Ensamblaje y anotación del genoma a escala cromosómica del cultivar tetraploide de papa Diacol Capiro adaptado a la región andina

Paula H. Reyes-Herrera[1,*], Diego A. Delgadillo-Duran[1], Mirella Flores-Gonzalez[2], Lukas A. Mueller[2], Marco A. Cristancho[3,+] and Luz Stella Barrero[1,*]

[1]Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), Bogotá, Colombia

[2]Boyce Thompson Institute, Ithaca, NY 14850, USA

[3]Vicerrectoría de Investigación y Creación, Universidad de los Andes, Bogotá, Colombia

[+]Dirección Actual: Centro Nacional de Investigaciones de Café (Cenicafé), Manizales, Colombia

*Autores de correspondencia: phreyes@agrosavia.co, lbarrero@agrosavia.co

Resumen

La papa (Solanum tuberosum) es un cultivo esencial para la seguridad alimentaria y está clasificada como el tercer cultivo más importante del mundo para el consumo humano. El cultivar Diacol Capiro ocupa la posición dominante en el cultivo colombiano, destinado principalmente a la industria de procesamiento de alimentos. Este cultivar autotetraploide altamente heterocigoto pertenece al grupo Andigenum y destaca por su adaptación a una amplia variedad de entornos que abarcan altitudes de 1800 a 3200 metros sobre el nivel del mar. Aquí se presenta un ensamblaje a escala cromosómica, denominado DC, para este cultivar. El ensamblaje se generó combinando secuenciación de consenso circular con ligadura de proximidad Hi-C para el andamiaje y representa 2,369 Gb con 48 pseudo-cromosomas que cubren 2,091 Gb y una tasa de anclaje del 88.26%. Las métricas del genoma de referencia, que incluyen un N50 de 50.5 Mb, una puntuación BUSCO del 99.38% y una puntuación del índice de ensamblaje LTR de 13.53, evidencian la alta calidad de ensamblaje alcanzada. Una anotación exhaustiva arrojó un total de 154.114 genes, y la puntuación BUSCO asociada del 95.8% para las secuencias anotadas certifica su completitud. El número de genes NLR (del inglés Nucleotide-Binding and Leucine-Rich-Repeat genes) predichos fue de 2107, con una gran representación de dominios que contienen NBARC (dominio de unión a nucleótidos compartido por Apaf-1, ciertos productos del gen R y CED-4) (99.85%). Un análisis comparativo del ensamblaje basado en la anotación con genomas de papa conocidos de alta calidad mostró métricas similares con diferencias en el número total de genes relacionadas con la ploidía. El ensamblaje genómico y la anotación de DC presentados en este estudio representan un activo valioso para la comprensión de la genética de la papa. Este recurso ayudará en las iniciativas de mejoramiento dirigidas y podrá contribuir a la creación de variedades de papa mejoradas, resistentes y más productivas, particularmente beneficiosas para los países de América Latina.

Palabras clave: ensamblaje del genoma; Solanum tuberosum; grupo Andigenum; ensamblaje tetraploide.

Introducción

La papa (Solanum tuberosum L.) es el cultivo, no cereal, más importante del mundo, con una producción mundial de 462 millones de toneladas en 2019. En Colombia es un alimento básico crucial para garantizar la seguridad alimentaria y representa la principal fuente de ingresos para aproximadamente 100.000 familias campesinas (Manrique-Carpintero et al. 2023; González-Orozco et al. 2023). El alto consumo y valor nutricional de la papa como cultivo de seguridad alimentaria ha sido ampliamente reconocido (Devaux et al. 2020). El mejoramiento genético de este cultivo ha surgido principalmente de dos grupos genéticos, el grupo Andigenum de tierras altas con variedades locales andinas y el grupo Chilotanum de tierras bajas con variedades locales chilenas, siendo el grupo Andigenum el más ampliamente cultivado (Spooner et al. 2007; Gavrilenko et al. 2013). Las papas "Andigenas" son las más importantes dentro del grupo Andigenum. Están adaptadas a la tuberización en días cortos, son autotetraploides (2n = 4x = 48) con herencia tetrasómica, y altamente heterocigóticas (Spooner et al. 2007).

Dentro del grupo Andigenum, el cultivar mejorado Diacol Capiro (también conocido como R12; registro del Instituto Colombiano Agropecuario (ICA) No. PAP-68-02), liberado en 1968, es uno de los más cultivados en Colombia (Romero et al. 2017; SIPSA 2013). Esta variedad se destaca por su excelente calidad para la industria de hojuelas y bastones de papa y también tiene buena calidad culinaria (Porras Rodríguez and Herrera Heredia 2015). Además, se cultiva desde hace varios años y se adapta a una gran variedad de regiones desde los 1800 hasta los 3200 metros sobre el nivel del mar en Colombia y en Ecuador, donde se distribuye en las zonas norte y central (Torres et al. 2011; Porras Rodríguez and Herrera Heredia 2015). Este cultivar de tubérculo de piel roja y pulpa crema (Figura 1) es tolerante a la marchitez bacteriana (causada por *Ralstonia solanacearum*) y a algunos virus; aunque es susceptible al tizón tardío (causado por *Phytophthora infestans*), a la sarna pulverulenta (causada por *Spongospora subterranea*) así como a otros estreses bióticos y abióticos (Andrade-Piedra and Torres 2011; Porras Rodríguez and Herrera Heredia 2015; Romero et al. 2017). Por lo tanto, la secuencia de su genoma puede contribuir a arrojar luz sobre los mecanismos moleculares para mejorar la tolerancia a estos rasgos.

El conocimiento del genoma de la papa abre grandes oportunidades para proseguir los esfuerzos de conservación y mejora genética. La papa ha sido un cultivo de gran interés para los investigadores de todo el mundo, que se esfuerzan por comprender mejor la complejidad de su genoma con el fin de liberar el potencial de las nuevas técnicas de mejoramiento (Ghislain and Douches 2020). Desde la publicación del primer genoma de la papa, el del doble monoploide DM1-3 516 R44 (PGSC 2011), se han publicado varios genomas más, incluidas la secuencia genómica de especies silvestres como *Solanum commersonii* (Aversano et al. 2015), *Solanum stenotomum* (Yan et al. 2021) y el clon endogámico M6 de *Solanum chacoense* (Leisner et al. 2018). Además, se publicaron los genomas de varias variedades autóctonas peruanas y chilenas, entre ellas dos del grupo Andigenum (Kyriakidou et al. 2020a,b), la papa diploide heterocigótica *S. tuberosum* del grupo Tuberosum RH89-039-16 utilizando una combinación de múltiples estrategias de secuenciación (Zhou et al. 2020), y una versión actualizada del DM1-3 516 R44 (ensamblaje DM v6.1) utilizando lecturas largas de Oxford Nanopore Technologies junto con un andamiaje de proximidad por ligamiento (Hi-C) para obtener un ensamblaje a escala cromosómica (Pham et al. 2020). Recientemente se ha secuenciado un genoma a escala cromosómica resuelto por haplotipos del cultivar

autotetraploide Otava, del grupo Tuberosum (SASA 2010), a partir de núcleos de polen utilizados para el agrupamiento de gametos (Sun et al. 2022; Campoy et al. 2020). Además, Hoopes et al. (2022) presentaron ensamblajes tetraploides de seis cultivares de papa de Norteamérica y Europa. Además, Bao et al. (2022) presentó un ensamblaje genómico resuelto por haplotipos de la papa tetraploide C88, cultivada en países asiáticos. Para mejorar nuestro conocimiento del genoma de la papa es muy deseable disponer de más genomas ensamblados procedentes de diferentes acervos genéticos mediante nuevas tecnologías de secuenciación y algoritmos de ensamblaje de genomas. Estas secuencias genómicas son fundamentales para el enriquecimiento del pan-genoma de la papa tetraploide, definido por primera vez por Hoopes et al. (2022). Esto se considera una tarea compleja debido a la extraordinaria diversidad del germoplasma de la papa (Ghislain and Douches 2020).

