UPOV
Internationaler Verband zum Schutz von Pflanzenzüchtungen

G

UPOV/INF/17/2

Original: englisch
Datum: 21. September 2021

RICHTLINIEN FÜR DIE DNS-PROFILIERUNG: AUSWAHL MOLEKULARER MARKER UND AUFBAU VON DATENBANKEN („BMT-RICHTLINIEN")

Dokument vom Rat
am 21. September 2021
auf dem Schriftweg angenommen

INHALTSVERZEICHNIS

A. EINLEITUNG

Dieses Dokument (BMT-Richtlinien) soll Anleitung zu harmonisierten Grundsätzen für die Verwendung molekularer Marker geben, um qualitativ hochwertige molekulare Daten für eine Reihe von Anwendungen zu erzeugen. In diesem Dokument werden nur molekulare DNS-Marker berücksichtigt.

Die BMT-Richtlinien sollen ferner den Aufbau von Datenbanken mit molekularen Profilen von Pflanzensorten behandeln, die möglicherweise mit verschiedenen Techniken in verschiedenen Labors erzeugt werden. Ziel ist es zudem, hohe Anforderungen zu stellen an die Qualität von Markern und an das Bestreben, reproduzierbare Daten anhand dieser Marker zu erzeugen, wenn sich Ausrüstungen und/oder Reaktionschemikalien ändern. Spezifische Vorsichtsmaßnahmen sind zu treffen, um qualitativ hochwertige Eingaben in eine Datenbank sicherzustellen.

B. ALLGEMEINE GRUNDSÄTZE

Die DNS-Profilierung einer Pflanzensorte erfordert einen Satz molekularer Marker und eine Methode, diese festzustellen. Zwei verschiedene molekulare Markersätze, die mit derselben Methode festgestellt wurden, führen bei einer bestimmten Sorte zu zwei unterschiedlichen DNS-Profilen. Dagegen wird die Feststellung der spezifischen Allele eines gegebenen molekularen Markersatzes mit zwei verschiedenen Methoden voraussichtlich zu identischen DNS-Profilen führen. Eine Standardisierung der Detektionsmethode und -technologie ist nicht erforderlich, solange die Qualitätskriterien erfüllt werden und die gewonnenen DNS-Profile konsistent sind. Die Technologie, die zur Feststellung gegebener Markersätze eingesetzt wird, sollte den Genotyp einer bestimmten Sorte nicht beeinflussen.

Bei den molekularen Markersätzen, den Methoden zur Markerfeststellung und dem anschließenden Aufbau der Datenbank lassen sich fünf verschiedene Phasen unterscheiden:

1. Auswahl molekularer Marker
2. Auswahl der Detektionsmethode
3. Validierung und Harmonisierung des Nachweisverfahrens
4. Aufbau der Datenbank
5. Datenaustausch

Diese verschiedenen Phasen werden in diesem Dokument eingehender beschrieben. Es wird davon ausgegangen, dass diese Phasen unabhängig sind vom Entwicklungsstand der Genotypisierungstechnologien und von künftigen Verbesserungen der Hochdurchsatz-Sequenzierung.

1. Auswahl der molekularen Marker

1.1 Sortensätze für das Auswahlverfahren

Für die DNS-Profilierung von Pflanzensorten und den Aufbau von Datenbanken sollten die molekularen Marker zielorientiert ausgewählt werden. Für die Einleitung des Markerauswahlverfahrens wird eine geeignete Anzahl von Sorten benötigt (Entwicklungsserie), um die Vielfalt widerzuspiegeln, die innerhalb der Gruppe/Pflanze/Art /bzw. des Typs beobachtet wird, für die bzw. den die Marker unterscheidend sein sollen. Eine weitere Auswahl erfolgt mittels Profilierung zusätzlicher Sorten (Validierungsserie), um die Leistung der Marker zu messen. Als Kriterien für die Auswahl der Validierungsserie kommen in Frage:

a) genetisch sehr ähnliche Sorten oder Linien, NILs, RILs
b) Elternlinien und Nachkommen
c) genetisch nah verwandte, jedoch morphologisch unterschiedliche Arten (z. B. Mutanten)
d) einige morphologisch ähnliche Sorten mit unterschiedlichem Stammbaum
e) verschiedene Partien derselben Sorte
f) unterschiedliche Herkunft innerhalb derselben Sorte

