UPOV

Union internationale pour la protection des obtentions végétales

F

Comité technique **TC/56/13**

Cinquante-sixième session **Original :** anglais
Genève, 26 et 27 octobre 2020 **Date :** 9 octobre 2020

EXAMEN DU DOCUMENT UPOV/INF/17 "DIRECTIVES CONCERNANT LES PROFILS D'ADN : CHOIX DES MARQUEURS MOLÉCULAIRES ET CONSTRUCTION D'UNE BASE DE DONNÉES Y RELATIVE ('DIRECTIVES BMT')"

Document établi par le Bureau de l'Union

Avertissement : le présent document ne représente pas les principes ou les orientations de l'UPOV

RÉSUMÉ

1. L'objet du présent document est d'examiner une proposition de révision du document UPOV/INF/17 "Directives concernant les profils d'ADN : choix des marqueurs moléculaires et construction d'une base de données y relative ('Directives BMT')", figurant dans le document UPOV/INF/17/2 Draft 4 et dans l'annexe du présent document.

2. Le TC est invité

a) à examiner la proposition de révision du document UPOV/INF/17 sur la base du document UPOV/INF/17/2 Draft 4,

b) à demander aux TWP d'examiner un projet de révision du document UPOV/INF/17/1 (document UPOV/INF/17/2 Draft 5) à leurs sessions de 2021 et

c) à noter qu'un projet de révision du document UPOV/INF/17 (UPOV/INF/17/2 Draft 6) serait proposé pour adoption par le Conseil à sa cinquante-cinquième session prévue le 29 octobre 2021, sous réserve de l'accord du TC à sa cinquante-septième session et du CAJ à sa soixante-dix-huitième session prévue en 2021.

3. Le présent document est structuré comme suit :

4. Les abréviations suivantes sont utilisées dans le présent document :

BMT :	Groupe de travail sur les techniques biochimiques et moléculaires, notamment les profils d'ADN
CAJ :	Comité administratif et juridique
TC :	Comité technique
TWA :	Groupe de travail technique sur les plantes agricoles
TWC :	Groupe de travail technique sur les systèmes d'automatisation et les programmes d'ordinateur
TWF :	Groupe de travail technique sur les plantes fruitières
TWO :	Groupe de travail technique sur les plantes ornementales et les arbres forestiers
TWP :	Groupe(s) de travail technique(s)
TWV :	Groupe de travail technique sur les plantes potagères

INFORMATIONS GÉNÉRALES

5. Les informations générales sur cette question sont fournies dans le document TC/55/7 “Techniques moléculaires”.

6. À sa cinquante-cinquième session[1], le TC a approuvé la proposition présentée par le BMT, à sa dix-huitième session[2], de demander à la France, aux Pays-Bas et à l’Union européenne d’établir un nouveau projet de document UPOV/INF/17 (document UPOV/INF/17/2 Draft 3) pour examen à la dix-neuvième session du BMT (voir le paragraphe 181 du document TC/55/25 “Compte rendu”).

7. À leurs sessions de 2020, les TWP ont examiné le document TWP/4/7 *“Molecular Techniques”*. Ils ont pris note des modifications à apporter au document UPOV/INF/17, approuvées par le BMT, telles qu’elles figurent dans l’annexe II du document TWP/4/7. Les TWP ont noté que le TC était convenu d’inviter la France, les Pays-Bas et l’Union européenne à élaborer une nouveau projet de document UPOV/INF/17 pour examen par le BMT à sa dix-neuvième session (voir les paragraphes 13 et 14 du document TWV/54/12 *“Report”*; les paragraphes 84 et 85 du document TWO/52/11 *“Report”*; les paragraphes 58 et 59 du document TWA/49/7 *“Report”*; les paragraphes 13 et 14 du document TWF/51/10 *“Report”*; et les paragraphes 66 et 67 du document TWC/38/11 *“Report”*).

8. Le programme d’élaboration de documents d’orientation et d’information pertinents prévoit la présentation d’un nouveau projet du document UPOV/INF/17 pour adoption éventuelle par le Conseil, à sa cinquante-cinquième session prévue en 2021, comme indiqué dans le document TC/56/14 “Élaboration de documents d’orientation et d’information”.

FAITS NOUVEAUX SURVENUS À LA DIX-NEUVIÈME SESSION DU BMT

9. À sa dix-neuvième session[3], le BMT a examiné les documents BMT/19/3 Rev. *“Revision of document INF/17”* et UPOV/INF/17/2 Draft 3 (voir les paragraphes 11 et 12 du document BMT/19/15 *“Report”*).

10. Le BMT est convenu que le projet d’orientations reproduit dans l’annexe du présent document devait être proposé au Comité technique comme point de départ pour la future révision du document UPOV/INF/17.