El presente estudio tiene como objetivo conseguir un ensamblaje del genoma a escala cromosómica en 48 pseudo-cromosomas para el cv. Diacol Capiro (ensamblaje DC) mediante la combinación de lecturas largas de PacBio Circular Consensus Sequencing (CCS) y datos de andamiaje Hi-C. El ensamblaje DC es el primer ensamblaje a escala cromosómica de una variedad del grupo Andigenum cultivada en la región andina de Colombia y Ecuador. La calidad del ensamblaje del genoma se comparó con los ensamblajes a escala cromosómica de la papa disponibles actualmente disponibles, que incluyen el monoploide DM v6.1 (Pham et al. 2020) y el haplotipo del genoma diploide RH39-039-16 (Zhou et al. 2020). Además, la calidad del ensamblaje DC se comparó con la de dos ensamblajes del grupo Andigenum ADG1-CIP 700921 y ADG2-CIP 702853 (Kyriakidou et al. 2020b). Además, se logró la anotación del ensamblaje DC y se comparó con otras seis anotaciones de patata de alta calidad disponibles públicamente en SpudDB. También se realizaron análisis adicionales del contenido de genes de unión a nucleótidos y de repetición rica en leucina (NLR) y de la distribución de dominios.

Materiales y métodos

Material vegetal

Los tubérculos de Diacol Capiro (Figura 1) se sembraron bajo condiciones de invernadero en el Centro de Investigación Tibaitata de la Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA) utilizando tierra solarizada mezclada con turba en una proporción de 3:1 en macetas de 20 kilogramos con una temperatura que osciló entre 19 y 27°C. El material fue suministrado por un programa de extensión de AGROSAVIA, conocido como El Plan de Semillas que proporciona materiales a los agricultores. Los entrenudos con meristemas laterales obtenidos de tallos de una planta se introdujeron en condiciones *in vitro*. Éstos se sembraron independientemente en medio Murashige y Skoog (MS) (Murashige and Skoog 1962) suplementado con sales basales MS, vitaminas, 3% de sacarosa y 0,7% de fitoagar y se colocaron en cámara de crecimiento a 16-20°C y fotoperiodo de 6-h-luz/8-h-oscuridad durante cuatro semanas. A partir de los brotes obtenidos, se sembraron diferentes explantes en tubos independientes y se colocaron en cámara de crecimiento durante seis semanas. Cincuenta y cinco plántulas in vitro se endurecieron durante 20 días y luego se sembraron en una mezcla de escoria de carbón: cáscara quemada en proporción 1:1 con fertilización-irrigación semanal en condiciones de invernadero.

Extracción de ADN y preparación de librerías para secuenciación

Antes del inicio de la floración, unas ocho semanas después del endurecimiento, se recogieron 100 gramos de tejido de hojas jóvenes que aún no habían extendido completamente los foliolos y se congelaron inmediatamente en nitrógeno líquido, almacenándose después a -80°C. El tejido se envió al Arizona Genomics Institute (Tucson, Arizona, EE.UU.) para la extracción de ADN genómico de alto peso molecular (HMW) y la construcción de bibliotecas HiFi para la secuenciación PacBio CCS de tres líneas CCS. Además, un gramo del mismo tejido foliar se sometió a la construcción de bibliotecas de captura de conformación de cromatina de alto rendimiento (Hi-C) utilizando el Proximo™Hi-C Plant Kit (Phase Genomics) y se secuenció en el Illumina HiSeq4000 bajo el modo paired-end 150 bp por Phase Genomics (Seattle,Washington, US). Además, se utilizaron 100 mg para la extracción de ADN con el DNeasy Plant Mini Kit (QIAGEN) seguida de la preparación de bibliotecas con el Nextera™ DNA Flex Kit en formato Dual Index (Illumina) que se enviaron para secuenciación paired-end en el sistema Illumina HiSeq4000 a Macrogen (Seúl, Corea del Sur). Los datos HiFi consistieron en 5,3 M de lecturas, N50 de 16.754 bases (b), y un total de 89 Gb secuenciados. El conjunto de datos de Illumina consistía en 1,49 G de lecturas paired-end con un total de 226 Gb; estas lecturas no se utilizaron en el ensamblaje, sólo se utilizaron para la validación técnica.

Estimación del tamaño, la ploidía y la heterocigosidad

Las estimaciones para el genoma DC se obtuvieron utilizando los conjuntos de datos HiFi e Illumina. Para ambos conjuntos de datos, se utilizó KMC3 para obtener estadísticas de k-meros (Kokot et al. 2017), mientras que GenomeScope 2.0 y Smudgeplot se utilizaron para obtener el perfil del genoma (Ranallo-Benavidez et al. 2020).

Ensamblaje del genoma de novo

Las lecturas HiFi se introdujeron en varios ensambladores *de novo*: HiCanu (Nurk et al. 2020), Hifiasm (Cheng et al. 2021) e IPA de PacBio (Biosciences 2020), ya que estos algoritmos están diseñados para ensamblar lecturas HiFi largas. Además, se utilizó el ensamblador Flye (Kolmogorov et al. 2019) porque tiene resultados notables para el ensamblaje de genomas utilizando lecturas largas (Reyes-Herrera et al. 2023). El mejor ensamblaje se obtuvo con HiCanu en base a la puntuación BUSCO 5.1.0 (Manni et al. 2021) (utilizando los linajes embryophyta_odb10 y solanales_odb10) así como las métricas: N50, L50, longitud total, contig más grande y número de contigs totales. Estas métricas fueron obtenidas por QUAST v5.0.2 (Gurevich et al. 2013) (Ver Tabla 1).

Las lecturas HiFi e Illumina se utilizaron por separado para estimar el tamaño del genoma, la heterocigosidad y para confirmar la ploidía (Figuras suplementarias S1, S2 y S3). A partir de las lecturas HiFi, se estimó un tamaño del genoma haploide de 679 Mb utilizando GenomeScope (Ranallo-Benavidez et al. 2020). Los parámetros utilizados fueron longitud k-mer = 21 y ploidía = 4, lo que corresponde a una longitud total de 2.716 Mb. La longitud de ensamblaje HiCanu de 2.456 Mb fue la más cercana a la longitud de ensamblaje tetraploide esperada (ver Tabla 1).

Para el proceso de andamiaje, se utilizó HiCExplorer (Ramírez et al. 2018) para filtrar los datos Hi-C. Como resultado, el 11,08 % del conjunto de datos inicial eran contactos válidos. A pesar de esto, el 30% de las lecturas no se mapearon de forma única debido a la poliploidía y las regiones repetitivas que fueron filtradas por HiCExplorer. Sin embargo, más del 40% de las lecturas emparejadas se encontraban a más de 10 Kb de distancia y

se confirmó que correspondían a uniones Hi-C reales. Por lo tanto, se utilizaron dos conjuntos de datos: los contactos válidos y las lecturas no mapeadas de forma única (en lo sucesivo denominadas Hi-C crudas).

A continuación, utilizamos ambos conjuntos de datos y el flujo de trabajo ALLHiC para el andamiaje porque era adecuado para un genoma autopoliploide (Zhang et al. 2018, 2019). El flujo de trabajo ALLHiC implicó cinco pasos: poda, partición, rescate, optimización y construcción. En el paso de poda, ALLHiC identificó un conjunto de contigs alélicos basados en la sintenia con los genomas ensamblados a escala cromosómica de referencia DM v6.1 (Pham et al. 2020) y RH39-039-16 (Zhou et al. 2020). Además, ALLHiC proporcionó un corrector para detectar y corregir contigs mal unidos utilizando señales Hi-C. Se generaron ocho ensamblajes variando los datos de entrada y los parámetros que incluían: los datos Hi-C utilizados para la corrección y el andamiaje, el ensamblaje inicial del genoma y los genomas de referencia en el paso de poda (véase la Tabla 2). Los contactos válidos filtrados y las lecturas Hi-C crudas se utilizaron como entrada para los datos Hi-C. En la Tabla 2 se detallan los datos utilizados para las estrategias de andamiaje.

La mayoría de los ensamblajes resultantes tenían 12 grupos homólogos con conjuntos de grandes scaffolds más un conjunto de contigs adicionales. Para cada grupo homólogo, se trazaron las diferencias de tamaño de los cuatro scaffolds más grandes (Figura 2). Ninguno de los ensamblajes mostró una clara superioridad sobre los demás, por lo que se exploró una estrategia híbrida. Se seleccionaron tres estrategias de ensamblaje (st2, st3 y st6), que dieron como resultados conjuntos de cuatro scaffolds con tamaños similares en la mayoría de los cromosomas. Esto se reflejó en la menor desviación estándar del tamaño de los andamiajes con puntuaciones BUSCO iguales (Tabla suplementaria S1).

