//UCIENCIA/23
CONVENCIÓN CIENTÍFICA INTERNACIONAL

Temática: Bioinformática en la investigación agropecuaria.

Identificación de genes candidatos del genoma del cacao asociados a caracteres morfológicos de importancia agronómica

Identification of candidate genes of the cocoa genome associated with morphological characters of agronomic importance

Elaine Hernández Pereira [1] MSc. Mario Pupo Meriño [2] Ángel Rafael Ramirez Ramírez [3]
Igor Bidot Martinez[4] Pierre Bertin [5]

[1] Universidad de Ciencias Informáticas . Carretera a San Antonio de los Baños Km 2½, Rpto. Torrens, La Habana. elainehp@uci.cu

[2] Universidad de Ciencias Informáticas . Carretera a San Antonio de los Baños Km 2½, Rpto. Torrens, La Habana. mpupom@uci.cu

[3] Universidad de Guantánamo, Facultad Agroforestal, Cuba. aramirez@cug.co.cu

[4] Universidad de Guantánamo, Facultad Agroforestal, Cuba. ibidot@cug.co.cu

[5] Universidad de Católica de Lobaina, Bélgica. pierre.bertin@uclouvain.be

[*] Autor para correspondencia: elainehp@uci.cu

Resumen

El cacao, proveniente de la semilla del árbol tropical cacaotero de naturaleza diploide, es ampliamente valorado por su uso en la comercialización y producción del chocolate, convirtiéndose en un insumo agrícola de gran importancia económica a nivel mundial. En Cuba, se llevó a cabo un estudio integral para la caracterización morfológica del cacao tradicional cubano, lo que permitió la identificación de genes candidatos relacionados con caracteres morfológicos relevantes en la agricultura. La investigación se basó en la integración de herramientas bioinformáticas para el procesamiento de datos, que incluyó la secuenciación ddRADseq, el control de calidad y la anotación de genes. Se identificaron 249 genes candidatos, destacando una porción significativa del genoma formada por elementos

transponibles y los transportadores de iones metálicos, específicamente de zinc, como fundamentales para el crecimiento saludable de la planta para su desarrollo metabólico y estructural. También se descubrieron genes codificados por proteínas ricas en leucina del sitio de unión a nucleótidos en respuesta a la Phytophthora, un hongo que afecta al cacao. Además, se evaluaron los genes relacionados con los flavonoides y la glucosilación, que tienen un gran potencial en las industrias de procesamiento de alimentos y en el mejoramiento de la calidad del producto. Finalmente, se estableció una relación entre las variables agromorfológicas y la variabilidad genética y morfoagronómica, lo que demuestra la importancia de la investigación para el desarrollo de la producción agrícola.

Palabras clave: ddRADseq, gen candidato, carácter morfológico, herramientas

Abstract

Cocoa, from the seed of the diploid tropical cocoa tree, is widely valued for its use in the marketing and production of chocolate, becoming an agricultural input of great economic importance worldwide. In Cuba, a comprehensive study was carried out for the morphological characterization of traditional Cuban cacao, which allowed the identification of candidate genes related to relevant morphological characters in agriculture. The research was based on the integration of bioinformatics tools for data processing, which included ddRADseq sequencing, quality control, and gene annotation. 249 candidate genes were identified, highlighting a significant portion of the genome made up of transposable elements and metal ion transporters, specifically zinc, as essential for healthy plant growth for its metabolic and structural development. Genes encoding leucine-rich nucleotide binding site proteins were also discovered in response to Phytophthora, a fungus that affects cocoa. In addition, genes related to flavonoids and glycosylation, which have great potential in food processing industries and in improving product quality, were evaluated. Finally, a relationship was established between the agromorphological variables and the genetic and morphoagronomic variability, which demonstrates the importance of research for the development of agricultural production.

Keywords: *ddRADseq, candidate gene, morphological character, tools.*

Introducción

La investigación sobre las plantas es una de las prácticas más antiguas del hombre, pues antes de nuestra era, los mejoramientos y estudios, eran muy lentos para alcanzar resultados fiables. En la actualidad, se puede apreciar la evolución de las especies naturales, al igual que los genes que influyen en su desarrollo. La manipulación de dichos genes para crear vegetales genéticamente mejorados**,** forma parte de la genética vegetal, lo que apunta a ser un hito de la agricultura moderna, determinado por la selección de plantas con comportamientos superiores en cuanto a productividad, calidad y resistencia a plagas y enfermedades. Con esto, se espera contribuir a aumentar la producción agrícola como en plantaciones de cacao (*Theobroma cacao L.*), por ser un insumo agrícola altamente rentable e importante económicamente a nivel mundial (Jorge León, 2000).

