Biblioteca Digital
F C E N - U B A

BIBLIOTECA CENTRAL LUIS F LELOIR
BIBLIOTECA CENTRAL LELOIR
FACULTAD DE CIENCIAS EXACTAS Y NATURALES UBA

Tesis Doctoral

Biología computacional aplicada al análisis genómico de dos cultivares élite (Solanum lycopersicum: Heinz y M82) y una especie silvestre de tomate (Solanum pennellii: LA0716)

Lichtenstein, Gabriel

2018-06-29

Este documento forma parte de la colección de tesis doctorales y de maestría de la Biblioteca Central Dr. Luis Federico Leloir, disponible en digital.bl.fcen.uba.ar. Su utilización debe ser acompañada por la cita bibliográfica con reconocimiento de la fuente.

This document is part of the doctoral theses collection of the Central Library Dr. Luis Federico Leloir, available in digital.bl.fcen.uba.ar. It should be used accompanied by the corresponding citation acknowledging the source.

Cita tipo APA:
Lichtenstein, Gabriel. (2018-06-29). Biología computacional aplicada al análisis genómico de dos cultivares élite (Solanum lycopersicum: Heinz y M82) y una especie silvestre de tomate (Solanum pennellii: LA0716). Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires.

Cita tipo Chicago:
Lichtenstein, Gabriel. "Biología computacional aplicada al análisis genómico de dos cultivares élite (Solanum lycopersicum: Heinz y M82) y una especie silvestre de tomate (Solanum pennellii: LA0716)". Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. 2018-06-29.

EXACTAS UBA
Facultad de Ciencias Exactas y Naturales

UBA
Universidad de Buenos Aires

Dirección: Biblioteca Central Dr. Luis F. Leloir, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. Intendente Güiraldes 2160 - C1428EGA - Tel. (++54 +11) 4789-9293

Contacto: digital@bl.fcen.uba.ar

UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales

Biología Computacional aplicada al análisis genómico de dos cultivares élite (*Solanum lycopersicum*: Heinz y M82) y una especie silvestre de tomate (*Solanum pennellii*: LA0716)

Tesis presentada para optar por el título de
Doctor de la Universidad de Buenos Aires
en el área de Ciencias Biológicas

Lic. Gabriel Lichtenstein

Director de Tesis: Dr. Fernando Carrari
Director Conjunto: Dr. Björn Usadel
Consejero de Estudios: Dr. Esteban Hopp

Lugar de trabajo: Laboratorio de Fisiología del Metabolismo Vegetal,
Instituto de Biotecnología, CICVyA, INTA de Castelar

Buenos Aires, 10 de marzo de 2018

Biología Computacional aplicada al análisis genómico de dos cultivares élite (*Solanum lycopersicum*: Heinz y M82) y una especie silvestre de tomate (*Solanum pennellii*: LA0716)

Resumen

Mediante la aplicación de diferentes enfoques de Biología Computacional, en esta Tesis se abordaron tres proyectos genoma de tomate: i) el genoma mitocondrial del cultivar *élite* Heinz se secuenció, ensambló y anotó; ii) se ancló la secuencia completa del genoma de la especie silvestre de tomate *Solanum pennellii*, el padre donante de la conocida población de líneas de introgresión (IL), al mapa genético y iii) mediante genómica comparativa y análisis de *QTL* (*loci* de caracteres cuantitativos) se identificaron genes candidatos asociados a variaciones en la fotosíntesis y en caracteres relacionados con el crecimiento de las plantas. Los resultados del ensamblado del genoma mitocondrial del tomate mostraron que las inserciones nucleares de origen mitocondrial (del inglés, *numts*: *nuclear mitochondrial DNA segments*) están presentes aún en los tomates contemporáneos. Este hallazgo fue validado por hibridación fluorescente *in situ* la cual mostró un número particularmente elevado de *numts* en el cromosoma 11 del cultivar Heinz. En la segunda parte de esta Tesis se presenta la secuencia completa del genoma de *S. pennellii* la cual fue ensamblada *de novo* en 4.591 *scaffolds,* de los cuales, el 97,1% fueron anclados (mediante una base de datos interna de 16.940 marcadores moleculares) al mapa genético del tomate. Por último, en el capítulo final de esta Tesis, el potencial como herramientas para la investigación y el mejoramiento de cultivos de las secuencias genómicas aquí presentadas, se demostró por el genoma del cultivar *S. lycopersicum* M82, ensamblado por referencia. En esta parte, se identificaron 87 genes candidatos involucrados en la determinación de *loci* de caracteres cuantitativos responsables de variaciones en la fotosíntesis y en caracteres relacionados con el crecimiento de la planta presentes en 11 regiones (*BIN*s) de las líneas de introgresión de tomate.

Palabras claves: Tomate – *Solanum* – Genómica – Bioinformática – Mitocondria

Computational Biology applied to genomic analyses of two tomato elite cultivars (*Solanum lycopersicum*: Heinz & M82) and a wild species (*Solanum pennellii*: LA0716)

Abstract

By the application of different Computational Biology approaches, in this Thesis, three tomato genome projects were boarded and analyzed: i) the mitochondrial genome from the elite Heinz cultivar was sequenced, assembled and annotated; ii) the genome sequence of *Solanum pennellii*, the donor parent of the well-known tomato introgression lines population (IL), was studied and anchored to a tomato genetic map to construct high-quality pseudo-chromosomes; and lastly iii) by a comparative genomics approach the assembled-by-reference genome of the M82 elite cultivar and recurrent IL parent was contrasted to the donor parental genome. In this regard, quantitative trait *loci* (*QTL*) containing candidate genes where selected for a comparative analysis of chromosomic regions (*BIN*s) linked to variations in photosynthesis- and plant growth-related traits. In the first chapter of this Thesis, an in-depth analysis of the tomato mitochondrial genome showed that nuclear DNA segments of mitochondrial origin (*numts*) are still present in the contemporary tomatoes. This *in silico* discovery was further cross-validated by fluorescence *in situ* hybridization (FISH) which confirmed a particularly high number of *numts* in Heinz cultivar chromosome 11. In the following chapter, the complete *de novo* genome sequence of *S. pennellii* is presented with a total of 4,591 *scaffolds* of which 97.1% were anchored (with an in-house database of 16,940 molecular markers) to the tomato genetic map. Indeed, the use of these genomes as tools for research and crop improvement was showcased by the assembled-by-reference M82 genome introduced in the last chapter of this Thesis. On this subject, structural and functional annotated *loci* distributed over 11 *BIN*s were compared and evaluated: 4,916 and 5,886 *loci* in M82 and LA0716, respectively. This study resulted in the selection of 87 candidate genes underpinning the genetic bases of *QTL* for variations in tomato photosynthesis and plant growth.

Keywords: Tomato – *Solanum* – Genomics – Bioinformatics – Mitochondria

Publicaciones

Parte de los resultados presentados en esta Tesis Doctoral han sido divulgados en las siguientes publicaciones con referato:

1. de Oliveira Silva FM., Lichtenstein G., Alseekh S., Rosado-Souza L., Conte M., Suguiyama VF., Lira BS., Fanourakis D., Usadel B., Bhering LL., DaMatta FM., Sulpice R., Araújo WL., Rossi M., de Setta N., Fernie AR., Carrari F., Nunes-Nesi A. The Genetic Architecture of Photosynthesis and Plant Growth Related Traits in Tomato. **Plant Cell Environ. 2017 Oct 18**. doi: 10.1111/pce.13084. [Epub ahead of print] PubMed PMID: 29044606.

2. Lichtenstein G., Conte M., Asis R., Carrari F. Chloroplast and Mitochondrial Genomes of Tomato. In: Causse M., Giovannoni J., Bouzayen M., Zouine M. (eds) The Tomato Genome. Compendium of Plant Genomes. **Springer, Berlin, Heidelberg. 2016 Nov 24;** 111:137. doi:10.1007/978-3-662-53389-5_7

3. Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sørensen I, Lichtenstein G, Fich EA, Conte M, Keller H, Schneeberger K, Schwacke R, Ofner I, Vrebalov J, Xu Y, Osorio S, Aflitos SA, Schijlen E, Jiménez-Goméz JM, Ryngajllo M, Kimura S, Kumar R, Koenig D, Headland LR, Maloof JN, Sinha N, van Ham RC, Lankhorst RK, Mao L, Vogel A, Arsova B, Panstruga R, Fei Z, Rose JK, Zamir D, Carrari F, Giovannoni JJ, Weigel D, Usadel B, Fernie AR. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. **Nat Genet. 2014 Sep**; 46(9):1034-8. doi: 10.1038/ng.3046. Epub 2014 Jul 27. PubMed PMID: 25064008.

4. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. **Nature. 2012 May 30**; 485(7400):635-41. doi: 10.1038/nature11119. PubMed PMID: 22660326; PubMed Central PMCID: PMC3378239.

Agradecimientos

A mi director, el Prof. Dr. Fernando Carrari, gracias por haberme dado la oportunidad de formar parte de este gran grupo de trabajo, allá por el año 2010. Gracias por haberme guiado a través de numerosos desafíos durante el proceso de mi formación humana y profesional corrigiendo mi curso siempre hacia el objetivo - realizar un doctorado en Bioinformática, algo visionario y extraño en su momento, pero lo logramos. Fer, gracias por haber compartido conmigo tus conocimientos y la pasión por lo que haces y amas, por haberme ayudado a sentar las bases prácticas y académicas y por darme los recursos para fortificarlas y, animarme ahora, a caminar por mi cuenta y emprender mi camino. Bajo tu dirección he encontrado el coraje para desafiarme y salir de mi zona de confort, para aprender cosas nuevas, cosas interdisciplinarias, innovadoras, y por momentos, más allá de la frontera del conocimiento. En todo este recorrido, sin duda, manteniendo al equipo unido, la Lic. Mariana Conte, ecuánime guerrera, estuvo ahí para apoyarme, enseñarme con su ejemplo y disciplina a trabajar en grupo, y mientras también a forjar mi templanza en épocas de escases y abundancia. Mari, jeje, sé que no te gusta mucho que te diga así y que me lo dejas pasar 😊, gracias por todo, por tus consejos científicos y humanos, por las horas de catarsis, por estar siempre presente. Gracias por dejarme aprender de vos y junto a vos, de Linux, de ADN y de secuencias y también sobre la lucha por los derechos y la igualdad laboral, jeje te quiero Mari. En fin, muchas gracias a ambos dos, ha sido un honor haber hecho mi doctorado junto a ustedes.

A mi Director Conjunto, el Prof. Dr. Björn Usadel, gracias por haberme dado la oportunidad de pasar una estadía de 4 meses formando parte del *AG Usadel* en el increíble *Max-Planck-Institut für Molekulare Pflanzenphysiologie* de Potsdam-Golm en Alemania, una de las mejores experiencias de mi vida. A Anthony *Tony* Bolger, bajo su ala, se despertó en mí el entusiasmo por aprender programación y divertirme en el proceso, sos mi Jedi Master, sábelo, gracias a ti, cada vez comprendo un poco más a Gödel, Escher & Bach. Gracias a vos Tony, y a tu esposa, Marie Bolger, por haberme dado tanto (incluyendo cervezas y desayunos irlandeses) y por hacerme sentir parte de sus vidas, los admiro chicos. A Federico Giorgi, espero verte otra vez che, tal vez en Stormwind o Ironforge 😊, tu buen humor siempre fue contagioso y motivador, gracias Fede, sos un gran coach de la vida. A la Dra. Malgorzata Maggie Ryngajllo, gracias por mostrarme lo más lindo de Berlín, por compartir todo, aceptarme en tu vida y por enseñarme tanto, Buziaki Czarnuszca. A Diana Pese, tu hospitalidad y tu familia (Big Nathan and Little Jasmín) me mostraron la calidez

del hogar y la vida en Alemania, y nunca voy a olvidarme de ustedes. En fin, a todo el equipo Usadel, que me recibió y me dio herramientas e incontables momentos para seguir en mi carrera científica en bioinformática: Axel Nagel, Alex Finck, Aleksandar Vasilvski, Heike Riegler, Marc Lohse y Faysal El-Khettabi. Ich danke dir vielmals für alles 😊

A mi consejero de estudios: el Prof. Dr. Esteban Hopp y a los integrantes del Comité de Seguimiento de Tesis: la Prof. Dra. Irene Ruth Heinz y el Prof. Dr. Adrián Turjanski. Gracias por su tiempo y sus excelentes aportes que me ayudaron a avanzar hasta la meta final.

Agradezco también al Max Planck y al CONICET por otorgarme becas de posgrado que me permitieron dedicar exclusivamente a mis estudios Doctorales. También agradezco al Instituto de Biotecnología del INTA Castelar y a la FCEN de la UBA, por abrirme las puertas para que pueda realizar mi Tesis.

A todos los que pasaron por el Grupo Carrari y me apoyaron tanto: Carla Coluccio Leskow, Luisa Fernanda Bermúdez, Laura Kamenetzky, Mariana Conte, Guadalupe Domínguez, Leandro Quadrana, Tomas Duffy, Estanislao Burgos, Julia Verónica Sabio García, Mariana G. López y Marina Insani.

A mis amigos y compañeros del Instituto de Biotecnología del INTA: Sebastián Asurmendi, Diego Zavallo, Luis De Haro, Cintia Acuña, Sebastián Moschen, Jeremias Zubrzycki, Laura Radonic, Pamela Victoria Villalba, Yamila Agrofoglio, Carla Valeria Filippi, Flavia Darqui, Lula Peluffo, Carlos Manacorda, Sole Lucero, Vanesa Nahirñak, Gabriela Conti, Martín González de Urreta, Andrea Venturuzzi, Ariel Nagel, Cecilia Rodriguez, Cintia Rendón, las hermanas Corina y Maqui Fusari, Guillermo Maroniche, Gabriela Llauger, Emiliano Ben Guerrero, Perla Arrua, Federico Ehrenbolger y a cada persona con la se cruzó mi camino y que no llego a nombrar aquí pero que está presente.

Al equipo de sistemas de Guillermo Sutz Kaelin, gracias por enseñarme tanto sobre sistemas, redes, servidores, administración y muchas cosas más, por ejemplo, salvarme las papas más de una vez con nuestro querido servidor *Solanum*.

A los grupos de Brasil de María Magdalena Rossi y de Adriano Nunes-Nesi y Franklin de Oliveira da Silva que me dieron la oportunidad de tener un Capítulo III en esta tesis, muchas gracias por todo.

A mis compañeras de Biología Molecular: Jessi Allende, Naty Ferressini Gerpe y Flor. Gracias por haberme apoyado y estar presentes en esos momentos, gracias por su amistad. A Carlos David Bruque por tus experiencias y a Agustín Dramis, por darme tu apoyo, tus ejercicios, especificaciones y Python con budín casero 😊 para superar Introducción a la Computación. A Esteban Mocskos y a Maria Elena Buemi, gracias por ayudarme con las aguas de ICB y llegar a buen puerto.

A los incondicionales de siempre, mis pilares de la vida, Sebastián Bulacios Keena y Aime Rojas, Ernesto Arizaga y Miranda Nebbia, Nicolás Fernández, Francisco Goicoechea, Matías y Julieta Coppola, y Luis y Valeria Dejo. No tengo palabras, tendría que hacer otra Tesis para agradecerles, gracias por estar toda mi vida.

A mis amigos de la Academia Rusconi.

A María Luján Brenda Sá Dono, gracias por amarme todo lo que pudiste y por enseñarme de todo. A tus viejos, Rochi y Guri, gracias por dejarme ser parte de la familia, sin su aliento no estaría escribiendo estas líneas. Gracias por creer en mí.

A mi mentor, el que me enseñó a trabajar y me mostro que podía perseguir mi vocación, mi ejemplo máximo, el Prof. Dr. Máximo Barón y su familia.

A María Elena Grosso, mi madre, la incondicional, la luchadora, la perseverante, la de mil batallas sin jamás rendirse, siempre estuviste y vas a estar en mí alma tatuada con fuego sagrado, gracias por criarme, creer en mí, apoyarme, cuidarme y bancarme tanto. No tengo palabras, te amo mamá.

A mis hermanos Diego, Eric y Ron Lichtenstein. Son lo mejor de lo mejor y lo saben, es un honor ser su hermano mayor, en la unión esta la fuerza y me lo han demostrado una y otra vez, mis bros, *my brothers in arms*, mis pequeños dragones, gracias por ser mis hermanos, alentarme, levantarme y enseñarme tantas cosas donde quiera que estemos en este mundo. A mi hermana prima Deborah Lichtenstein y su familia (Cristina, Paul y Thibaud), desde lejos, desde cerca, siempre están. Gracias por todo prima linda, un beso gigante. A mi cuña, Pol-y, gracias por ser mi familia, por esa luz que suena en los corazones.

A mi padre, Mario Félix Lichtenstein, de alguna manera todo esto fue posible gracias a vos, siempre ausente, pero a la vez presente, lo que no nos mata nos fortifica, hay que surfear arriba de las olas y no llorar sobre leche derramada. Gracias por enseñarme a ganar, perder todo y volver a ganar. Esto también pasará.

Al Dr. Alberto Tardivo, gracias por darnos tu apoyo a mí y a mi familia, el que vino sin que lo llamen se fue sin que lo echen, dejaste tu huella y agradezco todo lo que hiciste por nosotros y por hacernos más fuertes.

A los del IMPaM, gracias por dejarme completar mi Tesis y darme la oportunidad de trabajar con ustedes en Bioinformática. Gracias por darme un lugar, por su paciencia, por su apoyo y por todo lo que me enseñan y me dan día a día.

A las mejores compañer@s del mundo entero, las que pase lo que pase, en el aquí y ahora, *semper fidelis*, mi manada: a mi princesa Cindy bar*BIN*cha negra, a mi India Indiana de chocolato, a la única e inigualable Kiara Lua Grosso, al Enzo loco corazón de dragón, y a Mora zuni, que en paz descanse.

Al universo y a cada persona que desde su lugar y en su momento me ayudó y me ayuda a crecer como profesional y como persona.

Dedicatoria

A todos los que me acompañaron en este recorrido.

Índice

Abreviaturas 1

CAPÍTULO 0: INTRODUCCIÓN GENERAL 3

0.1 | El Tomate: una especie modelo con importancia económica 3
0.2 | Genomas de plantas 27
0.3 | Definición del Problema, Hipótesis y Objetivo General 31

CAPÍTULO Ⅰ: EL GENOMA MITOCONDRIAL DE *S. LYCOPERSICUM* (CV HEINZ) Y SU RELACIÓN CON EL GENOMA NUCLEAR 35

I.1 | Introducción 35
I.2 | Resultados 40
I.3 | Discusión 50

CAPÍTULO Ⅱ: EL GENOMA DEL TOMATE SILVESTRE (*SOLANUM PENNELLII*), PARENTAL DONANTE DE LA POBLACIÓN DE LÍNEAS DE INTROGRESIÓN MÁS ESTUDIADA DE LA FAMILIA *SOLANACEAE* 55

II.1 | Introducción 55
II.2 | Resultados 59
II.3 | Discusión 65

CAPÍTULO Ⅲ: ANÁLISIS DE REGIONES GENÓMICAS (*QTL*) E IDENTIFICACIÓN DE GENES CANDIDATOS ASOCIADOS A LA REGULACIÓN DE PROCESOS FISIOLÓGICOS ESENCIALES COMO LA FOTOSÍNTESIS, RESPIRACIÓN Y VÍAS METABÓLICAS ASOCIADAS 68

III.1 | Introducción 68
III.2 | Resultados 72
III.3 | Discusión 80

CAPÍTULO IV: CONCLUSIONES GENERALES DE LA TESIS 83

CAPÍTULO V: MATERIALES & MÉTODOS (M&M) 86

V.1 | Capítulo Ⅰ 86
V.2 | Capítulo Ⅱ 96
V.3 | Capítulo Ⅲ 112

CAPÍTULO VI: FINANCIAMIENTO 117

CAPÍTULO VII: BIBLIOGRAFÍA 118

CAPÍTULO VIII: ANEXOS 146

VIII.1 | Capítulo Ⅰ 146
VIII.3 | Capítulo Ⅲ 151

Abreviaturas

FASTA = es un formato basado en texto para representar secuencias de nucleótidos o secuencias de péptidos, en el que los nucleótidos o aminoácidos se representan utilizando códigos de una sola letra.

FASTQ = Es un formato de archivo de secuenciación que contiene nucleótidos y también los valores de calidad de cada uno (representados con caracteres *ASCII*), generalmente utilizado para las lecturas obtenidas por *Illumina*®.

SEQ = un sufijo que indica un archivo en formato FASTA.

QUAL = un sufijo que indica un archivo de valores de calidad PHRED.

PHRED = este puntaje es una medida de la calidad de las bases obtenidas por la secuenciación automatizada de ADN.

GFF3 = tipo de archivo con formato tabular para almacenar datos de características genéticas (del inglés: *gene feature format versión 3*).

LCB = Bloque Local Colineal (del inglés: *Local Colineal Block*).

AFLP = Longitud del polimorfismo del fragmento amplificado (del inglés: *Amplified Fragment Length Polymorphism).*

DArT = Tecnología de matrices de diversidad (del inglés: *Diversity Arrays Technology*).

EST = Etiqueta de secuencia expresada (del inglés: *Expressed Sequence Tag).*

NGS = Siguiente generación de secuenciación (del inglés: *Next Generation Sequencing).*

QTL = *Loci* de caracteres cuantitativos (del inglés: *Quantitative Trait Loci).*

RNAseq = Secuenciación de transcriptos de ARN (del inglés: *Sequencing of RNA Transcripts*).

SNP = Polimorfismo de nucleótido único (del inglés: *Single Nucleotide Polymorphism).*

Introducción

Capítulo 0: Introducción General

0.1 | El Tomate: una especie modelo con importancia económica

0.1.1 | Breve reseña histórica del tomate

Aunque todavía es un tema de debate, el origen del tomate se remonta a las regiones andinas de América del Sur. Es sorprendente ver lo largos e intrincados que fueron los viajes que llevaron a esta especie desde la cordillera del Perú hacia México, y luego desde ahí alcanzó las costas del Caribe donde se embarcó hacía las costas de Asia Sudoriental para luego arribar al sur de Italia y viajar hasta el norte de Europa, para finalmente regresar y ser introducido en América del Norte (revisado por **Colvine y Branthôme, 2016**).

Pero los viajes del tomate no sólo han sido geográficos. Si bien este se ha consumido y ha sido cultivado mucho antes de la era cristiana (500 años A.C en México), primero tuvo que convencer a las diversas culturas que su fruto era un alimento seguro. Por su forma acorazonada y por su color rojo logró conquistar a los árabes que lo descubrieron en España, y algún tiempo después fue considerado como un fruto afrodisíaco por el herborista italiano Pietro Andrea Gregorio Mattioli, quien le dio el nombre de *pomo d'oro* en el año 1544. Sin embargo, cerca del mismo tiempo, la planta de tomate fue considerada como "altamente tóxica" por el médico y herborista inglés, John Gerard. Por ese entonces las sospechas que levantó lo relegaron a la condición de una planta ornamental escondida en el fondo de los jardines a lo largo de los siglos XVI, XVII y principios del XVIII. Lo más que se puede decir es que por su color y sabor el tomate se utilizaba para condimentar y realzar algunas sopas por alrededor de la época del 1730 en Inglaterra, mientras que, al mismo tiempo en el otro lado del Atlántico, los científicos desalentaron fuertemente su consumo debido a su semejanza con las tóxicas plantas de "Hierba Mora" y "Mandrágora", ambas miembros de la familia *Solanácea* (revisado por **Colvine y Branthôme, 2016**).

Para que el tomate fuera definitivamente considerado como un alimento fue necesaria la influencia política del presidente Thomas Jefferson y su fuerza de convicción para que en 1809 se produjera una revolución cultural e industrial y, casi 30 años después, otra vez en los Estados Unidos, tuvo lugar una escalada mediática acerca de las virtudes del tomate por parte del New York Times y otros medios de prensa. Pero, aun así, pasarían otros treinta años más hasta que, en 1869, Henry John Heinz fundara su compañía en Pittsburgh cuyo nombre y producto insignia (el *kétchup*) permanecerían inextricablemente ligados al tomate (revisado por **Colvine y Branthôme, 2016**).

0.1.2 | El auge del cultivo y su importancia económica mundial

Si miramos los últimos 50 años, el tomate, hoy día cultivado en casi todas las latitudes, ha experimentado un espectacular crecimiento. En este sentido, las cantidades de tomates producidos mundialmente se han multiplicado 6,18 veces, pasando de producir poco más de 27 millones de toneladas en 1961 a más de 170 en 2014 (ver Figura y Tabla 1).

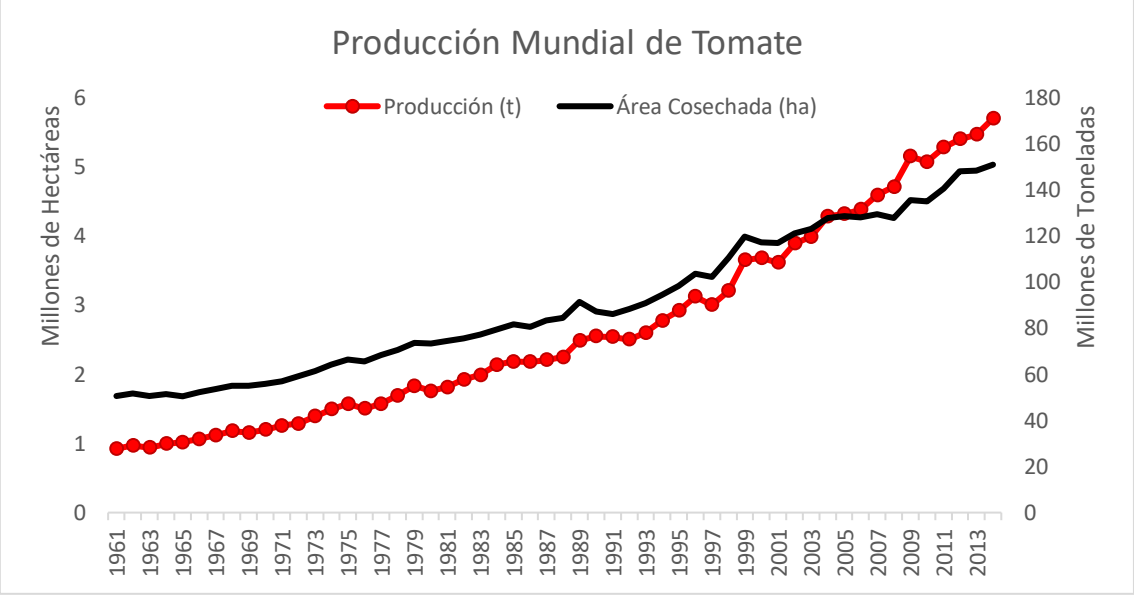


Figura 1: Producción mundial de tomates desde el año 1961 al 2014 (eje horizontal). El área cosechada (expresada en Millones de hectáreas (ha)) se indica sobre el eje vertical izquierdo (línea negra) y la cantidad de cosecha producida (expresada en Millones de toneladas (t)) se indica sobre el eje vertical derecho (línea roja). Fuente: FAOSTAT, 2017.

Tabla 1: Incremento mundial en la producción y área cosechada de tomates, al comparar los años 1961 y 2014, respectivamente. Fuente: FAOSTAT, 2017.

Año	Producción (t)	Área cosechada (ha)
1961	27.617.540	1.680.452
2014	170.750.767	5.023.810
Fracción de Incremento	6,18	2,99

t = toneladas
ha = hectáreas

Este incremento se observa aún más pronunciado en Asia y especialmente en China, el productor más grande del mundo, donde se han multiplicado por 11 los volúmenes de producción entre los años 1961 y 2014 (4.825.890 t en 1961 y 52.722.967 t en 2014), aportando, hoy día, más de un cuarto de la producción mundial. A su vez, en India la producción se ha multiplicado por 40 en el mismo periodo (464.000 t en 1961 y 18.735.910 en 2014) (ver Figura 2). Estas cifras, sin embargo, sólo tienen en cuenta la producción comercial y excluyen la agricultura familiar y la producción de subsistencia que puede ser bastante significativa en determinadas regiones (FAOSTAT, 2017).

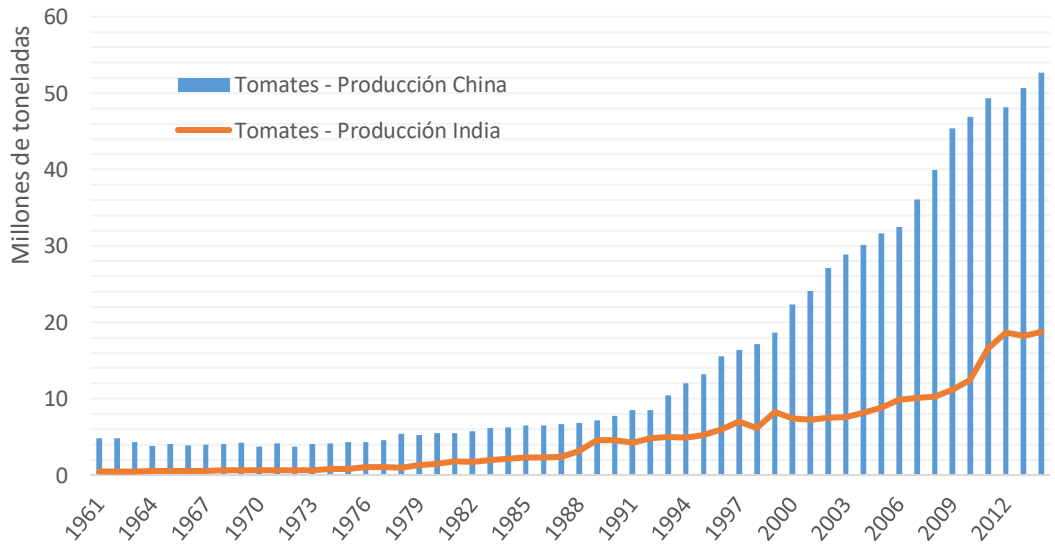
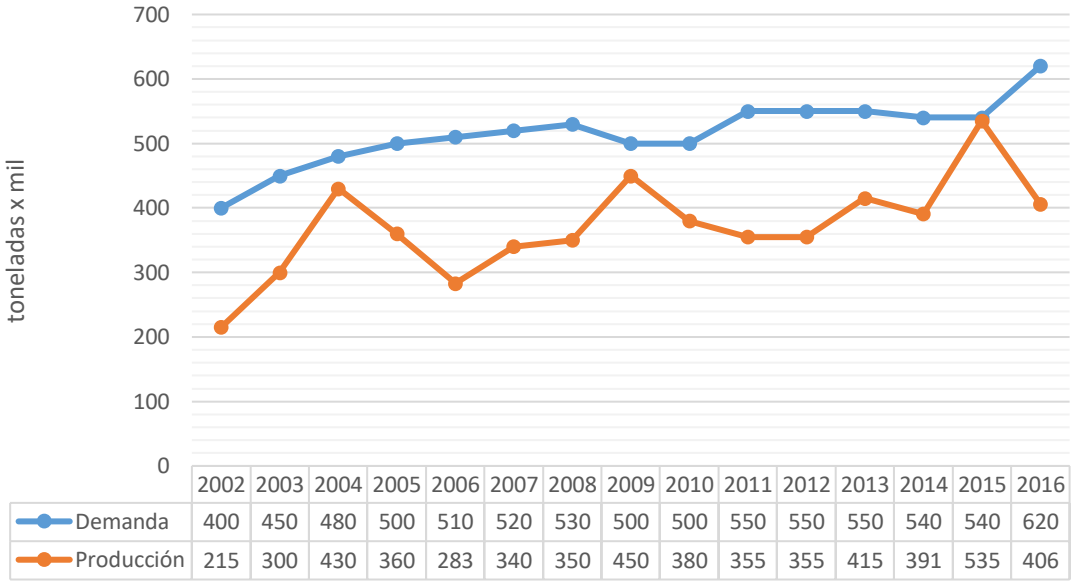


Figura 2: Producción de tomates en China (barras azules) e India (línea naranja) desde el año 1961 al 2014 expresado en millones de toneladas. Fuente: FAOSTAT, 2017.

Las razones de este crecimiento acelerado radican en una dramática mejora en la productividad agrícola, lo que permitió expandir la producción más allá de lo que se esperaría basándose sólo en aumentos de superficie existentes. Las cifras de rendimiento a nivel mundial (hectogramos por hectárea, hg/ha) obtenidas de FAOSTAT, 2017 (164.346 hg/ha en 1961 frente a 339.883 hg/ha en 2014) dan una idea aproximada de los asombrosos avances de la biotecnología y la agronomía moderna.

0.1.3 | La producción de tomate en Argentina

Según el "Informe de progresos del programa para el aumento de la competitividad de la industria del tomate" publicado por la "Asociación Tomate 2000", durante la campaña 2015-2016, la producción de tomate en Argentina alcanzó las 405.000 toneladas en 7.904 hectáreas cultivadas, siendo el rendimiento nacional medio de 51,2 t/ha. En este mismo informe se muestran un aumento progresivo de la producción y de la demanda a nivel nacional. Se puede observar que en el periodo 2002-2016 la producción se ha duplicado. Sin embargo, la demanda prácticamente no alcanza a satisfacerse y año a año, se incrementa (ver Figura 3). Cabe destacar la importante diferencia que se observa en la balanza comercial del sector, entre exportaciones e importaciones de productos industrializados que contienen tomate; por ejemplo, en 2015 se exportaron 10.599 t por un valor de 8,1 millones de dólares mientras que las importaciones sumaron un volumen de 23.750 t por un valor de 28,2 millones dólares para poder así satisfacer parcialmente la demanda de tomate (**Asociación Tomate 2000, 2016**).



	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Demanda	400	450	480	500	510	520	530	500	500	550	550	550	540	540	620
Producción	215	300	430	360	283	340	350	450	380	355	355	415	391	535	406

Figura 3: Producción (línea naranja) y Demanda (línea azul) de tomates en Argentina, observaciones medidas en toneladas x mil durante las campañas de los años 2002 a 2016. Fuente: Asociación Tomate 2000 (2002 y 2016).

0.1.4 | El tomate como un organismo modelo para estudios científicos

El espectacular crecimiento en la producción (y la consecuente relevancia económica) de la especie ha sido posible por los avances en el conocimiento acerca de la fisiología del cultivo y las bases genéticas de los principales caracteres de interés que derivan en tecnologías más eficientes para la producción. Esto quiere decir que el tomate se ha convertido en un muy buen modelo para el estudio científico (**Kimura y Sinha, 2008**). En este sentido, el tomate es un excelente modelo para estudiar el desarrollo y la maduración de los frutos carnosos y frescos. Cabe destacar, que el tomate es un fruto climatérico, con un requisito absoluto de la fitohormona etileno para que este madure. Esta dependencia del etileno ha posicionado la maduración del fruto del tomate como un sistema modelo para estudiar la regulación de la síntesis y percepción de esta fitohormona (**Klee y Giovannoni, 2011**). A su vez, este fruto es fuente de muchas moléculas con alto valor nutritivo y otras esenciales en la dieta humana (**Goff y Klee, 2006**). Se han reportado un sinnúmero de estudios que tratan de elucidar como se regula la acumulación de estos metabolitos (por ejemplo: **Akihiro et al., 2008; Osorio et al., 2011; Saito et al., 2008; Takayama & Ezura, 2015**). Más aún, se ha demostrado

que el tomate es una excelente plataforma para expresar transgenes de forma estable, motivo por el cual se lo considera un buen modelo para ensayos de ingeniería genética (**Kobayashi et al., 2003**; **Omura et al., 2007**; **Sun et al., 2007**).

A su vez, su genoma diploide, homocigota y relativamente pequeño (950 Mpb) organizado en 12 cromosomas, representa un modelo de estudio alternativo al de *Arabidopsis thaliana*, principalmente con relación a los procesos de formación y maduración de frutos carnosos y climatéricos (**Giovannoni, 2004**; **Carrari et al., 2006**; **Carrari y Fernie, 2006**). Teniendo en cuenta lo antedicho, en contraste con otras plantas modelo, como *Oryza sativa* (el arroz) y *Arabidopsis thaliana*, el tomate tiene varias características interesantes que lo diferencian de los demás. Por ejemplo, produce frutos carnosos importantes para la dieta humana (Figura 4 A) y además tiene brotes simpodiales, siendo la única planta modelo con hojas del tipo compuestas (Figura 4 B) y con una extensa cantidad de datos genómicos disponibles (**Kimura y Sinha, 2008**).

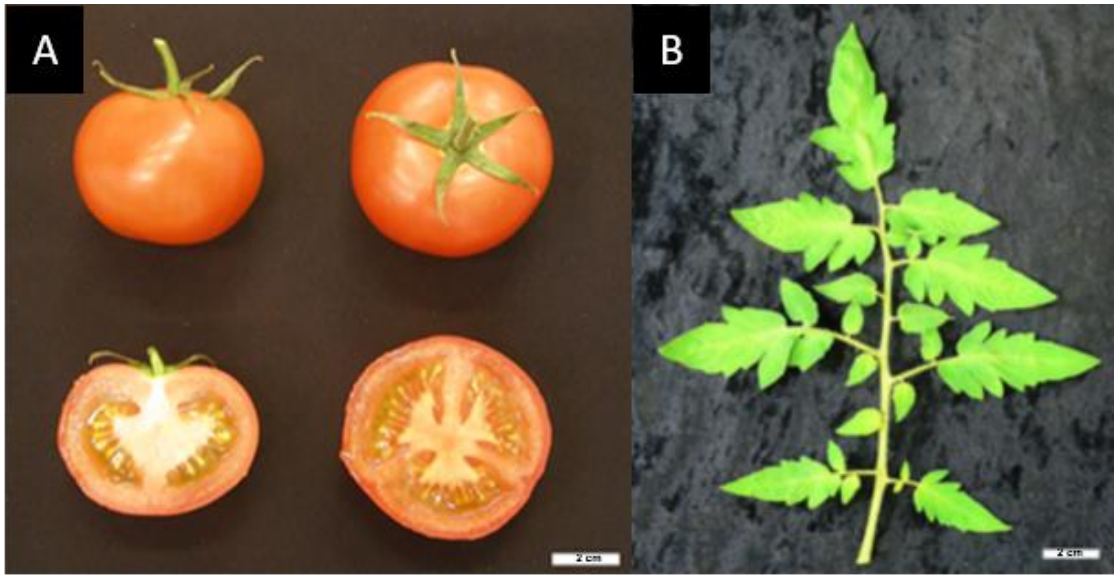


Figura 4: Fruto (A) y hojas (B) del tomate cultivado *Solanum lycopersicum*. Fuente: **Kimura y Sinha (2008)**

La mayoría de estos caracteres son de relevancia agronómica y son blanco de arduos esfuerzos para su domesticación y cultivo. En este sentido, los fenotipos de buena parte de estos caracteres han sido determinados no por un *locus* específico sino por los efectos colectivos de muchos *loci* (*loci* de caracteres cuantitativos, o *QTL* de sus siglas en inglés). En este sentido, el tomate es una planta ideal para el análisis de *QTL* porque presenta una gran variedad fenotípica de estos caracteres tanto en tomates de variedades cultivadas como en especies silvestres filogenéticamente cercanas y además se han generado variedades que combinan todo tipo de formas, colores y tamaños (**Bai y Lindhout, 2007**). Como consecuencia de todo esto, el tomate se ha convertido en un sistema modelo para el estudio de la evolución y maduración del fruto debido a (i) el desarrollo y caracterización de amplias poblaciones de mutantes (**Dellapenna et al., 1987**; **Mazzucato et al., 1998**; **Giovannoni, 2007**; **Kimura y Sinha, 2008**), (ii) la disponibilidad de herramientas genómicas, incluidas colecciones de *ESTs* en el portal http://solgenomics.net (**Bombarely et al., 2010**), mapas genéticos (**Causse et al., 2004**; **Yamamoto et al., 2005**), (iii) transformación genética eficiente y estable (**Ellul et al., 2003**; **Barone et al., 2008**) y otros métodos transitorios para evaluar la función génica (**Orzaez, 2005**; **Quadrana et al., 2011**); y por último (iv) existencia de repositorios con información centralizada y accesible a través de internet en sitios y bases de datos públicas (**Mueller et al., 2005**; **Hueso Estornell y Leandro, 2010**)]. De manera general, estas son algunas de las características que lo convierten en un buen blanco para la investigación tanto básica como aplicada a las industrias alimenticias y agropecuarias (**Ezura, 2013**).

0.1.5 | Marcadores moleculares y mapas genéticos

En 1980 el grupo de investigación del Dr. Botstein propuso la construcción de un mapa de ligamiento genético (del inglés: *linkage map*) en seres humanos basado en marcadores del tipo RFLPs, polimorfismos para el largo de los fragmentos de restricción, de sus siglas en inglés (**Botstein et al., 1980**). El éxito de su idea se ha verificado por el hecho de que los mapas genéticos por RFLPs ya se han construido, no sólo para los seres humanos, sino para una amplia variedad de otros organismos

(**O'Brien, 1993**). La mayoría de los mapas de RFLPs que habían sido publicados en esa época fueron de una densidad baja o moderada (es decir, un espaciamiento medio de marcadores ≥ 5 centimorgans (cM)) y estaban compuestos por entre 50 y 300 marcadores por genoma. Si bien estos mapas eran herramientas útiles para muchos trabajos genéticos, presentaban limitaciones inherentes que serían superadas por el desarrollo de mapas de alta densidad en los que los marcadores están espaciados a intervalos menores a lo largo del genoma. Los mapas de alta densidad pueden servir a una serie de propósitos en la investigación básica y aplicada. En primer lugar, son una herramienta clave para caminar/recorrer el cromosoma (del inglés: *chromosome walking*). En este sentido, los mapas de ligamiento de alta densidad pueden proporcionar puntos de partida para recorrer los cromosomas y acercarse prácticamente a cualquier gen del genoma. En segundo lugar, los mapas de alta densidad tienen una aplicación directa en la producción de plantas y animales ya que prácticamente aseguran que cualquier gen de interés estará estrechamente ligado al menos a un marcador molecular. Dichos ligamientos pueden explotarse para la selección basada en marcadores de genes deseables en programas de mejoramiento (**Burr et al., 1988**; **Young y Tanksley, 1989**). Finalmente, los mapas de ligamiento de alta densidad proporcionan una mayor probabilidad de que todo el genoma esté completamente cubierto con marcadores moleculares (es decir, que no haya grandes espacios sin marcas en el mapa). Este último punto es especialmente importante cuando se utilizan estos mapas para detectar y caracterizar *QTL* subyacentes en los que se necesita asegurar que todo el genoma se ha estudiado uniformemente (**Botstein et al., 1980**; **Paterson et al., 1988**). Teniendo en cuenta lo antedicho, para la especie que nos ocupa, se desarrolló una estrategia de mejoramiento asistido por marcadores y se construyeron mapas genéticos de alta densidad como Tomato-ExPen-1992 (**Tanksley et al., 1992**), Tomato-ExPen-2000 (**Fulton et al., 2002**), Kazusa-F2-2000 (**Shirasawa et al., 2010**) y TraitGenetics-ExPen-2000 (**Sim et al., 2012**). Los mapas Tomato-ExPen-1992 y 2000 fueron utilizados como punto de partida para diseñar la estrategia del secuenciamiento completo del genoma de la especie cultivada *Solanum lycopersicum* (**The Tomato Genome Consortium, 2012**) y juntos fueron los mapas utilizados en esta Tesis para el ordenamiento de

las secuencias de la especie silvestre de tomate *Solanum pennellii* que se presenta en el Capítulo II.

0.1.6 | La población de líneas de introgresión del tomate

Como en todas las especies cultivadas, uno de los objetivos más importantes en mejoramiento genético es aumentar el rendimiento. Además, se busca la obtención de variedades resistentes a factores bióticos y abióticos que no limiten los cultivos. Sin embargo, en los últimos años, el mejoramiento genético del tomate ha ampliado su actuación basada en las demandas generadas por el mercado de consumo. Para la industria, se busca aumentar los niveles de solidos solubles (medidos en grados *Brix*), compuestos principalmente por azúcares y ácidos orgánicos. Para el mercado de consumo fresco, se busca mejorar principalmente los metabolitos determinantes del aroma y del sabor del fruto, así como de la calidad nutricional (**Asprelli et al., 2017**). Sin embargo, para abordar estos desafíos, los programas de mejoramiento se enfrentan con la restricción de una estrecha base genética existente sobre el tomate, esta baja variabilidad genética es un factor limitante a la hora de buscar incrementar la productividad y el rendimiento de los cultivares elite como Heinz (LA1706) cuyo genoma mitocondrial será abordado en el Capítulo I de esta Tesis, y como el cultivar elite M82 (LA3475), cuyo genoma nuclear será discutido en el Capítulo II de esta Tesis. En este sentido, se estima que actualmente sólo el 2% de la variabilidad observada en el antiguo género *Lycopersicon* se encuentra en la especie cultivada del tomate (**Fray y Grierson, 1993**). Frente a este panorama, fuentes de germoplasma silvestre, filogenéticamente lejanas, como *Solanum pennellii* (ver Figura 5), han resultado sustancialmente útiles para la incorporación de nuevos caracteres en los tomates cultivados. En este sentido, existen varios ejemplos de la incorporación de genes de especies silvestres en tomates cultivados que, confieren resistencia a diferentes patógenos del cultivo como el gen *MI-1* de resistencia al nematodo del nudo de la raíz (*Meloidogyne spp*) detectado en el cromosoma 6 del tomate silvestre (**Seah et al., 2004**), otro ejemplo es el gen *PTO* que confiere resistencia a la infección bacteriana por *Pseudomonas syringae-pv* (**Martin et al., 1993**) o como

los factores de iniciación de la traducción *EIF4E* y *EIF4G* que confieren resistencia a infecciones por *Potyvirus* (**Piron et al., 2010**).

