

Comité Técnico

TC/56/13

**Quincuagésima sexta sesión
Ginebra, 26 y 27 de octubre de 2020**

**Original: Inglés
Fecha: 9 de octubre de 2020**

REVISIÓN DEL DOCUMENTO UPOV/INF/17 “DIRECTRICES PARA LOS PERFILES DE ADN: SELECCIÓN DE MARCADORES MOLECULARES Y CREACIÓN DE UNA BASE DE DATOS (“DIRECTRICES BMT”)”

Documento preparado por la Oficina de la Unión

Descargo de responsabilidad: el presente documento no constituye un documento de política u orientación de la UPOV

RESUMEN

1. El presente documento tiene por objeto examinar la propuesta de revisión del documento UPOV/INF/17 “Directrices para los perfiles de ADN: selección de marcadores moleculares y creación de una base de datos (“Directrices BMT”)” expuesta en el documento UPOV/INF/17/2 Draft 4 y en el Anexo del presente documento.

2. Se invita al TC a:

a) examinar la propuesta de revisión del documento UPOV/INF/17 sobre la base del documento UPOV/INF/17/2 Draft 4;

b) solicitar a los TWP que examinen un proyecto de revisión del documento UPOV/INF/17/1 (documento UPOV/INF/17/2 Draft 5) en sus sesiones de 2021; y

c) tomar nota de que un proyecto de revisión del documento UPOV/INF/17 (UPOV/INF/17/2 Draft 6) se someterá a la aprobación del Consejo en su quincuagésima quinta sesión, que se celebrará el 29 de octubre de 2021, previo acuerdo del TC en su quincuagésima séptima sesión y del CAJ en su septuagésima octava sesión, previstas en 2021.

3. El presente documento se estructura del modo siguiente:

RESUMEN	1
ANTECEDENTES.....	2
NOVEDADES ACONTECIDAS EN LA DECIMONOVENA SESIÓN DEL BMT	2
PRÓXIMOS PASOS	2

ANEXO: PROPUESTA DEL BMT PARA LA REVISIÓN DEL DOCUMENTO UPOV/INF/17

4. En el presente documento se utilizan las abreviaturas siguientes:

- BMT: Grupo de Trabajo sobre Técnicas Bioquímicas y Moleculares, y Perfiles de ADN en particular
- CAJ: Comité Administrativo y Jurídico
- TC: Comité Técnico
- TWA: Grupo de Trabajo Técnico sobre Plantas Agrícolas
- TWC: Grupo de Trabajo Técnico sobre Automatización y Programas Informáticos
- TWF: Grupo de Trabajo Técnico sobre Plantas Frutales
- TWO: Grupo de Trabajo Técnico sobre Plantas Ornamentales y Árboles Forestales
- TWP: Grupo(s) de Trabajo Técnico
- TWW: Grupo de Trabajo Técnico sobre Hortalizas

ANTECEDENTES

5. Los antecedentes de esta cuestión figuran en el documento TC/55/7 “Técnicas moleculares”.
6. En su quincuagésima quinta sesión,¹ el TC suscribió la propuesta, formulada por el BMT en su decimoctava sesión,² de que Francia, los Países Bajos y la Unión Europea elaboren un nuevo proyecto del documento UPOV/INF/17 (documento UPOV/INF/17/2 Draft 3) para que se examine en la decimonovena sesión de este Grupo de Trabajo (véase el párrafo 181 del documento TC/55/25 “Informe”).
7. En sus sesiones de 2020, los TWP examinaron el documento TWP/4/7 “*Molecular Techniques*” (Técnicas moleculares). Los TWP tomaron nota de las modificaciones del documento UPOV/INF/17 acordadas por el BMT, que figuran en el Anexo II del documento TWP/4/7. Asimismo, tomaron nota de que el TC había convenido en invitar a Francia, los Países Bajos y la Unión Europea a que elaboren un nuevo proyecto del documento UPOV/INF/17 para que sea examinado en la decimonovena sesión del BMT (véanse los párrafos 13 y 14 del documento TWV/54/12 “*Report*” (Informe), los párrafos 84 y 85 del documento TWO/52/11 “*Report*”, los párrafos 58 y 59 del documento TWA/49/7 “*Report*”, los párrafos 13 y 14 del documento TWF/51/10 “*Report*” y los párrafos 66 y 67 del documento TWC/38/11 “*Report*”).
8. En el programa para la elaboración de orientaciones y material de información pertinente se prevé someter un nuevo proyecto del documento UPOV/INF/17 a la aprobación del Consejo en su quincuagésima quinta sesión, prevista en 2021, según consta en el documento TC/56/14 “Elaboración de orientaciones y material de información”.