PROCHAINES ÉTAPES

11. Le projet d’orientations approuvé par le BMT, reproduit à l’annexe du présent document, est présenté dans le document UPOV/INF/17/2 Draft 4 avec changements apparents par rapport au texte du document UPOV/INF/17/1. Le TC souhaitera peut-être examiner le document UPOV/INF/17/2 Draft 4 comme point de départ pour une révision du document UPOV/INF/17.

12. Le TC souhaitera peut-être demander aux TWP d’examiner un projet de révision du document UPOV/INF/17/1 (document UPOV/INF/17/2 Draft 5) à leurs sessions de 2021, sur la base des modifications à apporter au document UPOV/INF/17/2 Draft 4, approuvées par le TC à sa cinquante-sixième session.

13. Sur la base des conclusions adoptées par le TC à sa cinquante-sixième session et par les TWP à leurs sessions de 2021, un projet de révision du document UPOV/INF/17 (UPOV/INF/17/2 Draft 6) serait proposé pour adoption par le Conseil à sa cinquante-cinquième session prévue le 29 octobre 2021, sous réserve de l’accord du TC à sa cinquante-septième session et du CAJ à sa soixante-dix-huitième session prévue en 2021.

---

[1] Tenue à Genève les 28 et 29 octobre 2019.

[2] Tenue à Hangzhou (Chine) du 16 au 18 octobre 2019.

[3] Accueillie par les États-Unis d’Amérique et tenue par des moyens électroniques du 23 au 25 septembre 2020.

14. Le TC est invité

a) à examiner la proposition de révision du document UPOV/INF/17 sur la base du document UPOV/INF/17/2 Draft 4,

b) à demander aux TWP d'examiner un projet de révision du document UPOV/INF/17/1 (document UPOV/INF/17/2 Draft 5) à leurs sessions de 2021 et

c) à noter qu'un projet de révision du document UPOV/INF/17 (UPOV/INF/17/2 Draft 6) serait proposé pour adoption par le Conseil à sa cinquante-cinquième session prévue le 29 octobre 2021, sous réserve de l'accord du TC à sa cinquante-septième session et du CAJ à sa soixante-dix-huitième session prévue en 2021.

[L'annexe suit]

ANNEXE

PROPOSITION DU BMT CONCERNANT LA RÉVISION DU DOCUMENT UPOV/INF/17

TABLE DES MATIÈRES

A. INTRODUCTION

Le présent document (Directives BMT) contient des directives sur des principes harmonisés relatifs à l'utilisation de marqueurs moléculaires qui serviront à produire des données moléculaires de haute qualité destinées à diverses applications. Seuls les marqueurs moléculaires d'ADN sont examinés dans le présent document.

Les Directives BMT ont également pour but de permettre l'élaboration de bases de données contenant des profils moléculaires de variétés végétales, qui peuvent être produits dans différents laboratoires à l'aide de diverses techniques. En outre, l'objectif est de définir des exigences élevées en ce qui concerne la qualité de marqueurs et l'intérêt de générer des données reproductibles à l'aide de ces marqueurs dans des situations où le matériel ou les réactifs chimiques peuvent varier. Des précautions particulières doivent être prises pour assurer la qualité des données saisies dans la base de données."

B. PRINCIPES GÉNÉRAUX

Aux fins de l'établissement du profil d'ADN d'une variété végétale, un ensemble de marqueurs moléculaires et une méthode de détection des marqueurs sont requis. Deux ensembles distincts de marqueurs moléculaires détectés au moyen d'une même méthode donneront deux profils d'ADN distincts pour une variété donnée. En revanche, deux méthodes distinctes pour détecter les allèles spécifiques d'un marqueur moléculaire donné sont censées donner des profils d'ADN identiques. La normalisation de la méthode de détection et de la technologie n'est pas requise, dès lors que les résultats remplissent les critères de qualité et que les profils d'ADN résultants sont cohérents. Quelle que soit la technologie utilisée pour détecter les ensembles de marqueurs définis, le génotype d'une variété particulière ne devrait pas être affecté.

Les ensembles de marqueurs moléculaires, les méthodes de détection de marqueurs et le processus ultérieur d'élaboration de bases de données peuvent être divisés en cinq phases distinctes :

1. Sélection des marqueurs moléculaires
2. Sélection de la méthode de détection
3. Validation et harmonisation de la méthode de détection

4. Construction de la base de données
5. Échange des données

Le présent document décrit ces différentes phases de façon plus détaillée. Il est considéré que ces phases sont indépendantes du stade de développement des technologies pour l'établissement de descriptions de génotypes et des améliorations futures en matière de séquençage à haut débit.

