# Résumé exécutif du rapport sur l'utilisation des données de séquençage de ressources génétiques pour l'alimentation et l'agriculture

Fondation pour la recherche sur la biodiversité

*Accéder au rapport complet : https://www.fondationbiodiversite.fr/wp-content/uploads/2019/11/FRB-Rapport-DSI-2019.pdf*

Membres Fondateurs de la FRB :

# SYNTHÈSE

Les sciences du vivant ont bénéficié des évolutions considérables des technologies de séquençage observées depuis la seconde moitié du XX$^{e}$ siècle. Les données de séquençage produites par ces techniques deviennent des éléments essentiels de la biologie moléculaire la faisant ainsi entrer dans l'ère de la génomique et amenant à une relation de plus en plus étroite entre génétique et numérique. L'informatique est utilisée dans deux cas, le traitement de l'information et sa transmission (cf. figure 1.).

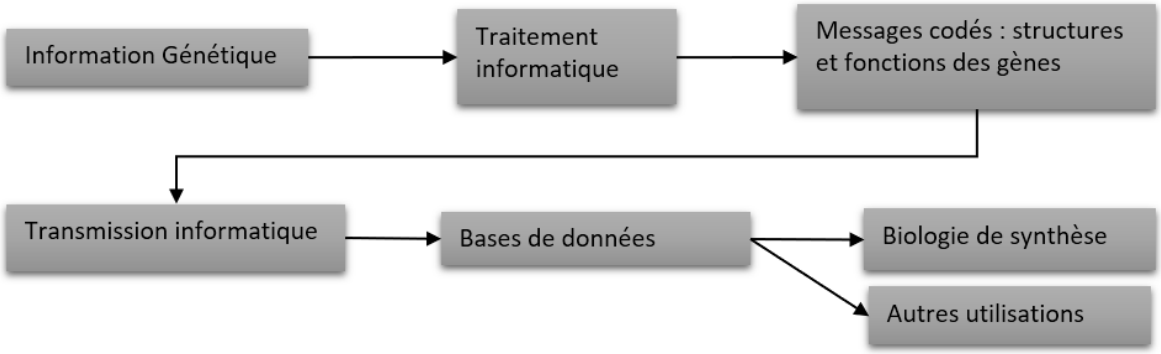



*Figure 1 : Une relation étroite entre génétique et numérique*[1].

La Convention sur la diversité biologique (CDB) signée à Rio de Janeiro le 5 juin 1992, a instauré un régime contractuel d'accès et de partage des avantages (APA) entre fournisseurs et utilisateurs de ressources génétiques fondé sur la réaffirmation de la souveraineté des États (Article 15 de la CDB). Le Protocole de Nagoya relatif à la CDB, a pour ambition de garantir l'accès aux ressources génétiques et le partage juste et équitable des avantages découlant de leur utilisation. L'utilisation croissante des bases de données pose la question de leur statut. C'est le cœur du débat qui anime les discussions des Parties signataires de la CDB, du TIRPAA ou encore de la Convention sur le droit de la mer.

Dans le cadre de la 14$^{e}$ Conférence des Parties de la Convention sur la diversité biologique (CDB) qui s'est déroulée 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 pour l'alimentation et l'agriculture (végétales, animales, aquatiques, forestières, microorganismes, invertébrés)[2].

L'importance et le rôle de ces ressources en termes de sécurité alimentaire et nutritionnelle ou encore d'adaptation et de mitigation des changements climatiques, a été soulignée par la Commission sur les ressources génétiques pour l'alimentation et l'agriculture (CRGAA). Dans ce cadre, l'étude a 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 (RGAA) et de déterminer une typologie d'utilisateurs et de pratiques.

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

[2] 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 des intervenants, sont disponibles sur le site de la FRB.

**PRINCIPAUX RÉSULTATS :**

Les sorties de l'étude ont donc été de :

- proposer une appellation pour remplacer le terme « information de séquençage numérique » (« *Digital sequence information* ») en « données numériques de séquences de ressources génétiques » (« *digital data on genetic resource sequences* ou *digital sequence data* ») ;
- définir une typologie suivant la chronologie du processus de séquençage : donnée brute, donnée nettoyée, donnée analysée ;
- énumérer les principales applications de ces données numériques.

L'« information numérique de données de séquençage » est un concept récent qui ne bénéficie pas encore de définition consensuelle au niveau international, alors même que les enjeux et opportunités de cette « ressource » dématérialisée font l'objet de discussions intenses notamment dans le contexte du développement des nouvelles technologies de manipulation du génome.

La difficulté à définir ce qu'est une donnée de séquençage est d'abord due au fait qu'elle revêt une réalité multiple. Ces données de séquençages de RGAA présentent ainsi plusieurs intérêts : de l'étude de la diversité génétique à la caractérisation génétique. Ces intérêts varient selon le type de ressource génétique considéré.

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. Les RGAA d'intérêt économique ont aussi bénéficié des programmes de sélection génétique qui aujourd'hui s'étendent peu à peu à l'ensemble des RGAA. La découverte de marqueurs d'intérêt (sexe, résistance à un parasite, etc.) a démultiplié l'intérêt de la technique par la sélection précoce des descendants et le raccourcissement du processus de sélection.

Des données de séquençages sont par exemple à la base des travaux :

- de caractérisation pour la conservation de races locales avicoles (ex. projet BioDivA) ;
- sur le contrôle des maladies dans la filière conchylicole (ex. projet VIVALDI) ;
- de suivi épidémiologique des abeilles ou de lutte contre le syndrome d'effondrement des colonies (ex. projet BEEHOPE)
- d'amélioration variétale chez des espèces cultivées (ex. projets SUNRISE et Genius) ;
- d'étude de la diversité microbienne au sein de la filière laitière (ex. CNIEL) ;
- pour faciliter la domestication de la levure pour l'industrie agro-alimentaire (ex. projet Bakery) ;
- de développement rapide de nouvelles stratégies de sélection et de production de variétés forestières.

.

# CONTRIBUTEURS

**COORDINATION ET RÉDACTION**
Charlotte Navarro, Fondation pour la recherche sur la biodiversité
Jean Lanotte, ministère de l'agriculture et de l'agriculture
Comité de pilotage de l'étude
Ana Deligny, traduction

**CITATION**
Fondation pour la recherche sur la biodiversité (2019), *Résumé exécutif du rapport de l'étude sur l'utilisation des données de séquençage des ressources génétiques pour l'alimentation et l'agriculture*, Paris, France : FRB, 26p.



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.

