



Análisis Genómico y Reconstrucción Metabólica de la Variedad Colombia de *Solanum tuberosum* L. Grupo Phureja

Oscar Alexis Quintero López

Facultad de Ingeniería
Departamento de Ingeniería de Sistemas e Industrial
Bogotá D.C., Colombia
2025

Análisis Genómico y Reconstrucción Metabólica de la Variedad Colombia de *Solanum tuberosum* L. Grupo Phureja

Oscar Alexis Quintero López

Tesis presentada como requisito parcial para optar por el título de:
Magíster en Bioinformática

Director:

PhD. Andrés Mauricio Pinzón Velasco
Profesor Asociado - Instituto de Genética
Universidad Nacional de Colombia

Codirector:

PhD. Luis Francisco Becerra Galindo
Profesor Asociado - Facultad de Ciencias y Educación
Universidad Distrital Francisco José de Caldas

Línea de investigación:

Biología de Sistemas

Grupo de investigación:

GiBBS

BIOMOLc

Universidad Nacional de Colombia
Facultad de Ingeniería
Departamento de Ingeniería de Sistemas e Industrial
2025

Cita 01.

Todd C. Miller *et al.*

Commit que incorpora la "lecture" a sudo — 12 ene 2004, versión 1.6.8p2

"We trust you have received the usual lecture from the local System Administrator. It usually boils down to these three things:
#1) Respect the privacy of others.
#2) Think before you type.
#3) With great power comes great responsibility."

Mensaje «lecture» de sudo

Declaración

Me permito afirmar que he realizado ésta tesis de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados el presente texto. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los he reconocido en el presente trabajo. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de tesis.

Bogotá D.C., February 13, 2026

Oscar Alexis Quintero López

Agradecimientos

A mis padres y a toda mi familia, por su amor incondicional, su ejemplo de trabajo y su confianza permanente. Cada paso de este camino ha estado sostenido por su apoyo, sus palabras de aliento y los valores que me inculcaron desde siempre. Este logro es también suyo; sin ustedes, nada de esto habría sido posible.

Al profesor Andrés, por abrirme las puertas, por su paciencia y guía generosa, y por enseñarme a mirar la ciencia con rigor y humildad. Al profesor Francisco, por la oportunidad, por su acompañamiento cercano y por todo lo aprendido bajo su tutela. Gracias a ambos por la confianza, por las conversaciones que encendieron ideas y por el tiempo dedicado a formar no solo a un investigador, sino a una mejor persona.

Extiendo mi gratitud a quienes, de una u otra forma, me brindaron apoyo en los momentos decisivos: gracias por las lecturas atentas, los consejos honestos y la compañía que hace posible lo difícil.

Dedico este trabajo a mi familia, con la esperanza de que cada página sea un reflejo de su cariño y de todo lo que me han dado.

Listado de símbolos y abreviaturas

Abreviatura	Significado
AIEA	Agencia Internacional de Energía Atómica
ATP	Adenosín trifosfato
BAM	Binary Alignment Map (Mapa de alineamiento binario)
BLAST	Basic Local Alignment Search Tool (Herramienta básica de búsqueda de alineamientos locales)
BLASTn	Basic Local Alignment Search Tool for nucleotides
BLASTp	Basic Local Alignment Search Tool for proteins
BUSCO	Benchmarking Universal Single-Copy Orthologs
CDS	Coding DNA Sequence (Secuencia codificante de ADN)
COBRA	COntstraint-Based Reconstruction and Analysis
COBRApy	COBRA para Python
COL	Proyecto Técnico de Cooperación (código de proyecto AIEA)
DNA	Ácido desoxirribonucleico
EC	Enzyme Commission (Comisión de Enzimas)
FASTA	Formato de texto para secuencias biológicas
FASTQ	Formato de archivo para secuencias y calidades
FBA	Flux Balance Analysis (Análisis de balance de flujos)
FVA	Flux Variability Analysis (Análisis de variabilidad de flujo)
GEMs	Genome-Scale Metabolic Models (Modelos metabólicos a escala genómica)
GFF	General Feature Format (Formato general de características)
GO	Gene Ontology (Ontología de genes)
GPR	Gene-Protein-Reaction (Relación gen-proteína-reacción)
GTF	Gene Transfer Format (Formato de transferencia de genes)
HiFi	High Fidelity (Alta fidelidad) - tecnología de secuenciación
JSON	JavaScript Object Notation
KAAS	KEGG Automatic Annotation Server
KEGG	Kyoto Encyclopedia of Genes and Genomes (Enciclopedia de genes y genomas de Kyoto)
MoMA	Minimization of Metabolic Adjustment
NCBI	National Center for Biotechnology Information (Centro Nacional para la Información Biotecnológica)
ORF	Open Reading Frame (Marco de lectura abierto)
PacBio	Pacific Biosciences - plataforma de secuenciación
PDB	Protein Data Bank (Banco de datos de proteínas)
Pfam	Protein families database (Base de datos de familias de proteínas)

PMN	Plant Metabolic Network (Red Metabólica de Plantas)
QV	Quality Value (Valor de calidad)
RAVEN	Reconstruction, Analysis and Visualization of Metabolic Networks
SBO	Systems Biology Ontology (Ontología de biología de sistemas)
SBML	Systems Biology Markup Language (Lenguaje de marcado para biología de sistemas)
SNP	Single Nucleotide Polymorphism (Polimorfismo de nucleótido único)

Resumen

Análisis Genómico y Reconstrucción Metabólica de la Variedad Colombia de *Solanum tuberosum* L. Grupo Phureja

El cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja constituye un recurso fitogenético estratégico para Colombia, con aproximadamente 18,000 hectáreas cultivadas anualmente que representan el 15% de la producción nacional de papa criolla. Sin embargo, la ausencia de un genoma de referencia bien ensamblado y anotado, así como de un modelo metabólico a escala genómica específico para este cultivar diploide, limita significativamente la comprensión mecánica de los determinantes moleculares de su fenotipo y restringe el desarrollo de programas de mejoramiento genético eficientes.

Se realizó el ensamblaje *de novo* del genoma completo del cultivar 'Criolla Colombia' utilizando tecnología PacBio HiFi (15.2 Gb de datos, lecturas promedio de 8 kb), seguido de corrección de errores con Inspector v1.2, control de calidad con BUSCO y BlobToolKit, y andamiaje cromosómico con RagTag. Se ejecutó la predicción estructural de genes con AUGUSTUS y anotación funcional con eggNOG-mapper, alcanzando cobertura del 79.4% de las proteínas predichas con asignaciones KEGG, COG y números EC. La reconstrucción del modelo metabólico a escala genómica (GEM) se implementó mediante COBRApy y ModelSEEDpy, con curación iterativa, enriquecimiento vía APIs de KEGG, validación con MEMOTE y pruebas de viabilidad mediante análisis de balance de flujos (FBA).

Se obtuvo un ensamblaje diploide de alta calidad con dos haplotipos diferenciados (1.66 Gb, heterocigosidad 1.36%), completitud BUSCO superior al 95% en todos los niveles taxonómicos, contaminación mínima (<4.2%), y especificidad taxonómica del 96% en Solanaceae. La anotación funcional identificó 39,127 genes codificantes de proteínas con alta sintenia cromosómica respecto al genoma de referencia DM1-3 516 R44 v6.1. Se desarrolló el primer modelo metabólico específico del cultivar 'Criolla Colombia', multicompartimental y estequiométricamente consistente, con 1,063 reacciones bioquímicas, 901 metabolitos únicos y viabilidad computacional confirmada mediante FBA (valor objetivo de biomasa: 361.32).

Esta investigación generó el primer genoma completo y modelo metabólico específico del cultivar 'Criolla Colombia', llenando una brecha crítica en el conocimiento de la papa criolla diploide. La integración genómica-metabólica lograda establece un marco metodológico reproducible para la caracterización funcional de cultivos andinos y demuestra la factibilidad de generar recursos de calidad internacional para genotipos locales. Los recursos constituyen herramientas fundamentales hacia un mejoramiento genético asistido por modelos, con potencial de impacto directo en la optimización de la papa criolla y la seguridad alimentaria regional.

Palabras clave: Ensamblaje genómico de alta fidelidad; Anotación funcional del genoma; Reconstrucción metabólica a escala genómica; Modelado computacional del metabolismo; Simulación de flujos metabólicos; *Solanum tuberosum* Grupo Phureja; Análisis de biomasa vegetal; Biología de sistemas vegetal.

Abstract

Genomic Analysis and Metabolic Reconstruction of the Colombia Variety of *Solanum tuberosum* L. Grupo Phureja

The 'Criolla Colombia' cultivar of *Solanum tuberosum* L. Grupo Phureja constitutes a strategic phyto-genetic resource for Colombia, with approximately 18,000 hectares cultivated annually representing 15% of the national criolla potato production. However, the absence of a well-assembled and annotated reference genome, as well as a genome-scale metabolic model specific to this diploid cultivar, significantly limits the mechanistic understanding of the molecular determinants of its phenotype and restricts the development of efficient genetic improvement programs.

De novo assembly of the complete genome of the 'Criolla Colombia' cultivar was performed using PacBio HiFi technology (15.2 Gb of data, average reads of 8 kb), followed by error correction with Inspector v1.2, quality control with BUSCO and BlobToolKit, and chromosomal scaffolding with RagTag. Structural gene prediction was executed with AUGUSTUS and functional annotation with eggNOG-mapper, achieving coverage of 79.4% of predicted proteins with KEGG, COG, and EC number assignments. Genome-scale metabolic model (GEM) reconstruction was implemented using COBRApy and ModelSEEDpy, with iterative curation, enrichment via KEGG APIs, validation with MEMOTE, and viability testing through flux balance analysis (FBA).

A high-quality diploid assembly with two differentiated haplotypes was obtained (1.66 Gb, heterozygosity 1.36%), BUSCO completeness greater than 95% at all taxonomic levels, minimal contamination (<4.2%), and 96% taxonomic specificity in Solanaceae. Functional annotation identified 39,127 protein-coding genes with high chromosomal synteny relative to the reference genome DM1-3 516 R44 v6.1. The first metabolic model specific to the 'Criolla Colombia' cultivar was developed, multicompartmental and stoichiometrically consistent, with 1,063 biochemical reactions, 901 unique metabolites, and computational viability confirmed through FBA (biomass objective value: 361.32).

This research generated the first complete genome and metabolic model specific to the 'Criolla Colombia' cultivar, filling a critical gap in the knowledge of diploid criolla potato. The achieved genomic-metabolic integration establishes a reproducible methodological framework for functional characterization of Andean crops and demonstrates the feasibility of generating international-quality resources for local genotypes. The resources constitute fundamental tools toward model-assisted genetic improvement, with potential for direct impact on criolla potato optimization and regional food security.

Keywords: High-fidelity genome assembly; Functional genome annotation; Genome-scale metabolic reconstruction; Computational metabolic modeling; Flux balance analysis; *Solanum tuberosum* Group Phureja; Plant biomass analysis; Plant systems biology.

Esta tesis de maestría se sustentó el 16 de diciembre de 2025 a las 9:00 a.m., y fue evaluada por los siguientes jurados:

PhD. Emiliano Barreto Hernandez
Profesor
Instituto de Biotecnología
Universidad Nacional de Colombia

PhD. Alejandro Caro Quintero
Profesor
Departamento de Biología
Universidad Nacional de Colombia

Lista de figuras

4-1	Morfología de una planta de <i>Solanum tuberosum</i> L.	8
5-1	Estructura jerárquica de un modelo SBML. El elemento central contiene seis componentes principales organizados por función: componentes estructurales (<i>Compartments, Species, Reactions</i>) que definen la arquitectura del modelo; elementos de configuración (<i>Parameters, Rules</i>) que establecen parámetros y restricciones matemáticas; y componentes dinámicos (<i>Events</i>) para simulaciones temporales. Esta organización modular facilita la interpretación y manipulación de GEMs por diferentes herramientas computacionales.	35
7-1	Análisis de sintenia global entre el genoma de referencia DM1-3 516 R44 v6.1 y el ensamblaje del cultivar 'Criolla Colombia'. El gráfico de puntos muestra la colinearidad cromosómica con regiones de alta conservación (líneas diagonales continuas) y eventos de reorganización. La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.	56
7-2	Evaluación de completitud BUSCO para transcritos predichos en diferentes grupos taxonómicos del genoma de <i>Solanum tuberosum</i> L. Grupo Phureja. Las barras apiladas muestran el porcentaje de BUSCOs completos únicos (S), completos duplicados (D), fragmentados (F) y faltantes (M).	58
7-3	Evaluación de completitud BUSCO para proteínas predichas en diferentes grupos taxonómicos del genoma de <i>Solanum tuberosum</i> L. Grupo Phureja. Las barras apiladas muestran el porcentaje de BUSCOs completos únicos (S), completos duplicados (D), fragmentados (F) y faltantes (M).	59
7-4	Distribución de lecturas mapeadas y no mapeadas durante la detección de contaminantes. El gráfico muestra que el 99.12% de las secuencias corresponden al genoma nuclear (unmapped), mientras que el 0.88% representa material potencialmente contaminante identificado durante el control de calidad.	61
7-5	Distribución taxonómica tipo Sankey desde la raíz filogenética hasta Viridiplantae. El diagrama muestra la clasificación jerárquica de las secuencias anotadas, confirmando su pertenencia al reino vegetal y la ausencia de contaminación significativa por otros grupos taxonómicos.	62
7-6	Distribución taxonómica detallada dentro de Solanaceae hasta <i>Solanum</i> . El diagrama de Sankey ilustra la clasificación específica de las secuencias dentro de la familia Solanaceae, mostrando una alta especificidad hacia el género <i>Solanum</i> como se esperaba para <i>Solanum tuberosum</i> L. Grupo Phureja.	63
7-7	Clasificación taxonómica tipo sunburst específica para <i>Solanum tuberosum</i> . La representación circular muestra la distribución jerárquica de las secuencias anotadas, desde los niveles taxonómicos superiores hasta la especie, confirmando la alta especificidad de la anotación para <i>Solanum tuberosum</i> L. Grupo Phureja.	64
7-8	Estructura del Repositorio de Modelado GEMs	71

E-1 Análisis de sintenia del chr01 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad:  1.0-0.75,  0.75-0.5,  0.5-0.25,  0.25-0.	97
E-2 Análisis de sintenia del chr02 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad:  1.0-0.75,  0.75-0.5,  0.5-0.25,  0.25-0.	98
E-3 Análisis de sintenia del chr03 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad:  1.0-0.75,  0.75-0.5,  0.5-0.25,  0.25-0.	99
E-4 Análisis de sintenia del chr04 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad:  1.0-0.75,  0.75-0.5,  0.5-0.25,  0.25-0.	100
E-5 Análisis de sintenia del chr05 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad:  1.0-0.75,  0.75-0.5,  0.5-0.25,  0.25-0.	101
E-6 Análisis de sintenia del chr06 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad:  1.0-0.75,  0.75-0.5,  0.5-0.25,  0.25-0.	102
E-7 Análisis de sintenia del chr07 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad:  1.0-0.75,  0.75-0.5,  0.5-0.25,  0.25-0.	103
E-8 Análisis de sintenia del chr08 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad:  1.0-0.75,  0.75-0.5,  0.5-0.25,  0.25-0.	104
E-9 Análisis de sintenia del chr09 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad:  1.0-0.75,  0.75-0.5,  0.5-0.25,  0.25-0.	105
E-10 Análisis de sintenia del chr10 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad:  1.0-0.75,  0.75-0.5,  0.5-0.25,  0.25-0.	106
E-11 Análisis de sintenia del chr11 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad:  1.0-0.75,  0.75-0.5,  0.5-0.25,  0.25-0.	107
E-12 Análisis de sintenia del chr12 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad:  1.0-0.75,  0.75-0.5,  0.5-0.25,  0.25-0.	108

Lista de tablas

4-1	Clasificación Taxonómica de la Papa Criolla (<i>Solanum tuberosum</i> L. Grupo Phureja 'Criolla Colombia')	7
6-1	Análisis de Calidad y Rendimiento de ADN Extraído de Muestras de <i>Solanum tuberosum</i> L. Grupo Phureja 'Criolla Colombia'	38
7-1	Características estructurales comprensivas del ensamblaje del cultivar 'Criolla Colombia'	54
7-2	Análisis cromosómico integrado del cultivar 'Criolla Colombia'	55
7-3	Resultados de posicionamiento cromosómico y validación del scaffolding	57
7-4	Validación de calidad comprensiva del ensamblaje del cultivar 'Criolla Colombia'	60
7-5	Resultados de la predicción estructural de genes con AUGUSTUS	65
7-6	Resultados generales de la anotación funcional con eggNOG-mapper	66
7-7	Distribución de las principales categorías COG en la anotación funcional.	66
7-8	Archivos de salida generados por eggNOG-mapper v2.1.12.	66
7-9	Características comprensivas del modelo GEM del cultivar 'Criolla Colombia' versión 1.9.	68
7-10	Comparación estructural entre modelos metabólicos de <i>S. tuberosum</i>	69
7-11	Métodos de expansión de cobertura génica en el modelo Criolla v1.9	69
7-12	Mejoras funcionales y estructurales implementadas en el modelo Criolla v1.9.	70
7-13	Características de especificidad del cultivar en el modelo Criolla Colombia v1.9	70
7-14	Estructura de directorios principales del repositorio GEM_Creole	72
7-15	Sistema de reportes y documentación del repositorio.	72
7-16	Etapas del pipeline secuencial principal de modelado metabólico	73
7-17	Pipelines de integración con bases de datos externas	74
7-18	Sistema de validación y control de calidad del repositorio	75
7-19	Sistema de gestión de versiones y trazabilidad computacional	76
7-20	Implementación de principios FAIR en el repositorio GEM_Creole.	76
A-1	Análisis Granulométrico y Características Básicas del Suelo	86
A-2	Complejo de Intercambio Catiónico y Bases Intercambiables	87
A-3	Acidez Intercambiable y Porcentajes de Saturación.	87
A-4	Plan de Fertilización Edáfica para Papa Criolla	88

A-5 Enmiendas y Correctivos Aplicados	88
---	----

Contenido

Agradecimientos	ii
Listado de símbolos y abreviaturas	iii
Resumen	v
Abstract	vi
Lista de figuras	viii
Lista de tablas	x
Contenido	xii
1 Introducción	1
2 Planteamiento del Problema	4
2.1 Pregunta de Investigación	4
3 Objetivos	5
3.1 Objetivo General	5
3.2 Objetivos Específicos	5
4 Estado del Arte	6
4.1 <i>Solanum tuberosum</i> L. Grupo Phureja	6
Origen, distribución y taxonomía 6 • Comparativa genómica: genoma referencia DM1-3 516 R44 vs. 'Criolla Colombia' 9 • Importancia agronómica y bases genéticas de rasgos clave 10	
4.2 Ensamblaje y Anotación de Nuevos Genomas de Plantas	11
Tecnologías de secuenciación en plantas 11 • Herramientas de ensamblaje y evaluación de calidad 12 • Anotación estructural y funcional: pipelines, bases de datos 13	
4.3 Metabolitos y Rutas de Interés Agronómico	14
Síntesis de carbohidratos y formación de almidón en tubérculos 14 • Biosíntesis de compuestos nutraceuticos y fitohormonas 15 • Mecanismos de respuesta a estrés abiótico (salinidad, sequía, temperatura) 16	
4.4 Aplicaciones en Fitomejoramiento Asistido por Modelos	18
Predicción de fenotipos metabólicos y optimización de flujo para mayor biomasa 18 • Ejemplos en especies	

modelo (Arabidopsis, maíz) y transferibilidad a papa criolla 19

5	Marco Teórico	21
5.1	Fundamentos de Bioinformática y Biología de Sistemas	21
	Concepto de sistema biológico y propiedad de emergencia en rutas metabólicas 21 • Modelos metabólicos a escala genómica (GEMs): definición, ventajas y limitaciones 22 • Principios de la modelación in silico: balance estequiométrico y análisis de flujo (FBA) 23	
5.2	Reconstrucción de Modelos Metabólicos a Escala Genómica.	23
	Fuentes de reacciones y metabolitos 24 • Estrategias de ensamblaje automático vs. curación manual de rutas 24 • Formulación de la reacción de biomasa: componentes, definición de coeficientes y justificación biológica 25 • Herramientas de modelado: COBRA Toolbox, COBRAPy, RAVEN, libSBML 26	
5.3	Validación y Evaluación de Calidad del Modelo GEM.	27
	Control de consistencia estequiométrica y termodinámica 27 • Benchmarks de calidad: MEMOTE y BlobToolKit para metadatos y métricas de modelo 29 • Integración de datos ómicos para refinar y validar predicciones 30	
5.4	Buenas Prácticas y Estándares en Reconstrucción Metabólica	32
	Normas MIRIAM para anotación y documentación de modelos 32 • Uso de Identificadores Únicos (InChI, ChEBI, Rhea) 33 • Formatos de Intercambio y Estandarización: SBML y Ontologías 34 • Versionado, trazabilidad y repositorio de datos (Git, FAIR data) 35	
6	Materiales y Métodos	37
6.1	Materiales.	37
	Muestras biológicas 37 • Reactivos y kits 38 • Recursos computacionales 39 • Equipos e infraestructura 39 • Software y bases de datos 40	
6.2	Métodos	40
	Ensamblaje <i>de novo</i> 41 • Evaluación de Ensamblaje 42 • Corrección y limpieza 43 • Anotación Estructural y Funcional 44 • Construcción de andamiaje cromosómico 45 • Reconstrucción metabólica (GEM) 48	
7	Resultados	53
7.1	Ensamblaje del Genoma	53
	Características Estructurales del Ensamblaje 53 • Análisis Cromosómico Detallado 54 • Análisis de Sintenia 55 • Resultados del Andamiaje Cromosómico 56 • Evaluación de Completitud BUSCO 57 • Validación de Calidad Integrada 59 • Anotación Funcional 64	
7.2	Reconstrucción Metabólica.	67
	Características del Modelo Metabólico 67 • Análisis Comparativo con Modelos Previos 69	
7.3	Repositorio Git	71
	Arquitectura y Organización del Repositorio 71 • Workflow de Modelado 73 • Sistema de Validación y Control de Calidad 74 • Gestión de Versiones y Trazabilidad 75 • Cumplimiento con Principios FAIR 76	
8	Discusión de resultados	78
8.1	Síntesis integradora y respuesta a la pregunta de investigación	78
8.2	Cumplimiento de objetivos	78
8.3	Calidad y validez de la base genómica	79
8.4	Evaluación del modelo metabólico y comparación con el estado del arte.	79
8.5	Implicaciones biológicas y agronómicas	80
8.6	Posicionamiento en el contexto científico internacional	81
8.7	Limitaciones y amenazas a la validez.	81

8.8	Perspectivas y trabajo futuro	81
9	Conclusiones	83
10	Recomendaciones	84
A	Análisis Físicoquímico de Suelos	86
A.1	Caracterización Básica del Suelo	86
A.2	Complejo de Intercambio Catiónico	86
A.3	Acidez Intercambiable y Saturación	87
A.4	Plan de Fertilización Aplicado	87
	Fertilizantes Aplicados 88 • Enmiendas y Correctivos 88	
A.5	Interpretación y Recomendaciones	88
B	Comandos Detallados de Ensamblaje	89
B.1	Preparación de Datos	89
B.2	Ensamblaje con Hifiasm	89
B.3	Evaluación de Ensamblaje	89
C	Comandos Detallados de Anotación	91
C.1	Predicción Estructural con AUGUSTUS	91
C.2	Anotación Funcional con eggNOG-mapper	91
C.3	Corrección de Errores	92
C.4	Control de Calidad y Filtración	92
D	Comandos Detallados de Scaffolding	94
D.1	Andamiaje con RagTag	94
D.2	Evaluación de continuidad del scaffolding	94
D.3	Validación de sintenia con D-GENIES	95
E	Análisis de Sintenia por Cromosoma	96
E.1	Cromosoma 1	97
E.2	Cromosoma 2	98
E.3	Cromosoma 3	99
E.4	Cromosoma 4	100
E.5	Cromosoma 5	101
E.6	Cromosoma 6	102
E.7	Cromosoma 7	103
E.8	Cromosoma 8	104
E.9	Cromosoma 9	105
E.10	Cromosoma 10	106
E.11	Cromosoma 11	107
E.12	Cromosoma 12	108

F Recursos Computacionales y Hardware	109
F.1 Clúster de alto rendimiento BYU.	109
F.2 Servidor del Instituto de Genética UNAL.	109
F.3 Secuenciador PacBio.	109
F.4 Equipos de laboratorio.	110
G Condiciones Ambientales de Laboratorio	111
Referencias Bibliográficas	112

1 Introducción

La seguridad alimentaria global enfrenta desafíos sin precedentes en el siglo XXI, donde el crecimiento poblacional, el cambio climático y la degradación de recursos naturales demandan estrategias innovadoras para incrementar la producción agrícola de manera sostenible (FAO, 2024). En este contexto, los cultivos andinos representan un reservorio genético invaluable, adaptado a condiciones ambientales extremas y con alto valor nutricional, constituyendo una alternativa estratégica para la diversificación y sostenibilidad de los sistemas productivos (Zagorščak et al., 2024). La papa criolla (*Solanum tuberosum* L. Grupo Phureja 'Criolla Colombia') ejemplifica estos recursos fitogenéticos, siendo fundamental para la seguridad alimentaria, la economía rural y la cultura agrícola de Colombia y la región andina, con aproximadamente 18,000 hectáreas cultivadas anualmente que representan el 15% de la producción nacional de papa criolla (Ñústez López & Rodríguez Molano, 2024).

Los avances en genómica funcional y biología de sistemas han revolucionado el entendimiento de los cultivos, permitiendo el desarrollo de modelos predictivos que integran información genómica con fenotipos complejos (Weckwerth, 2011). La reconstrucción metabólica a escala genómica (GEMs, por sus siglas en inglés) emerge como una herramienta poderosa de la bioinformática y la biología de sistemas, construyendo modelos computacionales que representan la red completa de reacciones bioquímicas de un organismo a partir de su información genética (Gao & Zhao, 2024). Estos modelos permiten simulaciones *in silico* del comportamiento metabólico bajo diversas condiciones, facilitando el diseño racional de estrategias de mejoramiento genético. Mientras que en plantas modelo como *Arabidopsis thaliana* ya se han establecido GEMs que utilizan algoritmos de optimización para predecir distribuciones de flujo metabólico (Williams et al., 2010), actualmente no existe un modelo equivalente para la papa criolla, limitando nuestra capacidad para comprender y optimizar integralmente su metabolismo.

A pesar de su relevancia estratégica, el conocimiento sobre la biología molecular del cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja presenta limitaciones críticas que restringen el desarrollo de programas de mejoramiento genético eficientes (Crookshanks et al., 2001; Palit et al., 2020). Específicamente, no existe un genoma de referencia bien ensamblado y anotado para este cultivar diploide, ni un modelo metabólico a escala genómica que permita comprender los mecanismos moleculares que determinan su fenotipo. Esta ausencia de recursos genómicos y metabólicos limita significativamente: (i) la identificación de genes y rutas asociados a rasgos agronómicos clave como rendimiento, calidad nutricional y tolerancia a estreses; (ii) la detección de cuellos de botella metabólicos que condicionan la productividad; y (iii) la evaluación reproducible de hipótesis mediante simulaciones computacionales (Haggart et al., 2011; Reijnders et al., 2014). Consecuentemente, las decisiones de mejoramiento se apoyan predominantemente en criterios fenotípicos y conocimiento transferido entre especies, con riesgo de sesgos por divergencia evolutiva y limitaciones en la comprensión mecanística del fenotipo.

La complejidad genómica de Solanaceae, caracterizada por alto contenido repetitivo, variantes estructurales

y heterocigosidad, exige datos de secuenciación de alta fidelidad y metodologías especializadas para el ensamblaje y anotación (Cheng et al., 2021). Simultáneamente, la integración de 157,407 anotaciones génicas heterogéneas y la construcción de modelos metabólicos específicos de cultivar demanda pipelines reproducibles, criterios de calidad rigurosos y validación experimental de las predicciones generadas mediante análisis de flujo metabólico (FBA) (Lieven et al., 2020). Superar estos desafíos técnicos y conceptuales requiere un enfoque colaborativo y un firme compromiso con la ciencia abierta, promoviendo la compartición de datos y modelos para acelerar el progreso científico y tecnológico en el ámbito agrícola.

En respuesta a esta problemática, el presente trabajo plantea la integración del ensamblaje y caracterización genómica del cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja con una reconstrucción metabólica a escala genómica que explique los mecanismos moleculares que determinan su fenotipo. Esta investigación se desarrolla en el marco del Proyecto Técnico de Cooperación COL5026 de la Agencia Internacional de Energía Atómica (AIEA), titulado "Fortalecimiento de Capacidades en Técnicas Nucleares para el Mejoramiento de Cultivos", el cual proporciona el contexto institucional y técnico para implementar estudios que fortalezcan el uso de plataformas de secuenciación de tercera generación (HiFi-PacBio) y modelado metabólico basado en restricciones (FBA/COBRA) en la optimización de cultivos estratégicos. La colaboración establecida con el Prof. Jeffrey Maughan de la Universidad Brigham Young (BYU, Utah, Estados Unidos) ha sido fundamental para el acceso a infraestructura computacional de alto rendimiento y expertise en ensamblaje genómico, facilitando la transferencia de metodologías especializadas y demostrando el carácter multidisciplinario y global de la investigación contemporánea en genómica funcional.

El objetivo general de esta tesis es describir el metabolismo que determina las características fenotípicas del cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja a través de una reconstrucción metabólica a escala genómica. Para alcanzar este propósito, se plantearon cuatro objetivos específicos: (i) identificar la secuencia genómica completa del cultivar mediante ensamblaje *de novo* con tecnología PacBio HiFi; (ii) caracterizar funcionalmente la secuencia genómica obtenida mediante anotación estructural y funcional comprehensiva; (iii) construir una representación computacional del metabolismo basada en la información derivada del genoma anotado; y (iv) precisar el grupo de reacciones involucradas en el aumento de biomasa del cultivar. La hipótesis de investigación postula que la integración del ensamblaje genómico de alta calidad con anotación funcional comprehensiva permite la construcción de un modelo metabólico a escala genómica consistente y funcionalmente viable, que explica los mecanismos moleculares determinantes del fenotipo y habilita la predicción *in silico* de estrategias de optimización metabólica.

La metodología sigue un enfoque secuencial que integra tecnologías de vanguardia en genómica y biología de sistemas. Primero, se realizó el ensamblaje *de novo* del genoma utilizando lecturas PacBio HiFi procesadas con Hifiasm, seguido de corrección de errores con Inspector, control de calidad mediante BUSCO y BlobTools, y andamiaje cromosómico guiado por referencia con RagTag. Posteriormente, se ejecutó la predicción estructural de genes con AUGUSTUS y la anotación funcional con eggNOG-mapper, proporcionando una cobertura del 79.4% de las proteínas predichas con asignaciones KEGG, COG y números EC. Finalmente, se implementó la reconstrucción del modelo metabólico utilizando COBRApy y ModelSEEDpy, con curación iterativa, enriquecimiento mediante APIs de KEGG, validación mediante MEMOTE y pruebas de viabilidad FBA. La reproducibilidad se garantiza mediante la implementación de un repositorio Git estructurado conforme a principios FAIR, con trazabilidad completa del proceso de construcción, versionado de modelos y criterios cuantitativos para promoción de borradores a versiones curadas.

La importancia científica de esta investigación radica en la generación del primer genoma completo y modelo metabólico específico del cultivar 'Criolla Colombia', llenando una brecha crítica en el conocimiento de la papa criolla diploide y estableciendo un precedente metodológico para la genómica funcional en cultivos andinos. La originalidad del trabajo se fundamenta en la integración sin precedentes de ensamblaje

genómico de alta fidelidad con reconstrucción metabólica computacional, creando un marco metodológico reproducible que habilita la comprensión mecanística del fenotipo y facilita el diseño racional de estrategias de mejoramiento asistido por modelos. El impacto potencial abarca tres niveles complementarios: (1) fortalecimiento de la investigación básica mediante la generación del primer genoma completo del cultivar y un modelo metabólico con 1,063 reacciones bioquímicas; (2) desarrollo de herramientas aplicadas para programas de mejoramiento asistido por marcadores moleculares y modelos predictivos; y (3) optimización del manejo agronómico en las 18,000 hectáreas de papa criolla cultivadas anualmente en Colombia.

El enfoque de biología de sistemas adoptado busca establecer las bases para un mejoramiento genético asistido por modelos, integrando información multi-ómica y modelado a escala genómica para identificar rutas y genes determinantes asociados a características agronómicas como el rendimiento, la calidad nutricional y la tolerancia a estreses. La reconstrucción metabólica del cultivar 'Criolla Colombia' ofrece la oportunidad de comprender con mayor profundidad las rutas metabólicas implicadas en procesos vitales, tales como la síntesis de compuestos de interés nutricional y la respuesta a factores de estrés abiótico (Weckwerth, 2011). A través de la modelación computacional de estas rutas, es posible detectar cuellos de botella y puntos críticos de regulación dentro del metabolismo de la planta, permitiendo intervenir de manera dirigida para favorecer tanto la productividad como la resiliencia frente a condiciones ambientales adversas.

Los recursos generados se constituyen como herramientas fundamentales para estudios de genética cuantitativa, mapeo de rasgos, cruzamientos y retrocruzamientos orientados al mejoramiento, posibilitando la identificación de loci asociados a rasgos de interés mediante metodologías de mapeo y asociación genómica (et al. Hardigan, 2017; Li et al., 2021). La disponibilidad de un genoma de referencia bien anotado permitirá identificar regiones génicas clave, analizar la heredabilidad de rasgos agronómicos y diseñar estrategias de selección más precisas, facilitando programas de mejoramiento genético dirigidos a aumentar el rendimiento en distintas zonas del país mediante la integración de enfoques de biología de sistemas y herramientas de análisis masivo de datos (Blanchard, 2004).

Esta tesis se organiza en ocho capítulos que proporcionan una hoja de ruta comprensiva: el estado del arte en genómica de papa y modelado metabólico vegetal (Capítulo 2), el marco teórico en ensamblaje genómico y reconstrucción metabólica (Capítulo 3), los materiales y métodos empleados (Capítulo 4), los resultados del ensamblaje genómico y caracterización del modelo metabólico (Capítulo 5), la discusión de resultados en el contexto científico internacional (Capítulo 6), y las conclusiones y recomendaciones para trabajo futuro (Capítulos 7-8). El modelo aquí presentado corresponde a una versión preliminar, concebida como un punto de partida abierto a la comunidad científica para su validación, refinamiento y ampliación futura, alineándose con las tendencias internacionales en ciencia abierta y colaboración interdisciplinaria.

En síntesis, esta investigación busca llenar un vacío concreto en la literatura científica y en la práctica agrícola, integrando una problemática local de gran relevancia con un enfoque de vanguardia en bioinformática y biología de sistemas. Los avances logrados resultan fundamentales para enfrentar los retos que plantea el cambio climático y la creciente demanda de alimentos, asegurando una agricultura más competitiva y sostenible. La naturaleza abierta y reproducible del trabajo promueve la colaboración científica y facilita la extensión de estas metodologías a otros cultivos andinos estratégicos, contribuyendo a la seguridad alimentaria y sostenibilidad agrícola en la región.

2 Planteamiento del Problema

La papa andina, representada por el cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja (conocida localmente como papa criolla), es estratégica para la economía rural y presenta potencial de exportación por su calidad organoléptica. Sin embargo, para esta variedad no existe un genoma de referencia bien anotado ni un modelo metabólico a escala genómica específico, lo que limita la comprensión mecanística de los determinantes de su fenotipo y restringe el diseño racional de estrategias de mejoramiento y manejo agronómico en los cerca de 18 000 ha cultivados anualmente en Colombia.

La ausencia de estos recursos dificulta: (i) priorizar rutas y genes asociados a rasgos agronómicos clave (rendimiento, calidad, tolerancia a estreses), (ii) identificar cuellos de botella metabólicos que condicionan la generación de biomasa y la síntesis de compuestos de interés, y (iii) evaluar de forma reproducible hipótesis mediante simulaciones *in silico* basadas en flujos metabólicos. En consecuencia, las decisiones de mejoramiento y manejo se apoyan predominantemente en criterios fenotípicos y conocimiento transferido entre especies, con riesgo de sesgos por divergencia evolutiva y por ensamblajes fragmentados de lecturas cortas que no resuelven regiones repetitivas ni la variabilidad estructural.

Existen, además, limitaciones técnicas y prácticas: la complejidad genómica de Solanaceae (contenido repetitivo, variantes estructurales) exige datos de alta fidelidad y cobertura; la predicción génica en especies no modelo y la integración de 157 407 anotaciones heterogéneas requieren validaciones de continuidad y completitud (por ejemplo, BUSCO) y controles de calidad taxonómica y funcional; la transferencia de anotaciones funcionales introduce incertidumbre; y la curación de reconstrucciones metabólicas específicas de cultivar demanda *pipelines* reproducibles y criterios de calidad (consistencia estequiométrica, balances de carga, pruebas de crecimiento y de “essentiality”). Si bien plataformas de tercera generación (HiFi-PacBio), marcos COBRA/FBA y recursos como eggNOG permiten abordar estas brechas, su articulación estandarizada y reproducible, orientada a este cultivar, sigue siendo un reto operativo que esta tesis aborda.

2.1 Pregunta de Investigación

La pregunta de investigación que guía este estudio es: ¿Cómo integrar el ensamblaje y la caracterización del genoma del cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja con una reconstrucción metabólica a escala genómica que explique los mecanismos que determinan su fenotipo?

3 Objetivos

3.1 Objetivo General

Describir el metabolismo que determina las características fenotípicas del cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja, a través de una reconstrucción metabólica a escala genómica.

3.2 Objetivos Específicos

- Identificar la secuencia genómica completa del cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja.
- Caracterizar funcionalmente la secuencia genómica obtenida del cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja.
- Construir una representación computacional del metabolismo del cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja, basada en la información derivada de su genoma anotado.
- Precisar el grupo de reacciones involucradas en el aumento de biomasa del cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja.

4 Estado del Arte

La papa criolla (*Solanum tuberosum* L. Grupo Phureja) es un grupo de variedades de papa diploide originarias de los Andes, ampliamente cultivadas desde el occidente de Venezuela hasta el centro de Bolivia (Ghislain et al., 2006). Este grupo, anteriormente considerado una especie independiente (*Solanum phureja*), ha sido re-clasificado taxonómicamente como parte de *Solanum tuberosum* L. dentro del cultivar-grupo Phureja (Huamán & Spooner, 2002; Ochoa Neves et al., 1990). Las papas del Grupo Phureja presentan gran importancia agrícola y cultural en países andinos como Colombia, donde constituyen un alimento básico y un producto de alto valor gastronómico. A continuación, se revisan los avances recientes en el conocimiento agronómico, genético y tecnológico de la papa criolla, incluyendo su clasificación, características, importancia socioeconómica, mejoramiento genético y desafíos fitosanitarios.

4.1 *Solanum tuberosum* L. Grupo Phureja

4.1.1 Origen, distribución y taxonomía

Los cultivares del Grupo Phureja de *Solanum tuberosum* L. derivan de una domesticación temprana en el altiplano andino, y constituyen linajes monofiléticos de papas diploides cultivadas en toda la cordillera, desde Venezuela hasta el norte de Argentina (Juyó et al., 2015; Lucas-Aguirre et al., 2025; Manrique-Carpintero et al., 2023). El análisis de marcadores SNP en la Colombian Central Collection (CCC) confirmó la presencia de dos subpoblaciones principales: los diploides del Grupo Phureja y los tetraploides del grupo Andígena, destacando la gran diversidad genética asociada a la domesticación andina (Berdugo-Cely et al., 2017; Berdugo-Cely et al., 2021).

En particular, el sur de Colombia (Nariño, Cundinamarca, Boyacá) se identifica como un centro de diversidad secundaria para este germoplasma, corroborado por el alto grado de heterocigosidad y subestructura observada en accesiones diploides (Berdugo-Cely et al., 2021; Dolničar & Bohanec, 2000; Juyó et al., 2015). Esta variabilidad genética convierte al Grupo Phureja en una fuente valiosa para estudios evolutivos y programas de mejoramiento.

4.1.1.1 Clasificación Taxonómica

Huamán and Spooner (2002) propusieron reclasificar *Solanum phureja* definida por Ochoa Neves et al. (1990), como parte de *Solanum tuberosum* L. Grupo Phureja, motivados por una baja diferenciación fenética entre clados cultivados diploides. Ese enfoque fue reforzado por evaluaciones genéticas subsiguientes, que consolidaron la unidad taxonómica en un solo grupo cultivado con grados de polimorfismo y carácter fenotípico similares (Berdugo-Cely et al., 2021; Bohórquez-Quintero et al., 2022; Diaz-Valencia et al., 2021; Manrique-Carpintero et al., 2023; Peña et al., 2015). Así, desde la categoría orden: Solanales, familia: Solanaceae, y género: Solanum, se reconoce el grupo cultivar *Solanum tuberosum* L. Grupo Phureja (Huamán & Spooner, 2002), y así a su vez el cultivar *Solanum tuberosum* L. Grupo Phureja 'Criolla Colombia' (Rodríguez et al., 2009). La jerarquía taxonómica completa del cultivar 'Criolla Colombia' del Grupo Phureja, desde el reino Plantae hasta su condición como cultivar dentro del grupo, se presenta en la tabla 4-1.

