



**Ensamblaje a nivel cromosómico y anotación del genoma de aislados de
Colletotrichum sp., agente causal de antracnosis en *Solanum betaceum* Cav.
(tomate de árbol) y *Solanum quitoense* Lam. (naranjilla).**

Acosta Villamarín, Sofía Micaella

Maestría en Biología Computacional

Trabajo de titulación, previo a la obtención del título de Magister en Biología Computacional

Ing. Francisco Javier Flores Flor, Ph.D.

2025

ÍNDICE DE CONTENIDOS

ÍNDICE DE CONTENIDOS.....	2
ÍNDICE DE FIGURAS.....	5
ÍNDICE DE TABLAS.....	6
ÍNDICE DE ANEXOS.....	7
RESUMEN.....	8
ABSTRACT	9
CAPÍTULO I: INTRODUCCIÓN	10
1.1 Planteamiento del problema	10
1.2 Justificación del problema	11
1.3 Objetivos de la investigación	12
1.3.1 Objetivo General	12
1.3.2 Objetivos Específicos	12
CAPÍTULO II: REVISIÓN DE LA LITERATURA	13
2.1 Cultivos de tomate de árbol y naranjilla en Ecuador.....	13
2.1.1 Importancia económica y social.....	14
2.1.2 Principales problemas fitosanitarios	14
2.2 El género <i>Colletotrichum</i>	15
2.2.1 Taxonomía y clasificación.....	15
2.2.2 Morfología y ciclo de vida	16
2.2.3 Estrategias de infección y adaptación al hospedero	16
2.2.4 Variabilidad genética y plasticidad fenotípica	17

	3
2.3 Antracnosis	18
2.3.1 Sintomatología.....	18
2.3.2 Métodos de control y manejo actuales.....	19
2.4 Avances en la caracterización de <i>Colletotrichum</i>	21
2.4.1 Métodos tradicionales de identificación	21
2.4.2 Métodos moleculares.....	22
2.4.3 Análisis genómicos	22
2.4.4 Genómica de <i>Colletotrichum</i>	23
2.5 Herramientas bioinformáticas y genómica aplicada a hongos fitopatógenos	24
2.5.1 Importancia del análisis genómico.....	24
2.5.2 Secuenciación de nueva generación (NGS)	24
2.5.3 Ensamblaje genómico	26
2.5.4 Anotación genómica	27
CAPÍTULO III: METODOLOGÍA	29
3.1 Obtención de las secuencias	29
3.2 Procesamiento de Secuencias Crudas	30
3.3 Ensamblaje a Nivel Cromosómico del Genoma	31
3.3.1 Ensamblaje de novo con SPAdes.....	31
3.3.2 Selección y preparación de la referencia	31
3.3.3 Ensamblaje a Nivel Cromosómico con Ragout	31
3.3.4 Evaluación del ensamblaje.....	32

3.4 Anotación del Genoma.....	32
3.4.1 Anotación Estructural del Genoma	32
3.4.2 Anotación Funcional del Genoma	34
3.5 Comparación de los Ensamblajes a Nivel Cromosómico vs. Nivel <i>Contigs</i>	34
3.5.1 Comparación Estructural de los Ensamblajes.....	34
3.5.2 Comparación Funcional	35
CAPÍTULO IV: RESULTADOS Y DISCUSIÓN	36
4.1 Procesamiento de Secuencias Crudas	36
4.2 Ensamblaje a Nivel Cromosómico del Genoma	37
4.2.1 Ensamblaje de novo con SPAdes.....	37
4.2.2 Ensamblaje a Nivel Cromosómico con Ragout	38
4.2.3 Evaluación del Ensamblaje.....	39
4.3 Anotación del Genoma.....	42
4.3.1 Anotación Estructural del Genoma	42
4.3.2 Anotación Funcional del Genoma	44
4.4 Comparación de los Ensamblajes	51
4.4.1 Comparación Estructural de los Ensamblajes.....	51
4.4.2 Comparación Funcional de los Ensamblajes	55
CAPÍTULO V: CONCLUSIONES	59
CAPÍTULO VI: RECOMENDACIONES.....	60
CAPÍTULO VII: REFERENCIAS	61

CAPÍTULO IX: ANEXOS	76
---------------------------	----

ÍNDICE DE FIGURAS

Figura 1 <i>Cultivo de Tomate de Árbol y Naranjilla en Ecuador</i>	13
Figura 2 <i>Ciclo de Vida General de Colletotrichum spp.</i>	17
Figura 3 <i>Síntomas de Antracnosis en Frutos de Tomate de Árbol y Naranjilla</i>	19
Figura 4 <i>Esquema General del Flujo de la Secuenciación de Nueva Generación (NGS)</i>	25
Figura 5 <i>Bloques de sintonía entre cromosomas homólogos de especies cercanas</i>	27
Figura 6 <i>Proceso de ensamblaje y anotación de un genoma</i>	28
Figura 7 <i>Pipeline metodológico empleado en el presente estudio</i>	29
Figura 8 <i>Gráficos de calidad de lecturas después del filtrado</i>	36
Figura 9 <i>Representación visual de los scaffolds ensamblados de Col_ToNa</i>	38
Figura 10 <i>Representación visual de los scaffolds ensamblados de Col_To</i>	39
Figura 11 <i>Representación visual de los scaffolds ensamblados de Col_Na</i>	39
Figura 12 <i>Familias Pfam más abundantes en Col_ToNa</i>	47
Figura 13 <i>Familias Pfam más abundantes en Col_To</i>	48
Figura 14 <i>Familias Pfam más abundantes en Col_Na</i>	48
Figura 15 <i>Suma acumulada de las longitudes de los contigs por ensamblaje de Col_ToNa</i>	53
Figura 16 <i>Grafo de ensamblaje de novo con SPAdes del aislado Col_ToNa</i>	53
Figura 17 <i>Grafo de ensamblaje cromosómico con Ragout del aislado Col_ToNa</i>	54
Figura 18 <i>Histograma de longitudes de contigs (>10 Kb) de los ensamblajes de Col_ToNa</i>	54
Figura 19 <i>Cobertura de anotación funcional en Col_ToNa</i>	57

ÍNDICE DE TABLAS

Tabla 1 <i>Aislados de Colletotrichum analizados</i>	30
Tabla 2 <i>Resumen por aislado del control de calidad y filtrado de lecturas</i>	37
Tabla 3 <i>Métricas de ensamblaje de novo (SPAdes) por aislado</i>	37
Tabla 4 <i>Métricas de ensamblaje asistido por referencia (Ragout) por aislado</i>	40
Tabla 5 <i>Contigs no anclados y Ns incorporadas después de Ragout</i>	40
Tabla 6 <i>Evaluación de completitud génica mediante BUSCO</i>	42
Tabla 7 <i>Resumen de la anotación estructural y el efecto del filtrado de isoformas</i>	43
Tabla 8 <i>Completitud de la anotación estructural con BUSCO antes de depurar isoformas</i>	43
Tabla 9 <i>Completitud de la anotación estructural con BUSCO tras depurar isoformas</i>	44
Tabla 10 <i>Anotación funcional por homología (DIAMOND/Swiss-Prot) por aislado</i>	46
Tabla 11 <i>Anotación funcional por dominios (Pfam/HMMER) por aislado</i>	47
Tabla 12 <i>Comparación de métricas de ensamblaje (SPAdes vs. Ragout) entre aislados</i>	52
Tabla 13 <i>Genes codificantes por aislado (Ragout vs. SPAdes)</i>	55

ÍNDICE DE ANEXOS

Anexo A Descarga del genoma de referencia de <i>Colletotrichum scovillei</i> (NCBI dataset)	76
Anexo B Preprocesamiento de secuencias	76
Anexo C Ensamblaje asistido por referencia con Ragout: receta .rcp y ejecución	76
Anexo D Evaluación del ensamblaje cromosómico frente a la referencia con QUAST	77
Anexo E Comparación de ensamblajes SPAdes vs. Ragout con QUAST	77
Anexo F Evaluación de completitud del ensamblaje con BUSCO	77
Anexo G Comparación de métricas (SPAdes vs. Ragout) - Aislado Col_ToNa	78
Anexo H Comparación de métricas (SPAdes vs. Ragout) - Aislado Col_To	78
Anexo I Curva de suma acumulada de las longitudes de los contigs - Aislado Col_To	79
Anexo J Histograma de longitudes de contigs (>10 Kb) - Aislado Col_To	79
Anexo K Grafo de ensamblaje de novo (SPAdes) - Aislado Col_To	80
Anexo L Grafo de ensamblaje cromosómico (Ragout) - Aislado Col_To	80
Anexo M Comparación de métricas (SPAdes vs. Ragout) - Aislado Col_Na	81
Anexo N Curva de suma acumulada de las longitudes de los contigs - Aislado Col_Na	81
Anexo O Histograma de longitudes de contigs (>10 Kb) - Aislado Col_Na	82
Anexo P Grafo de ensamblaje de novo (SPAdes) - Aislado Col_Na	82
Anexo Q Grafo de ensamblaje cromosómico (Ragout) - Aislado Col_Na	83
Anexo R Ejecución de la anotación estructural con BRAKER3	83
Anexo S Métricas de la anotación estructural: genes, transcritos y proteínas (GTF/FAA)	83
Anexo T Completitud de la anotación estructural con BUSCO	84
Anexo U Selección de la isoforma más larga por gen (AGAT)	84
Anexo V Extracción de CDS y proteínas a partir de GFF3 y genoma (gffread)	84
Anexo W Anotación por homología (DIAMOND BLASTp vs UniProt/Swiss-Prot)	84
Anexo X Anotación de dominios proteicos (Pfam-A; pfam_scan.pl/HMMER)	84

RESUMEN

La antracnosis es una enfermedad causada por hongos del género *Colletotrichum*, que impacta gravemente cultivos de interés agrícola y económico en Ecuador, como el tomate de árbol y naranjilla, limitando su rendimiento y vida poscosecha. En este estudio se realizó el ensamblaje a escala cromosómica y la anotación de tres aislados locales de *Colletotrichum* como base para una caracterización funcional más precisa. Se procesaron lecturas de Illumina mediante el control de calidad y filtrado, luego, se generó un borrador *de novo* con SPAdes, y, posteriormente, se efectuó el reordenamiento cromosómico por sintenia frente a una referencia cercana (*C. scovillei*) mediante Ragout. La calidad estructural se evaluó con QUAST y BUSCO. La anotación estructural se realizó con BRAKER3, se depuró a una isoforma por gen (AGAT + gffread) y se extrajeron las proteínas y CDS correspondientes. La anotación funcional combinó homología curada (DIAMOND/Swiss-Prot) y detección de dominios (Pfam/HMMER).

Los ensamblajes resultantes se consolidaron en 14-16 *scaffolds* por aislado, con N50 = 4.5-5.0 Mb, L50 = 5 y completitud BUSCO = 99.6-99.7%, compatibles con borradores a escala cromosómica (pseudocromosomas). BRAKER3 predijo ~13,700-13,900 genes por aislado, y los conjuntos de proteínas no redundantes mostraron completitud BUSCO ~99.9%. La anotación funcional alcanzó 59% de cobertura por homología y 70% por dominios, recuperando un núcleo funcional consistente (transportadores MFS/ABC, reguladores Zn_clus/Fungal_trans, P450/SDR, quinasas, proteínas HET), congruente con la biología hemibiótrofa del género *Colletotrichum*. En comparación con ensamblajes a nivel de *contigs*, el ordenamiento cromosómico redujo la fragmentación, mejoró continuidad y favoreció modelos génicos más íntegros e interpretables.

ABSTRACT

Anthrachnose is a disease caused by fungi of the genus *Colletotrichum* which severely affects crops of agricultural and economic importance in Ecuador, such as tree tomato and naranjilla, limiting yield and postharvest life. In this study, chromosome-scale assembly and annotation were performed for three local *Colletotrichum* isolates as a basis for more precise functional characterization. Illumina reads were subjected to quality control and filtering; a *de novo* draft assembly was then generated with SPAdes, and subsequently chromosome-scale synteny-guided reordering against a closely related reference (*C. scovillei*) was carried out using Ragout. Structural quality was assessed with QAST and BUSCO. Structural annotation was performed with BRAKER3, pruned to one isoform per gene (AGAT + gffread), and the corresponding proteins and CDS were extracted. Functional annotation combined curated homology (DIAMOND/Swiss-Prot) and domain detection (Pfam/HMMER).

The resulting assemblies were consolidated into 14–16 scaffolds per isolate, with N50 = 4.5–5.0 Mb, L50 = 5, and BUSCO completeness = 99.6–99.7%, consistent with chromosome-scale drafts (pseudochromosomes). BRAKER3 predicted ~13,700–13,900 genes per isolate, and the non-redundant protein sets showed BUSCO completeness of ~99.9%. Functional annotation achieved 59% coverage by homology and 70% by domains, recovering a consistent functional core (MFS/ABC transporters, Zn_clus/Fungal_trans regulators, P450/SDR, kinases, HET protein), in line with the hemibiotrophic biology of *Colletotrichum*. Compared with contig-level assemblies, chromosome-scale ordering reduced fragmentation, improved continuity, and favored more complete and interpretable gene models.

CAPÍTULO I: INTRODUCCIÓN

1.1 Planteamiento del problema

La antracnosis, conocida comúnmente como “ojo de pollo”, es una enfermedad causada por hongos del género *Colletotrichum*. Afecta tallos, hojas e inflorescencias, aunque el mayor impacto se observa en los frutos, donde, según las condiciones ambientales, puede provocar su momificación o pudrición (Castro-Quezada et al., 2019; Ochoa et al., 2010). Este patógeno se encuentra entre los diez hongos fitopatógenos de mayor relevancia a nivel mundial (Ciofini et al., 2022).

La caracterización de *Colletotrichum* representa un desafío debido a su alta variabilidad morfológica, patogénica y genética, lo que dificulta el desarrollo de estrategias de manejo efectivas y sostenibles (Casarrubias Carillo et al., 2003; Rodríguez-Ortega et al., 2018; Ruiz-Campos et al., 2022). Tradicionalmente, su identificación se ha basado en criterios morfológicos, pruebas de patogenicidad y sensibilidad a fungicidas. Sin embargo, estos métodos no siempre permiten diferenciar con precisión especies estrechamente relacionadas (Casarrubias Carillo et al., 2003).

A pesar de los avances en genómica, en Ecuador no se han desarrollado estudios de ensamblaje y anotación del genoma de *Colletotrichum* a nivel de cromosomas en aislados locales. Esto limita la identificación de genes asociados con virulencia, patogenicidad y adaptación al hospedero; lo cual restringe el alcanzar una comprensión profunda sobre el patógeno, obstaculizando el diseño de estrategias de control más efectivas. Como consecuencia, persisten pérdidas considerables en la producción de cultivos de interés agrícola como el tomate de árbol y la naranjilla.

1.2 Justificación del problema

En Ecuador, el cultivo y la comercialización de tomate de árbol y naranjilla son parte fundamental del sector agrícola, tanto por su valor económico como por su relevancia social. Se estima que alrededor de 10,000 productores dependen de estas cadenas productivas. Además, su exportación representa cerca del 5.3% del Valor Agregado Bruto Agropecuario, con destinos principales en Estados Unidos, España, Reino Unido y Países Bajos (Ministerio de Agricultura y Ganadería, 2024a, 2024b).

En 2023, las exportaciones de tomate de árbol alcanzaron 1,145 toneladas, generando USD 3 millones, representando 0.03% de las exportaciones agropecuarias del país. Sin embargo, se reportó la pérdida de 34 hectáreas de cultivos, de las cuales, el 82% fueron afectadas por plagas y enfermedades (Ministerio de Agricultura y Ganadería, 2024b). En el caso de la naranjilla, se exportaron 110 toneladas, generando USD 332,000 (0.004% de las exportaciones agropecuarias); sus pérdidas alcanzaron las 29 hectáreas, con un 22% atribuidas a plagas (Ministerio de Agricultura y Ganadería, 2024a).

Entre los principales problemas fitosanitarios que enfrentan estos cultivos se encuentra la antracnosis, con pérdidas de rendimiento estimadas entre el 60 y el 80% (Castro-Quezada et al., 2019; Ochoa et al., 2010). Aunque se han realizado estudios sobre su aislamiento, caracterización morfológica y comportamiento patogénico (Guevara-Suarez et al., 2022), el conocimiento genómico del patógeno en Ecuador sigue siendo limitado.

El análisis genómico a nivel de cromosomas de aislados locales de *Colletotrichum* podría proporcionar información de gran interés sobre su diversidad genética, mecanismos de infección y estrategias de adaptación (García Romero, 2019). Esto permitiría optimizar la identificación de genes clave y respaldar el desarrollo de estrategias de control más precisas, contribuyendo a la sostenibilidad de cultivos de alto valor agrícola y económico.

1.3 Objetivos de la investigación

1.3.1 Objetivo General

Realizar el ensamblaje a nivel cromosómico y la anotación del genoma de aislados de *Colletotrichum* sp., causante de antracnosis en *Solanum betaceum* (tomate de árbol) y *Solanum quitoense* (naranjilla).

1.3.2 Objetivos Específicos

- Ensamblar a nivel cromosómico el genoma de aislados de *Colletotrichum* sp. utilizando herramientas bioinformáticas.
- Anotar los genomas ensamblados de *Colletotrichum* sp. mediante la identificación de genes codificantes, elementos repetitivos y/o regiones funcionales.
- Evaluar la calidad y precisión del ensamblaje a nivel cromosómico y anotación de los genomas de *Colletotrichum* sp. por comparación con un ensamblaje convencional descrito en estudios previos.

CAPÍTULO II: REVISIÓN DE LA LITERATURA

2.1 Cultivos de tomate de árbol y naranjilla en Ecuador

El tomate de árbol (*Solanum betaceum*; Figura 1A) es una solanácea originaria de los Andes sudamericanos, cultivada principalmente en Colombia, Ecuador y Perú. En Ecuador, su producción está mayoritariamente en manos de pequeños y medianos agricultores, concentrados en la región Sierra, especialmente en las provincias de Carchi, Imbabura, Pichincha, Tungurahua, Cotopaxi, Chimborazo y Azuay (Marcial Acosta et al., 2023; Ministerio de Agricultura y Ganadería, 2024b).

La naranjilla (*Solanum quitoense*; Figura 1B), también perteneciente a la familia Solanaceae, es originaria de los bosques subtropicales húmedos de la cordillera de los Andes en Ecuador, Colombia y Perú. En el país, su cultivo comercial se desarrolla principalmente en provincias como Napo, Pastaza y Morona Santiago; y en menor escala en Tungurahua, Pichincha, Imbabura, Carchi y Santo Domingo de los Tsáchilas (Ministerio de Agricultura y Ganadería, 2024a; Revelo M. et al., 2010).

Figura 1

Cultivo de Tomate de Árbol y Naranjilla en Ecuador



Nota. Tomate de árbol (A). Naranjilla (B). Tomado de *Fichas técnicas INIAP*, por INIAP, 2024.

Ambos cultivos son parte esencial de la agrobiodiversidad ecuatoriana y constituyen una fuente importante de ingresos y sustento para miles de familias, lo que resalta su relevancia económica y social.

2.1.1 Importancia económica y social

El tomate de árbol es uno de los cultivos más relevantes de la Sierra ecuatoriana, con una producción anual estimada de 39,725 toneladas. De esta cifra, aproximadamente 30,500 toneladas se destinan a la industria agrícola y la exportación, mientras que unas 8,600 toneladas se comercializan en mercados locales y regionales. En los últimos años, sus exportaciones han incrementado, generando ingresos significativos para las provincias productoras. Este cultivo impulsa la economía familiar agrícola, fomenta la creación de empleo y contribuye a la economía local al mejorar las condiciones productivas y comerciales de los agricultores (Díaz Granda et al., 2017; Ministerio de Agricultura y Ganadería, 2024b).

La naranjilla, por otro lado, presenta una producción cercana a 28,000 toneladas anuales. Es una fruta de alto valor para la agroindustria y se estima que su cultivo beneficia directamente a unas 12,000 familias, en su mayoría pequeños agricultores que emplean mano de obra familiar, reforzando su papel como generador de empleo rural. Además, al ser un cultivo tradicional, posee un valor cultural significativo en las regiones productoras, aunque enfrenta retos asociados al manejo fitosanitario y a la limitada incorporación de tecnologías para mejorar la productividad (Castro López & Herrera, 2019; Ministerio de Agricultura y Ganadería, 2024a).

2.1.2 Principales problemas fitosanitarios

La productividad de los cultivos de tomate de árbol y naranjilla se ve amenazada por diversas plagas y enfermedades. Entre estas, la antracnosis, causada por hongos del género *Colletotrichum*, se considera una de las más perjudiciales, con pérdidas de rendimiento estimadas entre el 60 y el 80% en Ecuador. Su impacto es grave porque compromete tanto la

producción en campo como la calidad poscosecha, y esto reduce la competitividad en los mercados y pone en riesgo la sostenibilidad de las plantaciones (Castro-Quezada et al., 2019; Dofuor et al., 2023).