El cacao es la semilla del árbol tropical cacaotero de naturaleza diploide y hojas perennes de la familia malváceas. Se cultiva por millones de agricultores en todo el mundo y es un medio de subsistencia para más de 40 millones de personas. Se le otorga una gran importancia económica en la producción y comercialización del chocolate, y representa un recurso biocultural para el mundo (Jorge León, 2000) (Ricaño Rodríguez ,2018). En Cuba, las plantaciones de cacao se asentaron con mayor facilidad en la región oriental, en el macizo Nipe-Sagua-Baracoa, actualmente el mayor productor de cacao de Cuba. En la agricultura se constituyó un producto de subsistencia diversificado, orientado básicamente a satisfacer el consumo y el comercio interno, con exportaciones insignificantes (Igor Bidot Martínez, 2021).

Recientemente, gracias a la colaboración entre universidades cubanas y la Universidad Católica de Lovaina de Bélgica, se logró la caracterización genética del cacao tradicional cubano con 33 descriptores morfológicos cualitativos. Además, se amplificó el ácido desoxirribonucleico (ADN) con los 15 microsatélites estándares y se obtuvieron los datos ddRADseq (del inglés: double digestion Restrictionsite-Associated DNA sequencing). Según los resultados de la colaboración, se demuestra la alta variabilidad morfológica y la baja variabilidad genética del cacao, asociado al resultado de su evolución. Con una estrecha base genética como resultado de un cuello de botella y antecedentes de siete grupos genéticos y sus híbridos. La clasificación del cacao tradicional cubano en el grupo trinitario, le atribuye potencialidades para su empleo en el mejoramiento genético del cacao comercial presente en Cuba, debido a sus caracteres agronómicos más destacados para su agroproductividad como son: la calidad de la cocoa/chocolate, los relacionados con el rendimiento, la productividad, y la resistencia a Phytophthora, causante de la pudrición negra de la mazorca, la enfermedad más común en el país, que afectan fundamentalmente los frutos del cacao del P. palmivora y P. tropicalis (Igor Bidot Martínez, 2021).

A partir de los datos, junto con el apoyo de las entidades participantes como la Facultad de Biología de la Universidad de la Habana, la Universidad de Ciencias Informáticas y la Universidad de Guantánamo, se traza el siguiente objetivo: Identificar, con el empleo de herramientas bioinformáticas, genes candidatos del genoma del cacao, que puedan explicar caracteres morfológicos de importancia agronómica como la resistencia a enfermedades, calidad y productividad, con el uso de datos de ddRADseq de variantes de cacao cubanas. Esto permitirá seleccionar clones de cacao con potencialidades genéticas y morfológicas para el mejoramiento del cultivo, que contribuyan al incremento de las producciones y la calidad del chocolate. Además, posibilitará la formación de recursos humanos especializados en metodologías y tecnologías para la generación, manejo e interpretación de datos biológicos a gran escala. Este proyecto aportaría al desarrollo estratégico del cacao en Cuba, a la vez que podría contribuir a incrementar el valor agregado a dichas producciones.

Se busca vincular características de variabilidad genética y morfoagronómicas con las variables agromorfológicas, para poder explicar las relaciones que se encuentran entre ellos. Al mismo tiempo, se necesita identificar cuáles son los genes responsables de esta variabilidad, qué funciones cumplen en el metabolismo, el comportamiento de la planta y los procesos biológicos en los que participa. Entonces, existe un elevado cúmulo de información con las bases de datos internacionales, los datos ddRADseq y una amplia caracterización morfológica, con énfasis en aspectos de interés agroproductivos. Sin embargo, no se han identificado los genes candidatos en el genoma del cacao asociadas con características de interés agrícola como la resistencia a enfermedades, calidad, productividad.

Materiales y métodos

Este artículo presenta el diseño de un flujo de trabajo utilizado para la identificación de genes candidatos en el genoma del cacao (Cornejo, 2018). Con una base de datos que incluye 1 244 420 354 secuencias de ADN pertenecientes a 423 muestras y 406 plantas de cacao diferentes, el estudio cubre un amplio espectro de clonaciones alojadas en el Banco de Germoplasma de Cacao de Cuba. Los datos se organizan en tres bibliotecas analizadas, con la secuenciación ddRADseq, cada una de las cuales contiene 140 clones de cacao de ambas colecciones. Los investigadores decidieron enfocarse únicamente en la tercera biblioteca para el análisis de datos y la identificación de los genes candidatos para esta investigación.

En la tercera biblioteca, se llevó a cabo una evaluación del control de calidad a fin de identificar y excluir cualquier tipo de error que pudiera afectar la interpretación de los análisis. Para ello, se usó la herramienta Trim-galore! con el objetivo de analizar la probabilidad de error y, en conjunto con Cutadapt, se trataron las secuencias para detectar las regiones de bajo nivel de calidad y las secuencias de adaptadores innecesarias (Krueger F., 2017). Una vez confirmada la calidad de las secuencias a través del control de calidad, se procedió a utilizar Process_Radtags de Stacks para examinar las lecturas y desmultiplexar los datos (Catchen, 2013).