NOVEDADES ACONTECIDAS EN LA DECIMONOVENA SESIÓN DEL BMT

9. En su decimonovena sesión,³ el BMT examinó los documentos BMT/19/3 Rev. “*Revision of document INF/17*” (Revisión del documento INF/17) y UPOV/INF/17/2 Draft 3 (véanse los párrafos 11 y 12 del documento BMT/19/15 “*Report*”).
10. El BMT convino en que el proyecto de orientación que se reproduce en el Anexo del presente documento se proponga al Comité Técnico como punto de partida de una futura revisión del documento UPOV/INF/17.

PRÓXIMOS PASOS

11. El proyecto de orientación acordado por el BMT, que se reproduce en el Anexo del presente documento, se recoge en el documento UPOV/INF/17/2 Draft 4, con indicación de las modificaciones respecto del texto del documento UPOV/INF/17/1. Se propone al TC que examine el documento UPOV/INF/17/2 Draft 4 como punto de partida de la revisión del documento UPOV/INF/17.
12. Se invita al TC a solicitar a los TWP que examinen un proyecto de revisión del documento UPOV/INF/17/1 (documento UPOV/INF/17/2 Draft 5) en sus sesiones de 2021, teniendo en cuenta las modificaciones del documento UPOV/INF/17/2 Draft 4 que pueda acordar el TC en su quincuagésima sexta sesión.
13. Sobre la base de las conclusiones que alcancen el TC en su quincuagésima sexta sesión y los TWP en sus sesiones de 2021, un proyecto de revisión del documento UPOV/INF/17 (UPOV/INF/17/2 Draft 6) se someterá a la aprobación del Consejo en su quincuagésima quinta sesión, que se celebrará el 29 de octubre de 2021, previo acuerdo del TC en su quincuagésima séptima sesión y del CAJ en su septuagésima octava sesión, previstas en 2021.

14. Se invita al TC a:

- a) examinar la propuesta de revisión del documento UPOV/INF/17 sobre la base del documento UPOV/INF/17/2 Draft 4;

¹ Celebrada en Ginebra los días 28 y 29 de octubre de 2019.

² Celebrada en Hangzhou (China) del 16 al 18 de octubre de 2019.

³ Organizada por los Estados Unidos de América y celebrada por medios electrónicos del 23 al 25 de septiembre de 2020.

b) *solicitar a los TWP que examinen un proyecto de revisión del documento UPOV/INF/17/1 (documento UPOV/INF/17/2 Draft 5) en sus sesiones de 2021; y*

c) *tomar nota de que un proyecto de revisión del documento UPOV/INF/17 (UPOV/INF/17/2 Draft 6) se someterá a la aprobación del Consejo en su quincuagésima quinta sesión, que se celebrará el 29 de octubre de 2021, previo acuerdo del TC en su quincuagésima séptima sesión y del CAJ en su septuagésima octava sesión, previstas en 2021.*

[Sigue el Anexo]

PROPUESTA DEL BMT PARA LA REVISIÓN DEL DOCUMENTO UPOV/INF/17

ÍNDICE

A.	INTRODUCCIÓN.....	1
B.	PRINCIPIOS GENERALES.....	1
1.	Selección de los marcadores moleculares	2
1.1	<i>Conjuntos de variedades para el proceso de selección</i>	2
1.2	<i>Marcadores moleculares: consideraciones respecto del rendimiento</i>	2
2.	Selección del método de detección	3
2.1	<i>Métodos de determinación del perfil de ADN: consideraciones generales</i>	3
2.2.	<i>Acceso a la tecnología</i>	3
3.	Validación y armonización del conjunto de marcadores y el método de detección.....	3
3.1	<i>Validación y armonización: consideraciones generales</i>	3
3.2	<i>Consideraciones respecto del rendimiento: validación de los marcadores y los métodos</i>	3
3.3	<i>Consideraciones respecto de la coherencia</i>	4
4.	Creación de una base de datos específica para una especie.....	4
4.1	<i>Recomendaciones para diseñar bases de datos</i>	4
4.2	<i>Requisitos del material vegetal</i>	5
4.2.1	<i>Fuente del material vegetal</i>	5
4.2.2	<i>Tipo de material vegetal</i>	5
4.2.3	<i>Tamaño y tipo de la muestra (muestras en bloque o individuales)</i>	5
4.2.4	<i>Muestra de referencia de ADN</i>	5
4.3	<i>Tratamiento de los datos de secuencias</i>	5
4.4	<i>Tipo de base de datos</i>	6
4.5	<i>Modelo de base de datos</i>	6
4.6	<i>Diccionario de datos</i>	7
4.7	<i>Acceso a los datos y titularidad</i>	7
5.	Intercambio de datos.....	7
5.1	<i>Situaciones de intercambio de datos</i>	7
5.2	<i>Métodos de intercambio de datos</i>	7
6.	Resumen	8
C.	LISTA DE SIGLAS.....	8

A. INTRODUCCIÓN

El presente documento (Directrices BMT) tiene por finalidad ofrecer orientaciones sobre principios armonizados para el uso de marcadores moleculares con objeto de generar datos moleculares de gran calidad para diversas aplicaciones. En el presente documento solo se contemplan los marcadores moleculares de ADN.