2.2 El género *Colletotrichum*

El género *Colletotrichum* comprende un amplio número de hongos fitopatógenos que afectan una gran diversidad de cultivos, incluyendo hortalizas, legumbres, cereales y especies tropicales y subtropicales en todo el mundo (Guevara-Suarez et al., 2022; O'Connell et al., 2012). En la región norte de Sudamérica, especialmente en Colombia, Ecuador, Perú y Venezuela, estos patógenos han ocasionado pérdidas significativas en cultivos de alto valor agrícola y comercial, generando un impacto considerable (Castro-Quezada et al., 2019).

2.2.1 Taxonomía y clasificación

Taxonómicamente, *Colletotrichum* es un género complejo y de difícil clasificación. Desde su primera descripción en 1,831, se han registrado 942 especies, aunque estudios recientes sugieren que solo alrededor de 257 son aceptadas (Guevara-Suarez et al., 2022). En la última década, su taxonomía ha experimentado cambios significativos debido al uso de datos moleculares y análisis filogenéticos, ya que las características fenotípicas por sí solas no son suficientemente precisas para su identificación (Ruiz-Campos et al., 2022).

Los estudios moleculares han revelado la existencia de al menos once complejos de especies, definidos como grupos morfológicamente indistinguibles, pero genéticamente diferenciados (Ruiz-Campos et al., 2022). Actualmente, la clasificación integra la evaluación morfológica con el análisis multilocus de marcadores moleculares como ITS para separar taxones a nivel de complejos; además, genes como GAPDH facilitan una identificación preliminar entre especies (Guevara-Suarez et al., 2022; F. Liu et al., 2013; Weir et al., 2012).

2.2.2 Morfología y ciclo de vida

El cultivo, *Colletotrichum* forma colonias planas con márgenes definidos y micelio aéreo de densidad variable, con tonalidades que van del blanco al gris o rosado, dependiendo del medio y las condiciones ambientales. El reverso puede mostrar pigmentaciones grisáceas, verdosas o anaranjadas por la esporulación. Microscópicamente, presenta conidióforos hialinos, cortos y no ramificados, que producen conidios unicelulares, hialinos y de forma cilíndrica a fusiforme, dispuestos en masas de color salmón o rosado (Castro-Quezada et al., 2019; Prada & Basto, 2011).

Los conidios son las estructuras reproductivas responsables de la generación y dispersión del inóculo, estos marcan el inicio de su ciclo de vida en el hospedero (Figura 2). El ciclo comienza cuando estos entran en contacto con la superficie vegetal, germinan y forman apresorios (estructuras especializadas que facilitan la penetración mediante presión mecánica y la secreción de enzimas degradativas). Una vez dentro de los tejidos, el hongo desarrolla una fase biotrófica inicial, caracterizada por hifas bulbosas que colonizan células vivas sin destruirlas. Posteriormente, se dirige a una fase necrotrófica, en la que crecen hifas delgadas y ramificadas que degradan rápidamente los tejidos, provocando necrosis y el desarrollo de síntomas visibles en la planta (Ma et al., 2022; O'Connell et al., 2012).

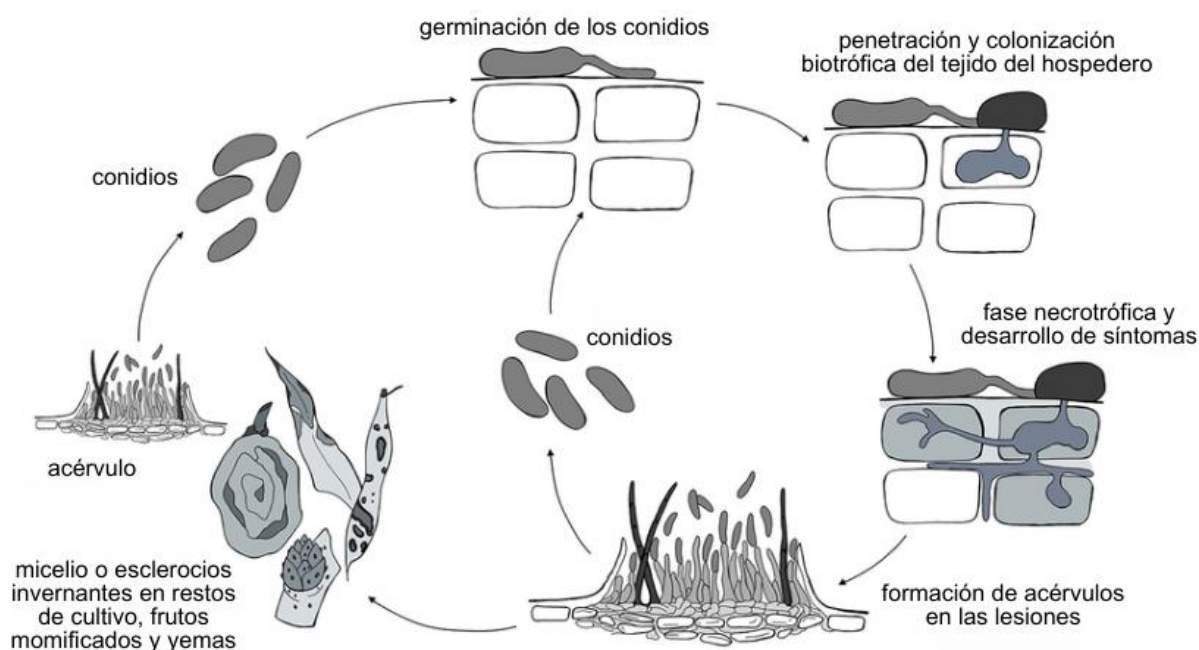
2.2.3 Estrategias de infección y adaptación al hospedero

Colletotrichum utiliza una estrategia de infección hemibiotrófica en la que alterna fases biotróficas y necrotróficas (Ma et al., 2022; O'Connell et al., 2012). La eficacia de este proceso radica en su capacidad para modular la respuesta inmune de la planta mediante genes efectores, que facilitan la colonización y expansión del patógeno (Rodríguez-López et al., 2009). Algunos aislados muestran una marcada especificidad hacia ciertos hospederos, mientras que otros presentan un amplio rango de plantas susceptibles, lo que influye

directamente en su distribución, virulencia y dificultad de control (Casarrubias Carillo et al., 2003; Falconí et al., 2013; Gan et al., 2021).

Figura 2

Ciclo de Vida General de *Colletotrichum* spp.



Nota. El hongo sobrevive como micelio o esclerocios en restos vegetales y produce conidios que, tras germinar y formar apresorios, penetran al hospedero. La infección progresa de una base biotrófica a una necrotrófica, generando lesiones que originan nuevos acérvulos y conidios responsables de ciclos secundarios de infección. Adaptado de “*Development of a model for Colletotrichum diseases with calibration for phylogenetic clades on different host plants*”, por Salotti, et al., 2023, *Frontiers in Plant Sciences*.

2.2.4 Variabilidad genética y plasticidad fenotípica

La variabilidad genética en *Colletotrichum* se manifiesta tanto en diferencias moleculares como en la capacidad de las poblaciones para adaptarse a condiciones ambientales diversas. Esta plasticidad fenotípica se refleja en variaciones de agresividad,

resistencia a fungicidas y capacidad de infectar múltiples especies vegetales. Estos factores favorecen su supervivencia, además de que dificultan el manejo integrado y la implementación de programas de control sostenibles (Casarrubias Carillo et al., 2003; Falconí et al., 2013; Gan et al., 2021).

2.3 Antracnosis

La antracnosis, también conocida como “ojo de pollo”, es considerada por muchos productores como la enfermedad más importante del tomate de árbol y una de las de mayor incidencia en la naranjilla en Ecuador. Se encuentra ampliamente distribuida en las zonas productoras y puede ocasionar pérdidas significativas tanto en campo como en poscosecha (Castro-Quezada et al., 2019; León F. et al., 2004). Además de reducir el rendimiento, impacta negativamente en la calidad comercial del fruto, limitando su vida útil y su valor en el mercado (Dofuor et al., 2023; Revelo M. et al., 2010). La enfermedad presenta mayor incidencia durante la época lluviosa, bajo condiciones de temperatura promedio entre 13 a 15 °C y humedad relativa superior al 90%, que favorecen la germinación y dispersión del patógeno (Castro-Quezada et al., 2019; León F. et al., 2004).

2.3.1 Sintomatología

Los síntomas de la antracnosis se observan principalmente en los frutos (Figura 3), aunque también pueden afectar tallos, ramas, hojas e inflorescencias (Castro-Quezada et al., 2019; Ochoa et al., 2010). En frutos en desarrollo, las primeras manifestaciones son decoloraciones y pequeñas lesiones de aspecto aceitoso, que progresivamente evolucionan a manchas oscuras, hundidas y de bordes bien definidos. En el centro de estas lesiones se forma un polvillo de color rosado o salmón, característico de la esporulación del hongo. A medida que la infección avanza, las lesiones se juntan hasta cubrir gran parte de la superficie, provocando la momificación del fruto (León F. et al., 2004).

En frutos maduros o cercanos a la cosecha, las lesiones se expanden rápidamente, comprometiendo el aspecto y la textura del fruto (Castro-Quezada et al., 2019). En hojas, la enfermedad se presenta como manchas con anillos concéntricos oscuros y bordes bien definidos, mientras que en inflorescencias provoca la caída prematura de flores y manchas en los pétalos (Ochoa et al., 2010). La infección puede permanecer latente hasta después de la cosecha, manifestándose durante el transporte y almacenamiento, lo que compromete la calidad, clasificación, empaque y comercialización del producto (Castro-Quezada et al., 2019; Dofuor et al., 2023; Ochoa et al., 2010).

Figura 3

Síntomas de Antracnosis en Frutos de Tomate de Árbol y Naranjilla



Nota. Tomate de árbol (A). Naranjilla (B). Tomado de “*Ojo de pollo y su control*”, por Instituto Nacional Autónomo de Investigaciones Agropecuarias (INIAP), 2010.

2.3.2 Métodos de control y manejo actuales

El manejo de la antracnosis se basa en la detección temprana y la reducción de fuentes de inóculo. El monitoreo frecuente de la plantación permite identificar focos iniciales y retirar manualmente los frutos afectados, evitando que actúen como reservorios del patógeno. En zonas húmedas o durante la época lluviosa, la aplicación de fungicidas puede disminuir la

incidencia, sin embargo, su efectividad rara vez supera el 60%, por lo que se recomienda integrarlos en un manejo integrado para lograr un control más sostenible (León F. et al., 2004).

Dentro de este enfoque, las estrategias tradicionalmente aplicadas incluyen tratamientos físicos y el uso de fungicidas químicos. Entre los métodos físicos se encuentran el calentamiento por vapor, el secado con aire forzado y la inmersión en agua caliente, capaces de inhibir el desarrollo de *Colletotrichum*, aunque con riesgo de afectar la calidad organoléptica y nutricional del fruto. El control químico emplea compuestos cúpricos, estrobilurinas, ditiocarbamatos, benzimidazoles y triazoles. Si bien estos productos reducen el inóculo y retrasan el desarrollo fúngico, el uso repetido de moléculas con el mismo modo de acción ha se relaciona con la aparición de cepas resistentes y supone riesgos para la salud humana y el medio ambiente (Ciofini et al., 2022).

Las regulaciones internacionales sobre límites máximos de residuos han registrado o prohibido ciertos fungicidas, obligando a los productores a invertir en tecnologías y prácticas que cumplan con dichos estándares (Ciofini et al., 2022). Esto ha impulsado la búsqueda de alternativas innovadoras y sostenibles, con menor toxicidad y bajo impacto ambiental. Entre ellas, el control biológico mediante bacterias, hongos antagonistas y levaduras ha mostrado reducciones considerables en la incidencia y severidad de la enfermedad. Así mismo, extractos vegetales y aceites esenciales de especies han evidenciado actividad antifúngica, aunque requieren optimización para su uso a escala comercial (Ciofini et al., 2022).

Otras alternativas incluyen el desarrollo estrategias basadas en biotecnología, como la edición génica mediante “Nuevas Tecnología de Mejoramiento” (NBT), que emplean sistemas como CRISPR-Cas9 y nucleasas específicas para introducir modificaciones genómicas precisas que incrementen la resistencia de las plantas. Sin embargo, su aplicación contra *Colletotrichum* está limitada por el conocimiento incompleto sobre los mecanismos genéticos que regulan la respuesta de las plantas a este patógeno (Ciofini et al., 2022).

A pesar de los avances en el desarrollo de estrategias de control químicas, biológicas y biotecnológicas, su eficacia a largo plazo depende de una identificación precisa del agente causal. Dada la complejidad taxonómica y la alta variabilidad genética de *Colletotrichum*, comprender su diversidad y mecanismos de patogenicidad es clave para diseñar medidas de manejo más efectivas (Casarrubias Carillo et al., 2003; Castro-Quezada et al., 2019; Falconí et al., 2013).

Dado que la efectividad de estas estrategias depende en gran medida de una correcta identificación del agente causal, los avances en la caracterización de *Colletotrichum* resultan esenciales para un manejo más preciso y sostenible de la antracnosis.

2.4 Avances en la caracterización de *Colletotrichum*

En la región interandina de Ecuador, *Colletotrichum acutatum* y *C. gloeosporioides* se han identificado como los principales agentes causales de la antracnosis en tomate de árbol y naranjilla, mientras que en la zona sur también se ha reportado *C. tamarilloi*, perteneciente al complejo *C. acutatum* (Caicedo et al., 2017; Castro-Quezada et al., 2019; Guevara-Suarez et al., 2022). Además, se ha registrado la presencia de *C. fructicola*, *C. scovillei* y *C. coccodes* en cultivos de otras especies de solanáceas del país (Cajas Salazar, 2020).

2.4.1 Métodos tradicionales de identificación

Tradicionalmente, la identificación de *Colletotrichum* se ha basado en la caracterización morfológica, como la observación del crecimiento de las colonias, la forma y tamaño de los conidios, y la presencia de setas; así como en la evaluación de la reacción de diferentes hospederos (Casarrubias Carillo et al., 2003; Castro-Quezada et al., 2019; Falconí et al., 2013). Sin embargo, la alta variabilidad fenotípica entre aislados y la existencia de especies morfológicamente indistinguibles limitan la precisión de este enfoque (Casarrubias Carillo et al., 2003). Por ejemplo, el complejo *C. acutatum* puede ser o no patógeno, mientras que *C.*

asianum, *C. boninense*, *C. tamarilloi* y *C. theobromicola* son patógenos no específicos y pueden infectar diversos hospederos (Castro-Quezada et al., 2019).

2.4.2 Métodos moleculares

El uso de secuencias de regiones como ITS y genes como GAPDH, ACT y TUB2 ha permitido diferenciar complejos de especies e identificar aislados a nivel molecular. En Colombia, estos marcadores han confirmado la presencia de especies pertenecientes a los complejos *C. acutatum*, *C. boninense* y *C. gloeosporioides* en frutos tropicales (Castro-Quezada et al., 2019). En Ecuador, estudios combinados de morfología y secuencias genéticas han confirmado la prevalencia de *C. acutatum* en la Sierra Norte (Pichincha, Cotopaxi, Imbabura) y de *C. tamarilloi* en la Sierra Sur (Azuay, Loja) (Caicedo et al., 2017; Falconí et al., 2013).

2.4.3 Análisis genómicos

Aunque los marcadores moleculares han representado un avance importante, resultan insuficientes para caracterizar completamente la diversidad genética y los factores asociados a la patogenicidad (Falconí et al., 2013). El análisis genómico ofrece una resolución superior para identificar genes asociados a virulencia y adaptación. En Ecuador, Bermeo (2024) analizó genomas de aislados locales de *Colletotrichum* patógenos de tomate de árbol y naranjilla; identificando 11,623 genes codificantes, de los cuales solo seis fueron clasificados como posibles efectores. El alto porcentaje de genes hipotéticos evidencia la necesidad de optimizar los procesos de ensamblaje y anotación para mejorar la clasificación funcional.

El ensamblaje a nivel cromosómico en especies de *Colletotrichum* ha demostrado ser útil para detectar reorganizaciones genómicas asociadas a la evolución y patogenicidad, como se ha documentado en *C. higginsianum* y *C. gloeosporioides* (Dallery et al., 2017; Gan et al., 2021; H. Wang et al., 2023). Este enfoque podría permitir una mejor anotación

genómica de aislados ecuatorianos, la identificación de genes clave y el diseño de estrategias más eficaces para el manejo de la antracnosis.

2.4.4 Genómica de *Colletotrichum*

Los estudios genómicos en especies del género *Colletotrichum* han evidenciado una notable diversidad en el tamaño, la organización cromosómica y el contenido génico, atribuida a la plasticidad evolutiva de este grupo de hongos fitopatógenos (Casarrubias Carillo et al., 2003; Falconí et al., 2013; Gan et al., 2021). En general, los genomas nucleares de especies representativas como *C. higginsianum*, *C. gloeosporioides* y *C. scovillei* presentan tamaños comprendidos entre 45 y 60 Mb, con un contenido de GC cercano al 50-54% (Adhikari et al., 2025; Dallery et al., 2017; Hsieh et al., 2022; Huo et al., 2021). El número de cromosomas suele oscilar entre 11 y 13, dentro de los cuales suelen incluirse minicromosomas de menos de 1 Mb, además de un genoma mitocondrial independiente de entre 30 y 35 Kb (Baroncelli et al., 2021; Becerra et al., 2023; H. Wang et al., 2023).

La anotación de genomas en distintas especies de *Colletotrichum* ha reportado entre 11,000 y 16,000 genes codificantes (Adhikari et al., 2025; Lu et al., 2022; Y. Wang et al., 2023). Entre estos se destacan familias de genes efectores, enzimas degradadoras de pared celular y proteínas involucradas en la detoxificación, que participan en la colonización y adaptación a diferentes hospederos (Lu et al., 2022). Asimismo, la abundancia de elementos transponibles y la ocurrencia de reordenamientos cromosómicos son considerados factores clave en la evolución del género, pues favorecen la rápida diversificación de especies, la adquisición de nuevos rangos de hospederos y la aparición de resistencia a fungicidas (Y. Wang et al., 2023)

2.5 Herramientas bioinformáticas y genómica aplicada a hongos fitopatógenos

2.5.1 Importancia del análisis genómico

La genómica juega un papel fundamental en el estudio de la variabilidad y patogenicidad de los hongos fitopatógenos. Permite identificar la diversidad genética derivada de elementos transponibles y secuencias repetitivas, que contribuyen a la adaptación del hongo y a su capacidad para evadir las defensas del hospedero (García Romero, 2019). También posibilita la caracterización de genes asociados a patogenicidad, como los que intervienen en la formación de apresorios y la producción de toxinas, claves para comprender los mecanismos de virulencia (Rodríguez-López et al., 2009).

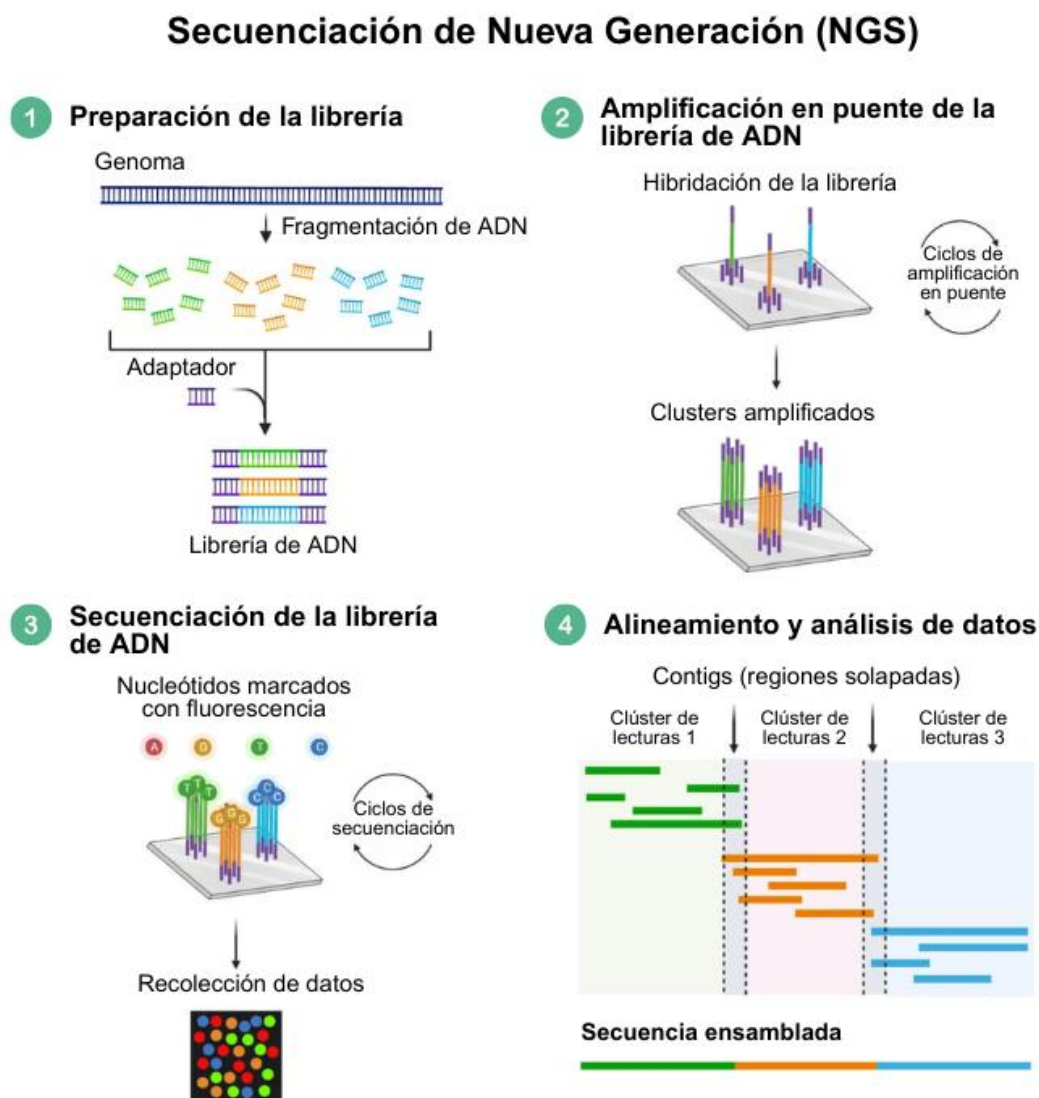
En *Colletotrichum*, este tipo de análisis han facilitado la identificación de genes efectores y reorganizaciones cromosómicas asociadas a la adaptación a distintos hospederos, aportando información fundamental para el desarrollo de estrategias de control (Dallery et al., 2017; Gan et al., 2021; Y. Wang et al., 2023).