1.2 Molekulare Marker – Leistungsaspekte

Die folgenden allgemeinen Aspekte für die Auswahl eines spezifischen Markers oder Markersatzes sollen ungeachtet der Verwendung der Marker geeignet sein:

a) Laborinterne und laborübergreifende Wiederholbarkeit, Solidität und Reproduzierbarkeit bei der Auswertung der Daten;

b) Mögliche Quellen für molekulare Marker
- Molekulare Marker, die aus öffentlichen Quellen stammen
- Molekulare Marker, die aus nichtöffentlichen Quellen stammen oder durch Aussortieren und Auswählen von handelsüblichen artspezifischen Chips und Arrays gewonnen wurden
- Molekulare Marker, die aus neu generierten Sequenzdaten ausgewählt wurden;

c) nach Möglichkeit Vermeiden von Markern mit „Nullallelen“ (d. h. ein Allel, das das Fehlen eines PCR-Produkts auf molekularer Ebene bewirkt), was ebenfalls nicht wesentlich, jedoch ratsam ist;

d) Zulassen einer problemlosen, objektiven und eindeutigen Auswertung von Markerprofilen. Diese leistungsstarken Marker werden gegenüber komplexen Markerprofilen, die zu Mehrdeutigkeit neigen, bevorzugt. Klare Schwarz-Weiß-Antworten erleichtern zudem die Harmonisierung;

e) Ko-dominante Marker werden im Allgemeinen gegenüber dominanten Markern bevorzugt, da sie eine höhere Unterscheidungskraft haben;

f) Marker können in kodierenden und/oder nichtkodierenden Regionen lokalisiert sein; und

g) Die Verwendung molekularer Marker ist artspezifisch und sollte die Besonderheiten der Vermehrung der Art berücksichtigen.

Es wird eingeräumt, dass spezifische Anwendungen bestimmte zusätzliche Kriterien erfordern könnten:

i. Die Anzahl der Marker sollte im Verhältnis stehen zur jeweils geforderten Genauigkeit des Genotyps. Die Anzahl der zum Erreichen der erforderlichen Auflösung oder Unterscheidungskraft einzusetzenden Marker ist abhängig vom Markertyp (dominant/kodominant, bi-/multiallelisch), der Art und der Qualität der Markerleistung;

ii. Die Einbeziehung des Genombereichs und des Kopplungs-Ungleichgewichts sollte die Ziele widerspiegeln. Das Bekanntsein der physischen und/oder genetischen Position des ausgewählten Markers im Genom ist zwar nicht wesentlich, ermöglicht aber eine sinnvolle Markerauswahl.

2. Auswahl der Detektionsmethode

2.1 DNS-Profilierungsverfahren – allgemeine Überlegungen

2.1.1 Wichtige Überlegungen zur Wahl eines DNS-Profilierungsverfahrens zur Gewinnung hochwertiger molekularer Daten sind:

a) Reproduzierbarkeit der Datengenerierung in den Laboren und Detektionsplattformen sowie zwischen diesen (verschiedenen Ausrüstungstypen);
b) Wiederholbarkeit im Zeitablauf;
c) Unterscheidungskraft;
d) Zeit- und Arbeitsintensität des Verfahrens;
e) Belastbarkeit hinsichtlich der zeitlichen Gegebenheiten und der Bedingungen (Empfindlichkeit gegenüber subtilen Änderungen von Ablauf oder Bedingungen);
f) Flexibilität des Verfahrens; Möglichkeit, die Anzahl der Proben und/oder der Marker zu variieren;
g) Auswertung der Daten ist von der Ausrüstung unabhängig;
h) Nachhaltigkeit der Datenbanken;
i) Zugänglichkeit der Methodik;
j) nicht abhängig von besonderen Maschinen, Chemikalien, Zulieferern, Partnern oder Produkten;
k) zur Automatisierung geeignet;
l) für Multiplexing geeignet; und
m) kostengünstig; Kosten, Zahl der Proben und Zahl der Marker stehen zueinander im Verhältnis.

