'intérieur d'autres génomes plus complexes (Bonneuil, Thomas, 2012).

Arabidopsis Thaliana

Une structure emblématique : le GIS biotechnologies vertes

La France est le premier exportateur mondial et le premier producteur européen dans le secteur des semences[56]. Le GIS Biotechnologies vertes est spécialisé dans l'innovation végétale. Inscrit dans l'ère de la génomique, ses thèmes de recherche du GIS Biotechnologies vertes sont :

- La diversité génétique et la pré-sélection ;
- La lutte génétique contre les pathogènes et les parasites ;
- La maîtrise de la recombinaison pour accélérer l'innovation variétale ;
- L'édition des génomes ;
- Les métabolites secondaires des plantes (bioéconomie) ;
- La photosynthèse, architecture de la plantes et racines.

Le schéma de pré-sélection des plantes ci-après illustre les priorités de recherche dans les programmes public-privés. Les espèces qui y apparaissent ont un intérêt agronomique fort. Les traits recherchés répondent aux enjeux économiques et d'adaptation des plantes à l'environnement. La collecte, la conservation et la caractérisation sont trois étapes essentielles pour le travail de sélection variétale. Les nouvelles techniques de séquençage permettent une meilleure compréhension du fonctionnement de la plante, une identification des zones d'intérêt et donc une meilleure caractérisation.

[56] D'après la synthèse de la réflexion prospective et propositions du GIS Biotechnologies Vertes.

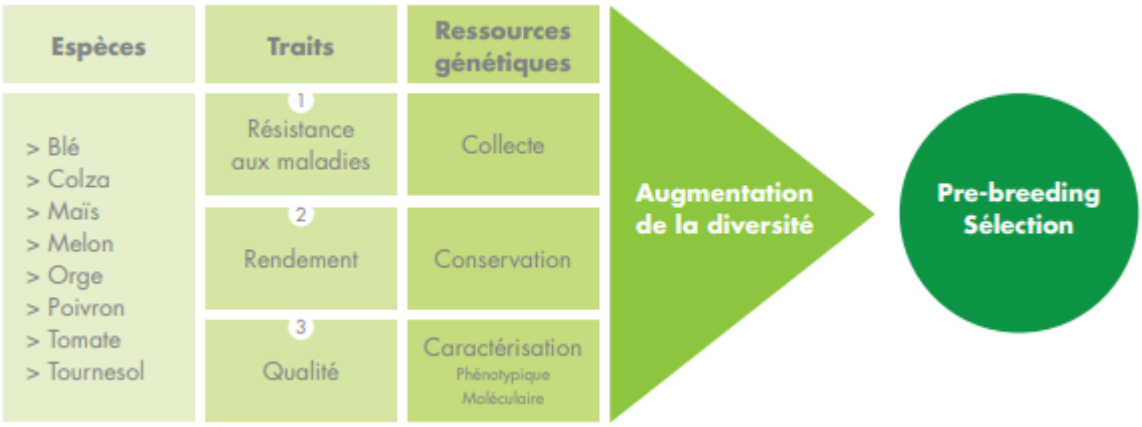


Figure 22 : Schéma synthétique de recherche sur la diversité génétique et la pré-sélection au sein du GIS BV

a. Le Programme Investissement d'Avenir (PIA) SUNRISE : du séquençage complet des génomes au programmes d'amélioration variétale

Contexte : Dans un contexte général de changement climatique, l'agriculture doit s'adapter aux nouvelles contraintes environnementales et notamment à la raréfaction de l'eau. Le tournesol, par sa faible exigence hydrique, est une des solutions pour faciliter l'adaptation de la filière végétale à ces évolutions. Améliorer sa résistance et ses caractéristiques agronomiques en conditions de sécheresse représente donc aujourd'hui un enjeu environnemental majeur. La production mondiale de graines oléagineuses, notamment de tournesol, doit également faire face à une demande croissante pour l'alimentation humaine (diversification des huiles), l'alimentation animale (richesse en protéines de ses tourteaux) et pour le développement des biocarburants et de la chimie verte.

Objectif :

Le Programme d'Investissements d'avenir Biotechnologies et Biosources SUNRISE a pour objectifs de réaliser le séquençage du génome complet (3,6 Gbases) qui servira de base à de nouveaux projets de sélection variétale et le reséquençage de génomes de 300 variétés de tournesol pour l'identification de marqueurs d'intérêts agronomiques.

Partenaires :

Le PIA SUNRISE a débuté en 2012 et s'achèvera en 2020, il réunit de nombreux partenaires. Le projet SUNRISE est coordonné par l'Inra et implique 9 laboratoires publics de recherche et 7 partenaires privés (5 entreprises semencières, une entreprise de biotechnologie, un institut technique). Ces partenaires collaborent aussi avec des partenaires étrangers (Canada et Etats-Unis notamment) travaillant sur le génome du tournesol depuis une dizaine d'années dans le cadre du consortium international pour les ressources génétiques de tournesol (ICSG). Le projet est régi par un accord de consortium qui organise notamment les règles relatives au partage des données. Il prévoit la construction d'une base de données initiale pour initier de nouveaux projets d'amélioration des variétés.

Techniques :

Le réseau de plateformes mobilisé est GenoToul qui contient les données générées dans un serveur spécifique (data center de l'INRA de Toulouse). Le reséquençage haut-débit type Illumina (2e génération) permet d'établir des indications de polymorphismes, elles sont accessibles et visualisables sur un site web (*genome browser*).

Sur l'accès aux données :

L'accord de consortium définit le cadre juridique pour l'accès aux données des différents partenaires et leur divulgation publique. En pratique, les données partagées concernent les résultats des assemblages de génomes et les polymorphismes issus du reséquençage car le transfert des données brutes est lourd, étant

donné le volume généré (de l'ordre du Téra octet) et des durées d'analyse très longues (de l'ordre de plusieurs mois). Le libre accès des résultats du séquençage du génome entier offre une base structurante pour de nouveaux projets en interne ou en partenariat avec des tiers. Concernant les données de polymorphismes, celles-ci sont plus stratégiques pour les différents partenaires : les parties privées ayant émis certaines réserves, des négociations ont lieu dès qu'un besoin de publication se présente pour les chercheurs. Les bases privilégiées lors de la publication sont celles de l'EMBL.

Résultats :

Un des résultats des recherches du projet SUNRISE a été de mettre en évidence la variabilité génétique du tournesol pour les processus de photosynthèse et de transpiration foliaire de la plante dans un contexte de déficit hydrique. Ces résultats pourront être intégrer aux modèles de culture. L'identification des gènes de tolérance à la sécheresse permettra d'améliorer les programmes de sélection et de mettre sur le marché de nouvelles variétés adaptées au changement climatique.

Sur la propriété des données :

Plus précisément, l'accord de consortium de ce PIA introduit la société Genoplante-Valor comme chargée de la propriété de certains résultats issus de SUNRISE, afin de réduire la fragmentation de la propriété intellectuelle.

b. Le projet de reséquençage massif chez le riz : plusieurs milliers de génomes pour exploiter la diversité du riz pour renforcer la sécurité alimentaire

L'IRRI (*International Rice Research Institute*, un centre du CGIAR situé aux Philippines) et le BGI (*Beijin Genomic Institut*, Chine) ont lancé, en 2010, un programme de reséquençage massif du riz avec, comme objectif, de séquencer l'ensemble des collections, représentant plus de cent mille accessions.

Les résultats de la première étape, qui a porté sur 3000 variétés asiatiques, conclue en 2014, a été publiée dans tous ses détails en 2018, avec la participation d'universités américaines et du Cirad.

Objectifs :

L'initiative globale de séquençage des collections de ressources génétiques du riz a pour objet l'étude de la diversité génétique et l'identification de gènes à valeur adaptative pour le développement de nouvelles variétés plus adaptées aux besoins de l'agriculture moderne face aux changements environnementaux.

Partenaires :

Pour la première phase, le séquençage des 3000 variétés de riz a été conduit par un consortium regroupant des institutions chinoises, l'IRRI, des universités américaines et le Cirad. Il a bénéficié de financements chinois et de la Fondation Bill & Melinda Gates. D'autres projets sont en cours, avec des financements variés, dont le projet IRIGIN financé par l'ANR via France Génomique. Le projet IRIGIN (*International RIce Genomic Initiative*). Ce projet constitue la contribution française à l'initiative de séquençage massif des riz. Coordonné par l'IRD, en collaboration avec le CIRAD, le CIAT (Centre internationale pour l'agriculture tropicale en Colombie, et centre du CGIAR) et *AfricaRice* (organisme panafricain du CGIAR), IRIGIN s'intègre dans le programme de recherche « *Rice* » du CGIAR. Il met l'accent sur du matériel en cours de sélection et sur les espèces africaines, dont la forme cultivée est adaptée à des conditions extrêmes, en particulier de forte température.

Résultats :

Plusieurs dizaines de millions de marqueurs moléculaires permettent une caractérisation très fine de la structure de la diversité le long du génome. On observe également une forte variation du contenu en gènes. Seules 60% des familles de gènes connues chez le riz sont communes à toutes les variétés ; les autres sont présentes, ou absentes, selon les variétés et constituent des sources potentielles d'adaptation. Les caractérisations phénotypiques sont en cours afin de repérer des variations au niveau de gènes importants sur le plan fonctionnel. Le but est de transférer et d'appliquer les découvertes de la génomique à l'amélioration du riz et d'autres céréales.

Sur la mise à disposition des données :

Les données de séquençage de la première phase sont disponibles sur le site internet[57] dans le cadre de *l'International Rice Informatics Consortium* (IRIC), consortium ouvert qui regroupe de nombreux acteurs[58]. Les institutions françaises participent en développant des outils nouveaux (comme le manager de workflow TOGGLe) accessibles sur la plateforme Southgreen.

Sur des formes de renforcement de capacités/transfert de technologies :

Les institutions membres de l'IRIC restent pour l'instant cantonnées aux pays développés. Dans le cadre de leurs mandats, le Cirad et l'IRD travaillent en coopération avec divers partenaires de pays en développement, les associent aux travaux de recherche et organisent des formations ciblées pour chercheurs et étudiants.

c. Le projet Genius, « Ingénierie cellulaire : Amélioration et innovation technologiques pour les plantes d'une agriculture durable », outils pour une modification ciblée des caractères agronomiques

Objectifs :

Ce projet vise à répondre aux enjeux actuels d'agriculture durable, à travers l'étude de plusieurs caractères de 12 espèces différentes[59] (neuf cultivées et trois modèles) pour la réduction des intrants (résistance à des pathogènes chez la tomate, le pommier, le peuplier, le colza), l'adaptation au changement climatique (tolérance à la salinité chez le riz), l'utilisation de la biomasse végétale (qualité de l'amidon chez la pomme de terre), etc.

Partenaires :

Genius (2012-2019) est l'un des projets Investissement d'Avenir financés par l'ANR dans le cadre l'action Biotechnologies et Bioressources et labellisés par le GIS Biotechnologies vertes. Les 14 partenaires réunissent des acteurs de la recherche publique (Inra, Cirad, universités de Lyon-III) ainsi que des entreprises privées (Biogemma, Germicopa, Société Nouvelle Pépinières & Roseraies Georges Delbard et Vilmorin).

Techniques :

Le projet consiste à appliquer à ces espèces des outils pour modifier des gènes de manière plus ciblée.

Les méthodes de modifications ciblées des gènes sont de plus en plus performantes. Lors de la conception du projet Genius en 2011, les partenaires se sont focalisés sur les méganucléases et les TALENs, alors en plein essor. Depuis, ils ont su adapter le programme de travail pour tenir compte de l'apparition de la technologie Cas9-CRISPR en 2012. Il est important de préciser que ces modifications partent de l'étape préliminaire d'observation des caractères d'intérêt sur le phénotype des plantes au champ ou en laboratoire. Sur la photographie ci-dessous, les caractères d'intérêt observés sont par exemple la couleur des feuilles pour la pomme ou la longueur des tiges pour le riz. Le point de départ est au niveau des ressources génétiques et des données de séquences qui sont générées à partir de ces ressources génétiques.

Résultats :

Des recherches ont porté sur les outils de sélection. Pour le maïs, cela a permis l'obtention de gamètes diploïdes pour la propagation asexuée des cultures. Un autre exemple de recherche du projet Genius consiste à travailler sur la qualité des produits, ici la pomme de terre, via la modification de gène(s) permettant la production d'un amidon composé uniquement d'amylopectine, utile pour l'industrie alimentaire et de la colle. Le tableau ci-après présente les exemples de recherche dans le cadre du projet Genius.

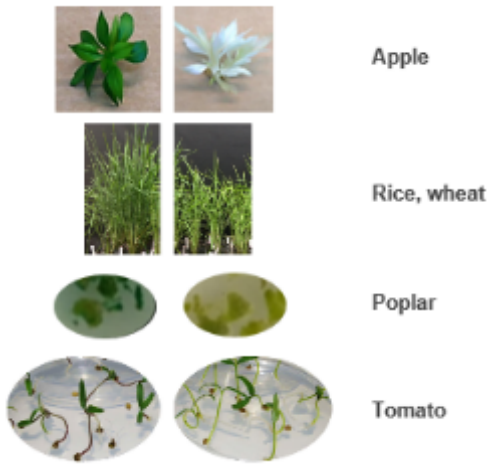
[57] http://snp-seek.irri.org/

[58] Arizona Genomics Institute (USA), Cornell University (USA), International Center for Tropical Agriculture -CIAT (Colombia), International Rice Research Institute - IRRI (Philippines), National Institute of Agrobiological Sciences (NIAS), Tsukuba, Japan, The Genome Analysis Center (UK), NIAB - National Institute of Agricultural Botany (United Kingdom), King Abdullah University of Science and Technology (KAUST), Bayer Crop Science, Syngenta, Institut de recherche pour le développement (IRD - France), Cirad - Agricultural Research for Development (France), University of Western Australia, Kongju National University (Korea), National Taiwan University, Louisiana State University (USA).

[59] Il s'agit de plantes cultivées couvertes par le TIRPAA.

Tableau 7 : Exemples de recherche dans le cadre du projet Genius

Thème de recherche	Modification de gènes	Finalités
Outils de sélection	Maïs, gamètes diploïdes	Propagation asexuée des cultures
Qualité des produits	Pomme de terre, amidon composé uniquement d'amylopectine	Industrie alimentaire et de la colle
Temps de floraison	Pomme, floraison très précoce	Cycle de vie raccourci et l'adaptation au changement climatique
Adaptation au stress abiotique	Riz, tolérance à la salinité	Pour la culture sur des terres marginales et l'adaptation au changement climatique
Résistance aux maladies	Tomate, résistance au potyvirus	Pour la protection des plantes et la réduction de pesticides



Vers des plantes cultivées présentant des caractères de culture et de qualité améliorés pour l'alimentation humaine et animale et d'autres utilisations, Peter Rogowsky, 2018

La mise à disposition des résultats :

Comme le projet Genius ne produit pas des données ou séquences à haut débit, il ne possède pas de base de données dédiée. Les modifications apportées aux génomes et, le cas échéant, les phénotypes associés, sont décrits dans des publications au fil de l'eau.

4. LES UTILISATIONS DE DONNÉES DE SÉQUENÇAGE DE RESSOURCES GÉNÉTIQUES DE MICROORGANISMES :

a. Les utilisations de données de séquençage de ressources génétiques de microorganismes du sol : les perspectives offertes par la métagénomique

Objectif :

Le concept d'holobionte répond à un besoin de perception du vivant plus holiste que celle communément adoptée. Le concept d'holobionte correspond à la plante et son cortège symbiotique (champignons, bactéries, etc.). L'importance de la symbiose mycorhizienne pour la croissance et la survie des plantes est au cœur des projets portant sur ce concept. L'objectif est la prise en compte du microbiote pour envisager demain une agriculture durable.

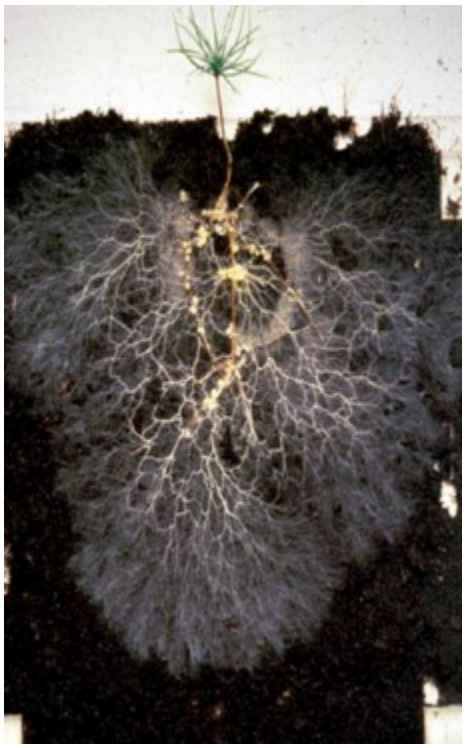
Partenaires :

L'unité ECOBIO « Ecosystèmes, Biodiversité, Evolution »[60] de l'université Rennes I, porte ce projet d'étude autour de l'holobionte, en partenariat avec l'université d'Amsterdam et de Wageningen, ainsi que l'Institut Max Planck en Cologne.

