

UPOV/INF/17/2 Draft 6**Original:** englisch**Datum:** 10. Juni 2021

zur Prüfung auf dem Schriftweg

**ENTWURF
(ÜBERARBEITUNG)**

RICHTLINIEN FÜR DIE DNS-PROFILIERUNG: AUSWAHL MOLEKULARER MARKER UND AUFBAU VON DATENBANKEN („BMT-RICHTLINIEN“)*vom Verbandsbüro erstelltes Dokument**auf dem Schriftweg zu prüfen vom**Technischen Ausschuss, Verwaltungs- und Rechtsausschuss und Rat im Jahr 2021**Haftungsausschluss: dieses Dokument gibt nicht die Grundsätze oder eine Anleitung der UPOV wieder*

Anmerkung zum Entwurf

~~Durchstreichen~~ (in Grau hervorgehoben) zeigt die Streichung von Wortlaut des Dokuments [UPOV/INF/17/1](#) an.

Unterstreichen (in Grau hervorgehoben) gibt Einfügungen in den Wortlaut von [UPOV/INF/17/1](#) an.

INHALTSVERZEICHNIS

A. EINLEITUNG	3
B. ALLGEMEINE GRUNDSÄTZE	3
1. Auswahl einer auf molekularen Markern beruhenden Methode.....	3
2. Auswahl molekularer Marker.....	4
2.1 Allgemeine Kriterien.....	4
2.2 Kriterien für spezifische Typen molekularer Marker.....	4
3. Zugang zur Technik.....	5
4. Zu analysierendes Material.....	6
4.1 Quelle des Pflanzenmaterials.....	6
4.2 Art des Pflanzenmaterials.....	6
4.3 Probengröße.....	6
4.4 DNS-Referenzprobe.....	6
5. Normung der Analyseprotokolle.....	7
5.1 Einleitung.....	7
5.2 Qualitätskriterien.....	7
5.3 Evaluierungsphase.....	7
5.4 Auswertung der molekularen Daten.....	8
6. Datenbanken.....	9
6.1 Typ der Datenbank.....	9
6.2 Datenbankmodell.....	9
6.3 Liste der Datenbankfelder.....	10
6.4 Beziehung zwischen den Tabellen.....	10
6.5 Datentransfer in die Datenbank.....	12
6.6 Datenzugriff/-eigentum.....	12
6.7 Datenanalyse.....	12
6.8 Validierung der Datenbank.....	12
7. Zusammenfassung.....	13
GLOSSAR	14
Mikrosatelliten oder einfache Sequenzwiederholungen (<i>Simple Sequence Repeats, SSR</i>).....	14
Einzel-Nukleotid-Polymorphismen (<i>Single Nucleotide Polymorphisms, SNP</i>).....	14
Cleaved Amplified Polymorphic Sequences (<i>CAPS</i>).....	14
Sequenzcharakterisierte amplifizierte Regionen (<i>Sequence-Characterized Amplified Regions (SCAR)</i>).....	15
Primerverlängerung („Pig tailing“).....	15
Nullallel.....	15
„Stotter“-Banden.....	15

A.	EINLEITUNG	3
B.	ALLGEMEINE GRUNDSÄTZE	3
1.	Auswahl der molekularen Marker	4
1.1	Sortensätze für das Auswahlverfahren	4
1.2	Molekulare Marker – Leistungsaspekte	5
2.	Auswahl der Detektionsmethode	6
2.1	DNS-Profilierungsverfahren – allgemeine Überlegungen	6
2.2	Zugang zur Technik	7
3.	Validierung und Harmonisierung von Markersatz und Detektionsmethode	7
3.1	Validierung und Harmonisierung – allgemeine Überlegungen	7
3.2	Leistungsaspekte – Validierung von Markern und Verfahren	7
3.3	Konsistenzaspekte	7
4.	Aufbau einer artspezifischen Datenbank	8
4.1	Empfehlungen für die Gestaltung der Datenbank	8
4.2	Anforderungen an das Pflanzenmaterial	8
4.3	Verarbeitung der Sequenzdaten	11
4.4	Typ der Datenbank	11
4.5	Datenbankmodell	11
4.6	Liste der Datenbankfelder	12
4.7	Datenzugriff / -eigentum	13
5.	Datenaustausch	14
5.1	Szenarien für den Datenaustausch	14
5.2	Verfahren für die Datenübertragung	14
6.	Zusammenfassung	14
C.	LISTE DER AKRONYME	17
<u>ANLAGE SZENARIEN FÜR DEN DATENAUSTAUSCH UND ÜBERTRAGUNGSMETHODEN</u>		