2.5.2 Secuenciación de nueva generación (NGS)

La secuenciación genómica ha evolucionado desde el método de Sanger hacia tecnologías de nueva generación (NGS) (Figura 4), que permiten generar grandes volúmenes de datos con mayor rapidez y menor costo. Las plataformas de lectura corta, como *Illumina* e *Ion Torrent*, ofrecen alta precisión, mientras que las de lectura larga, como *Oxford Nanopore* y *PacBio*, facilitan la resolución de regiones repetitivas y estructuralmente complejas (Brenes-Guillén, 2019). En hongos fitopatógenos, la combinación de lecturas cortas y largas ha demostrado mejorar la continuidad y calidad de los ensamblajes, favoreciendo el análisis detallado de su organización genómica (Dallery et al., 2017; Tsushima & Shirasu, 2022).

Figura 4

Esquema General del Flujo de la Secuenciación de Nueva Generación (NGS)



Nota. El proceso incluye: (1) preparación de la librería mediante fragmentación del ADN e incorporación de adaptadores; (2) amplificación en puente de la librería; (3) secuenciación de nucleótidos marcados con fluorescencia durante múltiples ciclos; y (4) análisis bioinformático, en el que las lecturas se alinean y ensamblan en *contigs* para la reconstrucción de la secuencia genómica. Adaptado de “*Next Generation Sequencing (NGS) and its application in genomics*” por Aseem, 2023.

2.5.3 Ensamblaje genómico

El ensamblaje genómico consiste en la reconstrucción de la secuencia de un genoma a partir de fragmentos de ADN obtenidos mediante secuenciación masiva (Brenes-Guillén, 2019). Existen dos estrategias principales: *de novo*, que no requiere información previa, y la basada en referencia, que utiliza un genoma previamente ensamblado como guía (Aguilar-Bultet & Falquet, 2015). A nivel computacional, los ensamblajes se basan en algoritmos como los grafos de Bruijn, empleados en lecturas cortas de Illumina, como SPAdes (Prjibelski et al., 2020) o Velvet (Zerbino & Birney, 2008); o los modelos overlap-layout-consensus (OLC), utilizados en lecturas largas de PacBio y Nanopore, como CANU (Koren et al., 2017) o Flye (Kolmogorov et al., 2019; Pallarés Zazo, 2021; Prjibelski et al., 2020; Rivas & Mcewen, 2011).

2.5.3.1 Construcción de pseudocromosomas.

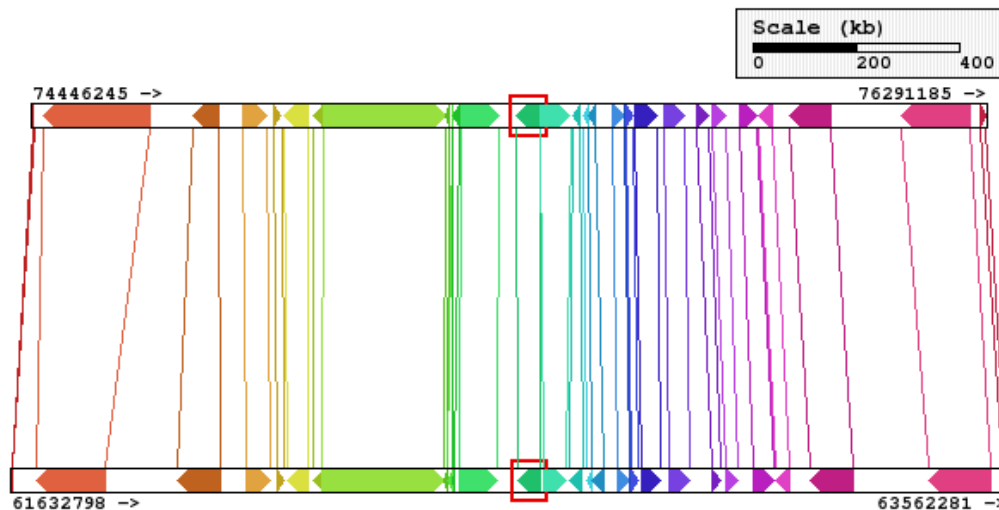
Tradicionalmente, el ensamblaje genómico a partir de lecturas cortas como Illumina, alcanza el nivel de *contigs* o *scaffolds*. Esto se debe a que las lecturas, de tan solo 100-300 pb, no logran resolver adecuadamente regiones altamente repetitivas o estructuralmente complejas del genoma, lo que genera fragmentación y limita la obtención de secuencias continuas a escala cromosómica (Dida & Yi, 2021a; Treangen & Salzberg, 2012).

El salto hacia pseudocromosomas puede lograrse mediante el uso de datos adicionales (ej. Hi-C, mapas ópticos) o por el ordenamiento por sintenia con genomas de referencia cercanos, utilizando herramientas como Ragout o RaGOO (Alonge et al., 2019; Kolmogorov et al., 2018). Los bloques de sintenia (Figura 5) se definen como regiones cromosómicas que conservan un orden común de genes homólogos derivados de un ancestro compartido (D. Liu et al., 2018). Este principio es utilizado por algoritmos de ordenamiento que colocan y orientan *scaffolds* fragmentados en función de la organización cromosómica de un genoma de referencia filogenéticamente cercano (Alonge et al., 2019; Kolmogorov et al., 2014, 2018).

En hongos como *Colletotrichum*, la aplicación de la sintenia resulta útil para construir pseudocromosomas confiables para identificar regiones divergentes asociadas con la patogenicidad y adaptación al hospedero (Gan et al., 2021).

Figura 5

Bloques de sintenia entre cromosomas homólogos de especies cercanas



Nota. Ejemplo general de sintenia entre dos cromosomas de especies distintas que comparten un gen en común (recuadro rojo). Los marcadores homólogos (flechas de colores), conectados por líneas, ilustran regiones conservadas entre genomas, denominados bloques de sintenia, que sirven de base para ordenar *scaffolds* en pseudocromosomas. Tomado de (Sinha & Meller, 2007).

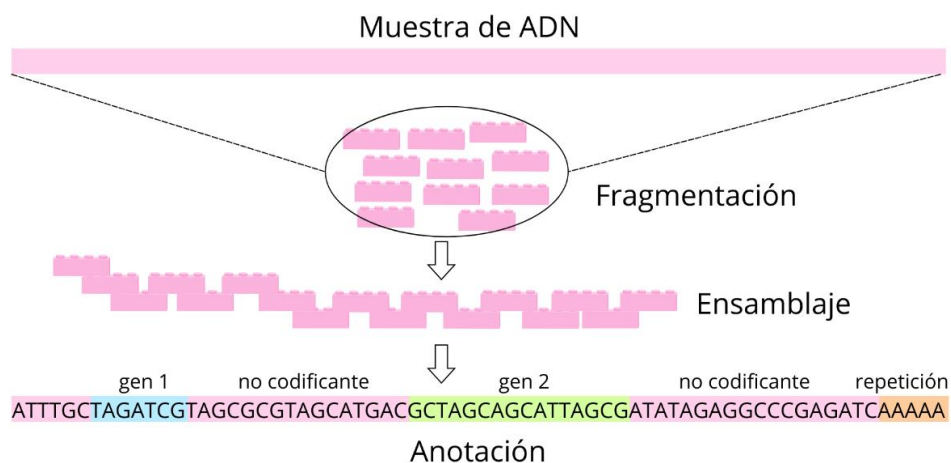
2.5.4 Anotación genómica

La anotación genómica es el proceso de identificar y caracterizar los elementos funcionales de un genoma ensamblado (Figura 6). Se clasifica en anotación estructural, que delimita genes, exones, intrones y regiones reguladoras; y anotación funcional, que asigna posibles funciones biológicas a las secuencias predichas (Aguilar-Bultet & Falquet, 2015; Brenes-Guillén, 2019; Yandell & Ence, 2012).

La anotación estructural combina predicciones *ab initio* mediante modelos estadísticos (ej. AUGUSTUS, GeneMark), junto con alineamiento con proteínas homólogas y datos transcriptómicos (RNA-Seq), que validan y refinan los modelos génicos (Brúna et al., 2021; Yandell & Ence, 2012). La anotación funcional complementa este proceso al asignar roles biológicos a las secuencias identificadas. Para ello se emplea la comparación con bases de datos de proteínas y dominios, como UniProtKB, eggNOG, Pfam e InterPro, que permiten inferir homología y función (Yandell & Ence, 2012). Además, se utilizan clasificaciones jerárquicas como Gene Ontology (GO) para describir procesos biológicos, funciones moleculares y componentes celulares, junto con la integración en rutas metabólicas a través de KEGG, lo que facilita el estudio de la organización funcional y metabólica del genoma (Huerta-Cepas et al., 2019; Kanehisa & Goto, 2000; Thomas, 2017).

Figura 6

Proceso de ensamblaje y anotación de un genoma



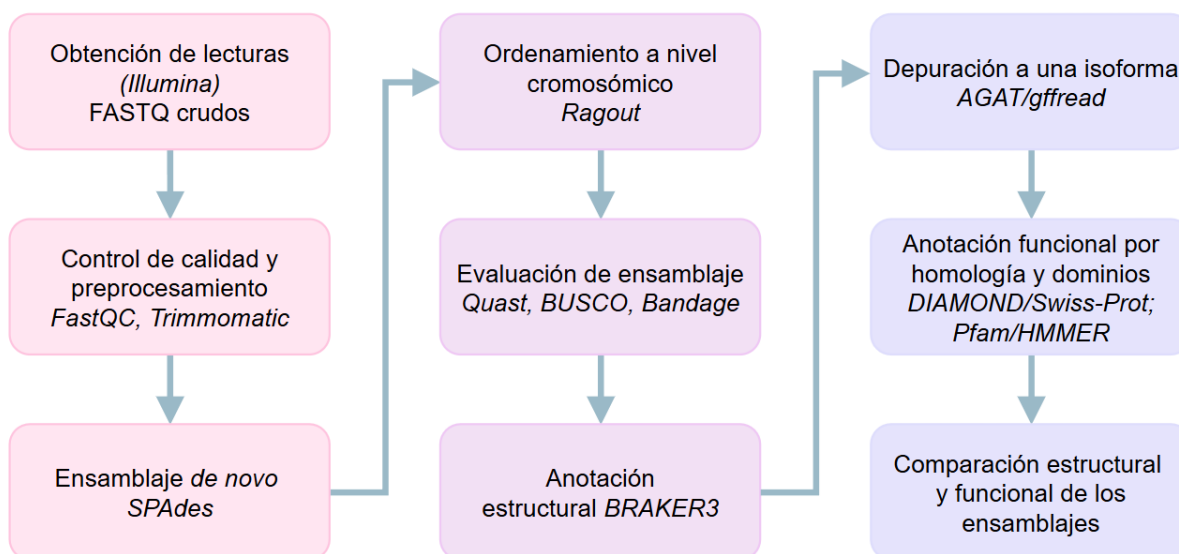
Nota. Representación general del flujo en genómica: el ADN se fragmenta y las lecturas obtenidas se ensamblan en secuencias más largas, que posteriormente son anotadas para identificar genes, regiones no codificantes y repeticiones. Adaptado de “*Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations*” por (Fuentes-Pardo & Ruzzante, 2017).

CAPÍTULO III: METODOLOGÍA

El flujo general de trabajo (Figura 7), partió del control de calidad y filtrado de lecturas *Illumina* (FastQC, Trimmomatic), se generó un ensamblaje *de novo* con SPAdes y luego se ordenaron y orientaron los fragmentos a escala cromosómica mediante Ragout por sintenia con una referencia cercana (*Colletotrichum scovillei*). Se evaluó la continuidad y estructura con QUAST/Bandage y la completitud génica con BUSCO. Se realizó la anotación estructural con BRAKER3 guiada por proteínas de OrthoDB, se depuró a una isoforma por gen y se extrajeron los CDS/proteínas (AGAT/gffread). Finalmente, se efectuó la anotación funcional por homología (DIAMOND/ Swiss-Prot) y por dominios (Pfam/HMMER), y se comparó la calidad estructural y funcional frente a un ensamblaje convencional a nivel de *contigs* reportado en un estudio previo.

Figura 7

Pipeline metodológico empleado en el presente estudio



3.1 Obtención de las secuencias

Las secuencias analizadas corresponden a tres aislados de *Colletotrichum* recolectados en diferentes localidades de la región interandina del Ecuador, en función del cultivo

hospedero: tomate de árbol y/o naranjilla. La secuenciación fue realizada por la empresa *BMKGENE*, utilizando la plataforma de *Illumina*, la cual generó lecturas *paired-end* (*forward* y *reverse*) en formato FASTQ. La información de procedencia y hospedero de los aislados se presentan en la Tabla 1.

Tabla 1

Aislados de Colletotrichum analizados

Código de identificación del aislado	Procedencia	Hospedero infectado
<i>Col_ToNa</i>	GET-Tumbaco	Tomate de árbol y naranjilla
<i>Col_To</i>	Tandapi	Tomate de árbol
<i>Col_Na</i>	Río Verde-Baños	Naranjilla

3.2 Procesamiento de Secuencias Crudas

Las secuencias crudas de los tres aislados de *Colletotrichum* (*Col_ToNa*, *Col_To*, *Col_Na*) fueron importadas en la plataforma KBase (<https://www.kbase.us/>) en formato *non-interleaved paired-end*.

El control de calidad inicial se realizó mediante la herramienta FastQC 0.12.1 (Andrews, 2010), evaluando parámetros como la calidad promedio por base (*Phred score*), contenido GC, la presencia de adaptadores y la presencia de nucleótidos indefinidos (N).

Posteriormente, se aplicó un filtrado con *Trimmomatic v0.39* (Bolger et al., 2014), aplicando parámetros para eliminar bases de baja calidad y lecturas excesivamente cortas, manteniendo la cobertura adecuada para el ensamblaje (*Leading minimum quality: 3, Trailing minimum quality: 3, minimum read length: 75, sliding window: 4:20*). El recorte de adaptadores no fue necesario, ya que el reporte de calidad inicial no evidenció su presencia.

Finalmente, se aplicó nuevamente FastQC sobre las lecturas filtradas, para confirmar la mejora en la calidad y cobertura adecuada para el ensamblaje.

3.3 Ensamblaje a Nivel Cromosómico del Genoma

3.3.1 Ensamblaje de novo con SPAdes

El ensamblaje preliminar se llevó a cabo con *SPAdes Genome Assembler v3.15.5* (Prjibelski et al., 2020) en la plataforma KBase. Este ensamblador, basado en grafos *de Bruijn*, está optimizado para lecturas cortas de Illumina y permite generar *contigs* y *scaffolds* a partir de datos de secuenciación *de novo*. Se obtuvo un borrador genómico en formato FASTA que constituyó el *target* para el ensamblaje asistido por referencia *Ragout*, dado que esta herramienta no trabaja con lecturas crudas, sino con ensamblajes preliminares ya construidos (Kolmogorov et al., 2018).

3.3.2 Selección y preparación de la referencia

Como genoma de referencia, se empleó *Colletotrichum scovillei* ([NCBI RefSeq Assembly: GCF_011075155.1](#)), identificado como especie filogenéticamente cerca a los aislados de interés (*Col_To*, *Col_Na*, *Col_ToNa*) de acuerdo con análisis previos (Bermeo Salazar, 2024). Para garantizar la compatibilidad con los algoritmos de detección de bloques de sintenia, tanto la referencia como el *target* fueron preprocesados mediante:

- Conversión de secuencias a mayúsculas.
- Sustitución de caracteres no estándar (\neq A,C,T,G,N) por "N".
- Simplificación de encabezados FASTA, eliminando espacios y caracteres especiales.

3.3.3 Ensamblaje a Nivel Cromosómico con Ragout

El ensamblaje a nivel cromosómico se realizó con *Ragout v2.3* (Reference-Assisted Genome Ordering UTility) (Kolmogorov et al., 2018), herramienta que ordena y orienta *contigs* en cromosomas mediante grafos de adyacencia de bloques de sintenia entre la referencia y el ensamblaje *target*. El proceso se ejecutó en un entorno Linux a través de la consola de comandos de Visual Code Studio (VCS).

Se ejecutó el modo “legacy recipe” definiendo una referencia (.references = cscov) y el ensamblaje objetivo o *target* (.target = tona). Para la detección de bloques de sintenia, se empleó el *backend Sibelia* (Pham et al., 2013) con tres escalas de resolución (5 kb, 500 pb y 100 pb). Ragout infirió el orden y orientación de los fragmentos, generando scaffolds cromosómicos en formato FASTA (Kolmogorov et al., 2018).

3.3.4 Evaluación del ensamblaje

La evaluación se realizó con un conjunto de herramientas complementarias. Se utilizó QUAST v5.2.0 (Mikheenko et al., 2018) para obtener métricas de continuidad (*N50*, *L50*, *número de scaffolds*, *longitudes totales*, *%GC*, *bases N*), y estadísticas de alineamiento respecto al genoma de referencia (*genome fraction*, *misassemblies*, *mismatches*). Además, se ejecutó BUSCO v5.0.0 (Manni et al., 2021) para evaluar la completitud génica usando el linaje *Sordariomycetes* (sordariomycetes_odb10), ya que el género *Colletotrichum* pertenece a esta clase (Schoch et al., 2020). Por último, se empleó Bandage v0.8.1 (Wick et al., 2015) para la visualización de los grafos de ensamblaje producidos por SPAdes y Ragout.

El procedimiento completo de ensamblaje asistido por referencia y evaluación fue replicado para los tres aislados (*Col_To*, *Col_Na*, *Col_ToNa*).

3.4 Anotación del Genoma

3.4.1 Anotación Estructural del Genoma

La anotación estructural de los genomas ensamblados se realizó con BRAKER3 (Gabriel et al., 2024), un pipeline automatizado que integra evidencias *ab initio* y basadas en homología para la predicción de genes. Este software combina tres componentes principales: GeneMark-ETP, que utiliza modelos probabilísticos entrenados con evidencia externa para la predicción de genes; AUGUSTUS, que refina los modelos y produce anotaciones génicas completas; y ProtHint, que genera pistas de alineamientos proteína-genoma a partir de bases de datos de proteínas de referencia. La integración de estas herramientas permite identificar de

manera precisa exones, intrones y marcos de lectura abiertos en genomas fúngicos. Dado que el proceso es altamente demandante en recursos computacionales, la ejecución se realizó en una supercomputadora.

Como evidencia proteica se utilizó la partición Fungi.fa.gz de OrthoDB v12 (https://bioinf.uni-greifswald.de/bioinf/partitioned_odb12/) (Tegenfeldt et al., 2025), una base de datos de ortólogos curada y ampliamente utilizada para anotación genómica y evaluación de completitud con BUSCO (Manni et al., 2021). Esta partición contiene proteínas representativas de especies fúngicas y proporciona alineamientos de alta calidad que incrementan la sensibilidad y precisión en la detección de genes homólogos (Tegenfeldt et al., 2025).

Como resultado, BRAKER3 generó archivos de anotación en formato GTF (coordenadas génicas y estructura de transcritos), así como secuencias CDS y proteínas en formato FASTA, los cuales fueron empleados para la anotación funcional y análisis comparativo.

A partir de los archivos de predicción generados por BRAKER en formato GTF, se realizó una conversión a formato GFF3 utilizando AGAT (Herramienta de Anotación Genómica). Posteriormente se aplicó el script `agat_sp_keep_longest_isoform.pl` para conservar únicamente la isoforma más larga por locus génico, evitando redundancias en la predicción de transcritos alternativos. Con el archivo resultante se empleó `gffread` para extraer las secuencias de las regiones codificantes (CDS) y sus respectivas traducciones proteicas en formato FASTA. Este procedimiento generó dos conjuntos depurados: un archivo CDS y un archivo de proteínas, no redundantes y representativos de una isoforma por gen.