Las Directrices BMT también tienen el propósito de abordar la creación de bases de datos de perfiles moleculares de variedades vegetales, obtenidos posiblemente en laboratorios diferentes y mediante el empleo de tecnologías también diferentes. Además, el objetivo del presente documento es lograr un alto grado de calidad de los marcadores y estimular el deseo de obtener datos reproducibles con la ayuda de esos marcadores aun cuando se den situaciones en las que los equipos y los reactivos químicos sean diferentes. Es preciso tomar precauciones específicas para asegurar la calidad de las entradas que se incluyen en la base de datos.

B. PRINCIPIOS GENERALES

Para determinar el perfil de ADN de una variedad vegetal, se requiere un conjunto de marcadores moleculares y un método para detectarlos. Dos conjuntos distintos de marcadores moleculares detectados mediante el mismo método darán lugar a dos perfiles diferentes de ADN para una misma variedad. Por el contrario, es previsible que empleando dos métodos distintos para detectar los alelos específicos de un determinado conjunto de marcadores moleculares se obtengan perfiles de ADN idénticos. No es necesario estandarizar el método de detección ni la tecnología, siempre que el rendimiento cumpla los criterios de calidad y los perfiles

de ADN obtenidos sean coherentes. Sea cual sea la tecnología utilizada para detectar conjuntos definidos de marcadores, el genotipo de una variedad determinada no debe verse afectado.

Los conjuntos de marcadores moleculares, los métodos de detección de marcadores y el proceso subsiguiente de desarrollo de la base de datos pueden subdividirse en cinco fases:

1. Selección de los marcadores moleculares
2. Selección del método de detección
3. Validación y armonización del método de detección
4. Creación de la base de datos
5. Intercambio de datos

En el presente documento se describen estas fases con mayor detalle. Las fases se consideran independientes del grado de desarrollo de las tecnologías de genotipado y de futuras mejoras en la secuenciación de alto rendimiento.

1. Selección de los marcadores moleculares

1.1 *Conjuntos de variedades para el proceso de selección*

Para determinar el perfil de ADN de las variedades vegetales y crear la base de datos, los marcadores moleculares han de seleccionarse en función del objetivo. Para iniciar el proceso de selección de marcadores, se necesita un número conveniente de variedades (conjunto de desarrollo) que refleje en la mayor medida posible la diversidad observada en el grupo, el cultivo, la especie o el tipo para el que los marcadores deben ser discriminatorios. La selección se refina determinando el perfil de otras variedades (conjunto de validación) a fin de medir el rendimiento de los marcadores. Para elegir el conjunto de validación pueden aplicarse los criterios siguientes:

- a) variedades o líneas de gran semejanza genética, NIL, RIL;
- b) líneas parentales y su descendencia;
- c) variedades genéticamente cercanas pero morfológicamente distintas (por ejemplo, mutantes);
- d) algunas variedades morfológicamente cercanas pero con genealogía diferente;
- e) distintos lotes de la misma variedad;
- f) orígenes diferentes de la misma variedad.

1.2 *Marcadores moleculares: consideraciones respecto del rendimiento*

A continuación se indican los criterios generales para seleccionar un marcador o un conjunto de marcadores específicos; estos criterios son aplicables independientemente del uso al cual se destinen los marcadores:

- a) repetibilidad, solidez y reproducibilidad en distintos laboratorios a efectos de evaluación de los datos;
- b) Posibles fuentes de marcadores moleculares
 - Marcadores moleculares procedentes de recursos públicos
 - Marcadores moleculares procedentes de recursos privados, evaluación y selección de chips y matrices comerciales específicos para determinadas especies
 - Marcadores moleculares seleccionados a partir de datos de secuencias de reciente obtención;
- c) evitar, en la medida de lo posible, marcadores con alelos “nulos” (alelo cuyo efecto es la ausencia de algún producto de la PCR a nivel molecular) –nuevamente, no es esencial pero sí aconsejable;
- d) Ha de permitir una evaluación fácil, objetiva e irrefutable de los perfiles de los marcadores. Los marcadores con buen rendimiento son preferibles a los perfiles complejos de marcadores que son susceptibles de interpretación. Las respuestas dicotómicas claras también facilitan la armonización;
- e) En general, los marcadores codominantes son preferibles a los marcadores dominantes porque su poder de discriminación es mayor;
- f) Marcadores ubicados en regiones codificantes y/o no codificantes; y
- g) El uso de marcadores moleculares es específico para una especie determinada, por lo que deben tenerse en cuenta las características de reproducción o multiplicación de esta.