A continuación, se examinaron las variaciones entre las estrategias de ensamblaje seleccionadas (st2, st3 y st6) mediante dotplots respecto a la referencia DMv6.1 (Figuras suplementarias S4, S5 y S6). Sin embargo, ninguna estrategia ha resultado ser superior a las demás, como indican la tasa de mapeo y las estadísticas QUAST descritas en la Tabla 3. Los pasos de ensamblaje y andamiaje no son determinísticos; cada estrategia produce un resultado diferente. Para lograr el mejor ensamblaje posible del genoma, estos resultados se evaluaron para construir un ensamblaje híbrido, que comprende el mejor conjunto posible de pseudo-cromosomas para cada grupo homólogo. Para ello, se compararon los cuatro scaffolds más grandes de las tres estrategias para cada cromosoma, y se obtuvo una configuración híbrida (combinando los scaffolds de st2, st3 y st6) (véase el pipeline de la Figura 3). Se utilizó el software Minimap2 (Li 2018) para alinear los scaffolds entre las tres estrategias (st2, st3 y st6) para cada cromosoma y así poder identificar scaffolds similares (paso ii, Figura 3). A continuación, se construyeron conjuntos híbridos utilizando un scaffold seleccionado entre los grupos similares y el grupo diferente (pasos iii y iv, Figura 3). El mejor conjunto se seleccionó en función de tres criterios (véanse los pasos v y vi, Figura 3) (1) la puntuación BUSCO para embryophyta_odb10 (como se muestra en las Figuras suplementarias S7-S18); (2) la mediana de la tasa de alineación con DM v6.1 (como se muestra en la Figura suplementaria S19); y (3) los dotplots a DM v6.1 (como se muestra en las Figuras suplementarias S4, S5 y S6). El procedimiento para construir cromosomas híbridos fue exitoso para los cromosomas 1, 3, 4 y 7. Para todos los demás cromosomas, los scaffolds de una sola de las estrategias (st2, st3 o st6) fue la mejor configuración siguiendo los tres criterios (véanse los detalles de la selección en la tabla suplementaria S2). La Tabla 3 muestra los resultados de la comparación de las tres estrategias de ensamblaje seleccionadas (st2, st3, st6) y el ensamblaje resultante denominado DC.

Tras el andamiaje con ALLHIC, se detectó una diferencia de aproximadamente 370 Mbp de longitud entre el ensamblaje Canu inicial y el ensamblaje DC. Era de esperar una diferencia de longitud atribuida al andamiaje. Sin embargo, ALLHIC no incluyó los contigs con señales Hi-C débiles, y aunque el pipeline tenía un paso de rescate, el 15% de los contigs seguían sin recuperarse. Los contigs perdidos se recuperaron en el andamiaje ALLHIC y se comprobó su alineación con el ensamblaje DC. La cobertura de un contig se definió como la parte del contig alineada dividida por la longitud del contig. Los contigs recuperados que tenían una cobertura de contig inferior a la mediana de la distribución de cobertura de contig se añadieron al ensamblaje. Además, se utilizó Juicebox (Durand et al. 2016) para comprobar y fijar manualmente el mapa de calor correspondiente para cada cromosoma.

Detección de contaminantes

El flujo de trabajo Blobtools v2 (Challis et al. 2020) se utilizó para detectar contaminantes y realizar una evaluación de la calidad del ensamblaje DC. Este proceso utiliza la cobertura del ensamblaje, así como los resultados de BUSCO, Blast y DIAMOND v2.0.11 (Buchfink et al. 2021a). Se identificaron dos contigs como Proteobacteria y se eliminaron del ensamblaje.

Evaluación del ensamblaje DC

En primer lugar, se mapeó el conjunto de datos de Illumina al ensamblaje para evaluar la integridad y la precisión utilizando BWA-MEM v0.7.12 (Li 2013) y se utilizó SAMtools v1.8 (Li et al. 2009) para obtener estadísticas de alineación. En segundo lugar, se utilizó BUSCO v5.1.0 (Manni et al. 2021) para evaluar la completitud del espacio génico para dos linajes: el embryophyta_odb10 y el solanales_odb10. En tercer lugar, se utilizó el índice de ensamblaje de repeticiones terminales largas (LTR) (LAI) para evaluar la continuidad del ensamblaje utilizando LTRharvest v1.6.1 (Ellinghaus et al. 2008), LTR_FINDER_parallel v1.1, LTR_retriever v2.9.0 (Ou y Jiang 2017) y LAI (Ou et al. 2018). En cuarto lugar, se utilizó Meryl para contar los k-mers y Merqury (Rhie et al. 2020) para realizar una evaluación sin referencias. Este enfoque consistió en comparar el ensamblaje con lecturas sin ensamblar basadas en k-meros. En quinto lugar, el ensamblaje de DC se comparó con los ensamblajes del grupo Andigenum, ADG1 y ADG2 (Kyriakidou et al. 2020b). En sexto lugar, la calidad del ensamblaje del genoma se comparó con dos ensamblajes a escala cromosómica actualmente disponibles para la papa: (1) el ensamblaje monoploide doble DMv6.1 del grupo Phureja de *S. tuberosum* que corresponde a una versión actualizada (Pham et al. 2020) de la primera versión del genoma de la papa (PGSC 2011) y, (2) un ensamblaje resuelto por haplotipos para una papa diploide del grupo Tuberosum (Zhou et al. 2020).

Anotación

Se construyó una librería de repeticiones utilizando RepeatModeler v1.0.8 (Smit y Hubley 2008). La librería de repeticiones se examinó con ProtExcluder v1.1 (Campbell et al. 2014) para excluir las proteínas existentes. La librería de repeticiones obtenida se utilizó para enmascarar el genoma tetraploide de la papa. Se predijeron genes codificantes de proteínas en el genoma enmascarado de la papa. Se utilizó el pipeline de anotación Maker v3.01.02 (Cantarel et al. 2008) para producir modelos génicos de consenso. Diferentes recursos como la predicción génica ab-initio,

la búsqueda de homología y el ARN-seq se proporcionaron en un proceso iterativo a Maker. Las proteínas revisadas de Viridiplantae de Uniprot se utilizaron como evidencia de homología junto con las proteínas DMv6.1 (PGSC 2011) y las proteínas de tomate ITAG4.0 (Hosmani et al. 2019). Los datos de ARN-seq para disponibles públicamente en el NCBI a través del SRA (Tabla suplementaria S3) de diferentes etapas de la vida se mapearon al genoma usando HiSAT2 v2.1.0 (Kim et al. 2015) y se proporcionaron como evidencia. Hubo una limitación que dificultó la identificación de genes específicos exclusivos del cultivar DC, debido a la ausencia de datos de RNA-Seq específicos de este cultivar. No obstante, la mayoría de las muestras de ARN-seq analizadas correspondían al grupo Solanum tuberosum Andigenum (Bozan et al. 2023; Ponce et al. 2022; Lin et al. 2015), al que pertenece el cultivar DC.

Además, la predicción génica *ab-initio* se obtuvo de Augustus (3.4.0) (Stanke et al. 2008) a través del pipeline BRAKER v2.1.4 (Hoff et al. 2019). El conjunto final de modelos de genes de consenso resultó de tres rondas de Maker utilizando entradas anteriores. BUSCO se utilizó para evaluar la integridad de la predicción de genes. La anotación funcional de las secuencias proteicas predichas se logró utilizando el software Eggnog (Cantalapiedra et al. 2021; Buchfink et al. 2021b) que alineó las secuencias proteicas contra bases de datos públicas, incluyendo SwissProt, TrEMBLE y KEGG, y datos de ortología eggNOG (Huerta-Cepas et al. 2019). Se obtuvieron valores adicionales para las bases de datos InterPro y Gene Ontology (GO) mediante InterProScan (Quevillon et al. 2005).

Se empleó la herramienta NLR-Annotator (Steuernagel et al. 2020) para identificar los dominios candidatos de unión a nucleótidos y de repetición de ricos en leucina (NLR) candidatos. Esta herramienta ejecuta un análisis *in silico* de las regiones anotadas del ensamblaje DC. El mismo enfoque se utilizó para analizar las anotaciones de seis ensamblajes disponibles en la base de datos SpudDB (http://spuddb.uga.edu/index.shtml). Estos incluyen tres ensamblajes tetraploides: Atlantic (Atlantic_v3 annotation) (Hoopes et al. 2022), Cooperation-88 (C88.v1 annotation) (Bao et al. 2022), y Otava (Otava.v1 annotation) (Sun et al. 2022). Además, se incluyeron un ensamblaje diploide RH89-039-16 (Zhou et al. 2020) y dos ensamblajes monoploides DMv6.1 (Pham et al. 2020) y M6 (anotación M6_v5) (Leisner et al. 2018).