**Relecteurs du résumé :** Robin Goffaux (FRB), Jean Lanotte (MAA), Hélène Soubelet (FRB).

# Table des matières

# I. HISTOIRE DE LA BIOLOGIE MOLÉCULAIRE À L'ESSOR DE LA GÉNOMIQUE

L'évolution dans le domaine des sciences du vivant est liée aux progrès technologiques. Les données de séquençage deviennent des éléments essentiels et la communauté scientifique s'organise pour permettre le partage de ces données à travers des bases de données en accès libre.

## A. NOTIONS ESSENTIELLES POUR 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[3] ou Adénine, Cytosine, Guanine, Uracile pour l'ARN[4]). 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 biochimique. Une séquence est donc une représentation de l'enchaînement des constituants élémentaires des molécules d'ADN (bases) symbolisés par les lettres A, C, G, et T[5].

La biologie moléculaire a notamment pour objet l'étude des macromolécules biologiques comme les acides nucléiques, dont l'ADN, et les protéines pour l'étude desquelles, des outils sont développés et utilisés par d'autres disciplines (ex. la génétique évolutive utilise des outils de biologie moléculaire). Cette discipline 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). Les techniques rattachées à la biologie moléculaire sont de l'ordre de l'exploration du vivant et de l'informatique[6].

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[7], passant d'une phase descriptive à une phase d'expérimentation fonctionnelle. L'analyse du génome est un des éléments essentiels dans l'étude du vivant et de son fonctionnement.

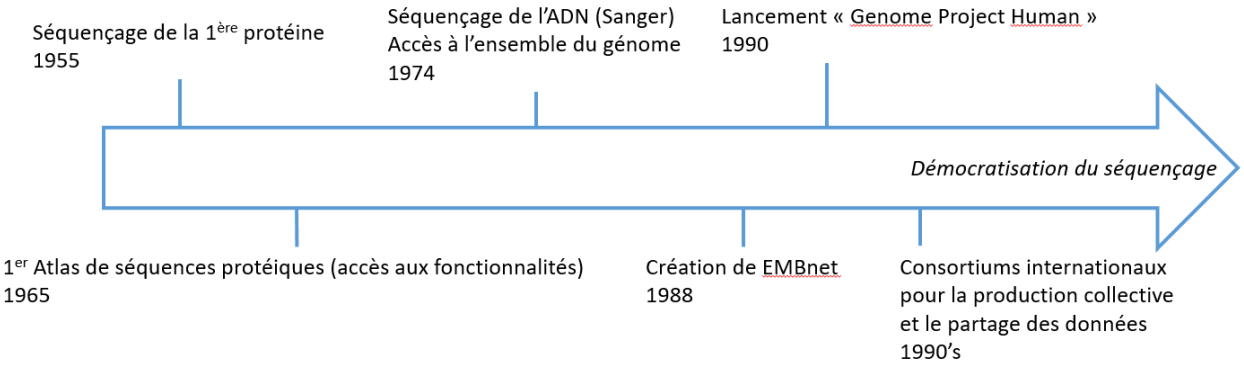



*Figure 2 : Les grandes étapes historiques de l'évolution de la génomique et notamment les évolutions des techniques de séquençage*[8]

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

[4] 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.

[5] 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.

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

[7] 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.

[8] Idem.

**Le séquençage** est l'ensemble des manipulations permettant de déterminer la séquence d'une molécule d'ADN, d'ARN, ou d'une protéine. Les bioinformaticiens 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 (cf. figure 2). La taille des génomes varie fortement en fonction des espèces considérées.

**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.

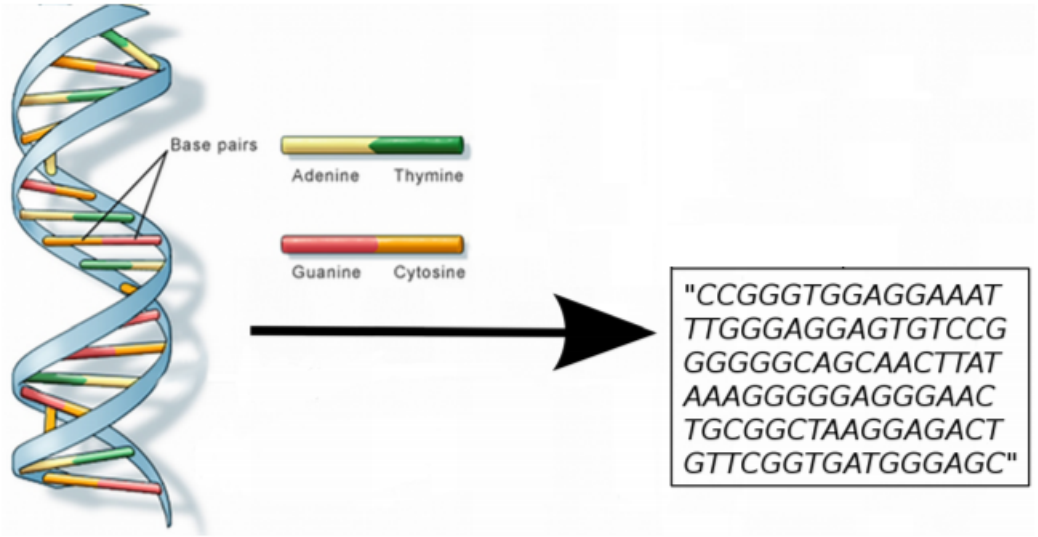



*Figure 3 : Aspect biologique et aspect informatique de l'ADN (P. Leleux, 2014)*

## B. LA PRODUCTION DES DONNÉES DE SÉQUENÇAGE ET LEUR MISE EN BASE DE DONNÉES

Les activités des chercheurs sur la question des données 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 d'autres disciplines ; la diffusion des connaissances qui correspond à des opérations standardisées[9]. Ces activités forment un cycle marqué par l'acquisition d'informations, leur stockage dans les bases de données et leur traitement informatique (cf. figure 4).