Techniques :

Les évolutions en génomique, métagénomique, etc. permettent de développer les connaissances autour de ces symbiontes. Notamment, l'information génomique permet la compréhension et la modélisation des interactions impliquées dans la production alimentaire et le fonctionnement des écosystèmes.

[60] Unité pluridisciplinaire d'écologie dont l'axe fédérateur a pour objet la biodiversité des écosystèmes continentaux et insulaires, de la molécule à l'écosystème.



Réseau mycorhisien dense (Vandenkoonhuyse, 2018)

Résultats :

Les macroorganismes ne sont pas des individus mais la résultante de l'assemblage d'un hôte avec son cortège symbiotique. Le microbiome des plantes est un environnement complexe dont les interactions entre ses composants ont des conséquences sur la survie de la plante (adaptation à de nouveaux habitats). La photo ci-contre illustre le réseau très dense des mycorhizes arbusculaires présent chez 80 % des plantes terrestres.

b. Les projets précompétitifs de l'industrie laitière : étude de la diversité microbienne

Pour la filière laitière, le Centre national interprofessionnel de l'économie laitière (CNIEL) possède une banque de souches (la collection « MIL ») composée de flores d'intérêt, de flores pathogènes, et de virus de bactéries. Cette collection est gérée par Actalia, le centre technique et d'expertise qualifié Institut technique agro-industriel (ITAI [61]) par le ministère de l'agriculture.

Les données de séquençage sont mobilisées dans des projets précompétitifs pour l'amélioration de la production et de la transformation laitière. Le but est d'améliorer la compréhension des écosystèmes microbiens, leur diversité et leurs fonctionnalités pour différents intérêts (connaissance, transfert d'outils, suivi d'écosystèmes, constituer des banques de référence, etc.). Les caractéristiques du fromage (couleur, acidification, texture, flaveur, etc.) sont générées par ces écosystèmes et contribuent à leur qualité.

[61] Les Instituts techniques agro-industriels (ITAI) sont des organismes de recherche technologique, d'expertise, d'assistance technique et de formation, au service des entreprises et en particulier des PME. Positionnés à la jonction du monde de la recherche, des entreprises et des organismes professionnels, ils jouent un rôle majeur dans la diffusion, le transfert et la valorisation des résultats de la recherche auprès des petites et moyennes entreprises.

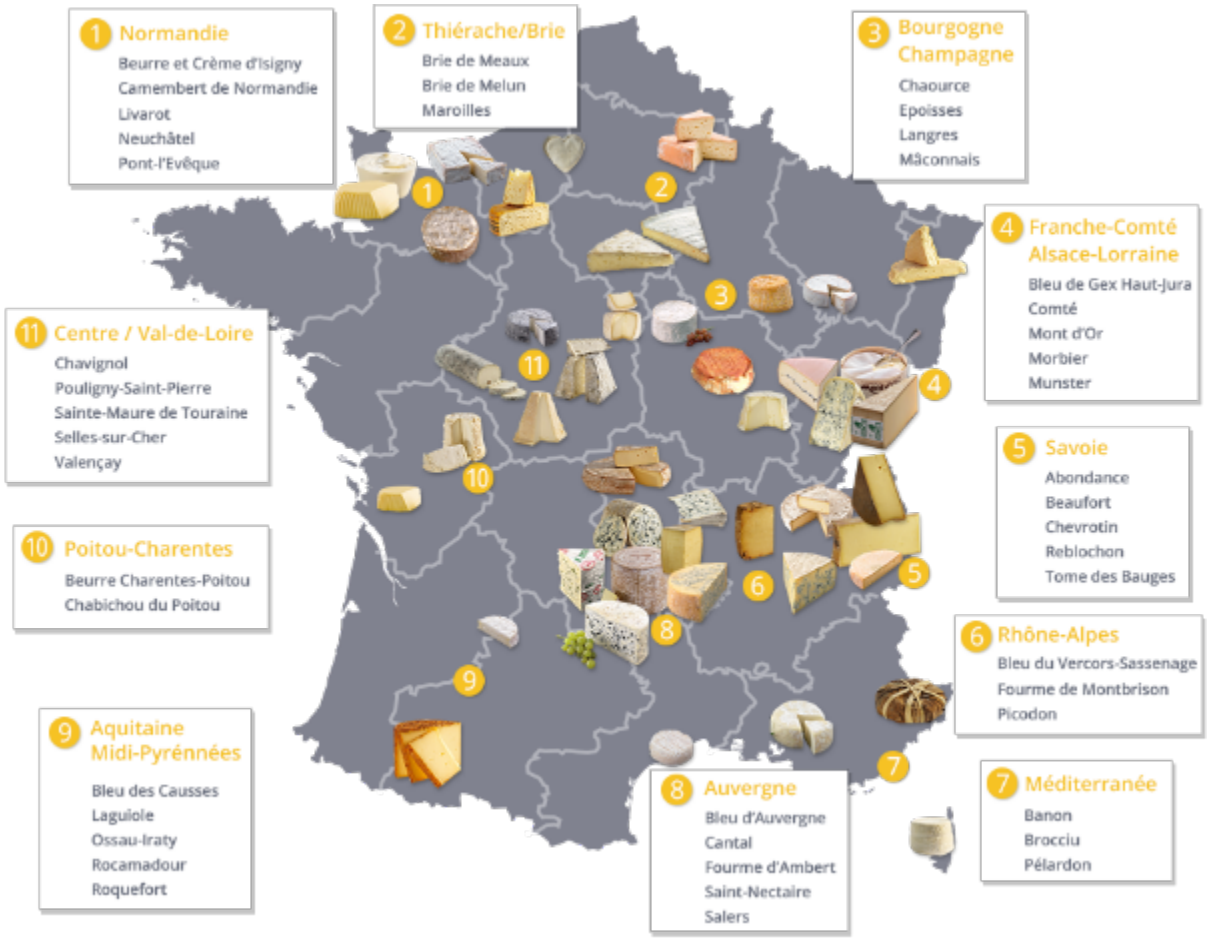


Figure 23 : Les 45 fromages AOP français, présentation « acquisition et utilisation des données de séquençage dans les projets soutenus par le CNIEL », Frédéric Gaucheron.

Objectif :

Le projet abordé a pour objectif d'établir un catalogue des communautés microbiennes présentes dans l'ensemble des fromages bénéficiant d'une appellation d'origine protégée (AOP[62]) française qui sont issues de la combinaison de pratiques variées de production laitière et de transformation du fromage. La composante microbiologique dans l'élaboration et la typicité des fromages est importantes mais complexe et varie d'un fromage à l'autre.

Technique :

L'approche métagénomique associée à l'utilisation de nouvelles techniques de séquençage haut débit est mobilisée pour ce projet.

Résultats :

Ce projet va alimenter les connaissances sur la diversité des communautés microbiennes naturelles qui sont perdues progressivement dans les laits et les fromages par la pression sanitaire.

c. Le projet Emissage : cas d'utilisation des données de séquençage pour le suivi épidémiologique au sein de la filière animale

Le projet Emissage répond aux enjeux de sécurité sanitaire dans les filières animales. La surveillance des dangers microbiologiques dans les chaînes alimentaires se fait par l'évolution des techniques en génomique. Le centre technique Actalia, l'Institut du porc (Ifip) et l'Agence nationale de sécurité de l'alimentation, de l'environnement et du travail (Anses) sont des acteurs incontournables dans l'utilisation des données de séquençage à cette fin.

[62] L'Appellation d'origine protégée (AOP) désigne un produit dont toutes les étapes de production sont réalisées selon un savoir-faire reconnu dans une même aire géographique, qui donne ses caractéristiques au produit. C'est un signe européen qui protège le nom du produit dans toute l'Union européenne (Site internet de l'Institut national de l'origine et de la qualité, Inao).

Salmonella enterica

Objectif :

Au sein du projet EMISSAGE (2018-2022), se concentre sur l'espèce (*Salmonella enterica*) et trois de ses sérovars[63] : *S. Typhimurium*, son variant monophasique et *S. Mbandaka*, représentant des situations épidémiologiques différentes. L'objectif du projet est d'assurer le suivi épidémiologique (distribution, facteurs, contrôle) des souches d'intérêts laitiers.

Le projet Emissage a pour but de :

- développer un outil informatique de traitement de données génomique brutes,
- créer une base de données génétiques qui puisse être partagée entre les partenaires du projet,
- développer des marqueurs épidémiologiques et génétiques pour optimiser la surveillance de ces sérovars dans les contextes terrain étudiés. Et ainsi faciliter pour les opérateurs des filières la surveillance sanitaire des Salmonella par l'utilisation de ces outils.

Partenaires :

Le projet EMISSAGE s'appuie sur l'UMT ASIICS[64] dont les travaux portent sur deux pathogènes d'intérêt majeur en filière laitière et porcine que sont Listeria monocytogenes et Salmonella enterica. L'UMT ASIICS a pour but le transfert d'outils d'analyse des données, de connaissances et de compétences vers les instituts techniques, afin qu'ils gagnent en savoir-faire génomique pour le déploiement des méthodes basées sur le séquençage génomique global[65] (WGS).

Techniques :

Des campagnes de prélèvements de souches de salmonelles dans les filières (abattoirs porcins, exploitations et sites de transformation laitiers) ont été réalisées et le séquençage WGS de ces souches a permis le recueil de leurs données épidémiologiques.

Résultats :

Ce projet permet d'identifier les sérovars pour une meilleure surveillance et l'étude de la dissémination et des voies de contamination, afin de mettre en place des mesures sanitaires adaptées.

d. Projet Bakery : domestication de la levure pour l'industrie agro-alimentaire

Contexte :

Le projet Bakery (2014-2018) a pour but d'étudier la diversité et les interactions d'un écosystème agroalimentaire Blé/Homme/Levain à faible intrant pour une meilleure compréhension de la durabilité de la filière boulangerie. Ce projet d'une durée de 3 ans est lié à un PIA.

Objectif :

Ce projet de recherche pluridisciplinaire et participatif vise à (i) décrire la diversité socio-culturelle des pratiques de boulangerie et la perception qu'en ont les consommateurs (ii) étudier les effets des variétés de blé, du terroir et des pratiques des boulangers sur la diversité du microbiome levain, la qualité sensorielle et nutritionnelle du pain ainsi que les préférences des consommateurs (iii) analyser les interactions microbiennes au sein du levain et leurs conséquences sur le fonctionnement du levain et sur la qualité du pain (iv) intégrer toutes les données pour identifier les déterminants de la diversité biologique et socio-culturelle dans la chaîne de boulangerie, (v) envisager des stratégies pour la conservation de la diversité biologique et de la diversité socio-culturelle en boulangerie.

[63] Un sérovar, ou sérotype, est la propriété antigénique permettant d'identifier une cellule ou un virus.

[64] L'UMT ASIICS est spécialisé sur la gestion sanitaire. Ses partenaires principaux sont Actalia (centre technique et d'expertise agroalimentaire), l'Anses (Agence nationales de sécurité sanitaire de l'alimentation, de l'environnement et du travail) et l'Ifip (Institut du porc).

[65] Le séquençage génomique global est très utilisé dans le domaine de l'épidémiologie. Il permet de différencier les souches d'agents pathogènes d'origine alimentaire en étudiant le génome microbien.

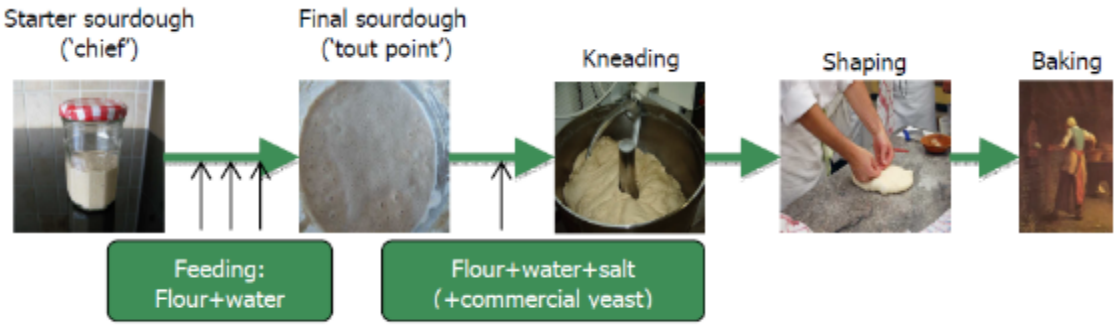


Figure 24 : Les étapes de production du pain,
(programme systèmes alimentaires durables: projet Bakery, ANR 2013)

Partenaires :

Le projet réuni sept partenaires dont les Centres internationaux de ressources microbiennes dédiés aux levure (CIRM levures[66]) et aux bactéries d'intérêt alimentaire (CIRM BIA) de l'Inra, l'Institut technique national de l'agriculture biologique (ITAB), l'université de Bretagne occidentale et l'Ecole nationale vétérinaire, agroalimentaire et de l'alimentation de Nantes-Atlantique.

Techniques :

Dans un premier temps, des enquêtes ont été menées auprès de 30 boulangers et agriculteurs, en particulier des boulangers français qui fabriquent des pains au levain, en utilisant des farines résultant des pratiques agro-écologiques. Ces enquêtes ont permis une récolte d'informations sur les pratiques des boulangers et sur l'origine des semences de blé pour les boulangers et agriculteurs, ainsi qu'une récolte des échantillons de farine, de levain et de pain. Dans un deuxième temps, une enquête a été menée auprès de consommateurs. En laboratoire, les analyses du microbiome des graines, de la farine et du levain ont mobilisé les techniques de séquençage métagénomique et la phylogénie (analyse des loci[67] d'ADNr). La caractérisation biochimique ainsi que l'analyse sensorielle seront réalisées pour les levains et les pains.

Résultats :

Les premiers résultats montrent que la boulangerie à faible intrant en France héberge une diversité d'espèces microbiennes importante et originale comparée à la diversité observée ailleurs dans le monde. De nouvelles espèces de levure ont été découvertes. Plusieurs espèces de bactéries lactiques ont été détectées en boulangerie pour la première fois. Les différents types de boulanger (paysans boulangers, artisans boulangers et PME) hébergent des communautés microbiennes différentes, ce qui montre l'importance de maintenir une diversité socio-culturelle.

La levure est un champignon microscopique unicellulaire (saccharomyces). Elle est utilisée depuis des millénaires à l'état sauvage et depuis le XXème siècle, elle est domestiquée et fabriquée pour l'industrie agro-alimentaire. Les applications les plus emblématiques sont les levures « ferments », les levures « aliments », levures « bénéfices santé » et les levures pour la production de biocarburants.

Encadré 6 : La levure, ressource génétique de microorganismes

5. LES UTILISATIONS DE DONNÉES DE SÉQUENÇAGE DE RESSOURCES GÉNÉTIQUES FORESTIÈRES (RGF) : APPORT DES DONNÉES DE SÉQUENÇAGE ET DE GÉNOTYPAGE DANS LA GESTION ET L'UTILISATION DES RESSOURCES GÉNÉTIQUES FORESTIÈRES

Partenaires :

Les partenaires impliqués dans les projets ayant pour vocation la conservation et l'amélioration et décrits ci-après sont l'Inra, l'Office national des forêts (ONF), l'Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture (IRSTEA), l'Institut technologique forêt cellulose bois-construction ameublement (FCBA) organisé notamment, au sein du GIS peuplier « Amélioration, sélection et

[66] Le CIRM-Levures a été constitué par le regroupement de nombreuses collections de levures (Technologie Laitière de l'INA-PG, INRA de Montpellier, de Colmar, de Poligny, du Rheu etc) autour de la Collection de Levures d'Intérêt Biotechnologique (CLIB) créée en 1991 à Thivernal-Grignon.

[67] Un locus est la position fixe d'un gène ou d'un marqueur sur un chromosome.

protection du peuplier » et de la Commission sur les ressources forestières qui travaille sur la diversité génétique des principales espèces forestières françaises.

Techniques :

L'adoption des technologies de Génotypage haut débit (GHD) et de Séquençage haut débit (SHD) a permis des percés scientifiques sur les arbres forestiers et ouvre des perspectives d'application en termes de gestion et conservation des RGF. Pour le peuplier noir notamment, L'Inra a investi au travers d'actions incitatives (AIP sequençage, AIP Bioressouces, BRG, ECOGER ...) dans le séquençage Sanger, puis dans le séquençage numérique et la construction de Biopuces dans le cadre de projets européens (EVOLTREE, NOVELTREE) qui ont généré des données et des outils permettant de mener des recherches innovantes sur l'adaptation et la valorisation des ressources génétiques de peuplier noir.