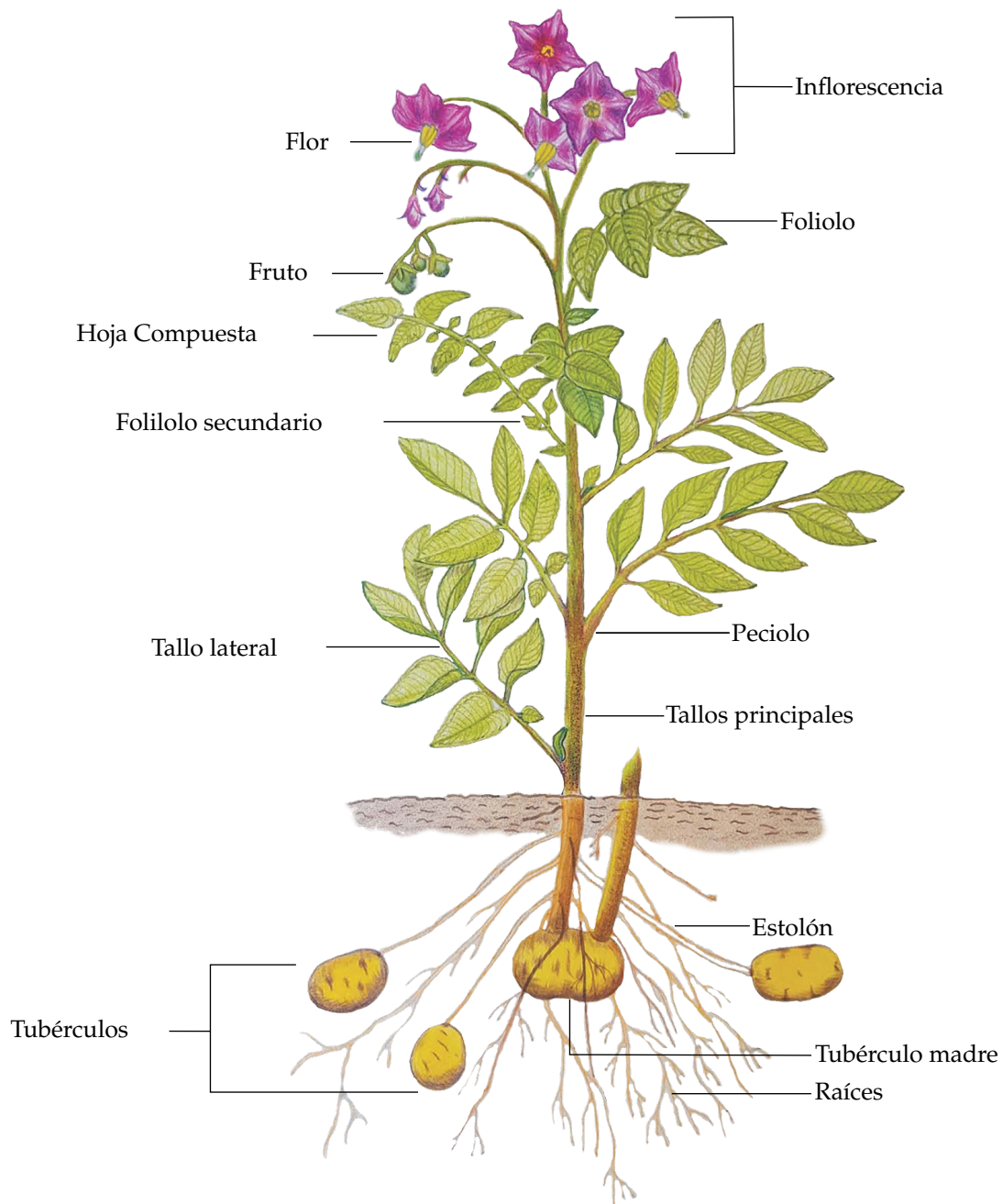
Reino	Plantae
División	Magnoliophyta
Clase	Magnoliopsida
Orden	Solanales
Familia	Solanaceae
Género	<i>Solanum</i>
Especie	<i>Solanum tuberosum</i> (Linnaeus, 1753)
Grupo	<i>Solanum tuberosum</i> Grupo Phureja (Huamán & Spooner, 2002)
Cultivar	<i>Solanum tuberosum</i> L. Grupo Phureja 'Criolla Colombia' (Rodríguez et al., 2009)

Tabla 4-1: Clasificación Taxonómica de la Papa Criolla (*Solanum tuberosum* L. Grupo Phureja 'Criolla Colombia')

4.1.1.2 Morfología

El Grupo Phureja se caracteriza por tallos erectos, hojas compuestas, inflorescencias cimosas y pedicelos articulados, en concordancia con descripciones anatómicas comparadas (Huamán & Spooner, 2002). Los tubérculos son de forma redondeada, piel y pulpa amarilla, sin periodo de dormancia post-cosecha, lo que propicia una brotación inmediata (Lucas-Aguirre et al., 2025; Rodríguez et al., 2009). Adicionalmente, estudios sobre morfofisiología en Phureja del sur de Colombia reportan tubérculos amarillos redondos, compactos y leves sin latencia, adaptados a fotoperíodo corto (Andrade-Díaz, 2024; Bohórquez-Quintero et al., 2022; Juyó et al., 2015; Lucas-Aguirre et al., 2025).

Estas características morfológicas y reproductivas (ausencia de dormancia) son fundamentales para la planificación de modelos GEM focalizados en rasgos de desarrollo vegetativo, almacenamiento de biomasa y adaptaciones metabólicas al fotoperíodo. La figura 4-1 ilustra estos caracteres morfológicos típicos, mostrando las principales estructuras anatómicas de una planta de *Solanum tuberosum* L., incluyendo el sistema radicular, tallo aéreo, hojas compuestas, inflorescencias y el desarrollo subterráneo de tubérculos, características típicas del grupo Phureja (véase Figura 4-1).



Nota. Adaptado de *Colores y Sabores de Mi Tierra: Papas Nativas Cultivadas en Boyacá* (p. 36), (Ojeda Pérez et al., 2021). Derechos de autor © 2021 por UPTC. <https://doi.org/10.19053/9789586605175>

Figura 4-1: Morfología general de una planta de *Solanum tuberosum* L., mostrando las estructuras vegetativas y reproductivas características.

4.1.2 Comparativa genómica: genoma referencia DM1-3 516 R44 vs. 'Criolla Colombia'

El primer genoma de referencia de papa criolla fue publicado en 2011 utilizando un clon monoploide duplicado derivado de *Solanum tuberosum* L. Grupo Phureja (accesión DM1-3 516 R44) (Hardigan et al., 2016). Este genoma de 844 Mb se ensambló inicialmente con lecturas cortas, lo que dejó regiones sin resolver. Avances tecnológicos recientes permitieron mejorar significativamente la calidad de este ensamble: una versión actualizada con lecturas largas de Nanopore y andamiaje Hi-C alcanzó 741.6 Mb secuenciados (87.8% del tamaño estimado) con 99% de la secuencia anclada en los 12 cromosomas (Pham et al., 2020). El nuevo ensamblaje redujo drásticamente la fragmentación (N50 de contig aumentó 595 veces) y contiene 33 mil genes codificantes de proteína anotados con alta completitud. Estos resultados establecen a DM1-3 516 R44 como un genoma de calidad referencia para papa diploide.

La variedad colombiana 'Criolla Colombia' pertenece igualmente al Grupo Phureja, pero a diferencia de DM1-3 516 R44 (una línea altamente endogámica y homocigota), es un cultivar heterocigoto seleccionado a partir de poblaciones locales de papa amarilla (Rodríguez et al., 2009). Por lo tanto, su genoma mantiene la estructura básica de 12 cromosomas y se espera un tamaño de aproximadamente $\sim 0.8\text{--}0.9$ Gb, pero con mayor variabilidad alélica. Estudios de diversidad muestran que las papas diploides nativas (Phureja, Stenotomum, Goniocalyx) comparten un acervo génico común y monofilético, aunque con alelos adaptativos particulares en cada región (Sturaro, 2025). Se espera que 'Criolla Colombia' conserve la mayoría de genes presentes en DM1-3 516 R44, pero con diferencias en secuencias reguladoras y en número de copias de ciertos genes relacionados con adaptaciones locales (resistencia a patógenos, calidad nutricional, etc.). De hecho, la descripción original de este cultivar indica que es una selección clonal de papas amarillas de Colombia con hábito de crecimiento erecto y tubérculos redondos de piel y pulpa amarilla. En rendimiento presenta $\sim 13\text{--}15$ toneladas/ha y desafortunadamente es susceptible a la gota (tizón tardío causado por *Phytophthora*) (Rodríguez et al., 2009), lo que sugiere que alelos de resistencia presentes en otras accesiones podrían estar ausentes o inactivos en su genoma.

Adicionalmente, la comparación con genomas de papa tetraploide provee contexto sobre diferencias genómicas relevantes. Un reciente ensamblaje de referencia de papa autotetraploide ($4n=48$ cromosomas) reveló un tamaño de ~ 2.67 Gb distribuidos en cuatro haplotipos por cromosoma. Este genoma tetraploide contiene más de 126 000 modelos génicos, correspondientes a $\sim 31\,500$ grupos de ortólogos; cerca del 19% de los genes presentan cuatro copias (una por haplotipo), mientras que otros genes han perdido duplicados y aparecen con una, dos o tres copias según reordenamientos y deleciones (Wang et al., 2022). Comparado con el diploide Phureja, el tetraploide exhibe expansión de algunas familias génicas (por ejemplo, genes de reparación de ADN como *MLH3* y *MSH6/7*), posiblemente como respuesta al incremento de ploidía (Pham et al., 2020; Wang et al., 2022). También se observaron variaciones estructurales asociadas a fenotipos agronómicos: genes de vías de flavonoides y glicósidos correlacionados con la coloración de la pulpa del tubérculo en diversas variedades comerciales (Wang et al., 2022).

Esto demuestra que, aunque el núcleo de genes esenciales es compartido, existen diferencias genómicas cuantitativas entre una papa diploide Phureja y las papas tetraploides cultivadas. En resumen, el genoma de 'Criolla Colombia'—una papa diploide andina— comparte la mayor parte de la información genética básica con la referencia DM1-3 516 R44 (Pham et al., 2020), pero estudios comparativos con otras papas sugieren que presentará variaciones alélicas únicas y posiblemente cambios en número de copias génicas que reflejan su adaptación y origen particular en los Andes colombianos.

4.1.3 Importancia agronómica y bases genéticas de rasgos clave

Las papas del Grupo Phureja, incluyendo la papa 'Criolla Colombia', tienen una gran importancia agronómica en los Andes por su corto ciclo de cultivo, altos contenidos nutricionales y valor gastronómico. A nivel agronómico, destacan por su rápida tuberización y falta de latencia prolongada de los tubérculos: a diferencia de las papas tetraploides de zonas templadas, las diploides andinas carecen de una verdadera dormancia post-cosecha (Ghislain et al., 2006). Esto significa que los tubérculos de papa criolla brotan poco tiempo después de la cosecha, lo cual permite varias cosechas al año en climas favorables, aunque dificulta su almacenamiento prolongado.

Genéticamente, la ausencia de dormancia en Phureja se ha asociado a niveles menores de ácido abscísico en tubérculos y a alelos particulares en loci reguladores de la latencia; estudios de ligamiento han identificado QTLs en el cromosoma 7 vinculados a la dormancia en cruces *Phureja* × *Chacoense*, indicando una base multigénica pero con algunos efectos mayores (Dogramaci et al., 2024). La heredabilidad de la dormancia en Phureja es alta (estimada > 0.7), reflejando que este rasgo distintivo está fuertemente controlado por la genética aditiva en estas poblaciones diploides (Haynes et al., 2019).

Otro rasgo clave es la calidad culinaria y nutricional de la papa criolla. Los cultivares de *Phureja* suelen presentar mayor materia seca y un almidón de cocción más suave que las papas comerciales tetraploides, lo cual redundo en texturas y sabores muy apreciados. Adicionalmente, muchas papas criollas tienen pulpa de color amarillo anaranjado intenso, indicativa de altos niveles de carotenoides (pigmentos nutraceuticos). Se ha reportado que accesiones de *Solanum tuberosum* L. Grupo Phureja pueden acumular de 3 a 5 veces más carotenoides totales que variedades blancas convencionales, predominando las xantófilas como luteína y zeaxantina. Curiosamente, en estas papas nativas se observó una correlación inversa entre β -caroteno y carotenoides totales: los genotipos de pulpa blanca tenían más β -caroteno que los de pulpa amarilla, en los cuales predominan otras xantófilas. Desde el punto de vista genético, la alta variación en contenido de carotenoides tiene heredabilidad moderada–alta y ha permitido identificar genes estructurales asociados. Por ejemplo, alelos del gen *beta-carotene hydroxylase* (que desvía β -caroteno hacia xantófilas) difieren entre papas amarillas y blancas, afectando el perfil final de pigmentos. Este amplio rango de variabilidad fenotípica en pigmentos hace del germoplasma *Phureja* un recurso valioso para biofortificación de vitamina A mediante mejoramiento convencional y biotecnológico (Sturaro, 2025).

En cuanto a la resistencia a enfermedades y adaptación, las papas criollas han sido tradicionalmente más susceptibles a algunas plagas como la gota (tizón tardío). 'Criolla Colombia' es un ejemplo: presenta alta sensibilidad a *P. infestans* (Rodríguez et al., 2009), lo que supone un desafío para su manejo agronómico. Sin embargo, la variabilidad genética presente en Phureja y parientes silvestres ofrece oportunidades para mejorar estos cultivares. Muchas diploides andinas no poseen ciertos genes R (de resistencia mayor) que sí están en especies silvestres; por ello, programas de mejoramiento han recurrido a introgresar resistencia desde parientes silvestres polinizando tetraploides con diploides (esquemas $4 \times - 2 \times$) para luego recuperar genomas diploides resistentes. Adicionalmente, la tolerancia a estrés abiótico en estas papas nativas no ha sido muy desarrollada comparada con variedades mejoradas: por ejemplo, su tolerancia a sequía o calor suele ser limitada debido a la adaptación original a climas de montaña con humedad adecuada. No obstante, su plasticidad genética podría explotarse en selección asistida. Se ha reconocido que los genes involucrados en rutas de señalización hormonal (ABA, etileno) y en protección celular (como chaperonas de estrés) varían entre genotipos andinos (Obidiegwu, 2015).

En general, la importancia agronómica de la papa criolla radica en su rápido ciclo y calidad de tubérculo, mientras que las bases genéticas de sus rasgos distintivos (ausencia de dormancia, coloración nutritiva, susceptibilidad a enfermedades) están siendo dilucidadas mediante estudios genómicos y de mapeo.

Esto permite orientar estrategias de mejoramiento para combinar el valor culinario de estos cultivares con mayores rendimientos y resistencia, por medio de cruces diploides asistidos por marcadores y eventualmente edición génica de loci clave una vez que el genoma de estos materiales esté completamente caracterizado (Sturaro, 2025).

4.2 Ensamblaje y Anotación de Nuevos Genomas de Plantas

La secuenciación y ensamblaje de genomas vegetales ha avanzado vertiginosamente en la última década, permitiendo descifrar genomas complejos de gran tamaño. En esta sección se revisan las tecnologías de secuenciación más empleadas en plantas, las herramientas para ensamble y evaluación, y los enfoques para anotación estructural/funcional de nuevos genomas, con énfasis en cultivos como *Solanum tuberosum* L.(papa).

4.2.1 Tecnologías de secuenciación en plantas

Las plantas suelen poseer genomas grandes y ricos en repeticiones, lo que planteó retos considerables en la era de la secuenciación Sanger. La introducción de la secuenciación de nueva generación (NGS) a partir de 2005 (principalmente plataformas Illumina de lecturas cortas de alto rendimiento) revolucionó la genómica vegetal al abaratar y agilizar la generación de datos. Sin embargo, las lecturas cortas (100–250 pb) dificultan ensamblar regiones repetitivas extensas típicas de genomas vegetales (como retrotransposones) y pueden fragmentar el ensamblaje. Por ello, para muchos genomas de plantas iniciales se adoptó una estrategia híbrida: secuenciación *shotgun* de genómica nuclear complementada con mapas de ligamiento, *contigs* BACs o mapas físicos para ordenar los fragmentos. Un ejemplo fue el primer genoma de papa (*Solanum tuberosum* DM1-3 516 R44) en 2011, ensamblado con millones de lecturas cortas pero quedando con cientos de brechas sin resolver (Pham et al., 2020).

En años recientes, las tecnologías de secuenciación de tercera generación han transformado la capacidad de ensamblaje al producir lecturas largas (de kilobases a incluso megabases). Plataformas como PacBio (SMRT sequencing) y Oxford Nanopore permiten leer moléculas de ADN intactas, abarcando regiones repetitivas y estructuralmente complejas. En plantas, donde los contenidos de ADN repetitivo pueden exceder el 50% del genoma, estas lecturas largas son cruciales para obtener *contigs* continuos. Por ejemplo, la actualización del genoma de referencia de papa *Phureja* con Nanopore logró aumentar el N50 de *contigs* > 500 veces al cerrar la mayoría de las brechas previas (Pham et al., 2020). De forma similar, otros cultivos (maíz, trigo) han pasado de ensambles altamente fragmentados con NGS a genomas casi completos gracias a PacBio de alta fidelidad (HiFi) o Nanopore de ultra larga lectura.

Otra innovación importante es la secuenciación por proximidad, como las librerías de mates de largo rango (10× Genomics) y la captura de conformación cromosomal (Hi-C). Estas técnicas no leen directamente el genoma, pero proporcionan información espacial que ayuda a ordenar y orientar los *contigs* en *superscaffolds* correspondientes a cromosomas completos. En el caso de papa, se utilizó Hi-C para anclar ~ 731 Mb del ensamble DM1-3 516 R44 v6.1 a los 12 cromosomas físicos, incluyendo la identificación de centrómeros mediante señal de secuencias repetitivas asociadas. Para genomas poliploides, también se emplean estrategias de separación de haplotipos (*haplotype phasing*). Por ejemplo, en papa tetraploide se ha logrado ensamblar los cuatro subgenomas por separado utilizando algoritmos que distinguen variantes heterocigotas y distribuyen las lecturas largas a distintos haplotipos. Tecnologías complementarias como

la cartografía óptica (BioNano Genomics) han sido aplicadas en algunos genomas vegetales para detectar macro-reordenamientos y validar la continuidad de ensamblajes (Pham et al., 2020).

4.2.2 Herramientas de ensamblaje y evaluación de calidad

El ensamblaje de un genoma vegetal típico requiere algoritmos capaces de manejar grandes volúmenes de datos y alta repetitividad. Las herramientas de ensamblaje se dividen en dos categorías generales: las orientadas a lecturas cortas (basadas en grafos de De Bruijn) y las diseñadas para lecturas largas (basadas en solapamiento y consenso). Para lecturas cortas de Illumina, programas clásicos como SOAPdenovo, ALLPATHS-LG o SPAdes se emplearon en muchos ensamblajes iniciales de plantas, construyendo grafos de *k*mers. Sin embargo, estos suelen generar *contigs* cortos en presencia de repeticiones mayores al tamaño de lectura. En cambio, para lecturas largas, ensambladores como Canu, Flye, FALCON o, más recientemente, HiCanu (optimizado para lecturas HiFi) son preferidos. Estos algoritmos detectan solapamientos reales entre lecturas largas para formar *contigs* extensos que a menudo abarcan genes completos y elementos repetitivos completos. En poliploides heterocigotos, algunos ensambladores incorporan rutinas para separar haplotipos durante el proceso (por ejemplo, Hifiasm tiene modo para tetraploides), evitando colapsar variantes alélicas distintas en una misma representación consensuada.

Tras obtener un ensamblaje preliminar, es esencial evaluar su calidad. Las métricas básicas incluyen el tamaño total ensamblado (que idealmente se acerca a la estimación del tamaño del genoma por métodos como citometría de flujo) y la continuidad, frecuentemente resumida en el N_{50} (longitud mínima tal que el 50% del genoma ensamblado está en *contigs* o *scaffolds* de esa longitud o mayor). Un N_{50} alto indica pocas fragmentaciones. No obstante, una métrica más informativa para genomas eucariotas es la completitud genómica estimada mediante conjuntos de genes ortólogos conservados. La herramienta BUSCO (Benchmarking Universal Single-Copy Orthologs) se ha vuelto estándar para cuantificar qué porcentaje de genes muy conservados (p. ej., ortólogos presentes en más del 90% de las plantas) aparecen completos en el ensamblaje (Pham et al., 2020). Un valor BUSCO alto (> 95%) sugiere que la mayoría de la información codificante está presente, aunque queden posibles brechas en regiones no génicas. Otras evaluaciones incluyen verificar la colinearidad con genomas de referencia si existen (por ejemplo, alineando *contigs* contra un cromosoma conocido para detectar inversiones o ausencias) y la identificación de posibles contaminantes (secuencias ajenas, típicamente microbianas, que puedan haberse ensamblado inadvertidamente). En el caso de *Solanum tuberosum* DM1-3 516 R44 v6.1, el ensamblaje final mostró un índice de calidad de ensamblaje LTR (LAI) de 13.5, que lo ubica en categoría de genoma de referencia (LAI > 10) por su nivel de completitud en elementos transponibles de tipo retroviral.

Un paso adicional para genomas poliploides es cuantificar la representación de cada haplotipo. Por ejemplo, en papa tetraploide se evaluó cuántos genes estaban presentes en 4 copias distintas vs. colapsados en menos copias (Wang et al., 2022). Esta evaluación ayuda a detectar regiones donde el ensamblaje fusionó haplotipos (subestimando el número real de alelos) o, por el contrario, los separó por completo. Idealmente, un buen ensamblaje poliploide presentará la mayor parte de genes con sus cuatro alelos separados, salvo en zonas muy conservadas donde las secuencias prácticamente idénticas pueden colapsarse. En resumen, la calidad de un ensamblaje de nuevo genoma vegetal se determina no solo por métricas de continuidad sino por su integridad genética. Las herramientas actuales permiten refinar ensamblajes automáticamente (pulido de secuencias con algoritmos como Pilon para corregir errores puntuales usando lecturas cortas, o Racon/Medaka en el caso de lecturas largas) y validar exhaustivamente la presencia de componentes esperados (genes esenciales, repeticiones centroméricas, organelos, etc.). Un ensamblaje confiable sienta las bases para la anotación y estudios posteriores sin sesgos significativos.

4.2.3 Anotación estructural y funcional: pipelines, bases de datos

Una vez ensamblado el genoma, el siguiente desafío es la anotación, es decir, identificar las posiciones y estructuras de los genes (anotación estructural) y asignarles funciones putativas (anotación funcional). En plantas, la anotación estructural enfrenta complicaciones como la presencia de intrones largos, familias génicas multigénicas y pseudogenes, así como la necesidad de distinguir genes muy similares derivados de duplicaciones. El enfoque moderno para anotación de genes es integrador o por evidencia múltiple. Herramientas como MAKER o BRAKER combinan predicciones *de novo* con alineamiento de secuencias de evidencia (transcriptomas o proteínas conocidas). Inicialmente, se ejecutan predictores *de novo*, como AUGUSTUS, GeneMark, GlimmerHMM que utilizan modelos entrenados para especies cercanas para esbozar posibles exones y genes en la secuencia genómica. Luego, se incorporan datos experimentales: alineamiento de lecturas de RNA-Seq o transcriptomas montados contra el genoma para delimitar exones verdaderos, así como comparación con proteínas de referencia (p. ej., de *Arabidopsis thaliana* o tomate) para validar marcos completos. En el caso de papa DM1-3 516 R44, la disponibilidad de lecturas de transcritos de longitud completa (cDNA de Nanopore) permitió anotar $\sim 32,917$ genes de alta confianza con un conjunto de modelos génicos mucho más completo que el obtenido con métodos previos (Pham et al., 2020). El uso de transcriptomas completos ayudó a definir UTRs, variantes de *splicing* y genes pequeños potencialmente pasados por alto.

Tras obtener la lista de genes candidatos, la fase funcional asigna identidades a cada gen. Esto típicamente involucra varias bases de datos:

- (1) Comparación contra proteínas conocidas (BLAST contra UniProt/SwissProt o contra proteínas de *Arabidopsis thaliana*) para transferir nombres o descripciones a genes con homología significativa.
- (2) Identificación de dominios proteicos conservados mediante InterProScan, que integra múltiples bases de datos de motivos (Pfam, PRINTS, PANTHER, PROSITE, etc.) y sugiere posibles funciones o familias.
- (3) Mapeo de los genes a vías metabólicas o categorías funcionales usando bases de datos como KEGG (vía la asignación de números de enzima EC y rutas metabólicas) o *Gene Ontology* (GO) para términos de función molecular, proceso biológico y componente celular.

Por ejemplo, para un nuevo genoma de papa criolla, se esperaría mapear sus genes metabólicos contra rutas como la de biosíntesis de almidón o carotenoides, comparando con esquemas conocidos en modelos de plantas. Existen también recursos específicos como PlantCyc y el Solanaceae Genome Network, que proporcionan referencias de vías metabólicas en solanáceas para anotar rutas especializadas (por ejemplo, alcaloides esteroidales en papa).

La calidad de la anotación funcional se mide a menudo por el porcentaje de genes con alguna asignación. En especies bien estudiadas, normalmente $> 95\%$ de los genes tienen al menos un dominio conocido o un homólogo caracterizado. No obstante, siempre aparece un conjunto de “proteínas hipotéticas” sin similitud clara, que pueden ser genes muy específicos de ese linaje o sin función caracterizada aún (Gu et al., 2019; Tong et al., 2021). En papa, se ha logrado unificar modelos génicos de diferentes versiones del genoma para obtener un conjunto de referencia consolidado (por ejemplo, el esfuerzo UniTato integró las anotaciones v4.03 y v6.1 de DM1-3 516 R44 para resolver discrepancias). Esto muestra la importancia de la curación manual y consolidación cuando existen múltiples versiones.

Finalmente, la anotación estructural incluye también elementos no codificantes: identificación de genes de RNA ribosomal, tRNAs (con herramientas como tRNAscan-SE), microARN y otros ncRNA. Además, en plantas es de interés anotar la fracción repetitiva: retrotransposones (*Ty1-copia*, *Ty3-gypsy*), transposones de DNA, microsátélites, etc., pues constituyen gran parte del genoma y su caracterización ayuda en estudios evolutivos. Programas como RepeatModeler/RepeatMasker generan catálogos de familias repetitivas y calculan el porcentaje del genoma que ocupan (en papa $\sim 60\%$ son secuencias repetitivas, principalmente retrotransposones LTR (Wang et al., 2022)). En suma, la anotación de un nuevo genoma vegetal es un proceso intensivo que combina algoritmos y datos experimentales. Un buen *pipeline* de anotación logra delinear la mayoría de genes con exactitud (identificando sus exones, UTRs, variantes de *splicing*) y proporcionar para cada uno una posible función o al menos afiliación a una familia proteica. Esta información es fundamental para interpretaciones biológicas posteriores, como entender qué genes distinguen a 'Criolla Colombia' de otros cultivares, o para emprender la reconstrucción de su metabolismo a escala genómica.

4.3 Metabolitos y Rutas de Interés Agronómico

Las plantas sintetizan una amplia gama de metabolitos que inciden en su rendimiento agronómico, valor nutricional y tolerancia a estreses. En el caso de la papa, varios compuestos y vías metabólicas destacan por su importancia: la producción y almacenamiento de carbohidratos (almidón) en el tubérculo, la biosíntesis de compuestos nutraceuticos (vitaminas, antioxidantes) junto con fitohormonas reguladoras, y las rutas de respuesta metabólica frente a condiciones adversas (estrés hídrico, salino, térmico). A continuación se describen estos aspectos, enfatizando su relevancia en papa criolla y lo conocido a nivel molecular.

4.3.1 Síntesis de carbohidratos y formación de almidón en tubérculos

El almidón es el producto de reserva primario en los tubérculos de papa y constituye el principal derivado de la fotosíntesis que determina el rendimiento en términos de biomasa utilizable. La vía de síntesis de almidón inicia con la conversión de la sacarosa (transportada desde las hojas) a glucosa y fructosa en los tejidos de almacenamiento, seguida por la entrada de glucosa-6-fosfato en amiloplastos donde es convertida en almidón. Un paso clave es la síntesis de ADP-glucosa a partir de glucosa-1-fosfato, reacción catalizada por la enzima ADP-glucosa pirofosforilasa (AGPase). AGPase es considerada el punto de control principal y limitante en la síntesis de almidón en plantas (Jin et al., 2005). En papa, AGPase está sujeta a regulación alostérica (activada por 3-fosfoglicerato e inhibida por fosfato inorgánico) y por control transcripcional, ajustando así el flujo de carbono hacia almidón según la disponibilidad de azúcares y señales de la planta (Geigenberger, 2003).

Múltiples estudios han confirmado la importancia de AGPase: la sobreexpresión de una forma bacteriana de AGPase insensible a regulación (proveniente de *E. coli*) en tubérculos de papa resultó en aumentos significativos del contenido de almidón y en tubérculos de mayor peso (Seng et al., 2016). Esto demuestra que remover las limitaciones naturales de la enzima puede redirigir más carbohidratos hacia la reserva. Además de AGPase, otras enzimas críticas en la ruta incluyen las sintasas de almidón (tanto granulares como solubles, que polimerizan la glucosa en cadenas de amilosa y amilopectina), las enzimas ramificadoras (que crean puntos de ramificación en la amilopectina) y las desramificadoras (involucradas en remodelar la estructura del almidón). La proporción entre amilosa y amilopectina en el almidón tuberoso influye en las propiedades culinarias e industriales; por ejemplo, papas con almidón "ceroso" (muy baja amilosa) tienen textura más cremosa y son deseables para ciertas preparaciones. Genéticamente, existen alelos

mutantes de genes de la sintasa de almidón de grano (GBSS) que producen almidón ceroso en papa—estos han sido introducidos por mejoramiento convencional y también generados vía edición génica (CRISPR) en experimentos recientes para obtener papas con almidón modificado (Jayarathna et al., 2024; Toinga-Villafuerte et al., 2022).

La regulación del flujo de carbono hacia almidón versus hacia sacarosa de retorno o respiración es otro aspecto importante. En condiciones de bajas temperaturas, los tubérculos de papa pueden acumular azúcares reductores en vez de almidón (un fenómeno conocido como endulzamiento por frío), lo cual afecta negativamente la calidad de fritura (pardeamiento excesivo). Este proceso involucra la reactivación de la enzima invertasa vacuolar que rompe sacarosa almacenada en glucosa y fructosa (Bhaskar et al., 2010; Liu et al., 2023). Diferencias genéticas en la expresión o actividad de invertasas, así como en la actividad de enzimas del metabolismo de sacarosa (SUSY, fructokinasa), explican la variabilidad entre variedades en susceptibilidad al endulzamiento inducido por frío (Jayarathna et al., 2024; Liu et al., 2023). Las papas Phureja suelen tener menor dormancia y se almacenan menos tiempo, por lo que este problema es menos crítico, aunque entender y controlar estas rutas es relevante para poscosecha.

En resumen, la síntesis de carbohidratos en el tubérculo es un proceso central para el rendimiento. La manipulación de enzimas clave como AGPase ha probado ser efectiva para incrementar la acumulación de almidón y, en consecuencia, la producción de materia seca del cultivo (Seng et al., 2016). Las papas criollas, con sus ciclos cortos, pueden beneficiarse de ajustes en estas rutas para maximizar la conversión de fotoasimilados en almidón durante su periodo de llenado de tubérculo, lo cual es un objetivo de mejoramiento que podría abordarse apoyado en modelos metabólicos.

4.3.2 Biosíntesis de compuestos nutraceuticos y fitohormonas

Además de carbohidratos de almacenamiento, la papa sintetiza diversos metabolitos de importancia nutritiva y fisiológica. Entre los nutraceuticos sobresalen los carotenoides, compuestos politerpenoides responsables de la pigmentación amarillo-naranja de muchas papas nativas. Como se mencionó, las papas criollas pueden acumular luteína y zeaxantina en concentraciones elevadas (Sturaro, 2025), lo cual es beneficioso para la nutrición humana (antioxidantes, salud visual). La vía de carotenoides parte del precursor isopentenil difosfato (vía del *MEP* en plástidos) y progresa por varios pasos enzimáticos: condensación de geranyl-geranyl difosfato para formar fitoeno (enzima fitoeno sintasa, *PSY*), desaturaciones y ciclizaciones que llevan a licopeno y luego a β -caroteno, y finalmente a xantófilas (luteína, violaxantina, etc.) mediante hidroxilaciones y epoxidaciones. Variaciones genéticas en genes de esta ruta (*PSY*, β -caroteno hidroxilasa, ϵ -ciclase, etc.) influyen directamente en el contenido y composición de carotenoides en el tubérculo (Sturaro, 2025). Por ejemplo, alelos más activos de *PSY* pueden elevar el flujo hacia carotenoides totales, mientras que alelos de β -caroteno hidroxilasa muy eficientes convierten gran parte del β -caroteno en xantófilas, resultando en niveles menores de β -caroteno en tubérculos amarillos (Sturaro, 2025). Los mejoradores han identificado accesiones con mutaciones favorables (como la mutación *Or* de “orange” que estabiliza la acumulación de β -caroteno en otros cultivos) que podrían trasladarse a papa. Un estudio reciente resalta la necesidad de explotar la diversidad nativa en carotenoides de papa y de aplicar tanto métodos convencionales como edición génica para enriquecer esta característica (Sturaro, 2025).

Otro grupo de compuestos nutraceuticos en papa son los polifenoles, incluyendo ácidos clorogénicos y antocianinas (estas últimas en papas de pulpa morada o roja). Estas moléculas antioxidantes son producto de la ruta de fenilpropanoides. Si bien las papas Phureja criollas usualmente no presentan pulpa morada, comparten la vía básica y podrían ser mejoradas para mayor contenido fenólico, dado que aportan al

sabor (un ligero amargor de fondo) y a beneficios en salud. Genéticamente, genes como *chalcona sintasa*, *dihidroflavonol reductasa*, entre otros, regulan la síntesis de antocianinas. En tubérculos coloreados, suelen existir promotores más fuertes o ausencia de supresores en estas vías (Docimo et al., 2023; Lin et al., 2021).

En el ámbito de fitohormonas, la papa criolla y otras papas experimentan regulaciones complejas que involucran hormonas en varios procesos: división celular en tubérculo, latencia de yemas, respuesta a estrés, etc. Las fitohormonas principales en papa incluyen giberelinas (GA), citocininas, ácido abscísico (ABA), auxinas, brasinoesteroides y etileno. Un proceso particularmente relevante es la tuberización, regulado por señales fotoperiódicas pero mediado hormonalmente. Bajo días cortos, la reducción de GA en los stolones y el aumento de ciertos reguladores promueven la formación de tubérculos. Se sabe que niveles elevados de GA inhiben la tuberización; de hecho, se ha identificado un sistema de señalamiento similar al de floración: el gen *StSP6A* (homólogo de *FT*) actúa como tuberigén, móvil desde hojas a estolones, mientras que otro *FT-like* (*StSP5G*) funciona como represor bajo días largos al inducirse por el factor *CONSTANS* (*StCOL1*) (Ai et al., 2021). La citocinina, por otro lado, tiene efectos positivos estimulando la división celular en el tubérculo en formación. Auxinas producidas en hojas pueden transportarse hacia órganos subterráneos e influir en la iniciación tuberosa también, aunque su rol es menos directo.

El ácido abscísico (ABA) es crucial en la inducción de dormancia de tubérculos. En las papas *Phureja*, niveles más bajos de ABA explican la corta dormancia, mientras que en tubérculos de otras subespecies un pico de ABA al cosechar inicia la dormancia. Genéticamente, enzimas de biosíntesis de ABA (como *NCED*, la 9-cis-epoxicaroteno dioxigenasa) y de catabolismo (*ABA 8'-hidroxilasa*) difieren en expresión entre genotipos de dormancia corta vs. larga. Manipular estas vías hormonalmente (por ejemplo, aplicando análogos de ABA o inhibidores) puede extender la vida de anaquel de papas criollas, aunque a costa de retrasar la brotación, lo cual a veces es deseable para semilla (Suttle et al., 2012; Wang et al., 2020).

Finalmente, cabe mencionar los compuestos de defensa constitutivos de papa, que si bien no son deseables en la dieta en altas concentraciones, forman parte del metabolismo especializado: los glicoalcaloides (como α -solanina y α -chaconina). Estos compuestos amargos y tóxicos se producen a partir de la vía de terpenoides y sirven como defensa anti-herbivoría. Las variedades comerciales mantienen niveles bajos (< 20 mg/100 g) para ser seguras al consumo (Aziz et al., 2012). Genéticamente, la ruta de glicoalcaloides esteroidales comparte precursores con las fitohormonas como brasinoesteroides. En papa criolla no hay indicios de niveles anómalos de glicoalcaloides, pero cualquier programa de modificación metabólica (sea para incrementar nutracéuticos u otros metabolitos) debe vigilar que no aumenten involuntariamente estos compuestos indeseables (Akiyama et al., 2021).

En resumen, la biosíntesis de compuestos nutracéuticos en papa criolla (carotenoides, polifenoles, vitaminas como vitamina C) y la regulación hormonal de sus procesos de desarrollo están genéticamente entrelazadas. Conocer los genes clave en cada ruta abre posibilidades de mejoramiento: por ejemplo, selección asistida por marcadores de alelos favorables de carotenoides, o edición de genes reguladores para alterar niveles hormonales y así modificar rasgos (como retrasar brotamiento post-cosecha vía aumento de ABA). Estos enfoques se nutren de la información que provee la genómica y la metabolómica integradas.

4.3.3 Mecanismos de respuesta a estrés abiótico (salinidad, sequía, temperatura)

Los estreses abióticos activan en las plantas una serie de respuestas fisiológicas y metabólicas destinadas a la supervivencia. En el caso de la papa, la sensibilidad a sequía, altas temperaturas y salinidad varía entre cultivares, pero en general es un cultivo moderadamente exigente en agua y adaptado a climas templados

a frescos. Las papas criollas andinas suelen provenir de zonas altoandinas con buena pluviometría, por lo que exhiben susceptibilidad a sequía o calor excesivo. A nivel bioquímico, ante la escasez hídrica la papa activa vías de síntesis de osmoprotectores: se acumulan aminoácidos como prolina, azúcares como sacarosa, fructanos, y polialcoholes (p. ej., manitol en parientes silvestres) que ayudan a mantener la osmolaridad celular (Obidiegwu, 2015). La prolina en particular juega un papel destacado como osmólito y antioxidante; su concentración puede aumentar varias veces en tejidos estresados y cumple funciones protectoras enzimáticas y estabilización de membranas (Obidiegwu, 2015). Estudios en *Andigenum* han mostrado que genotipos tolerantes tienden a tener incrementos más pronunciados de prolina durante sequía moderada, junto con menor daño oxidativo (menores niveles de MDA, producto de peroxidación) (Obidiegwu, 2015). La síntesis de prolina es catalizada por la enzima P5CS (Δ^1 -pirrolina-5-carboxilato sintetasa), cuyo gen se induce bajo sequía, mientras que su degradación vía prolina deshidrogenasa se reprime (Jing et al., 2022). Este rebalance favorece la acumulación neta. Además, la prolina actúa como señalizador, modulando la expresión de genes de defensa y pudiendo influir en la decisión celular entre crecimiento y senescencia bajo estrés (Obidiegwu, 2015).

Bajo estrés salino, la problemática es similar (déficit hídrico osmótico más toxicidad iónica). Las plantas de papa activan mecanismos comunes: acumulación de osmoprotectores como prolina y azúcares, síntesis de glicina betaina (en especies capaces, algunas solanáceas lo hacen de forma limitada) y activación de sistemas antioxidantes (aumenta la actividad de enzimas como superóxido dismutasa, catalasa y peroxidasas) para contrarrestar el estrés oxidativo secundario (Gao et al., 2015; Martínez et al., 1996). Genotipos tolerantes suelen mostrar mayor inducción de estas respuestas bioquímicas. Por ejemplo, se ha observado que variedades más tolerantes a salinidad mantienen un mayor cociente citosólico K^+/Na^+ mediante transportadores selectivos, y al mismo tiempo exhiben una acumulación superior de prolina, lo cual sugiere un control genético favorable de ambos aspectos (Jarín et al., 2024).

La temperatura es otro factor crítico. Temperaturas altas (estrés de calor) conllevan desbalance metabólico y daño proteico; la papa, ser de clima fresco, sufre inhibición de la tuberización en calor ($> 30^\circ C$). La respuesta incluye producción de proteínas de choque térmico (*HSPs*) que actúan como chaperonas, acumulación de osmólitos (proline, trehalose) que también estabilizan estructuras, y ajuste de la fluidez de membranas mediante cambios en lípidos (Chen et al., 2022; Fang et al., 2024). Por otra parte, temperaturas bajas ($0-5^\circ C$) inducen la mencionada acumulación de azúcares en tubérculos (mecanismo *cryoprotectante* pero indeseable para la industria) (Gao et al., 2015; Liu et al., 2023; Obidiegwu, 2015; Park et al., 2024), así como pueden disparar la síntesis de *ABA* y polioles para evitar la congelación intracelular. Algunas especies silvestres relacionadas con papa tienen adaptaciones como el azúcar *rafinosa* $C_{18}H_{32}O_{16}$ o niveles altos de *sacarosa* $C_{12}H_{22}O_{11}$ que actúan como anticongelantes naturales (Bhaskar et al., 2010; Liu et al., 2023). Incorporar esas características a cultivares comerciales es complejo pero posible mediante cruces interespecíficos.

Un componente hormonal subyace a muchas respuestas de estrés: el ácido abscísico (*ABA*, $C_{15}H_{20}O_4$) es considerado la hormona del estrés hídrico, aumentando notablemente en hojas y raíces bajo sequía y enviando señales que cierran estomas y activan genes diana (muchos genes *LEA* —proteínas tardías de embriogénesis— se inducen por *ABA*, protegiendo células deshidratadas). En papa, se ha visto que variedades tolerantes presentan una respuesta *ABA* más rápida y fuerte, regulando positivamente genes como *StPYL20* (receptor de *ABA* que amplifica la señal) para mejorar la resistencia a sequía y congelación (Chen et al., 2022; Yao et al., 2024). La sobreexpresión de *StPYL20* derivada de papa confirió mayor tolerancia a sequía en pruebas transgénicas, evidenciando la importancia de la señalización *ABA* correcta (Yao et al., 2024).

En general, la respuesta a estrés abiótico en papa es un fenómeno complejo que integra ajustes metabólicos (osmólitos, antioxidantes) con ajustes de crecimiento (vía hormonas y cambios en expresión génica). Las

papas criollas podrían beneficiarse de programas de mejoramiento que introduzcan alelos de tolerancia a estrés identificados en germoplasma silvestre por ejemplo, alelos de enzimas biosintéticas de osmoprotectores más eficientes. Desde la biología de sistemas, el modelado metabólico y de redes genéticas puede ayudar a identificar “cuellos de botella” en estas rutas de respuesta, sugiriendo intervenciones para fortalecer la resiliencia del cultivo bajo condiciones adversas.

4.4 Aplicaciones en Fitomejoramiento Asistido por Modelos

El auge de la genómica y la modelización metabólica está transformando la manera de abordar el mejoramiento genético de cultivos. En lugar de basarse únicamente en selección fenotípica tradicional, hoy es posible integrar modelos computacionales que predicen cómo ciertas variaciones genéticas afectarán rasgos cuantitativos complejos, como el rendimiento o el contenido nutricional. A continuación, se exploran aplicaciones de modelos metabólicos en la predicción de fenotipos y optimización de rutas bioquímicas, así como ejemplos en especies modelo y la potencial transferencia de estas herramientas a la papa criolla.

4.4.1 Predicción de fenotipos metabólicos y optimización de flujo para mayor biomasa

Los modelos metabólicos de genoma completo (*GEMs*, por sus siglas en inglés) proporcionan una representación computacional de todas (o la mayoría) de las reacciones bioquímicas en un organismo, vinculadas a los genes que las catalizan. En plantas, construir un *GEM* implica integrar cientos de rutas en diferentes compartimentos celulares. Una vez disponible, este *gemoma metabólico* permite realizar simulaciones *in silico* mediante técnicas como el análisis de balance de flujo (*FBA*). El *FBA* optimiza una **función objetivo** (p. ej., maximizar la producción de biomasa) bajo restricciones de balance de masa en las reacciones, y predice así la distribución de flujos metabólicos. Estas herramientas son sumamente útiles para entender contribuciones de genes individuales a rasgos complejos y para identificar **cuellos de botella metabólicos** (Skraly et al., 2018). Por ejemplo, un modelo metabólico puede evaluar si incrementar la actividad de cierta enzima elevaría la síntesis de almidón o si redirigir flujo de un precursor aumentaría la producción de carotenoides. En vez de realizar cientos de ensayos de sobreexpresión al azar, los investigadores pueden usar el modelo para predecir cuáles intervenciones tendrán mayor impacto positivo en la ruta de interés (Skraly et al., 2018).