Como es sabido, determinados usos pueden imponer otras consideraciones adicionales, entre ellas:

- a) El número de marcadores debe estar en equilibrio con la exactitud del genotipo que se requiera en función del objetivo. El número de marcadores para obtener la resolución o el poder de discriminación necesarios depende del tipo de marcador (dominante o codominante, bialélico o multialélico), de la especie y de la calidad del rendimiento de los marcadores;
- b) La cobertura genómica y el desequilibrio de ligamiento han de corresponderse con los objetivos. Conocer la posición física y/o genética de los marcadores seleccionados en el genoma no es indispensable, pero ayuda a seleccionar marcadores adecuados.

2. Selección del método de detección

2.1 *Métodos de determinación del perfil de ADN: consideraciones generales*

2.1.1 A continuación se señalan consideraciones importantes a la hora de seleccionar métodos de determinación del perfil de ADN:

- a) reproducibilidad de la producción de datos en distintos laboratorios y plataformas de detección (tipos de equipo diferentes);
- b) repetibilidad en el tiempo;
- c) poder de discriminación;
- d) tiempo y mano de obra;
- e) consistencia del rendimiento en cuanto a tiempo y condiciones (sensibilidad a leves variaciones en el protocolo o las condiciones);
- f) flexibilidad del método, posibilidad de emplear distinto número de muestras o de marcadores;
- g) la interpretación de los datos obtenidos no depende del equipo utilizado;
- h) viabilidad de las bases de datos;
- i) accesibilidad de la metodología;
- j) no depender de un aparato, unos reactivos químicos, un proveedor, unos socios o unos productos determinados;
- k) susceptible de automatización;
- l) apto para la multiplexación; y
- m) rentable (existe un equilibrio entre los costos, el número de muestras y el número de marcadores).

2.2. *Acceso a la tecnología*

Algunos marcadores moleculares y materiales están a libre disposición del público. Pero debido a la gran inversión que probablemente deberá hacerse para obtener marcadores de gran calidad, los marcadores y otros métodos o materiales pueden estar amparados por derechos de propiedad intelectual. La UPOV ha elaborado directrices, que han de seguirse, para la utilización de los productos y las metodologías que son objeto de derechos de propiedad intelectual. Es recomendable abordar todos los aspectos concernientes a los derechos de propiedad intelectual al principio de cualquier trabajo de desarrollo.

3. Validación y armonización del conjunto de marcadores y el método de detección

3.1 *Validación y armonización: consideraciones generales*

Los marcadores moleculares y los métodos de detección han de ser sólidos y dar lugar a perfiles de ADN coherentes. El rendimiento de los marcadores moleculares y los métodos de genotipado se evalúa mediante un proceso de validación. Cuando se trata de bases de datos compartidas, el proceso de armonización incluye la evaluación de la coherencia de los perfiles de ADN en distintos laboratorios utilizando equipos y reactivos químicos diferentes. El uso de marcadores y métodos validados permitirá obtener resultados armonizados.

3.2 *Consideraciones respecto del rendimiento: validación de los marcadores y los métodos*

El conjunto de marcadores seleccionado debe ser apto para cumplir su función. Ha de medirse la exactitud. Para determinar la idoneidad de un método y un conjunto de marcadores de ADN, deben tenerse en cuenta varios aspectos:

- a) la capacidad de discriminación o informativa;

- b) la repetibilidad: un mismo operador obtiene resultados idénticos en un ensayo con el mismo método e idénticas muestras de ensayo, en el mismo laboratorio y con el mismo equipo, en intervalos breves de tiempo;
- c) la reproducibilidad: distintos operadores obtienen resultados idénticos en un ensayo con el mismo método e idénticas muestras de ensayo, en el mismo o en distintos laboratorios y con equipos diferentes;
- d) la solidez: es una medida de su capacidad para no resultar afectado por desviaciones pequeñas pero deliberadas respecto de las condiciones experimentales descritas en los parámetros del procedimiento y proporciona una indicación de su fiabilidad durante el uso normal; y
- e) la tasa de error.

Las definiciones de las características de rendimiento están basadas en la norma ISO 16577:2016.