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

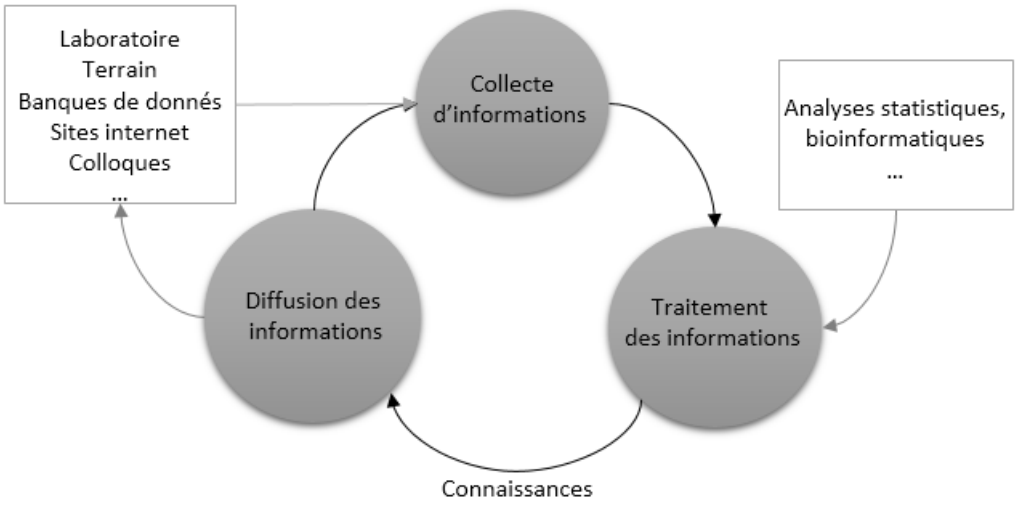



*Figure 4 : Le cycle de l'information scientifique et technique*[10]

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*) en 2003. Composé de 3 milliards de bases, ce séquençage a coûté 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 visent à décrypter la fonction des gènes nouvellement détectés chez un grand nombre d'espèces vivantes. La génétique fonctionnelle rapproche les approches biochimiques et physiologiques aux analyses de génomes entiers. Ceci induit un fort besoin de stockage et de diffusion de ces données qui sont capitalisées dans des bases de données de plus en plus volumineuses.

Ainsi par exemple, pour la mise en commun de leurs données de séquençage, le centre national pour l'information relative à la biotechnologie (NCBI), le laboratoire européen de biologie moléculaire (EMBL) et la banque de données ADN du Japon (DDBJ) ont monté une collaboration portant sur leurs bases de données. Cette collaboration couvre le spectre des lectures de données brutes, des annotations fonctionnelles et des informations contextuelles relatives aux échantillons et aux configurations expérimentales.

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 5) :

```
>gnl|ti|1586495440 name:1047100384971 mate:1902487597
TTGCAAGCTTAGTATTACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCCTTCTTGCC
AAAGACAATGCACCGCGGGACATTGCTGTACCAATCACCTTTTGATCCACTTCCTACCGAATGGATGCAA
AAATCAGTTTTAAATAGACAAAGGCATGTGGGAGAGGCGATCTTAGGGTTCCCTCTAGATCTACAGGGTG
ACCTAGTTGATGCGAATGGAGAGACTTGTAGAAGTTTTGATCGGTCAGGTTTATGTACTAGTTTCTCTAA
ATCTGCATCTACAGGTAATGATCTTTTACTTGGTAAAAAAAAAAAAAAAAAAAAAAAAAAGTACTCTGCGT
TGATACCACTGCTTAAGGGCGAATTCGCGGCCGCTAAATTCAATTCGCCCTATAGTGAGTCGTATTACAA
TTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCA
GCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGC
GCAGCCTATACGTACGGCAGTTTAAGGTTTACACCTATAAAAGAGAGAGCCGTTATCGTCTGTTTGTGGA
TGTACAGAGTGATATTATTGACACGCCGGGGCGACGGATGGTGATCCCCCTGGCCAGTGCACGTCTGCTG
TCAGATAAAGTCTCCCGTGAACTTTACCCGGTGGTGCATATCGGGGATGAAAGCTGGCGCATGATGACCA
CCGATATGGCCAGTGTGCCGGTCTCCGTTATCGGGGAAGAAGTGGCTGATCTCAGCCACCGCGAAAATGA
CATCAAAAACGCCATTAACCTGATGTTCTGGGGAATATAAATGTCAGCATGAGATTATCAAAAAGGATCT
TCACCTAGATCCTTTTCACGTAGAAAGCCAGTCCGCAGAAACGTGCTGACCCCTGATGAATGTCAGCTAC
TGGGCTATCTGGACAAGGGAAAACGCAAGCGCAAAGAGAAAGCAGTAGC
```

*Figure 5 : Données de séquençage brute de l'espèce Arabidopsis thaliana (Sequence Read Archive, 2019)*

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

- les lectures de séquences issues d'électrophorèse capillaire qui sont les chromatogrammes[11] de séquence d'ADN (cf. figure 6) :

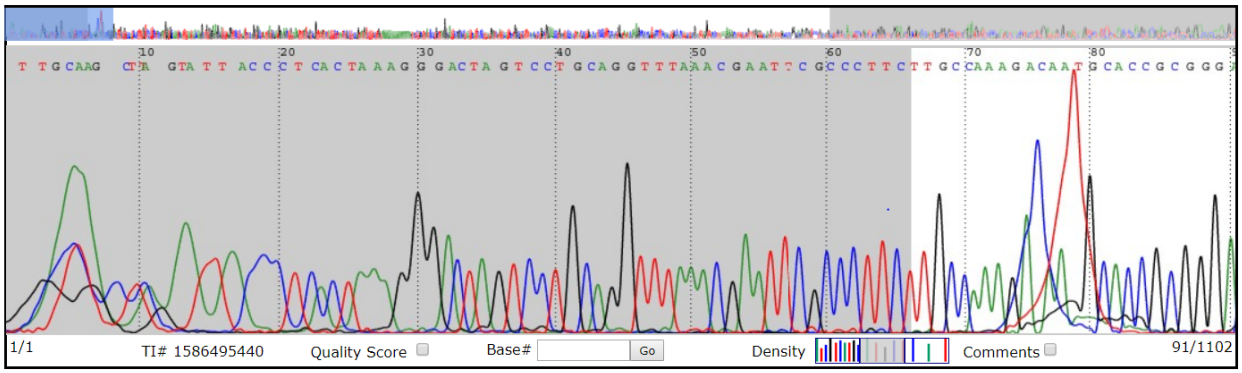



*Figure 6 : lecture capillaires ou chromatogrammes de séquence d'ADN de l'espèce Arabidopsis thaliana (Trace Archive, 2019)*

- les séquences annotées qui sont les régions fonctionnelles identifiées, souvent les gènes codant les protéines (cf. figure 7) :