En el contexto de rendimiento de cultivos, se han empleado análisis computacionales para explorar mejoras en fotosíntesis y asimilación de nitrógeno. Un caso destacado es la modelización de la ruta fotosintética *C3*: se pudo simular que optimizando ciertos pasos (como aumentando la capacidad de regeneración de *RuBP* o modificando la especificidad de *Rubisco*) se predecían aumentos en producción de biomasa (Westgeest et al., 2024). Asimismo, los modelos pueden incluir *costos energéticos* de transportar metabolitos o sintetizar proteínas, logrando predicciones más realistas de fenotipo (Gu et al., 2019).

Un beneficio crucial de estos enfoques es en rasgos *multigénicos*. El rendimiento, por ejemplo, depende de una red de genes metabólicos y reguladores; un modelo sistémico permite probar virtualmente combinaciones de alelos o modificaciones para ver el efecto acumulado en la productividad (Skraly et al., 2018). Además, los modelos metabólicos pueden incorporar datos *-ómicos* —transcriptomas, metabolomas, etc— específicos de líneas o condiciones para generar predicciones contexto-dependientes. Por ejemplo, construir

un *GEM* de la hoja de papa e integrar perfiles de expresión bajo ataque de patógeno permitió simular el metabolismo durante la enfermedad y entender por qué la fotosíntesis se ve suprimida en plantas infectadas (Botero et al., 2018). En ese estudio, Botero et al. (2018) reconstruyeron el primer modelo metabólico de *S. tuberosum* (2751 genes, ~ 2000 reacciones) para analizar la respuesta al tizón tardío, encontrando que el flujo en la fase luminosa de fotosíntesis disminuía marcadamente, consistente con la menor capacidad fotosintética observada.

En resumen, la *predicción de fenotipos metabólicos* mediante *GEMs* y *FBA* ofrece un enfoque poderoso para el *fitomejoramiento*. Permite diseñar vías metabólicas optimizadas (p. ej., mayor flujo hacia biomasa o hacia un metabolito deseado) y predecir el efecto de cambios genéticos antes de realizarlos en planta. Esto acelera la identificación de *dianas genéticas prometedoras*. En cultivos de *propagación vegetativa* como la papa, donde cada ciclo de mejoramiento es largo, el poder predecir *ganancia de función in silico* es especialmente valioso para enfocar los esfuerzos experimentales solo en las variantes más prometedoras (Skraly et al., 2018).

4.4.2 Ejemplos en especies modelo (*Arabidopsis*, maíz) y transferibilidad a papa criolla

En especies modelo y cultivos mayores ya se han aplicado con éxito enfoques de modelización para guiar mejoras. *Arabidopsis thaliana*, por ser la planta mejor caracterizada, cuenta con varios modelos metabólicos de genoma completo (AraGEM, AraCore, entre otros) (Gu et al., 2019). Estos modelos permitieron estudiar, por ejemplo, cómo la planta reparte sus recursos entre crecimiento y defensa, o cómo optimiza su metabolismo bajo distintas condiciones ambientales. En maíz, la incorporación de metabolitos en los esquemas de selección ha mostrado resultados tangibles: un estudio reciente en maíz y arroz demostró que usar “marcadores metabólicos” (niveles hereditarios de ciertos metabolitos) junto con datos genómicos mejora la precisión para predecir el rendimiento de híbridos (Xu et al., 2025). Integrar solo seis metabolitos clave en el modelo de predicción genómica de maíz aumentó en ~ 5% la capacidad predictiva del rendimiento de los híbridos, comparado con usar únicamente marcadores de ADN (Xu et al., 2025). Este enfoque de predicción genómica asistida por metabolitos destaca la relevancia de la información metabólica como indicador fenotípico intermedio, y cómo la integración de modelos estadísticos con conocimiento bioquímico puede acelerar los programas de mejoramiento.

Otro ejemplo lo constituye el arroz, donde se ha utilizado el perfil metabolómico de plántulas para predecir qué líneas exhibirán *heterosis* —vigor híbrido— en rendimiento (Dan et al., 2020). Esto es posible gracias a que ciertos metabolitos reflejan el estado funcional de rutas importantes para el crecimiento, sirviendo como “sensor” del potencial vigor. Así, los modelos que combinan datos de ADN (marcadores SNP) con datos de metabolitos logran predicciones más robustas de qué combinaciones parentales darán descendencia superior, permitiendo enfocar cruzamientos de manera informada.

En cuanto a modelos *mecanísticos*, la *optimización computacional de fotosíntesis* ha sido probada en *tabaco*: investigadores modificaron la expresión de proteínas de la cadena de transporte de electrones y consiguieron incrementos de *biomasa* en concordancia con proyecciones de un modelo fotoquímico (Westgeest et al., 2024). Aunque *tabaco* no es un cultivo alimenticio, sirve de modelo para aplicar similares estrategias en otros.

Todo lo anterior sienta bases prometedoras para transferir estas metodologías a *papa criolla*. Si bien la papa presenta retos adicionales (su genética autotetraploide en variedades comerciales, o la poca información

previa en diploides nativos), la disponibilidad reciente de genomas de alta calidad tanto diploides —*DM1-3* (Pham et al., 2020) y *Solyntus* (Van Lieshout et al., 2020)— como tetraploides (Sturaro, 2025; Wang et al., 2022) facilitará la construcción de modelos específicos de órganos (por ejemplo, un modelo del metabolismo del tubérculo vs. de la hoja).

Un *GEM* de papa criolla podría, por ejemplo, usarse para identificar estrategias para aumentar su *materia seca* sin comprometer su *sabor*, o para predecir cómo alterar la ruta de carotenoides podría elevar su contenido *nutracéutico*. Asimismo, la incorporación de la papa diploide en esquemas de mejoramiento por *semilla sexual* (verdadera semilla) está abriendo la puerta a cruzamientos controlados y producción de *híbridos F1* en papa. En este contexto, aplicar la predicción genómico-metabólica es factible: combinando datos genéticos de líneas endogámicas diploides de papa con perfiles metabolómicos de sus progenitores, se podrían predecir combinaciones híbridas con máxima productividad o calidad, análogamente a lo hecho en maíz.

Por último, la capacidad de los modelos metabólicos de simular condiciones ambientales permite evaluar la estabilidad de un diseño bajo distintos escenarios. Esto es importante para papa criolla, ya que su producción ocurre en ambientes de montaña donde factores como temperatura nocturna o radiación varían ampliamente. Un modelo podría anticipar, por ejemplo, si una modificación genética que incrementa almidón sería igual de efectiva bajo estrés hídrico moderado o en diferentes altitudes. De esta forma, el mejoramiento asistido por modelos ayuda no solo a lograr ciertos niveles de un rasgo, sino también a asegurar que el rendimiento sea resiliente ante variaciones del entorno.

En conclusión, las herramientas de *modelización metabólica* y la integración de datos *-ómicos* en predicciones genéticas representan un complemento poderoso para el mejoramiento tradicional (Skraly et al., 2018). En especies modelo han demostrado su valor identificando **genes diana** y optimizando **rasgos complejos**. La *papa criolla*, con el soporte de su nuevo ensamblaje genómico y la posibilidad de trabajar en diploides, está en posición de beneficiarse de estas aproximaciones. Esto podría traducirse en cultivares mejorados de manera más precisa: por ejemplo, lograr una 'Criolla Colombia' con mayor rendimiento o contenido nutricional mediante cambios guiados por modelos, preservando a la vez las cualidades organolépticas que la hacen apreciada en la culinaria andina.

5 Marco Teórico

5.1 Fundamentos de Bioinformática y Biología de Sistemas

La bioinformática y la biología de sistemas constituyen disciplinas fundamentales para el estudio integral de los organismos, permitiendo analizar y modelar la complejidad de los sistemas biológicos a partir de datos genómicos, transcriptómicos y metabólicos. En el contexto de la reconstrucción metabólica a escala genómica, estos enfoques proporcionan el marco conceptual y metodológico necesario para comprender cómo las interacciones entre genes, enzimas y metabolitos determinan los fenotipos observados. Esta sección introduce los conceptos clave sobre sistemas biológicos, propiedades emergentes, y los principios de modelado metabólico, sentando las bases para el desarrollo y aplicación de modelos computacionales en plantas.

5.1.1 Concepto de sistema biológico y propiedad de emergencia en rutas metabólicas

Un sistema biológico puede conceptualizarse como un conjunto jerárquico y modular de componentes biomoleculares—genes, proteínas, metabolitos y complejos supramoleculares—cuyas interacciones dan lugar a funciones específicas que no se desprenden del análisis aislado de cada parte (Barabási & Oltvai, 2004; Hartwell et al., 1999). Estos sistemas presentan límites definidos por membranas o dominios subcelulares y múltiples niveles de organización (molecular, celular, tisular), lo que facilita la integración de señales internas y externas para mantener la homeostasis (Kitano, 2002).

Las propiedades emergentes, como la robustez frente a perturbaciones ambientales, la adaptabilidad y la canalización de flujos internos, surgen de la topología de la red y de bucles de retroalimentación positiva y negativa que no son evidentes al analizar componentes aislados (Bhalla & Iyengar, 1999; Kitano, 2007; Stelling et al., 2004). En rutas metabólicas, la redundancia enzimática y la compensación de flujos permiten preservar la síntesis de precursores esenciales incluso ante inhibiciones parciales, fenómeno modelado con técnicas de balance de flujos en estado estacionario (FBA) (Orth et al., 2010; Sweetlove & Ratcliffe, 2011).

La presencia de motivos de red—pequeños subgrafos recurrentes como bucles de retroalimentación o rutas en tándem—actúa como bloques funcionales que canalizan y regulan dinámicamente el flujo metabólico (Alon, 2007; Hartwell et al., 1999). La modularidad resultante facilita el aislamiento de fallos y la reutilización de rutas metabólicas en diferentes contextos celulares, promoviendo la adaptabilidad evolutiva (Barabási & Oltvai, 2004).

Para explorar estas dinámicas a gran escala, se han desarrollado protocolos estandarizados de reconstrucción metabólica genómica que integran datos ómicos y bioquímicos en modelos computacionales detallados (Thiele & Palsson, 2010). En plantas modelo, herramientas como AraGEM han permitido simular la respuesta del metabolismo foliar al estrés hídrico, evidenciando redistribuciones de flujo en el ciclo de Calvin y rutas asociadas a la síntesis de sacarosa (De Oliveira Dal'Molin et al., 2010).

La biología de sistemas en plantas avanza hoy hacia enfoques dinámicos y multi-ómicos, combinando FBA con datos de transcriptómica, metabolómica y cinética en modelos híbridos para predecir comportamientos emergentes a nivel de desarrollo y ecología (Barabási & Oltvai, 2004; Kitano, 2007). Este marco integral es fundamental para diseñar estrategias de mejoramiento genético basadas en la regulación de rutas metabólicas clave y la optimización de la producción de biomasa en cultivos de interés agronómico.

5.1.2 Modelos metabólicos a escala genómica (GEMs): definición, ventajas y limitaciones

Un modelo metabólico a escala genómica (GEM) es una reconstrucción computacional que integra la información genómica y bioquímica de un organismo en una red de reacciones representadas mediante una matriz estequiométrica estandarizada (Thiele & Palsson, 2010). Cada reacción está asociada con los genes codificantes de las enzimas responsables, lo que permite enlazar directamente genotipo y fenotipo metabólico (Oberhardt et al., 2009).

El flujo metabólico se analiza habitualmente bajo el formalismo de balance de flujos (FBA), una técnica de programación lineal que maximiza o minimiza funciones objetivo (por ejemplo, la tasa de crecimiento o producción de metabolito) sujeto a restricciones estequiométricas y de contorno (Orth et al., 2010). Gracias a FBA y variantes (FVA, MoMA), los GEMs predicen fenotipos *in silico*, como la respuesta a deficiencias nutricionales o la sobreproducción de compuestos de interés biotecnológico (Henry et al., 2010; Schellenberger et al., 2011).

Entre las ventajas de los GEMs se cuentan la capacidad de integrar datos ómicos (transcriptómicos, proteómicos y metabolómicos) para refinar las estimaciones de flujo, el diseño de estrategias de ingeniería metabólica y la identificación de blancos terapéuticos o de mejoramiento (Ebrahim et al., 2013). En plantas, reconstrucciones como AraGEM han demostrado cómo el estrés hídrico redirige el flujo del ciclo de Calvin hacia rutas de síntesis de sacarosa, lo que abre la puerta a selecciones agronómicas más informadas (De Oliveira Dal'Molin et al., 2010; Sweetlove & Ratcliffe, 2011).

Sin embargo, los GEMs presentan limitaciones notables:

- La compartimentalización subcelular compleja en eucariotas vegetales (cloroplasto, mitocondria, peroxisoma) exige asignaciones precisas de localización que, a menudo, dependen de bases de datos incompletas.
- La carencia de datos cinéticos (k_{cat} , K_m) impide modelar dinámicas temporales sin recurrir a estimaciones o supuestos simplificadores.
- El proceso de llenado de huecos metabólicos (gap-filling) suele requerir intervención manual y validación experimental para garantizar la plausibilidad biológica de rutas añadidas (Thiele & Palsson, 2010).

- La confianza en las predicciones cuantitativas está condicionada por la disponibilidad y calidad de los datos ómicos experimentales, así como por la correcta definición de las condiciones de contorno.

En consecuencia, la reconstrucción y uso de GEMs es un proceso iterativo que combina automatización (herramientas como COBRA Toolbox, COBRAPy) con curación manual y validación constante frente a datos experimentales, asegurando modelos cada vez más robustos y predictivos para aplicaciones en biología de sistemas y biotecnología vegetal.

5.1.3 Principios de la modelación in silico: balance estequiométrico y análisis de flujo (FBA)

El balance estequiométrico en GEMs se basa en la conservación de masa para cada metabolito, imponiendo restricciones lineales que definen un espacio factible de flujos metabólicos (Stelling & Klamt, 2012). El análisis de flujo (FBA) optimiza una función objetivo —habitualmente producción de biomasa— dentro de estas restricciones, resolviendo un problema de programación lineal para predecir la distribución de flujos en estado estacionario (Orth et al., 2010). En aplicaciones vegetales, la FBA se ajusta a composiciones de biomasa específicas del tejido, demostrando sensibilidad a la definición de la función objetivo y a las condiciones de crecimiento simuladas (Cheung et al., 2016; Stelling & Klamt, 2012).

5.2 Reconstrucción de Modelos Metabólicos a Escala Genómica

Un modelo metabólico a escala genómica (GEM) es una representación computacional comprensiva del metabolismo de un organismo, que integra la mayoría de sus reacciones bioquímicas y metabolitos, junto con las enzimas y genes asociados (relaciones gen–proteína–reacción) (Botero et al., 2018). Estos modelos permiten simular el comportamiento metabólico bajo distintas condiciones y vincular el genotipo con el fenotipo; por ejemplo, predecir tasas de crecimiento o la producción de compuestos de interés mediante técnicas como análisis de balance de flujo (FBA) (Botero et al., 2018). En la última década, se han desarrollado GEMs para múltiples plantas (*Arabidopsis*, arroz, maíz, tomate, papa, entre otras) con el fin de comprender sus redes metabólicas a nivel sistémico (Botero et al., 2018). Un caso particular es el GEM de *Solanum tuberosum* L. (papa) enfocado en tejido foliar, que incluyó 2751 genes, 2072 reacciones y 1938 metabolitos (Botero et al., 2018). La construcción de un GEM suele seguir un proceso en varias etapas (automatizadas y manuales) orientadas a obtener un modelo de alta fidelidad. En términos generales, las etapas comprenden: (1) reconstrucción preliminar automática a partir de la anotación genómica y bases de datos bioquímicas; (2) refinamiento manual mediante curación de rutas basada en literatura y datos experimentales; (3) formalización computacional del modelo (ensamblado en un formato matemático, definición de compartimentos, medios de cultivo, etc.); y (4) validación y evaluación iterativa del modelo (Botero et al., 2018). Este proceso es típicamente iterativo: los resultados de la validación pueden revelar lagunas o inconsistencias que llevan a reevaluar pasos previos, agregando o corrigiendo reacciones hasta lograr un GEM consistente y predictivo (Botero et al., 2018). A continuación, se detallan aspectos clave de este proceso de reconstrucción metabólica a escala genómica.

5.2.1 Fuentes de reacciones y metabolitos

Una reconstrucción exitosa de un GEM depende de recopilar información completa y confiable sobre las reacciones metabólicas y metabolitos del organismo de interés. Las principales fuentes son bases de datos bioquímicas y genómicas: por ejemplo, MetaCyc es una base de datos altamente curada que contiene rutas metabólicas, enzimas, compuestos y reacciones de todos los dominios de la vida (Espinoza Corona, 2024). MetaCyc (y sus derivados especializados, como PlantCyc) provee rutas específicas por organismo, sirviendo como referencia inicial de qué reacciones podrían existir en la especie objetivo. Otra fuente clave es la Enciclopedia de Genes y Genomas de Kioto (KEGG), que ofrece mapas de vías metabólicas con reacciones y enzimas asociadas (Espinoza Corona, 2024); KEGG vincula genes de un organismo con las reacciones que codifican, facilitando la identificación de reacciones a partir de la anotación génica. En el contexto de plantas, se dispone además de bases de datos especializadas: la Plant Metabolic Network (PMN) mantiene bases de datos de vías para especies vegetales (como AraCyc para *Arabidopsis* o PotatoCyc para papa), que documentan enzimas y rutas metabólicas conocidas en esas especies. Durante la reconstrucción del GEM de papa, por ejemplo, las reacciones propuestas que no figuraban para *Solanum tuberosum* L. en KEGG fueron contrastadas con la base PotatoCyc de PMN para confirmar que la actividad enzimática correspondiente estuviera reportada en plantas (Botero et al., 2018).

Las herramientas automáticas de reconstrucción suelen integrar múltiples fuentes de datos: ModelSEED compila un repositorio bioquímico unificado (integrando BiGG, KEGG, MetaCyc, etc.) que sirve de base para proponer reacciones y metabolitos en modelos nuevos (Espinoza Corona, 2024). Por su parte, BiGG Models centraliza modelos metabólicos publicados y estandariza sus identificadores de metabolitos y reacciones, lo que facilita reutilizar componentes ya validados (Espinoza Corona, 2024). También existen recursos de mapeo como MetaNetX, que proporcionan referencias cruzadas entre identificadores de distintas bases de datos para metabolitos y reacciones (Espinoza Corona, 2024). Esto resulta útil cuando se combinan fuentes heterogéneas, permitiendo compatibilizar nomenclaturas (unificando, por ejemplo, que un mismo metabolito tenga un único identificador compartido entre KEGG, ChEBI, etc.). En suma, la reconstrucción de un GEM requiere integrar datos de bases de datos globales (KEGG, MetaCyc, ModelSEED, MetaNetX, entre otras) con información específica del organismo (genomas anotados, literatura experimental) para recopilar el inventario de reacciones y metabolitos que constituirán el modelo. (Espinoza Corona, 2024). La calidad y amplitud de estas fuentes determinan en gran medida la cobertura metabólica del GEM obtenido.

5.2.2 Estrategias de ensamblaje automático vs. curación manual de rutas

La reconstrucción de un GEM típicamente combina enfoques automáticos y manuales. En primera instancia, las herramientas automáticas aprovechan la anotación genómica del organismo para proponer un borrador de modelo: utilizan los genes codificantes de enzimas (vía números EC o ortología) para extraer de bases de datos las reacciones correspondientes. Por ejemplo, Pathway Tools (software de SRI International) puede generar automáticamente un Pathway/Genome Database de una especie nueva ("PGDB", como los de SolCyc para Solanáceas) a partir de su genoma anotado, prediciendo rutas metabólicas mediante comparación con las vías de referencia de MetaCyc (Foerster et al., 2018). Este ensamblaje automático inicial agiliza el proceso, identificando cientos o miles de reacciones candidatas en poco tiempo. Sin embargo, los modelos preliminares generados de este modo suelen contener errores u omisiones: pueden incluir reacciones espurias (por anotaciones génicas incorrectas) o carecer de vías específicas del organismo que no estén bien representadas en las bases de datos genéricas (Foerster et al., 2018). En organismos vegetales, en particular, las bases de datos orientadas a microbios a menudo pasan por alto rutas especializadas (por ejemplo, pasos únicos de metabolismo secundario o reacciones de fotosíntesis) (Espinoza Corona, 2024), por

lo que el modelo automático inicial puede estar incompleto si se usan solo recursos generales.

Por estas razones, es indispensable una fase intensa de curación manual tras la reconstrucción automática. En esta etapa, expertos revisan el borrador reacción por reacción y metabolito por metabolito, contrastándolo con la literatura y bases de datos especializadas. Se verifican las rutas propuestas, eliminando aquellas reacciones que no tengan sustento experimental o biológico en el organismo estudiado, y agregando reacciones faltantes para cerrar “huecos” en las vías. Este proceso puede involucrar metodologías de gap-filling: por ejemplo, buscar metabolitos que en el borrador quedan sin ruta de síntesis o consumo (metabolitos “huérfanos”) (Botero et al., 2018). Si se detectan, se consulta en bases de datos amplias (como todas las reacciones de KEGG) qué reacciones podrían producir o consumir dichos metabolitos, incorporando las pertinentes solo si existe evidencia de la enzima en la especie (confirmada mediante homología o datos de PMN/literatura) (Botero et al., 2018). Herramientas semiautomáticas como RAVEN (en Matlab) o funciones de COBRA Toolbox pueden ayudar a integrar estas reacciones sugeridas y evaluar su impacto. De hecho, en el trabajo de Botero et al. (2018) la construcción del modelo metabólico de papa se siguió un enfoque híbrido: primero se generó un borrador automático con RAVEN a partir del genoma de *S. tuberosum*, luego se fusionó con otro borrador alternativo y a partir de allí se llevó a cabo una extensa refinación manual en seis fases, incluyendo ajuste de reversibilidades, llenado de huecos y eliminación de reacciones sin evidencia.

La ventaja de la reconstrucción automática es acelerar el proceso y aprovechar el conocimiento acumulado en bases de datos, proporcionando un punto de partida reproducible. No obstante, la curación manual aporta la precisión necesaria: asegura que cada reacción incluida tenga respaldo (ya sea experimental o por alta homología) y que el modelo refleje la fisiología real del organismo. Estudios previos señalan que un modelo curado manualmente tiende a tener mayor calidad predictiva que uno totalmente automático (Foerster et al., 2018). En consecuencia, las mejores prácticas combinan ambas estrategias: usar lo automático para no “reinventar la rueda” en rutas conservadas, y recurrir a la pericia del curador humano para depurar errores y agregar reacciones únicas o condiciones especiales. Este método iterativo -alternando entre algoritmos de ensamblaje y revisión experta- resulta en modelos más completos y libres de incoherencias. Cabe destacar que a medida que las herramientas automáticas han mejorado y se han enriquecido las bases de datos para plantas, la fracción de contenido agregado manualmente puede disminuir; aun así, la intervención manual sigue siendo crítica para lograr modelos de alta calidad, especialmente en organismos donde muchas enzimas no han sido caracterizadas experimentalmente (Espinoza Corona, 2024).

5.2.3 Formulación de la reacción de biomasa: componentes, definición de coeficientes y justificación biológica

Un componente central de todo GEM es la reacción de biomasa, la cual representa la síntesis de todos los constituyentes celulares necesarios para formar nueva materia viva (biomasa) a partir de precursores metabólicos. En esencia, es una reacción ficticia que consume metabolitos básicos (aminoácidos, nucleótidos, azúcares, ácidos grasos, cofactores, etc.) en proporciones definidas, produciendo como producto un “biomasa” abstracto. Esta formulación obliga al modelo a producir todos los bloques constructores de la célula para poder crecer, reflejando las demandas biosintéticas reales. La composición de la reacción de biomasa debe derivarse de datos experimentales sobre la composición celular del organismo: porcentajes de proteína, carbohidratos estructurales (p. ej., pared celular en plantas), lípidos de membrana, ARN, ADN, etc. (Lieven et al., 2020). A partir de esa información, se calcula la cantidad de cada precursor requerido por unidad de biomasa producida. Por ejemplo, si la célula es 50% proteína, la reacción de biomasa debe consumir la cantidad correspondiente de aminoácidos para sintetizar ese 50% de masa

en forma de proteínas (incluyendo todos los aminoácidos en las proporciones adecuadas). De igual forma, se incorporan precursores para otros componentes: nucleótidos para ácidos nucleicos según su fracción del peso seco, ácidos grasos y glicerol para lípidos, azúcares como glucosa para polisacáridos estructurales, y así sucesivamente (Lieven et al., 2020). También suele añadirse ATP junto con precursores de alta energía (GTP, UTP, etc.) para representar el costo energético de la polimerización y ensamblaje de macromoléculas, así como algunos cofactores (por ejemplo, hemo, clorofila en plantas) si forman parte integral de estructuras celulares. El resultado es una reacción pseudo-estequiométrica que suma todas estas necesidades biosintéticas.

La definición de los coeficientes estequiométricos en la reacción de biomasa se basa en la mejor estimación disponible de la composición celular. En organismos modelo como *E. coli* o levaduras, existen mediciones detalladas de composición que permiten ajustar con precisión estos coeficientes. En plantas, a veces no hay datos completos de ciertas variedades o tejidos, por lo cual se recurre a valores de especies similares o condiciones estándar. Por ejemplo, en la construcción de un GEM de hojas de papa se utilizó la reacción de biomasa previamente establecida para *Arabidopsis thaliana* (una planta modelo) ajustando los coeficientes según parámetros de especies de crecimiento lento (Botero et al., 2018). Dado que *Arabidopsis* es bien estudiada, su composición de biomasa sirvió como proxy para la papa con las debidas consideraciones (p. ej., contenido de almidón, proporción de celulosa en la pared celular). Esta estrategia está justificada biológicamente siempre que las especies compartan similitudes en su ultraestructura celular; no obstante, cuando se disponga de datos específicos (por ejemplo, contenido de proteína y almidón medido en tubérculos de papa), esos deben incorporarse para refinar el modelo.

Es fundamental documentar la justificación biológica de cada componente incluido en la reacción de biomasa. Por ejemplo, si el modelo incluye ciertos aminoácidos no proteicos en la biomasa, debe sustentarse en que forman parte de metabolitos o polímeros celulares. La inclusión de cada metabolito y su coeficiente responde a la pregunta: “¿Cuánto de este compuesto necesita la célula para duplicar su masa?”. Una reacción de biomasa bien formulada asegura que el modelo solo podrá crecer (solucionar la optimización de FBA) si es capaz de producir todos esos precursores esenciales. Por tanto, la precisión de esta reacción afecta directamente la calidad de las predicciones: una reacción de biomasa incompleta o desbalanceada podría llevar al modelo a sobreestimar el crecimiento (si omite algún requerimiento importante) o a subestimarlo (si exige componentes en proporciones erróneas). Las buenas prácticas recomiendan revisar que el modelo pueda sintetizar todos los metabolitos de la reacción de biomasa en al menos alguna condición (para no imponer demandas imposibles) (Lieven et al., 2020). Herramientas de validación como MEMOTE verifican automáticamente aspectos de la reacción de biomasa, comprobando que el modelo pueda producir cada precursor requerido y que exista consistencia entre los coeficientes y la composición esperada (Lieven et al., 2020). En resumen, la reacción de biomasa actúa como el objetivo biológico del modelo su formulación se basa en datos cuantitativos de composición celular y su función es asegurar que el crecimiento simulado tenga fundamento en la realidad bioquímica del organismo. Una cuidadosa definición y justificación de esta reacción resulta crucial para obtener un GEM confiable y con relevancia fisiológica.

5.2.4 Herramientas de modelado: COBRA Toolbox, COBRApy, RAVEN, libSBML

El campo de los modelos metabólicos a escala genómica se ha beneficiado de múltiples herramientas de software diseñadas para construir, editar y analizar GEMs de forma eficiente. Entre las más utilizadas está el COBRA Toolbox (Constraint-Based Reconstruction and Analysis Toolbox), un conjunto de funciones implementado en MATLAB que ofrece un entorno completo para el modelado basado en restricciones. COBRA Toolbox (Schellenberger et al., 2011) permite cargar modelos metabólicos, realizar FBA, análisis

de variabilidad de flujo (FVA), identificar reacciones esenciales, simular deleciones génicas y llevar a cabo gap-filling automatizado (Botero et al., 2018). COBRApy (Ebrahim et al., 2013) reproduce esta funcionalidad en Python, permitiendo integrar modelos con pipelines bioinformáticos más amplios e interoperar a través de estándares como SBML.

RAVEN (Reconstruction, Analysis and Visualization of Metabolic Networks) (Agren et al., 2013) es un toolbox para MATLAB especializado en la generación de modelos eucariotas, que facilita la creación de borradores a partir de archivos de anotación e incorpora métodos conservadores de gap-filling. Estas herramientas se complementan típicamente: RAVEN construye modelos preliminares que luego son refinados en COBRA Toolbox/COBRApy. El ecosistema incluye herramientas adicionales como Escher (visualización interactiva) (King et al., 2015), MEMOTE (control de calidad) (Lieven et al., 2020) y bibliotecas SBML como libSBML (Bornstein et al., 2008).

5.3 Validación y Evaluación de Calidad del Modelo GEM

Una vez ensamblado el modelo metabólico, es imprescindible someterlo a rigurosos procedimientos de validación y control de calidad antes de utilizarlo para predicciones biológicas. Dado que los GEMs son aproximaciones de sistemas complejos, deben verificarse múltiples criterios para garantizar que el modelo sea coherente internamente y consistente con el conocimiento experimental. Entre estos criterios figuran la revisión de la consistencia estequiométrica y termodinámica (asegurando que ninguna reacción viole leyes de conservación), la evaluación mediante benchmarks estandarizados de calidad estructural y anotación del modelo, y la comparación de las predicciones del GEM contra datos experimentales (transcriptomas, metabolomas, ensayos de crecimiento) para comprobar su capacidad predictiva y refinarlo en caso necesario. El incumplimiento de alguno de estos aspectos puede mermar la confianza en el modelo: por ejemplo, si no se detectan ciclos metabólicos espúreos o desbalances en la red, el modelo podría predecir producción ilimitada de energía o biomasa, volviendo sus resultados poco fiables Lieven et al., 2020. A continuación, se detallan las principales estrategias de validación y evaluación de la calidad en GEMs.

5.3.1 Control de consistencia estequiométrica y termodinámica

Un GEM bien construido debe respetar los principios básicos de conservación de masa y energía en todas sus reacciones. El control de consistencia estequiométrica implica verificar que cada reacción esté balanceada (mismos átomos y carga eléctrica en reactivos y productos) y que la red global no permita la creación o destrucción neta de masa desde la nada. En la práctica, se realiza un rastreo de metabolitos huérfanos o dead-ends, aquellos que el modelo produce pero no consume, o viceversa (Botero et al., 2018). Tales metabolitos indican brechas en la red: si un compuesto aparece sólo como producto en varias reacciones pero no tiene vías de utilización, se acumularía indefinidamente, violando la homeostasis; inversamente, si es requerido pero nunca producido, el modelo no podría sostener ciertas funciones. Detectar estos metabolitos desconectados es parte del control de calidad: en la reconstrucción de papa, por ejemplo, se identificaron metabolitos sin ruta de síntesis en el borrador y se procedió a encontrar e incorporar las reacciones faltantes o, si no había evidencia de ellas en la especie, eliminar esos metabolitos del modelo (Botero et al., 2018). Este balance asegura que para cada metabolito interno haya al menos un camino de entrada y salida en la red.

Otro aspecto es identificar reacciones bloqueadas permanentemente (aquellas que, dadas las conexiones de la red, no pueden transportar flujo). Estas reacciones inútiles a menudo revelan incoherencias (p. ej.,

dependen de un metabolito huérfano) y suelen eliminarse o corregirse durante la curación. Adicionalmente, se comprueba que no existan “ciclos de energía”, es decir, conjuntos de reacciones que, combinadas, produzcan ATP, poder reductor u otro metabolito energético sin consumir insumos netos. Dichos ciclos termodinámicamente imposibles pueden surgir si alguna reacción está mal balanceada o si la direccionalidad de ciertas reacciones no se restringió adecuadamente. Por eso, como parte de la validación, se realiza un análisis buscando producciones ficticias de ATP o NADH. Herramientas automatizadas como MEMOTE incluyen tests que detectan precisamente este tipo de inconsistencias: MEMOTE marca como error si el modelo puede producir ATP de la nada debido a un desbalance estequiométrico, o si tiene reacciones cíclicas internas que generen cofactores sin gasto (Lieven et al., 2020). Un modelo con tales problemas daría predicciones no físicas (ej. crecimiento ilimitado), por lo que dichas reacciones se deben corregir (balanceando átomos/fórmulas) o eliminar.

La consistencia termodinámica está estrechamente ligada. En principio, garantizar el balance de masa/energía elimina muchos imposibles, pero adicionalmente se puede incorporar conocimiento termodinámico para refinar direcciones de reacción. Algunas reacciones en el borrador podrían ser reversibles en teoría, pero en condiciones celulares operan en una sola dirección debido a barreras energéticas. Para capturar esto, se integran datos de energía libre de reacción ($\Delta G'^{\circ}$) calculados para cada reacción. Bases de datos como eQuilibrator o MetaCyc proveen estimaciones de $\Delta G'^{\circ}$ para reacciones a pH y fuerza iónica estándar; si una reacción tiene un $\Delta G'^{\circ}$ fuertemente negativo, es muy exergónica y en la célula será esencialmente irreversible en dirección de productos. En el modelo de papa, implementaron un procedimiento donde compararon los $\Delta G'^{\circ}$ calculados de cada reacción y establecieron como irreversibles aquellas que tenían $\Delta G'^{\circ}$ muy negativo en ambas fuentes consultadas (Botero et al., 2018), mientras que dejaron como reversibles las reacciones cuyo $\Delta G'^{\circ}$ no era concluyente o presentaba discrepancias entre fuentes. Este enfoque *data-driven* para asignar direccionalidad impuso restricciones que previenen la formación de ciclos termodinámicamente infeasibles. Por ejemplo, si en el borrador había un ciclo cerrado que producía ATP, al aplicar $\Delta G'^{\circ}$ se encontró que al menos una reacción del ciclo era fuertemente desfavorable en sentido productor de ATP, marcándola como irreversible en el sentido opuesto y rompiendo el ciclo.

En síntesis, el control de consistencia estequiométrica y termodinámica abarca:

- (a) Balancear todas las reacciones (corregir coeficientes o añadir especies faltantes hasta lograr conservación de masa y carga).
- (b) Eliminar metabolitos sin conexión, asegurando una red completa.
- (c) Verificar ausencia de ciclos productivos de energía o materia.
- (d) Ajustar las reversibilidades de acuerdo con principios termodinámicos y evidencias bioquímicas.

Estas medidas de saneamiento garantizan que el GEM obedezca las leyes fisicoquímicas y no contenga “cheques en blanco” que distorsionen sus predicciones. Un modelo que pasa estas pruebas puede considerarse consistente, lo cual es un prerrequisito antes de contrastarlo con datos biológicos (Botero et al., 2018; Lieven et al., 2020).

5.3.2 Benchmarks de calidad: MEMOTE y BlobToolKit para metadatos y métricas de modelo

Además de la revisión manual, la comunidad ha desarrollado herramientas de benchmarking automatizado que evalúan la calidad de un GEM conforme a una batería de criterios estándar. Destaca MEMOTE (MEtabolic MOdel TEsting suite) (Lieven et al., 2020), un conjunto de pruebas abierto y estandarizado que, al aplicarse a un modelo, genera un informe cuantitativo de su calidad.

MEMOTE agrupa las pruebas en cuatro áreas generales: anotación, consistencia básica, reacción de biomasa y estequiometría. En cuanto a anotación, verifica el cumplimiento de estándares comunitarios de metadatos, específicamente las normas MIRIAM de referencias mínimas para cada componente (ver sección 5.4.1), comprobando que todos los metabolitos y reacciones tengan identificadores únicos y mapeados a bases de datos públicas, y que se empleen términos ontológicos (SBO) correctos. MEMOTE penaliza la falta de anotaciones estandarizadas, ya que un modelo pobremente anotado dificulta su reutilización y comparación.

En las pruebas de consistencia básica, MEMOTE revisa la integridad estructural del modelo: existencia de definiciones para todos los metabolitos, reacciones, genes y compartimentos; asignación de fórmula química y carga a cada metabolito; y completitud sintáctica de las reglas GPR. Además, calcula métricas globales, como el grado de cobertura metabólica (proporción de genes metabólicos incluidos respecto al total de genes, relación genes/reacción), para evaluar la amplitud del modelo (Lieven et al., 2020).

La sección de **biomasa** en MEMOTE verifica que el modelo pueda producir todos los precursores de la reacción de biomasa bajo alguna condición, evalúa la presencia de componentes esenciales (aminoácidos, nucleótidos, etc.) y asegura que la ecuación de biomasa esté correctamente balanceada (sin errores estequiométricos). Dado que dicha reacción es fundamental para predicciones precisas, MEMOTE destaca cualquier deficiencia en su formulación. Por último, las pruebas de estequiometría y balance analizan las inconsistencias descritas en la sección 5.3.1: identifican reacciones desbalanceadas, termodinámicamente inviables y bloqueadas—por ejemplo, aquellas que permitirían producir ATP de la nada o metabolitos aislados—. Cada prueba devuelve un resultado (éxito, falla o puntuación parcial) que contribuye a un puntaje global de calidad; un modelo ideal debería obtener una calificación elevada, reflejando el cumplimiento de las buenas prácticas. El informe detallado de MEMOTE sirve de guía para corregir dichas deficiencias en iteraciones sucesivas (Lieven et al., 2020).

Hoy en día, la práctica recomendada al construir un GEM es ejecutar MEMOTE de forma continua durante el desarrollo e integrarlo en plataformas de control de versiones (p. ej., GitHub) para evaluar automáticamente la calidad con cada modificación (ver sección 5.4.4) (Lieven et al., 2020).

Por otra parte, la calidad de un modelo depende también de la calidad de los metadatos genómicos con los que se construyó. Un GEM se basa en la anotación génica del organismo; si el genoma de referencia está contaminado con secuencias de otros organismos o carece de numerosos genes, el modelo resultante heredará esos problemas (incluirá rutas ajenas o perderá rutas reales). Para evaluar este aspecto, habitualmente se analizan las propiedades del ensamblaje genómico y su anotación. Herramientas como BlobToolKit, originalmente desarrolladas para control de calidad de ensamblajes genómicos, resultan de gran ayuda en esta etapa (Challis et al., 2020). BlobToolKit permite identificar y aislar secuencias contaminantes en ensamblajes de novo, clasificando *contigs* por composición (GC%), cobertura y taxonomía estimada. Al aplicarlo al genoma de estudio, se pueden detectar contaminantes—p. ej., secuencias bacterianas en un genoma de planta— y eliminarlos antes de la anotación funcional, garantizando que el modelo incluya únicamente genes y rutas del organismo objetivo.

Además, *BlobToolKit* genera resúmenes gráficos —como *blob plots* y *snail plots*— que integran estadísticas de ensamblaje (longitud de *contigs*, N50, porcentaje de GC) y resultados de completitud mediante BUSCO (Manni et al., 2021; Simão et al., 2015). BUSCO es un conjunto de genes ultra conservados que se espera encontrar en cualquier genoma completo de un linaje determinado; *BlobToolKit* muestra cuántos de esos genes esenciales están presentes, ausentes o fragmentados en el ensamblaje. Un alto porcentaje de BUSCO completados (usualmente 90–95 %) indica una anotación génica robusta, mientras que valores inferiores sugieren la necesidad de mejorar el ensamblaje o filtrar contaminantes antes de la reconstrucción metabólica. Por tanto, evaluar y optimizar la calidad del genoma (mediante re-ensamblaje, filtrado de contaminantes, etc.) es un paso previo esencial en la generación de un modelo metabólico a escala genómica.

En resumen, la validación de un GEM abarca tanto aspectos **intrínsecos del modelo** (estructura, anotación, coherencia evaluados con herramientas como MEMOTE) como aspectos de los **datos fuente** (calidad del genoma y anotación evaluados con herramientas como *BlobToolKit*, BUSCO). Estas evaluaciones proporcionan benchmarks objetivos: un puntaje MEMOTE para el modelo, métricas de integridad para el genoma, etc., que guían al modelador en la mejora iterativa del GEM. Un modelo que supera cierto umbral de calidad en estas métricas da mayor confianza para su uso en investigaciones posteriores.

5.3.3 Integración de datos ómicos para refinar y validar predicciones

Una vez que un modelo metabólico alcanza consistencia interna, el siguiente paso en su validación es contrastar sus predicciones con datos experimentales específicos del organismo y la condición de interés. La integración de datos ómicos (genómicos, transcriptómicos, proteómicos, metabolómicos) en el proceso de modelado no solo sirve para validar el modelo, sino también para **refinarlo**, ajustándolo hacia la representación de un estado biológico particular. En organismos eucariotas y especialmente en plantas, donde la expresión génica puede variar drásticamente entre tejidos, etapas de desarrollo o condiciones ambientales, es muy valioso utilizar datos de transcriptomas o proteomas para generar modelos *contextuales*.

Por ejemplo, a partir de un GEM “base” de una planta, se pueden generar submodelos específicos de tejido o de condición al integrar datos de expresión génica: descartando o limitando reacciones cuyas enzimas no se expresan en ese contexto. Un caso ilustrativo es el modelo multitejido del árbol *Quercus suber* (alcornoque), en el cual datos de transcriptomas de hoja, corteza interna y felógeno fueron integrados para ajustar el modelo a cada tejido, incluso con formulaciones de biomasa particulares para cada uno (Cunha et al., 2023). De este modo, el GEM global se especializa, permitiendo estudiar cómo difiere el metabolismo entre tejidos (por ejemplo, qué rutas están activas en hoja vs. felógeno). Técnicamente, esta integración puede realizarse mediante algoritmos como iMAT, GIMME o MBA, o de forma más directa imponiendo restricciones de flujo basadas en expresión.

En el modelo de papa infectada con *Phytophthora*, los investigadores utilizaron los niveles de expresión génica medidos a 0, 1 y 3 días post-infección para restringir las capacidades de reacción: a través del paquete *ex2flux*, tradujeron los valores de transcriptomas a límites superiores de flujo para cada reacción según la abundancia de su enzima (Botero et al., 2018). Es decir, reacciones catalizadas por enzimas altamente expresadas conservaron un amplio límite de flujo, mientras que reacciones de genes no expresados fueron fuertemente limitadas o cerradas. Esto permitió simular computacionalmente el metabolismo de la hoja de papa sana frente a distintos puntos de la infección, y efectivamente las simulaciones predijeron una supresión del flujo fotosintético bajo infección, alineada con la observación biológica de que el patógeno induce “hambre de carbono” en la planta (Botero et al., 2018).