Quelques exemples emblématiques peuvent être cités pour le pin maritime et le peuplier noir.

a. Pin maritime : Les marqueurs moléculaires au service de l'amélioration génétique

Contexte :

L'amélioration génétique du pin maritime a débuté dans les années 1960 par la sélection, en forêt, d'arbres présentant une supériorité pour les caractères d'intérêt sylvicole (croissance, rectitude du tronc, branchaison, résistance aux pathogènes). Ces arbres « élites » ont été conservés par greffage dans des parcs à clones ; ils constituent la population de base du programme d'amélioration. Deux cycles successifs de croisement/sélection ont permis de réaliser des gains génétiques significatifs diffusés au sein de variétés (gains d'environ 30 % pour la croissance et la rectitude du tronc). Chez le pin maritime, un cycle de sélection dure actuellement plus de 20 ans. En simplifiant, les arbres candidats à la sélection sont tout d'abord évalués à partir des performances de leurs descendants puis les meilleurs candidats sont croisés entre eux afin de générer de la variabilité génétique pour la génération suivante. En parallèle, les meilleurs individus sont greffés pour établir des vergers à graines qui fourniront, au bout de 8 à 10 ans, les semences pour les futures plantations. Le développement de méthodes de génotypage performantes et à moindre coût permet aujourd'hui d'envisager de nouvelles stratégies de sélection et de production de variétés beaucoup plus courte.

Résultats :

(i) Identifier l'origine géographique des RGF

Une biopuce à ADN a été développée sur la base du séquençage du transcriptome du pin maritime (Plomion *et al.*, 2016). Avec plus de 12 000 marqueurs moléculaires, elle a permis de caractériser la diversité génétique des principaux écotypes de pins maritimes. Sur la base de la structuration de cette diversité nucléotidique, les marqueurs qui différencient (en termes de fréquence allélique) les différentes provenances géographiques ont été identifiés. Il est ainsi possible de diagnostiquer la provenance géographique d'un peuplement ou d'un lot de graines.

(ii) Étudier le régime de reproduction dans les vergers

Les vergers de pin maritime sont constitués d'arbres élites en pollinisation libre. Grâce au développement d'une puce de 80 SNPs, il est désormais possible de mieux connaitre le régime de reproduction de ces vergers, à savoir les contributions parentales réelles et le pourcentage de pollution pollinique (i.e. pollinisation des arbres élites par des arbres extérieurs au verger et donc de qualité génétique a priori inférieure). Ces études doivent permettre d'optimiser le design et la gestion des vergers à graines afin de maximiser le gain génétique.

(iii) Certifier les RGF du programme d'amélioration : augmenter les gains génétiques des variétés

Un kit de 80 marqueurs moléculaires a été développé afin de caractériser chaque arbre et chaque copie clonale du programme d'amélioration (soit environ 7 000 spécimens). Ces cartes génétiques permettent de corriger les erreurs d'identité et de pedigree accumulés au cours de cycles de sélection (estimé au minimum à 10 %). Ceci permettra d'estimer plus précisément les valeurs génétiques des individus pour augmenter les gains génétiques des futures variétés.

(iv) Simplifier les cycles de sélection

Les chercheurs ont montré qu'il était possible de sélectionner quelques dizaines de marqueurs moléculaires pour identifier de façon unique chaque individu et reconstituer son pedigree, via l'identité de sa mère et de son père. Il devient alors envisageable de simplifier les cycles de sélection en substituant aux croisements bi-parentaux, des croisements de type « *polycross* » où une mère est croisée avec un mélange de plusieurs pollens. Cette stratégie présente aussi l'avantage de favoriser le brassage génétique dans la population d'amélioration. Le pedigree des arbres, indispensable pour évaluer leur valeur génétique avec précision, est alors reconstitué, *a posteriori*, grâce aux marqueurs moléculaires.

(v) Prédire la valeur d'un arbre par la génomique pour accélérer les cycles de sélection

Une autre approche rendue possible par l'utilisation des marqueurs moléculaires consiste à construire un modèle de prédiction calibré dans une population génotypée pour un grand nombre de marqueurs moléculaires (plusieurs milliers) et caractérisée finement pour ses performances (croissance, rectitude du tronc, etc.). Ce modèle statistique permet alors de prédire la valeur génétique d'un arbre à partir des marqueurs moléculaires sans attendre que ses performances soient mesurées à l'âge adulte, soit un gain de temps considérable. Afin d'évaluer la possibilité d'utiliser cette stratégie chez le pin maritime, des centaines d'arbres de la population d'amélioration ont été génotypés avec plusieurs milliers de marqueurs moléculaires. La performance réelle des arbres a ensuite été comparée avec celle prédite par le modèle statistique. Les résultats montrent que le niveau de précision est équivalent à celui basé sur la connaissance des pedigrees. A l'avenir, cette méthodologie devra être améliorée pour en augmenter la précision et essayer de sélectionner les meilleurs arbres en intra-famille afin d'accélérer les gains génétiques par unité de temps.

b. Le peuplier noir, une espèce d'intérêt économique et écologique

Contexte :

Le peuplier noir (*Populus nigra* L.) est une espèce forestière pionnière qui constitue avec les saules une composante majeur des écosystèmes des forêts riveraines en Europe et en Asie. D'autre part, cette espèce est utilisée comme espèce parentale pour la création de cultivars hybrides utilisés en populiculture, c'est-à-dire dans les cultures en peuplement artificiels de peupliers.

L'annonce du séquençage du génome du premier arbre, un peuplier américain (*Populus trichocarpa)* (Tuskan et al., 2006) a été le point d'inflexion de la courbe de croissance des ressources génomiques chez les peupliers.

L'existence du peuplier noir est menacée par modification de la dynamique des fleuves et par la coexistence avec des peupliers cultivés et ornementaux. Ces menaces posent un problème de conservation in situ des ressources génétiques naturelles.

Résultats :

(i) Une étape dans le programme de conservation : caractériser la diversité génétique

Une biopuce développée pour caractériser la diversité à l'échelle européenne (Faivre-Rampant *et al.*, 2016) permet de génotyper environ 8 000 marqueurs moléculaires ; les résultats majeurs sont :

- La diversité est structurée par des barrières physiques majeurs (Alpes), selon les grands bassins fluviaux mais des flux de gènes et un brassage génétique à longue distance reste important (ex. Loire/Allier) ;

- Une des caractéristiques dans les bassins fluviaux fortement régulés (ex : Rhin) est l'existence de nombreux individus possédant un patrimoine génétique commun (clone) ;

- Une introgression par le compartiment cultivé, c'est-à-dire le transfert de gènes d'une espèce vers le pool génétique d'une autre espèce, est détectée dans la plupart des populations étudiées.

(ii) Intégration de l'information génomique dans le programme d'amélioration

La création de variétés hybrides de peuplier repose sur des cycles de croisement et sélection chez les espèces parentales (sélection récurrente) et sur la sélection des hybrides obtenus avec les meilleurs parents. Les cycles de croisements produisent potentiellement de nombreux descendants, mais les capacités d'évaluation sont actuellement limitées. De plus, il faut 15 à 20 ans d'étapes d'évaluation/sélection multicritères pour obtenir une variété. Le séquençage et le génotypage haut débit permet de lever les limitations et d'accélérer les cycles de sélection :

- Reconstitution de pedigree, contrôle d'identité et traçabilité : La biopuce décrite plus haut a permis de corriger certaines erreurs de généalogie dans les populations d'amélioration de manière très précise et de mieux estimer les apparentements, une étape indispensable pour l'évaluation de la valeur génétique des individus. Elle a démontré également un fort potentiel pour contrôler l'identité des cultivars de peuplier *P. deltoides* x *P. nigra* inscrits au catalogue européen.

- Prédiction génomique de la valeur d'un arbre : De manière analogue au pin maritime, des milliers de marqueurs présents sur la biopuce citée précédemment ont été utilisés pour prédire la performance d'un millier d'arbres d'une population d'entrainement constituée des plusieurs familles avec des liens de parenté. De même, la précision de la prédiction génomique reste équivalente à celle qui peut être faite avec l'information du pedigree (apparentement généalogique). Toutefois, l'augmentation du nombre de marqueurs pourrait apporter un avantage à la prédiction génomique.

- Augmentation de la densité de marquage par imputation : l'un des facteurs touchant la précision de l'évaluation génomique est la densité de marqueurs. Malgré la réduction des coûts de séquençage, il n'est pas envisageable de séquencer actuellement tous les candidats à la sélection pour obtenir un génotypage haute densité. Un autre moyen d'augmenter la densité de marqueurs disponibles sur l'ensemble de la population d'amélioration est une méthode statistique dite par imputation. A partir d'individus bien choisis (parents par exemple) que l'on séquence entièrement, il est possible de compléter le génotype de tous les individus de la population d'amélioration (génotypés uniquement avec la biopuce) et ceci avec une bonne précision. Ainsi, on peut passer de 8 000 marqueurs à 1 400 000 marqueurs pour tous les individus.

Perspectives :

Les coûts de séquençage haut débit vont en diminuant, ce qui donne accès au polymorphisme à l'échelle du génome pour un plus grand nombre d'individus. Ces technologies évoluent également vers le séquençage de long fragment, une évolution qui ouvre l'accès à des polymorphismes autres que les mutations ponctuelles : les variants de structure. Ils sont encore peu étudiés chez les espèces forestières (Pinosio *et al.*, 2016) mais potentiellement importants en liaison avec les variations phénotypiques.

6. UTILISATIONS DE DONNÉES DE SÉQUENÇAGES DE RESSOURCES GÉNÉTIQUES D'INVERTÉBRÉS : POUR UNE GESTION DURABLE DE L'APICULTURE

a. Le projet BEEHOPE pour lutter contre le syndrome d'effondrement des abeilles

Abeille noire de Chizé (Biodiversa)

Le taux d'extinction actuel des espèces dans la biosphère serait comparable à celui des dernières extinctions massives[68]. La réduction de la richesse en espèces et de la diversité génétique s'accompagne de la détérioration d'un grand nombre de services écosystémiques tels que la pollinisation par les animaux (zoogamie). Plusieurs facteurs biotiques (pathogènes, espèces exotiques, par exemple) et abiotiques (perte et fragmentation de l'habitat, produits agrochimiques, changement climatique, etc.) sont probablement impliqués dans cette perturbation de la pollinisation et dans le déclin des espèces pollinisatrices entraînant une perte de diversité génétique.

L'abeille domestique illustre particulièrement bien ces problèmes : elle revêt une importance primordiale en matière écologique et agronomique ; pourtant, des pertes de colonies ont été récemment signalées dans le monde entier à des taux alarmants. L'abeille domestique est un insecte d'importance agroenvironnementale. Son activité de recherche de nourriture dans un rayon de 12 km autour de la ruche le met en contact avec une grande variété de polluants, y compris de pesticides. Depuis environ 20 ans, on observe que l'abeille domestique est soumise à un déclin constant pour lesquels pesticides et agents pathogènes semblent représenter les principaux contributeurs. Cependant, des études récentes suggèrent que les déclins actuels d'abeilles mellifères dans les ruchers européens peuvent également être causés par les échanges commerciaux et européens d'abeilles domestiques par (i) l'introduction de colonies non adaptées et artificiellement maintenues (ii) la propagation de pathogènes envahissants véhiculés par les abeilles allochtones.

Objectif :

Le projet BEEHOPE vise à mieux comprendre l'écologie de l'abeille noire (*Apis mellifera mellifera*) afin de pouvoir instaurer une gestion durable de l'apiculture. L'abeille noire est une espèce qui a été délaissée par les apiculteurs au profit d'espèces plus productives, bien qu'elle soit parfaitement adaptée aux climats et paysage d'Europe du Nord. Le projet a pour objectif de récolter des données sur l'abeille noire pour l'étude de ses caractères adaptatifs. Une zone de sept kilomètres de diamètre a été délimitée autour du premier rucher pour éviter que des espèces d'abeilles importées ne s'installent.

Partenaires :

Six partenaires européens, dont le CNRS de Chizé, participent au projet. Ils collectent et partagent les données recueillies sur l'abeille noire en impliquant les citoyens localement.

Techniques :

L'évaluation de la diversité génétique est réalisée à l'aide de marqueurs moléculaires (mitochondriales et microsatellites[69]). Le séquençage NGS est utilisé pour i) créer un nouveau système génétique basé sur les marqueurs moléculaires, (ii) créer un profil de marqueur moléculaire exclusif pour la population d'abeilles incluse dans chaque centre de conservation, utile pour l'affectation de l'origine, (iii) créer un ensemble de fragments génomiques montrant les signatures des balayages sélectifs associés à une adaptation locale.

Résultats :

À l'aide des habitants du territoire, un grand nombre d'essaims d'abeilles autour du CNRS de Chizé a été récupéré. Une identification génétique sur une zone de 7 km autour du CNRS est nécessaire pour avoir un état des lieux de la génétique des abeilles du territoire. Une partie des résultats montre que, malgré les efforts de protection de l'abeille noire, celle-ci présente des niveaux d'introgression élevé (8 % contre 30 % pour les populations non protégées). Certaines populations protégées nécessitent encore des ajustements

[68] Franck P., L. Garnery, A. Loiseau B.P. Oldroyd, H.R. Hepburn,M. Solignac, J.M. Cornuet (2001) Genetic diversity of the Honey bee in Africa: microsatellite and mitochondrial data Heredity 86 : 420-430

[69] Les microsatellites correspondent à des séquences constituées d'unités répétées de 1 à 4 nucléotides sur le génome.

dans les stratégies de gestion pour purger davantage les allèles étrangers identifiés à l'aide de marqueurs moléculaires (Pint *et al.*, 2014).

Mise à disposition des résultats :

Bien que tous les résultats soient publiés, il est prévu de réaliser un transfert d'informations vers les instances officielles concernées par l'évaluation des risques pour l'environnement. Il est prévu d'afficher les résultats sur un site web pour obtenir une diffusion large et non restreinte.

CONCLUSION

Ce rapport est le fruit de l'enquête commandée par le ministère de l'agriculture et de l'alimentation sur l'utilisation des données de séquençage de ressources génétiques pour l'alimentation et l'agriculture.

Ce rapport a mis en évidence la réalité multiple que couvre les données de séquençage, qui gagnent en intérêt à mesure qu'elles sont traitées, analysées et croisées avec d'autres données.

Deux résultats importants découlent des entretiens réalisés dans le cadre de l'étude. D'une part, nous proposons une terminologie pour traduire l'acronyme « DSI » « *digital sequence information* » (« information de séquençage numérique sur les ressources génétiques ») utilisé par la CDB : « données numériques de séquences de ressources génétiques » (« *Digital sequence data* » ou « *Digital data of genetic resource sequences* »). D'autre part, nous proposons une typologie simple suivant le protocole bioinformatique de traitement des données issues directement du séquenceur : donnée brute, donnée nettoyée, donnée analysée.

Le développement et la diffusion des nouvelles techniques de séquençage ont révolutionné les outils dans le domaine de la biologie moléculaire. Aujourd'hui, les enjeux sont liés au traitement des données produites, à leur transfert, à leur stockage ainsi qu'à leur statut juridique. Le domaine de l'informatique a acquis une place fondamentale dans les équipes de recherche.

Tous les secteurs sont concernés, plus largement que celui de l'alimentation et de l'agriculture, ce qui mériterait une approche coordonnée pour réfléchir au devenir des données de séquençage, en termes techniques et juridiques. Bien que le libre accès aux données soit prôné par les programmes de recherche européen et français, il existe déjà des droits sur l'accès aux données et aux bases de données. On assiste à une contradiction entre une volonté de promouvoir un accès libre aux données et une volonté de contrôler l'accès à l'information contenue dans les lots de données créés dans le cadre de projets de recherche partenariaux. De plus, la recherche privée n'est pas tenue de mettre à disposition les données qu'elle produit dès lors que l'acquisition de ces données n'a pas été obtenue grâce à un financement majoritairement public, cependant, la recherche privée bénéficie de l'accès aux bases de données publiques sans restriction et surtout sans avoir participé à son financement.

Au cours de l'enquête menée auprès des infrastructures de recherche et des entreprises dans le secteur de l'agriculture et de l'alimentation, différentes utilisations des données de séquençage ont été mises en évidence, des cas majoritaires tels que l'étude de diversité génétique ou la caractérisation du génome. D'autres techniques utilisent les données de séquençage de RGAA comme la sélection assistée par marqueurs et plus récemment les nouvelles techniques de sélection (« *New Breeding Technics* ») dont l'édition de génome. Les différents cas d'utilisation permettent de mettre en avant leur utilité face aux enjeux de sécurité alimentaire et d'adaptation au changement climatique, mais les risques liés à l'utilisation des nouvelles techniques d'édition du génome ont été soulignés. Ces utilisations varient selon le type de ressource génétique considéré.