2.2 Zugang zur Technik

Einzelne molekulare Marker und Materialien sind öffentlich verfügbar. Es dürften jedoch hohe Investitionen erforderlich sein, um hochqualitative Marker zu erzielen. Infolgedessen können Marker und andere Verfahren und sonstiges Material durch Rechte des geistigen Eigentums geschützt sein. Die UPOV erarbeitete eine Anleitung zur Nutzung der Produkte oder Methodiken, die Gegenstand von Rechten des geistigen Eigentums sind. Diese sollte befolgt werden. Es ist zu empfehlen, dass Angelegenheiten bezüglich der Rechte des geistigen Eigentums zu Beginn jeder Entwicklungsarbeit behandelt werden.

3. Validierung und Harmonisierung von Markersatz und Detektionsmethode

3.1 Validierung und Harmonisierung – allgemeine Überlegungen

Molekulare Marker und Detektionsmethoden müssen solide sein und zu konsistenten DNS-Profilen führen. Die Leistung der molekularen Marker und der Verfahren zur Genotypisierung wird im Rahmen des Validierungsprozesses evaluiert. Bei gemeinsamen Datenbanken wird die Übereinstimmung der DNS-Profile im Rahmen des Harmonisierungsprozesses in verschiedenen Laboren evaluiert, wobei unterschiedliche Ausrüstungen und Chemikalien verwendet werden. Die Verwendung validierter Marker und Verfahren wird zu harmonisierten Ergebnissen führen.

3.2 Leistungsaspekte – Validierung von Markern und Verfahren

Der gewählte Markersatz sollte für den Zweck geeignet sein. Die Genauigkeit muss gemessen werden. Um die Eignung eines Verfahrens oder DNS-Markersatzes festzustellen, sind diverse Punkte zu berücksichtigen:

a) Unterscheidungsvermögen/Informativität;
b) Wiederholbarkeit; wobei identische Prüfungsergebnisse anhand derselben Methode, an identischen Prüfgegenständen, im gleichen Labor, vom selben Labortechniker, unter Verwendung derselben Geräte und Ausstattung innerhalb kurzer Zeitabstände erzielt werden;

c) Reproduzierbarkeit; wobei Prüfergebnisse anhand derselben Methode, an identischen Prüfgegenständen, im selben Labor oder in verschiedenen Laboren, von verschiedenen Labortechnikern, unter Verwendung unterschiedlicher Geräte und Ausstattung erzielt werden;

d) Robustheit; ein Maß für seine Fähigkeit, von kleinen, aber gewollten Abweichungen von den in den Verfahrensparametern beschriebenen Versuchsbedingungen unbeeinflusst zu bleiben, und ein Hinweis auf seine Zuverlässigkeit bei normaler Verwendung; und

e) Fehlerrate.

Definitionen der Leistungsmerkmale basieren auf: ISO/16577: 2016

3.3 Konsistenzaspekte

Um eine Konsistenz der Ergebnisse zu erreichen, sollte beim Prozess der Harmonisierung von Markern und Verfahren zwischen verschiedenen Laboren bei gemeinsamer Datenbank (Ringtest) berücksichtigt werden:

a) Zur Prüfung der laborübergreifenden Konsistenz sollen in allen Laboren vorgegebene Referenz-Sortensammlungen, die ein breites Spektrum von Allelen abdecken, verwendet werden;

b) Einbeziehung von Duplikaten, Unterproben und individuellen Exemplaren einer Art zur Prüfung der Konsistenz der DNS-Profile und zur Einschätzung der Fehlerquote zwischen den Laboren;

c) Vereinbarungen zur Auswertung molekularer Daten. Die Notwendigkeit der Erstellung eines Protokolls für die Allel-/Bandenauswertung zwischen den Laboren ist abhängig vom verwendeten Markertyp (z. B. wesentlich bei SSR-Markern). Das Protokoll könnte sich mit der Auswertung folgender Daten befassen:

i. seltene Allele (d. h. diejenigen an einem spezifischen Locus, die mit einer Häufigkeit unter einem vereinbarten Schwellenwert (in der Regel 5-10 %) in einer Population) auftreten);

ii. Null-Allele (ein Allel, dessen Effekt das Fehlen eines PCR-Produkts auf molekularer Ebene ist);