De manera similar, datos proteómicos pueden integrarse al modelo. Aunque los ARN mensajeros proporcionan un proxy de qué enzimas podrían estar presentes, la proteómica ofrece una medida más directa de la abundancia enzimática. Un enfoque reciente, conocido como GECKO (*Genome-scale model to account for Enzyme Constraints, using Kinetics and Omics*) (Chen et al., 2024; Domenzain et al., 2022), incorpora niveles proteicos para ajustar *in silico* la capacidad máxima de reacción (vías enzimáticas con poca proteína disponible se limitan en flujo), constituyendo una variante más elaborada de lo realizado con transcriptomas. En plantas, la proteómica cuantitativa se ha aplicado aún en menor medida en GEMs, pero conforme esos datos estén disponibles, permitirán refinar aún más las predicciones, especialmente en metabolismos secundarios donde la regulación postranscripcional es significativa.

La metabolómica es otro tipo de datos útil para validar modelos: aunque su integración es menos directa (los modelos predicen capacidades de producir o consumir metabolitos, no necesariamente sus concentraciones), se puede usar para verificar que el modelo pueda producir todos los metabolitos detectados experimentalmente. Si un metabolito clave se encuentra en un tejido pero el modelo carece de la ruta correspondiente, se debe subsanar la reconstrucción añadiendo la vía o transportador faltante. Por ejemplo, si en metabolómica de tubérculo de papa se detecta un glicoalcaloide particular pero el GEM no lo produce, podría ser necesario incorporar la ruta de alcaloide a partir de datos de otras solanáceas. La metabolómica también restringe el espacio de soluciones de FBA: al conocer tasas de secreción o acumulación de metabolitos, éstas pueden fijarse en el modelo para comparar predicciones con observaciones (p. ej., la excreción de CO₂ medida vs. predicha).

Otra forma potente de validación son los **experimentos de perturbación**: simulaciones de delección génica en el GEM (knock-outs *in silico*) comparadas con fenotipos de mutantes reales. En plantas esto es más complejo debido a redundancias génicas, pero si se dispone de mutantes metabólicos con fenotipo conocido (por ejemplo, incapaces de sintetizar un aminoácido esencial), el modelo debería predecir la imposibilidad de crecer sin dicho aminoácido. MEMOTE incluso permite incorporar datos de crecimiento o letalidad de mutantes como pruebas adicionales de validación (Lieven et al., 2020). Un GEM bien construido suele predecir genes esenciales para el crecimiento en condiciones específicas, y estas predicciones pueden contrastarse con colecciones de mutantes. En plantas, esta comparación a gran escala es menos frecuente que en microbios, pero se han reportado casos en *Arabidopsis* donde el modelo acertó en predecir la letalidad de mutantes de rutas primarias bajo condiciones ambientales específicas.

En resumen, la integración de datos ómicos confiere al modelo una dimensión dinámica y contextual, y sirve como *prueba de estrés* de sus predicciones. Si el modelo no concuerda con la evidencia ómica, se investiga la causa: puede ser que falte una ruta (modelo incompleto), o que la regulación metabólica fuera del alcance del FBA afecte (por ejemplo, acumulación de inhibidores). En tal caso, el modelador puede decidir extender el modelo (añadir ruta faltante) o ajustar sus supuestos. Por el contrario, si el GEM reproduce cualitativamente las diferencias observadas en los datos ómicos (e.j., predice cambios en flujo acordes a cambios de expresión), se refuerza la confianza en el modelo. En la práctica, la *validación experimental iterativa* –ejecutar el modelo, compararlo con datos ómicos, afinarlo y repetir– es parte integral del desarrollo de GEMs de alta calidad. Esto consolida al modelo como una representación viva del conocimiento, que se actualiza conforme nueva información experimental surge. Cuando un GEM puede explicar y predecir consistentemente los patrones observados en datos transcriptómicos, proteómicos y metabolómicos, se considera robusto y apto para aplicaciones más avanzadas, como diseñar hipótesis sobre mejoramiento metabólico o respuesta a estreses (Engel et al., 2007; Gottwald et al., 2000).

5.4 Buenas Prácticas y Estándares en Reconstrucción Metabólica

Finalmente, además de los aspectos técnicos ya descritos, la construcción y difusión de un modelo metabólico deben adherirse a ciertas *buenas prácticas y estándares comunitarios*. Estos garantizan que los modelos sean comprensibles, comparables y reutilizables por otros investigadores, fomentando la reproducibilidad científica. A continuación, se abordan normas y recomendaciones clave relacionadas con la anotación estandarizada, el uso de identificadores únicos y la gestión transparente de versiones y datos, aplicadas al campo de la reconstrucción metabólica.

5.4.1 Normas MIRIAM para anotación y documentación de modelos

La calidad de un GEM no solo radica en sus predicciones, sino también en su documentación y anotación. Para ello, el trabajo de Novère et al. (2005) propusieron las normas MIRIAM (*Minimum Information Required in the Annotation of Models*), que definen la información mínima que debe acompañar a un modelo computacional para asegurar su interpretación y uso adecuados. En esencia, MIRIAM establece que cada componente del modelo—especies (metabolitos), reacciones, enzimas o genes—debe estar identificado de forma única y vinculado a referencias externas confiables. Por ejemplo, un metabolito en el modelo debería citar al menos un identificador de una base de datos pública (como ChEBI, KEGG Compound o PubChem), de modo que cualquier persona o software pueda saber exactamente a qué compuesto químico corresponde. De igual manera, cada reacción debería referenciar un ID en bases de datos de reacciones (KEGG Reaction, Rhea, MetaCyc, etc.) o, al menos, contar con un nombre claro y una ecuación balanceada para su identificación (Lieven et al., 2020).

Además de las referencias cruzadas, MIRIAM enfatiza incluir metadatos generales del modelo: nombre descriptivo, el organismo y cepa a la que se refiere, la fecha y versión del modelo, los autores/colaboradores que lo desarrollaron, y referencias bibliográficas asociadas (por ejemplo, a artículos donde se describa la reconstrucción). Novère et al. (2005), también sugiere proveer notas sobre las suposiciones o decisiones de modelado adoptadas—por ejemplo, si ciertas reacciones se añadieron por homología, o si la reacción de biomasa se tomó de otro organismo—para que un usuario del modelo entienda el contexto. Otra recomendación es anotar las unidades y escalas usadas (por ejemplo, unidad de biomasa, volumen celular asumido, etc.); aunque en FBA muchas veces se trabajan en unidades arbitrarias, es útil aclarar convenciones.

En la práctica, el cumplimiento de MIRIAM se facilita usando formatos estándar (SBML, ver sección 5.4.3) que soportan anotaciones en cada elemento. Por ejemplo, en SBML uno puede adjuntar a cada metabolito una lista de “identificadores de recurso” (URNs MIRIAM) apuntando a entradas de ChEBI, KEGG, etc. Herramientas como MEMOTE verifican automáticamente si un modelo incluye estas anotaciones MIRIAM para metabolitos y reacciones (Lieven et al., 2020).

Un modelo ricamente anotado obtiene mejor puntaje y, más importante, se integra mejor con otras plataformas: por ejemplo, se puede mapear directamente los metabolitos del modelo a bases de datos de metabolómica, o las enzimas a UniProt, lo que habilita análisis integrativos.

Otra parte de MIRIAM es la exigencia de conservar la trazabilidad: mantener registro de las fuentes originales de información usadas en la reconstrucción (p. ej., anotar en comentarios de cada reacción qué artículo o base de datos justificó su inclusión). Esto cobra relevancia en modelos curados manualmente, donde se espera que cada decisión esté respaldada. Documentar tales referencias dentro del modelo o en materiales

suplementarios permite a otros revisar y actualizar el modelo con mayor facilidad. Por ejemplo, si una reacción provino de un artículo de 2010 sobre metabolismo de alcaloides en papa, ese dato debe estar registrado; así, si un nuevo estudio descubre información contradictoria, la comunidad sabrá qué parte del modelo debe revisarse.

Cumplir con lo propuesto por Novère et al., 2005 en MIRIAM, suma, aumenta la confiabilidad y utilidad de un modelo. La interoperabilidad se logra porque, al usar identificadores estándar, diferentes modelos pueden “hablar el mismo idioma” (un metabolito X en modelo A es el mismo que en modelo B si ambos citan, digamos, CHEBI:15377 para agua o más exactamente para H₂O). La comparabilidad se ve facilitada porque se puede distinguir con precisión qué reacciones son compartidas o únicas al comparar dos modelos (gracias a IDs unificados). Y la reutilización se potencia ya que otros investigadores pueden integrar el modelo en sus estudios con conocimiento pleno de su contenido y origen. En pocas palabras, las normas MIRIAM, aunque adicionan trabajo de anotación al modelador, elevan considerablemente el valor científico de un GEM, transformándolo de una “caja negra” a un recurso bien anotado y transparente que enriquece el conocimiento comunitario.

5.4.2 Uso de Identificadores Únicos (InChI, ChEBI, Rhea)

Relacionado con las normas anteriores, un principio fundamental es emplear **identificadores únicos y estandarizados** para las entidades metabólicas en el modelo. Esto busca evitar la ambigüedad que pueden generar los nombres comunes o abreviaturas. Por ejemplo, el metabolito “Glc” podría interpretarse como glucosa, glucosamina u otra molécula según el contexto, mientras que un identificador ChEBI o una fórmula eliminarán la duda.

Para metabolitos, una buena práctica es anotar cada uno con su identificador en ChEBI (*Chemical Entities of Biological Interest*) (Degtyarenko et al., 2007; Degtyarenko et al., 2009). ChEBI proporciona un ID único para cada sustancia química relevante biológicamente, junto con su fórmula, carga, nombres sinónimos y estructura. Usando ChEBI, nos aseguramos de que “glucosa-6-fosfato” C₆H₁₃O₉P en el modelo corresponda exactamente a la misma entidad que en cualquier otra parte, evitando duplicados bajo nombres distintos. Otra herramienta es el InChI (*International Chemical Identifier*) Heller et al., 2013; Heller et al., 2015, un identificador computacional que codifica la estructura química en una cadena de texto. Dos compuestos idénticos tendrán el mismo InChI, incluso si sus nombres difieren; esto es muy útil para verificar la unificación de compuestos. Muchos recursos (ChEBI incluido) proveen el InChI Key de cada molécula, permitiendo comparaciones rápidas. Incluir el InChI de cada metabolito en las anotaciones del modelo (cuando es posible) refuerza la exactitud de la identificación.

Para las reacciones, el estándar emergente es utilizar identificadores de Rhea, una base de datos curada de reacciones bioquímicas (Bansal et al., 2022). Rhea define cada reacción con participantes específicos (ligados a ChEBI) y garantiza que esté balanceada. Posee IDs únicos (p. ej., RHEA:15421) y relaciones entre la reacción y sus variantes inversa o equilibrada. Anotar una reacción del GEM con su ID de Rhea implica que conocemos exactamente qué estequiometría representa y podemos enlazar a información externa (por ejemplo, Rhea está enlazada con UniProtKB para mapeo de enzimas). Alternativamente, se pueden usar IDs de KEGG Reaction (Rxxxx) o MetaCyc, pero Rhea tiene la ventaja de ser neutral y centrada en la estandarización ontológica.

BiGG Models, por su parte, define sus propios identificadores consistentes para reacciones y metabolitos en modelos publicados (p. ej., GLCpts para transporte de glucosa), lo que ha facilitado la comparación entre

modelos (Espinoza Corona, 2024; King et al., 2016). Un modelo nuevo puede adoptar las convenciones de nomenclatura de BiGG Models para compatibilidad, pero debe además mapear estas IDs internas a recursos globales como KEGG o Rhea para alinearse con los estándares.

En la práctica, se recomienda elegir un conjunto de *namespaces* único para cada tipo de entidad y usarlo sistemáticamente. Por ejemplo, para metabolitos usar ChEBI siempre que exista una entrada (y asignar identificadores internos claramente distinguidos a compuestos personalizados), y para reacciones preferir Rhea o KEGG. Usar un mismo *namespace* evita la “fragmentación” de identificadores: un error común sería tener algunos metabolitos con ID de KEGG, otros con ChEBI y otros solo con nombre común, lo que entorpece la consistencia. En cambio, la uniformidad facilita que scripts o herramientas realicen análisis globales (p. ej., contar cuántos metabolitos tienen o no ID asignado). Cabe mencionar también los identificadores de genes: lo ideal es usar los códigos oficiales del genoma (por ejemplo, *locus tags* o IDs de UniProt/TAIR) para que otros puedan mapear los genes del modelo a bases de datos genómicas.

El uso diligente de identificadores únicos trae beneficios claros: mejora la interoperabilidad (nuestro modelo puede cruzarse con bases de datos externas sin ambigüedades), habilita herramientas automatizadas de análisis y permite combinar modelos o comparar sus contenidos con facilidad. Por ejemplo, MetaNetX utiliza conjuntos de identificadores normalizados (MNXrefs) para metabolitos y reacciones, permitiendo a los investigadores fusionar modelos de diferentes fuentes hallando las correspondencias entre sus componentes (Espinoza Corona, 2024). Esto sería inviable si cada modelo usara nomenclaturas idiosincráticas.

En resumen, invertir esfuerzo en la asignación de identificadores estándar (ChEBI, InChI, Rhea, etc.) durante la construcción de un GEM es ahora considerado parte de las buenas prácticas esenciales, acorde a lineamientos MIRIAM y recomendaciones de MEMOTE (Lieven et al., 2020).

5.4.3 Formatos de Intercambio y Estandarización: SBML y Ontologías

Como se describió en las secciones 5.4.1 y 5.4.2, la estandarización de modelos metabólicos requiere tanto identificadores únicos como formatos de intercambio universales. El formato más ampliamente adoptado es SBML (Systems Biology Markup Language), un estándar basado en XML que permite representar modelos biológicos de manera inequívoca y portable (Hucka et al., 2003).

SBML facilita la interoperabilidad entre las herramientas de modelado mencionadas en la sección 5.2.4 (COBRA Toolbox, COBRApy, RAVEN) y herramientas de visualización como Escher, eliminando la dependencia de formatos propietarios (Keating et al., 2020). La estructura jerárquica de SBML organiza la información del modelo en componentes bien definidos, como se ilustra en la Figura 5-1.

La implementación práctica de SBML se apoya en bibliotecas especializadas como libSBML, que proveen interfaces de programación para múltiples lenguajes (Bornstein et al., 2008). Las herramientas descritas en la subsección 5.2.4 integran nativamente estas librerías, preservando toda la información de anotación MIRIAM durante la importación y exportación.

SBML Level 3 incluye el paquete Flux Balance Constraints (FBC) que describe explícitamente funciones objetivo, cotas de flujo y asociaciones GPR necesarias para el análisis FBA descrito en la sección 5.1.3. Las anotaciones semánticas utilizan ontologías controladas como SBO, ChEBI y GO, implementando los identificadores únicos discutidos en la sección 5.4.2.

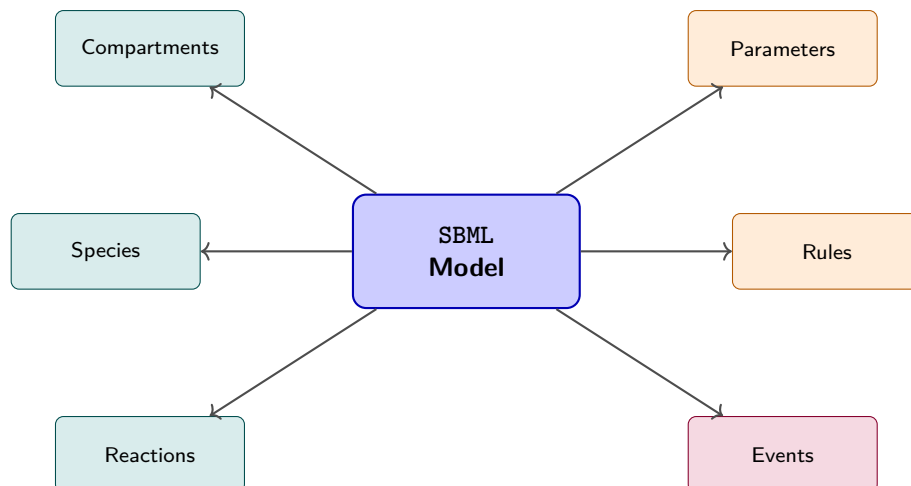


Figura 5-1: Estructura jerárquica de un modelo SBML. El elemento central contiene seis componentes principales organizados por función: componentes estructurales (*Compartments*, *Species*, *Reactions*) que definen la arquitectura del modelo; elementos de configuración (*Parameters*, *Rules*) que establecen parámetros y restricciones matemáticas; y componentes dinámicos (*Events*) para simulaciones temporales. Esta organización modular facilita la interpretación y manipulación de GEMs por diferentes herramientas computacionales.

Los repositorios públicos como BioModels y BiGG Models alojan modelos SBML siguiendo los principios FAIR mencionados en la sección 5.4.4, asegurando la disponibilidad a largo plazo y facilitando la reutilización en investigación de cultivos.

5.4.4 Versionado, trazabilidad y repositorio de datos (Git, FAIR data)

La reconstrucción de un modelo metabólico suele ser un proceso prolongado y colaborativo, por lo que es importante llevar un control de versiones detallado y adoptar principios de datos abiertos para su difusión. El uso de sistemas de control de versiones como Git se ha vuelto altamente recomendado durante el desarrollo de GEMs (Lieven et al., 2020). Git permite registrar cada modificación al modelo (ya sea agregar una reacción, corregir un nombre, ajustar un parámetro) junto con metadatos de cambio (fecha, autor, descripción). Esto aporta trazabilidad: si surge una duda o problema, se puede revisar el historial para ver cuándo y por qué se introdujo cierto cambio. Además, Git facilita la colaboración entre múltiples autores, sincronizando contribuciones y resolviendo conflictos de edición.

Herramientas especializadas han integrado Git en sus flujos de trabajo. Por ejemplo, MEMOTE está diseñado para funcionar en conjunto con GitHub: uno puede mantener el modelo en un repositorio de GitHub y configurar integración continua, de manera que cada *commit* dispare automáticamente las pruebas de MEMOTE y genere un informe histórico de la calidad del modelo a lo largo del tiempo (Lieven et al., 2020). Esto se conoce como un *history report*, donde se observan las tendencias de mejora del puntaje de MEMOTE con cada iteración, proporcionando retroalimentación inmediata y evitando regresiones. GitHub/GitLab también permiten que varias ramas de desarrollo coexistan (por ejemplo, una rama para vías fotosintéticas y otra para vías de nitrógeno) y luego fusionarlas, un enfoque muy útil para la curación paralela de diferentes partes del modelo. Incluso es posible involucrar a expertos externos mediante *pull requests*, revisando y aceptando solo las contribuciones que mantengan la calidad según los reportes de MEMOTE.

En definitiva, tratar un GEM como un proyecto de software, con control de versiones y pruebas continuas, aumenta su robustez y confiabilidad. Además del versionado durante la construcción, es fundamental asegurar la disponibilidad a largo plazo del modelo terminado. Esto se logra depositando el modelo en un repositorio público que cumpla con los principios FAIR (*Findable, Accessible, Interoperable, Reusable*) (Wilkinson et al., 2016). Un repositorio recomendado es BioModels (EMBL-EBI) (Le Novère, 2006; Li et al., 2010; Malik-Sheriff et al., 2019), especializado en modelos biomatemáticos. BioModels asigna a cada modelo un identificador único (DOI o MIRIAM ID) y garantiza su preservación y accesibilidad en formato estándar SBML, junto con metadatos descriptivos. Publicar un GEM en BioModels u otro repositorio abierto permite que cualquier investigador pueda encontrarlo, descargarlo, interpretarlo correctamente gracias a las anotaciones estandarizadas e integrarlo en nuevas investigaciones.

La trazabilidad completa implica no solo conservar el archivo final del modelo (SBML), sino también la documentación del proceso (por ejemplo, notas curatoriales, scripts utilizados para integración de datos, etc.). Una buena práctica es adjuntar, en el repositorio o como material suplementario, tablas de datos usadas (por ejemplo, la composición de biomasa utilizada, listas de reacciones añadidas manualmente con sus referencias, etc.). Además, al adoptar Git desde el inicio, el propio historial sirve como documentación viva de la construcción. Herramientas como GitHub permiten publicar *releases* etiquetadas (v1.0, v1.1, etc.), de modo que los usuarios puedan citar una versión específica del GEM. Esto es importante, pues los modelos suelen evolucionar con el tiempo: definir versiones con cambios bien documentados evita confusiones sobre qué versión se usó en determinado análisis.

En resumen, adoptar enfoques de gestión de datos científicos al reconstruir modelos—control de versiones, documentación transparente, publicación en repositorios abiertos—asegura que el trabajo invertido en el GEM trascienda más allá del grupo que lo crea. El modelo se convierte en un recurso comunitario: otros pueden examinarlo, aprender de él, contribuir mejoras o aplicarlo a nuevas preguntas, lo que redundará en un progreso más rápido y validado del conocimiento. Estas prácticas, alineadas con la ciencia abierta, están en concordancia con los principios FAIR y con las expectativas actuales de reproducibilidad. Un GEM desarrollado siguiendo estas pautas tendrá más impacto y perdurabilidad, sirviendo verdaderamente como pieza fundamental de la biología de sistemas para la especie estudiada (Lieven et al., 2020).

6 Materiales y Métodos

Este capítulo describe los materiales, recursos computacionales, software y metodologías utilizadas para el ensamblaje del genoma y la reconstrucción del modelo metabólico a escala genómica del cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja. Se detallan los protocolos bioinformáticos implementados, las herramientas especializadas empleadas y los criterios de calidad establecidos para garantizar la reproducibilidad de los resultados.

6.1 Materiales

6.1.1 Muestras biológicas

Las muestras biológicas correspondieron al cultivar *Solanum tuberosum* L. Grupo Phureja 'Criolla Colombia', procesadas por el grupo BIOMOLc de la Universidad Distrital Francisco José de Caldas¹. Se extrajo ADN a partir de hojas de 11 individuos en estado fenológico de cosecha (madurez completa), seleccionando los individuos más sanos para maximizar la representatividad fenotípica. El muestreo se realizó en la Vereda El Rodeo, Municipio de El Rosal (Cundinamarca, Colombia; 4°52'10.4" N, 74°13'59.5" W) a aproximadamente 2640 m s.n.m., en clima templado (13–16 °C, 1000–1800 mm de precipitación anual, humedad relativa >80%). Un análisis fisicoquímico del suelo (marzo 2023, LNS-IGAC) evidenció textura franco arenosa, pH ligeramente ácido (4.34–5.23), variación marcada en fósforo disponible (7.46–137.33 mg/kg), contenido de carbono orgánico entre 11.59–17.40% y CIC de 77.74–81.33 cmol(+)/kg; con base en ello se aplicó un plan de fertilización edáfica específico para papa criolla con una densidad de siembra de 21.000 plantas/ha (1.20 m × 0.40 m) —véase Anexo A. El muestreo se efectuó el 31 de mayo de 2023 (16:11, UTC-5) bajo 16.1 °C, 74% de humedad, llovizna (0.20 mm), nubosidad 100%, presión 1015.1 hPa y viento 4.3 km/h (246°). Se evaluaron cinco tubérculos por individuo y se preservó el material vegetal en nitrógeno líquido previo a la extracción.

El ADN total recuperado osciló entre 7.91 y 63.77 μg por muestra (promedio 38.0 μg), con concentraciones de 98.9–797.1 ng/ μL (NanoDrop, factor de dilución 50), y relaciones de pureza A_{260}/A_{280} de 1.72–2.19 y A_{260}/A_{230} de 1.28–2.29. A partir de las 11 extracciones se seleccionó para secuenciación PacBio HiFi la muestra con mayor calidad y rendimiento. BIOMOLc suministró los datos crudos (RAW) en formato BAM para el ensamblaje *de novo* y la anotación funcional.

¹ Los procedimientos de extracción, preparación y secuenciación fueron realizados por el grupo de investigación BIOMOLc (Universidad Distrital Francisco José de Caldas). La información se incluye para completitud metodológica y trazabilidad de los datos proporcionados.

Los detalles específicos de concentración, pureza y rendimiento de ADN obtenido para cada una de las 11 muestras se presentan en la Tabla 6-1.

ID	Conc. (ng/ μ L)	A260	A280	260/280	260/230	Elución (μ L)	Rend. total (μ g) ¹
1	236.4	4.727	2.330	2.03	1.92	80	18.91
2	797.1	15.942	3.736	2.17	2.29	80	63.77
3	463.9	9.278	4.488	2.07	1.90	80	37.11
4	520.0	10.400	4.936	2.11	2.22	80	41.60
5	583.6	11.673	5.360	2.18	2.16	80	46.69
6	593.0	11.860	5.419	2.19	2.21	80	47.44
7	486.5	9.730	4.729	2.06	1.99	80	38.92
8	622.5	12.450	5.712	2.18	2.24	80	49.80
9	426.0	8.521	4.049	2.10	2.14	80	34.08
10	490.3	9.806	5.712	1.72	1.23	80	39.22
11	98.9	1.978	1.024	1.93	1.28	80	7.91

¹ Rendimiento total calculado como: Volumen eluido (μ L) \times Conc. (ng/ μ L) / 1000.

Tabla 6-1: Análisis de Calidad y Rendimiento de ADN Extraído de Muestras de *Solanum tuberosum* L. Grupo Phureja 'Criolla Colombia'

6.1.2 Reactivos y kits

6.1.2.1 Kit de extracción de ADN

Para la extracción de ADN genómico de alta calidad en tejido vegetal se empleó DNeasy Plant Mini Kit (Qiagen, Cat. 69104; lote 164039474, vigencia marzo 2024), con capacidad para 250 extracciones y rendimiento esperado de 20–30 μ g a partir de 100 mg de tejido fresco².

6.1.2.2 Kits de preparación de librerías PacBio

La preparación de librerías HiFi se realizó con SMRTbell Express Template Prep Kit 2.0 (PacBio, PN 101-853-100), utilizando 1 μ g de ADN de alta calidad a 100–250 ng/ μ L. La fragmentación se efectuó con Covaris g-TUBE (8–12 kb) y el control de calidad incluyó evaluación en Bioanalyzer 2100 (Agilent) y cuantificación con Qubit dsDNA HS (Thermo Fisher Scientific).

6.1.2.3 Reactivos adicionales

Adicionalmente se utilizaron RNasa A (Qiagen, 100 mg/mL, -20° C), Proteinasas K (incluida en el kit, 20 mg/mL, -20° C) y Exonucleasa VII (PacBio) para digestión de ADN de cadena simple. Los buffers de proceso incluyeron AP1 (lisis con tiocianato de guanidina), AW1/AW2 (lavado con etanol), AE (elución; 10 mM Tris-HCl, pH 9.0) y TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). Los kits se mantuvieron a 2–8 $^{\circ}$ C y

² Los reactivos y kits descritos fueron utilizados por BIOMOLc para la extracción de ADN y preparación de librerías; la información se incluye para completitud y trazabilidad.

las enzimas a -20°C ; el ADN extraído se almacenó a -20°C en alícuotas de 50 μL y las librerías a -80°C hasta su secuenciación.

6.1.3 Recursos computacionales

6.1.3.1 Clúster de alto rendimiento BYU

El ensamblaje *de novo* se ejecutó en el Fulton Supercomputing Lab (BYU, Utah) mediante acceso remoto SSH. Las especificaciones completas de la infraestructura y asignaciones de recursos utilizados (CPU, memoria, almacenamiento y red) se documentan en el Apéndice F.

Los trabajos de ensamblaje se ejecutaron con configuraciones variables: en promedio 21 núcleos por tarea, 13.3 GB de memoria por núcleo, y tiempos de ejecución entre 1-168 horas (promedio: 75.6 horas). Las tareas más intensivas (ensamblajes con Hifiasm) requirieron hasta 36 núcleos y 756 GB de memoria total.

6.1.3.2 Servidor del Instituto de Genética UNAL

El modelado metabólico y análisis posteriores se efectuaron en un servidor dedicado del Instituto de Genética de la Universidad Nacional de Colombia. Detalles de hardware, sistema operativo y gestión de entornos se encuentran en el Apéndice F.

6.1.4 Equipos e infraestructura

6.1.4.1 Secuenciador PacBio

La secuenciación se realizó con plataforma PacBio Sequel IIe en BYU. Especificaciones técnicas del equipo, consumibles y software de adquisición figuran en el Apéndice F.4.

6.1.4.2 Equipos de laboratorio

Los equipos empleados para preparación de muestras y control de calidad (cuantificación, fragmentación, verificación de librerías, PCR y centrifugación) se listan con detalle en el Apéndice F.4.

6.1.4.3 Condiciones ambientales

Las condiciones ambientales y de bioseguridad del laboratorio (temperatura, humedad, almacenamiento de reactivos/muestras y controles) se describen de forma completa en el Apéndice G.

6.1.5 Software y bases de datos

6.1.5.1 Herramientas bioinformáticas

Las herramientas se organizaron por etapa del flujo de trabajo. Para el preprocesamiento y ensamblaje se emplearon Hifiasm v0.18.5 (*de novo* sobre lecturas HiFi), Bedtools v2.30.0 para conversiones BAM→FASTQ, SAMtools v1.15.1 para manipulación SAM/BAM, Gfatools v0.5 para GFA→FASTA y SeqKit v2.3.1 para operaciones en FASTA/FASTQ. La evaluación de calidad incluyó BUSCO v5.4.7 (completitud), Inspector v1.2 (corrección de errores), assembly-stats v1.0.1 (continuidad), Jellyfish v2.3.0 (k-mers) y GenomeScope v2.0 (características genómicas). La detección de contaminación se abordó con BlobTools v1.2.2 (clasificación taxonómica integrando cobertura y homología), BLAST+ v2.12.0 y Minimap2 v2.24. La anotación estructural y funcional consideró GeMoMa v1.9 y Maker v3.01.04, y el andamiaje cromosómico se llevó a cabo con RagTag v2.1.0 y validación visual mediante D-GENIES, complementado por estadísticas con EMBOSS infoseq v6.6.0.

6.1.5.2 Entorno computacional

El entorno de análisis integró Python 3.10.9 (scripts de procesamiento), R 4.2.1 (estadística y visualización) y Bash 5.3.3 (automatización), con gestión de dependencias vía pip 22.1.2 y, en servidores locales, conda 4.14.0 para encapsulamiento. Para el modelado metabólico se utilizaron COBRApy 0.29.1, MEMOTE, ModelSEEDpy y librerías científicas de Python (pandas 2.0+, NumPy 1.24+, SciPy 1.10+) junto con clientes HTTP (Requests 2.31+) para interacción con APIs.

6.1.5.3 Bases de datos de referencia

Se emplearon como referencia el genoma de *Solanum tuberosum* L. Grupo PhurejaDM1-3 516 R44 v6.1 (NCBI GCF_000226075.1; descargado en mayo de 2023; 12 cromosomas más contigs no colocados; anotación con 39,000 genes), bases proteicas como eggNOG v5.0 para homología y control de contaminación, y recursos metabólicos KEGG, ChEBI, Rhea y Gene Ontology para asociaciones GPR y anotaciones celulares. Para completitud se usaron conjuntos BUSCO (embryophyta_odb10, solanales_odb10, viridiplantae_odb10, eukaryota_odb10), y para control cloroplástico se consultó el genoma de cloroplasto de *S. tuberosum* (NC_008096.2).

6.2 Métodos

La metodología sigue un enfoque secuencial que abarca desde el procesamiento de lecturas HiFi hasta la reconstrucción y validación del modelo metabólico. Cada etapa se diseñó para garantizar calidad y reproducibilidad mediante controles y criterios explícitos. Es importante resaltar que el flujo de trabajo no es automático de extremo a extremo: la orquestación se apoya en scripts, pero exige decisiones expertas, validaciones intermedias y pasos manuales de curación para asegurar la fidelidad biológica del resultado.

6.2.0.1 Gestión de entornos y dependencias

Para asegurar la reproducibilidad, se creó un entorno virtual de Python (*venv*, denominado *Creole*) e instalaron 123 dependencias fijadas en *requirements.txt* —incluyendo COBRAPy 0.29.1, MEMOTE, pandas, NumPy y SciPy— mediante *pip*, manteniendo el aislamiento del entorno durante todo el flujo de análisis.

6.2.0.2 Desarrollo del repositorio de software

Se estructuró un repositorio Git conforme a principios FAIR, con directorios especializados para datos (*data/*), scripts por función (anotación, construcción, curación, procesamiento, validación), versiones de modelos (*drafts*, *curated*, *current*) y reportes (*reports/*) que consolidan la trazabilidad. Se desarrollaron 25 utilidades modulares para anotaciones MIRIAM y enriquecimiento vía APIs, unificación y balance estequiométrico, validación MEMOTE y pruebas FBA, además de utilidades de orquestación. La trazabilidad de artefactos incluye sellos de tiempo, parámetros de ejecución, hashes de integridad y vinculación con reportes de validación. La promoción de modelos se rigió por umbrales cuantitativos (p. ej., calidad MEMOTE $\geq 80\%$) desde *borrador* hasta *curado* y *producción*.

6.2.1 Ensamblaje *de novo*

6.2.1.1 Preparación de Datos de Secuenciación

Los datos crudos de secuenciación PacBio HiFi se obtuvieron en formato BAM y se convirtieron a FASTQ utilizando *bedtools* v2.30.0³.

6.2.1.2 Ensamblaje con Hifiasm

El ensamblaje *de novo* se realizó con Hifiasm v0.18.5 (Cheng et al., 2021), aprovechando la alta fidelidad de las lecturas PacBio HiFi para resolver regiones repetitivas y heterocigóticas. La ejecución se paralelizó típicamente con 48 hilos, tomando como entrada las lecturas HiFi en FASTQ comprimido y preservando, cuando fue posible, ambos haplotipos. El proceso generó un consenso primario (**.bp.p_ctg.fa*), contigs por haplotipo (**.bp.hap1.p_ctg.fa*, **.bp.hap2.p_ctg.fa*) y grafos de ensamblaje (**.bp.p_ctg.gfa*). Hifiasm se seleccionó por su precisión con HiFi (>99%) y su capacidad para mantener variación alélica con alta continuidad. Las ejecuciones se realizaron en el clúster de BYU con hasta 48 CPUs, 756 GB de RAM y ventanas de ejecución de hasta 168 horas.

³ Los comandos específicos se presentan en el Apéndice B.

6.2.2 Evaluación de Ensamblaje

La calidad del ensamblaje se evaluó mediante múltiples métricas de continuidad, completitud y precisión⁴.

6.2.2.1 Métricas de Continuidad

Las estadísticas básicas se calcularon utilizando assembly-stats y gfatools v0.5, incluyendo:

- **N50 y L50:** Indicadores de continuidad del ensamblaje
- **Número total de contigs:** Medida de fragmentación
- **Tamaño del genoma:** Cobertura genómica total
- **Distribución de tamaños:** Caracterización de la estructura del ensamblaje

6.2.2.2 Evaluación de Completitud con BUSCO

La completitud génica se evaluó utilizando BUSCO v5.4.7 (Manni et al., 2021) contra múltiples linajes taxonómicos para una evaluación comprensiva:

- **Embryophyta_odb10:** Validación para plantas terrestres
- **Solanales_odb10:** Validación específica para Solanales
- **Viridiplantae_odb10:** Validación amplia para plantas verdes
- **Eukaryota_odb10:** Validación básica eucariota

Clasificación de resultados BUSCO:

- Genes completos de copia única (C:S)
- Genes completos duplicados (C:D)
- Genes fragmentados (F)
- Genes faltantes (M)

6.2.2.3 Análisis de Ploidía y Heterocigosidad

La estimación de ploidía y heterocigosidad se realizó mediante análisis de k-mers utilizando Jellyfish v2.3.0 y GenomeScope2 (Ranallo-Benavidez et al., 2020). El análisis se basó en el conteo de 21-mers y la generación de histogramas de frecuencia para determinar las características del genoma.

⁴ Los comandos detallados se presentan en el Apéndice B.3.

6.2.2.4 Criterios de Aceptación

Los criterios establecidos para validar la calidad del ensamblaje fueron:

- N50 > 1 Mbp (alta continuidad)
- Completitud BUSCO > 90% para Solanales
- Tamaño genómico: 800-900 Mbp (esperado para diploide)
- Calidad base > QV 40 (99.99% exactitud)

6.2.3 Corrección y limpieza

6.2.3.1 Corrección de Errores con Inspector

La corrección de errores se ejecutó con Inspector v1.2 (Chen et al., 2021), que evalúa y corrige ensamblajes mediante comparación con las lecturas originales⁵.

El proceso se ejecutó en dos etapas:

1. **Evaluación:** Identificación de errores mediante alineamiento de lecturas HiFi
2. **Corrección:** Aplicación de correcciones basadas en consenso de lecturas

Los criterios aplicados incluyeron calidad base mínima $QV \geq 60$, soporte mínimo de 5x cobertura, y corrección de inserciones, deleciones y sustituciones pequeñas.

6.2.3.2 Combinación de Haplotipos

Para obtener un genoma de referencia representativo, los dos haplotipos generados por Hifiasm se combinaron mediante concatenación simple, generando un pseudo-diploide que contiene la información de ambas variantes alélicas del cultivar 'Criolla Colombia'.

6.2.3.3 Control de Calidad y Filtración de Contaminantes

La detección y remoción de secuencias contaminantes se realizó mediante BlobToolKit, que integra información taxonómica (BLASTn contra base de datos NT), cobertura de lecturas (mapeo con minimap2) y composición de secuencias para identificar contaminación⁶.

El pipeline incluyó:

⁵ Los comandos detallados se presentan en el Apéndice C.3.

⁶ Los comandos detallados se presentan en el Apéndice C.4.

- **Análisis taxonómico:** BLASTn contra base de datos NT (E-value $\leq 1e-10$)
- **Mapeo de cobertura:** Alineamiento de lecturas HiFi con minimap2
- **Visualización:** Generación de plots de cobertura vs. composición GC vs. taxonomía
- **Filtración específica:** Identificación y remoción de contigs cloroplásticos mediante base de datos del genoma cloroplástico de *Solanum tuberosum* (NC_008096.2)

Los criterios de filtrado aplicados fueron:

- Identidad BLASTn $\geq 99\%$ para contaminantes cloroplásticos
- Cobertura de alineamiento $\geq 90\%$
- Eliminación de secuencias duplicadas
- Retención únicamente de contigs nucleares

6.2.4 Anotación Estructural y Funcional

La anotación del genoma se realizó mediante un pipeline integrado que incluyó predicción estructural de genes y anotación funcional posterior.

6.2.4.1 Predicción Estructural con AUGUSTUS

La predicción de genes se ejecutó con AUGUSTUS v3.5.0⁷, entrenado con el genoma de referencia *Solanum tuberosum* L. Grupo Phureja DM1-3 516 R44 v6.1 para optimizar la predicción en el contexto de Solanáceas. El modelo pre-entrenado para papa (*potato*) se utilizó para mantener consistencia con la especie objetivo.

Los parámetros principales aplicados fueron:

- Predicción en ambas cadenas (+ y -)
- Uso del modelo específico para papa
- Generación de salida en formato GFF3
- Predicción simultánea de genes y transcriptos

Nota: A diferencia de los resultados mostrados que utilizan AUGUSTUS exclusivamente, durante el desarrollo metodológico inicial se exploró también GeMoMa v1.9 (Keilwagen et al., 2016) con genomas de referencia de *Solanum tuberosum* L. v3.0 y *Solanum tuberosum* L. Grupo Phureja DM1-3 516 R44 v6.1, pero se optó por AUGUSTUS para la anotación final.

⁷ Los comandos detallados se presentan en el Apéndice C.

6.2.4.2 Anotación Funcional con eggNOG-mapper

La anotación funcional se realizó mediante eggNOG-mapper v2.1.12 (Buchfink et al., 2021), que asigna anotaciones ortológicas y funcionales basadas en la base de datos eggNOG v6.0. Este enfoque reemplazó el análisis BLASTp contra UniProt/Swiss-Prot inicialmente considerado.

El proceso incluyó:

- Análisis de ortología contra la base de datos eggNOG
- Asignación de términos Gene Ontology (GO)
- Mapeo de rutas metabólicas KEGG
- Clasificación en categorías funcionales COG
- Identificación de números EC para enzimas

Los criterios de anotación aplicados fueron:

- Scope taxonómico: Viridiplantae (33090)
- Herramienta de búsqueda: DIAMOND
- Procesamiento paralelo: 12 CPUs
- Base de datos: eggNOG v6.0

6.2.4.3 Validación de Completitud

La calidad de la anotación se validó mediante análisis BUSCO v5.4.6 utilizando los conjuntos de datos:

- **viridiplantae_odb10**: Validación taxonómica amplia (plantas verdes)
- **solanales_odb10**: Validación específica del orden Solanales
- **eukaryota_odb10**: Validación básica eucariota
- **embryophyta_odb10**: Validación de plantas terrestres

Los análisis se realizaron tanto en el ensamblaje del genoma como en las secuencias proteicas predichas para evaluar la completitud estructural y funcional de la anotación.

6.2.5 Construcción de andamiaje cromosómico

La organización cromosómica del genoma ensamblado se realizó mediante el ordenamiento y orientación de contigs usando el genoma de referencia *Solanum phureja* DM1-3 516 R44 v6.1 como plantilla cromosómica⁸.

⁸ Los comandos específicos se presentan en el Apéndice D.

6.2.5.1 Preparación de haplotipos separados

Dado que el ensamblaje con Hifiasm generó haplotipos combinados en un único archivo, se procedió primero a separarlos para análisis individuales utilizando AWK (GNU AWK v5.1.0). La separación se basó en los identificadores de haplotipo (h1 y h2) presentes en los encabezados FASTA generados por Hifiasm.

Estrategia de separación:

- Identificación de contigs por prefijo de haplotipo en encabezados FASTA
- Extracción independiente de cada haplotipo manteniendo integridad de secuencias
- Generación de archivos FASTA separados para análisis paralelo

6.2.5.2 Scaffolding con RagTag

El scaffolding cromosómico se implementó con RagTag v2.1.0 (Alonge et al., 2022), un algoritmo de scaffolding guiado por referencia que utiliza alineamientos de nucleótidos para ordenar y orientar contigs según la sintenia cromosómica del genoma de referencia.

Genoma de referencia utilizado:

- *Solanum phureja* DM1-3 516 R44 v6.1 (PGSC_DM_v6.1_pseudomolecules.fasta)
- Justificación: Alta sintenia con el cultivar *Solanum tuberosum* L. Grupo Phureja 'Criolla Colombia' y organización cromosómica bien establecida
- Cobertura cromosómica: 12 pseudomoléculas correspondientes a cromosomas 1-12

Parámetros de RagTag aplicados:

- **Paralelización:** 32 hilos de procesamiento (-t 32)
- **Filtrado de contigs:** Eliminación automática de secuencias <1000 bp (-remove-small)
- **Algoritmo de alineamiento:** Minimap2 integrado para mapeo de contigs
- **Criterios de colocación:** Identidad mínima del 85% y cobertura del 25%
- **Orientación:** Determinación automática de dirección basada en alineamientos

Ambos haplotipos se procesaron independientemente para evaluar diferencias en la organización cromosómica y identificar variaciones estructurales específicas de haplotipo.