Traditionnellement, les ressources génétiques animales et aquatiques ont bénéficié davantage de projets de sélection faisant intervenir la génomique, et ce, pour des raisons économiques et techniques. Les ressources génétiques de microorganismes ont aussi bénéficié très tôt de programmes de recherche en génomique, cela s'explique par la taille de leur génome, plus petit et donc facile à manipuler. Les ressources génétiques végétales bénéficient aujourd'hui de programmes variés, allant de la caractérisation génétique à l'ajout de caractères intéressants pour l'agriculture. Dernièrement, la recherche intégrant les ressources génétiques forestières entre dans l'ère de la génomique, car les techniques sont plus accessibles en termes techniques et de coût.

Des projets de grande ampleur et interdisciplinaires sont envisagés dans un futur proche pour la recherche de gènes « perdus » des ancêtres des espèces domestiquées et sélectionnées, présentant un intérêt pour comprendre les clefs de l'évolution et de l'adaptation de la vie des plantes sur terre.

Ce rapport est un état des lieux, cependant de nombreuses questions ont été soulevées lors des entretiens et des champs de recherche restent à explorer (cf. annexe 18).

RÉFÉRENCES BIBLIOGRAPHIQUES

Aubertin C. (2018). Le protocole de Nagoya à l'épreuve de la recherche sur la biodiversité, In Hommes-Milieux. Vers un croisement des savoirs pour une méthodologie de l'interdisciplinarité, Adélie Pomade (dir.), Presses Universitaires de Rennes.

Bonneuil C., Thomas F., Petitjean O. (collab.) (2012). Semences : une histoire politique : amélioration des plantes, agriculture et alimentation en France depuis la Seconde Guerre mondiale. Paris: Fondation Charles Léopold Mayer, 216 p.

Chiron G., Chapuis H., Tixier-Boichard M., Restoux G., Rognon X., Lubac-Paye S., Vieaud A., Seigneurin F., Petitjean F., Guemene D. (2018). Quelle stratégie pour une politique de conservation des races locales avicoles ? (Biodiva). Innovations Agronomiques, 63, 357-371.

Christine (ed.), Juhé-Beaulaton D. (ed.), Boutrais Jean (ed.), Roussel B. (ed.) Patrimoines naturels au Sud: territoires, identités et stratégies locales. Paris (FRA) ; Paris : IRD ; MNHN, 53-70. (Colloques et Séminaires). Patrimoines Naturels au Sud : Territoires, Identités et Stratégies Locales : Séminaire, Paris (FRA), 2004. ISBN 2-7099-1560-X

Cirad, INRA, IRD (2011). Lignes directrices pour l'accès aux ressources génétiques et leur transfert. Edition : Délégation à la communication, Cirad.

Duclos A., Bed'hom B., Acloque H., Pain B. (2017). Modifications ciblées des génomes : apports et impacts pour les espèces d'élevage, INRA Prod. Anim 30 (1), 3-18.

FAO (2016). Eléments relatifs à l'accès et au partage des avantages, éléments visant à faciliter la concrétisation au niveau national de l'accès et du partage des avantages dans les différents sous-secteurs des ressources génétiques pour l'agriculture et l'alimentation.

Fondation pour la recherche sur la biodiversité (2017). L'APA Pas à Pas, Mise en œuvre du protocole de Nagoya et des réglementations d'accès aux ressources génétiques et aux connaissances traditionnelles associées et de partage des avantages issus de leur utilisation (APA) dans le cadre des activités de recherche et de développement, 144 pages.

Gallezot G. (2002)."La recherche in silico", In : Chartron G. (sous la dir.) "Les chercheurs et la documentation numérique : nouveaux serviceset usages", Edition du cercle de la Librairie, Collections.

Gaspin Christine (2015). « Les données de la recherche dans le domaine des sciences du vivant : évolution et perspectives à la lumière des nouvelles technologies du numérique et d'exploration du vivant», Présentation à Toulouse.

GIS Biotechnologies vertes (2016). Synthèse de la réflexion prospective et propositions du GIS Biotechnologies Vertes (GIS BV).

Groupe de travail « Cahier des charges informatique, bio-analyse/bioinformatique, bases de données mutations » dans le cadre du Réseau NGS Diagnostic (2016). Recommandations générales pour la gestion informatique des données et des analyses de séquençage à haut débit pour les laboratoires de diagnostic moléculaire de maladies génétiques.

Guillain P.-E. Livoreil B., Silvain J.-F. (2017). Communiqué de la Fondation pour la recherche sur la biodiversité, « COP 13 : biologie de synthèse et séquences numériques au cœur des débats ».

Ifremer (2017). La recherche européenne pour une conchyliculture durable et compétitive., premier bilan un an après le démarrage du projet Vivaldi.

Inra (2014). Rapport du groupe sur les données génomiques et les ressources génétiques « Partage des données relatives aux ressources génétiques et génomiques : états des lieux, analyse stratégique et besoins d'accompagnement », Inra.

Karger Elizabeth (2018). Options for benefit sharing : the case of digital sequence information on genetic resources, Master thesis of Global Chnage Ecology, University of Bayreuth.

Leleux Philippe (2014). Rapport de stage de recherche, Assemblage de génomes à l'aide des réseaux de fonction de coûts, Université Paul Sabatier, Unité MIAT, Inra.

Maxime Rotival. (2011) ; Approches intégrées du génome et du transcriptome dans les maladies complexes humaines. Génétique. Université Paris Sud - Paris XI.

Nations-Unies (1992). Convention sur la diversité biologique. https://www.cbd.int/doc/legal/cbd-fr.pdf

Nations-Unies (2010). Protocole de Nagoya sur l'accès aux ressources génétiques et le partage juste et équitable des avantages découlant de leur utilisation, relatif à Convention sur la diversité biologique. https://www.cbd.int/abs/doc/protocol/nagoya-protocol-fr.pdf

Peterlongo Peter (2016). Lire les lectures : analyse de données de séquençage. Bio-informatique, Université rennes1.

Pinto A., Henriques D., Chávez Galarza J-C., Kryger P., Garnery L., Van der Zee R., Dahle B., Soland-Reckeweg G., De la Rua P., Dall'Olio, R., Carreck, N., Johnston J. (2014). Genetic integrity of the Dark European honeybee (Apis mellifera mellifera) from protected populations: A genome-wide assessment using SNPs and mtDNA sequence data. Journal of Apicultural Research. 53. 269-278. 10.3896/IBRA.1.53.2.08.

Rey Alexandrine (2017). Le traitement de l'information génétique par le droit. L'exemple de l'information liée à la biodiversité, Thèse, Université de Montpellier.

Jacques Saliba (1999). « Le clonage en question : science, éthique, représentation sociale », Socio-anthropologie [En ligne], 5 | 1999, mis en ligne le 22 juillet 2005,

Secrétariat de la Convention sur la diversité biologique (2002). Lignes directrices de Bonn sur l'accès aux ressources génétiques et le partage juste et équitable des avantages résultant de leur utilisation. Montréal: Secrétariat de la Convention sur la diversité biologique.

Smouts M.C. (2005). Du patrimoine commun de l'humanité aux biens publics globaux. In : Cormier Salem Marie-Christine (ed.), Juhé-Beaulaton D. (ed.), Boutrais Jean (ed.), Roussel B. (ed.) Patrimoines naturels au Sud : territoires, identités et stratégies locales. Paris (FRA) ; Paris : IRD ; MNHN, 53-70. (Colloques et Séminaires). Patrimoines Naturels au Sud : Territoires, Identités et Stratégies Locales : Séminaire, Paris (FRA), 2004. ISBN 2-7099-1560-X

Weissenbach J. (2000). Texte de la 27ème conférence de l'Université de tous les savoirs réalisée le 27 janvier 2000, Le séquençage du génome humain : comment et pourquoi.

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ACRONYMES

AHTEG : Groupe spécial d'experts techniques
APA : Accès et partage des avantages juste et équitable découlant de l'utilisation des ressources génétiques
ATTM : accord-type de transfert de matériel (*Standard material transfer agreement*, SMTA)
CDB : Convention sur la diversité biologique
CEA : Commissariat à l'énergie atomique et aux énergies alternatives
CERBM-GIE : Centre Européen de Recherche en Biologie et en Médecine – Groupement d'intérêt économique
CIRAD : Centre de coopération international en recherche agronomique pour le développement
CNRGV : Centre national de ressources génomiques végétales
CNRS : Centre nationale de la recherche scientifique
COP : Conférence des Parties à la Convention sur la diversité biologique
COP-MOP : Conférence des Parties siégeant en tant que réunion des Parties à un protocole
DSII : *Digital Sequence Information,* Information de séquençage numérique
DTA : *Data Transfer Agreement*, Accord de transfert de données
FAO : *Food and Agriculture Organisation*, Organisation des Nations Unies pour l'alimentation et l'agriculture
FRB : Fondation pour la recherche sur la biodiversité
GCRAI : Groupe consultatif pour la recherche agricole internationale (CGIAR, *Consultative Group on International Agricultural Research*)
GIS : Groupement d'intérêt scientifique
INRA : Institut national de recherche agronomique
INSERM : Institut national de la santé et de la recherche médicale
IRD : Institut de recherche pour le développement
ITAI : Instituts techniques agro-industriels
MAT : *Mutually Agreed Terms*, Contrat de commun accord
NGS : *Next generation sequencing*, Séquençage nouvelle génération désignant notamment le séquençage à haut débit
OMS : Organisation mondiale de la santé
PIA : Programme d'investissement d'avenir
PIC : *Prior Informed Consent,* Consentement préalable en connaissance de cause
PIP : *Pandemic Influenza Preparedness*, Préparation en cas de grippe pandémique
RGAA : Ressources génétiques pour l'alimentation et l'agriculture
SBSTTA : Organe subsidiaire chargé de fournir des avis scientifiques, techniques et technologiques pour la mise en œuvre de la CDB
TGCC : Très grand centre de calcul
TIRPAA : Traité international sur les ressources phytogénétiques pour l'alimentation et l'agriculture
WGS : *Whole genome sequecing,* stratégie de séquençage global

GLOSSAIRE

ADN : Acide désoxyribonucléique. Molécules de taille importante contenant les instructions (gènes) et qui constituent les chromosomes (J. Weissenbach, 2000).

ARN : Eacide ribonucléique. Molécule qui transporte l'information contenue dans le patrimoine génétique (ADN) jusqu'aux ribosomes qui sont chargés de la "traduire" en protéines ayant des fonctions précises (Peterlongo, 2016).

Assemblage : Ensemble de séquences approximant le mieux possible la séquence d'un génome (France génomique).

Analyse structurale du génome : Analyse de la structure physique et organisationnelle des molécules d'ADN.

Biologie moléculaire : Étude des macromolécules biologiques comme les acides nucléiques, dont l'ADN, et les protéines. Les techniques rattachées à la biologie moléculaire sont de l'ordre de l'exploration du vivant et de l'informatique (Gallezot, 2002).

Clonage : Reproduction de cellules génétiquement identiques (Atlan, 1998).

CNV : *Copy Number Variation*, forme de polymorphisme qui présent des variations dans le nombre de copies de certaines parties du chromosome.

CRISPR : Acronyme de « Clustered Regularly Interspersed Short Palindromic Repeats », séquences répétées palindromiques courtes (30 bases) regroupées et régulièrement intercalées avec d'autres séquences qualifiées de « spacers », vestiges de séquences de bactériophages ou de plasmides, d'environ 36 bases.

Diversité génétique : La diversité génétique est la résultante de la sélection, la mutation, la migration, la dérive génétique et/ou la recombinaison. Tous ces phénomènes provoquent des changements dans les fréquences de gènes et d'allèles, conduisant à l'évolution des populations (Institut international de ressources phytogénétiques, IPGRI, and Cornell University, 2003).

Electrophorèse : Technique permettant de séparer les fragments d'ADN suivant leur taille (Guesdon, Universalis).

Gène : Unité biologique qui transmet des informations de l'hérédité et contrôle l'apparition d'un trait. Cette définition fait référence à Mendel, botaniste germanophone du XIXème siècle.

Génome : Ensemble des gènes contenus dans les cellules des êtres vivants (Weissenbach, 2000).

Génomique : Étude du génome. Elle se compose de deux volets complémentaires, l'analyse structurale du génome et la génomique fonctionnelle.

Génomique fonctionnelle : Étude de la fonction des gènes, la régulation de leur expression et leurs interactions. Elle s'intéresse aux molécules d'ARNm et/ou aux protéines résultant de l'expression des gènes.

Génotype : Composition génétique particulière d'un organisme.

***In silico* :** Dans le domaine de la biologie moléculaire, les recherches ne sont plus seulement *in vivo* ou *in vitro*, mais ont recours de plus en plus à la modélisation informatique. Le néologisme « *in sillico* » souligne

l'importance des technologies de l'information et de la communication (TIC) dans le développement de cette discipline et en désigne surtout deux champs spécifiques : la génomique et la bio-informatique (Gallezot, 2002).

Marqueur génétique : Caractère mesurable qui peut détecter une variation dans la séquence nucléique ou protéique. Les marqueurs génétiques peuvent être de différentes natures, morphologiques, protéiques (biochimiques), ADN (marqueurs moléculaires) (Institut international de ressources phytogénétiques, IPGRI, and Cornell University, 2003).

Microbiote : Ensemble des microorganismes présents dans un environnement ou un organisme.

Nucléotides : Molécules élémentaires de l'ADN et constitués d'un sucre (le désoxyribose), d'un résidu phosphate et d'une base azotée (adénine, guanine, cytosine et thymine).

Phénotype : Combinaison de caractères individuels résultant d'un génotype et son interaction avec l'environnement. Le phénotypage consiste à observer des organismes et les caractères qu'ils expriment dans leur environnement.

***Polymerase Chain Reaction* (PCR) :** Technique qui permet d'amplifier plusieurs millions de fois tout fragment d'ADN, aussi appelée « photocopieuse à gènes » (Chevassus-au-Louis, Universalis).

Polymorphisme : Concerne n'importe quelle différence entre les individus. On peut l'observer au niveau de l'aspect général des individus, au niveau des protéines, au niveau génétique également (Université Pierre et Marie Curie).

Protéomique : Étude de l'ensemble des protéines (identification, fonction, structure).

Puce à ADN ou de génotypage : Outils de la génomique permettant l'identification de SNP et donc de gènes ou de régions intéressantes du génome.

Séquençage : Le séquençage est l'ensemble des manipulations permettant de déterminer la séquence d'une molécule d'ADN, d'ARN, d'une protéine, etc. (Weissenbach, 2000).

SNP ou « marqueurs moléculaires » : *Single Nucleotide Polymorphism*, différence au niveau d'un nucléotide dans une séquence d'ADN, c'est un type de marqueur moléculaire très utilisé en génomique.

Transcriptomique : Étude de l'ensemble des ARN, souvent plus particulièrement les ARN messagers (ARNm), qui sont utilisés comme intermédiaires pour la production de protéines (Lopez-Maestre, 2017).