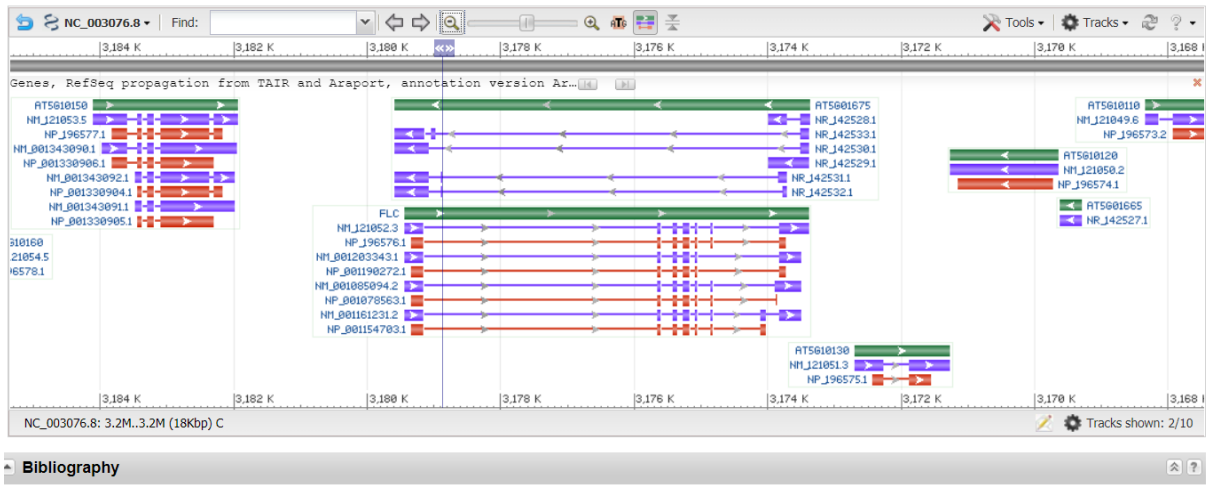



*Figure 7 : Les séquences annotées du gène Flowering Locus C codant pour la protéine MADS-BOX (floraison) de l'espèce Arabidopsis thaliana (GenBank, 2018)*

Certaines bases de données associent aux séquences des informations encore plus précises : BioSample par exemple ajoute 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).

[11] Diagramme résultant d'une chromatographie, technique permettant de séparer les composants chimiques présents dans un mélange. Dans le cas de la figure 6, les différents fragments issus de la dégradation de l'ADN d'*Arabidopsis thaliana* montrent des longueurs variables et des acides nucléiques différents en fin de chaine. En identifiant ces différentes longueurs et la base nucléique qui les terminent, on retrace l'enchainement de ces bases.

# II. QUE RECOUVRE LE TERME DE « DONNÉES DE SÉQUENÇAGE » OU « INFORMATION NUMÉRIQUE DE DONNÉES DE SÉQUENÇAGE » ?

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). Elles varient selon le regard et les traitements apportés par les différents types d'acteurs qui les gèrent et les utilisent.

## A. LES DONNÉES FACTUELLES ET LES DONNÉES TEXTUELLES

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.

Les données factuelles ou représentations de séquences nucléotidiques, sont issues des expérimentations (« paillasse ») ou des banques de séquences internationales. Leur description suit une notice standardisée où les dépositaires peuvent spécifier des champs pour le renseignement 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 technique utilisée pour la production de la séquence ;
- Les propriétés de la séquence (champ FEATURES) : annotations qui décrivent la séquence, 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'enregsitrement 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*

Les données textuelles renvoient à la littérature qui utilise de ces données (articles de revue, ouvrages scientifiques, actes de colloque, etc.) 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 sont souvent mieux décrites dans les articles scientifiques que dans les banques. L'enjeu d'exploitation des informations présentes dans ces bases de données est majeur.

### B. LES TYPOLOGIES PROPOSÉES

Dans le cadre de cette étude, une typologie opérationnelle des données de séquençage a été retenue. Premièrement, les **données brutes** sont les données issues des séquenceurs mais ne sont pas conservées. Deuxièmement, les fichiers textes communément appelés « séquences », sont les **données « propres » ou « nettoyés »** par des procédés informatiques qui consistent à enlever les parties de lecture de mauvaise qualité[12]. Enfin, **les données analysées qui consiste en** l'assemblage[13] permettant l'obtention de nouveaux fichiers texte. Cette typologie suit le protocole bioinformatique d'analyse des données de séquençage haut débit (nommé *pipeline*). Les types de fichiers générés au cours du processus sont différents en fonction de la finalité du projet et des utilisateurs. [14].


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

D'une façon plus détaillée, 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.

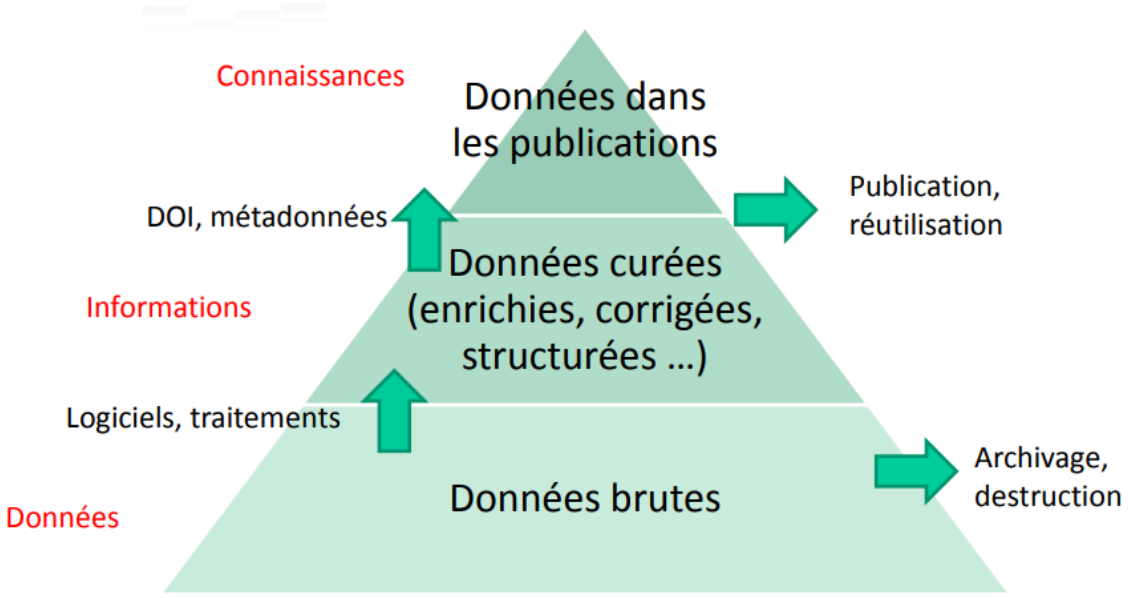



*Figure 12 : Les états de la donnée au sens large, Inra*

Sur le plan technique, les technologies évoluent très vite et les coûts d'analyse, de gestion et de stockage de la donnée deviennent plus importants que le coût de production. Il n'existe pas de modèle économique pour envisager de répondre à ce problème.

