

## **Nueva estrategia para el diseño de colecciones nucleares para estudios de asociación genética**

A.M. Vargas<sup>1</sup>, M.T. de Andrés<sup>1</sup> y J. Ibáñez<sup>2</sup>.

1 Instituto Madrileño de Investigación y Desarrollo Rural, Agrario y Alimentario (IMIDRA), Finca El Encín, Ctra. A2, 28800 Alcalá de Henares, Madrid.

2 Instituto de Ciencias de la Vid y del Vino (CSIC, Universidad de La Rioja, Gobierno de La Rioja), Edificio CCT, C/Madre de Dios 51, 26006 Logroño, La Rioja.

### **Resumen**

**Las colecciones nucleares son ampliamente utilizadas para diferentes fines, tanto en investigación como para conservación de germoplasma. En la actualidad, el proyecto RF2012-27, que aglutina a la práctica totalidad de colecciones nacionales de vid, tiene entre sus objetivos la creación de una colección nuclear de vid que represente la diversidad genética de la vid en España. Existen diferentes posibilidades para generar colecciones nucleares, en función del objetivo, pero resulta de interés producir colecciones nucleares que puedan servir para una amplia gama de aplicaciones. Uno de los métodos más comunes utiliza la estrategia de maximización alélica (M-strategy), basado en la selección, en una colección global, de aquellas accesiones que maximizan la riqueza alélica de la sub-colección o colección nuclear. Sin embargo, las clases minoritarias a menudo aparecen en una frecuencia muy baja en la colección nuclear, lo que puede representar un problema para algunas aplicaciones, como los estudios de asociación genética. En este tipo de estudios se intenta relacionar una determinada característica observable o medible (fenotipo) con un determinado cambio o polimorfismo en la secuencia de ADN (genotipo). En estos estudios, la existencia de clases fenotípicas y/o genotípicas con baja frecuencia puede dar lugar a la aparición de falsos positivos, asociaciones espurias que complican mucho el análisis. Asimismo, reducen la potencia del análisis, o capacidad para detectar asociaciones reales, dando lugar a la aparición de falsos negativos. En este trabajo se propone y evalúa una nueva metodología para construir colecciones nucleares utilizando la estrategia de maximización en pasos secuenciales, para aumentar la frecuencia de las clases minoritarias.**

**Palabras clave:** colección nuclear, estrategia M, marcadores moleculares, marcadores morfológicos

### **INTRODUCCIÓN**

El mantenimiento de colecciones de germoplasma es de gran importancia para conservar la diversidad genética en especies cultivadas, las cuales en la mayoría de los casos han sufrido una intensa erosión genética como consecuencia de diferentes factores agrícolas y sociales, como la importación de variedades extranjeras o enfermedades tales como la filoxera en el caso de la vid. Dado el elevado número de accesiones presentes en algunas colecciones, resulta complicado realizar una caracterización completa de las mismas, por lo que para numerosas aplicaciones resulta de gran utilidad el establecimiento de colecciones nucleares, que en un reducido número de accesiones representan el mayor porcentaje posible de diversidad existente en la colección global.

Actualmente, uno de los objetivos del proyecto RF2012-27, es la creación de una colección nuclear que represente la diversidad genética de la vid en España.

En los últimos años se han desarrollado diferentes métodos para la construcción de colecciones nucleares, en función de la ausencia o presencia de estratificación en la colección. En colecciones no estratificadas una opción es la selección de accesiones al azar (Estrategia R), pero actualmente, el criterio más usado es la maximización alélica o Estrategia M (Schoen y Brown, 1993), que examina las posibles colecciones nucleares y selecciona la que maximiza el número total de clases en los marcadores usados, que pueden ser moleculares y/o morfológicos. Este método permite definir primero el número mínimo de accesiones necesarias para capturar la diversidad deseada, para posteriormente identificarlas combinaciones de accesiones del tamaño requerido que incluyen el mayor número de clases para los marcadores empleados. Dado que el número de posibles combinaciones de accesiones crece factorialmente con el tamaño de las colecciones nuclear y global, no es posible en la mayoría de casos explorar todas las colecciones nucleares posibles, por lo que los software diseñados para la construcción de colecciones nucleares implementan algoritmos para mejorar los resultados a la vez que mantienen un tiempo de ejecución razonable. Como consecuencia, diferentes ejecuciones del programa pueden dar lugar a diferentes colecciones nucleares, no siempre de la misma calidad. Por otro lado, es importante tener en cuenta el objetivo de la colección, ya que diferentes aplicaciones podrían requerir distinto número de accesiones, diferente fracción de representación de la variabilidad total, o diferentes distribuciones fenotípicas, y por tanto, deberían usarse distintos métodos para su construcción.