ANNEXES

ANNEXE 1: CARACTÉRISTIQUES PROPRES AUX RESSOURCES GÉNÉTIQUES POUR L'AGRICULTURE ET L'ALIMENTATION (ANNEXE E DU DOCUMENT CGRFA-14/13/RAPPORT)

		Groupe de travail ressources zoogénétiques	Groupe de travail ressources.	Groupe de travail ressources. phytogénétiques
A. Rôle des Ressources génétiques pour l'alimentation et l'agriculture dans la sécurité alimentaire	A.1 Les ressources génétiques pour l'alimentation et l'agriculture font partie intégrante des systèmes de production agricole et alimentaire et sont essentielles pour parvenir à la sécurité alimentaire et au développement durable du secteur alimentaire et agricole		+	+
	A.2 Les ressources génétiques des plantes, animaux, invertébrés et microorganismes tissent au sein des écosystèmes agricoles un réseau interdépendant de diversité génétique.		+	
B. Rôle de la gestion humaine	B.1 L'existence de la plupart des ressources génétiques pour l'alimentation et l'agriculture est étroitement liée à l'activité humaine et nombre d'entre elles peuvent être considérées comme des formes de ressources génétiques modifiées par l'homme.		-	
	B.2 Le maintien et l'évolution de nombreuses ressources génétiques pour l'alimentation et l'agriculture supposent une intervention constante de l'homme, et leur utilisation durable pour la recherche, le développement et la production est un moyen important d'assurer leur conservation.	+	-	
C. Échanges internationaux et interdépendance	C.1 Tout au long de l'histoire, les ressources génétiques pour l'alimentation et l'agriculture ont fait l'objet d'échanges intenses entre communautés, pays et régions, souvent durant de longues périodes, et	+	-	+

	une large part de la diversité génétique aujourd'hui utilisée dans l'alimentation et l'agriculture est d'origine exotique.			
	C.2 Les pays sont interdépendants en matière de ressources génétiques pour l'alimentation et l'agriculture ; ils fournissent certaines ressources génétiques et en reçoivent d'autres.		+	
	C.3 Les échanges internationaux de ressources génétiques pour l'alimentation et l'agriculture jouent un rôle fondamental dans le fonctionnement du secteur, et ils devraient encore se développer.	+	+	+
D. Nature du processus d'innovation	D.1 En matière de ressources génétiques pour l'alimentation et l'agriculture, le processus d'innovation suit généralement un schéma progressif et il est issu des contributions apportées par une large gamme d'acteurs, notamment les communautés autochtones et locales, les agriculteurs, les chercheurs et les obtenteurs en des lieux et à des moments différents.	+	+	+
	D.2 La plupart des produits issus des ressources génétiques pour l'alimentation et l'agriculture ne sont pas développés à partir d'une seule ressource génétique mais à partir de plusieurs ressources génétiques pour l'alimentation et l'agriculture à différentes étapes du processus d'innovation.		-	+
	D.3 La plupart des produits mis au point à l'aide de ressources génétiques pour l'alimentation et l'agriculture peuvent à leur tour servir de ressources génétiques pour de nouveaux travaux de recherche-développement, d'où la difficulté d'opérer une distinction nette entre les		+	+

	fournisseurs et les destinataires de ressources génétiques pour l'alimentation et l'agriculture.			
	D.4 De nombreux produits agricoles sont commercialisés sous une forme permettant de les utiliser comme ressources biologiques et comme ressources génétiques.	-	+	
E. Détenteurs et utilisateurs de ressources génétiques pour l'alimentation et l'agriculture	E.1 Les ressources génétiques pour l'alimentation et l'agriculture sont détenues et utilisées par des parties prenantes nombreuses et variées. Il existe des communautés distinctes de fournisseurs et d'utilisateurs qui interviennent dans les différents sous-secteurs des ressources génétiques pour l'alimentation et l'agriculture.	+	-	+
	E.2 Les différentes parties prenantes qui gèrent et utilisent les ressources génétiques pour l'alimentation et l'agriculture sont interdépendantes.		+	
	E.3 Une part importante des ressources génétiques pour l'alimentation et l'agriculture est détenue par le secteur privé.	+	-	
	E.4 Une part importante des ressources génétiques pour l'alimentation et l'agriculture est détenue, et est accessible, ex situ.	-	-	
	E.5 Une part importante des ressources génétiques pour l'alimentation et l'agriculture est conservée in situ et au niveau de l'exploitation dans diverses conditions financières, techniques et juridiques.	+	+	
F. Pratiques en matière d'échanges de ressources génétiques pour l'alimentation et l'agriculture	F.1 Les ressources génétiques pour l'alimentation et l'agriculture sont échangées au titre de pratiques établies, dans des communautés existantes de fournisseurs et d'utilisateurs.	+	+	+

	F.2 La recherche-développement engendre d'importants transferts de matériel génétique entre différentes parties prenantes, tout au long de la chaîne de valeur.	+	-	
G. Avantages découlant de l'utilisation des ressources génétiques pour l'alimentation et l'agriculture	G.1 Globalement, les avantages apportés par les ressources génétiques pour l'alimentation et l'agriculture sont très importants, mais il est difficile d'estimer, au moment de la transaction, les avantages attendus de l'utilisation d'un échantillon déterminé de ressources génétiques pour l'alimentation et l'agriculture.		+	+
	G.2 L'utilisation des ressources génétiques pour l'alimentation et l'agriculture peut aussi apporter d'importants avantages non monétaires.		+	
	G.3 L'utilisation des ressources génétiques pour l'alimentation et l'agriculture peut aussi apporter d'importants avantages non monétaires.		+	

Note : Parmi les caractéristiques recensées par le Groupe de travail technique ad hoc sur l'accès aux ressources génétiques pour l'alimentation et l'agriculture et le partage des avantages en découlant, les groupes de travail techniques intergouvernementaux sur les ressources phytogénétiques et zoogénétiques et sur les ressources génétiques forestières ont mis en évidence les caractéristiques qui, pour leurs sous-secteurs respectifs, présentent un intérêt particulier (marquées d'un signe [+] dans le tableau ci-dessus) et celles qui présentent un intérêt moindre ou ne présentent pas d'intérêt particulier (marquées d'un signe [-]).

ANNEXE 2 : LISTE DES STRUCTURES MOBILISÉES POUR L'ENQUÊTE

Structures	Statut
ACTALIA	Institut technique
ACTIA	Réseau d'instituts techniques
Allice	Coopérative d'élevage
Anses	EPCA
Apis-gene	Société
Bayer	Entreprise semencière
Biogemma	Société
Cirad	Centre de recherche
CNIEL	Filière lait
CNRS	Centre de recherche
Confédération paysanne	Syndicat
Dupont	Entreprise biotechnologie
Florimond-Desprez	Entreprise semencière

Genoplante-Valor SAS	Société PI
Genopole	Biocluster
Geves	GIP
GNIS	Interprofession
Idele	Institut technique
Ifip	Institut technique
Ifremer	EPIC
IFV	Institut technique
Inra	Centre de recherche
Institut de biologie François Jacob	Institut de biologie
IRD	Centre de recherche
Lesaffre	Syndicat
Limagrain	Entreprise
Ministère de l'agriculture et de l'alimentation	Institution
Ministère de l'environnement	Institution
Ministère de la recherche	Institution
MNHN	Centre de recherche
RAGT	Entreprise semencière
Réseau semence paysanne	Association
Spygen	Société
Sysaaf	Syndicat
Unicoque	Coopérative
URGI	Plateforme bioinformatique

ANNEXE 3 : LES TRAITÉS INTERNATIONAUX

- La Convention pour la diversité biologique (CDB)

La Convention pour la diversité biologique est signée le 5 juin 1992 au Sommet de la Terre à Rio. Le Secrétariat de la CDB est basé à Montréal au Canada. Il a pour mission d'aider les gouvernements à mettre en œuvre la Convention sur la diversité biologique, d'organiser des réunions, de recueillir et diffuser des informations. La CDB est considérée comme le principal instrument international relatif au Développement durable (DD) du fait de ses trois objectifs. Un de ses principes fondamentaux est la souveraineté nationale de chaque État sur les ressources génétiques de son territoire.
La CDB définit de grands objectifs pour la protection de la biodiversité avec notamment pour intention de partager le cout de la préservation de la biodiversité entre les détenteurs et les utilisateurs des ressources génétiques.

Les trois objectifs de la Convention sont les suivants :

- Objectif 1 : la conservation de la diversité biologique.
- Objectif 2 : l'utilisation durable de la diversité biologique.
- Objectif 3 : l'accès et le partage juste et équitable des avantages (APA) découlant de l'utilisation des ressources génétiques.

Ces objectifs découlent de deux principes, d'une part, la souveraineté nationale de chaque État sur les ressources génétiques de son territoire et d'autre part la propriété des détenteurs sur leurs connaissances traditionnelles. Ce second principe ayant pour corollaire un accès facilité à ces connaissances mais un accès

conditionné, d'une part, au consentement préalable des détenteurs, et d'autre part, à des conditions convenues d'un commun accord entre détenteurs et utilisateurs.

Le Protocole de Nagoya, constitue le cadre juridique de la mise en œuvre du troisième objectif cité plus haut. Il est contraignant, et oblige les signataires à le transcrire dans leurs législations. Il vise :
Le partage juste et équitable des avantages découlant de l'utilisation des ressources génétiques est un mécanisme de reconnaissance des droits sur les ressources génétiques et les connaissances traditionnelles associées. Il est donc un instrument de la conservation de la diversité biologique et de l'utilisation durable de ses éléments constitutifs.

- Le Protocole de Nagoya

Le Protocole de Nagoya relatif à l'accès aux ressources génétiques et au partage juste et équitable des avantages découlant de leur utilisation a été adopté en 2010 lors de la 10e Conférence des Parties de la Convention sur la diversité biologique. Il précise le cadre international du mécanisme d'accès et de partage juste et équitable des avantages (APA) que les États signataires ont la responsabilité de traduire dans leur droit national. Ratifié par 70 États, le Protocole de Nagoya est entré en vigueur le 12 octobre 2014. Les informations sur l'état de sa mise en œuvre par les États sont disponibles sur le site du Centre d'échange d'informations sur l'accès et le partage des avantages (*ABS-Clearing House Mechanism*).

- Le SBSTTA (*Subsidiary Body on Scientific, Technic and Technological Advice*)

Le SBSTTA est un organe subsidiaire international chargé de fournir des avis scientifiques, techniques et technologiques relatifs à la mise en œuvre de la Convention pour la diversité biologique (CBD).

Le SBSTTA s'appuie sur des "points focaux" nationaux pour solliciter les experts. Dans ce cadre, il développe des liens entre les institutions "point focal national" et les organismes experts nationaux et supranationaux compétents. Les points focaux assurent aussi la liaison avec le Secrétariat de la CBD au nom des États Membres.

Le rôle de point focal SBSTTA-France est assuré conjointement par le MNHN (Muséum national d'Histoire naturelle) et la FRB. Ces deux institutions sont chargées de préparer les éléments de langage de la délégation française à la CDB pilotée par le Ministère chargé de l'environnement et le Ministère des affaires étrangères.

- Le Centre d'échange d'informations sur l'accès et le partage des avantages (Centre d'échange d'informations APA)[70]

C'est une plateforme d'échange d'informations sur l'accès et le partage des avantages instaurée par l'article 14 du Protocole pour faciliter la mise en œuvre du Protocole de Nagoya. Il favorise la sécurité et la transparence juridiques des procédures d'accès et de partage des avantages et assure en outre le contrôle de l'utilisation des ressources génétiques tout au long de la chaîne de valeur, notamment à l'aide d'informations pertinentes reconnues à l'échelle internationale. En hébergeant des informations pertinentes relatives à l'APA, le Centre d'échange APA offrira l'opportunité de mettre en relations des utilisateurs et les fournisseurs de ressources génétiques et des connaissances traditionnelles reliées.

- Le TIRPAA :

Le Traité international sur les ressources phytogénétiques pour l'alimentation et l'agriculture a été adopté par la 31e réunion de la Conférence de l'Organisation des Nations Unies pour l'alimentation et l'agriculture, le 3 novembre 2001. Il est entré en vigueur en 2004.

Le Traité vise à :

  - reconnaître l'énorme contribution des agriculteurs à la diversité des cultures qui nourrissent le monde ;

[70] Site officiel de l'ABS *clearing house* : https://absch.cbd.int/

- mettre en place un système mondial permettant de fournir un accès aux matériels phytogénétiques aux agriculteurs, aux sélectionneurs de végétaux et aux scientifiques ;
- s'assurer que les bénéficiaires partagent les avantages qu'ils tirent de l'utilisation de ces matériels génétiques avec les pays d'où ils proviennent

Les questions de statuts juridiques sur les ressources phytogénétiques s'étendent progressivement aux autres ressources génétiques pour l'agriculture et l'alimentation (animales, microorganismes, etc.). Cependant il n'existe pas de traité pour les ressources animales car les pays n'ont pas souhaité mutualiser leurs ressources. Pour ce qui est des microorganismes, les questions de sécurité sanitaire et de manipulations génétiques restent primordiales et imposent la mise en place de règles spécifiques.

- Le règlement européen n° 511/2014 du 16 avril 2014 transpose le troisième volet du protocole de Nagoya.

Il impose aux utilisateurs

- des obligations de tracer, documenter, conserver et transférer les informations sur les ressources génétiques et connaissances traditionnelles associées.
- de posséder un certificat de conformité internationalement reconnu du pays fournisseur partie au Protocole de Nagoya et ayant adopté des mesures sur l'APA, ou, à défaut de certificat, des éléments précis de traçabilité et d'attestation du respect des obligations, l'ensemble de cette documentation devant être conservé vingt ans après la fin de l'utilisation.
- de déclarer qu'ils ont fait preuve de la diligence nécessaire (« due diligence ») à deux étapes clé : à la réception de financements externes pour les travaux de recherche (que ces financements soient privés ou publics) et lors du développement final d'un produit mis au point à partir d'une ressource génétique ou d'une connaissance traditionnelle associée.

Le règlement impose aux États membres de :

- désigner une ou plusieurs autorités compétentes pour assurer le respect de ces règles,
- organiser des contrôles sur les utilisateurs
- et établir des règles nationales en matière de sanctions.

Ce règlement et son règlement d'exécution ont nécessité l'adoption de dispositions complémentaires au niveau national, notamment pour définir les sanctions pénales applicables et pour rendre les dispositions du règlement européen applicables dans les territoires français d'outre-mer à statut particulier au regard du droit de l'Union européenne (Nouvelle-Calédonie, Polynésie française, Wallis-et-Futuna). En France ces dispositions ont été prises par le titre IV de la loi du 8 août 2016 qui porte également sur la mise en œuvre des deux premiers volets du Protocole. Cette loi a également supprimé le régime spécifique qui était applicable au parc amazonien de Guyane. Le régime général de l'APA s'applique donc à ce territoire.

ANNEXE 4 : LES NOUVELLES TECHNIQUES DE SÉLECTION DES PLANTES (NPBT)

Les nouvelles techniques de sélection des plantes (NPBT)

- La cisgénèse / Intragénèse
- Le greffage sur porte-greffe transgénique
- L'agro-infiltration
- La méthylation de l'ADN assistée par ARN (ou « RdDM » selon l'acronyme anglais « *RNAdependent DNA methylation* »)
- La sélection inverse
- La génomique de synthèse
- La mutagénèse dirigée par oligonucléotides (ou « ODM » selon l'acronyme anglais de « *Oligonucleotide Directed Mutagenesis* »)
- Les nucléases dirigées (ou « SDN » selon l'acronyme anglais « *Site Directed Nuclease* »)

Les techniques d'édition de génome (ODM et SDN)

Techniques d'édition de génome (ODM et SDN): repose sur les mécanismes de réparation de l'ADN (jonction des extrémités non-homologues, mécanismes de recombinaison homologue er réparation des mésappariements).

Mutagénèse dirigée par Oligonucléotide (ODM) :

- Modification précise et choisie d'une ou de quelques paires de bases du génome de la plante.
- Oligonucléotide homologue à la séquence cible, mais porte les modifications voulues à recopier dans le génome.
- Mécanismes de réparation de l'ADN.
- Mutations crées transmises à la descendance.

Nucléase dirigée :

- Enzymes de type nucléases dirigées (SDN): capacité de couper les brins d'ADN au niveau d'une cible prédéfinie après hybridation spécifique de cette nucléase.
- Coupure active les mécanismes naturels de réparation de l'ADN pour créer une mutation non prédéfinie (insertions/délétions) ou modifier précisément l'ADN par insertion spécifique d'une matrice ADN.

4 familles de nucléases dirigées utilisées dans les laboratoires de recherche :

- Méganucléases, ZFNs, TALENs, CRISPR/Cas9.
- Modifications transmises à la plante.

Source : CTPS/CS/NBT/022018 NBTs et Techniques d'édition du génome : Impacts potentiels sur l'offre variétale et les activités du CTPS.

ANNEXE 5 : CONCLUSIONS PRINCIPALES DE L'AHTEG POUR LE SBSTTA EN JUILLET 2018

Lors de la réunion du mois de juillet 2018, l'AHTEG a émis un cadre de recommandations concernant les données de séquençage de ressources génétiques pour la préparation de la prochaine Conférence des parties qui aura lieu au mois de novembre 2018. Il faut noter que des questions restent en suspens et que la majorité des décisions demeurent à négocier. Parmi ces décisions :

- Le terme « informations de séquençage numérique n'est pas le plus approprié pour désigner les divers types d'informations sur les ressources génétiques mais sert de substitut provisoire jusqu'à l'adoption d'un nouveau terme ;

- L'information sur les données de séquençage numérique sur les ressources génétiques joue un rôle important pour la conservation de la diversité génétique et l'utilisation durable de ses éléments constitutifs ;

- L'utilisation de l'information de séquençage numérique sur les ressources génétiques et l'accès libre à cette information contribue à la recherche scientifique ;

- Les pays ne sont pas égaux face à l'utilisation, la production et l'analyse de l'information de séquençage numérique. Les parties sont encouragées à renforcée les capacités et le transfert de technologies ;

- Les Parties, les autres gouvernements, les peuples autochtones, les communautés locales et les parties prenantes concernées sont invitées à présenter leurs points de vue et des informations susceptibles de clarifier le concept d'information de séquençage numérique ;

- Les Parties et autres gouvernements sont invités à présenter les dispositions législatives internes et les autres mesures relatives à l'information de séquençage numérique ;

- La production, l'utilisation et la gestion de l'information de séquençage numérique sont dynamiques et influencées par les évolutions technologiques et scientifiques. Il est nécessaire d'analyser les faits survenant dans le domaine aux fins d'examiner les répercussions potentielles sur les objectifs de la CBD et le Protocole de Nagoya.