iii. „schwache" Banden (d. h. Banden, bei denen die Intensität unter einen vereinbarten Schwellenwert für die Erfassung fällt, der entweder empirisch oder automatisch festgelegt wird und dessen Auswertung anfechtbar sein kann);

iv. fehlende Daten (d. h. Loci, für die aus welchem Grund auch immer für eine oder mehrere Sorten keine Daten erfasst wurden); und

v. monomorphe Banden oder nicht-informative Allel-Scorewerte (diejenigen Allele/Banden, die bei jeder analysierten Sorte auftreten, d. h. in einer bestimmten Sortensammlung nicht polymorph sind).

4. Aufbau einer artspezifischen Datenbank

Die in einer Datenbank gespeicherten Daten sowie die Art und Weise der Speicherung sollten das datenproduzierende Verfahren widerspiegeln. Daher sollte der Aufbau einer Datenbank verschiedene Stufen der Datenproduktion berücksichtigen (d. h. Rohdaten, Sequenzdaten ...). In der Datenbank sollten die Endergebnisse, z. B. das DNS-Profil, sowie die Art seiner Gewinnung, sowohl in Bezug auf die Beschreibung des Laborverfahrens als auch der Rechenschritte gespeichert sein.

4.1 Empfehlungen für die Gestaltung der Datenbank

Bei der Gestaltung der Datenbank sollten folgende Aspekte berücksichtigt werden:

a) Die Architektur der Datenbank sollte flexibel sein und z. B. sowohl flache Dateien als auch komprimierte Archivformate speichern können.

b) Für die Laborversuche, die Datenverarbeitung und die Allel-Scores sind separate Tabellen und Einträge erforderlich.

c) Speicherung von Informationen auf verschiedenen Stufen, z. B. Allel-Scorewerte und Auswertungsregeln, die einer Entscheidung zugrunde liegen, und Verknüpfungen zu den Rohdaten (tiff-Dateien, bam-Dateien).

d) Dateien zum Variantenaufruf im VCF- oder BCF-Format entsprechend der Standardversion 4.2 oder höher. Die Header-Einträge sollten Namen und Version der verschiedenen Scripte enthalten, die für Kartierung und Filterung der Sequenzabschnitte sowie für Aufruf und Filterung der Varianten verwendet werden, und zwar dergestalt, dass der Bioinformatiker die Analyse wiederholen kann.

e) Bei Wiederholungsproben, bei denen das DNS-Profil nicht übereinstimmt, muss der Eintrag gekennzeichnet oder gegebenenfalls herausgefiltert werden. Die in solchen Fällen angewandten Regeln sind in einem öffentlich zugänglichen Code Repository zu dokumentieren, das Verweise auf die Variantenaufruf-Datei enthält. Die Häufigkeiten könnten auch für heterogene Sorten verwendet werden.

f) Validierung der VCF- und/oder BCF-Daten mittels einschlägiger Spezifikationen.

g) Leicht austauschbare Daten (z. B. API).

4.2 Anforderungen an das Pflanzenmaterial

Die Art der Quelle des Materials und die Anzahl der Proben, die in der Datenbank zu speichern und auszutauschen sind, sollten geprüft werden.

4.2.1 Quelle des Pflanzenmaterials

Das zu analysierende Pflanzenmaterial sollte eine authentische, repräsentative Probe der Sorte sein und gegebenenfalls aus dem Muster der für die Prüfung im Hinblick auf die Erteilung von Züchterrechten oder auf die amtliche Eintragung verwendeten Sorte stammen. Die Verwendung dieser Muster erfordert gegebenenfalls die Genehmigung der zuständigen Behörde, des Züchters und/oder des Erhaltungszüchters. Das Pflanzenmaterial, dem die Proben entnommen werden, sollte rückverfolgbar sein, falls sich einige der Pflanzen im Nachhinein als nicht repräsentativ für die Sorte erweisen.