---

[12] Le processus de nettoyage des données brutes permet d'éliminer des séquences dites contaminants. Ces séquences peuvent se former lors de la préparation du séquençage et peuvent poser un problème lors de l'étape d'assemblage.

[13] L'assemblage est l'étape d'alignement ou de fusion des fragments de séquence permettant de reconstruire la séquence originale. Il peut être comparer 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).

[14] 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.

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

## A. TYPOLOGIE DES UTILISATIONS DE DONNÉES DE SÉQUENÇAGE POUR LES RGAA

Les projets incluant du séquençage de microorganismes sont les plus nombreux en raison de la petite taille de leurs génomes, 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 à génome complexe font l'objet de plus en plus de nouveaux projets. Au travers des cas de figures rencontrés lors de ce travail plusieurs objectifs ont motivé la production de données de séquençage de ressources génétiques : connaissance fondamentale de la composition et du fonctionnement du génome, identification et exploration de la diversité génétique, amélioration génétique par différentes techniques de sélection reposant sur l'étude de séquences génétiques ou de modification *in vitro* du génome.

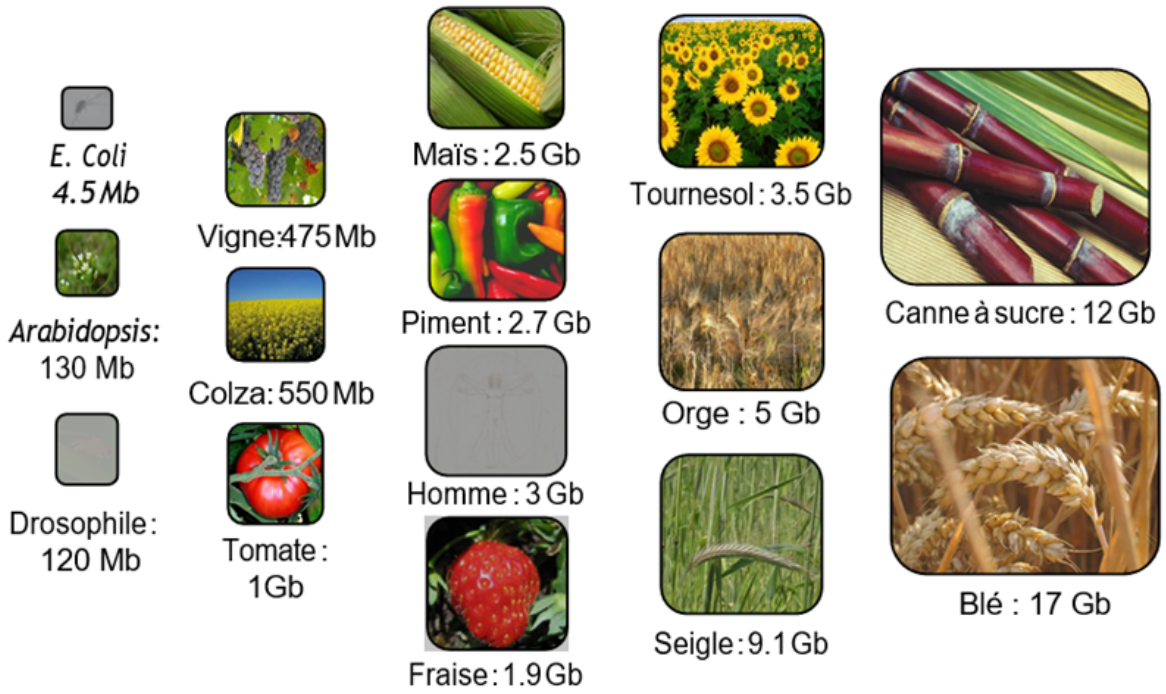



*Figure 13 : La taille des génomes, Centre national de ressources génétiques végétales (CNRGV)*

Les données de séquençage de ressources génétiques pour l'alimentation et l'agriculture sont d'abord et avant tout utilisées pour la taxonomie et l'amélioration dans le processus de sélection.

Par exemple, en termes de ressources génétiques animales, les bovins ont traditionnellement bénéficié de programmes de sélection génétique. 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, seules productrices de caviar.

Le séquençage permet ainsi une meilleure connaissance du fonctionnement moléculaire des organismes en vue d'affiner les programmes de sélection.

De manière générale, pour les espèces présentant des intérêts économiques moindres, les études de diversité génétique sont les plus fréquentes. Elles mesurent le degré de variétés des gènes au sein d'une même espèce. La cartographie génétique consiste à situer, le long des chromosomes, des séquences d'ADN connues.

D'autres techniques sont utilisées comme la sélection assistée par marqueurs permettant le suivi des gènes et le tri précoce des ressources considérées après croisement naturel. Plus récemment, l'édition de génome permet, notamment, la modification ciblée de caractère agronomique.

Le déploiement de ces techniques dépend des applications commerciales prévues, mais aussi des limites techniques liées à la taille des génomes (plus un génome est grand, plus il est complexe, et plus il est difficile de comprendre son fonctionnement).

**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

*Tableau : Utilisations principales des données de séquençage de génomes de RG, citées par les personnes interrogées, typologie extraite des entretiens de la présente étude*

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).

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

### i. Le projet BioDivA : caractérisation génétique pour la conservation des races locales avicoles.

*Figure 14 : Poule grise du Vercors (Association Ouantia Grise du Vercors)*

Contexte : 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.

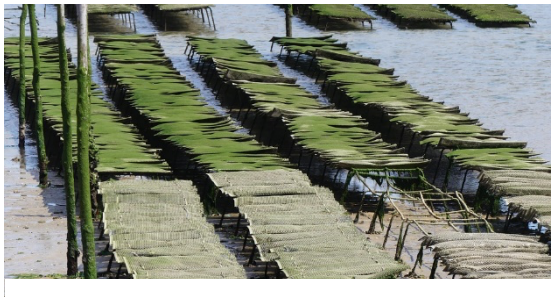
Objectif : Le projet BioDivA, financé le ministère de l'agriculture et de l'alimentation et la Région Rhône-Alpes, a débuté en 2013 pour une durée de trois ans. Afin de répondre ce problème, le projet a pour

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.
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, 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*.

