



UNIVERSIDAD NACIONAL DEL SUR

TESIS DE DOCTOR EN AGRONOMÍA

“Identificación de genes vinculados al modo reproductivo por medio de la secuenciación genómica de distintos genotipos de *Eragrostis curvula*”

Ing. Agr. José Carballo

Bahía Blanca

Argentina

2021

## **Prefacio**

Esta Tesis se presenta como parte de los requisitos para optar al grado Académico de Doctor en Agronomía, de la Universidad Nacional del Sur y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otra. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el ámbito del Departamento de Agronomía durante el período comprendido entre el 4 de abril del 2017 (Res. N°454) y la fecha actual, bajo la dirección de la Dra. Ingrid Garbus y la co-dirección del Dr. Mario Caccamo.

José Carballo



**UNIVERSIDAD NACIONAL DEL SUR**  
Secretaría General de Posgrado y Educación Continua

La presente tesis ha sido aprobada el **26/03/2021**, mereciendo la calificación de **.10.(.....Diez.....)**

## Agradecimientos

A los organismos que me brindaron el respaldo económico para que pueda desarrollar mi doctorado, en la primera etapa la Agencia Nacional de Promoción Científica y Tecnológica, y para la finalización de mi doctorado El Consejo Nacional de Investigaciones Científicas y Técnicas.

A mis directores, Ingrid Garbus y Mario Caccamo. Gracias Ingrid por guiarme durante todo este tiempo y mostrarme desde el primer día como debida recorrer este camino si quería seguir progresando. También quiero agradecer a Mario por abrirme las puertas del instituto NIAB y por confiar en mí para incluirme en distintos proyectos.

Quiere agradecerle a Viviana Echenique por haber confiado en mí desde el primer día que entre a su oficina y por todo su apoyo en este tiempo. Los que trabajamos en el grupo de Viviana conocemos muy bien su enorme capacidad para ayudarnos a resolver los problemas que nos surgen y lo importante que son los recursos que genera para cumplir nuestros objetivos personales y los objetivos del grupo.

Gracias a los tres, Ingrid, Mario y Viviana, por todo el aporte académico que me brindaron en todo este tiempo, ayudando a mi formación y a que pueda concretar esta Tesis. Gran parte de los conocimientos y capacidades que adquirí desde que comencé mi doctorado fueron sin duda fruto de sus contribuciones.

Les agradezco a todos los integrantes del laboratorio de Biotecnología de CERZOS por haberme ayudado a concretar este trabajo y por todos los gratos momentos que pasamos.

A todos los colaboradores Nacionales e Internacionales.

A mis padres y hermanos por todo el apoyo y la formación que me dieron durante toda mi carrera y también al resto de mi familia por siempre darme una mano cuando lo necesite.

A Graciela, Lizi, Juan, Luz y Marcos por acompañarme en todo este recorrido.

A mis amigos por la ayuda incondicional.

Finalmente quiero agradecer a mi familia, Gachi y Clarita, sin su apoyo no podría haber concretado este objetivo, no alcanzan las palabras para decirles lo importantes que son y que fueron para mí durante todo este tiempo.

Gracias!

José

## Resumen

*Eragrostis curvula* es una gramínea estudiada por su tolerancia al estrés por sequía y por tener la capacidad de reproducirse por apomixis, la cual es definida como un mecanismo que permite la clonación natural por medio de semillas resultando la descendencia genéticamente idéntica a la planta madre. *E. curvula* tiene un número básico de cromosomas de 10 y contiene genotipos con distintos niveles de ploidias. Una de las características la especie es que los genotipos diploides son siempre sexuales mientras que los poliploides son generalmente apomícticos aunque existen tetraploides sexuales. Desde hace décadas la apomixis ha sido estudiada con el fin de identificar las bases moleculares y genéticas para ser transferidas a otros cultivos de importancia económica. Las técnicas de secuenciación de ADN son una herramienta sumamente útil para descubrir genes y regiones asociadas a distintos caracteres biológicos, así como para realizar análisis filogenéticos y de sintenia entre especies. En el trabajo de esta tesis se propuso obtener secuencias genómicas de *E. curvula* utilizando tecnologías de última generación con el fin de identificar los genes relacionados a la apomixis y otros genes de interés como los de la tolerancia al estrés por sequía y la calidad forrajera.

El primer genoma de la especie secuenciado fue el genotipo diploide sexual Victoria que fue originado del tetraploide apomíctico Tanganyika INTA. Las plataformas utilizadas para su secuenciación fueron PacBio, Chicago y Hi-C y fue ensamblado a través del software FALCON con el que se obtuvo un N50 de ~43 mega pares de bases (Mb) y 1.143 contigs. Los genes anotados fueron 56.469 los cuales fueron utilizados para determinar la historia evolutiva de la especie. Además se caracterizaron las secuencias de los genes de las vías de la lignina relacionados con la calidad forrajera y los genes de la familia WRKY, vinculados a la tolerancia al estrés por sequía. Por otro lado se encontraron genes de *E. curvula* ortólogos a los expresados en etapas específicas de la vía sexual de *Oryza sativa*.

Los genomas tetraploides apomícticos secuenciados fueron el del cultivar Tanganyika INTA obtenido mediante la plataforma Illumina y el de Don Walter combinando las tecnologías Chromium 10X y Oxford Nanopore. El ensamblado de Tanganyika INTA con el software Masurca resultó en un N50 de 4.715 pb y 293.300 contigs, mientras que en el genoma de Don Walter ensamblado con Supernova+DBG2OLC se obtuvo un N50 de 224.390 pb y 7.542 contigs. Por medio de un análisis comparativo se pudo identificar la región vinculada a la apomixis en todos los genomas en la cual se encontraron genes que podrían tener un rol activo en la regulación del modo reproductivo de acuerdo con la literatura.

A través de esta tesis se pudieron secuenciar, ensamblar y anotar tres genomas de *E. curvula* lo cual representa un avance significativo en la caracterización de la especie. La identificación de regiones asociadas a la apomixis y sus posibilidades de expresión y regulación permitirá, en una instancia futura, manipular el carácter e intentar la introgresión del mismo en especies sexuales de interés agrícola por medio de ingeniería genética.

## **Abstract**

*Eragrostis curvula* is forage grass studied for its tolerance to draught stress and because its characteristic reproductive pathway, the apomixis. Apomixis is an asexual way of reproduction by seeds in which the progeny result genetically identical to the mother plant. The basic number of chromosomes of *E. curvula* is 10 and has different ploidy levels being the diploids always sexual while the polyploids are mainly apomictic despite there are fully sexual tetraploids. Apomixis has been studied in order to identify its molecular and genetic basis to transfer this trait to economically import crops. The DNA sequencing is a powerful tool to detect genes and/or regions associated to different traits and disclose the evolutionary history of the species. In this Thesis the latest generation technologies have been used to sequencing and assembly the *E. curvula* genome to identify the genes related to apomixis and other interesting traits like draught tolerance and forage quality.

The first sequenced genome of the species was the diploid sexual genotype Victoria originated from the apomictic tetraploid Tanganyika INTA. The technologies used for the sequencing were a combination of PacBio, Chicago and Hi-C and the assembler with best performance was FALCON. The final N50 obtained here was ~43 Mb and the number of contigs was 1,143. The 56,469 genes annotated in this genome were used to unravel the evolutionary history of *E. curvula*. Furthermore we identify genes in the lignin pathway associated to the forage quality and genes of the WRKY family which are related to al multiple abiotic stress like drought stress. The genes related to the sexual pathway were evaluated through a comparative analysis with specific development stages in *Oryza sativa*.

The tetraploids genomes of *E. curvula* sequenced here were derived from Tanganyika INTA, sequenced with Illumina, and Don Walter using a combination of Chromium 10x and Oxford Nanopore. The Tanganyika INTA assembly was performed with Masurca software obtaining an N50 of 4,715 bp and 293,000 contigs, while Don Walter was assembled with Supernova+DBG2OLC rendering an N50 of 224,390 bp and 7,542 contigs. Interestingly a region linked to apomixis was found through a comparative analysis between the genomes

of *E. curvula* that could have an active role in the regulation of the reproductive pathway based on the literature.

In this Thesis were sequenced, assembled and annotated three genomes of *E. curvula* representing a significant step forward in terms of characterisation of the species. The identification of the regions linked to apomixis, its expression and regulation could allow handling the trait and, in future instances, to introduce apomixis to sexual economically important crops.

**Publicación derivada de esta Tesis:**

Carballo J, Santos BA, Zappacosta D, Garbus I, Selva JP, Gallo CA, Díaz A, Albertini E, Caccamo M, Echenique V. A high-quality genome of *Eragrostis curvula* grass provides insights into Poaceae evolution and supports new strategies to enhance forage quality. *Scientific Reports*. 2019 Jul 15;9(1):1-5.

**Otras publicaciones logradas en este periodo asociadas a mi formación en bioinformática**

Garbus I, Romero JR, Selva JP, Pasten MC, Chinestra C, Carballo J, Zappacosta DC, Echenique V. De novo transcriptome sequencing and assembly from apomictic and sexual *Eragrostis curvula* genotypes. *PloS one*. 2017 Nov 1;12(11):e0185595.

Zappacosta DC, Gallardo JA, Carballo J, Meier MS, Rodrigo JM, Gallo CA, Selva JP, Stein J, Ortiz JP, Albertini E, Echenique VC. A high-density linkage map of the forage grass *Eragrostis curvula* and localization of the diplospory locus. *Frontiers in plant science*. 2019;10:918.

Garbus I, Selva JP, Pasten MC, Bellido AM, Carballo J, Albertini E, Echenique V. Characterization and discovery of miRNA and miRNA targets from apomictic and sexual genotypes of *Eragrostis curvula*. *BMC genomics*. 2019 Dec 1;20(1):839.

Selva JP, Zappacosta D, Carballo J, Rodrigo JM, Bellido A, Gallo CA, Gallardo J, Echenique V. Genes Modulating the Increase in Sexuality in the Facultative Diplosporous Grass *Eragrostis curvula* under Water Stress Conditions. *Genes*. 2020 Sep;11(9):969.

**Esta tesis se realizó en el marco de los siguientes proyectos:**

PICT Raíces 2017-0879. Genómica estructural para acceder a la región condicionante de la apomixis en *Eragrostis curvula*. IR: Dra. Viviana Echenique (CERZOS) y Dr. Mario Caccamo (NIAB, UK). 1.080.000 pesos.

Subsidio CONICET Proyectos de Investigación Plurianuales (PIP) 2015 - 2017. “Evaluación del rol de los pequeños RNAs en la regulación epigenética de la apomixis diplospórica en pasto llorón (*Eragrostis curvula*).” Código de proyecto: 11220150100963 CO. Director: Ingrid Garbus: Co-director: Viviana Echenique. Monto: \$450.000. Fecha de inicio: 01/01/2016

H2020-MSCA-RISE-2014. ID 645674. Plant Reproduction for Crop Improvement (PROCROP). Research and Innovation Staff Exchange (RISE) H2020-MSCA-RISE-2014. Proyecto de Cooperación Internacional Financiado por la UE. IR: Emidio Albertini, UNIPG, Perugia, Italia. Nodos Participantes: Italia (UNIPG, UNIMIL, Polo GGB), Francia (IRD), Argentina (UNR/ICAR, CERZOS). Monto total proyecto 738.000 euros. Nodo CERZOS: IR Viviana Echenique. 128.000 euros.

Subsidio de la Agencia Nacional de Promoción Científica y Tecnológica PICT-2013-2019. “Identificación de pequeños RNAs y su participación en la regulación epigenética de la apomixis diplospórica en pasto llorón (*E. curvula*)”. Resolución ANPCyT N° 214/14. Director: Ingrid Garbus Monto: \$191.730. Fecha de inicio: 10/10/2014. PIP

H2020-MSCA-RISE-2015.ID: 691109. Exploring the molecular control of seed yield in crops (EXPOSEED). Proyecto de Cooperación Internacional Financiado por la UE. IR: Raffaella Battaglia, CREA ES Fiorenzuola. Nodos Participantes: Italia (CREA, UNIMIL), España (SCIC), Alemania (KWS Lochow GMBH), Holanda (Keygene NV), Australia (University of Adelaide), México (CINVESTAV, Argentina (CERZOS). Monto total proyecto: 798.000 euros. Nodo CERZOS: IR Viviana Echenique. 63.000 euros.

PICT Raíces 2014 - 1243. Secuenciación del genoma de *Eragrostis curvula* a fin de identificar genes relacionados con el modo reproductivo y calidad del forraje. IR: Dra. Viviana Echenique (CERZOS) y Dr. Mario Caccamo (TGAC, UK) 525.000 pesos.

## Índice

Prefacio .....	2
Agradecimientos.....	3
Resumen .....	4
Abstract.....	5
Introducción general .....	11
<i>Eragrostis curvula</i> (Pasto llorón) .....	11
Apomixis.....	14
Origen natural de la apomixis .....	17
Herencia de la apomixis y enfoques para su estudio .....	19
Importancia de la Apomixis.....	22
Secuenciación genómica .....	24
Recursos genómicos de <i>Eragrostis</i> .....	28
Objetivo general.....	31
Los objetivos específicos de esta tesis son los siguientes.....	32
Hipótesis .....	33
Capítulo 1 Secuenciación y ensamblado del genoma diploide sexual de <i>Eragrostis curvula</i> .....	34
Introducción .....	34
Materiales y métodos.....	37
Material vegetal.....	37
Extracción de ADN .....	37
Preparación de las bibliotecas.....	38
Ensamblado .....	38
Validación por DArT-seq .....	41
Anotación de elementos repetitivos .....	42
Anotación de genes .....	42
Análisis de sintenia evolución de <i>E. curvula</i> .....	43
Identificación de genes relacionados con la tolerancia al estrés por sequía .....	45
Identificación de genes relacionados con la calidad forrajera .....	45
Identificación de genes relacionados a las vías reproductivas.....	46
Resultados.....	47
Ensamblado y validación del genomas de Victoria.....	47
Anotación de elementos repetitivos y genes .....	51
Evolución de <i>E. curvula</i> dentro de la familia de las Poaceae .....	54

Relación filogenética de los genotipos de <i>E. curvula</i> .....	61
Genes relacionados a las vías de las ligninas.....	62
Factores de transcripción WRKY en Victoria .....	64
Genes relacionados a la reproducción.....	66
Discusión .....	74
Capítulo 2 : Secuenciación y ensamblado de los genotipos apomícticos de <i>E. curvula</i> Don Walter y Tanganyika INTA.....	78
Introducción .....	78
Materiales y métodos.....	80
Material Vegetal y extracción.....	80
Secuenciación.....	80
Citometría de Flujo .....	81
Ensamblado.....	81
Anotación de elementos repetitivos .....	82
Anotación de genes .....	82
Análisis filogenético de la subfamilia Chloridoideae .....	83
Análisis comparativo de los cultivares de <i>E. curvula</i> y la región ligada a la apomixis .....	83
Resultados.....	85
Tamaño genómico y ploidía .....	85
Secuenciación y ensamblado .....	86
Anotación de elementos repetitivos .....	89
Anotación de genes .....	90
Análisis filogenético de la subfamilia Chloridoideae .....	92
Análisis comparativo de los cultivares de <i>E. curvula</i> y la región ligada a la apomixis .....	93
Discusión .....	100
Conclusión general .....	105
Referencias.....	106
ANEXO.....	129

## Introducción general

### *Eragrostis curvula* (Pasto llorón)

Dentro de las familias de las Poaceae (también conocida como gramíneas), subfamilia Chloridoideae, tribu Eragrostideae se encuentra el género *Eragrostis*. Este género contiene el 67% de la tribu contando con 350 especies (Peterson et al., 1997). El centro de distribución geográfica se encuentra en las sabanas del sur de África y su domesticación ha llevado a que hoy en día se encuentre ampliamente distribuido por todo el mundo, encontrándose en el continente americano alrededor de 120 especies (Phillips, 1982). Este género se caracterizan por contener plantas anuales o perennes con morfología C4, inflorescencia en panícula, ligua con pelos o raramente membranosa. Muchas de las especies de este género tienen importancia económica como forrajera y solo *Eragrostis tef* es utilizada para consumo humano en su región de origen, Etiopía.

Dentro del género *Eragrostis* las especies con mayor distribución e importancia económica son *E. tef* y *E. curvula* (Figura 1.1). *E. tef* es la especie más cultivada en Etiopía y la tercera en términos de producción de grano. Además de ser una pastura también es utilizada para el consumo directo de su grano y su harina es el principal ingrediente de los panes utilizados en la región occidental de África (Bultosa, 2007).

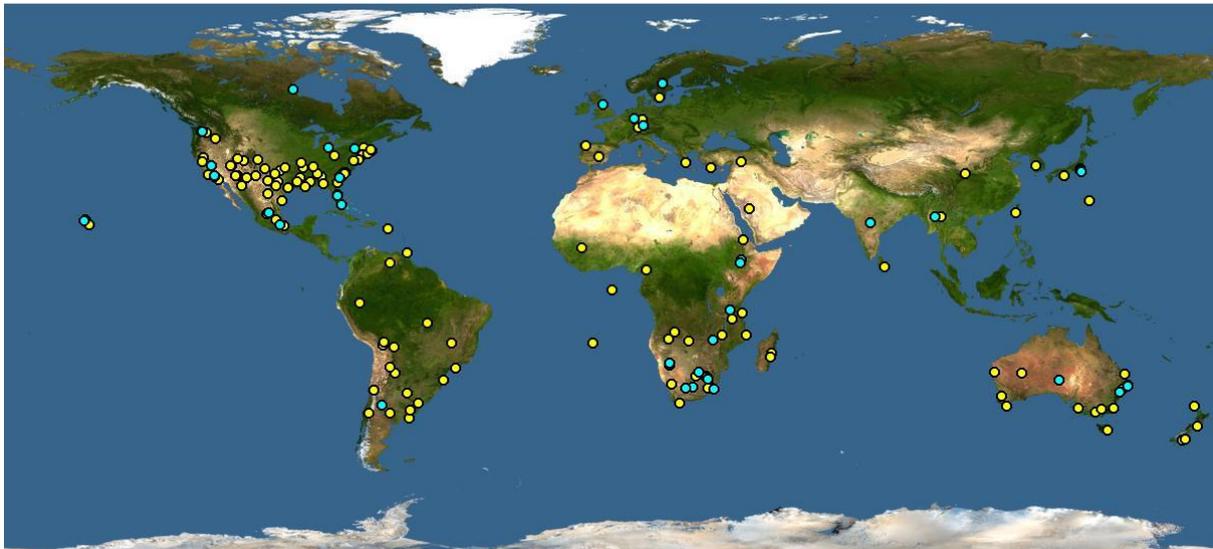


Figura 1.1: Mapa mundial con la distribución geográfica de *E. tef* (círculos celestes) y *E. curvula* (círculos amarillos). Se puede observar como de la región Africana se extendió hacia todo el mundo. Fuente: <https://www.discoverlife.org/>

*E. curvula* tiene crecimiento primavero-estival y es utilizada como forraje principalmente en Australia, Estados Unidos, Sudáfrica y Argentina. Esta especie C4 tiene mayor eficiencia fotosintética que las gramíneas C3, esto significa que con la misma intensidad de luz la tasa fotosintética es mayor. Dentro de sus características morfológicas se destacan su lígula pilosa y aurícula vestigial, inflorescencia en panoja laxa de 15-40 cm de largo, semilla muy pequeña 0,5 mm, con ramificaciones distantes y una altura de hasta 1.5 m (Figura 1.2).

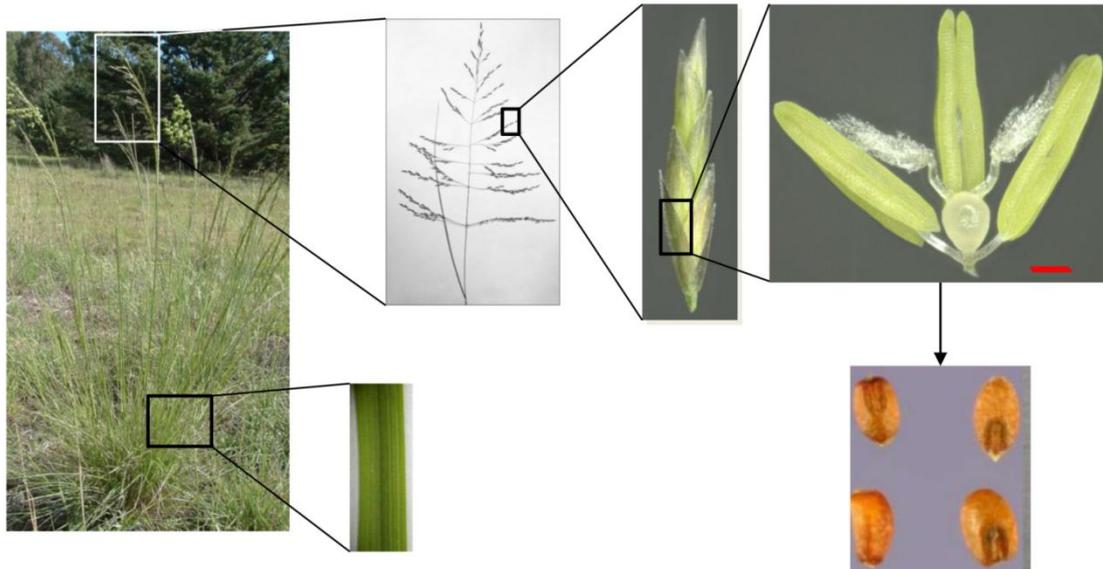


Figura 1.2: Órganos vegetativo y reproductivo de *E. curvula*. En el recuadro blanco se enfoca una panoja, que es ampliada en una espiguilla, luego se observan los órganos reproductivos (anteras y pistilo) y finalmente la semilla madura. En el recuadro negro se observa la hoja.

Antiguamente la clasificación de las especies se realizaba por caracteres morfológicos ya que no existían las herramientas moleculares actuales. Por la numerosa cantidad de variedades que se encuentran dentro de la especie *E. curvula*, en 1961 se propuso agruparlas en 5 grupos de acuerdo a sus características morfológicas, *curvula*, *robusta*, *robusta verde*, *robusta intermedia* y *chloromelas* (Leigh, 1961). Hoy esta clasificación ha entrado en desuso y otras herramientas como marcadores SSR se utilizan para clasificarla más precisamente (Zappacosta et al. 2011).

Su capacidad para desarrollarse en regiones con suelos arenosos y régimen hídrico bajo la posicionó como la primera alternativa para las zonas marginales de Argentina. Es así que en el último censo agropecuario nacional en 2018 del INDEC (<https://cna2018.indec.gob.ar/>) ocupó el quinto puesto, detrás de otras pasturas no aptas para condiciones marginales como la alfalfa. Las provincias de mayor producción de pasto llorón son San Luis, La Pampa, Buenos Aires y Córdoba y entre todas ellas suman el 94% (252.105,59 Has.) del total del pasto llorón implantado en el país (Figura 1.3). En los últimos 16 años la superficie

de este pasto ha decrecido de 786.749 a 268.966 Has. Al igual que la superficie de todas las pasturas perennes del país, este decrecimiento fue debido al cambio del uso de la tierra pasando de la ganadería a la agricultura.

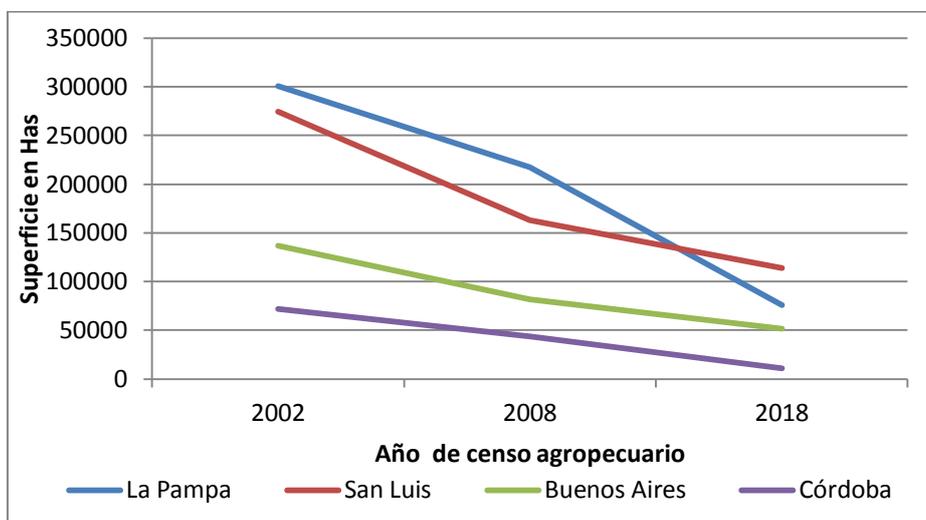


Figura 1.3: Superficie implantada con pasto llorón en los años 2002, 2008 y 2018 en las provincias de La Pampa, San Luis, Buenos Aires y Córdoba. Se puede ver como va disminuyendo su superficie por el cambio en el uso de la tierra.

*E. curvula*, al igual que todo el género *Eragrostis*, tiene un número básico de cromosomas de 10. Dentro de la especie se encuentran individuos con diferente ploidía (2X, 4X, 6X, 7X y 8X) siendo los más frecuentes los tetraploides. La mayor colección de cultivares en Argentina se encuentra en CERZOS CONICET la cual cuenta con materiales de distintas ploidía y modos reproductivos disponibles para el estudio de esta especie (Tabla 1.1).

Cultivar	Ploidía	Modo Reproductivo
Victoria	2X	Sexual
PI299920	2X	Sexual
PI208214	2X	Sexual
PI219919	2X	Sexual
PI219928	2X	Sexual
Tanganyika USDA	4X	apomíctico obligado
Tanganyika INTA	4X	apomíctico facultativo
Ermelo	4X	apomíctico facultativo
Morpa	4X	apomíctico facultativo
OTA-S	4X	Sexual
Catalina	4X	No caracterizado
Don Walter	4X	apomíctico facultativo
Don Eduardo	6X	apomíctico obligado

Don Luis	6X	No caracterizado
Don Pablo	7X	apomíctico obligado
Kromdraai	6X	apomíctico facultativo
Don Juan	8X	apomíctico obligado

Tabla 1.1: Cultivares con diferentes ploidias y modos reproductivos disponibles en el CERZOS CONICET para el estudio del pasto llorón

## Apomixis

La palabra apomixis en Griego (apo= ausencia mixis=mezcla) hace referencia a la falta de recombinación. Fue presentada por Winkler (Winkler, 1908) y su primera definición fue “apomixis es la sustitución de la reproducción sexual por otro tipo de reproducción asexual que no involucra la fusión celular o nuclear (i.e. fertilización)”. En la actualidad la definición más utilizada es la que define apomixis como “proceso que permite la clonación natural por medio de semillas resultando la descendencia genéticamente idéntica a la planta madre” (Nogler 1984, Grossniklaus 1998).

La apomixis en angiospermas fue clasificada de la siguiente manera (Fagerlind, 1940<sup>a</sup> y luego adaptado por Maheshwari 1950):

**Apomixis no recurrente:** Ocurre meiosis en la célula madre de la megáspora y se produce un embrión haploide. El nuevo embrión puede ser formado por la célula huevo o por otra célula del gametofito. La planta resultante es haploide y tiene la mitad de los cromosomas.

**Apomixis recurrente:** No se completa la meiosis. El embrión producido tiene la misma cantidad de cromosomas que la célula que le dio origen y surge de la célula madre de la megáspora o de una célula adyacente.

**Apomixis adventicia o embrionaria:** El embrión surge de una célula somática de la nucela o del tegumento interno.

**Apomixis vegetativa:** Son llamados apomícticas vegetativas aquellas especies que producen bulbos vegetativos u otros propágulo en el lugar de flores o inflorescencias. Algunos ejemplos son *Polygonum viviparum* y *Agave sp.*

Las definiciones modernas de apomixis no contemplan la reproducción vegetativa como apomícticas y solo tiene en cuenta la reproducción por semillas, clasificando los distintos tipos de apomixis de la siguiente manera (Nogler, 1984, Grossniklaus, 1998):

1. **Esporoítica:** El nuevo embrión prolifera directamente de una célula somática de la nucela o del tegumento interno del ovulo.

2. **Gametofítica:** El embrión se desarrolla a partir de células no reducidas.

2.1 Diplospórica: El embrión se origina directamente desde la célula madre de la megáspora.

2.2 Apospórica: El embrión se origina de una célula somática adyacente a la célula madre de la megáspora.

Tres componentes son fundamentales para que se produzca apomixis de acuerdo a esta última clasificación:

-Apomeiosis: ausencia de meiosis en la célula que da origen a la célula huevo, por lo tanto no existe reducción ni recombinación celular.

-Partenogénesis: Desarrollo de un embrión sin fertilización.

-Pseudogamia: Fertilización solo del endosperma. Aunque es raro existe en algunas especies la formación autónoma del endosperma como el caso de *Hieracium praealtum*.

En un ciclo reproductivo sexual de las angiospermas al comienzo de la megasporogénesis (formación de la megáspora) la arqueospora se convierte en la célula madre de la megáspora la cual se recombina y multiplica en dos células (diada), y se reduce formando cuatro células (tétrada) por medio de la meiosis (Figura 1.4 A). Tres células de las tétradas degeneran y una se convierte en la megáspora funcional. Al inicio de la megagametogénesis (formación de gameto femenino) la megáspora pasa por tres mitosis produciendo un saco embrionario formado por ocho núcleos que luego de la diferenciación celular se forman los núcleos polares ( $2n$ ) (que normalmente confluyen), dos sinérgidas ( $n$ ), 3 antípodas ( $n$ ) y la ovocélula o célula huevo. Finalmente conjuntamente con la degradación de las células antipodales y las sinérgidas se produce la característica doble fertilización de las angiospermas uniéndose un núcleo del polen con el núcleo polar formando el endosperma  $3n$  y otro con la célula huevo formando el embrión  $2n$ .

En organismo que se reproduce por apomixis gametofítica, dependiendo si es diplospórico o apospórico, la célula que dará origen al saco embrionario no reducido será la célula madre de la megáspora o una célula adyacente a la célula madre de la megaspora respectivamente, la cual luego tres mitosis forma un saco con un número variable de núcleos, dependiendo del tipo de apomixis, con células no reducidas ni recombinadas (apomeiosis) (Figura 1.4 B y C). En la diferenciación celular, en algunos tipos de apomixis apospóricas como el tipo *Hieracium*, se distinguen el núcleo polar ( $4n$ ), 2 sinérgidas ( $2n$ ), 3 células antipodales ( $2n$ ) y la ovocélula ( $2n$ ). Finalmente el núcleo polar se fecunda con un núcleo del polen (pseudogamia) para formar el endosperma ( $5n$ ). El embrión producido por

partenogénesis ( $2n$ ) no requiere fecundación. Por este mecanismo la semilla producida contiene la misma información genética que la planta madre.

Según el tipo de desarrollo del saco embrionario, la apomixis se puede dividir en nueve tipos (Savidan et al., 2001). *E. curvula* tiene su propio tipo característico denominado apomixis diplospórica “tipo Eragrostis”. La principal diferencia de este mecanismo es que tiene dos mitosis, formando el núcleo polar, dos sinérgidas, y la ovocélula, todas estas células son  $2n$  (Figura 1.4 C) (Meier et al., 2017). Cuando se produce a pseudogamia el endosperma final tiene un núcleo  $3n$  mientras que el embrión  $2n$ . Gracias a este mecanismo, los individuos sexuales y apomícticos de *E. curvula* contienen el mismo balance de núcleos entre el embrión y el endosperma ( $2n:3n$ ).

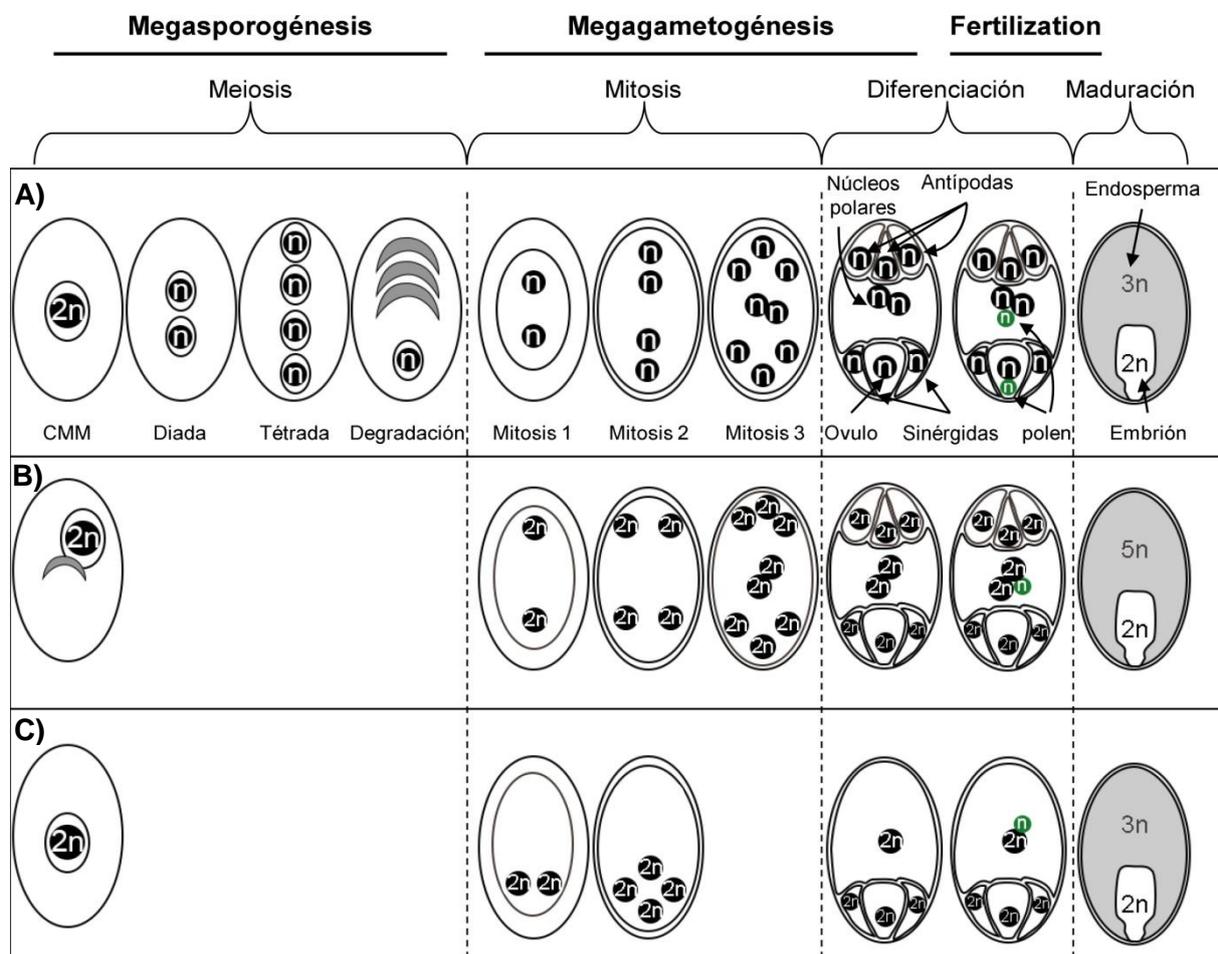


Figura 1.4: Embriogénesis en organismos A) sexuales, B) apomícticos apospóricos, C) apomícticos diplospóricos tipo Eragrostis. En la vía sexual y diplospóricas se puede observar que la relación embrión y endosperma es la misma.