El mapeo de asociación es una técnica actualmente extendida para la búsqueda de correlaciones entre genotipos y fenotipos en una población, y el uso de colecciones nucleares puede ser de gran utilidad para aumentar la eficacia de esta aplicación. Los estudios de asociación en plantas suelen presentar problemas relacionados con la distribución de frecuencias de clases en la colección seleccionada: es común la obtención de asociaciones espurias debido a la presencia de clases únicas o minoritarias. Ciertas clases de caracteres de interés comercial en vid presentan baja frecuencia, como el sabor moscatel o la apirenia, y la construcción de colecciones nucleares basada en la maximización de la diversidad total puede disminuir aún más la frecuencia de esas clases minoritarias. En este trabajo, proponemos una nueva metodología basada en la Estrategia M para obtener colecciones nucleares más adecuadas para mapeo de asociación, tratando de mejorar el poder de análisis mediante la maximización de las clases minoritarias.

## **MATERIAL Y MÉTODOS**

El material empleado en este trabajo consistió en 246 genotipos únicos pertenecientes a la colección de uva de mesa del Banco de Germoplasma de El Encín (IMIDRA). Dichas accesiones fueron caracterizadas mediante 37 descriptores fenológicos y morfológicos (17 cualitativos y 20 cuantitativos) y mediante 20 marcadores SSR (Ibáñez et al., 2009).

### **Construcción de colecciones nucleares**

En primer lugar, se evaluó la redundancia de la colección mediante los métodos R (al azar) y M (Maximización), implementados en el software MStrat 4.1 (Gouesnard et al., 2001), con el fin de establecer el número de accesiones necesario para representar un alto porcentaje de la variabilidad existente en la colección global.

A continuación se construyeron tres colecciones nucleares de 96 entradas usando la estrategia M. Para obtener cada una se generaron 100 réplicas de colecciones nucleares con MStrat usando 100 iteraciones y aplicando dos criterios de optimización: en primer lugar, la riqueza de clases, y en segundo lugar, el Índice de Shannon. Cada colección nuclear (CN) se construyó de acuerdo a los siguientes métodos:

*Método 1:* CN obtenida en un paso. Procedimiento: Generar 100 réplicas de CNs de 96 entradas con MStrat → Seleccionar la CN con mayor Índice de Shannon (CN96).

*Método 2:* Tres sub-colecciones, cada una obtenida en un paso. Procedimiento: Generar 100 réplicas de CNs de 32 entradas con MStrat → Seleccionar la sub-colección con el mayor Índice de Shannon → Eliminar de la colección global las accesiones incluidas en esta sub-colección y repetir el proceso dos veces más → Unir las tres sub-colecciones de 32 entradas para obtener una CN de 96 entradas (CN32Sx3).

*Método 3:* Tres sub-colecciones, cada una obtenida en 4 pasos. Procedimiento: Generar 100 réplicas de CNs de 32 entradas con MStrat → Seleccionar las accesiones más frecuentemente representadas en las 100 réplicas y fijarlas para el siguiente paso → repetir el proceso dos veces más → Generar 100 réplicas que incluirán todas las accesiones fijadas → Seleccionar la sub-colección con el mayor número de clases y, en caso de empate, con el mayor Índice de Shannon → Eliminar de la colección global las accesiones incluidas en esta sub-colección y repetir el proceso dos veces más → Unir las tres sub-colecciones de 32 entradas para obtener una CN de 96 entradas (CN32Fx3).

