Selección de colecciones núcleo de maíz (*Zea mays* L.) utilizando un algoritmo que permite maximizar la riqueza alélica y diversas métricas de distancia genética

Darío Alberto Micheli [1,2][0009-0009-6135-8659], Ignacio Torrent [3][0009-0009-9918-0120], Manuela Carrere Gómez [1,4][0009-0009-5774-598X], Roberto Lorea [1][0009-0000-3319-7047], María Laura Federico [1,5][0000-0002-9171-3617]

[1] Instituto Nacional de Tecnología Agropecuaria-Estación Experimental Pergamino (INTA-EEA Pergamino)
[2] Universidad Nacional del Noroeste de la Provincia de Buenos Aires (UNNOBA)
[3] *Bayer CropScience*, Estación Fontezuela, Buenos Aires, Argentina
[4] CITNOBA (UNNOBA - UNSAdA - CONICET)
[5] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
federico.marialaura@inta.gob.ar

Resumen. Las colecciones núcleo (CN) son subconjuntos de accesiones que representan una colección más amplia, conservando su diversidad genética y evitando la redundancia. En programas de mejoramiento genético de cultivos, las CN son importantes herramientas para tener disponibles, ya que permiten caracterizar y utilizar los recursos genéticos sin la necesidad de evaluar la totalidad del germoplasma. En este contexto, el presente trabajo tuvo como objetivo evaluar la implementación de un protocolo de selección de CN multipropósito, de máxima variabilidad genética, a fin de asistir al PMG de Maíz del INTA EEA-Pergamino en futuros estudios de mapeo por asociación a genoma completo y predicción genómica. Para ello, se utilizó un algoritmo basado en un método avanzado de búsqueda local estocástica, *Core Hunter 3*, capaz de utilizar matrices genotípicas y fenotípicas para maximizar diversos parámetros de distancia y diversidad genética simultáneamente. Se seleccionaron 7 CN del mismo tamaño (n=115) con distintos objetivos de optimización en base a datos genotípicos, y la combinación de éstos con datos fenotípicos desde un panel de mejoramiento (PPMG) conformado por 484 líneas endocriadas de maíz. En todos los casos, las CN seleccionadas retuvieron el 100% de los alelos presentes (CV=1), exhibiendo similares valores promedio para heterocigosidad esperada (HE=0.40) y heterocigosidad observada (HO=0.02) que el PPMG. Luego de evaluar la distribución de 3 caracteres fenotípicos y la representatividad poblacional en las CN seleccionadas por el algoritmo, determinamos que las CN seleccionadas con el objetivo *Accesion To Nearest Entry* (A-NE) cumplen con los requisitos para ser utilizadas como CN multipropósito en futuras investigaciones.

Palabras clave: SNP, Colección Núcleo, Diversidad Genética.

Selection of maize (*Zea mays* L.) core colletions using an algorithm that maximizes allele richness and diverse genetic distance metrics

Abstract. Core collections (CC) are subsets of accessions that represent a broader collection, preserving their genetic diversity and avoiding redundancy. In plant breeding programs, CC are valuable tools since they enable the characterization and use of genetic resources without the need to evaluate the entire germplasm. In this context, the present work aimed to evaluate the implementation of a protocol to select multipurpose CC, with maximum genetic variability,

to assist the INTA EEA-Pergamino Maize breeding program in future genome-wide association mapping (GWAS) and genomic prediction (GP) studies. *Core Hunter 3*, an algorithm based on an advanced stochastic local search method, capable of incorporating genotypic and phenotypic matrices, was used to maximize various distance and genetic diversity parameters simultaneously. Seven CC of the same size (n=115) were selected with different optimization objectives based on genotypic data and its combination with phenotypic data, from a breeding panel (BP) consisting of 484 maize inbred lines. All selected CC retained 100% of the alleles present (CV=1), exhibiting similar average values for expected heterozygosity (HE=0.40) and observed heterozygosity (HO=0.02) as the BP. After evaluating the distribution of three phenotypic traits and the population representativeness of the CC selected by the algorithm, we determined that the CC selected with the A-NE objective meet the requirements to be used as multipurpose CC in future research investigations.