La apomixis es un carácter polifilético que está distribuido en 33 de las 460 familias Angiospermas (Carman, 2008). Dentro de la familia de las gramíneas, se encuentran

especies apospóricas (Ej. *Paspalum notatum*) y diplospóricas, como *E. curvula*. En esta especie se pueden encontrar materiales con distinto modo reproductivo y, de acuerdo a esta característica, se los clasifica de la siguiente manera:

**Apomícticos obligados:** 100% de los sacos embrionarios dentro de un individuo derivan de apomeiosis.

**Apomícticos facultativos:** Dentro del mismo individuo conviven sacos derivados de apomeiosis y meiosis.

**Sexuales:** 100% de los sacos embrionarios dentro de un individuo son desarrollados en forma meiotica.

Ninguno de los materiales diploides de la especie (Tabla 1.1) se reproduce por apomixis, por lo que indicaría que existe una relación entre la ploidía y el modo reproductivo. Los cultivares tetraploides son generalmente facultativos aunque recientemente se comprobó que cruzando el cultivar sexual OTA con el cultivar apomíctico facultativo Don Walter se puede obtener una progenie con individuos sexuales y apomícticos (facultativos y obligados) con una proporción 1:1 (Zappacosta et al., 2019). Esta particularidad de contener todo el rango de expresión desde 100% de apomixis hasta 100% de sexualidad, pasando por individuos con número variable de sacos sexuales y apomícticos en la misma especie la convierte en un organismo ideal para el estudio de la apomixis diplospórica.

### **Origen natural de la apomixis**

Como se destacó en el punto anterior, la vía sexual y las vías apomícticas producen individuos genéticamente variables o clónales respectivamente. Sin embargo muchos de los componentes regulatorias de estas vías son compartidos (Koltunow, 1993). Por esta característica y por su naturaleza polifilética es que la amplia mayoría de las teorías enfocan a la apomixis como una variación del proceso sexual, existiendo solo un trabajo con la hipótesis inversa (Albertini et al. 2019).

Algunos autores (Lovell et al., 2017, Lynch et al., 1993) proponen que la acumulación de mutaciones en los genes relativos a la sexualidad produce pérdidas de funciones en esta vía, esta hipótesis surge de los análisis en poblaciones naturales de *Boechera* en la que los individuos apomícticos tenían mayor cantidad de mutaciones deletéreas que los sexuales.

Otros autores (Asker y Jerling 1992, Hojsgaard et al., 2014, Koch et al., 2003) proponen que la apomixis se habría generado espontáneamente por medio de la hibridación natural.

Existe vasta evidencia de plantas apomícticas originadas por hibridación de un individuo apomíctico que actúa como dador de polen y uno sexual que actúa como receptor (Paun et al., 2006, Šarhanová et al., 2017, Weber 2016), incluso en materiales diploides posiblemente derivados de tetraploides de *Boechera* se ha comprobado su origen híbrido (Beck et al., 2012). En todos los casos los materiales apomícticos resultan extremadamente heterocigotos comparados con sus pares sexuales, sugiriendo que provienen de combinaciones dispares. Los estudios mencionados han ayudado al entendimiento de los mecanismos y genes que actúan durante la hibridación pero no se sabe aún cuales son los factores clave que actúan en el cruzamiento de los individuos.

Otras hipótesis se enfocan en la poliploidización (relacionado con la hibridación) basado en que aproximadamente el 80% de las plantas apomícticas son poliploides (Carman 2001). Para que ocurra la poliploidización y obtener un nuevo juego de cromosomas, generalmente ocurren cambios previos que afectan la meiosis y el apareamiento en los cromosomas (como mutaciones), produciendo gametos no reducidos (Otto 2007). Esta ruta es común para la generación de poliploides en la naturaleza en las que la producción de gametos no reducidos ocurre en promedio 25,52% en híbridos y 0,56% en no híbridos (Ramsey et al., 2002). Cuando se producen híbridos poliploides, muy frecuentemente experimentan severas irregularidades en la meiosis las que afectan al nuevo individuo produciendo citotipos aneuploides, desbalanceados, inviables o, como ocurre en apomixis, apomeióticos (Ramsey et al., 1998).

Otras teorías acerca de la apomixis se basan en la epigenética. La epigenética se refiere a las modificaciones heredables en la expresión de los genes que no derivan de cambios en el nivel de nucleótidos en el ADN. Las modificaciones más comunes son 1) Metilación del ADN (incorporación de grupos metilo a la molécula de ADN): la metilación sobre promotores y regiones repetitivas produce el silenciamiento de genes, también ocurre sobre el cuerpo de los genes pero su efecto aún se está del todo dilucidado; 2) micro ARN (miARN): pequeñas secuencias de ARN no codificante que silencian los genes principalmente por regulación post transcripcional. Existen hipótesis que la apomixis sería un cambio epigenético reversible que silenciaría a las vías sexuales (Olmedo-Monfil et al., 2010). La coexistencia dentro de un mismo individuo de sacos sexuales y apomícticos indicaría que la apomixis se sobrepone reversiblemente a la sexualidad (Kumar 2017). Estos mecanismos epigenéticos actuarían reprimiendo la expresión de genes como respuesta a un factor de estrés ambiental o genómico (ejemplo poliploidización e hibridación), por lo cual podrían desencadenar una desregulación de la vía sexual (Grimanelli et al., 2012, Rodrigo et al., 2019). Estos cambios son heredables y estables en el tiempo. Existe evidencia de que los miARNs regulan a genes involucrados en la vía reproductiva de las plantas (Garbus et al.,

2019, Amiteye et al., 2013, Ortiz et al., 2019, Long et al., 2016). Otros estudios abordan la regulación epigenética del modo reproductivo por medio metilación del ADN encontrando genes silenciados por la metilación (Podio et al., 2014, García-Aguilar et al., 2010, Selva et al. 2017).

Todas estas teorías que han surgido a través de décadas de numerosa investigación parecieran estar relacionadas entre sí y ser válidas en conjunto, es decir, no serían mecanismo independientes y aislados. De esta manera, por ejemplo, se podría hipotetizar que en una especie dada, mutaciones en determinados genes producirían la pérdida de función de la meiosis dando lugar a la formación de gametos no reducidos, que por hibridación dieron origen a un organismo poliploide que podría ser facultativamente apomíctico (debido a heredar deficiencias en la meiosis) y el porcentaje de sacos sexuales y apomícticos podría estar regulados por mecanismos epigenéticos. Este ejemplo es muy simplificado dado que las plantas actuales son producto de millones de años de evolución y los mecanismos implicados aún no están totalmente resueltos.

### **Herencia de la apomixis y enfoques para su estudio**

Con el objetivo de poder determinar los mecanismos que regulan la apomixis, se han desarrollado muchos métodos y estrategias sobre varias especies con distintos tipos de apomixis. El desarrollo de nuevas tecnologías ha proporcionado más herramientas para poder responder a las preguntas que surgen de las investigaciones y se ha avanzado en términos de genes/regiones candidatos y modelos, pero aún no se ha podido dilucidar completamente como es regulada la apomixis. Una de las razones por las que todavía no está resuelto este mecanismo es porque todavía no están establecidas las bases genéticas y moleculares de los pasos claves en las vías sexuales. Según Grossniklaus (2001), solo a través de equipos multidisciplinarios que estudien la regulación de los mecanismos sexuales y apomícticos será posible aplicar la ingeniería genética en apomixis. De hecho, los mayores avances en términos de transferencia de la apomixis sintética se obtuvieron estudiando las vías sexuales (Khanday et al., 2019).

La herencia de la apomixis puede ser explicada muy bien por la genética mendeliana. Sin embargo, las complicaciones para el análisis se acumulan en especies apomícticas principalmente por interacciones epistáticas entre genes, componentes que se expresan esporofíticamente y no gametofíticamente, poliploidía, distorsión en la segregación y supresión de la recombinación (Ozias-Akins et al., 2007). El uso de distintas técnicas de mapeo molecular ha ayudado a identificar en algunas especies genes y regiones que estarían asociadas a la apomixis. En muchos estudios en especies diplospóricas y

apospóricas se determinó que la apomixis es dominante sobre la sexualidad (Nogler 1975, Savidan et al., 1980, Leblanc et al., 1995, Pessino et al., 1997, Gustine et al., 1997) y que los componentes de la apomixis (principalmente partenogénesis y apomeiosis) cosegregarían juntos como un carácter simple. Un modelo de referencia que ha servido para muchas especies es el de *Pennisetum* y *Cenchrus* donde por medio de la cruce de un genotipo sexual y otro apomítico construyeron una población de mapeo F1 en la cual evaluaron la segregación del modo reproductivo e identificaron zonas que segregan en forma conjunta en los individuos apomíticos. La hibridación *in situ* fluorescente con marcadores generados específicamente, permitió detectar la región en la que se encuentran, que fue secuenciada mediante la estrategia de cromosoma artificial bacteriano (BACs) (Ozias-Akins et al., 1993, Sherwood et al., 1994, Ozias-Akins et al., 1998, Goel et al., 2003, Conner et al., 2008). Dentro de esta región, se encontraron dos copias de BABY BOOM (*BBM*), un gen cuya expresión ectópica en *Arabidopsis thaliana* había funcionado como gatillo para el paso del crecimiento vegetativo al desarrollo embrionario (Boutillier et al. 2002). Luego se comprobó que la expresión de este gen en plantas sexuales transgénicas de *Pennisetum squamulatum* previa a la fertilización, induce la partenogénesis y produce descendencia haploide (Conner et al., 2015). La introgresión de este gen en arroz y maíz tuvo el mismo efecto produciendo embriones haploides (Conner et al., 2017). Para poder detectar regiones condicionantes de la apomixis en otras especies se ha seguido la misma lógica que el modelo de *Pennisetum*, pero con otras técnicas como mapeo óptico, genotipado por secuenciación, mapeo comparativo entre otras (Zappacosta et al., 2019, Labombarda et al., 2002, van Dijk et al., 2004).

Otro tipo de análisis para poder identificar directamente cuales son los genes que regulan la apomixis, ha sido la comparación en la expresión génica de genotipos apomíticos con genotipos sexuales. Para poder determinar el perfil transcripcional de los órganos reproductivos se han utilizado técnicas como display diferencial (Poirier et al., 1997), SuperSAGE (análisis serial de la expresión de genes) (Matsumura et al., 2006) y secuenciación masiva (Katz et al., 2010). La expresión de transcritos varía ampliamente según el tejido analizado, por este motivo a lo largo del tiempo se ha apuntado cada vez más específicamente al estudio de la apomixis, pasando de tejidos florales hasta los análisis actuales de célula específica. Análisis a nivel de tejido reproductivos fueron realizados por ejemplo en *Panicum* (Chen et al., 1999), *Paspalum* (De Oliveira et al., 2020, Laspina et al., 2008), *Pennisetum* (Singh et al., 2007), *Eragrostis* (Cervigni et al., 2008, Garbus et al., 2017). Otros estudios de transcriptómica, como el caso de *Paspalum* (Ortiz et al., 2017) y *Pennisetum* (Sahu et al., 2012) se han basado en extracciones de ARN en distintos estadios

cronológicos del desarrollo de la semilla para poder detectar sobre qué componentes actúan los genes.

De los estudios realizados con el fin de caracterizar tanto la vía apomíctica como la sexual se han podido identificar algunos genes con roles claves sobre los componentes de la apomixis. Con respecto a la apomeiosis, el knock-out de tres genes pudo convertir exitosamente la mitosis en meiosis por lo cual llamaron al mutante MiMe. Esta mutación fue desarrollada inicialmente en *A. thaliana* sobre dos genes que afectan la segregación de la cromatina (*Atrec8* y *osd1*) y uno que actúa sobre la recombinación (*Atspo11-1*). La descendencia de este mutante no recombina ni reduce sus núcleos, por lo que duplica su ploidía en cada generación (D'Erfurth et al., 2009). Luego en *O. sativa* se pudo determinar que además de *Atspo11-1*, existen otros genes que afectan la recombinación como *prd1*, *prd2* y *pair1*. De esta forma en *O. sativa* se desarrolló un nuevo mutante OsMiMe, *osd1*, *rec8* y *pair1*, el cual por autopolinización produjo la totalidad de individuos tetraploides demostrando que ambos gametos no estaban reducidos (Mieulet et al., 2016). En cuanto a la partenogénesis, los genes BABY BOOM-Like, que se encuentran en copias múltiples en la región identificada en *Pennisetum* (Conner et al., 2008), han resultado tener una alta expresión al momento de la formación del embrión, encontrándose además que su expresión ectópica (en condiciones normales su expresión es inducida por el núcleo del polen) produce genotipos haploides en *Zea mays* y *O. sativa*. Hasta el momento, *BBM1* es el gen mayormente caracterizado para la partenogénesis. La pseudogamia no ha sido estudiada tanto como los otros componentes de la apomixis pero existen teorías que los genes que inician la partenogénesis producen un recubrimiento en la célula huevo que no permite que sea fertilizada por el núcleo del polen (Vielle et al., 1995).

En la búsqueda de transcriptos involucrados en la activación cigótica en el ciclo sexual de arroz, fue descubierto que *BBM1* no se expresaba en la célula huevo previo a la fertilización y que luego su expresión aumenta significativamente (Khanday et al., 2019). Por medio de la introgresión de este gen con un promotor específico de la célula huevo se obtuvo un embrión sin fertilización. Combinando el triple mutante MiMe con *BBM1* fue posible reemplazar la meiosis y generar partenogénesis con lo que la progenie es idéntica a la planta que le dio origen. Aunque este hallazgo ha representado un gran avance, la baja eficiencia de este modelo (entre un 15%-26% de semillas apomícticas en distintas generaciones) no permite aun utilizarlo como herramienta genética.

A pesar de todos estos esfuerzos para poder dilucidar los mecanismos apomícticos, hasta el momento no se ha podido determinar claramente las bases moleculares y regulatorias tanto epigenéticas como genéticas. Algunos autores (Hojsgaard 2020) creen que el objetivo

será concretado utilizando un modelo específico para cada especie (debido a las diferencias interespecíficas de los tipos de apomixis) y que para concretarlo será fundamental la comparación genómica entre genotipos sexuales y apomícticos de la misma especie.

### **Importancia de la Apomixis**

Como se dijo anteriormente, la apomixis es un carácter polifilético ampliamente distribuido en angiospermas, sin embargo ninguno de los cultivos con mayor importancia económica cuenta con este modo reproductivo. La introducción de este carácter contiene ventajas que afectan desde los pequeños productores hasta las grandes empresas de mejoramiento, es decir, toda la cadena productiva.

Cuando se cruzan dos líneas endogámicas (líneas autopolinizadas con alto grado de homocigosis seleccionadas previamente por alguna característica deseada), una dadora de polen y otra receptora, se produce un híbrido que es superior a los padres con mayor rendimiento y biomasa. Este fenómeno se denomina heterosis o vigor híbrido y es el medio por el cual se obtienen los híbridos comerciales de algunos cultivos (Figura 1.5A). Para poder llegar a un híbrido comercial se deben seleccionar los padres de una población segregante con características deseadas. Luego se deben producir líneas homocigotas de esos padres por autopolinización o por la técnica de doble haploides en la que se dobla la ploidía del embrión o del polen haploide para obtener un individuo 100% homocigoto. Una vez conseguidos los padres homocigotos, estos se multiplican para tener la cantidad de individuos necesaria para la hibridación. Finalmente se cruzan los individuos para producir la semilla híbrida que será sembrada para obtener el producto final (forraje, grano, fruto, etc). Este proceso lleva aproximadamente entre 7 y 10 años y la cruce del híbrido para obtener la semilla comercial se realiza anualmente (MacRobert et al., 2015). Si la cosecha del híbrido comercial se sembrara, resultaría en una población segregante con individuos genética y morfológicamente heterogéneos sin las características deseadas. Incorporado la apomixis se podría mantener el vigor híbrido por muchas generaciones sin tener que hacer la cruce todos los años, ahorrando tiempo y aumentando superficie disponible para la producción de alimentos (figura 1.5B).

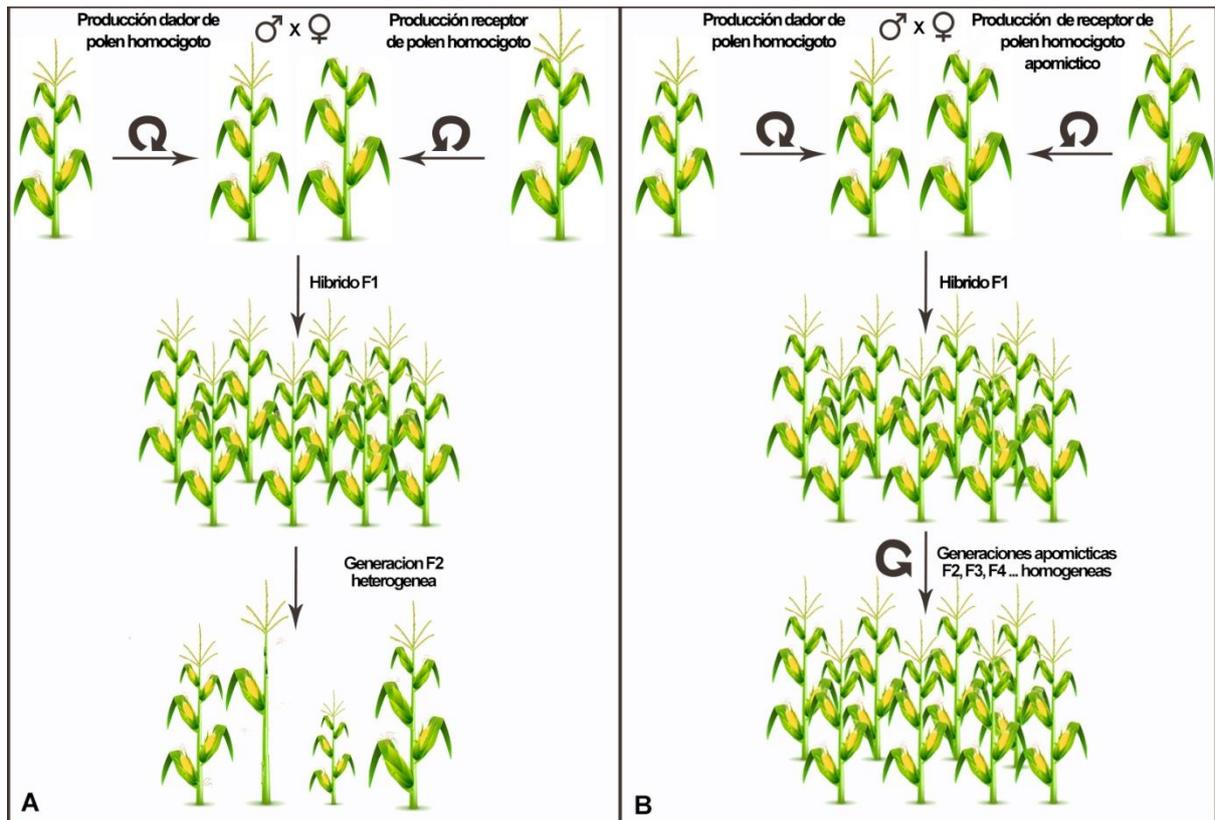


Figura 1.5: Producción híbrida de maíz con el fin de producir cultivares comerciales con vigor híbrido. A) Sistema tradicional de cruce entre padres homocigotos. B) Potencial sistema para la fijación del vigor híbrido en el cual se cruza un dador de polen que contiene el carácter de la apomixis produciendo progenie con el vigor híbrido de forma clonal en las próximas generaciones.

El sistema tradicional para la obtención de híbridos se utiliza principalmente en maíz (segundo cereal producido a nivel mundial) y girasol y no ha podido utilizarse en otros cultivos como arroz y trigo por el costo que implica desarrollar las cruzas, ya que sus órganos reproductivos dificultan la técnica dado que no se encuentran separados como en maíz. El desarrollo de nuevas especies híbridas traería un aumento de la producción mundial exponencial y ha sido llamada desde “la nueva revolución verde” hasta “El santo grial de la agricultura” (Salller et al., 2019, Calzada et al., 1996).

En el caso de los cultivos que ya se producen por hibridación, por medio de la apomixis se podrá diseñar el perfil necesario para cada cultivar según el tipo de región e introducir nuevas variedades en zonas en las que actualmente no se producen por que el desarrollo del híbrido es muy costoso relativo a los beneficios que traería. La disminución del costo de obtención de híbridos podría, por ejemplo, generar cultivares con mayor tolerancia a sequía para introducirlos en regiones con bajo régimen hídrico.

La introgresión apomixis autónoma además podría prescindir de la polinización. En muchos cultivos la polinización es un aspecto crítico para la obtención de frutos de calidad, tanto su sincronización (pistilo preparado para recibir el polen cuando está siendo liberado de las

anteras) como dispersión (factores climáticos como heladas y sequías pueden eliminar la dispersión del polen) pueden verse severamente afectados impactando directamente sobre la calidad y el rendimiento.

En especies como la papa que se reproducen vegetativamente a través de tubérculos, la dispersión de virus y bacterias se produce rápidamente. La producción de semillas clonales en estas especies disminuiría rápidamente estas enfermedades, bajando los costos de producción y la utilización de agroquímicos (Ramulu et al., 1999).

El valor total del mercado semillero fue estimado en 48.1 billones de dólares (Bonny et al., 2017) y, según la FAO (<http://www.fao.org/3/ca6929en/ca6929en.pdf>), el maíz representa 18 billones de dólares este total y la soja 8 billones. Estos números representan el alto costo de producir semillas híbridas con el sistema actual y permiten inferir el impacto que tendría la introducción de la apomixis solamente en maíz. Además, se estima que la fijación de la heterosis hará los programas de mejoramiento de especies como arroz más rápidos y más baratos con beneficios estimados en 2,5 billones de dólares por año. Finalmente, el cambio de propagación de vegetativo a semillas de cultivos como papa y mandioca podría ahorrar a los productores 3,2 billones de dólares por año (Spillani et al. 2004).

## **Secuenciación genómica**

Se denomina secuenciación genómica a la determinación del orden (secuencia) de nucleótidos del genoma de un organismo los cuales son traducidos a las letras T, G, A y C por sus nombres tirosina, guanina, adenina y citosina respectivamente. En el mismo año que se descubrió la estructura de doble hélice de ADN (Watson y Crick 1953) fue secuenciada la primera molécula que fue la proteína insulina por el método de partición de cromatografía. En esta metodología las proteínas son fragmentadas al azar, los fragmentos son leídos individualmente y luego por superposición se obtiene la secuencia completa consenso (Sanger y Thompson 1953). Si bien la primera secuenciación fue de aminoácidos, los principios fueron mantenidos para la secuenciación en 1965 de 76 nucleótidos del ARN de transferencia de levadura *Saccharomyces cerevisiae* (Holley et al., 1965). Luego se secuenciaron los marcadores de peso molecular utilizados actualmente como el fago  $\lambda$  y *Escherichia coli* (Wu y Kaiser 1968, Gilbert y Maxam 1973) y en 1976 el primer genoma completo del fago MS2 de 3.569 pares de bases (pb) (Fiers et al., 1976). En 1975 se desarrolló el primer método de secuenciación de Sanger y Coulson llamado "plus and minus" (Sanger y Coulson 1975) en el cual se realizaban cuatro pares de reacciones que contienen un primer, ADN polimerasa, cuatro nucleótidos de los cuales uno está marcado

en cada reacción y puede ser detectado por exposición a una placa de rayos X sobre el gel de poliacrilamida. En 1977 se termina de pulir este método y será llamado “método de Sanger” con un rendimiento revolucionario para la época en el que se podía obtener 400 bases en un día (Sanger et al., 1977). En la década del 80 para acelerar el proceso de secuenciación se desarrolla el método de secuenciación shotgun (Staden 1979) en el que por medio de vectores se clonan al azar moléculas largas de ADN que se secuencian en paralelo produciendo lecturas de ADN llamadas reads, que luego se ensamblan por solapamiento. Por este método en 1985 se pudo secuenciar el genoma completo del virus de la mononucleosis (virus Epstein-Barr) que tiene 172.282 pb. En 1982 se funda la primera base de datos, el GenBank, con 680.338 bases que al final de la década tendría 42.495.893 (<https://www.ncbi.nlm.nih.gov/genbank/statistics/>). En el año 1987 se desarrolla una nueva técnica que más adelante se conocería como ABI PRISM, capaz de automatizar parcialmente la secuenciación de ADN, con ventajas significativas sobre los métodos anteriores dentro de las cuales se pueden mencionar: aumento del rendimiento de secuenciación en una sola corrida, cambio de radioisótopos por fluoróforos y almacenamiento de las secuencias en computadoras (Smith et al., 1986). El título del artículo publicado en 1988 por Sanger “secuencias, secuencias y secuencias” (“Sequences, Sequences, and Sequences”) resume la explosión de la tecnología de la época. Con todos estos adelantos, en 1990 se forma el proyecto mundial para secuenciar el genoma de referencia humano que finalizaría en 2003 (Sawicki et al., 1993).

El proyecto del genoma humano generó interés en nuevos sectores para desarrollar tecnologías de secuenciación. En 2003, 454 Life Science introduce al mercado un secuenciado basado en la pirosecuenciación que produce lecturas llamadas lecturas de 500 pb con una precisión del 99% y un rendimiento de 25 millones de bases en 4 horas con un costo de la sexta parte del método de Sanger (Nyrén 1987). En 2006, Solexa (luego adquirida por Illumina) introduce la tecnología “pair-end” (se secuencian las dos terminales de la cadena de ADN) que produce lecturas de menor longitud que 454 pero con mayor rendimiento 1 Gpb (giga pares de bases) en cada secuenciación (Bentley et al., 2008). Estas tecnologías se denominaron NGS “new generation sequencing” y nuevas aplicaciones surgieron como secuenciación de exomas, secuenciación de RNA (RNA-Seq), inmunoprecipitación de la cromatina seguido de secuenciación (ChIP-seq), secuenciación epigenética, SNP-Seq entre otras.

La última tecnología de secuenciación fue llamada TGS “Third Generation Sequencing”. Esta tecnología se destaca por no necesitar amplificación del ADN previo a la secuenciación, poder secuenciar fragmentos largos y tener un rendimiento elevado (Braslavsky et al., 2003). Los mayores representantes de esta tecnología son Pacific

Biosciences (PacBio) y Oxford nanopore (ONT). Estas tecnologías pueden producir librerías de promedio entre 10 y 20 Kb y lecturas de hasta 50 Kb pero carecen de buena calidad, en promedio Illumina tiene un error de 1 cada 1000 bases contra 1 cada 10 de PacBio Sequel, sin embargo estos errores son al azar pudiendo realizar un consenso cuando hay suficiente cobertura (Carneiro et al 2012). En la actualidad PacBio desarrollo la tecnología HiFi que permite tener la misma precisión que Illumina con lecturas largas. La secuenciación ONT produce lecturas de mayor longitud, teniendo el record de 2.3 Mpb pudiendo cubrir mayor repeticiones que PacBio. Cabe destacar que ONT también desarrollo el secuenciador portable más pequeño hasta el momento con una preparación de librería muy sencilla que puede ser utilizada en lugares remotos y ser utilizado incluso en teléfonos celulares (Quick et al., 2016).

Las ventajas de utilizar TGS permiten acceder a todas las regiones de los genomas, tener cobertura homogénea, y aumentar la precisión estructural del ensamblado. La falta de precisión puede ser corregida con mayor cobertura o por medio de consensos alineando lecturas cortas. El desarrollo de nuevas tecnologías de secuenciación a lo largo del tiempo trajo aparejado una disminución de los costos (Wetterstrand 2019) disminuyendo de 10 millones a 0,010 dólares por Mpb del 2000 al 2019, con lo que trajo aparejado una masiva secuenciación de todo tipo de especies (Figura 1.6).

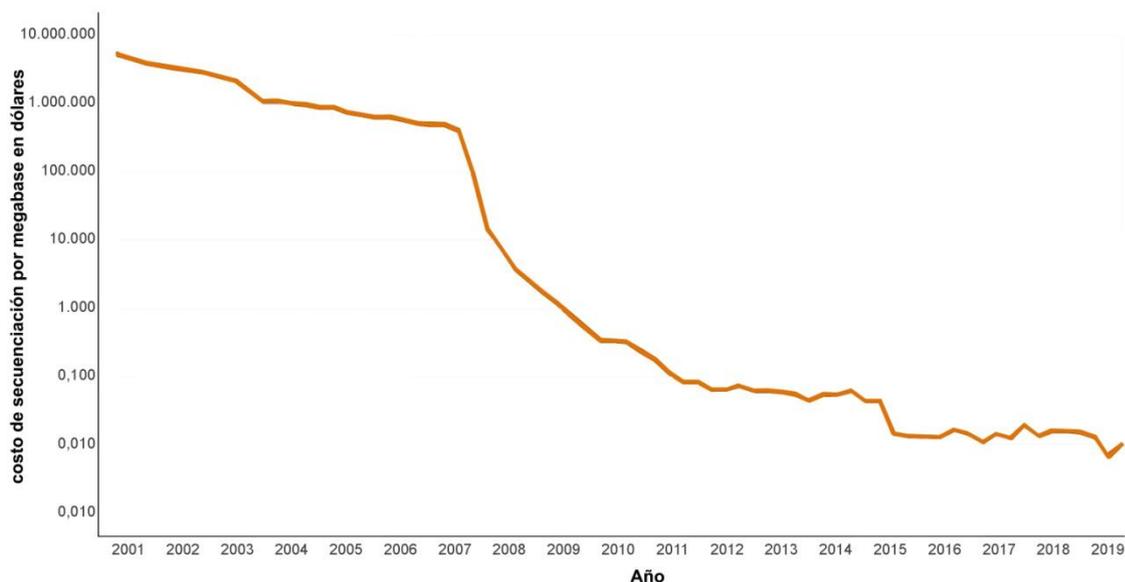


Figura 1.6: Costo de secuenciación en dólares por megabase desde la primer secuenciación del genoma humano. Puede observarse que hasta el 2007 la caída del costo fue lineal y con las nuevas tecnologías desarrolladas el descenso desde este punto comenzó a ser exponencial (Wetterstrand, 2019 [www.genome.gov/sequencingcostsdata](http://www.genome.gov/sequencingcostsdata))

Complementando el desarrollo de nuevas tecnologías de secuenciación surgieron métodos para poder obtener cromosomas enteros a menor costo. Las técnicas de “scaffolding” unen y orientan contigs (fragmentos ensamblados) por medio de ligamientos entre zonas genómicas distantes. La técnica de mapeo óptico comercializada por Bionano fue la primera en desarrollarse (Schwartz et al., 1993). Para esta técnica se utilizan fragmentos de ADN mayores a 150 Kb que luego son marcados por medio de enzimas que llevan distintos colorantes fluorescentes. Finalmente, con una cámara de alta resolución se lee el resultado que contiene el orden de óptico de los fluorescentes. Otro método utilizado para el scaffolding es Hi-C (Burton et al., 2013) que captura la conformación de los cromosomas utilizando ligación por proximidad. Estas ligaciones son cortadas con enzimas y religadas para poder ser secuenciadas con Illumina y mapeadas con el genoma para detectar regiones ligadas entre distintos contigs y así poder ensamblar los cromosomas.