1. Sélection des marqueurs moléculaires

1.1 Ensembles de variétés pour le processus de sélection

Pour les profils d'ADN des variétés végétales et la construction de bases de données, les marqueurs moléculaires devraient être sélectionnés en fonction de l'objectif à atteindre. Pour débuter le processus de sélection des marqueurs, un nombre approprié de variétés (ensemble de développement) est nécessaire pour refléter au mieux la diversité observée au sein du groupe, de la plante, de l'espèce ou du type pour lequel les marqueurs doivent avoir un pouvoir de discrimination. Une nouvelle sélection est réalisée à partir des profils de variétés supplémentaires (ensemble de validation) pour mesurer la performance des marqueurs. Les critères pour le choix de l'ensemble de validation peuvent être :

a) variétés ou lignées génétiquement très proches, NIL, RIL
b) lignées parentales et descendance
c) variétés génétiquement proches mais morphologiquement distinctes (p. ex. mutants)
d) certaines variétés morphologiquement proches avec des généalogies différentes
e) différents lots d'une même variété
f) différentes origines de la même variété

1.2 Marqueurs moléculaires – considérations relatives aux performances

Les critères généraux suivants utilisés pour sélectionner un marqueur spécifique ou un ensemble de marqueurs s'appliquent quelle que soit leur utilisation :

a) la répétabilité, la reproductibilité et la robustesse au sein d'un laboratoire et d'un laboratoire à l'autre en terme de notation des données;

b) les sources possibles de marqueurs moléculaires
   – marqueurs moléculaires tirés de ressources publiques
   – marqueurs moléculaires tirés de ressources non publiques, présélection et sélection de puces et de matrices spécifiques d'une espèce disponibles dans le commerce
   – marqueurs moléculaires sélectionnés à partir de données relatives à la séquence nouvellement générées;

c) éviter, dans la mesure du possible, des marqueurs avec des allèles "null" (c'est-à-dire des allèles dont l'effet se manifeste par une absence de produit PCR au niveau moléculaire), ce qui n'est à nouveau pas indispensable, mais conseillé;

d) permettre une notation aisée, objective et indiscutable des profils de marqueurs. Ces marqueurs performants sont préférables aux profils de marqueurs complexes qui sont délicats en termes d'interprétation. Des réponses claires et tranchées permettent également de faciliter l'harmonisation;

e) les marqueurs codominants sont généralement préférables aux marqueurs dominants du fait de leur pouvoir de discrimination plus élevé;

f) les marqueurs peuvent être situés dans des régions codantes ou non codantes; et

g) l'utilisation de marqueurs moléculaires est spécifique d'une espèce et devrait tenir compte des particularités de reproduction ou de multiplication de l'espèce.

Il est reconnu que certaines utilisations particulières peuvent imposer l'application de considérations supplémentaires, notamment (mais pas uniquement) les suivantes :

a) le nombre de marqueurs devrait être équilibré par rapport à la précision du génotype requise pour atteindre l'objectif. Le nombre de marqueurs pour atteindre la résolution ou le pouvoir de discrimination

requis dépend du type de marqueur (dominant/codominant; bi/multiallélique), de l'espèce et de la qualité de la performance du marqueur;

b) La couverture du génome et du déséquilibre de liaison devrait refléter les objectifs. Connaître la position physique ou génétique des marqueurs sélectionnés sur le génome n'est pas indispensable, mais permet un bon choix des marqueurs;

2. Sélection de la méthode de détection

2.1 Méthodes relatives aux profils d'ADN – considérations générales

2.1.1 Les critères ci-après sont importants dans la sélection des méthodes relatives aux profils d'ADN pour générer des données moléculaires de qualité :

a) reproductibilité de la production de données entre laboratoires et plateformes de détection (différents types de matériel);
b) possibilité de répétition dans le temps;
c) pouvoir de discrimination;
d) temps et main d'œuvre requis;
e) robustesse des performances dans le temps et les conditions (sensibilité aux changements subtils du protocole ou des conditions);
f) flexibilité de la méthode, possibilité de faire varier le nombre d'échantillons ou le nombre de marqueurs;
g) l'interprétation des données produites est indépendante du matériel;
h) pérennité des bases de données;
i) accessibilité de la méthode;
j) indépendante d'un appareil, d'une composition chimique, d'un fournisseur, de partenaires ou de produits spécifiques;
k) possibilité d'automatisation;
l) possibilité de multiplexage; et
m) rentabilité (équilibre des coûts, du nombre d'échantillons et du nombre de marqueurs).

2.2. Accès à la technologie

Certains marqueurs et matériel moléculaires sont accessibles au public. Cela étant, l'obtention de marqueurs de grande qualité suppose vraisemblablement un investissement important, de sorte qu'il est probable que certains marqueurs et autres méthodes ou matériel soient protégés par des droits de propriété intellectuelle. L'UPOV a mis au point des orientations pour l'utilisation de produits ou de méthodes qui font l'objet de droits de propriété intellectuelle et il convient de les suivre. Il est recommandé de régler les questions concernant les droits de propriété intellectuelle avant d'entreprendre des travaux d'amélioration.