3.3 Consideraciones respecto de la coherencia

Para obtener resultados coherentes, el proceso de armonización de los marcadores y los métodos en distintos laboratorios, cuando se trata de una base de datos compartida (ensayo interlaboratorios (*ring test*)), ha de contemplar lo siguiente:

- a) La utilización, en todos los laboratorios, de una colección definida de variedades que represente una gran diversidad de alelos, como referencia para evaluar la coherencia entre ellos;
- b) La inclusión de duplicados, submuestras, plantas de una variedad para verificar la coherencia de los perfiles de ADN y calcular la tasa de error en distintos laboratorios;
- c) Acuerdos sobre la evaluación de los datos moleculares. La necesidad de elaborar un protocolo para valorar alelos o bandas en distintos laboratorios dependerá del tipo de marcador utilizado (por ejemplo, resulta indispensable si se utilizan marcadores SSR). Dicho protocolo deberá contemplar el modo de valorar los elementos siguientes:
 - i. alelos raros (alelos en un locus determinado que aparecen con una frecuencia que está por debajo de un umbral acordado (comúnmente, entre el 5 y el 10%) para una población);
 - ii. alelos nulos (alelos cuyo efecto es la ausencia de un producto PCR a nivel molecular);
 - iii. bandas “tenues” (bandas con intensidad inferior a un umbral de detección acordado, establecido ya sea empírica o automáticamente, y cuya valoración puede ser cuestionada);
 - iv. datos ausentes (cualquier locus para el cual, por alguna razón, no hay datos registrados en una o más variedades); y
 - v. bandas monomórficas o alelos no informativos (alelos o bandas que aparecen en todas las variedades analizadas, es decir, que no son polimórficos en una determinada colección de variedades).

4. Creación de una base de datos específica para una especie

Los datos que se almacenan en una base de datos y el modo de almacenarlos deben ser un reflejo de su proceso de obtención. Por lo tanto, al crear la base de datos deben considerarse distintos grados de tratamiento de estos (es decir, datos brutos, datos de secuencias...). En la base de datos han de almacenarse los resultados finales, por ejemplo, el perfil de ADN y su proceso de obtención, indicando la descripción del método de laboratorio y los pasos computacionales.

4.1 Recomendaciones para diseñar bases de datos

A la hora de diseñar bases de datos han de tenerse en cuenta los aspectos siguientes:

- a) La arquitectura de la base de datos ha de ser flexible; por ejemplo, se deben poder almacenar archivos planos y archivos comprimidos.
- b) Se requieren cuadros y entradas independientes para el trabajo experimental de laboratorio, el tratamiento de los datos y la valoración de los alelos.

c) Debe almacenarse información a distintos niveles, por ejemplo, la valoración de los alelos y las reglas de interpretación que sustentan la decisión, así como los enlaces a los datos brutos (archivos TIFF, archivos BAM) que se generaron.

d) Para los datos de secuenciación, se empleará la versión 4.2 o una superior del formato estándar VCF o BCF. En las entradas de las cabeceras debe figurar el nombre y la versión de los distintos *scripts* utilizados para mapear las lecturas de secuencias, filtrar las lecturas e identificar y filtrar las variantes, de manera que un bioinformático pueda repetir el análisis.

e) Si se trata de réplicas cuyos perfiles de ADN no coinciden, el registro ha de señalarse o descartarse, si procede. Las reglas aplicadas en estos casos han de quedar documentadas en un depósito de códigos de dominio público, al que se hará referencia en el archivo VCF. En el caso de las variedades heterogéneas, también pueden utilizarse frecuencias.

f) Validación de los datos de los archivos VCF y/o BCF conforme a las especificaciones pertinentes.

g) Datos de fácil intercambio (por ejemplo, API).

4.2 *Requisitos del material vegetal*

Debe tenerse en cuenta la fuente del material, el tipo de material y cuántas muestras se han de almacenar y compartir en la base de datos.

4.2.1 Fuente del material vegetal

El material vegetal a analizar deberá ser una muestra auténtica y representativa de la variedad y, a ser posible, habrá sido obtenido a partir de la muestra de la variedad examinada a los efectos de los derechos de obtentor o del registro oficial. Para usar esas muestras se requiere el permiso de la autoridad pertinente, bien sea el obtentor o el conservador, según corresponda. Se debería poder identificar el origen del material vegetal del que se tomaron las muestras por si posteriormente se demostrara que algunas de ellas no eran representativas de la variedad.

4.2.2 Tipo de material vegetal

El tipo de material vegetal a muestrear para la extracción del ADN y el procedimiento empleado para el muestreo dependerán en gran medida del cultivo o la especie vegetal de que se trate. Por ejemplo, en el caso de las variedades de propagación por semilla pueden utilizarse estas como fuente del ADN, mientras que, en el caso de las variedades de multiplicación vegetativa, el ADN puede extraerse del material foliar. Sea cual fuere la fuente del material, el método empleado para el muestreo y la extracción del ADN debería estar documentado. Además, es menester verificar mediante análisis de ADN la coherencia de los resultados obtenidos por los métodos de muestreo y extracción.