Paralelo a la explosión de las distintas técnicas de secuenciación, se fueron desarrollando las bases de la bioinformática con el objetivo de lidiar con todas las secuencias y datos que fueron resultando de los distintos proyectos. En 1980 (Stade 1980) se desarrolla el primer programa bioinformático y se acuña la palabra contig definiéndola como un conjunto de lecturas de un gel que se relacionan entre ellas por solapamiento de su secuencia. Hoy se define como el resultado del ensamblado de las lecturas ya que no se utilizan más los gels para la secuenciación y los métodos de ensamblado no necesariamente son por solapamiento de toda la longitud de las lecturas. Los primeros ensamblados estaban basados en algoritmos denominados OLC (por sus siglas en inglés Overlapping Layout Consensus) que alineaban las secuencias entre si y tomaban el mejor resultado formando un consenso. En 1985, surge “The National Center for Biotechnology Information” (NCBI), una base de datos que permite realizar alineamiento con sus secuencias por medio de BLAST (Basic Local Alignment Search tool) y de esta forma poder rastrear secuencias ya conocidas en otros proyectos (Altschul et al., 1990). Con NGS el volumen de datos en términos de espacio informático y recursos para procesarlos llevó el algoritmo de OLC al de Bruijn graphs (DBG) el cual utiliza k-mers (cortes de las lecturas en fragmentos más pequeños) que utiliza para ensamblar las lecturas más rápidamente y con menos recursos. El último algoritmo desarrollado para lidiar con los errores de TGS fue Hierarchical Genome Assembly Process (HGAP), que básicamente utiliza las lecturas más largas y las alinea con las más cortas para crear un consenso previo al ensamblado, este método es principalmente utilizado por el software FALCON (Chin et al., 2013). FALCON está optimizado para genomas diploides de gran tamaño y divide su proceso en tres etapas: 1) las lecturas más cortas son alineadas con las más largas para generar secuencias

consenso de alta precisión llamadas pre-reads. Estas secuencias pueden ser pensadas como lecturas aún más largas pre-ensambladas y corregidas; 2) Las pre-reads son alineadas entre ellas y se produce el ensamblado en contigs; 3) finalmente el resultado de FALCON corresponde al genoma haploide de la especie, es decir, no puede distinguir entre haplotipos cuando los genomas son muy homocigotos, cuando el grado de heterocigosidad es alto este programa intenta buscar secuencias homologas por medio de alineamientos de regiones distantes a la región con heterosis para finalmente producir el genoma haploide. Otro algoritmo desarrollado para TGS es el del software Canu (Koren et al., 2017) que utiliza OLC con modificaciones especiales para lecturas largas. Al igual que en FALCON, el proceso está dividido en tres pasos: carga de lecturas en una base de datos, cómputo de k-mers, cómputo de solapamiento y carga de solapamiento en base de datos. Luego realiza diferentes acciones con los solapamientos: 1) corrección: se utilizan los solapamientos para corregir las lecturas y disminuir la tasa de error general; 2) filtrado: utiliza los solapamientos para decidir que regiones de cada read son de alta precisión; 3) consenso: utiliza los solapamientos más consistentes para producir el ensamblado final. Canu, a diferencia de FALCON, está recomendado para genomas poliploides.

En la actualidad (2020) la mayoría de los procesos de secuenciación implican la combinación de tecnologías de secuenciación para poder producir un genoma con alta calidad (baja tasa de error en sus bases) y contigüidad (baja fragmentación de los cromosomas). Según la tecnología incorporada será el programa bioinformático que se utilizará. El objetivo de la secuenciación de genomas puede variar según se siga un enfoque filogenético o biotecnológico, pero en todos los casos se busca obtener el catálogo completo de los genes y elementos no codificantes que componen a un organismo y el orden de los mismos. Es por medio de la anotación que se consigue este catálogo alineando genes anteriormente secuenciados de la misma especie o de especies cercanas y por medio de algoritmos que basados en las características estructurales de los elementos modelan los genes.

### **Recursos genómicos de *Eragrostis***

Como se dijo anteriormente *E. curvula* tiene un número básico de cromosomas de 10, sin embargo, los genotipos tetraploides evaluados hasta el momento son alotetraploides segmentales (Poverene 1988) y su Cx (tamaño del genoma haploide) varía entre 616 y 636 Mpb dependiendo el genotipo evaluado lo que corresponde a alrededor de 0.65 pg (Meier et al 2011). A pesar de no haberse evaluado específicamente la heterocigosidad de *E. curvula*

es esperable un alto grado de heterosis basado en la poca domesticación de la especie y a los ensayos de marcadores realizado (Zappacosta et al., 2011).

Con el objetivo de determinar las bases genéticas y moleculares de la apomixis diplospórica, en *E. curvula* se han desarrollado a la largo del tiempo distintos esfuerzos de secuenciación. En el 2008 se secuenciaron cuatro genotecas de cDNA construidas a partir de: i) panoja de una planta tetraploide apomítica, ii) panoja de una planta tetraploide sexual iii) panoja de una planta diploide sexual iv) hoja de un tetraploide apomítica (Cervigni et al., 2008). Se obtuvieron 12.295 secuencias sumando las cuatro librerías, con las cuales se pudieron diseñar marcadores SSR. Si bien se consiguió solo una fracción del total de los transcriptos de *E. curvula*, esta fue la primera base de datos de transcriptómica de la especie. Luego se desarrolló un estudio de expresión comparando dos réplicas biológicas de espiguillas de un material tetraploide sexual (OTA) y dos réplicas biológicas de un material tetraploide apomítico (Tanganyika) (Garbus et al., 2017). Se realizó la secuenciación de cDNA con la tecnología 454 y se obtuvieron 49.568 isotigs (lecturas de secuencias de cDNA ensambladas) y se desarrollaron SSR que pudieron ser validados en otros cultivares. Con un análisis de expresión diferencial se pudieron identificar 6.203 transcriptos sobreexpresados en el cultivar apomítico y 3.060 sobreexpresados en el cultivar sexual. Muchos de los genes encontrados están vinculados a la reproducción pero el gran número de transcriptos diferenciales dificultó la selección para su estudio. Este trabajo representó un gran avance dentro de los recursos genómicos hasta el momento, ya que amplió en gran medida la base de datos anterior e identificó genes sobreexpresados para la apomixis. Con los que respecta a la expresión, existe otro grupo de trabajo en Reino Unido en el laboratorio Sainsbury liderado por Mathew Moscou y Nick Talbot estudiando la resistencia de *E. curvula* a el patógeno *Magnaphorte oryzae* que secuenció 17 transcriptomas de distintos cultivares (trabajo no publicado) secuenciados con Illumina para poder desarrollar análisis comparativos.

En el 2019, fue finalizado el primer mapa de ligamiento de la población de mapeo obtenida de la de la cruce de Don Walter (apomítico facultativo) y OTA (sexual) (Zappacosta et al., 2019). Esta población fue genotipada con la tecnología GBS (genotyping by sequencing) la cual permitió obtener marcadores para cada uno de los individuos los cuales fueron utilizados para construir el mapa de ligamiento para cada uno de los progenitores. Dentro del grupo de ligamiento 3 del cultivar Don Walter fue posible delimitar una región de marcadores sin recombinación ligada a la apomixis.

Dentro del género *Eragrostis* se han generado recursos genómico para otras especies. Con el objetivo mejorar las características nutritivas se secuenciaron dos versiones del genoma

de *E. tef* y transcriptomas de distintos tejidos (VanBuren et al., 2019, Cannarozzi et al., 2014, Bai et al., 1999). La especie *Eragrostis nindensis* es utilizada como herramienta para el estudio de la tolerancia al estrés por sequía y su genoma fue recientemente secuenciado junto con transcriptomas de distintos tejidos (Pardo et al., 2020).

## Objetivo general

El descubrimiento de plantas sexuales (Voigt y Bashaw 1972, Cardone et al. 2006) y apomícticas facultativas en *E. curvula* (Voigt y Bashaw 1973, 1976, Brix 1974) demostró que el sistema reproductivo del pasto llorón es sumamente complejo. Los individuos diploides ( $2n = 2x = 20$ ) son muy infrecuentes en la naturaleza (Poverene y Voigt 1997) y los materiales tetraploides ( $2n = 4x = 40$ ) son generalmente apomícticos aunque se han obtenido genotipos 100% sexuales por medio de cruzamientos.

En las angiospermas es carácter heredable, pero la base genética es inesperadamente compleja. El mapeo genético de la aposporía y la diplosporía ha confirmado la herencia dominante tanto ligada como no ligada de la apomeiosis y la partenogénesis. En muchos de los estudios de mapeo se ha observado estrecha vinculación de múltiples marcadores y la supresión de la recombinación de por lo menos uno de los dos loci (Ozias-Akins y van Dijk, 2007; Albertini y col., 2010; Pupilli y Barcaccia, 2012).

A pesar de su amplia distribución en las plantas con flores el carácter apomixis no está presente en muchas especies de interés agronómico, a excepción de varias gramíneas forrajeras (Carneiro et al. 2006). Esto ha llevado a que la mayoría de los estudios genéticos sobre el mismo deban ser realizados en especies silvestres altamente heterocigotas y poco caracterizadas genéticamente.

La genómica está produciendo una nueva revolución en el área del mejoramiento vegetal, ya que permite analizar, de manera más precisa, la relación genotipo-fenotipo para un elevado número de caracteres y conocer la estructura y función de los genes involucrados (Tester y Langridge, 2010; Feuillet et al., 2011).

El objetivo general de esta Tesis es obtener por primera vez secuencias genómicas de buena de calidad de *Eragrostis curvula* que nos permitan dilucidar la arquitectura genómica de esta gramínea. Esto nos brindará acceso a un catálogo completo y ordenado de genes y a un casi ilimitado número de marcadores moleculares para asistir al mejoramiento. Entre los genes que revisten mayor interés en *E. curvula*, se encuentran los relacionados con calidad de forraje, con resistencia a estrés abiótico y, muy especialmente, los genes presentes en la región condicionante de la apomixis. Identificar estas regiones y sus posibilidades de expresión permitirá, en una instancia futura, manipular el carácter e intentar la introgresión del mismo en especies sexuales de interés agrícola por medio de ingeniería genética.

**Los objetivos específicos de esta tesis son los siguientes**

- Secuenciar y ensamblar el genotipo diploide sexual de *Eragrostis curvula* cv. Victoria
- Secuenciar y ensamblar los genotipos apomícticos facultativos de *E. curvula* cv. Don Walter y cv. Tanganyika INTA.
- Identificar los elementos repetitivos de los genomas.
- Realizar una anotación exhaustiva de los genes de cada genotipo.
- Identificar y clasificar genes relacionados con la tolerancia al estrés y a la calidad forrajera.
- Anclar la región asociada a la apomixis en trabajos anteriores y compararla entre los genotipos apomícticos y los sexuales
- Encontrar la relación filogenética de la especie dentro del género *Eragrostis*, la subfamilia Chloridoideae y la familia Poaceae.
- Identificar los genes regulatorios de las vías sexuales caracterizados en otras especies vinculadas principalmente a la meiosis, fertilización de la ovocélula y formación del embrión.
- Identificar las regiones o región vinculada a la apomixis dentro de todos los genotipos.
- Determinar la expresión de los genes dentro de la región utilizando datos transcriptómicos.

## Hipótesis

- El uso complementario de las nuevas tecnologías de secuenciación permiten ensamblar genomas repetitivos y poliploides como el de *E. curvula*.
- La secuenciación de tres genomas de *E. curvula*, obtenidos a partir de genotipos diferentes (diploide sexual, tetraploide apomíctico), permite obtener el catálogo cercano al total de los genes de la especie.
- La anotación de los genes de los genomas de *E. curvula* permite clasificarlos en relación a la tolerancia al estrés por sequía y a la calidad forrajera.
- En los genomas apomícticos existe una región apomíctica la cual regula el carácter.
- La apomixis se produce a través de la desregulación de la sexualidad, la expresión de genes no existentes en genotipos sexuales y la regulación epigenética.

## Capítulo 1 Secuenciación y ensamblado del genoma diploide sexual de *Eragrostis curvula*

### Introducción

La incorporación de *E. curvula* en los sistemas productivos de cría se debe a su persistencia, producción, capacidad de carga, baja susceptibilidad a plagas y enfermedades, alta tolerancia al frío invernal y por su alta tolerancia a sequía, pudiendo sobrevivir con un mínimo de entre 250 y 350 mm por año (Kent 2019). Sin embargo su digestibilidad 25 días luego del rebrote de primavera es del 60% y cae al 50% en 70 días (Vera et al., 1973). En resumen, su capacidad para sobrevivir bajo condiciones de escases de agua, la posiciona como la principal alternativa en zonas marginales pero su baja digestibilidad limita su expansión a regiones donde otras especies tienen mayor calidad.

Uno de los compuestos que limita la calidad es la lignina, un polímero orgánico compuesto por monómeros denominados unidades H, G y S que es el mayor componente recalcitrante (resistente a la degradación por microorganismos) de la biomasa de las plantas. Es una sustancia hidrofóbica que forma la pared secundaria de los forrajes, por lo que no permite que las bacterias de los rumiantes la degraden. La lignina forma un entramado tridimensional con la celulosa y hemicelulosa por lo tanto, cuando aumenta su contenido, “secuestra” estos componentes no dejándolos disponibles para ser digeridos, disminuyendo de este modo el porcentaje de digestibilidad. El silenciamiento de genes y enzimas implicadas en la biosíntesis de ligninas por medio de ingeniería genética, puede disminuir entre 10% y 50% su contenido (Poovaiah et al., 2014). Sin embargo, en muchos casos afecta a otras funciones como la producción de biomasa y resistencia al estrés. El mejoramiento de la digestibilidad sin alterar otras funciones se puede lograr silenciando enzimas que cambien la composición relativa de los monómeros de la lignina H, G y S y el contenido de celulosa (Tu et al. 2010, Chen et al. 2003). Por ejemplo, el silenciamiento de la enzima caffeoyl-CoA 3-O-methyltransferase con ARN de interferencia disminuye en maíz el contenido de lignina en un 22,4%, aumenta el contenido de celulosa en un 23,3% y aumenta la relación de las unidades S/G un 53,08%, mejorando la digestibilidad sin afectar otras funciones (Li et al. 2013). En trabajos anteriores en *E. curvula* se pudieron aislar e identificar cuatro alelos de esta enzima (Díaz et al., 2010) representando un potencial gen para su mejoramiento. Otra enzima recientemente identificada en las vías de las ligninas es caffeoyl shikimate esterase (CSE) (Vanholme et al., 2013) cuyo silenciamiento disminuye el contenido total de lignina, aumenta los azúcares solubles y cambia la relación entre las unidades de monómeros. La identificación y caracterización de la secuencia que codifica

CSE en *E. curvula* podría ser el paso inicial para trabajar en el efecto del silenciamiento de este gen en plantas forrajeras, ya que hasta el momento solo fue testeado en *A. thaliana* (Vanholme et al., 2013 y Vargas et. al 2016) y *Populus* (Saleme et al., 2017).

La resistencia a distintos estreses es otra de las características importantes de *E. curvula*. Los cultivos están expuestos a diferentes estreses bióticos y abióticos que afectan su normal desarrollo y su la productividad. El estrés por sequía es considerado el de mayor impacto sobre la productividad, disminuyendo el rendimiento más que cualquier otro estrés (Lambers et al., 2008). El cambio climático además de aumentar la temperatura (causando mayor evapotranspiración), disminuye la frecuencia de precipitaciones, por lo tanto, los nuevos programas de mejoramiento están siendo dirigidos a la identificación y selección de genes relacionados al estrés por sequía con el fin de obtener cultivares con mayor tolerancia (Tester y Langridge 2010). La familia de factores de transcripción WRKY ha sido estudiada ampliamente por su rol en el crecimiento y desarrollo de las plantas y su respuesta ante al estrés. Por ejemplo, en *A. thaliana*, la activación de la expresión del gen AtWRKY57 incrementa la tolerancia al estrés (Jiang et al., 2012). En maíz, ZmWRKY106 y ZmWRKY40 se han hallado sobreexpresados frente a condiciones de estrés por sequía y su sobreexpresión por transgénesis en *A. thaliana* le confiere mayor tolerancia a este estrés. Debido a la excelente tolerancia a la sequía de *E. curvula*, la identificación, clasificación y búsqueda del rol de estos factores de transcripción podrían ser relevantes para describir las bases moleculares que le confieren resistencia e introgresarla en otros cultivos de mayor importancia económica.

Las dos características de *E. curvula* mencionadas anteriormente, tolerancia a estrés por sequía y calidad forrajera, son muy importantes y se les ha prestado especial atención en trabajos de investigación (Balsamo et al., 2006, Díaz et al 2010., Gargano et al., 2001, Lauric et al., 2012). Sin embargo, el principal motivo por el que se estudia *E. curvula* es por su modo reproductivo. *E. curvula* se ha convertido en un modelo para el estudio de la apomixis diplospórica. Una de las principales razones es que su relación entre los núcleos del endosperma y embrión son constantes ( $2n$  embrión,  $3n$  endosperma) en individuos sexuales y apomícticos, y que la segregación de la apomixis en esta especie ofrece todo el rango de modos reproductivos para su estudio en materiales tetraploides, desde apomícticos obligados, apomícticos facultativos hasta sexuales. Si bien se ha identificado una región que estaría vinculada a la apomixis (Zappacosta et al., 2019), hasta el momento esta región no ha sido secuenciada y, por lo tanto, los genes dentro de la región no se han podido identificar.

Por medio de la secuenciación genómica se podría obtener respuesta para todas las problemáticas relacionadas a *E. curvula*, i.e., identificación de genes con el objetivo de comprender las bases moleculares de la tolerancia a la sequía, la calidad forrajera y especialmente el modo reproductivo. Además, la secuenciación genómica por si misma constituye una herramienta para el diseño de análisis evolutivos, diseño de marcadores y desarrollo de cultivares transgénicos. *E. curvula* por otra parte, ofrece la posibilidad de secuenciar cultivares diploides con genomas más pequeños, por lo tanto, con menos dificultades para el ensamblado y la identificación de elementos repetitivos. De esta forma, a través de la secuenciación, ensamblado y anotación de los elementos codificantes y no codificantes del genoma diploide sexual de cv. Victoria se buscará caracterizar la especie a nivel genómico y realizar avances en la caracterización de los genes vinculados la calidad forrajera, estrés y modo reproductivo.

## Materiales y métodos

### Material vegetal

La secuenciación fue realizada a partir de un individuo del cultivar diploide ( $2n=2x=20$ ) sexual Victoria (planta UNST1122) mantenido en el invernáculo de CERZOS-CONCIET en Bahía Blanca. Este cultivar surgió de la regeneración del cultivo de inflorescencias inmaduras *in vitro* a partir del material apomítico facultativo tetraploide ( $2n=4x=40$ ) Tanganyika INTA (Cardone et al., 2006). Victoria no puede ser autopolinizada para obtener un genoma homocigoto por ser autoincompatible, por lo tanto se espera obtener un genoma heterocigoto por tener el background genético de Tanganyika INTA, el cual proviene de la selección de cultivares naturalizados en Argentina. Finalmente para validar los marcadores SSR se utilizaron los cultivares PI208214 (2x), PI299919 (2x), PI299920 (2x), PI299928 (2x), PI574506 (OTA, 4x), PI234217 (Tanganyika, 4x) del germoplasma de USDA y Tanganyika (4x), Don Walter (4x), Don Pablo (7x), Don Luis (7x) y Victoria (2x) del germoplasma de INTA

### Extracción de ADN

La extracción de ADN para la secuenciación del cultivar Victoria mediante PacBio, DArT y para la amplificación de SSR se realizó a partir de 80 mg de tejido de hoja utilizando en el método CTAB optimizado para *E. curvula* (Meier et al., 2011). El protocolo fue adaptado para obtener moléculas de ADN largas, teniendo especial cuidado en pasos críticos para evitar la fragmentación. De manera semejante se obtuvo el ADN de los cultivares utilizados para el análisis de marcadores SSR. La extracción de ADN de los marcadores DArT fue realizado con la misma técnica a partir de una población segregante de la cruce entre el cultivar tetraploide facultativo Don Walter y el cultivar sexual OTA donde se añadieron dos muestras del cv. Victoria para su secuenciación.

Para el diseño de la librería de Dovetail Chicago, hojas frescas de cv. Victoria fueron enviadas en hielo seco a Scotts Valley, California, Estados Unidos donde se extrajeron ~500 ng de ADN de alto peso molecular con una longitud media de 100 kb. El ADN fue cruzado y ligado para la reconstrucción *in vitro* de la cromatina, que luego fue fijada en formaldehído (Putnam et al., 2016). Para Hi-C la cromatina es fijada con formaldehído en el núcleo que luego es cruzada, ligada y extraída (Lieberman-Aiden et al., 2009). Las estrategias Chicago y Hi-C difieren en la molécula inicial, ya que la primera utiliza ADN de

alto peso molecular y, la segunda, cromatina endógena lo que varía la proximidad de la ligación de los fragmentos (cientos de bases en Chicago y miles de bases en Hi-C).

### **Preparación de las bibliotecas**

Tecnología PacBio: La secuenciación de lecturas largas del cultivar Victoria se realizó con la tecnología "Single Molecule Real Time" (SMRT), a través del sistema de secuenciación Sequel en el centro de investigación genética de la Universidad de Liverpool en Reino Unido. Luego de la evaluación de la calidad del ADN, se diseñaron dos bibliotecas con distintos insertos, una de 10 Kb y otra con 20 Kb. El tamaño inserto de ADN de cada bibliotecas se seleccionó con el selector de campo pulsante BluePippin (<https://sagescience.com/products/bluepippin/>). Luego de reparar los os extremos 3' y 5' y ligar los adaptadores el ADN fue sembrado en las celdas del secuenciador PacBio Sequel donde se encuentra fijada la enzima polimerasa (Eld et al., 2009).

Tecnología Hi-C y Chicago: Una vez que se produjo el cruzamiento y la fijación de la cromatina como se mencionó anteriormente se digirió con la enzima *DpnII* y el extremo 5' del corte se rellenó con nucleótidos biotinilados, luego los extremos libres fueron ligados. Finalmente, las ligaciones por proximidad de la cromatina fueron revertidas y se purificó el ADN para remover el sobrante de biotina. El ADN fue cortado en fragmentos de ~350 bases, se le ligaron los adaptadores Illumina en ambos extremos y se secuenciaron los fragmentos de ADN con Illumina HiSeq X pair-end de 150 bases (Putnam et al., 2016).

Tecnología DArT-seq: Para reducir la complejidad del ADN las muestras de cada individuo fueron digeridas con las enzimas *PstI* y *MseI*. Los extremos cohesivos de estos cortes fueron ligados con un adaptador Illumina y una secuencia de entre 5 y 8 bases llamada barcode que permite identificar a cada individuo luego de la secuenciación. En este caso se utilizaron 96 barcodes que corresponden uno a cada muestra de ADN diferente incluyendo dos del cultivar Victoria. Finalmente, estos fragmentos fueron secuenciados con Illumina single end de 77 bases. La preparación de librerías, el análisis bioinformático y el SNP calling (determinación de la posición de los marcadores SNP) fueron provistos por el laboratorio de servicios de análisis genéticos en CIMMYT (México) (Ren et al., 2015).

### **Ensamblado**

Una de las ventajas de la tecnología PacBio y de los softwares disponibles para su ensamblado es que no es necesario un filtrado o tratamiento previo al ensamblado, por lo

que se utilizan las lecturas crudas tal cual son enviadas por el proveedor de secuenciación. Las lecturas de PacBio fueron ensambladas con los programas Canu (Koren et al., 2017) y FALCON (Chin et al., 2013) con los parámetros descritos en la tabla 2.1. Para determinar la mejor versión del ensamblado final, se evaluaron los siguientes indicadores: N50, bases totales ensambladas, número de contigs y tamaño promedio de contig. Además, para evaluar si el ensamblado estaba completo en términos de genes, se utilizó el software BUSCO (Seppey et al., 2019). Este software cuenta con una base de datos de genes de referencia que se encuentran en el 99% de las especies vegetales en una sola copia y, al contrastarlos con un genoma o transcriptoma, arroja el porcentaje genes completos (en una y múltiples copias), genes fragmentados y genes no encontrados que se han alcanzado con la secuenciación (Seppey et al., 2019).

<b>Version</b>	<b>Length_cutoff</b>	<b>Length_cutoff_pr</b>
Falconv5.4	5 kb	7,5 kb
Falconv5.3	5 kb	7,5kb
Falconv5.2	5 kb	7,5 kb mincov5
Falconv5.1	5 kb	5 kb mincov8
Falconv5.0	5 kb	5 kb
Falconv4.5	7,5 kb	7,5 kb mincov5
Falconv4.4	7,5 kb	7,5 kb mincov5
Falconv4.3	7,5 kb	5 kb micov5
Falconv4.1	7,5 kb	5 kb
Falconv4.2	7,5 kb	7,5 kb
Falconv3.1	10 kb	5 kb
Falconv3.2	10 kb	7,5 kb
Falconv2.1	12 kb	5 kb
Falconv2.2	12 kb	7,5 kb
Falconv1.1	16,7 kb	1 kb
Falconv1.2	16,7 kb	2,5kb
Falconv1.3	16,7 kb	5 kb
Falconv1.4	16,7 kb	7.5 kb
Falconv1.5	16,7 kb	10 kb
<b>Version</b>	<b>ErrorRate</b>	<b>minOverlap</b>
Canuv4	0,03	1000 bp
Canuv3	0,01	500 bp
Canuv2	0,045	1000 bp

Version	Length_cutoff	Length_cutoff_pr
Canuv1	0,045	500 bp

Tabla 2.1: Parámetros utilizados para las distintas versiones del ensamblado de las lecturas crudas del genoma del cultivar cv. Victoria. El ensamblado se continuo con la versión que resulto con mayor N50 y genes BUSCO y menor cantidad de contigs. Length\_cutoff: tamaño de lectura de minino utilizado para la corrección del ensamblado. Length\_cutoff\_pr: Tamaño de lectura de minino para pre ensamblado. ErrorRate: tasa de error tolerada para ensamblar. minOverlap: solapamiento minino para el ensamblado.

Debido a que la lecturas de PacBio tienen aproximadamente una tasa de error del 15%, una vez ensambladas es necesario hacer un pulido (Weirather et al., 2017). Para esto, se alinearon las lecturas crudas con el ensamblado utilizando el software Pbalgn (<https://github.com/PacificBiosciences/pbalgn>) con los parámetros por default y luego se realizó el pulido con el software Arrow (<https://github.com/PacificBiosciences/GenomicConsensus>), también con los parámetros por default. Este software utiliza el alineamiento de Pbalgn y contrasta cada base en el ensamblado contra las bases que alinearon en esa posición seleccionando la de mayor cobertura.

Para obtener el ensamblado final, en pasos consecutivos se integraron las librerías de Chicago y Hi-C con el software Dovetail Hi-Rise (Koch et al., 2016) al ensamblado de PacBio. Este proceso comienza mapeando las lecturas de Chicago al ensamblado PacBio y cuando uno de los pares de cada lectura es mapeada en un contig distinto y esto sucede además con otros pares, se unen los dos contigs formando un scaffold (unión de contigs previamente ensamblados por medio de alguna evidencia que valide que se encuentran adyacentes). Finalmente, se realiza el mismo procedimiento con la librería de Hi-C y los scaffolds de Chicago+Pacbio obteniéndose el ensamblado final. En la Figura 2.1 se esquematizan mediante un diagrama de flujo los programas y los datos utilizados durante el ensamblado.

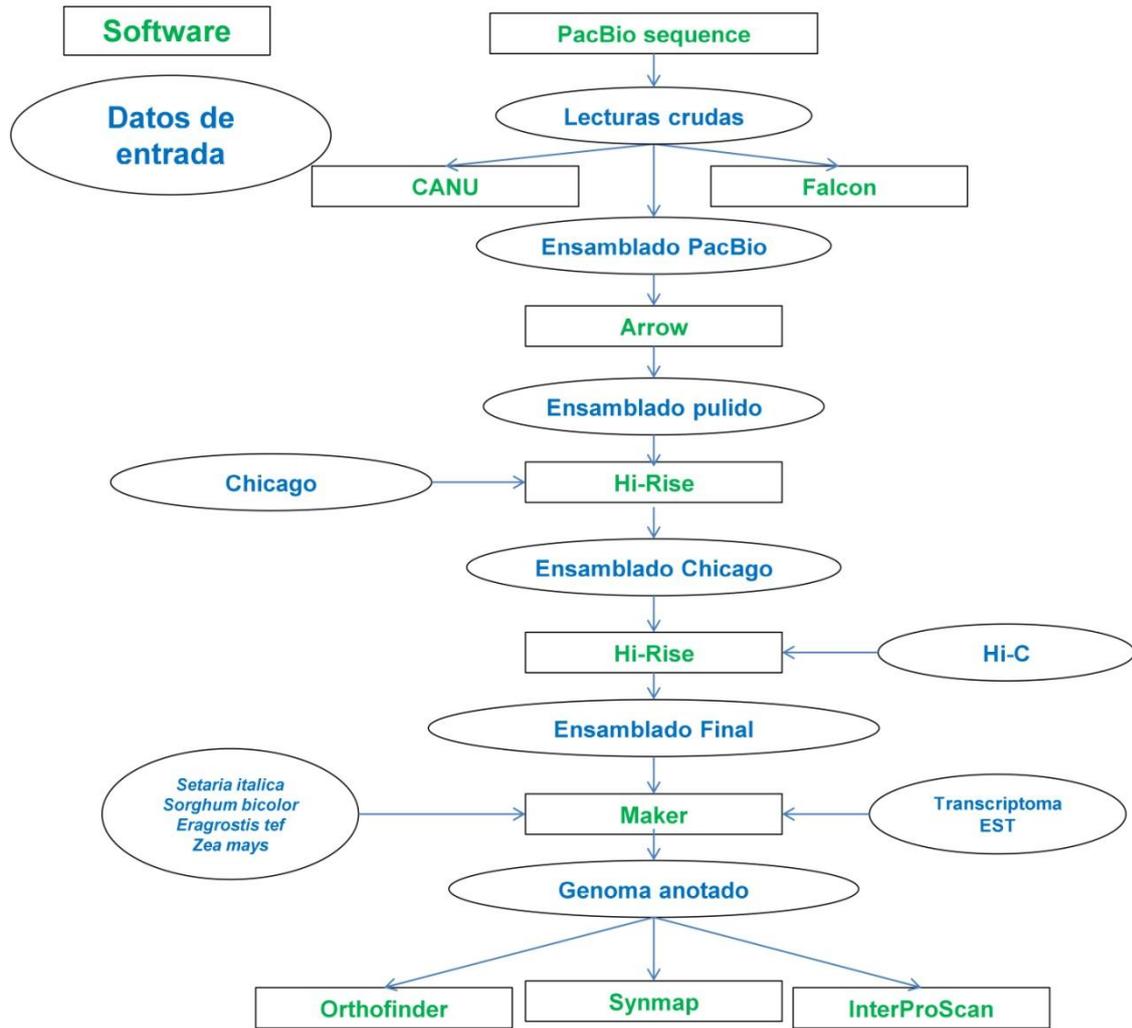


Figura 2.1: Proceso de ensamblado y anotación del genoma de Victoria. En verde están indicados los programas utilizados y en azul los datos de entrada.

## Validación por DArT-seq

Con el objetivo de validar el ensamblado final con otra fuente de secuenciación, las lecturas de DArT-seq correspondientes a las muestras de Victoria fueron mapeadas con el software Bowtie (Langmead et al., 2012) contra el genoma ensamblado utilizando los parámetros “end-to-end” y “-k 10”. El mapeo de las lecturas fue graficado con una ventana de 500 kb para identificar regiones con mayor densidad de marcadores.

### **Anotación de elementos repetitivos**

Para la identificación de elementos repetitivos se creó una biblioteca que contiene elementos identificados *de novo* con los softwares RepeatModeller (Smit et al., 2008) y TransposonePSI (Altschul y Koonin, 1998) y elementos identificados por homología con la base de datos RepBase23 (Boa et al., 2015) mediante el software RepeatMasker (Smit et al., 2013). Cuando la misma secuencia fue identificada por los dos softwares, los elementos *de novo* fueron combinados con el programa USEARCHv7 (Edgar, 2010) y luego fueron clasificados con el programa RepeatClassifier (Smit et al., 2008). Finalmente, RepeatMasker fue utilizado con los parámetros “-s” y “-lib” con las bases de datos construidas previamente para identificar, cuantificar y enmascarar los elementos repetitivos.

### **Anotación de genes**

La anotación de los genes del genoma del cultivar Victoria se hizo sobre el ensamblado enmascarado (reemplazo de nucleótidos por “Ns” en las regiones donde se encuentran elementos repetitivos) para evitar la anotación de genes relacionados con elementos repetitivos. La búsqueda de los genes se realizó con el software MAKER (Campbell et al., 2014). Este programa actúa “dirigiendo” a otros programas con el fin de identificar genes en base a la homología con proteínas y transcriptomas y en base a *Ab initio* (Figura 2.2). El término *Ab initio* hace referencia a el modelado de genes por medio de algoritmos que buscan características estructurales como codones start y stop y patrones específicos en los intrones y exones de los genes. De esta forma se ejecuta MAKER tres veces, la primera buscando genes por homología y las otras dos ejecuciones son iteraciones *Ab initio*. El modelado de genes por homología se realiza por medio dos programas, Exonerate (Slater et al., 2005) para el alineamiento de transcriptos y BLASTP (Altschul et al., 1990) para el alineamiento de proteínas. En este caso, se utilizaron transcriptomas previamente secuenciados de distintos órganos y cultivares de *E. curvula* (Garbus et al., 2019, Cervigni et al., 2008) y proteínas de *E. tef* (Cannarozzi et al., 2014), *Setaria italica* (Zhang et al., 2012), *Sorghum bicolor* (Paterson et al., 2009) y *Z. mays* (Jiao et al., 2017). Para el

modelado *Ab initio* MAKER utiliza los programas Augustus (Stanke et al., 2006) y SNAP (Johnson et al., 2008) que necesitan ser entrenados con genes de alta confianza como los genes BUSCO. Finalmente MAKER compara los modelos de genes (*Ab initio* y por homología) en cada sitio y toma el de mayor confianza.

La clasificación de la anotación de los genes en base a sus dominios y funciones se realizó con el software InterProScan versión 5 (Finn et al., 2017) el cual tiene integrado 14 bases de datos con distintas características. La salida de este programa fue introducida en el programa Blast2go (Conesa y Götz, 2008) para hacer la clasificación en base a los términos de Gene Ontology (<http://geneontology.org>).

La anotación se encuentra disponible en la plataforma online ensembl ([http://plants.ensembl.org/Eragrostis\\_curvula](http://plants.ensembl.org/Eragrostis_curvula)) y en NCBI bajo al acceso GCA\_007726485.1.