4.2.2 Art des Pflanzenmaterials

Die Art des Pflanzenmaterials, dem Proben zu entnehmen sind, und das Verfahren für die Entnahme von Proben des Materials für die DNS-Extraktion werden weitgehend von der betreffenden Pflanze oder Art abhängen. Bei samenvermehrten Sorten beispielsweise kann der Samen als Quelle der DNS verwendet werden, während die DNS bei vegetativ vermehrten Sorten aus dem Blattmaterial extrahiert werden kann. Welches auch immer die Quelle des Materials ist, es sollte das Verfahren für die Probenentnahme und die DNS-Extraktion dokumentiert werden. Zudem sollte überprüft werden, dass die Verfahren für die Probeentnahme und die Extraktion bei der DNS-Analyse übereinstimmende Ergebnisse zeitigen.

4.2.3 Probengröße und Art (Massen- oder Einzelproben)

Es ist wesentlich, dass die für die Analyse entnommenen Proben für die Sorte repräsentativ sind. Die Besonderheiten der Vermehrung sollten beachtet werden (vergleiche Allgemeine Einführung).

4.2.4 DNS-Referenzprobe

Es könnte eine DNS-Referenzsammlung aus dem Pflanzenmaterial, dem Proben entnommen werden, angelegt werden. Das Verfahren zur Probeentnahme sollte der empfohlenen Vorgehensweise folgen, und die DNS-Extraktion sollte einigen Qualitätskriterien genügen. Beide müssen dokumentiert werden.

Die DNS-Proben sollten so gelagert werden, dass ein Zerfall verhindert wird (z. B. durch Lagerung bei -80°C). Der Transport von DNS-Referenzproben ist in Dokument TGP/5: Abschnitt 1 beschrieben.

4.3 Verarbeitung der Sequenzdaten

Ein ausführliches Protokoll der Datenverarbeitungspipeline kann Folgendes beinhalten:

a) Art und Version der Tools;
b) die für das Tool verwendete Befehlszeile einschließlich Schwellenwerten;
c) Reproduzierbarkeitszählwerte;
d) Möglichkeit, die Daten weiterzugeben und gemeinsam zu verarbeiten;
e) Abgleich-Rohdaten (BAM oder CRAM-Dateien) sollten nach Möglichkeit gespeichert werden;
f) Es muss eine VCF-Datei pro Sorte vorhanden sein, mehrprobenbasierte VCF-Dateien sind nicht geeignet;

g) Wenn VCF-Dateien gespeichert werden, sollten alle Positionen (sowohl Varianten als auch Nicht-Varianten) sowie deren Tiefe gespeichert werden;

h) Sowohl der heuristische als auch der probabilistische Ansatz sollten erwogen und im Hinblick auf Detektionsmethoden verglichen werden;

i) Die Datenbanken sollten die Ein- und Ausgabe von Variantenaufrufdaten in standardisiertem Format (VCF oder BCF) unterstützen;

j) Die Datenverarbeitungspipeline sollte eine ausführliche Protokolldatei generieren, die zusammen mit den Variantenaufrufdaten zu speichern ist;

k) Nach Möglichkeit sollten Rohdaten gespeichert werden, so dass die Datenverarbeitung mit neuen oder aktualisierten Tools wiederholt werden kann; und