Una vez terminada la desmultiplexación de los datos, se pasó a asignar las lecturas secuenciadas al genoma de referencia de cacao Matina 1-6 con el uso de la herramienta Bwa-mem2 (Kim, 2023). Se eligieron las opciones por defecto de la herramienta, excepto la penalización por errores de apareamiento, que se fijó en -B 6 (-B 4 por defecto) siguiendo las recomendaciones de Cornejo et al. (2018). Para analizar los datos de alineación y detectar posibles sesgos de mapeo de los datos, se utilizó QualiMap BamQC, la cual proporcionó una vista general de los datos. La tasa de saturación de las características detectadas fue analizada en relación a la profundidad de secuenciación, lo que

permitió determinar si era posible detectar más características al aumentar la profundidad de secuenciación y calcular las coberturas correspondientes a las características genómicas.

Una vez confirmada la calidad de mapeo, se hizo el llamado de variantes con BBtools (Bushnell, 2017). Posteriormente, se eliminaron posibles falsos positivos en la lista de variantes generadas con la herramienta Lofreq (Wilm, 2012). Esta técnica facilitó identificar posiciones variantes que estaban marcadas por un sesgo significativo en el hilo del que se derivan las lecturas de apoyo y la alineación de registros SAM/BAM de lectura. Se complementó este análisis con la herramienta Samtools, la cual posibilitó trabajar los datos de alineación de secuencias para construir en cada posición del genoma un corte vertical a través de todas las lecturas que cubren la posición ("pileup") (Danecek, 2021). Para calcular las probabilidades de genotipo, se escojió una técnica que representó cuán consistentes eran los datos observados con los posibles genotipos diploides.

Para detectar posibles variantes genéticas bayesianas y encontrar pequeños polimorfismos, se configuró los filtros de entrada de la herramienta FreeBayes (Nancy F., 2016). Se excluyeron las alineaciones del análisis con una calidad de mapeo inferior a 20. Posteriormente, la herramienta Bcftools registró los genotipos más probables de todas las muestras en cada sitio de variante, se convirtió así un archivo VCF de muestras múltiples (Danecek, 2021). Para eliminar variantes con patrones de herencia incompatibles con la herencia observada del fenotipo en los individuos afectados, se empleó el método de exclusión. Una vez finalizado el proceso de mapeo de las lecturas, el llamado de variantes y su posterior filtrado, se utilizó la herramienta SnpEff para anotar las variantes y predecir los efectos de las variantes genéticas del ADN del cacao. Se seleccionó el efecto "Alto" (High) y se mantuvieron los parámetros por defecto. Los resultados obtenidos a través de SnpEff incluyen una amplia variedad de métricas útiles, tales como la distribución de variantes en las características genéticas o los cambios en los codones (Cingolani, 2012).

En la plataforma String, se llevó a cabo una recopilación e integración sistemática de las interacciones proteína-proteína de los genes encontrados, con el fin de enriquecer su funcionamiento en su red. Este proceso permitió la visualización completa de la red, así como predicciones futuras acerca de las funciones de los genes (Damian, 2023). Posteriormente, se investigó individualmente cada uno de estos genes con el propósito de determinar si habían sufrido alguna mutación que pudiera estar asociada con una característica morfológica específica. Para identificar los posibles genes candidatos en el genoma del cacao, se usó un método basado en genómica comparativa, donde se estudiaron las similitudes y diferencias entre cada uno de ellos. Para obtener asociaciones significativas de los caracteres morfológicos de importancia agronómica con los genes candidatos encontrados, se aprovechó la información recopilada en Uniprot y Panther (Coudert, 2023) (Paul D., 2022).

Resultados y discusión

La Figura 1 presenta una representación visual de los gráficos de calidad por base obtenidos a partir del programa TimGalore!, los cuales muestran las secuencias primarias de la biblioteca, así como las secuencias obtenidas después de ser procesadas. Al observar las secuencias sin procesar, se evidenció un patrón esperado en cuanto a la calidad, con una disminución en la misma al inicio seguida de un incremento progresivo hasta alcanzar valores máximos y una posterior disminución uniforme. Además, se observaron variaciones en la distribución del contenido de nucleótidos por posición para las secuencias primarias.