ANNEXE 6 : SCHÉMA SIMPLIFIÉ DE L'OBJET D'ÉTUDE DE LA BIOLOGIE MOLÉCULAIRE

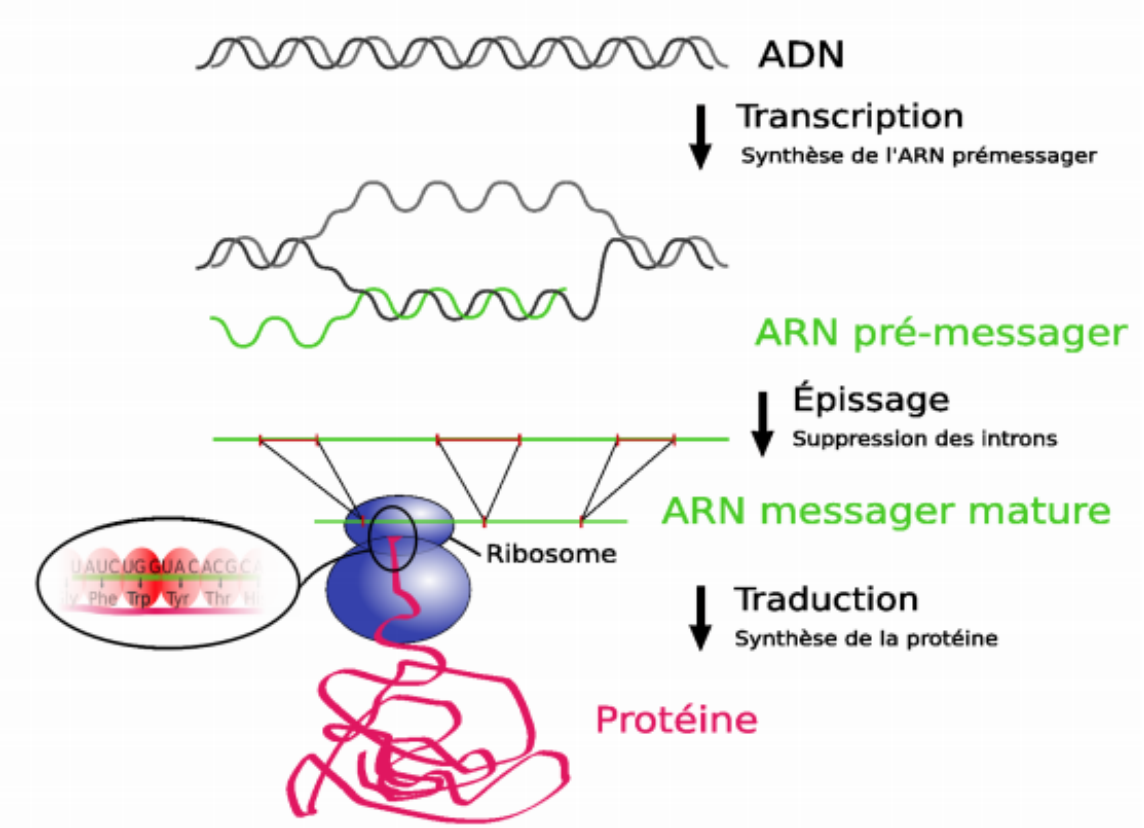


FIGURE 1.1 – Le *dogme central* de la biologie moléculaire

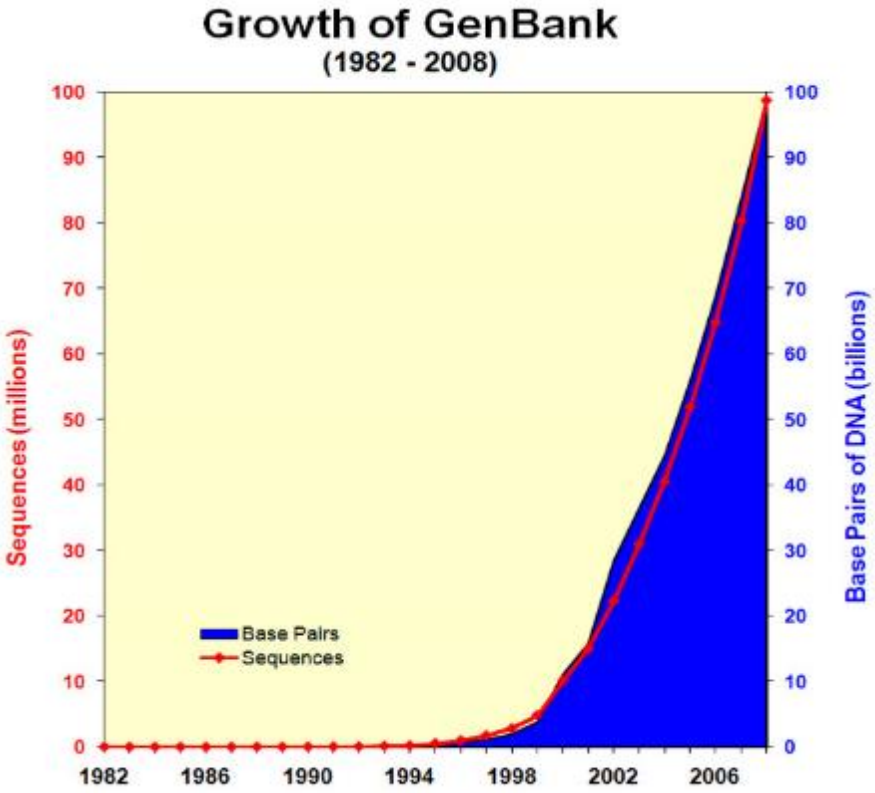
Source : Peterlongo, 2016.

ANNEXE 7 : ÉVOLUTION DES TECHNOLOGIES DE SÉQUENÇAGE

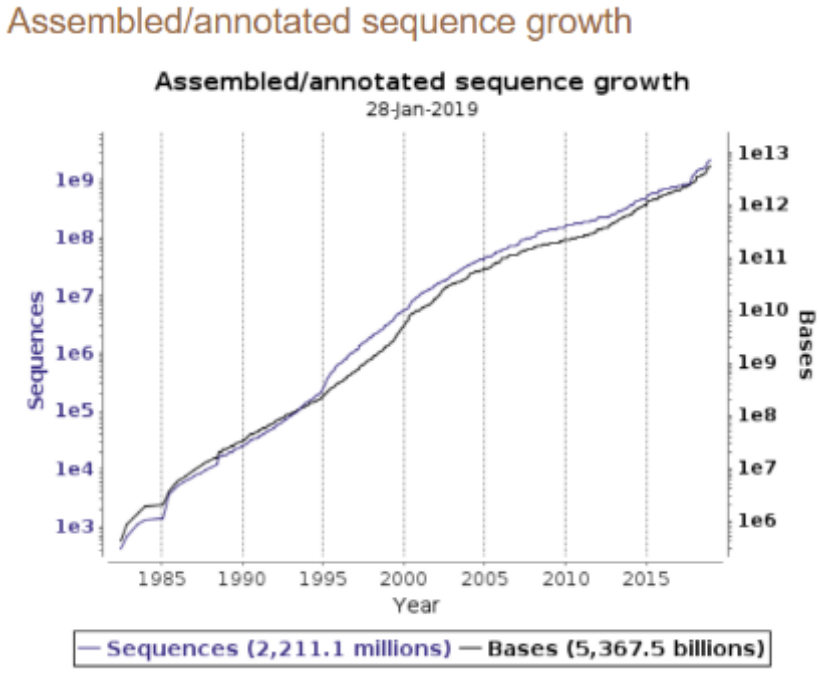
Société	Thermo (ABI)	Illumina	Pacific Biosciences (PacBio)	Oxford Nanopore Technologies (ONT)
Génération	Méthode Sanger *Référence Initiale*	NGS (2ème génération)	NGS (génération 2.5 ou 3...)	NGS (3ème ou 4ème génération...)
Capacité de lecture	1 000 pb	100-300 pb (séquence courte)	+ 20 000 pb (séquence longue)	+ de 100 000 pb (séquence longue)
Détection	Optique : fluorescence			
		Optique: fluorescence	Optique: fluorescence	Ampérométrique (mesure de la variation de l'intensité de courant à potentiel imposé)
Capacité de séquençage maximum	1 Mbases	3 TBases par jour (NovaSeq)	10 à 20 Gbases par SMRT (Single Molecule, Real-Time) Nombre de SMRT par jour à voir avec utilisateurs	Objectif: 1 à 2 TBases par jour (PromethION)
Technologie		Séquençage par synthèse (technologie SBS)	Séquençage par synthèse	Nanopore - Temps réel
Taux d'erreur	< 0,1%	< 1%	10 - 15% (erreurs aléatoires)	≤ 10% (erreurs systématiques)
Particularités		Nécessite peu d'ADN en comparaison avec les technologies PacBio et ONT Amplification des banques avant séquençage	Séquençage de la molécule unique: pas de PCR Les longues lectures requièrent de l'ADN de haut poids moléculaire	Séquençage de la molécule unique: pas de PCR Les longues lectures requièrent de l'ADN de haut poids moléculaire Technologie permettant de séquencer l'ARN directement sans passer par l'ADN complémentaire
Coût investissement machine Haut Débit		NovaSeq: élevé (1 million de $)	Sequel: moyen (environ 400 k$)	PromethION: faible

Caractéristiques principales des technologies disponibles pour le séquençage au sein du Genoscope, (Entretien avec Arnaud Lemainque, responsable du Genoscope, août 2018)

ANNEXE 8 : CROISSANCE DU NOMBRE DE SÉQUENCE STOCKÉE DANS GENBANK (NCBI)

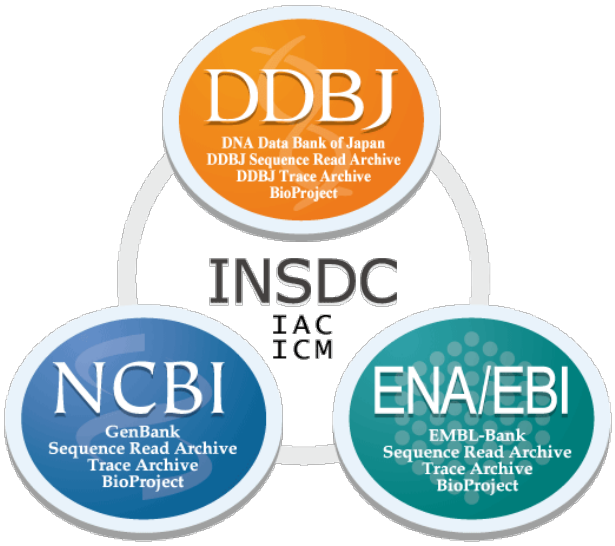


« En décembre 2018, GeneBank stockait 211 millions de séquences correspondant à 285 milliards de bases » (https://www.ncbi.nlm.nih.gov/genbank/statistics/).



Croissance du nombre de séquences annotée à l'ENA, EMBL-EBI
Source : https://www.ebi.ac.uk/ena/about/statistics

ANNEXE 9 : CONSTITUTION DE LA COLLABORATION INTERNATIONALE DES BASES DE DONNÉES DE NUCLÉOTIDES (INSDC)



ANNEXE 10 : LISTE DES PLATEFORMES DE GÉNOMIQUE ET DE BIOINFORMATIQUE EN FRANCE

Nom de la plateforme	Appartenant à France génomique	Label GIS Ibisa	Certification ISO9001	Certification NFX 50-900	Autres certifications
Biogenouest à Rennes		■	■		
CGFB à Bordeaux	■	■	■	■	
eBIO en Ile de France	■		■		
GenoBIRD	■	■	■		
GenomEast à Strasbourg	■	■	■	■	■
Genomic Paris Centre		■	■	■	
GenoScreen à Lille			■		
GenoSol			■		
Gentyiane à Clermont-Ferrand	■	■	■	■	■
GeT-Genotoul à Toulouse	■	■	■	■	■
GeT-PlaGe (bioinfo)	■	■	■	■	■
GPTR Genotypage			■		
Institut de Génomique (ancien CNS et CNG) regroupe le Genoscope et le CNRGH	■				■
MIGALE en Ile de France	■		■		
Microscope	■		■	■	
Montpellier MGX Genomix	■	■	■		
PACA-Bioinformatique	■	■			
Plateforme de séquençage de l'I2BC		■	■		
POPS de l'iPS2 en Ile de France	■	■	■		
South Green		■			
TGML à Marseille	■	■	■	■	
Tilling et recherche translationnelle		■	■		
UCAGenomiX Génomique fonctionnelle Nice-Sophia Antipo	■	■	■		
URGI en Ile de France	■		■		

Plateformes spécialisées dans le génome humain					
CGFB de Bordeaux					
Bilille à Lille (Bioinformatique)					
BIOINFO- Curie					
BIOINFO Pasteur					
Biomics à Grenoble					
Genomic Paris Centre IBENS					
Microbiogénomique à Marseille					
PRABI – Rhône-Alpes					
ProfileXpert à Lyon					
P3S à Paris					
Réseau de plateforme LIGAN à Lille					

ANNEXE 11 : DESCRIPTION DES PRINCIPAUX PÔLES DE GÉNOMIQUE EN FRANCE

Le Genopole[71] regroupe des laboratoires, des centres de recherche et des entreprises travaillant dans le domaine de la biotechnologie. Ce regroupement est dédié à la recherche en génomique, génétique et biotechnologie. Les quatre axes stratégiques concernent la génomique pour l'environnement et la santé, la thérapie génique, la biotechnologie industrielle et la bioinformatique. Le Genopole est aujourd'hui le troisième biocluster européen en termes d'équipement et de volume de données produites. Ce volume est en croissance constante et engendre un besoin de bioinformaticiens pour traiter ces données.

Le Genoscope est l'ancien Centre national de séquençage (CNS). Il a intégré l'Institut de génomique du CEA il y a une dizaine d'années, qui est le partenaire et coordinateur de France génomique. Il constitue aujourd'hui un des plus grands pôles de séquençage sur le territoire français. Il repose sur un système de gestion de l'information de laboratoire qui permet le suivi des activités réalisées sur l'échantillon jusqu'à la production de séquence. Le stockage des données générées par le Genoscope est assuré par le Très grand centre de calcul (TGCC), mis à disposition dans le cadre de France Génomique.

Le pôle de génomique GenoToul de Toulouse fédère des plateformes de bioinformatique, éthique, cohorte, biostatistique, imagerie, protéomique, phénotypage, bio banque, etc. L'Inra est responsable de la bioinformatique et le CNRS est responsable de la protéomique et de l'imagerie. L'infrastructure de stockage du pôle a une capacité de 3 Pétaoctets. GenoToul concentre environ 200 banques de données indexées.

Get-PlaGe est la plateforme de génomique et de transcriptomique reconnue et utilisée à l'échelle nationale. Elle propose notamment les techniques de séquençage nouvelle génération (NGS) haut débit ou séquençage Sanger, de transcriptomique, de génotypage et d'analyses bioinformatiques et statistiques de données.

Le Centre national ressources phytogénétiques (CNRGV) de Toulouse est un centre de ressources biologiques et génomiques dédié aux ressources génétiques de plantes

[71] Créé en 1998 par l'État, la Région Ile-de-France et l'AFM-Téléthon (Association française contre les myopathies), le Genopole est situé à Evry en Essonne. Il réunit 86 entreprises de biotechnologies, 19 laboratoires académiques de recherche, 24 plateformes technologiques mutualisées (site officiel du Genopole).

modèles et des plantes cultivées. Plus de 40 espèces sont étudiées et près de 22 millions d'échantillons biologiques sont stockés. Les outils proposés dans ce centre permettent de caractériser la biodiversité végétale et comprendre comment les plantes s'adaptent à leur environnement à travers l'analyse de leurs génomes. Le centre est impliqué notamment dans les projets de séquençage complet et d'annotation du génome du blé[72] ou de la canne à sucre. Il concentre aussi les outils d'analyse du génome les plus récentes dont le projet « *Catch my interest* » est le plus emblématique. Les résultats de ces projets constituent des innovations pour l'amélioration des variétés en permettant une meilleure compréhension de la biologie des espèces.

Le pôle MGX-Montpellier GenomiX regroupe quatre plateaux de génomique de Montpellier : le plateau UM2, le plateau IRB, le plateau de génotypage et le plateau IFG/IGH (spécialisé dans le séquençage très haut débit, la biostatistique et la bioinformatique). Ce dernier assure l'étude du transcriptome[73] de nombreuses espèces végétales, animales et de procaryotes.