La calidad y completitud de la anotación estructural se evaluó con BUSCO (Manni et al., 2021), empleando el modo *protein* sobre las secuencias de aminoácidos predichas. Se utilizó el *dataset* `fungi_odb10`, que contiene 758 genes ortólogos altamente conservados, grupo

taxonómico al que pertenece *Colletotrichum*. BUSCO clasificó cada ortólogo en categorías de completo (de copia única o duplicado), fragmentado o ausente, permitiendo estimar la completitud y precisión de la anotación generada por BRAKER3. Este procedimiento se realizó antes y después de la depuración de las anotaciones con AGAT+gffread.

3.4.2 Anotación Funcional del Genoma

La anotación funcional de los conjuntos proteicos no redundantes (~13,000 proteínas por aislado) se abordó mediante dos enfoques complementarios. En primer lugar, se realizó la búsqueda de homologías utilizando DIAMOND BLASTp v2.1.10 (Buchfink et al., 2021) contra la base de datos UniProt/SwissProt (Boutet et al., 2007), empleando el parámetro -k 1 para recuperar únicamente el mejor alineamiento por proteína. Este procedimiento permitió cuantificar el número de secuencias con anotación por similitud y obtener descripciones funcionales a partir de proteínas homólogas caracterizadas en bases de datos curadas.

Para la identificación de dominios conservados se empleó la base de datos Pfam-A (Finn et al., 2013), ampliamente utilizada en la anotación de proteínas fúngicas. El análisis se ejecutó mediante el script pfam_scan.pl en combinación con HMMER v3.4 (Finn et al., 2011), aplicando un nivel de significancia de $e\text{-value} \leq 1e\text{-}5$ tanto a nivel de secuencia como de dominio. Este enfoque permitió detectar dominios y motivos estructurales que aportan información funcional independiente de la anotación por homología.

3.5 Comparación de los Ensamblajes a Nivel Cromosómico vs. Nivel *Contigs*

3.5.1 Comparación Estructural de los Ensamblajes

Los ensamblajes obtenidos por SPAdes y Ragout fueron comparados mediante métricas reportadas por QUAST y BUSCO, visualización de grafos en Bandage y la distribución de longitudes de *contigs* con Compare Assembled Contig Distribution.

En primer lugar, se analizaron los reportes generados por QUASt v5.2.0 (Mikheenko et al., 2018), contrastando las métricas clásicas de ensamblaje: número de *contigs*, longitudes, N50/L50, porcentaje de GC y bases indefinidas “N”; obtenidas tanto en el ensamblaje convencional *de novo* con SPAdes como en el ensamblaje asistido por referencia con Ragout. Esto permitió establecer diferencias en la continuidad y estructuración cromosómica alcanzada entre ambos enfoques.

Posteriormente, se realizó una inspección visual de la organización de los grafos de ensamblaje en Bandage v0.8.1 (Wick et al., 2015). Para ello, se cargaron los archivos FASTA resultantes de cada ensamblaje. Esta comparación cualitativa permitió evidenciar diferencias en la complejidad del grafo y continuidad de *contigs*, proporcionando una validación visual del impacto del reordenamiento cromosómico.

Finalmente, se empleó Compare Assembled *Contig* Distribution v1.1.2 disponible en la plataforma KBase. Esta herramienta permitió comparar de manera gráfica la distribución de longitudes de *contigs* entre ensamblajes, facilitando la identificación de mejoras en continuidad y reducción en el número de fragmentos tras la aplicación de Ragout.

3.5.2 Comparación Funcional

La comparación se planteó para contrastar los resultados funcionales del pipeline empleado en el presente estudio (que partió de un ensamblaje a nivel cromosómico y una anotación estructural/funcional estandarizada), con los resultados obtenidos en un estudio previo realizado por Bermeo Salazar (2024), en el que los aislados fueron ensamblados *de novo* con SPAdes y anotados con *funannotate*. La comparación se diseñó como un modelo descriptivo basado en métricas reportadas por cada trabajo. Toda la información se organizó posteriormente en una tabla de resumen por aislado y por tipo de métrica, acompañada de una descripción de coincidencias y diferencias entre estudios.

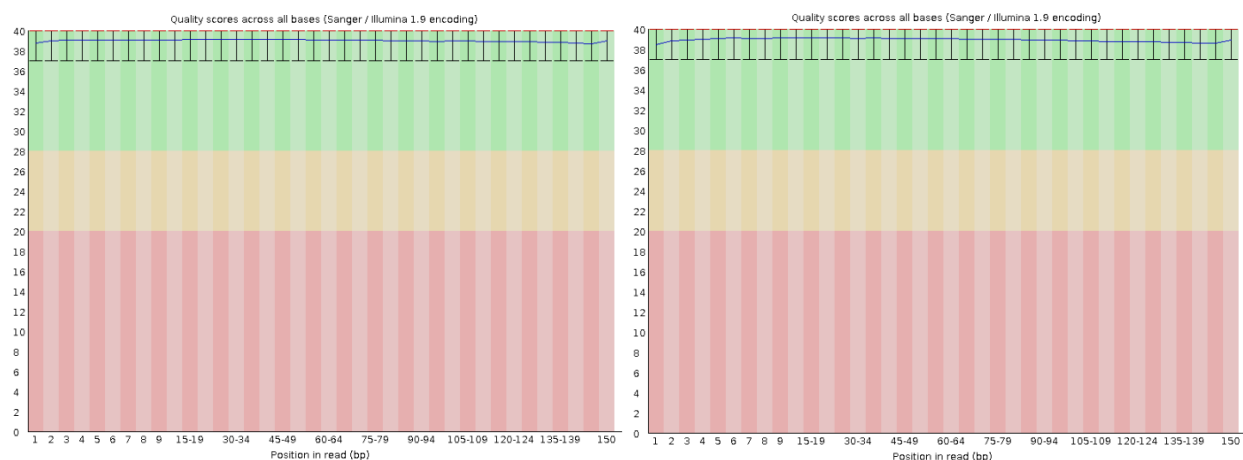
CAPÍTULO IV: RESULTADOS Y DISCUSIÓN

4.1 Procesamiento de Secuencias Crudas

El control de calidad con FastQC mostró lecturas de alta calidad en los tres aislados (*Col_To*, *Col_Na*, *Col_ToNa*), con valores promedio de Phred ~ 39 , sin adaptadores, ni secuencias sobrerrepresentadas y con un contenido de GC estable dentro del rango esperado para hongos como *Colletotrichum* (40-50%) (Adhikari et al., 2025; Huo et al., 2021). Tras el recorte, los cambios fueron mínimos: se observó una ligera reducción en la longitud de algunos extremos (ajuste del rango 100-150 pb a 75-150 pb) sin deterioro de la calidad por base ni del GC, lo que confirma que el preprocesamiento no comprometió la información útil para en ensamblaje (Figura 8).

Figura 8

Gráficos de calidad de lecturas después del filtrado



Nota. Resultados de la calidad de las lecturas para las secuencias *Forward* y *Reverse* respectivamente del aislado *Col_ToNa*. Phred Score ~ 39 . Gráficas generadas por FastQC.

La Tabla 2 resume el leve descenso en bases totales (2.9-3.0 \rightarrow 2.8—2.9 Gpb) tras el filtrado, manteniéndose Phred ≈ 39 y GC ≈ 49 -50%. Esto indica que el *trimming* removió fragmentos deteriorados sin afectar la calidad global ni la cobertura efectiva, condición clave

para la construcción de grafos de ensamblaje robustos que sirven de base tanto para ensamblajes *de novo* como para los basados en referencia (Bolger et al., 2014).

Tabla 2

Resumen por aislado del control de calidad y filtrado de lecturas

Aislado	Total de bases crudas (Gpb)	Total de bases filtradas (Gpb)	GC (%)	Calidad media (Phred Score)
<i>Col_ToNa</i>	3.0	2.9	50	~39
<i>Col_To</i>	2.9	2.8	49	~39
<i>Col_Na</i>	3.0	2.8	49	~39

Nota. Los valores reportados en la tabla son el promedio de los datos de las lecturas *forward* y *reverse* para cada aislado. Los valores de GC y Phred se mantuvieron constantes antes y después del filtrado.

4.2 Ensamblaje a Nivel Cromosómico del Genoma

4.2.1 Ensamblaje *de novo* con SPAdes

El ensamblaje inicial con SPAdes presentó tamaños genómicos de 52.4-54.2 Mb, acordes con lo descrito para especies de *Colletotrichum* (O'Connell et al., 2012), aunque con un nivel elevado de fragmentación: *Col_ToNa* 1,170 *contigs* (N50=0.93; L50=19), *Col_To* 430 *contigs* (N50=0.44; L50=39) y *Col_Na* 1,027 *contigs* (N50=0,46; L50=36) (Tabla 3). A pesar de que el %GC (~51%) y el número de N por 100 kb (3.6-6.7) fueron bajos, la fragmentación limita la reconstrucción de genes largos y la evaluación de la sintenia, lo que hace necesario optimizar el ensamblaje mediante estrategias de asistencia por referencia (Dida & Yi, 2021b).

Tabla 3

Métricas de ensamblaje de novo (SPAdes) por aislado

Aislado	# <i>Contigs</i>	<i>Contig</i> más largo	Longitud total	N50	L50	GC (%)	N / 100 kpb
<i>Col_ToNa</i>	1170	2,246,602	52,697,920	929,323	19	50.92	5.89
<i>Col_To</i>	430	1,262,157	52,414,135	444,786	39	51.02	3.75
<i>Col_Na</i>	1027	1,807,121	54,179,120	464,458	36	50.89	6.7

Nota. Los datos fueron obtenidos de los reportes de QUAST.

4.2.2 Ensamblaje a Nivel Cromosómico con Ragout

El ensamblaje preliminar *de novo* con SPAdes se ordenó y orientó con Ragout, generando *scaffolds* de escala cromosómica a partir de bloques de sintenia con una referencia cercana. Ragout implementa un enfoque de reordenamiento genómico que aprovecha múltiples referencias, y deduce el orden y la orientación de los fragmentos para producir ensamblajes a nivel de cromosomas (Kolmogorov et al., 2014, 2018).

En este caso, el ensamblaje con Ragout consolidó el genoma en 14-16 *scaffolds* por aislado (Figuras 9-11), con *scaffolds* máximos de 6.5-6.6 Mb, N50 de 4.5-5 Mb y L50 = 5 (Tabla 4). Estos valores indican que la mitad del genoma queda contenida en solo cinco fragmentos, un patrón consistente con ensamblajes a escala cromosómica (Dallery et al., 2017; Kolmogorov et al., 2018). Además, menos del 2% de la secuencia total quedó en fragmentos no ubicados (<1Mb), lo que refleja que prácticamente todo el genoma fue organizado en *megascaffolds*.

Figura 9

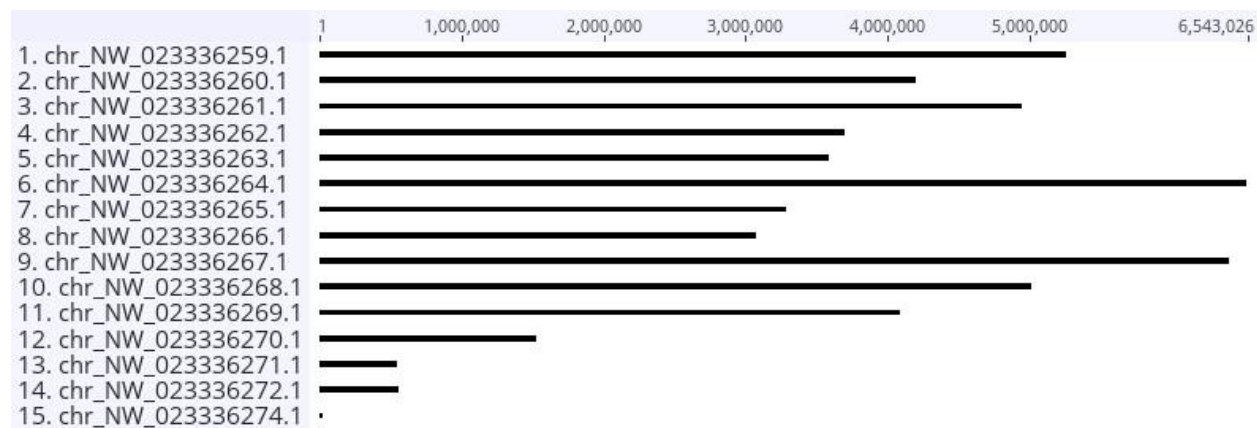
Representación visual de los *scaffolds* ensamblados de Col_ToNa



Nota. Gráfico obtenido del Software Geneious Prime.

Figura 10

Representación visual de los scaffolds ensamblados de Col_To



Nota. Gráfico obtenido del Software Geneious Prime.

Figura 11

Representación visual de los scaffolds ensamblados de Col_Na



Nota. Gráfico obtenido del Software Geneious Prime.

4.2.3 Evaluación del Ensamblaje

Para la valoración estructural se empleó QUASt, que resume contigüidad (N50/L50), tamaño y composición (longitud total, %GC) y *gaps* (N por 100 kpb) (Mikheenko et al., 2018). Las métricas reportadas para el ensamblaje asistido por referencia (Tabla 4) muestran valores

de N50/L50 y *scaffold* máximo consistentes con ensamblajes cromosómicos (Dallery et al., 2017; Kolmogorov et al., 2018; Mikheenko et al., 2018). El incremento de N por 100 kpb refleja *gaps* generados por el *scaffolding* y no necesariamente pérdidas de información. Esto debido a que, las métricas basadas en alineamiento, calculan solo sobre las regiones alineadas a la referencia y dependen de su cercanía filogenética (Mikheenko et al., 2018).

Tabla 4

Métricas de ensamblaje asistido por referencia (Ragout) por aislado

Aislado	# Scaffolds	Scaffold más largo	Longitud total	N50	L50	GC (%)	N/ 100 kpb
<i>Col_ToNa</i>	14	6,571,457	52,129,356	4,559,604	5	50.89	694.98
<i>Col_To</i>	15	6,543,026	52,883,064	4,953,199	5	51.14	2206.49
<i>Col_Na</i>	16	6,519,681	53,865,182	5,000,376	5	51.09	2897.77

Nota. Los datos fueron obtenidos de los reportes de QUAST de los ensamblajes cromosómicos con Ragout contra la secuencia de referencia *C. scovillei*.

Tabla 5

Contigs no anclados y N's incorporadas después de Ragout

Aislado	Genoma alineado (%)	Contigs no anclados	Longitud no anclada (pb)	N's introducidas	N's por evento
<i>Col_ToNa</i>	99.97	17	175,099	360,118	2,058
<i>Col_To</i>	99.98	20	450,617	1,164,924	4,289
<i>Col_Na</i>	99.95	90	988,718	1,557,658	4,955

Nota. Los datos fueron obtenidos de los reportes de QUAST de los ensamblajes *de novo* con SPAdes contra los ensamblajes cromosómicos con Ragout.

Los indicadores basados en alineamiento de QUAST muestran que los tres ensamblajes muestran una fracción del genoma alineada de 19.70-20.35%, con 0 reensamblajes erróneos y un radio de duplicación de 1.00 en todos los casos, lo que indica una estructura general coherente respecto a la referencia. La baja fracción de alineamiento no implica un ensamblaje incompleto, sino que refleja la divergencia interespecífica con la

referencia utilizada (*C. scovillei*), dado que QUAST se basa en alineamientos nucleotídicos estrictos, solo las regiones conservadas se contabilizan (Mikheenko et al., 2018). Esto se contrasta con los registros de Ragout, que evidenciaron una cobertura de bloques de sintenia cercana al 80%, suficiente para guiar el reordenamiento de pseudocromosomas (Kolmogorov et al., 2018).

La Tabla 5 cuantifica el contenido del ensamblaje *de novo* que no fue incorporado en los pseudocromosomas y el contenido de N's introducido durante el ensamblaje con Ragout. En los tres aislados, los fragmentos no anclados fueron escasos (17-90 contigs) y su longitud acumulada fue <2% del genoma, mientras que las N's introducidas variaron entre 360,118 y 1,557,658, valores esperados en un ordenamiento por bloques de sintenia (Kolmogorov et al., 2018). Además, la métrica de N's por evento (total de N's dividido para el número de gaps) resume el tamaño promedio de las brechas: ~2.06 kb en *Col_ToNa* y ~4.29-4.96 kb en *Col_To* y *Col_Na*. Estos resultados indican que Ragout integró la mayor parte del contenido de SPAdes y que las discrepancias se limitan a regiones difíciles (como *gaps*), sin comprometer la estructura general del genoma reportada en la Tabla 4.

En las regiones alineadas, las tasas por base tienen un valor aproximado de 3,800 mismatches/100 kpb y 390 indels/100 kpb. El bloque de alineamiento más largo tiene un valor medio de 122 kb y la longitud total alineada de 10.4 Mb. Estas cifras describen el grado de alienabilidad frente a la referencia elegida y dependen de su proximidad filogenética y del contenido repetitivo compartido (Mikheenko et al., 2018).

La completitud génica evaluada con BUSCO (dataset Sordariomycetes_odb10, n=3,817) refuerza lo observado en las métricas estructurales: un ensamblaje contiguo y completo. Los tres ensamblajes presentan una completitud muy alta (C = 99.6-99.7%), con casi todos los ortólogos en copia única (S = 99.5-99.6%) y duplicación mínima (D = 0.1%). Los fragmentados (F = 0.1%) y faltantes son mínimos (M = 0.2-0.3%) (Tabla 6). Este patrón (C/S

altos + D/F/M bajos) es el esperado en ensamblajes de calidad cromosómica y favorecen una anotación fiable debido a una menor cantidad de genes partidos (Manni et al., 2021).

Tabla 6

Evaluación de completitud génica mediante BUSCO

Aislado	Completo (C %)	De copia única (S%)	Duplicados (D%)	Fragmentados (F%)	Faltantes (M%)
<i>Col_ToNa</i>	99.7	99.6	0.1	0.1	0.3
<i>Col_To</i>	99.6	99.5	0.1	0.1	0.3
<i>Col_Na</i>	99.6	99.5	0.1	0.1	0.3

Nota. Los datos fueron obtenidos de los reportes de BUSCO.

Los resultados sostienen que los ensamblajes obtenidos corresponden a borradores de calidad cromosómica (pseudocromosómicos). El número de *scaffolds* (14-16) coinciden con el rango de cromosomas descrito para especies del género *Colletotrichum* (Dallery et al., 2017; O'Connell et al., 2012). Aunque la fracción de alineamiento nucleotídico frente a *C. scovillei* fue relativamente baja (~20%), esto refleja divergencia interespecífica más que deficiencias del ensamblaje. De hecho, los reportes de Ragout muestran que cerca del 80% del genoma fue cubierto por bloques de sintenia, suficiente para guiar un reordenamiento robusto. Por lo tanto, aún sin alcanzar resolución telómero a telómero, los resultados capturan la organización cromosómica global y constituyen una base confiable para la posterior anotación.

4.3 Anotación del Genoma

4.3.1 Anotación Estructural del Genoma

Predicción inicial (BRAKER3) (Tabla 7). BRAKER3 predijo entre 13,702 y 13,877 genes por aislado, valores que se encuentran dentro del rango reportado para especies de *Colletotrichum* (12,000-15,000 genes) (Dallery et al., 2017; Gan et al., 2021; O'Connell et al., 2012). Se obtuvieron ~26,400-26,700 transcritos (~1.9 isoformas por gen). El número de proteínas fue equivalente al número de transcritos, lo que indica que no hubo pérdida en la

traducción de ORFs. La longitud promedio de las proteínas fue de ~505 aminoácidos, valor consistente entre aislados y congruente con lo descrito en otros genomas del género. Es

Tabla 7

Resumen de la anotación estructural y el efecto del filtrado de isoformas

Aislado	Genes predichos	Proteínas/ Transcritos (antes)	Proteínas/ Transcritos (después)	Longitud media proteína (aa)
<i>Col_ToNa</i>	13,702	26,486	13,702	505.08
<i>Col_To</i>	13,802	26,711	13,802	504.61
<i>Col_Na</i>	13,877	26,757	13,877	505.01

Nota. “antes/después” se refiere al proceso de depuración de isoformas manteniendo una isoforma por locus. Tras el filtrado: # transcritos = # proteínas = # genes.

Complejidad inicial (BUSCO) (Tabla 8). La evaluación de la complejidad de las anotaciones estructurales mediante BUSCO (dataset fungi_odb10; 758 ortólogos universales), evidenció valores de 99.9-100% en los tres aislados. Este nivel de complejidad es consistente al reportado en otros genomas de referencia del género, como *C. higginsianum* y *C. graminicola* (Gan et al., 2021; H. Wang et al., 2023). En cuanto al tipo de ortólogos detectados, se observó un alto porcentaje de genes duplicados (92.9-93.9%), y un bajo número de ortólogos en copia única (6.1-7.0%). Esto puede deberse, por un lado, a redundancias introducidas por BRAKER3, dado que la herramienta tiende a predecir múltiples isoformas por gen (Brúna et al., 2021; Gabriel et al., 2024).

Tabla 8

Complejidad de la anotación estructural con BUSCO antes de depurar isoformas

Aislado	Completo (C %)	De copia única (S%)	Duplicados (D%)	Fragmentados (F%)	Faltantes (M%)
<i>Col_ToNa</i>	99.9	6.6	93.3	0.1	0.0
<i>Col_To</i>	100.0	6.1	93.9	0.0	0.0
<i>Col_Na</i>	99.9	7.0	92.9	0.0	0.1

Nota. Los datos fueron obtenidos de los reportes de BUSCO.