4.2.3 Tamaño y tipo de la muestra (muestras en bloque o individuales)

Es indispensable que las muestras tomadas para el análisis sean representativas de la variedad. Deberán tomarse en consideración las particularidades de la reproducción sexuada o la multiplicación vegetativa (véase la Introducción General).

4.2.4 Muestra de referencia de ADN

Se puede crear una colección de muestras de referencia de ADN con el material vegetal muestreado. El método de muestreo debe ajustarse a los procedimientos recomendados y se deben establecer criterios de calidad para la extracción de ADN. Es necesario documentar tanto el muestreo como la extracción.

Las muestras de ADN deberían almacenarse de forma tal que se evite su degradación (por ejemplo, a -80°C). La transferencia de las muestras de referencia de ADN se describe en la sección 1 del documento TGP/5.

4.3 *Tratamiento de los datos de secuencias*

Un registro detallado del proceso de tratamiento de los datos puede incluir la siguiente información:

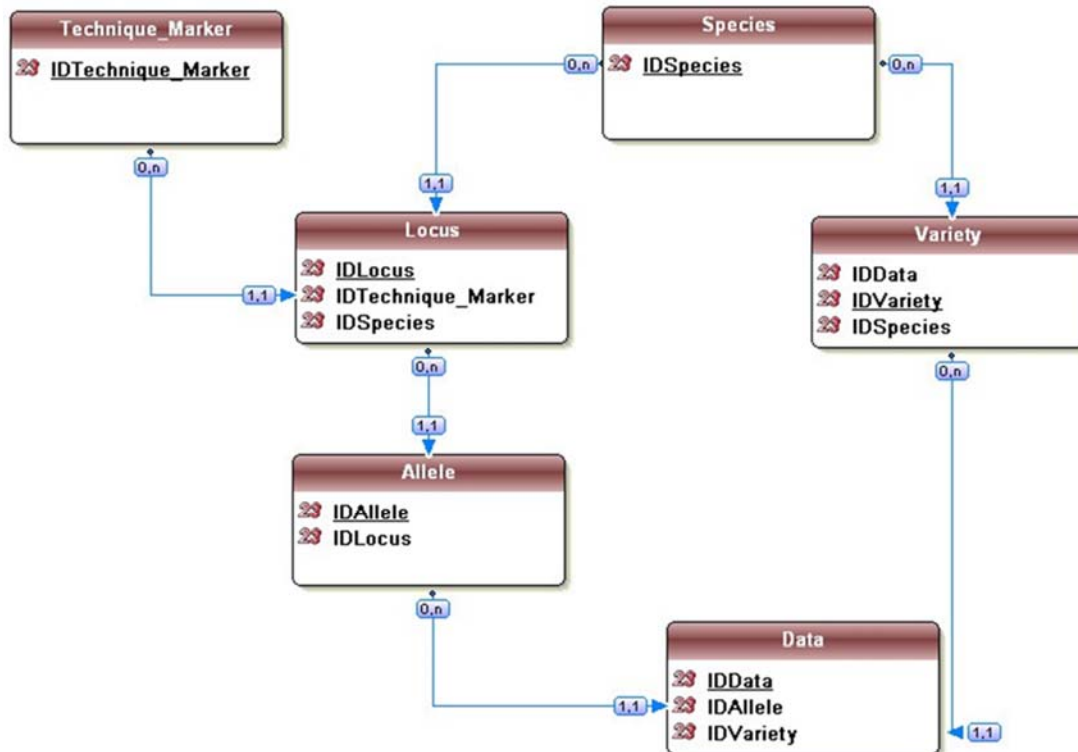
- a) tipo de herramientas y versiones utilizadas;
- b) línea de comandos utilizada para la herramienta y umbrales;
- c) recuentos de reproducibilidad;
- d) posibilidad de intercambiar datos y proceso de intercambio;
- e) si es posible, deben almacenarse los datos brutos de los alineamientos (archivos BAM o CRAM);
- f) los archivos VCF para varias muestras no son adecuados, debe figurar un archivo VCF por variedad;
- g) si se almacenan los archivos VCF, se deben registrar todas las posiciones (de variantes y no variantes) y su profundidad;
- h) para los métodos de detección se han de considerar y comparar enfoques tanto heurísticos como probabilísticos;
- i) las bases de datos deben facilitar la entrada y salida de los datos de identificación de variantes en un formato estándar (VCF o BCF);
- j) el proceso de tratamiento de los datos debe quedar reflejado en un detallado archivo de registro que se almacenará junto con los datos de identificación de variantes;
- k) si es posible, deben almacenarse los datos brutos para que el tratamiento de los datos pueda repetirse con nuevas herramientas o actualizaciones; y
- l) debe registrarse un valor p o la incertidumbre para un alelo determinado.