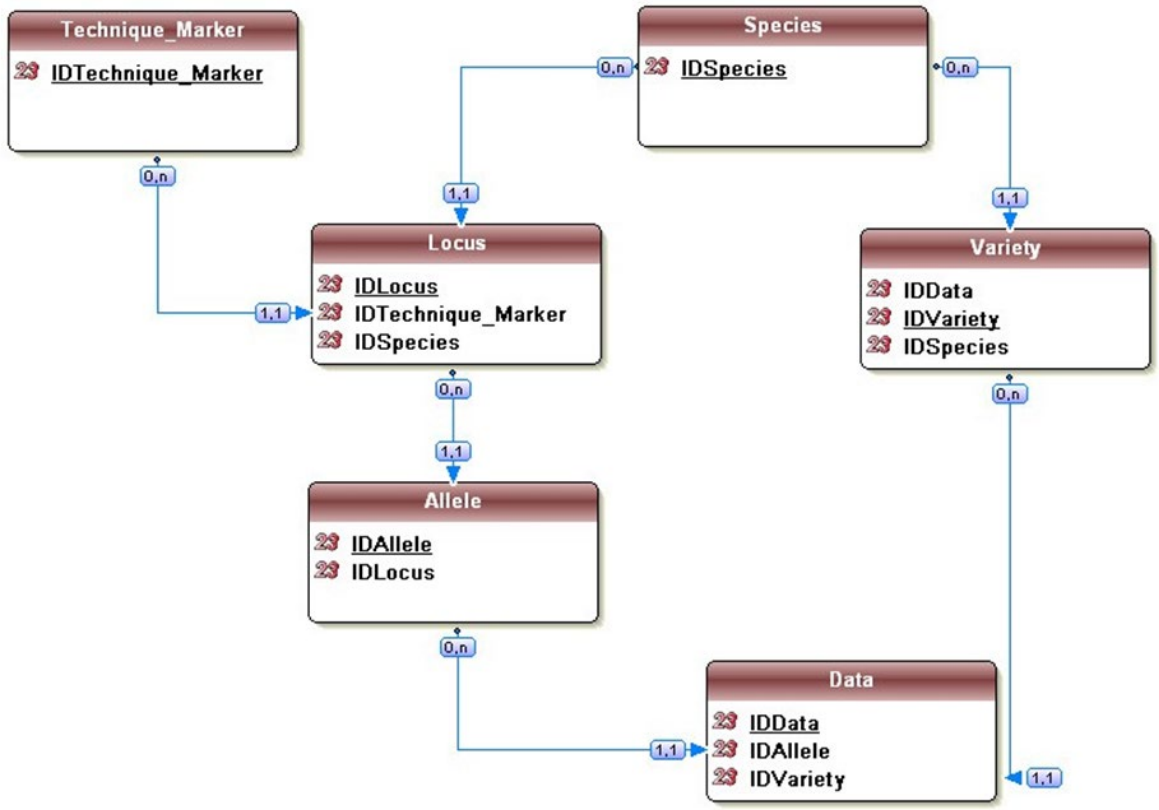
l) Es sollte in Bezug auf ein gegebenes Allel ein p-Wert oder ein Unsicherheitshinweis gespeichert werden.

4.4 Typ der Datenbank

Molekulare Daten können auf zahlreiche Arten gespeichert werden. Deshalb ist es wichtig, dass eine Datenbankstruktur entwickelt wird, die mit allen beabsichtigten Verwendungen der Daten kompatibel ist.

4.5 Datenbankmodell

Das Datenbankmodell sollte von IT-Datenbankexperten zusammen mit den Nutzern der Datenbank festgelegt werden. Das Datenbankmodell sollte mindestens sechs Kernobjekte enthalten: Art, Sorte, Marker-Detektionsmethode, Marker; Locus und Allel. Bei mittels Sequenzierungsdaten gewonnenen Varianten können die VCF-Dateien in einer relationalen oder einer Nicht-SQL-Datenbank gespeichert werden. In diesem Fall gibt jeder Datenbanksatz in Bezug auf eine Variante eine definierte Genomversion sowie Chromosom, Position und Referenzallel an.



4.6 Liste der Datenbankfelder

4.6.1 In einer Datenbank wird jedes der Objekte zu einer Tabelle, in der Felder festgelegt sind, beispielsweise:

a) Markertyp: gibt den Code oder den Namen des Verfahrens oder den Typ des verwendeten Markers an, z. B. SSR, SNP usw.

b) Position des Referenzgenoms oder Locus-Code: Vorzugsweise sollten Genom-Assembly-Version, Chromosom und Position angegeben werden, wenn ein Referenzgenom für die betreffende Art verfügbar ist, z. B. SL2.50ch05:63309763 für Tomate *Solanum lycopersicum* Assembly-Version 2.50 auf Chromosom 5 Position 63309763. Falls kein Referenzgenom verfügbar oder der Standort unbekannt ist, kann ein Name oder ein Code für den Locus für die betreffende Art benutzt werden, z. B. gwm 149, A2 usw.

c) Genotyp: Für SNP-Profile sollte die Allelzusammensetzung von SNP oder MNP angegeben werden, z. B. A/T oder A/A. Für andere Verfahren gibt der Genotyp den Namen oder den Code des Allels eines gegebenen Locus für die betreffende Art an, z. B. 1, 123 usw.