Efecto del filtrado de isoformas. Tras la depuración de las isoformas, el número de genes se mantuvo constante ($Col_ToNa=13,702$; $Col_To=13,802$; $Col_Na=13,877$), mientras que los transcritos y proteínas se redujeron a uno por gen, quedando iguales al número de loci. Antes del filtrado, BRAKER predijo ~ 1.93 isoformas por gen, después del filtrado, una isoforma. Los conjuntos de proteínas mantuvieron una completitud superior al 99.9%, con una drástica reducción de duplicados (0.3%, Tabla 9), confirmando que las duplicaciones iniciales correspondían a isoformas redundantes generadas por BRAKER3.

Tabla 9

Completitud de la anotación estructural con BUSCO tras depurar isoformas

Aislado	Completo (C %)	De copia única (S%)	Duplicados (D%)	Fragmentados (F%)	Faltantes (M%)
<i>Col_ToNa</i>	99.9	99.6	0.3	0.1	0.0
<i>Col_To</i>	100.0	99.7	0.3	0.0	0.0
<i>Col_Na</i>	99.9	99.6	0.3	0.0	0.1

Nota. Los datos fueron obtenidos de los reportes de BUSCO.

Consideración sobre el enmascaramiento. En este estudio no aplicamos enmascaramiento previo. En *Colletotrichum*, el contenido repetitivo suele ser bajo a moderado ($\sim 4\%$ para *C. scovillei* y $\sim 7\%$ en *C. higginsianum*), por lo que, sumado al alto valor de completitud obtenido con BUSCO, se infirió que el omitir el enmascaramiento no afectó la calidad de la anotación (Dallery et al., 2017; H. Wang et al., 2023).

4.3.2 Anotación Funcional del Genoma

Anotación por homología con DIAMOND (UniProt/Swiss-Prot).

Se comparó el conjunto proteico no redundante (la isoforma más larga por gen) de cada aislado contra UniProtKB/Swiss-Prot mediante DIAMOND en modo top-hit. En los tres genomas, $\sim 59\%$ de proteínas tuvieron al menos un hit curado, con valores muy similares entre aislados (Tabla 10). Este rendimiento es consistente con el uso de Swiss-Prot como referencia

confiable y curada, y con la sensibilidad de DIAMOND para asignaciones rápidas y exactas a nivel de proteína (Bateman et al., 2023; Buchfink et al., 2014).

La fracción restante (~41%) careció de homología detectable a nivel de proteína en Swiss-Prot, lo cual resulta esperado en hongos fitopatógenos por la presencia de proteínas específicas de linaje y secuencias muy divergentes. En estos casos, la evidencia por dominios (por ejemplo, Pfam con HMMER) es una opción viable para inferir función a partir de firmas conservadas (Eddy, 2011; Mistry et al., 2021).

A nivel funcional, los tres conjuntos proteicos comparten un núcleo metabólico y de interacción hospedero-patógeno dominado por enzimas redox del metabolismo secundario (monooxigenasas citocromo P450, oxidorreductasas y deshidrogenadas) (Shin et al., 2018); transferasas (metiltransferasas y glicosiltransferasas), que modifican metabolitos propios y del hospedero; transportadores (superfamilias MFS y ABC), vinculados con el intercambio nutrimental y la tolerancia a compuestos del medio (Costa et al., 2014; Perlin et al., 2014); hidrolasas/proteasas asociadas a remodelación de la pared celular y colonización; y componentes de señalización (quinasas) (Fu et al., 2022). Adicionalmente, se detectó un subconjunto de proteínas HET (incompatibilidad heterocariótica) asociado con compatibilidad vegetativa (Paoletti, 2016).

La abundancia de P450 y enzimas redox concuerda con la biología hemibiótrofa de *Colletotrichum*, que requiere ajustar finamente el metabolismo secundario para neutralizar defensas del hospedero y producir metabolitos durante la infección. Se han descrito P450 con roles en virulencia, desarrollo y degradación de xenobióticos en hongos filamentosos (Gan et al., 2021; O'Connell et al., 2012; Shin et al., 2018). Los transportadores MFS y ABC se alinean con estrategias de adquisición de recursos y tolerancia de fitoquímicos y antimicrobianos, funciones implicadas en la interacción planta-patógeno (Costa et al., 2014; Perlin et al., 2014). Las hidrolasas/proteasas respaldan la capacidad de acceder a polímeros de la pared vegetal y

modular la inmunidad del hospedero, mientras que las quinasas regulan la diferenciación infecciosa y la patogenicidad en *Colletotrichum* (Bradley et al., 2022; Fu et al., 2022; Jashni et al., 2015). Finalmente, la presencia de proteínas HET sugiere mecanismos de compatibilidad vegetativa con impacto potencial en la estructura poblacional y la estabilidad de linajes (Daskalov, 2023; Paoletti, 2016).

Tabla 10

Anotación funcional por homología (DIAMOND/Swiss-Prot) por aislado

Aislado	Proteínas analizadas	UniProtKB/Swiss-Prot	Proteínas anotadas (%)	Principales funciones/hits
<i>Col_ToNa</i>	13,702	8082	59.0	Metiltransferasas, proteínas HET (compatibilidad vegetativa),
<i>Col_To</i>	13,802	8105	58.7	transportadores MFS, enzimas de biosíntesis secundaria, factores de transcripción y enzimas regulatorias
<i>Col_Na</i>	13,877	8130	58.6	

Nota. Proteínas anotadas por DIAMOND/Swiss-Prot, usando la isoforma más larga por gen. Se reportan proteínas con ≥ 1 hit (coincidencias).

Anotación por dominios Pfam con HMMER.

Se identificaron dominios conservados en los tres conjuntos proteicos usando Pfam-A mediante pfam_scan.pl y HMMER. Este enfoque, basado en perfiles HMM curados, es ampliamente aceptado para inferir función a partir de dominios y motivos conservados incluso cuando no hay homología clara a nivel de proteína completa. En los tres aislados, ~70% de las proteínas presentó ≥ 1 dominio Pfam y se obtuvieron ~14,000 asignaciones por aislado, evidenciando una cobertura uniforme por dominios (Tabla 11).

Las familias Pfam dominantes son consistentes entre aislados y definen un núcleo funcional conservado para la biología del patógeno: transporte (MFS_1, Sugar_tr, ABC_tran), regulación transcripcional específica de hongos (Zn_clus, Fungal_trans, Fungal_trans2),

señalización (Pkinase), y alorreconocimiento/compatibilidad vegetativa (HET) (Figuras 12-14). Este perfil sugiere funciones comunes para la adquisición de nutrientes, manejo de compuestos del hospedero y el control fino de la expresión génica durante la infección (MacPherson et al., 2006; Paoletti, 2016; Perlin et al., 2014; Shin et al., 2018).

Tabla 11

Anotación funcional por dominios (Pfam/HMMER) por aislado

Aislado	Genes predichos	Proteínas con ≥ 1 Pfam	Cobertura Pfam (%)	Pfam hits
<i>Col_ToNa</i>	13,702	9,608	70.1	14,179
<i>Col_To</i>	13,802	9,637	69.8	14,204
<i>Col_Na</i>	13,877	9,697	69.9	14,308

Nota. Anotación por dominios pfam_scan.pl/HMMER contra Pfam-A (isoforma más larga por gen). Se reportan proteínas con ≥ 1 dominio y el total de dominios detectados (Pfam hits).

Figura 12

Familias Pfam más abundantes en Col_ToNa

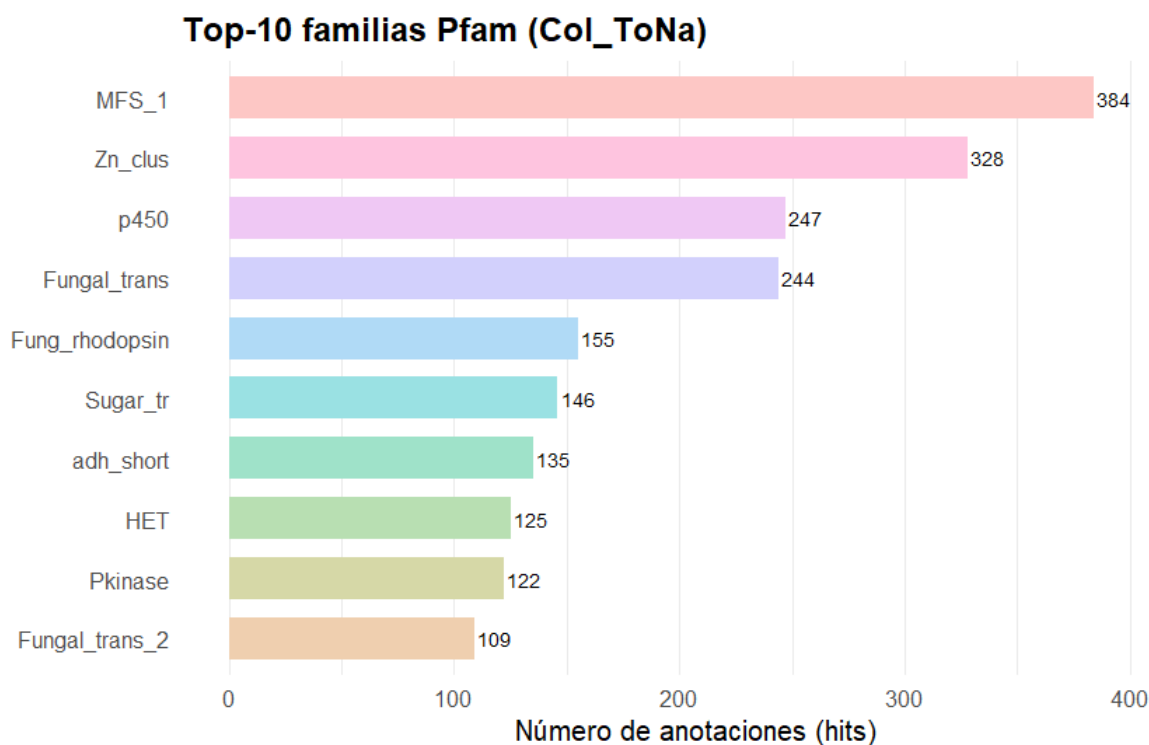
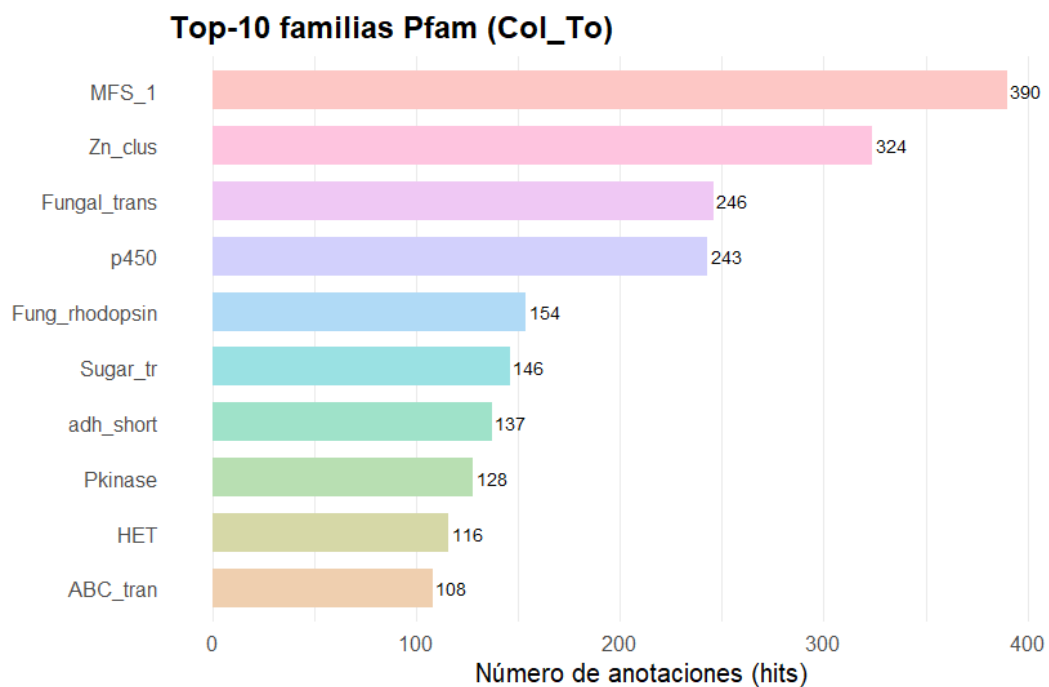
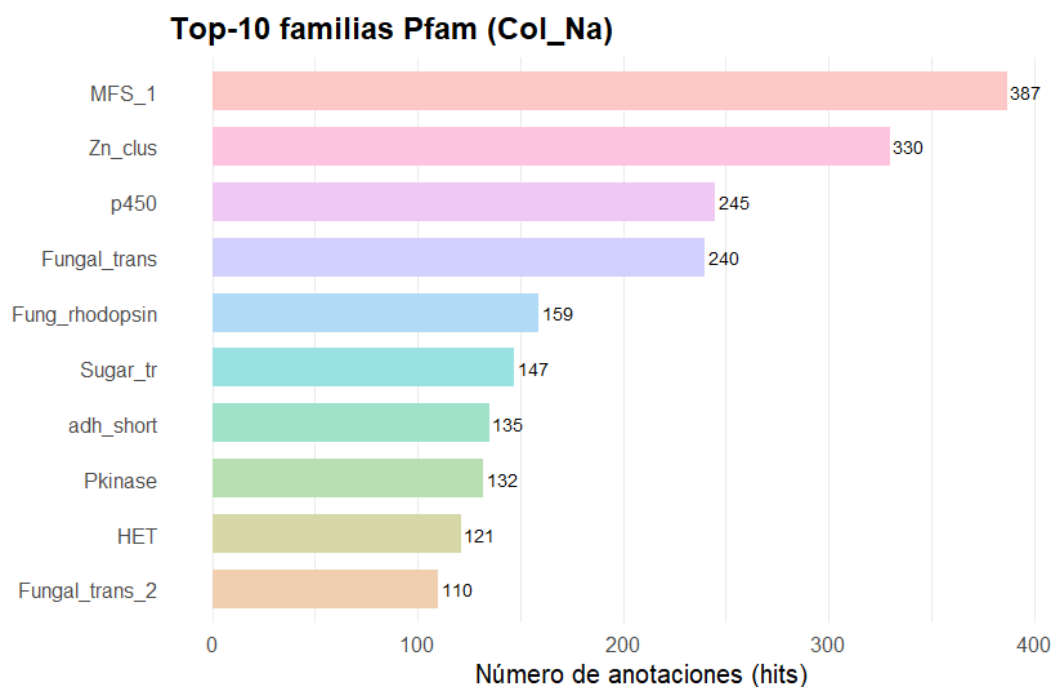


Figura 13*Familias Pfam más abundantes en Col_To***Figura 14***Familias Pfam más abundantes en Col_Na*

Transportadores (MFS_1, Sugar_tr, ABC_tran). Los MFS y ABC son las superfamilias de transporte con mayor presencia en hongos, median la importación de nutrientes y eflujo de xenobióticos/fitoquímicos, funciones clave para tolerar la defensa del hospedero. Su abundancia es común en fitopatógenos y se asocia a virulencia y supervivencia en ambientes hostiles (Costa et al., 2014; Perlin et al., 2014).

Regulación transcripcional fúngica (Zn_clus, Fungal_trans, Fungal_trans2). Estos factores son exclusivos en hongos y controlan numerosos clústeres biosintéticos y vías metabólicas. Su representación alta concuerda con la necesidad de modular dinámicamente el metabolismo secundario y la expresión de efectores a lo largo del ciclo infeccioso (X. Liu et al., 2023; MacPherson et al., 2006; Zhang et al., 2023).

Metabolismo secundario y redox (p450, adh_short). Las monooxigenasas citocromo P450 y enzimas SDR (adh_short) participan en biosíntesis y modificación de metabolitos, homeostasis redox y degradación de compuestos del hospedero. Varias de estas enzimas se han visto implicadas directamente en virulencia y adaptación ecológica (Chen et al., 2014; Shin et al., 2018).

Señalización (Pkinase). La expansión de quinasas concuerda con el control de diferenciación infecciosa (como la formación de apresorios) y de la patogenicidad en *Colletotrichum*. Homólogos de la quinasa PMK1 han demostrado ser reguladores centrales en la virulencia de distintas especies del género (Jiang et al., 2022; Lian et al., 2019; Takano et al., 2007).

Compatibilidad vegetativa (HET). Los dominios HET median incompatibilidad vegetativa (alorreconocimiento) y pueden limitar la fusión de hifas entre genotipos no compatibles, con efectos sobre estructura poblacional, flujo citoplasmático y estabilidad de linajes (Daskalov, 2023; Paoletti, 2016; Paoletti & Clavé, 2007).

El conjunto de dominios predominantes (MFS_1, Sugar_tr, ABC_tran, P450 y SDR) indica funciones de intercambio con el hospedero y manejo de compuestos bioactivos. Una mayor presencia de MFS_1/Sugar_tr y transportadores ABC se asocia con mayor capacidad de captura de nutrientes y eflujo de moléculas tóxicas (xenobióticos y fitoquímicos), rasgos vinculados a virulencia y tolerancia a estrés químico (Costa et al., 2014; Perlin et al., 2014). A su vez, más P450/SDR sugiere mayor potencial de detoxificación y biosíntesis/modificación de metabolitos secundarios, asociados a mecanismos de respuesta a ROS durante la infección (Chen et al., 2014; Shin et al., 2018).

En cuanto a la regulación y señalización (Zn_clus, Fungal_trans, Pkinase), la abundancia observada respalda un ajuste adaptativo de la transcripción en clústeres biosintéticos y efectores, y un control de vías MAPK asociadas a diferenciación infecciosa (Jiang et al., 2022; X. Liu et al., 2023; MacPherson et al., 2006; Takano et al., 2007; Zhang et al., 2023). Este perfil resulta acorde al modo hemibiotrófico del género, que alterna una fase biotrófica (crecimiento en células vivas) y una fase necrotrofa (degradación de tejidos), donde se requiere neutralizar fitoquímicos, reciclar ROS y ajustar el intercambio de metabolitos a lo largo del ciclo infectivo (Gan et al., 2013; O'Connell et al., 2012).

El perfil descrito es coherente con lo descrito para *Colletotrichum* y *Sordariomycetes* fitopatógenos cercanos, por lo que lo interpretamos como señal biológica y no como ruido de base porque se produce en los tres aislados, aparece con dos enfoques independientes de anotación (Pfam/HMMER y DIAMOND/Swiss-Port), además, involucra familias funcionalmente congruentes para el modo hemibiotrófico del género (Gan et al., 2013; O'Connell et al., 2012; Perlin et al., 2014).

4.4 Comparación de los Ensamblajes

4.4.1 Comparación Estructural de los Ensamblajes

La comparación entre el ensamblaje *de novo* (SPAdes) y el ensamblaje asistido por referencia (Ragout) evidencia mejoras sustanciales en la continuidad y organización a escala cromosómica de los tres aislados de *Colletotrichum*. Ragout reduce drásticamente el número de fragmentos y concentra la longitud genómica en pocos *scaffolds* multimegabase, mientras que SPAdes mantiene cientos a miles de *contigs* de menor tamaño.

La comparación de métricas (Tabla 12) muestra esta tendencia: en *Col_ToNa*, Ragout alcanzó un N50 de 4.55 Mb y un L50 de 5, mientras que SPAdes presentó un N50 de 0.92 Mb y un L50 de 19; es decir, la mitad del genoma quedó contenida en solo cinco *scaffolds*, en contraste con los diecinueve necesarios en SPAdes. Así mismo, Ragout presentó un fragmento más largo de 6.57 Mb, aproximadamente el triple del valor de 2.24 Mb obtenido con SPAdes. Resultados similares se observaron en *Col_To* y *Col_Na*, donde Ragout incrementó el N50 hasta ~5 Mb y redujo el L50 a 5, frente a valores inferiores a 0.5 Mb y L50 de 36-39 en SPAdes, lo que confirma la superioridad estructural del ensamblaje cromosómico en los tres aislados.

El contenido de N aumenta tras Ragout (de 0.006 – 0.007% en SPAdes a 0.69-2.90% en Ragout). Este incremento es el comportamiento esperado del *scaffolding*: en los estándares de NCBI/AGP, las N representan brechas explícitas cuando el orden y orientación están resueltos pero la secuencia intermedia aún es desconocida (NCBI, 2021). Además, herramientas de *scaffolding* cromosómico como Ragout, unen *contigs* y colocan cadenas de N entre ellos para identificar secuencias faltantes, sin alterar la secuencia conocida o “inventar” bases (Kolmogorov et al., 2014). Por lo tanto, un mayor % de N no supone pérdidas de información, sino zonas a cerrar con herramientas complementarias que permitan resolver los *gaps* y acercar el número de *scaffolds* al recuento cromosómico (Elbers et al., 2019).