A. EINLEITUNG

Dieses Dokument (BMT-Richtlinien) soll Anleitung zur Entwicklung von harmonisierten Methoden-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 der Marker 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 einer auf molekularen Markern beruhenden Methode

1.1 Wichtige Kriterien für die Auswahl einer Methode sind:

- a) Reproduzierbarkeit der Datengenerierung zwischen Labors und Detektionsmethoden (verschiedene Ausrüstungstypen);
- b) Wiederholbarkeit im Zeitablauf;
- c) Unterscheidungskraft;
- d) Möglichkeiten für die Entwicklung von Datenbanken, und
- e) Zugänglichkeit der Methode

1.2 Da verbesserte Techniken und neue Ausrüstungen verfügbar werden, ist es für die weitere Nachhaltigkeit der Datenbanken wichtig, daß die Interpretation der erzeugten Daten unabhängig von den für ihre Generierung genutzten Ausrüstungen ist. Dies ist beispielsweise der Fall bei DNS-Sequenzierungsdaten. Anfänglich wurden radioaktiv markierte Primer und Sequenzierungsgele für die Generierung dieser Daten verwendet, während dies heute mit fluoreszierenden Farbstoffen und anschließender Auftrennung mit weitgehend automatisierten Kapillargel-Elektrophoresesystemen mit hoher Durchsatzleistung erfolgt.

1.3 Trotz dieser Unterschiede sind die mit den verschiedenen Verfahren erzeugten Daten übereinstimmend und unabhängig von den für ihre Generierung genutzten Verfahren. Dies kann auch für Daten gelten, die beispielsweise anhand von DNS-Mikrosatelliten (einfache Sequenzwiederholungen (*Simple Sequence Repeats*, SSR)) oder Polymorphismen in einem einzelnen Nukleotid (*Single Nucleotide Polymorphisms*, SNP)) erzeugt werden. Diese Wiederholbarkeit und Reproduzierbarkeit sind wichtig beim Aufbau und Betrieb und für die Lebensdauer der Datenbanken und äußerst wichtig bei der Einrichtung einer zentral geführten Datenbank, die mit verifizierten Daten aus verschiedenen Quellen bestückt ist.

1.4 Die molekularen Verfahren, die für die Sortenprofilierung problemlos anwendbar sind, werden durch die Anforderung eingeschränkt, daß die Daten wiederholbar, reproduzierbar und übereinstimmend sein müssen. Obwohl verschiedene Multilocus-DNS-Profilierungsverfahren erfolgreich für die Forschung genutzt wurden, kann bei vielen davon Kodominanz nicht problemlos erfaßt werden, und die Reproduzierbarkeit von komplexeren Bandenmustern zwischen Labors, die verschiedene Ausrüstungen benutzen, kann problematisch sein.

1.5 Diese Faktoren verursachen im Kontext der Sortenprofilierung Schwierigkeiten. Infolgedessen befasst sich dieses Dokument insbesondere mit Überlegungen und Empfehlungen zu angemessen definierten und erforschten Anwendungen von SSR (Mikrosatelliten) und, im Hinblick auf die Zukunft, zu Sequenzierungsinformationen (d. h. *Single Nucleotide Polymorphisms*, SNP). Weitere Verfahren, die sich auf DNS-Sequenzinformationen stützen, wie *Cleaved Amplified Polymorphic Sequences* (CAPS) und *Sequence Characterized Amplified Regions* (SCAR) können die obigen Kriterien ebenfalls erfüllen, doch ist ihre Anwendung bei der DNS-Profilierung von Pflanzensorten noch nicht untersucht.

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:

a) 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;

b) 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 molekularer Marker

2.1 Allgemeine Kriterien

Folgende allgemeine Kriterien für die Wahl eines spezifischen Markers oder Markersatzes sollen für molekulare Marker, ungeachtet der Verwendung der Marker, geeignet sein, obwohl eingeräumt wird, daß spezifische Anwendungen bestimmte zusätzliche Kriterien erfordern könnten:

- a) nutzbares Polymorphiegrad;
- b) Wiederholbarkeit innerhalb und Reproduzierbarkeit zwischen Labors in Bezug auf die Auswertung der Daten;
- c) bekannte Verteilung der Marker im ganzen Genom (d. h. Kartenposition), was zwar nicht entscheidend, jedoch eine wertvolle Information ist und hilft, die Auswahl von gekoppelten Markern zu vermeiden, und
- d) nach Möglichkeit Vermeiden von Markern mit "Nullallelen" (d. h. ein Allel, dessen Wirkung das Fehlen eines PCR-Produkts auf Molekularebene ist), was ebenfalls nicht entscheidend, jedoch ratsam ist.