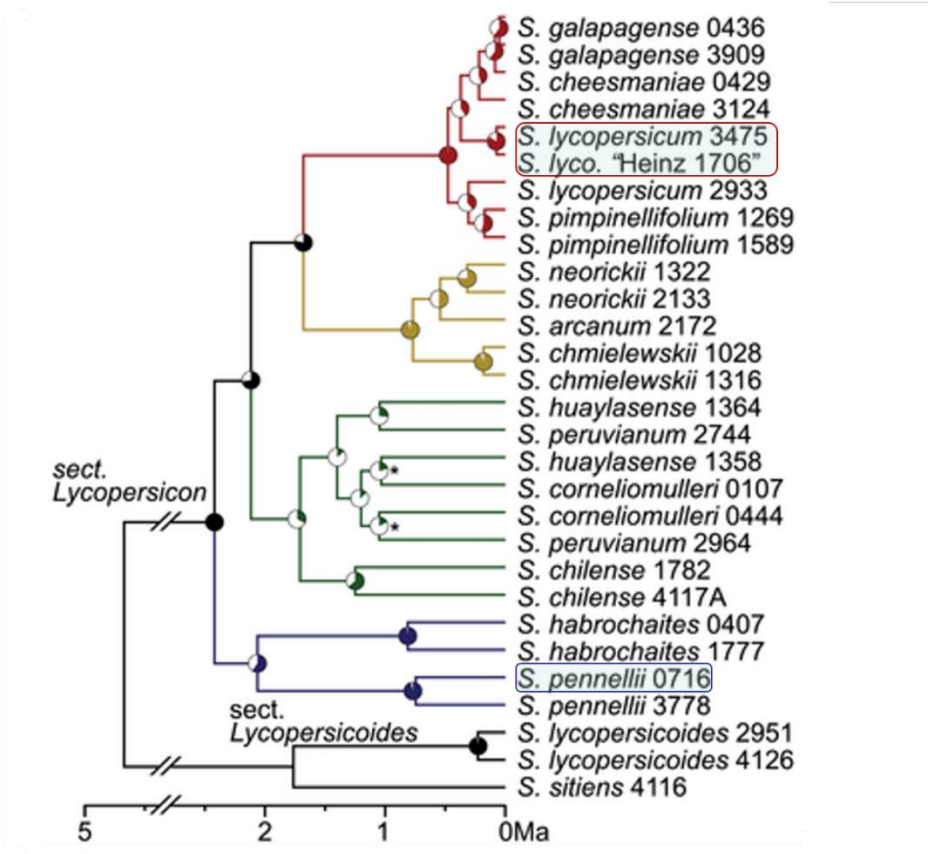


Figura 5: La filogenia del género *Solanum,* sección *Lycopersicon*. En el recuadro rojo se observan agrupados los cultivares Heinz (LA1706) y M82 (LA3475) de la especie *Solanum lycopersicum.* Por otro lado, en el recuadro purpura, se observa la especie silvestre *Solanum pennellii* (LA0716) la cual se encuentra a una distancia filogenéticamente lejana de la especie cultivada. Fuente: (**Pease et al., 2016**)

En resumen, con el fin de aumentar la variabilidad genética que ha ido perdiendo el tomate cultivado, en 1995, los investigadores Eshed y Zamir de la Universidad Hebrea de Jerusalén desarrollaron una colección de 76 líneas de introgresión (*ILs*), en las cuales, segmentos definidos del genoma de la especie silvestre de tomate *Solanum pennellii* sustituyen regiones homólogas en el fondo genético de la especie cultivada *Solanum lycopersicum* (M82) (ver Figura 6). Como resultado, estás 76 *ILs* cubren todo el genoma de la especie cultivada y los segmentos introgresados están

bien delimitados por medio de marcadores moleculares, lo cual resultó en 107 unidades superpuestas entre las *ILs*. Estas unidades, llamadas *BINs,* contienen segmentos del alelo silvestre compartido entre las *ILs* contiguas y abarcan todo el genoma (**Eshed y Zamir, 1995**; **Pan et al., 2000**) (ver Figura 7).

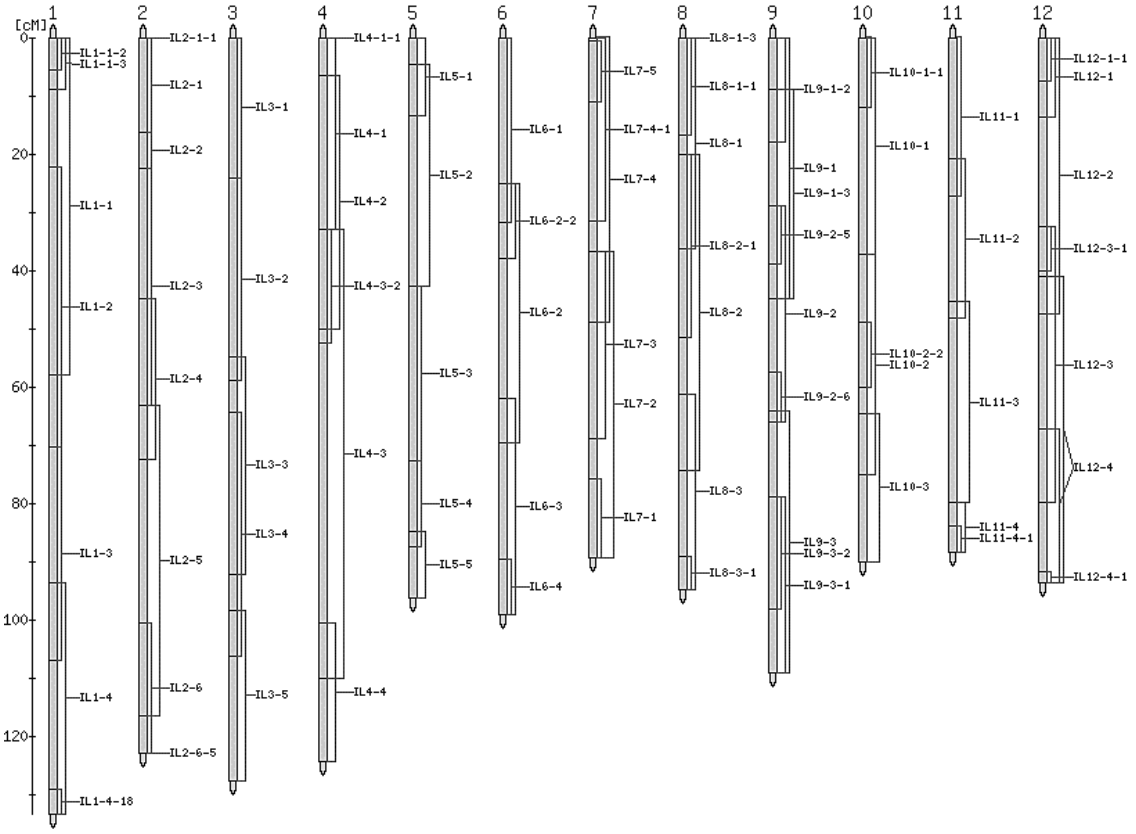


Figura 6: Mapa ilustrativo de la población de 76 líneas de introgresión provenientes del cruzamiento entre la especie silvestre *S. pennellii* (LA0716: como padre donante) y la especie domesticada *S. lycopersicum* (cv. M82: como padre recurrente), desarrolladas por Eshed y Zamir (1995). Sobre los diagramas de los 12 cromosomas de *S. lycopersicum* (barras grises oscuras) se encuentran delimitados por marcadores moleculares segmentos definidos del genoma de la especie silvestre *S. pennellii* (corchetes etiquetados) enumerados desde la primera línea (IL-1-1) que contiene una introgresión sobre el cromosoma 1, hasta la última, con una introgresión sobre el cromosoma 12 (IL-12-5). Fuente: *Tomato IL map*, http://solgenomics.net.

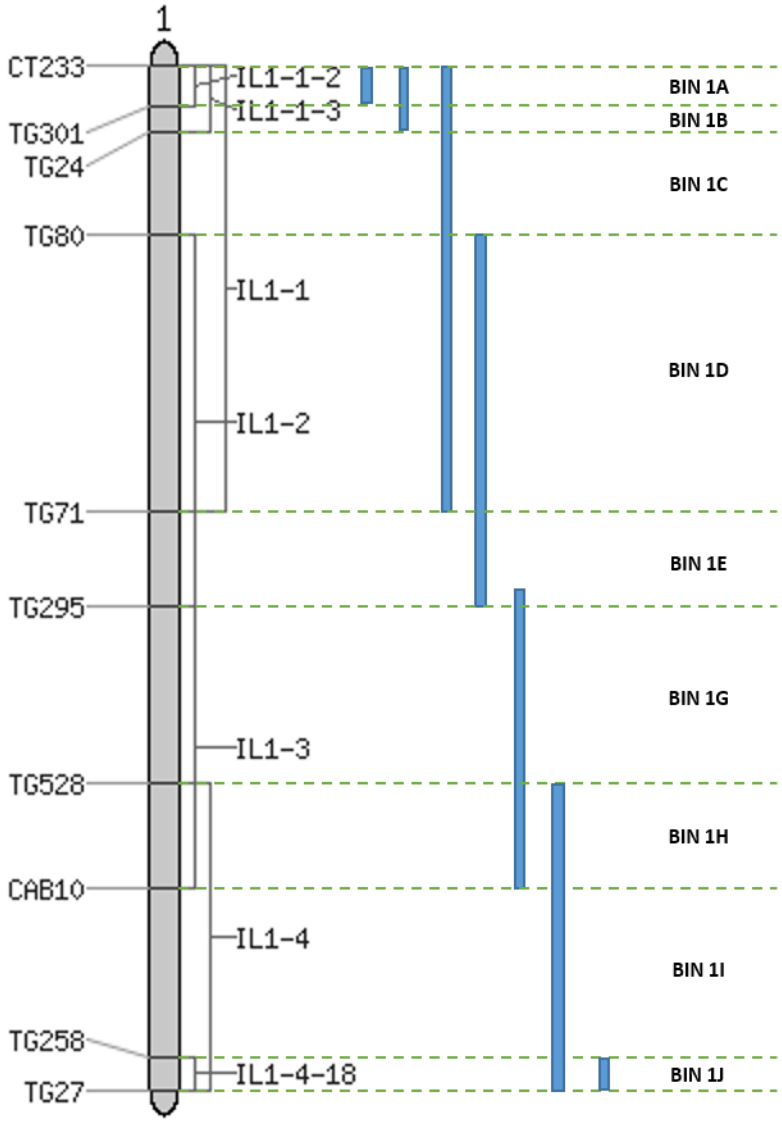


Figura 7: Mapa ilustrativo de las 7 líneas de introgresión que abarcan al cromosoma 1 (IL1-1 a la IL-1-4). Los segmentos introgresados de *S. pennellii* en el fondo genético de *S. lycopersicum* de cada IL se ilustran con barras de color celeste. En este cromosoma se observan 9 segmentos entre las *ILs* (demarcados por líneas punteadas de color verde) que se denominan *BIN*s, para este caso desde el *BIN* 1A al 1J se cubre todo el cromosoma 1. Fuente: **Pan et al. (2000)**.

En este sentido, la población de *ILs* de tomate ha sido ampliamente caracterizada con el fin de identificar variaciones para caracteres de interés agronómico. Utilizando esta población se han mapeado más de 2.000 *QTL* (**Lippman et al., 2007**). Algunos ejemplos los constituyen el mapeo de los determinantes genéticos de metabolitos relacionados con la calidad nutricional y la presencia de antioxidantes en el fruto (**Rousseaux et al., 2005**), la detección de 130 *QTL* para 38 caracteres relacionados con la calidad organoléptica y 81 *QTL* asociados al tamaño, contenido de ácidos orgánicos y azúcares del fruto (**Causse, 2002**; **Causse et al., 2004**) y más recientemente en nuestro laboratorio se identificaron regiones genómicas asociadas a los contenidos de vitamina E en los frutos (**Almeida et al., 2011**) y se describieron los mecanismos de regulación de uno de los *QTL*s mayores (**Quadrana et al., 2013**, **2014**). Debido a su estructura genómica, las *ILs* se presentan como una herramienta muy interesante para elucidar las bases genéticas detrás de los *QTL*, ya que todas las diferencias fenotípicas entre las 76 líneas y el cultivar elite M82 (padre recurrente) están asociadas al fragmento introgresado de la especie silvestre *Solanum pennellii* y/o a la interacción de estas con el resto del genoma de la especie cultivada. El primer ejemplo que demostró la utilidad de estos materiales para la identificación de los determinantes genéticos de caracteres de calidad de los frutos fue publicado por (**Fridman et al. 2000, 2004**). En estos trabajos se identificó a la IL 9-2-5 con incrementos significativos (aprox. 25%) en los grados *Brix*. Luego del análisis genético y el mapeo de alta resolución de este *QTL*, se identificó un gen que codifica para una enzima de tipo invertasa de pared celular; *LIN5*. La comparación entre los alelos (silvestre y cultivado) permitió concluir que el *QTL* estaba determinado por el cambio de un único nucleótido en el tercer exón del gen, siendo ese cambio el causante de modificaciones en las propiedades cinéticas de la enzima. Está invertasa es clave en la carga de sacarosa en el floema, siendo este el mayor factor involucrado en el incremento del contenido de azúcar en el fruto (**Baxter et al., 2005**).

Históricamente, la caracterización de *QTL*, como fue descripta más arriba, era realizada por medio de largos procesos de mapeo genético, seguidos de asignaciones físicas para finalmente llegar al clonado posicional del/los gen/es. La disponibilidad de información sobre secuencias genómicas y datos de expresión

génica, así como de mapas de alta densidad, permitió proponer la estrategia de "genes candidatos" como una alternativa al clonado posicional. Este enfoque consiste en la identificación de genes que cosegregan con los caracteres de interés y que, por la función de sus productos, pueden contribuir o determinar las variaciones observadas (**Grandillo y Cammareri, 2016**).

La candidatura de un gen puede ser propuesta a partir de un conjunto de evidencias tales como: i) si el gen es expresado en las mismas condiciones temporales y espaciales que el *QTL*; ii) si los alelos presentes en los parentales de la población de mapeo presentan patrones de expresión diferenciales o; iii) si los alelos parentales presentan polimorfismos no silenciosos. No obstante, la causalidad debe ser probada por medio de análisis funcionales utilizando diversas herramientas de genética reversa; muchas de ellas basadas en mutagénesis o en el silenciamiento génico específico por medio de tecnologías basadas en ARN de interferencia (del inglés: *RNAi*) (**Rothan y Causse, 2007**).

0.1.7 | Secuenciación de ácidos nucleicos: perspectiva histórica y últimos avances

La determinación del orden de los residuos de ácidos nucleicos en muestras biológicas es un componente integral para una amplia variedad de aplicaciones en investigación. Durante los últimos cincuenta años un gran número de investigadores se han abocado al desarrollo y aplicación de tecnologías para facilitar la secuenciación de moléculas de ADN y ARN. Durante esta corta escala de tiempo se han experimentado enormes cambios en el campo de la secuenciación, pasando de la secuenciación de oligonucleótidos cortos a millones de bases, del desafío para deducir la secuencia codificante de un solo gen hasta la rápida y ampliamente disponible secuenciación de genomas completos (**Heather y Chain, 2016**). Dado que en esta Tesis se utilizaron y analizaron datos de secuenciación de ácidos nucleicos obtenidos por distintas técnicas, a continuación, se describen brevemente las mismas a modo de reseña.

0.1.8 | Secuenciación de primera generación

Método de Sanger, secuenciación por didesoxinucleótidos trifosfato (ddNTPs) como terminadores de la cadena de ADN.

Desde que Watson y Crick resolvieron la famosa estructura tridimensional del ADN en 1953 (**Watson y Crick, 1953**), el gran avance que cambió para siempre el progreso de la tecnología de secuenciación del ADN llegó en 1977, con el desarrollo de la técnica de secuenciación por terminador de cadena utilizando didesoxinucleótidos, comúnmente llamada "secuenciación de primera generación" o simplemente "método de Sanger" (**Sanger et al., 1977**). Está técnica hace uso de análogos químicos de los desoxirribonucleótidos (dNTPs: Adenina, Guanina, Timina y Citosina - A, C, T y G, respectivamente), que son los didesoxinucleótidos (ddNTPs: ddGTP, ddATP, ddTTP y ddCTP) que al carecer del grupo hidroxilo 3' que se requiere para la extensión de las cadenas de ADN, no permiten la formación del enlace con el fosfato 5' del siguiente dNTP (**Chidgeavadze et al., 1984**) y terminan la reacción. La mezcla de ddNTPs marcados en una reacción de extensión de la cadena de ADN, utilizando una fracción de la concentración de dNTPs tradicionales, da lugar a que se produzcan cadenas de ADN de cada longitud posible, porque a medida que los didesoxinucleótidos se incorporan al azar mientras que la hebra se extiende, se detiene la progresión de la síntesis. Teniendo esto en cuenta, mediante la realización de cuatro reacciones en paralelo, conteniendo cada base de ddNTP individual, es posible procesar los resultados en cuatro carriles de geles resolutivos (como los de poliacrilamida), y puede utilizar la autorradiografía para inferir cuál es la secuencia de nucleótidos en la hebra original, ya que habrá una banda radioactiva en el carril correspondiente en esa posición del gel. Debido a su precisión y robustez, este método se convertiría en la tecnología más utilizada para secuenciar el ADN en los primeros años del auge de la biología molecular (1980 en adelante) hasta el nuevo milenio.

Se realizaron una serie de mejoras en la tecnología de secuenciación de Sanger original; estas implicaron principalmente la sustitución del fosforo- o tritio-radiomarcado por detección fluorométrica (evitando los efectos indeseados del uso de radioisótopos) y una mejor detección mediante electroforesis basada en

capilares. Ambas mejoras contribuyeron al desarrollo de máquinas de secuenciación de ADN cada vez más automatizadas (**Smith et al., 1985; Ansorge et al., 1986, 1987; Prober et al., 1987; Kambara et al., 1988; Luckey et al., 1990; Swerdlow y Gesteland, 1990**) y por consiguiente se produjo una oleada de máquinas comerciales de secuenciación de ADN (**Hunkapiller et al., 1991**) que se utilizaron para secuenciar los genomas de especies cada vez más complejas. El ejemplo emblemático es la publicación del primer borrador completo del genoma humano (**Lander et al., 2001**; **Venter et al., 2001**) obtenido enteramente utilizando esta tecnología.

Estas máquinas de secuenciación de ADN de primera generación producen lecturas de una longitud ligeramente inferior a una Kilobase (Kb). Para conseguir fragmentos más largos, se utilizan técnicas como la "secuenciación por fragmentación aleatoria" (denominada en inglés: *shotgun sequencing*), en la que los fragmentos de ADN solapantes se clonan y secuencian por separado y luego las distintas lecturas se ensamblan *in silico* en una larga secuencia contigua, llamada "*contig*" en el léxico de la bioinformática (**Staden, 1979**; **Anderson, 1981**). Esto es posible justamente por la secuenciación de un gran número de fragmentos solapantes, cuyo orden no es *a priori* conocido.

El desarrollo de técnicas moleculares como la reacción en cadena de la polimerasa (*PCR*, de sus siglas en inglés) (**Saiki et al., 1985**, **1988**) y las tecnologías de ADN recombinante (**Jackson et al., 1972**; **Cohen et al., 1973**) ayudaron a la revolución de la genómica, proporcionando medios para generar altas concentraciones de ADN puro necesarias para la secuenciación. Eventualmente, los secuenciadores por didesoxinucleótido terminador-de-cadena más recientes, -como la gama ABI PRISM producida por *Applied Biosystems*® (**Smith et al., 1986**), permitieron la secuenciación simultánea de cientos de muestras (**Ansorge, 2009**)- y fueron los que se utilizaron para producir el primer borrador del genoma humano (citado más arriba) en menor tiempo del estimado originalmente. Por otra parte, la secuenciación por Sanger también se utilizó para secuenciar genomas de varias especies de plantas como *Arabidopsis thaliana* (**The Arabidopsis Genome Initiative. et al., 2000**), arroz (*Oryza sativa* (**Goff et al., 2002a**; **Yu et al., 2002b**)), sorgo (*Sorghum bicolor* (**Paterson et al., 2009**)), vid (*Vitis vinifera*

(**Jaillon et al., 2007**)), álamo (*Populus trichophora* (**Tuskan et al., 2006**)) y soja (*Glycine max*; (**Schmutz et al., 2010**)), entre otras especies. Sin embargo, el tiempo que toma este método de secuenciación y los costos asociados a éste, evitaron la posibilidad de secuenciar un gran número de especies de plantas, especialmente aquellas con genomas grandes y complejos (por ejemplo el trigo, cuyo genoma se estima en un tamaño de ~17 Gpb (**The International Wheat Genome Sequencing Consortium, 2014**; **Montenegro et al., 2017**).

0.1.9 | Secuenciación de segunda generación (*SGS*)

Breve reseña de las tecnologías de alto rendimiento

Paralelamente al desarrollo de los esfuerzos de secuenciación de Sanger a gran escala, se desarrolló otra técnica que preparó el escenario para la primera ola de secuenciadores de ADN de la siguiente generación (*NGS*, por sus siglas en inglés: *Next Generation Sequencing*). Este método difiere marcadamente de los métodos existentes en el sentido de que no se deduce la identidad de nucleótidos mediante el uso de dNTPs u oligonucleótidos marcados con radioactividad o fluorescencia antes de visualizarlos con electroforesis (**Heather y Chain, 2016**). En su lugar, se utiliza un método luminiscente para medir la síntesis de pirofosfato: Esto consiste en un proceso de dos enzimas en el cual i) la ATP sulfurilasa se utiliza para convertir pirofosfato en ATP, y ii) luego este se utiliza como substrato para la luciferasa, por lo tanto, la producción de luz será proporcional a la cantidad de pirofosfato disponible (**Nyrén y Lundin, 1985**). Este enfoque se utilizó para inferir secuencias midiendo la producción de pirofosfato a través del sistema a medida que cada nucleótido es lavado por turnos sobre el molde de ADN que está fijado a una fase sólida (**Hyman, 1988**).

Esta técnica, llamada pirosecuenciación, iniciada por Pål Nyrén y sus colegas, poseía una serie de características que se consideraron beneficiosas: podría realizarse utilizando nucleótidos naturales (en lugar de los dNTP pesadamente modificados utilizados en los protocolos de terminación de cadena) y observados en tiempo real (en lugar de requerir largas electroforesis) (**Nyrén, 1987**; **Ronaghi**

et al., 1996, **1998**). Mejoras posteriores incluyeron la unión del ADN a perlas/esferas paramagnéticas seguido de la degradación enzimática de dNTPs que no pudieron incorporarse durante la síntesis, eliminando así la necesidad de largas etapas de lavado. La principal dificultad que plantea esta técnica es averiguar cuántos casos del mismo nucleótido hay en fila en una posición dada (p. ej.: AAAAA): la intensidad de la luz liberada corresponde a la longitud del homopolímero, pero el ruido produce una lectura no lineal por encima de cuatro o cinco nucleótidos idénticos (**Ronaghi, 1998**). Más tarde, la pirosecuenciación finalmente se licenció a "*454 Life Sciences®*", una compañía de biotecnología fundada por Jonathan Rothburg, donde evolucionó como el primer gran éxito comercial de la tecnología de *NGS* (**Heather y Chain, 2016**).

Posteriormente, esta tecnología se amplió a la nueva versión del instrumento *454*, el GS FLX Titanium (www.454.com/products/). Paralelamente, varias compañías lanzaron sistemas *NGS* competidores que incluyeron a "Solexa 1G" (más tarde llamado "*Genome Analyzer*"), GA, GA II, HiSeq 2000, HiSeq 1000, Hi ScanSQ y MiSeq de la compañía *Illumina®* (www.illumina.com) y SOLiD 3 y SOLiD 4 por *Applied Biosystems®* (www.appliedbiosystems.com). Actualmente, estas tecnologías se denominan sistemas de Secuenciación-de-Segunda-Generación (o *SGS*, de sus siglas en inglés) (**Thudi et al., 2012**). Aunque todos estos sistemas se pueden utilizar para una multitud de aplicaciones, en particular, para la investigación en genómica de plantas (**Varshney et al., 2009**), *Illumina®* y *Roche/454®* han proporcionado las plataformas *SGS* más ampliamente adoptadas como es evidente por el gran número de publicaciones aparecidas en los últimos años citándolas (**Thudi et al., 2012**). Sin embargo, cabe mencionar, que la necesidad de un proceso de amplificación previo puede introducir errores en la secuencia de la hebra utilizada como templado, así como también se puede introducir sesgo de amplificación. Por otra parte, en comparación con la secuenciación por el método de Sanger, el promedio de longitud de las lecturas producidas por plataformas *SGS* es, hoy día, más corto que el método de Sanger (< 1000 pb) (**Metzker, 2010**; **Schatz et al., 2010**). Basado en el rendimiento obtenido, excepto por la máquina MiSeq, todas las demás plataformas de secuenciación de *Illumina®* (GA, GA II, HiSeq 2000, HiSeq 1000 y Hi ScanSQ), *Applied Biosystems®*

(SOLiD 3 y SOLiD 4) y *Roche/454®* (GS 20 / FLX, GS FLX Titanium, GS FLX Titanium XLR70 y GS Junior) se consideran plataformas *SGS* de alto rendimiento ya que son capaces de generar de tres a cuatro órdenes de magnitud más secuencia de ADN que la secuenciación basada en el método de Sanger o de primera generación (**Thudi et al., 2012**).

Siguiendo la evolución de lo antedicho, el dispositivo *"GS FLX" de Roche/454®* producía más de un millón de lecturas de hasta 1.000 bases por cada 10 horas de trabajo, con un rendimiento total de 400 a 600 Megabases (Mb). Así, el secuenciador 454 tenía las lecturas cortas de mayor longitud entre todas las plataformas SGS. Por otro lado, el secuenciador *"Genome Analyzer"* de *Illumina®* producía más de cien millones de lecturas cortas de alta calidad (de hasta 150 bases de longitud) tomando 3 a 5 días de trabajo, obtenía así varias Gigabases (Gb) de secuencia (www.illumina.com). Hasta la fecha, la mayoría de los artículos de *NGS* publicados han descrito métodos utilizando los datos de secuencia producidos por el "*Genome Analyzer"* (**Thudi et al., 2012**). De hecho, en la actualidad, el *Illumina* "*Genome Analyzer*" HiSeq 2000 es capaz de producir lecturas individuales de 2100 bases de lecturas-pareadas (del inglés, paired-end reads) y genera unos 200 Gb de secuencias cortas por ciclo. Como resultado, la división *Roche/454®* fue cerrada por Roche en el año 2013 cuando su tecnología se volvió no competitiva (**Hollmer, 2013**).

La gran cantidad de datos generados por estas tecnologías *SGS* de alto rendimiento plantea un desafío para el almacenamiento, transferencia y posterior análisis de los mismos (**Thudi et al., 2012**). Esto es especialmente cierto para las lecturas más cortas generadas por los sistemas *Illumina* y *ABI SOLiD* que hacen que los procesos de alineamiento de lecturas y ensamblaje de secuencias sean un gran desafío (**Schatz et al., 2010**). Sin embargo, las tecnologías de *SGS* se están utilizando para una variedad de aplicaciones incluyendo secuenciación *de novo* de genomas, análisis de transcriptomas, expresión diferencial de genes, descubrimiento de marcadores moleculares y muchas otras aplicaciones en especies vegetales de

importancia estratégica como el cacao (**Argout et al., 2011**), el garbanzo/*chickpea* (**Garg et al., 2011**; **Hiremath et al., 2011**), las arvejas/*pigeon peas* (**Dubey et al., 2011**; **Varshney et al., 2011**; **Kudapa et al., 2012**), la palmera de dátiles/*date palm* (**Al-Dous et al., 2011**) y los guisantes/*peas* (**Franssen et al., 2011**).

La recientemente descrita "revolución genómica", impulsada en gran parte por estos notables cambios en la tecnología de secuenciación de nucleótidos, ha alterado drásticamente los costos asociados. Las capacidades de los secuenciadores de ADN han crecido a un ritmo aún más rápido que el observado en la revolución informática descrita por la ley de Moore: la complejidad de los microchips (medidos por el número de transistores por unidad de costo) se duplica aproximadamente cada dos años; en 2010 se duplicó cada cinco meses (**Stein, 2010**). Las diversas tecnologías derivadas son diversas en sus características químicas, capacidades y especificaciones, proporcionando a los investigadores una diversa caja de herramientas con la cual diseñar experimentos. Sin embargo, en los últimos años la plataforma de secuenciación de *Illumina*® ha sido la más exitosa, hasta el punto de monopolizar el mercado (**Schadt et al., 2010**) y, por lo tanto, probablemente puede considerarse como la mayor contribución a la segunda generación de secuenciadores de ADN (**Heather y Chain, 2016**).

Lecturas simples (del inglés, *SE: single-end*) y lecturas pareadas (del inglés, *PE: paired-end* y *MP: mate-pair*)

Es imposible resolver repeticiones de un largo X a menos que tengas lecturas más largas que X (ver Figura 8). Para el ensamblaje de lecturas cortas (< 100 pb) en *contigs* y *scaffolds* es crítico el uso de protocolos de secuenciación de lecturas pareadas, las cuales se introdujeron por primera vez a principios de la década de 1990 para su uso en la secuenciación de Sanger (**Roach et al., 1995**). En el contexto de la secuenciación por fragmentación aleatoria utilizada para el genoma humano, la importancia de utilizar lecturas con extremos pareados para extender *contigs*

por sobre regiones repetitivas del ADN fue enfatizada por Weber y Myers (**Weber y Myers, 1997**).

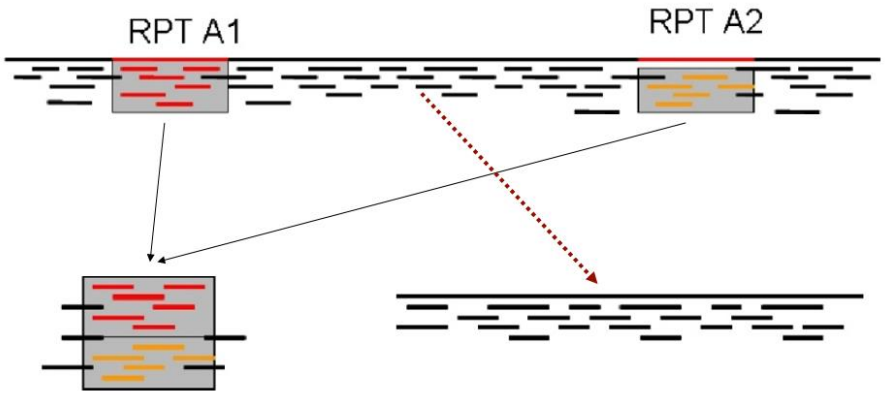


Figura 8: Efecto de las repeticiones sobre el ensamblado de lecturas. El elemento repetitivo RPT A es colapsado en un único "*contig*" de corta longitud (abajo a la izq.) y el resultado son dos *contigs* truncados en vez de ensamblar una sola secuencia contigua flanqueada por repeticiones. Fuente: www.slideshare.net/AustralianBioinformatics/torsten-seemann-de-novo-genome-assembly.

Este enfoque se convirtió en el paradigma dominante para la secuenciación de genomas en la última década. El emparejamiento de lecturas permite el ensamblado de regiones repetitivas distribuidas a lo largo de todo el genoma, dado que las lecturas que derivan de una repetición particular a menudo se pueden inferir si sus lecturas pareadas se encuentran en secuencias flanqueantes únicas (**Chapman et al., 2011**). Con las cortas lecturas de *SGS*, las ventajas del enfoque de extremos pareados se acentúa (**Chaisson et al., 2009**), y esta estrategia es aprovechada por los programas de ensamblado de lecturas de *SGS* que se han desarrollado recientemente (incluyendo *EULER-SR* (**Chaisson y Pevzner, 2008**), *Velvet* (**Zer*BIN*o y Birney, 2008**; **Zer*BIN*o et al., 2009**), *ALLPATHS* (**Butler et al., 2008**; **MacCallum et al., 2009**), *ABySS* (**Simpson et al., 2009**) y *SOAPdenovo* (**Li et al., 2010**)). Todos estos ensambladores utilizan grafos de *De Bruijn* para representar el problema del ensamblado (**Pevzner et al., 2001**), en el que las lecturas se descomponen en palabras superpuestas de longitud k (k-meros), donde k es una fracción de la longitud de la lectura a ensamblar (**Chapman et al., 2011**). A continuación, se detallarán las dos estrategias utilizadas en esta Tesis para generar lecturas pareadas (*PE* y *MP*).

Lecturas pareadas del tipo *PE*

Como se mencionó anteriormente, la secuenciación por lecturas pareadas es una estrategia para obtener secuencias de los dos extremos del mismo fragmento de ADN y generar datos de alta calidad que puedan alinearse con alto grado de confiabilidad. Esta estrategia facilita la detección de reordenamientos genómicos y elementos de secuencia repetitivos. La secuenciación por lecturas pareadas permite, por ejemplo, que se secuencien ~150 pb de ambos extremos de un fragmento de ADN de 600 pb y debido a que se conoce *a priori* el tamaño del fragmento y por ende la distancia que habrá entre cada lectura pareada (del inglés: *insert-size, insert-length* o *library-size*), los algoritmos de alineamiento pueden usar esta información para mapear las lecturas sobre regiones repetitivas de manera más precisa (ver Figura 9). Esto da lugar a un mejor alineamiento de las lecturas, especialmente a través de regiones repetitivas difíciles de secuenciar del genoma como son las regiones de transposones, elementos satélites en tándem o en la presencia de familias génicas (genes parálogos) (***Illumina*, 2017**).

Figura 9: Alineamiento de lecturas pareadas. La secuenciación pareada permite secuenciar ambos extremos de un fragmento de ADN. Como *a priori* se conoce la distancia entre cada lectura *PE* (línea punteada), los algoritmos de alineamiento pueden usar esta información para mapear las lecturas sobre repeticiones (RPT A1 y A2) con mayor precisión evitando colapsar y truncar estas regiones, como se observa en la Figura 8. Fuente: (***Illumina*, 2017**).

Lecturas pareadas del tipo *MP*

La secuenciación de lecturas pareadas antes descrita permite que se secuencien alrededor de 150 nt de ambos extremos del fragmento de ADN y debido a que se conoce la distancia que hay entre cada lectura, los algoritmos de alineamiento y ensamblado pueden usar esta información para mapear las lecturas sobre regiones

repetitivas de manera precisa. En este sentido, la limitante radica en que el tamaño del fragmento a secuenciar por *SGS* no producirá un resultado óptimo si este fragmento es de un tamaño mayor a unos 800 pb (***Illumina*, 2017**). Teniendo en cuenta lo antedicho se diseñó un método que permite secuenciar por lecturas pareadas grandes fragmentos de ADN de alrededor de 2 - 40 Kpb, y a este tipo de lecturas se las denominó *MP* (del inglés: *Mate-Pair*). Mediante la unión de adaptadores a los extremos terminales del fragmento de ADN, a secuenciar, la molécula lineal se circulariza y luego se vuelve a fragmentar y se purifica el fragmento de interés, que ahora pasa a tener un tamaño ≤ 800 pb y se puede proceder con la secuenciación de manera similar que con el método de lecturas *PE* antes descrito. De esta manera se logran secuenciar los extremos que sabemos *a priori* estaban a una distancia de ~ 2 – 40 Kpb (ver Figura 10).

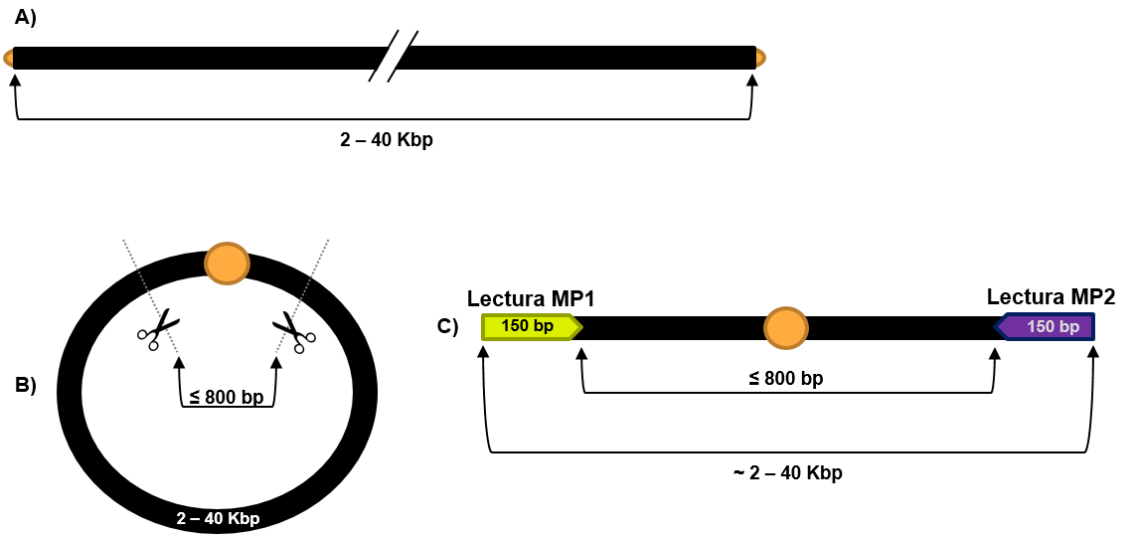


Figura 10: A) Fragmento lineal de ADN a secuenciar de entre 2 a 40 Kpb con adaptadores *Illumina® MP* en sus extremos (semi-círculos naranjas). **B)** Fragmento de ADN circularizado y luego fragmentado con un tamaño ≤ 800 pb quedando enfrentados los extremos de la molécula lineal a secuenciar. **C)** Fragmento ≤ 800 pb secuenciado por el método de lecturas *PE* con la salvedad de que se sabe de antemano que la distancia entre los pares-de-lecturas es de ~ 2 – 40 Kpb. Fuente: (***Illumina*, 2017**).

Lecturas simples (*SE*)

Este tipo de secuencias se obtienen a partir de secuenciar fragmentos de ADN en una sola dirección (ver Figura 11) y su tamaño depende de la tecnología utilizada (*Illumina®*, *Roche®/454*, Sanger, etc.). Por otro lado, las lecturas *SE* pueden

obtenerse durante el preprocesamiento de las lecturas *PE* y *MP* donde se filtran por calidad y contaminación pudiendo entonces quedar lecturas desemparejadas (del inglés: *Unpaired* o simplemente *U*).

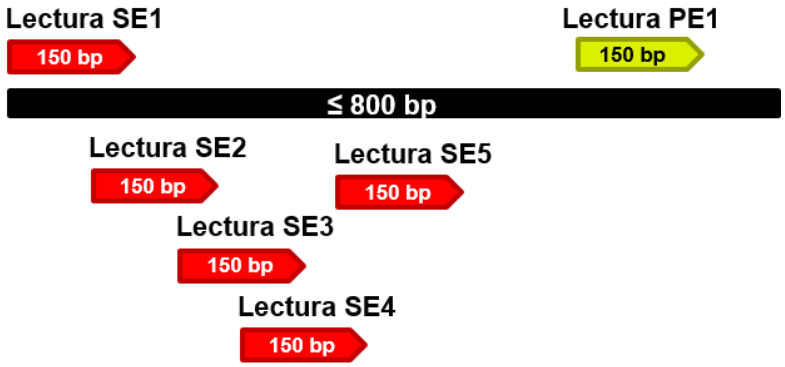


Figura 11: Fragmento Lineal de ADN con un tamaño menor a 800 pb (línea negra) secuenciado por lecturas *SE* (en color rojo) y una lectura *MP* que quedó desemparejada luego de ser filtrada por calidad en el preprocesamiento (en color verde). Fuente: (***Illumina*, 2017**)

El uso de una combinación de secuenciación de fragmentos de ADN de tamaños cortos y largos resulta en una cobertura máxima del genoma para ser ensamblado *de novo*. Debido a que los fragmentos más grandes pueden estar emparejados a través de mayores distancias con lecturas *MP*, estos proporcionan una mayor capacidad para desenmarañar secuencias altamente repetitivas y regiones donde se han producido grandes reordenamientos estructurales. Por otro lado, los fragmentos más cortos secuenciados a profundidades más altas (*SE* y *PE*) pueden llenar huecos (del inglés, *Gaps*) que no hayan sido cubiertos por las lecturas *MP* secuenciadas a profundidades más bajas. Por lo tanto, una biblioteca diversa de fragmentos cortos y largos da como resultado un mejor ensamblado *de novo*, dando lugar a menos *gaps*, *contigs* y *scaffolds* más grandes y por lo tanto una mayor precisión de la secuencia consenso final (***Illumina*, 2017**).

0.2 | Genomas de plantas

En el año 2000 la secuenciación del genoma de la planta modelo *Arabidopsis thaliana* (**The Arabidopsis Genome Initiative. et al., 2000**) fue un hito importante no sólo para la investigación de plantas, sino también para la secuenciación de genomas en general. Fue uno de los primeros genomas de organismos multicelulares que se completaron, y fue secuenciado por un gran consorcio multinacional para hacer frente a este enorme desafío. Como consecuencia, esta herramienta, hasta entonces sin precedentes, estimuló y aceleró la investigación genómica en plantas. El enfoque elegido para esta iniciativa se basó en la superposición de cromosomas artificiales bacterianos (*BACs*, de sus siglas en inglés) que representan el camino de mosaicos más corto posible (del inglés: *minimal tiling path*) para cubrir cada brazo del cromosoma. Las secuencias de *BACs* se ensamblaron individualmente y se organizaron de acuerdo con el mapa físico de *Arabidopsis* lo que culminó con la creación de una secuencia genómica borrador de muy alta calidad (**The Arabidopsis Genome Initiative. et al., 2000**). El esfuerzo y el tiempo prohibitivamente altos asociados con este enfoque limitaron su aplicabilidad a unos pocos genomas pequeños. Sin embargo, sólo dos años después, el genoma del arroz (*Oryza sativa*), aproximadamente tres veces mayor al de *Arabidopsis*, se elucidó utilizando el mismo enfoque basado en *BACs*. Esta fue la primera planta de relevancia agronómica cuyo genoma se secuenció completamente (**Goff et al., 2002b**; **Yu et al., 2002a**). El álamo fue el primer árbol de interés agronómico que se secuenció, y para éste no se utilizó el enfoque basado en *BACs* sino que se empleó la estrategia de fragmentación aleatoria del genoma completo (del inglés: *whole genome shotgun sequencing, WGS*) (**Tuskan et al., 2006**; **Wullschleger et al., 2013**). Aquí, todo el genoma se fragmenta al azar en partes más pequeñas y luego estas se secuencian y ensamblan. Si bien esta estrategia reduce significativamente el tiempo de preparación y su costo, los datos son más difíciles de ensamblar y típicamente resulta en una secuencia borrador más fragmentada. Para los siguientes genomas de plantas de cultivo se emplearon ambas estrategias, equilibrando las diversas ventajas y desventajas. En el año 2007, la secuencia del genoma de la vid (V*itis vinifera*) fue publicada por dos grupos independientes, ambos basados en la estrategia de *WGS*. Mientras que un equipo

se centró en una línea endogámica (**Jaillon et al., 2007**), el otro analizó un genotipo heterocigota utilizado para la producción de vino (**Velasco et al., 2007**).

Aunque la estrategia de *WGS* redujo el tiempo y el esfuerzo requerido para ensamblar un genoma, la generación de secuencias requería de mucho tiempo y su precio aún era muy elevado debido al alto costo que en ese entonces tenía la secuenciación por el método de Sanger. Es por eso que la incorporación de las tecnologías de *NGS* mejoraron dramáticamente la relación entre producción y costo en la secuenciación de genomas (**Bolger et al., 2014a**). La tecnología *NGS* permitió abarcar una amplia gama de aplicaciones 'post Sanger', ya que con la adopción de esta tecnología es posible secuenciar múltiples fragmentos de ADN en paralelo para producir un número mucho mayor de lecturas, aunque generalmente estas son de una longitud más corta y de una calidad inferior a las obtenidas por Sanger. Como fue antes mencionado, la primera tecnología *NGS, Roche/454®*, se basó en pirosecuenciación y fue inicialmente adoptada para proyectos genoma de tamaño bacteriano debido a la baja producción de datos y a las lecturas relativamente cortas que se obtenían. Pero las subsiguientes mejoras en la tecnología de *Roche/454®* permitieron que esta se aplique a genomas más complejos, tanto sea en conjunto con la secuenciación Sanger o mediante *Roche/454®* solamente (**Bolger et al., 2014a**).

Siguiendo en línea con lo antedicho, otra de las primeras tecnologías *NGS* fue la secuenciación por síntesis desarrollada por *Illumina®*. Este enfoque *NGS* generó rendimientos de datos mucho mayores que *Roche/454®*, aunque a costa de lecturas aún más cortas. Pero, desde su introducción, tanto el número de lecturas, como en particular la longitud de estas ha mejorado dramáticamente hoy día. En este sentido, la secuenciación por el método de *Illumina®* se pudo adoptar, por ejemplo, para ensamblar el genoma del pepino (*Cucumis sativus*) (**Huang et al., 2009**), donde se combinó con la secuenciación convencional por Sanger tomando un enfoque híbrido, lo que demostró que el uso de esta tecnología era factible para secuenciar el genoma completo de una planta. Esto repercutió en que la aplicación de *NGS* para dilucidar genomas de plantas se convirtiese en una tendencia cada

vez más fuerte. Por ejemplo, *Roche/454®* se aplicó en combinación con Sanger para ensamblar el genoma de la manzana (*Malus x Domestica Borkh*) (**Velasco et al., 2010**), y luego, en otro caso, Sanger se sustituyó por *Roche/454®* como la fuente de datos primaria para los genomas del cacao (*Theobroma cacao*) (**Argout et al., 2011**) y del melón (*Cucumis melo L.*) (**Garcia-Mas et al., 2012**), aunque, sin embargo, para adquirir información estructural sobre largas distancias genómicas y repeticiones aún era necesario utilizar secuenciación de puntas de *BACs* mediante el método de Sanger (**Bolger et al., 2014a**).

