Ensamblaje del genoma de la variedad CC 01-1940, un híbrido de caña de azúcar de gran productividad en la industria Colombiana.

Jhon Henry Trujillo Montenegro[1,2], María Juliana Rodríguez Cubillos[3], Cristian Darío Loaiza[1], Manuel Quintero[1], Héctor Fabio Espitia-Navarro[1], Fredy Antonio Salazar Villareal[1], Carlos Arturo Viveros Valens[1], Andrés Fernando González Barrios[3], José De Vega[4], Jorge Duitama[5] and John Riascos[1*]

[1]Centro de Investigación de la Caña de Azúcar de Colombia, CENICAÑA, Cali, Colombia.

[2] Research Group in Bioinformatics, Department of Computer Science, Faculty of Engineering, Universidad Del Valle, Cali, Colombia.

[3]Grupo de diseño de productos y procesos. Department of Chemical and Food Engineering. Faculty of Engineering. Universidad de los Andes, Bogotá, Colombia.

[4] Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK.

[5]Systems and Computing Engineering Department, Universidad de los Andes, Bogotá, Colombia.

Email Contacto: jhtrujillo@cenicana.org

La caña de azúcar es uno de los cultivos económicos más importantes del mundo, sin embargo, debido a la complejidad en su genoma se había limitado la investigación genética en este cultivo. Las principales dificultades radican en un alto nivel de ploidia, la presencia aneuploidias (x = 10–13, 2n = 100–130), altos niveles de heterocigosidad, alto contenido repetitivo y un tamaño del genoma cerca de 10 Gbp. No obstante, gracias a los avances en las tecnologías de secuenciación es posible llevar a cabo el ensamblaje de este tipo de genomas. Contar con un genoma de referencia permitirá al centro de Investigación de la Caña de Azúcar de Colombia, CENICAÑA, avanzar en la implementación de mejores estrategias de mejoramiento genético apoyadas en el uso de marcadores moleculares. Para lograr esto, CENICAÑA ha generado el primer ensamblaje del genoma completo monoploide a partir de un híbrido colombiano de caña de azúcar perteneciente a la variedad CC 01-1940. Para este ensamblaje se empleó un total de 100.5 Gbp de lecturas largas PacBio, 107.2 Gbp de lecturas cortas pareadas de Illumina, y 102.87 Gbp de lecturas Hi-C. Inicialmente, se generó un ensamblaje base en términos de contigs utilizando las lecturas PacBio y la herramienta Flye-Assembler. Posteriormente, este ensamblaje fue mejorado utilizando los datos HI-C y la herramienta ALLHIC, para generar un ensamblaje en términos de pseudo-cromosomas. Las lecturas cortas de Illumina fueron empleadas para corregir errores de ensamblaje por medio de la herramienta NGSEP. La caracterización del contenido genético se llevó a cabo utilizando las metodologías MAKER y Tuxedo, mientras que la caracterización del contenido repetitivo se realizó utilizando la herramienta RepeatMasker

v4.1.1. Siguiendo esta metodología, se generó un ensamblaje monoploide de la variedad CC 01-1940 que se caracteriza por tener un total de 10 pseudo-cromosomas, 44 scaffolds, 35,035 contigs, un total de 903.1 Mbp ensamblado y un N50 de 34.94 Mbp. Considerando únicamente los 10 pseudo-cromosomas, el N50 alcanza a 55.79 Mbp y la longitud de ensamblaje resuelto se reduce a 498,3 Mbp. En cuanto a su caracterización genética se identificaron un total de 63,724 modelos de genes y un total de 1,442,953 de secuencias repetitivas. Este trabajo describe los esfuerzos desarrollados por CENICAÑA hacia un ensamblaje a nivel cromosómico del genoma altamente complejo de la caña de azúcar. Mejoramos el genoma más reciente disponible para el híbrido R570, el cual solo incluye las regiones ricas en genes. Igualmente, se logró utilizar este ensamblaje como genoma de referencia logrando una tasa de mapeo dos veces mayor que con otros genomas de referencias de caña de azúcar, lo cual se traduce en mejoras para los análisis de población y mapeo de diferentes características de interés.