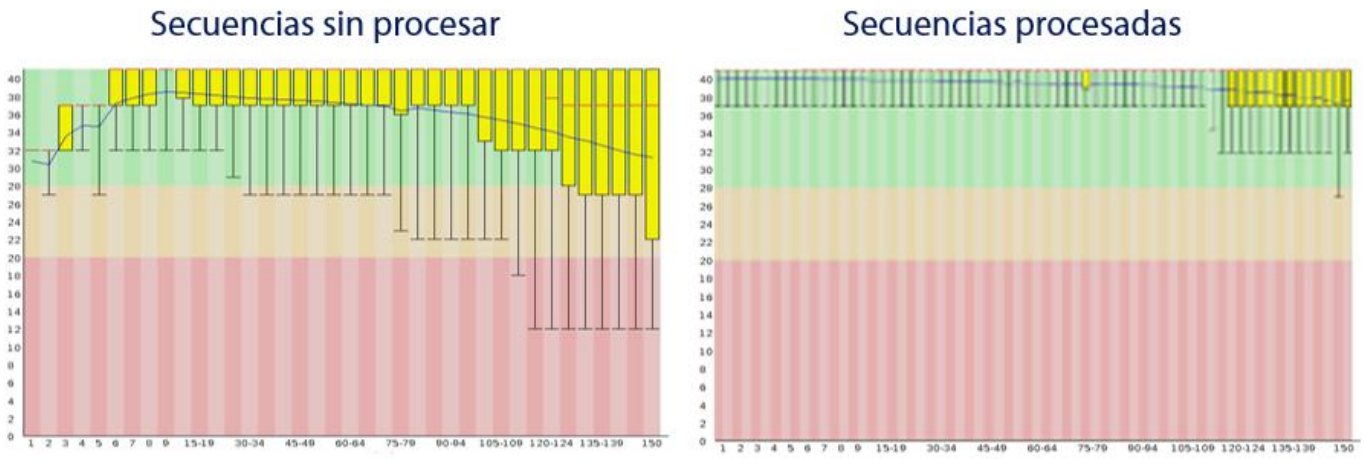

Figura 1 Valores promedio de calidad por posición de los nucleótidos en las secuencias de ADN de la biblioteca anterior y después del procesamiento.

En la Figura 1, se pueden observar tres zonas de calidad: verde para la óptima, naranja para la intermedia y roja para la baja. Estas zonas son fundamentales para la interpretación de los resultados y la toma de decisiones en análisis posteriores. Gracias a este análisis, se pudo mejorar aún más la calidad de las secuencias. Al procesar las secuencias primarias, se notaron cambios significativos en los valores de calidad por posición, así como en la distribución del contenido de nucleótidos por posición. Este procesamiento permitió obtener información más precisa y detallada de las secuencias, lo que resulta muy valioso para futuros análisis y estudio.

En la evaluación de la calidad de los datos de alineación proporcionados en un archivo BAM, generados en el mapeo de las lecturas al genoma de referencia con Bwa-mem2, se utilizó QualiMap BamQC. El programa realizó informes de calidad que representan la uniformidad y rango de cobertura de secuenciación, incluyendo el número de bases de referencia cubiertas por lecturas mapeadas en varias profundidades y la tasa de cobertura en ubicaciones genómicas comunes. La tasa de cobertura proporciona una medida de la cantidad de referencias secuenciadas, con un 55% de fracción de referencia secuenciada reflejada en la cobertura. Se representó el porcentaje de bases recortadas en las lecturas para la detección de las mismas (25%-50%). Para ello, se compararon con una distribución del genoma recalculada, permitiendo verificar cambios en el contenido de GC (Contenido guanina citocina) de un 30%.

Selección de variantes

Una anotación fundamental de variantes implica categorizar cada una de ellas según su relación con las secuencias de codificación en el genoma, así como su capacidad para modificar dichas secuencias y afectar el producto del gen. Para ello, se compararon las variantes con otras variantes conocidas y anotaciones del genoma, lo que permitió predecir información relevante acerca de las mismas. Para llevar a cabo este análisis, se utilizó SNPeff, obteniendo un total de 269 279 variantes (antes de aplicar el filtro) en un genoma de longitud total de 324 879 930, y una longitud efectiva de 324 731 338. En la Tabla 1 se resumen los tipos de variantes destacadas, incluyendo SNP, inserción (INS) y supresión (DEL), así como su cuenta total.

Tabla 1 Tipos de variantes destacadas.

Número de variantes por tipo

Type	Total
SNP	256,959
MNP	0
INS	6,654
DEL	5,666
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	269,279

Enriquecimiento funcional

Una variante genética es un cambio permanente en la secuencia de ADN que forma un gen, lo que puede afectar uno o varios componentes básicos del ADN en dicho gen. En el estudio, se analizó un conjunto de variables con características genéticas para predecir su efecto y anotar las variantes. Se seleccionó la variación de 'Alto' impacto en la secuencia de ADN de la muestra para obtener información y así identificar los genes del genoma del cacao. Para integrar todas las asociaciones conocidas y previstas entre proteínas del cacao, incluyendo la detección automatizada de funcionalidades enriquecidas de los 249 genes identificados de la variación de mejor impacto, se utilizó la base de datos String. En la Figura 2 se presenta una red que cubre todas las proteínas mapeadas, sus 18 interconexiones encontradas y una agrupación jerárquica de la propia red String, dividida en 3 grupos más pequeños estrechamente vinculados.