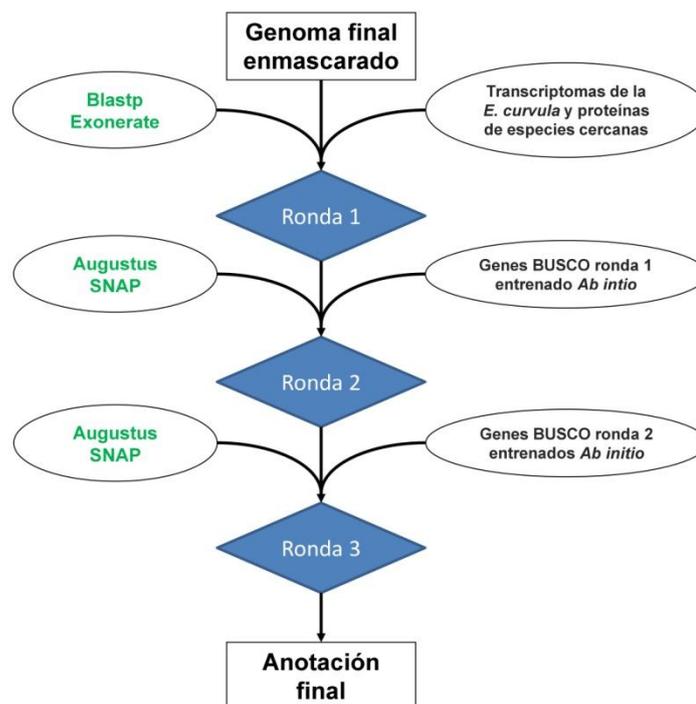


Figura 2.2: Pasos para la anotación final de los genes de cv. Victoria dirigida por el software MAKER en tres rondas sucesivas. A la izquierda en verde se encuentran los programas utilizados en cada ronda y a la derecha la información utilizada para el modelado de los genes. En la anotación final se obtiene el modelo con mayor confianza para cada sitio.

## Análisis de sintenia evolución de *E. curvula*

La conservación del orden de los genes dentro de bloques cuando se comparan dos cromosomas se denomina sintenia, es decir, dos genomas evolutivamente cercanos son más sinténicos que uno que no lo es (Stein et al., 2013). Por medio del software SyMAP (Soderlund et al., 2011) se analizó la sintenia entre el ensamblado de *E. curvula* y el de *E. tef* (VanBuren et al., 2019), *S. bicolor* (Paterson et al., 2009), *Z. mays* (Jiao et al., 2017), *O. sativa* (Kawahara et al., 2013) y *Oropetium thomaem* (VanBuren et al., 2015).

Los genes sinténicos (genes ortólogos dentro de un bloque sinténico) fueron listados de a pares y se calculó la tasa de sustitución sinónima (Ks) con BioPerl utilizando el método de Nei–Gojobori. Finalmente para identificar eventos de duplicación genómica, se graficó y corrigió el Ks por el método de Wang (Wang et al., 2015).

La relación evolutiva de Victoria dentro de la familia poaceae se evaluó mediante la homología con proteínas que surgieron de las anotaciones de las especies *S. italica*, *Triticum aestivum*, *E. tef*, *O. sativa*, *Z. mays* (Jiao et al., 2017), *S. bicolor* (Paterson et al., 2009), *Brachypodium distachyon* (Vogel et al., 2010), *O. thomaeum* (VanBuren et al., 2015), *Musa itinerans* (Wu et al., 2016) y *Panicum hallii* (Lovell et al., 2016). Para identificar los genes ortólogos entre todas estas especies se utilizó el programa OrthoFinder (Emms et al., 2015). Con el alineamiento de las proteínas de este programa se creó un árbol filogenético utilizando el modelo de Jones Taylor Thornton con el software MEGA X (Kumar et al., 2018) La calibración de tiempo de divergencia fue realizado por el método de Prasad (Prasad et al., 2005) donde se estima que la divergencia entre *Z. mays* y *O. sativa* es de 70 millones de años y *M. itinerans* fue utilizada como “outgroup” (especie lo suficientemente cercana para tener genes ortólogos pero que no agrupe con ninguna otra del análisis). La expansión y contracción de las familias de genes se calculó con el software CAFEv3.0 (Han et al., 2013) con múltiple “birth-death” lambda con el objetivo de tener en cuenta las duplicaciones genómicas a lo largo de la evolución.

La identificación de SSR en el genoma de Victoria se realizó con el software SSR locator (Da Maia et al., 2008) con el fin, por un lado de validar el ensamblado y, por otro, de evaluar la relación filogenética entre distintos cultivares. Así con la herramienta Primer3 (Koressaar et al., 2007), se diseñaron 15 pares de primers seleccionados al azar que fueron amplificados sobre 10 genotipos. Para realizar el análisis filogenético se incluyeron marcadores desarrollados en trabajos previos para aumentar la precisión del análisis (Garbus et al., 2017). En las reacciones de PCR, realizadas por el grupo de trabajo del CERZOS, se utilizó 1 µl de 10 mM dNTPs mix, 1X de solución buffer, 0.5 µl de cada primer (100 pmol/µl), 1,6 U Taq polimerasa y 60 ng de ADN genómico en 20 µl en volumen de reacción. El programa de PCR consistió en una desnaturalización inicial a 94C° durante 3

minutos seguido de 40 ciclos a 94 C° por 30 seg a la temperatura óptima de annealing para cada par de primers, 72 C° por 30 seg, 72 C° de extensión final por 5 min. La validación en Victoria se realizó sobre gel de agarosa 1.5% y la presencia y ausencia de las bandas se determinó sobre gel de poliacrilamida al 6%.

### **Identificación de genes relacionados con la tolerancia al estrés por sequia**

Debido a que los genes WRKY representan la familia de genes más estudiada relacionada a la tolerancia al estrés por sequía, se decidió analizar esta familia en *E. curvula*. De acuerdo a su estructura, los genes fueron clasificados en grupo I, II y III y en los subgrupos (IIa, IIb, IIc, IId and IIe) (Eulgem et al., 2007). Primero se buscaron aquellos genes que contienen el dominio WRKY en la anotación Pfam incluida en las bases de datos de InterProScan. Luego con la herramienta MEME (Bailey et al., 2009), se identificó el número de dominios WRKY y fueron clasificados en el grupo I los genes con dos dominios WRKY; en el grupo II aquellos genes con un solo dominio WRKY y el dominio “zink finger” terminal H-X<sub>1</sub>-H; en el grupo III genes con un solo dominio WRKY y el dominio “zink finger” terminal H-X<sub>1</sub>-C. Los subgrupos del grupo II fueron clasificados de acuerdo a los restantes dominios. Finalmente a través de MUSCLE (Edgar et al., 2004), se realizó un alineamiento múltiple y se construyó un árbol filogenético con el modelo de “maximum likelihood” con el software MEGA X.

### **Identificación de genes relacionados con la calidad forrajera**

Los genes que participan en la vía de la lignina fueron extraídos desde la anotación KEGG (Kanehisa et al., 2019) incluida en InterProScan. La base de datos KEGG contiene el esquema de la vía completa pero los genes de referencia para algunas especies como maíz y arroz no están presentes. Por este motivo, que no esté resaltado en la vía no necesariamente significa que no esté en la especie. Recientemente fue descubierto que el silenciamiento de la enzima caffeoyl shikimate esterases (CSE) disminuye el contenido de lignina (Vanholme et al., 2013) en *A. thaliana*. Existen dos clases de genes que codifican la para la proteína CSE, la clase II está ampliamente distribuida en todo el reino vegetal mientras que la clase I solo se encuentra en algunas especies. Estos genes fueron buscados e identificados en la anotación de Victoria y a partir de ellos fueron diseñados primers de PCR para su amplificación en *E. tef* y *E. curvula*. El programa utilizado consistió en una desnaturalización a 94 C° durante 2 minutos, seguido de 38 ciclos a 94 C° por 15 seg, 50 C° por 20 seg y 72 C° por 80 seg y una extensión final de 5 min a 72 C°. Los

amplicones fueron analizados en gel de agarosa al 1.5 % y clonados en el vector pGEM-T Easy (Promega). La secuencia resultante fue alineada con otras especies.

### **Identificación de genes relacionados a las vías reproductivas**

Para la identificación de los genes de *Victoria* involucrados en la reproducción se recurrió a la información transcriptómica disponible en *O. sativa* antes y después de la doble fertilización (Anderson et al., 2017, Anderson et al., 2013). El primer paso consistió en detectar todos los genes ortólogos con alto grado de confianza entre *O. sativa* y *Victoria* con la plataforma Ensembl (<http://plants.ensembl.org>). Por otro lado se mapearon al ensamblado del genoma las lecturas previamente secuenciadas de un cultivar 100% apomíctico y otro sexual (Garbus et al., 2017) con el software Hisat2 (Kim et al., 2016) con los parámetros "--very-sensitive, --dta y -a". Luego con la el paquete de R "GenomicFeatures" (Lawrence et al. 2013) se detectaron las regiones en las cuales alineaban las lecturas (exón, intron, CDS, UTR, etc.) y por medio de paquete "GenomicAlignments" (Lawrence et al. 2013) se construyó una matriz con las lecturas que mapearon en transcritos. Finalmente con el paquete EdgeR (Robinson et al., 2011) se normalizó la matriz y se realizó el análisis de expresión diferencial tomando como sobre expresados y reprimidos aquellos transcritos con un fold change mayor o menor a 2 (equivalente a log2 fold change mayor y menor que 1) respectivamente y un FDR (tasa de falsos positivos) menor a 0,05 (Benjamini et al., 2000). Finalmente, se analizó la expresión temporal y espacial los genes sobreexpresados y reprimidos en apomixis que tenían genes ortólogos expresados durante la reproducción de *O. sativa*.

## Resultados

### Ensamblado y validación del genomas de Victoria

Luego de la extracción de ADN genómico en el laboratorio de CERZOS-CONICET, dos muestras ADN del cultivar Victoria fueron enviadas para su secuenciación a la Universidad de Liverpool. Inicialmente el servicio de secuenciación analizó la calidad por medio del equipo *fragment analyzer* indicando que las muestras tenían una longitud máxima de 41.800 y 44.178 pb (Figura 2.3) y una concentración 0,78 y 0,62  $\mu\text{g}/\mu\text{l}$  respectivamente. Tanto la cantidad como la longitud del ADN estuvieron dentro del rango de valores recomendados, por lo tanto se prosiguió con la construcción de dos genotecas. La primera fue diseñada con un inserto de 10 kb y se obtuvieron 6.223.627 lecturas con una longitud promedio de 5.296 pb mientras que en la segunda el inserto fue de 20 kb, la cantidad de lecturas fueron 3.309.811 con una longitud promedio fue 7.018 pb. El total de bases obtenidas (56.186 millones de bases) correspondió a una cobertura de aproximadamente 90X de la longitud del genoma haploide.

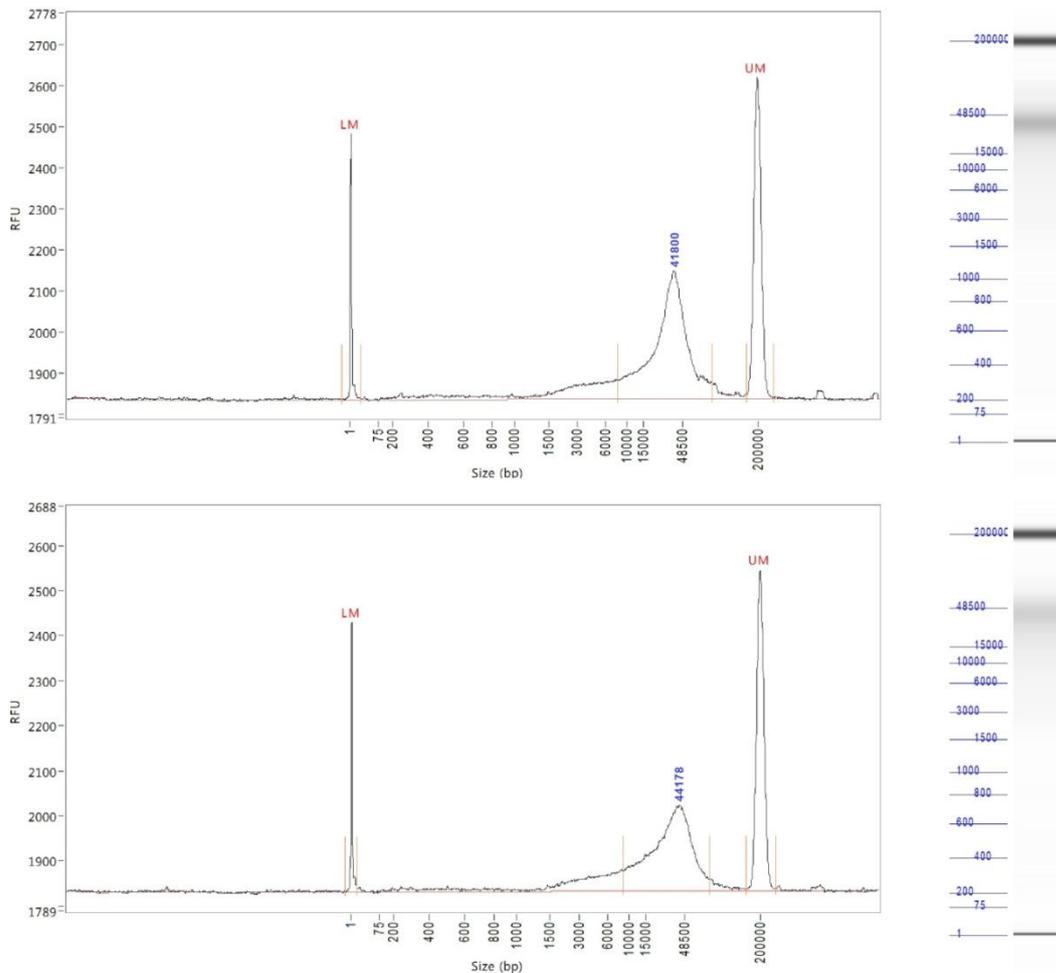


Figura 2.3: Análisis de calidad de las dos muestras de ADN extraído para la secuenciación del genoma del cultivar Victoria. En el eje Y se encuentran las unidades de fluorescencia relativa y en el X el tamaño de fragmento de ADN. LM (lower marker) y UM (upper marker) corresponden a marcadores de 1 y 200.000 pb respectivamente. El pico (41.800 y 44.176 pb) representa la longitud en la cual se encuentra la máxima cantidad de ADN.

El ensamblado de estas lecturas fue llevado a cabo con el programa FALCON utilizando los parámetros descritos en la versión falconv4.5 (tabla 2.1), que fue la que permitió obtener el mayor N50, el menor número de contigs y arrojó la mayor cantidad de genes completos BUSCO. Este ensamblado produjo una longitud de 601 Mb distribuida en 3.516 contigs con un N50 de 378.697 pb, una longitud promedio de 171.108,24 pb y 88,4% de los genes esperados por BUSCO (tabla 2.2). En la próxima etapa se realizó el pulido con el programa Arrow y el principal cambio que surgió de este proceso fue el aumento en el porcentaje de genes BUSCO de 88,4 a 96 %. El ensamblado de las lecturas de PacBio con FALCON y pulido con Arrow junto con las 432 millones reads pair-end 2x150 bp obtenidas en la secuenciación de Chicago correspondió a la entrada para el primer “scaffolding” realizado con el programa Hi-RISE. En esta etapa mejoraron ampliamente las métricas debido

principalmente al aumento del N50 a 791.258 pb y la disminución de la cantidad de contigs a 1.884. Para obtener el ensamblado final se integraron 333 millones de lecturas pair-end 2x150 pb secuenciadas con la metodología Hi-C. Si bien el número de contigs disminuyó a 1.143 el N50 alcanzó una escala cromosómica incrementando hasta 43.411.000 bp. El resultado final del análisis BUSCO arrojó 96,4% de genes completos de los cuales el 80,4% fueron de copia simple, 1,2% fragmentados y solo 2,6% ausentes. Todas las métricas evaluadas en la tabla 2.2 demostraron que el genoma está prácticamente completo, que contiene contigs de tamaño cromosómico y que en dentro de ellos se encuentra casi el total de los genes de la especie, por lo tanto, representando un genoma de excelente calidad. En la Figura 2.1 se puede observar cómo se fueron incorporando las distintas secuenciaciones y los programas utilizados para el ensamblado.

	<b>PacBio</b>	<b>Pulido</b>	<b>Chicago</b>	<b>Chicago + Hi-C</b>
<b>Longitud pb</b>	601.616.585	600.872.314	602.350.000	602.432.814
<b>N50 pb</b>	378.697	380.299	791.258	43.411.000
<b>Nro. contigs</b>	3.516	3.118	1.884	1.143
<b>Promedio pb</b>	171.108.24	192.710.81	319.718.81	527.062.82
<b>BUSCO</b>	C:88,4%	C:96,0%	C:96,1%	C:96,4%

Tabla 2.2: Métricas evaluadas para la determinación de la calidad de los distintos ensamblados de *E. curvula*. Longitud en pares de bases, N50 en pares de bases, número final de contigs en cada ensamblado, longitud promedio de los contigs y genes completos (C) BUSCO,

Con el objetivo de validar el ensamblado con una fuente de secuenciación externa, dos muestras del cultivar Victoria fueron sometidas al protocolo de DArT-seq resultando en 245.626 y 246.752 “single end” lecturas de 69 pb. Luego del análisis bioinformático realizado por el proveedor en CIMMyT (México) se obtuvieron 6.307 marcadores SNP. El 95,5% del total de los marcadores obtenidos pudieron ser mapeados al genoma ensamblado y, como era de esperarse, los contigs más largos contienen mayor cantidad de marcadores (Tabla 2.3, Figura 2.4). Sin embargo, la distribución de este mapeo no fue homogénea encontrándose en los siete contigs más largos mayor cantidad de marcadores en las regiones distales 5' y 3' (Figura 2.5). Debido a que tecnología DArT-seq está diseñada para evitar regiones repetitivas y apuntar principalmente a genes, se podría hipotetizar que los siete contigs de mayor longitud corresponden a cromosomas ya que la distribución del mapeo tiene una tendencia de parábola invertida con menor cantidad de marcadores en el centro.

<b>Nombre</b>	<b>Longitud</b>	<b>Genes</b>	<b>DArT</b>
<b>Contig10</b>	59.696.155	6.315	694
<b>Contig12</b>	57.091.500	4.450	489

<b>Contig38</b>	57.060.119	5.759	623
<b>Contig6</b>	55.809.479	5.036	550
<b>Contig3</b>	45.712.337	4.943	440
<b>Contig1</b>	43.444.889	3.164	278
<b>Contig28</b>	42.455.315	3.888	397
<b>Contig30</b>	29.802.391	3.198	447
<b>Contig8</b>	28.838.055	3.393	326
<b>Contig50</b>	26.185.977	2.345	247
<b>Contig25</b>	23.821.556	2.503	286
<b>Contig4</b>	21.583.994	1.943	249
<b>Contig44</b>	8.977.321	705	75
<b>Contig33</b>	4.567.900	320	25

Tabla 2.3: Catorce contigs con mayor longitud del ensamblado del genoma del cultivar Victoria con el número de genes anotados y el número de marcadores DArT localizados en cada contig.

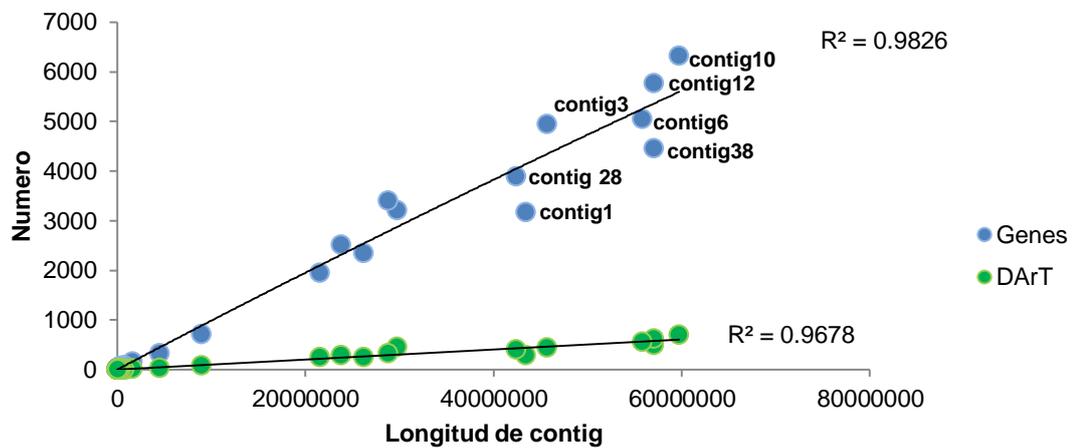


Figura 2.4: Gráfico de la regresión entre el número de genes (círculos celestes) y marcadores DArT (círculos verdes) en cada contig con la longitud de los mismos. Los siete puntos más distantes corresponderían a cromosomas.

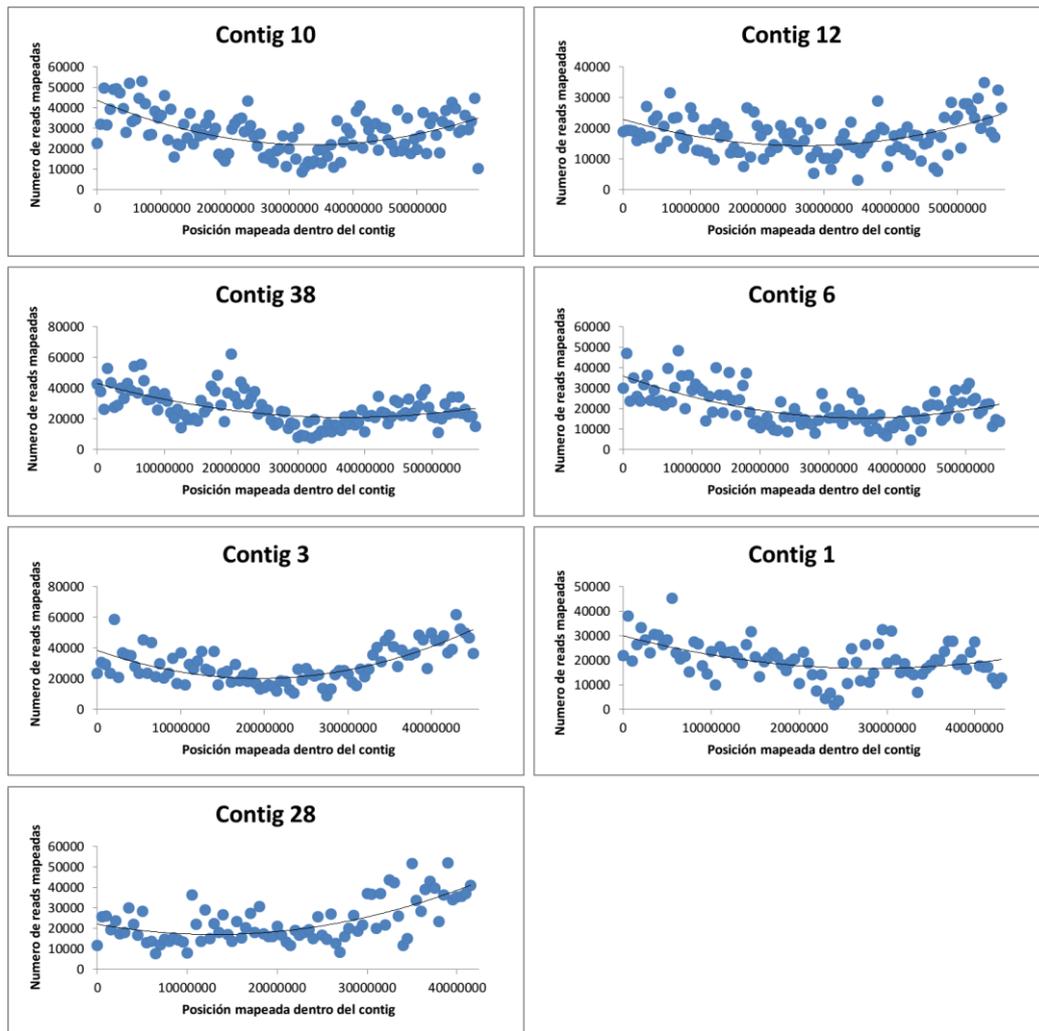


Figura 2.5: Distribución de los marcadores SNP a lo largo de los siete contigs. Cada círculo azul representa 500 marcadores en una ventana de 500 kb.

### Anotación de elementos repetitivos y genes

El análisis para la identificación de elementos repetitivos arrojó que el 28,7% del genoma ensamblado se encuentra cubierto por estas estructuras (Figura 2.6). Al igual que la mayoría de las gramíneas los retrotransposones (RT) fueron el principal elemento identificado seguido por elementos no clasificados identificados *de novo*. Dentro de los RTs, los más representados fueron los LTRs de las familias *Gypsy* y *Copia* con un 13,62% y 3,14% respectivamente que corresponde a un índice 4,3:1. Este valor se encuentra muy aproximado al informado para *E. tef* (4,27:1) (Gebre et al., 2016) y superior al de *S. italica* (3,08:1), *Z.mays* (1,91:1) y *S. bicolor* (3,67:1) (Zhang et al., 2012, Jiao et al., 2017, Paterson et al., 2009).

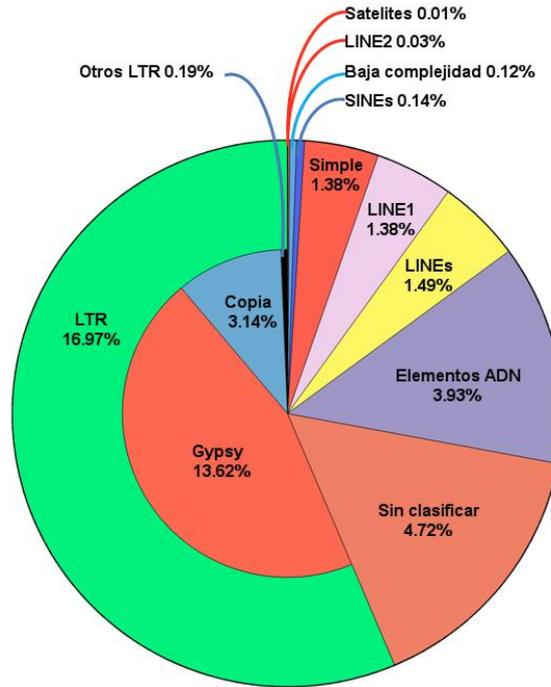


Figura 2.6: Elementos repetitivos identificados en el genoma ensamblado del cultivar Victoria. En verde se observan los más abundantes, los LTRs, y en el círculo interno la proporción de LTR clasificados como *Gypsy* y *Copia*.

Previo a la anotación de los genes, las repeticiones identificadas en el genoma fueron enmascaradas para evitar la anotación de genes vinculados a elementos repetitivos y para realizar más eficientemente el proceso. Posteriormente, luego de tres iteraciones con el programa MAKER (Figura 2.2) se obtuvieron 56.469 genes con una longitud promedio de 1.424 pb y 93,4% de los genes BUSCO completos. La anotación funcional de estos genes se realizó por medio del software InterProScan, encontrando 35.713 genes con homología en al menos una base de datos (tabla 2.4). Además se realizó un alineamiento con la base de datos TrEMBL encontrando homología con 18.955 genes. En la base de datos GO se encontraron 29.429 genes vinculados con términos GO. En las categorías principales de GO, i.e., componentes celulares (Cellular component), función molecular (molecular function), y proceso biológico (biological process) se encontraron 16.710, 33.820 y 33.601 genes respectivamente. Es importante remarcar que un mismo gen puede tener múltiples términos, incluso de la misma categoría principal. Cada categoría principal fue clasificada en el nivel 2 según gene ontology con el programa blast2go (Figura 2.7).

<b>Bases de datos InterProScan</b>	
<b>Base de datos</b>	<b>Genes con homología</b>
<b>SFLD</b>	292
<b>ProDom</b>	539
<b>Hamap</b>	1.008
<b>PIRSF</b>	2.215
<b>TIGRFAM</b>	3.353
<b>PRINTS</b>	5.834
<b>Coils</b>	6.971
<b>ProSitePatterns</b>	7.657
<b>SMART</b>	12.257
<b>CDD</b>	13.001
<b>ProSiteProfiles</b>	15.495
<b>MobiDBLite</b>	26.335
<b>SUPERFAMILY</b>	29.770
<b>Gene3D</b>	29.980
<b>Pfam</b>	35.713
<b>PANTHER</b>	44.532
<b>Kegg</b>	10.005
<b>GO</b>	29.429
<b>Base de datos TrEMBL</b>	
<b>TrEMBL</b>	18.955

Tabla 2.4: Número de genes con homología en las bases de datos incluidas en el InterProScan y TrEMBL.

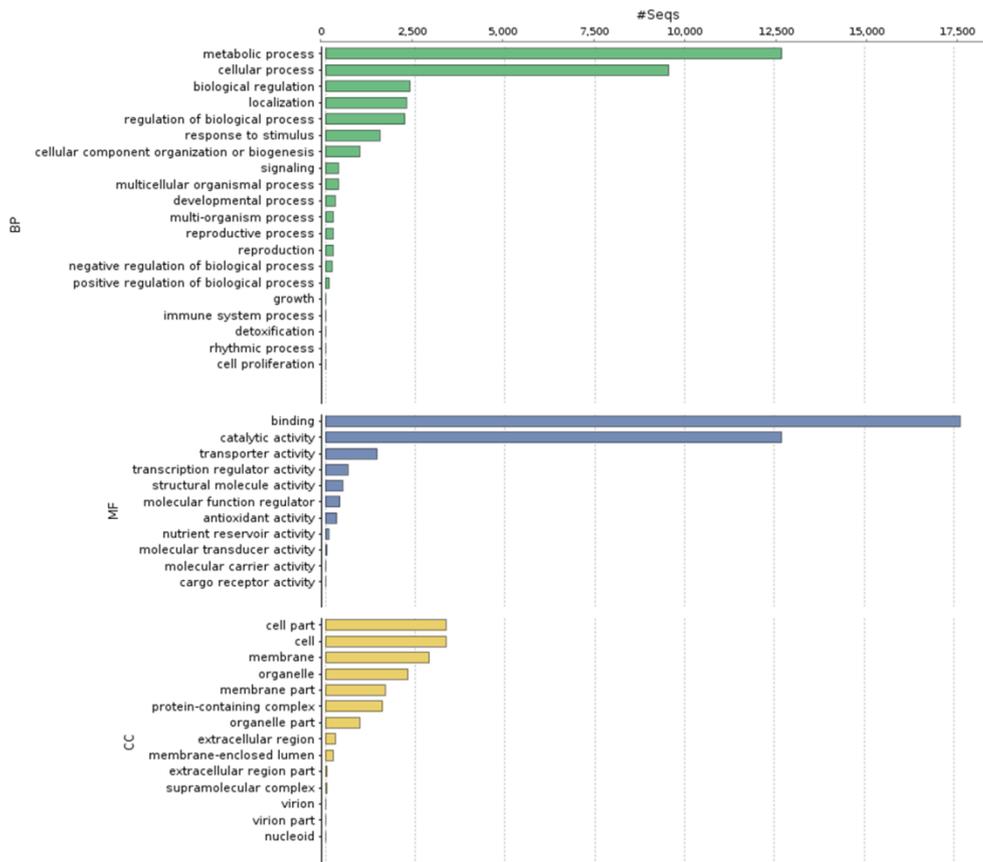


Figura 2.7: distribución de los genes en los términos de cada categoría principal. BP: biological Process; MF: Molecular Function; CC: Cellular component. La clasificación se realizó en el nivel 2 con el programa Blas2go.

## Evolución de *E. curvula* dentro de la familia de las Poaceae

Dentro de la familia de las poaceae existe un evento de duplicación común a todas las especies denominado paleoduplicación (Kishimoto et al., 1994). Por medio del análisis de Ks (tasa de sustitución sinónima de nucleótidos) es posible aproximar el tiempo de esta paleoduplicación, el tiempo de duplicaciones específicas de cada especie y la relación evolutiva entre especies basado en los cambios que ocurren en los genes en términos de Ks tomando como referencia evidencias fósiles que datan la diversificación de *O. sativa* hace 70 millones de años (M.A.) (Prasad et al., 2005). De esta manera, haciendo comparación entre pares de genes ortólogos, se pudo determinar la divergencia de *E. curvula* con *O. sativa*, *Z. mays*, *S. bicolor*, *O. thomaeum* y *E. tef*. Así, se comprobó que *E. tef* es la especie que tiene la mayor cantidad de pares de genes ortólogos a *E. curvula*, mostrando el Ks más bajo entre las especies comparadas y demostrando que se habría diversificado hace 14,5 M.A. (Figura 2.8 a). De las especies analizadas la primera que se diversificó fue, como era esperado, *O. sativa* hace 70 M.A., luego le siguieron *Z. mays* y *S. bicolor* contemporáneamente hace 50 M.A. y finalmente una de las especies más lejanas a

*E. curvula* de la subfamilia Chloridoideae, *O. thomaeum* hace 31 M.A. El análisis de Ks entre los pares de genes homólogos dentro de cada especie permitió estimar la paleoduplicación y las duplicaciones particulares de cada especie (Figura 2.8 b). El aumento de los pares de genes con Ks alrededor de 0.7 indicó que la duplicación de todas las gramíneas habría ocurrido hace 80-90 M.A., (figura 2.8 b gWGD) la de *Z. mays* hace 20 M.A., mientras que las de *E. tef* y *E. curvula* hace 9,6 y 4,5 M.A., respectivamente. A pesar de haber tenido una duplicación *Z. mays* continua en estado diploide ( $2n=2x=20$ ) debido a que esta duplicación fue segmental, es decir, se duplicaron sus cromosomas y se reensamblaron quedando 2 juegos de cromosomas con mayor longitud (Gaut et al., 1997). Por otro lado, la duplicación de *E. tef* tiene un origen híbrido, se cruzaron dos especies compatibles y se duplicó la cantidad de cromosomas formando un genotipo alotetraploide ( $2n=4x=40$ ) (Ingram et al., 2003).