El desarrollo de técnicas de preparación de bibliotecas para obtener lecturas pareadas (*PE* y *MP*) de insertos de longitud moderada (3-20 Kpb) también ayudó a compensar la menor longitud de las lecturas *NGS*, en comparación con Sanger. Sin embargo, no fue hasta la obtención del genoma de la frutilla (*Fragaria vesca*) (**Shulaev et al., 2011**) que el primer genoma de una planta se logró secuenciar utilizando solamente *NGS*, combinando lecturas de las plataformas *454*, *Illumina* y *SOLiD*. Poco tiempo después, la secuenciación por *Illumina* se estableció como la plataforma *NGS* dominante para la secuenciación de genomas, proporcionando la mayor parte de los datos en proyectos genomas como el repollo chino (*Brassica rapa*) (**Wang et al., 2011**), la papa (*Solanum tuberosum*) (**Xu et al., 2011**), la banana (*Musa acuminata*) (**D'Hont et al., 2012**), el garbanzo (*Ciser arietinum*) (**Varshney et al., 2013**), la naranja (*Citrus sinensis*) (**Xu et al., 2013**), y la sandía (*Citrullus lanatus*) (**Guo et al., 2012**), e incluso también el gran genoma del abeto (*Picea abies*) (**Nystedt et al., 2013**) (ver Figura 12 para una línea de tiempo que pone las secuencias genómicas en orden).

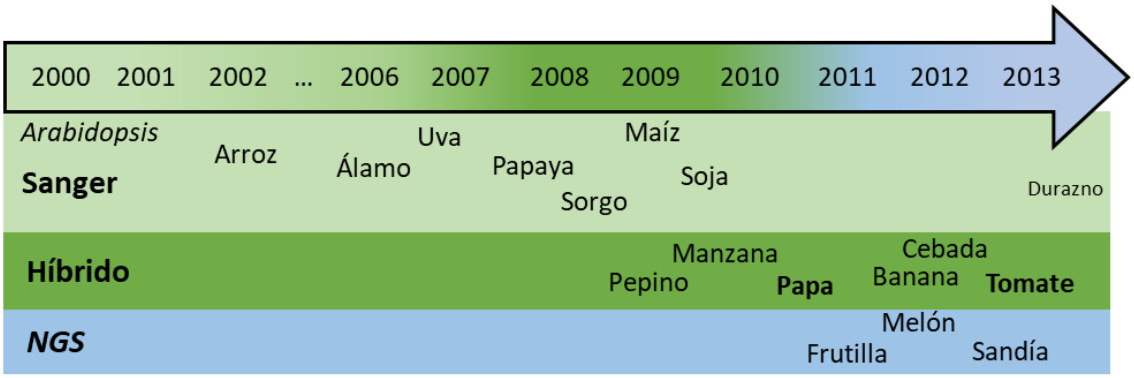


Figura 12: Línea de tiempo que ilustra la aparición de los genomas de plantas y las distintas tecnologías (Sanger y/o *NGS*) utilizadas para obtenerlos. Figura adaptada de **Bolger et al. (2014a)**

Si bien todavía hay muchos cultivos para los cuales aún falta secuenciar el genoma y, por otro lado, muchas mejoras que realizar sobre los genomas que ya han sido ensamblados, está claro que la obtención de genomas de plantas resulta una herramienta muy útil para distintos tipos de estudios. Por ejemplo, uno de los principales objetivos detrás de la obtención del genoma de una planta de interés radica en la necesidad de utilizar marcadores moleculares de alta densidad a los que se les pueda asignar rápidamente rasgos agronómicamente deseables y así poder identificar genes candidatos dentro de *loci* de interés (**Bolger et al., 2014a**). Esto puede verse en el creciente uso de genomas para el mapeo de *QTL* de rasgos deseables, incluso si sólo se dispone de secuencias de genomas parciales (**Paux et al., 2012**). Estos rasgos, una vez caracterizados, pueden ser introducidos en variedades de élite. De esta manera, el genotipado de alto rendimiento mediante *NGS* se ha transformado en una opción eficaz a la hora de realizar mejoramiento asistido por marcadores, tanto para la selección genómica, como para un manejo más eficiente de los recursos genéticos, permitiendo una resolución mucho mayor de la que sería posible de otra manera (**Cabezas et al., 2011**).

0.3 | Definición del Problema, Hipótesis y Objetivo General

En los últimos años ha resultado una pregunta abierta cuánto del aporte de la genómica ha sido aprovechado para contestar preguntas básicas de la biología de las plantas y cuánto de estas respuestas ha contribuido a resolver problemas prácticos, por ejemplo, de conservación y mejoramiento de las especies. En este sentido, es posible ejemplificar con algunos trabajos que dan cuenta de la extrema utilidad de las secuencias de los genomas completos de las especies de plantas con interés agronómico. Teniendo en cuenta lo antedicho i) una vez que el genoma del arroz estuvo disponible, se utilizó inmediatamente para ayudar a dilucidar un *QTL* importante para la producción de granos de arroz que se descubrió era una citoquinina oxidasa (**Bolger et al., 2014c**). Este trabajo demostró que un *QTL* asociado a la productividad del arroz, *GN1A*, puede reducirse a un gen de la citoquinina oxidasa / deshidrogenasa (OSCKX2), una enzima que degrada la hormona citoquinina. Cambios en los niveles de expresión de *OSCKX2* causan la acumulación de citoquininas en los meristemas de la inflorescencia y aumenta el número de órganos reproductivos, lo que aumenta el rendimiento del grano (**Ashikari et al., 2005**). Por otro lado, ii) del mismo modo, la disponibilidad de la secuencia del genoma del maíz (*Zea mays*) (**Schnable et al., 2009**) permitió desarrollar potentes mapas de haplotipos (**Gore et al., 2009**) y encontrar *QTL* asociados a la acumulación de biomasa y la obtención de bioenergía utilizando todo el genoma junto a métodos de predicción metabólica (**Riedelsheimer et al., 2012**). Además del rendimiento del cultivo (semilla, grano, etc.), el control de la floración y el tiempo de maduración es un interés recurrente, ya que representa un rasgo agronómico importante para la adaptación a diferentes regímenes de luz y latitudes geográficas. Por lo tanto, no es sorprendente que iii) el genoma de la soja (*Glycine max*) se haya utilizado para desentrañar el *locus* de maduración *E1* que tiene un gran impacto en el tiempo a floración (**Xia et al., 2012**). Este representa un carácter importante para la adaptación a fotoperiodos. Por último, iv) la secuencia del genoma de la papa (*Solanum tuberosum*) ha ayudado a identificar un factor de transcripción que regula la madurez de la planta y su ciclo de vida, se demostró que el gen *STCDF1.2* pertenece a la familia de los factores de transcripción DOF (del inglés, *DNA-binding with one finger*) el cual regula la

tuberización y la duración del ciclo de vida de las plantas, actuando como mediador entre el reloj circadiano y la señal de tuberización móvil STSP6A. (**Kloosterman et al., 2013**). Con todo, es posible afirmar que la secuenciación de genomas grandes y complejos de especies cultivadas y silvestres, facilitada por el desarrollo de las tecnologías de secuenciación y la expansión de la bioinformática, ha proporcionado nuevas oportunidades para el mejoramiento de cultivos. Los desafíos actuales incluyen entender cómo la variación genética se traduce en una mayor desempeño fenotípico en el campo (**Bevan y Uauy, 2013**). Estos enfoques, basados en el desarrollo de la genómica, están teniendo una profunda influencia en la mejora de los cultivos para la producción de alimentos y combustibles (**Sequencing Project et al., 2005**). En resumen, la información genómica se ha convertido en una herramienta fundamental para el diseño de estrategias racionales de mejoramiento de especies vegetales y para el manejo de la diversidad presente y la conservación de esta misma.

Dos hipótesis generales surgen a partir de analizar el estado del conocimiento acerca de los genomas vegetales y su explotación para el desarrollo racional de estrategias de mejoramiento.

1) La primera de ellas se enfoca en entender la estructura del genoma mitocondrial y su relación con el genoma nuclear y predice que el mismo presenta una estructura multipartita y un dinamismo producto de la interacción con el genoma nuclear de la propia célula. Esto puede ser evidenciado tanto por el número y tamaño de las estructuras genómicas que se puedan ensamblar como por la detección de inserciones en el genoma nuclear de origen mitocondrial (del inglés, *numts: nuclear mitochondrial DNA).*

2) La segunda de las hipótesis que sustentan este trabajo de Tesis refiere a la variabilidad que se genera a partir de cruzamientos interespecíficos como es el caso de las especies que componen la sección *Lycopersicum* del género *Solanum.* Para el caso de las *IL* de tomate presentado previamente (Figura 6 y 7), dicha variación es el resultado del reemplazo de los alelos domesticados por aquellos provenientes

del padre salvaje. Para abordar está hipótesis, el primer paso es conocer la estructura del genoma de la especie silvestre.

En el marco de estas dos hipótesis el objetivo general de esta Tesis es generar conocimiento acerca de las estructuras genómicas del tomate, particularmente del genoma mitocondrial de la especie cultivada (*Solanum lycopersicum*) y su relación con el genoma nuclear (Capítulo I). A su vez, se aborda el estudio del genoma nuclear de una especie silvestre de tomate (*Solanum pennellii*) (Capítulo II) ensayando la contribución de alelos exóticos aportados por esta especie sobre la asimilación y distribución del carbono en plantas de tomate (Capítulo III).

Capítulo I

Capítulo I: El genoma mitocondrial de *S. lycopersicum* (cv Heinz) y su relación con el genoma nuclear

I.1 | Introducción

I.1.1 | El genoma mitocondrial del tomate

Los organismos fotosintéticos superiores poseen muchos tipos de células, las cuales exhiben una gran compartimentalización interna. Esta característica hace que el estudio de las diferentes vías metabólicas que tienen lugar a lo largo de la vida de las células vegetales sea altamente complejo. En este sentido, las mitocondrias (del griego: *mitos* -hilo- y *chondros* -granulo-) y cloroplastos (o plástidos) (del griego: *khloros* -verde- y *plastós* -forma-) son las organelas intracelulares que contienen la maquinaria necesaria para la respiración y la fotosíntesis de la célula, respectivamente. Estas organelas también participan en la biosíntesis de metabolitos esenciales tales como aminoácidos, nucleótidos, lípidos y almidón (**Lichtenstein et al., 2016**). En las plantas, los cloroplastos y las mitocondrias son las "Centrales de Energía" celulares. Los primeros capturan la energía lumínica del sol y las otras desarman esta energía capturada en paquetes más pequeños de adenosina trifosfato (ATP) que luego se utilizan como fuente de energía química para alimentar todo el trabajo celular. Por lo tanto, para una clara comprensión de los procesos fisiológicos a nivel de toda la planta se requiere comprender a fondo las interacciones que ocurren entre estas organelas con el resto de los compartimentos celulares. En mamíferos estas interacciones entre organelas implican principalmente la transferencia de proteínas y metabolitos. Sin embargo, la transferencia de genes desde las mitocondrias y cloroplastos hacía al núcleo es otro tipo de interacción esencial en las células vegetales. Por ejemplo, en un trabajo realizado en maíz (*Zea mays*) 51 inserciones nucleares de origen mitocondrial fueron identificadas utilizando una biblioteca de 20 cósmidos conteniendo ADNmt (570 Kb) del cultivar elite B73 que fueron utilizados como sondas para un experimento de hibridación fluorescente *in-situ* (FISH) con los 10 cromosomas del maíz (**Lough et al., 2008**). Aunque las mitocondrias y los cloroplastos mantienen parte de sus genomas ancestrales, los procesos de

transferencia de genes con el núcleo están operando continuamente (**Lichtenstein et al., 2016**).

Durante las últimas décadas del siglo XIX las mitocondrias se observaron por primera vez en una variedad de tipos celulares como hilos de gránulos llamados sarcosomas, bioblastos o condriosomas (**Schmidt, 1913**). Por otra parte, **Nägeli (1846**), descubrió en las células vegetales que los cloroplastos se multiplicaban por división. A principios del siglo XX, se publicaron los primeros informes de herencia no-mendeliana en plantas superiores basados en estudios de hojas variegadas (**Correns, 1908**). Estos trabajos mostraron que solo pocas hojas variegadas (verde y blanco) fueron causadas por factores heredados de una manera no-mendeliana. Más tarde, otros análisis de variegación en plantas superiores revelaron que los determinantes genéticos de estos caracteres estaban asociados con cloroplastos, lo que sugirió que estos orgánulos podían albergar información genética. Estas observaciones llevaron al botánico ruso Mereschkowski a ser el primero en hablar de la teoría endosimbiótica (**Mereschkowski, 1905**). En la segunda década del siglo XXI, **Wallin (1923**) amplió esta idea y corroboró la hipótesis sobre el origen mitocondrial. Muchos libros de texto describen esta teoría en detalle, por lo que en este capítulo no nos detendremos en detalle en este aspecto.

Posteriormente a estos descubrimientos fueron documentadas las observaciones de (**Nass y Nass, 1963**) y (**Schatz et al., 1964**). Mediante el uso de dos enfoques diferentes, estos autores informaron por primera vez y al mismo tiempo que las mitocondrias de las células de embrión de pollo (*Gallus gallus domesticus*) y de la levadura (*Saccharomyces cerevisiae*) contenían una significativa cantidad de ADN (ADNmt). En cuanto a observaciones en plantas superiores, a principios de la década de 1960 se realizaron estudios que demostraron que la esterilidad masculina citoplasmática (*CMS*) es un rasgo hereditario materno, señalando la existencia de estructuras de ADN que eran únicas dentro de las mitocondrias de las células vegetales de diferentes especies cultivadas (**Leaver y Gray, 1982**).

I.1.2 | El genoma mitocondrial y su naturaleza multipartita

El ADN mitocondrial posee repeticiones "directas" distribuidas a lo largo de todo el genoma. La interacción y la recombinación de estas repeticiones da como resultado la formación de pequeñas moléculas circulares y lineales de ADN, lo que se conoce como "genoma multipartito". Esta continua interacción y recombinación entre repeticiones "directas" puede producir, en tiempos evolutivos cortos, reorganizaciones en regiones de ADN codificante, resultando en polimorfismos que pueden ser detectados incluso en especies de plantas cercanas.

Con respecto al ADNmt del tomate, no fue hasta 1992 que el grupo de **Melchers et al., (1992**) publicó el primer mapa físico del genoma mitocondrial, y años más tarde, se informó un modelo de su tamaño y organización basándose en el mapeo-físico de ADNmt obtenido por hibridación (**Shikanai et al., 1998**). Este modelo propuso que la estructura genómica de la mitocondria del tomate presentaría un tamaño de 450 Kb y una estructura compuesta por 5 partículas sub-genómicas de diferentes tamaños organizadas en 2 cadenas circulares y en 3 cadenas lineales (ver Figura 13). Según los autores estas 5 partículas coexisten en un rango dinámico regulado de alguna manera por sitios de alta actividad de recombinación homóloga (**Shikanai et al., 1998**).

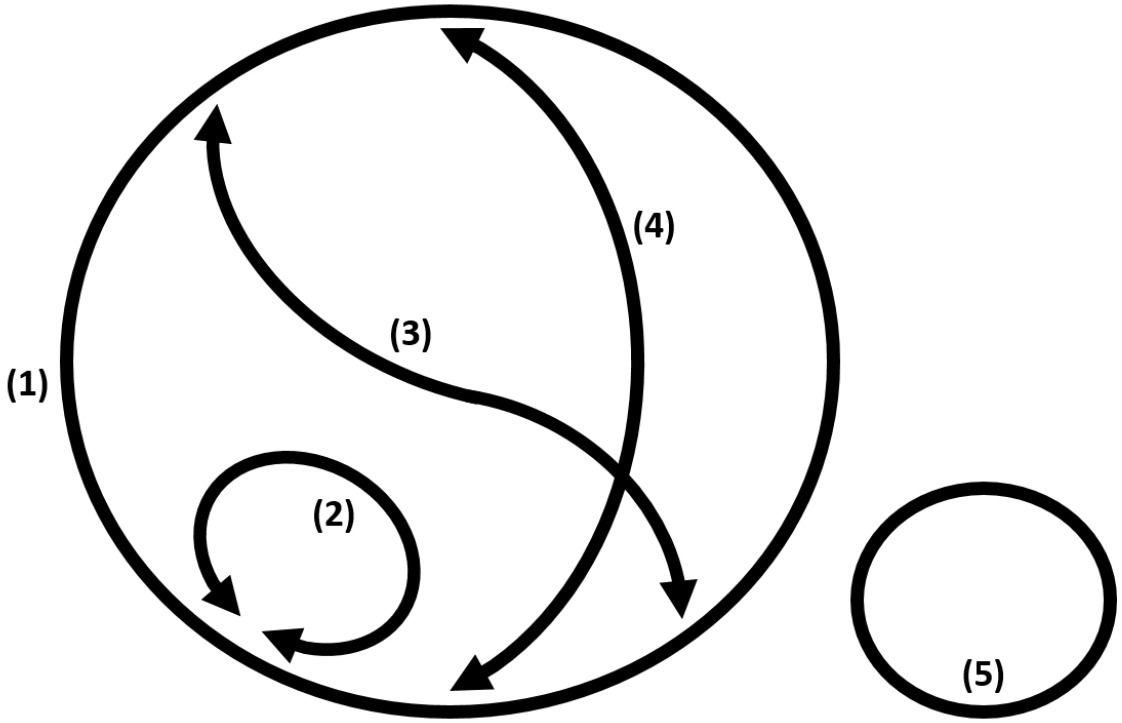


Figura 13: Estructura multi-partita propuesta para el genoma de la mitocondria de *S. lycopersicum*. De las cinco estructuras propuestas, tres de ellas son lineales (2, 3 y 4) y dos circulares (1 y 5). Figura tomada y adaptada de **Shikanai et al.,(1998**)

I.1.3 | Inserciones Nucleares de Origen Mitocondrial (*numts*)

Tanto *in vitro* como *in vivo*, los *numt*s se integran al ADN nuclear a través de uniones no-homólogas de extremos terminales (*NHEJ*: del inglés, *non-homologous end joining*,) en aquellas regiones donde haya roturas de doble hebra (*DSBs:* del inglés, *double-strand breaks*) (**Hazkani-Covo et al., 2010**). La frecuencia de inserción de *numts* en 85 genomas eucariotas secuenciados revelaron que el contenido de *numts* está fuertemente correlacionado con el tamaño del genoma, lo que sugiere que la tasa de inserción podría estar limitada por la frecuencia de *DSBs*. En humanos, *numts* polimórficos vinculan genotipos mitocondriales heredados por la madre a los haplotipos de ADN nuclear durante el pasado, ofreciendo nuevas oportunidades para asociar marcadores nucleares con marcadores mitocondriales que permiten trazar una escala temporal sobre estos eventos (**Hazkani-Covo et al., 2010**). Dos estudios previos sugirieron que el tamaño del genoma nuclear y la longitud acumulada de *numts* se correlacionaron

positivamente (**Smith et al., 2011**; **Michalovova et al., 2013**). En plantas, análisis de secuencias cromosómicas completas de 9 especies vegetales mostraron que sus genomas son ricos en ADN de origen mitocondrial (**Yoshida et al., 2017**). Por ejemplo, *A. thaliana* contiene el número más bajo de segmentos mitocondriales insertos en el núcleo, 41 *numts* (282 Kb) y en contrapunto *Z. mays* contiene el número más elevado, 1456 *numts* (964 Kb). (**Yoshida et al., 2017**). Estos estudios previos sugirieron que el tamaño del genoma nuclear y la longitud acumulada de *numts* se correlacionan positivamente (**Smith et al., 2011**; **Michalovova et al., 2013**). En este capítulo de Tesis, se aborda la detección de *numts* en el genoma del cultivar Heinz de *S. lycopersicum*, por un método computacional (bloques de colinealidad) y por hibridación fluorescente *in-situ* (*FISH* y *GISH*).

En el marco de la primera hipótesis que planteo en esta Tesis, los objetivos específicos que se abordaron en este capítulo comprenden:

1. Obtener secuencias de alta calidad mediante la fragmentación aleatoria de ADN mitocondrial de tomate de alta pureza.
2. Ensamblar *contigs* y *scaffolds* a partir de los fragmentos secuenciados.
3. Anotar estructural y funcionalmente las regiones codificantes del genoma mitocondrial.
4. Buscar inserciones mitocondriales en el genoma nuclear del tomate.
5. Comprobar la existencia de *numts* mediante hibridación *in situ* –FISH-.

En este capítulo se describen los resultados obtenidos al secuenciar y analizar el genoma de la mitocondria de *S. lycopersicum* (cv Heinz). A su vez, se analiza su estructura en comparación con genomas secuenciados de otras especies de *Embryophyta*. A partir de la detección de *numts* en el genoma del cultivar Heinz, tanto aplicando análisis *in silico* (bloques de colinealidad) como por hibridación fluorescente *in situ* (FISH y GISH), se resumen los hallazgos sobre la funcionalidad de este genoma junto con su dinámica en relación con los recientes eventos de intercambio de ADN con el núcleo, un proceso que parece seguir siendo operativo.

I.2 | Resultados

El genoma mitocondrial de *Solanum lycopersicum* (cultivar Heinz) se obtuvo mediante la tecnología de secuenciación por didesoxinucleótido terminador (método de Sanger) aplicada sobre fragmentos de hebras de ADN previamente fragmentadas de forma aleatoria. Las lecturas de secuencia resultantes de este método (del inglés, *reads*) fueron ensambladas *de novo* en secuencias contiguas mediante el algoritmo de superposición, consenso y maquetado (del inglés, *overlap-consensus-layout*). El listado de *contigs* fue ordenado y orientado mediante el uso de lecturas pareadas (*PE*), obteniéndose así *contigs* interconectados por brechas de nucleótidos genéricos y contiguos simbolizados con la letra "N" a modo de espaciadores (del inglés, *gaps*), formando así andamios de *contigs* (del inglés, *scaffolds*). El conjunto de *contigs* y *scaffolds* conforman el borrador del genoma mitocondrial del tomate (SOLYC_MT v1.50) que se presenta en este Capítulo y el cual fue almacenado en una serie de archivos disponibles en el banco de datos genómicos del NCBI con el número de acceso AFYBoooooooo; y también en la página web de acceso público que se creó específicamente para este proyecto: www.mitochondrialgenome.org/.

I.2.1 | Secuencias de ADN mitocondrial

Los datos fueron producidos mediante tres equipos de secuenciación capilar (ABI 3130, ABI 3500 y ABI 3730, respectivamente) en la "Unidad de Genómica del INTA de Castelar". Lo que resultó en un conjunto de datos de secuenciación compuesto por 7.051 cromatogramas almacenados en archivos con el formato propietario "*ab1*".

I.2.2 | Preprocesamiento de datos

Recorte de puntas y vectores

Los 7.051 cromatogramas fueron preprocesados con algoritmos para recortar bases de baja calidad de los extremos 5´ y 3' de las secuencias (del inglés, *end-clipping*) con el parámetro umbral establecido como PHRED ≥ 20 en una ventana de 25 nt. Luego se buscaron y recortaron fragmentos contaminantes pertenecientes a los vectores de clonado (del inglés, *vector-trimming*) utilizados (pMOSBlue y pZErO-2) con el parámetro umbral para recortado establecido como % de identidad ≥ 90, solapamiento mínimo = 30 nt y descarte de lecturas < 100 nt; (ver M&M, sección V.1.4). Esto redundó en un total de 4.875 secuencias de alta calidad que superaron los parámetros establecidos. Como resultado se descartó un 32% de las lecturas y el restante 68% se utilizó para trabajar en las siguientes etapas.

Filtrado de contaminantes

El objetivo de este paso fue remover todas aquellas secuencias que no sean identificables como secuencias mitocondriales del clado *embryophyta*. Se construyó una base de datos conteniendo todos los genomas mitocondriales de *Plantae* disponibles públicamente hasta al momento; lo que resultó en una base de datos de 48 genomas mitocondriales (plant_MT_dB), la cual se publicó en el sitio de internet http://www.mitochondrialgenome.org/BLAST/. Utilizando como semilla las 4.875 lecturas mencionadas previamente, se realizó una búsqueda por similitud de secuencias contra la base de datos plant_MT_dB. Como resultado, un 30% de las lecturas fue descartado por considerarse secuencias no mitocondriales y las restantes 3.388 lecturas que superaron los parámetros de corte (e-valor = 1E-10; % identidad ≥ 80 y % de cobertura ≥ 12) se utilizaron para trabajar en las siguientes etapas.

Ordenamiento de lecturas pareadas

Los fragmentos que fueron purificados de la genoteca mitocondrial tienen entre 1.000 y 40.000 nt de longitud y fueron secuenciados de forma pareada por ambos extremos de cada fragmento (F y R, respectivamente). La longitud promedio de las lecturas Sanger fue de 1.024 nt. Por lo tanto, como los insertos mitocondriales fueron secuenciados por ambas puntas, se obtuvieron lecturas pareadas a una distancia conocida (1 – 4 Kb). Las lecturas pareadas, luego se utilizaron para ordenar, orientar y expandir *contigs* en *scaffolds* durante la etapa de ensamblado, permitiendo obtener secuencias más largas y resolver regiones repetitivas en tándem. Teniendo esto en cuenta, cabe destacar que durante el preprocesamiento aquellas lecturas que dejaron de ser pareadas (F y R) se clasifican como lecturas desemparejadas (del inglés, *unpaired* o U). En este sentido, con la herramienta SangerPairedOmatic (ver M&M, sección V.1.6) las 3.388 lecturas se distribuyeron en tres archivos del tipo FASTQ dando como resultado: i) 891 secuencias SolycMT.F; ii) 891 secuencias SolycMT.R y iii) 1.606 secuencias SolycMT.U.

Contigs obtenidos por secuenciación *Roche/454®*

Adicionalmente a las lecturas obtenidas por secuenciación Sanger, descriptas arriba, se obtuvieron un conjunto de 577 *contigs* resultantes de explorar el borrador del genoma del cultivar Heinz (ensamblado a partir de lecturas *NGS Roche/454®*), mediante una búsqueda por similitud de secuencias contra la base de datos *plant_MT_dB*. El largo promedio de estos *contigs* fue de 1.150 nt y resultaron en un total de 665,439 nucleótidos, que se añadieron al conjunto de datos obtenido por nuestro laboratorio por el método de Sanger, para ensamblar el genoma mitocondrial del cultivar Heinz (ver Tabla 2).

Tabla 2: Número de lecturas y cantidad de nt resultantes de las librerías de ADN mitocondrial de *S. lycopersicum* (cv Heinz) secuenciadas por el método de Sanger y *Roche/454®*.

Tecnología	Lecturas	Nucleótidos
Sanger	3.388	3.186.430
Roche/454®	577	665.439
Total	3.965	3.851.869

I.2.3 | Ensamblado del genoma mitocondrial de tomate

Obtención de *contigs*

Las secuencias se ensamblaron mediante el algoritmo de superposición, consenso y maquetado implementando el software *CAP3* (**Huang y Madan, 1999**) lo que dio como resultado 173 *contigs* (ver Tabla 3).

Tabla 3: Número de *contigs* obtenidos después del ensamblado con *CAP3*, se detalla el número de *contigs*, nucleótidos totales, largo promedio, largo máximo (*contig* más largo), el largo del *contig* N50, y la cantidad de nucleótidos genéricos (Ns) que hay en el ensamblado.

contigs	nucleótidos	largo promedio	largo máximo	N50	cantidad de Ns
173	347.278	3.546,60	12.591	4.590	845

Armado de *scaffolds* y cerrado de *gaps*

Para el armado de los *scaffolds* se buscó información que permita ligar dos o más *contigs* mediante el aprovechamiento de lecturas pareadas (F y R) pertenecientes al mismo inserto mitocondrial, ya que se conocía de antemano, que las lecturas *PE* estaban a una distancia estimada de 1.000 a 4.000 nt. Esto se realizó mediante alineamiento de las 3.388 lecturas Sanger con el programa *Bowtie* (**Langmead et al., 2009**) seguido de una búsqueda con la herramienta *Tablet* (**Milne et al., 2013**) para visualización de genomas y alineamientos (ver M&M, sección V.1.8). Una vez maquetados los *scaffolds* se diseñaron primers sobre las puntas de los *contigs* que conforman los *scaffolds*, y se amplificaron mediante *PCR*. Los

productos de amplificación por PCR fueron secuenciados por Sanger y se cerraron los *gaps* (brechas de segmentos continuos de nucleótidos N) entre los *contigs* enlazados por lecturas pareadas y se obtuvieron así *scaffolds* completos (ver Figura 14). Cuando el *gap* entre *contigs* enlazados no pudo ser cerrado por completo se unieron por medio de una secuencia espaciadora de 100 Ns contiguas.

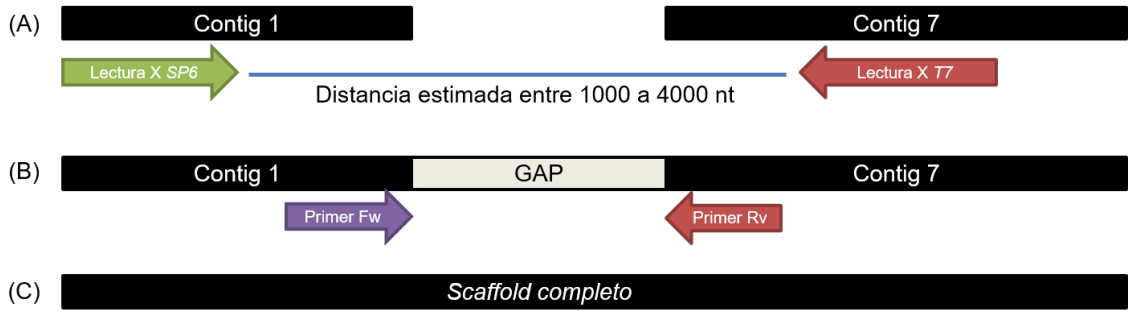
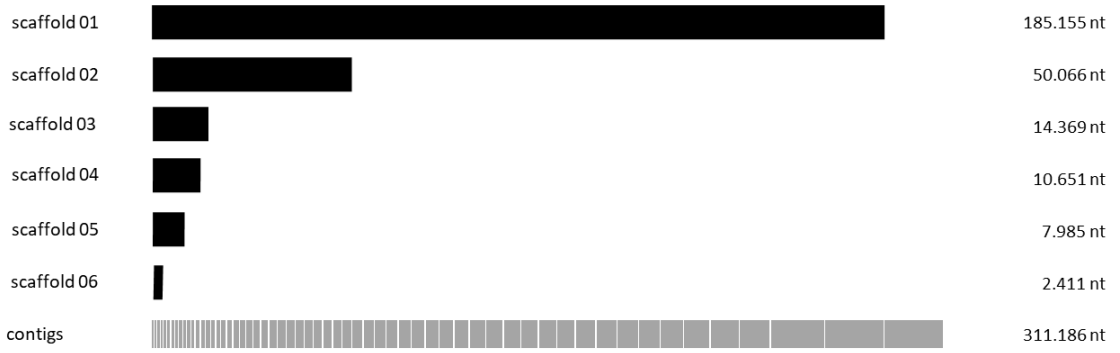


Figura 14: Diagrama del protocolo utilizado para armar *scaffolds* utilizando (A) lecturas pareadas ligadas en *contigs* distintos, (B) primers diseñados sobre las puntas de los *contigs* y (C) secuenciación por Sanger para cerrado de gaps.

Como resultado final se obtuvo un ensamblado de 578.247 nt distribuidos en 162 secuencias contiguas (6 *scaffolds* y 156 *contigs;* ver Figura 15). Los cuales en conjunto representan el primer borrador del genoma mitocondrial del tomate. El mismo está disponible públicamente en el sitio de internet www.mitochondrialgenome.org y al mismo tiempo todas las secuencias fueron depositadas en las bases de datos públicas e internacionales DDBJ, EMBL y GenBank bajo el BioProject ID: 67471 y el número de entrada AFYB00000000.



ensamblado	*scaff+cont* (≥ 0 nt)	tamaño total (≥ 0 nt)	*scaff+cont* (≥ 1000 nt)	tamaño total (≥ 1000 nt)
Solyc_MT_v1.50	162	578.247	89	526.701
GC (%)	**N50**	**N75**	**L50**	**L75**
44	14.369	3.212	3	29

Figura 15: Diagrama ilustrativo del genoma mitocondrial del tomate (v1.50). En la parte superior se observan *scaffolds* (barras negras) y *contigs* ordenados de menor a mayor tamaño (barras grises) y el número de nucleótidos (nt) contenidos en las distintas pseudomoléculas. En la tabla inferior se presentan algunos valores indicativos de la calidad del ensamblado obtenidos con la herramienta *Quast* (**Gurevich et al., 2013**).

Como se puede observar, el ensamblado del genoma mitocondrial del tomate (cv Heinz) quedó compuesto por 6 unidades físicas (*scaffolds*) y varias partículas sub-genómicas (*contigs*). La presencia de los 6 *scaffolds* obtenidos sugieren una estructura mitocondrial multipartita y que estaría en concordancia con el mapa físico de la mitocondria de tomate reportado por **Shikanai et al. (1998)** (ver Figura 13) y con los trabajos reportados previamente para las especies *Nicotiana tabacum* (**Sugiyama et al., 2005**) y *Oryza sativa* (**Tian et al., 2006**).

I.2.4 | Predicción y anotación de genes

El ensamblado del genoma mitocondrial del tomate (Solyc_MT_v1.50) se anotó de forma estructural y funcional. Para el primer abordaje se realizó la predicción de estructuras genéticas *ab initio* mediante la herramienta *FGENESH* (**Salamov y Solovyev, 2000**) y resultó en la predicción de 90 marcos abiertos de lectura (del inglés: *ORFs, open reading frames*). Por otra parte, para la anotación del tipo funcional el objetivo fue asignar funciones a las estructuras predichas. Esto se realizó mediante un alineamiento reciproco con la herramienta *Artemis* (**Rutherford et al., 2000**) entre Solyc_MT_v1.50 y el genoma mitocondrial de

referencia de la especie cercana *Nicotiana tabacum* (familia *Solanaceae*), el cual ya se encontraba previamente anotado y validado (**Sugiyama et al., 2005**). Se estudió a cada uno de los ortólogos alineados con *Artemis* y se seleccionaron aquellos alineamientos con un alto porcentaje de similitud de secuencia (≥ 80%), este análisis resultó en la anotación de 57 genes clasificados en tres grupos: i) 36 genes codificantes, ii) 3 genes de ARN ribosomal y iii) 18 genes de ARN de transferencia. En resumen, de los 90 *ORFs* predichos por *FGENESH*, 57 fueron asignados a una función biológica y los restantes 33 *ORFs* fueron anotados como predicciones *in silico* o proteínas hipotéticas (ver Tabla Anexa 1, en Anexos VIII.1).

I.2.5 | Segmentos nucleares de origen mitocondrial (*numts*)

Con el fin de identificar posibles inserciones nucleares de origen mitocondrial (*numts*) se realizó una búsqueda de bloques de secuencias conservadas de forma local y colineal (*local-colineal-blocks, LCBs*) entre el ensamblado mitocondrial y el nuclear, utilizando dos métodos diferentes: i) Por un lado, se realizó una búsqueda *in-silico* mediante la herramienta *Mauve progressive Aligner* (**Darling et al., 2004**) (ver M&M, sección V.1.9). En este sentido, mediante el análisis de 118 *LCBs* entre el genoma mitocondrial y nuclear, se pudieron identificar 92 *numts* (**The Tomato Genome Consortium, 2012**). En la Figura 16 se puede observar como los *numts* varían en número, tamaño y posición en todos los cromosomas, excepto en el cromosoma 6, donde no se detectaron inserciones de *numts*. Por otra parte, el número de *numts* varió entre tres y once en los cromosomas 4 y 8, mientras que el número más alto, veinticinco, se detectó en el cromosoma 11. Por otra parte, ii) de manera independiente, se realizaron experimentos de hibridación *in-situ* fluorescente (*FISH*) con una mezcla equimolar de 40 clones mitocondriales con tamaños de inserción ≥ 2,5 Kpb marcados con dioxigenina, también se híbrido el clon del *BAC* Le_HBa323E19 de la genoteca nuclear de *S. lycopersicum* (cv Heinz) marcado con biotina para identificar el brazo largo del cromosoma 11 y el trasfondo de los cromosomas nucleares fueron marcados con DAPI (4', 6-diamino-2-fenilindol). El resultado del experimento de FISH de ADN mitocondrial secundó el análisis realizado *in-silico* con el *Mauve Progressive Aligner* (Figura 17).

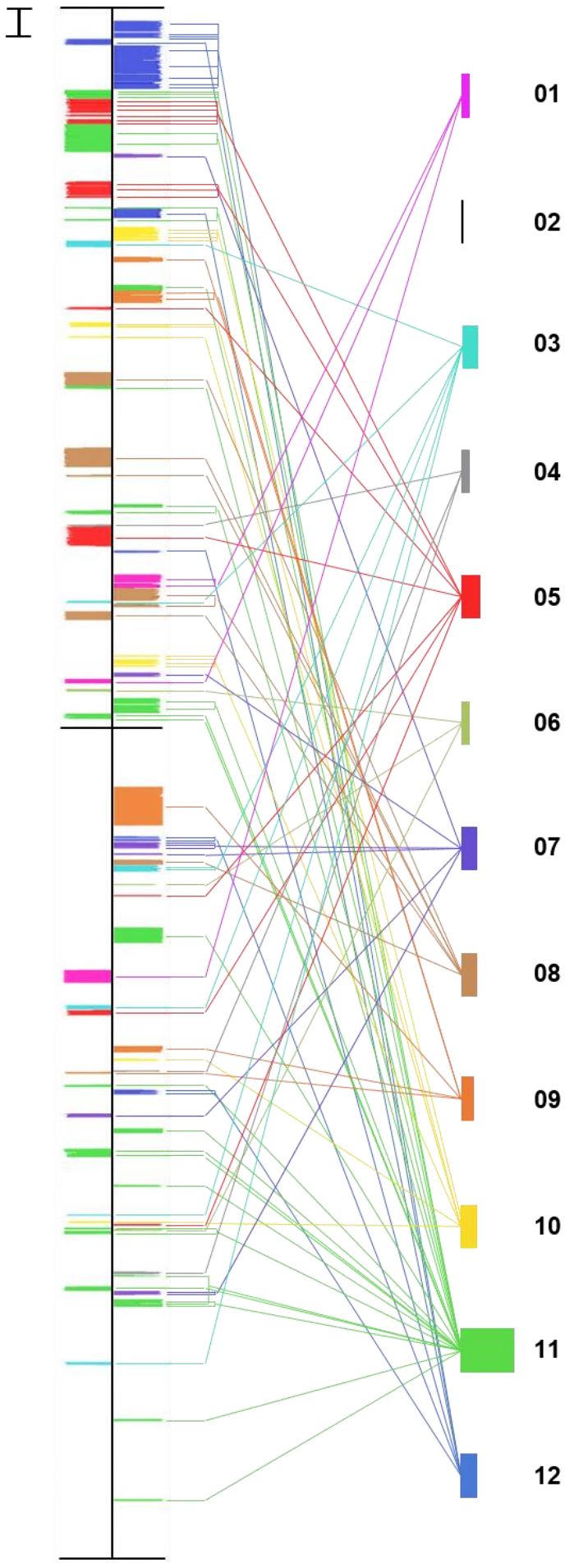
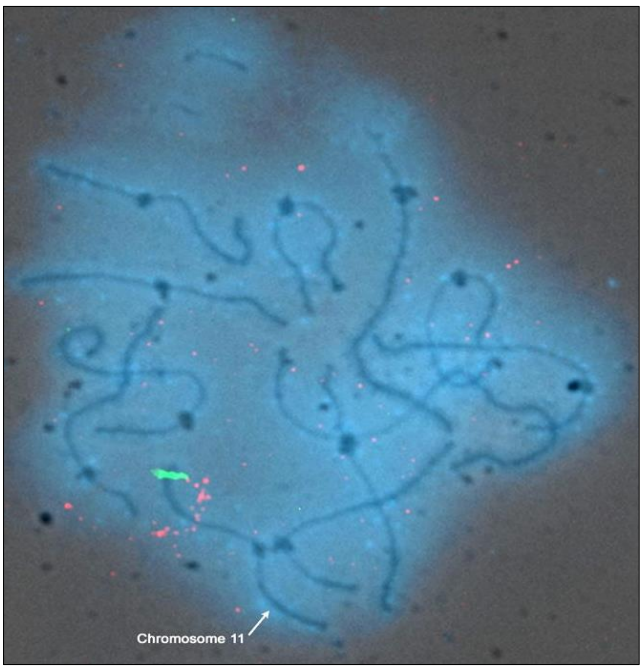


Figura 16: Inserciones nucleares de origen mitocondrial (*numts*) en el tomate *S. lycopersicum* cv Heinz. Se muestra una representación esquemática de un mapa lineal del genoma mitocondrial. La línea negra vertical representa el ensamblado del genoma mitocondrial del tomate (v1.50) y la barra de escala (arriba a la izquierda) representa 10 Kpb. Las líneas negras horizontales agrupan al genoma mitocondrial en *scaffolds* (segmento superior) y *contigs* (segmento inferior). Los bloques de colores que se extienden a través del genoma representan secuencias colineales conservadas (LCBs) predichas por la herramienta "*Mauve Progressive Aligner*" utilizando los cromosomas nucleares (v2.40) como referencia. Los *LCBs* sentido se muestran a sobre la derecha de la línea negra y los *LCBs* anti-sentido sobre la izquierda. El alto y ancho de los bloques representan el grado de conservación y la longitud, respectivamente. Las líneas que conectan cada bloque mitocondrial con los cromosomas nucleares (barras numeradas del 01 al 12) representan eventos de inserción mitocondrial. El grosor de los cromosomas nucleares es proporcional al número de *numts*.



Foci	*locus*	Localización celular
Rojo	Clones MT	Mitocondria
Azul	ADN rico en A-T	Cromosomas Nucleares
Verde	Le_HBa323E19	Cromosoma 11

Figura 17: Hibridación fluorescente *in-situ*: Se utilizaron como sondas *FISH* i) una mezcla equimolar de 40 clones mitocondriales (clones MT) con tamaños de inserción ≥ 2,5 Kpb etiquetados con dioxigenina (*foci* color rojo), ii) el clon del *BAC* Le_HBa323E19 etiquetado con biotina para identificar el brazo largo del cromosoma 11 (*foci* color verde) y iii) los cromosomas nucleares se etiquetaron con Actina (*foci* de color azul). Los *foci* rojos se encuentran en su mayoría sobre el brazo largo del cromosoma 11 y los *foci* dispersos marcan otros sitios de inserción de ADN mitocondrial en otros cromosomas. Este patrón de hibridación concuerda con las inserciones mitocondriales que se muestran en el análisis *in-silico* de la Figura 16.

I.2.6 | Resumen de los principales resultados

1. Los datos de secuenciación del ADN mitocondrial de *S. lycopersicum* (cv Heinz) permitieron obtener el primer borrador del genoma completo de esta organela.
2. El genoma mitocondrial de tomate está compuesto de varias partículas sub-genómicas y la presencia de 6 moléculas más grandes (identificados a partir de los 6 *scaffolds* y 156 *contigs*, ensamblados *in-silico).* Esto apoya la hipótesis de una estructura multipartita que está de acuerdo con el mapa físico de la mitocondria de este especie comunicado por **Shikanai et al (1998)** y con los modelos sugeridos en trabajos previos en *Nicotiana tabacum* (**Sugiyama et al., 2005**) y *Oryza sativa* (**Tian et al., 2006**).
3. Se anotaron de forma estructural y funcional 57 genes con significativa homología con aquellos anotados previamente para el genoma mitocondrial de *Nicotiana tabacum* (**Sugiyama et al., 2005**) y a su vez se predijeron *in-silico* y *de novo* 33 estructuras hipotéticas codificantes.
4. Los resultados del análisis *in-silico* como aquellos surgidos de experimentos de *FISH* sugieren que existe transferencia horizontal de material genético entre el genoma mitocondrial y nuclear y que estos pueden ser detectados en forma de *numts*.