3. Validation et harmonisation d'un ensemble de marqueurs et d'une méthode de détection

3.1 Validation et harmonisation – considérations générales

Les marqueurs moléculaires et les méthodes de détection doivent être robustes et donner lieu à des profils d'ADN cohérents. Les performances des marqueurs moléculaires et des méthodes utilisées pour l'établissement de descriptions de génotypes sont évaluées dans un processus de validation. Dans le cas de bases de données partagées, la cohérence des profils d'ADN dans les différents laboratoires est évaluée dans le processus d'harmonisation à l'aide de différents matériels et compositions chimiques. L'utilisation de marqueurs et de méthodes validés aboutira à des résultats harmonisés.

3.2 Considérations relatives aux performances – validation des marqueurs et des méthodes

L'ensemble de marqueurs sélectionné devrait être adapté à son objet. Sa précision devrait être mesurée. Pour déterminer si une méthode et un ensemble de marqueurs d'ADN conviennent, plusieurs éléments doivent être pris en considération :

a) pouvoir de discrimination/pouvoir informatif;
b) possibilité de répétition : lorsque des résultats d'essai identiques sont obtenus avec la même méthode, sur des objets identiques, dans le même laboratoire, par le même opérateur, en utilisant le même équipement dans de courts intervalles de temps;

c) reproductibilité : lorsque les résultats des essais identiques sont obtenus avec la même méthode, sur des objets identiques, dans le même laboratoire ou entre différents laboratoires, avec différents opérateurs, en utilisant des équipements différents;

d) robustesse : mesure de sa capacité à ne pas être affecté par des écarts faibles mais délibérés par rapport aux conditions expérimentales décrites dans les paramètres de la procédure, donne une indication de sa fiabilité en utilisation normale; et

e) taux d'erreur.

Les définitions des caractères relatifs à la performance sont fondées sur la norme ISO 16 577:2016

3.3 Considérations relatives à la cohérence

Pour obtenir une cohérence des résultats, le processus d'harmonisation des marqueurs et des méthodes entre les différents laboratoires dans le cas d'une base de données partagée (test d'étalonnage) doit inclure ce qui suit :

a) Utilisation d'une collection de variétés définie représentant un large éventail d'allèles comme référence dans tous les laboratoires pour vérifier la cohérence entre les laboratoires

b) Inclusion d'échantillons doubles, sous-échantillons, plantes individuelles d'une variété pour vérifier la cohérence des profils d'ADN et estimer le taux d'erreur entre les laboratoires

c) Accords sur la notation des données moléculaires. La nécessité de mettre au point un protocole de notation des allèles/bandes entre les laboratoires dépend du type de marqueur utilisé (essentiel pour les SSR). Le protocole pourrait permettre de déterminer comment noter les éléments suivants :

i. allèles rares (c'est-à-dire allèles à locus spécifique apparaissant dans une population à une fréquence inférieure à un seuil convenu (habituellement 5 à 10%);

ii. allèles à fréquence nulle (un allèle dont l'effet consiste en l'absence d'un produit PCR à l'échelle moléculaire);

iii. bandes "faibles" (c'est-à-dire des bandes dont l'intensité est inférieure à un seuil de détection convenu, fixé soit empiriquement, soit automatiquement, et dont la notation peut être remise en question);

iv. données manquantes (c'est-à-dire tout locus pour lequel aucune donnée n'est enregistrée, pour quelque raison que ce soit, dans une ou plusieurs variétés); et

v. bandes monomorphes ou notations d'allèles non informatives (allèles/bandes apparaissant dans chaque variété analysée, c'est-à-dire qui ne sont pas monomorphes dans une collection particulière de variétés).

4. Construction d'une base de données spécifique à une espèce

Les données qui sont stockées dans une base de données et la façon dont elles sont stockées doivent fournir des indications sur le processus de production des données. Par conséquent, la construction d'une base de données devrait tenir compte des différents niveaux de traitement des données (c.-à-d. données brutes, données séquentielles, etc.). La base de données devrait contenir les résultats finaux, p. ex. le profil d'ADN ainsi que des indications sur la façon dont il a été établi avec une description de la méthode employée par le laboratoire et les étapes de calcul.

4.1 Recommandations pour la conception d'une base de données

Dans la conception de bases de données, il convient de tenir compte des éléments suivants :

a) L'architecture de la base de données devrait être flexible, par exemple permettre de stocker à la fois des fichiers plats et des fichiers d'archives compressés.

b) Des tableaux et des entrées distincts sont nécessaires pour le travail expérimental en laboratoire, le traitement des données et la notation des allèles.

c) Le stockage d'informations à différents niveaux, par exemple notations des allèles et toute règle d'interprétation ayant présidé à la décision, et les liens vers les données brutes (fichiers tiff, fichiers bam) qui ont été produites.

d) Pour les données de séquençage, des fichiers de détection des variants au format VCF ou BCF correspondant à la version standard 4.2 ou à une version supérieure. Les entrées d'en-tête doivent contenir le nom et la version des différents scripts utilisés pour la cartographie des lectures, le filtrage des lectures, la détection des variants et le filtrage des variants, de sorte qu'un bioinformaticien puisse répéter l'analyse.

e) Dans le cas d'échantillons répétés, lorsque le profil d'ADN ne correspond pas, l'enregistrement doit être marqué ou filtré, le cas échéant. Les règles appliquées dans ces cas doivent être documentées dans un référentiel de codes accessible au public établi à partir du fichier de détection des variants. Des fréquences pourraient également être utilisées pour des variétés hétérogènes.

f) Validation des données VCF et BCF par rapport aux exigences pertinentes.

g) Données faciles à partager (p. ex. API).