ANNEXE 12 : RÉFLEXIONS DU GROUPE DE TRAVAIL MOBILISÉ POUR CETTE ÉTUDE SUR LA TERMINOLOGIE « DONNÉES DE SÉQUENÇAGE » OU « INFORMATION NUMÉRIQUE DE SÉQUENÇAGE »

Sur la formulation : « Information de séquençage *numérique* » :

C'est la donnée qui est numérique, pas le séquençage. Donc privilégier : « Donnée/Information *numérique* de séquençage ».

Sur les termes « information » VS « donnée » :

On ne peut opposer « information » et « donnée » mais il est utile de les distinguer.

Le terme « information » est plus général que « donnée ». Mais toute information peut se traduire en données et toute donnée est une information. Il est donc préférable d'utiliser le terme « données », plus restrictif.

Il est plus précis de parler de « donnée » que d' « information » car une donnée peut se décrire avec un standard (unité de quantité, contrôle qualité, etc.) ce qui n'est pas le cas d'une information.

Une donnée peut être considérée comme une représentation numérique d'une information.

Cependant, si on adjoint le terme « numérique » à « information », l'ambiguïté est levée, « information numérique » paraît être nécessairement une donnée.

Sur le terme « numérique » :

Il faut garder « numérique » pour limiter les formes de données concernées.

Sur les termes « séquençage » VS « séquence » :

En considérant la typologie : donnée brute (issue directement du séquenceur) / donnée curée / donnée analysée, les données de séquence pourraient constituer un ensemble plus vaste que les données de séquençage, car intégrant les séquences annotées, etc. Dans l'esprit de la discussion, il n'y a pas lieu de restreindre ce périmètre. Donc, il est préférable d'utiliser le terme « séquence » à « séquençage ».

Les données de séquences brutes représentent une information élémentaire qui n'est pas utilisable sans un contexte documenté (les métadonnées).

Sur la formulation « *sur* les ressources génétiques » :

[72] http://presse.inra.fr/Communiques-de-presse/assemblage-de-la-sequence-du-genome-complet-du-ble-accessible-sur-une-plateforme-Inra

[73] Le transcriptome est l'ensemble des ARN issus de la transcription du génome.

On fait des recherches *sur* les ressources génétiques mais on séquence *des* ressources génétiques. Le terme « sur » peut induire des ambigüités. Par exemple, si je dis : « je séquence des microorganismes associés à une plante (qui est la ressource génétique) », cela équivaut à dire : « je fais du séquençage sur cette ressource génétique » et non pas « de cette ressource génétique ». Il est donc préférable d'utiliser « *de* ressources génétique ».

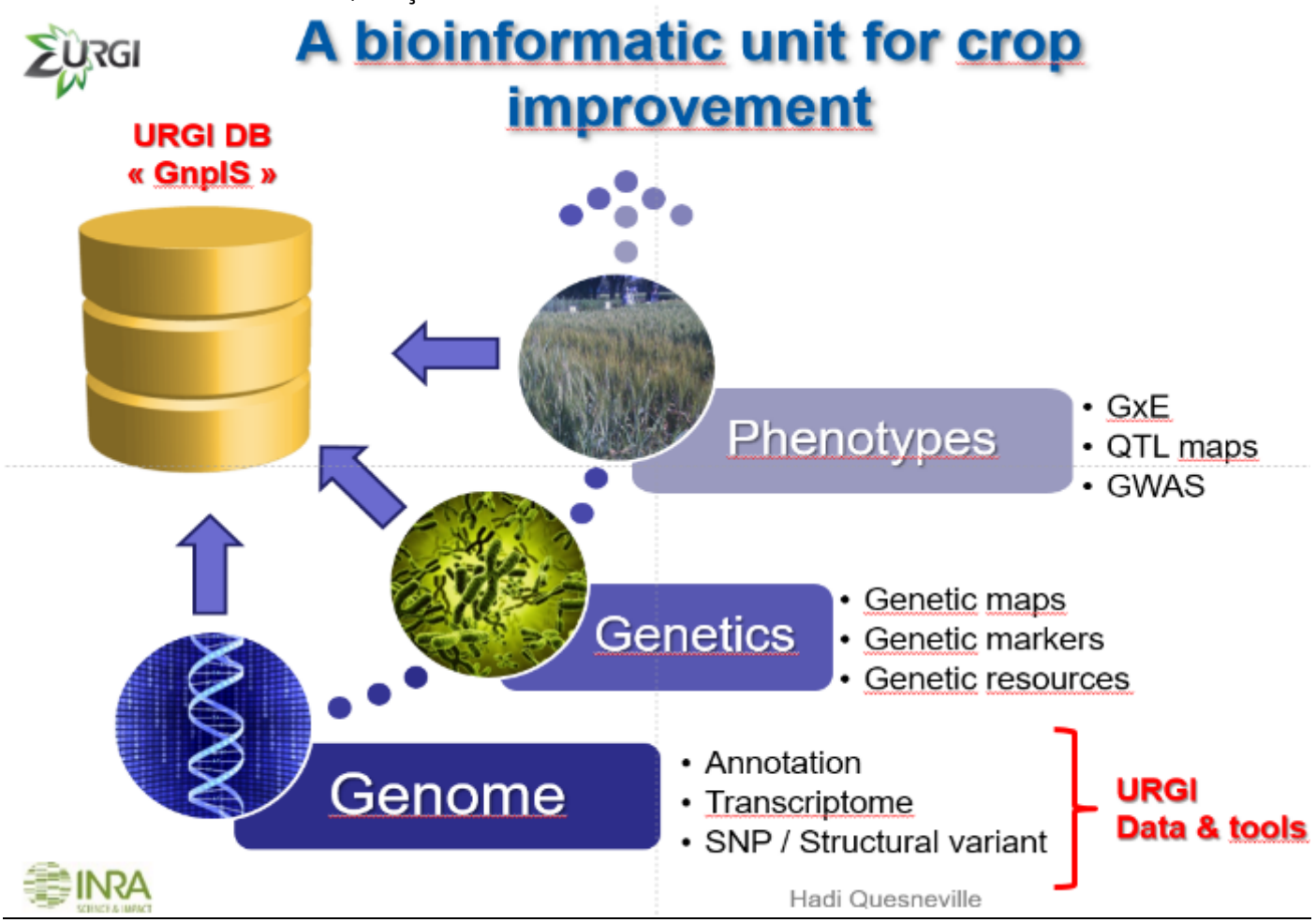
En conclusion :
Il est préférable d'utiliser la terminologie suivante : données numériques de séquences de ressources génétiques.
En anglais, cela pourrait se traduire par :
Digital sequence data
Digital data of genetic resource sequences

ANNEXE 13 : TYPOLOGIE RÉALISÉE PAR LE GROUPE DE TRAVAIL SUR LE PARTAGE DES DONNÉES RELATIVES AUX RESSOURCES GÉNÉTIQUES ET GÉNOMIQUES : ÉTATS DES LIEUX, ANALYSES STRATÉGIQUES ET BESOINS D'ACCOMPAGNEMENT DE L'INRA

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
elaborée	variants structuraux	VCF (V4.1+)
elaborée	patrons de méthylation	bed ?
elaborée	annotations des gènes	gff3, asn, embl
elaborée	orthologues, paralogues, familles de gènes	Tables xml
elaborée	cartes (génétiques, QTLs, physiques)	acp, text formatted
elaborée	données passeport populations/souches	Voir bases FAO
elaborée	données passeport croisements temporaires	?
elaborée	données passeport banques génomiques	?

ANNEXE 14 : LE SYSTÈME D'INFORMATION DE LA PLATEFORME BIOINFORMATIQUE URGI POUR LA GESTION DES DONNÉES DE SÉQUENÇAGE



ANNEXE 15 : LES RÉGLEMENTATIONS SUR LES DONNÉES

Conséquences des droits de propriété intellectuelle et de brevetabilité sur l'accès aux données :
En 1961 est signé la Convention de l'UPOV[74] qui établit le Certificat d'obtention végétale (COV). Le but de la convention est de protéger les obtentions végétales au moyen d'un système harmonisé de propriété industrielle, afin d'encourager le développement de nouvelles variétés végétales. Le COV est un titre de propriété intellectuelle permettant l'accès à l'information génétique, autorisant le croisement et la sélection de variétés couvertes par ce système de protection individuelle à condition que les variétés qui en sont issues en soient suffisamment distinctes (le « D » des critères DSH à remplir pour l'inscription de ces variétés au catalogue et obtention d'un COV).

Par ailleurs, la directive relative à la protection des inventions biotechnologiques du 6 juillet 1998 énonce qu'une matière biologique isolée de son environnement naturel ou produite à l'aide d'un procédé technique peut être l'objet d'une invention, même lorsqu'elle préexistait à l'était naturel. La logique technique prédomine donc dans le droit des brevets.

Concernant les bases de données bio-informatiques qui attestent d'investissements, il existe une directive[75] qui permet leur protection juridique et l'interdiction d'extraire ou de réutiliser leur contenu.

La question de l'articulation entre propriété intellectuelle, droit des brevets et principes de la CDB se pose à nouveau dans un contexte de numérisation de l'information génétique. Le partage juste et équitable des avantages découlant de l'exploitation des ressources génétiques doit être réalisé en tenant compte de tous les droits sur les ressources et les techniques.

Les négociations sur ce sujet visent à établir un équilibre entre promotion du libre accès aux données, la recherche d'une réglementation sur l'accès aux données et les droits de propriété intellectuelle et de brevetabilité.

Depuis l'entrée en vigueur du Protocole de Nagoya en octobre 2014, les ressources biologiques, dont sont issues les séquences numériques qui proviennent des États signataires ayant mis en place des mesures d'accès aux ressources génétiques, sont couvertes par des obligations d'APA. Les **accords de transfert de matériels** (ATM) peuvent stipuler des conditions particulières relatives à l'utilisation envisagée des ressources génétiques.

En d'autres termes, les négociations sur l'accès aux ressources génétiques et donc au matériel génétique peut entrainer **des négociations sur l'accès aux informations génétiques**. Cela pose donc la question de savoir si les résultats obtenus par la recherche à partir des informations génétiques sont soumis aux régimes APA des territoires concernés par les négociations.

Plus largement, la question soulevée est celle de la dépendance entre les ressources génétiques et l'information génétique qui en découle. Déjà dans le Code la propriété intellectuelle, la notion de dépendance peut s'établir entre l'information génétique brute (non-travaillée) et l'information génétique traitée (brevet, COV) ou collectée (par séquençage ou extraction à partir d'une base de données).

Pour certaines parties, l'accès libre aux données est considéré comme une forme de partage des avantages, et l'Union européenne prône cet accès libre. Néanmoins, l'accès aux données est par ailleurs conditionné par les droits de propriété intellectuelle et de brevetabilité qui sont des droits sur les ressources et les techniques.

Le libre accès aux bases de données, une forme de partage des résultats issus de la recherche ?

Les dispositions établies par l'article 15.7 sur le partage juste et équitable des avantages découlant de l'utilisation des ressources génétiques et de ses dérivés permettent aux différentes parties d'imaginer un cadre très large. **Aussi, donner accès au pays fournisseurs des ressources génétiques à une base de données où sont stockées les informations extraites de ses ressources, constituent pour certaines parties une forme de partage des avantages**. En effet, la bio-informatique et l'évolution des techniques de modification génétique ont favorisé la constitution de bases de données génétiques. L'accès à ces bases de

[74] L'UPOV est une organisation internationale dont le siège est à Genève. Elle a été fondée en 1961 à Paris, au moment où la convention du même nom a été signée. Cette convention est entrée en vigueur en 1968. Elle a été révisée en 1972, 1978 et 1991.

[75] Directive 96/9/CE du 11 mars 1996.

données constitue un avantage important pour l'innovation dans la mesure où on dispose des outils et des capacités pour l'utilisation.

La directive européenne, ainsi que la loi française pour une République numérique de 2016, encouragent un partage des résultats issus de programmes de recherche par leur publication sur des bases de données de publiques. Notamment, le programme Horizon 2020 comporte l'obligation d'assurer le libre accès aux publications issues des recherches qu'il aura contribuées à financer, sous peine de sanctions financières.

Il convient de préciser que malgré ce mouvement de libre accès[76], les conditions de publication des résultats de recherche sont conditionnées par les relations entre les partenaires des projets. De manière générale, l'information génétique valorisée par les travaux de la recherche est appropriée dans le cadre de la directive de la Cour de justice de la commission européenne (CJCE) relative à la protection des inventions biotechnologiques du 6 juillet 1998. D'autres dispositifs sont à prendre en compte pour comprendre que les mécanismes de partage des résultats issu de l'utilisation de données de séquençage.

ANNEXE 16 : TABLEAU DES EXEMPLES D'UTILISATIONS DES DONNÉES DE SÉQUENÇAGE DE RGAA RÉCOLTÉS LORS DE L'ENQUÊTE

[76] Le mouvement du libre accès désigne l'ensemble des initiatives prises pour une mise à disposition des résultats de la recherche au plus grand nombre, sans restriction d'accès, que ce soit par l'auto archivage ou par des revues en libre accès.

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
1	1011 génomes de levures 2013-2019	France génomique	Université de Strasbourg, IRCAN, Genocscope	RG microorganismes : levures milieux naturels et présentes dans l'alimentation	1011 isolats naturels d'un organisme modèle eucaryote	Monde : origine écologique et géographique différente	Séquençage complet au Genoscope 12Megabases ; Caractérisation phénotypique	Carte génétique très détaillée chez la levure Saccharomyces cerevisiae Diversité génétique et phénotypique	INSDC : données brutes; LGD (bdd spécifique à la levure) : données analysées	Accès
2	AATTOL 2011-2016	ANR	Cirad, Cidres	RG de bovins et de microorganismes (parasite)	Cidres	Afrique de l'Ouest	Séquençage 2ème génération d'ARN	Caractérisation de bases moléculaires de la trypanotolérance chez les bovins	Cluster de calcul SouthGreen	EMBL après publication

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
3	Alive 2018-2020	PIA (ADEME), Spygen	Publics et privés : AFB, Université de Montpellier, WWF, ect.	Tous les types de RG	Genbank, pays partenaires, chercheurs internationaux bdd propres	Monde	Metabarcoding	Création d'une base de données DSI et RG d'échantillons environnementaux	En réflexion	Opendata
4	Bakery 2014-2018	PIA	CIRM-levures, CIRM-BIA,ITAB, universités	RG levures	Entreprises privées de la filière boulangerie	France	Métagenomique	Diversité génétique des communautés microbiennes	CIRM	?
5	BEEHOPE	ANR	Six partenaires européens dont le CNRS de Chizé	RG d'invertébrés	Apiculteurs	France	Marqueurs moléculaires; séquençage NGS	Analyse génétique ; Protection des abeilles du territoire (abeille noire)	?	Publication

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
6	BiodivA 2012-2016	Ministère de l'agriculture et la Région Rhône-Alpes (Casdar)	l'UMR Gabi de l'Inra, le Sysaaf, Itavi, le Centre de sélection de Béchanne et Labogena	RG avicoles	Eleveurs	France	Génotypage	Caractérisation de la diversité génétique	?	Publication
7	Catch My Interest 2016-	FEDER	Institut Carnot Plante2Pro, FEDER, Inra UMR LIPM, CNRGV	RG végétales de Tournesol résistant et non-résistant	?	France	Capture de la molécule d'ADN par Crispr-Cas9; Séquençage de cette région; Annotation de la séquence	Caractériser des zones d'intérêt agronomique sur le génome « marqueurs diagnostics » – région qui confère au Tournesol une résistance au parasite Orobranche	?	?