I.3 | Discusión

Los datos de secuencia del ADN mitocondrial del tomate *S lycopersicum* (cv Heinz) obtenidos por fragmentación genómica aleatoria y por la metodología clásica de terminador didesoxinucleótido (Sanger) permitieron el ensamblado de un borrador de alta calidad del genoma de esta organela utilizando el algoritmo de superposición, consenso y maquetado. En trabajos previos, el grupo de **Sugiyama et al., (2005)** logró ensamblar el genoma mitocondrial del tabaco (*Nicotiana tabacum*) y luego compararon el resultado con los genomas mitocondriales de *Arabidopsis thaliana* (**Unseld et al., 1997**), *Beta vulgaris* (**Kubo et al., 2000**), *Oryza sativa* (**Notsu et al., 2002**) y *Brassica napus* (**Handa, 2003**). Utilizando la misma estrategia, pusimos a punto el método para nuestro conjunto de datos y durante el proceso logramos obtener el conocimiento básico necesario para abordar un proyecto de ensamblado *de novo* como el del genoma nuclear de la especie silvestre *Solanum pennellii* que se describe en el Capítulo II de esta Tesis. En cuanto al borrador del genoma mitocondrial de tomate este fue publicado en el año 2012 junto al Consorcio Internacional del Genoma del Tomate (**The Tomato Genome Consortium, 2012**). En este trabajo un grupo de más de 300 personas y 14 países contribuyeron a obtener un genoma completo (nuclear, cloroplástico y mitocondrial) de alta calidad del cultivar elite de *Solanum lycopersicum* (cv. Heinz, LA1706). Por otro lado, en el mismo trabajo también se publicó un genoma borrador de la especie silvestre más cercana (*Solanum pimpinellifolium*). Cabe destacar, que, si bien el tomate moderno es un organismo diploide, este trabajo ayudo a develar como el genoma nuclear del tomate se triplicó en dos ocasiones previas. En uva (*Vitis vinifera*), ya se había observado un evento similar (**Jaillon et al., 2007**). Por su parte, en tomate (*S. lycopersicum*), se observó que el primer evento de triplicación ocurrió hace unos 130 millones de años y que por otro lado, de forma más reciente, un segundo evento de triplicación ocurrió hace unos 60 millones de años, y este, tuvo importantes implicancias para el desarrollo de la fruta actual (**Hill, 2012**). De la misma manera, cabe mencionar, que la calidad del genoma mitocondrial del tomate resultó comparable con las 5 especies de angiospermas modelos que habían sido previamente secuenciadas por otros grupos (ver Tabla 4).

Tabla 4: Características de los genomas mitocondriales de angiospermas

Característica	*NT*	*AT*	*BN*	*BV*	*OS*	***SL***
Tamaño total (nt)	430.597	366.924	221.853	368.799	490.520	578.247
Número de genes	60	55	53	52	56	57
Estructura multipartita	Si				Si	Si

Nicotiana tabacum (NT), Arabidopsis thaliana (AT), Brassica napus (BN), Beta vulgaris (BV), Oryza sativa (OS) y Solanum lycopersicum (SL)

.

I.3.1 | Estructura multipartita

La presencia de estructuras mitocondriales de tamaños variables (organización multipartita) que presenta la mitocondria del tomate (6 *scaffolds*) está de acuerdo con el mapa físico publicado anteriormente por **Shikanai et al., (1998)** donde se presentan la estructura multipartita de la mitocondria de *S lycopersicum* (cv Heinz) (ver Figura 13 y Figura 15). Esta organización del ADNmt de tomate es comparable con aquellas reportadas para *Nicotiana tabacum* (**Sugiyama et al., 2005**) y *Oryza sativa* (**Tian et al., 2006**). En este sentido, nuestro resultado aporta evidencia que refuerza la observación hecha por (**Fauron y Casper, 1995**) que sugiere que la organización de los condromas de angiospermas se caracteriza por la presencia de genomas multi-partitos debido a la alta frecuencia de recombinación mediada por las secuencias repetitivas que hay en estas estructuras genómicas. El modelo tradicional del genoma circular de la mitocondria de las angiospermas (del inglés: *MC, master circle*) fue construido en base a mapas de restricción, en donde es posible acomodar toda la información genética (**Tian et al., 2006**). En contraste, una investigación realizada por microscopia electrónica demostró que el ADNmt de células en cultivo de *Chenopodium album* está compuesto principalmente por moléculas lineares de diversos tamaños *in vivo*, junto a estructuras del tipo roseta y del tipo sigma (**Backert y Börner, 2000**). Debido a la alta cantidad de estructuras mitocondriales de tamaño dinámico que se observaron durante el transcurso del crecimiento celular, se ha postulado que representan estructuras intermediarias de la replicación. Por otro lado, estructuras similares también se han observado al microscopio óptico en células BY-2 de *Nicotiana tabacum* (**Oldenburg y Bendich,**

1996). Por lo tanto hay diferencias contrapuestas entre las formas mitocondriales derivadas del mapeo por fragmentación y las observaciones al microscopio (**Sugiyama et al., 2005**). De todos modos, aunque esta discrepancia aún no ha sido resuelta, los dos tipos de evidencia sugieren que la organización estructural del ADNmt es altamente dinámica (**Lichtenstein et al., 2016**). Más aún, se postula que la estructura multi-partita puede actuar como un ensamble de genes redundantes y así modular el número de copias genómicas en los condromas de las plantas. Por lo tanto, la estructura multipartita es un factor de relevancia a la hora de analizar los *contigs* y *scaffolds* del ensamblado del genoma mitocondrial del tomate. La dinámica de estas estructuras genómicas podría alterar el fenotipo de la planta de tomate como lo sugieren dos reportes anteriores realizados por (**Kanazawa y Hirai, 1994**; **Janska et al., 1998**). Estos autores demuestran que la esterilidad citoplasmática masculina (*CMS*) en *Nicotiana tabacum* y *Phaseolus vulgaris* está relacionada con la ocurrencia de estructuras multipartitas, la heteroplasmia y la deriva génica.

I.3.2 | Segmentos nucleares de origen mitocondrial (*numts*)

De acuerdo con la teoría endosimbiótica (**Margulis y Bermudes, 1985**), la mitocondria y su genoma son los restos de un antepasado eubacteriano de vida libre, probablemente una alfa-proteobacteria, que fue engullida por una célula huésped eucariótica que luego estableció una relación simbiótica dentro de ella (**Gray, 1999**). El huésped proporcionó el grueso del genoma nuclear, y la gran mayoría de los genes del genoma endosimbiótico se perdieron o se transfirieron al genoma nuclear en una etapa temprana de la evolución. Por lo tanto, sólo una pequeña fracción de los genes originales se encontrarían en el ADNmt moderno. Una de las características que distingue al ADNmt de las plantas superiores del de los animales y otros organismos es que aunque la transferencia de los genes mitocondriales al núcleo y su activación funcional cesó en el antepasado común de los animales, en las plantas con flor, la pérdida de genes mitocondriales y la transferencia de genes han sido procesos continuos y frecuentes (**Palmer et al., 2000**). En la Figura 16 se puede observar como los *numts* varían en número, tamaño y posición en todos los cromosomas del cultivar Heinz, por ejemplo, se observa

que el número de *numts* varió entre dos y once en los cromosomas 4 y 8, mientras que el número más alto, veinticinco, se detectó en el cromosoma 11, y en el cromosoma 6, no se detectaron inserciones de *numts*. Para comprender mejor estos números se podría ampliar este estudio y buscar *numts* en el genoma de variedades silvestres de tomate como *Solanum pennellii* y *Solanum pimpinellifolium* para elucidar el rol de los *numts* sobre el proceso de especiación del tomate. En este sentido, cabe destacar que en nuestro trabajo el hallazgo de 92 *numts* en el genoma del cv Heinz es compatible con el hecho de que tanto secuencias mitocondriales y cloroplásticas han sido identificadas previamente en los genomas nucleares de la espinaca *Spinacia oleracea* (**Scott y Timmis, 1984**; **Cheung y Scott, 1989**), el tomate *Solanum lycopersicum* (**Pichersky y Tanksley, 1988**; **Pichersky et al., 1991**), el tabaco *Nicotiana tabacum* (**Ayliffe y Timmis, 1992**), la papa *Solanum tuberosum* (**du Jardin, 1990**) y también en miembros de la familia *Chenopodiaceae (Beta vulgaris, C. album, Chenopodium quinoa, Atriplex cinerea,* y *Enchyleana tomentosa*) (**Ayliffe et al., 1988**). Estudios contemporáneos demuestran que el escape de material genético desde las organelas al núcleo ocurre con más frecuencia de lo que generalmente se cree (**Timmis et al., 2004**). Si esta continua dinámica estructural que presentan los *numts* (llamados "poltergeists moleculares" por **Hazkani-Covo et al., (2010**)) del condroma del tomate tiene consecuencias directas sobre el fenotipo y el vigor de la planta aún es una pregunta abierta.

Capítulo II

Capítulo II: El genoma del tomate silvestre (*Solanum pennellii*), parental donante de la población de líneas de introgresión más estudiada de la familia *Solanaceae*

II.1 | Introducción

II.1.1 | Características generales de la especie silvestre de tomate *Solanum pennellii*

Solanum pennellii es una especie de tomate silvestre endémica de las regiones andinas de América del Sur, donde ha evolucionado naturalmente para adaptarse y prosperar en su árido hábitat nativo (ver Figura 18). Teniendo esto en cuenta, *S. pennellii* presenta gran resistencia al estrés hídrico, a la salinidad y a la alcalinidad del suelo, así como también a insectos y patógenos. En este sentido, debido a su alta tolerancia al estrés biótico y abiótico y a su inusual morfología, es un importante donante de germoplasma ya que es sexualmente compatible con el tomate domesticado *Solanum lycopersicum* (**Rick y Tanksley, 1981**).

Figura 18: *Solanum pennellii* LA0716 en su hábitat natural sobre la cordillera Andina. Foto tomada de *The C.M. Rick Tomato Genetics Resource Center (TGRC)*.

El tomate cultivado tiene un fruto rojo, jugoso y de sabor dulzón, además sus hojas son compuestas, de proporciones grandes y con puntas aguzadas (ver Figura 4 y Figura 19). En contraste, si observamos a su pariente silvestre (*S. pennellii*) éste presenta un fruto pequeño de coloración verde con un sabor displicente al paladar y sus hojas compuestas son pequeñas y redondeadas (**Lockhart, 2013**) (ver Figura 19 y Figura 20).

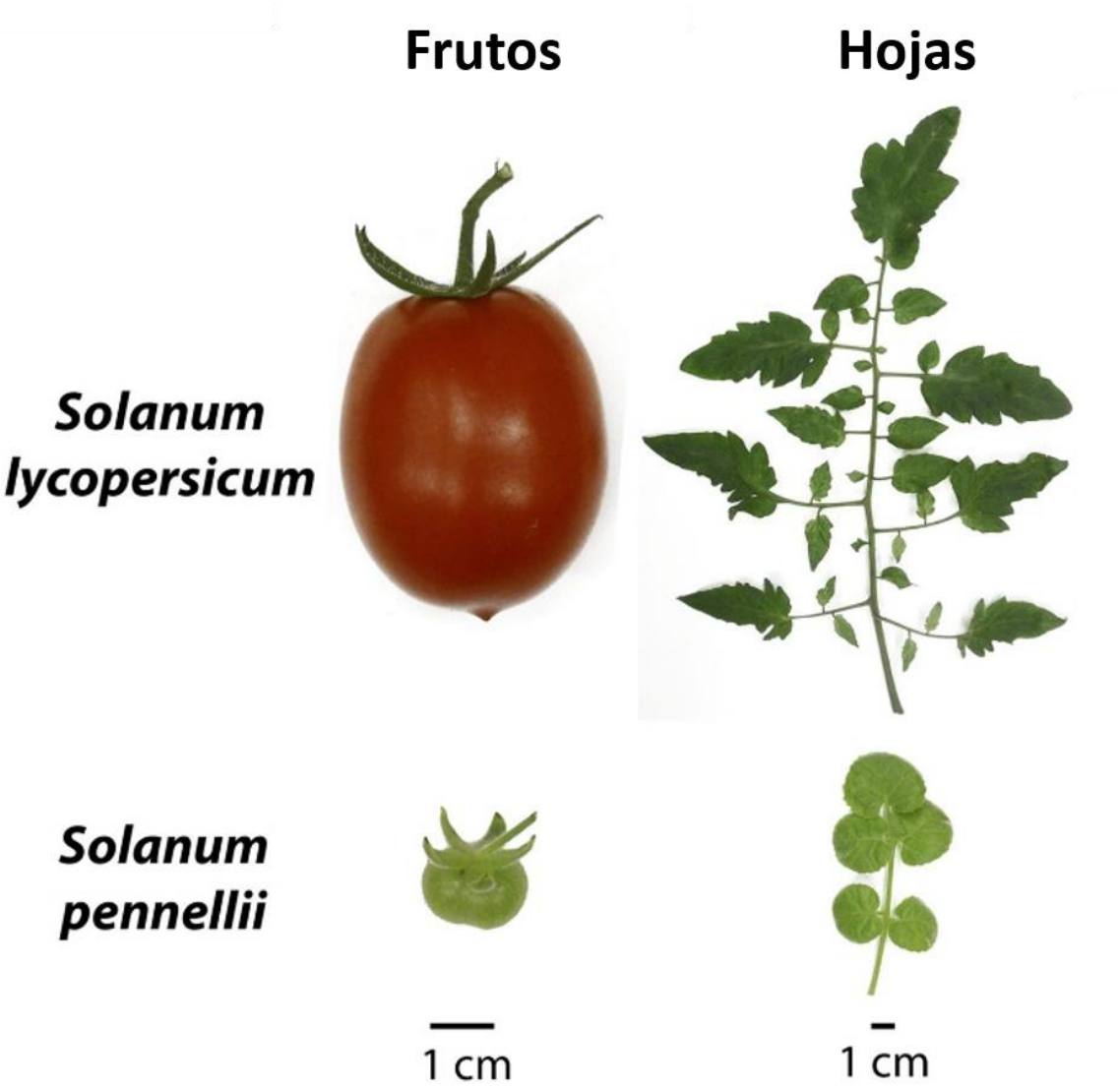


Figura 19: Diferencias morfológicas en el fruto y hojas entre el tomate domesticado (*S. lycopersicum*) y su pariente silvestre (*S. pennellii*). Fuente: (**Lockhart, 2013**).

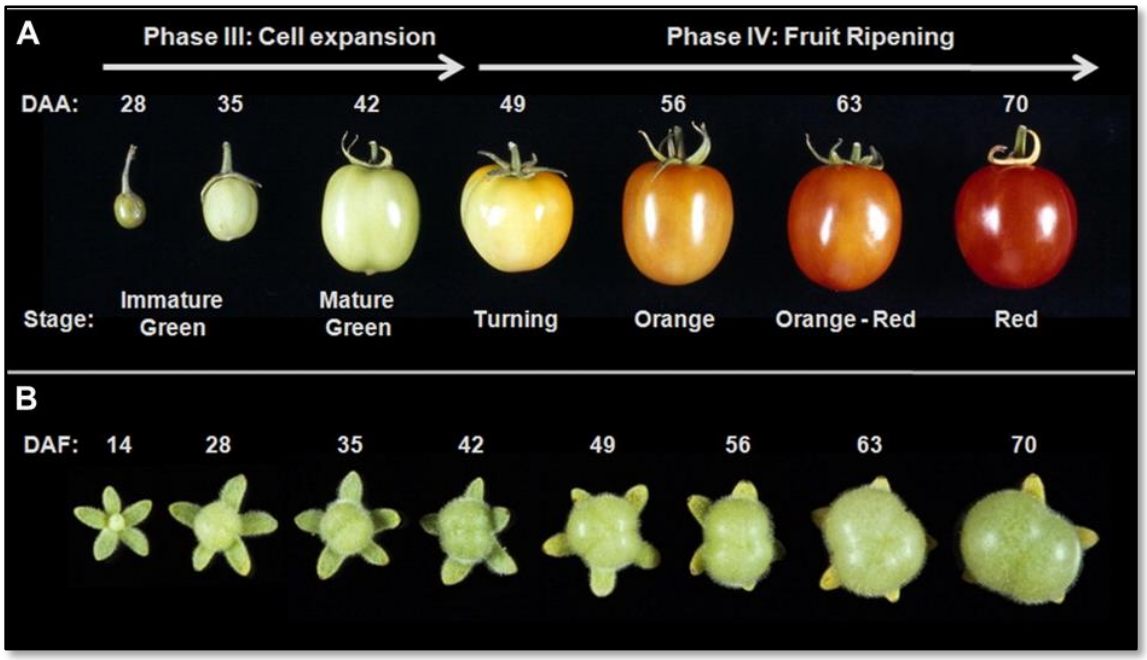


Figura 20: Características del crecimiento de los frutos cosechados en diferentes momentos durante el desarrollo de la fruta roja de *S. lycopersicum* (A) y de la fruta verde de *S. pennellii* (B). Figura tomada de **Steinhauser et al., (2010)**.

Cabe destacar, que hoy en día, en comparación con su pariente silvestre, el tomate domesticado tiene una variabilidad genética muy estrecha y con limitado margen para el mejoramiento. En buena parte, esta limitación se debe a varios cuellos de botella genéticos, por ejemplo: el efecto fundador, la deriva génica y las selecciones naturales y artificiales que ocurrieron durante la domesticación y la evolución de los cultivares modernos (**Rick et al., 1976**). Sin ir más lejos, los tomates introducidos por primera vez en Europa por los exploradores españoles proporcionaron toda la base genética de los cultivares modernos y, por consiguiente, los cultivares europeos y estadounidenses actuales resultan muy similares entre sí. En este sentido, se estima que sólo un 5% de la variación genética total dentro del género *Solanum* se puede encontrar hoy día dentro de la especie cultivada *lycopersicum* (**Miller y Tanksley, 1990**) y en consecuencia los genes detrás de muchas características agrícolas deseables no se encuentran en esta especie (**Foolad, 2007**). Sin embargo, se ha demostrado que las especies de tomate silvestre son una buena fuente de genes y caracteres deseables para el mejoramiento de cultivares elite. En este sentido, cabe mencionar que las especies

de tomate con mayor variabilidad genética hoy día son: *S. chilense*, *S. hirsutum*, *S. peruvianum* y *S. pennellii*. Por otra parte, las especies que presentan un estrecho margen de variabilidad son: *S. lycopersicum, S. cheesmanii* y *S. pimpinellifolium* (**Miller y Tanksley, 1990**; **Bretó et al., 1993**). Teniendo en cuenta que la radiación del clado de tomate se ha estimado en unos 7,8 millones de años (**Clint Nesbitt et al., 2002**) y que hay una divergencia evolutiva de 2,7 millones de años entre *S. lycopersicum* y *S. pennellii* (**Kamenetzky et al., 2010**) los cruzamientos entre estas plantas distantemente relacionadas (ver Figura 5, sección 0.1.6) han conducido a mejoras sustanciales en el aumento de variabilidad genética y rendimiento agronómico del tomate. En particular, la población de *ILs* obtenida del cruzamiento entre *S. pennellii* (LA0716) y *S. Lycopersicum* (cv M82) se ha utilizado para definir numerosos *QTL*s para el aumento del rendimiento, la composición química, la morfología, la tolerancia al estrés abiótico y la heterosis extrema (**Arumuganathan y Earle, 1991**; **Kamenetzky et al., 2010**). Aunque los estudios genéticos hasta la fecha han demostrado ser informativos, solo unos pocos genes subyacentes a *QTL* específicos han sido clonados, en gran parte debido a la falta de la secuencia completa del genoma de *S. pennellii*. De esta manera, para apoyar los análisis de *QTL* en pos del mejoramiento del cultivo de tomate, decidimos proceder con la secuenciación del genoma completo del parental silvestre de las *ILs* de tomate (**Bolger et al., 2014a**).

II.2 | Resultados

II.2.1 | Secuenciación, ensamblado, andamiaje y anclaje

La fragmentación aleatoria de ADN de alto peso molecular (*whole genome shotgun*) extraído de hojas de la especie de tomate silvestre *S. pennellii* redundó en bibliotecas genómicas (ver M&M, sección V.2.1) agregadas según el tamaño de los fragmentos obtenidos. Es decir, en bibliotecas de fragmentos de 150, 300 y 600 pb (*Illumina® PE,* ver M&M, sección V.2.3), y en bibliotecas de fragmentos grandes de 3, 5, 8, 10, 20, 30 y 40 Kpb (*Illumina® MP,* ver M&M, sección V.2.3) y por último en fragmentos de ~120 Kpb que fueron clonados en *BACs* (Sanger *PE,* ver M&M, sección V.2.3). Las puntas de estos fragmentos de ADN fueron secuenciadas y las lecturas obtenidas fueron preprocesadas por calidad y secuencias contaminantes (ver M&M, sección V.2.4), lo que resulto en un total de 2,66 billones de lecturas y 258 billones de pb (ver Tabla 5). En este sentido, las lecturas de las bibliotecas *Illumina® PE* se utilizaron para ensamblar *contigs* con la herramienta *SOAPdenovo* (ver M&M, sección V.2.5), estas bibliotecas brindaron una alta cobertura de secuencia (~190X) sobre el tamaño estimado del genoma de *S. pennellii* (1.2 billones de pb) lo que favoreció el solapamiento de las lecturas para reconstruir secuencias contiguas mediante el algoritmo de De Bruijn. Esto resultó en el ensamblado de 4.315.954 *contigs* con un N50 de 2.176 (ver Tabla 6). Luego, estos resultados se complementaron mediante el uso de las lecturas pareadas de las bibliotecas de fragmentos más grandes (*Illumina® MP* y Sanger *PE*) y la herramienta *SSpace,* con la cual se ensamblaron *scaffolds* mediante el andamiaje de los *contigs* obtenidos (ver M&M, sección V.2.6). El nuevo ensamblado resultó en un total de 407.506 *scaffolds* con un N50 de 1.603.317 (ver Tabla 7). De este ensamblado se seleccionaron solo aquellos *scaffolds* con un tamaño superior a 2 Kpb (ver M&M, sección V.2.6), lo que resulto en un ensamblado preliminar de alta calidad contenido en 4.579 *scaffolds* (ver Tabla 6) que abarcan 942.595.034 de pb o el 79% del tamaño estimado del genoma. Luego, los 4.579 *scaffolds* fueron anclados al mapa genético *Tomato-EXPEN 2000* (ver M&M, sección V.2.8) utilizando una selección de 2.746 marcadores moleculares unívocos, con los cuales un 67% de los *scaffolds* fue posicionado y orientado sobre el mapa genético (ver Tabla 7). Para

finalizar la construcción de las pseudomoléculas basándose en la alta sintenia entre *S pennellii* y *S lycopersicum* (ver sección 0.1.6) se ordenaron y orientaron los *scaffolds* restantes alineando estos contra el genoma del cultivar Heinz v2.40 mediante la herramienta *MUMmer,* posicionado así el 97% de los *scaffolds* del genoma de *Solanum pennellii* sobre las 12 pseudomoléculas (ver M&M, sección V.2.8). En este sentido, el 3% restante de los *scaffolds* que quedaron sin anclar fueron agregados a una pseudomolécula artificial a la cual se denominó cromosoma cero (*cho*) hasta que haya información disponible que permita posicionarlos en el mapa (ver Tabla 6, Anclado v 1.90). En resumen, esto resulto en la construcción de 12 super *scaffolds* o pseudomoléculas (cromosomas) que conformaron el borrador final (v 1.90) del genoma nuclear de la especie silvestre de tomate *Solanum pennellii* (ver Figura 21). Este montaje final compuesto por unas 942,6 Mpb (**Bolger et al., 2014a**), es consistente con el tamaño genómico de 1.200 Mpb determinado anteriormente por citometría de flujo (**Arumuganathan et al., 1991**).

Tabla 5: Resumen de las lecturas totales y cantidad de pares de bases obtenidas de la secuenciación y preprocesamiento de las bibliotecas de ADN de *S. pennellii.*

Secuenciación	∑ lecturas	∑ pb
Illumina® PE	1,73 Bn	165 Bn
Illumina® MP	935 M	93 Bn
Sanger PE	23.488	15,8 M
Total	2,66 Bn	258 Bn

Tabla 6: Resumen de los *contigs* empalmados en *scaffolds* utilizando el software *SSPACE* a partir de los datos de distancia entre las puntas de las librerías *MP* y *PE* de *Illumina*®.

Etapa	N50	*contigs* o *scaffolds* totales	nucleótidos totales	Ns
contigs	2.176	4.315.954	1.117.562.721	0
scaffolds	1.603.317	407.506	1.021.472.455	125.806.430
scaffolds > 2 Kpb	1.741.129	4.579	942.595.034	67.190.021
Anclado (v1.90)	1.741.129	13	942.595.034	67.190.021

Tabla 7: *scaffolds* anclados al mapa genético EXPEN2000 mediante marcadores moleculares unívocos (para más detalle, ver M&M, sección V.2.8).

marcadores unívocos	*scaffolds* anclados	nucleótidos anclados	% anclado del genoma
2.746	518	630.866.022	67%

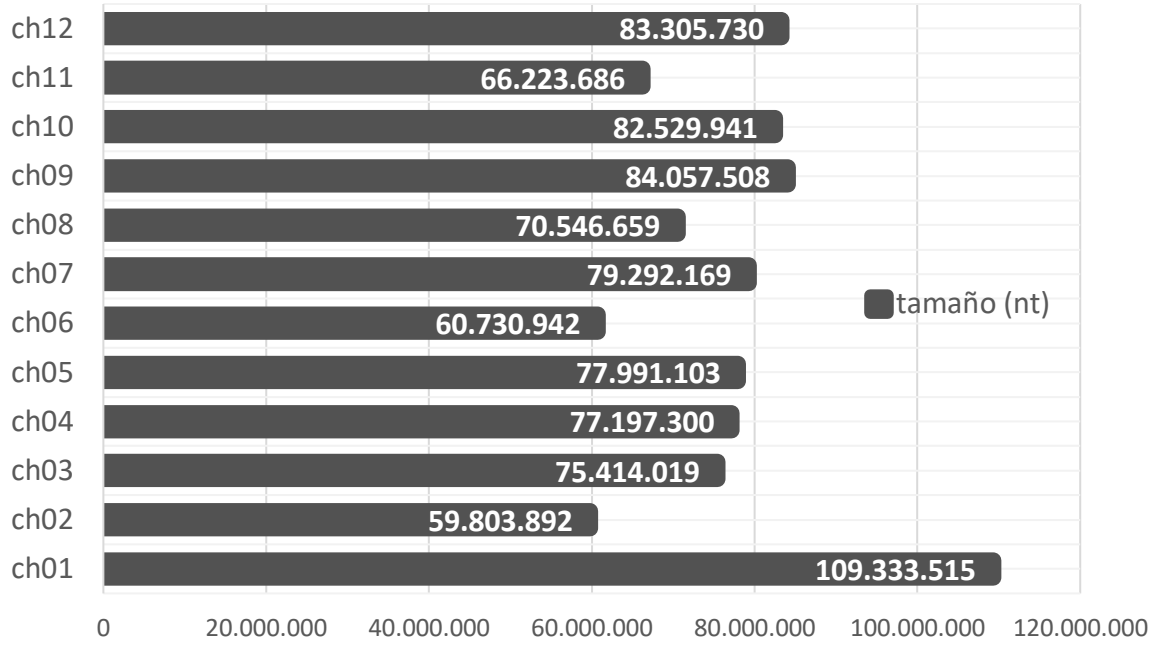


Figura 21: Esquema representativo de las 12 pseudomoléculas (cromosomas) del genoma ensamblado de *S. pennellii* (LA0716) v1.90. El tamaño total del genoma fue de 948.595.034 nt y se representa distribuido en 12 barras horizontales de color gris. En el eje horizontal se representa el tamaño en nucleótidos y en el eje vertical la identificación del cromosoma ensamblado (ch01 – ch12, se omite el ch0).

La comparación entre los genomas de *S. pennellii, S. lycopersicum* cv Heinz y *S. tuberosum* reveló que el borrador del genoma de *S. pennellii* es, a pesar de haber sido ensamblado solamente a partir de datos de secuenciación de segunda generación (*NGS*), estadísticamente comparable con el borrador del genoma de *S. tuberosum*. En este sentido, el genoma de *S. pennellii* tiene un mayor tamaño que el de la papa (942 Mpb vs 715 Mpb) y además también presenta un N50 moderadamente mayor (1,74 Mpb vs. 1,35 Mpb), aunque por otro lado presenta un mayor porcentaje de indeterminaciones ("Ns") en su secuencia (7,1 % vs 6,2 %). Por

otra parte, cabe destacar que el genoma de *S. lycopersicum*, fue obtenido aplicando tanto la técnica basada en clones como la técnica de fragmentación aleatoria de genoma completo (*WGS*) y más aún se utilizó una com*bin*ación de tecnologías de secuenciación de Sanger y *NGS*, lo que resultó en que el ensamblado del cultivar Heinz sea el más completo del grupo, con un N50 de 16.5Mpb. En este sentido, si fragmentamos los 4.697 *scaffolds* en los *contigs* que los componen y comparamos el tamaño final de los *contigs* de cada uno de los genomas, *S. pennellii* presenta un "*contig* N50" de 45,7 Kpb y queda ubicado entre *S. tuberosum* (31,4 Kpb) y *S. lycopersicum* cv Heinz (86,9 Kpb). Por último, los datos de las librerías de secuencias crudas de la especie silvestre *S. pennellii* (LA0716) fueron depositados en la base de datos pública del "*European Bioinformatics Institute* (EBI)" bajo el número de proyecto PRJEB5809. Por otra parte, la secuencia del genoma ensamblado v1.90 fue depositada en el "*European Nucleotide Archive* (ENA)" bajo el número de entrada HG975439–HG975452. ↜

II.2.2 | Anotación estructural de genes

Se utilizó la herramienta *Augustus* (**Stanke et al., 2004**) entrenada para la especie cercana *S. lycopersicum*. En este sentido, *Augustus* fue suministrado con datos experimentales de secuencias de *RNAseq* y de *ESTs* de *S. pennellii* para asistir en las predicciones. Esto resultó en la predicción de 44.966 estructuras génicas codificantes (*CDS*) y sus respectivas traducciones a secuencias de aminoácidos (AA) disponibles como archivos del tipo GFF3 (ver abreviaturas). Estos archivos pueden descargarse vía *FTP* desde el portal de internet público y especializado en solanáceas: ftp://ftp.solgenomics.net/genomes/Solanum_pennellii

II.2.3 | Anotación funcional de genes

Los 44.966 transcriptos identificados durante la anotación estructural del genoma de *S. pennellii* se procesaron con la herramienta *Mercator* (**Lohse et al., 2014**) para realizar la anotación funcional de genes mediante la asignación de categorías *MapMan* (términos de ontología génica (GO) construidos específicamente para genomas de plantas). Como resultado *Mercator* clasifico 52.074 proteínas en

grupos, de las cuales se asignaron 25.428 (aproximadamente el 44,6%) en categorías *MapMan* organizadas por función biológica (ver Figura 23), las restantes 26.646 proteínas (aproximadamente el 53,5%) se agruparon en la categoría 35 (*not assigned*).

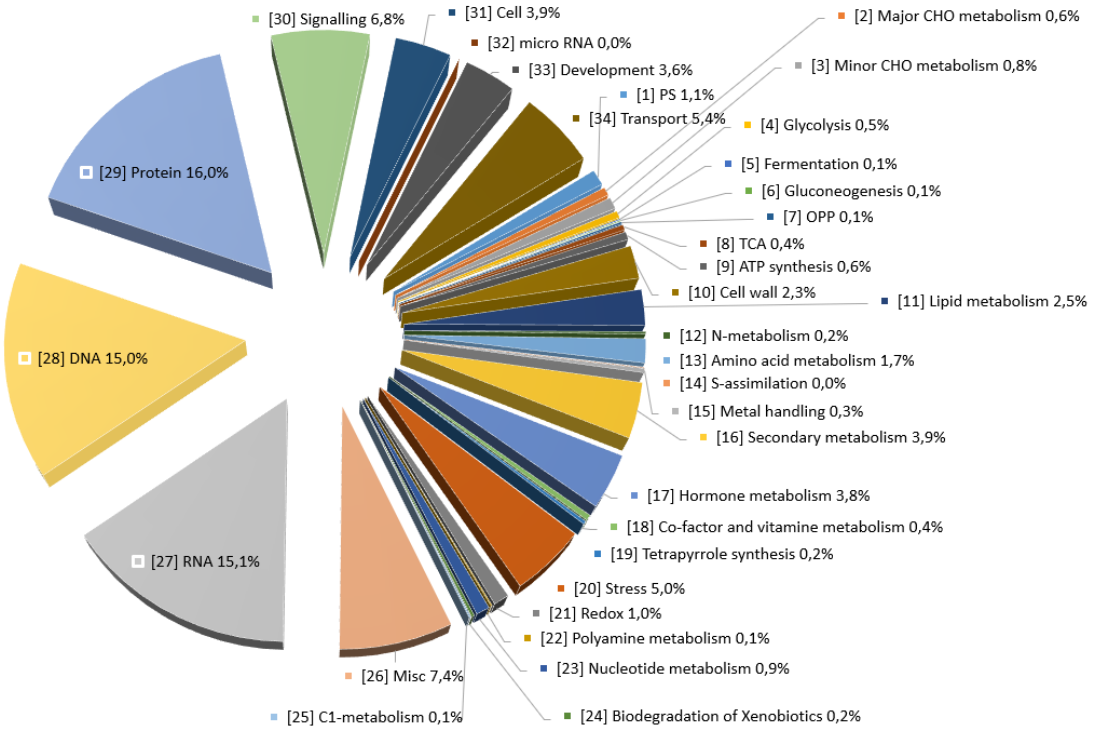


Figura 23: Categorías funcionales asignadas a 25.428 proteínas de *S. pennellii* con la herramienta *Mercator*. Los modelos génicos (CDS) de *S. pennellii* obtenidos con la herramienta *Augustus* se sometieron a la predicción funcional de *Mercator*. Las anotaciones resultantes se clasificaron por términos *MapMan* según su función biológica (**Bolger et al., 2014a**). Entre corchetes se aprecia el *MapMan BIN* (0 - 35) que identifica a cada jerarquía de la ontología *MapMan*, luego el nombre funcional de cada *BIN* y el porcentaje que este representa del total de proteínas asignadas a estas funciones. En este gráfico se omitió el *BIN [35] not assined 53,5%* para una mayor legibilidad y dar foco a las proteínas funcionalmente anotadas mediante *Mercator*.

II.2.4 | Resumen de los principales resultados

1. Se obtuvieron genotecas de *S. pennellii* de diversos tamaños de insertos (300 pb – 40 Kpb) que fueron secuenciadas utilizando la tecnología *Illumina® PE* y *MP*.
2. La información de secuencia de estas genotecas redundó en lecturas de alta calidad que se utilizaron para el ensamblado de *contigs* y *scaffolds*.
3. A partir de la base de datos de marcadores moleculares construida, con información posicional unívoca sobre el mapa genético EXPEN2000, fue posible anclar (ordenar y orientar) los *scaffolds* en 12 pseudo-cromosomas/moléculas.
4. Se anotaron las estructuras génicas de *Solanum pennellii*; éstas fueron caracterizadas utilizando búsquedas por similitud de secuencias con la herramienta de anotación en base a ontologías génicas *Mercator,* luego las proteínas fueron clasificadas y agrupadas según sus funciones biológicas en los niveles superiores de los términos (*BIN*) ontológicos *MapMan*.
5. Los resultados de este capítulo permitieron obtener el primer borrador del genoma completo de la especie silvestre de tomate *S. pennellii*, el padre donante de la población de *ILs* más utilizada para el mapeo de *QTL* en esta especie.

II.3 | Discusión

La biodiversidad natural proporciona una oportunidad para investigar la multitud de caracteres que afectan la supervivencia de las plantas, la aptitud reproductiva y su potencial de rendimiento. En la investigación agronómica, las especies silvestres relacionadas con las plantas cultivadas ofrecen un medio para comprender las bases genéticas de los eventos de domesticación pasados y sobre todo para identificar los polimorfismos que proporcionan una base para la obtención de futuras variedades superiores (**Ofner et al., 2016**). El valor de las especies silvestres para entender la biología de los cultivos se destacó poco después del redescubrimiento de las leyes de Mendel (**Bessey, 1906**), y pasó casi un siglo antes de que se ejecutaran por primera vez programas de introgresión a gran escala para caracteres simples y complejos (**Zamir et al., 2001**). Previo a la obtención de los genomas de los parentales de las *ILs* (*S. pennellii* x *S. lycopersicum*) la identificación de *loci* de caracteres cuantitativos (*QTL*) y los genes responsables de los caracteres de interés agronómico se realizaba por medio de marcadores moleculares, los cuales permitieron que el genoma de la población de *ILs* sea segmentado en 107 *BIN*s (**Eshed y Zamir, 1995**; **Pan et al., 2000**) (ver Figura 7, sección 0.1.6). Esta población de líneas casi isogénicas ha sido ampliamente caracterizada con el fin de identificar diferencias en caracteres de interés agronómico, siendo mapeados a la fecha más de 2.000 *QTL* (**Lippman et al., 2007**). Por ejemplo, como se mencionó anteriormente (ver sección 0.1.6), el grupo de **Rousseaux et al., (2005)** estudió metabolitos relacionados a la calidad nutricional y presencia de antioxidantes en el fruto de esta población. Además de este estudio, el grupo de la Dra. Causse asigno 130 *QTL* a 38 caracteres relacionados con la calidad organoléptica y 81 *QTL* asociados al tamaño, contenido de ácidos orgánicos y azúcares de fruto (**Causse, 2002**; **Causse et al., 2004**). Debido a su estructura genómica, las *ILs* son una herramienta de alto valor para develar los determinantes genéticos detrás de los caracteres cuantitativos, ya que todas las diferencias fenotípicas entre los linajes introgresados y *S. lycopersicum* (cv. M82) están determinadas sólo por el fragmento de *S. pennellii*. Esto fue claramente demostrado por el estudio realizado por **Fridman et al., (2000** y **2004)**, en el que se identificó que la IL 9-2-5 presentaba un incremento del 25% en el contenido de sólidos solubles de los frutos

(° *Brix*). Después del mapeo genético y físico de alta resolución, un fragmento de la región introgresada de *S. pennellii* fue secuenciada. En ella se identificó un gen que codifica para una invertasa de pared celular, LIN5. La comparación de los alelos silvestres y cultivados permitió concluir que el *QTL* está determinado por el cambio de un único nucleótido en el tercer exón del gen, siendo que ese cambio modificaba las propiedades cinéticas de la enzima. Esta invertasa es clave en la carga de sacarosa en el floema, siendo este el mayor factor involucrado en el incremento del contenido de azúcar en el fruto (**Baxter et al., 2005**). En este sentido, el genoma completo del parental donante presentado en este Capítulo de Tesis (**Bolger et al., 2014**) proporciona una valiosa herramienta para acelerar el descubrimiento de *QTL* y genes clave en vías metabólicas de interés agronómico y nutricional, como son la fotosíntesis y la respiración (este tema se aborda en más profundidad en el Capítulo III de esta Tesis). En un ejemplo reciente de nuestro laboratorio, utilizando el genoma secuenciado de *S. pennellii,* se logró identificar los mecanismos que determinan la acumulación de vitamina E en el fruto del tomate cultivado (**Quadrana et al., 2014**). Como perspectiva, una comprensión más completa de estos genes candidato, junto con aquellos involucrados en las vías de biosíntesis de flavonoides y vitaminas, facilitará enormemente la obtención de frutos de mejor sabor y calidad nutritiva que es lo que hoy día el consumidor mayormente demanda. Por último, los resultados presentados en este Capítulo demuestran la importancia de la conservación de la biodiversidad en las ecorregiones silvestres como fuentes esenciales de germoplasma e información genética para el mejoramiento agrobiotecnológico.

Capítulo III

Capítulo III: Análisis de regiones genómicas (*QTL*) e identificación de genes candidatos asociados a la regulación de procesos fisiológicos esenciales como la fotosíntesis, respiración y vías metabólicas asociadas

III.1 | Introducción

La vida en la tierra depende de la capacidad de las plantas para convertir la energía solar en biomasa. En este sentido, nuestro conocimiento sobre los determinantes moleculares y fisiológicos detrás del desarrollo y maduración de las plantas con frutos carnosos, como la respiración y la fotorrespiración, son esenciales para mejorar el rendimiento y la calidad de la producción de tomate. Con esto en mente, determinar las variaciones naturales que se dan sobre este tipo de procesos esenciales, pero en diferentes años de cosecha, pueden revelar las bases de los caracteres cuantitativos asociados con el rendimiento de las plantas; y de esta manera, proporcionar información sobre los mecanismos involucrados en la adaptación de la planta a un entorno variable. La comprensión de estos mecanismos de respuesta a un entorno dinámico, pavimentarán el camino hacia modelos que puedan predecir la productividad de los cultivos dentro de un contexto de cambio climático global (**Zou et al., 2007**; **Mozdzer y Zieman, 2010**). En este sentido, la mejora de la eficiencia fotosintética puede lograrse a través de varias estrategias diferentes, por ejemplo, la mejora de la captura de la luz o de la conversión de la energía lumínica, el aumento de la tasa de fijación y conversión de carbono o eludiendo la fotorrespiración (**Ort et al., 2015**). Sin embargo, aumentar de forma independiente la actividad fotosintética del órgano fuente sin relación con un aumento en el transporte de azúcar, que fortalezca al órgano destino, lleva a la acumulación de carbohidratos en las hojas, lo que, en contrapartida, da lugar a una disminución de la fotosíntesis y a una menor eficiencia de la misma en términos de unidad de espacio. Como resultado, las estrategias para mejorar la fotosíntesis deben situarse en el contexto de las complejas relaciones que hay entre todos los órganos de la planta (**Rossi et al., 2015**).

En este capítulo de Tesis vamos a recurrir al hecho de que existe una variación natural en la fotosíntesis de las plantas dentro y entre las especies (**Hikosaka, 2010**; **Lawson et al., 2012**), y que por lo tanto, puede explorarse para identificar determinantes genéticos que permitan una mejora adicional en el rendimiento de los cultivos (**Flood et al., 2011**; **Lawson et al., 2012**). La variación genética asociada a variaciones en la eficiencia fotosintética, determinada por parámetros como la clorofila y la fluorescencia, se ha investigado previamente en varias especies, incluyendo: *Arabidopsis thaliana* (**El-Lithy et al., 2005**; **Rooijen et al., 2015**), trigo (*Triticum aestivum*) (**Bertin y Gallais, 2001**; **Yang et al., 2007**; **Zhang et al., 2009**; **Driever et al., 2014**), arroz (*Oryza sativa*) (**Gu et al., 2012**) y soja (*Glycine max*) (**Yin et al., 2010**). Sorprendentemente, a pesar del creciente número de estudios basados en variaciones naturales, los factores genéticos involucrados en la regulación de la fotosíntesis siguen siendo, en gran medida, desconocidos (**Flood et al., 2011**). Esto se debe probablemente a que las variantes alélicas generalmente no se encuentran en los genes fotosintéticos centrales conservados, sino en genes que codifican "factores auxiliares" o "modificadores" de la vía de la fotosíntesis (**Flood et al., 2011**). De lo antedicho, se deduce que el uso de la variación natural para estudiar la función de los genes que codifican para tales reguladores podría proporcionar una mayor comprensión de los mecanismos de regulación de la fotosíntesis. Es importante destacar que tales reguladores no han sido identificados. Esto es probable debido a su redundancia o su gran número y por otra parte el hecho de que los reguladores individuales sólo pueden explicar una pequeña fracción de la variación en la fotosíntesis. En contraste, la variación natural en genes que regulan el crecimiento de las plantas se ha descrito en un gran número de especies, incluyendo a la planta modelo *Arabidopsis thaliana* (**Cross et al., 2006**; **Meyer et al., 2007**; **Sulpice et al., 2009**) como también en plantas de cultivo (**Semel et al., 2006**; **Liang et al., 2009**; **Driever et al., 2014**; **Rosado-Souza et al., 2015**; **Wen et al., 2016**). Sin embargo, similar a la situación que describimos anteriormente para la fotosíntesis, aunque estos estudios revelaron una variación natural considerable en el crecimiento de la planta, se sabe relativamente poco acerca de su base genética, probablemente debido a que, por

ejemplo, la tasa de crecimiento es un integrador de numerosos factores interdependientes (**Koornneef et al., 2004**).

Como se mencionó en capítulos anteriores, en el tomate, se han realizado numerosos estudios, utilizando la población de *ILs* de *S. pennellii* x M82 (**Eshed y Zamir, 1995**), en un intento por entender la composición de su fruto (**Frary et al., 2000**; **Fridman et al., 2000**; **Liu et al., 2003a**; **Schauer et al., 2006**; **Tieman et al., 2006**; **Do et al., 2010**; **Schilmiller et al., 2010**; **Quadrana et al., 2014**). Estudios realizados con esta población de *ILs*, han permitido identificar cientos de *QTL* relacionados con caracteres del fruto de tomate (**Semel et al., 2006**), como la composición de metabolitos (**Causse et al., 2004**; **Schauer et al., 2006**; **Almeida et al., 2011**), el color del fruto (**Liu et al., 2003b**) y también con actividades enzimáticas producidas durante el desarrollo del fruto (**Steinhauser et al., 2010**). Esta población también se ha utilizado para identificar *QTL* involucrados en la morfología radicular y foliar (**Lockhart, 2013**; **Ron et al., 2013**) indicando varios genes candidatos (**Kamenetzky et al., 2010**), que han sido clonados inequívocamente (**De Godoy et al., 2013**; **Quadrana et al., 2014**).

Por otro lado, caracteres relacionados con la fotosíntesis también se han estudiado en tomate, como por ejemplo, los determinantes morfológicos y anatómicos de la conductancia mesofílica en especies de tomate silvestre (**Muir et al., 2014**). Además, se han mapeado *QTL* en la población de *ILs* para: la composición de isótopos de carbono en hojas (**Xu et al., 2008**), los rasgos foliares asociados con la adaptación a la variación en el suministro de agua y caracteres involucrados con la respuesta anatómica relacionada a la conductancia estomática (g_s) (**Fanourakis et al., 2015**). Así, esta colección de *ILs*, ampliamente estudiada, junto con la disponibilidad de la secuencia del genoma de las especies parentales (*S. pennellii* y M82) (Capítulo II de esta Tesis y **Bolger et al., 2014a**), constituyen una potente herramienta para la identificación de determinantes genéticos involucrados en la fisiología, el crecimiento y el desarrollo de la planta de tomate y sus frutos.