### **Evaluación de las colecciones nucleares**

Se compararon las tres colecciones obtenidas con los diferentes métodos y su calidad se evaluó en función del Índice de Shannon y de las frecuencias de las clases, prestando especial atención a la frecuencia de las clases minoritarias.

## **RESULTADOS**

El objetivo de este trabajo fue maximizar la representación de clases minoritarias en colecciones nucleares. Para ello, se desarrollaron tres métodos con diferentes pasos secuenciales para la construcción de dichas colecciones.

### **Redundancia y eficiencia de la Estrategia M**

El set de datos empleado contenía 246 accesiones que incluían un total de 461 clases diferentes para los 57 marcadores empleados. La variabilidad capturada por la estrategia M fue un 14% superior a la capturada por la estrategia R. El análisis de redundancia determinó que 54 accesiones (22% del total de accesiones) son suficientes para representar el total de 461 clases. Una colección nuclear de 32 entradas representa hasta un 94% de la diversidad existente en la colección total.

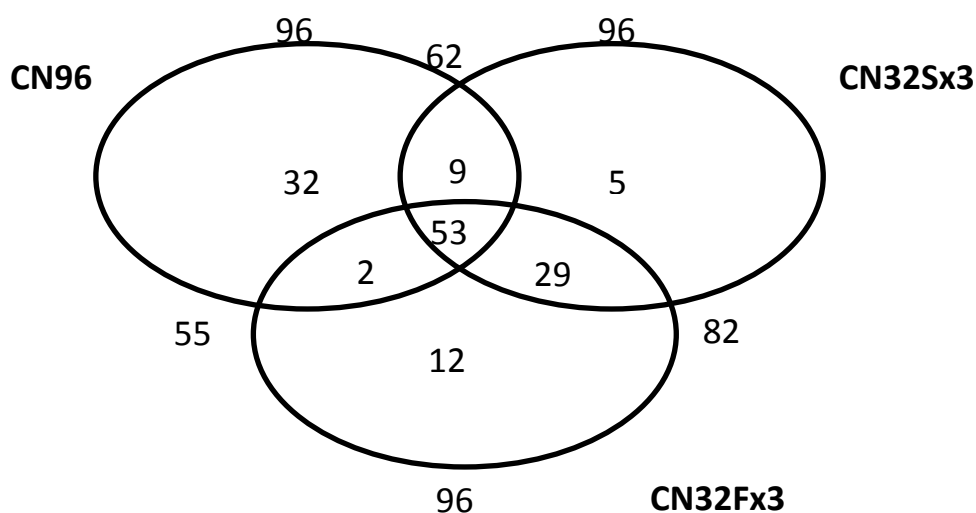
### **Construcción de las colecciones nucleares**

Las 100 réplicas generadas mediante el método 1 incluyeron todas las clases existentes en la colección completa, de manera que la selección de CN96 se basó en el Índice de Shannon. Este criterio fue también el empleado para seleccionar las tres sub-colecciones construidas mediante el método 2 (CN32S). Cada sub-colección CN32F fue obtenida en 4 pasos usando el método 3. Mediante los primeros tres pasos se fueron seleccionando las accesiones que aparecían con mayor frecuencia en las 100 réplicas obtenidas con MStrat. El número de accesiones seleccionadas y fijadas en cada paso varió

entre 4 y 11, que aparecían en las 100 colecciones con una frecuencia promedio del 84% en el paso 1, del 73% en el paso 2 y del 57% en el paso 3.

### Composición de las colecciones nucleares

Las accesiones compartidas entre las tres colecciones nucleares fueron 53 (55%); sin embargo, 82 accesiones (85%) resultaron comunes entre CN32Sx3 y CN32Fx3, mientras dichas colecciones compartieron sólo el 65% y el 57% de sus accesiones con CN96 (Figura 1).