Keywords: SNP, Core Collection, Genetic Diversity.

1 Introducción

Las colecciones núcleo (CN) (Frankel, 1984) son subconjuntos de accesiones que representan la diversidad genética de una colección más amplia, lo que permite caracterizar y utilizar los recursos genéticos sin evaluar la totalidad del germoplasma (Soleimani et al., 2020). La selección de CN no es una tarea fácil, se debe minimizar la duplicación de accesiones y maximizar la diversidad genética (Brown, 1989). Además, el parámetro que define la calidad de una CN es su objetivo o propósito. No es lo mismo, captar accesiones con resistencia a una enfermedad o alto rendimiento que representar el patrón de diversidad genética en la colección (Odong et al., 2013).

En los últimos años, ante el incremento en el volumen de datos genotípicos y fenotípicos disponibles, se impulsó una búsqueda exhaustiva de métodos que sean capaces de seleccionar CN mediante la implementación de algoritmos eficientes. En este sentido, *Core Hunter 3*, un algoritmo basado en un método avanzado de búsqueda local estocástica, es capaz de utilizar matrices genotípicas y fenotípicas para maximizar diversos parámetros de distancia y diversidad genética simultáneamente (De Beukelaer et al., 2018). En este trabajo, evaluamos su implementación en un protocolo de selección de CN multipropósito, de máxima variabilidad genética, a fin de asistir al Programa de Mejoramiento Genético (PMG) de Maíz del INTA EEA-Pergamino en futuros estudios de mapeo por asociación a genoma completo y predicción genómica.

2 Materiales y Métodos

El panel de mejoramiento (PPMG) conformado por 484 líneas endocriadas de maíz fue genotipado con un panel *DArTag* de mediana densidad (3305 SNP) desarrollado por CIMMYT-CGIAR. Los materiales fueron sembrados siguiendo un Diseño en Bloques Completos al Azar (DBCA) con dos repeticiones en 3 localidades el año 2019. Tres

caracteres fenológicos fueron evaluados: (i) altura de planta (AP), (ii) altura de inserción de espiga (AE) y (iii) días a floración (DAF). Los BLUPs (BLUP, *Best Linear Unbiased Predictor*) fueron calculados usando *META-R* (Alvarado et al., 2020).

La matriz genotípica filtrada (461 líneas, 2199 SNP) por frecuencia del alelo menor (MAF) igual o mayor 0.05 y heterocigosidad menor a 20% fue recodificada utilizando la función *snp.recode* del paquete *ASRgenomics* (Gezan et al., 2024). La selección y evaluación de CN se llevó a cabo utilizando las funciones *coreHunterData*, *sampleCore* y *evaluateCore* del paquete *corehunter* (De Beukelaer et al., 2018). La diversidad genética de *Nei* (DG), heterocigosidad observada (HO), heterocigosidad esperada (HE) y coeficiente de endogamia (F) se estimaron con la función *popgen* del paquete *snpReady* (Granato et al., 2018). La pertenencia de cada línea a distintas subpoblaciones se basó en la estructura poblacional disponible del PPMG (K=12). La prueba de bondad de ajuste Chi-cuadrado se realizó con la función *chisq.test* (B=10000, $p<0.05$). Estos análisis se efectuaron en entorno *R* (R, 2015).