Tabla 12

Comparación de métricas de ensamblaje (SPAdes vs. Ragout) entre aislados

Aislado	Ensamblaje	# Contigs/ Scaffolds	Longitud total (Mb)	Contig más largo (Mb)	N50 (Mb)	L50	Gaps (%)	GC (%)
Col_ToNa	SPAdes	1,170	52.69	2.25	0.92	19	0.006	50.92
	Ragout	14	52.13	6.57	4.55	5	0.69	50.89
Col_To	SPAdes	430	52.41	2.25	0.44	39	0.004	51.02
	Ragout	15	52.88	6.54	4.95	5	2.21	51.14
Col_Na	SPAdes	1,027	54.17	2.25	0.46	36	0.007	50.89
	Ragout	16	53.87	6.52	5.00	5	2.90	51.09

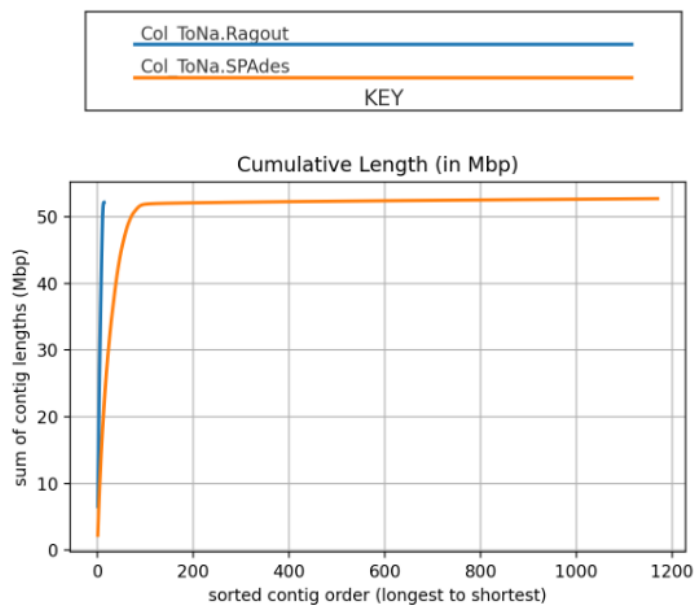
Nota. Comparación resumida de métricas de ensamblaje obtenidas con SPAdes (ensamblaje *de novo*) y Ragout (ensamblaje asistido por referencia). Los datos fueron obtenidos de los reportes de QUAST.

Las curvas de suma acumulada de longitudes obtenidas en Compare Assembled *Contig* Distribution (Figura 15), evidencian que Ragout (línea azul), alcanza la longitud total del genoma con muy pocos fragmentos, mientras que SPAdes (línea naranja) requiere centenares para aproximarse a la misma longitud. En Bandage, los grafos pasan de distribuciones dispersas con SPAdes (Figura 16) a trayectorias lineales y continuas con Ragout (Figura 17), en concordancia con la mejora en N50/L50. Los histogramas de longitudes (>10 Kb) (Figura 18) confirman el desplazamiento hacia *scaffolds* multimegabase tras del ordenamiento

Estos resultados confirman Ragout permitió reorganiza el ensamblaje preliminar de SPAdes en un ensamblaje a nivel cromosómico, reduce la fragmentación y mejora la continuidad del genoma en los tres aislados. El aumento de N es intrínseco al proceso de *scaffolding* y no implica pérdida de contenido, más bien, delimita brechas que pueden abordarse en futuros estudios.

Figura 15

Suma acumulada de las longitudes de los contigs por ensamblaje de Col_ToNa



Nota. Gráfica obtenida del análisis con Compare Assembled *Contig* Distribution en KBase.

Figura 16

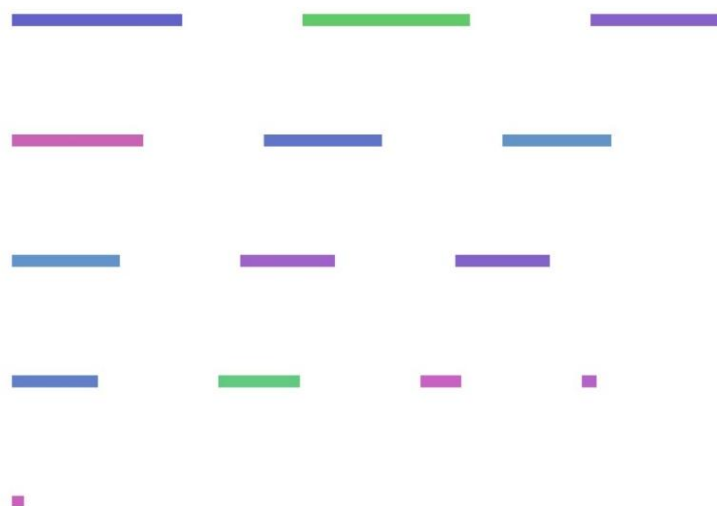
Grafo de ensamblaje de novo con SPAdes del aislado Col_ToNa



Nota. Gráfica obtenida del software Bandage.

Figura 17

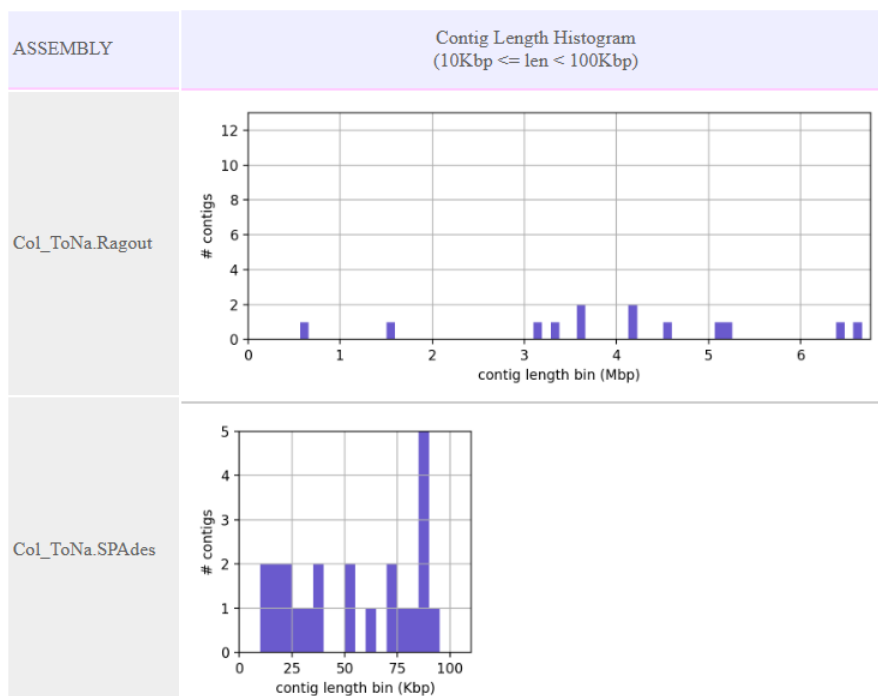
Grafo de ensamblaje cromosómico con Ragout del aislado Col_ToNa



Nota. Gráfica obtenida del software Bandage.

Figura 18

Histograma de longitudes de contigs (>10 Kb) de los ensamblajes de Col_ToNa



Nota. Gráficas obtenidas de Compare Assembled Contig Distribution.

4.4.2 Comparación Funcional de los Ensamblajes

La comparación se orientó a valorar si partir de un ensamblaje a nivel cromosómico (Ragout) permitió obtener una anotación funcional más informativa e interpretable que la lograda sobre un ensamblaje de contigs con SPAdes (Bermeo Salazar, 2024). Este enfoque se planteó porque la continuidad del ensamblaje condiciona la integridad de los modelos génicos, y por lo tanto, la detección de dominios y la asignación de funciones. En genomas eucariotas, ensamblajes más continuos reducen la fragmentación de genes entre *scaffolds* y mejoran la calidad de la anotación resultante (Yandell & Ence, 2012). Además, Ragout ordenó y orientó los fragmentos guiados por sintenia con referencias cercanas, lo que elevó la continuidad del genoma antes de la anotación (Kolmogorov et al., 2014, 2018). La Tabla 13 resume el número de genes codificantes por aislado en ambos pipelines.

Tabla 13

Genes codificantes por aislado (Ragout vs. SPAdes)

Aislado	Ensamblaje a nivel cromosómico (Presente estudio)	Ensamblaje <i>de novo</i> (Bermeo Salazar, 2024)
<i>Col_ToNa</i>	13,702	11,623
<i>Col_To</i>	13,802	11,953
<i>Col_Na</i>	13,877	12,658

Nota. Los recuentos del presente estudio corresponden a modelos depurados a una isoforma por gen tras el ensamblaje a nivel cromosómico (Ragout). Los valores de Bermeo provienen de ensamblajes *de novo* con SPAdes y reflejan los genes finales tras la depuración de contaminación.

En este estudio, tras depurar isoformas a una por gen, los conjuntos proteicos mantuvieron completitud BUSCO ~99.9% con duplicados ~0.3%, lo que sugiere que la depuración eliminó redundancias técnicas sin sacrificar ortólogos esenciales y dejó un conjunto proteico no redundante adecuado para la anotación funcional. Este comportamiento respalda que las diferencias observadas en la Tabla 1 derivaron de mejoras en continuidad y coherencia

estructural previas a la anotación. A nivel de contenido, las anotaciones en este estudio mostraron un núcleo funcional consistente (transportadores MFS/ABC, reguladores Zn_clus/Fungal_trans, enzimas P450/SDR, quinasas y HET), que repitió entre aislados y es congruente con la biología hemibiótrofa de *Colletotrichum*.’

La Figura 19 ilustra la cobertura de anotación funcional en *Col_ToNa*. Se observa que la anotación basada en dominios (Pfam/HMMER) y en homología curada (Swiss-Prot/DIAMOND) cubrió una fracción amplia del conjunto proteico (59-70%), mientras que en el estudio de Bermeo solo se informó la cobertura KEGG (36%). Aunque estas métricas no son estrictamente equivalentes (Pfam y Swiss-Prot evalúan dominios y similitud de proteína completa; KEGG asigna ortólogos a rutas), el contraste indicó que los perfiles HMM de Pfam aportaron evidencia funcional cuando la homología de longitud completa fue insuficiente y que una mayor continuidad en el ensamblaje favoreció modelos génicos más íntegros, y en consecuencia, una interpretación funcional más clara (Buchfink et al., 2021; Finn et al., 2013; Mistry et al., 2021).

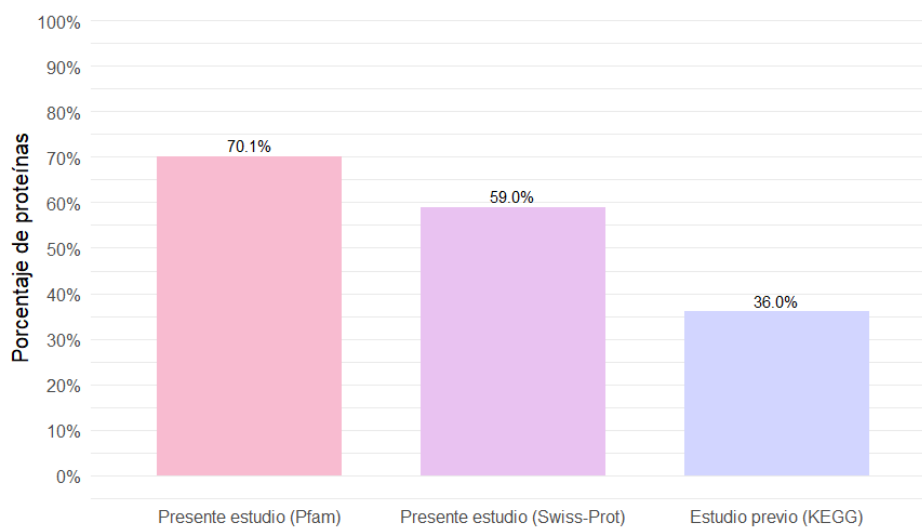
Por otra parte, en el reporte de Bermeo se describió una presencia marcada de proteínas hipotéticas, incluyendo casos en los que no se detectaron dominios conservados, lo que limitó la inferencia funcional pese al cribado con herramientas de efectores (Bermeo Salazar, 2024). Este patrón se reportó con ejemplos puntuales (por ejemplo, una proteína hipotética con dominio NPP1, otra con dominio de unión a celulosa y una sin dominios detectables), destacándose la dificultad de anotar efectores y proteínas específicas de linaje cuando las secuencias quedan parcialmente representadas o poco conservadas (Bermeo Salazar, 2024). Dado que su ensamblaje permaneció a nivel de contigs, es probable que esa fracción de genes hipotéticos correspondiera a ORFs incompletos o fragmentados debido a cortes entre *scaffolds* o anotación menos informada (Yandell & Ence, 2012).

En este estudio, el uso de perfiles HMM de Pfam aportó evidencia funcional basada en dominios a gran parte del conjunto proteico, precisamente donde la homología a nivel proteína

completa resulta limitante por divergencia de secuencias o especificidad del linaje (Eddy, 2011; Mistry et al., 2021). La detección sistemática de dominios (como MFS_1, Zn_clus, P450, adh_short, Pkinase y HET), permitió asignar funciones generales y dar soporte a interpretaciones biológicas aun cuando ciertos genes carecieron de top-hits curados en Swiss-Prot. (Finn et al., 2013; Mistry et al., 2021).

Figura 19

Cobertura de anotación funcional en Col_ToNa



Nota. Porcentaje de proteínas con ≥ 1 dominios Pfam y con top-hit en Swiss-Prot (presente estudio, ensamblaje cromosómico) y cobertura KEGG (estudio previo, ensamblaje *de novo*) en *Col_ToNa*. Las métricas no son equivalentes, se muestran como complementos para comparar la cobertura de los pipelines.

En general, los resultados indican que valió la pena introducir el paso adicional de reordenamiento cromosómico previo a la anotación, ya que produjo modelos génicos más íntegros, facilitó la detección de dominios y la descripción de un núcleo funcional repetible entre aislados, además, redujo la cantidad de etiquetas inespecíficas (“hipotéticas”) como resultado de ORFs truncados o poco informativos. Este resultado se alinea con buenas prácticas de

anotación que recomiendan maximizar la continuidad del ensamblaje antes de ejecutar predictores y flujos funcionales (Brúna et al., 2021; Gabriel et al., 2024; Yandell & Ence, 2012).

Finalmente, es importante señalar que parte de las diferencias entre ambos trabajos pudo atribuirse a elecciones de pipeline (por ejemplo, predictores y bases consultadas), por lo que la comparación se interpretó de forma descriptiva. Mientras que en este estudio se documentó de manera uniforme un perfil funcional sustentado por dominios, en el de (Bermeo Salazar, 2024), se enfatizó la dificultad de asignar función a un subconjunto amplio de proteínas hipotéticas. Aun así, el conjunto de los resultados reportados sugiere que, para *Colletotrichum*, el salto a ensamblajes de escala cromosómica constituye una inversión metodológica rentable para mejorar la calidad y la interpretabilidad de la anotación funcional.

CAPÍTULO V: CONCLUSIONES

- El reordenamiento cromosómico con Ragout reorganizó el ensamblaje *de novo* con SPAdes en 14-16 *scaffolds* por aislado, con N50 \approx 4.5-5.0 Mb y L50 = 5; la completitud del ensamblaje (BUSCO) fue C \approx 99.6-99.7%, sustentando calidad cercana a escala cromosómica.
- La concordancia entre métricas (N50/L50, BUSCO) y visualizaciones (grafos lineales en Bandage y curvas acumuladas en QUAST) respalda la validez del pseudoensamblaje a escala cromosómica y su consistencia entre aislados.
- La anotación estructural con BRAKER3 y la depuración a una isoforma por gen con AGAT + gffread, predijo \sim 13,700-13,900 genes por aislado. Los conjuntos proteicos resultantes alcanzaron \sim 99.9% de completitud (BUSCO).
- La anotación funcional alcanzó el 59% de cobertura por homología (DIAMOND/Swiss-Prot) y 70% por dominios (Pfam/HMMER), recuperando un núcleo funcional consistente entre aislados (transportadores MFS/ABC, reguladores Zn_clus, Fungal_trans, P450/SDR, quinasas, HET), congruente con la biología hemibiótrofa del género *Colletotrichum*.
- En comparación con el ensamblaje a nivel de *contigs*, el enfoque cromosómico redujo la fragmentación, mejoró la continuidad y favoreció modelos génicos más íntegros y comparables, aumentando la interpretabilidad de la anotación.
- Los recursos generados (ensamblajes cromosómicos FASTA, anotaciones GFF/GTF) habilitan análisis posteriores (efectores, CAZymes, rutas metabólicas y variación estructural), y constituyen una base técnica para estudios de estrategias fitosanitarias.

CAPÍTULO VI: RECOMENDACIONES

- Incorporar lecturas largas (Oxford Nanopore/PacBio) o Hi-C y realizar un ensamblaje híbrido para refinar la captura de la conformación cromosómica y alcanzar ensamblajes telómero a telómero (T2T), facilitando una caracterización funcional más robusta.
- Enmascarar el contenido repetitivo antes de la anotación estructural, reportar el porcentaje de repeticiones y documentar el impacto de este paso adicional sobre el número de genes y la cobertura de la anotación funcional.
- Completar la anotación funcional combinando homología y dominios, como la integración de InterProScan y el mapeo a rutas en KEGG para facilitar comparaciones consistentes entre aislados.
- Profundizar el perfilado de funciones clave asociadas a patogenicidad, enfocándose en metabolitos secundarios, secretoma y efectores.
- Realizar ortología comparativa y construir un pangenoma entre los tres aislados y referencias cercanas, relacionando la variación en el contenido de efectores y CAZymes con el hospedero y los fenotipos de virulencia reportados.
- Implementar control de contaminación sistemático antes y después del ensamblaje, documentando el material removido y su impacto en las métricas generales (tamaño del genoma, N50, L50, BUSCO, número de genes y proteínas).
- Fortalecer la reproducibilidad de este estudio mediante la fijación de versiones y parámetros del pipeline, unificando la nomenclatura de los archivos resultantes, y dejando disponibles los datos en repositorios para su verificación y reuso.

CAPÍTULO VII: REFERENCIAS

- Adhikari, T. B., Muzhinji, N., Philbrick, A. N., & Louws, F. J. (2025). Genomic insights and comparative analysis of *Colletotrichum* species associated with anthracnose fruit rot and crown rot of strawberry in North Carolina. *Frontiers in Microbiology*, *16*, 1515894. <https://doi.org/10.3389/FMICB.2025.1515894/BIBTEX>
- Aguilar-Bultet, L., & Falquet, L. (2015). Secuenciación y ensamblaje de novo de genomas bacterianos: una alternativa para el estudio de nuevos patógenos. *Revista de Salud Animal*, *37*(2), 125–132. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0253-570X2015000200008&lng=es&nrm=iso&tlng=es
- Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., Lippman, Z. B., & Schatz, M. C. (2019). RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, *20*(1), 1–17. <https://doi.org/10.1186/S13059-019-1829-6/TABLES/2>
- Andrews, S. (2010). *FastQC: A Quality Control tool for High Throughput Sequence Data*. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Baroncelli, R., Pensec, F., Lio, D. Da, Bouffleur, T., Vicente, I., Sarrocco, S., Picot, A., Baraldi, E., Sukno, S., Thon, M., & Le Floch, G. (2021). Complete Genome Sequence of the Plant-Pathogenic Fungus *Colletotrichum lupini*. *Molecular Plant-Microbe Interactions*, *34*(12), 1461–1464. <https://doi.org/10.1094/MPMI-07-21-0173-A/ASSET/IMAGES/LARGE/MPMI-07-21-0173-AT1-1640671635110.JPEG>
- Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Garmiri, P., da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., ...