2.2 Kriterien für spezifische Typen molekularer Marker

2.2.1 Mikrosatellitenmarker

2.2.1.1 Die Analyse der einfachen Sequenzwiederholungen (SSR oder Mikrosatelliten: siehe Glossar) mittels Polymerase-Kettenreaktion (PCR) wird heute allgemein genutzt und hat verschiedene Vorteile:

2.2.1.2 SSR-Marker sind kodominant ausgeprägt, in der Regel einfach zu erfassen und können problemlos kartiert werden. Sie wurden in verschiedenen Labors verwendet und analysiert und können unter spezifischen experimentellen Bedingungen robust und wiederholbar sein. Zudem können sie mit automatisierten nicht-radioaktiven DNS-Sequenzierern mit hoher Durchsatzleistung mittels der Gel-Elektrophorese oder der Kapillar-Elektrophorese analysiert werden, und mehrere können gleichzeitig analysiert werden (Multiplexing).

2.2.1.3 Für eine effektive Mikrosatellitenanalyse ist die Auswahl hochqualitativer Marker von entscheidender Bedeutung. Hierzu ist u. a. zu beachten:

- a) Grad des "Stotterns" (Generierung einer Serie von einer oder mehreren Banden, die sich in der Größe durch eine Wiederholungseinheit unterscheiden);
- b) (n+1)-Peaks; die Taq-Polymerase fügt häufig 1 bp zum Ende eines Fragments hinzu. Dies kann durch Verwendung von verlängerten ("pigtailed") Primern verhindert werden (siehe Glossar);
- c) Größe des Amplifikationsprodukts;
- d) effektive Trennung zwischen den verschiedenen Allelen mit geeigneten Detektionsmethoden;
- e) zuverlässige und reproduzierbare Auswertung der Allele mit verschiedenen Detektionsmethoden;
- f) Polymorphiegrad zwischen Sorten (es ist zu beachten, daß dies die Untersuchung einer erheblichen Anzahl Sorten voraussetzt);
- g) Vermeiden von Kopplung.

2.2.1.4 Für die Auswertung von SSR in verschiedenen Labors und die Nutzung verschiedener Detektionssysteme ist es entscheidend, daß Referenzallele (d. h. Sortenserien) festgelegt und in alle Analysen einbezogen werden. Diese Referenzallele sind notwendig, weil sich die Molekulargewichtsstandards in den verschiedenen derzeit verfügbaren Detektionsmethoden unterschiedlich verhalten und daher für die Allelidentifikation nicht geeignet sind.

2.2.1.5 Die in einem bestimmten Labor verwendeten Primer sollten durch einen gesicherten Lieferanten synthetisiert werden, um die Möglichkeit verschiedener DNS-Profile infolge der Verwendung von Primern, die durch verschiedene Quellen synthetisiert wurden, zu reduzieren.

2.2.2 Einzel-Nukleotid-Polymorphismen (Single Nucleotide Polymorphism, SNP)

Einzel-Nukleotid-Polymorphismen (Single Nucleotide Polymorphisms, SNP: siehe Glossar) können durch DNS-Sequenzierung nachgewiesen werden. Dies ist ein Routineverfahren, das im allgemeinen eine sehr hohe Wiederholbarkeit im Zeitablauf sowie Reproduzierbarkeit zwischen den Labors zeigt. Der Nachweis spezifischer SNP kann jedoch anhand einer Reihe von Verfahren vorgenommen werden, von denen viele noch nicht Routine sind. Die SNP haben bei diploiden Pflanzen naturgemäß lediglich zwei allelische Stufen, während diese bei Polyploiden variieren können, bei denen es Dosiereffekte geben wird. Der einfache Aufbau der SNP macht die Auswertung der SNP verhältnismäßig unkompliziert und zuverlässig. Das bedeutet auch, daß möglicherweise zahlreiche Marker entweder einzeln oder in Multiplexen analysiert werden müssen, um die effiziente und effektive Profilierung eines bestimmten Genotyps zu ermöglichen.