3 Resultados

Se seleccionaron 7 CN del mismo tamaño (n=115, 25% del PPMG) con distintos objetivos de optimización en base a datos genotípicos, y la combinación de éstos con datos fenotípicos (+F). Los objetivos de optimización utilizados maximizaron: (i) la representatividad (A-NE, *Accession To Nearest Entry*), (ii) la diversidad (E-NE, *Entry To Nearest Entry* y E-E, *Entry to Entry*) y (iii) la riqueza alélica (HE, *Average Expected Heterozygosity*). La diversidad genética de las CN seleccionadas se evaluó utilizando distintos criterios propuestos por Odong et al. (2013). En todos los casos, las CN seleccionadas retuvieron el 100% de los alelos presentes (*Coverage*, CV=1), exhibiendo similar diversidad genética de *Nei* (DG=0.40), heterocigosidad promedio esperada (HE=0.40) y observada (HO=0.02) y coeficiente de endogamia (F=0.96) que el PPMG (Tabla 1). Solo las 2 CN seleccionadas con el objetivo E-NE, que maximiza la diversidad pero no la representatividad, reportaron diferencias significativas ($p<0.05$) con el PPMG respecto de su estructura poblacional (Tabla 1). Se observa que la CN seleccionada con el objetivo E-NE no posee accesiones para 2 de las 12 subpoblaciones previamente descriptas y la CN seleccionada con el objetivo E-NE+F posee diferencias en la proporción de accesiones presentes en las subpoblaciones respecto del PPMG.

Tabla 1. Índices de diversidad genética y estructura poblacional de las CN respecto del PPMG

	PPMG	A-NE	A-NE+F	E-NE	E-NE+F	E-E	E-E+F	HE
DG	0.40	0.40	0.40	0.40	0.40	0.41	0.41	0.41
HE	0.40	0.40	0.40	0.40	0.40	0.41	0.41	0.41
HO	0.02	0.02	0.02	0.01	0.02	0.01	0.01	0.02
F	0.96	0.95	0.95	0.97	0.96	0.98	0.97	0.96
X^2	-	17.50	7.38	45.36	22.74	17.31	15.46	14.17
p	-	0.13	0.83	2.00E-04*	0.03*	0.14	0.22	0.29

Diversidad genética de *Nei* (DG), heterocigosidad esperada (HE), heterocigosidad observada (HO), coeficiente de endogamia (F).

*Diferencia estadísticamente significativa ($p<0.05$) usando la prueba de bondad de ajuste Chi-cuadrado en una Simulación de *Monte Carlo* de 10000 iteraciones entre las CN y el PPMG.

En la Figura 1 (A-C) se ejemplifica la conservación de la diversidad genética y representatividad de las 12 subpoblaciones en las CN A-NE y A-NE+F. Respecto de la

distribución de los fenotipos (AP, AE y DAF) observados en las distintas CN, las seleccionadas con los objetivos E-E y HE se diferenciaron del PPMG (Fig. 1 D-F).