4.2 Conditions requises pour le matériel végétal

La source et le type de matériel, ainsi que le nombre d'échantillons à stocker et partager dans la base de données, doivent être pris en considération.

4.2.1 Source du matériel végétal

Le matériel végétal à analyser doit être un échantillon authentique et représentatif de la variété et, lorsque c'est possible, être obtenu à partir de l'échantillon de la variété utilisé pour l'examen aux fins de l'octroi des droits d'obtenteur et de l'enregistrement officiel. L'utilisation de ces échantillons devra faire l'objet, selon le cas, d'une autorisation de l'autorité compétente, de l'obtenteur ou du conservateur. Les plantes sur lesquelles les échantillons sont prélevés devraient pouvoir être retrouvées dans le cas où il apparaîtrait par la suite que certaines d'entre elles ne sont pas représentatives de la variété.

4.2.2 Type de matériel végétal

Le type de matériel végétal à échantillonner et la procédure d'échantillonnage à suivre en vue de l'extraction d'ADN dépendront, dans une large mesure, de l'espèce végétale concernée. Ainsi, dans le cas des variétés reproduites par voie sexuée, la semence peut servir de source d'ADN, alors que, dans le cas des variétés multipliées par voie végétative, l'ADN peut être extrait à partir des feuilles. Quelle que soit la source du matériel végétal, la méthode d'échantillonnage et d'extraction de l'ADN devrait être référencée. En outre, il conviendra de vérifier que les méthodes d'échantillonnage et d'extraction permettent d'obtenir des résultats d'analyse ADN stables.

4.2.3 Taille et type de l'échantillon (échantillons globaux ou individuels)

Il est essentiel que les échantillons prélevés pour l'analyse soient représentatifs de la variété. Il convient de prendre en considération les particularités de la reproduction ou multiplication de la variété (voir l'Introduction générale).

4.2.4 Échantillon d'ADN de référence

Une collection d'ADN de référence peut être établie à partir du matériel végétal échantillonné. La méthode d'échantillonnage devrait suivre les procédures recommandées et les critères de qualité pour l'extraction d'ADN doivent être établis. Tous deux doivent être documentés.
Les échantillons d'ADN doivent être conservés dans des conditions empêchant leur dégradation (p. ex. stockage à -80 °C). Le transfert d'échantillons d'ADN de référence est décrit dans la section 1 du document TGP/5.

4.3 Traitements des données relatives aux séquences

Un journal détaillé du pipeline de traitement de données pourrait indiquer :

a) le type et les versions des outils;
b) la ligne de commande utilisée pour l'outil, y compris les seuils;
c) la reproductibilité (comptage);