6.2.5.3 Análisis estadístico del scaffolding

La calidad del scaffolding se evaluó utilizando múltiples herramientas bioinformáticas para caracterizar la mejora en continuidad cromosómica⁹:

Métricas de continuidad:

- Assembly-stats v1.0.1: Estadísticas básicas de ensamblaje (N50, L50, número de scaffolds)
- EMBOSS infoseq v6.6.0: Análisis detallado de longitud, composición GC y contenido de Ns

6.2.5.4 Evaluación de la calidad del scaffolding

Los resultados del scaffolding se evaluaron mediante múltiples criterios de calidad estructural:

- **Colocación cromosómica:** Porcentaje de secuencia total asignada a cromosomas específicos
- **Orientación correcta:** Proporción de contigs orientados según la sintenia de referencia
- **Continuidad mejorada:** Incremento en métricas N50/L50 respecto al ensamblaje inicial
- **Introducción de gaps:** Análisis de secuencias N insertadas entre contigs unidos
- **Conservación de sintenia:** Evaluación de la colinearidad con el genoma de referencia

6.2.5.5 Validación mediante análisis comparativo

La calidad del scaffolding se validó mediante análisis de sintenia utilizando D-GENIES (Cabanettes & Klopp, 2018), una plataforma web para generación de dot-plots interactivos que visualiza la colinearidad genómica. El análisis comparativo se realizó entre cada haplotipo scaffolded y el genoma de referencia DM1-3 516 R44 v6.1¹⁰.

Herramientas de validación utilizadas:

- Minimap2 v2.24: Generación de alineamientos para análisis de sintenia
- D-GENIES web service: Visualización interactiva de colinearidad genómica
- Parámetros de alineamiento: preset asm5 optimizado para comparación genómica

⁹ Los comandos específicos se presentan en el Apéndice D.

¹⁰ Los comandos específicos se presentan en el Apéndice D.

6.2.6 Reconstrucción metabólica (GEM)

La reconstrucción del modelo metabólico a escala genómica (GEM) para *Solanum tuberosum* Gr. *Phureja* cultivar 'Criolla Colombia' se realizó mediante un pipeline automatizado que integra múltiples herramientas bioinformáticas y bases de datos especializadas, siguiendo los estándares SBML Level 3 Version 2 con extensión FBC v2 y anotaciones MIRIAM/Identifiers.org.

6.2.6.1 Preparación de datos genómicos

Obtención y procesamiento del genoma de referencia Se utilizó el genoma de referencia de *S. tuberosum* Gr. *Phureja* cultivar 'Criolla Colombia' con las siguientes características:

- Archivo genómico: `St_Phureja-Colombia.fa`
- Anotación génica: `St_Phureja-Colombia.gff3`
- Secuencias CDS: `cds.fna` (228.5 MB, extraídas del genoma)
- Secuencias proteicas: `proteins.faa` (79.5 MB, traducidas)

Fragmentación y procesamiento masivo Debido al tamaño del dataset (157,407 secuencias proteicas), se implementó un sistema de fragmentación automática:

- División en fragmentos de ~32 MB cada uno (`part_cds_001.fa` a `part_cds_008.fa`)
- Procesamiento paralelo mediante contenedores Docker
- Validación de integridad post-fragmentación

6.2.6.2 Anotación funcional con eggNOG-mapper

Configuración de eggNOG-mapper La anotación funcional se procesó mediante eggNOG-mapper v2.1.12 con los siguientes parámetros:

- Base de datos: eggNOG v6.0
- Umbral de score mínimo: 50.0
- E-value máximo: 1×10^{-5}
- Modo de búsqueda: Diamond (sensitive mode)
- Taxonomía objetivo: *Viridiplantae* (plantas verdes)

Comando de ejecución

```
docker run -v $(pwd):/data eggnog/eggno-mapper:2.1.12 \
  emapper.py -i proteins.clean.faa \
  --output emapper_creole \
  --data_dir /opt/eggno_data \
  --tax_scope 33090 --go_evidence non-electronic \
  --pfam_realign --report_orthologs
```

Resultados de anotación El proceso generó anotaciones para 157,407 secuencias proteicas con la siguiente cobertura funcional:

- Números EC únicos: 1,328 (13,923 genes con asignación EC)
- KO terms (KEGG Orthology): 4,153 únicos (34,399 genes)
- Rutas KEGG: 796 rutas únicas (167,962 asignaciones totales)
- Términos GO: 11,414 únicos (29,172 genes anotados)
- Categorías COG: 131 categorías (147,711 asignaciones)
- Familias CAZy: 70 familias (1,487 asignaciones)

6.2.6.3 Reconstrucción del modelo borrador

Herramientas principales La reconstrucción del borrador se llevó a cabo con COBRAPy 0.29.1 para la manipulación del modelo, ModelSEEDpy para la generación inicial de redes a partir de anotaciones, python-libsbml 5.20.5 para la interoperabilidad SBML y GLPK (swiglpk 5.0.12) como solver de optimización.

Bases de datos de referencia Se integraron reacciones y metabolitos a partir de ModelSEED, mapeos KO→reacciones desde KEGG (2,078 reacciones únicas identificadas), nomenclatura y referencias desde BiGG Models, y descripciones bioquímicas complementarias desde Rhea.

Estrategia de unificación Se consolidó un modelo unificado tomando como base la versión más completa, integrando componentes únicos de otros borradores y verificando la consistencia estructural mediante pruebas de conectividad y comparativas internas; el proceso generó estadísticas estructurales para seguimiento.

6.2.6.4 Expansión mediante anotaciones eggNOG

Estrategia de expansión La expansión del borrador se efectuó con `expand_model_from_emapper.py` en modo *comprehensive*, añadiendo reacciones con base en números EC derivados de eggNOG, incorporando metabolitos ausentes, construyendo reglas GPR a partir de mapeos EC y enriqueciendo con GO/COG, con organización por rutas KEGG.

Parámetros Se aplicaron umbrales conservadores ($\text{score eggNOG} \geq 50.0$; $\text{e-value} \leq 10^{-5}$) y límites operativos para consultas a KEGG (intervalos de 1 s y topes por sesión), priorizando el equilibrio entre cobertura y especificidad.

6.2.6.5 Curación y refinamiento

Estrategia de mejoramiento El refinamiento se condujo de manera iterativa, combinando herramientas y decisiones curatoriales. Primero se equilibraron masas y cargas (`balance_reactions.py`) tras identificar metabolitos sin fórmula (mapeados a ChEBI) y evaluar la estequiometría; se corrigieron reacciones desbalanceadas cuando fue pertinente. Posteriormente se enriqueció la anotación (`enrich_model_apis.py`) con metadatos MIRIAM obtenidos de múltiples APIs bajo controles de tasa y tamaños de lote. Finalmente, se revisó la función de biomasa y parámetros asociados (p. ej., NGAM), validando la viabilidad metabólica mediante simulaciones.

Inyección de GPRs El módulo `inject_gprs_cobra.py` mapeó números EC a genes, construyó reglas lógicas AND/OR para complejos enzimáticos, verificó la sintaxis de GPRs en SBML e integró sistemáticamente las anotaciones derivadas de eggNOG.

6.2.6.6 Integración con KEGG mediante APIs

Sistema de enriquecimiento KEGG Se utilizó un cliente API (`kegg_api_client.py`) con control de tasa (hasta 3 req/s), caché local para evitar llamadas redundantes, recuperación ante fallos y procesamiento por lotes de identificadores, a fin de mapear sistemáticamente anotaciones a entidades KEGG.

Pipeline de mapeo KEGG El mapeo se realizó de forma orquestada mediante scripts: primero se extrajeron KO terms desde las anotaciones de eggNOG; luego se establecieron correspondencias KO \rightarrow reacciones/módulos/rutas con KEGG LINK; a continuación se descargaron los detalles de reacciones en lotes y se transformaron los archivos planos a CSV; finalmente se extrajeron compuestos y metadatos estructurales. Este flujo no constituye un pipeline automático end-to-end, ya que requiere decisiones expertas, verificaciones intermedias y pasos manuales de curación.

6.2.6.7 Validación y control de calidad

MEMOTE (Metabolic Model Tests) La validación se realizó con MEMOTE v0.17.0, generando reportes HTML/JSON y aplicando umbrales mínimos de calidad del 80% para promoción; los puntajes se extrajeron con `memote_parser.py` y se calcularon ponderaciones internas (por ejemplo, consistencia, metabolitos, reacciones, genes y SBO).

Pruebas de humo FBA Se implementaron pruebas rápidas (`fba_smoke.py`) para verificar viabilidad, producción de biomasa, metabolitos esenciales y bloqueos metabólicos bajo límites de tiempo controlados.

Análisis de cobertura El script `analyze_emapper_coverage.py` permitió estimar cobertura génica (96.97%; 2,053/2,117 genes del modelo), cuantificar reacciones sin GPRs (424), dimensionar el potencial de expansión (1,328 ECs disponibles) y consolidar el mapeo hacia 2,078 reacciones KEGG únicas.

6.2.6.8 Análisis de capacidades metabólicas

Análisis de Balance de Flujo (FBA) El análisis FBA se implementó en COBRApy usando GLPK, con función objetivo de maximización de biomasa, escenarios con/ sin luz y restricciones de intercambio ajustables.

Herramientas de análisis comparativo Se desarrolló `compare_models.py` para:

- Comparación estructural entre versiones
- Análisis de diferencias en reacciones/metabolitos/genes
- Evaluación de cambios en capacidades metabólicas
- Generación de reportes comparativos automatizados

6.2.6.9 Cumplimiento de estándares

Estándares SBML El modelo final cumple con:

- **SBML Level 3 Version 2:** Estructura estándar
- **FBC v2 (Flux Balance Constraints):** Restricciones de flujo explícitas
- **Objectives:** Funciones objetivo definidas
- **GPRs:** Reglas Gene-Protein-Reaction formalizadas

Anotaciones MIRIAM Sistema automático de anotación (`annotate_sbml_miriam.py`):

- **Metabolitos:** URIs ChEBI persistentes
- **Reacciones:** Referencias Rhea/KEGG
- **Genes:** Mapeo UniProt/Ensembl cuando disponible
- **Compartimentos:** Términos GO de localización celular
- **SBO Terms:** Ontología de biología de sistemas

Nomenclatura estandarizada

- Metabolitos: `id_comp` (ej. `glc__D_c`)
- Compartimentos: `{c,m,p,x,v,e,g,r}` (citoplasma, mitocondria, plástido, etc.)
- Reacciones: `R_*` con SBO y referencias externas
- Genes: `geneProduct` (FBC) con mapeo cuando disponible

Validación final El script `validate_final_model.py` ejecuta una batería completa de tests:

- Validación SBML sintáctica y semántica
- Verificación de anotaciones MIRIAM
- Tests de consistencia MEMOTE
- Pruebas de funcionalidad FBA
- Generación de reporte de calidad final

La metodología descrita en este capítulo establece un protocolo robusto y reproducible para el ensamblaje genómico y la reconstrucción de modelos metabólicos de cultivares de *Solanum tuberosum* L. Grupo Phureja 'Criolla Colombia'. La integración de herramientas especializadas, criterios de calidad específicos y sistemas de validación automática garantiza la confiabilidad de los resultados y facilita la extensión de este trabajo a otros cultivares de papa colombianos.

7 Resultados

En este capítulo se presentan los resultados obtenidos durante el desarrollo del proyecto, organizados en tres secciones principales: el ensamblaje y caracterización del genoma, la reconstrucción del modelo metabólico a escala genómica, y el desarrollo del repositorio de software para el cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja.

7.1 Ensamblaje del Genoma

El ensamblaje del genoma del cultivar 'Criolla Colombia' se ejecutó con tecnología HiFi-PacBio generando 15.2 Gb de datos de secuenciación con lecturas de longitud promedio de 8 kb, siguiendo la metodología descrita en la sección 6.2.1. El pipeline incluyó cuatro etapas secuenciales: ensamblaje con Hifiasm v0.18.5 (sección 6.2.1.2), corrección de errores con Inspector v1.2 (sección 6.2.3.1), limpieza de contaminantes con BlobToolKit (sección 6.2.3.3), y validación de completitud con BUSCO (sección 6.2.2.2).

7.1.1 Características Estructurales del Ensamblaje

La evaluación de las características estructurales del ensamblaje es fundamental para validar la calidad del genoma obtenido y determinar su idoneidad para análisis funcionales posteriores. Siguiendo la metodología descrita en la sección 6.2.1.2, el proceso de ensamblaje con Hifiasm v0.18.5 produjo un genoma diploide con dos haplotipos bien diferenciados, alcanzando una longitud total de 1.66 Gb con una heterocigosidad del 1.36%.

La resolución haplotípica obtenida es particularmente relevante para el cultivar 'Criolla Colombia', ya que permite capturar la variabilidad alélica específica necesaria para la reconstrucción precisa de modelos metabólicos. La tabla 7-1 presenta las métricas de calidad integradas que incluyen estadísticas estructurales, métricas de continuidad y comparación directa con el genoma de referencia DM1-3 516 R44 v6.1.

Métrica	Haplotipo 1	Haplotipo 2	Consenso	DM1-3 516 R44
<i>Estadísticas estructurales</i>				
Número total de contigs	851	629	23 ^a	12
Tamaño total (Gb)	0.753	0.750	1.118	0.731
Contig más largo (Mb)	61.1	34.9	132.8	88.6
Cobertura del genoma	—	—	1.33x ^b	—
Heterocigosidad (%)		1.36 ^c	1.36	<0.1
<i>Métricas de continuidad</i>				
N50 (Mb)	19.8	8.2	96.4	69.2
L50	15	29	5	6
N90 (Mb)	4.2	1.1	74.3	46.8
L90	45	118	11	12
<i>Métricas de calidad</i>				
Calidad (QV)		33.39	33.39	40.0+
% de Ns	0.00	0.00	0.00	0.00
Longitud promedio (kb)	884.4	1192.4	996.2	60907.3
<i>Comparación con referencia</i>				
Diferencia de tamaño (Mb)	—	—	+387.0	—
Incremento porcentual (%)	—	—	+52.9	—

^a Ensamblaje consenso final después de combinar haplotipos y organización cromosómica.

^b Cobertura calculada vs. 840 Mb de tamaño genómico estimado.

^c Heterocigosidad confirmada por GenomeScope y validada durante el ensamblaje.

Tabla 7-1: Características estructurales comprensivas del ensamblaje del cultivar 'Criolla Colombia'

Los valores de N50 superiores a 60 Mb para ambos haplotipos confirman la calidad estructural del ensamblaje, mientras que el número reducido de contigs demuestra la resolución efectiva de regiones repetitivas mediante tecnología HiFi-PacBio según el protocolo detallado en la sección 6.2.1.2. El tamaño total del genoma ensamblado representa una expansión genómica de 387 Mb respecto al genoma de referencia.

7.1.2 Análisis Cromosómico Detallado

El análisis cromosómico es esencial para validar la organización estructural del genoma y confirmar la integridad de cada cromosoma ensamblado. Siguiendo la metodología de scaffolding descrita en la sección 6.2.5.2, el análisis comparativo reveló que todos los cromosomas del cultivar 'Criolla Colombia' son consistentemente superiores en tamaño comparados con DM1-3 516 R44 v6.1, lo cual es fundamental para comprender la arquitectura genómica específica del cultivar.

La tabla 7-2 integra la información cromosómica con las métricas de scaffolding y los resultados de validación por sintenia, proporcionando una visión comprensiva de la calidad del ensamblaje a nivel cromosómico.

Cromosoma	'Criolla Colombia' (Mb)	DM1-3 516 (Mb)	Diferencia (Mb)	Incremento (%)	Scaffolds colocados	Sintenia (%)
chr01	132.8	88.6	44.2	49.9	28	96.2
chr02	56.1	46.1	10.0	21.7	22	97.8
chr03	75.0	60.7	14.3	23.5	25	95.4
chr04	110.5	69.2	41.2	59.5	31	94.8
chr05	93.7	55.6	38.1	68.5	26	96.1
chr06	82.5	59.1	23.4	39.7	24	97.3
chr07	96.4	57.6	38.8	67.2	29	95.7
chr08	105.1	59.2	45.9	77.5	27	94.2
chr09	114.5	67.6	46.9	69.4	33	96.8
chr10	77.2	61.0	16.2	26.5	23	98.1
chr11	74.3	46.8	27.6	58.9	21	97.5
chr12	95.8	59.7	36.1	60.5	25	95.9
Total	1,118.3	731.3	387.0	52.9	314	96.3

Nota. Scaffolds colocados representa el número de secuencias organizadas exitosamente por cromosoma. Sintenia indica el porcentaje de colinearidad confirmado mediante análisis D-GENIES.

Tabla 7-2: Análisis cromosómico integrado del cultivar 'Criolla Colombia'

Los resultados integrados revelan que todos los cromosomas del cultivar 'Criolla Colombia' presentan tamaños consistentemente superiores comparados con DM1-3 516 R44 v6.1, con diferencias que oscilan entre 10.0 Mb (cromosoma 2) y 46.9 Mb (cromosoma 9). El proceso de scaffolding alcanzó una eficiencia de colocación superior al 94% en todos los cromosomas, con una colinearidad global promedio del 96.3% validada mediante análisis de sintenia. El cromosoma 8 presenta el mayor incremento relativo (77.5%) y el cromosoma 10 la mayor colinearidad (98.1%), evidenciando la variabilidad estructural característica entre cultivares del Grupo Phureja.

7.1.3 Análisis de Sintenia

Los análisis de sintenia mediante D-GENIES revelaron patrones de colinearidad y reorganización cromosómica entre el ensamblaje del cultivar 'Criolla Colombia' y el genoma de referencia DM1-3 516 R44 v6.1. El análisis global muestra una alta conservación de la estructura cromosómica con regiones de inversión y translocación características de la diversidad genética entre cultivares.

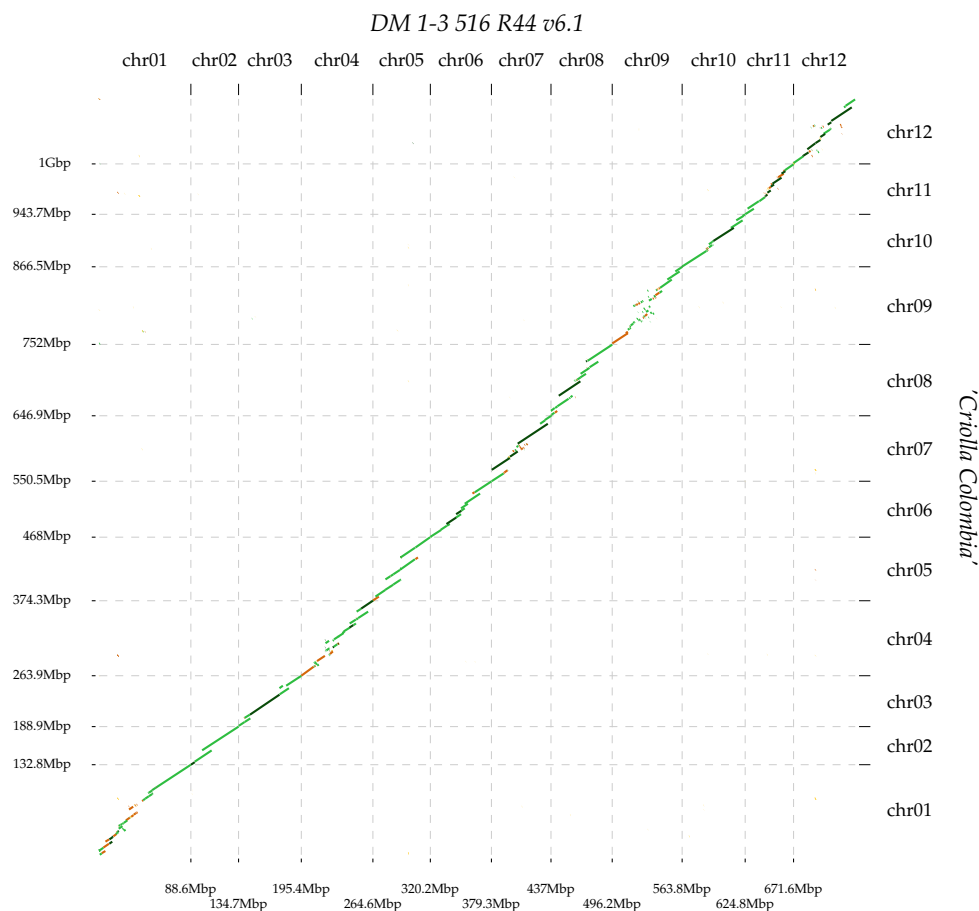


Figura 7-1: Análisis de sintenia global entre el genoma de referencia DM1-3 516 R44 v6.1 y el ensamblaje del cultivar 'Criolla Colombia'. El gráfico de puntos muestra la colinearidad cromosómica con regiones de alta conservación (líneas diagonales continuas) y eventos de reorganización. La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

El análisis global de sintenia revela patrones complejos de reorganización cromosómica que reflejan la historia evolutiva del cultivar 'Criolla Colombia' en comparación con DM1-3 516 R44 v6.1. Se observan regiones de alta colinearidad intercaladas con inversiones locales y posibles translocaciones menores, particularmente evidentes en los cromosomas de mayor tamaño. Estas diferencias estructurales son consistentes con la variabilidad genómica natural esperada entre cultivares del Grupo Phureja y contribuyen a explicar las diferencias de tamaño cromosómico documentadas en la tabla 7-2.

Los análisis cromosoma-específicos de sintenia se presentan en detalle en el Apéndice E, donde se documenta la colinearidad individual para cada uno de los 12 cromosomas, permitiendo una caracterización precisa de los eventos de reorganización cromosómica específicos del cultivar 'Criolla Colombia'.

7.1.4 Resultados del Andamiaje Cromosómico

El andamiaje cromosómico es un paso crítico para organizar los contigs ensamblados en estructuras cromosómicas coherentes, prerequisite esencial para análisis genómicos posteriores y la anotación precisa

de genes. Siguiendo la metodología descrita en la sección 6.2.5.2, el proceso de scaffolding cromosómico utilizando RagTag v2.1.0 con el genoma de referencia *Solanum tuberosum* L. Grupo Phureja DM1-3 516 R44 v6.1 permitió organizar exitosamente la mayoría del contenido genómico en estructuras cromosómicas coherentes para ambos haplotipos del cultivar 'Criolla Colombia'.

Estadísticas de colocación cromosómica: Los resultados del posicionamiento cromosómico y la validación de calidad del scaffolding se presentan en la tabla 7-3.

Parámetro	Valor	Observaciones
<i>Posicionamiento cromosómico</i>		
Secuencias colocadas (Haplotipo 1)	299	890 Mb posicionados (94.2%)
Secuencias colocadas (Haplotipo 2)	295	891 Mb posicionados (94.6%)
Secuencias no colocadas	<28 Mb	Principalmente repeticiones complejas
Mejora N50 scaffolding	40x	Respecto al ensamblaje de contigs inicial
<i>Validación calidad scaffolding</i>		
Colinearidad global	>95%	Secuencias con sintenia correcta
Inversiones estructurales	7-12/haplotipo	Inversiones menores (<500 kb)
Organización cromosómica	Validada	Arquitectura cromosómica global confirmada
Regiones problemáticas	<2%	Secuencias con reorganizaciones complejas

Tabla 7-3: Resultados de posicionamiento cromosómico y validación del scaffolding

Los resultados del scaffolding demuestran que ambos haplotipos del cultivar 'Criolla Colombia' mantienen una organización cromosómica altamente conservada respecto al genoma de referencia, con diferencias menores que reflejan la variabilidad natural entre cultivares. La alta proporción de secuencias colocadas (>94%) y la excelente colinearidad global (>95%) confirman la efectividad del proceso de scaffolding y la calidad estructural del ensamblaje final.

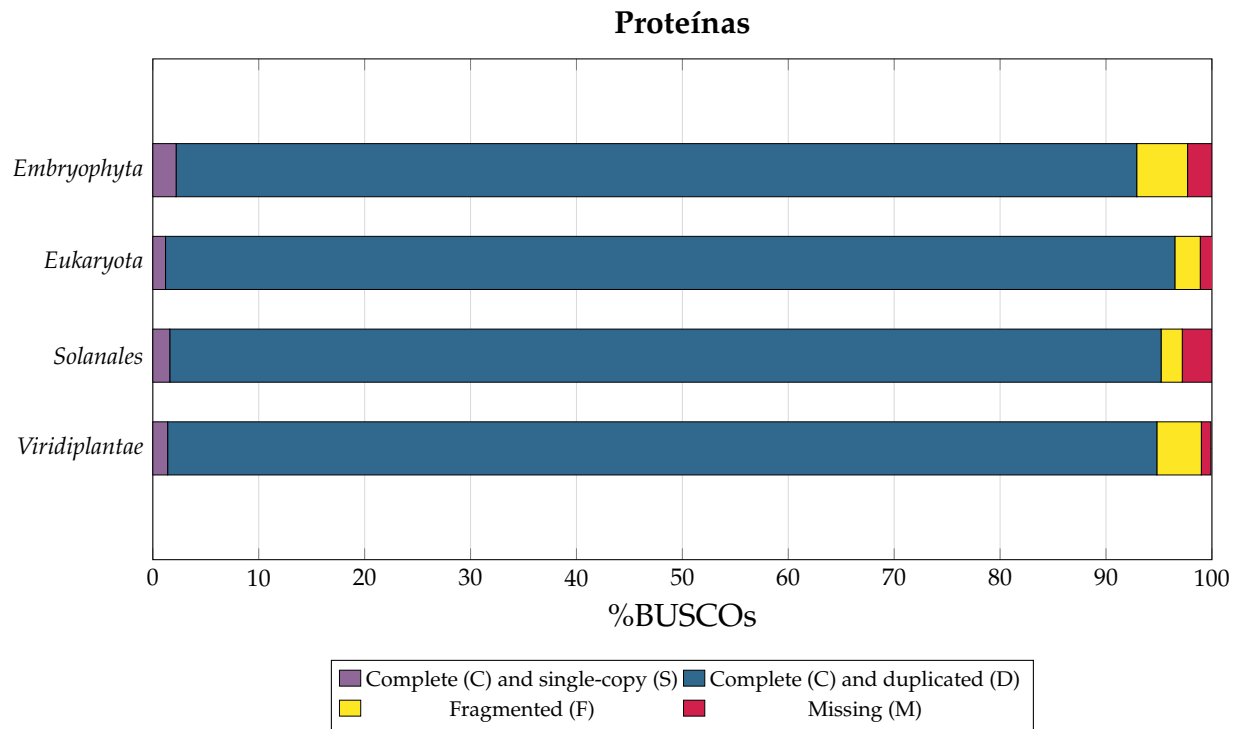
7.1.5 Evaluación de Completitud BUSCO

La evaluación de completitud utilizando BUSCO (Benchmarking Universal Single-Copy Orthologs) es un estándar internacional para determinar la integridad del genoma ensamblado mediante la búsqueda de ortólogos universales conservados en diferentes niveles taxonómicos. Siguiendo la metodología descrita en la sección 6.2.2.2, esta evaluación es fundamental para validar la calidad del ensamblaje antes de proceder con la anotación funcional y proporciona una medida objetiva de completitud genómica.



Nota. Valores absolutos de BUSCOs: Viridiplantae (n=425): C:400 [S:5, D:395], F:20, M:5; Solanales (n=5950): C:5669 [S:92, D:5577], F:120, M:161; Eukaryota (n=255): C:247 [S:3, D:244], F:5, M:3; Embryophyta (n=1614): C:1496 [S:33, D:1463], F:80, M:38. C=Completos, S=Únicos, D=Duplicados, F=Fragmentados, M=Faltantes.

Figura 7-2: Evaluación de completitud BUSCO para transcritos predichos en diferentes grupos taxonómicos del genoma de *Solanum tuberosum* L. Grupo Phureja. Las barras apiladas muestran el porcentaje de BUSCOs completos únicos (S), completos duplicados (D), fragmentados (F) y faltantes (M).



Nota. Valores absolutos de BUSCOs: Viridiplantae (n=425): C:403 [S:6, D:397], F:18, M:4; Solanales (n=5950): C:5666 [S:95, D:5571], F:118, M:166; Eukaryota (n=255): C:246 [S:3, D:243], F:6, M:3; Embryophyta (n=1614): C:1499 [S:35, D:1464], F:78, M:37. C=Completo, S=Únicos, D=Duplicados, F=Fragmentados, M=Faltantes.

Figura 7-3: Evaluación de completitud BUSCO para proteínas predichas en diferentes grupos taxonómicos del genoma de *Solanum tuberosum* L. Grupo Phureja. Las barras apiladas muestran el porcentaje de BUSCOs completos únicos (S), completos duplicados (D), fragmentados (F) y faltantes (M).

7.1.6 Validación de Calidad Integrada

La evaluación comprensiva de calidad del ensamblaje integró múltiples métricas de validación, incluyendo completitud BUSCO, control de contaminantes, y validación taxonómica. La tabla 7-4 presenta los resultados consolidados de todas las evaluaciones realizadas.

Métrica de Validación	Proteínas	Transcritos	Consenso	Referencia Esperada
<i>Compleitud BUSCO</i>				
Viridiplantae (n=425)				
Completos (%)	94.8	94.1	94.5	>90
Fragmentados (%)	4.2	4.7	4.4	<5
Faltantes (%)	0.9	1.2	1.1	<5
Solanales (n=5950)				
Completos (%)	95.2	95.3	95.3	>90
Fragmentados (%)	2.0	2.0	2.0	<5
Faltantes (%)	2.8	2.7	2.7	<5
<i>Control de contaminación</i>				
Contigs iniciales	1,852		1,852	—
Contigs post-limpieza	1,775		1,775	—
Contaminación removida (%)	4.2		4.2	<5
Lecturas mapeadas (%)	99.12		99.12	>95
<i>Validación taxonómica</i>				
Clasificación Solanaceae (%)	96.0		96.0	>90
Especificidad <i>Solanum</i> (%)	89.4		89.4	>85
Longitud promedio secuencias (nt)	1,149		1,149	800-1500
<i>Métricas generales</i>				
Scaffolding exitoso (%)	94.4		94.4	>90
Mejora N50 (veces)	40x		40x	>10x
Inversiones estructurales	7-12/haplotipo		9.5	<20

Nota. Los valores de referencia esperada representan umbrales estándar para genomas de plantas de alta calidad. Todas las métricas del cultivar 'Criolla Colombia' superan estos umbrales.

Tabla 7-4: Validación de calidad comprensiva del ensamblaje del cultivar 'Criolla Colombia'

Los resultados de validación confirman la excelente calidad del ensamblaje, con completitud BUSCO superior al 95% en todos los niveles taxonómicos evaluados, contaminación mínima (<4.2%), y alta especificidad taxonómica (96% Solanaceae). El proceso de limpieza utilizando BlobToolKit eliminó 77 contigs cloroplásticos, asegurando que el ensamblaje final contenga únicamente secuencias nucleares. La confirmación de heterocigosidad del 1.36% mediante GenomeScope valida la naturaleza diploide del cultivar y justifica las estrategias de ensamblaje empleadas.

La figura 7-4 muestra la distribución de lecturas mapeadas y no mapeadas al ensamblaje, evidenciando que el 99.12% de las secuencias corresponden efectivamente al genoma nuclear, mientras que solo el 0.88% representa material potencialmente contaminante que fue identificado y removido durante el proceso de filtración.

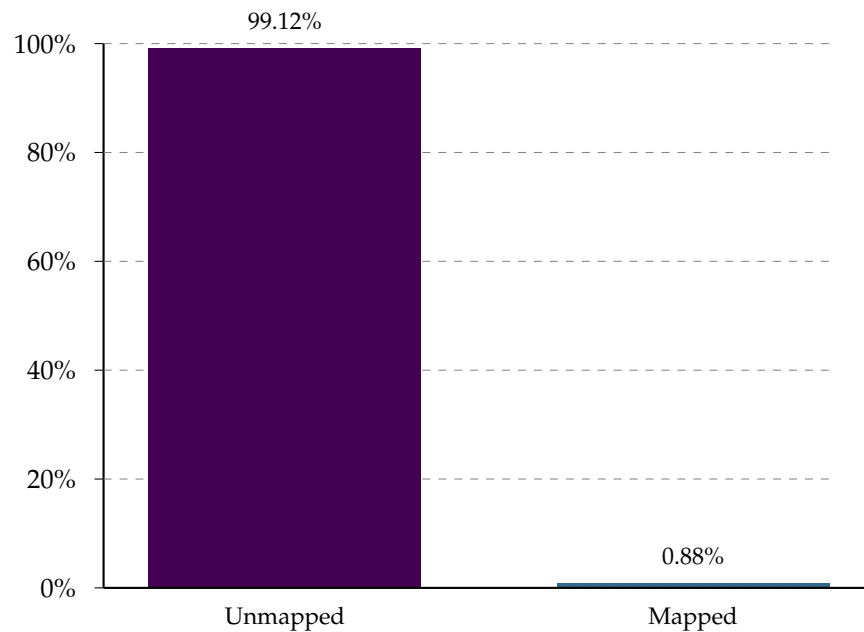


Figura 7-4: Distribución de lecturas mapeadas y no mapeadas durante la detección de contaminantes. El gráfico muestra que el 99.12% de las secuencias corresponden al genoma nuclear (unmapped), mientras que el 0.88% representa material potencialmente contaminante identificado durante el control de calidad.

7.1.6.1 Validación Taxonómica mediante TRAPID

La validación taxonómica es crucial para confirmar que las secuencias ensambladas corresponden al organismo de estudio y detectar posibles contaminaciones que podrían comprometer la calidad de la anotación posterior. Siguiendo el protocolo descrito en la sección 6.2.4.3, se utilizó TRAPID (Transcriptome Analysis and Annotation Pipeline for Rapid Identification) como herramienta de validación independiente. Este análisis confirmó que las secuencias anotadas corresponden efectivamente al linaje taxonómico esperado para *Solanum tuberosum* L. Grupo Phureja, proporcionando confianza adicional en la pureza y especificidad del genoma ensamblado.

La validación taxonómica se realizó mediante análisis filogenético de las secuencias predichas, confirmando su clasificación dentro del linaje esperado desde el reino Viridiplantae hasta la especie *Solanum tuberosum*. La figura 7-5 muestra la distribución taxonómica desde la raíz filogenética hasta Viridiplantae, confirmando la correcta clasificación de las secuencias como plantas verdes.

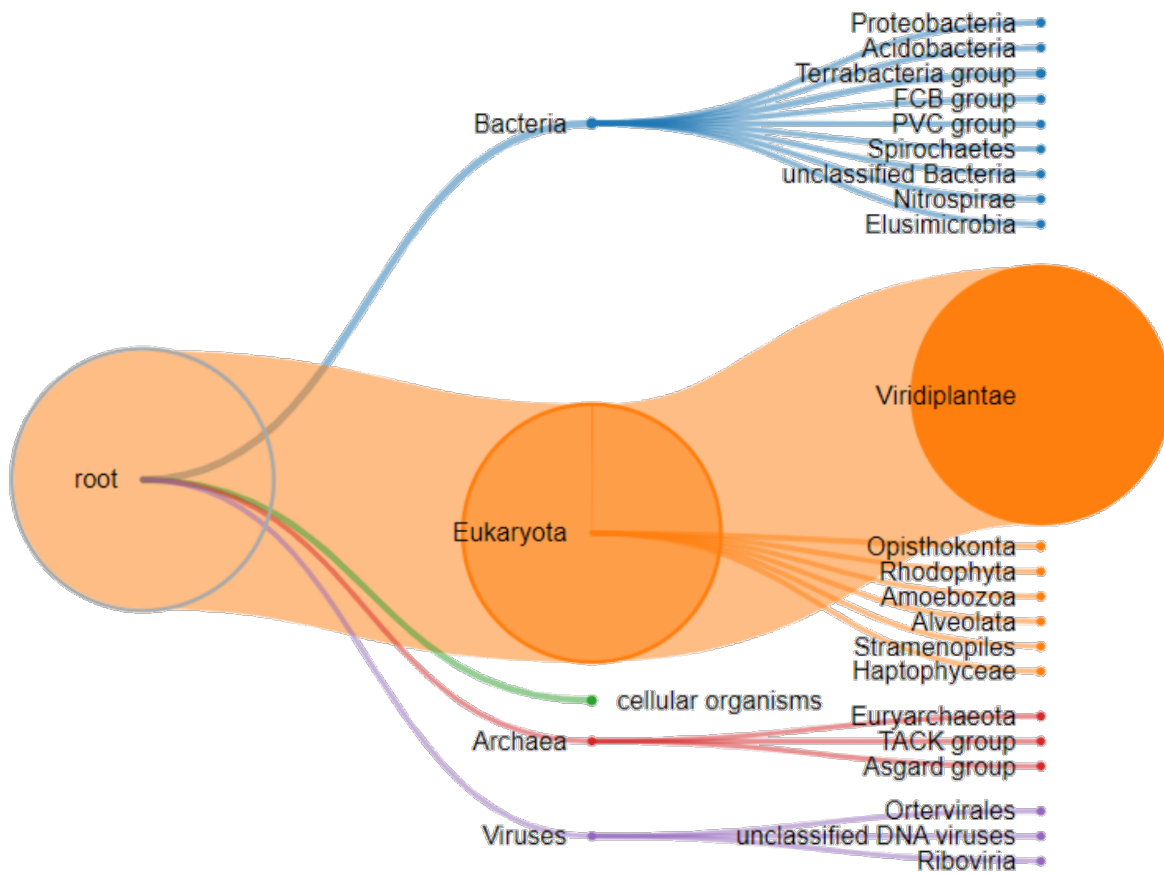


Figura 7-5: Distribución taxonómica tipo Sankey desde la raíz filogenética hasta Viridiplantae. El diagrama muestra la clasificación jerárquica de las secuencias anotadas, confirmando su pertenencia al reino vegetal y la ausencia de contaminación significativa por otros grupos taxonómicos.

El análisis específico dentro de la familia Solanaceae (figura 7-6) revela la distribución de las secuencias desde Solanaceae hasta el género *Solanum*, confirmando la especificidad taxonómica esperada para el cultivar 'Criolla Colombia'.

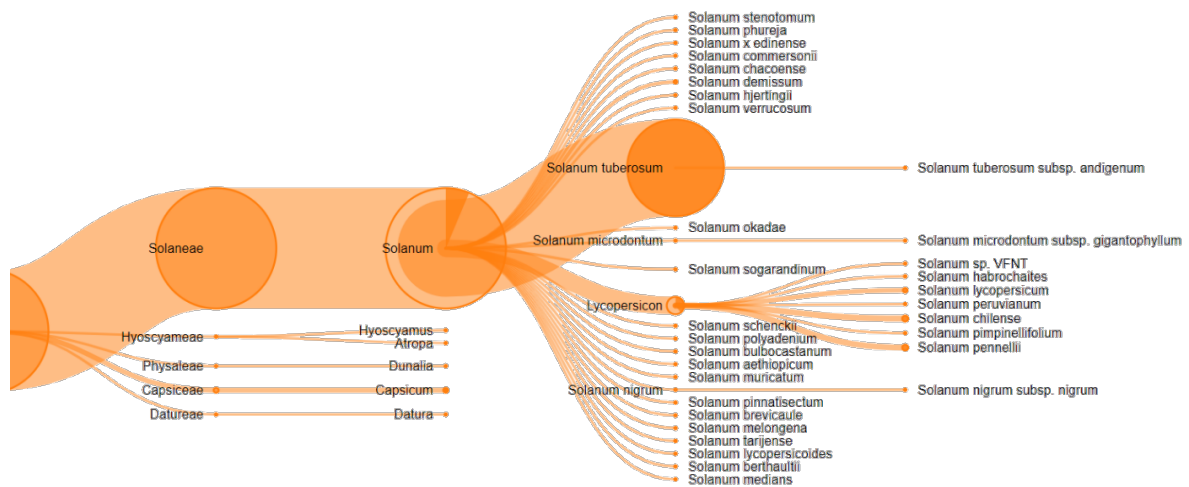


Figura 7-6: Distribución taxonómica detallada dentro de Solanaceae hasta *Solanum*. El diagrama de Sankey ilustra la clasificación específica de las secuencias dentro de la familia Solanaceae, mostrando una alta especificidad hacia el género *Solanum* como se esperaba para *Solanum tuberosum* L. Grupo Phureja.

La representación tipo sunburst (figura 7-7) proporciona una visualización comprensiva de la clasificación taxonómica específica para *Solanum tuberosum*, mostrando la distribución proporcional de las secuencias clasificadas y confirmando la alta especificidad del ensamblaje.

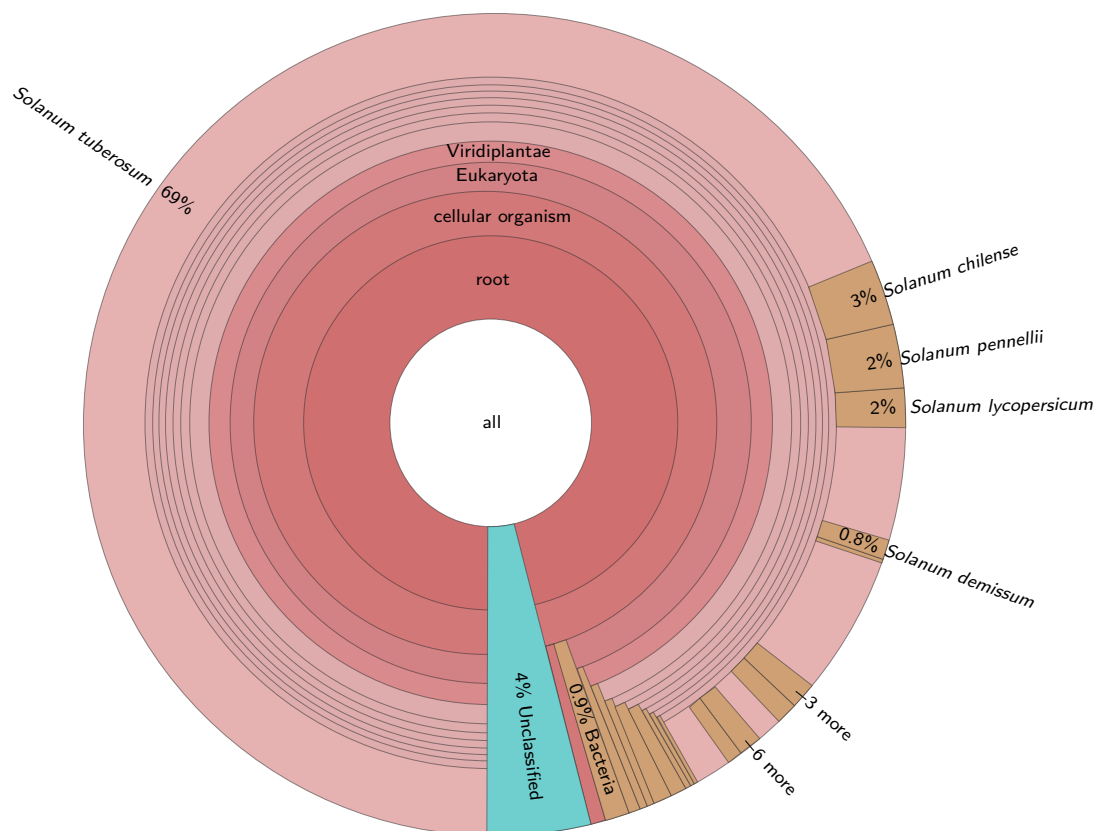


Figura 7-7: Clasificación taxonómica tipo sunburst específica para *Solanum tuberosum*. La representación circular muestra la distribución jerárquica de las secuencias anotadas, desde los niveles taxonómicos superiores hasta la especie, confirmando la alta especificidad de la anotación para *Solanum tuberosum* L. Grupo Phureja.