De esta manera, el objetivo principal de este Capítulo es identificar genes clave entre los *loci* de caracteres cuantitativos relacionados con los procesos de

fotosíntesis, respiración y crecimiento de la planta de tomate. Para lo cual, se analizaron dos experimentos independientes realizados con esta población de *ILs* y el cultivar M82 como control (ver M&M, sección V.3.1). Estos experimentos consistieron en cultivar las plantas durante dos años consecutivos (la primavera del año 2012 y la del año 2013) en condiciones controladas de invernáculo y registrar parámetros fotosintéticos y de crecimiento en plantas de 4 semanas, previo a la floración de las mismas (ver Figura 24).

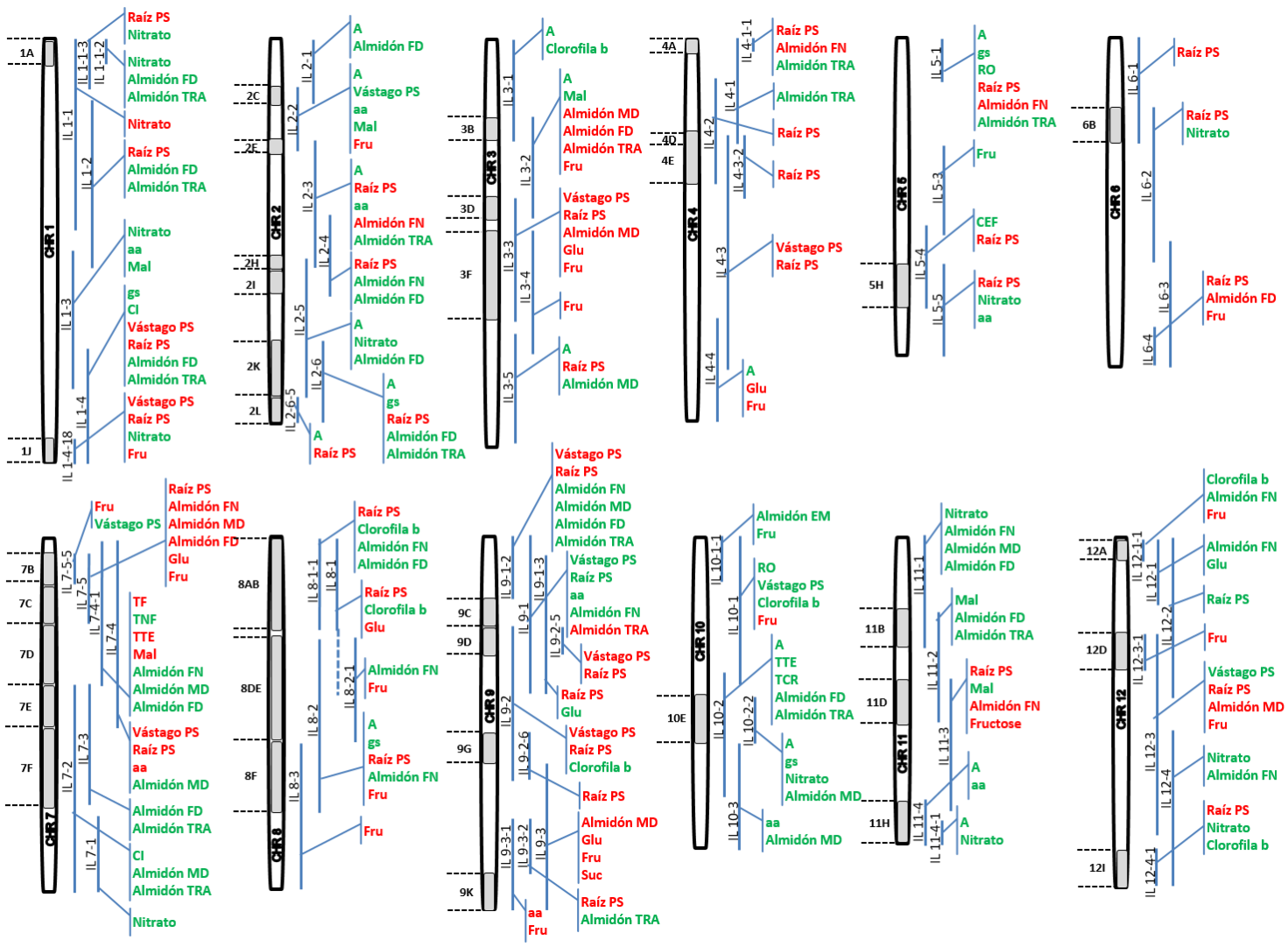
Figura 24: Las 71 *ILs* y el cv M82 se cultivaron en condiciones semi-controladas en un invernadero ubicado en la localidad de Viçosa (642 m de altitud, 20º45' de latitud sudoriental y 42º51' de longitud oeste) en el sudeste de Brasil, con una intensidad lumínica media de 400 µmol de fotones m-2 s-1 diarios.

III.2 | Resultados

En este capítulo, se abordó el análisis de caracteres cuantitativos asociados a la eficiencia fotosintética en una población de *ILs* de tomate (71 *ILs* y el padre recurrente M82), incluyendo los contenidos de las hojas de un conjunto de metabolitos y parámetros relacionados con el crecimiento. A partir de la comparación de estos parámetros entre las distintas *ILs* con el padre recurrente se identificaron 208 *QTL*, entre los que se destacan la acumulación de materia seca en raíces y tallos y niveles de metabolitos primarios claves (aminoácidos, glucosa, sacarosa, fructosa y malato), se realizaron análisis genómicos de las regiones involucradas y se identificaron 87 genes candidatos involucrados en la regulación de estos caracteres.

III.2.1 | *QTL* ligados a fotosíntesis, metabolismo y crecimiento

Para mapear los *QTL* a intervalos genómicos más pequeños se identificaron las superposiciones (denominadas *BIN*s) entre los segmentos introgresados en las 71 *ILs* (ver Figura 7 en sección 0.1) para aquellos caracteres que mostraron diferencias significativas en los caracteres medidos en comparación con el padre recurrente M82 (ver M&M, sección V.3.4). Mediante la aplicación de este criterio se identificaron 33 *BIN*s asociados a *QTL* y distribuidos en regiones superpuestas (ver Figura 25 y Tabla 8). En este sentido, utilizando los marcadores moleculares que flanquean a las regiones mencionadas (ver Figura 7 en sección 0.1), se obtuvieron las secuencias de cada una de estas regiones de los genomas ensamblados de los parentales *S lycopersicum* (cv M82) y *S. pennellii* (ver Tabla Anexa 2, en Anexos VIII.3). Luego, estos 33 fragmentos de cada genoma fueron anotados de forma estructural y *ab-initio* mediante la herramienta *Augustus* (**Stanke et al., 2004**) para obtener los *loci* codificantes dentro de cada alelo (ver Tabla 8 y M&M, sección V.3.4). Cabe destacar, que en total se detectaron 1.676 (15%) más genes de *S. pennellii* que de *S. lycopersicum* entre todos los *BIN*s, en particular se observó que el *BIN* 3D y 4DE (asociados al contenido de almidón y biomasa radical, respectivamente) mostraron 47% y 12% más *genes* en *S. pennellii* que en M82, respectivamente. Por otro lado, sólo cuatro *BIN*s: 2C, 4A, 9C y 11D, mostraron el mismo número de genes predichos en las dos especies parentales (ver Tabla 8).



i) Intercambio de gases y parámetros de fluorescencia de la clorofila *a*:
A - Fotosíntesis
gs - Conductancia estomática
TF - Tasa fotorrespiratoria
RO - Respiración en oscuridad
CI - Concentración intercelular de CO_2
CNF - Coeficiente de enfriamiento no-fotoquímico
CEF - Coeficiente de enfriamiento fotoquímico
TTE - Tasa de transporte de electrones (TTE)
TCR - Tasa de crecimiento relativo (TCR)
Vástago PS - Peso seco del vástago
Raíz PS - Peso seco de la raíz
ii) Compuestos que contienen nitrógeno y carbono en las hojas:
Mal – Malato
Fru – Fructosa
Suc - Sacarosa
Glu - Glucosa
Clorofila *b*
aa - Aminoácido
Niveles de Almidón al final del día (Almidón FD),
del medio día (Almidón MD) y
de la noche (Almidón FN)
Almidón TRA - Tasa de recambio de almidón.

Figura 25: Mapa de 33 *BIN*s (barras grises) asociados a *QTL* de interés mostrando cambios significativos en parámetros i) fotosintéticos (Intercambio de gases y parámetros de fluorescencia de la clorofila *a* y ii) metabólicos relacionados con crecimiento vegetal (Compuestos que contienen nitrógeno y carbono en las hojas) medidos en dos temporadas (2012 y 2013) sobre 71 *ILs* (barras azules) con un fondo genético de *S. lycopersicum* (cv. M82). Se muestran los caracteres incrementados (verdes) o disminuidos (rojos) significativamente en contraste con M82 (según prueba de *Dunnett*, $p < 0{,}05$). Figura tomada y adaptada de: **de Oliveira Silva et al. (2017)**.

Tabla 8: Contenido génico de los 33 *BIN*s superpuestos a los *QTL* analizados. Tabla tomada y adaptada de: **de Oliveira Silva et al. (2017)**.

		# de *genes* predichos	
Caracteres*	***BINs***	***S. lycopersicum***	***S. pennellii***
A	2C	1	1
	2H	166	177
	3B	44	49
	10E	319	324
	11H	62	68
A y aa	2E	163	168
A y Almidón FD	2K	138	149
A y Raíz PS	2L	125	130
Raíz PS	4DE	1.140	1.292
	5H	95	104
	6B	249	280
	7BC	263	262
	9D	54	56
	9G	74	76
Raíz y Vástago PS	1J	202	208
	9C	75	75
Raíz PS y Clorofila *b*	8AB	3.461	3.942
Almidón FD	2I	305	314
	11B	100	104
Almidón MD	7E	472	449
Almidón FN	12A	117	127
Almidón TRA	4A	0	0
	7F	168	165
Almidón MD y Fru	3D	1.808	2.663
Almidón FN y Fru	8DE	505	516
Mal	11D	17	17
Nitrato	1A	64	70
	12I	21	24
Fru	3F	380	393
	7B	279	291
	8F	122	126
	9K	120	119
	12D	303	349
Total		11.412	13.088

* Las abreviaturas de los caracteres fueron tomadas de la Figura 25.

Teniendo en cuenta lo antedicho, para todos los análisis genómicos posteriores nos centramos en 11 de estos *BIN*s, ya que los mismos parecen contener la base genética para cambios fisiológicos y metabólicos en los caracteres de mayor interés (ver Figura 25). En este sentido, con el objeto de conocer más en detalle las funciones biológicas que subyacen a los *loci* predichos, se utilizó la herramienta *Augustus* para extraer las secuencias de proteínas de estos. Las proteínas predichas se anotaron funcionalmente mediante un alineamiento recíproco con el motor de búsqueda *BLASTP* (del inglés, *RBH: reciprocal best hit*) contra la base de datos de anotaciones del cultivar Heinz (ITAG 2.40) (**The Tomato Genome Consortium, 2012**) utilizando un parámetro de corte de E-valor = 1E-10. Luego, las proteínas fueron clasificadas y agregadas en grupos por función biológica utilizando una base de datos pública de ontologías *MapMan* de *S. lycopersicum* (**Lohse et al., 2014**). A continuación, para develar variantes en las bases genéticas que subyacen a estos caracteres, se evaluaron, mediante una tabla de contingencia con prueba de *Fisher* (p-valor < 0,05), los términos *MapMan* (ver M&M, sección V.3.4 para más detalle) asociados con las proteínas de los 11 *BIN*s de *S. pennellii versus* los de *S. lycopersicum* (ver Tabla Anexa 3, en Anexos VIII.3). Con estos resultados, a excepción del *BIN* 9C, se identificaron categorías funcionales *MapMan* sobrerrepresentadas en los *BIN*s de ambas especies (ver Tabla 9).

Tabla 9: Términos *MapMan* sobrerrepresentados (# de *loci* y p-valor) en los 11 *BIN*s seleccionados de las 71 *ILs* de *S. pennellii versus S. lycopersicum* cv. M82. Tabla tomada y adaptada de **de Oliveira Silva et al. (2017)**.

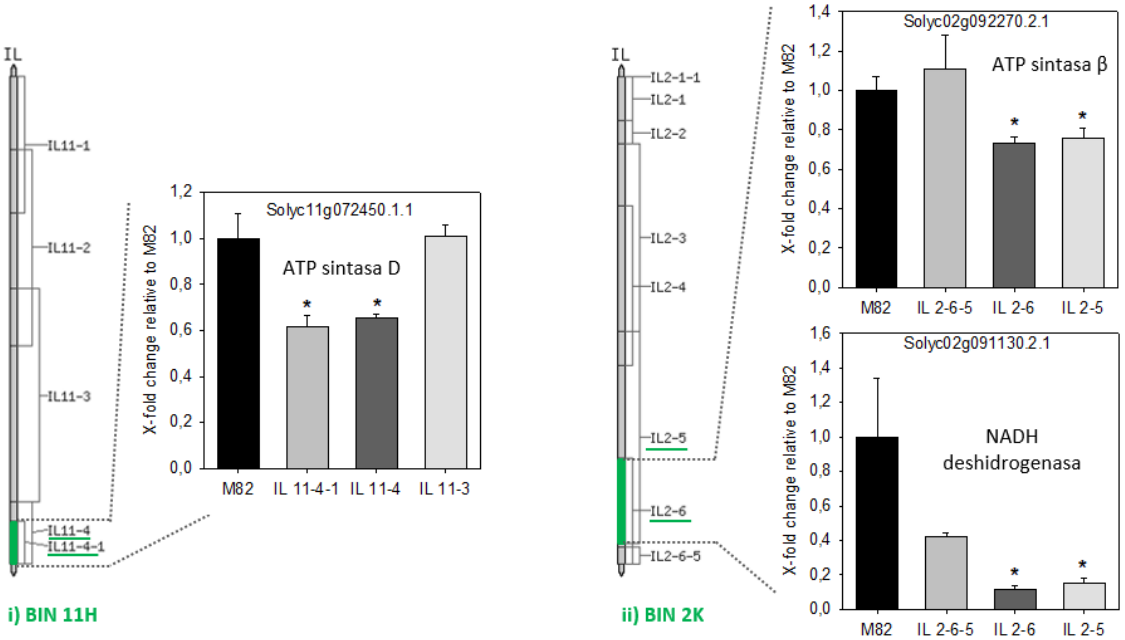
***BIN*s**	**Caracteres**	**Termino *MapMan* sobrerrepresentado**	***S. pennellii***		***S. lycopersicum***	
		Función biológica (código #)	**Σ de *loci***	**p-valor***	**Σ de *loci***	**p-valor***
1J	Raíz y Vástago PS	Desarrollo (33)			13	0,032
		Metabolismo de poliaminas (22)	2	0,013		
		Misceláneas (26)	22	0,038		
2L	*A* y Raíz PS	Biodegradación de xenobióticos (24)			2	0,011
		Misceláneas (26)	16	0,021		
4DE	Raíz PS	Señalización (30)			78	0,047
9G	Raíz PS	ARN (27)	18	0,016		
		Estrés (20)	8	0,024		
7BC	Raíz PS	Metabolismo mayor de CHO (2)			6	0,025
		Metabolismo lipídico (11)			10	0,036
		Fotosíntesis (1)			9	0,049
2K	*A* y Almidón FD	Pared celular (10)			7	0,047
		Metabolismo menor de CHO (3)	4	0,014		
11H	*A*	Tratamiento de metales (15)	2	0,019		
9D	Raíz y Vástago PS	TCA / transformación orgánica (8)	2	0,011		
11B	Almidón FD	Metabolismo de aminoácidos (13)	5	0,016		
		Sin asignar (35)	12	0,046		
11D	Malato	Tratamiento de metales (15)			2	0,001
		ARN (27)	6	0,018		

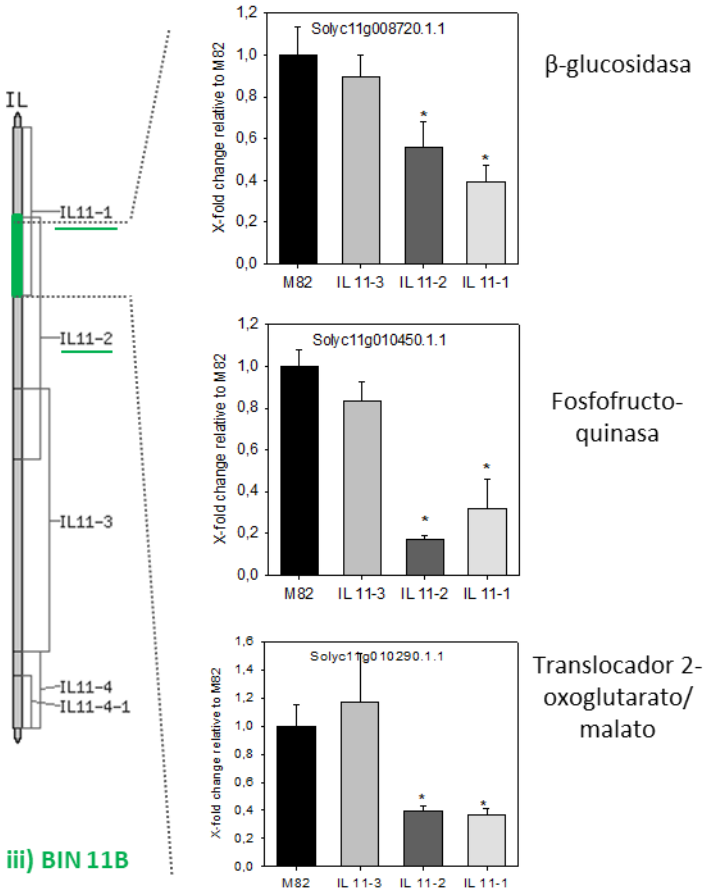
* La sobrerrepresentación se evaluó mediante una prueba de *Fisher* ($p < 0,05$) (ver Tabla Anexa 3, Anexos VIII.3).
Nota: El *BIN* 9C no presentó diferencias en ningún termino *MapMan* y éste no se muestra en la tabla.

III.2.2 | Identificación de genes candidatos

Las anotaciones funcionales del contenido génico de los 11 *BIN*s fueron revisadas y curadas manualmente. Es decir, como resultado de una búsqueda por similitud de secuencia utilizando el método de *BLASTP* reciproco (ver M&M, sección V.3.4 para más detalle), las proteínas predichas conllevan una descripción funcional curada y detallada (asignada por el consorcio *ITAG* 2.30). En este sentido, como primer criterio de selección se revisó la descripción funcional de cada uno de los *locus* predichos (ver Tabla 9, sección III.2.1). Si esta descripción indicaba una fuerte relación funcional con los mecanismos de fotosíntesis, metabolismo y crecimiento de la planta, el *locus* para la proteína anotada se seleccionó como gen candidato. De esta manera se obtuvo, analizando los 11 *BIN*s, un listado de 87 genes candidatos (debido a su tamaño limitante las Tablas correspondientes se adjuntan como material digital anexo, también disponibles como Tablas S1 y S4 en la publicación de **de Oliveira Silva et al., 2017**). La mayoría de estos candidatos codifican para enzimas (31) o genes implicados en los procesos de recepción/señalización (18). Sin embargo, varios de ellos también están involucrados en el metabolismo del ARN o asociados con el transporte celular. Con la excepción de los *BIN* 9D (*S. pennellii*) y 11B (ambos genotipos), más del 50 % de los genes candidatos seleccionados pertenecen a una categoría *MapMan* sobrerrepresentada, mientras que el único candidato seleccionado del *BIN* 11D no lo hace. Para tratar de inferir el rol que estos genes tienen sobre la regulación de los caracteres observados, mediante el uso de datos de *RNAseq*, obtenidos de los trabajos previos de **Chitwood et al., (2013)** y **Bolger et al., (2014a**), se estudiaron 3 *BIN*s (2K, 11H y 11B) y 6 genes seleccionados, para comprobar los niveles de expresión de los alelos silvestres de los mismos. En este sentido, en comparación con el cultivar M82, el *BIN* 11H y el *BIN* 2K mostraron una alta tasa de asimilación de CO_2, junto con reducidos niveles de expresión de los alelos salvajes de los genes que codifican las enzimas ATP sintasa y NADH deshidrogenasa. Otro de los ejemplos presentados muestra que el *BIN* 11B, asociado a altos niveles de almidón, correlaciona con niveles diferenciales de ARNm de los genes que codifican 3 enzimas clave; la β glucosidasa, la fosfofructoquinasa y un posible traslocador de 2-oxoglutarato/malato (ver Figura 26)

Figura 26: Datos de niveles de expresión (*RNAseq*) de 6 genes de interés en 3 regiones genómicas: i) *BIN* 11H (Alta tasa de asimilación de CO_2), ii) *BIN* 2K (Alta tasa de Asimilación de CO_2) y iii) *BIN* 11B (Altos niveles de Almidón). Figuras tomadas y adaptadas de (**Chitwood et al., 2013**; **de Oliveira Silva et al., 2018**).





III.2.3 | Resumen de los principales resultados

1. La evaluación de caracteres cuantitativos asociados a fotosíntesis, respiración y crecimiento en la población de *ILs* de tomate mostró alta variabilidad y se detectaron *QTL* para los mismos.

2. Las mediciones de 24 caracteres cuantitativos sobre la población de 71 *ILs* de Tomate (realizadas en 2 años independientes) permitieron identificar 11 *BIN*s con *QTL* relacionados a un aumento de la capacidad de fotosíntesis, respiración y crecimiento con respecto al cultivar elite M82.

3. Se identificaron 87 genes candidato y/o factores genéticos involucrados en la regulación de los *QTL* evaluados.

4. Estos resultados demuestran que los genomas parentales de las *ILs* (*S. pennellii* LA0716 y *S. Lycopersicum* M82) son una herramienta útil para realizar estudios de genómica comparativa con el fin de mejorar la producción de tomate.

III.3 | Discusión

En este Capítulo se analizaron dos experimentos independientes; en cada uno se ensayaron 71 líneas de introgresión de tomate y 1 cultivar control (M82). El primer experimento se sembró y muestreo durante el año 2012, y el segundo durante el 2013. Se midieron 24 parámetros cuantitativos de ambos grupos relacionados con los procesos de fotosíntesis, respiración y crecimiento de la planta de tomate y se los contrastó con los datos sobre el parental recurrente M82 (ver Figura 25). Se identificaron 87 genes candidatos en 11 regiones genómicas superpuestas entre las *ILs* (*BIN*s). Estas fueron identificadas como regiones introgresadas clave por contener las bases genéticas detrás de las variaciones medidas en los caracteres cuantitativos (ver Figura 25, sección III.2.1). Por ejemplo, en el presente estudio observamos un aumento en el nivel de asimilación fotosintética (*A*) de hasta un 19% (**de Oliveira Silva et al., 2018**). Cabe destacar, que esto se encuentra en el rango de los aumentos informados al utilizar enfoques transgénicos en tabaco (*Nicotiana tabacum*) sobre-expresando tanto la sedoheptulosa-1,7-bisfosfatasa, la fructosa-1,6-bisfosfato aldolasa, y el cotransportador de bicarbonato de sodio (*ictB*) (**Simkin et al., 2015**). El aumento de *A* también se reportó en una variedad de tomate modificada con disminuida expresión de la enzima malato deshidrogenasa mitocondrial (**Nunes-Nesi et al., 2005**). En este sentido, la evidencia más reciente indicaría que el malato desempeña un papel regulador clave en la función estomática en respuesta a cambios ambientales (**Nunes-Nesi et al., 2007**; **Araújo et al., 2011**; **Gago et al., 2016**; **Medeiros et al., 2016**). Este papel regulador explicaría en parte, los altos niveles de almidón encontrados en las *ILs* portadoras del *BIN* 11B (IL-11-1 e IL-11-2, ver Figura 26-iii, sección III.2.2). Teniendo esto en cuenta, se observa una disminución en la expresión del gen candidato anotado como translocador de 2-oxoglutarato/malato en el *BIN* 11B, el cual media el transporte de 2-oxoglutarato producido en el citosol o la mitocondria al cloroplasto en el intercambio de malato, proporcionando así el 2-oxoglutarato requerido para la asimilación de amoníaco por la glutamato sintasa plastídica (**Schneidereit et al., 2006**; **Eisenhut et al., 2015**). Por otra parte, también se ha demostrado que el translocador 2-oxoglutarato/malato desempeña un papel importante en la interfaz entre el metabolismo de carbono y nitrógeno en hojas de tabaco

(**Schneidereit et al., 2006**) y es importante para proteínas de almacenamiento y biosíntesis de aminoácidos en semillas de arveja (*Pisum sativum*) (**Riebeseel et al., 2009**). Estos hallazgos apuntan a un papel crucial del translocador de 2-oxoglutarato/malato en el metabolismo del carbono y, en particular, en la regulación del contenido de almidón.

Siguiendo con la caracterización de los genes candidatos, hoy en día, la mayor parte de la variación fotosintética se ha explicado por variaciones en la conductancia estomática (*gs*). Por ejemplo, estudios previos informaron que varios segmentos del genoma de *S. pennellii*, cuando se introgresan al fondo genético de *S. lycopersicum*, dan como resultado cambios en la composición de isótopos de carbono (δ13C) en relación con M82 (**Xu et al., 2008**). En el estudio de Xu et al., (2008) encontraron que la IL 5-4 mostraba la δ13C con una carga menos negativa y que correspondería a una mayor eficiencia del uso de agua. En contraste, encontraron que la IL 11-4 mostraba una δ13C con una mayor carga negativa que M82 y, curiosamente, esta línea junto con la línea superpuesta (IL 11-4-1), ambas conteniendo al *BIN* 11H, mostraron una *A* elevada en nuestro estudio (ver Figura 26-i, sección III.2.2). Por otro lado, uno de nuestros genes candidato, anotado como una subunidad de ATP sintasa, se identificó en éste mismo *BIN* (ver Figura 26-i, sección III.2.2). Sin ir más lejos, se sabe que este gen está altamente expresado en hojas de tomate y en etapas de desarrollo temprano de flores y tejidos de frutos (**Geisler et al., 2012**). Mas aún, la importancia de la ATP sintasa en las plantas ha quedado demostrada anteriormente. Por ejemplo, la inhibición a corto plazo de la ATP sintasa por oligomicina disminuye gradualmente la relación celular ATP/ADP y conduce a la reprogramación transcripcional, que consiste en una inducción temprana y masiva de vías respiratorias alternativas (**Geisler et al., 2012**). Además, estudios genéticos con mutantes de la subunidad δ de la ATP sintasa mitocondrial en *Arabidopsis* proporcionaron más apoyo a la importancia de la enzima en plantas, ya que conducen hacia ajustes a largo plazo en el metabolismo mitocondrial, crecimiento reducido y un desarrollo deficiente de gametofitos (**Geisler et al., 2012**).

En otro cromosoma, en la región genómica del *BIN* 2 K que se superpone con las *ILs* 2-5 y 2-6, donde se mapeó un *QTL* para *A* y niveles de almidón al final del día

(Almidón FD) (ver Figura 25, sección III.2.1), se detectaron reducciones significativas del ARNm de los alelos de *S. pennellii* para *loci* involucrados en el metabolismo energético de la célula ya que codifican para otra subunidad de una ATP sintasa y también para una NADH deshidrogenasa del Complejo I mitocondrial (ver Figura 26-ii, sección III.2.2). En plantas de tomate, se sabe que la NADH deshidrogenasa está menos expresada en las hojas, pero que tiene un mayor nivel de expresión en frutos maduros y en pimpollos florales. Además, se ha demostrado que el Complejo I, del cual esta enzima es parte, representa un importante sitio de entrada para los electrones de la cadena respiratoria, transfiriendo los electrones de NADH hacia a la ubiquinona (**Braun et al., 2014**). Por último, los resultados presentados en este Capítulo, junto con estudios previos sobre la misma población que permitieron la identificación de *QTL* para la expresión génica, niveles de metabolitos y la actividad enzimática, allanan el camino hacia la validación de genes candidatos que regulan los caracteres relacionados con la fotosíntesis en el tomate. En este sentido, este estudio aporta datos sobre los fundamentos mecanísticos de cómo las variantes de secuencia afectan la fotosíntesis y, en última instancia, al crecimiento de las plantas. Además, proporcionamos algunos indicios para la comprensión actual de los mecanismos fisiológicos y bioquímicos implicados en la regulación del metabolismo primario en las hojas de tomate.

Capítulo IV: Conclusiones Generales de la Tesis

A partir de secuenciar, ensamblar, ordenar y anotar el genoma mitocondrial del tomate cultivado y mediante análisis *in silico* e *in vivo*, los resultados obtenidos en esta Tesis permiten proponer que la estructura del genoma mitocondrial del tomate es dinámica y ésta organizada de forma multipartita en 5 o 6 subpartículas genómicas. Por otra parte, mediante un análisis comparativo de bloques de secuencias conservadas, se identificaron segmentos nucleares de origen mitocondrial. Este hallazgo fue validado por hibridación fluorescente *in situ* y se observó un número particularmente elevado de *numts* en el cromosoma 11. Esto implica que el dinamismo del genoma mitocondrial no se restringe a eventos de recombinación y re-arreglos intraorganelares, sino también a eventos de intercambio de material genético con otras organelas, como el núcleo. Estos resultados demuestran la existencia de transferencia intercompartimental en tomate y apoyan la idea de que estos eventos representan una fuerza motriz importante para la reubicación y remodelación de genes y genomas, contribuyendo a la amplia diversidad genética de la especie.

En la segunda parte de esta Tesis se presentó la secuencia completa del genoma de la especie silvestre *S. pennellii* (padre donante de las *ILs*) la cual fue secuenciada mediante tecnología de segunda generación (*Illumina®*) y ensamblada *de novo* mediante grafos de De Bruijn, en 4.579 *scaffolds,* de los cuales, el 97,1% fue anclado mediante una base de datos interna de 24.375 marcadores moleculares al mapa genético *Tomato EXPEN2000.* De esta manera, se posicionaron los *scaffolds* en 12 pseudomoléculas (cromosomas) conformando el genoma borrador de *S. pennellii*. El ordenamiento de estas secuencias y el anclaje de este genoma al mapa genético permitieron identificar genes candidatos para tolerancia a estrés (no abordados en esta Tesis) lo cual proporcionó pruebas de que los elementos transponibles tienen un papel importante en la evolución de estos caracteres. A su vez, la disponibilización de esta información representa una valiosa herramienta para mejorar el entendimiento de las bases genéticas que subyacen a la variabilidad

fenotípica de las Solanáceas, permitiendo de esta manera, eficientizar el proceso de mejoramiento genético.

Por último, en el capítulo final de la Tesis, el potencial como herramientas para la investigación y el mejoramiento de cultivos inherente a las secuencias genómicas presentadas, se demostró mediante la asociación de 24 caracteres cuantitativos, medidos sobre la población de *ILs*, a alelos exóticos presentes en las mismas. Este análisis resultó en la identificación de 87 genes candidatos distribuidos en 11 *BIN*s e involucrados en la determinación de *QTL* ligados a variaciones en fotosíntesis y caracteres relacionados con el crecimiento de las plantas. Esto demuestra que la disponibilidad de los genomas son una herramienta viable para realizar estudios de genómica comparativa con el fin de mejorar los caracteres productivos en especies de alto valor comercial, como el tomate.

Firma del Director de Tesis	
Firma del Director Conjunto	Firma del Alumno

Materiales & Métodos

Capítulo V: Materiales & Métodos (M&M)

V.1 | Capítulo I

V.1.1 | Extracción de ADN mitocondrial de *Solanum lycopersicum*

El aislamiento de ADN mitocondrial de alta pureza se realizó a partir de plántulas de tomate etioladas siguiendo los protocolos descritos previamente por **Scotti N. (2001)** y **Gianniny et al. (2004)**. Se realizaron una serie de centrifugaciones diferenciales a altas velocidades (> 20 mil RPM) para obtener un precipitado enriquecido en mitocondrias. Este fue purificado a través de un gradiente de *Percoll* (18%) y se midió en cada fracción obtenida la actividad enzimática de la citocromo-c-oxidasa (*CYT-C*), presente solo en mitocondrias, y de la fosfofructoquinasa (*PFP*) presente solo en el núcleo. Finalmente se seleccionaron las fracciones 27 a 31, que presentaron mayores porcentajes de actividad de *CYT-C* (Figura 27).

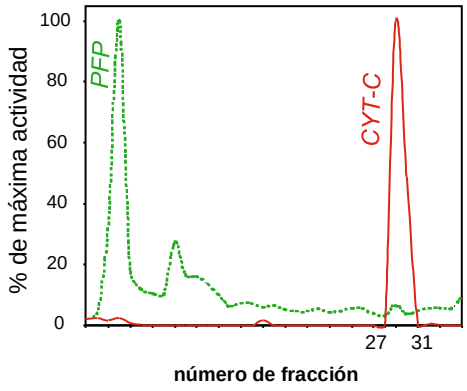


Figura 27: Actividades enzimáticas de citocromo-c-oxidasa (CYT-C) (marcador de mitocondrias) y de fosfofructoquinasa (PFP) (marcador nuclear) en 35 fracciones (de 1 ml cada una) de un gradiente de *Percoll* donde se cargaron 1 ml de un precipitado mitocondrial obtenidos a partir de 100 gramos de tejido fresco de hipocótilos de tomate (cv. Heinz). Las actividades enzimáticas se midieron como se describe en **Carnal y Black, (1983)** y **Melendez-Ferro et al., (2013)**. Los resultados se expresan en % de la fracción que presentó la mayor actividad (100 %).

Paralelamente, se realizó un control de calidad del ADN extraído de las fracciones 27 a 31 mediante PCR en diluciones seriadas de los ADN utilizando *primers* específicos para detectar la presencia de genes marcadores de cada fracción; Rubisco ($RBCL_{cloro}$) para cloroplastos, $Actina_{nucl}$ para núcleo y de la enzima mitocondrial ciclooxigenasa-2 ($COXII_{mito}$). Se pudo observar que efectivamente la muestra obtenida estaba enriquecida en ADN mitocondrial en 5 diluciones (Figura 28). Debido al alto nivel de $COXII_{mito}$ presente con respecto a los otros genes, se consideró que la muestra de ADN mitocondrial era de alta calidad, por lo que se prosiguió a la construcción de las genotecas bacterianas para su posterior secuenciación.

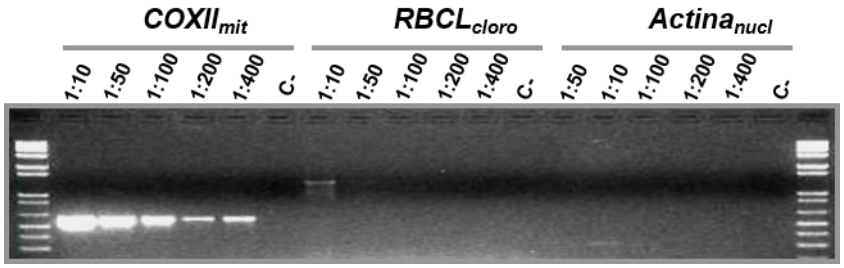


Figura 28: Diluciones seriadas (1:10, 1:50, 1:100,1:200, 1:400 y control sin diluir (C-)) de la muestra de ADN obtenida amplificadas por PCR para detectar la presencia de genes marcadores Rubisco ($RBCL_{cloro}$) para cloroplastos, $Actina_{nucl}$ para núcleo y la enzima mitocondrial ciclooxigenasa-2 ($COXII_{mito}$). Se observa que todas las diluciones están enriquecidas con el gen marcador de ADN mitocondrial y la presencia de ADN nuclear y de cloroplasto es casi imperceptible en comparación.

V.1.2 | Construcción de genotecas

Luego de la extracción, el ADN mitocondrial fue fragmentado de manera aleatoria mediante dos métodos distintos: i) por digestión con las enzimas de restricción Hind-III y Xho-I, y ii) por ruptura mecánica mediante sonicación. En ambos casos fueron purificados los fragmentos con peso molecular de entre 1 y 4 Kpb de las bandas del gel de agarosa correspondiente a cada método (ver Figura 29).

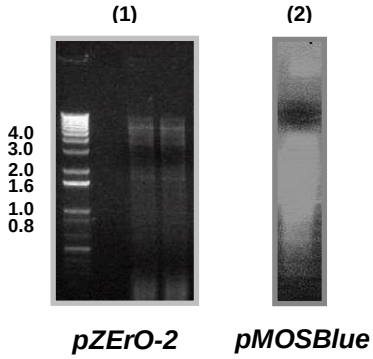


Figura 29: Imagen del gel de agarosa 1,5% obtenido luego de la fragmentación aleatoria 1) por enzimas de restricción y 2) por sonicación. Se purificaron los fragmentos de las bandas entre 1 y 4 Kpb para construir las genotecas en los plásmidos 1) pZErO-2 y 2) pMOSBlue.

Luego, los fragmentos (1 a 4 Kpb) obtenidos por el primer método fueron insertados en plásmidos pZErO-2 (*Invitrogen® Cloning Kit*) y en contraparte aquellos obtenidos por el segundo método fueron insertados en plásmidos pMOSBlue (*Blunt Ended Cloning Kit, Amersham Biosciences®*) (ver Figura Anexa 2, en Anexos VII.1). Posterioremente se transformaron bacterias competentes (*Escherichia coli,* cepas DH5 alfa y DH10 beta) con las ligaciones anteriormente mencionadas; los clones de las bibliotecas de pZErO-2 fueron crecidos y seleccionados por su resistencia a kanamicina (50 µgr/ml) en medio LB agar, mientras que los clones de las bibliotecas realizadas en pMOSBlue fueron seleccionados por su resistencia a tetraciclina (15 µgr/ml) y ampicilina (50 µgr/ml) en medio LB agar. Los clones con inserto fueron seleccionados a partir del *blue/white screening* adicionando al medio de cultivo el sustrato cromogénico X-Gal (200 µgr/ml) y el análogo de galactosa IPTG (Isopropyl β-D-1-thiogalactopyranoside) en 1 mM de concentración final.

Por último, los plásmidos de los clones positivos se purificaron mediante el método clásico de *miniprep* y se secuenciaron por el método de Sanger en el servicio de secuenciación de la Unidad Genómica del Instituto de Biotecnología del INTA de Castelar (http://inta.gob.ar/servicios/secuenciacion-de-acidos-nucleicos-adn-por-electroforesis-capilar). Cabe mencionar que los clones de la genoteca están

disponibles para la comunidad científica y pueden ser solicitados sin cargo mediante un formulario disponible en el sitio de internet del proyecto mitocondrial (http://www.mitochondrialgenome.org).

V.1.3 | Asignación de bases

Durante el preprocesamiento de datos (previo al ensamblado) el objetivo fue asignar las bases correspondientes (A, T, C, G o N) a los distintos picos de los electroferogramas contenidos en archivos con formato *ab1* (ver Figura 30). Con este fin, para asignar cada una de las bases a las lecturas se utilizó el algoritmo KB-*basecaller* del paquete informático provisto y licenciado por el fabricante de los equipos de secuenciación (*ABI DNA Sequencing Analysis Software v5.0*®).

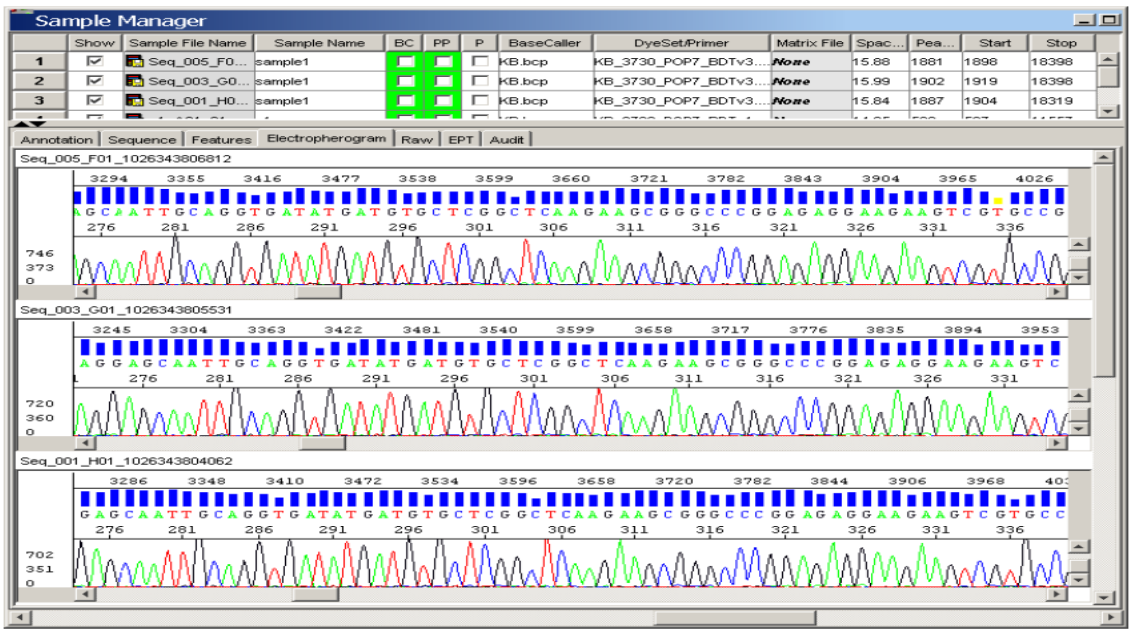


Figura 30: Archivos ab1 cargados en el administrador de muestras *ABI DNA Sequencing Analysis Software v5.0*® donde se realizó la asignación de bases (*basecalling*) de la genoteca de ADNmt.

En todos los casos se utilizó el archivo de parámetros de movilidad más reciente (*DyeSets/Primers POP7_BDTv3.mob*) y se asignó un nucleótido genérico (N) si la

calidad del pico del cromatograma era inferior a un valor de PHRED 10. En total se obtuvieron 7.051 lecturas con un largo promedio de 1.000 pb (ver Tabla 10).

Tabla 10: Resultados de la secuenciación de la genoteca de ADNmt de *S. lycopersicum* (cv Heinz).

Fuente	**Número de lecturas**
ABI 3130	6.717
ABI 3730	79
ABI 3500	228
Controles internos*	27
Lecturas descartadas**	273
Total	7.051

* corresponden a corridas de secuenciación para probar los reactivos y el equipamiento
** corresponden a lecturas que fueron descartadas por la baja calidad de los picos a lo largo de todo el cromatograma.

V.1.4 | Recorte de puntas y vectores

Luego de la asignación de bases, se procesaron todos los archivos "ab1" para recortar aquellas bases de baja calidad de los extremos 5' y 3' de cada una de las lecturas evaluadas mediante el programa *CodonCode Aligner 3.7.1* (**Richterich, 1998**). Se utilizaron los siguientes parámetros: i) recortar las bases del inicio y del final de las lecturas hasta que haya solo 6 nucleótidos con un valor de calidad PHRED ≤ 20 dentro de una ventana de 25 bases, y ii) después del recorte de puntas descartar aquellas lecturas con un tamaño < 100 nt. En paralelo, el *CodonCode Aligner* (**Richterich, 1998**) se configuró para ejecutar la tarea de recorte-de-vectores de clonado (del inglés, *vector-trimming*), donde se establecen parámetros de búsqueda para identificar si durante la secuenciación de los fragmentos de la genoteca pudo haber algún remanente del ADN colindante perteneciente a los plásmidos utilizados como vectores (*pMOSBlue y pZErO-2*), en este caso se utilizaron los siguientes parámetros para este algoritmo: i) Porcentaje de identidad (% Id) ≥ 90, ii) largo de solapamiento mínimo ≥ 30 nt, y iii) descartar lecturas con una longitud < 100 nt.

En resumen, luego del preprocesamiento de lecturas por *end-clipping y vector-trimming,* de los 7.051 cromatogramas ab1 se seleccionaron un total de 4.875 secuencias consideradas de alta calidad. Por lo que un 32% de las lecturas fue descartado y solo un 68% de los datos originales se utilizó en las siguientes etapas.

V.1.5 | Filtrado de contaminantes

El objetivo de este paso fue remover todas aquellas secuencias que no sean identificables como secuencias mitocondriales del clado *embryophyta*. Para concretar este objetivo se construyó una base de datos conteniendo todos los genomas mitocondriales de *Plantae* que había hasta el momento disponibles públicamente; lo que resulto en una base de datos personalizada de 48 genomas mitocondriales (plant_MT_dB) disponible en: www.mitochondrialgenome.org.

Acto seguido, se realizó una búsqueda por similitud de secuencias, usando la herramienta básica de alineamiento local *BLAST+* (**Camacho et al., 2009**); como semilla de búsqueda (del inglés: *query*) se utilizaron las antes mencionadas 4.875 lecturas (previamente convertidas de formato *ab1* a formato *fasta* y *qual*, mediante el *CodonCode Aligner*) y como sujeto (del inglés: *subject*) se proporcionó la base de datos plant_MT_dB. Los siguientes parámetros de corte se utilizaron para la corrida de *BLASTN* i) e-valor = 1E-10; ii) % id ≥ 80 y iii) % de cobertura ≥ 12. Como resultado, un 30% (1466/4875) de las lecturas no tuvieron similitud con la base de datos de plant_MT_dB y fueron descartadas. Por lo cual, luego del preprocesamiento, se obtuvieron 3.388 lecturas de alta calidad para ser utilizadas en los pasos siguientes.

V.1.6 | Ordenamiento de lecturas pareadas

Los fragmentos de la genoteca mitocondrial tienen entre 1 y 4 Kpb de longitud, pero en promedio el largo total de cada lectura Sanger fue de 1.024 nt. En este sentido, también hay que mencionar que los insertos mitocondriales fueron secuenciados por ambas puntas (sentido y antisentido, del inglés: *Fw: Forward and Rv: Reverse*), obteniéndose lo que se conoce como lecturas pareadas a una distancia conocida (del inglés: *paired-end reads*). Luego del ensamblado, las lecturas

pareadas, se utilizan para ordenar, orientar y expandir *contigs* en *scaffolds* utilizando la información de las lecturas *Fw* y *Rv* que sabemos *a priori* se encuentran a una distancie de entre 1 y 4 Kpb.