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

### ii. 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 entraînent de lourdes pertes économiques.



*Élevage d'huîtres ou ostréiculture*

Objectif : Le projet européen d'Horizon H2020 VIVALDI a démarré en 2016 pour 4 ans. Il 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[15]. 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 il 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 est prévu. Au cœur de ce réseau, VIVALDI contribue ainsi à partager l'information et les expériences de chacun sur les mortalités de coquillages pour un meilleur contrôle des maladies associées.

Les techniques : Séquençage du génome complet des bivalves.

[15] 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.

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[16].

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

Contexte : Le tournesol, par sa faible exigence hydrique, est une des solutions pour faciliter l'adaptation de la filière végétale aux effets des changements climatiques. 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 projet SUNRISE (2012-2020) a pour objectifs de réaliser le séquençage du génome complet 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.

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.

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 intégrés 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.

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

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

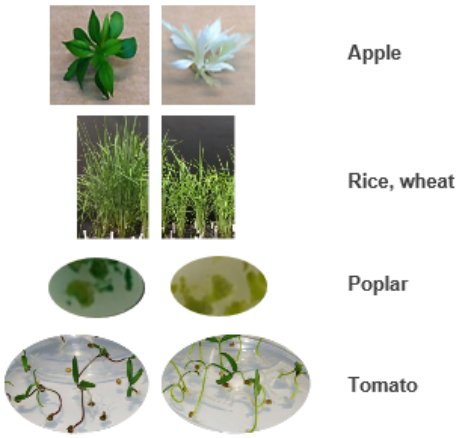



*Figure 15 : Vers des plantes cultivées présentant des caractères de cultures et de qualité améliorés pour l'alimentation humaine et animale et d'autres utilisations, Peter Rogowsky, 2018.*

Objectifs : Le projet Genius (2012-2019) vise à répondre aux enjeux actuels d'agriculture durable, à travers l'étude de plusieurs caractères de 12 espèces différentes (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. Ce projet est financé par des fonds publics (programme d'investissement d'avenir de l'État).

Techniques : Le projet consiste à appliquer à ces espèces des outils pour modifier des gènes de manière plus ciblée. 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. Les caractères d'intérêt observés sont par exemple la couleur des feuilles pour la pomme ou la longueur des tiges pour le riz.

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.

| 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 |

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

Contexte : Les caractéristiques du fromage (couleur, acidification, texture, flaveur, etc.) sont en partie déterminées par les écosystèmes microbiens, qui contribuent à leur qualité. Ainsi, la composante microbiologique dans l'élaboration et la typicité des fromages est importante mais demeure complexe et varie d'un fromage à l'autre.

Les données de séquençage sont utilisées pour mieux comprendre les écosystèmes microbiens, leur diversité et leurs fonctionnalités afin d'améliorer la production et la transformation laitière. En France, 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, un Institut technique agro-industriel (ITAI[17]).

Objectif : Le projet du CNIEL 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[18]) française qui sont issues de la combinaison de pratiques variées de production laitière et de transformation du fromage.

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.

[17] Les Instituts techniques agro-industriels (ITAI) sont des organismes privés 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.

[18] 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).

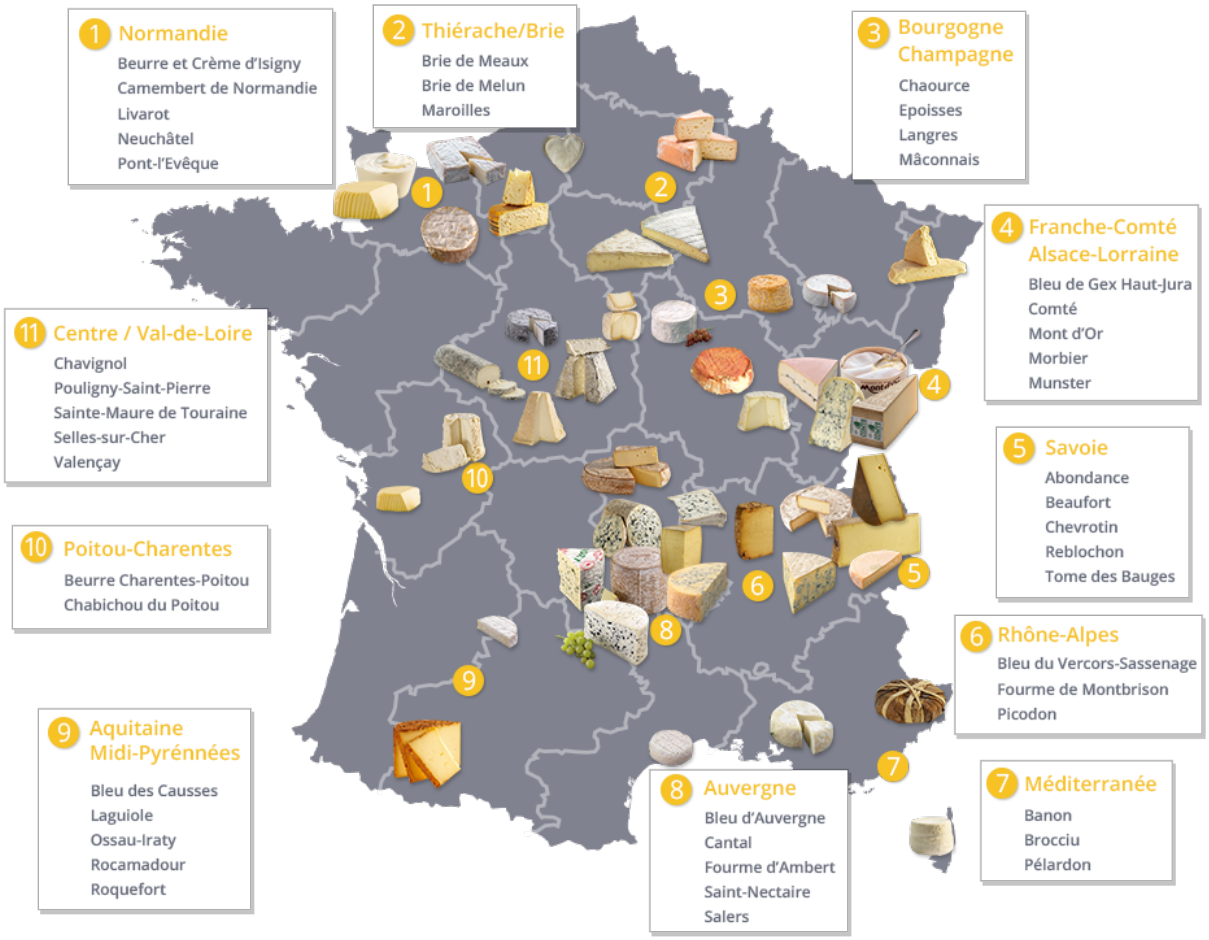



*Figure 15 : Les 45 fromages AOP français, présentation « Acquisition et utilisation des données de séquençage dans les projets soutenus par le CNIEL », Frédéric Gaucheron*

### vi. Le 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 est financé par des fonds publics (programme d'investissement d'avenir de l'État).