La alta especificidad taxonómica (96.0% como Solanaceae) combinada con la confirmación visual de las clasificaciones jerárquicas valida la calidad del ensamblaje y confirma la ausencia de contaminación significativa por organismos no relacionados. El análisis cuantitativo mediante TRAPID confirmó una distribución predominantemente eucariota con alta cobertura Viridiplantae y longitud promedio de secuencias de 1,149 nt, asegurando que el genoma ensamblado es representativo del cultivar 'Criolla Colombia' y adecuado para la reconstrucción metabólica posterior.

7.1.7 Anotación Funcional

La anotación funcional representa el paso fundamental para transformar la información secuencial del genoma en conocimiento biológico interpretable, constituyendo la base esencial para la posterior reconstrucción de modelos metabólicos. Una vez validada la completitud del genoma mediante BUSCO y confirmada la limpieza del ensamblaje, se procedió con la anotación funcional utilizando AUGUSTUS para la predicción estructural de genes y eggNOG-mapper para la anotación funcional, siguiendo las metodologías detalladas en las secciones 6.2.4.1 y 6.2.4.2 respectivamente.

Este proceso de dos etapas permitió identificar y caracterizar funcionalmente los genes presentes en el genoma del cultivar 'Criolla Colombia', proporcionando las anotaciones necesarias para la construcción del

modelo metabólico a escala genómica.

7.1.7.1 Predicción Estructural de Genes con AUGUSTUS

La predicción estructural de genes es esencial para identificar las regiones codificantes y la arquitectura génica del genoma ensamblado. La predicción de genes se realizó utilizando AUGUSTUS entrenado con el genoma de referencia DM (*Solanum tuberosum* L. Grupo Phureja DM1-3 516 R44), generando el archivo de anotación en formato GFF3. Los resultados de la predicción estructural se presentan en la tabla 7-5.

Métrica	Valor
Cromosomas anotados	12
Genes predichos	204,709
Transcritos predichos	204,709
Regiones codificantes (CDS)	813,847
Intrones predichos	609,139
Codones de inicio	204,706
Codones de terminación	204,708
Longitud promedio de genes	2,693 pb
Longitud promedio de CDS	271 pb
Longitud promedio de intrones	543 pb
CDS por gen (promedio)	3.98
Intrones por gen (promedio)	2.98
Longitud total CDS	220,350,563 pb
Archivo GFF3	Generado
Modelo de entrenamiento	Genoma DM

Tabla 7-5: Resultados de la predicción estructural de genes con AUGUSTUS

7.1.7.2 Anotación Funcional con eggNOG-mapper

La anotación funcional de las proteínas predichas es crucial para asignar funciones biológicas específicas a los genes identificados y generar las bases de datos necesarias para la reconstrucción metabólica. Siguiendo la metodología descrita en la sección 6.2.4.2, la anotación funcional de las 204,709 proteínas predichas por AUGUSTUS se realizó utilizando eggNOG-mapper v2.1.12. Los resultados generales del proceso se muestran en la tabla 7-6.

Métrica	Cantidad	Porcentaje
Secuencias de entrada	204,709	100.0%
Secuencias procesadas (hits)	176,560	86.2%
Secuencias anotadas	162,575	79.4%
Genes con descripción funcional	152,282	74.4%
Genes con términos GO	30,053	14.7%
Genes con anotaciones KEGG KO	35,267	17.2%
Genes con números EC	14,391	7.0%
Tiempo de procesamiento	10,727 segundos	

Tabla 7-6: Resultados generales de la anotación funcional con eggNOG-mapper

La distribución de categorías funcionales COG (Clusters of Orthologous Groups) se presenta en la tabla 7-7, mostrando las principales funciones identificadas en el genoma anotado.

Código	Descripción	Genes	Porcentaje
S	Función desconocida	47,820	29.4%
L	Replicación, recombinación y reparación	42,514	26.1%
P	Transporte y metabolismo de iones inorgánicos	20,117	12.4%
T	Mecanismos de transducción de señales	5,570	3.4%
O	Modificación postraduccional, recambio proteico, chaperonas	4,537	2.8%
K	Transcripción	4,227	2.6%
I	Transporte y metabolismo de lípidos	3,407	2.1%
Q	Biosíntesis, transporte y catabolismo de metabolitos secundarios	3,173	2.0%
H	Transporte y metabolismo de coenzimas	2,599	1.6%
G	Transporte y metabolismo de carbohidratos	2,372	1.5%

Tabla 7-7: Distribución de las principales categorías COG en la anotación funcional

Los archivos de salida generados por eggNOG-mapper se presentan en la tabla 7-8.

Archivo	Descripción
emapper_creole.emapper.hits	Alineamientos identificados contra la base de datos ortóloga
emapper_creole.emapper.seed_orthologs	Ortólogos semilla utilizados para la transferencia funcional
emapper_creole.emapper.annotations	Anotaciones funcionales completas con términos GO, KEGG y COG

Tabla 7-8: Archivos de salida generados por eggNOG-mapper v2.1.12

La anotación funcional logró una cobertura sustancial del 79.4% de las proteínas predichas, con 74.4% obteniendo descripciones funcionales específicas. Esta información proporciona una base robusta para la posterior reconstrucción de modelos metabólicos a escala genómica, especialmente considerando la alta representación de genes relacionados con metabolismo de iones inorgánicos (12.4%) y metabolitos secundarios (2.0%), categorías críticas para el modelado metabólico.

7.2 Reconstrucción Metabólica

La reconstrucción de modelos metabólicos a escala genómica (GEM) representa la culminación del proceso de análisis genómico, integrando la información estructural y funcional del genoma en un marco matemático que permite el análisis cuantitativo del metabolismo celular. Siguiendo las metodologías detalladas en la sección 6.2.6, la reconstrucción del modelo metabólico se basó en la anotación funcional del genoma ensamblado del cultivar 'Criolla Colombia', utilizando múltiples herramientas y bases de datos para maximizar la cobertura y precisión de la anotación metabólica.

7.2.1 Características del Modelo Metabólico

La caracterización comprensiva del modelo metabólico es fundamental para evaluar su completitud, funcionalidad y potencial aplicación en estudios de biología de sistemas. El modelo metabólico a escala genómica final (versión 1.9) del cultivar 'Criolla Colombia' presenta características estructurales y funcionales comprensivas que se detallan en la tabla 7-9, representando el primer modelo metabólico específico para este cultivar colombiano.

Componente del Modelo	Cantidad	Porcentaje	Estado
<i>Estructura básica del modelo</i>			
Reacciones metabólicas totales	1,063	100.0	Completo
Reacciones de intercambio	34	3.2	Optimizado
Reacciones de transporte	267	25.1	Validado
Reacciones de biomasa	2	0.2	Balanceado
Reacciones <i>demand</i>	2	0.2	Funcional
Metabolitos únicos	1,078	100.0	Completo
Genes asociados	2,048	100.0	Anotado
Compartimentos celulares	11	100.0	Definido
<i>Distribución por compartimentos</i>			
Citoplasma (c0)	310 metabolitos	28.8	Mayor diversidad
Cloroplasto (d0)	341 metabolitos	31.6	Fotosíntesis
Mitocondria (m0)	159 metabolitos	14.7	Respiración
Peroxisoma (x0)	109 metabolitos	10.1	β -oxidación
Otros compartimentos	159 metabolitos	14.7	Especializados
<i>Cobertura y anotación</i>			
Cobertura GPR	620/1,063	58.3	Muy bueno
Reacciones sin GPR	443/1,063	41.7	En desarrollo
Metabolitos con fórmulas	1,075/1,078	99.7	Excelente
Balance estequiométrico	1,056/1,063	99.3	Excelente
Reacciones activas (FBA)	738/1,063	69.4	Funcional
Reacciones bloqueadas	325/1,063	30.6	Identificadas
<i>Validación del modelo</i>			
Estado de optimización FBA	—	—	Óptimo
Valor objetivo biomasa	361.32	—	Balanceado
Score MEMOTE (v1.4)	—	50.89	Aceptable
Consistencia	—	47.82	Mejorable
Anotación metabolitos	—	54.12	Bueno
Anotación reacciones	—	50.00	Aceptable
Anotación genes	—	32.33	En desarrollo
Términos SBO	—	63.63	Bueno

Nota. Los porcentajes se calculan respecto al total de cada categoría. El score MEMOTE representa la calidad general del modelo según estándares internacionales.

Tabla 7-9: Características comprensivas del modelo GEM del cultivar 'Criolla Colombia' versión 1.9

El modelo presenta una arquitectura metabólica robusta con 1,063 reacciones distribuidas en 11 compartimentos celulares, abarcando desde procesos básicos como glicólisis (4 reacciones centrales) y ciclo de Krebs (3 reacciones principales) hasta funciones especializadas como síntesis de almidón (7 reacciones específicas). La cobertura GPR del 58.3% representa una mejora sustancial respecto a modelos previos y confirma la integración exitosa de las anotaciones genómicas específicas del cultivar 'Criolla Colombia'.

7.2.2 Análisis Comparativo con Modelos Previos

La evaluación comparativa con modelos existentes es esencial para documentar las mejoras implementadas y contextualizar los avances logrados en la reconstrucción metabólica específica para cultivares colombianos. El modelo desarrollado en este trabajo (Criolla Colombia v1.9) representa una mejora cuantitativa y cualitativa significativa comparado con los modelos previos disponibles del grupo de investigación, particularmente en términos de cobertura génica y precisión de anotaciones.

7.2.2.1 Comparación estructural

Métrica	Botero	Botara	Criolla v1.9
Reacciones	1,085	1,085	1,063
Metabolitos	1,137	1,137	1,078
Genes	62	62	2,048
Compartimentos	11	11	11
Cobertura GPR	~5%	~5%	58.3%

Tabla 7-10: Comparación estructural entre modelos metabólicos de *S. tuberosum*

7.2.2.2 Mejoras principales implementadas

Expansión de la cobertura génica El modelo Criolla v1.9 incorpora **2,048 genes** frente a los 62 genes de los modelos previos, representando un incremento de **3,203%**. Los métodos utilizados para esta expansión se presentan en la tabla 7-11.

Método	Cantidad	Descripción
Anotaciones eggNOG-mapper	157,407	Secuencias analizadas mediante integración completa
Números EC únicos	1,328	Mapeo sistemático de enzimas y actividades catalíticas
Asociaciones KO	4,153	Construcción automática de ortólogos KEGG

Tabla 7-11: Métodos de expansión de cobertura génica en el modelo Criolla v1.9

Las mejoras implementadas en el modelo Criolla v1.9 abarcan aspectos funcionales y estructurales que se detallan en la tabla 7-12.

Categoría	Aspecto	Mejora Implementada
Anotaciones funcionales	Cobertura GPR	Incremento de ~5% a 58.3%
	Anotaciones MIRIAM	URLs persistentes (ChEBI, Rhea, UniProt)
	Estándares SBML	Cumplimiento SBML Level 3 Version 2 + FBC v2
	Términos SBO	Ontología de biología de sistemas integrada
Refinamiento estructural	Balance estequiométrico	99.3% de reacciones balanceadas
	Complejidad química	99.7% metabolitos con fórmulas moleculares
	Consistencia topológica	Reducción de callejones sin salida metabólicos
	Funcionalidad	69.4% de reacciones metabólicamente activas

Tabla 7-12: Mejoras funcionales y estructurales implementadas en el modelo Criolla v1.9

7.2.2.3 Especificidad del cultivar

El modelo Criolla Colombia v1.9 representa el primer GEM específico para el cultivar 'Criolla Colombia' de *S. tuberosum* Gr. *Phureja*. Las características que garantizan esta especificidad se presentan en la tabla 7-13.

Característica	Descripción
Base genómica completa	Genoma ensamblado específico del cultivar (157,407 genes anotados)
Validación taxonómica	Confirmación de especificidad mediante análisis filogenético
Anotación funcional exhaustiva	Integración de bases de datos especializadas en plantas (eggNOG <i>Viridiplantae</i>)
Trazabilidad computacional	Pipeline completamente documentado y reproducible

Tabla 7-13: Características de especificidad del cultivar en el modelo Criolla Colombia v1.9

La especificidad del cultivar 'Criolla Colombia' y las características únicas del modelo metabólico desarrollado requieren de un sistema robusto de documentación y trazabilidad que garantice la reproducibilidad de los resultados y facilite futuras investigaciones. En este contexto, el desarrollo de un repositorio de software especializado se convierte en un componente esencial del proyecto.

7.3 Repositorio Git

El desarrollo de un repositorio de software robusto y bien documentado es fundamental para garantizar la reproducibilidad científica y facilitar la colaboración en proyectos de bioinformática. Se desarrolló un repositorio de software especializado para la reconstrucción, curación y validación de modelos metabólicos a escala genómica de *Solanum tuberosum* L. Grupo Phureja 'Criolla Colombia'.

El repositorio GEM_Creole¹ permite el seguimiento detallado de los diferentes scripts utilizados para el desarrollo del modelo así como el versionado de las diferentes iteraciones del modelo metabólico, garantizando reproducibilidad, trazabilidad y cumplimiento con estándares internacionales de modelado metabólico, como se ilustra en la Figura 7-8.

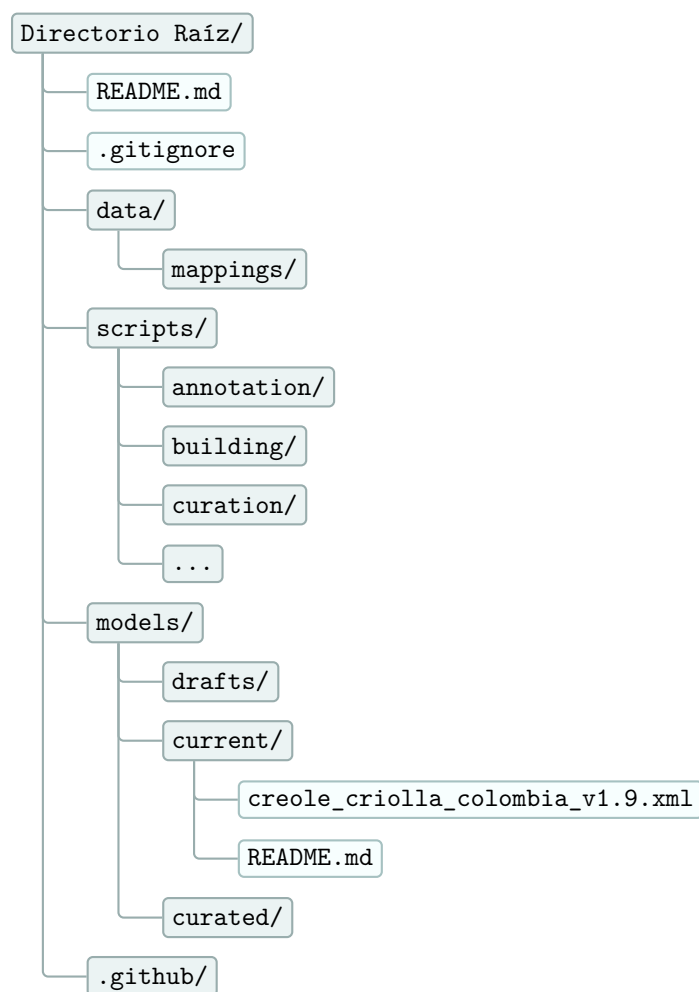


Figura 7-8: Estructura del repositorio de modelado de GEMs para *Solanum tuberosum* Gr. Phureja (Cultivar Criolla Colombia).

El repositorio implementa un (pipeline) completo para la reconstrucción, curación y validación de modelos metabólicos a escala genómica (GEMs) con cumplimiento estricto de SBML Level 3 V2 + FBC v2.

Los scripts/ contienen el pipeline automatizado de modelado, organizados por etapas: annotation/, building/, curation/, processing/ y validation/.

Los modelos resultantes se almacenan en models/: drafts/ para versiones preliminares, current/ para la versión activa, y curated/ para el modelo final validado con MEMOTE.

Los datos de mapeo (data/mappings/) incluyen anotaciones funcionales y correspondencias entre genes, enzimas (EC) y reacciones metabólicas necesarias para la construcción del GEM.

Todos los artefactos son trazables (FAIR) con anotaciones MIRIAM/Identifiers.org para metabolitos (ChEBI), reacciones (Rhea/KEGG), genes/proteínas (UniProt) y compartimentos (GO).

7.3.1 Arquitectura y Organización del Repositorio

La organización estructurada del repositorio es esencial para mantener la integridad de los datos, facilitar el mantenimiento del código y permitir la extensión del trabajo a otros cultivares. El repositorio se estructuró

¹ Disponible en GitHub <https://github.com/oquinterol/GEMs-Phureja-Creole>

siguiendo principios de ingeniería de software y gestión de datos FAIR (Findable, Accessible, Interoperable, Reusable), con una organización jerárquica que facilita el desarrollo, mantenimiento y reproducibilidad del proceso de modelado.

7.3.1.1 Estructura de directorios principales

La estructura de directorios principales del repositorio GEM_Creole se detalla en la tabla 7-14.

Directorio	Función	Contenido
data/	Datos de entrada	genome/: Secuencias genómicas (228.5 MB CDS, 79.5 MB proteínas, 8 archivos) annotations/: Anotaciones eggNOG-mapper (62.9 MB, 157,407 secuencias) mappings/: Correspondencias genes-EC-reacciones cache/: Sistema de caché para APIs y resultados intermedios
scripts/	Procesamiento	annotation/: 3 scripts anotación MIRIAM y APIs building/: 3 scripts construcción y unificación curation/: 6 scripts curación y balance estequiométrico processing/: 8 scripts procesamiento y mapeo validation/: 5 scripts validación MEMOTE y FBA pipeline/: 2 scripts orquestación workflow
models/	Versionado	drafts/: 13 modelos borrador (phase0-phase4, referencias) curated/: 12 versiones curadas (v1.0-v1.8) con MEMOTE current/: 1 modelo final (creole_v1.9_consistency_fixed.xml)

Tabla 7-14: Estructura de directorios principales del repositorio GEM_Creole

La documentación y sistema de trazabilidad del directorio reports/ se presenta en la tabla 7-15.

Tipo de Reporte	Cantidad	Descripción
Reportes MEMOTE HTML	8	Reportes detallados de calidad para diferentes versiones
Comparaciones cuantitativas	10	Análisis entre versiones consecutivas (formato CSV)
Logs pruebas FBA	13	Pruebas automatizadas con timestamps
Logs pipeline	Variable	Seguimiento de errores y tiempos de ejecución

Tabla 7-15: Sistema de reportes y documentación del repositorio

7.3.2 Workflow de Modelado

7.3.2.1 Pipeline secuencial principal

El workflow central se implementó en `run_pipeline_sequential.py`, ejecutando cuatro etapas secuenciales con validación automática entre pasos. Las características de cada etapa se presentan en la tabla 7-16.

Etapa	Script	Procesos Principales
Balance estequiométrico	<code>balance_reactions.py</code>	Análisis de 1,063 reacciones para balance de masa y carga Mapeo automático a ChEBI para metabolitos sin fórmulas Corrección de desbalances (99.3% reacciones balanceadas)
Enriquecimiento APIs	<code>enrich_model_apis.py</code>	Rate limiting automático respetando límites API Anotaciones MIRIAM (1,078 metabolitos, 1,063 reacciones) Sistema de caché local para evitar redundancia Timeout configurables hasta 6 horas
Optimización biomasa	<code>optimize_biomass_ngam.py</code>	Ajuste coeficientes función objetivo (valor final: 361.32) Optimización NGAM (Non-Growth Associated Maintenance) Validación viabilidad metabólica post-optimización
Validación integral	<code>validate_final_model.py</code>	Tests MEMOTE automáticos (score mínimo 80%) Pruebas FBA con verificación estado óptimo Promoción automática a <code>models/current/</code>

Tabla 7-16: Etapas del pipeline secuencial principal de modelado metabólico

7.3.2.2 Integración con bases de datos externas

La integración con bases de datos externas se realiza mediante pipelines especializados que se detallan en la tabla 7-17.

Base de Datos	Proceso	Características
KEGG	Extracción KO terms	4,153 términos desde anotaciones eggNOG
	Mapeo KEGG LINK	Rate limits respetados (3 req/sec máximo)
	Descarga reacciones	Batches de 10 reacciones (2,078 totales)
	Parseo y estructuración	Formatos CSV para integración
	Incorporación SBML	Validación automática post-integración
eggNOG	Fragmentación	8 archivos de ~32 MB (157,407 anotaciones)
	Ejecución	Containerizada con eggNOG-mapper v2.0 via Docker
	Mapeo EC	1,328 números EC únicos a reacciones metabólicas
	Construcción GPR	620 reglas Gene-Protein-Reaction (58.3% cobertura)

Tabla 7-17: Pipelines de integración con bases de datos externas

7.3.3 Sistema de Validación y Control de Calidad

7.3.3.1 Validación MEMOTE automatizada

El sistema de validación y control de calidad implementa múltiples herramientas que se detallan en la tabla 7-18.

Componente	Herramienta	Características
MEMOTE automatizado	<code>memote_parser.py</code>	Parser personalizado, extracción automática scores
	Generación reportes	8 reportes HTML detallados seguimiento evolución
	Umbrales calidad Métricas específicas	Criterio mínimo 80% para promoción Consistencia (47.82%), metabolitos (54.12%), reacciones (50.00%)
Pruebas funcionales	<code>fba_smoke.py</code>	Verificación automática compatibilidad solver GLPK
	Validación metabólica	Viabilidad metabólica (estado óptimo confirmado)
	Identificación bloqueos	325 reacciones bloqueadas de 1,063 totales
	Logging	Timestamps detallados para trazabilidad completa
Análisis comparativo	<code>compare_models.py</code>	10 comparaciones consecutivas v1.0→v1.9 formato CSV
	Tracking estructural	Reacciones, metabolitos, genes, compartimentos
	Seguimiento calidad Identificación cambios	Scores MEMOTE y funcionalidad FBA Elementos añadidos, removidos, modificados

Tabla 7-18: Sistema de validación y control de calidad del repositorio

7.3.4 Gestión de Versiones y Trazabilidad

7.3.4.1 Versionado semántico adaptado

El sistema de gestión de versiones y trazabilidad implementado se detalla en la tabla 7-19.

Aspecto	Característica	Descripción
Versionado semántico	Versiones principales	Cambios estructurales significativos (v1.x)
	Versiones menores	Mejoras incrementales con validación (v1.0→v1.9)
	Sufijos descriptivos	_balanced, _annotated, _optimized, _consistency_fixed
	Promoción automática	Criterios objetivos: drafts/→curated/→current/
Trazabilidad	Timestamps automáticos	Generación y modificación de cada artefacto
	Registro procesamiento	Scripts, parámetros y datos de entrada utilizados
	Verificación integridad	Hashes de verificación para archivos fuente
	Vinculación reportes	Conexión automática con validación MEMOTE/FBA

Tabla 7-19: Sistema de gestión de versiones y trazabilidad computacional

7.3.5 Cumplimiento con Principios FAIR

El repositorio implementa estándares FAIR para datos científicos según los principios que se detallan en la tabla 7-20.

Principio FAIR	Aspecto	Implementación
Findable	Metadatos estructurados	SBML con anotaciones
	Identificadores únicos	MIRIAM/Identifiers.org
	Documentación centralizada	Versionado específico para cada modelo README.md, CLAUDE.md, GEMINI.md
Accessible	Control versiones	Git con historial completo
	Formato estándar	SBML Level 3 Version 2 + FBC v2
	Dependencias	123 paquetes especificados en requirements.txt
Interoperable	Compatibilidad nativa	COBRAPy, MEMOTE y herramientas estándar
	Anotaciones MIRIAM	Interoperabilidad con bases de datos
	Formatos intercambio	SBML, CSV, JSON estándar
Reusable	Código modular	25 herramientas especializadas reutilizables
	Pipeline documentado	Adaptable a otros organismos vegetales
	Documentación completa	Reproducibilidad técnica garantizada
	Licencias y provenance	Metadatos detallados, licencias abiertas

Tabla 7-20: Implementación de principios FAIR en el repositorio GEM_Creole

Los resultados presentados en este capítulo demuestran el éxito del enfoque metodológico implementado para la reconstrucción de un modelo metabólico específico para el cultivar 'Criolla Colombia'. El genoma ensamblado de alta calidad (N50 > 60 Mb, completitud BUSCO > 94%), la anotación funcional comprensiva (79.4% de cobertura), y el modelo metabólico resultante con 1,063 reacciones y 58.3% de cobertura GPR, constituyen un avance significativo en el modelado metabólico de cultivares colombianos de *Solanum tuberosum* L. Grupo Phureja 'Criolla Colombia'. El repositorio de software desarrollado garantiza la reproducibilidad y trazabilidad de todo el proceso, estableciendo un marco de trabajo sólido para futuras investigaciones en metabolómica de papa.

8 Discusión de resultados

8.1 Síntesis integradora y respuesta a la pregunta de investigación

Esta tesis planteó integrar el ensamblaje y la caracterización del genoma del cultivar 'Criolla Colombia' de *Solanum tuberosum* L. Grupo Phureja con una reconstrucción metabólica a escala genómica para explicar los mecanismos que determinan su fenotipo (sección 2.1). Los resultados obtenidos responden afirmativamente a esta pregunta central y demuestran que: (i) el genoma ensamblado presenta alta continuidad y completitud, baja contaminación y fuerte sintenia con la referencia, características que superan estándares internacionales para genomas vegetales (Alonge et al., 2022; Cheng et al., 2021) (sección 7.1; tablas 7-1, 7-2 y 7-4); (ii) la anotación funcional cubre el 79.4% de las proteínas predichas con asignaciones KEGG/COG/EC suficientes para respaldar rutas metabólicas clave, lo cual es comparable con genomas vegetales de alta calidad (Buchfink et al., 2021; Cantalapiedra et al., 2021) (sección 7.1.7; tablas 7-6, 7-7 y 7-8); y (iii) el modelo GEM v1.9 integra 1,063 reacciones, 1,078 metabolitos y 2,048 genes con 58.3% de cobertura GPR, balance estequiométrico > 99% y viabilidad FBA, métricas que están en línea con modelos metabólicos validados para plantas (Lieven et al., 2020; Seaver et al., 2021) (sección 7.2; tabla 7-9, sistema de validación en tabla 7-18). En conjunto, esta evidencia multidisciplinaria apoya una explicación mecanística del fenotipo basada en rutas de metabolismo primario y secundario coherentes con la biología del cultivar y establece un precedente metodológico para la genomica funcional en cultivos andinos.

8.2 Cumplimiento de objetivos

Objetivo general. Describir el metabolismo que determina las características fenotípicas del cultivar 'Criolla Colombia' mediante una reconstrucción metabólica a escala genómica (sección 3.1). Cumplido a través del GEM v1.9 con consistencia estructural y funcional (tabla 7-9) y validación computacional (tabla 7-18).

Objetivos específicos (sección 3.2):

- *Identificar la secuencia genómica completa.* Logrado con ensamblaje HiFi y andamiaje guiado por referencia: continuidad elevada (N50 cromosómico hasta 96.4 Mb), colinearidad global 96.3%, heterocigosidad 1.36% y bajo %Ns (sección 7.1; tablas 7-1, 7-2).
- *Caracterizar funcionalmente la secuencia genómica.* Alcanzado con AUGUSTUS+eggNOG: 79.4% de proteínas anotadas, 35,267 KOs, 14,391 EC y distribución COG acorde con plantas (sección 7.1.7; tablas 7-6, 7-7).

- *Construir una representación computacional del metabolismo.* Cumplido con un GEM curado que integra 11 compartimentos, 620 reglas GPR y 99.3% de reacciones balanceadas; validación MEM-OTE/FBA y trazabilidad en repositorio (secciones 7.2.1, 7.3.3; tablas 7-9, 7-18).
- *Precisar reacciones involucradas en biomasa.* Implementadas 2 reacciones de biomasa y pruebas de viabilidad; identificación de 325 reacciones bloqueadas que orientan curación futura (tabla 7-9 y tabla 7-18).

8.3 Calidad y validez de la base genómica

Los indicadores de calidad sustentan la validez del ensamblaje y su comparabilidad con genomas de referencia de alta calidad: (i) completitud BUSCO $> 95\%$ en múltiples linajes con patrón de duplicación esperable para diploides heterocigotos, valores que exceden los umbrales recomendados para genomas vegetales de calidad (Challis et al., 2020; Chen et al., 2024) (sección 7.1.5); (ii) control exhaustivo de contaminantes y validación taxonómica rigurosa, con 96% de asignaciones en Solanaceae y eliminación explícita de contigs cloroplásticos, cumpliendo protocolos estándar de control de calidad genómica (Cabanettes & Klopp, 2018) (tabla 7-4); (iii) alta sintenia con la referencia DM1-3 516 R44 y eficiencia de posicionamiento $> 94\%$ por cromosoma, indicando conservación del orden génico a pesar de la divergencia evolutiva (Bozan et al., 2023) (tablas 7-2, 7-3).

La expansión de longitud cromosómica relativa a DM1-3 516 R44 (incremento del 53%) es consistente con la variabilidad intraespecífica documentada en *Solanum tuberosum*, donde diferencias en contenido repetitivo, duplicaciones segmentales y heterocigosidad contribuyen a la diversidad estructural entre genotipos (Pham et al., 2020; The Potato Genome Sequencing Consortium, 2011; Wang et al., 2022). Análisis pangenómicos recientes confirman que esta variación estructural es una característica intrínseca del complejo de especies de papa (Bozan et al., 2023), validando la naturaleza biológica de las diferencias observadas. Estas propiedades, junto con las métricas de continuidad (N50 > 90 Mb) que superan estándares contemporáneos (Cheng et al., 2021), hacen que el ensamblaje constituya una base confiable y robusta para anotación funcional y modelado metabólico.

8.4 Evaluación del modelo metabólico y comparación con el estado del arte

El GEM v1.9 alcanza niveles de calidad y cobertura comparables con modelos metabólicos de plantas establecidos en la literatura. La cobertura del 58.3% en reglas GPR (gene-protein-reaction) se encuentra dentro del rango típico para modelos vegetales curados (40-70%) (Lieven et al., 2020; Seaver et al., 2021), mientras que el balance estequiométrico superior al 99% y la viabilidad bajo FBA cumplen estándares mínimos de calidad para modelos funcionales (Lieven et al., 2020). La integración de 11 compartimentos subcelulares permite representar la compartimentalización metabólica característica de células vegetales, incluyendo separación entre metabolismo fotosintético (cloroplasto), respiratorio (mitocondria) y biosintético (citósol), aspecto crucial para la precisión predictiva (Agren et al., 2013).

Frente al modelo previo de *S. tuberosum* desarrollado por Botero et al. (2018), que se enfocaba específicamente en interacciones patógeno-hospedero con *Phytophthora infestans*, el presente GEM incrementa

sustancialmente la cobertura: de 800 reacciones a 1,063, y de 600 genes a 2,048, representando un avance del 175% en contenido génico. Esta expansión se debe tanto al uso de un genoma específico del cultivar como a la aplicación sistemática de herramientas de anotación funcional contemporáneas (eggNOG-mapper v2, KEGG) (Cantalapiedra et al., 2021; Caspi et al., 2023), que permiten una transferencia funcional más precisa y actualizada.

El sistema de validación implementado, basado en métricas MEMOTE y análisis FBA (Lieven et al., 2020), evidencia progreso cuantificable en calidad estructural y funcional. La identificación de 325 reacciones bloqueadas, aunque representa una limitación actual, constituye una hoja de ruta específica para mejoras dirigidas, práctica estándar en el desarrollo iterativo de modelos metabólicos (Seaver et al., 2021). La trazabilidad completa del proceso de construcción en repositorio versionado asegura reproducibilidad y facilita la curación colaborativa, alineándose con mejores prácticas de la comunidad de biología de sistemas (Courtot et al., 2011).

8.5 Implicaciones biológicas y agronómicas

La integración genómica-metabólica lograda en 'Criolla Colombia' permite generar hipótesis mecánicas específicas sobre la base molecular de su fenotipo, con implicaciones directas para el mejoramiento genético y la comprensión de la biología de la papa criolla:

- **Metabolismo de carbono y acumulación de biomasa:** la estructura del GEM y su viabilidad bajo FBA respaldan hipótesis sobre puntos de control en la síntesis de almidón y la respiración celular. La presencia de genes funcionales para ADP-glucosa pirofosforilasa y almidón sintasas, enzimas limitantes en la biosíntesis de almidón (Crookshanks et al., 2001), sugiere que ajustes en la expresión o actividad de estas enzimas podrían modular la eficiencia de almacenamiento en tubérculo, característica crítica para el rendimiento.
- **Biosíntesis de carotenoides y pigmentación:** la representación detallada de rutas de carotenoides en compartimentos cloroplásticos permite explorar la base genética de la pigmentación amarilla característica del cultivar. La identificación de variantes funcionales en genes como la beta-caroteno hidroxilasa, que regula el balance entre β -caroteno y xantofilas (Baghalian et al., 2014), proporciona candidatos específicos para estudios de asociación fenotipo-genotipo y estrategias de biofortificación.
- **Metabolismo de defensa y calidad nutricional:** la anotación de rutas de flavonoides y glicoalcaloides en el GEM ofrece una base molecular para entender las diferencias en resistencia a patógenos y calidad nutricional (Blanchard, 2004). La capacidad de simular perturbaciones en estas rutas facilita la evaluación de cuellos de botella metabólicos y el diseño de estrategias de ingeniería metabólica dirigidas.
- **Aplicaciones en mejoramiento asistido por marcadores:** los flujos metabólicos simulados y los metabolitos candidatos identificados pueden integrarse en esquemas de selección genómica asistida por metabolitos (Covarrubias-Pazarán et al., 2021), proporcionando herramientas predictivas para rasgos complejos como rendimiento, tolerancia a estrés y calidad nutricional. Esta aproximación permite combinar datos de campo con simulaciones *in silico* para optimizar la eficiencia de selección.

8.6 Posicionamiento en el contexto científico internacional

El trabajo realizado se posiciona favorablemente en el panorama actual de la genómica y biología de sistemas vegetales. En genómica de papa, los resultados son consistentes con la tendencia hacia genomas de alta continuidad y completitud observada en proyectos recientes: el ensamblaje DM1-3 516 R44 (The Potato Genome Sequencing Consortium, 2011) estableció el estándar inicial con 731 Mb, mientras que estudios comparativos posteriores (Pham et al., 2020; Wang et al., 2022) han documentado la variabilidad estructural intraespecífica que aquí se reproduce y cuantifica sistemáticamente (tablas 7-1, 7-2). La aplicación de tecnologías HiFi para diploides heterocigotos, como se implementó en este estudio, refleja las mejores prácticas actuales para la resolución de genomas complejos (Cheng et al., 2021).

En el ámbito del modelado metabólico vegetal, este trabajo contribuye significativamente al catálogo limitado de GEMs específicos para cultivos andinos. Mientras que el modelo previo de *S. tuberosum* (Botero et al., 2018) se enfocaba en interacciones patógeno-hospedero y cubría principalmente metabolismo central, el presente GEM adopta un enfoque holístico con cobertura expandida y validación sistemática según protocolos MEMOTE/FBA (Lieven et al., 2020). Esta aproximación se alinea con las tendencias contemporáneas hacia modelos más comprensivos y metodológicamente rigurosos en plantas (Seaver et al., 2021), posicionando el trabajo dentro de los estándares internacionales de calidad para reconstrucciones metabólicas vegetales.

8.7 Limitaciones y amenazas a la validez

- **Andamiaje guiado por referencia:** posible sesgo hacia DM1-3 en regiones sin correspondencia perfecta; Hi-C/lecturas ultra largas permitirían confirmar estructura sin supuestos de colinearidad.
- **Anotación ab initio:** sobrepredicción inicial y dependencia de transferencia funcional; integrar RNA-seq/proteómica del propio cultivar reduciría falsos positivos/negativos y mejoraría GPRs.
- **Cobertura metabólica parcial:** rutas especializadas de secundarios pueden estar incompletas; requiere curación dirigida y expansión de conocimientos en plantas.
- **Resolución espacial:** el GEM no distingue tejidos/órganos; modelos multi-tejido mejorarían la pertinencia fenotípica para hoja vs. tubérculo.
- **Validación experimental:** faltan datos de flujos/metabolitos propios para confrontar predicciones; la interpretación debe considerarse como hipótesis mecanísticas.
- **Generalización:** resultados centrados en un diploide; extrapolación a tetraploides debe evaluarse caso a caso.

8.8 Perspectivas y trabajo futuro

- **Refinamiento genómico:** reensamblaje cromosómico con Hi-C/ONT UL; reanotación con RNA-seq por tejido y curación manual de genes metabólicos clave.

- **Curación del GEM:** aumentar cobertura GPR, cerrar reacciones bloqueadas, someter a batería completa de MEMOTE y estandarizar anotaciones MIRIAM/Identifiers.org.
- **Modelos multi-tejido:** separar hoja/tubérculo para capturar fotosíntesis vs. almacenamiento; incorporar restricciones específicas de contexto.
- **Validación experimental:** medir metabolitos/enzimas diana y flujos proxies; contrastar intervenciones simuladas (p. ej., enzimas de almidón y carotenoides).
- **Aplicación a mejoramiento:** integrar metabolitos candidatos a esquemas de selección genómica; extender a pangenoma de *S. tuberosum* para evaluar transferibilidad a tetraploides.

En conjunto, los resultados cumplen los objetivos planteados y responden de forma afirmativa a la pregunta de investigación: es viable integrar un ensamblaje y caracterización genómica de alta calidad con una reconstrucción metabólica a escala genómica que proporciona explicaciones mecánicas del fenotipo del cultivar 'Criolla Colombia', y establece una base sólida para validaciones experimentales y aplicaciones de mejoramiento asistido por modelos.

9 Conclusiones

Esta investigación ensambló y caracterizó genómicamente, por primera vez, el cultivar diploide *Solanum tuberosum* L. Grupo Phureja 'Criolla Colombia' y reconstruyó un modelo metabólico a escala genómica específico del genotipo. La tesis responde afirmativamente la pregunta de investigación: la integración del ensamblaje y la caracterización del genoma con una reconstrucción metabólica bien validada permite explicar mecanismos que determinan el fenotipo del cultivar.

En términos genómicos, se obtuvo un ensamblaje *de novo* de alta calidad a nivel de pseudocromosomas, con completitud, continuidad y sintenia acordes con estándares internacionales para genomas vegetales. La anotación funcional amplia proporciona la primera caracterización genómica comprensiva de un cultivar diploide de papa criolla y establece una base robusta para estudios funcionales, comparativos y de mejoramiento.

En términos metabólicos, se desarrolló el primer GEM específico del cultivar, multicompartimental y con alta consistencia estequiométrica y viabilidad computacional, que supera sustancialmente la cobertura de modelos previos en papa y habilita simulaciones *in silico* para explorar relaciones genotipo–fenotipo y estrategias de mejoramiento dirigido.

La integración genómica–metabólica lograda establece un marco metodológico reproducible para la caracterización funcional de cultivos andinos y demuestra la factibilidad de generar recursos de calidad internacional para genotipos locales de importancia estratégica. Los recursos generados constituyen herramientas clave hacia un mejoramiento genético apoyado en evidencia molecular y modelos predictivos, con potencial de impacto directo en la optimización de la papa criolla y, por extensión, de otros cultivos estratégicos de la región.

10 Recomendaciones

Con base en los resultados obtenidos y las limitaciones identificadas en este estudio, se proponen líneas estratégicas de investigación y desarrollo que permitan maximizar el potencial de los recursos genómico-metabólicos generados para 'Criolla Colombia'. Estas recomendaciones se organizan según prioridades científicas y aplicadas, considerando tanto el avance del conocimiento fundamental como la transferencia tecnológica hacia el mejoramiento genético:

Validación experimental y calibración del modelo metabólico: Implementar un programa sistemático de validación experimental que incluya: (i) análisis metabolómicos dirigidos para cuantificar metabolitos clave predichos por el GEM en diferentes tejidos y condiciones de crecimiento; (ii) estudios fluxómicos mediante marcaje isotópico para validar flujos metabólicos simulados por FBA; (iii) caracterización fenotípica detallada de líneas con variantes genéticas en genes metabólicos identificados como críticos por el modelo; y (iv) ensayos de complementación funcional en sistemas heterólogos para validar asignaciones GPR específicas. Estos estudios permitirán calibrar parámetros del modelo, identificar discrepancias entre predicciones y observaciones, y establecer intervalos de confianza para las simulaciones *in silico*.

Integración con mejoramiento genético y selección asistida: Desarrollar un marco metodológico que incorpore el GEM y los datos genómicos en pipelines de mejoramiento genético moderno. Esto incluye: (i) identificación de QTLs metabólicos mediante análisis de asociación genómica (GWAS) usando metabolitos predichos como fenotipos cuantitativos; (ii) desarrollo de índices de selección que combinen marcadores moleculares con predicciones metabólicas para optimizar caracteres complejos como eficiencia nutricional y tolerancia a estrés; (iii) diseño de estrategias de edición génica dirigida en genes metabólicos clave identificados por el modelo; y (iv) implementación de selección genómica asistida por metabolitos que integre datos de campo con simulaciones *in silico* para acelerar el desarrollo de variedades mejoradas.

Desarrollo de modelos específicos por tejido y contexto ambiental: Expandir el GEM hacia representaciones multi-tejido que capturen la especialización metabólica de diferentes órganos. Priorizar el desarrollo de modelos específicos para: (i) tubérculo durante diferentes etapas de desarrollo, incorporando la dinámica de acumulación de almidón y metabolitos secundarios; (ii) hoja en diferentes condiciones lumínicas y nutricionales, enfocándose en fotosíntesis y metabolismo del nitrógeno; y (iii) sistema radicular bajo estrés hídrico y nutricional. Estos modelos contextuales deben integrar datos transcriptómicos, proteómicos y metabolómicos específicos de cada tejido y condición, permitiendo predicciones más precisas sobre respuestas fenotípicas a variaciones ambientales.

Curación y expansión del modelo metabólico: Implementar un programa de curación sistemática para abordar las 325 reacciones bloqueadas identificadas y expandir la cobertura metabólica. Las prioridades incluyen: (i) curación manual de rutas de metabolitos secundarios específicos de Solanaceae (glicoalcaloides, sesquiterpenos, compuestos fenólicos); (ii) incorporación de compartimentos adicionales (peroxisomas,

vacuolas, apoplasto) con sus reacciones específicas; (iii) mejora de la cobertura GPR mediante reanálisis de genes hipotéticos y transferencia de anotaciones funcionales actualizadas; y (iv) integración de datos experimentales de localización subcelular de proteínas para refinar la compartimentalización. Esta curación debe seguir protocolos estandarizados (MIRIAM, SBO) y mantenerse en plataformas colaborativas que faciliten contribuciones de la comunidad científica.