4.4 Tipo de base de datos

Los datos moleculares pueden almacenarse de muchas maneras diferentes y, en consecuencia, es importante que la estructura desarrollada para la base de datos sea compatible con todos los usos a los que pretendidamente se destinarán los datos.

4.5 Modelo de base de datos

La definición del modelo de base de datos estará a cargo de expertos en bases de datos de T.I., quienes contarán para ello con la colaboración de los usuarios de la base de datos. El modelo de base de datos habrá de contener, como mínimo, seis objetos básicos: *Species* (especie); *Variety* (variedad); *Marker detection method* (método de detección del marcador); *Marker* (marcador); *Locus* (locus); y *Allele* (alelo). Para las variantes obtenidas a partir de datos de secuenciación, los archivos VCF pueden almacenarse en una base de datos relacional o no SQL. En este caso, cada registro de una variante en la base de datos tiene una versión genómica, un cromosoma, una posición y un alelo de referencia definidos.



4.6 Diccionario de datos

4.6.1 En una base de datos, cada uno de los objetos configura un cuadro en el que se definen los campos. Por ejemplo:

a) Tipo de marcador: designa el código o el nombre de la técnica o el tipo de marcador utilizado, por ejemplo, SSR, SNP, etc.

b) Posición en el genoma de referencia o código del locus: De ser posible, se debe proporcionar una versión del ensamblaje del genoma, el cromosoma y la posición si se dispone de un genoma de referencia de la especie de que se trate, p. ej., SL2.50ch05:63309763 en el caso del tomate (*Solanum lycopersicum*), la versión 2.50 del ensamblaje del genoma, el cromosoma 5 y la posición 63309763. Si no se dispone de un genoma de referencia o se desconoce la posición, puede utilizarse el nombre o el código del locus de la especie en cuestión, por ejemplo *gwm 149*, *A2*, etc.

c) Genotipo: En el caso de los perfiles SNP, debe proporcionarse la composición alélica del *SNP* o el *MNP*, por ejemplo, *A/T* o *A/A*. Si se trata de otras técnicas, el genotipo designa el nombre o el código del alelo de un locus determinado de la especie de que se trate, por ejemplo, *1*, *123*, etc.

d) Profundidades alélicas o valor de los datos: En el caso de los SNP obtenidos de datos de secuenciación de nueva generación, estos deben indicar la profundidad de cobertura de los alelos, por ejemplo, *10/20* para un alelo *A/T*, en el que *A* se ha cubierto mediante 10 lecturas y *T* mediante 20 lecturas. De lo contrario, se designa el valor de los datos de una muestra determinada respecto de un alelo presente en un locus determinado, por ejemplo, *0* (*ausencia*), *1* (*presencia*), *0,25* (*frecuencia*), etc.

e) Variedad: Denominación de la variedad o referencia del obtentor: la variedad es el objeto para el que se han obtenido los datos.

f) Tipo de variedad: por ejemplo, línea endógama o híbrido.

g) Especie: la especie se designa mediante el nombre botánico o el nombre nacional común que, a veces, se refiere también al tipo de variedad (por ejemplo, variedades del tipo de verano o de primavera, etc.). Se recomienda utilizar el código UPOV para evitar los problemas que plantean los sinónimos.

4.6.2 En el "diccionario de datos" debe especificarse, en cada cuadro, el número de campos, su nombre y definición, y los posibles valores y normas que han de aplicarse.

4.7 Acceso a los datos y titularidad

Se recomienda abordar todas las cuestiones concernientes a la titularidad de los datos y el acceso a los datos de la base de datos en el momento de iniciar cualquier trabajo.

5. Intercambio de datos

5.1 Situaciones de intercambio de datos

A efectos de la cooperación, el modelo de datos debe admitir distintas situaciones, como el intercambio de datos generados a partir de un conjunto estándar de marcadores para un cultivo determinado (situación nº 1) y la búsqueda y consulta de datos de determinadas variedades generados a partir del mismo conjunto estándar de marcadores (situación nº 2). En el Anexo "Situaciones de intercambio de datos y métodos de transmisión de datos" se ofrecen detalles técnicos de ambas situaciones.

5.2 Métodos de intercambio de datos

5.2.1 Los datos a transmitir pueden contener información diversa, como loci, muestras, ADN y datos y perfiles de huellas genéticas. El método de transmisión de datos se determinará con arreglo al contenido que se ha de transmitir y debe tenerse en cuenta lo siguiente: a) la cantidad de datos;

b) la complejidad de los datos;

c) los requisitos para las funciones de búsqueda.

En el Anexo "Situaciones de intercambio de datos y métodos de transmisión de datos" se ofrecen detalles técnicos de ambas situaciones.

5.2.2 Los formatos de datos empleados habitualmente son zip, csv, json y xml. A continuación se indican las características de cada uno:

1) El formato zip admite diversos archivos de información sobre los datos en formato original y, gracias a su elevado índice de compresión de datos y facilidad de transmisión, resulta adecuado para datos complejos y de gran volumen.