d) Alleltiefen oder Datenwert: Für SNP, die aus den Sequenzierungsdaten der nächsten Generation gewonnen werden, sollte dies die Tiefe der Verteilung für Allele angeben, z. B. 10/20 für ein A/T-Allel, bei dem A durch 10 Lesungen und T durch 20 abgedeckt ist. Andernfalls gibt dies einen Datenwert für eine gegebene Probe auf einem gegebenen Locus-Allel an, z. B. 0 (Fehlen), 1 (Vorhandensein), 0,25 (Häufigkeit) usw.

e) Sorte: Sortenbezeichnung oder Anmeldebezeichnung: Die Sorte ist das Objekt, für das die Daten erlangt wurden.

f) Typ der Sorte: z. B. Inzuchtlinie oder Hybrid

g) Art: Die Art wird durch den botanischen Namen oder den landesüblichen Namen angegeben, der sich mitunter auch auf den Sortentyp bezieht (z. B. Verwendung, Winter/Sommertyp usw.). Die Verwendung des UPOV-Codes wird empfohlen, um Probleme mit Synonymen zu lösen.

4.6.2 In jeder Tabelle müssen die Zahl der Felder, ihr Name und ihre Definition, die möglichen Werte und die zu befolgenden Regeln in der „Liste der Datenbankfelder“ festgelegt werden.

4.7 Datenzugriff / -eigentum

Es wird empfohlen, dass alle Angelegenheiten bezüglich des Eigentums der Daten und des Zugriffs zu den Daten in der Datenbank zu Beginn der Arbeit behandelt werden.

5. Datenaustausch

5.1 Szenarien für den Datenaustausch

Zu Zwecken der Zusammenarbeit sollte das Datenmodell verschiedene Arten von Szenarien ermöglichen, einschließlich des Austauschs von Daten, die aus einem standardisierten Satz von Markern für eine bestimmte Kulturpflanze erzeugt wurden (Szenario 1), und der Suche und Einsicht von Daten ausgewählter Sorten, die aus demselben standardisierten Satz von Markern erzeugt wurden (Szenario 2). Technische Details zu beiden Szenarien sind in der Anlage beschrieben: Datenaustauschszenarien und Datenübertragungsmethoden.

5.2 Verfahren für die Datenübertragung

5.2.1 Die Übertragung von Fingerprintdaten kann eine Reihe von Informationen enthalten, wie z. B. Loci, Proben, DNS, Fingerprintdaten und Fingerprintprofile. Die Art der Datenübertragung muss durch den zu übertragenden Inhalt bestimmt werden und sollte Folgendes berücksichtigen:

a) Menge der Daten
b) Komplexität der Daten
c) Anforderungen für Abfrage- oder Suchfunktionen.

Technische Details zu den Datenübertragungsverfahren sind in der Anlage beschrieben: Datenaustauschszenarien und Datenübertragungsverfahren.

5.2.2 Zu den üblicherweise verwendeten Datenformaten gehören: zip, csv, json und xml. Ihre jeweiligen Eigenschaften sind wie folgt:

1) Das Zip-Format ermöglicht eine Vielzahl von Dateninformationsdateien im Originalformat und ist aufgrund seines hohen Datenkomprimierungsgrades und der einfachen Übertragung für große und komplexe Daten geeignet.

2) Das csv-Format eignet sich besser für Dateninformationen im einfachen Datenformat, was den Vorteil hat, dass weniger ungültige Daten vorliegen und die Verarbeitungsgeschwindigkeit höher ist.

3) Die Formate json und xml können komplexere Zeichendateninformationen und mehr redundante Informationen enthalten, bieten aber beide eine gute Lesbarkeit.