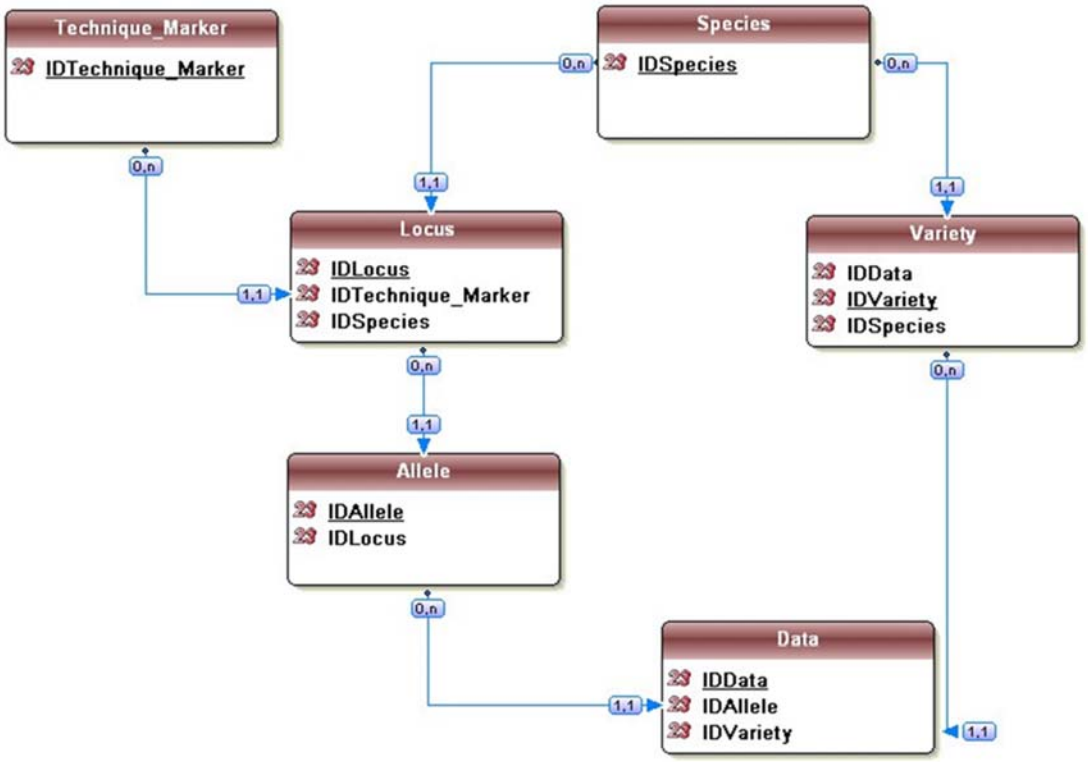
d) les possibilités de partage des données et processus;
e) les données d'alignement brutes (fichiers BAM ou CRAM) devraient être stockées si possible;
f) les fichiers VCF multiéchantillons ne conviennent pas, un fichier VCF par variété doit être présent;
g) si des fichiers VCF sont stockés, toutes les positions (variants et non variants) et leur profondeur devraient être stockées;
h) des approches à la foi heuristiques et probabilistes devraient être envisagées et comparées pour les méthodes de détection;
i) les bases de données devraient faciliter l'entrée et la sortie de données de détection des variants dans un format normalisé (VCF ou BCF);
j) le pipeline de traitement des données devrait aboutir à un fichier journal détaillé qui devrait être stocké conjointement avec les données de détection des variants;
k) si possible, les données brutes devraient être stockées de sorte que le traitement des données puisse être répété avec de nouveaux outils ou avec des outils mis à jour; et
l) la valeur p ou l'incertitude pour un allèle donné devrait être stockée.

4.4 Type de base de données

Il existe de nombreux moyens de stockage des données moléculaires. C'est pourquoi il importe que la structure de la base de données soit élaborée de façon à s'appliquer à toutes les utilisations prévues des données.

4.5 Modèle de base de données

Le modèle de base de données devrait être défini par des experts en bases de données informatiques, en liaison avec les utilisateurs. Chaque modèle devrait contenir au minimum six objets principaux : espèce, variété, méthode de détection des marqueurs, marqueur, locus et allèle. Pour les variants obtenus à partir de données de séquençage, les fichiers VCF peuvent être stockés dans une base de données relationnelle ou non-SQL. Dans ce cas, chaque enregistrement de la base de données correspondant à une variante présente une version génomique, un chromosome, une position et un allèle de référence définis.



4.6 Dictionnaire de données

4.6.1 Dans une base de données, chacun des objets devient un tableau à l'intérieur duquel des champs sont définis. Par exemple :

a) Type de marqueur : indique le code ou le nom de la technique ou le type de marqueur utilisé, p. ex. SSR, SNP, etc.

b) Position de génome de référence ou code locus : de préférence, une version d'assemblage de génome, un chromosome et une position devraient être fournis si un génome de référence est disponible pour l'espèce concernée, p. ex. SL2.50ch05:63309763 pour la tomate *Solanum lycopersicum*, version d'assemblage 2.50, en position 63309763 sur le chromosome 5. Si aucun génome de référence n'est disponible ou que l'emplacement est inconnu, le nom ou le code du locus pour les espèces concernées peut être utilisé, p. ex. gwm 149, A2, etc.

c) Génotype : pour les profils SNP, la composition allélique du SNP ou du MNP devrait être donnée, p. ex. A/T ou A/A. Pour d'autres techniques, le génotype indique le nom ou le code de l'allèle d'un locus donné pour les espèces concernées, p. ex. 1, 123, etc.

d) Profondeurs de l'allèle ou valeur des données : pour les SNP obtenus à partir de données de séquençage de nouvelle génération, cela devrait indiquer la profondeur de la couverture des allèles, p. ex. 10/20 pour un allèle A/T, A étant couvert par 10 lectures et T par 20. Autrement, cela indique une valeur de données pour l'échantillon sur un locus-allèle déterminé, p. ex. 0 (absence), 1 (présence), 0,25 (fréquence), etc.

e) Variété : dénomination de la variété ou référence de l'obtenteur : la variété est l'objet pour lequel les données sont obtenues.

f) Type de variété : p. ex. lignée endogame ou hybride.

g) Espèce : l'espèce est indiquée par le nom botanique ou le nom commun national, qui peut également renvoyer au type de variété (p. ex. utilisation, type hivernal/printanier, etc.). L'utilisation du code UPOV est recommandée pour éviter les problèmes liés aux synonymes.