Luego del preprocesamiento, las lecturas que pasaron los parámetros de corte fueron convertidas del formato de archivo *fasta* y *qual* al formato *fastq* mediante un script en *Perl* (qual_to_fastq.pl, https://github.com/gjuggler/greg-ensembl), obteniéndose el archivo SolycMT.fastq. Durante el preprocesamiento se eliminaron algunas lecturas de baja calidad por lo que algunas lecturas perdieron su par, para separar aquellas lecturas *Fw* que tuvieran su par *Rv* (*paired-reads*) de aquellas que hayan quedado desapareadas (del inglés, *unpaired*) se desarrolló una herramienta personalizada (SangerPairedOmatic.jar, **Bolger et al., 2014b**) que tomó como entrada el archivo *fastq* (SolycMT.fastq de 3.388 secuencias) y devolvió las lecturas organizadas y separadas en tres archivos: i) SolycMT.F.fastq de 891 secuencias pareadas *Fw*; ii) SolycMT.R.fastq de 891 secuencias pareadas *Rv* y iii) SolycMT.U.fastq de 1606 secuencias que quedaron desapareadas luego del recorte de puntas y filtrado de contaminantes.

V.1.7 | *Contigs* ensamblados con lecturas *Roche/454®*

Adicionalmente a las 3.388 secuencias de las genotecas mitocondriales secuenciadas por Sanger, el Consorcio Internacional del Genoma del Tomate (https://solgenomics.net/organism/Solanum_lycopersicum/genome) nos facilitó un conjunto de 616 *contigs* que resultaron de sondear el borrador del genoma nuclear, ensamblado a partir de lecturas *NGS Roche/454®*. Mediante una búsqueda por similitud de secuencias con plastos y organelas, solo 577 *contigs* pasaron el algoritmo de filtrado a través la base de datos de *plant_MT_dB*. El largo promedio de estos *contigs* fue de 1.150 nt y sumaron un total de 665.439 nucleótidos al set de datos para ensamblar el genoma mitocondrial del cultivar Heinz.

IV.1.8 | Ensamblado de la mitocondria

Para el ensamblado con *CAP3* se utilizaron los siguientes datos de entrada:

i) *sl_MT_assembly_set.fasta* | 3.388 lecturas, aportando un total de 3.851.869 nt

ii) *sl_MT_assembly_set.qual* | valores de calidad PHRED de cada base

iii) s*l_MT_assembly_set.con.* | especificación de lecturas pareadas (*constraints*)

CAP3 se ejecutó en un servidor *Dell PowerEdge R710,* adquirido por nuestro grupo (ver Anexos VIII.1.1), y se utilizaron 24 CPUs (intel xeon X5650 @ 2.67Ghz) y 64GB de memoria RAM, lo que dio como resultado 173 *contigs* (ver Tabla 3, sección I.2.3).

Para el armado de los *scaffolds* se buscó información que permita ligar dos o más *contigs* mediante el aprovechamiento de lecturas pareadas (*Fw* y *Rv*) pertenecientes al mismo inserto mitocondrial, ya que se conocía de antemano, que las lecturas *PE* estaban a una distancia estimada de 1.000 a 4.000 nt. Esto se realizó mediante alineamiento de las 3.388 lecturas Sanger con el programa *Bowtie* (**Langmead et al., 2009**) seguido de una búsqueda con la herramienta *Tablet* (**Milne et al., 2013**) para visualización de genomas y alineamientos (ver Figura 31).

Figura 31: Se observa en la interfaz gráfica del *Tablet* el Contig28 y las lecturas que lo componen en busca de evidencia de lecturas pareadas que estén en *contigs* separados.

Como resultado final se obtuvo un ensamblado de 578.247 nt distribuidos en 162 secuencias contiguas (6 *scaffolds* y 156 *contigs;* ver Figura 15, sección I.2.3). Los cuales en conjunto representan el primer borrador del genoma mitocondrial del tomate. La cobertura del ensamblado fue calculada en 7X (3.851.869 / 578.247). El mismo está disponible públicamente en el sitio de internet www.mitochondrialgenome.org y al mismo tiempo todas las secuencias fueron depositadas, mediante la herramienta *tbl2asn* (www.ncbi.nlm.nih.gov/genbank/tbl2asn2/), en las bases de datos públicas e internacionales DDBJ, EMBL y GenBank bajo el BioProject ID: 67471 y el número de entrada AFYB00000000.

V.1.9 | Segmentos nucleares de origen mitocondrial (*numts*)

Con el fin de identificar posibles inserciones nucleares de origen mitocondrial (*numts*, por sus siglas en inglés: *nuclear mitochondrial DNA segments*) se realizó una búsqueda de bloques de secuencias conservadas y libres internamente de reordenamientos genómicos (del inglés, *LCBs: local-colineal-blocks*) entre el ensamblado mitocondrial y el nuclear. Esto se realizó mediante una búsqueda *in-silico* con la herramienta *Mauve progressive Aligner* (**Darling et al., 2004**). De esta manera mediante el análisis de 118 *LCBs* (ver Figura 32) entre el genoma mitocondrial y el nuclear se pudieron identificar 92 *numts* (ver Figura 16, sección I.2.5) (**The Tomato Genome Consortium, 2012**).

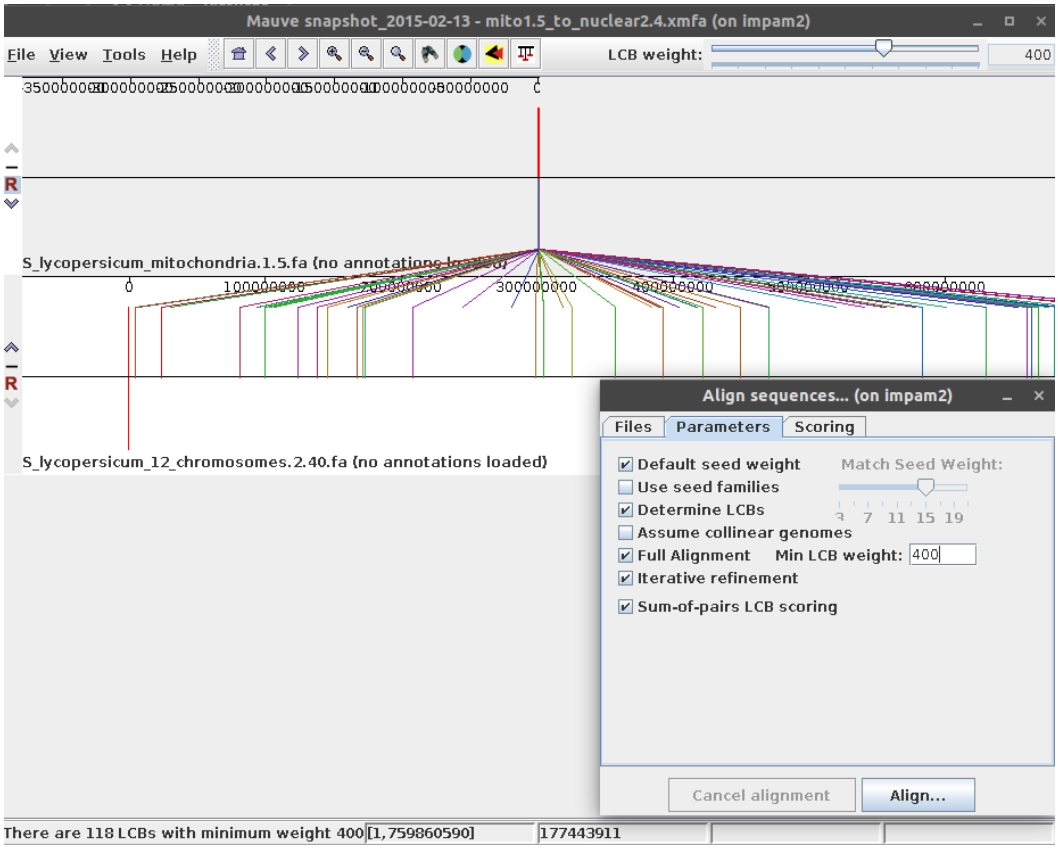


Figura 32: Alineamiento entre el genoma mitocondrial y nuclear de *S. lycopersicum* cv Heinz mediante el *Mauve progressive Aligner* con el parámetro *min LCB weight* = 400 (abajo a la derecha) mostrando los bloques colineales locales (*LCBs*) entre ambos genomas (barras de colores).

Hibridación *in situ* fluorescente

Los resultados obtenidos *in silico* fueron validados mediante hibridación *in situ* fluorescente. Se realizaron experimentos de hibridación *in-situ* fluorescente (*FISH*) con una mezcla equimolar de 40 clones mitocondriales con tamaños de inserción ≥ 2,5 Kpb etiquetados con dioxigenina, también se híbrido el clon del *BAC* Le_HBa323E19 de la genoteca nuclear de *S. lycopersicum* (cv Heinz) etiquetado con biotina para identificar el brazo largo del cromosoma 11 y el trasfondo de los cromosomas nucleares fueron marcados con DAPI (4', 6-diamino-2-fenilindol). Los resultados obtenidos de la localización por hibridación *in-situ* fluorescente (*FISH*) de ADN mitocondrial secundo el análisis realizado *in-silico* con el *Mauve Progressive Aligner* (Figura 17, sección I.2.5.)

V.2 | Capítulo Ⅱ

V.2.1 | El genoma de *Solanum pennellii* (LA0716)

Material vegetal

Las semillas de *Solanum pennellii* (LA0716) se obtuvieron del Centro de Recursos Genéticos del Tomate (TGRC), UC Davis, CA, EE. UU. Se sembraron directamente en suelo y las plantas resultantes se cultivaron en invernadero en macetas de 20 litros bajo condiciones controladas en invernáculo con un fotoperíodo de 16/8 h, temperatura: 25ºC ± 5°C, humedad: 60% e irradiancia: 300 ± 100 µmol m-2 s-1. Para las bibliotecas de lecturas *PE*, se recogieron hojas jóvenes y se congelaron directamente en nitrógeno líquido para realizar la extracción de ADN nuclear. Para las bibliotecas de lecturas *MP*, las hojas jóvenes se colocaron en *buffer* TE enfriado con hielo (10 mM Tris-HCl, 1 mM de ácido etilen-diamino-tetra-acético (EDTA)) e inmediatamente después se procesaron para la extracción de los núcleos y posterior extracción de ADN.

V.2.2 | Extracción de ADN nuclear de *Solanum pennellii*

Extracción de ADN nuclear para bibliotecas *NGS PE*

La extracción de ADN nuclear para las bibliotecas de fragmentos de ADN de bajo peso molecular (150, 300, y 600 pb) se realizó utilizando 0,5 g de hojas jóvenes que se trituraron a un polvo fino en nitrógeno líquido usando un mortero y un mazo y se transfirieron a un tubo de centrifugación de polietileno de 15 ml mantenido en hielo y conteniendo 10 ml de *buffer* de extracción de núcleos (10 mM de Tris-HCl (pH 9,5), 10 mM de EDTA (pH 8,0), 100 mM de KCl, 500 mM de Sacarosa, 4 mM de Espermidina, 1 mM de Espermina y 0,1 % de 2-Mercaptoetanol). El homogenato se mezcló minuciosamente con una punta de pipeta ancha (para minimizar el daño al ADN) y se filtró a través de dos capas de Miracloth (*CalBiochem*®) hacia un tubo de centrífuga de polietileno de 50 ml mantenido en hielo. Se añadieron 2 ml de *buffer* de lisis (10% de Triton X-100 en *buffer* de extracción de núcleos) a la suspensión filtrada y se mezcló suavemente durante 2 minutos sobre hielo. Los núcleos se sedimentaron por centrifugación a 2000 x g durante 10 minutos. Se añadieron 500 µl de *buffer* de extracción CTAB (Tris 100 mM (pH 7,5), NaCl 0,7 M, EDTA 10 mM, 2-Mercaptoetanol al 1% y Bromuro de cetiltrimetilamonio al 2%) al *pellet* de los núcleos, se mezcló e incubó durante 30 min a 60º C. La mezcla se extrajo con 350 µL de cloroformo/alcohol isoamílico (24:1). Finalmente se precipitó el ADN por Isopropanol y se lavó varias veces con EtOH al 75%.

Extracción de ADN nuclear para bibliotecas *NGS MP*

La extracción de ADN nuclear para las bibliotecas de fragmentos de ADN de alto peso molecular (3, 5, 8, 10, 20, 30 y 40 Kpb), se realizó utilizando una versión modificada del protocolo de **Peterson et al., (1997**). Entre 20 y 100 g de hojas jóvenes y capullos de flores se recogieron directamente en *buffer* TE (pH 7,0) enfriado sobre hielo. El material vegetal se retiró del *buffer* TE y se colocó en una licuadora (*Waring*®) junto con *buffer* de extracción previamente enfriado y recién preparado (1 mol de 2-metil-2,4-pentanodiol, 10 mM de PIPES, 10 mM de MgCl2, 4% (p/v) de PVP-10, 10 mM de Metabisulfito sódico, 25 mM de 2-Mercaptoetanol,

0,5% (p/v) de Dietilditiocarbamato sódico, 200 mM de L-lisina y 6 mM de Ácido etilenglicol tetraacético (EGTA), pH 7.0) y la mezcla se homogeneizó durante 30 segundos a toda velocidad. El material vegetal homogeneizado se exprimió a través de 4 capas de *Miracloth* (*CalBiochem*®) y luego se filtró por gravedad a través de 8 capas de *Miracloth* (*CalBiochem*®). Se añadió Tritón X-100 al 10% (v/v) al homogenato filtrado a una concentración final de 0,5% (v/v) y la mezcla se incubó en un baño de hielo con un suave balanceo durante 30 minutos. Luego, el homogenato se centrifugó a 800 x g durante 20 minutos a 4° C para sedimentar los núcleos. El sobrenadante se desechó y el sedimento se resuspendió suavemente usando un pincel pequeño empapado en *buffer* de extracción preenfriado y recién preparado. Puesto que el *pellet* contiene grandes cantidades de desechos celulares, se aplicó el siguiente proceso de lavado: El volumen se llevó hasta 10 ml con el mismo *buffer* y el sedimento resuspendido se mezcló suavemente. La mezcla se centrifugó a 600 x g durante 20 minutos a 4º C y el sobrenadante se desechó. Después de tres ciclos de lavado, el *pellet,* de un color grisáceo (no se observaron rastros de clorofila), se resuspendió usando un pequeño pincel empapado en *buffer* de extracción y se lo llevó hasta 5 ml usando el mismo *buffer*. Se añadió SDS al 20% (p/v) hasta una concentración final del 2% (p/v) para lisar los núcleos. El contenido se mezcló suavemente invirtiendo el tubo. Luego, el tubo se calentó en un baño térmico a 60° C durante 10 minutos y se dejó enfriar a temperatura ambiente. Se añadió perclorato sódico 5 M a una concentración final de 1 M y el tubo se invirtió suavemente. La mezcla se centrifugó a 400 x g durante 20 minutos a temperatura ambiente. El sobrenadante se retiró suavemente a un tubo nuevo usando puntas de pipeta anchas (estas puntas se usaron para el resto del procedimiento de extracción para minimizar el daño al ADN). Acto seguido, la extracción de ADN se realizó añadiendo un volumen igual de fenol/cloroformo/alcohol isoamílico (25:24:1) al tubo. Luego, el tubo se colocó horizontalmente sobre un balancín y se agitó suavemente durante 30 minutos. La mezcla se centrifugó a 3000 x g durante 10 minutos a temperatura ambiente y la fase superior se transfirió a un tubo nuevo. Se llevó a cabo una segunda extracción de fenol/cloroformo/alcohol isoamílico seguido por dos extracciones más usando solo cloroformo. La fase superior se transfirió a un nuevo tubo y se añadió acetato sódico 3 M a una concentración final

de 0,3 M. El tubo se invirtió suavemente varias veces para mezclar. El ADN se precipitó añadiendo 2 volúmenes de etanol enfriado sobre hielo y el tubo se invirtió varias veces. El ADN precipitado se transfirió a un tubo limpio usando una varilla de vidrio y se dejó evaporar el etanol. Finalmente, el ADN se disolvió en *buffer* TE y la concentración de ADN se midió usando *Nanodrop* (*Thermo Fischer*®) y *Qubit* 2.0 (*Invitrogen*®).

V.2.3 | Estrategias de secuenciación

Genotecas secuenciadas por *Illumina*® *(NGS)* y el método de Sanger.

Lecturas-pareadas (*PE*) por secuenciación *Illumina*®

Construcción de las bibliotecas

Para las genotecas *PE* de fragmentos pequeños, el ADN genómico se cortó en fragmentos de 150-600 pb utilizando el instrumento *Covaris* S2 (*Covaris*®). El ADN se volvió a suspender en 1x *buffer* TE (*Applied Biosystems*® P/N 4389764). La reparación final de fragmentos cizallados, la adición de un residuo A al extremo 3' de los fragmentos y la ligación de adaptadores se realizaron según las instrucciones de *Illumina*®. El ADN modificado con adaptadores se resolvió en un gel de agarosa al 2% (incluyendo 400 ng/ml de bromuro de etidio) en *buffer* TAE durante 90 minutos a 120 voltios. Se cortaron las bandas conteniendo fragmentos de 150, 300 y 600 pb bajo iluminación de un equipo *Dark Reader* (*Clare Chemical Research*®). Esto permitió construir 3 bibliotecas de lecturas *PE* basándose en la distancia que había entre las puntas terminales de los fragmentos (ver Tabla 11). Después se aisló el ADN con un kit de extracción en gel (*Qiagen*®) y se amplificó por PCR durante 12 ciclos con los *primers* de PCR suministrados por *Illumina* y se cuantificó con un fluorómetro *Qubit* 2.0 (*Invitrogen*®). El ADN se diluyó y almacenó a -20ºC como *stock* en *buffer* EB (10 mM de Tris y 0.1 mM de EDTA; pH 8.0; *Qiagen*®) suplementado con *Tween*-20 al 0,1%. La validación de las bibliotecas se realizó utilizando un equipo *Agilent*® *Bioanalyzer* procesando 1 µl de muestra en un chip DNA 1000 (*Agilent Technologies*®).

Secuenciación de las bibliotecas

La secuenciación de todas las bibliotecas *PE* se realizó en un equipo *Illumina® Genome Analyzer GA IIx* en el centro para secuenciación del "*Max Planck Institute for Developmental Biology Genome Center*" en Tübingen, Alemania. La generación del clúster se realizó en una *Cluster Station* (cat no. SY-301-2001; *Illumina®*), de acuerdo con las instrucciones del fabricante. La secuenciación de ADN se realizó en un equipo *Genome Analyzer GA IIx* (*Illumina®*) utilizando el *TruSeq SBS Sequencing Kit v5* de acuerdo con las instrucciones del fabricante; Los programas de control de secuenciación fueron el *SCS* versión 2.8 y el *RTA1.8.7*. Cabe destacar que el largo de las lecturas obtenidas con el equipo *Illumina® GA IIx* alcanzo en promedio un largo de lectura de 100 nucleótidos para ambas puntas (ver Tabla 11).

Tabla 11: Bibliotecas de lecturas *Illumina® PE* obtenidas a partir de la fragmentación sónica aleatoria del ADN genómico de *S. pennellii.*

Distancia entre lecturas *PE* (pb)	**Número de Lecturas (M)**
150	773
300	522
600	859
Total~	2 Bn

nt = nucleótidos
M = millones
Bn = billones

Lecturas-pareadas distantes (*MP*) por secuenciación hibrida *Illumina® - Roche/454®*

Construcción de las bibliotecas

Para la construcción de las genotecas con insertos de mayor tamaño (*MP)*, se adoptó un enfoque híbrido *Roche/454® - Illumina®*. El ADN se fragmentó en tamaños apropiados (3, 5, 8, 10, 20, 30 y 40 Kpb) utilizando el dispositivo de corte de ADN *HydroShear PlusTM* (*Digilab®*). Teniendo esto en cuenta, las bibliotecas de 3 y 5 Kpb se prepararon utilizando el kit de *Illumina® Mate-Pair* (número de

catálogo: 1004876, *Illumina*®) siguiendo las recomendaciones de la guía de preparación (número de catálogo: *PE*-930-1003 REV. A, noviembre de 2009). En cambio, para las bibliotecas con distancias más largas (8, 10, 20, 30 y 40 Kpb), se adoptó un enfoque híbrido *Illumina*® - *Roche/454*®, ya que al momento de la preparación de estas librerías el método de *Roche/454*® *Mate-Pair* lograba fragmentos circularizados de ADN más largos que con el Kit *Illumina*® *Mate-Pair*. Por lo que el ADN se sonicó en los tamaños de fragmento apropiados (8, 10, 20, 30 y 40 Kpb) utilizando el dispositivo de corte de ADN *HydroShear Plus* (*Digilab*®) y los fragmentos reparados se seleccionaron en un gel de agarosa al 0,5% que corrió durante 17 horas a 35 voltios durante la noche. Los extremos de estos fragmentos fueron circularizados utilizando el método *Cre-IoxP* descrito por *Roche/454*® (*GS FLX Titanium Paired End Library Preparation Method Manual,* oct. 2009) y luego las muestras se procesaron según el protocolo de *Illumina*® para librerías de lecturas-pareadas (*Illumina*® *PE sample preparation kit*, número de catálogo: *PE*-102-1001). Todos los procedimientos de lavado se realizaron tres veces con 500 µl de TE y con Tris-HCL 1x (*Qiagen*®). Los fragmentos se amplificaron durante 20 ciclos y se purificaron mediante *Agencourt AMPure beads XP* (*Beckman Coulter*®).

Secuenciación de las bibliotecas

Excepto en el caso de la última genoteca *MP* de 40 Kpb, todas las otras se secuenciaron como sigue. La generación del clúster se realizó en una *Cluster Station* (cat no. SY-301-2001; *Illumina*®), de acuerdo con las instrucciones del fabricante. La secuenciación de ADN se realizó en un equipo *Genome Analyzer GA IIx* (*Illumina*®) utilizando el *TruSeq SBS Sequencing Kit v5* de acuerdo con las instrucciones del fabricante; Los programas de control de secuenciación fueron el *SCS* versión 2.8 y el *RTA1.8.7*. Para la biblioteca final (40 Kpb), la secuenciación del ADN se realizó en un dispositivo *Illumina*® *HiSeq2000*, equipado con los programas *HCS* versión 1.5.15 y *RTA* versión 1.13. La generación del clúster se realizó en un *cBot* (receta: *PE_Amp_Lin_Block_Hyb_v8.0*, *Illumina*®) utilizando una célula de flujo v3 y reactivos de TruSeq *PE Cluster Kit v3* (*Illumina*®) de acuerdo con las instrucciones del fabricante. La secuenciación se realizó en modo pareado con una longitud de

lectura de entre 80 y 150 pb. Lo que resulto en la obtención de 7 bibliotecas del tipo *MP* (ver Tabla 12).

Tabla 12: Librerías de lecturas *MP*, obtenidas por el método híbrido *Illumina*® / Roche 454® secuenciadas con un instrumental *Illumina*® *"Genome Analyzer GA IIx"*.

Distancia entre lecturas *MP* (pb)	**Número de Lecturas (M)**
3.000	68
5.000	311
8.000	144
10.000	154
20.000	156
30.000	528
40.000	301
Total~	1.4 Bn

nt = nucleótidos
M = millones
Bn = billones

Lecturas-pareadas de *BACs* por secuenciación *Sanger*

Construcción de las bibliotecas

A partir de una biblioteca genómica construida con el vector de clonado pBeloBAC11 (52.992 clones) de la especie silvestre de tomate *S. pennellii* (LA0716) con un tamaño de inserto promedio de 120 Kpb (rango 40-320 Kpb) (**Kamenetzky et al., 2010**) se obtuvieron secuencias de los extremos de cada inserto (puntas) de un total de 11.798 clones.

Secuenciación de las bibliotecas

Para la secuenciación de puntas de BACs, se utilizaron 0,2-0,5 µg de ADN purificado y se realizaron reacciones de secuenciación con *ABI® BigDye Terminator* v3.1. Para la secuenciación de BACs por el método de Sanger se utilizó un secuenciador de ADN *ABI®* 3730 xl y el resultado fueron 23.494 secuencias (11.798 F y 11.696 R). Las secuencias se preprocesaron removiendo nucleótidos de baja calidad (PHRED < 20), filtrando fragmentos contaminantes del vector de clonado

y sólo se consideraron aquellas secuencias finales de BAC con largos mayores a 100 pb en ambos extremos. Por último, las lecturas se separaron en F, R y U (ver Tabla 13).

Tabla 13: Lecturas-pareadas de BACs de la genoteca de clones pBeloBACII de *S. pennellii.*

Distancia promedio entre lecturas-pareadas Sanger (pb)	Número de Lecturas	Dirección de lectura
120.000	10.615	F
120.000	10.615	R
---	2.264	U
Total	23.494	F, R y U

nt = nucleótidos
F = extremo 5' (*Forward*)
R = extremo 3' (*Reverse*)
U = desapareada (*Unpaired*)

Con un largo de lectura promedio de 1.000 nt, las 23.494 secuencias Sanger suman más de 23 millones de nt, lo que resultó en un 2% de cobertura del genoma total (estimado previamente en 1.2 Bn de nt). Debido a la baja cobertura estas lecturas no fueron utilizadas para el ensamblado de *contigs* directamente, pero debido a su alta calidad se las utilizó para la etapa de construcción de *scaffolds* y su posterior anclado al mapa genético.

V.2.4 | Preprocesamiento de lecturas por calidad y contaminación

Las bibliotecas *PE* y *MP* fueron preprocesadas utilizando el programa *TrimmOmatic* versión 0.13 (**Bolger et al., 2014b**), para eliminar secuencias de adaptadores de *Illumina*® y bases de baja calidad (PHRED < 20). Se retiraron bases de baja calidad de ambos extremos de las lecturas, se usó el recortador de ventana deslizante para eliminar la secuencia de baja calidad desde el extremo 3' de las lecturas, usando una puntuación de calidad media requerida de 20 PHRED sobre una ventana de 4 bases. Las lecturas de menos de 36 pb fueron eliminadas. El preprocesamiento de las bibliotecas *MP* se personalizó en función del protocolo de preparación de la genoteca. Para la primera genoteca de 5Kpb, que se creó con el protocolo de *MP* de *Illumina*®, el filtrado fue igual al de las *PE*, pero con una

ventana deslizante menos estricta (PHRED ≥ 10 en una ventana de 4 nucleótidos) y una longitud mínima aceptada de 24 bases. Para las bibliotecas híbridas de *Roche/454® - Illumina®,* el proceso de filtrado se modificó agregando la secuencia enlazadora *LoxP* (TCGTATAACTTCGTATAATGTATGCTATACGAAGTTATTACG) más su secuencia reversa complementaria al archivo de adaptadores de TruSeq2. La presencia de esta secuencia dentro de las lecturas es un artefacto de la preparación de la biblioteca que de no recortarla evitaría la alineación de muchas lecturas válidas. Por último la contaminación mitocondrial y cloroplástica se estimó mediante la alineación de las bibliotecas secuenciadas usando el programa *Bowtie* versión 0.12.7 (**Langmead et al., 2009**) con el genoma del cloroplasto de *S. lycopersicum* (*genbank* GI: 113531108) y con las secuencias mitocondriales (*genbank* GI: 378405840 - GI: 378406034) utilizadas como cebos.

Por otra parte, las secuencias de las genotecas de *BACs* secuenciadas por el método de Sanger fueron convertidas del formato tradicional (FASTA y QUAL, ver glosario) al formato de segunda generación (FASTQ) y luego se procesaron con el programa T*rimmOmatic* con los mismos parámetros utilizados para las genotecas *PE* y *MP* con excepción de que la búsqueda de adaptadores y vectores de clonado se realizó con el programa *CodonCode Aligner* (versión 1.6.3; *CodonCode Corporation®*). El resumen de los resultados de las bibliotecas *PE*, *MP* y Sanger *PE* se presenta en las Tablas 14 y 15. En este sentido, se puede observar que las bibliotecas *Illumina® PE* de 150, 300 y 600 pb suman un total de 2,15 billones de lecturas y 229,36 billones de pares de bases, que luego del preprocesamiento (*TrimmOmatic* y eliminación de contaminantes) resultaron en un total de 1,7 billones de lecturas y 165 billones de pb, por lo que un 19% de lecturas de la genoteca *Illumina® PE* fue descartado en esta etapa. Por otra parte, las bibliotecas de *Illumina® MP* (3, 5, 8, 10, 20, 30 y 40 Kpb) suman 1,66 billones de lecturas conteniendo unos 183,57 billones de nucleótidos, que luego del preprocesamiento se redujeron a 935 millones de lecturas con 93 billones de nucleótidos, por lo que se descartó un 43% de las bibliotecas *Illumina® MP* denotando una perdida significativa de información debido a la complejidad de la técnica utilizada para su preparación (Tabla 14).

Tabla 14: Lecturas *Illumina® PE* y *MP* de ADN de *S. pennellii* crudas y luego del preprocesamiento de calidad y contaminantes. Se muestra la sumatoria (Σ) del número de lecturas presentes en las genotecas secuenciadas y la cantidad de nucleótidos obtenidos.

	Illumina® PE	***Illumina® MP***
Σ lecturas	2,15 Bn	1,66 Bn
Σ nucleótidos	229,36 Bn	183,57 Bn
Σ lecturas preprocesadas	1,73 Bn	935 M
Σ nucleótidos preprocesados	165 Bn	93 Bn

M = millones
Bn = billones

Con respecto a las genotecas de BACs secuenciadas por el método de Sanger *PE* (52.992 clones) se obtuvieron 23.494 lecturas, cada una con un largo promedio cercano a 1.000 nt, lo que significó un total de 17.059.144 de pb que luego del preprocesamiento sumaron un total de 23.488 lecturas y 15.779.740 nucleótidos. Debido a la alta calidad de la secuenciación por el método de Sanger solo se descartó un 7,5 % de la genoteca Sanger *PE* (ver Tabla 15).

Tabla 15: Lecturas Sanger *PE* crudas y luego del preprocesamiento. Se muestran la sumatoria (Σ) del número de lecturas y de la cantidad de nucleótidos totales obtenidos.

	Sanger *PE*
Σ lecturas	23.494
Σ nucleótidos	17.059.144
Σ lecturas preprocesadas	23.488
Σ nucleótidos preprocesados	15.779.740

V.2.5 | Ensamblado de lecturas en *contigs*

Previo al ensamblado, los datos de lecturas *PE* fueron corregidos utilizando el programa *SOAP error corrector* versión 1.0, utilizando 8 microprocesadores, pero de lo contrario se usaron los parámetros predeterminados. El ensamblado y el andamiaje inicial se realizó con el programa *SOAPdenovo* versión 1.05 (**Luo et al., 2012**), utilizando las genotecas *PE* y las *MP* más cortas (3, 5 y 8 Kpb). El tamaño del K-mero a utilizar por el programa se estableció en 63 (-K 63), y se activó la resolución de repeticiones (del inglés, *read-repeat resolution*, -R), y el recorte de grafos de baja cobertura (del inglés, *graph trimming*, -D -d). Para mejorar el rendimiento, se utilizaron los 48 microprocesadores (-p 48) del GabiPD *cluster* (**Usadel et al., 2012**).

Las lecturas de *Illumina*® se ensamblaron en *contigs* mediante un algoritmo que recorre grafos de De Bruijn atravesando un camino euleriano implementado mediante la herramienta *SOAPdenovo* (**Luo et al., 2012**). En este sentido, las lecturas se alinean, se solapan y se extienden formando una secuencia consenso en fragmentos de mayor tamaño llamados *contigs*. De esta manera se obtuvieron un total de 4.315.954 *contigs* conteniendo 1.117.562.721 nucleótidos, un *contig* N50 de 2.176 pb y por sobre este quedaron 81.824 *contigs* (*contigs* ≥ N50) que contienen el 50% del ensamblado total, en esta etapa no se encuentran nucleótidos genéricos (Ns) en el ensamblado (ver Tabla 16).

Tabla 16: Resumen de los *contigs* ensamblados con el software *SOAPdenovo* a partir de los datos de secuenciación de ADN del tomate silvestre *S. pennellii*.

Etapa	N50	*contigs* > N50	*contigs* totales	nucleótidos totales	Ns
Contigs	2.176	81.824	4.315.954	1.117.562.721	0

V.2.6 | Andamiaje de *contigs:* obtención de *scaffolds*

Utilizando los 4.315.954 de *contigs* obtenidos con *SOAPdenovo* (**Luo et al., 2012**) y las genotecas de lecturas *MP* de distancia larga (10, 20, 30 y 40 Kpb) se realizó el andamiaje de *contigs* para armar los *scaffolds* utilizando el programa *SSpace* versión 1.0 (**Boetzer et al., 2011**) para obtener el borrador del genoma de *S. pennellii* v1.90 (**Bolger et al., 2014a**). Las métricas de las distintas etapas del ensamblado (N10 – N90) se detallan en la Tabla 17. Como resultado el *scaffold* más grande fue ligeramente mayor a 10 Mpb y el tamaño del *scaffold* medio ponderado (N50) fue de 1.7 Mpb y 156 fue la cantidad de *scaffolds* que conformaron el 50 % del tamaño total del genoma (n > N50). Por otra parte, El N90 fue de 437 Kpb, indicando que el 90 % del ensamblado está contenido en *scaffols* con un tamaño medio de 0.5 Mpb. La totalidad del conjunto presenta huecos (del inglés, *gaps*) que suman unas 67.2 Mpb representadas por "Ns" genéricas.

En total, durante el andamiaje se construyeron 407.506 *scaffolds* conteniendo 1.021.472.455 pb de las cuales 125.806.430 fueron Ns utilizadas para unir los *gaps* entre *contigs* (ver Tabla 7, sección II.2.1). El N50 luego del andamiaje ascendió a 1.603.317 pb y los *scaffolds* conteniendo el 50% del genoma fueron 177 (*scaffolds > N50*). En este sentido, cabe destacar, que para las siguientes etapas (p.ej: el anclado de los *scaffolds* al mapa genético), fueron seleccionados los 4.591 *scaffolds* que tenían un tamaño superior a 2.000 nt (*scaffolds > 2 Kpb*) ya que los *scaffolds* más pequeños fueron considerados de baja calidad o quimeras debido al bajo número de lecturas pareadas que soportaban el andamiaje de los *contigs* que los componían. Esta metodología también fue empleada en trabajos previos, como el genoma nuclear de *Solanum lycopersicum* v2.40 (**The Tomato Genome Consortium, 2012**), y por otra parte este pequeño tamaño de *scaffold* fue una limitante a la hora de realizar la anotación estructural de genes *in silico* mediante la herramienta *Augustus* (**Stanke et al., 2004**).

V.2.7 | Eliminación de contaminantes

Después de completar el ensamblado y el andamiaje, los *scaffolds* se alinearon con la base de datos de nucleótidos no redundantes (*NT*) del *NCBI* (descargada el 18 de mayo de 2013) utilizando el programa *BLASTN* (*Blast+* versión 2.2.28) con los parámetros por defecto, con excepción de un valor mínimo de e-valor de 1E-10. Los datos de taxonomía correspondientes a la secuencia de referencia más cercana se utilizaron para determinar si era probable que algún *scaffold* fuera de origen no vegetal, y por lo tanto debía de ser removido. Esto dio como resultado la eliminación de 12 pequeños *scaffolds* que comprendían 29.742 bases, dejando como resultado 4.579 *scaffolds* en la secuencia borrador del genoma de *S. pennellii* (*scaffolds > 2 Kpb decon*, ver Tabla 17).

Tabla 17: Estadísticas detalladas del ensamblado del genoma de *S. pennellii*. Las estadísticas se dan después de cada etapa de ensamblado con *SOAPdenovo*. Las estadísticas se calcularon utilizando la herramienta *statsContigAll* de http://code.google.com/p/curtain/, contra los *contigs* y/o *scaffolds* producidos en cada etapa. post *PE* se refiere al resultado después de usar solo las bibliotecas *PE* para el andamiaje. *Post MP* 3-5, 8-10, 20, 30 y 40 Kpb se refiere a los resultados después del andamiaje con la biblioteca *MP* correspondiente. Mín. y Máx. son los tamaños de secuencia mínimo y máximo y el total se refiere a la suma de las longitudes de todas las secuencias. Tabla tomada y adaptada de (**Bolger et al., 2014a**)

Etapa	N10*	N20*	N30*	N40*	N50*	N60*	N70*	N80*	N90*	Mín*	Máx*	Total*	Σ de *Contigs y/o Scaffolds*	n > N50	Ns
Contigs	17.180	10.976	7.148	4.370	2.176	779	246	108	68	64	80.781	1.117.562.721	4.315.954	81.824	0
post PE	114.924	78.931	59.405	45.150	33.782	24.683	16.565	9.015	2.281	100	391.544	907.124.276	464.970	6.982	10.971.966
post MP 3-5 Kpb	291.764	200.811	152.980	116.613	88.502	64.533	43.654	23.524	3.572	100	925.062	919.696.645	435.788	2.750	24.672.194
post MP 8-10 Kpb	726.647	506.420	390.275	300.355	227.029	170.160	119.359	71.492	19.034	100	1.973.027	982.759.790	413.693	1.162	87.631.776
post MP 20Kpb	2.215.469	1.476.079	1.168.714	944.145	738.109	566.421	394.021	238.393	45.648	100	4.240.888	1.006.163.183	409.235	386	110.997.002
post MP 30 Kpb	4.951.557	3.283.930	2.463.944	1.937.804	1.541.925	1.184.815	918.483	562.572	107.406	100	9.085.884	1.016.203.713	407.707	182	120.600.869
post MP 40 Kpb	5.003.978	3.370.668	2.482.965	2.019.941	1.603.317	1.212.905	944.646	600.486	98.078	100	10.126.651	1.021.472.455	407.506	177	125.806.430
scaffolds	4.970.060	3.341.192	2.463.860	1.997.983	1.590.935	1.196.704	935.772	594.885	95.443	100	10.011.355	1.012.612.203	407.506	177	67.624.937
scaffolds > 2 Kpb	5.114.738	3.416.689	2.670.319	2.113.751	1.741.129	1.353.889	1.059.177	763.066	437.042	2000	10.011.355	942.624.776	4.591	156	67.190.024
scaffolds > 2 Kpb decon	5.114.738	3.416.689	2.670.319	2.113.751	**1.741.129**	1.353.889	1.059.177	763.066	437.042	2000	10.011.355	**942.595.034**	**4.579**	**156**	67.190.021

* valores expresados como cantidad de nucleótidos

V.2.8 | Ensamblado de pseudomoleculas (cromosomas)

Búsqueda y organización de marcadores moleculares

De las 13.763 secuencias de marcadores moleculares disponibles en bases de datos mapeados en tomate, 6.328 se seleccionaron y utilizaron como marcas para anclar y asignarle una posición a los *scaffolds* de *S. pennellii* sobre el mapa genético de Tomate EXPEN2000 (ver Tabla 18). Los marcadores moleculares fueron recolectados de diferentes fuentes: i) 4.103 secuencias corresponden a marcadores mapeados del tipo *RFLPs, SSRs, PCRs* y *COS* descargados del portal público www.solgenomics.net; ii) 543 secuencias corresponden a secuencias de puntas de *BAC*s previamente mapeadas en nuestro laboratorio (**Kamenetzky et al., 2010**); iii) 333 secuencias corresponden a marcadores *DarT* mapeados por **Van Schalkwyk et al., (2012**) y iv) 8.784 secuencias corresponden a marcadores *SNP* mapeados por **Sim et al., (2012**).

Tabla 18: Marcadores moleculares con secuencia y posición (cM = *centiMorgan*) sobre el mapa genético Tomato EXPEN2000. En la tabla se indican la fuente, el tipo y cantidad de marcadores recopilados y el número de marcadores con homología de secuencia en los *scaffolds* de *Solanum pennellii* v1.90; los cuales aportan información posicional unívoca (número de cromosoma y ubicación en cM) para ensamblar el mapa físico de 12 pseudomoléculas.

Fuentes de marcadores moleculares	Tipo de marcadores	# de marcadores	# de marcadores en *scaffolds*
Tomato EXPEN 1992 y 2000	RFLPs, SSRs, PCRs and COS	4.103	2.516
Kamenetzky et al., 2010	BACs & Bac-ends	543	543
Schalkwyk et al., 2012	DArT	333	109
Sim et al., 2012	SNPs	8.784	3.160
Total		13.763	6.328

Por otra parte, se realizó una búsqueda por similitud de secuencia entre los marcadores moleculares de la Tabla 19 y la genoteca de clones de *BACs* de *Solanum pennellii* (descripta anteriormente) para marcarlos y asignarles información posicional, y de esta manera, utilizarlos como marcadores adicionales; ya que los *BACs* (secuenciados por el método de Sanger) son más largos que el promedio de

los marcadores moleculares recolectados, lo cual aumentó la posibilidad de anclar un mayor número de *scaffolds* y asignarlos al mapa genético para construir las pseudomoléculas (ver Tabla 19).

Tabla 19: Resumen de *scaffolds* y cantidad de nucleótidos anclados al mapa genético.

Fuentes de marcadores moleculares	Tipo de marcadores	Σ de marcadores	Σ de marcadores en *scaffolds*	Σ de marcadores unívocos	Σ de *scaffolds* anclados	Σ de nucleótidos anclados
Tomato EXPEN 1992 y 2000	*RFLPs, SSRs, PCRs & COS*	4.103	2.516	1.486	305	361.055.986
Kamenetzky et al., 2010	*BACs*	543	543	169	38	54.237.918
Schalkwyk et al., 2012	*DArT*	333	109	87	42	477.814.024
Sim et al., 2012	*SNPs*	8.784	3.160	837	76	113.217.110
Genoteca de BACs	*BAC-ends*	10.612	10.605	167	57	39.834.888
Total		24.375	16.933	2.746	518	630.866.022
Tamaño del genoma ensamblado de *S. pennellii* v1.90						942.624.776
Porcentaje del genoma anclado al mapa genético						~ 67%

Basándonos en la alta sintenia entre *S. pennellii* y *S. lycopersicum* (ver introducción general, sección 0.1.5) los *scaffolds* restantes se ordenaron y orientaron alineándolos contra el genoma del cultivar Heinz mediante la herramienta *MUMmer,* posicionado así el 97% del genoma de *Solanum pennellii* en el mapa genético (ver Tabla 20).

Tabla 20: Resumen de *scaffolds* anclados al mapa genético *Tomato* EXPEN 2000

Etapa	N50	Σ de *scaffolds*	Σ de nucleótidos anclados	Ns	% Anclado al mapa genético
Anclado	1.741.129	13	948.595.034	67.190.021	97 %

V.3 | Capítulo III

V.3.1 | Material vegetal y condiciones experimentales

Las semillas de las líneas de introgresión (*ILs*) de *S. pennellii* (**Eshed y Zamir, 1995**) y de *S. lycopersicum* cv M82 fueron gentilmente donadas por el Profesor Dani Zamir de la Universidad Hebrea de Rehovot en Israel. Las semillas fueron germinadas en medios de Murashige y Skoog (**Murashige y Skoog, 1962**) que contenían un 2% (p/v) de sacarosa y se cultivaron en una cámara de crecimiento (250 µmol fotones m-2 s-1,22°C) durante 45 días bajo un fotoperiodo de 16 hs de luz y 8 hs de oscuridad, antes de su transferencia al invernadero. Las plántulas enraizadas fueron aclimatadas como se describió anteriormente en **Carrari et al., (2003**) y posteriormente fueron transferidas a macetas de 1,16 lts. que contenían sustrato comercial *Plantmax*®, complementado al inicio del cultivo con 2,5 g de fertilizante químico (N: 4; P2O5:14; K2O: 8) por macetero. Las plantas se cultivaron en un invernadero ubicado en la localidad de *Viçosa* (642 m de altitud, 20º45' de latitud sudoriental y 42º51' de longitud oeste) en el sudeste de Brasil, con una intensidad lumínica media de 400 µmol de fotones m-2 s-1 diarios. Durante todo el período de crecimiento, las plantas se mantuvieron en condiciones naturalmente fluctuantes de intensidad lumínica, temperatura y humedad relativa del aire. El experimento, que comprende 71 *ILs* más el cultivar elite M82, se repitió en la primavera de dos años consecutivos (2012 y 2013). Seis plantas de cada uno de los genotipos (M82 e *ILs* derivadas) fueron dispuestas en un diseño de bloques completos aleatorizados (RCBD), considerando como unidad experimental una planta por maceta.