Resultados y discusión

Estimación del tamaño, la ploidía y la heterocigosidad

Se utilizaron conjuntos de datos HiFi e Illumina para estimar el tamaño, la ploidía y la heterocigosidad en el genoma de DC. Para el conjunto de datos HiFi se obtuvo una cobertura estimada de 100x, mientras que para el conjunto de datos Illumina se obtuvo una cobertura estimada de más de 300x. La longitud haploide estimada del genoma fue de 682 Mb y 679 Mb para los conjuntos de datos de Illumina y HiFi, respectivamente (una diferencia del 0,43% entre ambas longitudes estimadas, véanse las Figuras suplementarias S1-S2). Considerando un tamaño del genoma de 679 Mb para el genoma haploide y de 2,716 Gb para el genoma tetraploide, el ensamblaje DC cubrió el 87,5 % de la longitud estimada. La estimación de la longitud de repetición para cada genoma haploide osciló entre 296 y 300 Mb para el conjunto de datos de Illumina. Smudgeplot (Ranallo-Benavidez et al. 2020) confirmó que el ensamblaje DC procedía de un genoma tetraploide para ambos conjuntos de datos (véase la Figura suplementaria S3). La heterocigosidad estimada osciló entre el 4,13% y el 6,78% para el conjunto de datos Illumina y entre el 3,88% y el 5,82% para el conjunto de datos HiFi. Esto cae dentro de un rango similar a los ensamblajes de Andigenum ADG1 (3,52%) y ADG2 (7,75%) de los que se informó anteriormente (Kyriakidou et al. 2020b).

Evaluación del ensamblaje DC

El ensamblaje final de DC tenía una longitud total de 2,369,577,969 bases (Tablas 3 y 4), con un 88.26% de estas bases ancladas a 48 pseudo-cromosomas (véase el dotplot del ensamblaje en la Figura 4). El ensamblaje tetraploide tiene 1194 contigs (99.9% de los contigs más largos de 10 Kb). El 98.58% de las lecturas de illumina se alinearon y emparejaron correctamente con el ensamblaje DC. El linaje embryophyta_odb10 comprendía 1614 ortólogos de BUSCO, de los cuales 1604 estaban completos en el ensamblaje DC (lo que representa el 99.38%) (Tabla 4). Estos contenían 186 copias únicas y 1418 copias duplicadas. Los restantes ortólogos de este linaje eran cinco fragmentados y cinco ausentes. El linaje solanales_odb10 comprendía 5950 ortólogos BUSCO, de los cuales 5888 estaban completos en el ensamblaje DC representando el 98.96%, 577 copias únicas y 5311 duplicadas, mientras que cinco de los ortólogos restantes estaban fragmentados y 57 faltaban. El ensamblaje se construyó utilizando una estrategia híbrida, que implicaba la combinación de scaffolds de diferentes ensamblajes para cuatro cromosomas específicos. El objetivo principal era extraer el ensamblaje óptimo posible del conjunto de datos disponible. En particular, el ensamblaje DC mostró un desempeño superior al de los ensamblajes de entrada en varias métricas, incluida la calidad (medida por LAI), N50 y la longitud total. En todas las demás métricas, el desempeño del ensamblaje DC fue similar al de los ensamblajes originales, pero nunca demostró un desempeño inferior (Tabla 3). Aunque la estrategia híbrida resultó eficaz, es importante señalar que para mejorar aún más la calidad del ensamblaje es necesario incluir datos adicionales.

La Figura 5 ilustra cinco círculos concéntricos, cada uno de los cuales proporciona información específica sobre los scaffolds del ensamblaje de DC. El círculo A abarca la representación de los 48 pseudo-cromosomas. El círculo B muestra un diagrama de dispersión que representa el LAI. En particular, el ensamblaje DC obtuvo una puntuación LAI de 13.53, que se ajusta a los genomas de referencia de alta calidad (Tabla 4). Las puntuaciones LAI que oscilan entre 10 y 20 son en realidad indicativas de alta calidad. El ensamblaje muestra predominantemente un LAI que cae dentro del rango 10 < LAI < 20, con picos notables que incluso superan un LAI de 20. El círculo C presenta la alineación de lecturas de CENH3-ChIP-seq (Gong et al. 2012) con el ensamblaje DC, indicativo de centrómeros. Entre los 30 pseudocromosomas, es evidente un pico distinto, que sirve como indicador de los centrómeros. Sin embargo, en los pseudocromosomas restantes, hay una ausencia de dicho pico, lo que podría indicar la presencia de regiones centroméricas distintas. Esto concuerda con las observaciones del ensamblaje del genoma tetraploide C88 (Bao et al. 2022). Los dos últimos círculos (D, E) representan la distribución de genes y LTR (Long Terminal Repeat) dentro de ventanas de 5 Mb. Como se esperaba, hay un notable enriquecimiento de genes más allá de los centrómeros o en proximidad a los telómeros y una correspondiente reducción alrededor de los centrómeros consistente con la densidad génica previamente reportada (Hoopes et al. 2022; Bao et al. 2022). Por el contrario, la distribución

de LTRs parece ser más uniforme. Por último, una evaluación sin referencias utilizando Merqury para el ensamblaje de DC dio como resultado un valor de calidad (QV) de consenso de alta precisión de 65.458, y una integridad de k-mer del 92.39%.

Comparación del ensamblaje del genoma DC con los ensamblajes ADG1, ADG2 del grupo Andigenum y otros ensamblajes tetraploides

La Tabla 4 muestra las estadísticas del genoma para los ensamblajes ADG1, ADG2 y DC. El ensamblaje ADG2 se construyó sólo con lecturas de Illumina, el ensamblaje ADG1 utilizó lecturas de Illumina y PacBio RS II, y el ensamblaje DC de este estudio utilizó PacBio HiFi para construir el ensamblaje inicial junto con datos Hi-C para el andamiaje. Al incluir lecturas largas de PacBio HiFi, el número de contigs se redujo significativamente y el N50 aumentó (Tabla 4). La puntuación BUSCO aumentó a medida que el genoma está menos fragmentado. Para el linaje embryophyta_odb10, la puntuación BUSCO fue del 85.63%, 48.69% y 99.38% para los ensamblajes ADG1, ADG2 y DC, respectivamente. Sin embargo, la puntuación BUSCO completa y duplicada (embryophyta_odb) fue significativamente mayor para el ensamblaje DC, con un 87.79%, lo que indica que recuperó más de un haplotipo, en comparación con el 7.68% de ADG1 y el 5.63% de ADG2. Además, el ensamblaje DC tuvo la mejor puntuación LAI (13.53), que corresponde a genomas de referencia de alta calidad (entre 10 y 20), mientras que el ADG1 y el ADG2 tuvieron puntuaciones LAI de 9.06 y 7.29, que corresponden a genomas de calidad de ensamblaje borrador. La porción repetitiva fue menor en el genoma DC con un 49.92% en comparación con el 60.2% en ADG1 (Kyriakidou et al. 2020b), y el 65-68% de siete ensamblajes de genomas tetraploides (Sun et al. 2022; Hoopes et al. 2022). Las secuencias repetitivas presentes en el ensamblaje del genoma DC condujeron al enmascaramiento del 44% del ensamblaje. La porción repetitiva del ensamblaje de DC estaba más cerca del rango reportado en un análisis comparativo previo de secuencias repetitivas en especies de papa y tomate (Gaiero et al. 2019). Las clases predominantes de elementos transponibles (ET) fueron retrotransposones (37%) y transposones de ADN (5%) (Figura S21). En comparación con seis ensamblajes de genomas tetraploides de alta calidad, se observó que estas dos clases eran los ETs más abundantes, constituyendo el 26% y el 1.48%, respectivamente (Hoopes et al. 2022). Esta observación coincide con los resultados del estudio del superpangenoma de la papa, que incluía 296 diploides y poliploides, en el que se identificaron retrotransposones como la clase de ET predominante (Bozan et al. 2023). Las longitudes o los tamaños de los pares de bases de los LTR, los transposones de ADN y los ARN pequeños en DC se encontraban dentro de un rango comparable a los observados en el superpangenoma (véase la Figura S16 en (Bozan et al. 2023), mientras que las LINE y las repeticiones simples eran comparables a las registradas en los ensamblajes tetraploides mencionados anteriormente (véase la Tabla S9 en (Hoopes et al. 2022)).