4.6.2 Dans chaque tableau, le nombre de champs, leur nom et leur définition, les valeurs possibles et les règles à suivre doivent être définis dans le "dictionnaire de données".

4.7 Accès aux données et propriété des données

Avant toute chose, il est recommandé de régler les questions concernant la propriété des données et l'accès à la base de données.

5. Échange des données

5.1 Scénarios d'échange des données

À des fins de coopération, le modèle de données devrait permettre différents types de scénarios, notamment l'échange de données produites à partir d'un ensemble normalisé de marqueurs pour une plante spécifique (scénario 1), et la recherche et la consultation de données relatives à des variétés sélectionnées, générées à partir du même ensemble normalisé de marqueurs (scénario 2). Les détails techniques des deux scénarios sont décrits dans l'annexe "Scénarios d'échange de données et méthodes de transfert de données".

5.2 Méthodes d'échange de données

5.2.1 La transmission de données relatives à l'empreinte génétique peut inclure toute une série d'informations, comme des locus, des échantillons, l'ADN, des données relatives à l'empreinte et des profils d'empreinte génétique. La méthode de transmission des données doit être déterminée par le contenu à transférer et doit tenir compte des éléments suivants :

a) le volume de données,
b) la complexité des données,
c) les conditions relatives aux fonctions de requête ou de recherche.

Les détails techniques des méthodes de transfert des données sont décrits dans l'annexe "Scénarios d'échange de données et méthodes de transfert de données".

5.2.2 Les formats de données les plus utilisés sont les suivants zip, csv, json et xml. Leurs caractéristiques respectives sont les suivantes :

1) Le format zip permet d'obtenir divers fichiers d'information dans le format d'origine et, grâce à son taux de compression élevé et à sa facilité de transmission, il convient aux données volumineuses et complexes.

2) Le format csv est plus adapté pour les informations en format de données simple, qui a l'avantage de réduire le nombre de données non valides et d'accélérer la vitesse de traitement.

3) Les formats json et xml peuvent contenir des informations relatives aux données de caractères plus complexes et des informations plus redondantes, mais tous deux offrent une bonne lisibilité.

6. Résumé

On trouvera ci-après un résumé de la méthode qu'il est recommandé de suivre en vue de l'obtention de profils d'ADN de qualité des variétés, y compris du choix et de l'utilisation des marqueurs moléculaires et de la construction de bases de données moléculaires partagées et durables (c'est-à-dire de bases de données pouvant être alimentées à l'avenir par des données provenant de plusieurs sources, indépendamment de la technique utilisée).

a) envisager une méthode plante par plante;
b) déterminer les types et les sources de marqueurs acceptables;
c) déterminer les plateformes et l'équipement de détection acceptables;
d) convenir des laboratoires qui participeront à l'essai;
e) se mettre d'accord sur les questions de qualité;
f) vérifier la source du matériel végétal utilisé;
g) déterminer les marqueurs à utiliser lors de la phase préliminaire d'évaluation en commun, qui doit impliquer plusieurs laboratoires et des équipements de détection différents;
h) réaliser une évaluation;
i) mettre au point et approuver un protocole de notation des données moléculaires;
j) convenir de l'ensemble matériel/référence végétal(e) à analyser; et de la (des) source(s);
k) analyser la collection de variétés retenue, dans différents laboratoires et au moyen d'équipements de détection différents, en utilisant des échantillons doubles et en échangeant les échantillons/extraits d'ADN en cas de problème;
l) utiliser dans toutes les analyses des éléments de référence (variétés, échantillons d'ADN et allèles, le cas échéant);
m) vérifier toutes les étapes (y compris la saisie des données) – automatiser les opérations au maximum;
n) mener un essai "en aveugle" dans différents laboratoires à l'aide de la base de données;
o) adopter les procédures relatives à l'adjonction de nouvelles données.

C. LISTE DES SIGLES

API	Application Programming Interface (interface de programmation)
BAM	Binary Alignment Map
BCF	Binary Call Format
CRAM	Compressed Reference-oriented Alignment Map
MNP	Multiple Nucleotide Polymorphism
NGS	Next Generation Sequencing
NIL	Near Isogenic Line
RIL	Recombinant Inbred Line
SAM	Sequence Alignment Map
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
SSR	Simple Sequence Repeats
TIFF	Tagged Image File Format
VCF	Variant Call Format

[L'appendice de l'annexe suit]

APPENDICE DE L'ANNEXE

SCÉNARIOS D'ÉCHANGE DES DONNÉES ET MÉTHODES DE TRANSFERT DE DONNÉES

A : Scénarios d'échange des données

Scénario 1 : échange de données produites à partir d'un ensemble normalisé de marqueurs pour une plante spécifique

Afin d'échanger des données sur l'ensemble de marqueurs utilisés pour une plante spécifique, le service Web suivant peut être utilisé :
https://office.org/locus?upov_code={upovcode}&type={marker type}&method={observation method}