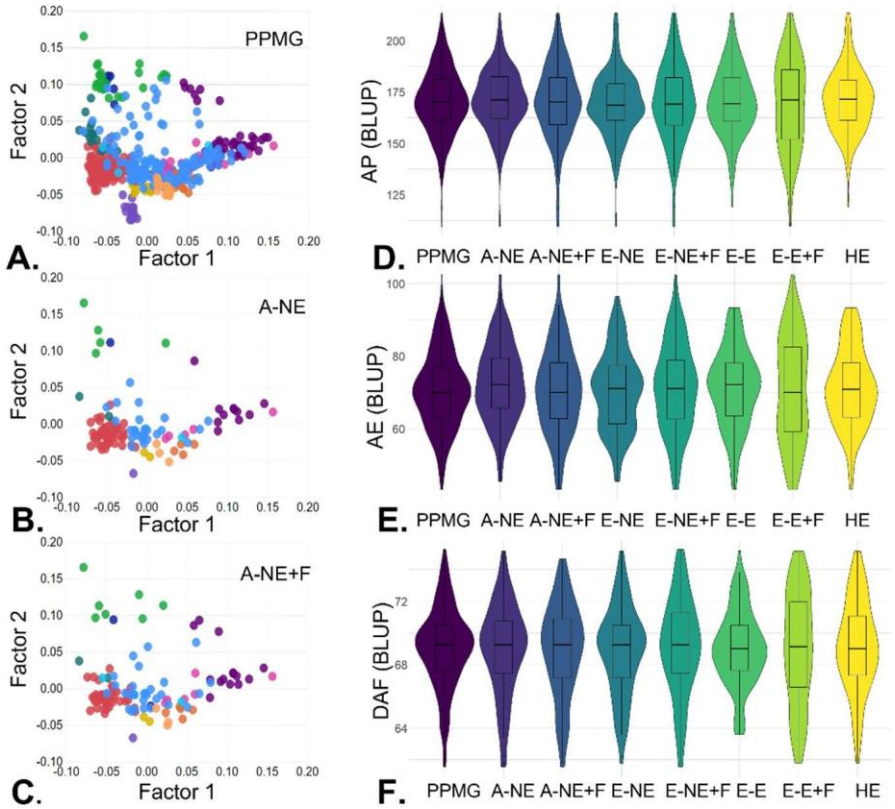

Fig. 1. Diversidad genética de las CN seleccionadas con el objetivo A-NE y distribución de los fenotipos en las 7 CN respecto del PPMG. A-C. Escalado multidimensional (EMD) lineal que ilustra la representatividad de las líneas seleccionadas en las CN A-NE y A-NE+F. Los distintos colores representan la pertenencia a cada una de las 12 subpoblaciones descriptas en el PPMG. D-F. Distribución de los fenotipos (BLUP) de altura de planta (AP), altura de inserción de espiga (AE) y días a floración (DAF) en las CN seleccionadas con distintos objetivos.

4 Conclusiones

La utilización de *Core Hunter 3* permitió seleccionar 7 CN que retuvieron la diversidad alélica presente en el PPMG, manteniendo en gran medida su estructura poblacional y variabilidad fenotípica. Siendo las CN seleccionadas con el objetivo A-NE, las que mejor maximizan la diversidad genética y mantienen la representatividad poblacional del PPMG. Estas CN podrían ser utilizadas como CN multipropósito constituyendo un punto de partida de futuras investigaciones que contribuirán a dar rápida respuesta a problemas que afecten el cultivo de maíz como consecuencia del cambio climático.

Referencias

Alvarado, G., Rodríguez, F. M., Pacheco, A., Burgueño, J., Crossa, J., Vargas, M., Pérez-Rodríguez, P., y Lopez-Cruz, M. A. (2020). META-R: A software to analyze data from multi-environment plant breeding trials. The Crop Journal, 8, 745-756.

Brown, A. H. D. (1989). Core collection: a practical approach to genetic resources management. Genome, 31, 818-824.

De Beukelaer, H., Davenport, G. F., y Fack, V. (2018). Core Hunter 3: flexible core subset selection. BMC Bioinformatics, 19(1), 203.

Frankel, O. H. (1984). Genetic Perspectives of Germplasm Conservation. En Arber, W.K., Llimensee, K., Peacock, W.J., y Stralinger, P. (Eds.), Genetic Manipulation: Impact on Man and Society (pp. 161-170). Cambridge University Press, Cambridge.

Gezan, S., de Oliveira, A. A., Galli, G., y Murray, D. (2024). ASRgenomics: An R package with complementary genomic functions. VSN International. Hemel Hempstead, United Kingdom.

Granato, I. S. C., Galli, G., y de Oliveira Couto, E. G. (2018). snpReady: a tool to assist breeders in genomic analysis. Molecular Breeding, 38, 102.

Odong, T. L., Jansen, J., van Eeuwijk, F. A., y van Hintum, T. J. (2013). Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *TAG*. Theoretical and applied genetics. 126(2), 289–305.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Soleimani, B., Lehnert, H., Keilwagen, J., Plieske, J., Ordon, F., Naseri Rad, S., Ganal, M., Beier, S., y Perovic, D. (2020). Comparison Between Core Set Selection Methods Using Different Illumina Marker Platforms: A Case Study of Assessment of Diversity in Wheat. Frontiers in Plant Science, 11, 1040.