- Zhang, J. (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), D523–D531. <https://doi.org/10.1093/NAR/GKAC1052>
- Becerra, S., Baroncelli, R., Bouffleur, T. R., Sukno, S. A., & Thon, M. R. (2023). Chromosome-level analysis of the *Colletotrichum graminicola* genome reveals the unique characteristics of core and minichromosomes. *Frontiers in Microbiology*, 14, 1129319. <https://doi.org/10.3389/FMICB.2023.1129319/EPUB>
- Bermeo Salazar, D. J. (2024). *Análisis genómico de un aislado Colletotrichum sp. causante de antracnosis en tomate de árbol y naranjilla*. Universidad de las Fuerzas Armadas ESPE.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. <https://doi.org/10.1093/BIOINFORMATICS/BTU170>
- Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., & Bairoch, A. (2007). UniProtKB/Swiss-Prot. *Plant Bioinformatics*, 406, 89–112. https://doi.org/10.1007/978-1-59745-535-0_4,
- Bradley, E. L., Ökmen, B., Doehlemann, G., Henrissat, B., Bradshaw, R. E., & Mesarich, C. H. (2022). Secreted Glycoside Hydrolase Proteins as Effectors and Invasion Patterns of Plant-Associated Fungi and Oomycetes. *Frontiers in Plant Science*, 13, 853106. <https://doi.org/10.3389/FPLS.2022.853106/XML>
- Brenes-Guillén, L. (2019). Ensamblaje y anotación de genomas microbianos. *Revista de Biología Tropical*.
- Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, 3(1), 1–11. <https://doi.org/10.1093/NARGAB/LQAA108>

- Buchfink, B., Reuter, K., & Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* 2021 18:4, 18(4), 366–368.
<https://doi.org/10.1038/s41592-021-01101-x>
- Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1), 59–60. <https://doi.org/10.1038/NMETH.3176>,
- Caicedo, J. D., Lalangui, K. P., Pozo, A. N., Cevallos, P. A., Arahana, V. S., & Méndez, K. S. (2017). Multilocus molecular identification and phylogenetic analysis of *Colletotrichum tamarilloi* as the causal agent of Tamarillo (*Solanum betaceum*) anthracnose in the Ecuadorian highlands. *European Journal of Plant Pathology*, 148(4), 983–996.
<https://doi.org/10.1007/S10658-017-1155-3/METRICS>
- Cajas Salazar, A. (2020). *Identificación de Colletotrichum spp. como agente causante de la antracnosis en hojas de uvilla (Physalis peruviana L.) en el centro de la región interandina del Ecuador*. Pontificia Universidad Católica del Ecuador.
- Casarrubias Carillo, U., Chavira González, M. M., Cruz Hernández, A., Cárdenas Soriano, E., Nieto Angel, D., & Guevara González, R. (2003). Variabilidad Genética de *Colletotrichum gloeosporioides* (Penz.) Penz. y Sacc. Aislado de Frutos de Papaya (*Carica papaya* L.) Mediante el Uso de Marcadores Moleculares RAPD. *Revista Mexicana de Fitopatología*, 21(3), 338–345. <https://www.redalyc.org/articulo.oa?id=61221315>
- Castro López, W. O., & Herrera, L. (2019). *La naranjilla (Solanum quitoense Lam.) en Ecuador*. <https://dspace.uclv.edu.cu/server/api/core/bitstreams/46f9f1d0-7562-4e18-9e86-54f453ca8624/content>
- Castro-Quezada, P., Bravo-Zúñiga, C., Cabrera-Cabrera, A., Quillay-Curay, N., Ramón-Montoya, M., Belesaca-Morocho, I., & Diaz-Granda, L. (2019). Caracterización morfológica

y molecular del agente causal de la antracnosis en tomate de árbol en Azuay y Loja.

Bosques Latitud Cero, 9(1)(1), 1–15.

Chen, W., Lee, M. K., Jefcoate, C., Kim, S. C., Chen, F., & Yu, J. H. (2014). Fungal Cytochrome P450 Monooxygenases: Their Distribution, Structure, Functions, Family Expansion, and Evolutionary Origin. *Genome Biology and Evolution*, 6(7), 1620–1634.

<https://doi.org/10.1093/GBE/EVU132>

Ciofini, A., Negrini, F., Baroncelli, R., & Baraldi, E. (2022). Management of Post-Harvest Anthracnose: Current Approaches and Future Perspectives. *Plants 2022*, Vol. 11, Page 1856, 11(14), 1856. <https://doi.org/10.3390/PLANTS11141856>

Costa, C., Dias, P. J., Sá-Correia, I., & Teixeira, M. C. (2014). MFS multidrug transporters in pathogenic fungi: do they have real clinical impact? *Frontiers in Physiology*, 5, 197.

<https://doi.org/10.3389/FPHYS.2014.00197>

Dallery, J. F., Lapalu, N., Zampounis, A., Pigné, S., Luyten, I., Amselem, J., Wittenberg, A. H. J., Zhou, S., de Queiroz, M. V., Robin, G. P., Auger, A., Hainaut, M., Henrissat, B., Kim, K. T., Lee, Y. H., Lespinet, O., Schwartz, D. C., Thon, M. R., & O'Connell, R. J. (2017). Gapless genome assembly of *Colletotrichum higginsianum* reveals chromosome structure and association of transposable elements with secondary metabolite gene clusters. *BMC Genomics*, 18(1), 1–22. <https://doi.org/10.1186/S12864-017-4083-X/FIGURES/7>

Daskalov, A. (2023). Emergence of the fungal immune system. *IScience*, 26(6), 106793.

<https://doi.org/10.1016/J.ISCI.2023.106793>

Díaz Granda, L., Canto Sáenz, M., Alegre Orihuela, J., Camarena Mayta, F., & Julca Otiniano, A. (2017). Sostenibilidad social de los subsistemas productivos de tomate de árbol (*Solanum betaceum* Cav) en el Cantón Guachapala, Provincia de Azuay - Ecuador.

Ecología Aplicada, 16(2), 99–104. <https://doi.org/10.21704/REA.V16I2.1013>

- Dida, F., & Yi, G. (2021a). Empirical evaluation of methods for de novo genome assembly. *PeerJ Computer Science*, 7, e636. <https://doi.org/10.7717/PEERJ-CS.636>
- Dida, F., & Yi, G. (2021b). Empirical evaluation of methods for de novo genome assembly. *PeerJ Computer Science*, 7, e636. <https://doi.org/10.7717/PEERJ-CS.636>
- Dofuor, A. K., Quartey, N. K. A., Osabutey, A. F., Antwi-Agyakwa, A. K., Asante, K., Boateng, B. O., Ablormeti, F. K., Lutuf, H., Osei-Owusu, J., Osei, J. H. N., Ekloh, W., Loh, S. K., Honger, J. O., Aidoo, O. F., & Ninsin, K. D. (2023). Mango anthracnose disease: the current situation and direction for future research. *Frontiers in Microbiology*, 14, 1168203. <https://doi.org/10.3389/FMICB.2023.1168203/FULL>
- Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, 7(10), e1002195. <https://doi.org/10.1371/JOURNAL.PCBI.1002195>
- Elbers, J. P., Rogers, M. F., Perelman, P. L., Proskuryakova, A. A., Serdyukova, N. A., Johnson, W. E., Horin, P., Corander, J., Murphy, D., & Burger, P. A. (2019). Improving Illumina assemblies with Hi-C and long reads: An example with the North African dromedary. *Molecular Ecology Resources*, 19(4), 1015. <https://doi.org/10.1111/1755-0998.13020>
- Falconí, C. E., Visser, R. G. F., & van Heusden, A. W. (2013). Phenotypic, Molecular, and Pathological Characterization of *Colletotrichum acutatum* Associated with Andean Lupine and Tamarillo in the Ecuadorian Andes. *https://Doi.Org/10.1094/PDIS-02-12-0175-RE*, 97(6), 819–827. <https://doi.org/10.1094/PDIS-02-12-0175-RE>
- Finn, R. D., Bateman, A., Clements, J., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2013). Pfam: the protein families database. *Nucleic Acids Research*, 42(Database issue), D222. <https://doi.org/10.1093/NAR/GKT1223>

- Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue), W29.
<https://doi.org/10.1093/NAR/GKR367>
- Fu, T., Shin, J. H., Lee, N. H., Lee, K. H., & Kim, K. S. (2022). Mitogen-Activated Protein Kinase CsPMK1 Is Essential for Pepper Fruit Anthracnose by *Colletotrichum scovillei*. *Frontiers in Microbiology*, 13, 770119. <https://doi.org/10.3389/FMICB.2022.770119/FULL>
- Fuentes-Pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. *Molecular Ecology*, 26(20), 5369–5406. <https://doi.org/10.1111/MEC.14264>
- Gabriel, L., Bruna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., & Stanke, M. (2024). BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *BioRxiv: The Preprint Server for Biology*.
<https://doi.org/10.1101/2023.06.10.544449>
- Gan, P., Hiroshima, R., Tsushima, A., Masuda, S., Shibata, A., Ueno, A., Kumakura, N., Narusaka, M., Hoat, T. X., Narusaka, Y., Takano, Y., & Shirasu, K. (2021). Telomeres and a repeat-rich chromosome encode effector gene clusters in plant pathogenic *Colletotrichum* fungi. *Environmental Microbiology*, 23(10), 6004–6018.
<https://doi.org/10.1111/1462-2920.15490>
- Gan, P., Ikeda, K., Irieda, H., Narusaka, M., O'Connell, R. J., Narusaka, Y., Takano, Y., Kubo, Y., & Shirasu, K. (2013). Comparative genomic and transcriptomic analyses reveal the hemibiotrophic stage shift of *Colletotrichum* fungi. *New Phytologist*, 197(4), 1236–1249.
<https://doi.org/10.1111/NPH.12085>,

- García Romero, I. A. (2019). Genómica de hongos fitopatógenos como herramienta para el desarrollo de estrategias de manejo. *Revista Colombiana de Biotecnología*, 21(2), 3–5. <https://doi.org/10.15446/REV.COLOMB.BIOTE.V21N2.84132>
- Guevara-Suarez, M., Cárdenas, M., Jiménez, P., Afanador-Kafuri, L., & Restrepo, S. (2022). Colletotrichum Species Complexes Associated with Crops in Northern South America: A Review. In *Agronomy* (Vol. 12, Issue 3). MDPI. <https://doi.org/10.3390/agronomy12030548>
- Hsieh, D. K., Chuang, S. C., Chen, C. Y., Chao, Y. T., Lu, M. Y. J., Lee, M. H., & Shih, M. C. (2022). Comparative Genomics of Three Colletotrichum scovillei Strains and Genetic Analysis Revealed Genes Involved in Fungal Growth and Virulence on Chili Pepper. *Frontiers in Microbiology*, 13, 818291. <https://doi.org/10.3389/FMICB.2022.818291/BIBTEX>
- Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., Von Mering, C., & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1), D309–D314. <https://doi.org/10.1093/NAR/GKY1085>
- Huo, J., Wang, Y., Hao, Y., Yao, Y., Wang, Y., Zhang, K., Tan, X., Li, Z., & Wang, W. (2021). Genome sequence resource for colletotrichum scovillei, the cause of anthracnose disease of chili. *Molecular Plant-Microbe Interactions*, 34(1), 122–126. <https://doi.org/10.1094/MPMI-03-20-0055-A/ASSET/IMAGES/LARGE/MPMI-03-20-0055-AT1-1608309937301.JPEG>
- Jashni, M. K., Mehrabi, R., Collemare, J., Mesarich, C. H., & de Wit, P. J. G. M. (2015). The battle in the apoplast: further insights into the roles of proteases and their inhibitors in plant–pathogen interactions. *Frontiers in Plant Science*, 6(AUG), 584. <https://doi.org/10.3389/FPLS.2015.00584>

- Jiang, L., Zhang, S., Su, J., Peck, S. C., & Luo, L. (2022). Protein Kinase Signaling Pathways in Plant-Colletotrichum Interaction. *Frontiers in Plant Science*, 12, 829645.
<https://doi.org/10.3389/FPLS.2021.829645>
- Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1), 27–30. <https://doi.org/10.1093/NAR/28.1.27>,
- Kolmogorov, M., Armstrong, J., Raney, B. J., Streeter, I., Dunn, M., Yang, F., Odom, D., Flicek, P., Keane, T. M., Thybert, D., Paten, B., & Pham, S. (2018). Chromosome assembly of large and complex genomes using multiple references. *Genome Research*, 28(11), 1720–1732. <https://doi.org/10.1101/GR.236273.118>
- Kolmogorov, M., Raney, B., Paten, B., & Pham, S. (2014). Ragout—a reference-assisted assembly tool for bacterial genomes. *Bioinformatics*, 30(12), i302–i309.
<https://doi.org/10.1093/BIOINFORMATICS/BTU280>
- Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5), 540–546.
<https://doi.org/10.1038/S41587-019-0072-8;SUBJMETA>
- Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive κ -mer weighting and repeat separation. *Genome Research*, 27(5), 722–736. <https://doi.org/10.1101/GR.215087.116/-/DC1>
- León F., J., Viteri D., P., & Cevallos A., G. (2004). *Manual del cultivo de tomate de árbol*.
<https://repositorio.iniap.gob.ec/bitstream/41000/827/4/iniapscm61.pdf>
- Liang, X., Wei, T., Cao, M., Zhang, X., Liu, W., Kong, Y., Zhang, R., & Sun, G. (2019). The MAP Kinase CfPMK1 Is a Key Regulator of Pathogenesis, Development, and Stress Tolerance

of *Colletotrichum fructicola*. *Frontiers in Microbiology*, 10(MAY), 1070.

<https://doi.org/10.3389/FMICB.2019.01070/FULL>

Liu, D., Hunt, M., & Tsai, I. J. (2018). Inferring synteny between genome assemblies: A systematic evaluation. *BMC Bioinformatics*, 19(1), 1–13. <https://doi.org/10.1186/S12859-018-2026-4/FIGURES/9>

Liu, F., Damm, U., Cai, L., & Crous, P. W. (2013). Species of the *colletotrichum gloeosporioides* complex associated with anthracnose diseases of Proteaceae. *Fungal Diversity*, 61(1), 89–105. <https://doi.org/10.1007/S13225-013-0249-2/METRICS>

Liu, X., Li, R. Q., Zeng, Q. X., Li, Y. Q., & Chen, X. A. (2023). A Novel Zn₂Cys₆ Transcription Factor, TopC, Positively Regulates Trichodin A and Asperpyridone A Biosynthesis in *Tolyposcladium ophioglossoides*. *Microorganisms*, 11(10), 2578. <https://doi.org/10.3390/MICROORGANISMS11102578/S1>

Lu, X., Miao, J., Shen, D., & Dou, D. (2022). Proteinaceous Effector Discovery and Characterization in Plant Pathogenic *Colletotrichum* Fungi. *Frontiers in Microbiology*, 13, 914035. <https://doi.org/10.3389/FMICB.2022.914035/BIBTEX>

Ma, W., Gao, X., Han, T., Mohammed, M. T., Yang, J., Ding, J., Zhao, W., Peng, Y. L., & Bhadauria, V. (2022). Molecular Genetics of Anthracnose Resistance in Maize. *Journal of Fungi*, 8(5), 540. <https://doi.org/10.3390/JOF8050540>

MacPherson, S., Larochelle, M., & Turcotte, B. (2006). A Fungal Family of Transcriptional Regulators: the Zinc Cluster Proteins. *Microbiology and Molecular Biology Reviews*, 70(3), 583–604. <https://doi.org/10.1128/MMBR.00015-06/ASSET/723086DB-40D6-46EA-9769-A724CD60C318/ASSETS/GRAPHIC/ZMR0030621290005.JPEG>

- 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>
- Marcial Acosta, L., Martínez Salinas, A., León Ruíz, J., Suárez Tapia, A., Viera Arroyo*, W., Marcial Acosta, L., Martínez Salinas, A., León Ruíz, J., Suárez Tapia, A., & Viera Arroyo*, W. (2023). DESARROLLO DEL FRUTO EN CULTIVARES DE TOMATE DE ÁRBOL (*Solanum betaceum* Cav.). *Perfiles*, 1(29), 30–39. <https://doi.org/10.47187/PERF.V1I29.200>
- Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., & Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, 34(13), i142–i150. <https://doi.org/10.1093/BIOINFORMATICS/BTY266>
- Ministerio de Agricultura y Ganadería. (2024a). Boletín situacional de naranjilla 2023. *Ministerio de Agricultura y Ganadería*. <https://sipa.agricultura.gob.ec/index.php/situacionales-agricolas/situacional-naranjilla>
- Ministerio de Agricultura y Ganadería. (2024b). *Boletín situacional de tomate de árbol 2023*. Ministerio de Agricultura y Ganadería. <https://sipa.agricultura.gob.ec/index.php/situacionales-agricolas/situacional-tomate-arbol-2021>
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1), D412–D419. <https://doi.org/10.1093/NAR/GKAA913>

NCBI. (2021). *Gapped Format for Genome Submissions*.

https://www.ncbi.nlm.nih.gov/genbank/wgs_gapped/?utm_source=chatgpt.com

Ochoa, J., Ellis, M., & Alwang, J. (2010). *Ojo de pollo de la naranjilla y su control*.

O'Connell, R. J., Thon, M. R., Hacquard, S., Amyotte, S. G., Kleemann, J., Torres, M. F., Damm, U., Buiate, E. A., Epstein, L., Alkan, N., Altmüller, J., Alvarado-Balderrama, L., Bauser, C. A., Becker, C., Birren, B. W., Chen, Z., Choi, J., Crouch, J. A., Duvick, J. P., ... Vaillancourt, L. J. (2012). Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. *Nature Genetics* 2012 44:9, 44(9), 1060–1065. <https://doi.org/10.1038/ng.2372>

Pallarés Zazo, J. (2021). *Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons*.

Paoletti, M. (2016). Vegetative incompatibility in fungi: From recognition to cell death, whatever does the trick. *Fungal Biology Reviews*, 30(4), 152–162.
<https://doi.org/10.1016/J.FBR.2016.08.002>

Paoletti, M., & Clavé, C. (2007). The fungus-specific HET domain mediates programmed cell death in *Podospora anserina*. *Eukaryotic Cell*, 6(11), 2001–2008.
https://doi.org/10.1128/EC.00129-07/SUPPL_FILE/FIG_S1.ZIP

Perlin, M. H., Andrews, J., & Toh, S. S. (2014). Essential letters in the fungal alphabet: ABC and MFS transporters and their roles in survival and pathogenicity. *Advances in Genetics*, 85, 201–253. <https://doi.org/10.1016/B978-0-12-800271-1.00004-4>,

Pham, S., Minkin, I., Pham, H., Starostina, E., & Vyahhi, N. (2013). C-Sibelia: an easy-to-use and highly accurate tool for bacterial genome comparison. *F1000Research*, 2, 258.
<https://doi.org/10.12688/F1000RESEARCH.2-258.V1>

- Prada, P., & Basto, G. (2011). *Manejo integrado de la antracnosis (Colletotrichum gloeosporioides (Penz), Penz y Sacc) en el cultivo de tomate de árbol (Cyphomandra betacea Sendt)*. Corporación Colombiana de Investigación Agropecuaria.
<https://repository.agrosavia.co/handle/20.500.12324/2376>
- Prijbelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics*, 70(1), e102.
<https://doi.org/10.1002/CPBI.102;PAGE:STRING:ARTICLE/CHAPTER>
- Revelo M., J., Viteri D., P., Vásquez C., W., Vaiverde, F., León F., J., & Gallegos, P. (2010). *Manual del cultivo ecológico de la naranjilla*.
- Rivas, E. M., & Mcewen, J. G. (2011). *ENSAMBLAJE, ANOTACIÓN Y ANÁLISIS INICIAL DE LOS GENOMAS DE Emmonsia parva Y Emmonsia crescens (Onygenales, Ascomycota)*.
- Rodríguez-López, E. S., González-Prieto, J. M., & Mayek-Pérez, N. (2009). La Infección de *Colletotrichum gloeosporioides* (Penz.) Penz. y Sacc. en Aguacatero (*Persea americana* Mill.): Aspectos Bioquímicos y Genéticos. *Revista Mexicana de Fitopatología*, 27.
https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-33092009000100007
- Rodríguez-Ortega, D., Vega-Jiménez, L., Murillo-Ilbay, Á., Peralta-Idrovo, E., & Rosas-Sotomayor, J. C. (2018). Variabilidad patogénica de *Colletotrichum lindemuthianum* y resistencia en germoplasma de *Phaseolus vulgaris* L. de Ecuador. *Agronomía Mesoamericana*, 29(1). <https://doi.org/10.15517/ma.v29i1.27511>
- Ruiz-Campos, C., Umaña-Rojas, G., & Gómez-Alpízar, L. (2022). Identificación multilocus de especies de *Colletotrichum* asociadas a la antracnosis. *Agronomía Mesoamericana*, 33(1), 1–18. <https://doi.org/10.15517/rac.v37i1.10710>

Schoch, C. L., Ciuffo, S., Domrachev, M., Hottton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., McVeigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database*, 2020.

<https://doi.org/10.1093/DATABASE/BAAA062>

Shin, J., Kim, J. E., Lee, Y. W., & Son, H. (2018). Fungal Cytochrome P450s and the P450 Complement (CYPome) of *Fusarium graminearum*. *Toxins*, 10(3), 112.

<https://doi.org/10.3390/TOXINS10030112>

Sinha, A. U., & Meller, J. (2007). Cinteny: Flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, 8(1), 1–9.

<https://doi.org/10.1186/1471-2105-8-82/FIGURES/5>

Takano, Y., Kikuchi, T., Kubo, Y., Hamer, J. E., Mise, K., & Furusawa, I. (2007). The *Colletotrichum lagenarium* MAP Kinase Gene CMK1 Regulates Diverse Aspects of Fungal Pathogenesis. *https://Doi.Org/10.1094/MPMI.2000.13.4.374*, 13(4), 374–383.

<https://doi.org/10.1094/MPMI.2000.13.4.374>

Tegenfeldt, F., Kuznetsov, D., Manni, M., Berkeley, M., Zdobnov, E. M., & Kriventseva, E. V. (2025). OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Research*, 53(D1), D516–D522.

<https://doi.org/10.1093/NAR/GKAE987>

Thomas, P. D. (2017). The Gene Ontology and the meaning of biological function. *Methods in Molecular Biology (Clifton, N.J.)*, 1446, 15. https://doi.org/10.1007/978-1-4939-3743-1_2

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics*, 13(1), 36–46.