Par exemple, pour obtenir des informations sur l'ensemble de marqueurs pour le maïs en utilisant la méthode SSR et CE, il faut accéder à l'URL suivante :
https://office.org/locus?upov_code=ZEAAA_MAY&type=SSR&method=CE

Le résultat serait :

{"techniqueid": "CN_SSR_ZEAA_MAY_CE_V_1",
"description": "Laboratory method description"
["locusid": "M01",
"alleles":
["alleleid": "238/256",
"examplevariety":
],
["alleleid": "238/271",
"examplevariety":
],
["alleleid": "246/246",
"examplevariety":
],
["alleleid": "246/248",
"examplevariety":
],
["alleleid": "246/250",
"examplevariety":
],
["alleleid": "246/254",
"examplevariety":
],
["alleleid": "246/256",
"examplevariety":
],
["alleleid": "246/260",
"examplevariety":
],
["alleleid": "246/277",
"examplevariety":
],
["alleleid": "246/284",
"examplevariety":
],
["alleleid": "246/288",
"examplevariety":
],
["alleleid": "248/250",
"examplevariety":
],
["alleleid": "248/256",
"examplevariety":
],
["alleleid": "248/271",
"examplevariety":
],
["alleleid": "248/290",
"examplevariety":
],
["alleleid": "250/250",
"examplevariety":
],
["alleleid": "250/252",
"examplevariety":
],
["alleleid": "250/256",
"examplevariety":
],
["alleleid": "250/275",
"examplevariety":
],
["alleleid": "252/256",
"examplevariety":
],
["alleleid": "252/260",
"examplevariety":
],
["alleleid": "252/271",
"examplevariety":
],
["alleleid": "252/273",
"examplevariety":
],
["alleleid": "252/282",
"examplevariety":
],
["alleleid": "254/254",
"examplevariety":
],
["alleleid": "254/271",
"examplevariety":
],
["alleleid": "254/284",
"examplevariety":
],
["alleleid": "254/286",
"examplevariety":
],
["alleleid": "256/256",
"examplevariety":
],
["alleleid": "256/264",
"examplevariety":
],
["alleleid": "256/266",
"examplevariety":
],
["alleleid": "256/271",
"examplevariety":
],
["alleleid": "256/284",
"examplevariety":
],
["alleleid": "256/286",
"examplevariety":
],
["alleleid": "258/258",
"examplevariety":
],
["alleleid": "264/284",
"examplevariety":
],
["alleleid": "271/292",
"examplevariety":
]
],

["locusid"="M02".
"alleles": […]
]} vi

Scénario 2 : recherche et consultation de données relatives à des variétés sélectionnées, générées à partir du même ensemble normalisé de marqueurs

Afin de rechercher et de consulter les données moléculaires d'une variété, le service Web suivant peut être utilisé :
https://office.org/variety?id={irn}&techniqueid={technique_code} vi

Par exemple,
https://office.org/variety?id=XU_30201800000140 &techniqueid= CN_SSR_ZEAA_MAY_CE_V_1 vi

Le résultat serait :

{"techniqueid": "CN_SSR_ZEAA_MAY_PAGE ",
"varietyid": " XU_30201800000140 ",
"computationalsteps": "xxxxxxxxxxxx"
"data":
[
"id": "M01",
"value" : "254/254"
],
[
"id": "M02",
"value" : "347/347"
],
[
"id": "M03",
"value" : "292/292"
],
[
"id": "M04",
"value" : "361/361"
],
…
} vi

B : Méthodes de transfert de données

Voici un exemple de création d'un paquet d'empreintes génétiques au format zip pour la transmission de données. Cette méthode doit d'abord utiliser des identifiants indépendants pour identifier les échantillons, l'ADN, les données d'empreinte génétique et le registre des empreintes génétiques. Ensuite, le fichier de données au format json contient tous les locus, tous les échantillons et toutes les informations relatives à l'ADN. Chaque donnée d'empreinte génétique est stockée séparément dans son propre fichier au format json. L'identifiant de l'empreinte génétique sera lié au locus correspondant des données d'empreinte génétique, et tous les fichiers de données d'empreinte génétique et les fichiers de spectre d'empreinte génétique seront stockés séparément dans le répertoire correspondant. La structure du paquet de données d'empreinte génétique est donc la suivante :

zip/markers.json
zip/samples.json
zip/dnas.json
zip/genes/gene_id_1.json
zip/genes/gene_id_2.json
.....
zip/genes/gene_id_n.json
zip/maps/map_id_1.png
zip/maps/map_id_2.png
.....
zip/maps/map_id_m.png

Le paquet d’empreintes génétiques en format zip peut être étendu pour inclure plus d’informations. Le cœur du paquet est le fichier de données d’empreinte génétique, qui constitue le cœur de la corrélation, de sorte que la corrélation entre les parties peut être correctement analysée, ce qui permet la transmission des données entre des systèmes différents.

[Fin de l’appendice de l’annexe et du document]