Comparación del ensamblaje del genoma DC con los ensamblajes a escala cromosómica DMv6.1 y RH89-039-16

Las estadísticas de ensamblaje de estos genomas se muestran en la Tabla 4. El número de contigs para los ensamblajes fue de 288 para DMv6.1 (un genoma monoploide), 3024 para RH89-039-16 (un genoma diploide) y 1194 para DC (un genoma tetraploide), siendo el número de contigs para el tetraploide aproximadamente cuatro veces mayor que el del monoploide, como era de esperar. El N50 fue de 59 Mb para DMv6.1, 66 Mb para RH89-039-16 y 50 Mb para DC. Sin embargo, en la puntuación BUSCO completa, los tres genomas tienen puntuaciones similares para el linaje embryophyta_odb10 con 99.38% para DMv6.1, 99.32% para RH89-039-16, y 99.38% para DC. La puntuación BUSCO indicaba completitud y duplicación, con un porcentaje de duplicación que aumentaba proporcionalmente con el nivel de ploidía, como se preveía. Así, para el monoploide DMv6.1 la puntuación fue del 1.92%, para el diploide RH89-039-16 la puntuación fue del 68.22% y para el DC fue del 87.79%. Para el linaje solanales_odb, la puntuación BUSCO completa también fue similar con un 98.67%, 98.87% y 98.96% para DMv6.1, RH89-039-16 y DC, respectivamente. El índice LAI que evalúa la calidad del ensamblaje del genoma fue de 6.47 para el RH89-039-16, que correspondía a la calidad del genoma borrador. En cambio, 13.56 y 13.53 para los ensamblajes DMv6.1 y DC, respectivamente, se situaron en la categoría de genoma de referencia. Una versión ampliada de este análisis comparativo, incluyendo dos ensamblajes adicionales Solyntus(van Lieshout et al. 2020) y M6 (Leisner et al. 2018) se encuentra en la Tabla suplementaria S4.

Anotación

Se identificaron un total de 154,114 predicciones génicas, el 86.16% de las cuales estaban ancladas a los 48 pseudo-cromosomas (véase la Figura suplementaria S22 para una distribución detallada por cromosoma). Las secuencias anotadas en el ensamblaje DC tuvieron una puntuación BUSCO del 95.8% para el transcriptoma de linaje embryophyta_odb10 (véase la Tabla 5). El rango de puntuaciones BUSCO para otras seis anotaciones de ensamblaje de alta calidad publicadas se situó entre el 90.1% para M6_v5 (Leisner et al. 2018) y el 99.2% para C88.v1 (Bao et al. 2022). Esto sitúa la anotación DC bien dentro del rango de calidad de los ensamblajes en términos de puntuación BUSCO. Por lo tanto, de estas comparaciones se puede inferir una anotación DC de alta calidad. Esto es digno de mención teniendo en cuenta que los tetraploides previamente anotados derivados de ensamblajes genómicos por fases utilizaron la secuencia genómica acoplada con datos de secuencia de ARN (Hoopes et al. 2022; Bao et al. 2022; Sun et al. 2022), mapeo S1 (Bao et al. 2022), o secuencia de polen de una sola célula (Sun et al. 2022) - todo lo cual facilita el ensamblaje genómico derivado de la anotación de papas auto-tetraploides. El enfoque utilizado en DC aprovecha los datos de anotación de alta calidad disponibles, lo que se traduce en una métrica BUSCO comparable que indica exhaustividad.

En cuanto al número de genes predichos, osciló entre 32,917 para el monoploide DMv6.1 y 269,097 para el tetraploide Atlantic. El número de genes DC fue más similar al del tetraploide C88 (154,114 frente a 150,853). El cultivar C88 tiene un pedigrí del grupo Andigena de *S. tuberosum* como planta paterna y una papa india como fuente materna (Bao et al. 2022), mientras que DC procede del mismo grupo Andigena. Por lo tanto, existen similitudes en el contenido génico entre estas dos variedades: El 95.49% de los genes predichos por C88 se alinearon con la anotación de DC. Además, el 89.13% de los genes predichos de DC se alinearon con la anotación de C88 con un porcentaje de posiciones idénticas superior al 70%. Aunque ambas anotaciones comparten la mayoría de los genes, las diferencias pueden atribuirse a especificidades únicas inherentes a cada variedad. En general, el número de genes aumentó con la ploidía como se preveía, ya que cualquier reducción de la ploidía conduciría a la pérdida de genes (Sun et al. 2022), y la papa tetraploide proporcionaría más copias de genes como copias de seguridad para los alelos defectuosos (Bao et al. 2022).

También se identificaron genes NLR predichos *in silico* en DC y en los otros seis ensamblajes (Tabla 5). El número de NLR candidatos fue de 2107 para DC y de 2186, 4065 y 2937 para los tetraploides Atlantic, C88 y Otava, respectivamente. Las disparidades en los números, en comparación con los comunicados para C88 (2262 genes NLR) en Bao et al. (2022), pueden deberse a la utilización de metodologías distintas en ambos análisis. El C88 utilizó el pipeline RGAugury (Li et al. 2016; Jupe et al. 2012) mientras que el DC utilizó NLR-Annotator (Steuernagel et al. 2020). Además, el M6_v5 mostró una mayor abundancia de NLR (1148) en comparación con los ensamblajes DMv6.1 (566) y RH89-039-16 (378). Así pues, el análisis reveló un mayor número de genes NLR predichos en los cuatro tetraploides (>2000 genes) seguidos de los diploides y los ensamblajes monoploides, lo que indica una reserva genética prometedora para abordar las respuestas inmunitarias innatas.

Los genes de dominio NLR pertenecen a una de las mayores familias multigénicas de las plantas (Ercolano et al. 2022). La clase más grande de genes de resistencia (R) de plantas codifica para un NBARC conservado (para dominio de unión a nucleótidos compartido por Apaf-1, ciertos productos del gen R y CED-4) fusionado a repeticiones ricas en leucina (proteínas NBARC-LRR), a veces en proximidad con otros elementos como dominios TIR (Receptor Toll/Interleucina-1) (Dangl y Jones 2001). La proporción de varias de estas combinaciones de dominios dentro de los genes NLR se visualiza utilizando diferentes colores en la Figura 6. Cuatro combinaciones principales de dominios de genes NLR son consistentemente prominentes en los siete ensamblajes: CC-NBARC (dominio en espiral-NBARC), CC-NBARC-LRR, NBARC y NBARC-LRR. Cada una de estas configuraciones representa más del 7% del total.

La combinación TIR-NBARC-LRR también se observa en los siete ensamblajes. Entre ellos, C88 exhibe la mayor proporción de esta combinación de dominios (16.1%), mientras que DC muestra la menor (0.33%) (Figura 6). La presencia de las otras cuatro combinaciones de dominios en las que interviene TIR (TIR, TIR-CC-NBARC-LRR, TIR-LRR, TIR-NBARC) disminuye en la mayoría de los ensamblajes. La suma de proporciones de estas cuatro combinaciones por ensamblaje oscila entre el 0.66% en DC y el 8.23% en Otava. Así, DC, una variedad del grupo Andígena conocida por ser susceptible al tizón tardío y a otros patógenos, posee una proporción modesta de combinaciones de dominios TIR. Por el contrario, C88, una variedad del mismo grupo resistente al tizón tardío (Bao et al. 2022), presenta una proporción importante de las combinaciones de dominios TIR predichas. El menor contenido de dominios TIR en DC u otras características distintivas podrían hacerla más susceptible, ofreciendo oportunidades de mejora mediante nuevas introgresiones. De hecho, se ha demostrado que C88, que porta los grupos de genes R1 (un CC-NBS-LRR (Ballvora et al. 2002) y R2) de resistencia al tizón tardío, contiene introgresiones silvestres dentro de los genes NLR, lo que podría explicar el origen de los genes de resistencia funcionales (Bao et al. 2022). También se observaron introgresiones silvestres en el primer pangenoma tetraploide de la papa, que incluía la variedad Atlantic que contiene el clúster R1. Este examen minucioso de los genes relacionados con la resistencia a enfermedades en el DC será crucial para emplear técnicas de edición genética con el fin de mejorar el cultivar (Tiwari et al. 2022).