6. Zusammenfassung

Nachstehend ist eine Zusammenfassung des für die hochwertige DNS-Profilierung von Sorten empfohlenen Vorgehens, einschließlich der Auswahl und Verwendung molekularer Marker und des Aufbaus gemeinsamer und nachhaltiger molekularer Datenbanken wiedergegeben (d. h. Datenbanken, die künftig aus einer Reihe von Quellen, unabhängig von der angewandten Technik, bestückt werden können).

a) Prüfung des Vorgehens nach Pflanzenart;
b) Einigung auf einen akzeptierten Markertyp und die Quelle;
c) Einigung auf zulässige Detektionsmethoden/-ausrüstungen;
d) Einigung auf die an der Prüfung zu beteiligenden Labors;
e) Einigung auf Qualitätsaspekte;
f) Überprüfung der Quelle des verwendeten Pflanzenmaterials;
g) Einigung auf die Marker, die in einer vorläufigen kollaborativen Evaluierungsphase verwendet werden sollen, in die mehr als ein Labor und verschiedene Detektionsmethoden einbezogen werden;
h) Durchführung einer Evaluierung;
i) Erstellung und Vereinbarung eines Protokolls für die Auswertung der molekularen Daten;
j) Einigung auf das Pflanzenmaterial/Referenzset, das zu analysieren ist, und auf die Quelle(n);
k) Analyse der vereinbarten Sortensammlung in verschiedenen Labors/verschiedenen Detektionsmethoden anhand von Doppelproben und Austausch von Proben/DNS-Extrakten, wenn Probleme auftreten;
l) Verwendung von Vergleichssorten (gegebenenfalls Sorten, DNS-Proben und Allelen) bei allen Analysen;
m) Überprüfung aller Stadien (einschließlich der Dateneingabe) – möglichst weitreichende Automatisierung;
n) Durchführung eines ‚Blindtests' in verschiedenen Labors anhand der Datenbank;
o) Annahme von Verfahren zur Hinzufügung neuer Daten.

C. LISTE DER AKRONYME

API	Application Programming Interface
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

[Anlage folgt]

ANLAGE

SZENARIEN FÜR DEN DATENAUSTAUSCH UND ÜBERTRAGUNGSMETHODEN

A: Szenarien für den Datenaustausch

Szenario 1: Austausch von Daten, die aus einem standardisierten Satz von Markern für eine bestimmte Pflanze erzeugt wurden

Um Daten über den für eine bestimmte Pflanze verwendeten Markersatz auszutauschen, kann folgender Webservice verwendet werden:
https://office.org/locus?upov_code={upovcode}&type={marker type}&method={observation method}

Um z. B. Marker-Set-Informationen für Mais anhand der SSR- und CE-Methode zu erhalten, sollte die folgende URL aufgerufen werden:
https://office.org/locus?upov_code=ZEAAA_MAY&type=SSR&method=CE

Das Ergebnis wäre:

{"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

Szenario 2: Suche und Ansicht von Daten ausgewählter Sorten, die aus demselben standardisierten Markersatz generiert wurden

Um molekulare Daten einer Sorte zu suchen und einzusehen, kann folgender Webdienst verwendet werden: https://office.org/variety?id={irn}&techniqueid={technique_code} vi

Zum Beispiel
https://office.org/variety?id=XU_30201800000140 &techniqueid= CN_SSR_ZEAA_MAY_CE_V_1 vi

Das Ergebnis wäre:

{"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: Verfahren für die Datenübertragung

Nachfolgend wird ein Beispiel für den Aufbau eines Fingerprintpakets in einem Zip-Format für die Datenübertragung gegeben. Dieses Verfahren muss zunächst unabhängige IDs verwenden, um Proben, DNS, Fingerprintdaten und Fingerprintatlas zu identifizieren. Danach enthält die Datendatei im json-Format alle Loci, Proben und DNS-Informationen. Alle Fingerprintdaten werden unabhängig in einer eigenen Datei im json-Format gespeichert. Die Fingerprint-ID wird an den entsprechenden Locus der Fingerprintdaten gebunden und alle Fingerprintdatenfiles und Fingerprintspektrumsdateien werden unabhängig voneinander in dem entsprechenden Verzeichnis gespeichert. Die Formatstruktur des Fingerprint-Datenpakets ist wie folgt:

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

Das Fingerprint-Paket im Zip-Format kann um weitere Informationen erweitert werden. Das Herzstück des Pakets ist das Fingerprintdatenfile, das den Kern der Korrelation darstellt, so dass die Korrelation zwischen den Teilen korrekt analysiert werden kann, was eine Datenübertragung über verschiedene Systeme hinweg ermöglicht.

[Ende der Anlage und des Dokuments]