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.

3-2.2 Zugang zur Technik

Einzelne molekulare Marker und Materialien sind öffentlich verfügbar. Es dürften jedoch hohe Investitionen erforderlich sein, um beispielsweise hochqualitative SSR-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 Anleitung sollte für diese Richtlinien 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 Variantenaufzuruf im VCF- oder BCF-Format entsprechend der Standardversion 4.2 oder höher. Die Header-Einträge sollten Namen und Version der verschiedenen Skripte 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 Variantenaufzuruf-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 ~~Zu analysierendes Material~~ Anforderungen an das Pflanzenmaterial

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

4.2.1 Quelle des Pflanzenmaterials

Das zu analysierende Pflanzenmaterial sollte eine authentische, repräsentative Probe der Sorte sein und ~~nach Möglichkeit 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 ~~von diesen Mustern des für die Prüfung im Hinblick auf die Erteilung von Züchterrechten oder auf die amtliche Eintragung eingereichten Materials~~ 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 ~~genormt und~~ 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. ~~Was die Repräsentativität für die Sorte betrifft, sollten d~~Die Besonderheiten der Vermehrung ~~sollten~~ beachtet werden (vergleiche Allgemeine Einführung). ~~Die Probengröße sollte unter Berücksichtigung geeigneter statistischer Verfahren bestimmt werden.~~

4.2.4 DNS-Referenzprobe

~~Es wird empfohlen, gemäß den Abschnitten 4.1 bis 4.3 e~~Es könnte eine DNS-Referenzprobensammlung aus dem Pflanzenmaterial ~~anzulegen, dem Proben entnommen werden, angelegt werden. Dies hat den Vorteil, daß die DNS-Referenzproben gelagert und anderen Labors zur Verfügung gestellt werden können. 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.~~

5. ~~Normung der Analyseprotokolle~~

5.1 ~~Einleitung~~

~~Dieses Dokument soll nicht detaillierte technische Protokolle für die Erstellung von DNS Profilen von Sorten bereitstellen. Grundsätzlich kann jede geeignete Analysemethodik angewandt werden, doch ist es wichtig, daß die Methodik angemessen validiert wird. Dies kann über ein international anerkanntes Validierungsverfahren oder durch Entwicklung eines methodenspezifischen Vorgehens erfolgen. In beiden Fällen gibt es einige zweckdienliche allgemeine Überlegungen.~~

~~Jedes für die Bestimmung eines Genotyps und den Aufbau von Datenbanken angewandte Verfahren sollte aus technischer Sicht einfach durchzuführen, zuverlässig und robust sein und eine problemlose und eindeutige Auswertung der Markerprofile in verschiedenen Labors ermöglichen. Dies setzt eine Standardisierung voraus, beispielsweise bei der Auswahl der Marker, der Referenzallele und der Allelbenennung/auswertung.~~

5.2 ~~Qualitätskriterien~~

5.2.1 ~~Es ist wichtig, bestimmte Qualitätskriterien zu berücksichtigen, die beispielsweise folgendes betreffen:~~

- a) ~~die Qualität der DNS;~~
- b) ~~die Methoden für die DNS-Extraktion~~
- c) ~~die Primersequenzen;~~
- d) ~~die bei PCR-basierten Methoden zu verwendende Polymerase~~
- e) ~~für PCR-basierte Methoden die Menge/Konzentration jeder PCR-Komponente und anderer Komponenten, und~~
- f) ~~die PCR-Durchlaufbedingungen.~~