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[19], 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.

[19] La levure est un champignon microscopique unicellulaire (saccharomyces). Elle est utilisée depuis des millénaires à l'état sauvage et depuis le XX$^{e}$ 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.

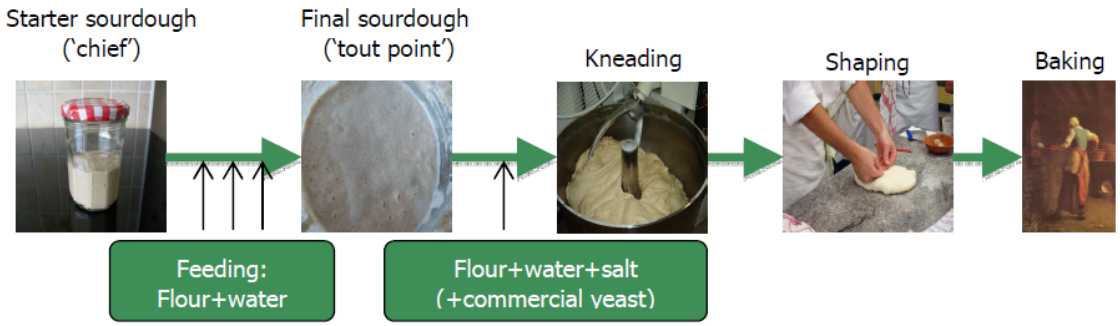



*Figure 16 : Les étapes de production du pain*
*(programme Systèmes alimentaires durables : projet Bakery, ANR 2013)*

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. 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 petites et moyennes entreprises) hébergent des communautés microbiennes différentes, ce qui montre l'importance de maintenir une diversité socio-culturelle.

### vii. Le 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. 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.

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.

Résultats : Sur la base de la structuration de la 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.

Le développement de puce a aussi permis d'améliorer la connaissance du régime de reproduction des vergers de pin maritime. Ces études doivent permettre d'optimiser le design et la gestion des vergers à graines afin de maximiser le gain génétique.

Un ensemble de 80 marqueurs moléculaires a été développé afin d'estimer plus précisément les valeurs génétiques de chaque arbre pour augmenter les gains génétiques des futures variétés.

Les chercheurs ont découvert les marqueurs moléculaires pour identifier de façon unique chaque individu et reconstituer son pedigree. Il devient alors envisageable de simplifier les cycles de sélection en substituant aux croisements biparentaux, 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.

Une autre approche rendue possible par l'utilisation des marqueurs moléculaires, et appliqué au cas du pin maritime, 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, de l'ordre d'une dizaine d'années.

### viii. Le projet BEEHOPE pour lutter contre le syndrome d'effondrement des abeilles et pour une gestion durable de l'apiculture

Contexte : Le taux d'extinction actuel des espèces dans la biosphère serait comparable à celui des dernières extinctions massives[20]. 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[21].

Objectif : Le projet de l'Agence nationale de la recherche (ANR) BEEHOPE a démarré en 2013. Il 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

[20] 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

[21] Résumé à l'intention des décideurs de l'évaluation de la plateforme intergouvernementale scientifique et politique sur la biodiversité et les services écosystémiques (IPBES) des pollinisateurs, de la pollinisation et de la production alimentaire. Copyright © 2016, Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) ISBN: 978-92-807-3568-0 Job Number: DEW/1990/NA

Nord. Le projet a pour objectif de récolter des données sur l'abeille noire pour l'étude de ses caractères adaptatifs.

Techniques : L'évaluation de la diversité génétique est réalisée à l'aide de marqueurs moléculaires. Le séquençage 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 : Malgré les efforts de protection de l'abeille noire, celle-ci présente des niveaux d'hybridation élevé (8 % contre 30 % pour les populations non protégées). Certaines populations protégées nécessitent encore des ajustements dans les stratégies de gestion[22] pour purger davantage les allèles étrangers identifiés à l'aide de marqueurs moléculaires[23].

[22] Dans les aires protégées des stocks de reproducteurs sélectionnés sont accouplés au sein de stations de reproduction isolées afin d'empêcher le flux de gènes de sources non désirées.

[23] Pinto, M & Henriques, Dora & Chávez Galarza, Julio César & Kryger, Per & Garnery, Lionel & Van der Zee, Romée & Dahle, Bjørn & Soland-Reckeweg, Gabriele & De la Rua, Pilar & Dall'Olio, Raffaele & Carreck, Norman & Johnston, J. (2014). Genetic integrity of the Dark European honey bee (*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.

# CONCLUSION

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 remplacer 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, en France, 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. 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 du rapport final).

# ANNEXE

Tableau récapitulatif de exemples d'utilisation de données de séquençage de ressources génétiques pour l'alimentation et l'agriculture.