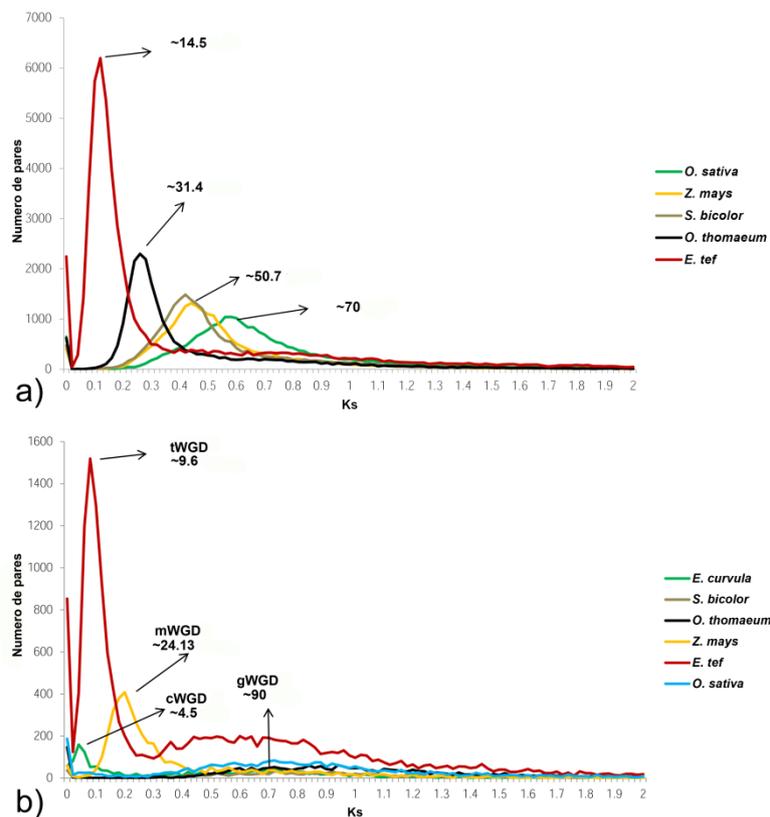


Figura 2.8: Distribución de Ks dentro de las especies seleccionadas. a) Ks entre *E. curvula* y las especies detalladas en cada pico indica en millones de años el tiempo de diversificación. B) Análisis de genes homólogos dentro de cada especie; en cada pico se encuentra el número en millones de años desde la duplicación genómica de todo su genoma. Ks: tasa de sustitución sinónima; tWGD: duplicación de *E. tef*, mWGD: duplicación de *Z. mays*, cWGD: duplicación de *E. curvula*, gWGD: duplicación genómica compartida por todas las gramíneas.

Complementario al análisis de Ks, se realizó un árbol filogenético, construido con el método de máxima verosimilitud (maximum likelihood) con el programa OrthoFinder, utilizando 1.185 grupos ortólogos los cuales contienen genes en una o dos copias en todas las especies analizadas (Figura 2.9). El número total de grupos ortólogos detectados fue 24.747, de los cuales 9.189 fueron comunes a todas especies consideradas en la Figura 2.9. Con el fin de evaluar el número de orthogroups dentro de cada subfamilia se realizó un diagrama de Venn con todas las especies analizadas dentro de las Chloridoideae, Panicoideae, Pooideae y Ehrhartoideae (Figura 2.10). Así el número de grupos ortólogos compartidos por la subfamilia Chloridoideae fueron 12.940 mientras que las Panicoideae y Pooideae fueron 13.621 y 15.532 respectivamente. Dentro de las subfamilias C4 (Chloridoideae y Panicoideae), 11.570 orthogroups fueron compartidos por estas especies y 2.051 fueron específicos de las Panicoideae y 1.370 de las Chloridoideae. En un análisis homólogo en las subfamilias C3 se encontraron 14.203 orthogroups comunes, 1.329 específicos de Pooideae y 2.601 de Ehrhartoideae.

El análisis anterior es considerado estático porque simplemente compara la relaciones de los las familias de genes (grupos ortólogos) entre las especies. Sin embargo, es posible calcular las contracciones y expansiones de los genes en cada nodo por medio del software CAFÉ. De esta manera se puede observar que luego de la duplicación compartida por todas las especies, las contracciones en la mayoría de los casos fueron superiores a las expansiones, es decir, las especies fueron perdiendo genes en el camino hacia la especificación (Figura 2.9). En el caso de *Z. mays* y *E. curvula* puede observarse que en el nodo final las expansiones fueron significativamente superiores a las contracciones, esto podría deberse a que ambos genotipos se encuentran en estado diploide y sufrieron una duplicación. Además puede observarse en *O. thomaeum*, la especie con menor tamaño de genoma dentro de las gramíneas, que el grado de contracción a lo largo de su historia fue siempre mayor que al de expansión y su tasa final de contracción:expansión fue de 7,14, mayor que todas las especies analizadas.

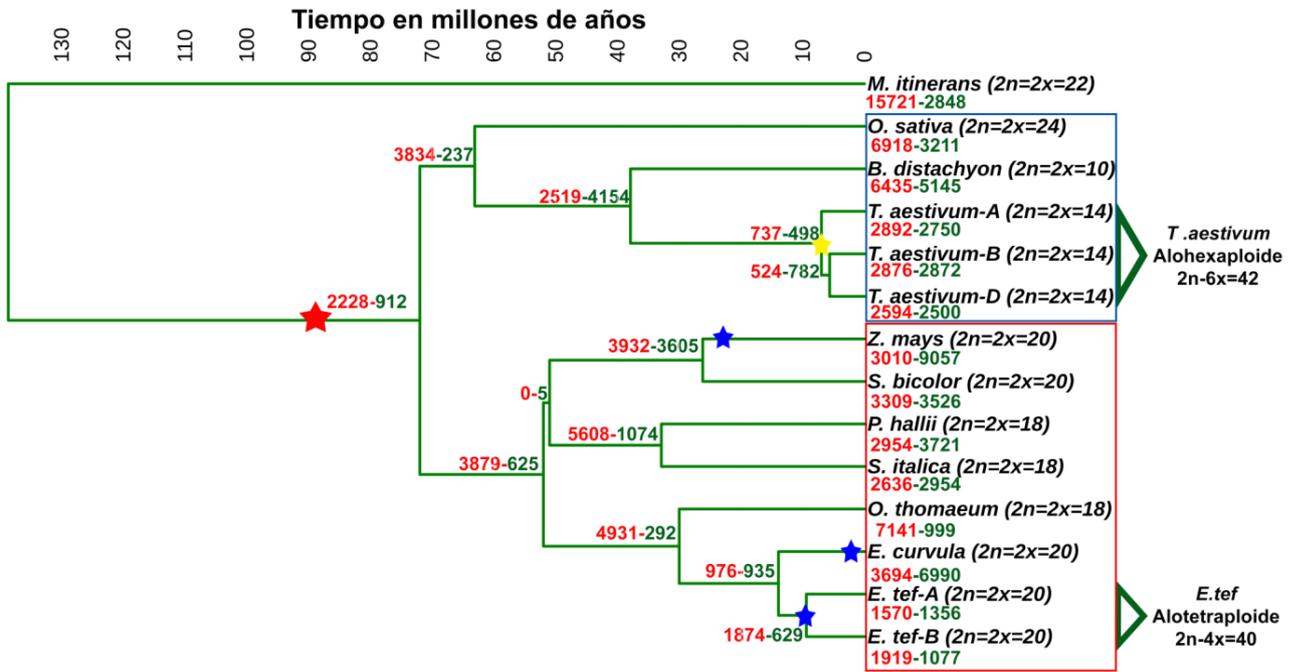


Figura 2.9: Árbol filogenético (método de maximum likelihood) con las gramíneas seleccionadas y *M. itinerans* como outgroup. La estrellas indican eventos de poliploidización: la roja la paleoduplicación compartida por todas las gramíneas, la amarilla la hexaploidización de *T. aestivum* y las azules las duplicaciones de *Z. mays*, *E. curvula* y *E. tef*. Las especies C4 se encuentran encuadradas en rojo y las C3 en azul. Los números rojos y verdes en cada rama representan contracciones y expansiones de genes, respectivamente. La escala en el margen superior de la figura indica la divergencia de cada rama en millones años.

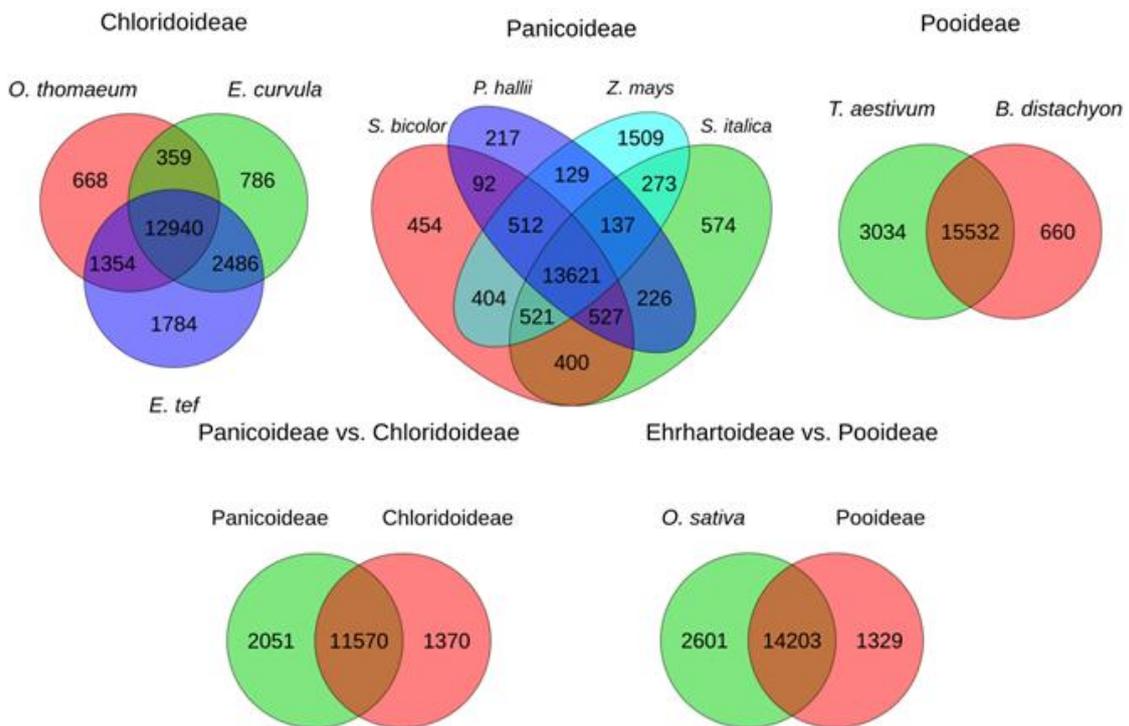


Figura 2.10: Diagramas de Venn de los grupos ortólogos correspondientes a las especies seleccionadas dentro de cada subfamilia. Las Panicoideae y Chloridoideae corresponden a las especies C4 analizadas mientras que *O. sativa* y Pooideae corresponden a las C3.

Para completar el análisis evolutivo, se realizó un estudio de sintenia con distintas especies. Así fue encontrado que el 79% del genoma de *Victoria* está cubierto por bloques sinténicos de *Z. mays* y *S. bicolor* (Tabla 2.5, Figura 2.11), distribuidos en 363 y 305 bloques respectivamente, mientras que para las otras tres especies mostradas en la Tabla 2.5, el porcentaje de cobertura del genoma de *Victoria* es cercano a 85%. Cuando analizamos la relación inversa, i.e. la cobertura de los genomas de estas especies con el genoma de *E. curvula*, observamos que tanto para *O. sativa* como para *E. tef*, especie significativamente más cercana filogenéticamente, el porcentaje de genoma cubierto es del 98%. Sin embargo, puede observarse que la cantidad de bloques compartidos entre las especies del género *Eragrostis* es menor, indicando que cada bloque comprende una mayor proporción del genoma y por lo tanto se encuentran más conservados. Se puede observar además que el cromosoma 3 de *O. sativa* es completamente sinténico al Contig3 de *E. curvula* y que éste además es completamente sinténico al cromosoma 4 A y B de *E. tef*, demostrando la conservación de este cromosoma por millones de años (Figura 2.11).

Finalmente, a través de este análisis se pudo comprobar que los siete contigs hipotetizados previamente como cromosomas están completamente cubiertos por los cromosomas 1, 2, 4, 5, 6, 9 y 10 A y B de *E. tef* los cuales son sinténicos con los contigs 10, 38, 3, 6, 12, 28 y 10 respectivamente (Figura 2.12). Más aun, no solo apoya esta hipótesis, si no que sugiere que los cromosomas 3, 7 y 8 estarían fragmentados en solo dos contigs cada uno demostrando el alto grado de contigüidad y completamiento conseguido en el ensamblado de *Victoria*.

<b>Especies</b>	<b>Bloques</b>	<b>Bloques invertidos</b>	<b>Anclajes</b>	<b>Cobertura <i>E. curvula</i></b>	<b>Cobertura de especie</b>
<b><i>Z. mays</i></b>	363	182	42,818	79%	94%
<b><i>S. bicolor</i></b>	305	149	44,051	79%	66%
<b><i>O. thomaeum</i></b>	251	124	35,121	85%	96%
<b><i>O. sativa</i></b>	262	131	36,698	84%	98%
<b><i>E. tef</i></b>	138	66	40,370	85%	98%

Tabla 2.5: Relación de sintenia entre *E. curvula* y las especies listadas. En la columna Cobertura *E. curvula* se muestra el porcentaje de genoma de *E. curvula* cubierto por las especies listadas en la primer columna, mientras que la columna cobertura especie corresponde al porcentaje de los genomas de las mismas especies cubiertas por *E. curvula*.

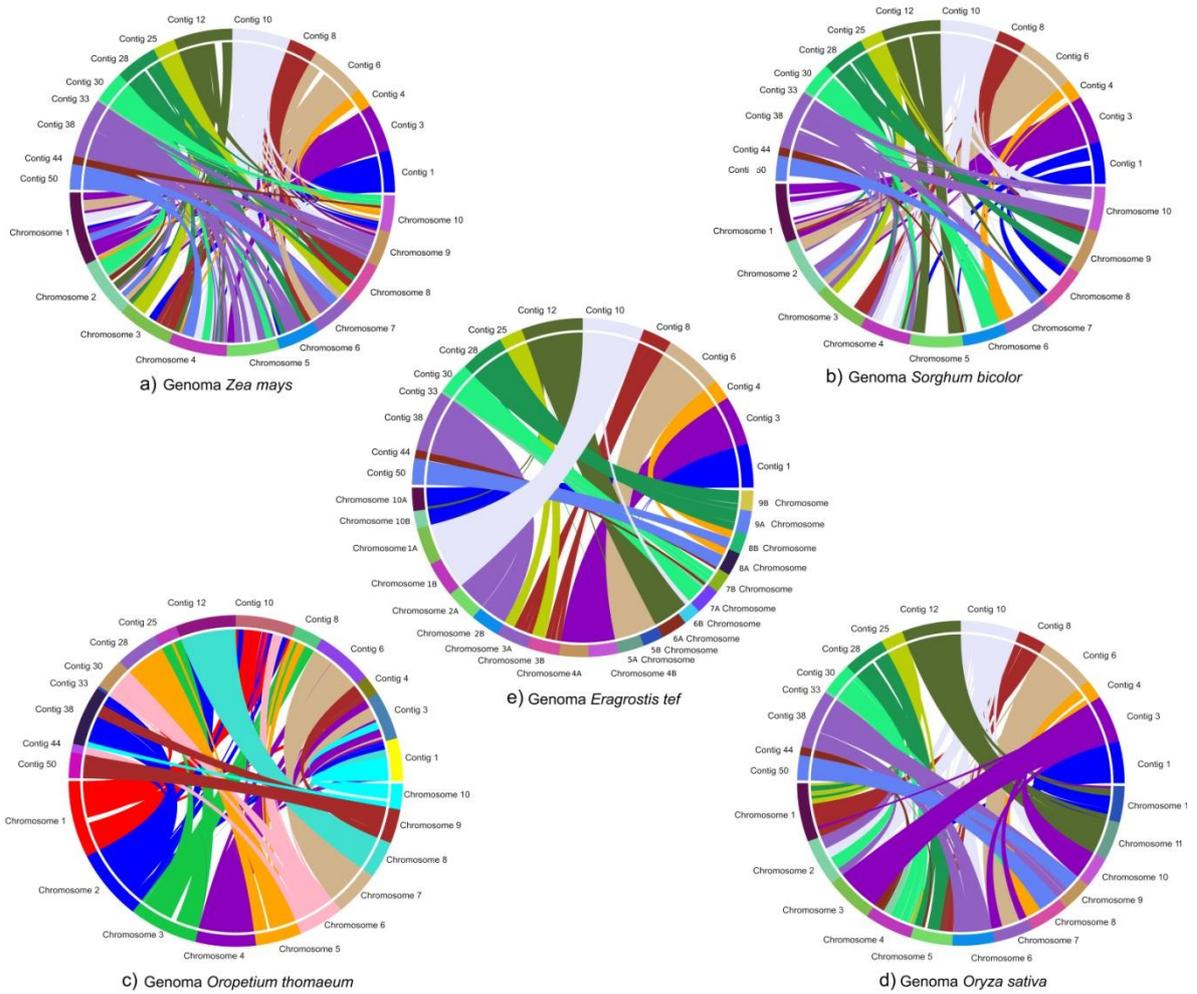


Figura 2.11: Gráfico circular mostrando la sintenia entre *Victoria* y las especies analizadas. Los catorce contigs más largos fueron graficados en *E. curvula*. Cada bloque sinténico está representado con un color distinto.

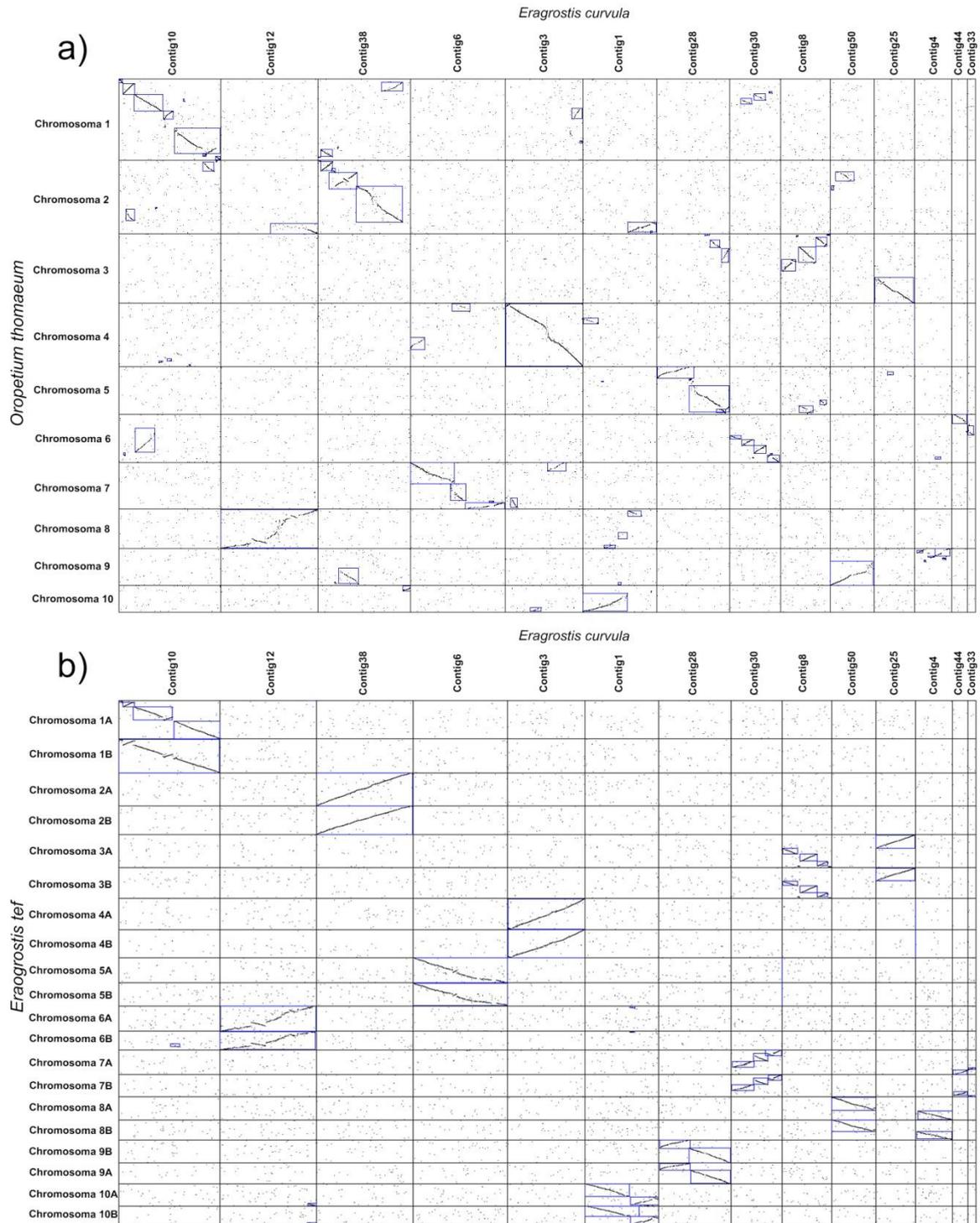


Figura 2.12: Análisis comparativo para la identificación de bloques sinténicos entre los contigs ensamblados de Victoria y los cromosomas de: a) *O. thomaeum*; b) *E. tef*. Los cuadros azules representan bloques sinténicos entre las especies y las líneas diagonales son el resultado de la unión de cada punto que representa un gen sinténico.

## Relación filogenética de los genotipos de *E. curvula*

Para el análisis de la relación filogenética dentro de la especie se utilizaron 35 pares de primers de marcadores SSR previamente diseñados sobre secuencias transcriptómicas (Garbus et al., 2017). El análisis in silico de estos pares de primers reveló que 28 de ellos presentaron 100% de identidad y cobertura con el ensamblado de Victoria. Además, 15 nuevos pares de primers fueron diseñados a partir del ensamblado, de los cuales 14 pudieron ser amplificados en ADN genómico (Figura 2.13, Tabla 1 ANEXO).

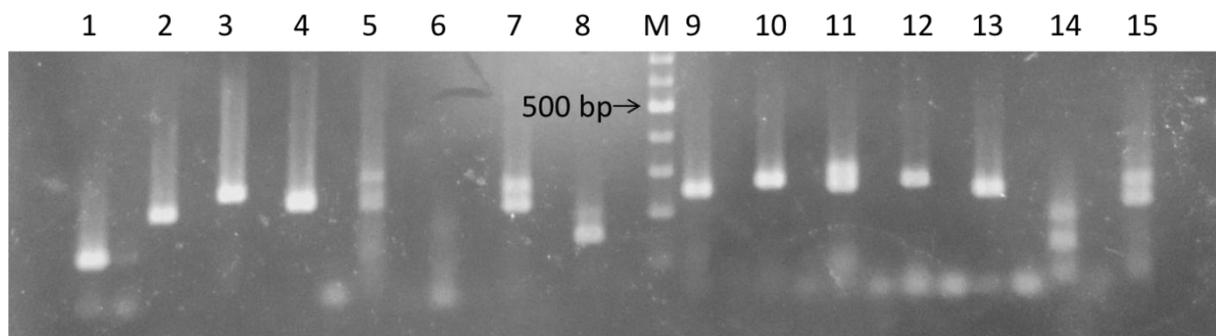


Figura 2.13: Amplificación de los marcadores SSR diseñados sobre el genoma de Victoria. Los números superiores corresponden a la nomenclatura dada a cada par de primers. M) marcador molecular de 100 pb.

Luego, estos marcadores fueron amplificados a partir de ADN de 11 materiales de *E. curvula* y con estos datos se construyó un árbol filogenético utilizando la matriz de Jaccard para determinar la distancia entre marcadores. En este árbol (Figura 2.14), el cultivar hexaploide Don Luis y el cultivar heptaploide Don Pablo se observan agrupados, sugiriendo que la divergencia respecto a los otros cultivares incluidos en el análisis podría estar relacionada con la poliploidía. Siguiendo esta tendencia, Victoria resultó cercana al cultivar diploide PI299920 y a Tanganyika INTA, material tetraploide que le dio origen. Otros materiales que pudieron ser agrupados fueron Don Walter y OTA los cuales en trabajos previos demostraron tener mayor afinidad para su cruzamiento (Zappacosta et al., 2019). Finalmente los materiales diploides PI208214 y PI299928 agruparon con mayor proximidad a la rama tetraploide de OTA y Don Walter.

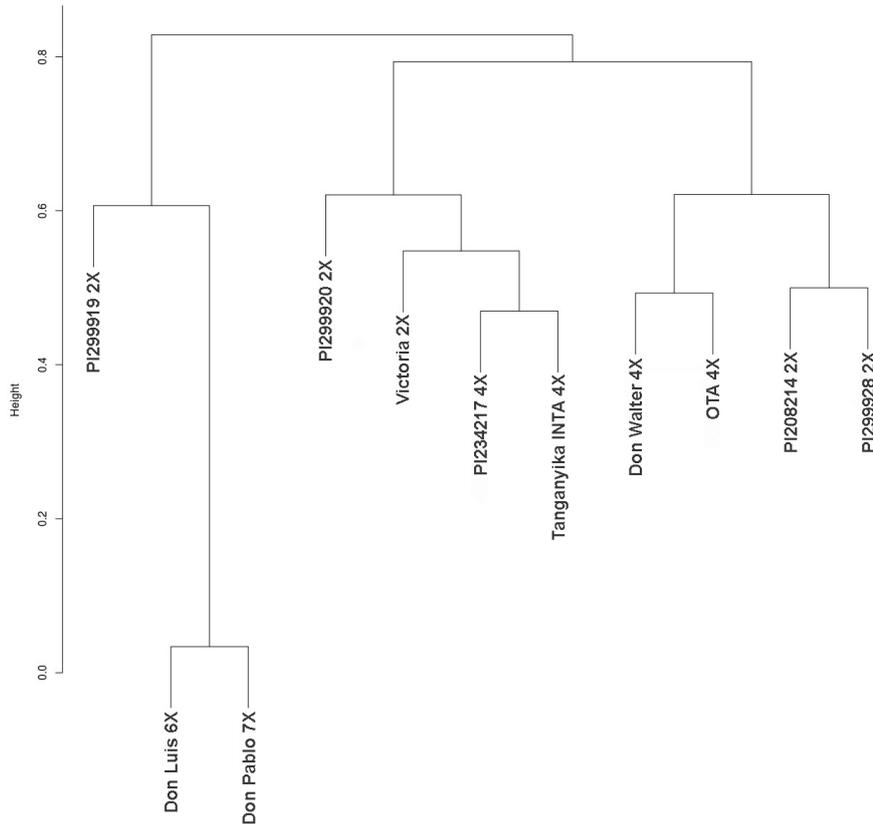


Figura 2.14: Árbol filogenético construido a partir del análisis de los datos obtenidos mediante marcadores SSR con distintos materiales de *E. curvula*.

### Genes relacionados a las vías de las ligninas

La vía de síntesis de lignina disponible en KEGG, denominada “phenylpropanoid biosynthesis”, comienza con alguno de los aminoácidos aromáticos fenilalanina, tirosina o triptófano y finaliza con la producción de los monómeros H, G y S los cuales forman el polímero de lignina. Dentro de la anotación del cultivar Victoria se encontraron todos los genes de referencia disponibles para la especie *Z. mays* que recorren la vía de principio a fin (Figura 2.15).

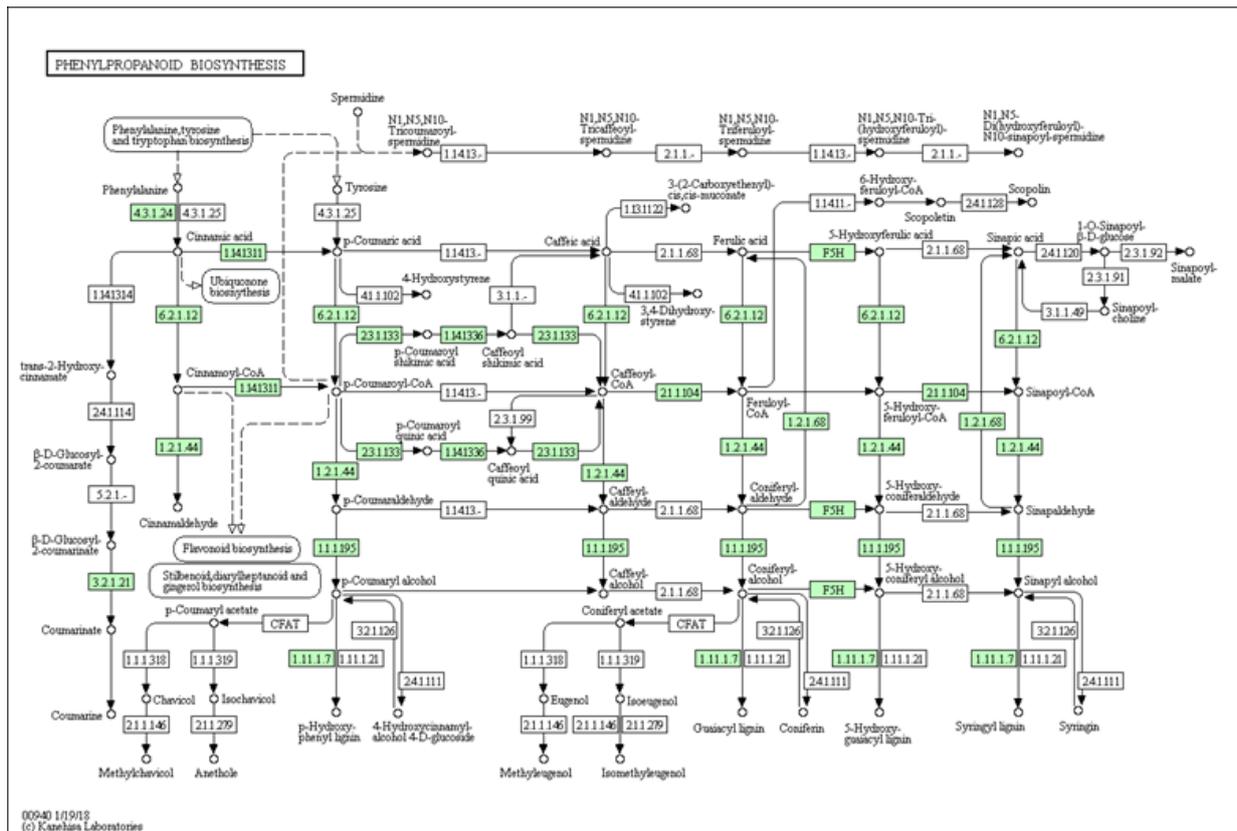


Figura 2.15: Vía de las ligninas, en verde se encuentran las enzimas detectadas en la anotación del cultivar *Victoria* que corresponden en la totalidad de las enzimas que KEGG utiliza como referencia para *Z. mays*.

De esta vía se seleccionó la enzima caffeoyl shikimate esterase (CSE) para ser analizada con mayor profundidad porque se correlaciona directamente con el contenido y la proporción final de monómeros de lignina afectando la calidad forrajera. Los genes codificantes para las enzimas CSE II se encuentran ampliamente distribuidos en todas las especies vegetales mientras que los codificantes para CSE I solo se encuentran en algunas especies (Tabla 2 ANEXO). Para su identificación en el genoma de *Victoria* fueron utilizadas como referencia la CSE I de *O. sativa* (XM\_002462989.2) y CSE II de *S. bicolor* (XM\_015768109.2). Por medio de primers de PCR diseñados sobre estos genes se pudieron amplificar, clonar y secuenciar los fragmentos esperados a partir de ADN de *Victoria*, los cuales presentaron un alineamiento perfecto con el genoma (Tabla 3 ANEXO). Además estos fragmentos fueron identificados en trabajos de transcriptómica previos demostrando que se encuentran activos en la especie (Garbus et al., 2019).

Los genes de *E. curvula* obtenidos en el ensamblado y validado por la clonación de CSE I y II fueron alineados contra distintas especies pudiendo determinarse que en *Z. mays*, *S. bicolor*, *S. italica*, *B. distachyon* y *T. aestivum* solo se encuentra CSE I, mientras que los dos tipos fueron identificados en *Victoria*, *O. thomaeum*, *P. halli* y *O. sativa*. Cuando se realizó el

análisis no pudo detectarse *in silico* sobre el genoma disponible de *E. tef* (Cannarozzi et al., 2014) si la clase I de CSE estaba presente, por este motivo se amplificó, secuenció y clonó en el cultivar Enatite confirmando su presencia. Luego con la secuenciación del nuevo genoma (VanBuren et al., 2020) pude ser corroborado *in silico* que ambos tipos de CSE se encuentran en *E. tef*.

### **Factores de transcripción WRKY en Victoria**

Una de las familias de factores de transcripción que más ha sido estudiada para la detección de candidatos a la tolerancia al estrés biótico y abiótico es WRKY. Debido a que *E. curvula* puede tolerar a condiciones de alta temperatura y radiación, así como a los suelos arenosos deficientes de agua se decidió caracterizar y clasificar esta gran familia en el genoma de Victoria. Indagando en la base de datos Pfam se pudo encontrar que 74 genes del total anotado en el genoma del cv. Victoria contienen motivos WRKY y Zinc "finger". Las 74 proteínas que codifican para estos genes fueron alineadas con el programa MUSCLE y con el software MEGA X se realizó un árbol filogenético con el modelo máxima verosimilitud, agrupándose 7 proteínas en el grupo I, 32 en el grupo II y 35 en el grupo III. Las proteínas del grupo II se clasificaron en las 5 subclases descritas de acuerdo a la presencia de motivos adicionales en las secuencias (Figura 2.16, Figura 1 ANEXO).