~~5.2.2 Die detaillierte Methode sollte in einem Protokoll erläutert werden.~~

~~5.3 Evaluierungsphase~~

~~5.3.1 Einleitung~~

~~Zur Auswahl geeigneter Marker und Erstellung ausreichender Laborprotokolle für eine gegebene Art wird eine vorläufige Evaluierungsphase empfohlen, an der mehr als ein Labor beteiligt ist (d. h. ein international anerkanntes Validierungsverfahren, z. B. ein Ringtest gemäß international vereinbarter Normen). Diese Phase sollte sich hauptsächlich mit der Auswahl eines Markersets befassen und wird in der Regel die Evaluierung bestehender Marker umfassen, die entweder veröffentlicht wurden oder über andere Quellen verfügbar sind. Die Zahl der zu bewertenden Marker wird variieren und hängt von den Möglichkeiten der jeweiligen Art ab. Die Marker sollten aus zuverlässigen Quellen (z. B. „peer-reviewed“ Veröffentlichungen) und von gesicherten Lieferanten stammen. Die endgültige Wahl einer zu evaluierenden Anzahl wird am Ende des Prozesses ein Gleichgewicht zwischen Kosten und der Anforderung sein, ein zufriedenstellendes Set vereinbarter Marker zu erzeugen. Ziel ist es, ein vereinbartes Markersets zu erzeugen, das in verschiedenen Labors, die potentiell unterschiedliche Arten von Ausrüstungen und verschiedene Quellen chemischer Reagenzien usw. verwenden, zuverlässig und reproduzierbar analysiert, ausgewertet und erfaßt werden~~

~~5.3.2 Wahl der Sorte~~

~~Als Grundlage für die Evaluierungsphase sollte eine angemessene Anzahl Sorten aufgrund der genetischen Variabilität innerhalb der Arten und des Typs der betreffenden Sorten ausgewählt werden. Die Wahl der Sorten sollte eine angemessene Diversitätsspanne reflektieren und nach Möglichkeit einige verwandte und einige morphologisch ähnliche Sorten einschließen, um die Beurteilung des Unterscheidungsniveaus in diesen Fällen zu ermöglichen.~~

~~5.3.3 Interpretation der Ergebnisse~~

~~Die folgende Evaluierungsphase sollte nach Möglichkeit ein international anerkanntes Validierungsverfahren einschließen, damit die gesamte Methode objektiv beurteilt werden kann. Marker, die in einem der an dieser Evaluierungsphase beteiligten Labors Schwierigkeiten verursachen, sollten für die spätere Verwendung zurückgewiesen werden. Da sich die meisten Fehler bei der Analyse großer Sortensammlungen aus Auswertungsfehlern zu ergeben scheinen, sollte der Aufbau von Datenbanken auf Doppelproben gestützt werden (z. B. verschiedene Teilproben von Samen derselben Sorte), die in mehr als einem Labor analysiert werden. Da die Teilproben (oder DNS-Extrakte aus diesen) im Falle einer Diskrepanz ausgetauscht werden können, ist dieses Verfahren bei der Feststellung von Fehlern bei der Probenentnahme oder von Fehlern aufgrund der Heterogenität innerhalb der Proben äußerst wirksam und eliminiert etwaige Laborartefakte.~~

~~5.4 Auswertung der molekularen Daten~~

~~In Verbindung mit der Evaluierungsphase sollte ein Protokoll für die Allel-/Bandenauswertung erstellt werden. Das Protokoll sollte sich mit der Art und Weise der Auswertung folgender Daten befassen:~~

- ~~a) seltene Allele (d. h. diejenigen an einem spezifische Locus, die mit einer Häufigkeit unter einem vereinbarten Schwellenwert (in der Regel 5-10 %) in einer Population) auftreten;~~
- ~~b) Nullallele (ein Allel, dessen Wirkung das Fehlen eines PCR-Produkts auf Molekularebene ist);~~
- ~~c) „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);~~
- ~~d) fehlende Daten (d. h. Loci, für die aus welchem Grund auch immer für eine Sorte oder Sorten keine Daten erfaßt wurden);~~
- ~~e) monomorphe Banden (diejenigen Allele/Banden, die bei jeder analysierten Sorte auftreten, d. h. in einer bestimmten Sortensammlung nicht polymorph sind).~~

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.