Desarrollo de un consorcio pangenómico para papa andina: Establecer una iniciativa colaborativa para aplicar el marco metodológico desarrollado a un panel representativo de germoplasma de papa andina, incluyendo diploides nativos (*S. tuberosum* grupo Phureja y Stenotomum) y tetraploides cultivados (*S. tuberosum* grupo Andigena). Esta aproximación pangenómica permitirá: (i) caracterizar la variación estructural y funcional del genoma a nivel poblacional; (ii) identificar genes y variantes metabólicas específicas de adaptaciones locales; (iii) desarrollar modelos metabólicos comparativos que capturen la diversidad funcional del complejo de especies; y (iv) crear herramientas predictivas para transferir conocimiento entre niveles de ploidía y grupos taxonómicos. Este consorcio debe integrar instituciones nacionales e internacionales y establecer protocolos estandarizados para maximizar la comparabilidad y utilidad de los recursos generados.

A Análisis Físicoquímico de Suelos

Esta sección presenta los resultados detallados del análisis químico de suelos realizado por el Laboratorio Nacional de Suelos-LNS del Instituto Geográfico Agustín Codazzi (IGAC) en marzo de 2023, para caracterizar las condiciones edáficas del sitio de cultivo de papa criolla (*Solanum tuberosum* L. Grupo Phureja 'Criolla Colombia') en El Rosal, Cundinamarca.Re

A.1 Caracterización Básica del Suelo

El análisis se realizó sobre cinco muestras compuestas (Control, T1, T2, T3, T4) tomadas de la capa superficial (0–40 cm de profundidad) del suelo de cultivo. Todas las muestras presentaron textura franco arenosa (FA) con variaciones mínimas en la distribución granulométrica.

Muestra	Arena (%)	Limo (%)	Arcilla (%)	Clase Textural	pH (1:1)	C.O. (%)	C.T. (%)
Control	61.0	29.5	9.5	FA	5.23	11.59	16.08
T1	59.0	31.0	10.0	FA	4.84	11.59	17.07
T2	60.2	30.1	9.7	FA	4.51	11.59	17.27
T3	59.9	30.3	9.8	FA	5.23	11.59	17.40
T4	60.7	29.7	9.6	FA	4.34	11.59	16.66
Promedio	60.2	30.1	9.7	FA	4.83	11.59	16.90

C.O. = Carbono Orgánico; C.T. = Carbono Total; FA = Franco Arenoso

Tabla A-1: Análisis Granulométrico y Características Básicas del Suelo

A.2 Complejo de Intercambio Catiónico

Los resultados del análisis del complejo de intercambio catiónico revelan niveles adecuados de bases intercambiables con variaciones importantes en la disponibilidad de fósforo entre las diferentes parcelas analizadas.

Muestra	CIC cmol(+)/kg	Ca cmol(+)/kg	Mg cmol(+)/kg	K cmol(+)/kg	Na cmol(+)/kg	B.T. cmol(+)/kg	S.B. (%)	P disp. mg/kg
Control	79.31	5.83	1.31	1.51	0.10	8.75	11.03	7.46
T1	81.33	5.87	1.18	1.12	0.13	8.30	10.21	38.54
T2	79.70	6.71	1.56	1.90	0.17	10.34	12.97	61.89
T3	77.74	7.36	1.46	1.51	0.12	10.45	13.44	8.50
T4	79.94	6.46	1.71	2.01	0.14	10.32	12.91	137.33
Promedio	79.60	6.45	1.44	1.61	0.13	9.63	12.11	50.74

CIC = Capacidad de Intercambio Catiónico; B.T. = Bases Totales; S.B. = Saturación de Bases;
P disp. = Fósforo disponible (Bray II)

Tabla A-2: Complejo de Intercambio Catiónico y Bases Intercambiables

A.3 Acidez Intercambiable y Saturación

El análisis de acidez intercambiable muestra niveles variables entre las muestras, con algunas parcelas presentando condiciones más favorables que otras para el desarrollo del cultivo.

Muestra	A.I. cmol(+)/kg	S.A.I. (%)	Ca/Mg	Mg/K	Ca/K	(Ca+Mg)/K
Control	1.41	13.86	4.45	0.87	3.86	4.73
T1	2.24	21.27	4.98	1.05	5.24	6.30
T2	3.31	24.26	4.30	0.82	3.53	4.35
T3	1.19	10.19	5.04	0.97	4.87	5.84
T4	3.59	25.81	3.78	0.85	3.21	4.07
Promedio	2.35	19.08	4.51	0.91	4.14	5.06

A.I. = Acidez Intercambiable; S.A.I. = Saturación de Acidez Intercambiable

Las relaciones catiónicas indican el balance nutricional del suelo

Tabla A-3: Acidez Intercambiable y Porcentajes de Saturación

A.4 Plan de Fertilización Aplicado

Basándose en los resultados del análisis de suelos, se implementó un plan específico de fertilización edáfica para papa criolla con las siguientes características:

A.4.1 Fertilizantes Aplicados

Fertilizante	1ª Aplicación (kg/ha)	2ª Aplicación (kg/ha)	Total (kg/ha)
Urea (46-0-0)	20	20	163
Superfosfato Triple (0-46-0)	70	20	652
Sulfato de Potasio (0-0-50)	10	10	100
Total fertilizantes	100	50	915

Densidad de siembra: 21.000 plantas/ha (1.20 m × 0.40 m)

1ª aplicación: 15 días después de la siembra en banda lateral

2ª aplicación: después de la desyerba, antes del aporque

Tabla A-4: Plan de Fertilización Edáfica para Papa Criolla

A.4.2 Enmiendas y Correctivos

Enmienda	Aplicación PSI (g/m lineal)	Equivalente (kg/ha)
Humus o compost maduro (mín. 2% N total)	100	800
Cal agrícola no magnesiana (mín. 70% CaCO ₃)	180	1500
Total correctivos y enmiendas	280	2300

PSI = Presiembra incorporado (15–30 días antes de la siembra)

Aplicación al fondo del surco en banda

Tabla A-5: Enmiendas y Correctivos Aplicados

A.5 Interpretación y Recomendaciones

Los suelos analizados presentan las siguientes características principales:

- **Textura:** Franco arenosa, favorable para el drenaje y desarrollo radicular
- **pH:** Ligeramente ácido (4.34–5.23), dentro del rango aceptable para papa
- **Materia orgánica:** Niveles adecuados de carbono orgánico (11.59%)
- **Fertilidad:** CIC moderada a alta (77.74–81.33 cmol(+)/kg)
- **Fósforo:** Niveles muy variables entre parcelas (7.46–137.33 mg/kg)
- **Balance catiónico:** Relaciones Ca/Mg apropiadas, algunas deficiencias de Mg

El plan de fertilización implementado consideró estas características para optimizar la nutrición del cultivo y corregir las deficiencias identificadas en el análisis.

B Comandos Detallados de Ensamblaje

Esta sección presenta los comandos específicos utilizados en el proceso de ensamblaje *de novo* del genoma.

B.1 Preparación de Datos

Conversión de BAM a FASTQ:

```
bedtools bamtobam -i sorted.bam -fq output.fastq
```

B.2 Ensamblaje con Hifiasm

Comando de ensamblaje principal:

```
/nobackup/scratch/grp/fslg_pws_module/software/hifiasm/hifiasm \  
-o hifiasm_v18_m84100_230620_204602_s3 \  
-t 48 \  
m84100_230620_204602_s3.hifi_reads.default.sorted.fastq.gz
```

B.3 Evaluación de Ensamblaje

Conversión de GFA a FASTA:

```
gfatools gfa2fa hifiasm_v18_m84100_230620_204602_s3.bp.p_ctg.gfa > assembly.fasta
```

Cálculo de estadísticas básicas:

```
assembly-stats assembly.fasta > assembly.stats
```

Evaluación BUSCO para múltiples linajes:

```
# Embryophyta
busco --mode genome \
--in hifiasm_v18_m84100_230620_204602_s3.bp.p_ctg.fa \
--lineage embryophyta_odb10 \
--out busco_embryophyta \
--cpu 4
```

```
# Solanales
busco --mode genome \
--in hifiasm_v18_m84100_230620_204602_s3.bp.p_ctg.fa \
--lineage solanales_odb10 \
--out busco_solanales \
--cpu 4
```

```
# Viridiplantae
busco --mode genome \
--in hifiasm_v18_m84100_230620_204602_s3.bp.p_ctg.fa \
--lineage viridiplantae_odb10 \
--out busco_viridiplantae \
--cpu 4
```

```
# Eukaryota
busco --mode genome \
--in hifiasm_v18_m84100_230620_204602_s3.bp.p_ctg.fa \
--lineage eukaryota_odb10 \
--out busco_eukaryota \
--cpu 4
```

Análisis de k-mers para heterocigosidad:

```
# Conteo de 21-mers
jellyfish count -m 21 -s 1000000000 -t 10 -C \
m84100_230620_204602_s3.hifi_reads.default.sorted.fastq.gz \
-o reads_solanum_phureja_col_trim_m21.jf
```

```
# Generación del histograma
jellyfish histo -t 10 reads_solanum_phureja_col_trim_m21.jf > \
reads_solanum_phureja_col_trim_m21.histo
```

```
# Análisis con GenomeScope2 (ejecutado en web)
# URL: http://qb.cshl.edu/genomescope/genomescope2.0/
```

C Comandos Detallados de Anotación

Esta sección presenta los comandos específicos utilizados en el proceso de anotación estructural y funcional del genoma.

C.1 Predicción Estructural con AUGUSTUS

Comando principal de AUGUSTUS:

```
augustus --species=potato \  
--gff3=on \  
--progress=true \  
--strand=both \  
Draft-St_Phureja-Colombia.fa > \  
Draft-St_Phureja-Colombia-chr.gff3
```

C.2 Anotación Funcional con eggNOG-mapper

Comando de eggNOG-mapper:

```
emapper.py -i proteins.clean.faa \  
--itype proteins \  
-m diamond \  
--data_dir /db \  
--cpu 12 \  
-o emapper_creole \  
--output_dir /work/reports \  
--tax_scope 33090 \  
--override
```

C.3 Corrección de Errores

Etapa 1: Evaluación de errores

```
inspector.py -c PacBio_HiFi \  
-r m84100_230620_204602_s3.hifi_reads.default.sorted.fastq \  
-i hifiasm_v18_m84100_230620_204602_s3.bp.p_ctg.fa \  
-o inspector_out_m84100_230620_204602_s3/ \  
--datatype pacbio-hifi
```

Etapa 2: Corrección de errores

```
inspector-correct.py \  
-i inspector_out_m84100_230620_204602_s3/ \  
--datatype pacbio-hifi \  
-o inspector_out_m84100_230620_204602_s3/corrected_asm/
```

C.4 Control de Calidad y Filtración

Fragmentación para análisis distribuido:

```
seqkit split -i -p 40 -O output_chunks \  
hifiasm_v18_m84100_230620_204602_s3.bp.hap_cat_1_2.p_ctg_correct.fa
```

Análisis BLASTn contra bases de datos de referencia:

```
export BLASTDB=/apps/blast/databases  
  
blastn -query assembly_chunk.fa \  
-db nt \  
-outfmt "6 qseqid sseqid pident length mismatch gapopen \  
qstart qend sstart send evalue bitscore staxids" \  
-evaluate 1e-10 \  
-max_target_seqs 10 \  
-num_threads 32 \  
-out blast_results.out
```

Mapeo de lecturas para análisis de cobertura:

```
minimap2 -t 32 -ax map-hifi assembly.fa reads.fastq.gz > mapped.sam  
samtools view -bS mapped.sam | samtools sort > mapped.sorted.bam  
samtools index mapped.sorted.bam
```

Análisis con BlobTools:

```
blobtools create \  
-i assembly.fa \  
-b mapped.sorted.bam \  
-t blast_results.out \  
-o blobtools_dataset
```

```
blobtools plot -i blobtools_dataset.json -o plots/
```

Remoción de contaminantes cloroplásticos:

```
cat blast_results.out | awk 'print $0 "\t" $5/$4*100' | \  
awk '($6>=99 && $7>=99)' | cut -f 1 | sort | uniq > contigs_to_remove.txt
```

```
sed 's/^/>/' contigs_to_remove.txt > contigs_to_remove_fasta.txt
```

```
seqkit grep -v -f contigs_to_remove_fasta.txt assembly.fa | \  
seqkit rmdup -s > assembly_clean.fa
```

D Comandos Detallados de Scaffolding

Esta sección resume los comandos utilizados para el andamiaje cromosómico (scaffolding) con RagTag y la validación de calidad y sintenia asociadas.

D.1 Andamiaje con RagTag

Anclaje de contigs al genoma de referencia DM1-3 516 R44 v6.1:

```
# Mapeo de contigs contra la referencia para guiar RagTag
minimap2 -x asm5 -t 16 DM1-3_v6.1.fa assembly_clean.fa > aln.paf

# Andamiaje principal
ragtag.py scaffold -o ragtag_out \
  -u -r -t 16 \
  DM1-3_v6.1.fa assembly_clean.fa

# Resultados principales
# ragtag_out/ragtag.scaffold.fasta      (secuencias scaffolded)
# ragtag_out/ragtag.scaffold.agp       (coordenadas de ensamblaje)
# ragtag_out/ragtag.scaffold.coords.paf
```

D.2 Evaluación de continuidad del scaffolding

Estadísticas por cromosoma y N50/L50:

```
assembly-stats ragtag_out/ragtag.scaffold.fasta > scaffold.stats
```

Asignación de scaffolds por cromosoma:

```
python utils/scaffold_per_chr.py \
```

```
--agp ragtag_out/ragtag.scaffold.agp \  
--out scaffolds_per_chr.tsv
```

D.3 Validación de sintenia con D-GENIES

Generación de archivo de alineamiento para dot-plot:

```
minimap2 -x asm5 -t 16 DM1-3_v6.1.fa \  
ragtag_out/ragtag.scaffold.fasta > scaffold_vs_ref.paf
```

Conversión a formato requerido y carga en D-GENIES:

```
# Convertir PAF a formato TSV si se requiere por la instancia local  
paftools.js view scaffold_vs_ref.paf > scaffold_vs_ref.tsv
```

```
# Cargar referencia y consulta en D-GENIES (web o local)  
# https://dgenies.toulouse.inra.fr/
```

Cálculo de colinearidad porcentual por cromosoma (opcional):

```
python utils/synteny_metrics.py \  
--paf scaffold_vs_ref.paf \  
--ref DM1-3_v6.1.fa \  
--qry ragtag_out/ragtag.scaffold.fasta \  
--out synteny_by_chr.tsv
```

E Análisis de Sintenia por Cromosoma

Esta sección presenta los análisis detallados de sintenia cromosoma-específicos entre el genoma de referencia DM1-3 516 R44 v6.1 y el ensamblaje del cultivar 'Criolla Colombia'. Cada figura muestra la colinearidad y eventos de reorganización para cada uno de los 12 cromosomas individualmente.

E.1 Cromosoma 1

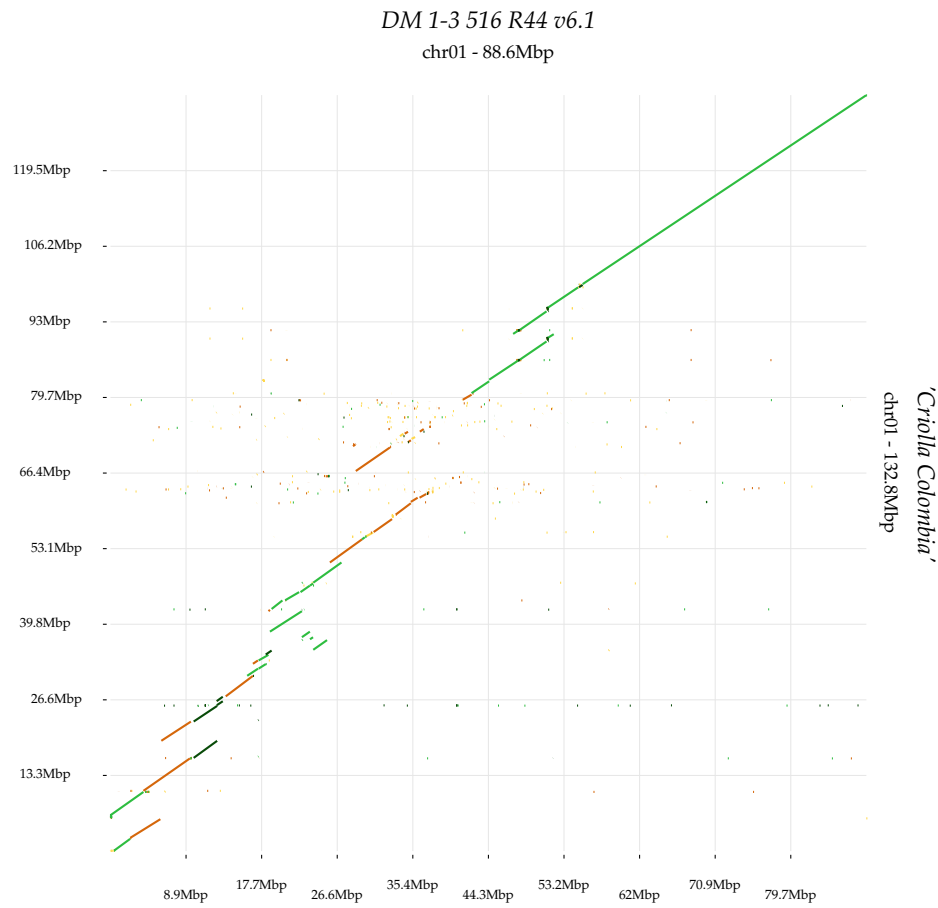


Figura E-1: Análisis de sintenia del chr01 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

E.2 Cromosoma 2

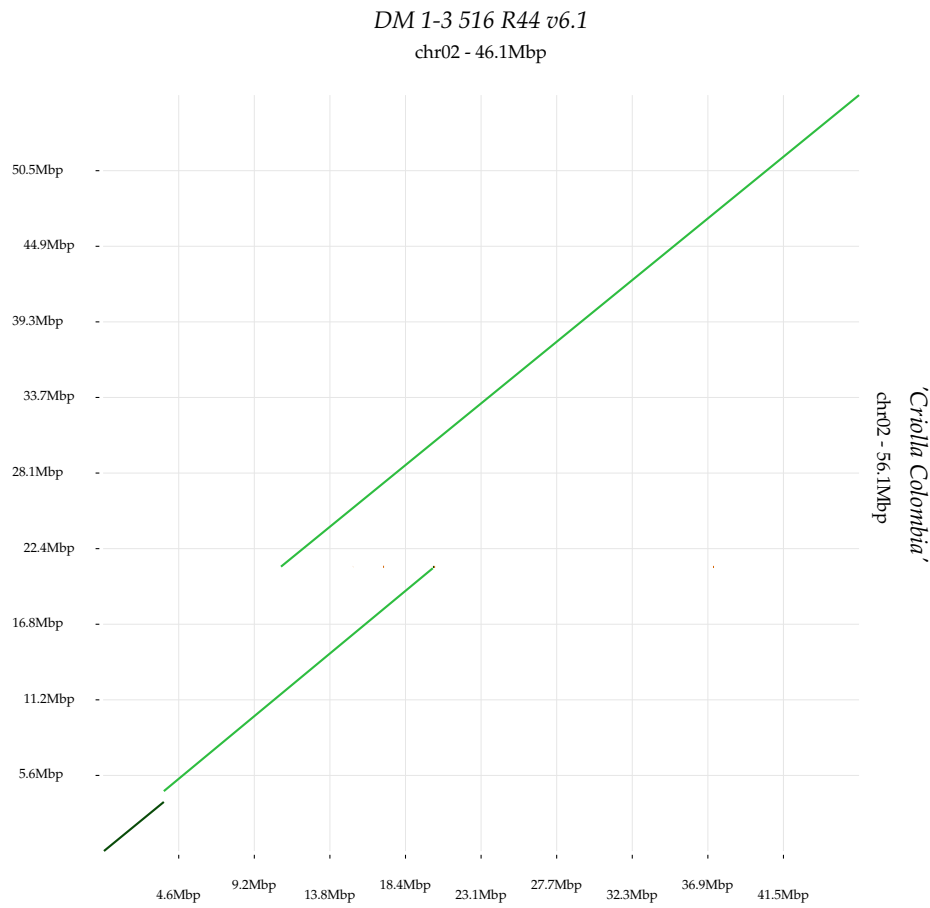


Figura E-2: Análisis de sintenia del chr02 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

E.3 Cromosoma 3

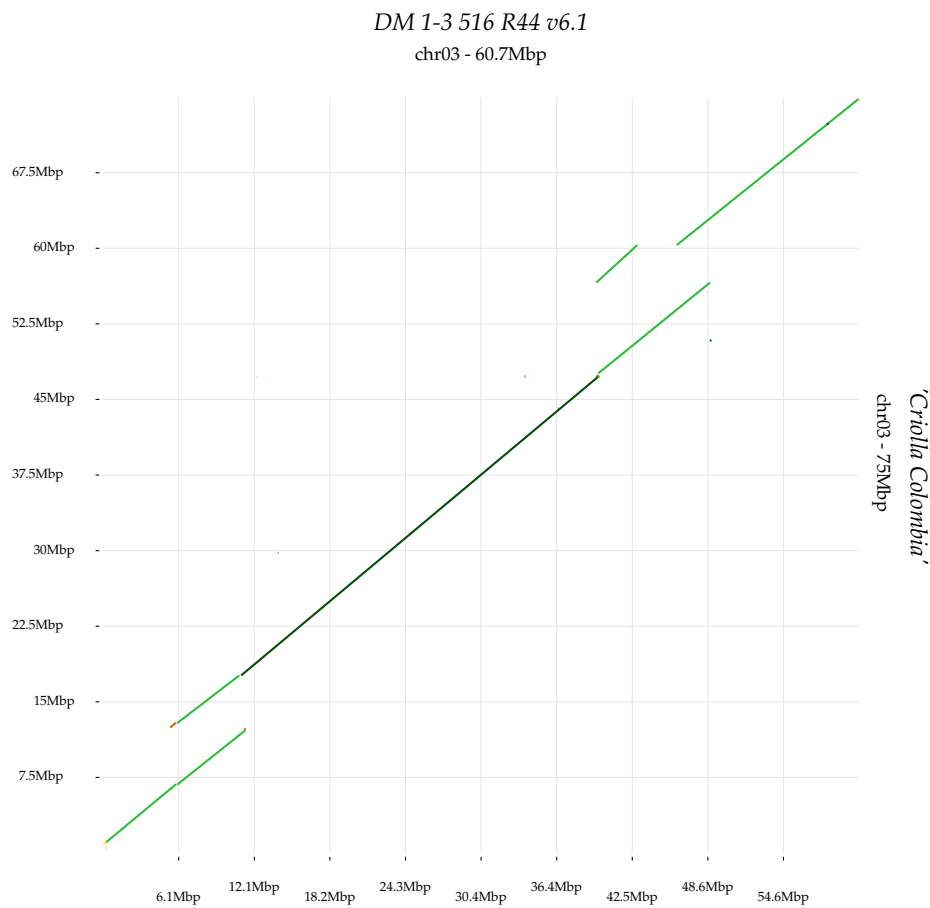


Figura E-3: Análisis de sintenia del chr03 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

E.4 Cromosoma 4

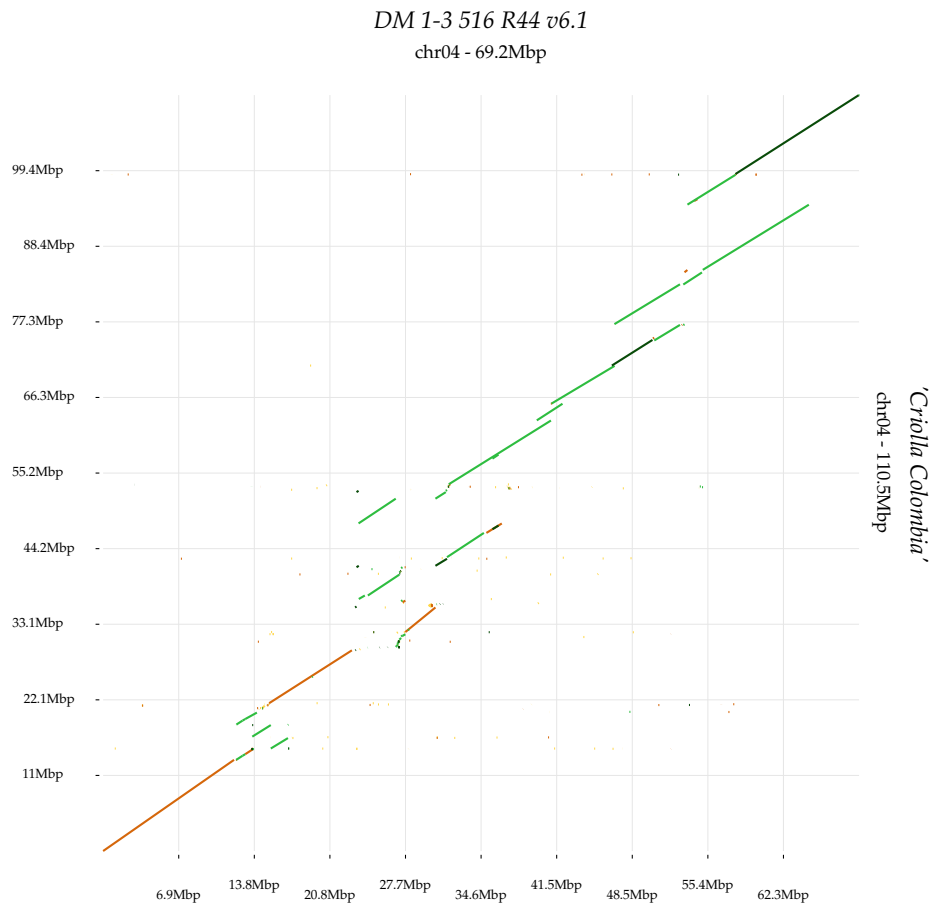


Figura E-4: Análisis de sintenia del chr04 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

E.5 Cromosoma 5

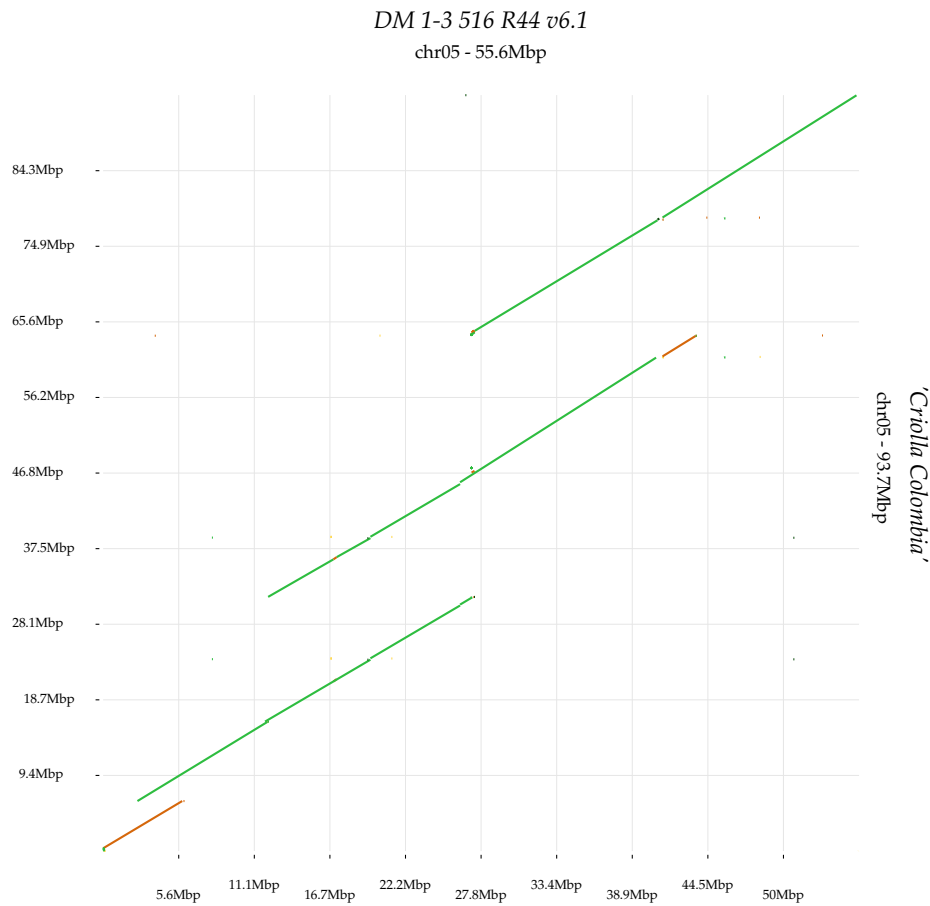


Figura E-5: Análisis de sintenia del chr05 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

E.6 Cromosoma 6

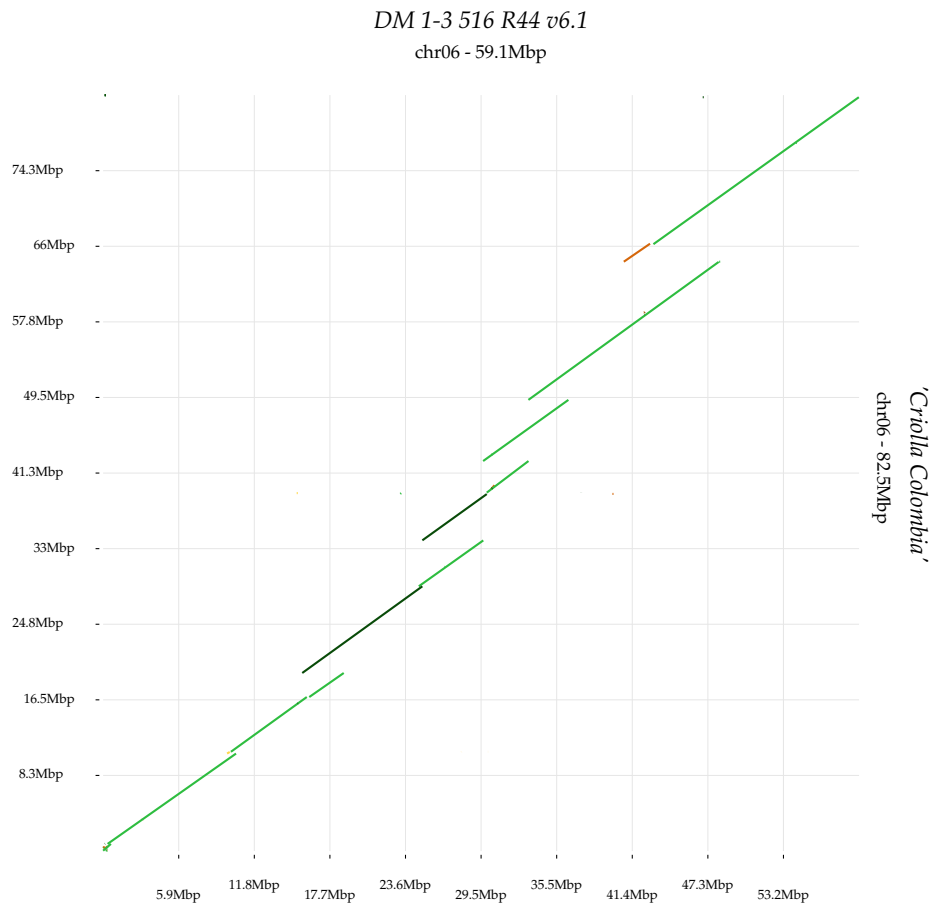


Figura E-6: Análisis de sintenia del chr06 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

E.7 Cromosoma 7

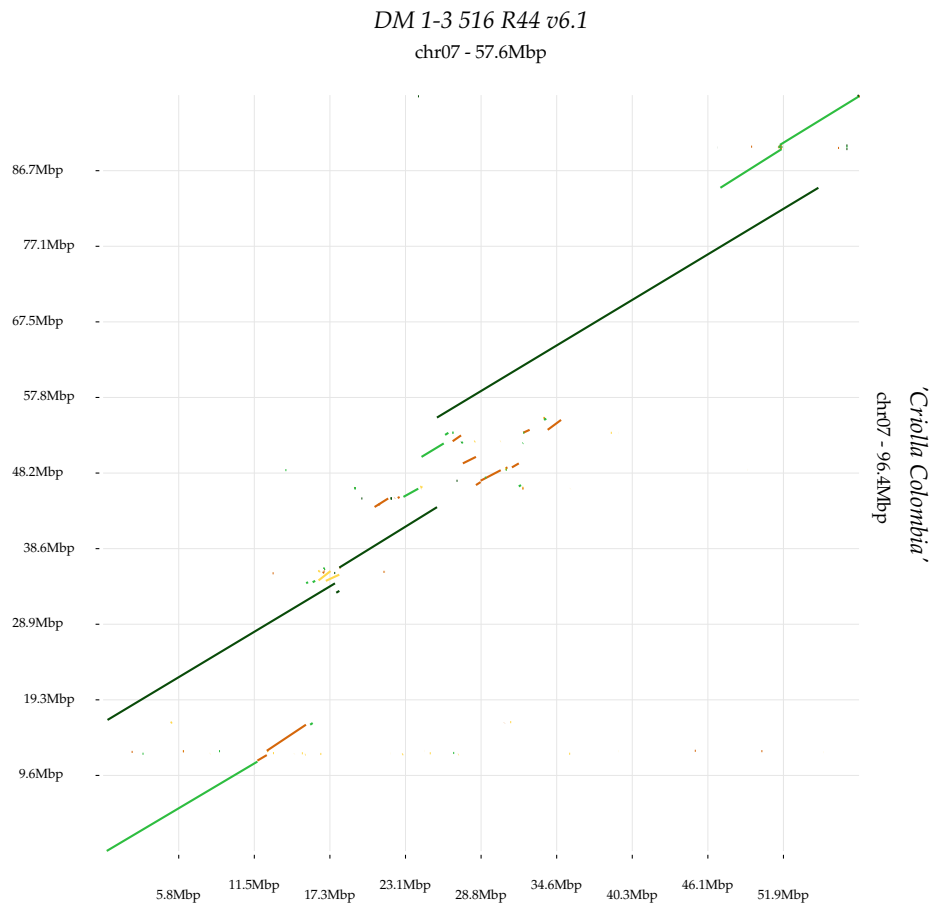


Figura E-7: Análisis de sintenia del chr07 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

E.8 Cromosoma 8

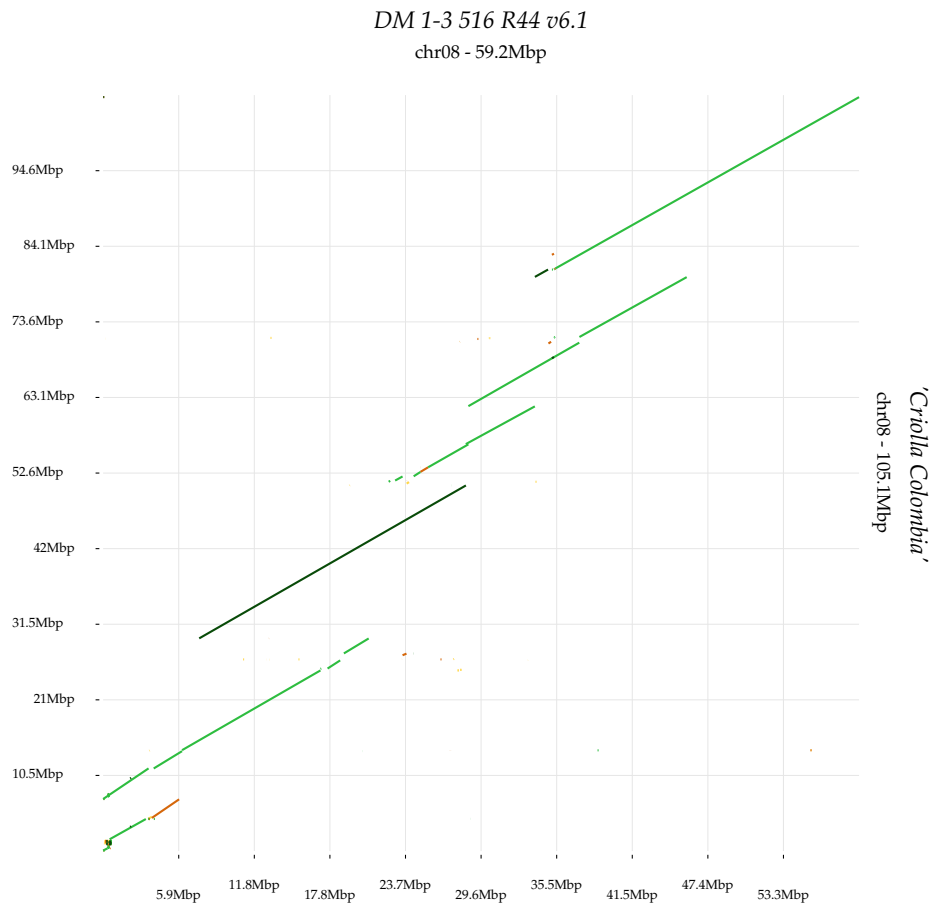


Figura E-8: Análisis de sintenia del chr08 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

E.9 Cromosoma 9

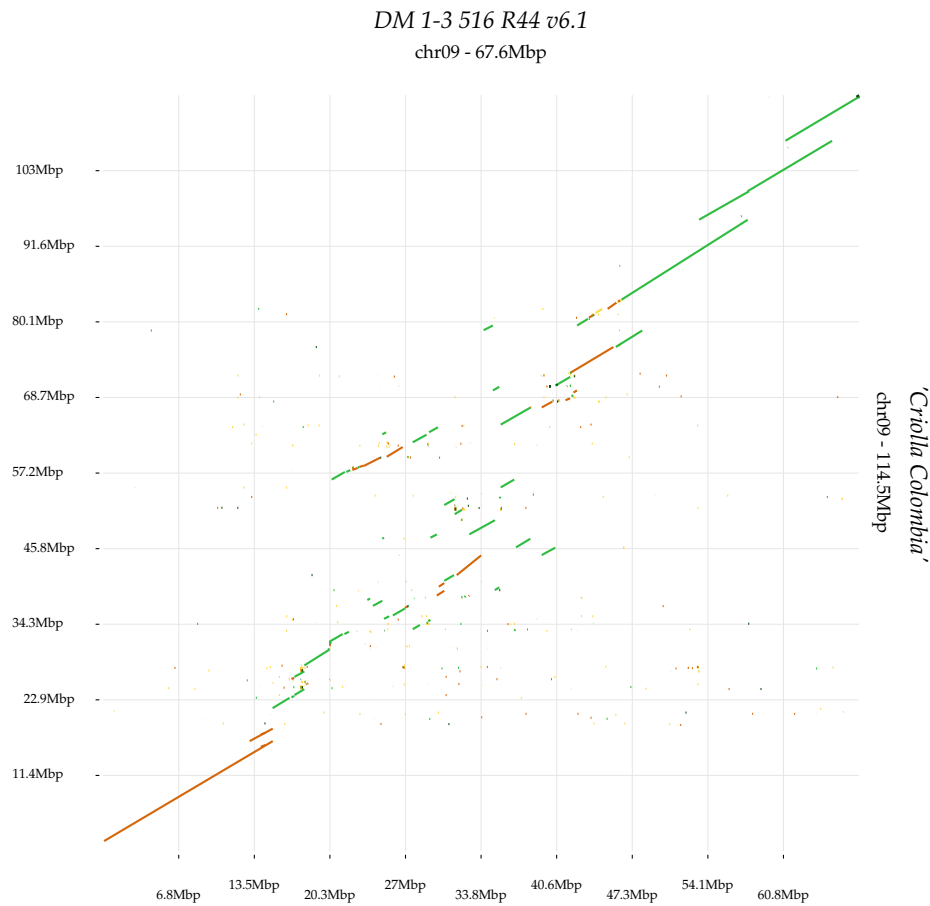


Figura E-9: Análisis de sintenia del chr09 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

E.10 Cromosoma 10

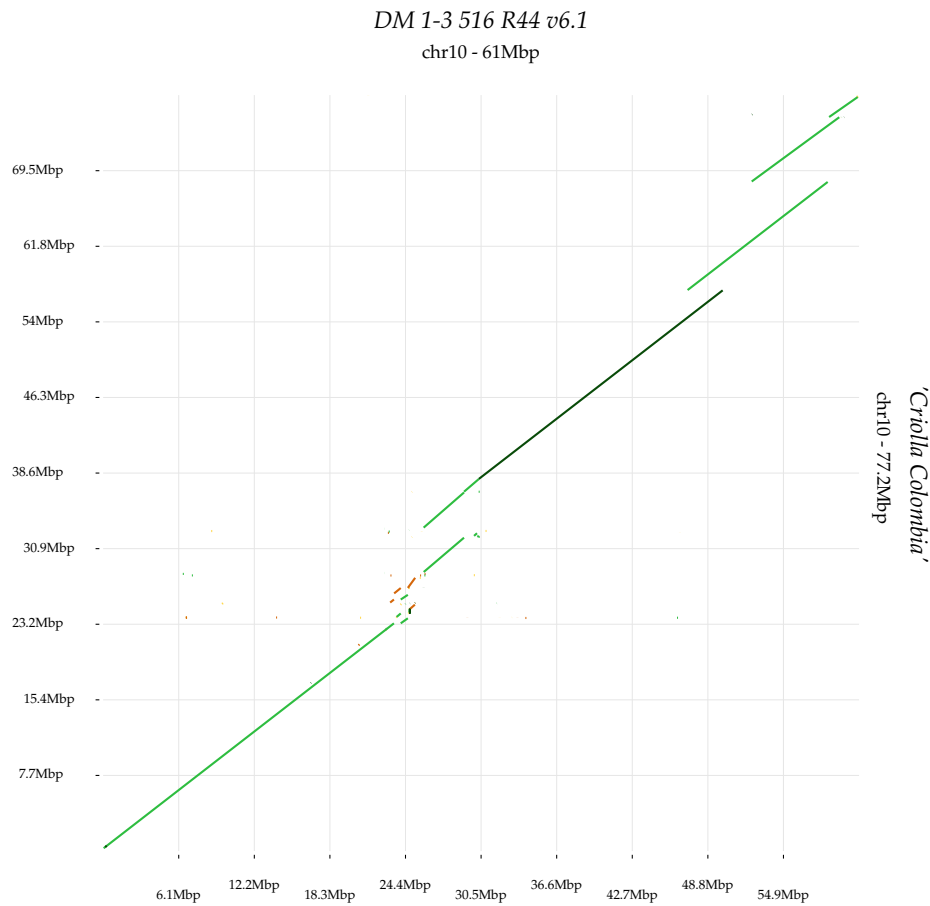


Figura E-10: Análisis de sintenia del chr10 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

E.11 Cromosoma 11

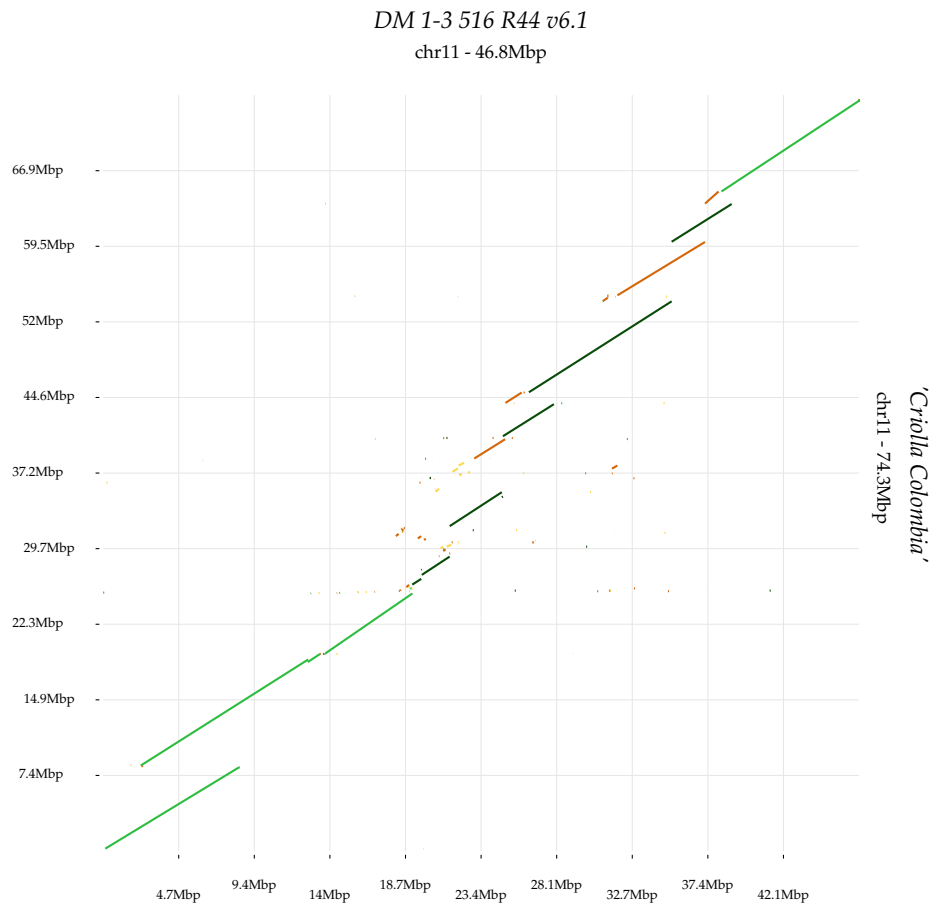


Figura E-11: Análisis de sintenia del chr11 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

E.12 Cromosoma 12

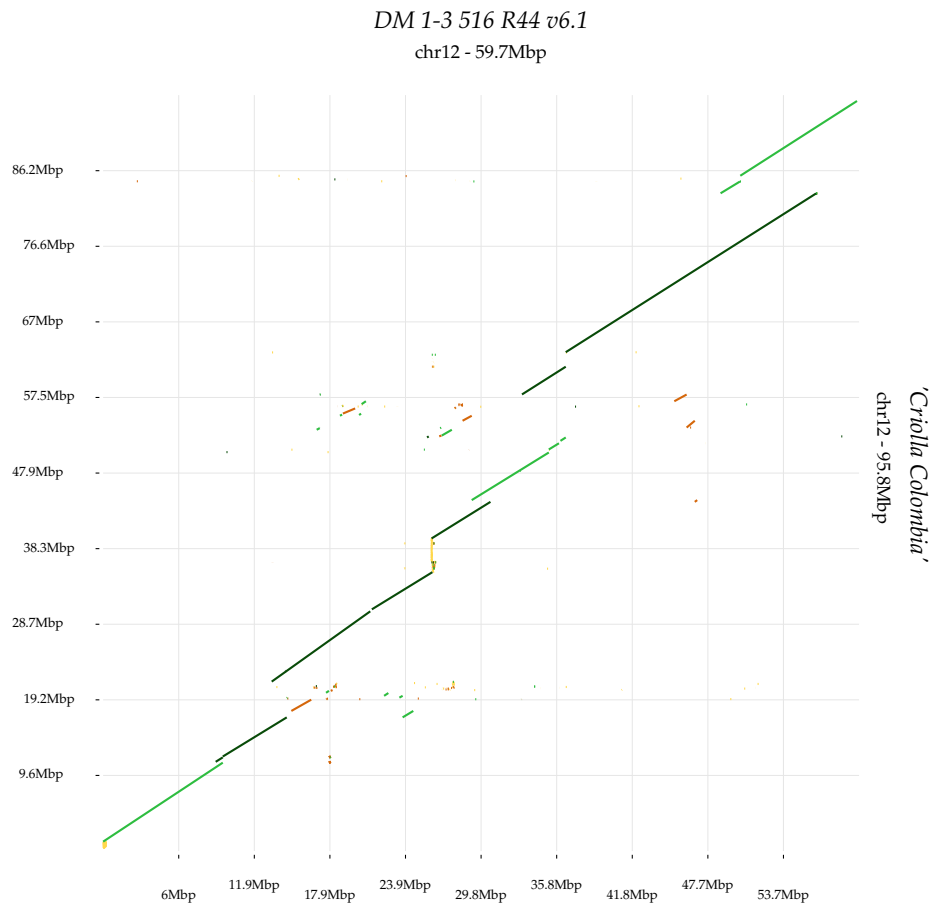


Figura E-12: Análisis de sintenia del chr12 entre DM1-3 516 R44 v6.1 (eje X) y el cultivar 'Criolla Colombia' (eje Y). La intensidad de color indica el porcentaje de identidad: ■ 1.0–0.75, ■ 0.75–0.5, ■ 0.5–0.25, ■ 0.25–0.