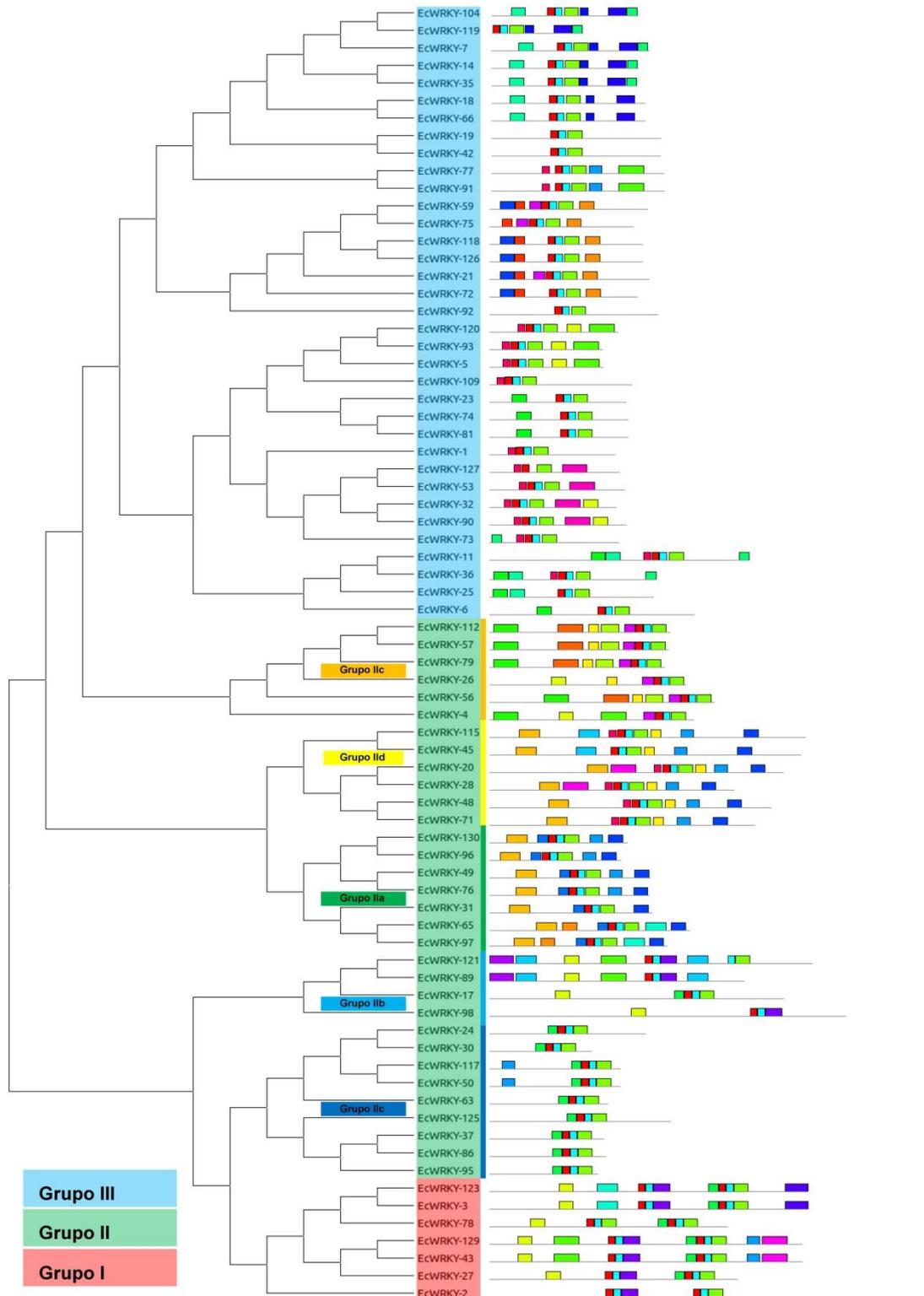


Figura 2.16: Árbol filogenético de los factores de transcripción WRKY encontrados en *E. curvula*. El color resaltado sobre el nombre corresponde al grupo I, II y III mientras que los subgrupos del grupo II se encuentran representados con una barra de distinto color a la derecha del nombre. La representación de la secuencia de la proteína con sus motivos se encuentra en la tabla 2.6

1		DGYQWRKYGQKVIKG
2		SPYPRSYRCTYKNC
3		PVKKQVQRSAEDPSLLIVTYEGEHTHGRP
4		PVKKKVERSIDGQITEIVYKGSNHHPKPPNRR
5		KTEELASLEAELRRVSEENRRLREMLTAVAASYRALQSHL
6		RKVREPRVVFQTRSEVDILD
7		KDLVEQMATALTSDPBFTAALAAITSRI
8		DGLVLNFGSSGSGHQQAACKKQKQYHQWQSVPPGQSSMMNFDQSWKGG
9		GPPKLVQPLSVQFQFQGNVAHRYPFHNQPKQKLAEMFKRSNSGINLKF
10		VTIPAGFSPSALLEPVLNTNSASPPSPT
11		RAAVESCHRVLALLSQPDPAQLKSIASETDEACAKFRKVVALLSDGGG
12		EMPKGQLNLEZILRCLBRALAALKQGT
13		AATISASAPGPTITLDTKNGAAGHA
14		DVEKVPEIEQSSPAPELAGMASPGWDPLDGCLDWE
15		RKMREKRSIKVPAISNKIADIPP
16		RKARVSWRARTYAPY
17		ERSIGIITSSNFDGGRKRSA
18		PENNPDAHSLLOLSASLTVKTESLAAAT
19		MASTTSAASMLMSGSVASLD
20		FHLIGGPVADPVDVQQAPKRRCTGKGEDGRGKCA
21		PHQFHWAGSSAPTSNGPVKMEVDRQSESSSSSTGALPRTLLPJGSSRT
22		REAAAAEVAIEYAIKSHQPLLLLHHEHS
23		LGDESSDFGZFSNYDYLGL
24		KTAAFKPIPPRQLGPLPYAAQQQPHWGHQQQNPAGEPV
25		TNPGKSTPPLSNLAPEGSSNDVKDGVVETKDSWPPSKTETA
26		RPKVSRYVVRADPADLSLVVK
27		NDEPPAPFVINFGSST
28		NQQQKAVVPPAQHQQQEAPPARQGLDGLPVQSEPGNAATGLGGSPEEK

Tabla 2.6: Tabla con los motivos encontrados en las secuencias de los factores de transcripción WRKY. En la fila uno representado en rojo se observa el dominio WRKY.

## Genes relacionados a la reproducción

Para poder determinar los genes de Victoria que tendrían alguna participación en la reproducción, se utilizó el programa hisat2 para mapear las lecturas de cada una de las dos muestras ARN de espiguilla del cultivar Tanganyika USDA (100% apomíctico) y del cultivar OTA (100% sexual) (Garbus et al., 2017) con el genoma de Victoria. Así, se mapearon 277.636 (88.69%) y 712.160 (89.22%) lecturas de las muestras de Tanganyika mientras que en las muestras de OTA se mapearon 457.648 (92.75%) y 718.157 (93.03%) lecturas. Las regiones en las cuales alinearon las lecturas en el genoma fueron predominantemente codificantes, sin embargo se encontraron también lecturas en regiones intergénicas (Figura 2.17).

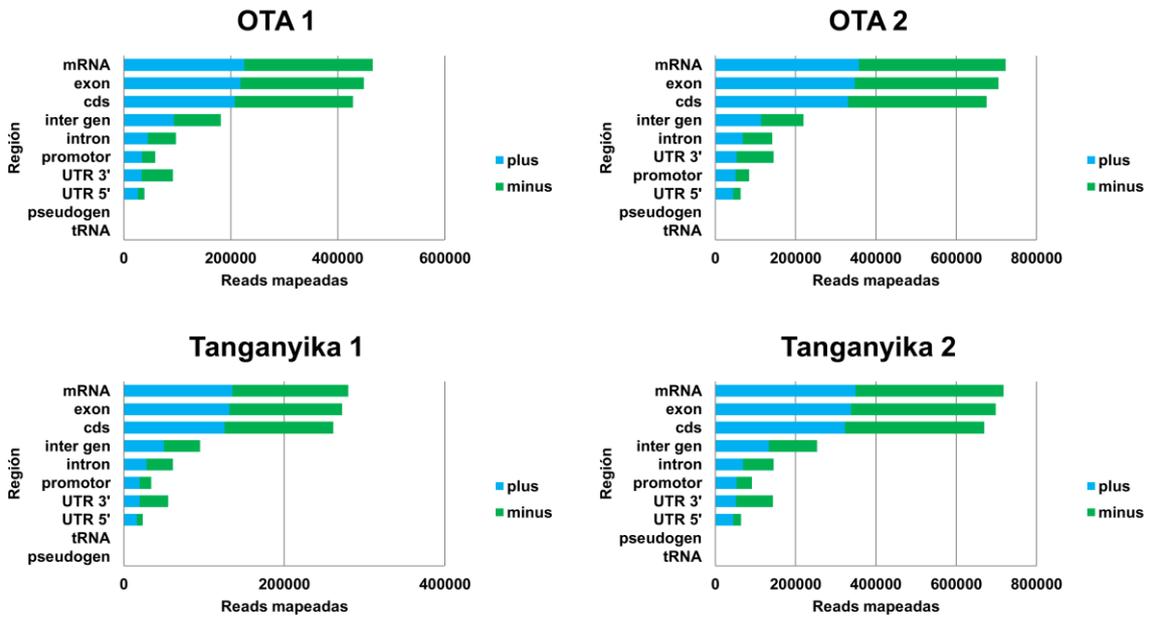


Figura 2.17: Distribución por región de las lecturas mapeadas de las dos muestras de Tanganyika y de OTA. Las barras azules “plus” indican las lecturas que mapearon en la misma dirección y las verdes “minus” las que mapearon en posición inversa.

Con esta información se realizó una matriz de expresión encontrando que 24.634 genes presentes en el genoma de Victoria están expresados en el transcriptoma de inflorescencia de OTA y Tanganyika USDA. Para determinar si las diferencias de expresión encontradas en los distintos genotipos están relacionadas con el modo reproductivo se realizó un análisis de componentes principales con la matriz de expresión indicando que el 90% de la varianza separa a las muestras según su modo reproductivo (Figura 2.18).

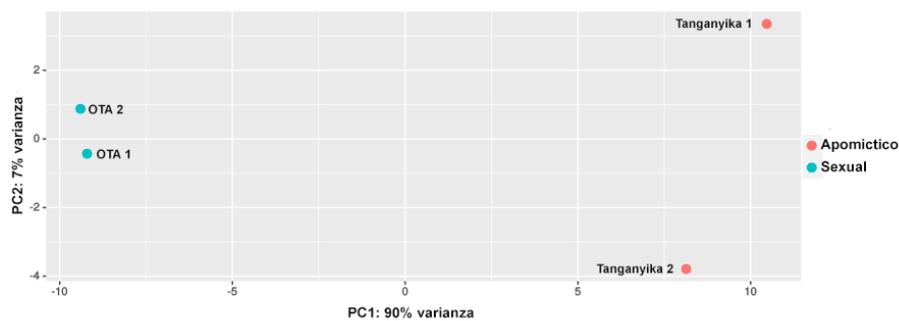


Figura 2.18: Análisis de componentes principales. Cada círculo corresponde a una muestra y el color indica el modo reproductivo. En los ejes se muestra en porcentaje de varianza de cada componente

De los 24.634 genes expresados, 968 se expresaron diferencialmente con un FDR < 0.05 y de los cuales 487 ( $\log_{2}FC > 1$ ) estuvieron sobreexpresados en el genotipo apomítico y 481 ( $\log_{2}FC < 1$ ) reprimidos. Con el fin de detectar a qué etapa de la vía reproductiva se encuentran asociados los genes diferenciales se identificaron sus pares ortólogos en los transcriptomas reproductivos de *O. sativa* previo y posterior a la polinización (Anderson et al., 2013, Anderson et al., 2017). Así, se encontró que de los 487 genes sobreexpresados en el transcriptoma apomítico 256 estaban expresados en alguno de los estadios de *O. sativa* (ovocélula, célula generativa, célula vegetativa, 2.5, 5 y 9 horas luego de la polinización) y de los 481 genes reprimidos 184 tuvieron algunos ortólogos expresados en alguna de las etapas reproductivas de *Oryza*. Luego se comparó la expresión diferencial de genes en *E. curvula* con el comportamiento de estos genes en los distintos estadios de *O. sativa* (Tabla 1.6). De esta manera se determinó que 246 de los 256 genes sobreexpresados en apomixis no estaban expresados diferencialmente o estaban reprimidos en alguna de las vías sexuales de arroz (es decir, corresponde a 0 y -1 en alguno de los estadios de arroz y 1 en apomixis, según la nomenclatura utilizada en la Tabla 2.7). Por otra parte, 176 de los 184 genes que estaban reprimidos en apomixis, resultaron sobreexpresados sin expresión diferencial en alguno de los estadios de arroz. (Figura 2.19)

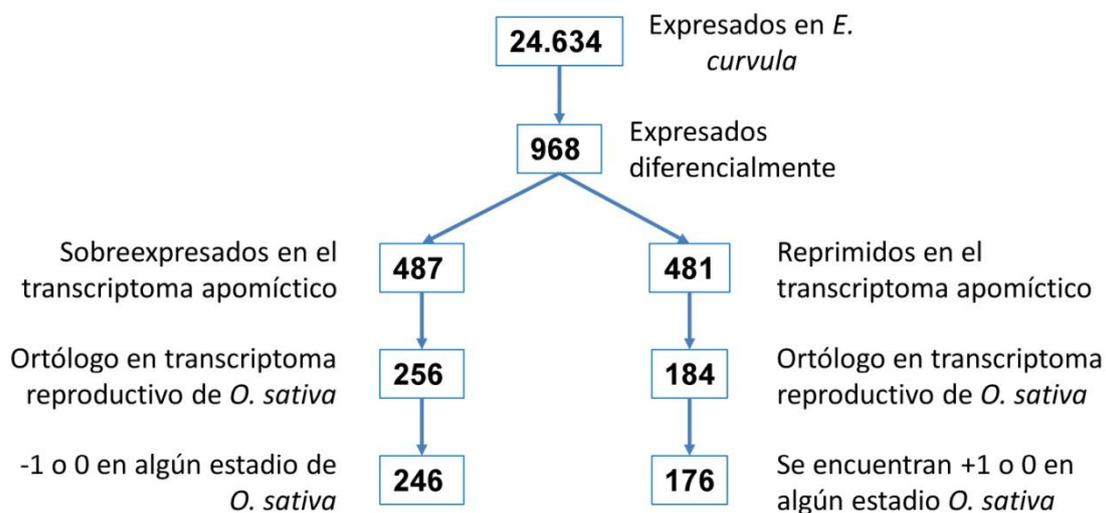


Figura 2.19: Diagrama de los genes surgidos del análisis de las lecturas de expresión de los genotipos Tanganyika USDA y OTA y su comparación con los genes expresados previo y luego de la polinización en arroz. +1, -1, 0 significa sobreexpresado, reprimido y expresado, respectivamente. El fold change tomado para los sobreexpresados y reprimidos fue mayor y menor a 2 respectivamente y el FDR menor a 0,05.

Los 176 genes reprimidos en el transcriptoma apomítico y sobreexpresados o expresados en alguno de los estadios de arroz fueron especialmente analizados ya que podrían ser el sustento de la hipótesis que en los genotipos apomíticos las vías sexuales son reprimidas. De esta manera, para dilucidar si la expresión de estos genes resultaba o no específica de cada etapa del desarrollo, se graficaron los datos obtenidos en un diagrama de Venn (Figura 2.20). Así, se encontró que dos, cinco y cuatro genes se encontraban reprimidos en la ovocélula, célula regenerativa y vegetativa del polen, respectivamente, mientras que en los estadios post polinización no hay genes expresados específicamente. Para poder determinar cuáles podrían ser los efectos de la represión de estos genes se buscó en la literatura aquellos trabajos donde fueron mutados o silenciados para discutir su posible rol en *E. curvula*.

Estadio		Victoria	
		1	-1
Ovocélula	1	119	50
	-1	64	107
	0	67	29
Regenerativa	1	67	61
	-1	76	35
	0	107	90
Vegetativa	1	80	114
	-1	126	62
	0	123	64
<b>Polinización</b>			
2.5 HLP	1	3	8
	-1	4	2
	0	77	76
5 HLP	1	21	16
	-1	33	36
	0	30	34
9 HLP	1	33	21
	-1	23	46
	0	28	19

Tabla 2.7: Número de genes de Victoria ortólogos a los reportados en los distintos estadios reproductivos de *O. sativa*. Estos genes corresponden a los 256 y 184 genes de la Figura 2.19. Como se puede observar el mismo gen aparece en más de un estadio. 1=sobre expresado, -1= reprimido, 0= sin expresión diferencial, HLP= horas luego de la polinización.

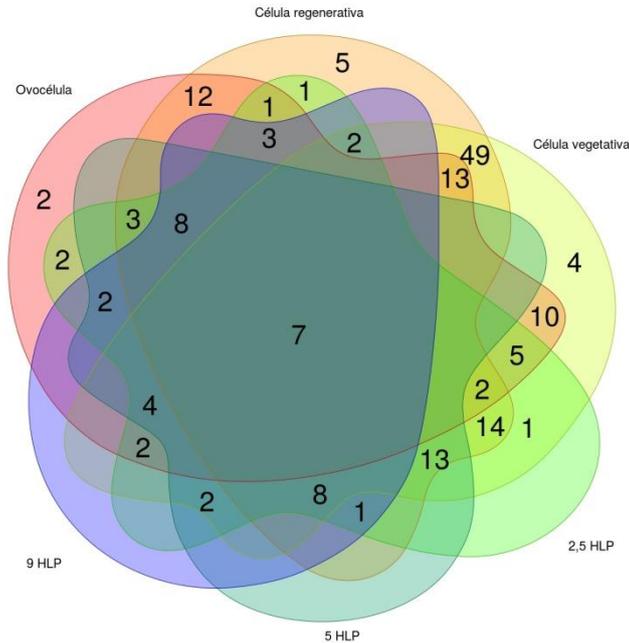


Figura 2.20: Distribución de los 176 genes reprimidos en en apomixis y no expresados diferencialmente o sobreexpresados en los transcriptomas de arroz previo (ovocélula, célula regenerativa y célula vegetativa) y posterior a la fertilización 2.5, 5 y 9 horas luego de la polinización (HLP)).

Los dos genes sobreexpresados en la ovocélula de arroz (reprimidos en apomixis en *E. curvula*) fueron 26S proteasome regulatory subunit 6 (*RPN6*) correspondiente a la accesión de NCBI de Victoria EJB05\_00524 y L-ascorbate peroxidase 4 (*APx4*) correspondiente a EJB05\_47722 (Tabla 2.8). Existen estudios que indican que *RPN6* se expresa en un complejo junto con *RPN5*, *RPN8* y *RPN11* (Fu et al., 2001) y que este complejo es esencial para la gametogénesis y el desarrollo esporofítico. La mutación del gen *RPN5* produce fenotipos estériles, sin embargo la autopolinización de estos individuos produce genotipos fértiles pero con un marcado asincronismo entre los individuos wild-type y las autopolinizadas, por ejemplo, cuando las plantas wild-type alcanzan estadio de corazón temprano las plantas mutantes se encuentran todavía en estadio globular (Book et al., 2009). El otro gen presente en la ovocélula, *APx4*, no ha sido caracterizado individualmente pero la mutación de su proteína chaperona ankyrin repeat-containing protein 2a (*AKR2A*) produce una reducción significativa en su expresión y el silenciamiento de este gen a través de mutantes y RNAi produce un marcado retraso de la floración y una pérdida total de la producción de semillas (Shen et al., 2010) (Tabla 2.8).

Los cinco genes detectados específicamente en la célula regenerativa del polen fueron: LRR receptor-like serine/threonine-protein kinase (EJB05\_31266), phosphatidylinositol-phospholipase C2 (EJB05\_00114), Galactokinase (EJB05\_03325), Mediator of RNA polymerase II transcription subunit 17 (EJB05\_31980) y UDP-glucose 4-epimerase 3 (EJB05\_42263) (Tabla 2.8).

El silenciamiento de LRR receptor-like serine/threonine-protein kinase (*RGF1*) causa que la expresión de Plethora 1 y Plethora 2 (*PLT1* y *PLT2*, de la familia de *BBM*) sea indetectable (Ou et al., 2016). Si bien estos genes hasta hace poco tiempo eran estudiados asociados al tejido radicular (Aida et al., 2004, Du et al., 2017), recientemente se encontró que un miembro de esta familia, *BBM1* (también llamado *PLT6*) produce partenogénesis cuando es expresado ectópicamente en el óvulo, actuando el polen como gatillo natural de este gen (Khanday et al., 2019). La mutación de los genes *PLT1* y *PLT2* produce anomalías principalmente en el número de células y en la organización final del embrión por cambios en el alargamiento y la división celular (Horstman et al., 2014, Aida et al., 2004).

El gen phosphatidylinositol-phospholipase C2 (*PLC2*) también tiene un rol importante en el desarrollo reproductivo, especialmente durante la megagametogénesis y el desarrollo del embrión, donde se lo encontró activamente expresado en *Arabidopsis* (Di Fino et al., 2017). Este gen es requerido para la progresión normal de la mitosis en los primeros estadios de la formación del gametófito femenino y también para la normal división celular luego de la fertilización del polen. Otros estudios encontraron además que afecta la germinación del polen (Li et al., 2015).

Otro gen que ha sido encontrado activamente expresado durante la floración es Galactokinase (Yang et al., 2009), que si bien se ha silenciado no se ha estudiado específicamente en los estadios reproductivos (Egert et al., 2012).

Mediator RNA polymerase II transcription subunit 17 (*MED17*) es un gen implicado en el cambio de la fase vegetativa a reproductiva y actúa como mediador para la síntesis de miRNA y secuencias largas no codificantes (Zhang et al. 2018). Si bien la regulación de las vías reproductiva por medio de secuencias no codificantes no está dilucidada, está comprobado que tanto los miRNA como las secuencias largas no codificantes tienen un rol activo (Ochogavía et al., 2018, Garbus et al., 2019), con lo cual, *MED17* podría tener un rol importante en la regulación de los mecanismos reproductivos.

Finalmente el gen UDP-glucose 4-epimerase 3 (*UGE3*) está relacionado al desarrollo del polen y la semilla post embriogénesis. En plantas con este gen mutado se encontró que el polen está severamente afectado provocando esterilidad masculina (Rösti et al., 2007).

Dentro de los genes reprimidos en apomixis y expresados en sexualidad en la célula vegetativa se encontraron: EJB05\_14522, Phosphatidylinositol 4-kinase gamma 4 (EJB05\_48586), Clathrin heavy chain 1 (EJB05\_53329) y Cation/H(+) antiporter 15 (EJB05\_53831) (Tabla 2.8).

Clathrin heavy chain 1 (*CHC1*) es un gen vinculado a la endocitosis y al transporte de auxina a los distintos órganos de la planta. Uno de los fenotipos observados en plantas con este gen mutado fue división celular aberrante en la parte basal de los embriones y anomalías en la distribución de las auxinas en el embrión (Kitakura et al., 2011).

El gen Phosphatidylinositol 4-kinase gamma 4 (*PIK4*) está vinculado a la formación del polen y su mutación produce una disminución significativa del número de granos de polen y consecuentemente menor número de semillas. Sin embargo se observó que la meiosis de en estados iniciales de la formación del polen no es afectada (Alves-Ferreira et al., 2007).

Cation/H(+) antiporter 15 (*CHX15*) es un gen de la familia *CHX* la cual forma un complejo entre todos sus integrantes que está asociado a la formación y del polen y del tubo polínico. *CHX15* está particularmente involucrado en el intercambio catiónico en el polen tricelular (polen maduro) (Sze et al., 2004). Recientemente se demostró también que cuando algunos genes de esta familia son mutados cambia la expresión de todos sus integrantes, se produce la germinación del polen pero solo la mitad de los óvulos es fertilizada y se produce un patrón donde solo es fertilizada la célula polar (como ocurre en apomixis) o la ovocélula pero no las dos (Padmanaban et al., 2017).

El último de los genes encontrado reprimido en apomixis y sobreexpresado en célula vegetativa de los individuos sexuales fue succinate dehydrogenase 1 (*SDH1*). El silenciamiento de este gen por medio de mutaciones y RNAi produce cambios en la segregación indicando defectos gametofíticos masculinos y femeninos. En el óvulo los mutantes presentan dos tipos de fenotipos, uno en el cual el saco embrionario es arrestado en el estadio binucleado y otro en el cual el polo nuclear no se fusiona. Por otro lado, el polen se ve severamente afectado, ya que solo el 50% de los granos resultan viables (León et al 2007).

Ubicación en <i>Oryza</i>	Accesión	Nombre corto	Nombre del gen
<b>Ovocélula</b>	EJB05_00524	<i>RPN6</i>	26S proteasome regulatory subunit RPN6
	EJB05_47722	<i>APx4</i>	L-ascorbate peroxidase 4
<b>Célula regenerativa</b>	EJB05_00114	<i>PLC2</i>	phosphatidylinositol-phospholipase C2
	EJB05_03325	<i>GATK</i>	Galactokinase
	EJB05_31266	<i>RGF1</i>	LRR receptor-like serine/threonine-protein kinase
	EJB05_31980	<i>MED17</i>	Mediator of RNA polymerase II transcription subunit 17
	EJB05_42263	<i>UGE3</i>	UDP-glucose 4-epimerase
<b>Célula vegetativa</b>	EJB05_14522	<i>SDH1</i>	succinate dehydrogenase 1
	EJB05_48586	<i>PIK4</i>	Phosphatidylinositol 4-kinase gamma 4
	EJB05_53329	<i>CHC</i>	Clathrin heavy chain 1
	EJB05_53831	<i>CHX15</i>	Cation/H(+) antiporter 15

Tabla 2.8: Genes que se encontraron reprimidos en el genotipo apomítico de *E. curvula* y no mostraron expresión diferencial o estuvieron sobreexpresados en los transcriptomas de arroz en la ovocélula, célula regenerativa y célula vegetativa.

## Discusión

La caracterización de genes relacionados específicamente con el desarrollo de embriones apomícticos tiene una relación muy estrecha con las vías sexuales, probablemente porque muchos de los componentes de sus vías son comunes, por lo que se cree que posiblemente el desarrollo apomíctico haya surgido como una variante del sexual (Grossniklaus 2001, Koltunow and Grossniklaus, 2003, Hojsgaard 2020). Aunque la herencia de la apomixis puede ser explicada por la genética mendeliana, también han sido descritas interacciones epistáticas, componentes que son expresados esporofíticamente o gametofíticamente, presencia de modificadores, distorsiones en la segregación y supresión de la recombinación, complejizando el estudio de su regulación y de su base genética (Ozias-Akins y van Dijk 2007). El desarrollo de los sacos apomícticos es el resultado de numerosas regulaciones finamente orquestadas espacial y temporalmente (Grimanelli., 2012, Khanday et al., 2019). Sin embargo, los posibles mecanismos que desregulan la sexualidad para el surgimiento del desarrollo apomíctico no se conocen y son ampliamente especulativos en la mayoría de los casos. Dentro de estos mecanismo se incluyen: la poliploidización (Carman 2001), hibridación (Asker y Jerling 1992, Hojsgaard et al., 2014, Koch et al., 2003) y regulaciones epigenéticas involucradas en el silenciamiento (Grimanelli, 2012, Garbus et al., 2019, Amiteye et al., 2013), entre otros. En las últimas décadas, los resultados logrados utilizando modelos de plantas sexuales han resultado sumamente fructíferos, develando aspectos claves en el control molecular y el desarrollo apomíctico dentro de los cuales se incluyen el control epigenético (Grimanelli et al., 2012) y regulación transcripcional de la partenogénesis (Conner et al., 2017; Anderson et al., 2017). Por otro lado se pudo obtener cultivos apomícticos sintéticos a través de ingeniería genética demostrando que, aunque en muchos de los casos la eficiencia no es alta, es sumamente factible la transformación de especies sexuales en apomícticas (Khanday et al., 2019).

El uso de varias estrategias de secuenciación de ADN genómico de última generación, combinado con el uso de programas específicos para su ensamblado, permitieron obtener en esta Tesis el primer genoma de la especie *E. curvula*. Dentro de los motivos que llevaron a la elección del cultivar sexual Victoria ( $2n=2x=20$ ) como el primero de la especie en ser secuenciado se puede mencionar su origen, dado que se obtuvo a partir del cultivo de inflorescencias de un cultivar tetraploide apomíctico facultativo y la menor complejidad de su genoma con respecto a los cultivares poliploides. Además, actualmente es aceptado que solo a través del conocimiento de la sexualidad podrá determinarse cuales son los

mecanismos que provocan su desregulación (Hojsgaard et al., 2020, Grossniklaus, 2001). Por estos motivos, la secuenciación y ensamblado de un genotipo sexual diploide de *E. curvula* es sumamente conveniente ya que la disminución de su complejidad favorece la obtención de ensamblados de mayor calidad mientras que su naturaleza sexual ofrece una referencia para luego ser contrastado con genotipos poliploides apomícticos los cuales contienen los genes y/o regiones condicionantes de la apomixis.

Uno de los desarrollos tecnológicos más recientes para obtener genomas a escala cromosómica es ligación por proximidad utilizada por las técnicas de Chicago y Hi-C. La combinación de estas tecnologías con lecturas de gran longitud como las de PacBio permitió obtener un ensamblado final de 602.432.814 pb distribuidas en 1.143 contigs. De los 10 cromosomas de *E. curvula*, 7 han sido ensamblados en un solo fragmento mientras que los otros 3 se encuentran en dos fragmentos cada uno. El N50 final conseguido con la integración de todas las tecnologías fue de 43 Mb lo cual indica el alto grado de contigüidad del genoma obtenido. En el ensamblado del genoma de *Durio zibethinus* (~738 Mb), especie en la que se ha recurrido a la misma estrategia de secuenciación, se obtuvo un N50 de 22,7 Mb (Teh et al, 2017), aproximadamente la mitad que en Victoria, reflejando que aquí se ha logrado optimizar el rendimiento de los programas de ensamblado disponibles. En otros proyectos como el de *Chenopodium quinoa* (1.45 Gb) (Jarvis et al., 2017), *M. truncatula* (465 Mb) (Zou et al., 2018) y *Arabis alpine* (370 Mb) (Jiao et al., 2017) utilizando otras metodologías de secuenciación como mapeo óptico y Hi-C se obtuvieron valores de N50 de 3,84 Mb, 12,5 Mb y 31 Mb, respectivamente. Todos estos resultados reflejan la complejidad de obtener un genoma de alta calidad, incluso combinando tecnologías y, como quedó evidenciado en el ensamblado de *E. curvula*, lo poderoso que resulta la unión de las tecnologías PacBio, Chicago y Hi-C.

El ensamblado de Victoria cubrió alrededor del 97% de la longitud en pb estimada de genoma y contiene 95,5% de los marcadores DArT, sugiriendo que el genoma está prácticamente completo. Además, fueron anotados 56.469 genes, pudiéndose identificar el 96,4% de los genes BUSCO completos, de los cuales el 16,1% estuvieron duplicados. Este número de genes duplicados junto con la cantidad de genes, la cual pareciera estar sobreestimada comparada con especies cercanas (*E. tef*  $2n=4x=40$ , 68.255 genes), nos da una idea de la redundancia de del genoma, es decir existen regiones que no pudieron ser colapsadas para formar el genoma haploide. El principal motivo por el cual no se colapsaron completamente los haplotipos del genoma diploide de Victoria es su nivel de heterocigosidad. Victoria proviene de cultivo de inflorescencias *in vitro* de un genotipo tetraploide muy poco domesticado (Cardone et al., 2006), por lo tanto se espera que el nivel de heterocigosidad sea alto. Por otro lado, la autoincompatibilidad de este genotipo no

permite obtener homocigosis por autocruzamiento. Las métricas cualitativas y cuantitativas de este genoma permitieron que fuera seleccionado para estar en la plataforma ensembl ([https://plants.ensembl.org/Eragrostis\\_curvula](https://plants.ensembl.org/Eragrostis_curvula)) la cual solo publica genomas de alta calidad.

El tipo y cantidad de elementos repetitivos varía según la especie, su número está asociado al tamaño de los genomas y a la complejidad intrínseca de cada especie (Dubin et al., 2018). El 28% del ensamblado del genoma de Victoria estuvo cubierto por elementos repetitivos un valor muy similar al reportado en *E. tef* (27.46%) (Vanburen et al., 2020). Por otro lado este número fue menor al obtenido en los genomas de *S. italica* (46%) (Zhang et al., 2012), *S. bicolor* (62%) (Paterson et al., 2012) y *Z. mays* (66%) (Jiao et al., 2017)

Este ensamblado además permitió determinar la ubicación evolutiva de *E. curvula* dentro de la familia Chloridoideae y sirvió de referencia para el diseño de primers de SSR para poder determinar la cercanía evolutiva entre los diferentes genotipos de la especie. Además se pudieron identificar 74 factores de transcripción WRKY los cuales pudieron ser clasificados en tres grupos principales y cinco subgrupos basados en la clasificación propuesta por otros autores (Eulgem et al., 2007). La identificación, clasificación y caracterización de esta familia constituye un paso muy importante para poder dilucidar las bases moleculares de la tolerancia al estrés por sequía de esta especie ya que permitirán determinar en trabajos de expresión su rol específico en esta característica tan importante.

Una de las principales limitantes de *E. curvula* como fuente de forraje es su baja digestibilidad, una característica que ha podido ser mejorada exitosamente modificando el contenido relativo de lignina por medio de ingeniería genética en *Paspalum dilatatum*, (Giordano et al., 2014). Las enzimas CSE I y CSE II han sido estudiadas por su importancia en las vías de síntesis de ligninas (Vanholme et al., 2013, Ha et al., 2016), ya que su silenciamiento ha demostrado que disminuye el contenido de este polímero. La clase II de esta enzima se encuentra ampliamente distribuida en todas las especies vegetales mientras que la clase I ha sido identificada solo en algunos grupos. De esta manera, se comprobó que *E. curvula* al igual que en *E. tef* y *O. thomaeum* contiene ambos tipos de enzimas siendo un potencial blanco para programas de mejoramiento de la especie.