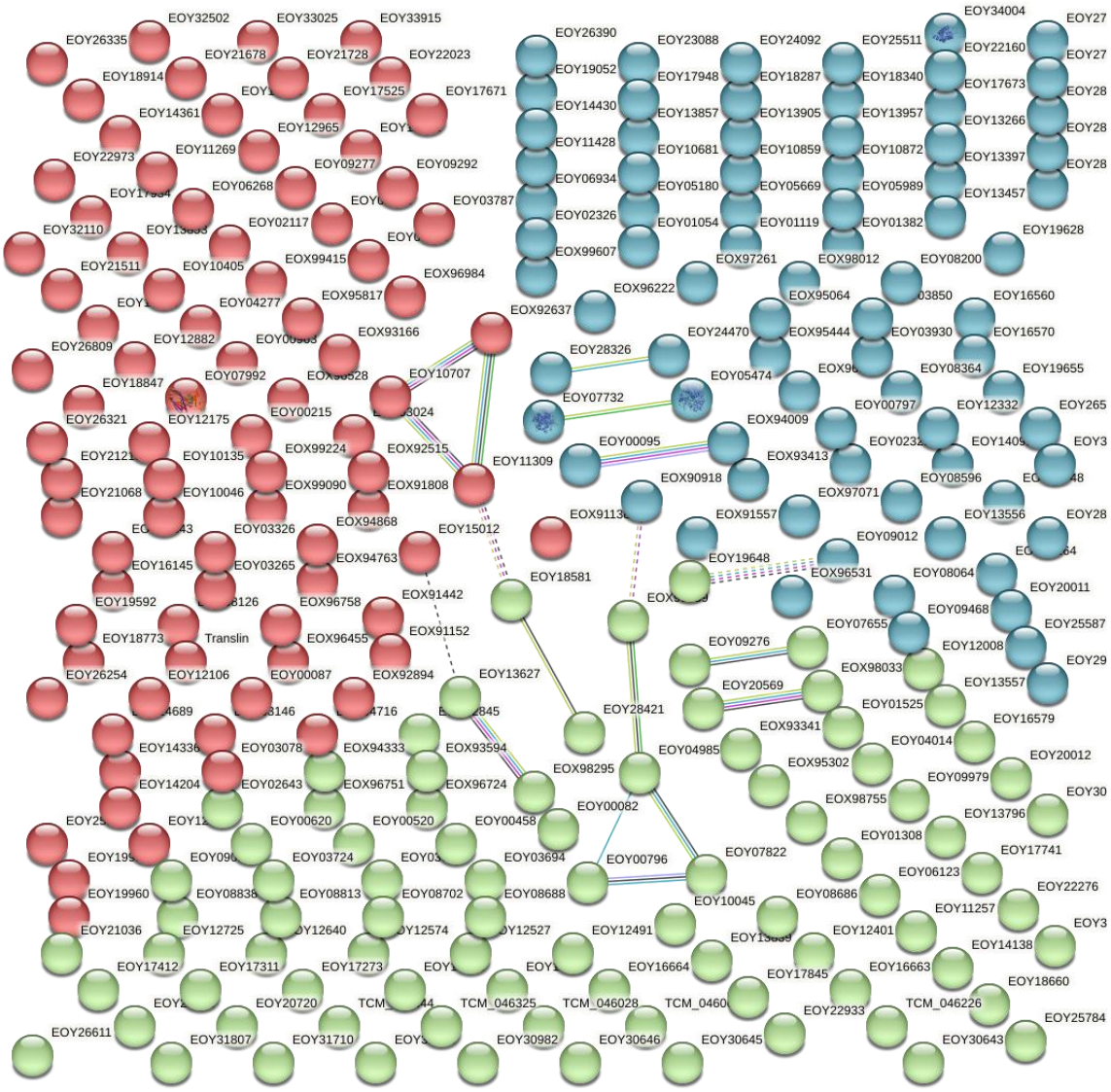

Figura 2 Enriquecimiento funcional de 249 genes del cacao.

A pesar de haber logrado un enriquecimiento funcional en la red, existen proteínas que aún no han sido estudiadas y que String no las reconoce, lo que llevó a clasificar 78 proteínas como no caracterizadas en el estudio. Sin embargo, al enfocarnos en las variaciones genéticas, se encontró que los eventos de transposición son una de las variaciones más comunes, con un nivel de enriquecimiento de 1.25. Estos eventos pueden provocar la activación, inactivación,

duplicación e incluso la aparición de un nuevo gen. Las secciones del genoma ricas en elementos transposicionales (TE) se asocian con alteraciones epigenéticas y diversos patrones de transcripción y acumulación de mutaciones. En conjunto, la presencia y las acciones de los TE promueven la variabilidad, la adaptabilidad y la resistencia a diferentes factores en los genes del cacao, como se muestra en la Tabla 2. La cuenta en red (count in network) se representa en la cantidad de proteínas que están anotadas con el primer número y el segundo número indica cuántas proteínas existen en total (en la red y en segundo plano). A partir de esta relación se puede comprobar que tan grande es el efecto de enriquecimiento (strength) y cuán significativo puede ser el enriquecimiento (false discovery rate).