También se realizó con éxito la anotación funcional para 143,486 predicciones génicas dentro del ensamblaje DC, que constituyen el 93.1% del total de genes identificados (para más detalles, véase la Tabla S5 suplementaria). Entre estas anotaciones funcionales, la distribución de ortólogos es la siguiente: 61.7% para *Solanum tuberosum*, 23.9% para *Solanum lycopersicum*, 4.51% para *Nicotiana tormentosis* y 4.33% para *Nicotiana sylvestris*. Estos ortólogos identificados representan colectivamente el 94.5% de las anotaciones funcionales y, como era de esperar, están representados principalmente por las especies de *Solanum* filogenéticamente relacionadas. Por último, utilizamos phylostratr (Arendsee et al. 2019) para estimar el filostrato de cada gen en la anotación DC y obtuvimos un total de 5349 de genes específicos de *Solanum tuberosum*, este conjunto es un recurso valioso también para las otras especies dentro del género *Solanum*.

Los retos que plantea la combinación de alelos deseables en genomas tetraploides heterocigóticos como el de la papa pueden reducirse mediante un conocimiento más profundo de su diversidad y complejidad genómica. El ensamblaje del genoma resuelto por haplotipos y la predicción de genes codificantes de proteínas de la DC, una variedad cultivada en la región andina de Colombia y Ecuador, contribuyen a este objetivo. El genoma de DC aumenta el repertorio de los recientes ensamblajes haplotipo-resueltos de alta calidad de tetraploides (Bao et al. 2022; Hoopes et al. 2022; Sun et al. 2022). El estudio de DC que aquí se presenta muestra un ensamblaje y una anotación del genoma de alta calidad en comparación con otros genomas disponibles.

El análisis adicional de genomas de papas tetraploides con haplotipos altamente divergentes, es decir, en un pangenoma, se considera una contribución clave para diseñar híbridos como nuevos esquemas de mejora basados en genomas de papa (Hoopes et al. 2022; Zhang et al. 2021). Además, el análisis de grupos de genes NLR en haplotipos es deseable para combinar genes de resistencia de haplotipos específicos como parte de estos esquemas de mejora. Los pangenomas recopilados abarcan un importante acervo genético derivado de 44 papas diploides (Tang et al. 2022), seis tetraploides de Norteamérica y Europa (Hoopes et al. 2022) y 296 diploides y poliploides de 60 especies (Bozan et al. 2023). En conjunto, aumentan la diversidad alélica y el repertorio génico de los pangenomas de papa. Sin embargo, a pesar de la amplia distribución de *S. tuberosum* Andigenum, que abarca desde el norte de Colombia hasta el centro de Bolivia (Spooner et al. 2010), la representación de accesiones tetraploides del Grupo Andigenum originarias de la región andina de Colombia es limitada (ver Tabla suplementaria S5). Por lo tanto, el ensamblaje DC sirve como la primera representación de accesiones tetraploides del Grupo Andigenum originarias de la región andina de Colombia.

Conclusión

En este estudio, se llevó a cabo un ensamblaje *de novo* de una variedad tetraploide de Andígena denominada DC (Diacol Capiro), combinando la secuenciación de alta fidelidad (HiFi) de PacBio y la secuenciación de proximidad por ligamiento (HiC). Para la validación técnica se utilizó la secuenciación de Illumina. Las estrategias híbridas de ensamblaje dieron como resultado un ensamblaje genómico con cuatro haplotipos resueltos. Este ensamblaje final de DC obtenido es un genoma de calidad de referencia comparable a otros genomas de alta calidad. Así, el monoploide DMv6.1 y el diploide RH89-039-16, que corresponden a genomas a escala cromosómica con haplotipos resueltos, tienen métricas de ensamblaje similares a las del DC. La anotación genética del DC también mostró métricas BUSCO comparables a las de otros genomas tetraploides, diploides y monoploides de alta calidad. Sin embargo, DC mostró diferencias en el contenido de NLR y en la combinación de dominios, con una representación reducida de dominios que contienen TIR. Queda por investigar si esto podría estar relacionado con su susceptibilidad al tizón tardío y otros patógenos. DC es la primera secuencia de genoma profundo de alta calidad de ensamblaje y anotación de una variedad dentro del grupo Andigenum cultivada en la región andina de Colombia y Ecuador, donde surgen los centros de diversidad de papa cultivada. Este genoma ensamblado funciona como un valioso instrumento para mejorar el pangenoma de la papa y formular estrategias innovadoras de mejoramiento

dirigidas a desarrollar variedades de papa superiores, resistentes y más productivas. Esto contribuye a una comprensión más profunda de la diversidad y la evolución del genoma en este destacado cultivo.

Disponibilidad de datos

Los datos crudos están disponibles en el Bioproyecto PRJNA770354, Biosample SAMN22215911 (SRA accession SRR16302871) en el NCBI. El ensamblaje del genoma también está disponible en el NCBI con el número de accesión JAJHQB000000000. Además, los flujos de trabajo y scripts personalizados para el ensamblaje del genoma, la corrección, el andamiaje y anotación utilizados en el proyecto están alojados en GitHub https://github.com/phrh/DiacolCapiro_Genome. Este documento también está disponible en español para continuar el debate sobre este tema en Colombia y, más ampliamente, en la región (Archivo S2).

Agradecimientos

A Félix Eugenio Enciso, Felipe Uribe y María del Socorro Cerón por la introducción y desarrollo *in vitro* de Diacol Capiro bajo condiciones controladas. A María del Socorro Cerón por proporcionar tubérculos de Diacol Capiro. A Paola Delgadillo por la recolección de tejido foliar fresco para la extracción de ADN. Al personal de los Laboratorios de Producción Vegetal y Genética Molecular de AGROSAVIA por apoyar la introducción *in vitro* y la preparación de bibliotecas Illumina, respectivamente.

Financiación

La financiación fue obtenida por el proyecto No. 1000941 del Ministerio de Agricultura de Colombia a AGROSAVIA en colaboración con la Universidad de los Andes. La financiación para acceder a la infraestructura del Sandbox de datos fue proporcionada por el Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia.
El salario de MC fue parcialmente financiado por la beca UKRI-BBSRC 'Capacity building for bioinformatics in Latin America' (CABANA), en nombre del Global Challenges Research Fund [BB/P027849/1].

Contribuciones de los autores

PRH: Conceptualización, obtención de financiación (HiFi, Sandbox MINTIC), metodología, análisis, redacción, revisión y edición. DDD: Metodología, análisis. MC: Conceptualización, adquisición de fondos (bioinformática), metodología, revisión y edición. LSB: Conceptualización, adquisición de fondos (HiC, Illumina), metodología, revisión y edición. MFG: Conceptualización, metodología, análisis, redacción, revisión y edición. LAM: Conceptualización, metodología, análisis, revisión y edición.

Conflictos de intereses

Ninguno declarado.

Bibliografía citada

Andrade-Piedra J, Torres L. 2011. Inventario de tecnologías e información para el cultivo de papa en ecuador. Information at https://cipotato.org/papaenecuador/2017/10/12/24-diacol-capiro/.

Arendsee Z, Li J, Singh U, Seetharam A, Dorman K, Wurtele ES. 2019. phylostratr: A framework for phylostratigraphy. Bioinformatics. 35:3617–3627.

Aversano R, Contaldi F, Ercolano MR, Grosso V, Iorizzo M, Tatino F, Xumerle L, Dal Molin A, Avanzato C, Ferrarini A *et al*. 2015. The Solanum commersonii Genome Sequence Provides Insights into Adaptation to Stress Conditions and Genome Evolution of Wild Potato Relatives. The Plant Cell. 27:954–968.

Ballvora A, Ercolano MR, Weiss J, Meksem K, Bormann CA, Oberhagemann P, Salamini F, Gebhardt C. 2002. The r1 gene for potato resistance to late blight (phytophthora infestans) belongs to the leucine zipper/nbs/lrr class of plant resistance genes. The Plant Journal. 30:361–371.