2) El formato csv es más adecuado para información sobre datos en formato sencillo, con la ventaja de que contiene menos datos no válidos y la velocidad de procesamiento es mayor.

3) Los formatos json y xml pueden contener información sobre datos más complejos e información más redundante, pero ambos muestran una buena legibilidad.

6. Resumen

A continuación se ofrece un resumen del enfoque recomendado para la determinación de alta calidad del perfil de ADN de las variedades, incluida la selección y utilización de marcadores moleculares, así como la creación de bases de datos moleculares compartidas y sostenibles (es decir, bases de datos que en el futuro puedan llenarse con datos procedentes de una diversidad de fuentes, independientemente de la tecnología empleada).

- a) considerar el enfoque cultivo a cultivo;
- b) acordar un tipo de marcador aceptable y la fuente del mismo;
- c) acordar las plataformas y los equipos de detección aceptables;
- d) acordar los laboratorios que participarán en la prueba;
- e) acordar las cuestiones relativas a la calidad;
- f) verificar la fuente del material vegetal empleado;
- g) acordar los marcadores a utilizar en una fase de evaluación preliminar en la que se contará con la colaboración de más de un laboratorio y se usarán equipos de detección diferentes;
- h) realizar una evaluación;
- i) elaborar y acordar un protocolo para la evaluación de los datos moleculares;
- j) acordar el material vegetal y el conjunto de referencias a analizar, y la fuente de los mismos;
- k) analizar la colección de variedades acordada, en distintos laboratorios y con distintos equipos de detección, utilizando muestras duplicadas y, si surgiesen problemas, intercambiando muestras o extractos de ADN;
- l) utilizar referencias (variedades, muestras de ADN o alelos, según corresponda) en todos los análisis;
- m) verificar todas las etapas (inclusive la entrada de datos) – lograr la mayor automatización posible;
- n) efectuar una “prueba a ciegas” de la base de datos en laboratorios diferentes;
- o) adoptar los procedimientos para la adición de nuevos datos.

C. LISTA DE SIGLAS

API	Interfaz de programación de aplicaciones (<i>Application Programming Interface</i>)
BAM	Mapa de alineamiento binario (<i>Binary Alignment Map</i>)
BCF	<i>Binary Call Format</i>
CRAM	<i>Compressed Reference-oriented Alignment Map</i>
MNP	Polimorfismo de múltiples nucleótidos (<i>Multiple Nucleotide Polymorphism</i>)
NGS	Secuenciación de nueva generación (<i>Next Generation Sequencing</i>)
NIL	Línea casi isogénica (<i>Near Isogenic Line</i>)
RIL	Línea recombinante endógama (<i>Recombinant Inbred Line</i>)
SAM	Mapa de alineamiento de secuencias (<i>Sequence Alignment Map</i>)
SNP	Polimorfismo de nucleótido único (<i>Single Nucleotide Polymorphism</i>)
SQL	Lenguaje de consulta estructurado (<i>Structured Query Language</i>)
SSR	Secuencias simples repetidas (<i>Simple Sequence Repeats</i>)
TIFF	Formato de archivo de imagen con etiquetas (<i>Tagged Image File Format</i>)
VCF	<i>Variant Call Format</i>

Situación nº 2: búsqueda y consulta de datos de determinadas variedades generados a partir del mismo conjunto estándar de marcadores

Para buscar y consultar datos moleculares de una variedad, se puede recurrir al siguiente servicio en línea:
https://office.org/variety?id={irn}&techniqueid={technique_code} vi

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

El resultado sería el siguiente:

```
{ "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: Métodos de transmisión de datos

A continuación se ofrece un ejemplo de creación de un paquete de datos de huella genética en formato zip para su transmisión. En primer lugar, han de asignarse identificadores independientes a las muestras, el ADN, los datos de huella genética y el atlas de la huella genética. De este modo, el archivo de datos en formato json contendrá toda la información sobre los loci, las muestras y el ADN. Cada conjunto de datos de huella genética se almacena por separado en un archivo json. El identificador de la huella genética estará vinculado al correspondiente locus de los datos de huella genética, y todos los archivos de datos de huella genética y del espectro de la huella genética se almacenarán por separado en el directorio correspondiente. El paquete de datos de huella genética tendrá la siguiente estructura:

```
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
```

El paquete de datos de huella genética en formato zip puede ampliarse para introducir más información. El archivo de datos de huella genética constituye el núcleo del paquete y de la correlación, por lo que puede analizarse correctamente la correlación entre las partes y los datos pueden transmitirse entre diferentes sistemas.

[Fin del Apéndice del Anexo y del documento]