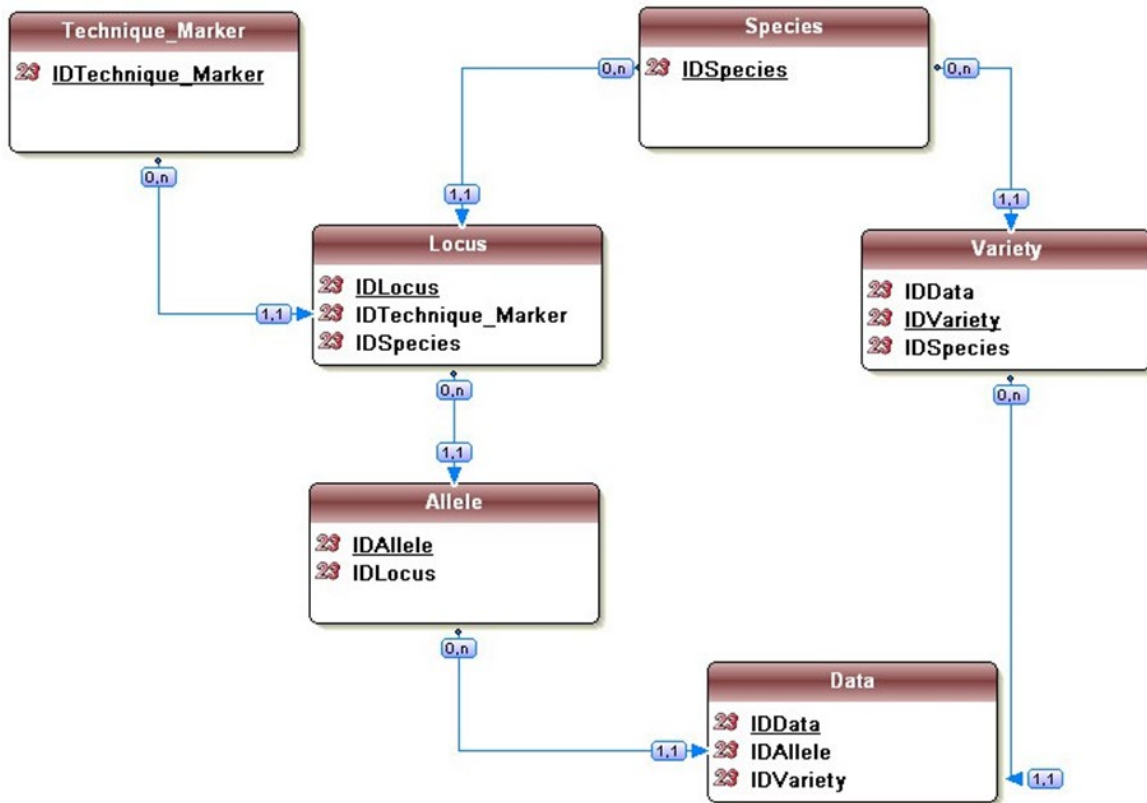
6. Datenbanken

6-14.4 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.

6-24.5 Datenbankmodell

Das Datenbankmodell sollte von IT-Datenbankexperten zusammen mit den Nutzern der Datenbank festgelegt werden. Das Datenbankmodell sollte mindestens sechs Kernobjekte enthalten: Art, Sorte, Sorte, Technik, und Allel-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.



6.34.6 Liste der Datenbankfelder

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

- Verfahrens-/Markertypcode:** gibt den Code oder den Namen des Verfahrens oder den Typ des verwendeten Markers an, z. B. SSR, SNP usw.
- Position des Referenzgenoms oder Locus-Code:** gibt an, 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.
- Allel-Code:** zeigt an 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.
- Datenwert:** 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.
- Sorte:** **Sortenbezeichnung oder Anmeldebezeichnung:** Die Sorte ist das Objekt, für das die Daten erlangt wurden.
- Typ der Sorte:** z. B. Inzuchtlinie oder Hybrid
- 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 würde empfohlen, um die Probleme mit Synonymen zu lösen und wäre daher für die Koordinierung von Vorteil.

4.6.3.2 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.

6.4 — Beziehung zwischen den Tabellen

6.4.1 — Die Verknüpfungen zwischen den Tabellen sind ein wichtiger Aspekt der Datenbankkonstruktion. Die Verknüpfungen zwischen den Tabellen lassen sich wie folgt veranschaulichen:

Tabelle	Verknüpfung	Tabelle	Beschreibung
Frau	0 — oder 1 bis n (0, n)	Kind	0: Eine Frau hat möglicherweise kein Kind 1 bis n: Eine Frau kann 1 bis n Kinder haben (sie ist dann eine Mutter)
Kind	1 bis 1 (1,1)	Frau	Ein bestimmtes Kind hat nur eine biologische Mutter

6.4.2 — Die nachstehende Tabelle gibt die Beziehung zwischen den sechs minimalen Kernobjekten an, wie im Datenbankmodell in Abschnitt 6.2 vorgeschlagen:

Tabelle	Verknüpfung	Tabelle	Beschreibung
Verfahren/Marker	0 — oder 1 bis n	Locus	0: Ein Verfahren/Marker kann in 'Verfahren/Marker' vorhanden sein, selbst wenn in der Datenbank noch kein Locus/Allel verwendet wird 1 bis n: Ein vorgegebener Markertyp kann 1 bis n nützliche Loci ergeben
Locus	1 bis 1	Verfahren/Marker	Ein gegebener Locus wird im Rahmen eines gegebenen Verfahrens/Markers definiert
Locus	1 bis n	Allel	Für jeden Locus kann 1 oder mehr als 1 Allel beschrieben werden
Allel	1 bis 1	Locus	Ein gegebenes Allel wird im Rahmen eines gegebenen Locus definiert
Allel	0 — oder 1 bis n	Daten	0: Ein gegebenes Allel kann definiert werden, jedoch ohne Daten 1 bis n: Ein gegebenes Allel kann in 1 bis n Daten gefunden werden
Daten	1 bis 1	Allel	Die Daten entsprechen einem gegebenen Allel
Sorte	0 — oder 1 bis n	Daten	0: Die Sorte hat keine Daten 1 bis n: Die Sorte hat Daten
Daten	1 bis 1	Sorte	Die Daten entsprechen einer gegebenen Sorte
Daten	1 bis 1	Art	Die Daten werden für eine gegebene Sorte und dann für die Art der Sorte erlangt
Art	0 — oder 1 bis n	Daten	0: Eine Art hat möglicherweise keine Daten 1 bis n: Eine Art kann 1 bis n Daten haben.

6.5 — Datentransfer in die Datenbank

Zur Reduzierung der Fehlerzahl beim Datentransfer und der Transkription ist es ratsam, den Datentransfer in die Datenbanken möglichst weitgehend zu automatisieren.

6.6.4.7 — 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.

6.7.5. Datenaustausch

Das Analyseverfahren wird durch den Zweck bestimmt, zu dem die Daten analysiert werden. In diesen Richtlinien werden daher keine ausdrücklichen Empfehlungen abgegeben.

6.8 Validierung der Datenbank

Nach Abschluß der ersten Phase der Datenbank wird empfohlen, einen 'Blindtest' durchzuführen, d. h. eine Reihe von Proben an verschiedene Labors zu schicken und diese zu ersuchen, das vereinbarte Protokoll zusammen mit der Datenbank anzuwenden, um die Proben zu identifizieren.

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 Fingerprintsdaten kann eine Reihe von Informationen enthalten, wie z. B. Loci, Proben, DNS, Fingerprintsdaten und Fingerprintsprofile. 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.

7.6 Zusammenfassung

Nachstehend ist eine Zusammenfassung des für die hochwertige DNS-Profilierung von Sorten empfohlenen Vorgehens, einschließlich der Auswahl und der Verwendung molekularer Marker im Hinblick auf den Aufbau zentraler, gemeinsamer und nachhaltiger molekularer Datenbanken für DNS-Profile von Sortenwiedergegeben (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 (vergleiche Abschnitt 5.2);
- f) Überprüfung der Quelle des verwendeten Pflanzenmaterials (vergleiche Abschnitt 4);
- 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 (vergleiche Abschnitt 2);
- h) Durchführung einer Evaluierung (vergleiche Abschnitt 5.3);

- i) Erstellung und Vereinbarung eines Protokolls für die Auswertung der molekularen Daten (~~vergleiche Abschnitt 5.4~~);
- 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 davon Verfahren zur Hinzufügung neuer Daten.

GLOSSAR

Mikrosatelliten oder einfache Sequenzwiederholungen (Simple Sequence Repeats, SSR)

Mikrosatelliten oder einfache Sequenzwiederholungen (SSR) sind doppelt wiederholte DNS-Sequenzen, in der Regel mit einer Wiederholungseinheit von 2 bis 8 Basenpaaren (z. B. GA, CTT und GATA). Bei vielen Arten wurde nachgewiesen, daß für einige Mikrosatelliten mehrere Allele vorhanden sind, die sich aus Unterschieden in der Wiederholungszahl dieser Wiederholungseinheit ergeben. Mikrosatelliten können mit PCR anhand spezifischer Primer analysiert werden. Dieses Verfahren wird als Mikrosatelliten mit sequenzmarkierten Loci (*Sequence-tagged-site Microsatellites*, STMS) bezeichnet. Die Allele (PCR-Produkte) können dann mittels Agarosegel- oder Polyacrylamidgel-Elektrophorese aufgetrennt werden. Zur Entwicklung von Mikrosatelliten mit sequenzmarkierten Loci werden Informationen über die Sequenz der DNS benötigt, die den Mikrosatelliten seitlich begrenzt. Diese Informationen können mitunter aus bestehenden Datenbanken für DNS-Sequenzen beschafft werden. Andernfalls müssen sie empirisch ermittelt werden.