Bao Z, Li C, Li G, Wang P, Peng Z, Cheng L, Li H, Zhang Z, Li Y, Huang W *et al*. 2022. Genome architecture and tetrasomic inheritance of autotetraploid potato. Molecular Plant. 15:1211–1226.

Biosciences P. 2020. Ipa hifi genome assembler. software at https://github.com/PacificBiosciences/pbipa.

Bozan I, Achakkagari SR, Anglin NL, Ellis D, Tai HH, Strömvik MV. 2023. Pangenome analyses reveal impact of transposable elements and ploidy on the evolution of potato species. Proceedings of the National Academy of Sciences. 120:e2211117120.

Buchfink B, Reuter K, Drost HG. 2021a. Sensitive protein alignments at tree-of-life scale using diamond. Nature methods. 18:366–368.

Buchfink B, Reuter K, Drost HG. 2021b. Sensitive protein alignments at tree-of-life scale using diamond. Nature methods. 18:366–368.

Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei J, Achawanantakun R, Jiao D, Lawrence CJ *et al*. 2014. Maker-p: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant physiology. 164:513–524.

Campoy JA, Sun H, Goel M, Jiao WB, Folz-Donahue K, Wang N, Rubio M, Liu C, Kukat C, Ruiz D *et al*. 2020. Gamete binning: chromosomelevel and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. Genome biology. 21:1–20.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggnog-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Molecular biology and evolution. 38:5825–5829.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M. 2008. Maker: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome research. 18:188–196.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit – Interactive Quality Assessment of Genome Assemblies. G3 Genes|Genomes|Genetics. 10:1361–1374.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods. 18:170–175.

Dangl JL, Jones JD. 2001. Plant pathogens and integrated defence responses to infection. nature. 411:826–833.

Devaux A, Goffart JP, Petsakos A, Kromann P, Gatto M, Okello J, Suarez V, Hareau G. 2020. Global food security, contributions from sustainable potato agri-food systems. The potato crop: Its agricultural, nutritional and social contribution to humankind. pp. 3–35.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016. Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. Cell Systems. 3:99–101.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. Ltrharvest, an efficient and flexible software for de novo detection of ltr retrotransposons. BMC bioinformatics. 9:1–14.

Ercolano MR, D'Esposito D, Andolfo G, Frusciante L. 2022. Multilevel evolution shapes the function of nb-lrr encoding genes in plant innate immunity. Frontiers in Plant Science. 13:1007288.

Gaiero P, Vaio M, Peters SA, Schranz ME, de Jong H, Speranza PR. 2019. Comparative analysis of repetitive sequences among species from the potato and the tomato clades. Annals of Botany. 123:521–532.

Gavrilenko T, Antonova O, Shuvalova A, Krylova E, Alpatyeva N, Spooner DM, Novikova L. 2013. Genetic diversity and origin of cultivated potatoes based on plastid microsatellite polymorphism. Genetic Resources and Crop Evolution. 60:1997–2015.

Ghislain M, Douches DS. 2020. The genes and genomes of the potato. The potato crop: Its agricultural, nutritional and social contribution to humankind. pp. 139–162.

Gong Z, Wu Y, Koblížková A, Torres GA, Wang K, Iovene M, Neumann P, Zhang W, Novák P, Buell CR *et al*. 2012. Repeatless and Repeat-Based Centromeres in Potato: Implications for Centromere Evolution . The Plant Cell. 24:3559–3574.

González-Orozco CE, Reyes-Herrera PH, Sosa CC, Torres RT, Manrique-Carpintero NC, Lasso-Paredes Z, Cerón-Souza I, Yockteng R. 2023. Wild relatives of potato (solanum l. sec. petota) poorly sampled and unprotected in colombia. Crop Science. .

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. Quast: quality assessment tool for genome assemblies. Bioinformatics. 29:1072–1075.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation with braker. Gene prediction: methods and protocols. pp. 65–95.

Hoopes G, Meng X, Hamilton JP, Achakkagari SR, Guesdes FdAF, Bolger ME, Coombs JJ, Esselink D, Kaiser NR, Kodde L *et al*. 2022. Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. Molecular plant. 15:520–536.

Hosmani PS, Flores-Gonzalez M, van de Geest H, Maumus F, Bakker LV, Schijlen E, van Haarst J, Cordewener J, Sanchez-Perez G, Peters S *et al*. 2019. An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, hi-c proximity ligation and optical maps. biorxiv. p. 767764.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ *et al*. 2019. eggnog 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research. 47:D309–D314.

Jupe F, Pritchard L, Etherington GJ, MacKenzie K, Cock PJ, Wright F, Sharma SK, Bolser D, Bryan GJ, Jones JD *et al*. 2012. Identification and localisation of the nb-lrr gene family within the potato genome. BMC genomics. 13:1–14.

Kim D, Langmead B, Salzberg SL. 2015. Hisat: a fast spliced aligner with low memory requirements. Nature methods. 12:357–360.

Kokot M, Długosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. Bioinformatics. 33:2759–2761.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nature biotechnology. 37:540–546.

Kyriakidou M, Achakkagari SR, López JHG, Zhu X, Tang CY, Tai HH, Anglin NL, Ellis D, Strömvik MV. 2020a. Structural genome analysis in cultivated potato taxa. Theoretical and Applied Genetics. 133:951–966.

Kyriakidou M, Anglin NL, Ellis D, Tai HH, Strömvik MV. 2020b. Genome assembly of six polyploid potato genomes. Scientific data. 7:1–6. Leisner CP, Hamilton JP, Crisovan E, Manrique-Carpintero NC, Marand AP, Newton L, Pham GM, Jiang J, Douches DS, Jansky SH *et al*. 2018. Genome sequence of m6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species solanum chacoense, reveals residual heterozygosity. The Plant Journal. 94:562–570.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34:3094–3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

Li P, Quan X, Jia G, Xiao J, Cloutier S, You FM. 2016. Rgaugury: a pipeline for genome-wide prediction of resistance gene analogs (rgas) in plants. BMC genomics. 17:1–10.

Lin T, Lashbrook CC, Cho SK, Butler NM, Sharma P, Muppirala U, Severin AJ, Hannapel DJ. 2015. Transcriptional analysis of phloemassociated cells of potato. BMC genomics. 16:1–24.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution. 38:4647–4654.

Manrique-Carpintero NC, Berdugo-Cely JA, Cerón-Souza I, Lasso-Paredes Z, Reyes-Herrera PH, Yockteng R. 2023. Defining a diverse core collection of the colombian central collection of potatoes: a tool to advance research and breeding. Frontiers in Plant Science. 14:1046400. Murashige T, Skoog F. 1962. A revised medium for rapid growth and bio assays with tobacco tissue cultures. Physiologia plantarum. 15:473–497.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. Hicanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome research. 30:1291–1305.

Ou S, Chen J, Jiang N. 2018. Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Research. 46:e126–e126. Ou S, Jiang N. 2017. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons . Plant Physiology. 176:1410–1422.

PGSC. 2011. Genome sequence and analysis of the tuber crop potato. Nature. 475:189–195.

Pham GM, Hamilton JP, Wood JC, Burke JT, Zhao H, Vaillancourt B, Ou S, Jiang J, Buell CR. 2020. Construction of a chromosome-scale long-read reference genome assembly for potato. GigaScience. 9. giaa100.

Ponce OP, Torres Y, Prashar A, Buell R, Orjeda G, Compton L. 2022. Transcriptome profiling shows a rapid variety-specific response in two andigenum potato varieties under drought stress. Frontiers in Plant Science. 13:1003907.

Porras Rodríguez PD, Herrera Heredia CA. 2015. Modelo productivo de la papa variedad diacol capiro para el departamento de antioquia.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. Interproscan: protein domains identifier. Nucleic acids research. 33:W116–W120.

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. 2018. High-resolution tads reveal dna sequences underlying genome organization in flies. Nature communications. 9:1–15.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. Genomescope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. Nature Communications. 11:1–10.

Reyes-Herrera PH, Torres-Bedoya E, Lopez-Alvarez D, Burbano-David D, Carmona SL, Bebber DP, Studholme DJ, Betancourt M, SotoSuarez M. 2023. Genome sequence data reveal at least two distinct incursions of the tropical race 4 variant of fusarium wilt into south america. Phytopathology® . 113:90–97.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome biology. 21:1–27.