**Figura 1.** Distribución de las accesiones entre las tres colecciones nucleares.

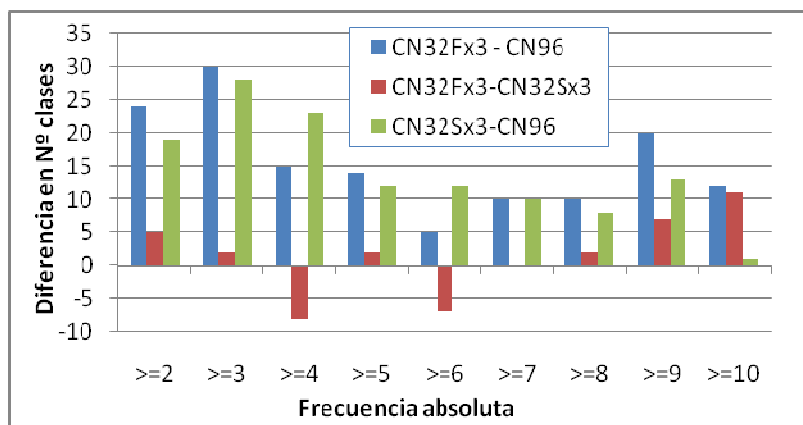
### Evaluación de la diversidad capturada por las colecciones nucleares

La colección obtenida mediante el método 1 presentó el total de clases existentes en la colección global, mientras las obtenidas mediante los métodos 2 y 3 contenían más del 99% del total de clases (Tabla 1). El número de clases únicas fue superior en CN96, mientras que el número de clases minoritarias fue mayor en las colecciones CN32x3, que presentaron menos clases de baja frecuencia (<4) y más clases minoritarias con frecuencia superior a 3.

**Tabla 1.** Índice de Shannon y número de clases representado en cada colección nuclear. Las clases minoritarias son las que presentan una frecuencia absoluta menor de 10 y no son únicas.

Evaluación de distribución de clases	Global	CN96	CN32Sx3	CN32Fx3
Índice de Shannon	85,65	88,54	90,57	90,25
Nº clases totales	461	461	459	457
Nº clases únicas	32	68	47	40
Nº total clases minoritarias (frecuencia >1 y <10)	124	179	197	191
Nº clases con frecuencia <4	94	136	111	117
Nº clases minoritarias con frecuencia >3 y <10	75	26	45	40

La Figura 2 muestra las diferencias en el número de clases presentes en cada CN con frecuencias absolutas mínimas entre 2 y 9. La ganancia en estas clases fue claramente superior en las colecciones CN32x3 frente a CN96, mientras que entre ambas CN32x3 las diferencias son mucho menores y tanto positivas como negativas.



**Figura 2.** Diferencias entre los tres métodos respecto a la ganancia de clases de menor frecuencia.

## DISCUSIÓN

En colecciones de germoplasma con elevado número de accesiones, una caracterización completa fenotípica y genotípica puede resultar muy costosa, por lo que generalmente es necesario trabajar con colecciones nucleares, que contienen una pequeña representación de la colección global manteniendo la máxima diversidad posible. Dependiendo de la aplicación de cada colección nuclear, pueden emplearse diferentes métodos para su construcción, y por tanto, también deberían emplearse diferentes criterios para la evaluación de diferentes tipos de colecciones, varios de los cuales han sido recientemente discutidos por Odong et al. (2013). Por ejemplo, para ciertos análisis estadísticos como los análisis de asociación genética es importante que cada clase esté representada varias veces para incrementar el poder del test, reduciéndose las asociaciones espurias por azar (falsos positivos) y los falsos negativos resultantes de una significación insuficiente.