Einzel Nukleotid Polymorphismen (Single Nucleotide Polymorphisms, SNP)

Polymorphismen eines einzelnen Nukleotids (SNP) (ausgesprochen wie "snips") sind Sequenzvariationen, die auftreten, wenn ein einzelnes Nukleotid (A, T, C oder G) in der Genomsequenz verändert ist. Ein SNP könnte beispielsweise die DNS-Sequenz AAGGCTAA in ATGGCTAA ändern. Damit eine Variation als SNP angesehen wird, muß sie im allgemeinen in mindestens 1 % der Population auftreten. Die potentielle Zahl der SNP-Marker ist sehr hoch, was bedeutet, daß es möglich sein sollte, sie in allen Teilen des Genoms nachzuweisen. SNP können sowohl in kodierenden (Gen-) Regionen als auch in nichtkodierenden Regionen des Genoms auftreten. Der Nachweis der SNP beinhaltet eine vergleichende Sequenzierung einer Anzahl Individuen aus einer Population. Potentielle SNP werden in der Regel eher aus verfügbaren Sequenz-Datenbanken durch Vergleich aneinander ausgerichteter Sequenzen identifiziert. Sie können zwar durch verhältnismäßig einfache PCR und Gel-Elektrophorese nachgewiesen werden, doch werden zur Zeit Verfahren mit hohem Datendurchlauf und Mikroanordnungen für die gleichzeitige automatische Auswertung Hunderte von SNP-Loci entwickelt.

Gespaltene Amplifizierte Polymorphe Sequenzen (CAPS)

Cleaved Amplified Polymorphic Sequences (CAPS) sind DNS-Fragmente, die durch PCR anhand spezifischer Primer mit 20-25 bp amplifiziert werden, gefolgt von Verdau mit einer Restriktions-Endonuklease. Dann werden die Längen-Polymorphismen, die sich aus der Variation beim Auftreten an den Restriktions-Loci ergeben, durch Gel-Elektrophorese der extrahierten Produkte identifiziert. Im Vergleich zu Markern wie RFLP sind die Polymorphismen wegen der geringen Größe der amplifizierten Fragmente (300-1800 bp) schwieriger zu identifizieren. Die CAPS-Analyse erfordert jedoch keinen Southern-blot-Hybridisierung und keinen radioaktiven Nachweis. CAPS wurden bisher in der Regel vorwiegend für Genkartierungsstudien angewandt.

Sequenzcharakterisierte amplifizierte Regionen (Sequence-Characterized Amplified Regions, SCAR)

Sequenzcharakterisierte amplifizierte Regionen (*Sequence-Characterized Amplified Regions* (SCAR)) sind DNS-Fragmente, die durch PCR anhand spezifischer Primer mit 15-30 bp amplifiziert werden, die aus zuvor identifizierten polymorphen Sequenzen erzeugt wurden. Durch die Verwendung längerer PCR-Primer umgehen die SCAR-Marker das Problem der geringen Reproduzierbarkeit. Sie sind in der Regel auch kodominante Marker. SCAR sind lokusspezifisch und wurden für Genkartierungsstudien und die markerunterstützte Selektion angewandt.

Primerverlängerung ("Pig-tailing")

Bei der SSR-Analyse ist das "pig-tailing" die Addition einer kurzen spezifischen Oligonukleotid-Sequenz zu den beim PCR-Verfahren verwendeten Primern als Mittel zur Verbesserung der Klarheit der Amplifikationsprodukte und zur Reduzierung der Artefakte.

Nullallel

Bei der SSR-Analyse ist ein "Nullallel" ein Allel an einem bestimmten Locus, dessen Wirkung das Fehlen eines PCR-Produkts ist.

„Stotter“-Banden

Bei der SSR-Analyse bedeuten „Stotterbanden“ das Auftreten einer Serie von einer oder mehreren Banden, die sich nach der PCR in der Größe durch 1 Wiederholungseinheit unterscheiden.
Mikrosatelliten oder einfache Sequenzwiederholungen (Simple Sequence Repeats, SSR).

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:

<pre> {"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", </pre>	<pre> "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", </pre>	<pre> ["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 </pre>
---	--	---

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={im}&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": "xxxxxxxxxxxxx"
  "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.