<https://doi.org/10.1038/NRG3117;SUBJMETA>

- Tsushima, A., & Shirasu, K. (2022). Genomic resources of *Colletotrichum* fungi: development and application. *Journal of General Plant Pathology* 2022 88:6, 88(6), 349–357.
<https://doi.org/10.1007/S10327-022-01097-Y>
- Wang, H., Huang, R., Ren, J., Tang, L., Huang, S., Chen, X., Fan, J., Li, B., Wang, Q., Hsiang, T., Liu, H., & Li, Q. (2023). The evolution of mini-chromosomes in the fungal genus *Colletotrichum*. *MBio*, 14(4), e0062923. https://doi.org/10.1128/MBIO.00629-23/SUPPL_FILE/MBIO.00629-23-S0010.XLSX
- Wang, Y., Xu, W. T., Lu, R. Sen, Chen, M., Liu, J., Sun, X. Q., & Zhang, Y. M. (2023). Genome Sequence Resource for *Colletotrichum gloeosporioides*, an Important Pathogenic Fungus Causing Anthracnose of *Dioscorea alata*. *Plant Disease*, 107(3), 893–895.
<https://doi.org/10.1094/PDIS-03-22-0567-A/ASSET/IMAGES/LARGE/PDIS-03-22-0567-AT1-1680200307289.JPEG>
- Weir, B. S., Johnston, P. R., & Damm, U. (2012). The *Colletotrichum gloeosporioides* species complex. *Studies in Mycology*, 73(1), 115–180. <https://doi.org/10.3114/SIM0011>
- Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, 31(20), 3350–3352.
<https://doi.org/10.1093/BIOINFORMATICS/BTV383>
- Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* 2012 13:5, 13(5), 329–342. <https://doi.org/10.1038/nrg3174>
- Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), 821. <https://doi.org/10.1101/GR.074492.107>
- Zhang, Y., Zhuang, X., Meng, J., Zan, F., Liu, Z., Qin, C., Hao, L., Wang, Z., Wang, L., Li, H., Li, H., & Ding, S. (2023). A Putative Zn(II)2Cys6-Type Transcription Factor FpUme18 Is

Required for Development, Conidiation, Cell Wall Integrity, Endocytosis and Full Virulence in *Fusarium pseudograminearum*. *International Journal of Molecular Sciences*, 24(13), 10987. <https://doi.org/10.3390/IJMS241310987/S1>

CAPÍTULO IX: ANEXOS

Anexo A

Descarga del genoma de referencia de Colletotrichum scovillei (NCBI dataset)

```
# Descarga genoma de referencia NCBI
datasets download genome accession GCF_011075155.1 --include genome --
filename cscov.zip

# Descomprimir el archivo descargado
unzip -o cscov.zip

# Copiar el FASTA a ~/tesis/data/ como cscov.fasta
mkdir -p ~/tesis/data
cp "$(find ncbi_dataset -type f -name '*.fna' | head -n1)"
~/tesis/data/cscov.fasta

#Verificar
ls -lh ~/tesis/data/cscov.fasta
```

Anexo B

Preprocesamiento de secuencias

```
cd ~/tesis/data/ # directorio de las secuencias FASTA

# ACGTN en mayúsculas
awk '/^>/{print;next}{gsub(/[^ACGTNacgtn]/,"N"); print toupper($0)}'
input.fasta > input.clean.fasta

# Simplificació de encabezados (quitar todo tras el primer espacio)
awk 'BEGIN{FS=" "}{if(/^>){sub(/ .*/,""); print $1} else {print}}'
input.clean.fasta > input.simple.fasta
```

Anexo C

Ensamblaje asistido por referencia con Ragout: receta .rcp y ejecución

```
# Definición de rutas
REF="/home/sofia/tesis/data/cscov.simple.fasta"
TGT="/home/sofia/tesis/data/coltona_spades.fasta"

# Crear la receta .rcp
cat > coltona_cscov.rcp <<'EOF'
.references = cscov
.target = coltona

cscov.fasta = /home/sofia/tesis/data/cscov.simple.fasta
coltona.fasta = /home/sofia/tesis/data/coltona_spades.fasta
EOF

# Ejecutar Ragout
rm -rf ragout/coltona_cscov
ragout coltona_cscov.rcp -o ragout/coltona_cscov -t 8 --refine --overwrite
```

Anexo D

Evaluación del ensamblaje cromosómico frente a la referencia con QUASt

```
# Definición de rutas y carpeta de salida
REF="/home/sofia/tesis/data/cscov.simple.fasta"
ASM="/home/sofia/tesis/ragout/coltona_cscov/coltona_scaffolds.fasta"
OUT="/home/sofia/tesis/quast/coltona_cscov"
THREADS=8
MINCONTIG=5000

mkdir -p "$OUT"

# QUASt Ragout vs Referencia
quast.py "$ASM" \
  -r "$REF" \
  -o "$OUT" \
  -t "$THREADS" \
  -m "$MINCONTIG" \
  --eukaryote
```

Anexo E

Comparación de ensamblajes SPAdes vs. Ragout con QUASt

```
# Definición de rutas y carpeta de salida
REF="/home/sofia/tesis/ragout/coltona_cscov/coltona_scaffolds.fasta"
ASM="/home/sofia/tesis/data/coltona_spades.fasta"
OUT="/home/sofia/tesis/quast/spades_vs_coltona_ragout"
THREADS=8
MINCONTIG=5000

mkdir -p "$OUT"

# QUASt SPAdes vs Ragout
quast.py "$ASM" \
  -r "$REF" \
  -o "$OUT" \
  -t "$THREADS" \
  -m "$MINCONTIG" \
  --eukaryote
```

Anexo F

Evaluación de completitud del ensamblaje con BUSCO

```
# modo genome; sordariomycetes_odb10
busco \
  -i /home/sofia/tesis/ragout/coltona_cscov/coltona_scaffolds.fasta \
  -o coltona_cscov_busco \
  -l sordariomycetes_odb10 \
  -m genome \
  -c 8
```

Anexo G

Comparación de métricas (SPAdes vs. Ragout) - Aislado Col_ToNa

ASSEMBLY	LONGEST CONTIG (bp)	Nx (Lx)		LENGTH (bp)	NUM CONTIGS	SUM LENGTH (bp)
		Nx	Lx			
Col_ToNa.Ragout	6571457	N50:	4559604	$\geq 10^6$	12	51530599
		L50:	(5)	$\geq 10^5$	13	52098201
		N75:	3649273	$\geq 10^4$	14	52129356
		L75:	(8)	$\geq 10^3$	14	52129356
		N90:	3148574	≥ 500	14	52129356
		L90:	(11)	≥ 1	14	52129356
Col_ToNa.SPAdes	2246602	N50:	929323	$\geq 10^6$	17	24753278
		L50:	(19)	$\geq 10^5$	76	50609701
		N75:	541425	$\geq 10^4$	99	51849063
		L75:	(38)	$\geq 10^3$	208	52073616
		N90:	272199	≥ 500	1170	52697920
		L90:	(59)	≥ 1	1170	52697920

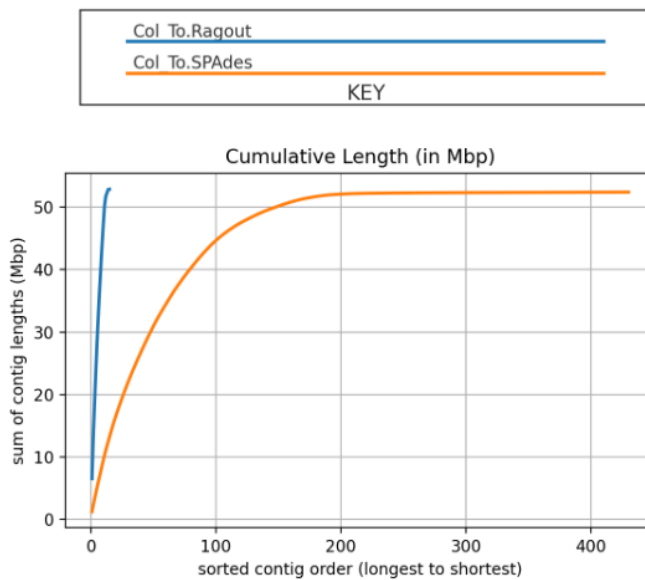
Anexo H

Comparación de métricas (SPAdes vs. Ragout) - Aislado Col_To

ASSEMBLY	LONGEST CONTIG (bp)	Nx (Lx)		LENGTH (bp)	NUM CONTIGS	SUM LENGTH (bp)
		Nx	Lx			
Col_To.Ragout	6543026	N50:	4953199	$\geq 10^6$	12	51732719
		L50:	(5)	$\geq 10^5$	14	52852160
		N75:	3709047	$\geq 10^4$	15	52883064
		L75:	(8)	$\geq 10^3$	15	52883064
		N90:	3083916	≥ 500	15	52883064
		L90:	(11)	≥ 1	15	52883064
Col_To.SPAdes	1262157	N50:	444786	$\geq 10^6$	1	1262157
		L50:	(39)	$\geq 10^5$	127	48253619
		N75:	260779	$\geq 10^4$	201	52101597
		L75:	(77)	$\geq 10^3$	272	52311061
		N90:	118785	≥ 500	430	52414135
		L90:	(117)	≥ 1	430	52414135

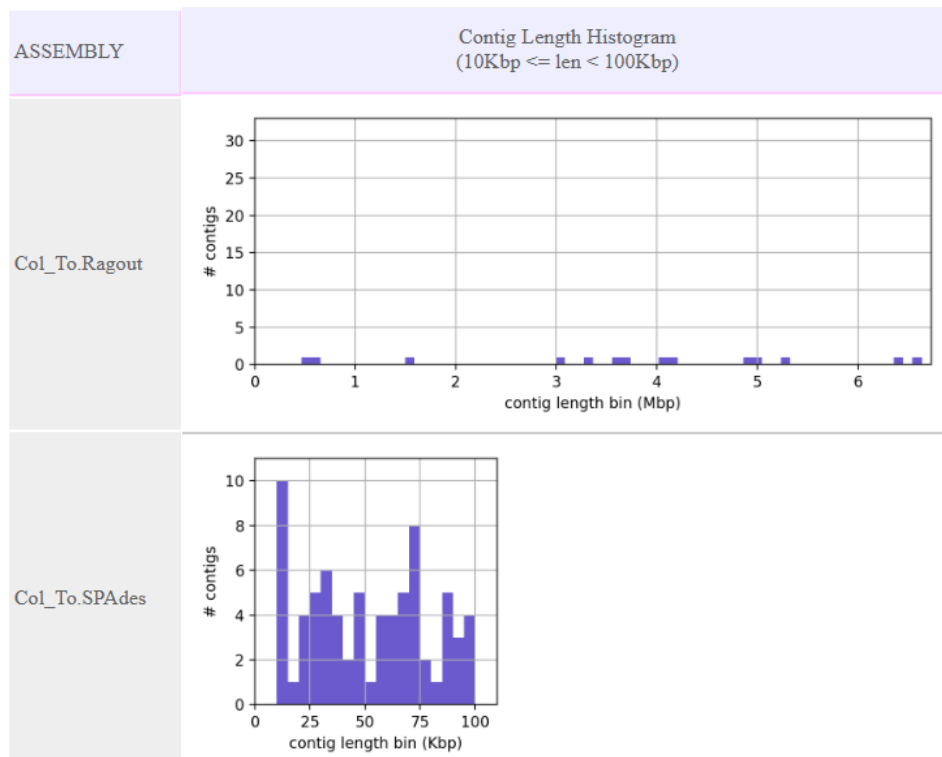
Anexo I

Curva de suma acumulada de las longitudes de los contigs - Aislado Col_To



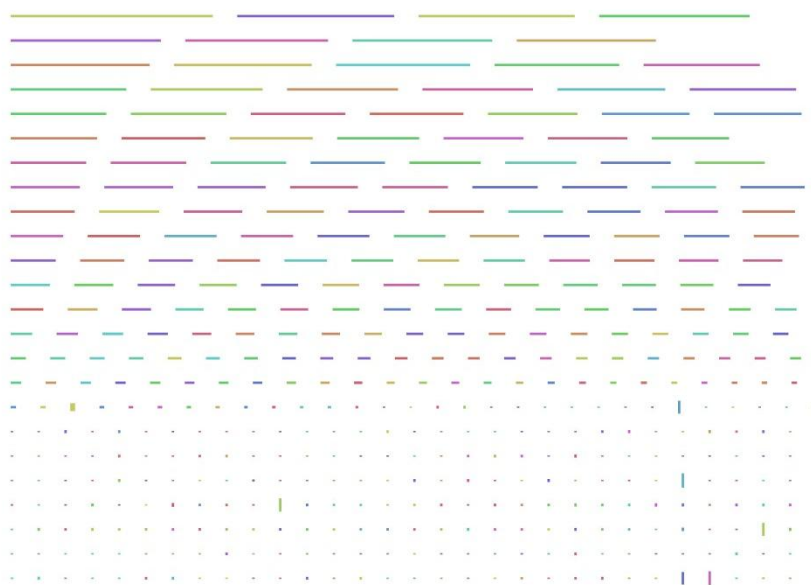
Anexo J

Histograma de longitudes de contigs (>10 Kb) - Aislado Col_To



Anexo K

Grafo de ensamblaje de novo (SPAdes) - Aislado Col_To

**Anexo L**

Grafo de ensamblaje cromosómico (Ragout) - Aislado Col_To



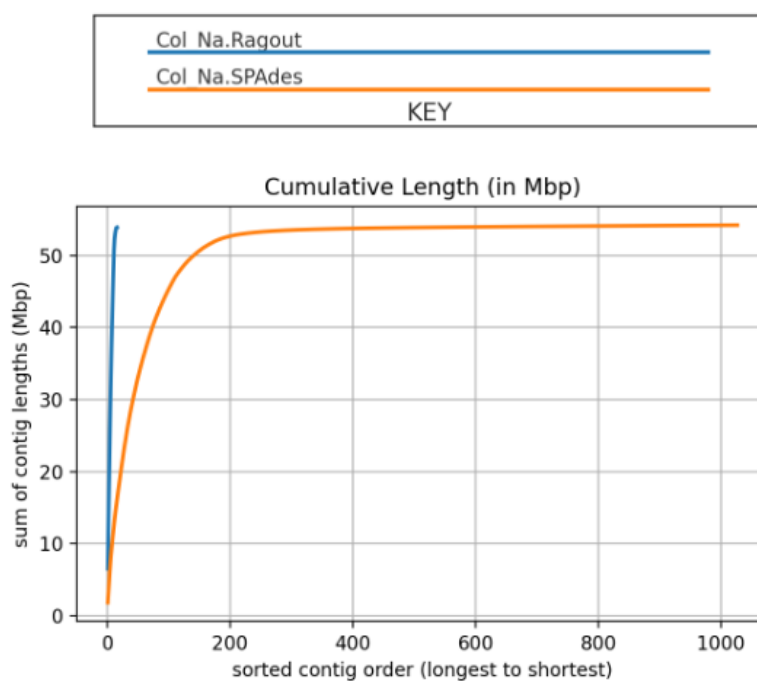
Anexo M

Comparación de métricas (SPAdes vs. Ragout) - Aislado Col_Na

ASSEMBLY	LONGEST CONTIG (bp)	Nx (Lx)		LENGTH (bp)	NUM CONTIGS	SUM LENGTH (bp)
		Nx	Lx			
Col_Na.Ragout	6519681	N50:	5000376	$\geq 10^6$	12	52281950
		L50:	(5)	$\geq 10^5$	15	53834879
		N75:	3658026	$\geq 10^4$	16	53865182
		L75:	(8)	$\geq 10^3$	16	53865182
		N90:	3128674	≥ 500	16	53865182
		L90:	(11)	≥ 1	16	53865182
Col_Na.SPAdes	1807121	N50:	464458	$\geq 10^6$	6	8138869
		L50:	(36)	$\geq 10^5$	125	48596375
		N75:	229241	$\geq 10^4$	229	53061128
		L75:	(77)	$\geq 10^3$	508	53851360
		N90:	97652	≥ 500	1027	54179120
		L90:	(127)	≥ 1	1027	54179120

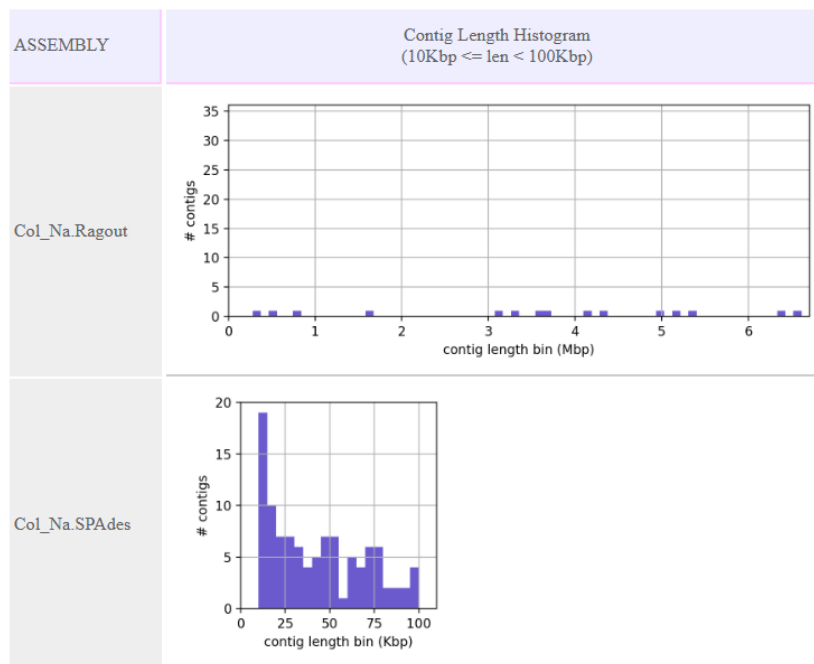
Anexo N

Curva de suma acumulada de las longitudes de los contigs - Aislado Col_Na



Anexo O

Histograma de longitudes de contigs (>10 Kb) - Aislado Col_Na



Anexo P

Grafo de ensamblaje de novo (SPAdes) - Aislado Col_Na



Anexo Q

Grafo de ensamblaje cromosómico (Ragout) - Aislado Col_Na



Anexo R

Ejecución de la anotación estructural con BRAKER3

```
# BRAKER3
braker.pl \
  --threads 35 \
  --species=col_tona \
  --workingdir=$(pwd) \
  --genome=/home/jovyan/execution/sofia/coltona_scaffolds.fasta \
  --prot_seq=/home/jovyan/data/sofia/fungi.fa \
  --verbosity 3
```

Anexo S

Métricas de la anotación estructural: genes, transcritos y proteínas (GTF/FAA)

```
# NOTA: Ubicarse en el directorio del archivo GTF de BRAKER

# número de genes anotados en el GTF
awk '$3=="gene"{n++} END{print n+0}' braker.gtf

# número de transcritos anotados en el GTF
awk '$3=="transcript"{n++} END{print n+0}' braker.gtf

# número de proteínas predichas (cuentas de cabeceras en FASTA)
grep -c '^>' braker.aa

# longitud media de proteínas predichas (aa)
awk 'BEGIN{L=0;N=0} /^>/{if(N>0) print L; L=0; N++; next} {gsub(/\r/, ""); L+=length($0)} END{if(N>0) print L}' braker.aa \
| awk '{s+=$1;c++} END{print s/c}'
```

Anexo T

Complejidad de la anotación estructural con BUSCO

```
# BUSCO en modo proteína para varios aislados
# modo protein; fungi_odb10
for iso in coltona colto colna; do busco -i "${iso}_braker.aa" -m protein -l
fungi_odb10 -o "busco_${iso}_fungi10" -c 8; done
```

Anexo U

Selección de la isoforma más larga por gen (AGAT)

```
# Conservar solo la isoforma más larga por gen
agat_sp_keep_longest_isoform.pl \
  -gff ~/tesis/annotation/agat_out/colna_braker_gff3 \
  -o ~/tesis/annotation/agat_out/colna_braker.longest.gff3
```

Anexo V

Extracción de CDS y proteínas a partir de GFF3 y genoma (gffread)

```
# Extraer CDS (fna) y proteínas (faa) desde GFF3 y genoma
gffread \
  ~/tesis/annotation/agat_out/coltona_braker.longest.gff3 \
  -g /home/bflores/braker_run/sofia/coltona_scaffolds.fasta \
  -x ~/tesis/annotation/gffread_out/coltona_CDS_longest.fna \
  -y ~/tesis/annotation/gffread_out/coltona_proteins_longest.faa \
  -F -Z
```

Anexo W

Anotación por homología (DIAMOND BLASTp vs UniProt/Swiss-Prot)

```
# Búsqueda de homologías contra Swiss-Prot
diamond blastp \
  -q ~/tesis/annotation/gffread_out/coltona_proteins_longest.faa \
  -d ~/tesis/databases/uniprot_sprot \
  -o ~/tesis/annotation/diamond_out/coltona_vs_uniprot.tsv \
  -f 6 qseqid sseqid pident length evalue bitscore stitle \
  -k 1 -e 1e-5 -p 8 \
  --tmpdir /tmp/diamond_tmp
```

Anexo X

Anotación de dominios proteicos (Pfam-A; pfam_scan.pl/HMMER)

```
# Búsqueda de dominios Pfam-A con HMMER
pfam_scan.pl \
  -fasta ~/tesis/annotation/gffread_out/coltona_proteins_longest.faa \
  -dir ~/tesis/pfam \
  -e_seq 1e-5 \
  -cpu 12 \
  -outfile ~/tesis/annotation/interproscan_out/coltona_Pfam.tsv
```