La evaluación de la redundancia constituye un paso previo a la construcción de la colección nuclear para determinar el número de accesiones necesarias para capturar la diversidad requerida, o, en caso de que el número de entradas sea predeterminado, para conocer cuánta diversidad es posible capturar. La ganancia de diversidad fue un 14% superior empleando el método M frente al R, muy similar a lo que previamente observaron Le Cunff et al. (2008) en 2.262 genotipos únicos de *V. vinifera* para 20 SSRs eliminando alelos con frecuencia inferior a 0,05 (15%), confirmando la superioridad de la estrategia M en vid, a pesar de su elevada heterocigosidad.

En este trabajo se propone una nueva metodología para la construcción de colecciones nucleares con el objetivo de incrementar su eficiencia y maximizar la representación de clases minoritarias, tratando de capturar un mínimo de tres representantes para la mayoría de clases. Para ello, dos nuevos métodos (2 y 3) se compararon con un método Standard (método 1) (Le Cunff et al., 2008). Las tres

colecciones comparten 53 de las 96 accesiones, por lo que todas las diferencias se deben al resto de accesiones diferentes en cada caso.

Los métodos 2 y 3, basados en la fusión de tres sub-colecciones independientes para obtener la colección nuclear definitiva dieron lugar a la obtención de menos clases únicas (indeseables) y menos clases con frecuencia baja (<4) (Tabla 1). Todo ello dio lugar a que mostraran un claro incremento en la frecuencia de clases minoritarias con respecto al método 1 (Figura 2). También se obtuvo un mayor Índice de Shannon con estos métodos, indicando una distribución de clases más equilibrada.

El método 3 consistió en fijar las accesiones más representadas en 100 réplicas a lo largo de tres pasos, intentando evitar accesiones que pudieran ser incluidas en la sub-colección nuclear por azar. El porcentaje de representación mínimo requerido para fijar las accesiones en el siguiente paso podría variar entre especies y/o colecciones, por lo que debe establecerse para cada caso, intentando incluir un número de entradas similar en cada uno (en torno a 8 accesiones en nuestro caso). Los resultados obtenidos con el método 2 fueron muy similares a los del método 3, e incluso superiores en algunos parámetros. De hecho, 82 entradas fueron comunes entre ambas colecciones. Dado que el método 2 requiere un solo paso para la construcción de cada sub-colección, es la opción más recomendable para la obtención de colecciones nucleares de forma general, aunque puede haber casos específicos donde el método 3 podría ser más apropiado, como para colecciones globales muy amplias con un elevado número de clases, en las que diferentes combinaciones de accesiones podrían dar lugar a colecciones nucleares de similar calidad, resultando conveniente añadir otro criterio de selección de accesiones (su repetitividad en 100 colecciones nucleares potenciales).

## CONCLUSIONES

En este trabajo proponemos una nueva metodología para generar colecciones nucleares en pasos secuenciales usando la estrategia M, con el fin de incrementar la representación de clases minoritarias en la colección nuclear y mejorar los resultados de los análisis de asociación genética con ciertos caracteres de interés comercial y agronómico cuyas clases deseadas ocurren en baja frecuencia.

## Referencias

- Gouesnard, B., Bataillon, T.M., Decoux, G., Rozale, C., Schoen, D.J., David, J.L. (2001). MSTRAT: An algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. *The Journal of Heredity*, 92, 93-94.
- Ibáñez, J., Vargas, A.M., Palancar, M., Borrego, J., de Andrés, M.T. (2009). Genetic Relationships among Table-Grape Varieties. *American Journal of Enology and Viticulture*, 60, 35-42.
- Le Cunff, L., Fournier-Level, A., Laucou, V., Vezzuli, S., Lacombe, T., Adam-Blondon, A., Boursiquot, J.M., This, P. (2008). Construction of nested genetic core collections to optimize the exploitation of natural diversity in *Vitis vinifera* L. subspecies *sativa*. *BMC Plant Biology*, 8:31.
- Odong, T., Jansen, J., Eeuwijk, F.v., Hintum, T.v. (2013). Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *Theoretical and Applied Genetics*, 126, 289-305.
- Schoen, D.J. y Brown, A.H.D. (1993). Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. *Proceedings of the National Academy of Sciences*, 90, 10623-10627.