Tabla 2 Enriquecimiento de los procesos biológicos con respecto a los 249 genes.

Biological Process (Gene Ontology)			
description	*count in network*	*strength*	*false discovery rate*
Transposition, rna-mediated	8 of 53	1.25	6.38e-05
DNA integration	14 of 118	1.14	1.71e-08
RNA-dependent DNA biosynthetic process	15 of 135	1.11	1.71e-08
DNA biosynthetic process	16 of 169	1.05	1.71e-08
DNA recombination	10 of 202	0.76	0.0078
Nucleic acid phosphodiester bond hydrolysis	17 of 374	0.73	6.38e-05
DNA metabolic process	19 of 566	0.59	0.00054
Nucleobase-containing compound biosynthetic process	17 of 635	0.5	0.0226
Aromatic compound biosynthetic process	20 of 885	0.42	0.0443

Identificación de genes candidatos

De los 249 genes identificados, el 19% fueron valorados en relación al crecimiento de la planta, el 27% en su resistencia contra patógenos, el 18% en su respuesta a estreses por condiciones ambientales y un 4% para posibles mejoras en la producción. Asimismo, se encontró que un 18% de los genes no estaban caracterizados y un 14% eran genes putativos (segmentos de ADN que se cree que son genes, pero cuya función exacta aún no se conoce) (Figura 3). Además, se estudió la concurrencia de estos genes, pero se encontró que los más específicos del Theobroma cacao resultaron ser los no caracterizados.

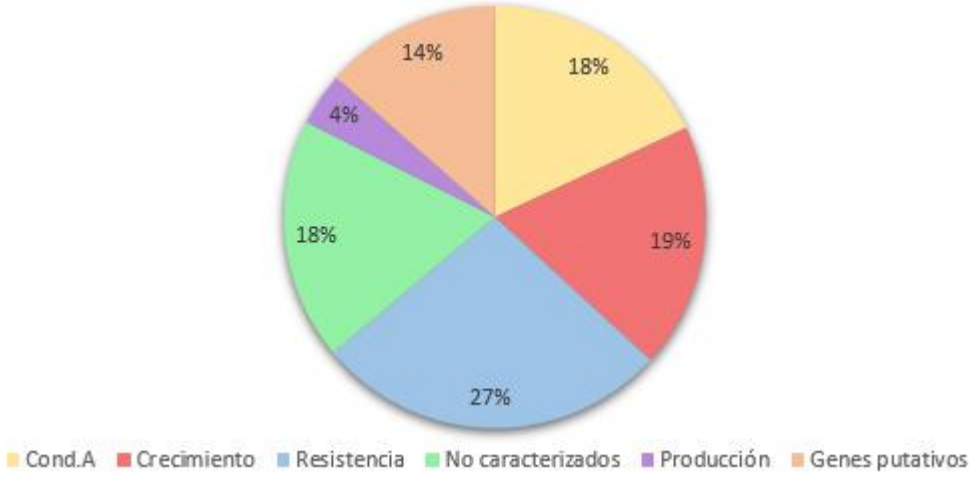

Figura 3 Características morfológicas relacionadas con los genes del cacao.

Cuando las plantas se encuentran bajo situaciones de estrés abiótico, se ha demostrado que los genes que pertenecen a las familias de transportadores y almacenamiento vacuolar de azúcares son altamente eficientes en su respuesta para mantener el metabolismo celular. Además, las proteínas modulares de unión de ARN también influyen en diversos aspectos de la expresión génica relacionados con el crecimiento de los orgánulos y el núcleo. Estas proteínas tienen la capacidad de alterar la secuencia, recambio, procesamiento y traducción del ARN, así como intervenir en el proceso de replicación del ADN y las ARN.

Los genes que controlan las características de color en el cacao son de gran interés debido a su importancia en la atracción de polinizadores y en la producción de pigmentos que protegen a la planta contra el estrés ambiental. Entre las familias de genes involucrados se encuentran las proteínas de repetición pentatricopeptídica, polimerasas de ADN/ARN y la familia de membrana MBOAT (membrane-bound O-acyltransferase). Cuando las plantas enfrentan situaciones de estrés, se produce una respuesta que induce la transcripción, traducción de genes y la estimulación de las anexinas, especialmente en respuesta a factores como la sequía, el frío, el estrés por calor, los rayos ultravioleta, la salinidad, el estrés oxidativo y el estrés mecánico.