Romero AP, Alarcón A, Valbuena RI, Galeano CH. 2017. Physiological assessment of water stress in potato using spectral information. Frontiers in Plant Science. 8:1608.

SASA. 2010. Ecpd european cultivated potato database. Database at https://www.europotato.org/varieties/view/Otava-E.

SIPSA D. 2013. El cultivo de la papa, solanum tuberosum alimento de gran valor nutritivo, clave en la seguridad alimentaria mundial. INSUMOS Y FACTORES ASOCIADOS A LA PRODUCCIÓN AGROPECUARIA. 15.

Smit A, Hubley R. 2008. Repeatmodeler v1.0.8 open-1.0 http://www. repeatmasker. org.

Spooner DM, Gavrilenko T, Jansky SH, Ovchinnikova A, Krylova E, Knapp S, Simon R. 2010. Ecogeography of ploidy variation in cultivated potato (solanum sect. petota). American journal of botany. 97:2049–2060.

Spooner DM, Núñez J, Trujillo G, del Rosario Herrera M, Guzmán F, Ghislain M. 2007. Extensive simple sequence repeat genotyping of potato landraces supports a major reevaluation of their gene pool structure and classification. Proceedings of the National Academy of Sciences. 104:19398–19403.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cdna alignments to improve de novo gene finding. Bioinformatics. 24:637–644.

Steuernagel B, Witek K, Krattinger SG, Ramirez-Gonzalez RH, Schoonbeek Hj, Yu G, Baggs E, Witek AI, Yadav I, Krasileva KV *et al*. 2020. The nlr-annotator tool enables annotation of the intracellular immune receptor repertoire. Plant Physiology. 183:468–482.

Sun H, Jiao WB, Krause K, Campoy JA, Goel M, Folz-Donahue K, Kukat C, Huettel B, Schneeberger K. 2022. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nature genetics. 54:342–348.

Tang D, Jia Y, Zhang J, Li H, Cheng L, Wang P, Bao Z, Liu Z, Feng S, Zhu X *et al*. 2022. Genome evolution and diversity of wild and cultivated potatoes. Nature. 606:535–541.

Tiwari JK, Buckseth T, Challam C, Zinta R, Bhatia N, Dalamu D, Naik S, Poonia AK, Singh RK, Luthra SK *et al*. 2022. Crispr/cas genome editing in potato: current status and future perspectives. Frontiers in Genetics. 13:827808.

Torres L, Montesdeoca F, Gallegos P, Castillo C, Asaquibay C, Valverde F, Andrade-Piedra J. 2011. Inventario de tecnologías e información para el cultivo de papa en ecuador. Centro Internacional de la Papa (CIP). 3.

van Lieshout N, van der Burgt A, de Vries ME, Ter Maat M, Eickholt D, Esselink D, van Kaauwen MP, Kodde LP, Visser RG, Lindhout P *et al*. 2020. Solyntus, the new highly contiguous reference genome for potato (solanum tuberosum). G3: Genes, Genomes, Genetics. 10:3489–3495.

Yan L, Zhang Y, Cai G, Qing Y, Song J, Wang H, Tan X, Liu C, Yang M, Fang Z *et al*. 2021. Genome assembly of primitive cultivated potato Solanum stenotomum provides insights into potato evolution. G3 Genes|Genomes|Genetics. 11. jkab262.

Zhang C, Yang Z, Tang D, Zhu Y, Wang P, Li D, Zhu G, Xiong X, Shang Y, Li C *et al*. 2021. Genome design of hybrid potato. Cell. 184:3873–3883.

Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, Zhu F, Jones T, Zhu X, Bowers J *et al*. 2018. Allele-defined genome of the autopolyploid sugarcane saccharum spontaneum l. Nature genetics. 50:1565–1573.

Zhang X, Zhang S, Zhao Q, Ming R, Tang H. 2019. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on hi-c data. Nature plants. 5:833–845.

Zhou Q, Tang D, Huang W, Yang Z, Zhang Y, Hamilton JP, Visser RG, Bachem CW, Buell CR, Zhang Z *et al*. 2020. Haplotype-resolved genome analyses of a heterozygous diploid potato. Nature genetics. 52:1018–1023.

Leyendas Figuras y Tablas

Tabla 1 Estadísticas de ensamblaje del genoma de las lecturas HiFi.

Figura 1 Tubérculos del cultivar Diacol Capiro.

Tabla 2 Datos utilizados para las estrategias (st) de scaffolding en ALLHiC: (1) Ensamblaje inicial, (2) Datos Hi-C utilizados para la corrección del genoma, (3) Ensamblaje corregido utilizado para el andamiaje, (4) Datos Hi-C utilizados para el andamiaje y, (5) Genoma de referencia utilizado para el paso de poda. Las estrategias seleccionadas aparecen resaltadas en amarillo.

Tabla 3 C Comparación de las estrategias de scaffolding st2, st3 y st6 con el ensamblaje DC. Las mejores métricas por línea se resaltan en verde, mientras que los problemas potenciales por montaje se resaltan en rojo.

Figura 2 Diagramas de caja muestran el rango de tamaños de los cuatro scaffolds más grandes para cada uno de los 12 cromosomas (Chr) utilizando diferentes estrategias de ensamblaje (st0-st7). También se incluye el ensamblaje DC. En la leyenda, las estrategias se separan en función del genoma de referencia utilizado en la etapa de poda. Las líneas horizontales superiores indican el tamaño de cada cromosoma en los genomas de referencia DMv6.1 (verde) y RH89-039-16 con dos haplotipos (azul y rojo).

Figura 3 Flujo de trabajo para la selección de pseudo-cromosomas utilizando los resultados de tres estrategias. (i) Selección de los cuatro scaffolds más grandes (A, B, C y D) para cada cromosoma (Chri) para las estrategias st2, st3 y st6. (ii) Alineamiento con Minimap2 entre los scaffolds de cada estrategia y comparación para identificar grupos similares y scaffolds diferentes. (iii) Identificación de los scaffold representativos, entre scaffolds similares, basándose en la puntuación BUSCO. En este ejemplo, Ast2, Cst2 y Dst2 obtuvieron la puntuación BUSCO más alta. (iv) Construcción de estrategias híbridas combinando scaffolds representativos y el grupo diferente, cada fila es una combinación alternativa de scaffolds para representar el Chri. (v) Comparaciones de estrategias híbridas basadas en la puntuación BUSCO, la mediana de la tasa de alineación con DMv6.1 y la contigüidad basada en el diagrama de puntos con DMv6.1. (vi) Selección final de los mejores candidatos como pseudo-cromosomas para Chri, en este caso Ast2Ast3Cst2Dst2 es una combinación de andamiajes de st2 y st3.

Figura 5 Diagrama de círculos para el ensamblaje DC. A. 48 pseudo-cromosomas. B. LAI sobre ventanas deslizantes de 100 kb∗1. C. Lecturas de CENH3ChIP-seq (Gong et al. 2012) alineadas con el ensamblaje DC, con cromosomas divididos en ventanas de 100 kb. Los picos resultantes detectados en este análisis son indicativos de posibles regiones centroméricas dentro del genoma. D. Histograma de predicciones de genes utilizando un tamaño de ventana de 5 Mb. E. Histograma para LTRs con un tamaño de ventana de 5 Mb.
*1 El verde indica las posiciones con LAI >= 20 (categoría de oro (Ou et al. 2018)), el azul las posiciones con 10<LAI<20, y el rojo las posiciones con LAI <10.

Tabla 4 Estadísticas de ensamblaje del genoma para los ensamblajes a escala cromosómica de *S. tuberosum* utilizados como referencia y los ensamblajes del grupo Andigenum de *S. tuberosum*

Table 5 Comparación de las anotaciones del ensamblaje del genoma de la papa, ordenadas alfabéticamente, basándose en múltiples factores, incluyendo la puntuación BUSCO (utilizando el transcriptoma de modo y el linaje embryophyta_odb10), el número de genes y la presencia de genes putativos NLR. La anotación DC se resalta en negrita.

Figura 6 Proporciones de genes NLR (Nucleotide-binding Leucine-rich Repeat) predichos en seis ensamblajes distintos del genoma de la papa. La representación de varias combinaciones de dominios dentro de los genes NLR se visualiza utilizando diferentes colores (véase la leyenda). Los porcentajes inferiores al 1% no se incluyeron en la figura para evitar solapamientos visuales y mantener la claridad.