V.3.2 | Mediciones de parámetros fotosintéticos

Todos los procedimientos de muestreo se llevaron a cabo en la primera hoja fuente completamente expandida en plantas de 4 semanas de edad. Las muestras de hojas para determinar el contenido de almidón se recogieron al final de la noche (Almidón FN), a media tarde (Almidón MD) y al final del día (Almidón FD). El crecimiento de la planta se estimó utilizando la altura de la planta (medida desde

la base hasta la punta de la yema terminal) y un componente temporal que proporciona un parámetro integrador denominado tasa de crecimiento relativo (*RGR*) según lo descrito por (**Hunt, 1982**). La materia seca acumulada en brotes y raíces se determinó para cada planta individual al final del experimento, cuatro semanas después de la transferencia al invernadero como se describió anteriormente (**Nunes-Nesi et al., 2005**) Se realizaron análisis de intercambio gaseoso sobre el segundo foliolo terminal de la tercera hoja completamente expandida desde el ápice en plantas de 4 semanas de edad utilizando dos equipos portátiles e idénticos para medir intercambio de gas de flujo abierto (Li 6400XT, *Li-Cor®*) equipados con una cámara de fluorescencia integrada (LI-6400-40; *Li-Cor®*), utilizando exactamente los mismos ajustes de calibración para ambos equipos. La tasa neta de asimilación de CO2 (*A*), la conductancia de las estomas a vapor de agua (*gs*) y la concentración interna de CO_2 (*Ci*) se midieron de 08:00 a 12:00 hs (tiempo solar) bajo una concentración externa de CO_2 de 400 µmol mol-1 aire, una temperatura media de 27 º C y un caudal de 300 µmol s^{-1} en ambos años. Todas las mediciones se llevaron a cabo bajo una luz artificial saturada de 1.000 µmol de fotones m^{-2} s^{-1} que fue suministrada por un diodo emisor de luz (*LED*) de color azul al 10% para maximizar la apertura estomatal. La eficiencia intrínseca del uso del agua (*WUE*) fue estimada según lo descrito por **Cavatte et al., (2012**). Las tasas de respiración en oscuridad (*Rd*) se determinaron de 21:00 a 23:30 h en la misma hoja que se usó previamente para determinar la asimilación neta de carbono. A partir de las mediciones com*bin*adas de fluorescencia e intercambio de gas, estimamos la tasa de fotorrespiración (*Pr*) según (**Valentini et al., 1995**). La misma hoja utilizada anteriormente para determinar *A*, fue adaptada a la oscuridad durante 1 hora e iluminada con haces de medición modulados y débiles (0,03 µmol m^{-2} s^{-1}) para obtener la fluorescencia inicial (F_o). Se aplicaron pulsos de luz blanca saturada de 8.000 µmol de fotones m^{-2} s^{-1} para determinar la emisión máxima de fluorescencia (F_m). Por otro lado, en las hojas adaptadas a la luz, se midió el rendimiento de fluorescencia (F_s) de estado estable y se midió el pulso saturado de luz blanca (8.000 µmol m^{-2} s^{-1}; 0,8 s) para determinar la fluorescencia máxima (F_m') adaptada a la luz. Luego, la luz actínica se apagó y se aplicó una iluminación de color rojo oscuro (2 µmol m^{-2} s^{-1}) para medir la fluorescencia inicial adaptada a la

luz (F_o') según **Perez et al., (2014)**. Por último, se determinaron los coeficientes de enfriamiento fotoquímico (*qP*) y de enfriamiento no fotoquímico (*NPQ*) y también la tasa de transporte de electrones (*ETR*), como se describió anteriormente por **Schreiber et al., (1995)** y **Perez et al., (2014)**.

V.3.3 | Determinación de los niveles de metabolitos

Para estas determinaciones las muestras de hojas se recogieron en medio del período de luz, utilizando la misma hoja previamente utilizada para determinar los parámetros fotosintéticos. El tejido fue congelado en nitrógeno líquido y almacenado a -80 °C hasta su posterior análisis. La extracción de metabolitos se realizó mediante la molienda del tejido en nitrógeno líquido seguido de la adición inmediata de 80% de etanol, según lo descrito por **Cross et al., (2006)**. Los niveles de almidón, sacarosa, fructosa y glucosa en el tejido foliar se determinaron exactamente como se describió anteriormente por **Fernie et al., (2001)**. Los niveles de malato y fumarato se determinaron exactamente como se detalla en **Nunes-Nesi et al., (2007)**, mientras que los contenidos de nitrato (NO_3) y clorofila *a* y *b* se determinaron como se detalló en **Sulpice et al., (2009)**. El contenido total de proteínas y aminoácidos se cuantificó como se informó anteriormente (**Cross et al., 2006**).

V.3.4 | Métodos para la selección de genes candidato

Análisis estadístico

Los análisis estadísticos se realizaron utilizando el programa *GENES* (**Cruz, 2013**) y el software estadístico R (www.r-project.org). Todos los conjuntos de datos de 2012 y 2013 se sometieron a un análisis de varianza de dos vías (ANOVA), que se realizó suponiendo efectos aleatorios para la interacción de bloques, genotipos, entorno y genotipo x entorno. El modelo lineal que se ajustó a los datos fue $Y_{ijk} = \mu + G_i + B_k + A_j + GA_{ij} + E_{ijk}$, donde Y_{ijk} es el valor observado de la i-ésima línea en el k-ésimo entorno y el j-ésimo bloque; μ es la gran media, G_i el efecto aleatorio de la línea i; B_k el efecto aleatorio del bloque k; A_j el efecto aleatorio del entorno j; GA_{ij} el efecto aleatorio de la interacción del genotipo i con el entorno j y E_{ijk} es el error experimental. Las medias se probaron mediante la prueba de *Dunnett* con un nivel de significancia del 5%. La correlación de *Pearson* se realizó para examinar las relaciones entre las variables. Los valores de *Pearson* r se convirtieron en *t-valores* para obtener los *p-valores*. Los resultantes *p*-valores se corrigieron utilizando el control de tasa de descubrimiento falso de **Benjamini y Hochberg (1995)**, utilizando un *script* en R del software *Rbio* (www.biometria.ufv.br).

Criterios para la selección de *QTL*

En este trabajo se define *QTL* como aquellas regiones genómicas para las cuales un caracter se asoció significativamente entre la *IL* y su parental recurrente (M82) mediante la prueba de *Dunnett* (p-valor <0.05). La selección de los *BINs* asociados a *QTL* para realizar los análisis genómicos comparativos se basó en el criterio de que las regiones genómicas estaban cubiertas por al menos dos *ILs* que mostraban una misma variación fenotípica con respecto al *loci* original de M82.

Anotación de genes y tabla de contingencia

Las secuencias de los 11 *BIN*s seleccionados se extrajeron de los genomas parentales (M82 y *S. pennellii*, previamente secuenciados en **Bolger et al., (2014a)**) utilizando los marcadores flanqueantes de cada *BIN* (ver Tabla Anexa 2, sección VIII.3). La

predicción estructural de los genes se realizó *de novo* mediante el uso del *software Augustus* (**Stanke et al., 2004**). Las proteínas predichas se anotaron funcionalmente por alineamiento reciproco con la base de datos de proteínas del cv Heinz (ITAG 2.30) utilizando el motor de búsqueda *BLASTP* (e-valor = 1E-10). Las proteínas que coincidían con los elementos transponibles se eliminaron del conjunto de datos. Las listas de genes no redundantes obtenidas para cada *BIN* se extrajeron mediante un *script* dedicado (Tabla S1, anexa en la versión digital) y luego se agruparon en términos *MapMan* utilizando la tabla de ontologías de Solyc_ITAG_2.30 (https://mapman.gabipd.org/). Esto se hizo por separado para los genes de M82 y *S. pennellii*. Para evaluar la sobrerrepresentación de cualquier término *MapMan* dentro de los 11 *BIN*s analizados, después de la categorización se construyó una tabla de contingencia entre el número de genes asignados a un término *MapMan* en cada *BIN* y se lo comparo con el total de genes asignados a un término *MapMan* en los genomas parentales (M82 y *S. pennellii*). La significancia ($p < 0.05$) fue evaluada mediante la prueba de *Fisher* (ver Tabla Anexa 3, sección VIII.3). Las anotaciones funcionales del contenido génico de los 11 *BIN*s fueron revisadas y curadas manualmente, las proteínas predichas conllevan una descripción funcional curada y detallada (asignada por el consorcio *ITAG* 2.30). En este sentido, como primer criterio de selección se revisó la descripción funcional de cada uno de los *locus* predichos y si esta descripción indicaba una fuerte relación funcional con los mecanismos de fotosíntesis, metabolismo y crecimiento de la planta, el *locus* para la proteína anotada se seleccionó como gen candidato. De esta manera se obtuvo, analizando los 11 *BIN*s, un listado de 87 genes candidatos (Tablas S1 y S4, anexas en la versión digital, tomadas de **de Oliveira Silva et al., 2017**).

Capítulo VI: Financiamiento

Los recursos financieros destinados a este proyecto fueron provistos por dos subsidios obtenidos por nuestro grupo de trabajo, del cual, el Dr. Fernando Carrari es el investigador principal y director de los mismos. A saber:

a) PIP Conicet 5.788: Caracterización funcional de genes relacionados con respuestas a estreses bióticos y abióticos mediante técnicas genómicas y transformación genética de leguminosas y solanáceas (2006-2009).

b) *Start-Up Grant* de la Sociedad Max Planck (Alemania): *Biodiversity exploitation: Identification of metabolic QTL using wild species tomato* (2007-2012).

Capítulo VII: Bibliografía

Las citas y referencias bibliográficas de esta Tesis fueron incorporadas mediante la herramienta *Mendeley Desktop* (**The Mendeley Support Team, 2011**) usando el estilo de la Revista Argentina de Antropología Biológica.

Akihiro T, Koike S, Tani R, Tominaga T, Watanabe S, Iijima Y, Aoki K, Shibata D, Ashihara H, et al. 2008. Biochemical mechanism on GABA accumulation during fruit development in tomato. Plant Cell Physiol. 49:1378-1389.

Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, et al. 2011. De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). Nat. Biotechnol. 29:521-527.

Almeida J, Quadrana L, Asís R, Setta N, De Godoy F, Bermúdez L, Otaiza SN, Corrêa Da Silva J V., Fernie AR, et al. 2011. Genetic dissection of vitamin e biosynthesis in tomato. J. Exp. Bot. 62:3781-3798.

Anderson S. 1981. Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic Acids Res. 9:3015-3027.

Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M. 1987. Automated DNA sequencing: Ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Res. 15:4593-4602.

Ansorge W, Sproat BS, Stegemann J, Schwager C. 1986. A non-radioactive automated method for DNA sequence determination. J. Biochem. Biophys. Methods [Internet] 13:315-23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3559035

Ansorge WJ. 2009. Next-generation DNA sequencing techniques. N. Biotechnol. 25:195-203.

Araújo WL, Nunes-Nesi A, Osorio S, Usadel B, Fuentes D, Nagy R, Balbo I, Lehmann M, Studart-Witkowski C, et al. 2011. Antisense Inhibition of the Iron-Sulphur Subunit of Succinate Dehydrogenase Enhances Photosynthesis and Growth in Tomato via an Organic Acid–Mediated Effect on Stomatal Aperture. Plant Cell [Internet] 23:600-627. Available from: http://www.plantcell.org/lookup/doi/10.1105/tpc.110.081224

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, et al. 2011. The genome of Theobroma cacao. Nat. Genet. [Internet] 43:101-108. Available from: http://www.nature.com/doifinder/10.1038/ng.736

Arumuganathan K, Earle ED. 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Report. 9:208-218.

Arumuganathan K, Slattery JP, Tanksley SD, Earle ED. 1991. Preparation and flow cytometric analysis of metaphase chromosomes of tomato. Theor. Appl. Genet. [Internet] 82:101-111. Available from: http://www.springerlink.com/index/10.1007/BF00231283

Ashikari M, Sakakibara H, Lin S, Yamamoto T, Takashi T, Nishimura A, Angeles ER, Qian Q, Kitano H, et al. 2005. Plant science: Cytokinin oxidase regulates rice grain production. Science (80-.). 309:741-745.

Asociación Tomate 2000. 2016. Informe progresos 2015-2016. Available from: https://inta.gob.ar/sites/default/files/inta_programa_para_el_aumento_de_la_competitividad_de_la_industria_del_tomate._informe_de_progresos_2016-2017.pdf

Asprelli PD, Sance M, Insani EM, Asis R, Valle EM, Carrari F, Galmarini CR, Peralta IE. 2017. Agronomic performance and fruit nutritional quality of an Andean tomato collection. En: Acta Horticulturae. Vol. 1159. . p 197-204.

Ayliffe MA, Timmis j. N, Steele Scott N. 1988. Homologies to chloroplast DNA in the nuclear DNA of a number of Chenopod species. Theor. Appl. Genet. 75:282-285.

Ayliffe MA, Timmis JN. 1992. Plastid DNA sequence homologies in the tobacco nuclear genome. Mol. Gen. Genet. [Internet] 236:105-12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1337369

Backert S, Börner T. 2000. Phage T4-like intermediates of DNA replication and recombination in the mitochondria of the higher plant Chenopodium album (L.). Curr. Genet. [Internet] 37:304-14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10853767

Bai Y, Lindhout P. 2007. Domestication and breeding of tomatoes: What have we gained and what can we gain in the future? Ann. Bot. 100:1085-1094.

Barone A, Chiusano ML, Ercolano MR, Giuliano G, Grandillo S, Frusciante L. 2008. Structural and Functional Genomics of Tomato. Int. J. Plant Genomics [Internet] 2008:1-12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18317508

Baxter CJ, Carrari F, Bauke A, Overy S, Hill SA, Quick PW, Fernie AR, Sweetlove LJ. 2005. Fruit carbohydrate metabolism in an introgression line of tomato with increased fruit soluble solids. Plant Cell Physiol. 46:425-437.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B [Internet] 57:289-300. Available from: http://www.jstor.org/stable/2346101

Bertin P, Gallais A. 2001. Genetic variation for nitrogen use efficiency in a set of recombinant inbred lines II - QTL detection and coincidences. Maydica 46:53-68.

Bessey CE. 1906. Crop improvement by utilizing wild species. J. Hered. os-2:112-118.

Bevan MW, Uauy C. 2013. Genomics reveals new landscapes for crop improvement. Genome Biol.

[Internet] 14:206. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-6-206

Boetzer M, Henkel C V., Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics [Internet] 27:578-9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21149342

Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sørensen I, et al. 2014a. The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat. Genet. [Internet] 46:1034-1038. Available from: https://doi.org/10.1038/ng.3046

Bolger AM, Lohse M, Usadel B. 2014b. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, Mayer KFX. 2014c. Plant genome sequencing - applications for crop improvement. Curr. Opin. Biotechnol. 26:31-37.

Bombarely A, Menda N, Tecle IY, Buels RM, Strickler S, Fischer-York T, Pujar A, Leto J, Gosselin J, et al. 2010. The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. Nucleic Acids Res. [Internet] 39:1149-1155. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20935049

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. [Internet] 32:314-31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/6247908

Braun H-P, Binder S, Brennicke A, Eubel H, Fernie AR, Finkemeier I, Klodmann J, König A-C, Kühn K, et al. 2014. The life of plant mitochondrial complex I. Mitochondrion [Internet] 19:295-313. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1567724914000208

Bretó MP, Asins MJ, Carbonell EA. 1993. Genetic variability in Lycopersicon species and their genetic relationships. Theor. Appl. Genet. 86:113-120.

Burr B, Burr FA, Thompson KH, Albertson MC, Stuber CW. 1988. Gene mapping with recombinant inbreds in maize. Genetics [Internet] 118:519-26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3366363

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. 2008. ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome Res. 18:810-820.

Cabezas JA, Ibáñez J, Lijavetzky D, Vélez D, Bravo G, Rodríguez V, Carreño I, Jermakow AM, Carreño J, et al. 2011. A 48 SNP set for grapevine cultivar identification. BMC Plant Biol. [Internet] 11:153. Available from: http://bmcplantbiol.biomedcentral.com/articles/10.1186/1471-2229-11-153

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+:

Architecture and applications. BMC Bioinformatics 10.

Carnal NW, Black CC. 1983. Phosphofructokinase activities in photosynthetic organisms : the occurrence of pyrophosphate-dependent 6-phosphofructokinase in plants and algae. Plant Physiol. [Internet] 71:150-5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16662776

Carrari F, Baxter C, Usadel B, Urbanczyk-Wochniak E, Zanor M-I, Nunes-Nesi A, Nikiforova V, Centero D, Ratzka A, et al. 2006. Integrated Analysis of Metabolite and Transcript Levels Reveals the Metabolic Shifts That Underlie Tomato Fruit Development and Highlight Regulatory Aspects of Metabolic Network Behavior. PLANT Physiol. [Internet] 142:1380-1396. Available from: http://www.plantphysiol.org/cgi/doi/10.1104/pp.106.088534

Carrari F, Fernie AR. 2006. Metabolic regulation underlying tomato fruit development. J. Exp. Bot. [Internet] 57:1883-97. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16449380

Carrari F, Nunes-Nesi A, Gibon Y, Lytovchenko A, Loureiro ME, Fernie AR. 2003. Reduced Expression of Aconitase Results in an Enhanced Rate of Photosynthesis and Marked Shifts in Carbon Partitioning in Illuminated Leaves of Wild Species Tomato. Plant Physiol. [Internet] 133:1322-1335. Available from: http://www.plantphysiol.org/cgi/doi/10.1104/pp.103.026716

Causse M. 2002. QTL analysis of fruit quality in fresh market tomato: a few chromosome regions control the variation of sensory and instrumental traits. J. Exp. Bot. [Internet] 53:2089-2098. Available from: https://academic.oup.com/jxb/article-lookup/doi/10.1093/jxb/erf058

Causse M, Duffe P, Gomez MC, Buret M, Damidaux R, Zamir D, Gur A, Chevalier C, Lemaire-Chamley M, et al. 2004. A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. J. Exp. Bot. [Internet] 55:1671-1685. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15258170

Cavatte PC, Oliveira ÁAG, Morais LE, Martins SCV, Sanglard LMVP, Damatta FM. 2012. Could shading reduce the negative impacts of drought on coffee? A morphophysiological analysis. Physiol. Plant. 144:111-122.

Chaisson MJ, Brinza D, Pevzner PA. 2009. De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res. 19:336-346.

Chaisson MJ, Pevzner PA. 2008. Short read fragment assembly of bacterial genomes. Genome Res. 18:324-330.

Chapman JA, Ho I, Sunkara S, Luo S, Schroth GP, Rokhsar DS. 2011. Meraculous: De novo genome assembly with short paired-end reads. PLoS One [Internet] 6:e23501. Available from: http://dx.plos.org/10.1371/journal.pone.0023501

Cheung WY, Scott NS. 1989. A contiguous sequence in spinach nuclear DNA is homologous to three separated sequences in chloroplast DNA. Theor. Appl. Genet. [Internet] 77:625-633. Available from:

http://download.springer.com/static/pdf/480/art%253A10.1007%252FBF00261234.pdf?auth66=1379865102_ee86751588c5c1fdcc3a224520a0e318&ext=.pdf

Chidgeavadze ZG, Beabealashvilli RS, Atrazhev AM, Kukhanova MK, Azhayev A V, Krayevsky AA. 1984. 2',3'-Dideoxy-3' aminonucleoside 5'-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. Nucleic Acids Res. [Internet] 12:1671-86. Available from: http://www.ncbi.nlm.nih.gov/pubmed/6322115%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC318607

Chitwood DH, Kumar R, Headland LR, Ranjan A, Covington MF, Ichihashi Y, Fulop D, Jiménez-Gómez J, Peng J, et al. 2013. A Quantitative Genetic Basis for Leaf Morphology in a Set of\nPrecisely Defined Tomato Introgression Lines. Plant Cell 25:2465-2481.

Clint Nesbitt T, Tanksley SD, Nesbitt TC, Tanksley SD. 2002. Comparative sequencing in the genus Lycopersicon. Implications for the evolution of fruit size in the domestication of cultivated tomatoes. Genetics [Internet] 162:365-79. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1462239&tool=pmcentrez&rendertype=abstract

Cohen SN, Chang ACY, Boyer HW, Helling RB. 1973. Construction of Biologically Functional Bacterial Plasmids In Vitro. Proc. Natl. Acad. Sci. [Internet] 70:3240-3244. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.70.11.3240

Colvine S, Branthôme FX. 2016. The Tomato: A Seasoned Traveller. En: Causse M., Giovannoni J., Bouzayen M. ZM, editor. The Tomato Genome. Compendium of Plant Genomes. Springer, Berlin, Heidelberg. p 1-5. Available from: http://link.springer.com/10.1007/978-3-662-53389-5_1

Correns VCL. 1908. Vererbungsversuche mit blass (gelb) grtinen und buntbl ittrigen Sippen bei Mirabilis Jalapa, Urticapilu life a und Zunaria annua. Zeitschrift für Indukt. Abstammungs und Vererbungslehre 1:291-329.

Cross JM, von Korff M, Altmann T, Bartzetko L, Sulpice R, Gibon Y, Palacios N, Stitt M. 2006. Variation of Enzyme Activities and Metabolite Levels in 24 Arabidopsis Accessions Growing in Carbon-Limited Conditions. PLANT Physiol. [Internet] 142:1574-1588. Available from: http://www.plantphysiol.org/cgi/doi/10.1104/pp.106.086629

Cruz CD. 2013. GENES - a software package for analysis in experimental statistics and quantitative genetics. Acta Sci. Agron. [Internet] 35:271-276. Available from: http://periodicos.uem.br/ojs/index.php/ActaSciAgron/article/view/21251

D'Hont A, Denoeud F, Aury J-MM, Baurens F-CC, Carreel FF, Garsmeur O, Noel B, Bocs SS, Droc GG, et al. 2012. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature [Internet] 488:213-217. Available from:

http://www.nature.com/doifinder/10.1038/nature11241

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394-1403.

Dellapenna D, Kates DS, Bennett AB. 1987. Polygalacturonase Gene Expression in Rutgers, rin, nor, and Nr Tomato Fruits. Plant Physiol. [Internet] 85:502-7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16665727

Do PT, Prudent M, Sulpice R, Causse M, Fernie AR. 2010. The Influence of Fruit Load on the Tomato Pericarp Metabolome in a Solanum chmielewskii Introgression Line Population. PLANT Physiol. [Internet] 154:1128-1142. Available from: http://www.plantphysiol.org/cgi/doi/10.1104/pp.110.163030

Driever SM, Lawson T, Andralojc PJ, Raines CA, Parry MAJ. 2014. Natural variation in photosynthetic capacity, growth, and yield in 64 field-grown wheat genotypes. J. Exp. Bot. 65:4959-4973.

Dubey A, Farmer A, Schlueter J, Cannon SB, Abernathy B, Tuteja R, Woodward J, Shah T, Mulasmanovic B, et al. 2011. Defining the transcriptome assembly and its use for genome dynamics and transcriptome profiling studies in Pigeonpea (Cajanus cajan L.). DNA Res. 18:153-164.

Eisenhut M, Hocken N, Weber APM. 2015. Plastidial metabolite transporters integrate photorespiration with carbon, nitrogen, and sulfur metabolism. Cell Calcium 58.

El-Lithy ME, Rodrigues GC, Van Rensen JJS, Snel JFH, Dassen HJHA, Koornneef M, Jansen MAK, Aarts MGM, Vreugdenhil D. 2005. Altered photosynthetic performance of a natural Arabidopsis accession is associated with atrazine resistance. J. Exp. Bot. 56:1625-1634.

Ellul P, Garcia-Sogo B, Pineda B, Ríos G, Roig L, Moreno V. 2003. The ploidy level of transgenic plants in Agrobacterium-mediated transformation of tomato cotyledons (Lycopersicon esculentum L.Mill.) is genotype and procedure dependent. Theor. Appl. Genet. [Internet] 106:231-238. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12582848

Eshed Y, Zamir D. 1995. An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics [Internet] 141:1147. Available from: http://www.genetics.org/cgi/content/abstract/141/3/1147

Ezura H. 2013. From randomly to inevitable: Accelerating tomato breeding by comprehensive tools and information. Breed. Sci. [Internet] 63:1-2. Available from: http://jlc.jst.go.jp/DN/JST.JSTAGE/jsbbs/63.1?lang=en&from=CrossRef&type=abstract

Fanourakis D, Giday H, Milla R, Pieruschka R, Kjaer KH, Bolger M, Vasilevski A, Nunes-Nesi A, Fiorani F, et al. 2015. Pore size regulates operating stomatal conductance, while stomatal

densities drive the partitioning of conductance between leaf sides. Ann. Bot. 115:555-565.

Fauron C, Casper M. 1995. The maize mitochondrial genome: dynamic, yet functional. Trends Genet. 11:228-235.

Fernie AR, Roscher A, Ratcliffe RG, Kruger NJ. 2001. Fructose 2,6-bisphosphate activates pyrophosphate: Fructose-6-phosphate 1-phosphotransferase and increases triose phosphate to hexose phosphate cycling heterotrophic cells. Planta 212:250-263.

Flood PJ, Harbinson J, Aarts MGM. 2011. Natural genetic variation in plant photosynthesis. Trends Plant Sci. [Internet] 16:327-335. Available from: http://dx.doi.org/10.1016/j.tplants.2011.02.005

Foolad MR. 2007. Genome mapping and molecular breeding of tomato. Int. J. Plant Genomics 2007.

Franssen SU, Shrestha RP, Bräutigam A, Bornberg-Bauer E, Weber AP. 2011. Comprehensive transcriptome analysis of the highly complex Pisum sativum genome using next generation sequencing. BMC Genomics [Internet] 12:227. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-227

Frary A, Nesbitt TC, Frary A, Grandillo S, Knaap E Van Der, Cong B, Liu J, Meller J, Elber R, et al. 2000. Locus Key to the Evolution of Tomato Fruit Size. Science (80-.). 289:85-88.

Fray RG, Grierson D. 1993. Molecular genetics of tomato fruit ripening. Trends Genet. [Internet] 9:438-43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8122312

Fridman E, Carrari F, Liu Y-S, Fernie AR, Zamir D. 2004. Zooming In on a Quantitative Trait for Tomato Yield Using Interspecific Introgressions. Science (80-.). [Internet] 305:1786-1789. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15375271

Fridman E, Pleban T, Zamir D. 2000. A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. Proc. Natl. Acad. Sci. U. S. A. [Internet] 97:4718-4723. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=18299&tool=pmcentrez&rendertype=abstract

Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD, Hoeven R Van Der, Eannetta NT, Tanksley SD. 2002. Identification , Analysis , and Utilization of Conserved Ortholog Set Markers for Comparative Genomics in Higher Plants. Society [Internet] 14:1457-1467. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12119367%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC150699

Gago J, Daloso D de M, Figueroa CM, Flexas J, Fernie AR, Nikoloski Z. 2016. Relationships of Leaf Net Photosynthesis, Stomatal Conductance, and Mesophyll Conductance to Primary Metabolism: A Multispecies Meta-Analysis Approach. Plant Physiol. [Internet] 171:265-279.

Available from: http://www.plantphysiol.org/lookup/doi/10.1104/pp.15.01660

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, Gonzalez VM, Henaff E, Camara F, Cozzuto L, et al. 2012. The genome of melon (Cucumis melo L.). Proc. Natl. Acad. Sci. [Internet] 109:11872-11877. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1205415109

Garg R, Patel RK, Tyagi AK, Jain M. 2011. De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA Res. [Internet] 18:53-63. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3041503&tool=pmcentrez&rendertype=abstract

Geisler DA, Papke C, Obata T, Nunes-Nesi A, Matthes A, Schneitz K, Maximova E, Araujo WL, Fernie AR, et al. 2012. Downregulation of the -Subunit Reduces Mitochondrial ATP Synthase Levels, Alters Respiration, and Restricts Growth and Gametophyte Development in Arabidopsis. Plant Cell [Internet] 24:2792-2811. Available from: http://www.plantcell.org/cgi/doi/10.1105/tpc.112.099424

Giovannoni JJ. 2004. Genetic regulation of fruit development and ripening. Plant Cell [Internet] 16 Suppl:S170-80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15010516

Giovannoni JJ. 2007. Fruit ripening mutants yield insights into ripening control. Curr. Opin. Plant Biol. [Internet] 10:283-289. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17442612

De Godoy F, Bermúdez L, Lira BS, De Souza AP, Elbl P, Demarco D, Alseekh S, Insani M, Buckeridge M, et al. 2013. Galacturonosyltransferase 4 silencing alters pectin composition and carbon partitioning in tomato. J. Exp. Bot. 64:2449-2466.

Goff SA, Klee HJ. 2006. Plant Volatile Compounds: Sensory Cues for Health and Nutritional Value? Science (80-.). [Internet] 311:815-819. Available from: http://www.sciencemag.org/cgi/content/abstract/sci;311/5762/815

Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, et al. 2002a. A draft sequence of the rice genome (sativa L. ssp. japonica). Science (80-.). [Internet] 296:92-100. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1068275

Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, et al. 2002b. A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica). Science (80-.). 296:92-100.

Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, et al. 2009. A first-generation haplotype map of maize. Science (80-.). 326:1115-1117.

Grandillo S, Cammareri M. 2016. Molecular Mapping of Quantitative Trait Loci in Tomato. En:

Springer, Berlin, Heidelberg. p 39-73. Available from: http://link.springer.com/10.1007/978-3-662-53389-5_4

Gray MW. 1999. Evolution of organellar genomes. Curr. Opin. Genet. Dev. [Internet] 9:678-87. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10607615

Gu J, Yin X, Struik PC, Stomph TJ, Wang H. 2012. Using chromosome introgression lines to map quantitative trait loci for photosynthesis parameters in rice (Oryza sativa L.) leaves under drought and well-watered field conditions. J. Exp. Bot. 63:455-469.

Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, et al. 2012. The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. Nat. Genet. [Internet] 45:51-58. Available from: http://www.nature.com/doifinder/10.1038/ng.2470

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 29:1072-1075.

Handa H. 2003. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (Brassica napus L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana. Nucleic Acids Res. [Internet] 31:5907-5916. Available from: http://www.nar.oupjournals.org/cgi/doi/10.1093/nar/gkg795

Hazkani-Covo E, Zeller RM, Martin W. 2010. Molecular Poltergeists: Mitochondrial DNA Copies (<italic>numts</italic>) in Sequenced Nuclear Genomes. PLoS Genet [Internet] 6:e1000834. Available from: http://dx.doi.org/10.1371/journal.pgen.1000834

Heather JM, Chain B. 2016. The sequence of sequencers: The history of sequencing DNA. Genomics [Internet] 107:1-8. Available from: http://dx.doi.org/10.1016/j.ygeno.2015.11.003

Hikosaka K. 2010. Mechanisms underlying interspecific variation in photosynthetic capacity across wild plant species. Plant Biotechnol. 27:223-229.

Hill R. 2012. Tomato genome sequence bears fruit. Nature [Internet]. Available from: http://www.nature.com/doifinder/10.1038/nature.2012.10751

Hiremath PJ, Farmer A, Cannon SB, Woodward J, Kudapa H, Tuteja R, Kumar A, Bhanuprakash A, Mulaosmanovic B, et al. 2011. Large-scale transcriptome analysis in chickpea (Cicer arietinum L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. Plant Biotechnol. J. 9:922-931.

Hollmer M. 2013. Roche to close 454 Life Sciences as it reduces gene sequencing focus. Fierce Biotech [Internet]. Available from: http://www.fiercebiotech.com/medical-devices/roche-to-close-454-life-sciences-as-it-reduces-gene-sequencing-focus

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, et al. 2009. The genome of the cucumber, Cucumis sativus L. Nat. Genet. [Internet] 41:1275-1281. Available from:

http://www.nature.com/doifinder/10.1038/ng.475

Huang X, Madan a. 1999. CAP3: A DNA sequence assembly program. Genome Res. [Internet] 9:868-77. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=310812&tool=pmcentrez&rendertype=abstract

Hueso Estornell L, Leandro. 2010. Estudio transcriptómico y metabolómico del desarrollo partenocárpico del fruto del tomate y aplicaciones biotecnológicas. Available from: http://digital.csic.es/handle/10261/24025

Hunkapiller T, Kaiser RJ, Koop BF, Hood L. 1991. Large-scale and automated DNA sequence determination. Science (80-.). 254:59-67.

Hunt R. 1982. Plant growth curves. The functional approach to plant growth analysis. Plant growth curves. Funct. approach to plant growth Anal.

Hyman ED. 1988. A new method of sequencing DNA. Anal. Biochem. 174:423-436.

Illumina. 2017. An Introduction to Next-Generation Sequencing Technology. Available from: http://res.illumina.com/documents/products/illumina_sequencing_introduction.pdf

Jackson DA, Symons RH, Berg P. 1972. Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of Escherichia coli. Proc. Natl. Acad. Sci. [Internet] 69:2904-2909. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.69.10.2904

Jaillon O, Aury J-MM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature [Internet] 449:463-467. Available from: http://www.nature.com/doifinder/10.1038/nature06148

Janska H, Sarria R, Woloszynska M, Arrieta-Montiel M, Mackenzie S. 1998. Stoichiometric shifts in the common bean mitochondrial genome leading to male sterility and spontaneous reversion to fertility. Plant Cell [Internet] 10:1163-80. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=144058&tool=pmcentrez&rendertype=abstract

du Jardin P. 1990. Homologies to plastid DNA in the nuclear and mitochondrial genomes of potato. Theor. Appl. Genet. [Internet] 79:807-812. Available from: http://link.springer.com/10.1007/BF00224249

Kambara H, Nishikawa T, Katayama Y, Yamaguchi T. 1988. Optimization of parameters in a DNA sequenator using fluorescence detection. Bio/Technology 6:816-821.

Kamenetzky L, Asís R, Bassi S, de Godoy F, Bermúdez L, Fernie AR, Van Sluys M-A, Vrebalov J,

Giovannoni JJ, et al. 2010. Genomic analysis of wild tomato introgressions determining metabolism- and yield-associated traits. Plant Physiol. [Internet] 152:1772-86. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20118271

Kanazawa A, Hirai A. 1994. Reversible Changes in the Composition of the Population of Mtdnas during Dedifferentiation and Regeneration in Tobacco. Genetics [Internet] 138:865-870. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1206234/

Kimura S, Sinha N. 2008. Tomato (Solanum lycopersicum): A model fruit-bearing crop. Cold Spring Harb. Protoc. [Internet] 3:pdb.emo105. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21356708

Klee HJ, Giovannoni JJ. 2011. Genetics and Control of Tomato Fruit Ripening and Quality Attributes. Annu. Rev. Genet. [Internet] 45:41-59. Available from: http://www.annualreviews.org/doi/10.1146/annurev-genet-110410-132507

Kloosterman B, Abelenda JA, Gomez MDMC, Oortwijn M, De Boer JM, Kowitwanich K, Horvath BM, Van Eck HJ, Smaczniak C, et al. 2013. Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature 495:246-250.

Kobayashi T, Kato-Emori S, Tomita K, Ezura H. 2003. Transformation of tomato with the melon 3-hydroxy-3-methylglutaryl coenzyme A reductase leads to increase of fruit size. Plant Biotechnol. [Internet] 20:297-303. Available from: http://jlc.jst.go.jp/JST.JSTAGE/plantbiotechnology/20.297?from=Google

Koornneef M, Alonso-Blanco C, Vreugdenhil D. 2004. Naturally occurring genetic variation in Arabidopsis thaliana. Annu. Rev. Plant Biol. [Internet] 55:141-172. Available from: http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.arplant.55.031903.141605?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3Dncbi.nlm.nih.gov%0Apapers2://publication/doi/10.1146/annurev.arplant.55.031903.141605%0Apapers2://publica

Kubo T, Nishizawa S, Sugawara a, Itchoda N, Estiati a, Mikami T. 2000. The complete nucleotide sequence of the mitochondrial genome of sugar beet (Beta vulgaris L.) reveals a novel gene for tRNA(Cys)(GCA). Nucleic Acids Res. [Internet] 28:2571-6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102699&tool=pmcentrez&rendertype=abstract

Kudapa H, Bharti AK, Cannon SB, Farmer AD, Mulaosmanovic B, Kramer R, Bohra A, Weeks NT, Crow JA, et al. 2012. A comprehensive transcriptome assembly of pigeonpea (Cajanus cajan L.) Using sanger and second-generation sequencing platforms. Mol. Plant 5:1020-1028.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, et al. 2001. Initial sequencing and analysis of the human genome. Nature [Internet] 409:860-921. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/11237011%5Cnhttp://www.nature.com/nature/journal/v409/n6822/pdf/409860a0.pdf

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol [Internet] 10:R25. Available from: file:///Users/slevy/Documents/Endnote_library/PDFs/bowtie_aligner.pdf%5Cnhttp://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19261174

Lawson T, Kramer DM, Raines CA. 2012. Improving yield by exploiting mechanisms underlying natural variation of photosynthesis. Curr. Opin. Biotechnol. 23:215-220.

Leaver CJ, Gray MW. 1982. Mitochondrial Genome Organization and Expression in Higher Plants. Annu. Rev. Plant Physiol. [Internet] 33:373-402. Available from: http://www.annualreviews.org/doi/abs/10.1146/annurev.pp.33.060182.002105

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 20:265-272.

Liang Y, Zhang K, Zhao L, Liu B, Meng Q, Tian J, Zhao S. 2009. Identification of chromosome regions conferring dry matter accumulation and photosynthesis in wheat (Triticum aestivum L.). Euphytica 171:145-156.

Lichtenstein G, Conte M, Asis R, Carrari F. 2016. Chloroplast and Mitochondrial Genomes of Tomato. Springer-Verlag Berlin Heidelb. [Internet] 7:111-137. Available from: http://link.springer.com/10.1007/978-3-662-53389-5

Lippman ZB, Semel Y, Zamir D. 2007. An integrated view of quantitative trait variation using tomato interspecific introgression lines. Curr. Opin. Genet. Dev. [Internet] 17:545-52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17723293

Liu Y-S, Gur A, Ronen G, Causse M, Damidaux R, Buret M, Hirschberg J, Zamir D. 2003a. There is more to tomato fruit colour than candidate carotenoid genes. Plant Biotechnol. J. [Internet] 1:195-207. Available from: http://doi.wiley.com/10.1046/j.1467-7652.2003.00018.x

Liu Y-S, Gur A, Ronen G, Causse M, Damidaux R, Buret M, Hirschberg J, Zamir D. 2003b. There is more to tomato fruit colour than candidate carotenoid genes. Plant Biotechnol. J. 1:195-207.

Lockhart J. 2013. A Quantitative Genetic Basis for Leaf Morphology is Revealed in a Set of Precisely Defined Tomato Introgression Lines. Plant Cell [Internet] 25:2379-2379. Available from: http://www.plantcell.org/cgi/doi/10.1105/tpc.113.250710

Lohse M, Nagel A, Herter T, May P, Schroda M, Zrenner R, Tohge T, Fernie AR, Stitt M, et al. 2014. Mercator: A fast and simple web server for genome scale functional annotation of plant sequence data. Plant, Cell Environ. 37:1250-1258.

Lough AN, Roark LM, Kato A, Ream TS, Lamb JC, Birchler JA, Newton KJ. 2008. Mitochondrial

DNA transfer to the nucleus generates extensive insertion site variation in maize. Genetics [Internet] 178:47-55. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2206095&tool=pmcentrez&rendertype=abstract

Luckey JA, Drossman H, Kostichka AJ, Mead DA, D'Cunha J, Norris TB, Smith LM. 1990. High speed DNA sequencing by capillary electrophoresis. Nucleic Acids Res. 18:4417-4421.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience [Internet] 1:18. Available from: http://www.gigasciencejournal.com/content/1/1/18%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3626529&tool=pmcentrez&rendertype=abstract

MacCallum I, Przybylski D, Gnerre S, Burton J, Shlyakhter I, Gnirke A, Malek J, McKernan K, Ranade S, et al. 2009. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. Genome Biol. [Internet] 10:R103. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-10-r103

Margulis L, Bermudes D. 1985. Symbiosis as a mechanism of evolution: status of cell symbiosis theory. Symbiosis [Internet] 1:101-124. Available from: http://europepmc.org/abstract/MED/11543608

Martin GB, Brommonschenkel SH, Chunwongse J, Frary A, Ganal MW, Spivey R, Wu T, Earle ED, Tanksley SD. 1993. Map-based cloning of a protein kinase gene conferring disease resistance in tomato. Science (80-.). [Internet] 262:1432-6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7902614

Mazzucato A, Taddei AR, Soressi GP. 1998. The parthenocarpic fruit (pat) mutant of tomato (Lycopersicon esculentum Mill.) sets seedless fruits and has aberrant anther and ovule development. Development [Internet] 125:107-14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9389668

Medeiros DB, Martins SCV, Cavalcanti JHF, Daloso DM, Martinoia E, Nunes-Nesi A, DaMatta FM, Fernie AR, Araújo WL. 2016. Enhanced Photosynthesis and Growth in *atquac1* Knockout Mutants Are Due to Altered Organic Acid Accumulation and an Increase in Both Stomatal and Mesophyll Conductance. Plant Physiol. [Internet] 170:86-101. Available from: http://www.plantphysiol.org/lookup/doi/10.1104/pp.15.01053

Melchers G, Mohri Y, Watanabe K, Wakabayashi S, Harada K. 1992. One-step generation of cytoplasmic male sterility by fusion of mitochondrial-inactivated tomato protoplasts with nuclear-inactivated Solanum protoplasts. Proc. Natl. Acad. Sci. U. S. A. [Internet] 89:6832-6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=49598&tool=pmcentrez&render

type=abstract

Melendez-Ferro M, Rice MW, Roberts RC, Perez-Costas E. 2013. An accurate method for the quantification of cytochrome C oxidase in tissue sections. J. Neurosci. Methods [Internet] 214:156-62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23376420

Mereschkowski C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biol Cent. 25:593-604.

Metzker ML. 2010. Sequencing technologies the next generation. Nat. Rev. Genet. 11:31-46.

Meyer RC, Steinfath M, Lisec J, Becher M, Witucka-Wall H, Torjek O, Fiehn O, Eckardt A, Willmitzer L, et al. 2007. The metabolic signature related to high plant growth rate in Arabidopsis thaliana. Proc. Natl. Acad. Sci. [Internet] 104:4759-4764. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0609709104

Michalovova M, Vyskot B, Kejnovsky E. 2013. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: Size, relative age and chromosomal localization. Heredity (Edinb). 111:314-320.

Miller JC, Tanksley SD. 1990. RFLP analysis of phylogenetic relationships and genetic variation in the genus Lycopersicon. Theor. Appl. Genet. [Internet] 80:437-48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24221000

Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shawand PD, Marshall D. 2013. Using tablet for visual exploration of second-generation sequencing data. Brief. Bioinform. 14:193-202.

Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee HT, Chan CKK, Visendi P, Lai K, Doležel J, et al. 2017. The pangenome of hexaploid bread wheat. Plant J. 90:1007-1013.

Mozdzer TJ, Zieman JC. 2010. Ecophysiological differences between genetic lineages facilitate the invasion of non-native Phragmites australis in North American Atlantic coast wetlands. J. Ecol. 98:451-458.

Mueller L a, Tanksley SD, Giovannoni JJ, van Eck J, Stack S, Choi D, Kim BD, Chen M, Cheng Z, et al. 2005. The Tomato Sequencing Project, the first cornerstone of the International Solanaceae Project (SOL). Comp. Funct. Genomics [Internet] 6:153-8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2447522&tool=pmcentrez&rendertype=abstract

Muir CD, Hangarter RP, Moyle LC, Davis PA. 2014. Morphological and anatomical determinants of mesophyll conductance in wild relatives of tomato (Solanum sect. Lycopersicon, sect. Lycopersicoides; Solanaceae). Plant, Cell Environ. 37:1415-1426.

Murashige T, Skoog F. 1962. A Revised Medium for Rapid Growth and Bio Assays with Tobacco Tissue Cultures. Physiol. Plant. 15:473-497.

Nägeli C. 1846. Über Polysiphonia und Herposiphonia. Zeitschrift für wissenschaftliche Bot. [Internet] 4:207–256. Available from: https://species.wikimedia.org/wiki/Carl_Wilhelm_von_Nägeli

Nass S, Nass MM. 1963. Intramitochondrial Fibers With Dna Characteristics. Ii. Enzymatic and Other Hydrolytic Treatments. J. Cell Biol. [Internet] 19:613-629. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2106332&tool=pmcentrez&rendertype=abstract

Notsu Y, Masood S, Nishikawa T, Kubo N, Akiduki G, Nakazono M, Hirai a, Kadowaki K. 2002. The complete sequence of the rice (Oryza sativa L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. Mol. Genet. Genomics [Internet] 268:434-45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12471441

Nunes-Nesi A, Carrari F, Gibon Y, Sulpice R, Lytovchenko A, Fisahn J, Graham J, Ratcliffe RG, Sweetlove LJ, et al. 2007. Deficiency of mitochondrial fumarase activity in tomato plants impairs photosynthesis via an effect on stomatal function. Plant J. 50:1093-1106.

Nunes-Nesi A, Carrari F, Lytovchenko A, Smith AMO, Loureiro ME, Ratcliffe RG, Sweetlove LJ, Fernie AR. 2005. Enhanced photosynthetic performance and growth as a consequence of decreasing mitochondrial malate dehydrogenase activity in transgenic tomato plants. Plant Physiol. [Internet] 137:611-22. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1065362&tool=pmcentrez&rendertype=abstract

Nyrén P. 1987. Enzymatic method for continuous monitoring of DNA polymerase activity. Anal. Biochem. 167:235-238.

Nyrén P, Lundin A. 1985. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. Anal. Biochem. 151:504-509.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-CC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. Nature [Internet] 497:579-584. Available from: http://www.nature.com/doifinder/10.1038/nature12211

O'Brien SJ. 1993. Genetic maps : locus maps of complex genomes. 6th ed. (O'Brien SJ, editor.). Cold Spring Harbor Laboratory Press. Available from: https://books.google.com.ar/books/about/Genetic_Maps.html?id=4di1AQAACAAJ&source=kp_cover&redir_esc=y

Ofner I, Lashbrooke J, Pleban T, Aharoni A, Zamir D. 2016. Solanum pennellii backcross inbred lines (BILs) link small genomic bins with tomato traits. Plant J. 87:151-160.

Oldenburg DJ, Bendich a. J. 1996. Size and Structure of Replicating Mitochondrial DNA in

Cultured Tobacco Cells. Plant Cell [Internet] 8:447-461. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=161112&tool=pmcentrez&rendertype=abstract

de Oliveira Silva FM, Lichtenstein G, Alseekh S, Rosado-Souza L, Conte M, Suguiyama VF, Lira BS, Fanourakis D, Usadel B, et al. 2018. The genetic architecture of photosynthesis and plant growth-related traits in tomato. Plant Cell Environ. 41:327-341.