Los genes encontrados relacionados con transportadores de iones metálicos, como el hierro (Fe), el zinc (Zn), el magnesio (Mg) y el cobre (Cu), son esenciales para el crecimiento saludable de la planta ya que se requieren para funciones tanto estructurales como catalíticas en las proteínas involucradas en su metabolismo y desarrollo. En particular, cabe resaltar la importancia del transporte de Zn para el metabolismo del cacao. De igual manera, los

transportadores MATE (familia de proteína extrusión multiantimicrobiana) funcionan como antiportadores en la tolerancia a la toxicidad del aluminio, la desintoxicación de metales pesados y el transporte de metabolitos secundarios, como antocianidinas, flavonoides y hormonas (Hekkehard Neuhaus, 2007).

La mayoría de los metabolitos secundarios se acumulan en forma de glicoconjugados y se producen como un mecanismo de defensa contra el estrés biótico y abiótico para proporcionar a las plantas armas antimicrobianas y antioxidantes. Estos genes, mayormente presentes en los cotiledones y endospermos de las semillas, constituyen del 2 al 10%, como es el caso de la familia de proteínas de las lectinas (proteínas de unión al azúcar que son extremadamente específicas para sus moléculas de azúcar). Además, los genes de resistencia en el cacao son codificados por proteínas ricas en leucina del sitio de unión a nucleótidos (NBS-LRR). Estas proteínas están involucradas en la detección de diversos patógenos, especialmente Phytophthora. Otro gen de gran importancia es el Gen TCM_032155, relacionado con la senescencia en las plantas. Este gen controla la ruptura de la pared celular durante la senescencia, lo que conduce a la descomposición de los tejidos vegetales y la liberación de nutrientes para ser reutilizados por la planta (Leah McHale, 2006) (Stefan Royaert, 2016).

En este contexto, las β-glucosidasas juegan un papel crucial en la liberación enzimática de compuestos aromáticos provenientes de precursores glucosídicos. Los genes con estas características, en la industria de los sabores, contienen hexosiltransferasas, pertenecientes a la familia de las glicosiltransferasas, son un grupo sumamente diverso de enzimas que transfieren un azúcar desde un donador activado a una molécula aceptora. Estas enzimas están implicadas en la síntesis y modificación de la amplia variedad de glucoconjugados presentes en las plantas.

La glucosilación desempeña un papel importante en la estabilización y percepción de aromas y sabores, especialmente en relación al amargor y dulzor. En consecuencia, la glucosilación es esencial para la estabilización de los derivados volátiles de terpenoides, flavonoides y otros compuestos fenólicos. Además, la expansión selectiva de algunas familias de genes durante la evolución -como los genes relacionados con los flavonoides- ha proporcionado una importante fuente de genes candidatos para mejorar los procesos industriales. De hecho, estos genes, con todas estas características, tienen un enorme potencial en la industria de procesamiento de alimentos, utilizándose como enzimas de sabor (Gopal Singh ,2015).

Conclusiones

En conclusión, se logró identificar genes candidatos en el genoma del cacao, lo que permitió conocer mejor la variedad tradicional cubana de este cultivo y realizar un estudio genético detallado. Se encontró que la mayoría de los

genes presentan elementos transponibles y se identificaron genes de resistencia codificados por proteínas ricas en leucina del sitio de unión a nucleótidos, involucrados en la detección de la Phytophthora. Se determinó que en situaciones de estrés abiótico, los genes de familias de transportadores y almacenamiento vacuolar de azúcares son más eficientes en la respuesta de la planta, por su removilización para el mantenimiento del metabolismo celular. Asimismo, se encontró que el estrés en las plantas induce la transcripción, traducción de genes y la estimulación de las anexinas, especialmente en respuesta a la sequía, el frío, el estrés por calor, los rayos ultravioleta, la salinidad, el estrés oxidativo y mecánico. Además, se identificó otro gen de gran importancia en el cacao, el Gen TCM_032155, relacionado con la senescencia en las plantas. La identificación de este gen es relevante para entender los procesos biológicos que ocurren en el cacao y cómo estas plantas se adaptan a diferentes condiciones ambientales. Finalmente, se identificaron genes relacionados con los flavonoides, que tienen un gran potencial en las industrias de procesamiento de alimentos y se utilizan como enzimas de sabor, lo que proporciona una fuente importante de genes candidatos para el mejoramiento del cacao. En conjunto, los resultados obtenidos en esta investigación pueden ser utilizados para el desarrollo de estrategias de mejoramiento genético en el cacao, así como para la implementación de prácticas agrícolas más eficientes y sostenibles.

Referencias

Jorge León (2000). *Botánica de los cultivos tropicales* . 3era edición. Instituto Interamericano de Cooperación para la Agricultura, San José, Costa Rica : Agroamerica.

Ricaño-Rodríguez (2018). *El estudio genómico del cacao(Theobroma cacao L.).* Agro Productividad. Vol. 11, no. 9.