Utilizando trabajos de transcriptómica de la especie y de *O. sativa* se pudieron encontrar genes candidatos para la regulación de la apomixis. Los genes reprimidos en sexualidad fueron principalmente investigados. Así se encontró una variedad de genes de los cuales algunos están implicados directamente en procesos similares a los que ocurren en genotipos apomícticos, como por ejemplo, *CHX15* relacionado a la fecundación de un solo núcleo (Padmanaban et al., 2017).y *RGF1* el cual regula directamente *PLT1* y *PLT2* que participan en la organización celular del óvulo (Ou et al., 2016). Este análisis fue hecho solo sobre

genes específicos de la ovocélula, la célula regenerativa y la célula vegetativa del polen, lo que no implica que algunos genes que actúan en múltiples estadios tengan funciones importantes en apomixis. Por otro lado, solo se tuvieron en cuenta los genes reprimidos en apomixis y con una condición diferente en arroz (sobreexpresados o no expresados diferencialmente) debido a que los genes sobreexpresados para la apomixis podrán ser analizados en genotipos apomícticos utilizando otras estrategias.

Finalmente, este genoma será utilizado para su comparación con genomas apomícticos para poder determinar si existen regiones condicionantes de la apomixis que se encuentran presentes solo en los genotipos con este carácter. En conclusión, utilizando diferentes tecnologías para ensamblar y validar el genoma de *E. curvula* se ha hecho un notable avance en su caracterización. Este ensamblado además constituye una valiosa herramienta para la búsqueda, validación y caracterización de genes para el estudio de diferentes características como la tolerancia a distintos estreses abióticos y la calidad forrajera.

## Capítulo 2 : Secuenciación y ensamblado de los genotipos apomícticos de *E. curvula* Don Walter y Tanganyika INTA

### Introducción

La disminución de los costos de secuenciación debido al desarrollo de nuevas tecnologías llevó en una revolución en cuanto a la secuenciación de especies vegetales, pasando de solo estar secuenciada *Arabidopsis thaliana* en el 2000 (Kaul et al., 2000) a encontrarse en la actualidad (Agosto 2020) 484 genomas secuenciados de los cuales 55 corresponden a gramíneas (<https://www.plabipd.de>). Los genomas de las especies vegetales representan desafíos mayores que otras especies, en primer lugar por el alto contenido de elementos repetitivos y en segundo lugar porque muchas de ellas tienen varios juegos de cromosomas, es decir son poliploides. Sin embargo, incluso con estas complicaciones se han podido secuenciar genomas de alta calidad gigantescos como el de *Triticum aestivum* ( $2n=6x=42$ , ~17 Gb) (Alonge et al., 2020) y con altos niveles de ploidía como el de *Fragaria x ananassa* ( $2n=8x=56$ , ~805 Mb) (Edger et al., 2019). Aunque las primeras versiones de algunos de estos genomas fueron construidos a partir de secuencias cortas (Clavijo et al., 2017), el desarrollo de la secuenciación de ADN por medio de lecturas largas simplificó mucho su ensamblado, especialmente debido a la mayor posibilidad de poder abarcar en una única read regiones con largas repeticiones. A pesar de todos estos avances, existen algunas especies, denominadas comúnmente huérfanas, que tienen un gran interés biológico pero todavía no han sido secuenciadas (Tadele et al., 2019). *Eragrostis curvula* es parte de este grupo debido a que, aunque existe un genoma de alta calidad (Capítulo 1, Carballo et al., 2019), por tratarse de un cultivar diploide no está presente la región de la apomixis, la característica de mayor interés en la especie

*Eragrostis curvula* es una especie forrajera gramínea que forma un complejo agámico con genotipos desde  $2x$  hasta  $8x$ , de estos, los diploides y tetraploides han sido los más estudiados (Cardone et al., 2006, Cervigni et al., 2008, Selva et al. 2017, Garbus et al., 2017). Hasta 1971 se creía que todos los genotipos de *E. curvula* eran tetraploides y se reproducían por apomixis diplospóricas, pero el hallazgo de genotipos diploides en África no solo demostró que existe variabilidad cromosómica, si no también que estos individuos diploides son siempre sexuales (Voigt et al., 1971, Voigt y Bashaw 1972). Dentro de los genotipos diploides, el cultivar sexual Victoria es en el que centramos la atención debido a que fue obtenido en nuestro laboratorio a partir del cultivar apomíctico Tanganyika INTA por medio de cultivo de inflorescencias *in vitro* (Cardone et al., 2006, Carballo et al., 2019). Un poco más adelante se descubriría que además existen genotipos tetraploides en los cuales

conviven sacos apomíticos y sexuales en la misma planta, denominados apomíticos facultativos (Voigt y Bashaw, 1976). Finalmente, con el registro del cultivar OTA (Voigt, 1976) se demostró la presencia de genotipos tetraploides con el 100% de los sacos sexuales. Esta variabilidad de modos reproductivos fue recientemente evaluada con la cruce del cultivar OTA con Don Walter, resultando la progenie en genotipos sexuales, apomíticos obligados y apomíticos facultativos (Zappacosta et al., 2019). Es importante remarcar que de los genotipos tetraploides, el cultivar apomítico facultativo Don Walter ha sido el único capaz de poder polinizar a un genotipo 100% sexual, por lo que últimamente se le ha dado mayor importancia dentro de este complejo. Tanto el tipo de segregación del modo reproductivo como los marcadores ligados a la apomixis encontrados en este último trabajo indicarían que existe una región que controla la apomixis en *E. curvula* de la misma manera que sucede con *Pennisetum* y *Cechrus* (Conner et al., 2008). Esta combinación de distintos modos reproductivos con la misma ploidía en una sola especie convierten a *E. curvula* en un modelo ideal para estudiar la apomixis diplospórica. Todo indicaría que los organismos tetraploides de *E. curvula* al igual *E. tef* serían alotetraploides, el único estudio cromosómico, aunque no es concluyente, lo clasifica como alotetraploide segmental (Poverene 1988).

Aunque se han hecho algunos avances para transferir la apomixis a plantas sexuales (Khanday et al., 2019, Wang et al., 2020, D'Erfurth et al., 2009), no se conocen exactamente ni los mecanismos ni las bases moleculares de ninguno de sus componentes (i.e., apomeiosis, partenogénesis, pseudogamia). Se cree que serían unos pocos los genes que regulan la apomixis o desregulan la sexualidad (Voigt y Bashaw, 1976, Asker y Jerling, 1992, Koltunow et al., 1995), sin embargo ninguno de los genes/regiones candidatos surgido de cientos de trabajos de investigación ha sido la clave para poder distinguir las diferencias entre las vías sexuales y apomíticas. La secuenciación de organismos apomíticos y su comparación con genotipos de la misma especie con modo de reproducción sexual podría ser una nueva forma de abordar el problema. La secuenciación de el genotipo sexual del Capítulo 1, junto con el mapa de ligamiento previamente obtenido (Zappacosta et al., 2019) serán utilizadas como complemento a secuenciaciones de genomas de apomíticos de *E. curvula* para poder detectar los genes y/o regiones que responsables de la regulación del modo reproductivo.

## **Materiales y métodos**

### **Material Vegetal y extracción**

Los materiales utilizados para extraer el ADN fueron los genotipos apomícticos facultativos Don Walter y Tanganyika INTA ( $2n=4x=40$ ). Don Walter fue seleccionado porque es el padre de una población de mapeo recientemente obtenida en el laboratorio (Zappacosta et al., 2019). Por otra parte, la elección de Tanganyika INTA se debió a que es el material que le dio origen al genotipo diploide Victoria previamente secuenciado.

La extracción de ADN del cultivar Tanganyika INTA se llevó a cabo con el método CTAB (Meier et al 2011). Para el cultivar Don Walter se tomaron dos muestras que fueron extraídas en distintos laboratorios, una en CERZOS-CONICET en Bahía Blanca, Argentina y otra NIAB-EMR en Reino Unido. La primera se realizó también por el método de CTAB teniendo en cuenta los pasos críticos para obtener fragmentos largos mientras que la segunda utilizó un protocolo optimizado para la obtención de fragmentos de ADN largos (Schalamun y Schwessinger, 2016)

### **Secuenciación**

La secuenciación del cultivar Don Walter se realizó por medio de dos bibliotecas, la primera fue realizada en la Universidad de Liverpool en Reino Unido utilizando la plataforma Chromium 10x genomics. De esta manera 1 ng de ADN fue introducido en “Chromium controller Chip” junto con los reactivos siguiendo las instrucciones del fabricante (<https://support.10xgenomics.com/de-novo-assembly/library-prep/doc/user-guide-chromium-genome-reagent-kit-v1-chemistry>). Brevemente, a cada fragmento de ADN se le unieron “barcodes” (secuencias únicas de 16 bases) junto con adaptadores de secuenciación que son retenidos al fragmento en una emulsión denominada GEM (Gel Bead in Emulsion). A cada fragmento que ingresó a una GEM se le une más de un barcode con la misma secuencia, que luego son secuenciados con una plataforma Illumina HiSeq X pair end 2x150 pb. De esta forma esta tecnología produce lecturas cortas pero los barcodes permiten detectar de qué molécula proviene cada read.

La segunda librería fue secuenciada en el instituto NIAB en East Malling en Reino Unido con la tecnología Oxford Nanopore. El tipo de librería utilizada para esta secuenciación fue 1D que a pesar de tener mayor tasa de error el rendimiento aumenta significativamente. La

biblioteca se preparó sobre los 50 ng de ADN extraídos que luego fue secuenciados en la plataforma Gridlon (Schalamun et al., 2017).

La secuenciación del cultivar Tanganyika INTA fue realizada en INDEAR (Rosario, Santa Fe, Argentina) con la plataforma Illumina Hiseq 1500 pair-end 2x250. La cuantificación de la muestra de ADN para la preparación de la librería Nextera DNA se realizó con Quant-iT PicoGreen® DNA Assay Kit (Invitrogen) de acuerdo a las instrucciones detalladas por el fabricante. Para la preparación de las bibliotecas se partió de 50 ng de ADN purificado, normalizado a 2,5 ng/ul. Se realizaron las etapas de tagmentación del ADN, purificación, ligación de adaptadores y amplificación de los fragmentos obtenidos y finalmente se realizó una última purificación previa a la secuenciación ([https://support.illumina.com/sequencing/sequencing\\_kits/nextera\\_dna\\_kit.html](https://support.illumina.com/sequencing/sequencing_kits/nextera_dna_kit.html) Nextera® DNA Library Prep Reference Guide).

### **Citometría de Flujo**

La determinación del tamaño de los genomas estudiados se realizó por medio de un análisis de citometría de flujo en el INTA Castelar. Para realizar este estudio se enviaron al laboratorio semillas de cada cultivar que fueron germinadas para el aislamiento y tinción con ioduro de propidio de los núcleos de células de tejido de hoja fresco (Galbraith et al., 1983). Esta solución es introducida en el citometro de flujo el cual mide la fluorescencia que puede ser extrapolada a pares de bases utilizando en la misma corrida un patrón con el tamaño de genoma conocido, en este caso se utilizó el *Secale cereale* que tiene un genoma de 7.917 Mb.

### **Ensamblado**

En el caso del genotipo Don Walter, las lecturas obtenidas de Chromium 10X fueron ensambladas con el software supernova (Weisenfeld et al., 2017), única opción para ensamblar *de novo* secuencias conseguidas con esta tecnología. Luego las lecturas obtenidas a través de Oxford Nanopore fueron integradas al ensamblado de supernova con el software DBG2OLC (Ye et al., 2016), explorando distintos parámetros. Este ensamblado fue pulido con una iteración de racon (Vaser et al., 2017) utilizando las lecturas largas de Oxford Nanopore y dos iteraciones de Pilon utilizando las lecturas de la secuenciación de Chromium 10X. Estos programas permiten alinear las lecturas al ensamblado y determinar la base que se presenta con mayor frecuencia en cada posición.

Para el ensamblado de las lecturas cortas de Tanganyika INTA se probaron los programas Masurca (Zimin et al., 2013) y Spades (Bankevich et al., 2012) encontrando mejores resultados con el primero.

La calidad de ambos genomas ensamblados fue evaluada por medio de los siguientes parámetros: N50, número de contigs, longitud final y número de genes BUSCO completos (Seppey et al., 2019).

### **Anotación de elementos repetitivos**

La anotación de elementos repetitivos se realizó por medio de una combinación de programas. Para todos ellos la entrada fueron las secuencias de los distintos ensamblados solo con los SSR enmascarados por el programa GMATA (Wang y Wang et al., 2016). Luego para detectar los LTR se utilizó LTRharvest (Ellinghaus et al., 2008) y LTR\_retriever (Ou et al., 2018), para los helitrones HelitronScanner (Xiong et al., 2014) y para los MITE mitehunter (Han et al., 2010). Finalmente, se construyó una biblioteca con todas estas estructuras y, junto con los contigs de los ensamblados finales, se corrió el programa RepeatMasker (Smit et al., 2013) para realizar una búsqueda basada en homología con la base de datos repbase 23 (Bao et al., 2015).

### **Anotación de genes**

La anotación de los genes de los genomas de Tanganyika INTA y Don Walter se hizo sobre la salida del programa LTR\_retriever. La búsqueda de los genes se realizó con el software MAKER (Campbell et al., 2014) al igual que en el genoma de Victoria (ver materiales y métodos capítulo 1). A diferencia del genoma diploide, en este caso se contaron con ensamblados transcriptómicos de 17 cultivares de *E. curvula* provistos por el grupo de investigación de Dr. Matthew Moscou del laboratorio de Sainsbury en Norwich, Reino Unido (Datos no publicados). Además se utilizaron transcriptomas previamente secuenciados de distintos órganos y cultivares de *E. curvula* (Garbus et al., 2019, Cervigni et al., 2008) y proteínas de *E. tef* (Cannarozzi et al., 2014), *Setaria italica* (Zhang et al., 2012), *Sorghum bicolor* (Paterson et al., 2009) y *Z. mays* (Jiao et al., 2017). Para el modelado *Ab initio* se utilizaron los programas Augustus (Stanke et al., 2006) y SNAP (Johnson et al., 2008) entrenados con los genes BUSCO. Luego de las tres iteraciones de MAKER los genes fueron alineados a la librería de elementos repetitivos creada en el paso anterior mediante el software Exonerate (Slater et al., 2005) y se descartaron aquellos genes con cobertura  $\leq 75\%$  e identidad  $\leq 95\%$ . Por último, los genes restantes fueron alineados con BLAST

(Altschul et al., 1990) a la base de datos TREMBL (<https://www.uniprot.org/>) para asignarles una función o descripción.

### **Análisis filogenético de la subfamilia Chloridoideae**

Utilizando el programa OrthoFinder (Emms et al., 2015) se detectaron los grupos de genes ortólogos presentes en los genomas disponibles de la subfamilia de las Chloridoideae. De esta manera los cultivares de *E. curvula* secuenciados en esta Tesis, Victoria, Don Walter y Tanganyika INTA junto con *E. tef* (VanBuren et al., 2020), *E. nindensis* (Pardo et al., 2020), *O. thomaeum* (VanBuren et al., 2015), *Zoysia japonica* (Tanaka et al., 2016) y *Z. mays* (Jiao et al., 2017) como outgroup fueron analizados para construir el árbol filogenético. Con el fin de incorporar otros cultivares de *E. curvula* se ensamblaron por separado los transcriptomas de OTA (100% sexual) y Tanganyika USDA (100% apomítico) previamente secuenciados (Garbus et al., 2017) utilizando el programa Trinity (Haas et al., 2013) con los parámetros por default. Los transcritos de cada uno de los cultivares fueron traducidos a proteínas por medio del protocolo por homología del programa TransDecoder (Haas et al., 2013) y se agregaron al análisis de OrthoFinder. Por último, se construyó un árbol filogenético mediante el método de Máximo likelihood con el alineamiento provisto por OrthoFinder y se analizó la distancia filogenética entre todas las especies y cultivares.

### **Análisis comparativo de los cultivares de *E. curvula* y la región ligada a la apomixis**

El mapa de ligamiento de la especie *E. curvula* (Zappacosta et al., 2019) surgió del cruzamiento de el genotipo Don Walter y OTA, es decir, un genotipo apomítico facultativo y un genotipo sexual. En el mapa de ligamiento de Don Walter se descubrió dentro del grupo de ligamiento 3 una región ligada a apomixis sin recombinación. Esta región está determinada por cuatro marcadores de 60 pb, los cuales fueron alineados con el programa BLAST (Altschul et al., 1990) a todos los genomas. Utilizando como referencia el genoma de Victoria el cual tiene mayor contigüidad se realizó un análisis comparativo de sintenia dentro de la región utilizando el software jvci (Tang et al., 2008). Así, se alinearon los genomas de Don Walter y Tanganyika INTA con Victoria, identificando bloques sinténicos con más de dos genes por medio del programa LAST (Kielbasa et al., 2011). Luego se filtraron los contigs que alinearon en la región donde se encuentran los marcadores en Victoria y se determinaron todos los contigs sinténicos entre los tres genomas en la región.

Finalmente dentro de estos contigs se analizaron los genes sinténicos y no sinténicos, ya que en muchos casos una parte del contig fue sinténica y la otra no.

Para determinar la relación de estos genes con el modo reproductivo se llevó a cabo, al igual que en el Capítulo 1, un análisis de expresión diferencial de los genes anotados en Don Walter y en Tanganyika INTA. Para ello se tomaron las lecturas previamente secuenciadas de un cultivar 100% apomíctico y un cultivar 100% sexual (Garbus et al., 2017) y se mapearon a cada uno de los genomas con el software Hisat2 (Kim et al., 2016). Luego se filtraron con los paquetes de R "GenomicFeatures" y "GenomicAlignments" (Lawrence et al. 2013) tomando en cuenta solo aquellas lecturas que alinearon con genes. Finalmente por medio del programa EdgeR (Robinson et al., 2011) se normalizó la matriz con las cuentas de lecturas mapeadas por gen y se realizó un análisis de expresión diferencial tomando como criterio para sobreexpresado y reprimido un fold change mayor y menor a 2, respectivamente (equivalente a log fold change mayor y menor que 1) y un FDR (tasa de falsos positivos) menor a 0,05 (Benjamini et al., 2000).

Los genes sobreexpresados y reprimidos para apomixis presentes dentro de la región delimitada por los marcadores se alinearon con el programa BLAST (Altschul et al., 1990) utilizando como referencia la base de datos UniProt. Aquellos genes con anotación funcional fueron investigados individualmente en la literatura para determinar su posible función dentro de las vías reproductivas.

## Resultados

### Tamaño genómico y ploidía

Para la medición del tamaño y la ploidía de los genomas de los cultivares Tanganyika INTA y Don Walter se realizó un análisis de citometría de flujo. Esta técnica mide la fluorescencia del Ioduro de propidio para estimar el contenido de ADN de los núcleos (mayor cantidad de ADN mayor fluorescencia) de cada cultivar que al ser contrastados con *Secale cereale*, la cual tiene un tamaño de genoma conocido (7.917 Mb), se puede determinar el tamaño del genoma en pb de cada cultivar. Además, se incluyó en el análisis a Victoria como referencia, ya que es un genotipo diploide. De esta forma, se estimó que el tamaño del haploide de Victoria es de 604,3 Mb mientras que el tamaño haploide de los genomas tetraploides fue 1.195,18 y 1.247,57 Mb para Don Walter y Tanganyika INTA respectivamente. Además, la fluorescencia de los materiales tetraploides fue aproximadamente el doble que el material diploide (Figura 3.1) dando aún más evidencia que los cultivares Don Walter y Tanganyika INTA tienen el doble de ploidía que Victoria.

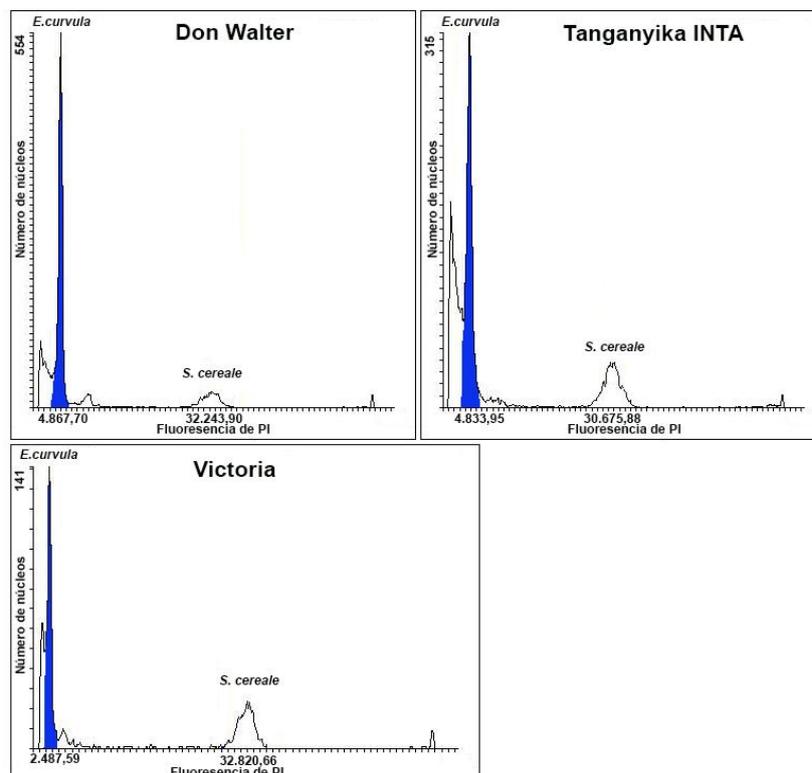


Figura 3.1: Histogramas obtenidos del clitómetro de flujo. El eje Y representa el número de núcleos y el eje X la fluorescencia del Ioduro de propidio para los núcleos. Se puede observar un mayor fluorescencia de *S. cereale* lo cual era esperado ya que su genoma tiene 7.917 Mb. En el caso de los

cultivares de *E. curvula* el tamaño se determinó sobre el pico azul ya que el resto de los picos corresponden a núcleos en activa división celular.

## Secuenciación y ensamblado

### Don Walter

La extracción de ADN del cultivar apomítico facultativo Don Walter realizada en el laboratorio de CERZOS-CONICET para la secuenciación con Chromium 10x resultó en una concentración de 0,27  $\mu\text{g}/\mu\text{l}$  y un tamaño de fragmento promedio de 34.368 pb (Figura 3.2). Por otro lado, en la extracción llevada a cabo en el laboratorio de NIAB EMR en Reino unido se obtuvo una concentración de 0,86  $\mu\text{g}/\mu\text{l}$  y un tamaño de fragmento promedio superior a 60.000 pb, ideal para la secuenciación de lecturas largas de Oxford Nanopore (Figura 3.3).

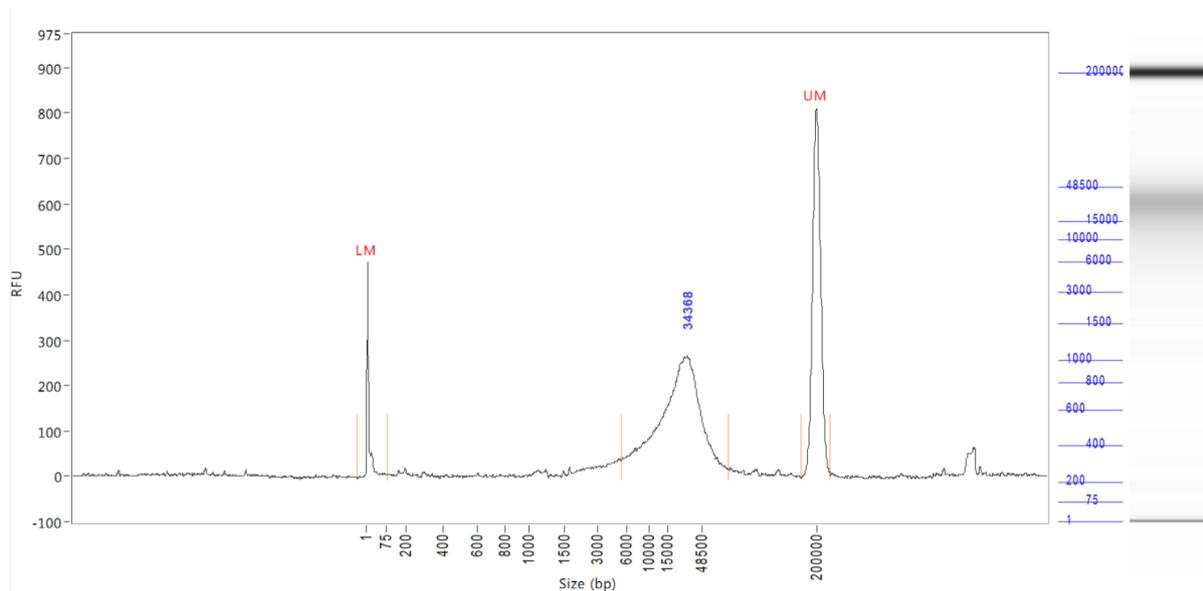


Figura 3.2: Análisis de calidad de la muestra ADN extraído para la secuenciación con Chromium 10X del cultivar Don Walter. En el eje y se encuentran las unidades de fluorescencia relativa (FU) y en el x el tamaño de fragmento de ADN. LM (lower marker) y UM (upper marker) corresponden a marcadores de 1 y 200.000 pb respectivamente.

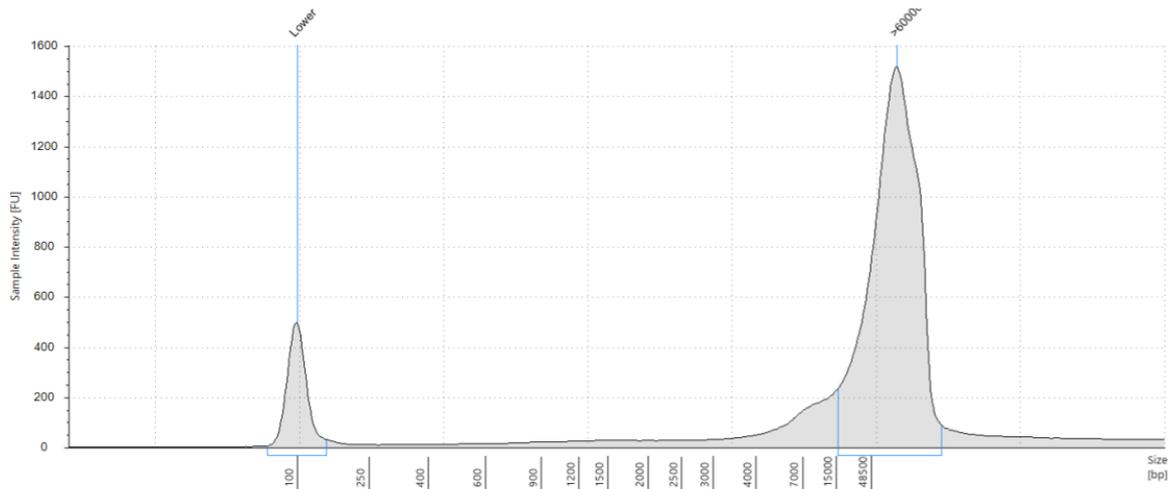


Figura 3.3: Análisis de calidad de la muestra de ADN extraído para la secuenciación con Oxford Nanopore del cultivar Don Walter. En el eje y se encuentran las unidades de fluorescencia relativa (FU) y en el x el tamaño de fragmento de ADN. Lower (lower marker) corresponde al marcador de 100 pb.

La secuenciación del ADN obtenido del cultivar Don Walter por medio de la plataforma 10X genomics resultó en 424.370.442 lecturas 2x150 pb, lo que corresponde a 106,09X de cobertura del genoma haploide. Estas lecturas fueron ensambladas *de novo* con el software supernova, que arroja como salida 3 archivos, llamados pseudohap, megabubbles y raw. Estos archivos se diferencian en el grado de “haploidización”, es decir en el grado de colapsado de los genomas haploides, siendo mayor en pseudohap y menor en raw (Tabla 3.1). El ensamblado con mayor contigüidad fue, como era esperado, la versión de pseudohap con un N50 de 9.198 pb y 137.814 secuencias. Sin embargo, el resultado de megabubbles fue muy parecido con un N50 de 8.480 pb y 153.492 secuencias y superándolo en longitud con 833.56 Mb ensambladas. A pesar de tener una contigüidad aceptable para el tipo de tecnología utilizada, el porcentaje de genes BUSCO fue muy pobre, llegando 67,4% de genes completos en el mejor de los casos (Tabla 3.1).

	Raw	Megabubbles	Pseudohap	Nanopore
<b>Longitud pb</b>	3.210.254.314	833.562.730	770.857.453	1.147.811.592
<b>N50 pb</b>	2.994	8.480	9.198	224.390
<b>Nro. contigs</b>	11.652.730	153.492	137.814	7542
<b>Promedio pb</b>	275.49	5430.65	5593.46	152189,28
<b>BUSCO</b>	C:20%	C:65,7%	C:67,4	96,9%

Tabla 3.1: Métricas del ensamblado de Don Walter con el software Supernova y el ensamblado final incorporando las lecturas de Nanopore con el software DBG2OLC. Longitud en pares de bases, N50

en pares de bases, número final de contigs en cada ensamblado, longitud promedio de los contigs y genes (C) completos BUSCO.

Para mejorar estos resultados, se secuenció una celda de Oxford Nanopore, resultando 1.864.546 lecturas equivalente a una cobertura de 9X. Estas secuencias se incorporaron al ensamblado megabubbles de supernova con el software DBG2OLC (parámetros: AdaptiveTh=0.01 y MinOverlap = 40) resultando un N50 de 224.390 pb, 1.147,8 Mb de longitud total y 7.542 contigs. Este ensamblado después de las rondas de pulido de racon pasó a tener el 80,7% de los genes completos BUSCO y luego de dos iteraciones de Pilon (Walker et al., 2014) se obtuvo un 96,9%, de los de los cuales 60.8%, están presentes en una sola copia y 36.1% en más de una. De esta manera todos los parámetros evaluados mejoraron notablemente con el agregado de solo una celda de Oxford Nanopore al ensamblado de las lecturas de Chromium 10x.

### Tanganyika INTA

La extracción de ADN del cultivar apomítico Tanganyika INTA se realizó en el laboratorio de CERZOS-CONICET. Para la preparación de las bibliotecas se partido de 50 ng de ADN purificado, normalizado a 2.5 ng/ul. Luego se llevó a cabo la tagmentación, purificación, ligación de adaptadores y amplificación de los fragmentos ADN. Finalmente, para chequear la calidad de la librería final se corrió 1ul de la muestra en el 2100 Bioanalyzer (Agilent Technologies), obteniendo un tamaño promedio de la genoteca de 767 bp (Figura 3.3).

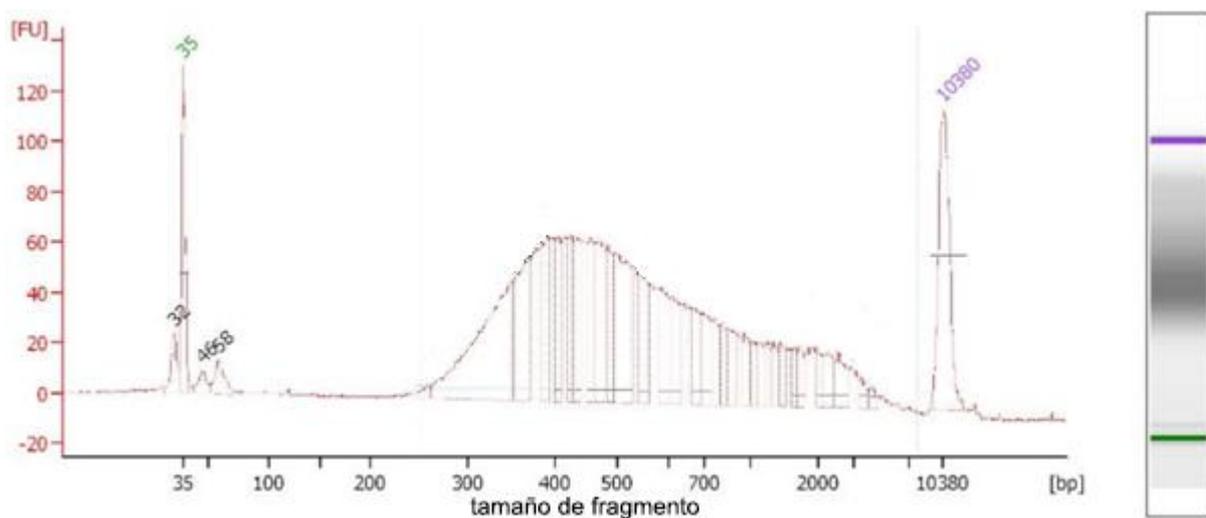


Figura 3.3: Análisis de calidad de la muestra ADN extraído para la secuenciación con Illumina del cultivar Tanganyika INTA. En el eje y se encuentran las unidades de fluorescencia relativa (FU) y en el x el tamaño de fragmento de ADN. Los dos marcadores en verde y violeta corresponden a 35 pb y 10380 pb, respectivamente.

Este ADN fue secuenciado mediante la plataforma Illumina 2x250 pb, resultando 266.042.616 lecturas lo que representa 110,85X de cobertura del genoma haploide. El ensamblado se realizó con el programa Masurca con la opción AUTO k-mer activada obteniendo 293.300 contigs con un N50 de 4.715 pb y una longitud total de 892.919.326 pb. El porcentaje de genes BUSCO completo para este ensamblado fue 83.2% (46% en una sola copia y 37,2% duplicados) mientras que el de fragmentados y perdidos fue de 7,8% y 9%, respectivamente. El porcentaje de genes BUSCO completos fue aceptable para un genoma poliploide ensamblado solo con lecturas cortas.