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
8	Divseek 2016	Crop Trust	68 partenaires : Africa Rice, Ag Research, AAC, ACPFG, AIT CATIE, CIAT, CGIAR, etc.	RPG	?	Banques de données privées et publiques internationales	Communication, séquençage	Faciliter la génération, l'intégration et le partage de données et d'informations liées aux ressources phytogénétiques	Divseek ne stock pas de données	Publique

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
9	ECOBIOPRO 2010-2013	ANR	ADIV, ADRIA, AERIAL, BIOCEANE, IFIP, IFREMER, Inra, ONIRIS, PFI	RG de bactéries, levures, moisissures	?	France, partenaires privés de l'agro-alimentaire (ex : viande hachée de veau)	Séquençage 2ème génération Illumina	Description et évolution des écosystèmes microbiens des produits carnés et de la mer; Bioprotection des aliments (développement de cultures protectrices)	?	GenBank (NCBI)

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
10	EMBARC YEASTIP	Consortium Européen du centre de ressources microbiennes (Type Projet FP7 de la CE)	Inra, CBS, DSMZ, CABI, etc.	Levures : environ 5000 séquences de RG de microorganismes	GenBank, DSMZ, CABI	?	Séquençage de marqueurs de référence	Obtenir par souche de référence une dizaine de marqueurs pour faciliter l'identification et la phylogénie; Taxonomie	Serveur Micalis	Webservice : bdd YeastIP accessible au public
11	FISHBOOST 2014-2017	CE	14 partenaires européens dont l'Inra, l'ifremer, le Sysaaf	RG aquatique	Inra, Adhérents du Sysaaf	Europe	Séquençage, génotypage	Sélection génomique	?	Publication

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
12	Food Microbiomes levure laitière Geotrichum candidum au sein de Saccharomycotina	ANR CNIEL	ANR, CNIEL, Producteurs de laits français et étrangers	Levure de référence française (6000 gènes) du fromage pont l'évêque	Génomes de levures du NCBI pour la comparaison des séquences	?	Séquençage génome entier	Mieux connaître le génome pour comprendre une adaptation éventuelle au milieu fromage	A Gant, en Belgique à (ORCAE) ; INSDC	Avant la publication
13	Generation Challenge program 2004-2013 (JC Glaszmann)	CGIAR	200 partenaires	RPG	?	Différents pays du monde	Séquençage	Amélioration des cultures (tolérance à la sécheresse)	Plateforme IBP	Licence Creative Commons Attribution-NonCommercial6ShareAlike

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
14	GeneRice (Génération et déploiement de variétés de riz efficaces en utilisation d'azote et éditées par génome) 2017-2019	Agropolis, Inra-Cirad, CFPRI	Inra, Cirad, FOFIFA, CIAT, UC Chile	RPG de riz, variété népalaise	Base de données du Cirad	Népal	Séquençage partiel d'une zone d'intérêt identifiée et amplifiée; Transformation génétique par Crispr-Cas9	Amélioration génétique assistée par la génomique d'un caractère agronomique complexe (l'efficacité d'utilisation de l'azote); Evaluation socio-économiques de nouvelles techniques d'amélioration des plantes	Cirad Montpellier	Bases de données publiques

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
15	Genius 2012-2019	PIA (GIS BV)	Inra, Cirad, Lyon3, Biogemma, Gemricopa, Société nouvelle Pépinières&Roseraies Georges Delbard et Vilmorin	RPG	Collections	France	Méganucléase, Talens, Crispr-cas9	Modification ciblée du génome pour l'adaptation au changement climatique	Ne produit pas de Données de séquençage	Publication

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
16	GnpIS 2002-aujourd'hui	ANR	Inra, Génoplante, Transplant, ELIXIR-Excelerate.	Espèces végétales et leurs champignons pathogènes	CRB Inra, Collections internationales	Monde	Bioinformatique	Créer un système d'information intégratif multi spécifique dédié aux parasites des plantes et des champignons . Liens entre structure du matériel génétique et traits agronomiques	Portail et Système d'information	Banques de données une partie en accès libre/une partie en accès restreint

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
17	International Wheat Genome Sequencing Consortium – IWGSC 2005-aujourd'hui	International	1500 membres Public-privé; 60 pays	RPG blé « Gold » issue du CNRGV	?	Monde	Séquençage du génome complet	Connaissance fondamentale et caractérisation de régions d'intérêt ; Faire une séquence génomique de haute qualité du blé tendre	INSCDC	Disponible sur URGI pf de bioinformatique
18	IRIC (International Rice Informatics Consortium) Projet Genomes riz 3000	France Génomique, IRRI (Philippines), BGI	IRD, Cirad, CIAT (Colombie), AfricaRice	RG végétale de variétés de riz	Collection de l'IRRI, de la France	Philippines, France, Afrique	Séquençage du plus de 3000 variétés	Diversité génétique. Sélection variétale	Création d'un portail web mondial	Licence Open source

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
19	MétaPDOcheese Projet pré-compétitif de l'industrie laitière	CNIEL, CNAOL et INRA	CNIEL, France génomique	RG microorganismes	Entreprises privées de l'industrie fromagère	fromages bénéficiant d'une AOP	Métagenomique; NTS	Diversité génétique des communautés microbiennes	Collection "MIL"	?
20	Projet ANR PEAKYEAST 2015-2018	ANR	Inra (plusieurs UMR dont STLO à Rennes, l'institut MICALIS de Jouy en Josas, SPO de Montpellier)	RG de microorganismes (levures Saccharomyces cerevisiae)	11 vignobles	?	Isolement, purification, Séquençage de l'ADNr 16S (Allemagne), phénotypage	Identification taxonomique ; Évolution de la levure du vin Saccharomyces cerevisiae vers son pic adaptatif ; Caractérisation des relations bactéries et levures	Logiciel propre	Séquences 16S sur base de données en accès libre

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
21	Projet Emissage 2018-2022	MAA (Casdar)	ACTALIA, Ifip, Anses, UMT ASIICS	RG microbiennes (trois sérovars de *Salmonella enterica)*	Abattoirs porcins, exploitations et sites de transformation du lait	France	Séquençage génomique global (WGS)	Surveillance sanitaire des Salmonella par les opérateurs des filières	Base de données partagées entre les partenaires sur le serveur de l'Anses (2019)	NCBI mais anonymes; Restitution des résultats
22	Projet Gaïa (n'a pas encore démarré)	ANR	International	RG végétales	?	Monde	Séquençage de génome	Explorer sur la surface de la planète la biodiversité, les hot spots (gènes d'intérêt autres que le rendement)	INSCDC	Publication Opendata

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
23	Projet IMAGE (Innovative Management of Animal Genetic Resources) 2016-2023	CE Programme Horizon 2020	28 Partenaires : 3 PME, 3 ONG, la FAO, 9 IR, etc.	RG animales	Cryo banques et CRB	Monde	Génomique, bioinformatique, nouvelle base de données	Améliorer les banques de gènes d'animaux pour la sélection variétale Nouvelle harmonisation des bases de données; Recherche de caractère adaptatifs	Création d'un système d'informations pour l'utilisation des collections génétiques	Publique

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
24	Projet investissement d'Avenir SUNRISE 2012-2019	PIA Labex TULIP	16 partenaires (6 entreprises semencières, une entreprise de biotechnologie, Inra, UPMC)	RG Tournesol (heliantus) + Espèce parasite orobanche (RPG)	Collections propres des RG de l'Inra (CRB Tournesol) et des collections privés	France; Population d'orobanche d'origine multiple (Europe)	Séquençage du génome complet de la lignée de tournesol XRQ ; Reséquençage de variétés	Décrypter le génome complet pour « accélérer les programmes de sélection variétale et nouvelles variétés adaptées aux changements et respectueuses de l'environnement »	Base de données propre en accès défini par l'AC ; et EMBL en accès libre	EMBL après publication

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
25	Projet privé : Identification de microorganismes	Partenaire privé de l'industrie agro-alimentaire	Laboratoire ACTALIA, pôle sécurité et aliments, Partenaires privés	RG de microorganismes (bactéries, levures, moisissures)	?	France, partenaires privés de l'agro-alimentaire	Séquençage 1ère génération par des prestataires (Eurofins en Allemagne)	Identification de microorganismes responsable d'une erreur dans le produit alimentaire attendu (yaourt qui gonfle, moisissure qui se développe)	Base de donnée privée de ACTALIA	X

26	RETHINk Tomate et Bio agresseurs Modèle agronomique 2018-2020	ANR LABEX TULIP	SYNGENTA, Inra	RPG Tomate (200 lignées de l'espèce sauvage Solanum pimpinellifolium), + RG 100 souches bactériennes (Ralstonia solanacearum)	Tomate : Majorité des lignées en accès libre au Tomato Genetic Resource Center (TGRC)- Collection privée de SYNGENTA Ralstonia : Collection gérée par l'UMR PVBMT (CIRAD)- La Réunion	Tomate : Amérique du sud Ralstonia : Amérique du sud, Afrique, Océanie	Séquençage génomique haut-débit Get-Plage, Univ. Chicago	Identification des bases génétiques de mécanismes de résistance durable aux pathogènes ; Générer de nouvelles variétés durables; Bases génétiques de la coévolution Tomate/Ralstonia en condition de stress abiotiques	Tomate : TGCR et collection privée SYNGENTA pour les graines/RG; Base de données public-privée (accès sur accord de collaboration) + baies de stockages locales pour données génomiques; Ralstonia : UMR-PVBMT et LIPM pour les souches; + baies de stockages locales pour données génomiques	Données brutes sur NCBI lors de la publication

	Nom du projet/initiative	Financeurs	Partenaires	Type de RG (phyto, zoo, aqua, forest, microorga et invert,)	Fournisseur de données ou du matériel (collection/ base de données existantes, extraction propre)	Origine géographique	Techniques	Finalité	Modalités de Stockage (outils et lieu) : où elles st hébergées et quel droit s'applique ?	Modalités de Diffusion
27	VIVALDI 2016-2020	IFREMER Entreprises H2020	21 partenaires	RG de mollusques et de bactéries		Différentes pays d'Europe, Israël, Norvège	Séquençage; Transcriptomique	Etude de l'impact des maladies chez les bivalves (classe de mollusques); Valorisation économique future	Traçabilité chacun déclare où sont les données (ordinateur, plateformes…)	Publications

ANNEXE 17 : LES SÉQUENÇAGES DE GÉNOMES DE RESSOURCES PHYTOGÉNÉTIQUES

Plant Names	Noms de plantes	Scientific Names	Genome Sequenced Year	Genome Size
Sunflower	Tournesol	*Helianthus annuus*	2017	3,5 GB
Wheat	Blé	*Triticum aestivum*	2017	15GB
Lupin	Lupin	*Lupinus angustifolius*	2016	600MB
Greater duckweed	Grande lentille d'eau	*Spirodela polyrhiza*	2014	158 MB
Loblolly pine	Pin de Loblolly	*Pinus taeda*	2014	20.15 GB
Pepper	Poivre	*Capsicum annuum*	2014	3.48 GB
Sugar beet	Betterave à sucre	*Beta vulgaris*	2014	590 MB
African oil palm	Palmier à huile africain	*Elaeis guineensis*	2013	1,8 GB
Capselle	Capselle	*Capsella rubella*	2013	130 MB
Carnivorous bladderwort plant	Plante carnivore	*Utricularia gibba*	2013	82 MB
Chickpea	Pois chiche	*Cicer arietinum*	2013	740 MB
Common Bean	Haricot Commun	*Phaseolus vulgaris*	2013	520 MB
Dwarf birch	Bouleau nain	*Betula nana*	2013	450 MB
Einkorn wheat	Blé Einkorn	*Triticum urartu*	2013	5 GB
Norway spruce	Spruce de Norvège	*Picea abies*	2013	20 GB
Rubber tree	Hévéa	*Hevea brailiensis*	2013	2.15 GB
Tobacco plant	Plante de tabac	*Nicotiana sylvestris*	2013	2.636 GB
White spruce	Épinette blanche	*Picea gluca*	2013	20.8 GB
Wild banana	Banane sauvage	*Musa balbisiana*	2013	438 MB
Banana	banane	*Musa acuminate*	2012	523 MB
Barley	Orge	*Hordeum vulgare*	2012	5.1 GB
Cassava	Manioc	*Manihot esculenta*	2012	760 MB
Domesticated Tomato	Tomate Domestique	*Solanum lycopersicum*	2012	900MB
Japanese apricot	Abricot japonais	*Prunus mume*	2012	280 MB
Melon	Melon	*Cucumis melo*	2012	450 MB
Musk melon	Melon musqué	*Cucumis melo*	2012	450 MB

Neem	Neem	*Azadirachta indica*	2012	370 MB
Watermelon	Pastèque	*Citrullus lanatus*	2012	353.5 MB
Arabidopsis	Arabidopsis	*Arabidopsis lyrata*	2011	207 MB
Date palm	Palmier dattier	*Phoenis dactylifera*	2011	658 MB
Hemp	Chanvre	*Cannabis sativa*	2011	820 MB
Medicago	Medicago	*Medicago turncatula*	2011	240 MB
Pigeon pea	Pigeon Pois	*Cajanus cajan*	2011	833 MB
Potato	Patate	*Solanum tuberosum*	2011	844 MB
Selaginella	Selaginella	*Selaginella moellendorffii*	2011	110 MB
Thellungiella	Thellungiella	*Thellungiella parvula*	2011	140 MB
Wild strawberry	Fraise sauvage	*Gragaria vesca*	2011	240 MB
Apple	Pomme	*Malus domestica*	2010	742.3 MB
Brachypodium	Brachypodium	*Brachypodium distachyon*	2010	272 MB
Peach	Pêche	*Prumus persica*	2010	227 MB
Soybean	Soja	*Glycine max*	2010	950 MB
Cucumber	Concombre	*Cucumis sativus*	2009	367 MB
Maize	Maïs	*Zea mays*	2009	2.5 GB
Sorghum	Sorgho	*Sorghum bicolor*	2009	700 MB
Papaya	Papaye	*Carica papaya*	2008	372 MB
Phuscomitrella	Phuscomitrella	*Phuscomitrella patens*	2008	480 MB
Wine Grape	Raisin De Vin	*Vitis vinifera*	2007	500 MB
Poplar	Peuplier	*Popuus trichocarpa*	2006	510 MB
Rice	Riz	*Oryza sativa*	2002	370 MB
Arabidopsis	Arabidopsis	*Arabidopsis thaliana*	2000	120 MB

Date de séquençage entier du génome et taille des génomes de ressources phytogénétiques.
(Centre National de Ressources Génomiques Végétales)

ANNEXE 18 : CHAMPS DE RECHERCHE LIÉS À L'ÉTUDE DE L'UTILISATION DES DONNÉES DE SÉQUENÇAGE DE RGAA À EXPLORER

Ce rapport est un état des lieux, cependant de nombreuses questions ont été soulevées lors des entretiens et des champs de recherche restent à explorer

- L'utilisation des données de séquençage concerne tous les domaines (anthropologie pour retracer l'histoire des populations humaines du passé par exemple, génétique, etc.). Le concept de données de séquençage est à envisager de manière transdisciplinaire car les multiples utilisations des données de séquençage permettraient d'ouvrir des discussions intersectorielles.

- En décembre 2018, *GenBank* stockait 211 millions de séquences correspondant à 285 milliards de bases. Etant donnée l'arrivée du séquençage complet pour tous les types de ressources génétiques, il est essentiel de disposer d'outils orientés « *Big Data* » pour le calcul, la gestion et le stockage. Des solutions externalisées ou locales ont été envisagées mais nécessitent qu'elles soient compatibles avec les infrastructures informatiques et que les compétences au sein des services bioinformatiques soient effectives.

- L'accès aux données de séquençage, produites dans le cadre de projets, est pour certaines parties déjà une forme de partage des avantages entre utilisateurs et fournisseurs, mais les capacités de réutilisation et d'analyse sont inégales d'un pays à l'autre. Une des pistes de réflexion proposée est l'étude des mesures d'accompagnement des partenaires pour l'utilisation des données de séquençage pour servir les objectifs de la CDB. Ces mesures d'accompagnement relèvent du renforcement de capacités, du transfert de compétences ou d'outils d'analyse et pourraient ou devraient être considérées comme la mise en œuvre du partage des avantages.

- Des réflexions ont émergé sur la mise en place d'un système multilatéral d'accès aux données de séquençage dans le cadre de cette étude et font échos à des réflexions dans d'autres pays comme en Allemagne. Ce système multilatéral pourrait être envisagé sur le modèle du TIRPAA. La mise en place d'un tel système implique d'approfondir les connaissances sur les possibilités d'assurer une traçabilité sur l'utilisation des données de séquençage. Des réflexions sont menées sur le mécanisme de la *block chain*[77].

- Le cadre juridique existant en France sur les données est riche. Une étude plus approfondie sur ce cadre aiderait à articuler une réglementation sur l'accès et le partage des avantages découlant de l'utilisation des données de séquençage de ressources génétiques.

- Sur le plan de la brevetabilité, le développement de la recherche en génomique permise par l'utilisation des données de séquençage a un impact fort. Le droit des brevets issu de la directive sur la protection des inventions biotechnologiques semble contribuer au malaise car il est encore très accueillant vis-à-vis des gènes natifs en Europe. Des inquiétudes ont émergé concernant la brevetabilité du vivant et la propriété intellectuelle (PI) permises par l'accès aux bases de données et à tout un travail fait par des générations de paysans. En France, la réglementation exclut la brevetabilité de l'information génétique native[78].

[77] Dans ce mécanisme, les transferts de données entre le fournisseur et l'utilisateur passe par un système décentralisé, partagé et public.

[78] Un gène natif est un gène naturel qui n'a fait l'objet d'aucune modification biotechnologique. Depuis 2016, un amendement interdisant le brevetage des « produits issus de procédés essentiellement biologiques » a été adopté en France.