Omura T, Watanabe S, Iijima Y, Aoki K, Shibata D, Ezura H. 2007. Molecular and genetic characterization of transgenic tomato expressing 3-hydroxy-3-methylglutaryl coenzyme A reductase. Plant Biotechnol. 24:107-115.

Ort DR, Merchant SS, Alric J, Barkan A, Blankenship RE, Bock R, Croce R, Hanson MR, Hibberd JM, et al. 2015. Redesigning photosynthesis to sustainably meet global food and bioenergy demand. Proc. Natl. Acad. Sci. [Internet] 112:8529-8536. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1424031112

Orzaez D. 2005. Agroinjection of Tomato Fruits. A Tool for Rapid Functional Analysis of Transgenes Directly in Fruit. PLANT Physiol. [Internet] 140:3-11. Available from: http://www.plantphysiol.org/cgi/doi/10.1104/pp.105.068221

Osorio S, Alba R, Damasceno CMB, Lopez-Casado G, Lohse M, Zanor MI, Tohge T, Usadel B, Rose JKC, et al. 2011. Systems Biology of Tomato Fruit Development: Combined Transcript, Protein, and Metabolite Analysis of Tomato Transcription Factor (nor, rin) and Ethylene Receptor (Nr) Mutants Reveals Novel Regulatory Interactions. Plant Physiol. 157:405-425.

Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu YL, Song K. 2000. Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc. Natl. Acad. Sci. U. S. A. [Internet] 97:6960-6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=34370&tool=pmcentrez&rendertype=abstract

Pan Q, Liu YS, Budai-Hadrian O, Sela M, Carmel-Goren L, Zamir D, Fluhr R. 2000. Comparative genetics of nucleotide binding site-leucine rich repeat resistance gene homologues in the genomes of two dicotyledons: Tomato and arabidopsis. Genetics 155:309-322.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, et al. 2009. The Sorghum bicolor genome and the diversification of grasses. Nature [Internet] 457:551-556. Available from: http://www.nature.com/doifinder/10.1038/nature07723

Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD. 1988. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature [Internet] 335:721-726. Available from: http://www.nature.com/doifinder/10.1038/335721a0

Paux E, Sourdille P, Mackay I, Feuillet C. 2012. Sequence-based marker development in wheat: Advances and applications to breeding. Biotechnol. Adv. 30:1071-1088.

Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. PLOS Biol. [Internet] 14:e1002379. Available from: http://dx.plos.org/10.1371/journal.pbio.1002379

Perez CE, Rodrigues FA, Moreira WR, DaMatta FM. 2014. Leaf gas exchange and chlorophyll a fluorescence in wheat plants supplied with silicon and infected with Pyricularia oryzae. Phytopathology [Internet] 104:143-149. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24047250

Peterson DG, Boehm KS, Stack SM. 1997. Isolation of milligram quantities of nuclear DNA from tomato (Lycopersicon esculentum), a plant containing high levels of polyphenolic compounds. Plant Mol. Biol. Report. 15:148-153.

Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. U. S. A. [Internet] 98:9748-53. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=55524&tool=pmcentrez&rendertype=abstract

Pichersky E, Logsdon JM, McGrath JM, Stasys R a. 1991. Fragments of plastid DNA in the nuclear genome of tomato: prevalence, chromosomal location, and possible mechanism of integration. Mol. Gen. Genet. [Internet] 225:453-8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1673221

Pichersky E, Tanksley SD. 1988. Chloroplast DNA sequences integrated into an intron of a tomato nuclear gene. Mol. Gen. Genet. [Internet] 215:65-8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2853831

Piron F, Nicolaï M, Minoïa S, Piednoir E, Moretti A, Salgues A, Zamir D, Caranta C, Bendahmane A. 2010. An induced mutation in tomato eiF4E leads to immunity to two potyviruses. PLoS One 5.

Prober JM, Trainor GL, Dam RJ, Hobbs FW, Charles W, Zagursky RJ, Cocuzza AJ, Jensen MA, Prober JM, et al. 1987. A System for Rapid DNA Sequencing with Fluorescent Chain-Terminating Dideoxynuclcotides. Science (80-.). [Internet] 238:336-341. Available from: http://www.jstor.org/stable/1700686%5Cnhttp://www.jstor.org/stable/1700686?seq=1&cid=pdf-reference#references_tab_contents%5Cnhttp://about.jstor.org/terms

Quadrana L, Almeida J, Asís R, Duffy T, Dominguez PG, Bermúdez L, Conti G, Corrêa da Silva J V., Peralta IE, et al. 2014. Natural occurring epialleles determine vitamin E accumulation in tomato fruits. Nat. Commun. [Internet] 5. Available from: http://www.nature.com/doifinder/10.1038/ncomms5027

Quadrana L, Almeida J, Otaiza SN, Duffy T, da Silva JVC, de Godoy F, Asís R, Bermúdez L, Fernie

AR, et al. 2013. Transcriptional regulation of tocopherol biosynthesis in tomato. Plant Mol. Biol. 81:309-325.

Quadrana L, Rodriguez MC, Lopez M, Bermudez L, Nunes-Nesi A, Fernie AR, Descalzo A, Asis R, Rossi M, et al. 2011. Coupling Virus-Induced Gene Silencing to Exogenous Green Fluorescence Protein Expression Provides a Highly Efficient System for Functional Genomics in Arabidopsis and across All Stages of Tomato Fruit Development. PLANT Physiol. [Internet] 156:1278-1291. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21531899

Richterich P. 1998. Estimation of errors in «Raw» DNA sequences: A validation study. Genome Res. 8:251-259.

Rick CM, Kesicki E, Fobes JF, Holle M. 1976. Genetic and biosystematic studies on two new sibling species of Lycopersicon from interandean Perú. Theor. Appl. Genet. 47:55-68.

Rick CM, Tanksley SD. 1981. Genetic variation in Solanum pennellii: Comparisons with two other sympatric tomato species. Plant Syst. Evol. 139:11-45.

Riebeseel E, Häusler RE, Radchuk R, Meitzel T, Hajirezaei M-R, Emery RJN, Küster H, Nunes-Nesi A, Fernie AR, et al. 2009. The 2-oxoglutarate/malate translocator mediates amino acid and storage protein biosynthesis in pea embryos. Plant J. [Internet] 61:350-363. Available from: http://doi.wiley.com/10.1111/j.1365-313X.2009.04058.x

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, et al. 2012. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat. Genet. 44:217-220.

Roach JC, Boysen C, Wang K, Hood L. 1995. Pairwise end sequencing: A unified approach to genomic mapping and sequencing. Genomics 26:345-353.

Ron M, Dorrity MW, de Lucas M, Toal T, Hernandez RI, Little SA, Maloof JN, Kliebenstein DJ, Brady SM. 2013. Identification of Novel Loci Regulating Interspecific Variation in Root Morphology and Cellular Development in Tomato. PLANT Physiol. [Internet] 162:755-768. Available from: http://www.plantphysiol.org/cgi/doi/10.1104/pp.113.217802

Ronaghi M. 1998. DNA SEQUENCING:A Sequencing Method Based on Real-Time Pyrophosphate. Science (80-.). [Internet] 281:363-365. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.281.5375.363

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. 1996. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. Anal. Biochem. [Internet] 242:84-89. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0003269796904327

Ronaghi M, Uhlén M, Nyrén P. 1998. A sequencing method based on real-time pyrophosphate. Science [Internet] 281:363, 365. Available from:

http://www.sciencemag.org/cgi/doi/10.1126/science.281.5375.363%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/9705713

Rooijen R van, Aarts MGM, Harbinson J. 2015. Natural genetic variation for acclimation of photosynthetic light use efficiency to growth irradiance in Arabidopsis. Plant Physiol. [Internet] 167:1412-29. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4378148&tool=pmcentrez&rendertype=abstract

Rosado-Souza L, Scossa F, Chaves IS, Kleessen S, Salvador LFD, Milagre JC, Finger F, Bhering LL, Sulpice R, et al. 2015. Exploring natural variation of photosynthetic, primary metabolism and growth parameters in a large panel of Capsicum chinense accessions. Planta 242:677-691.

Rossi M, Bermudez L, Carrari F. 2015. Crop yield: Challenges from a metabolic perspective. Curr. Opin. Plant Biol. 25:79-89.

Rothan C, Causse M. 2007. Natural and artificially induced genetic variability in crop and model plant species for plant systems biology. EXS [Internet] 97:21-53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17432262

Rousseaux MC, Jones CM, Adams D, Chetelat R, Bennett A, Powell A. 2005. QTL analysis of fruit antioxidants in tomato using Lycopersicon pennellii introgression lines. Theor. Appl. Genet. 111:1396-1408.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M-A, Barrell B. 2000. Artemis: sequence visualization and annotation. Bioinformatics [Internet] 16:944-945. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/16.10.944

Saiki R, Gelfand D, Stoffel S, Scharf S, Higuchi R, Horn G, Mullis K, Erlich H. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science (80-.). [Internet] 239:487-491. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.2448875

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science (80-.). [Internet] 230:1350-1354. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=2999980

Saito T, Matsukura C, Sugiyama M, Watahiki A, Ohshima I, Iijima Y, Konishi C, Fujii T, Inai S, et al. 2008. Screening for γ-aminobutyric Acid (GABA)-rich Tomato Varieties. J. Japanese Soc. Hortic. Sci. [Internet] 77:242-250. Available from: http://joi.jlc.jst.go.jp/JST.JSTAGE/jjshs1/77.242?from=CrossRef

Salamov AA, Solovyev V V. 2000. Ab initio gene finding in Drosophila genomic DNA. Genome

Res. 10:516-522.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. [Internet] 74:5463-5467. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.74.12.5463

Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. Hum. Mol. Genet. 19.

Van Schalkwyk A, Wenzl P, Smit S, Lopez-Cobollo R, Kilian A, Bishop G, Hefer C, Berger DK. 2012. Bin mapping of tomato diversity array (DArT) markers to genomic regions of Solanum lycopersicum × Solanum pennellii introgression lines. Theor. Appl. Genet. [Internet] 124:947-56. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3284683&tool=pmcentrez&rendertype=abstract

Schatz G, Haslbrunner E, Tuppy H. 1964. Deoxyribonucleic acid associated with yeast mitochondria. Biochem. Biophys. Res. Commun. [Internet] 15:127-132. Available from: http://ac.els-cdn.com/0006291X64903110/1-s2.0-0006291X64903110-main.pdf?_tid=a6affcb0-208d-11e3-8e80-00000aab0f6b&acdnat=1379528031_a49dfbfa99aa45a55eba1facd26558a1

Schatz MC, Delcher ALAL, Salzberg SLSL. 2010. Assembly of large genomes using second-generation sequencing. Genome Res. [Internet] 20:1165. Available from: http://genome.cshlp.org/content/20/9/1165.short

Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, et al. 2006. Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat. Biotechnol. [Internet] 24:447-54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16531992

Schilmiller A, Shi F, Kim J, Charbonneau AL, Holmes D, Daniel Jones A, Last RL. 2010. Mass spectrometry screening reveals widespread diversity in trichome specialized metabolites of tomato chromosomal substitution lines. Plant J. 62:391-403.

Schmidt EW. 1913. Pflanzliche Mitochondrien. Progressus rei Bot. 4:164-183.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature [Internet] 463:178-183. Available from: http://www.nature.com/doifinder/10.1038/nature08670

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, et al. 2009. The B73 maize genome: Complexity, diversity, and dynamics. Science (80-.). 326:1112-1115.

Schneidereit J, Häusler RE, Fiene G, Kaiser WM, Weber APM. 2006. Antisense repression reveals a crucial role of the plastidic 2-oxoglutarate/malate translocator DiT1 at the interface between

carbon and nitrogen metabolism. Plant J. [Internet] 45:206-224. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16367965

Schreiber U, Hormann H, Neubauer C, Klughammer C. 1995. Assessment of Photosystem II Photochemical Quantum Yield by Chlorophyll Fluorescence Quenching Analysis. Funct. Plant Biol. [Internet] 22:209-220. Available from: http://www.publish.csiro.au/paper/PP9950209

Scott NS, Timmis JN. 1984. Homologies between nuclear and plastid DNA in spinach. Theor. Appl. Genet. [Internet] 67:279-288. Available from: http://download.springer.com/static/pdf/47/art%253A10.1007%252FBF00317055.pdf?auth66=1379864976_e2040a6ecdcc19d3df6d453afe297329&ext=.pdf

Seah S, Yaghoobi J, Rossi M, Gleason CA, Williamson VM. 2004. The nematode-resistance gene, Mi-1, is associated with an inverted chromosomal segment in susceptible compared to resistant tomato. Theor. Appl. Genet. 108:1635-1642.

Semel Y, Nissenbaum J, Menda N, Zinder M, Krieger U, Issman N, Pleban T, Lippman Z, Gur A, et al. 2006. Overdominant quantitative trait loci for yield and fitness in tomato. Proc. Natl. Acad. Sci. U. S. A. [Internet] 103:12981-12986. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0604635103

Sequencing Project IRG, Matsumoto T, Wu J, Kanamori H, Katayose Y, Fujisawa M, Namiki N, Mizuno H, Yamamoto K, et al. 2005. The map-based sequence of the rice genome. Nature [Internet] 436:793-800. Available from: http://www.nature.com/doifinder/10.1038/nature03895

Shikanai T, Kaneko H, Nakata S, Harada K, Watanabe K. 1998. Mitochondrial genome structure of a cytoplasmic hybrid between tomato and wild potato. Plant Cell Rep. [Internet] 17:832-836. Available from: http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s002990050493

Shirasawa K, Asamizu E, Fukuoka H, Ohyama A, Sato S, Nakamura Y, Tabata S, Sasamoto S, Wada T, et al. 2010. An interspecific linkage map of SSR and intronic polymorphism markers in tomato. Theor. Appl. Genet. [Internet] 121:731-9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2909429&tool=pmcentrez&rendertype=abstract

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, et al. 2011. The genome of woodland strawberry (Fragaria vesca). Nat. Genet. [Internet] 43:109-116. Available from: http://www.nature.com/doifinder/10.1038/ng.740

Sim S-C, Durstewitz G, Plieske J, Wieseke R, Ganal MW, Van Deynze A, Hamilton JP, Buell CR, Causse M, et al. 2012. Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. PLoS One [Internet] 7:e40563. Available from:

http://dx.plos.org/10.1371/journal.pone.0040563

Simkin AJ, McAusland L, Headland LR, Lawson T, Raines CA. 2015. Multigene manipulation of photosynthetic carbon assimilation increases CO<inf>2</inf> fixation and biomass yield in tobacco. J. Exp. Bot. 66:4075-4090.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJMM, Birol I. 2009. ABySS: A parallel assembler for short read sequence data. Genome Res. [Internet] 19:1117-23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19251739

Smith DR, Crosby K, Lee RW. 2011. Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. Genome Biol. Evol. 3:365-371.

Smith LM, Fung S, Hunkapiller MW, Hunkapiller TJ, Hood LE. 1985. The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: Synthesis of fluorescent DNA primers for use in DNA sequence analysis. Nucleic Acids Res. 13:2399-2412.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. 1986. Fluorescence detection in automated DNA sequence analysis. Nature 321:674-679.

Staden R. 1979. A strategy of DNA sequencing employing computer programs. Nucleic Acids Res. 6:2601-2610.

Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: A web server for gene finding in eukaryotes. Nucleic Acids Res. 32.

Stein LD. 2010. The case for cloud computing in genome informatics. Genome Biol. [Internet] 11:207. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-5-207

Steinhauser M-C, Steinhauser D, Koehl K, Carrari F, Gibon Y, Fernie AR, Stitt M. 2010. Enzyme activity profiles during fruit development in tomato cultivars and Solanum pennellii. Plant Physiol. 153:80-98.

Sugiyama Y, Watase Y, Nagase M, Makita N, Yagura S, Hirai A, Sugiura M. 2005. The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. Mol. Genet. Genomics [Internet] 272:603-15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15583938

Sulpice R, Pyl E-T, Ishihara H, Trenkamp S, Steinfath M, Witucka-Wall H, Gibon Y, Usadel B, Poree F, et al. 2009. Starch as a major integrator in the regulation of plant growth. Proc. Natl. Acad. Sci. [Internet] 106:10348-10353. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0903478106

Sun HJ, Kataoka H, Yano M, Ezura H. 2007. Genetically stable expression of functional miraculin, a new type of alternative sweetener, in transgenic tomato plants. Plant Biotechnol. J. 5:768-

777.

Swerdlow H, Gesteland R. 1990. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. Nucleic Acids Res. 18:1415-1419.

Takayama M, Ezura H. 2015. How and why does tomato accumulate a large amount of GABA in the fruit? Front. Plant Sci. [Internet] 6. Available from: http://journal.frontiersin.org/Article/10.3389/fpls.2015.00612/abstract

Tanksley SD, Ganal MW, Prince JP, de Vicente MC, Bonierbale MW, Broun P, Fulton TM, Giovannoni JJ, Grandillo S, et al. 1992. High density molecular linkage maps of the tomato and potato genomes. Genetics [Internet] 132:1141-60. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1205235&tool=pmcentrez&rendertype=abstract

The Arabidopsis Genome Initiative., The Arabidopsis Genome Initiative, The Arabidopsis Genome Initiative., Arabidopsis Genome Initiative S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, et al. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature [Internet] 408:796-815. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11130711

The International Wheat Genome Sequencing Consortium (IWGSC). 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science [Internet] 345:1251788. Available from: http://www.sciencemag.org/content/345/6194/1250092.abstract

The Mendeley Support Team. 2011. Getting Started with Mendeley. Mendeley Deskt. [Internet]:1-16. Available from: http://www.mendeley.com

The Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature [Internet] 485:635-41. Available from: http://dx.doi.org/10.1038/nature11119%5Cnhttp://www.nature.com/nature/journal/v485/n7400/full/nature11119.html?WT.ec_id=NATURE-20120531%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/22660326%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3378239

Thudi M, Li Y, Jackson SA, May GD, Varshney RK. 2012. Current state-of-art of sequencing technologies for plant genomics research. Brief. Funct. Genomics [Internet] 11:3-11. Available from: https://academic.oup.com/bfg/article-lookup/doi/10.1093/bfgp/elr045

Tian X, Zheng J, Hu S, Yu J. 2006. The rice mitochondrial genomes and their variations. Plant Physiol. [Internet] 140:401-410. Available from: http://www.plantphysiol.org/content/140/2/401.short

Tieman DM, Zeigler M, Schmelz E a, Taylor MG, Bliss P, Kirst M, Klee HJ. 2006. Identification of loci affecting flavour volatile emissions in tomato fruits. J. Exp. Bot. [Internet] 57:887-96.

Available from: http://www.ncbi.nlm.nih.gov/pubmed/16473892

Timmis JN, Ayliffe M a, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. [Internet] 5:123-35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14735123

Tuskan G a., Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, et al. 2006. The genome of black cottonwood, Populus trichocarpa. Science (80-.). [Internet] 313:1596. Available from: http://www.sciencemag.org/cgi/content/abstract/313/5793/1596?ijkey=PZlbhgTWt6pMI&keytype=ref&siteid=sci

Unseld M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides. Nat. Genet. [Internet] 15:57–61. Available from: http://www.nature.com/ng/journal/v15/n1/abs/ng0197-57.html

Usadel B, Schwacke R, Nagel A, Kersten B. 2012. GabiPD – The GABI Primary Database integrates plant proteomic data with gene-centric information. Front. Plant Sci. 3:1-6.

Valentini R, Epron D, Deangelis P, Matteucci G, Dreyer E. 1995. In-Situ Estimation of Net Co2 Assimilation, Photosynthetic Electron Flow and Photorespiration in Turkey Oak (Q-Cerris L) Leaves - Diurnal Cycles under Different Levels of Water-Supply. Plant Cell Environ. [Internet] 18:631-640. Available from: internal-pdf://1995_valentini_pce-1474241034/1995_Valentini_PCE.pdf%5Cn%3CGo to ISI%3E://A1995RD74900003http://onlinelibrary.wiley.com/doi/10.1111/j.1365-3040.1995.tb00564.x/abstract

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MTA, Azam S, Fan G, et al. 2011. Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. Nat. Biotechnol. [Internet] 30:83-89. Available from: http://www.nature.com/doifinder/10.1038/nbt.2022

Varshney RK, Hiremath PJ, Lekha P, Kashiwagi J, Balaji J, Deokar AA, Vadez V, Xiao Y, Srinivasan R, et al. 2009. A comprehensive resource of drought- and salinity- responsive ESTs for gene discovery and marker development in chickpea (Cicer arietinum L.). BMC Genomics [Internet] 10:523. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-10-523

Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, et al. 2013. Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. Nat. Biotechnol. [Internet] 31:240-246. Available from: http://www.nature.com/doifinder/10.1038/nbt.2491

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SKSK, Troggio M, et al. 2010. The genome of the domesticated apple (Malus × domestica

Borkh.). Nat. Genet. [Internet] 42:833-839. Available from: http://www.nature.com/doifinder/10.1038/ng.654

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, et al. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One 2.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, et al. 2001. The sequence of the human genome. Science [Internet] 291:1304-51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11181995

Wallin IE. 1923. The University of Chicago. Am. Nat. [Internet] 57:255-261. Available from: http://www.jstor.org/stable/2456429

Wang XX, Wang HHH, Wang JJJ, Sun R, Wu J, Liu S, Bai Y, Mun J-HH, Bancroft I, et al. 2011. The genome of the mesopolyploid crop species Brassica rapa. Nat. Genet. [Internet] 43:1035-1039. Available from: http://www.nature.com/doifinder/10.1038/ng.919

Watson JD, Crick FHC. 1953. The Structure of DNA. Cold Spring Harb. Symp. Quant. Biol. [Internet] 18:123-131. Available from: http://symposium.cshlp.org/cgi/content/abstract/18/0/123

Weber JL, Myers EW. 1997. Human whole-genome shotgun sequencing. Genome Res. 7:401-409.

Wen W, Liu H, Zhou Y, Jin M, Yang N, Li D, Luo J, Xiao Y, Pan Q, et al. 2016. Combining Quantitative Genetics Approaches with Regulatory Network Analysis to Dissect the Complex Metabolism of the Maize Kernel. Plant Physiol. [Internet] 170:136-146. Available from: http://www.plantphysiol.org/lookup/doi/10.1104/pp.15.01444

Wullschleger SD, Weston DJ, Difazio SP, Tuskan GA. 2013. Revisiting the sequencing of the first tree genome: Populus trichocarpa. Tree Physiol. 33:357-364.

Xia Z, Watanabe S, Yamada T, Tsubokura Y, Nakashima H, Zhai H, Anai T, Sato S, Yamazaki T, et al. 2012. Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. Proc. Natl. Acad. Sci. [Internet] 109:E2155-E2164. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1117982109

Xu Q, Chen L-L, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao W-B, Hao B-H, et al. 2013. The draft genome of sweet orange (Citrus sinensis). Nat Genet [Internet] 45:59-66. Available from: http://view.ncbi.nlm.nih.gov/pubmed/23179022

Xu X, Martin B, Comstock JP, Vision TJ, Tauer CG, Zhao B, Pausch RC, Knapp S. 2008. Fine mapping a QTL for carbon isotope composition in tomato. Theor. Appl. Genet. 117:221-233.

Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, et al. 2011. Genome sequence and analysis of the tuber crop potato. Nature [Internet] 475:189-195. Available from: http://www.nature.com/doifinder/10.1038/nature10158

Yamamoto N, Tsugane T, Watanabe M, Yano K, Maeda F, Kuwata C, Torki M, Ban Y, Nishimura S, et al. 2005. Expressed sequence tags from the laboratory-grown miniature tomato (Lycopersicon esculentum) cultivar Micro-Tom and mining for single nucleotide polymorphisms and insertions/deletions in tomato cultivars. Gene [Internet] 356:127-34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15975739

Yang DL, Jing RL, Chang XP, Li W. 2007. Quantitative trait loci mapping for chlorophyll fluorescence and associated traits in Wheat (Triticum aestivum). J. Integr. Plant Biol. 49:646-654.

Yin Z, Meng F, Song H, He X, Xu X, Yu D. 2010. Mapping quantitative trait loci associated with chlorophyll a fluorescence parameters in soybean (Glycine max (L.) Merr.). Planta 231:875-885.

Yoshida T, Furihata HY, Kawabe A. 2017. Analysis of nuclear mitochondrial DNAs and factors affecting patterns of integration in plant species. Genes Genet. Syst. [Internet] 92:27-33. Available from: https://www.jstage.jst.go.jp/article/ggs/92/1/92_16-00039/_article

Young ND, Tanksley SD. 1989. Restriction fragment length polymorphism maps and the concept of graphical genotypes. Theor. Appl. Genet. [Internet] 77:95-101. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24232480

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, et al. 2002a. A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science [Internet] 296:79-92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11935017

Yu J, Hu S, Wang JJJJ, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y, et al. 2002b. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science [Internet] 296:79-92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11935017

Zamir D, Maloof JN, Foolad M, Zamir D, Barone A, Chiusano M, Ercolano M, Giuliano G, Grandillo S, et al. 2001. Improving plant breeding with exotic genetic libraries. Nat. Rev. Genet. [Internet] 2:983-989. Available from: http://www.nature.com/doifinder/10.1038/35103590

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. [Internet] 18:821-829. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18349386

Zerbino DR, McEwen GK, Margulies EH, Birney E. 2009. Pebble and rock band: Heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. PLoS One 4.

Zhang K, Fang Z, Liang Y, Tian J. 2009. Genetic dissection of chlorophyll content at different growth stages in common wheat. J. Genet. 88:183-189.

Zou J, Rogers WE, Siemann E. 2007. Differences in morphological and physiological traits between native and invasive populations of Sapium sebiferum. Funct. Ecol. 21:721-730.

Tomato IL map: https://solgenomics.net/cview/map.pl?map_id=il6

FAOSTAT: http://www.fao.org/faostat/es/#data

Anexos

Capítulo VIII: Anexos

VIII.1 | Capítulo I

VIII.1.1 | Servidor para Bioinformática

Se realizó la compra y puesta en funcionamiento de un servidor Dell R710 (ver Figura Anexa 1) para brindar servicios de análisis y soporte al grupo:

IP Fija: 172.22.160.24

VPN *FortiClient*: vpn.inta.gov.ar:10443

Sistema Operativo: Linux CentOS 7

***Hardware*:**

- 64GB RAM (DDR3 @ 1333Mhz)
- 24 threads (intel® dual Xeon CPU X5650 @ 2.67Ghz)
- 7TB HDD (RAID 5)
- 8TB Iomega® NAS HDD para *backups* incrementales

Figura Anexa 1: Vista frontal del Servidor Dell PowerEdge R710 *rackeable*. Imagen tomada de: https://www.dell.com/downloads/global/products/pedge/r710-spec-sheet.pdf

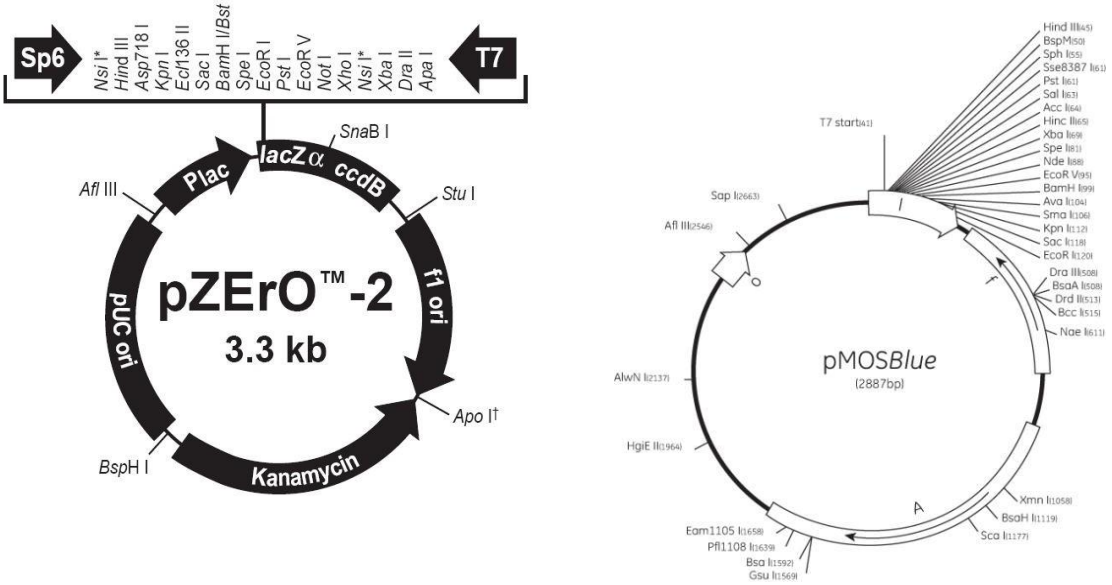


Figura Anexa 2: Plásmidos utilizados para construir las genotecas mitocondriales de *S. lycopersicum* (cv Heinz). pZErO-2 fue utilizada con los fragmentos provenientes de la digestión enzimática y pMOSBlue con los fragmentos obtenidos por digestión mecánica.

Tabla Anexa 1: Anotaciones sobre el genoma de la mitocondria del cv Heinz de *Solanum lycopersicum*

GEN	*CONTIG o SCAFFOLD*	ESTRUCTURA Y POSICIÓN*
Complejo I (NADH deshidrogenasa)		
NAD1	*scaffold 01*	exon1 (36692..37076),
	scaffold 02	exon2 complement (21471..22946)
	contig 40	exon3 complement (3802..3993)
	scaffold 02	exon4 complement (24354..24412)
	contig 27	exon5 (5308..5566)
NAD2	*scaffold 02*	exon1 and 2 complement join (38408..38560,36986..37378)
	scaffold 01	exon3 and 4 join (125519 ..128098), complement (177289..177861)
	contig 454_NT_84825	exon5 (62559..62718)
NAD3	*scaffold 02*	gene (3502..3858)
NAD4	*scaffold 01*	gene complement join (114506..114595,117226..117648,120782..121294,122707..123167)
NAD4L	*contig 27*	gene (6589..6891)
NAD5	*scaffold 01*	exon 1 and 2 join (82149..82376, 83221..84437)
	contig Slmi19004D09	exon3 (581..601)
	scaffold 05	exon 4 (8..304)
	contig 83	exon5 (340..488)
NAD6	*scaffold 03*	gene complement (3268..3915)
NAD7	*scaffold 01*	gene complement(16024..21324)
NAD7	*scaffold 01*	gene complement (16024..21324)
NAD9	*scaffold 02*	gene complement (32694..33266)
Complejo II (Succinato deshidrogenasa)		
SDH3	*scaffold 02*	gene complement (39135..39461)
SDH4	*contig 68*	gene (2045..2412)
Complejo III (Citocromo c reductasa)		
COB	*scaffold 01*	gene complement (105312..106493)

COMPLEX IV		
COX1	*scaffold 01*	gene complement (54036 ..55619)
COX2	*scaffold 01*	gene complement join (145846..146232, 147633..148999)
COX3	*contig 68*	gene (1320..2117)
Complejo IV (Citocromo C oxidasa)		
CCMB	*contig Slmi18005G01*	gene (1..153)
CCMC	*scaffold 04*	gene (91..144)
CCMFC	*contig 4*	gene join (21..679,1629..2179)
CCMFN	*scaffold 01*	gene complement join (8167 ..8182,49679..51047,51071..51402)
Complejo V (ATP sintasa)		
ATP1	*contig 2*	gene (866 a 2392)
ATP4	*contig 27*	gene (7080..7678)
ATP6	*contig 32*	gene (971..1665)
ATP8	*contig 68*	gene (217..685)
ATP9	*contig 7*	gene complement (12840..13073)
Proteínas ribosomales		
RPL2	*contig 7*	gene join (3041..3924,5837..5953)
RPL5	*scaffold 01*	gene complement (108145..108693)
RPL16	*contig 60*	gene (7539..8044)
RPS1	*contig 27*	gene complement (1302..1893)
RPS3	*contig 60*	gene join (4877..4950,6022..7638)
RPS4	*scaffold 03*	gene complement (4607..5656)
RPS10	*scaffold 01*	gene complement join (55782..55895,56671..56919)
RPS19	*contig 60*	gene (4579 ..4863)
RPS13	*scaffold 02*	gene (23277..23628)
RPS12	*scaffold 02*	gene (3907..4284)
RPS14	*scaffold 01*	gene complement (107774..108143)
ARNs Ribosomales		

RRN18	*scaffold 01*	rRNA complement (149582..151525)
RRN5	*scaffold 01*	rRNA complement (149299..149417)
RRN26	*scaffold 02*	rRNA complement (11751..15245)
Otros ORF		
MATR	*contig 64*	gene (1..1846)
ARNt		
TRNI(CAU)	*scaffold 01*	tRNA (4884..4957)
TRNC(GCA)	*scaffold 01*	tRNA (87452..87522)
TRNY(GUA)	*scaffold 01*	tRNA (124918..125000)
TRNQ(UUG)	*scaffold 01*	tRNA complement (65395..65466)
TRNG(GCC)	*scaffold 01*	tRNA complement (68743..68814)
TRNM(CAU)	*scaffold 01*	tRNA (80672..80744)
TRNE(UUC)	*scaffold 01*	tRNA complement (60178..60249)
TRNN(GUU)	*scaffold 01*	tRNA (124311..124382)
TRNH(GUG)	*scaffold 02*	tRNA (28882..28955)
TRNK(UUU)	*scaffold 02*	tRNA (47840..47912)
TRND(GUC)	*contig 18*	tRNA (842..915)
TRNS(GGA)	*contig 18*	tRNA (1603..1689)
TRNS(UGA)	*contig 27*	tRNA complement (12800..12882)
TRNP(UGG)	*contig 52*	tRNA (7555..7628)
TRNW(CCA)	*contig 52*	tRNA (7786..7859)
TRNS(GCU)	*contig 69*	tRNA (3831..3918)
TRNF(GAA)	*contig 69*	tRNA (4281..4354)
TRNP(UGG)	*contig 69*	tRNA (4604..4679)

* el tipo de estructura anotada, seguido y entre paréntesis está la posición (en pb) que contiene a la secuencia codificante sobre el *contig* o *scaffold*.

VIII.3 | Capítulo III

Tabla Anexa 2: Descripción física de las 11 regiones genómicas (*BIN*s) identificadas en este estudio.

BIN	Marcadores (1)		Cromosoma	Especie (2)	Coordenadas (nt) (3)		Tamaño del *BIN* (nt)
	Limite Norte	Limite Sur			Limite Norte	Limite Sur	
1J	SGN-M339	SGN-M179	ch01	M82	88.405.992	90.012.404	1.606.412
				S. pen	107.469.899	109.091.436	1.621.537
2K	SGN-M122	SGN-M46	ch02	M82	46.972.754	48.343.617	1.370.863
				S. pen	56.708.671	58.184.832	1.476.161
2L	SGN-M46	SGN-M100	ch02	M82	48.343.094	49.466.629	1.123.535
				S. pen	58.184.286	59.337.992	1.153.706
4DE	SGN-M169	SGN-M399	ch04	M82	4.835.317	54.473.808	49.638.491
				S. pen	5.052.493	67.132.547	62.080.054
7BC	SGN-M604	SGN-M85	ch07	M82	243.493	3.597.081	3.353.588
				S. pen	474.483	4.036.295	3.561.812
9C	SGN-M82	SGN-M89	ch09	M82	1.908.874	2.780.571	871.697
				S. pen	1.959.093	2.914.799	955.706
9D	SGN-M89	SGN-M8002	ch09	M82	2.779.816	3.442.909	663.093
				S. pen	2.914.068	3.554.228	640.160
9G	SGN-M310	SGN-M193	ch09	M82	60.595.984	61.658.753	1.062.769
				S. pen	76.506.918	77.571.165	1.064.247
11B	SGN-M582	SGN-M121	ch11	M82	2.651.274	3.708.436	1.057.162
				S. pen	2.718.613	3.836.493	1.117.880
11D	SGN-M125	SGN-M565	ch11	M82	48.434.778	48.609.237	174.459
				S. pen	61.071.210	61.271.448	200.238
11H	SGN-M34	SGN-M151	ch11	M82	52.536.512	53.256.293	719.781
				S. pen	65.305.916	66.065.401	759.485

(1) Los códigos de los marcadores son los asignados en el sitio htttp://solgenomics.net.
(2) M82: *Solanum lycopersicum* y *S. pen*: *Solanum pennellii* LA0716.
(3) Las coordenadas corresponden a las versiones del genoma 1.0 de M82 y 2.0 de *S.pen*.

Tabla Anexa 3: Tabla de contingencia de términos *MapMan* sobrerrepresentados de forma significativa según prueba de *Fisher* (p-valor < 0,05) en los 11 *BIN*s seleccionados.

***BIN* M82**	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
1J	29	protein	91	1E-24
1J	27	RNA	56	6E-08
1J	34	transport	46	4E-18
1J	17	hormone metabolism	39	3E-16
1J	35	not assigned	22	2E-03
1J	20	stress	17	2E-02
1J	33	development	13	3E-02
1J	10	cell wall	12	3E-03
1J	13	amino acid metabolism	10	7E-04
1J	3	minor CHO metabolism	6	2E-03
1J	23	nucleotide metabolism	4	3E-02
BIN S. Penn	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
1J	29	protein	96	6E-27
1J	27	RNA	58	3E-08
1J	34	transport	49	7E-20
1J	17	hormone metabolism	43	7E-19
1J	26	misc	22	4E-02
1J	35	not assigned	20	2E-04
1J	20	stress	18	1E-02
1J	10	cell wall	11	2E-02
1J	13	amino acid metabolism	10	9E-04
1J	3	minor CHO metabolism	6	2E-03
1J	23	nucleotide metabolism	5	7E-03
1J	22	polyamine metabolism	2	1E-02
***BIN* M82**	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
2K	29	protein	33	5E-03
2K	27	RNA	30	8E-03
2K	30	signalling	20	6E-05
2K	26	misc	17	2E-02
2K	34	transport	14	1E-02
2K	33	development	12	3E-03
2K	17	hormone metabolism	12	1E-02
2K	13	amino acid metabolism	10	3E-05

2K	10	cell wall	7	5E-02
2K	9	mitochondrial electron transport / ATP synthesis	5	3E-03
2K	14	S-assimilation	2	7E-04
2K	5	fermentation	2	1E-02
BIN S. Penn	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
2K	29	protein	36	4E-03
2K	27	RNA	33	4E-03
2K	30	signalling	20	2E-04
2K	26	misc	17	3E-02
2K	17	hormone metabolism	15	1E-03
2K	34	transport	15	1E-02
2K	33	development	11	2E-02
2K	13	amino acid metabolism	10	7E-05
2K	9	mitochondrial electron transport / ATP synthesis	5	4E-03
2K	3	minor CHO metabolism	4	1E-02
2K	14	S-assimilation	2	8E-04
2K	5	fermentation	2	1E-02
***BIN* M82**	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
2L	29	protein	38	1E-05
2L	27	RNA	30	1E-03
2L	30	signalling	15	4E-03
2L	31	cell	13	1E-04
2L	16	secondary metabolism	10	2E-03
2L	28	DNA	9	4E-02
2L	23	nucleotide metabolism	4	6E-03
2L	2	major CHO metabolism	4	2E-02
2L	24	biodegradation of xenobiotics	2	1E-02
2L	12	N-metabolism	2	4E-02
BIN S. Penn	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
2L	29	protein	40	5E-06
2L	27	RNA	31	1E-03
2L	26	misc	16	2E-02
2L	30	signalling	15	6E-03
2L	31	cell	12	6E-04
2L	28	DNA	9	5E-02

2L	16	secondary metabolism	8	2E-02
2L	23	nucleotide metabolism	4	7E-03
2L	2	major CHO metabolism	4	3E-02
2L	12	N-metabolism	2	4E-02
***BIN* M82**	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
4DE	35	not assigned	416	9E-43
4DE	29	protein	317	1E-29
4DE	12	N-metabolism	27	6E-19
4DE	28	DNA	90	5E-13
4DE	31	cell	80	9E-12
4DE	26	misc	129	3E-08
4DE	17	hormone metabolism	80	2E-06
4DE	20	stress	89	2E-06
4DE	27	RNA	205	9E-06
4DE	16	secondary metabolism	52	1E-04
4DE	3	minor CHO metabolism	18	4E-04
4DE	1	PS	7	2E-03
4DE	4	glycolysis	10	9E-03
4DE	34	transport	75	2E-02
4DE	25	C1-metabolism	4	3E-02
4DE	7	OPP	5	4E-02
4DE	30	signalling	78	5E-02
BIN S. Penn	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
4DE	29	protein	392	4E-46
4DE	12	N-metabolism	49	1E-44
4DE	35	not assigned	458	2E-43
4DE	28	DNA	126	2E-25
4DE	31	cell	101	1E-17
4DE	26	misc	163	4E-14
4DE	17	hormone metabolism	110	2E-13
4DE	27	RNA	266	3E-13
4DE	20	stress	121	4E-13
4DE	3	minor CHO metabolism	24	3E-06
4DE	16	secondary metabolism	61	1E-05
4DE	4	glycolysis	13	9E-04
4DE	34	transport	88	4E-03
4DE	1	PS	10	6E-03

4DE	7	OPP	6	2E-02
4DE	25	C1-metabolism	4	5E-02
***BIN* M82**	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
7BC	29	protein	92	5E-16
7BC	34	transport	41	1E-10
7BC	26	misc	38	1E-05
7BC	27	RNA	60	3E-05
7BC	3	minor CHO metabolism	9	5E-05
7BC	30	signalling	26	4E-03
7BC	20	stress	23	5E-03
7BC	33	development	17	2E-02
7BC	2	major CHO metabolism	6	3E-02
7BC	11	lipid metabolism	10	4E-02
7BC	19	tetrapyrrole synthesis	4	4E-02
7BC	1	PS	9	5E-02
BIN S. Penn	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
7BC	29	protein	87	9E-14
7BC	34	transport	43	8E-12
7BC	27	RNA	61	1E-05
7BC	26	misc	36	7E-05
7BC	3	minor CHO metabolism	8	3E-04
7BC	19	tetrapyrrole synthesis	6	2E-03
7BC	30	signalling	26	4E-03
7BC	33	development	17	2E-02
7BC	20	stress	20	3E-02
7BC	16	secondary metabolism	13	3E-02
***BIN* M82**	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
9C	27	RNA	20	2E-03
9C	4	glycolysis	3	3E-03
9C	18	Co-factor and vitamine metabolism	3	3E-03
9C	11	lipid metabolism	5	1E-02
9C	26	misc	11	2E-02
9C	17	hormone metabolism	7	3E-02
BIN S. Penn	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
9C	27	RNA	21	9E-04
9C	4	glycolysis	3	3E-03

9C	17	hormone metabolism	8	1E-02
9C	11	lipid metabolism	5	1E-02
9C	18	Co-factor and vitamine metabolism	2	3E-02
9C	26	misc	10	4E-02
***BIN* M82**	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
9D	29	protein	15	1E-02
9D	15	metal handling	2	1E-02
9D	21	redox	3	2E-02
9D	27	RNA	13	3E-02
9D	13	amino acid metabolism	3	4E-02
BIN S. Penn	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
9D	8	TCA / organic transformation	2	1E-02
9D	15	metal handling	2	1E-02
9D	27	RNA	14	2E-02
9D	29	protein	15	2E-02
9D	21	redox	3	2E-02
9D	13	amino acid metabolism	3	5E-02
***BIN* M82**	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
9G	29	protein	25	5E-05
9G	30	signalling	11	2E-03
9G	34	transport	10	4E-03
9G	26	misc	12	4E-03
9G	2	major CHO metabolism	3	3E-02
9G	21	redox	3	4E-02
BIN S. Penn	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
9G	29	protein	28	2E-06
9G	34	transport	11	1E-03
9G	30	signalling	11	3E-03
9G	2	major CHO metabolism	4	4E-03
9G	26	misc	12	5E-03
9G	27	RNA	18	2E-02
9G	20	stress	8	2E-02
9G	21	redox	3	5E-02
***BIN* M82**	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
11B	27	RNA	31	4E-06

11B	34	transport	15	1E-04
11B	29	protein	29	3E-04
11B	26	misc	16	1E-03
11B	33	development	9	9E-03
11B	28	DNA	8	2E-02
11B	11	lipid metabolism	5	4E-02
BIN S. Penn	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
11B	27	RNA	31	2E-05
11B	34	transport	17	2E-05
11B	29	protein	30	3E-04
11B	26	misc	17	7E-04
11B	13	amino acid metabolism	5	2E-02
11B	28	DNA	8	3E-02
11B	33	development	8	3E-02
11B	35	not assigned	12	5E-02
11B	11	lipid metabolism	5	5E-02
***BIN* M82**	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
11D	23	nucleotide metabolism	3	1E-04
11D	15	metal handling	2	1E-03
BIN S. Penn	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
11D	23	nucleotide metabolism	3	1E-04
11D	27	RNA	6	2E-02
***BIN* M82**	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
11H	27	RNA	20	1E-04
11H	31	cell	7	3E-03
11H	1	PS	5	5E-03
11H	29	protein	18	7E-03
11H	11	lipid metabolism	5	7E-03
11H	34	transport	8	1E-02
11H	30	signalling	8	2E-02
BIN S. Penn	**Código *MapMan***	**Término *MapMan***	**Cantidad de *loci***	**P-VALOR < 0,05**
11H	27	RNA	21	2E-04
11H	29	protein	18	2E-02
11H	31	cell	10	4E-05
11H	34	transport	9	7E-03

11H	30	signalling	8	3E-02
11H	1	PS	5	6E-03
11H	11	lipid metabolism	5	1E-02
11H	15	metal handling	2	2E-02