| | **Nom du projet/initiative** | **Partenaires** | **Type de RG (phytogénétique, zoogénétique, aquatique, forestière, microorganisme et invertébrés)** | **Finalité** |
|---|---|---|---|---|
| 1 | 1011 génomes de levures 2013-2019 | Université de Strasbourg, IRCAN, Genocscope | RG microorganismes : levures milieux naturels et présentes dans l'alimentation | Carte génétique très détaillée chez la levure Saccharomyces cerevisiae Diversité génétique et phénotypique |
| 2 | AATTOL 2011-2016 | Cirad, Cidres | RG de bovins et de microorganismes (parasite) | Caractérisation de bases moléculaires de la trypanotolérance chez les bovins |
| 3 | Alive 2018-2020 | Publics et privés : AFB, Université de Montpellier, WWF, etc. | Tous les types de RG | Création d'une base de données DSI et RG d'échantillons environnementaux |
| 4 | Bakery 2014-2018 | CIRM-levures, CIRM-BIA,ITAB, universités | RG levures | Diversité génétique des communautés microbiennes |
| 5 | BEEHOPE | Six partenaires européens dont le CNRS de Chizé | RG d'invertébrés | Analyse génétique ; Protection des abeilles du territoire (abeille noire) |
| 6 | BiodivA 2012-2016 | l'UMR Gabi de l'Inra, le Sysaaf, Itavi, le Centre de sélection de Béchanne et Labogena | RG avicoles | Caractérisation de la diversité génétique |
| 7 | Catch My Interest | Institut Carnot Plante2Pro, FEDER, Inra UMR LIPM, CNRGV<br><br>2016- | RG végétales de Tournesol résistant et non-résistant | Caractériser des zones d'intérêt agronomique sur le génome « marqueurs diagnostics » – région qui confère au Tournesol une résistance au parasite Orobranche |
| 8 | Divseek<br><br>2016 | 68 partenaires : Africa Rice, Ag Research, AAC, ACPFG, AIT CATIE, CIAT, CGIAR, etc. | RPG | Faciliter la génération, l'intégration et le partage de données et d'informations liées aux ressources phytogénétiques |
| 9 | ECOBIO-PRO 2010-2013 | ADIV, ADRIA, AERIAL, BIOCEANE, IFIP, IFREMER, Inra, ONIRIS, PFI | RG de bactéries, levures, moisissures | Description et évolution des écosystèmes microbiens des produits carnés et de la mer; Bioprotection des aliments |

| | | | | |
|---|---|---|---|---|
| | | | | (développement de cultures protectrices) |
| 10 | EMBARC YEASTIP | Inra, CBS, DSMZ, CABI, etc. | Levures : environ 5000 séquences de RG de microorganismes | Obtenir par souche de référence une dizaine de marqueurs pour faciliter l'identification et la phylogénie; Taxonomie |
| 11 | FISH-BOOST 2014-2017 | 14 partenaires européens dont l'Inra, l'Ifremer, le Sysaaf | RG aquatique | Sélection génomique |
| 12 | Food Microbiomes levure laitière *Geotrichum candidum* au sein de Saccharomycotina | ANR, CNIEL, Producteurs de laits français et étrangers | Levure de référence française (6000 gènes) du fromage pont l'évêque | Mieux connaître le génome pour comprendre une adaptation éventuelle au milieu fromage |
| 13 | Generation Challenge program 2004-2013 (JC Glaszmann) | 200 partenaires | RPG | Amélioration des cultures (tolérance à la sécheresse) |
| 14 | GeneRice (Génération et déploiement de variétés de riz efficaces en utilisation d'azote et éditées par génome) 2017-2019 | Inra, Cirad, FOFIFA, CIAT, UC Chile | RPG de riz, variété népalaise | 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 |
| 15 | Genius 2012-2019 | Inra, Cirad, Lyon3, Biogemma, Gemricopa, Société nouvelle Pépinières&Roseraies Georges Delbard et Vilmorin | RPG | Modification ciblée du génome pour l'adaptation au changement climatique |
| 16 | GnpIS<br><br>2002 – aujourd'hui | Inra, Génoplante, Transplant, ELIXIR-Excelerate. | Espèces végétales et leurs champignons pathogènes | 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 |

| | | | | |
|---|---|---|---|---|
| 17 | International Wheat Genome Sequencing Consortium (IWGSC) 2005-aujourd'hui | 1500 membres Public-privé; 60 pays | RPG blé « Gold » issue du CNRGV | Connaissance fondamentale et caractérisation de régions d'intérêt ; Faire une séquence génomique de haute qualité du blé tendre |
| 18 | IRIC (International Rice Informatics Consortium) Projet Genomes riz 3000 | IRD, Cirad, CIAT (Colombie), AfricaRice | RG végétale de variétés de riz | Diversité génétique. Sélection variétale |
| 19 | MétaPDOchee e Projet précompétitif de l'industrie laitière | CNIEL, France génomique | RG microorganismes | Diversité génétique des communautés microbiennes |
| 20 | Projet ANR PEAKYEAST 2015-2018 | Inra (plusieurs UMR dont STLO à Rennes, l'institut MICALIS de Jouy en Josas, SPO de Montpellier) | RG de microorganismes (levures Saccharomyces cerevisiae) | Identification taxonomique ; Évolution de la levure du vin Saccharomyces cerevisiae vers son pic adaptatif ; Caractérisation des relations bactéries et levures |
| 21 | Projet Emissage 2018-2022 | ACTALIA, Ifip, Anses, UMT ASIICS | RG microbiennes (trois sérovars de *Salmonella enterica*) | Surveillance sanitaire des Salmonella par les opérateurs des filières |
| 22 | Projet Gaïa (n'a pas encore démarré) | International | RG végétales | Explorer sur la surface de la planète la biodiversité, les hot spots (gènes d'intérêt autres que le rendement) |
| 23 | Projet IMAGE (Innovative Management of Animal Genetic Resources) 2016-2023 | 28 Partenaires : 3 PME, 3 ONG, la FAO, 9 IR, etc. | RG animales | Améliorer les banques de gènes d'animaux pour la sélection variétale Nouvelle harmonisation des bases de données; Recherche de caractère adaptatifs |
| 24 | Projet investissement d'Avenir SUNRISE 2012-2019 | 16 partenaires (6 entreprises semencières, une entreprise de biotechnologie, Inra, UPMC) | RG Tournesol (*heliantus*) + Espèce parasite orobanche (RPG) | Décrypter le génome complet pour « accélérer les programmes de sélection variétale et nouvelles variétés adaptées aux changements et respectueuses de l'environnement » |
| 25 | Projet privé : Identifica- | Laboratoire ACTALIA, pôle sé- | RG de microorganismes (bactéries, levures, moisissures) | Identification de microorganismes responsable d'une erreur dans |

| | | | | |
|---|---|---|---|---|
| | tion de microorganismes | curité et aliments, Partenaires privés | | le produit alimentaire attendu (yaourt qui gonfle, moisissure qui se développe) |
| 26 | RETHINk Tomate et Bio agresseurs Modèle agronomique 2018-2020 | SYNGENTA, Inra | RPG Tomate (200 lignées de l'espèce sauvage *Solanum pimpinellifolium*), + RG 100 souches bactériennes (*Ralstonia solanacearum*) | 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 |
| 27 | VIVALDI 2016-2020 | 21 partenaires | RG de mollusques et de bactéries | Etude de l'impact des maladies chez les bivalves (classe de mollusques); Valorisation économique future |