F Recursos Computacionales y Hardware

F.1 Clúster de alto rendimiento BYU

Para el ensamblaje *de novo* del genoma de *Solanum tuberosum* L. Grupo Phureja 'Criolla Colombia' se utilizaron los recursos del Fulton Supercomputing Lab (BYU, Utah, EE. UU.), con acceso remoto vía SSH. La infraestructura dispone de 34,948 núcleos de CPU en 614 nodos, más de 176 TB de memoria RAM, 351 GPUs, Red Hat Enterprise Linux 9.4, planificador SLURM, red InfiniBand de alta velocidad y un sistema de archivos distribuido superior a 6 PB. Los trabajos de ensamblaje se ejecutaron típicamente con 21 núcleos por tarea y 13.3 GB de memoria por núcleo, con tiempos de 1–168 h (promedio 75.6 h); las ejecuciones más intensivas (Hifiasm) alcanzaron 36 núcleos y 756 GB de RAM.

F.2 Servidor del Instituto de Genética UNAL

Los resultados del ensamblaje se resguardaron en el servidor de almacenamiento del grupo BIOMOLc (Universidad Distrital), y el modelado metabólico y análisis posteriores se realizaron en un servidor dedicado del Instituto de Genética de la Universidad Nacional de Colombia (laboratorio GiBBS, Prof. Andrés Pinzón), con CPU AMD de 72 núcleos, 256 GB de RAM, Ubuntu Server 18.04.6 LTS y conectividad Ethernet 1 Gbps; el aislamiento de dependencias se gestionó con pyenv-virtualenv.

F.3 Secuenciador PacBio

La secuenciación se realizó en BYU usando PacBio Sequel IIe, con química Sequel II Kit 2.0 (PN 101-820-200), SMRT Cell 8M (PN 101-389-001), SMRT Link v11.0.0.146574 y 30 h de adquisición en modo HiFi-CCS; los datos RAW (BAM) se almacenaron inicialmente en el FSL para procesamiento.

F.4 Equipos de laboratorio

Para la preparación de muestras y control de calidad se emplearon, entre otros, un NanoDrop 2000c (Thermo Fisher) para cuantificación/pureza (A_{260}/A_{280} , A_{260}/A_{230}), Covaris g-TUBE (PN 520079; 10,000 rpm, 1 min) para fragmentación a 8–12 kb, Agilent Bioanalyzer 2100 (DNA 12000, PN 5067-1508) para verificación de librerías, un termociclador Bio-Rad T100 para QC por PCR y una centrifuga Eppendorf 5424R (FA-45-24-11; 21,130×g).

G Condiciones Ambientales de Laboratorio

Durante la preparación experimental se mantuvieron 20–22 °C (± 2 °C) y 45–65% de humedad relativa, con trabajo estéril en cabina de flujo laminar. El ADN se conservó a –20 °C sin ciclos de descongelación, las librerías a –80 °C y los reactivos sensibles a 2–8 °C. Se dispusieron áreas separadas para pre y post-PCR, pipetas dedicadas con puntas con filtro y controles negativos por sesión.

Referencias Bibliográficas

- Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., & Nielsen, J. (2013). The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for *Penicillium chrysogenum* (C. D. Maranas, Ed.). *PLoS Computational Biology*, *9*(3), e1002980. <https://doi.org/10.1371/journal.pcbi.1002980>
- Ai, Y., Jing, S., Cheng, Z., Song, B., Xie, C., Liu, J., & Zhou, J. (2021). DNA methylation affects photoperiodic tuberization in potato (*Solanum tuberosum* L.) by mediating the expression of genes related to the photoperiod and GA pathways. *Horticulture Research*, *8*(1). <https://doi.org/10.1038/s41438-021-00619-7>
- Akiyama, R., Watanabe, B., Nakayasu, M., Lee, H. J., Kato, J., Umemoto, N., Muranaka, T., Saito, K., Sugimoto, Y., & Mizutani, M. (2021). The biosynthetic pathway of potato solanidanes diverged from that of spirosolanes due to evolution of a dioxygenase. *Nature Communications*, *12*(1). <https://doi.org/10.1038/s41467-021-21546-0>
- Alon, U. (2007). Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, *8*(6), 450–461. <https://doi.org/10.1038/nrg2102>
- Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., & Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology*, *23*(1), 258. <https://doi.org/10.1186/s13059-022-02823-7>
- Andrade-Díaz, D. (2024). Morpho-Physiological Characterization of Potato (*Solanum Tuberosum*) Genotypes of the Andigenum and Phureja Group from the Working Collection of the Universidad De Nariño. *International Journal of Life Science and Agriculture Research*, *03*(05). <https://doi.org/10.55677/ijlsar/v03i5y2024-09>
- Aziz, A., Randhawa, M. A., Butt, M. S., Asghar, A., Yasin, M., & Shibamoto, T. (2012). Glycoalkaloids (α -Chaconine and α -Solanine) Contents of Selected Pakistani Potato Cultivars and Their Dietary Intake Assessment. *Journal of Food Science*, *77*(3). <https://doi.org/10.1111/j.1750-3841.2011.02582.x>
- Baghalian, K., Hajirezaei, M.-R., & Schreiber, F. (2014). Plant Metabolic Modeling: Achieving New Insight into Metabolism and Metabolic Engineering. *The Plant Cell*, *26*(10), 3847–3866. <https://doi.org/10.1105/tpc.114.130328>
- Bansal, P., Morgat, A., Axelsen, K. B., Muthukrishnan, V., Coudert, E., Aimo, L., Hyka-Nouspikel, N., Gasteiger, E., Kerhornou, A., Neto, T. B., Pozzato, M., Blatter, M.-C., Ignatchenko, A., Redaschi, N., & Bridge, A. (2022). Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Research*, *50*(D1), D693–D700. <https://doi.org/10.1093/nar/gkab1016>
- Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–113. <https://doi.org/10.1038/nrg1272>
- Berdugo-Cely, J., Valbuena, R. I., Sánchez-Betancourt, E., Barrero, L. S., & Yockteng, R. (2017). Genetic diversity and association mapping in the Colombian Central Collection of *Solanum tuberosum* L. Andigenum group using SNPs markers (X.-Q. Li, Ed.). *PLOS ONE*, *12*(3), e0173039. <https://doi.org/10.1371/journal.pone.0173039>

- Berdugo-Cely, J. A., Martínez-Moncayo, C., & Lagos-Burbano, T. C. (2021). Genetic analysis of a potato (*Solanum tuberosum* L.) breeding collection for southern Colombia using Single Nucleotide Polymorphism (SNP) markers (T.-Y. Chiang, Ed.). *PLOS ONE*, *16*(3), e0248787. <https://doi.org/10.1371/journal.pone.0248787>
- Bhalla, U. S., & Iyengar, R. (1999). Emergent properties of networks of biological signaling pathways. *Science*, *283*(5400), 381–387. <https://doi.org/10.1126/science.283.5400.381>
- Bhaskar, P. B., Wu, L., Busse, J. S., Whitty, B. R., Hamernik, A. J., Jansky, S. H., Buell, C. R., Bethke, P. C., & Jiang, J. (2010). Suppression of the Vacuolar Invertase Gene Prevents Cold-Induced Sweetening in Potato. *Plant Physiology*, *154*(2), 939–948. <https://doi.org/10.1104/pp.110.162545>
- Blanchard, J. L. (2004). Bioinformatics and Systems Biology, rapidly evolving tools for interpreting plant response to global change. *Field Crops Research*, *90*(1), 117–131. <https://doi.org/10.1016/j.fcr.2004.07.015>
- Bohórquez-Quintero, M. d. I. A., Galvis-Tarazona, D. Y., Arias-Moreno, D. M., Ojeda-Peréz, Z. Z., Ochatt, S., & Rodríguez-Molano, L. E. (2022). Morphological and anatomical characterization of yellow diploid potato flower for effective breeding program. *Scientific Reports*, *12*(1). <https://doi.org/10.1038/s41598-022-20439-6>
- Bornstein, B. J., Keating, S. M., Jouraku, A., & Hucka, M. (2008). LibSBML: An API Library for SBML. *Bioinformatics*, *24*(6), 880–881. <https://doi.org/10.1093/bioinformatics/btn051>
- Botero, K., Restrepo, S., & Pinzón, A. (2018). A genome-scale metabolic model of potato late blight suggests a photosynthesis suppression mechanism. *BMC Genomics*, *19*(S8). <https://doi.org/10.1186/s12864-018-5192-x>
- Bozan, I., Achakkagari, S. R., Anglin, N. L., Ellis, D., Tai, H. H., & Strömviik, M. V. (2023). Pangenome analyses reveal impact of transposable elements and ploidy on the evolution of potato species. *Proceedings of the National Academy of Sciences*, *120*(31). <https://doi.org/10.1073/pnas.2211117120>
- Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366–368. <https://doi.org/10.1038/s41592-021-01101-x>
- Cabanettes, F., & Klopp, C. (2018). D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *6*, e4958. <https://doi.org/10.7717/peerj.4958>
- Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., & al., e. (2021). eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. <https://doi.org/10.1093/molbev/msab293>
- Caspi, R., Billington, R., Fulcher, C. A., & al., e. (2023). The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Research*, *51*(D1), D618–D627. <https://doi.org/10.1093/nar/gkac1076>
- Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit—interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics*, *10*(4), 1361–1374. <https://doi.org/10.1534/g3.119.400908>
- Chen, L., Zhou, F., Chen, Y., Fan, Y., Zhang, K., Liu, Q., Tu, W., Jiang, F., Li, G., Zhao, H., & Song, B. (2022). Salicylic Acid Improves the Constitutive Freezing Tolerance of Potato as Revealed by Transcriptomics and Metabolomics Analyses. *International Journal of Molecular Sciences*, *24*(1), 609. <https://doi.org/10.3390/ijms24010609>
- Chen, Y., Gustafsson, J., Tafur Rangel, A., Anton, M., Domenzain, I., Kittikunapong, C., Li, F., Yuan, L., Nielsen, J., & Kerkhoven, E. J. (2024). Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO Toolbox 3.0. *Nature Protocols*, *19*(3), 629–667. <https://doi.org/10.1038/s41596-023-00931-7>
- Chen, Y., Zhang, Y., Wang, A. Y., Gao, M., & Chong, Z. (2021). Accurate long-read de novo assembly evaluation with Inspector. *Genome Biology*, *22*(1), 312. <https://doi.org/10.1186/s13059-021-02527-4>

- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2), 170–175. <https://doi.org/10.1038/s41592-020-01056-5>
- Cheung, C. Y. M., Poolman, M. G., Ratcliffe, R. G., & Sweetlove, L. J. (2016). Flux balance analysis of plant metabolism: The effect of biomass composition and model structure on flux prediction. *Frontiers in Plant Science*, 7, 537. <https://doi.org/10.3389/fpls.2016.00537>
- Courtot, M., Juty, N., Knüpfer, C., Waltemath, D., Zhukova, A., Dräger, A., Dumontier, M., Finney, A., Golebiewski, M., Hastings, J., Hoops, S., Keating, S., Kell, D. B., Kerrien, S., Lawson, J., Lister, A., Lu, J., Machne, R., Mendes, P., ... Le Novère, N. (2011). Controlled vocabularies and semantics in systems biology. *Molecular Systems Biology*, 7(1). <https://doi.org/10.1038/msb.2011.77>
- Covarrubias-Pazarán, G., Martini, J. W. R., Quinn, M., & Atlin, G. (2021). Strengthening Public Breeding Pipelines by Emphasizing Quantitative Genetics Principles and Open Source Data Management. *Frontiers in Plant Science*, 12. <https://doi.org/10.3389/fpls.2021.681624>
- Crookshanks, M., Emmersen, J., Welinder, K. G., & Lehmann Nielsen, K. (2001). The potato tuber transcriptome: Analysis of 6077 expressed sequence tags. *FEBS Letters*, 506(2), 123–126. [https://doi.org/10.1016/S0014-5793\(01\)02888-5](https://doi.org/10.1016/S0014-5793(01)02888-5)
- Cunha, E., Silva, M., Chaves, I., Demirci, H., Lagoa, D. R., Lima, D., Rocha, M., Rocha, I., & Dias, O. (2023). The first multi-tissue genome-scale metabolic model of a woody plant highlights suberin biosynthesis pathways in *Quercus suber* (C. Kaleta, Ed.). *PLOS Computational Biology*, 19(9), e1011499. <https://doi.org/10.1371/journal.pcbi.1011499>
- Dan, Z., Chen, Y., Zhao, W., Wang, Q., & Huang, W. (2020). Metabolome-based prediction of yield heterosis contributes to the breeding of elite rice. *Life Science Alliance*, 3(1), e201900551. <https://doi.org/10.26508/lsa.201900551>
- De Oliveira Dal'Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., & Nielsen, L. K. (2010). AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in *Arabidopsis*. *Plant Physiology*, 152(2), 579–589. <https://doi.org/10.1104/pp.109.148817>
- Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M., & Ashburner, M. (2007). ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36, D344–D350. <https://doi.org/10.1093/nar/gkm791>
- Degtyarenko, K., Hastings, J., De Matos, P., & Ennis, M. (2009). ChEBI: An Open Bioinformatics and Cheminformatics Resource. *Current Protocols in Bioinformatics*, 26(1). <https://doi.org/10.1002/0471250953.bi1409s26>
- Diaz-Valencia, P., Melgarejo, L. M., Arcila, I., & Mosquera-Vásquez, T. (2021). Physiological, Biochemical and Yield-Component Responses of *Solanum tuberosum* L. Group Phureja Genotypes to a Water Deficit. *Plants*, 10(4), 638. <https://doi.org/10.3390/plants10040638>
- Docimo, T., Scotti, N., Tamburino, R., Villano, C., Carputo, D., & D'Amelia, V. (2023). Potato nutraceuticals: Genomics and biotechnology for bio-fortification. In C. Kole (Ed.), *Compendium of crop genome designing for nutraceuticals* (pp. 1183–1215). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-4169-6_48
- Dogramaci, M., Dobry, E. P., Fortini, E. A., Sarkar, D., Eshel, D., & Campbell, M. A. (2024). Physiological and molecular mechanisms associated with potato tuber dormancy (M. Considine, Ed.). *Journal of Experimental Botany*, 75(19), 6093–6109. <https://doi.org/10.1093/jxb/erae182>
- Dolničar, P., & Bohanec, B. (2000). Ploidy and morphological characteristics of *Solanum tuberosum* × *Solanum phureja* hybrids. *Pflügers Archiv - European Journal of Physiology*, 439(7), R9–R11. <https://doi.org/10.1007/bf03376504>
- Domenzain, I., Sánchez, B., Anton, M., Kerkhoven, E. J., Millán-Oropeza, A., Henry, C., Siewers, V., Morrissey, J. P., Sonnenschein, N., & Nielsen, J. (2022). Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. *Nature Communications*, 13(1). <https://doi.org/10.1038/s41467-022-31421-1>

- Ebrahim, A., Lerman, J. A., Palsson, B. O., & Hyduke, D. R. (2013). COBRApy: CONstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology*, 7(1), 74. <https://doi.org/10.1186/1752-0509-7-74>
- Engel, N., Van Den Daele, K., Kolukisaoglu, Ü., Morgenthal, K., Weckwerth, W., Pärnik, T., Keerberg, O., & Bauwe, H. (2007). Deletion of Glycine Decarboxylase in Arabidopsis Is Lethal under Nonphotosynthetic Conditions. *Plant Physiology*, 144(3), 1328–1335. <https://doi.org/10.1104/pp.107.099317>
- Espinoza Corona, M. I. (2024). *Desarrollo de un modelo metabólico a escala genómica para Macrocyctis: proceso, herramientas de generación automática y potencial de aplicación*. Retrieved July 16, 2025, from <https://repositorio.uchile.cl/handle/2250/204042>
Accepted: 2025-04-03T14:02:29Z.
- et al. Hardigan, M. A. (2017). Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proceedings of the National Academy of Sciences*, 114(46), E9999–E10008. <https://doi.org/10.1073/pnas.1714380114>
- Fang, G., Yang, S., Ruan, B., Ye, G., He, M., Su, W., Zhou, Y., Wang, J., & Yang, S. (2024). Research Progress on Physiological, Biochemical, and Molecular Mechanisms of Potato in Response to Drought and High Temperature. *Horticulturae*, 10(8), 827. <https://doi.org/10.3390/horticulturae10080827>
- FAO. (2024, May 30). *International Day of Potato: At inaugural celebration, FAO highlights crop's significance and further potential*. Newsroom. Retrieved June 16, 2025, from <https://www.fao.org/newsroom/detail/international-day-of-potato--at-inaugural-celebration--fao-highlights-crop-significance-and-further-potential/en>
- Foerster, H., Bombarely, A., Battey, J. N. D., Sierro, N., Ivanov, N. V., & Mueller, L. A. (2018). SolCyc: A database hub at the Sol Genomics Network (SGN) for the manual curation of metabolic networks in *Solanum* and *Nicotiana* specific databases. *Database*, 2018. <https://doi.org/10.1093/database/bay035>
- Gao, H.-J., Yang, H.-Y., Bai, J.-P., Liang, X.-Y., Lou, Y., Zhang, J.-L., Wang, D., Zhang, J.-L., Niu, S.-Q., & Chen, Y.-L. (2015). Ultrastructural and physiological responses of potato (*Solanum tuberosum* L.) plantlets to gradient saline stress. *Frontiers in Plant Science*, 5. <https://doi.org/10.3389/fpls.2014.00787>
- Gao, Y., & Zhao, C. (2024). Development and applications of metabolic models in plant multi-omics research. *Frontiers in Plant Science*, 15, 1361183. <https://doi.org/10.3389/fpls.2024.1361183>
- Geigenberger, P. (2003). Regulation of sucrose to starch conversion in growing potato tubers. *Journal of Experimental Botany*, 54(382), 457–465. <https://doi.org/10.1093/jxb/erg074>
- Ghislain, M., Andrade, D., Rodríguez, F., Hijmans, R. J., & Spooner, D. M. (2006). Genetic analysis of the cultivated potato *Solanum tuberosum* L. Phureja Group using RAPDs and nuclear SSRs. *Theoretical and Applied Genetics*, 113(8), 1515–1527. <https://doi.org/10.1007/s00122-006-0399-7>
- Gottwald, J. R., Krysan, P. J., Young, J. C., Evert, R. F., & Sussman, M. R. (2000). Genetic evidence for the *in planta* role of phloem-specific plasma membrane sucrose transporters. *Proceedings of the National Academy of Sciences*, 97(25), 13979–13984. <https://doi.org/10.1073/pnas.250473797>
- Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biology*, 20(1). <https://doi.org/10.1186/s13059-019-1730-3>
- Haggart, C. R., Bartell, J. A., Saucerman, J. J., & Papin, J. A. (2011). Whole-Genome Metabolic Network Reconstruction and Constraint-Based Modeling*. In *Methods in Enzymology* (pp. 411–433, Vol. 500). Elsevier. <https://doi.org/10.1016/B978-0-12-385118-5.00021-9>
- Hardigan, M. A., Crisovan, E., Hamilton, J. P., Kim, J., Laimbeer, P., Leisner, C. P., Manrique-Carpintero, N. C., Newton, L., Pham, G. M., Vaillancourt, B., Yang, X., Zeng, Z., Douches, D. S., Jiang, J., Veilleux, R. E., & Buell, C. R. (2016). Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated *Solanum tuberosum*. *The Plant Cell*, 28(2), 388–405. <https://doi.org/10.1105/tpc.15.00538>
- Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(S6761), C47–C52. <https://doi.org/10.1038/35011540>

- Haynes, K. G., Zotarelli, L., Christensen, C. T., & Walker, S. (2019). Early Generation Selection within a Diploid Hybrid *Solanum tuberosum* Groups Phureja and Stenotomum Population for the Intense Yellow-flesh Creamer Potato Market. *HortScience*, *54*(12), 2118–2124. <https://doi.org/10.21273/hortsci13576-18>
- Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., & Pletnev, I. (2013). InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, *5*(1). <https://doi.org/10.1186/1758-2946-5-7>
- Heller, S. R., McNaught, A., Pletnev, I., Stein, S., & Tchekhovskoi, D. (2015). InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, *7*(1). <https://doi.org/10.1186/s13321-015-0068-4>
- Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., & Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, *28*(9), 977–982. <https://doi.org/10.1038/nbt.1672>
- Huamán, Z., & Spooner, D. M. (2002). Reclassification of landrace populations of cultivated potatoes (*Solanum* sect. *Petota*). *American Journal of Botany*, *89*(6), 947–965. <https://doi.org/10.3732/ajb.89.6.947>
- Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H., ... Wang, J. (2003). The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, *19*(4), 524–531. <https://doi.org/10.1093/bioinformatics/btg015>
- Jarin, A., Ghosh, U. K., Hossain, M. S., Mahmud, A., & Khan, M. A. R. (2024). Glycine betaine in plant responses and tolerance to abiotic stresses. *Discover Agriculture*, *2*(1). <https://doi.org/10.1007/s44279-024-00152-w>
- Jayarathna, S., Péter-Szabó, Z., Nestor, G., Andersson, M., Vilaplana, F., & Andersson, R. (2024). Impact of mutations in starch synthesis genes on morphological, compositional, molecular structure, and functional properties of potato starch (S. K. Verma, Ed.). *PLOS ONE*, *19*(9), e0310990. <https://doi.org/10.1371/journal.pone.0310990>
- Jin, X., Ballicora, M. A., Preiss, J., & Geiger, J. H. (2005). Crystal structure of potato tuber ADP-glucose pyrophosphorylase. *The EMBO Journal*, *24*(4), 694–704. <https://doi.org/10.1038/sj.emboj.7600551>
- Jing, Q., Hou, H., Meng, X., Chen, A., Wang, L., Zhu, H., Zheng, S., Lv, Z., & Zhu, X. (2022). Transcriptome analysis reveals the proline metabolic pathway and its potential regulation TF-hub genes in salt-stressed potato. *Frontiers in Plant Science*, *13*. <https://doi.org/10.3389/fpls.2022.1030138>
- Juyó, D., Sarmiento, F., Álvarez, M., Brochero, H., Gebhardt, C., & Mosquera, T. (2015). Genetic Diversity and Population Structure in Diploid Potatoes of *Solanum tuberosum* Group Phureja. *Crop Science*, *55*(2), 760–769. <https://doi.org/10.2135/cropsci2014.07.0524>
- Keating, S. M., Waltemath, D., König, M., Zhang, F., Dräger, A., Chaouiya, C., Bergmann, F. T., Finney, A., Gillespie, C. S., Helikar, T., Hoops, S., Malik-Sheriff, R. S., Moodie, S. L., Moraru, I. I., Myers, C. J., Naldi, A., Olivier, B. G., Sahle, S., Schaff, J. C., ... Zucker, J. (2020). SBMLLevel 3: An extensible format for the exchange and reuse of biological models. *Molecular Systems Biology*, *16*(8). <https://doi.org/10.15252/msb.20199110>
- Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O., & Grau, J. (2016). GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods in Molecular Biology*, *1415*, 161–177. https://doi.org/10.1007/978-1-4939-3578-9_9
- King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., & Palsson, B. O. (2015). Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways (P. P. Gardner, Ed.). *PLOS Computational Biology*, *11*(8), e1004321. <https://doi.org/10.1371/journal.pcbi.1004321>

- King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., & Lewis, N. E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, 44(D1), D515–D522. <https://doi.org/10.1093/nar/gkv1049>
- Kitano, H. (2002). Systems biology: A brief overview. *Science*, 295(5560), 1662–1664. <https://doi.org/10.1126/science.1069492>
- Kitano, H. (2007). Towards a theory of biological robustness. *Molecular Systems Biology*, 3(1), 137. <https://doi.org/10.1038/msb4100179>
- Le Novère, N. (2006). BioModels Database: A free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research*, 34(90001), D689–D691. <https://doi.org/10.1093/nar/gkj092>
- Li, C., Donizelli, M., Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M. I., Snoep, J. L., Hucka, M., Le Novère, N., & Laibe, C. (2010). BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 4(1). <https://doi.org/10.1186/1752-0509-4-92>
- Li, H., Luo, W., Ji, R., Xu, Y., Xu, G., Qiu, S., & Tang, H. (2021). A comparative proteomic study of cold responses in potato leaves. *Heliyon*, 7(2), e06002. <https://doi.org/10.1016/j.heliyon.2021.e06002>
- Lieven, C., Beber, M. E., Olivier, B. G., Bergmann, F. T., Ataman, M., Babaei, P., Bartell, J. A., Blank, L. M., Chauhan, S., Correia, K., Diener, C., Dräger, A., Ebert, B. E., Edirisinghe, J. N., Faria, J. P., Feist, A. M., Fengos, G., Fleming, R. M. T., García-Jiménez, B., ... Zhang, C. (2020). MEMOTE for standardized genome-scale metabolic model testing. *Nature Biotechnology*, 38(3), 272–276. <https://doi.org/10.1038/s41587-020-0446-y>
- Lin, S., Singh, R. K., Moehninsi, & Navarre, D. A. (2021). R2R3-MYB transcription factors, StmiR858 and sucrose mediate potato flavonol biosynthesis. *Horticulture Research*, 8(1). <https://doi.org/10.1038/s41438-021-00463-9>
- Linnaeus, C. (1753). *Solanum tuberosum*. In *Species plantarum* (p. 185, Vol. 1). Laurentius Salvius. <https://www.ipni.org/n/821337-1>
- Liu, T., Kawochar, M. A., Begum, S., Wang, E., Zhou, T., Jing, S., Liu, T., Yu, L., Nie, B., & Song, B. (2023). Potato tonoplast sugar transporter 1 controls tuber sugar accumulation during postharvest cold storage. *Horticulture Research*, 10(4). <https://doi.org/10.1093/hr/uhad035>
- Lucas-Aguirre, J. C., Quintero-Castaño, V. D., Henao-Ossa, J. S., Barrón-García, O. Y., & Rodríguez-García, M. E. (2025). Influence of Germination Time on the Morphological, Structural, Vibrational, Thermal and Pasting Properties of Potato Starch from *Solanum tuberosum* Phureja Group. *Potato Research*, 68(2), 1375–1395. <https://doi.org/10.1007/s11540-024-09784-3>
- Malik-Sheriff, R. S., Glont, M., Nguyen, T. V. N., Tiwari, K., Roberts, M. G., Xavier, A., Vu, M. T., Men, J., Maire, M., Kananathan, S., Fairbanks, E. L., Meyer, J. P., Arankalle, C., Varusai, T. M., Knight-Schrijver, V., Li, L., Dueñas-Roca, C., Dass, G., Keating, S. M., ... Hermjakob, H. (2019). BioModels—15 years of sharing computational models in life science. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkz1055>
- Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (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(10), 4647–4654. <https://doi.org/10.1093/molbev/msab199>
- Manrique-Carpintero, N. C., Berdugo-Cely, J. A., Cerón-Souza, I., Lasso-Paredes, Z., Reyes-Herrera, P. H., & 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. <https://doi.org/10.3389/fpls.2023.1046400>
- Martinez, C. A., Maestri, M., & Lani, E. G. (1996). In vitro salt tolerance and proline accumulation in Andean potato (*Solanum* spp.) differing in frost resistance. *Plant Science*, 116(2), 177–184. [https://doi.org/10.1016/0168-9452\(96\)04374-9](https://doi.org/10.1016/0168-9452(96)04374-9)

- Novère, N. L., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., Crampin, E. J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J. L., Spence, H. D., & Wanner, B. L. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology*, *23*(12), 1509–1515. <https://doi.org/10.1038/nbt1156>
- Ñústez López, C. E., & Rodríguez Molano, L. E. (2024). *Papa criolla (Solanum tuberosum L. Grupo Phureja) : manual de recomendaciones técnicas para su cultivo en el departamento de Cundinamarca* (Primera). CORREDOR TECNOLÓGICO AGROINDUSTRIAL CTA-2. <https://repositorio.unal.edu.co/handle/unal/86779>
- Oberhardt, M. A., Palsson, B. Ø., & Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, *5*, 320. <https://doi.org/10.1038/msb.2009.77>
- Obidiegwu, J. E. (2015). Coping with drought: Stress and adaptive responses in potato and perspectives for improvement. *Frontiers in Plant Science*, *6*. <https://doi.org/10.3389/fpls.2015.00542>
- Ochoa Neves, C. M., Ugent, D., Correll, D. S., & Frey, F. (1990). *The potatoes of South America: Bolivia* (1. publ). Cambridge Univ. Press.
- Ojeda Pérez, Z. Z., Arias Moreno, D. M., Bohórquez Quintero, M. D. L. Á., Pacheco Díaz, J. E., & Araque Barrera, E. J. (2021, September 7). *Colores y sabores de mi tierra: Papas nativas cultivadas en Boyacá* (1st ed.). UPTC. <https://doi.org/10.19053/9789586605175>
- Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*, 245–248. <https://doi.org/10.1038/nbt.1614>
- Palit, P., Kudapa, H., Zougmore, R., Kholova, J., Whitbread, A., Sharma, M., & Varshney, R. K. (2020). An integrated research framework combining genomics, systems biology, physiology, modelling and breeding for legume improvement in response to elevated CO₂ under climate change scenario. *Current Plant Biology*, *22*, 100149. <https://doi.org/10.1016/j.cpb.2020.100149>
- Park, D.-J., Kim, D.-H., Yong, S.-H., Kim, S.-A., Park, K.-B., Cha, S.-A., Lee, J.-H., & Choi, M.-S. (2024). Establishment of Efficient Method for Evaluation of Heat Stress Tolerance of Herbaceous Plant Species and Selection of Heat-Tolerant Plants. *Horticulturae*, *10*(12), 1290. <https://doi.org/10.3390/horticulturae10121290>
- Peña, C., Restrepo-Sánchez, L.-P., Kushalappa, A., Rodríguez-Molano, L.-E., Mosquera, T., & Narváez-Cuenca, C.-E. (2015). Nutritional contents of advanced breeding clones of *Solanum tuberosum* group Phureja. *LWT - Food Science and Technology*, *62*(1), 76–82. <https://doi.org/10.1016/j.lwt.2015.01.038>
- Pham, G. M., Hamilton, J. P., Wood, J. C., Burke, J. T., Zhao, H., Vaillancourt, B., Ou, S., Jiang, J., & Buell, C. R. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience*, *9*(9). <https://doi.org/10.1093/gigascience/giaa100>
- Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1432. <https://doi.org/10.1038/s41467-020-14998-3>
- Reijnders, M. J., Van Heck, R. G., Lam, C. M., Scaife, M. A., Santos, V. A. M. D., Smith, A. G., & Schaap, P. J. (2014). Green genes: Bioinformatics and systems-biology innovations drive algal biotechnology. *Trends in Biotechnology*, *32*(12), 617–626. <https://doi.org/10.1016/j.tibtech.2014.10.003>
- Rodríguez, L. E., Ñústez, C. E., & Estrada, N. (2009). Criolla Latina, Criolla Paisa y Criolla Colombia, nuevos cultivares de papa criolla para el departamento de Antioquia (Colombia). *Agronomía Colombiana*, *27*(3), 289–303. Retrieved July 11, 2025, from <https://revistas.unal.edu.co/index.php/agrocol/article/view/13204>
- Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R., & Palsson, B. Ø. (2011). Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox v2.0. *Nature Protocols*, *6*(9), 1290–1307. <https://doi.org/10.1038/nprot.2011.308>

- Seaver, S. M. D., Liu, F., Zhang, Q., & al., e. (2021). The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Research*, 49(D1), D575–D588. <https://doi.org/10.1093/nar/gkaa746>
- Seng, S., Wu, J., Sui, J., Wu, C., Zhong, X., Liu, C., Liu, C., Gong, B., Zhang, F., He, J., & Yi, M. (2016). ADP-glucose pyrophosphorylase gene plays a key role in the quality of corm and yield of cormels in gladiolus. *Biochemical and Biophysical Research Communications*, 474(1), 206–212. <https://doi.org/10.1016/j.bbrc.2016.04.103>
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Skraly, F. A., Ambavaram, M. M., Peoples, O., & Snell, K. D. (2018). Metabolic engineering to increase crop yield: From concept to execution. *Plant Science*, 273, 23–32. <https://doi.org/10.1016/j.plantsci.2018.03.011>
- Stelling, J., & Klamt, S. (2012). Basic concepts and principles of stoichiometric modeling of metabolic networks. *Constraint-Based Reconstruction and Analysis*, 1–30. <https://doi.org/10.1002/biot.201200291>
- Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J., & Doyle, J. (2004). Robustness of Cellular Functions. *Cell*, 118(6), 675–685. <https://doi.org/10.1016/j.cell.2004.09.008>
- Sturaro, M. (2025). Carotenoids in Potato Tubers: A Bright Yellow Future Ahead. *Plants*, 14(2), 272. <https://doi.org/10.3390/plants14020272>
- Suttle, J. C., Abrams, S. R., De Stefano-Beltrán, L., & Huckle, L. L. (2012). Chemical inhibition of potato ABA-8'-hydroxylase activity alters in vitro and in vivo ABA metabolism and endogenous ABA levels but does not affect potato microtuber dormancy duration. *Journal of Experimental Botany*, 63(15), 5717–5725. <https://doi.org/10.1093/jxb/ers146>
- Sweetlove, L. J., & Ratcliffe, R. G. (2011). Flux-Balance Modeling of Plant Metabolism. *Frontiers in Plant Science*, 2. <https://doi.org/10.3389/fpls.2011.00038>
- The Potato Genome Sequencing Consortium. (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355), 189–195. <https://doi.org/10.1038/nature10158>
- Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1), 93–121. <https://doi.org/10.1038/nprot.2009.203>
- Toinga-Villafuerte, S., Vales, M. I., Awika, J. M., & Rathore, K. S. (2022). CRISPR/Cas9-Mediated Mutagenesis of the Granule-Bound Starch Synthase Gene in the Potato Variety Yukon Gold to Obtain Amylose-Free Starch in Tubers. *International Journal of Molecular Sciences*, 23(9), 4640. <https://doi.org/10.3390/ijms23094640>
- Tong, H., Küken, A., Razaghi-Moghadam, Z., & Nikoloski, Z. (2021). Characterization of effects of genetic variants via genome-scale metabolic modelling. *Cellular and Molecular Life Sciences*, 78(12), 5123–5138. <https://doi.org/10.1007/s00018-021-03844-4>
- Van Lieshout, N., Van Der Burgt, A., De Vries, M. E., Ter Maat, M., Eickholt, D., Esselink, D., Van Kaauwen, M. P. W., Kodde, L. P., Visser, R. G. F., Lindhout, P., & Finkers, R. (2020). Solyntus, the New Highly Contiguous Reference Genome for Potato (*Solanum tuberosum*). *G3: Genes | Genomes | Genetics*, 10(10), 3489–3495. <https://doi.org/10.1534/g3.120.401550>
- Wang, F., Xia, Z., Zou, M., Zhao, L., Jiang, S., Zhou, Y., Zhang, C., Ma, Y., Bao, Y., Sun, H., Wang, W., & Wang, J. (2022). The autotetraploid potato genome provides insights into highly heterozygous species. *Plant Biotechnology Journal*, 20(10), 1996–2005. <https://doi.org/10.1111/pbi.13883>
- Wang, Z., Ma, R., Zhao, M., Wang, F., Zhang, N., & Si, H. (2020). NO and ABA Interaction Regulates Tuber Dormancy and Sprouting in Potato. *Frontiers in Plant Science*, 11. <https://doi.org/10.3389/fpls.2020.00311>

- Weckwerth, W. (2011). Green systems biology — From single genomes, proteomes and metabolomes to ecosystems research and biotechnology. *Journal of Proteomics*, 75(1), 284–305. <https://doi.org/10.1016/j.jprot.2011.07.010>
- Westgeest, A. J., Vasseur, F., Enquist, B. J., Milla, R., Gómez-Fernández, A., Pot, D., Vile, D., & Violle, C. (2024). An allometry perspective on crops. *New Phytologist*, 244(4), 1223–1237. <https://doi.org/10.1111/nph.20129>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1). <https://doi.org/10.1038/sdata.2016.18>
- Williams, T. C., Poolman, M. G., Howden, A. J., Schwarzlander, M., Fell, D. A., Ratcliffe, R. G., & Sweetlove, L. J. (2010). A Genome-Scale Metabolic Model Accurately Predicts Fluxes in Central Carbon Metabolism under Stress Conditions. *Plant Physiology*, 154(1), 311–323. <https://doi.org/10.1104/pp.110.158535>
- Xu, Y., Yang, W., Qiu, J., Zhou, K., Yu, G., Zhang, Y., Wang, X., Jiao, Y., Wang, X., Hu, S., Zhang, X., Li, P., Lu, Y., Chen, R., Tao, T., Yang, Z., Xu, Y., & Xu, C. (2025). Metabolic marker-assisted genomic prediction improves hybrid breeding. *Plant Communications*, 6(3), 101199. <https://doi.org/10.1016/j.xplc.2024.101199>
- Yao, P., Cui, J., Zhang, C., Wei, J., Su, X., Sun, C., Bi, Z., Liu, Z., Bai, J., & Liu, Y. (2024). Overexpression of the Potato StPYL20 Gene Enhances Drought Resistance and Root Development in Transgenic Plants. *International Journal of Molecular Sciences*, 25(23), 12748. <https://doi.org/10.3390/ijms252312748>
- Zagorščak, M., Abdelhakim, L., Rodriguez-Granados, N. Y., Široká, J., Ghatak, A., Bleker, C., Blejec, A., Zrimec, J., Novák, O., Pěňčík, A., Baebler, Š., Borroto, L. P., Schuy, C., Županič, A., Afjehi-Sadat, L., Wurzinger, B., Weckwerth, W., Novak, M. P., Knight, M. R., ... Teige, M. (2024, July 23). *Integration of multi-omics data and deep phenotyping of potato enables novel insights into single- and combined abiotic stress responses*. <https://doi.org/10.1101/2024.07.18.604140>