Igor Bidot Martínez, Yurelkys Fernández Maura, Pierre Bertin, Heide-Marie Daniel, Stephan Declerck and Cony Decock (2021). *Diversidad morfológica, genética y fitopatología del cacao (Theobroma cacao L.) tradicional cubano*. Universidad de Guantánamo. Guantánamo, Cuba , Universidad Católica de Lovaina. Louvain-la-Neuve, Bélgica , Universidad de La Habana. La Habana, Cuba.

Osorio-Guarin, J. A., Quackenbush, C. R., & Cornejo, O. E. (2018). *Ancestry informative alleles captured with reduced representation library sequencing in Theobroma cacao*. *PLoS One*, 13(10), e0203973. doi:10.1371/journal.pone.0203973.

Krueger, F. (2017). *Trim Galore: a wrapper script to automate quality and adapter trimming.* http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Catchen, Julian, Hohenlohe, Paul A., Bassham, Susan, Amores, Angel and Cresko, William A (2013). *Stacks: an analysis tool set for population genomics*. Molecular Ecology. 2013. Vol. 22, no. 11.

Kim, Changdae, Koh, Kwangwon, Taehoon, Han, Daegyu and Seo, Jiwon (2023). *BWA-MEM-SCALE: Accelerating Genome Sequence Mapping on Commodity Servers. In : Proceedings of the 51st International Conference on Parallel Processing*. New York, NY, USA : Association for Computing Machinery. DOI 10.1145/3545008.3545033.

Bushnell, Brian, Rood, Jonathan and Singer, Esther(2017). *BBMerge – Accurate paired shotgun read merging via overlap*. DOI 10.1371/journal.pone.0185056.

Wilm, Andreas, Aw, Pauline Poh Kim, Bertrand, Denis, Yeo, Grace Hui Ting, Ong, Swee Hoe, Wong, Chang Hua, Khor, Chiea Chuen, Petric, Rosemary, Hibberd, Martin Lloyd and Nagarajan, Niranjan (2012). *LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets.* DOI 10.1093/nar/gks918.

Danecek, Petr, Bonfield, James K, Liddle, Jennifer, Marshall, John, Ohan, Valeriu, Pollard, Martin O, Whitwham, Andrew, Keane, Thomas, Mccarthy, Shane A, Davies, Robert M and LI, Heng (2021). *Twelve years of SAMtools and BCFtools. GigaScience*. DOI 10.1093/gigascience/giab008.

Hansen, Nancy F (2016). *Variant Calling From Next Generation Sequence Data*. New York, Methods in Molecular Biology. ISBN 978-1-4939-3576-5.

Cingolani, Pablo, Platts, Adrian, Wang, Le Lily, Coon, Melissa, Nguyen, Tung, Wang, Luan, Land, Susan J., LU, Xiangyi, Ruden and Douglas M (2012). *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff*. DOI 10.4161/fly.19695.

Damian, Kirsch, Rebecca, Koutrouli, Mikaela, Nastou, Katerina, Mehryary, Farrokh, Hachilif, Radja, Gable, Annika L., Fang, Tao, Doncheva, Nadezhda T., Pyysalo, Sampo, Bork, Peer, Jensen, Lars J. and Von Mering, Christian (2023) . *The String database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids.* DOI 10.1093/nar/gkac1000.

Coudert, Elisabeth, Gehant, Sebastien, Edouard, Pozzato, Monica, Baratin, Delphine, Teresa, Christian J A, Nicole, Alan (2023) *The Uniprot Consortium. Annotation of biologically relevant ligands in UniProtKB using ChEBI. Bioinformatics*. DOI 10.1093/bioinformatics/btac793.

Thomas , Paul D., Dust, Anushya, Tremayne, Laurent-Philippe and others (2022). *Panther: Making genome-scale phylogenetics accessible to all. Protein Science.* 2022. DOI 10.1002/pro.4218.

Hekkehard Neuhaus (2007)T*ransport of primary metabolites across the plant vacuolar membrane.* Technische Universitat Kaiserslautern, Postfach 3049, D-67653 Kaiserslautern, Germany

Leah McHale, Xiaoping Tan, Patrice Koehl and Richard W Michelmore(2006) *Plant NBS-LRR proteins: adaptable guards.* The Genome Center, University of California, Davis, CA 95616, USA. DOI:10.1186/gb-2006-7-4-212

Stefan Royaert,Johannes Jansen, Daniela Viana da Silva, Samuel Martins de Jesus Branco, Donald S. Livingstone III, Guiliana Mustiga, Jean-Philippe Marelli, Ioná Santos Araújo, Ronan Xavier Corrêa & Juan Carlos Motamayor (2016) *Identification of candidate genes involved in Witches' broom disease resistance in a segregating mapping population of Theobroma cacao L. in Brazil.*

Gopal Singh , A. K. Verma, Vinod Kumar(2015) *Catalytic properties, functional attributes and industrial applications of b-glucosidases.* DOI 10.1007/s13205-015-0328-z.