### **Anotación de elementos repetitivos**

La anotación de los elementos repetitivos en Don Walter a través de los distintos software cubrió en total 54,07% del genoma (Figura 3.4 A). Como era de esperarse los LTR-RT resultaron mayoritarios con un 26,53%, de los cuales 20,27% fueron *Gypsy* y 6,1% *Copia*, siendo la relación *Gypsy:Copia* 3,32:1, menor que la relación 4,3:1 encontrada en el genoma de Victoria (Capítulo 1). Por otro lado, el genoma de Tanganyika INTA presentó un 54,96 % de elementos repetitivos (Figura 3.4 B) encontrándose una relación entre los elementos *Gypsy:Copia* de 3,06:1, similar al reportado en Don Walter. Este cambio en la relación *Gypsy:Copia* entre el genoma diploide y los tetraploides podría estar relacionado con el aumento del contenido de elementos repetitivos comúnmente observado en genomas poliploides. Además en el genoma de Tanganyika INTA se encontró un 31,11% de elementos sin clasificar, posiblemente debido a que grado de fragmentación que presenta este ensamblado no le permite determinar a los softwares a que clase pertenece cada elemento.

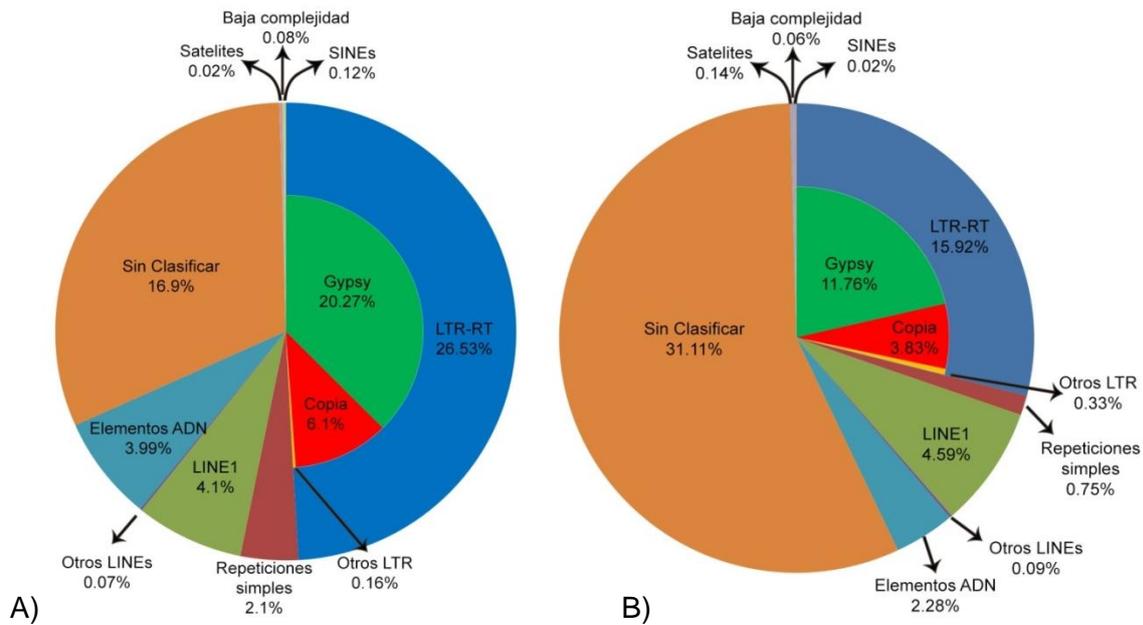


Figura 3.4: Distribución porcentual de los principales elementos repetitivos encontrados en el genoma de Don Walter (A) y de Tanganyika INTA (B). Los porcentajes representan la proporción de cada tipo de elemento respecto del total cubierto en cada genoma, que es del 54,02% en Don Walter y 54,96% en Tanganyika INTA.

### Anotación de genes

La anotación de los genomas ensamblados de los cultivares tetraploides representa un desafío tanto por su longitud, contenido de elementos repetitivos y grado de fragmentación. Sin embargo, los nuevos transcriptomas de *E. curvula* facilitados por el laboratorio Sainsbury de Reino Unido, obtenidos a partir de 17 cultivares distintos, ayudaron en gran medida a modelar los genes y obtener una anotación de alta calidad. Al igual que en el genoma de Victoria, se realizaron tres iteraciones por medio del programa MAKER, la primera utilizando toda la información transcriptómica disponible para la especie sumada a proteínas de especies cercanas mientras que la segunda y la tercera se realizaron a través de entrenamientos *ab Initio* para el modelado de genes. Como data set de entrenamiento para este modelado se utilizaron las secuencias de los genes BUSCO identificadas en cada genoma. Así, en el genoma de Don Walter se anotaron 103.808 genes, aproximadamente el doble que en el diploide Victoria. Se identificaron el 96,1% de los genes BUSCO, de los cuales 56% están en una sola copia y 40,1% en copias múltiples, demostrando el efecto de la ploidía sobre la anotación. Además, el porcentaje de genes BUSCO fragmentados fue 2,2% y el de no encontrados 1,7% (Tabla 3.2). El genoma de Tanganyika INTA fue anotado con el mismo método encontrándose 142.275 genes. Este aumento en el número de genes

respecto a Don Walter se debe principalmente al grado de fragmentación de este genoma donde se identificó un porcentaje genes BUSCO fragmentados de 8,7%, superior a los genes no encontrados que fue del 5,2%. Por otro lado los genes BUSCO completos fueron 86,1%, representando 46,9% los de copia única y 39,2% los de múltiple copia.

	Número de genes	Longitud promedio	BUSCO
<b>Victoria</b>	56.469	1.424,03 pb	C:93.4%[S:65.3%,D:28.1%],F:1.7%,M:4.9%
<b>Don Walter</b>	103.961	1.337,71 pb	C:96.1%[S:56.0%,D:40.1%],F:2.2%,M:1.7%
<b>Tanganyika INTA</b>	142.275	999,87 pb	C:86.1%[S:46.9%,D:39.2%],F:8.7%,M:5.2%

Tabla 3.2: Resultado de la anotación de los genes de los cultivares de *E. curvula* secuenciados. La longitud promedio de genes se encuentra en pares de bases (pb). Los genes BUSCO completos (C), simple copia (S), múltiple copia (D), fraccionados (F) y no encontrados (M) se expresan como porcentaje respecto de los 1440 genes de referencia.

A continuación, se utilizó el software OrthoFinder para determinar los genes ortólogos entre los tres cultivares analizados, tomando como datos de entrada las proteínas anotadas en cada genoma. De esta manera, se realizó un análisis comparativo en el cual se graficaron en un diagrama de Venn todos los genes comunes a los tres cultivares y aquellos específicos de cada cultivar (Figura 3.5). En este análisis se observó que el aproximadamente el 90% de los genes de Victoria se encuentran en los demás cultivares, y por otro lado que existe un alto número de genes específicos de los materiales tetraploides, llegando a 30.085 en Don Walter y 41.602 en Tanganyika INTA, probablemente por el grado de fraccionamiento de estos genomas. Finalmente se determinó que el número total de genes compartido entre todos los genomas fue de 132.956 representando el 43% del total de genes anotados.

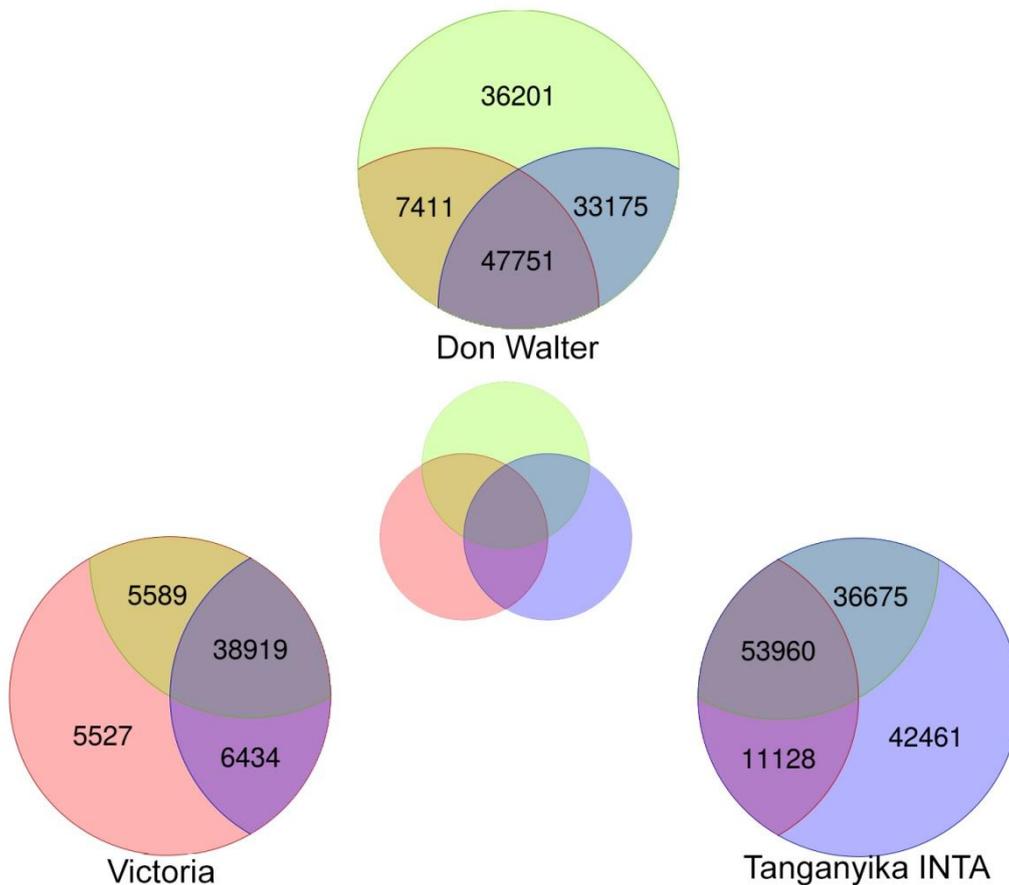


Figura 3.5: Diagrama de Venn de los cultivares, Victoria, Don Walter y Tanganyika INTA con los genes ortólogos identificados en con el programa OrthoFinder. Los círculos laterales representan los 56.469, 103.961 y 142.275 genes compartidos y específicos de Victoria, Don Walter y Tanganyika INTA respectivamente, mientras que el esquema central muestra las intersecciones considerada.

### Análisis filogenético de la subfamilia Chloridoideae

Dentro de la subfamilia chloridoideae, en las familias poaceae hasta el momento fueron secuenciados los genomas de las especies *E. tef*, *E. nindensis*, *Z japonica*, *O. thomaeum* y los tres cultivares de *E. curvula* descritos en esta Tesis: Victoria, Tanganyika INTA y Don Walter. La anotación de estas especies sumada a las 16.503 y 17.808 proteínas completas identificadas en el ensamblado transcriptómico de Tanganyika USDA y OTA fueron utilizadas como entrada del programa OrthoFinder. Así se encontraron genes y grupos de genes ortólogos (ortogrupos) entre las especies (Tabla 4 ANEXO). Los genes del cultivar Victoria fueron mayormente encontrados en otras especies quedando sin asignar solo 6,4% del total de los genes anotados. Por otro lado, los cultivares tetraploides Don Walter y Tanganyika INTA tuvieron un alto porcentaje de genes sin asignar representando un 21% y 19,8%, respectivamente. Este alto porcentaje podría ser por la falta de clasificación o la no identificación de proteínas asociadas elementos repetitivos (i.e., proteínas repetitivas

anotadas como genes) debido a la falta de contigüidad de los genomas. Por otro lado se espera que una parte de estos porcentajes sea real, ya que ambas especies tienen características que ninguno de los otros genomas anotados tiene.

Una de las ventajas de utilizar el software OrthoFinder es que en el resultado final se puede estimar los genes con baja cantidad de copias en todos los genomas y realizar un árbol filogenético. De esta manera alineando 1.920 genes y utilizando *Z. mays* como outgroup se construyó el árbol filogenético de la subfamilia Chloridoideae con el método maximum likelihood (Figura 3.6). Como era de esperarse, las especies de los géneros *Oropetium* y *Zoysia* se agruparon en un clado separado del género *Eragrostis*. *E. nindensis* fue la especie más alejada del género *Eragrostis* separándose del clado donde se encuentran todo el resto de las especies. Dentro de la especie *E. curvula*, los cultivares Victoria y Tanganyika INTA se agruparon por el hecho de tener el mismo background genético. Luego los más cercanos a este grupo fueron Don Walter y OTA, los progenitores de la población de mapeo recientemente obtenida en nuestro laboratorio mientras que Tanganyika USDA fue el más alejado.

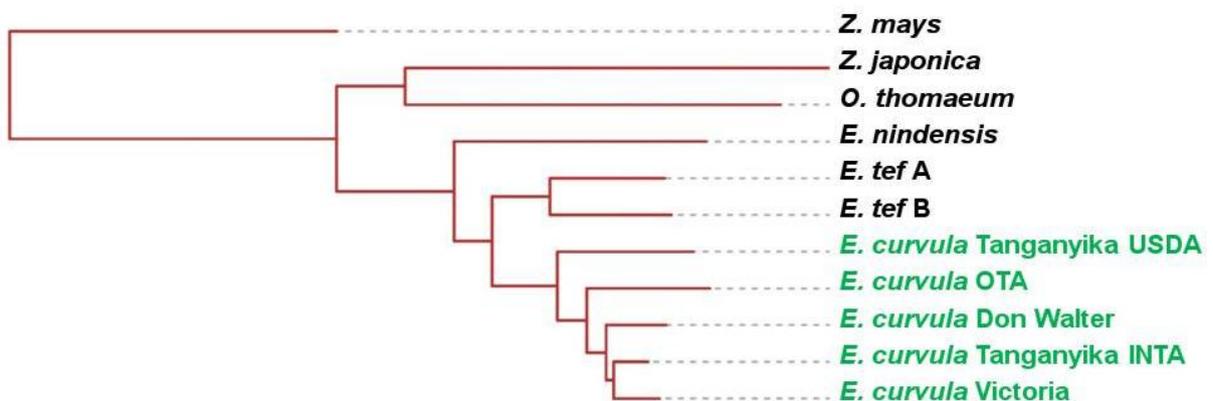


Figura 3.6: Árbol filogenético realizado con el modelo maximum likelihood sobre las chloridoideae incluyendo a *Z. mays* como outgroup. El árbol fue realizado sobre 1920 genes presentes en todas las especies. En verde se resaltan los cultivares pertenecientes a la especie *E. curvula*.

### **Análisis comparativo de los cultivares de *E. curvula* y la región ligada a la apomixis**

Para poder identificar genes condicionantes y/o regulatorios de la apomixis se compararon los genomas de *E. curvula* secuenciados hasta el momento en la región específica ligada a la apomixis (Zappacosta et al., 2019). Esta región surgió del análisis la población de 61

individuos híbridos productos del cruzamiento de Don Walter, como dador de polen, y OTA como receptor. Esta población junto a sus padres fue genotipada por secuenciación (GBS) y sus marcadores de 60 pb fueron utilizados para realizar el mapa de ligamiento de Don Walter y OTA. Dentro del grupo de ligamiento 3 en el mapa de ligamiento de Don Walter se encontraron cuatro marcadores ligados al carácter de la apomixis los cuales fueron mapeados en todos los genomas. Los cuatro marcadores llamados TP79423, TP107627, TP135456 y TP95591 alinearon en las bases 18.720.221, 24.698.616, 26.219.968 y 29.212.634 del Contig28, respectivamente, formando una única región delimitada por 10.492.089 pb en el genoma de Victoria respaldando el agrupamiento del mapa de ligamiento. Los primeros tres marcadores alinearon con los genes EJB05\_34569, EJB05\_35038 y EJB05\_35146 (nomenclatura de la base de datos NCBI) mientras que el marcador TP95591 alineó con una región intergénica.

La fragmentación y la ploidía del genoma de Don Walter no permitió ubicar todos los marcadores en una sola secuencia estando los marcadores TP79423, TP107627 y TP135456 en dos secuencias cada uno en regiones génicas mientras que el marcador TP95591 alineó con un sola secuencia en una región intergénica. Por otro lado, el mayor grado de fragmentación en el genoma de Tanganyika INTA impidió aún más determinar una sola región alineando los marcadores TP79423 y TP107627 con dos secuencias en regiones génicas, el marcador TP135456 con un gen y al igual que en el resto de los genomas el marcador TP95591 alineó en una región intergenica.

Con el fin de explorar los genes que en todos los genomas se ubican dentro de la región, se realizó un análisis de sintenia por bloques utilizando como referencia el genoma de Victoria debido a que contiene los cuatro marcadores ligados a la apomixis en el mismo contig. Para evitar errores en los alineamientos y asegurar que las secuencias de cada uno de los genomas correspondan a regiones ortológicas se tuvieron en cuenta los bloques con más de dos genes sinténicos. La región delimitada entre las bases 18.720.221 y 29.212.634 del contig28 de Victoria tiene anotados 767 genes, los cuales resultaron sinténicos con genes presentes en 59 contigs de Don Walter y 169 de Tanganyika INTA, que contienen en total 1.880 y 478 genes, respectivamente (Figura 3.7 y Tabla 3.3). Como puede observarse, en el ensamblado de Tanganyika INTA se obtuvieron mayor cantidad de contigs sinténicos que en el de Don Walter, sin embargo, es necesario tener presente que la longitud promedio de cada contig de Tanganyika INTA es de 12.019,90 pb, mientras que en Don Walter la longitud promedio de los contigs es de 292.842,67 pb. De los 1.880 genes anotados en los 59 contigs de Don Walter, 974 fueron ortólogos a Victoria, es decir, dentro de un mismo contig de Don Walter existen genes ortólogos al contig28 y genes que no lo son. De la

misma manera de los 478 genes presentes en 169 contigs de Tanganyika INTA, 398 fueron ortólogos al contig28 de Victoria (Tabla 3.3).

De los 767 genes de Victoria en la región, 156 no fueron detectados en ningún bloque sinténico, lo cual no significa que no estén en otros genomas, sino que probablemente la corta longitud de los contigs de los genomas tetraploides no permita encontrar genes adyacentes para formar un bloque, o podría deberse a cambios estructurales reales dentro del genoma. Para investigar este comportamiento se buscaron específicamente en el análisis de OrthoFinder los 156 genes de la región de Victoria no encontrados en los bloques sinténicos de Tanganyika INTA y Don Walter. Con esta estrategia pudieron identificarse 95 de estos genes, apoyando la idea de que la ausencia de genes adyacentes es debida a la corta longitud de los contigs obtenidos en estos ensamblados.

Otro factor a tener en cuenta en este análisis es el mayor grado de redundancia que existe en los genomas tetraploides con respecto a los diploides. De esta manera se encontraron que los 974 genes de Don Walter son ortólogos a 578 genes de Victoria lo que indica que en muchos casos hay más de un gen de Don Walter cubriendo un único gen de Victoria. En Tanganyika INTA sucede algo similar, encontrando 398 genes ortólogos a 359 genes de Victoria, en este caso el número de genes duplicados por gen de Victoria es menor por la menor cantidad de genes sinténicos entre estos genomas.

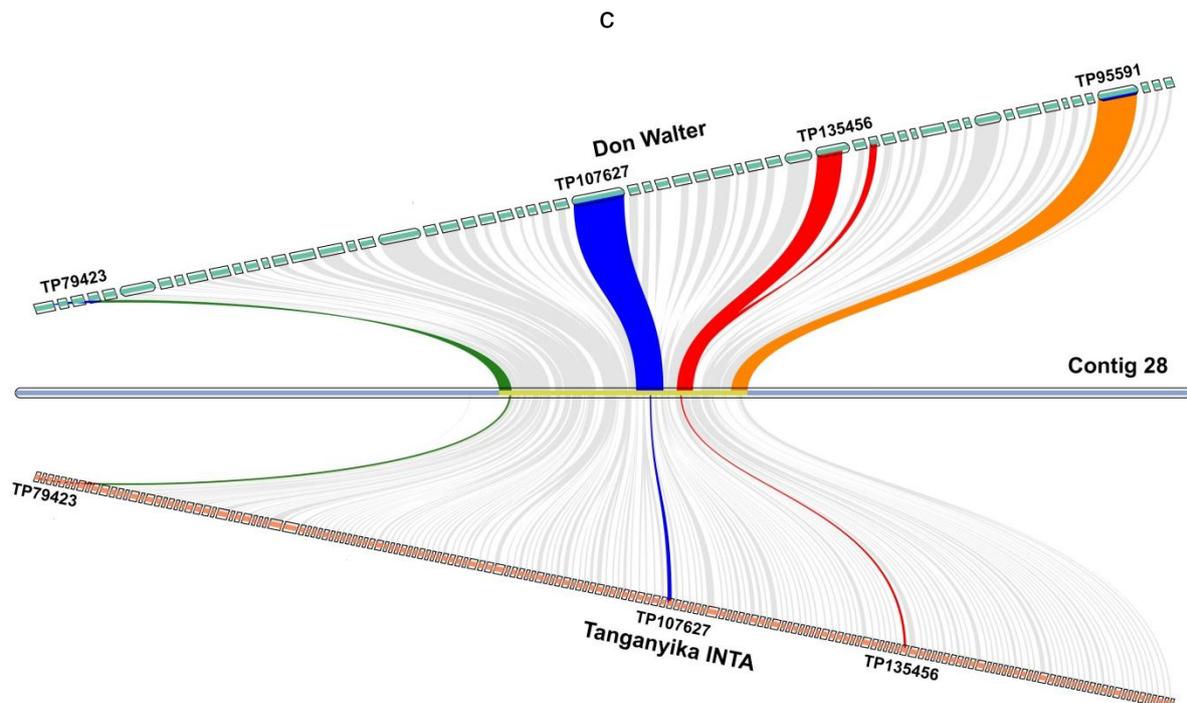


Figura 3.7: Contigs sinténicos a la región delimitada en amarillo en Victoria delimitada por los cuatro marcadores en verde, azul, rojo y naranja. En la parte superior se encuentran los 59 contigs de Don Walter donde alinearon los cuatro marcadores y en la inferior los 169 contigs de Tanganyika INTA donde alinearon tres marcadores.

	<b>Don Walter</b>	<b>Tanganyika INTA</b>
<b>Contigs en región</b>	59	169
<b>Longitud promedio de contigs</b>	292.842,67	12.019.9
<b>Genes en contig</b>	1880	478
<b>Genes ortólogos en la región</b>	974	398
<b>Genes no ortólogos en la región</b>	906	80

Tabla 3.3: Número de contigs que alinean a la región delimitada en la figura 3.7 y cantidad de genes ortólogos y no ortólogos presente en esos contigs.

Para evaluar *in silico* la expresión de los genes de la región delimitada por los marcadores, se utilizaron las lecturas de trabajos transcriptómicos previos (Garbus et al., 2017) de los cultivares Tanganyika USDA (100% apomíctico) y OTA (100% sexual) y se hizo un análisis de expresión diferencial. Utilizando esta información se evaluaron todos los genes presentes en la región de Victoria y todos los genes (otólogos y no ortólogos) presentes en los 59 y 169 contigs de Don Walter y Tanganyika INTA en la región (Tabla 3.4). De esta manera se encontraron genes de la región que estuvieron tanto sobreexpresados como reprimidos en Tanganyika USDA, es decir, genes con alta expresión en el genotipo apomíctico y baja en el sexual, y viceversa. Así de los genes de la región en Don Walter, 39 resultaron diferenciales (22 sobreexpresados y 17 reprimidos) en el genotipo apomíctico, mientras que en Tanganyika INTA se encontraron 15 genes (11 sobreexpresados y 4 reprimidos) y finalmente 11 de los genes encontrados en Victoria estuvieron reprimidos en Tanganyika USDA.

		<b>Sobreexpresados</b>	<b>Reprimidos</b>
<b>Don Walter</b>	<b>Ortólogos</b>	16	13
	<b>No ortólogo</b>	6	4
<b>Tanganyika INTA</b>	<b>Ortólogo</b>	9	4
	<b>No ortólogo</b>	2	0
<b>Victoria</b>	<b>Ortólogos</b>	0	11

Tabla 3.4: Genes sobreexpresados y reprimidos para apomixis que se encuentran en los contigs sinténicos a la región delimitada por los marcadores de la Figura 3.7. EL análisis se realizó tanto sobre los genes ortólogos a la región como los no ortólogos en los genomas de Tanganyika INTA y Don Walter.

Todos los genes presentes en la tabla 3.4 fueron anotados funcionalmente con la base de datos UniProt. El subconjunto de 11 genes encontrados en Victoria que fueron reprimidos en el transcriptoma del genotipo apomíctico fue especialmente analizado ya que podría estar relacionado a la pérdida de alguna función en la vía sexual. De los 11 genes, siete no tuvieron alineamiento en UniProt mientras que los restantes tuvieron alineamientos con Cation/H(+) antiporter 19 (*CHX19*), Serine carboxypeptidase (*SCPL27*), Receptor-like protein kinase FERONIA (*FER*) y Histone-lysine N-methyltransferase (*H3-K9-HMTase 4*).

La familia de *CHX* está muy ligada al desarrollo reproductivo y *CHX15* también fue encontrado reprimido en el análisis comparativo con arroz en el Capítulo 1. Particularmente la pérdida de función de los genes *CHX19* junto con *CHX17* y *CHX18* produce la fertilización de los núcleos polares o la ovocélula, pero no los dos, es decir, uno de los núcleos del polen no llega a su objetivo. De estos tres genes solo *CHX19* se encuentra expresado en la ovocélula y en el tubo polínico, mientras que el resto se expresa en el endosperma y tubo polínico. La mutación de *CHX18* y *CHX19* produce algunos genotipos que son polinizados pero la formación del embrión es retardada, lo cual puede provocar un asincronismo al momento de la fertilización (Padmanaban et al 2017). *SCPL27* no ha sido caracterizado funcionalmente, pero se ha comprobado su expresión durante el desarrollo reproductivo (Fraser et al., 2005). El tercer gen encontrado reprimido en Victoria, *FER*, tiene doble rol, asegura la entrega del núcleo germinativo del polen y previene la llegada de otros núcleos. En el 80% de los genotipos con el gen *FER* mutado no se induce la rotura del tubo polínico y en el 50% se obtuvieron la entrada de múltiples tubos (Duan et al., 2020). Finalmente *H3-K9-HMTase 4* es responsable de la metilación de la citosina de muchos genes especialmente en sitios distintos a CG. Varios estudios (Jackson et al., 2002, Lippman et al., 2003, Malagnac et al., 2002) han demostrado que la mutación de este gen trae aparejado la expresión de otros genes, por lo tanto en vez de cambiar o reprimir una función, como generalmente ocurre cuando se silencia un gen, desmetila genes permitiendo que se expresen activamente. Si bien no ha sido relacionado específicamente a genes vinculados a la reproducción no se podría descartar que sea un regulador epigenético de algún gen relacionado a las vías reproductivas.

Dentro de los genes de los genomas tetraploides ortólogos y no ortólogos a la región, también se analizó la anotación funcional de los genes reprimidos (Tabla 3.4). De los 17 genes hallados en la región en Don Walter que se encontraron reprimidos en el transcriptoma de Tanganyika USDA (13 ortólogos + 4 no ortólogos), 9 tuvieron anotación. Del mismo modo, de los 4 genes reprimidos en Tanganyika INTA solo uno no alineó con la base de datos UniProt. Una vez, más dentro de estos genes reprimidos identificados tanto en los contigs de la región de Don Walter como en la de Tanganyika INTA, se encontró

reprimido el gen *CHX19* previamente mencionado por lo cual se ha convertido en un candidato de gran importancia por su expresión y función. Además, al igual que en Victoria, también se encontró reprimido en la región en Don Walter el gen responsable de la metilación *H3-K9-HMTase 4*. El resto de los genes reprimidos en Tanganyika USDA y encontrados en los genotipos tetraploides no han sido relacionados al modo reproductivo debido a que no han sido estudiados o su función no ha sido vinculada a las vías reproductivas.

Dentro de los genes que se encontraron sobreexpresados en Tanganyika USDA tanto en los ortólogos como no ortólogos se investigaron aquellos que de acuerdo a su anotación funcional y la literatura tendrían relación directa con las vías reproductivas. Así de los 16 genes encontrados ortólogos en Don Walter, 4 no tuvieron anotación y solo los siguientes 4 de los anotados se encontraron relacionados a las vías reproductivas:

EMBRYONIC FACTOR 1 (*FAC1*) es esencial para el pasaje de ovocélula a embrión y en su mutación se observaron ovocélulas letales las cuales no pueden formar semillas viables (Xu et al 2005).

PROLIFERA (*PRL*) es un gen preferentemente expresado en el alelo materno durante el desarrollo gametofítico que fue encontrado expresado en los núcleos polares y en la ovocélula pero no en los núcleos del polen en *A. thaliana* (Springer et al., 2000). La mutación de este gen en ocasiones produce genotipos con sacos embrionarios sin fertilizar pero con un endosperma formado en forma autónoma lo que significa que este gen podría ser esencial para mantener el equilibrio de núcleos del endosperma y embrión (Springer et al., 2000).

PATELLIN6 (*PATL6*) se expresa en la embriogénesis temprana y su principal función es está asociada la división mitótica. La mutación de este gen produce división celular y cambios en la división celular en el ovulo lo que lleva a una desorganización de los núcleos (Tejos et al., 2018).

EMBRYO-DEFECTIVE 1006 (*EMB*) pertenece a una familia de proteínas de cloroplastos indispensables para el desarrollo embrionario. Si bien este gen no se ha caracterizado individualmente se ha comprobado que participa en el desarrollo embrionario y pigmentación del grano (Meinke et al., 2019).

De los genes sobreexpresados para la apomixis y sin ortólogos en Don Walter el único gen vinculado a la reproducción fue Serine/threonine-protein kinase (*MPS1*) previamente descrito *E. curvula* asociado a la regulación de la megagametogénesis. (Selva et al., 2020)

De los genes que se sobreexpresaron en Tanganyika INTA el que tiene mayor relación con la reproducción es ANAPHASE-PROMOTING COMPLEX/CYCLOSOME SUBUNIT 1 (*APC1*). Este gen es parte de una familia que a través de la degradación por medio de la ubiquitinación actúa como regulador y cuya función está vinculada a la especificación celular, la morfogénesis y la división celular (Dante et al 2014). *APC1* está expresada activamente durante la gametogénesis y embriogénesis y su mutación provoca defectos como arresto de embriones, desorganización celular, degradación del embrión o el núcleo polar entre otros. Dentro de estos fenotipos el más frecuente el arresto de embriones (40%) por lo que pareciera que este gen es fundamental para el desarrollo reproductivo (Wang et al., 2013). El otro gen de interés en este conjunto es FAR-RED- IMPAIRED RESPONSE1 (*FAR1*) quien junto con ED ELONGATED HYPOCOTYL3 (*FHY3*) y miR156-SPL regulan la floración (Xie et al., 2020). El miRNA miR156-SPL fue encontrado diferencialmente expresado en genotipos apomícticos en trabajos previos (Garbus et al. 2019) dando aún más evidencia de que este conjunto de genes y miRNA estaría regulando conjuntamente el desarrollo reproductivo.

## Discusión

Los avances tecnológicos en términos de secuenciación y ensamblado de genomas han llevado a una revolución que cambió el modo de abordar y resolver enigmas biológicos. Tomando como referencia el reino vegetal, en los últimos 10 años se han ingresado 1549 nuevos genomas a la base de datos NCBI (<ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/plant/>) y su crecimiento continua siendo exponencial, como ejemplo, los genomas ingresados entre el año 2019 y septiembre del 2020 ya superaron el número de genomas disponibles en la historia previa de la base de datos. El aumento de la cantidad de genomas secuenciados está muy correlacionado con la disminución de los costos de secuenciación, sin embargo, esta caída de costos no fue a través de la optimización de técnicas antiguas si no por el desarrollo de nuevas tecnologías que además aumentan significativamente la calidad final de los ensamblados (Wetterstrand, 2019 [www.genome.gov/sequencingcostsdata](http://www.genome.gov/sequencingcostsdata)).

Utilizando distintas tecnologías se han podido secuenciar y ensamblar los genomas tetraploides de dos cultivares apomícticos de *E. curvula*. De esta manera, a través del trabajo realizado en esta Tesis se ha logrado reportar tres genomas de la especie, el del genotipo diploide sexual Victoria ( $2n=2x=20$ ) (Carballo et al., 2019) y los genotipos tetraploides apomícticos ( $2n=4x=40$ ) Don Walter y Tanganyika INTA. En particular estos dos últimos genomas constituyen el primer registro en la literatura de la secuenciación y ensamblado de genomas de una especie apomíctica.

Utilizando un enfoque novedoso, se secuenció con la tecnología Chromium 10X y Oxford Nanopore el genoma de Don Walter. En el ensamblado final se obtuvieron 7.542 contigs distribuidos en 1.147,8 Mb con un N50 de 224.390 pb y 96% de los genes BUSCO. Este ensamblado híbrido representa una alternativa muy ventajosa en términos económicos significando menos de la mitad del costo de secuenciar el genoma diploide utilizando únicamente la tecnología PacBio. Para el segundo genoma secuenciado, Tanganyika INTA, se utilizó la tecnología Illumina con una longitud de lecturas pair-end de 250 pb. El ensamblado de estas lecturas dio como resultado un genoma con poca contigüidad, obteniéndose 892,9 Mb distribuidas en 293.300 contigs con un N50 de 4.725 pb. Estas métricas son debidas a que los ensamblados realizados a partir de lecturas cortas no permiten cubrir extensivamente a los elementos repetitivos (Treangen et al., 2012). La decisión de secuenciar el genoma de Tanganyika INTA estuvo fundamentada en el hecho de que este cultivar dio origen, a través del cultivo *in vitro* de inflorescencias inmaduras, al genotipo diploide sexual Victoria (Cardone et al., 2006) por lo que, teniendo distinto modo

reproductivo, comparten el mismo background genético. De esta manera se podrían asociar las diferencias con genes o regiones génicas involucradas en la apomixis. Por lo tanto, el ensamblado logrado ha cumplido con las expectativas a pesar de no poseer una alta contigüidad.

La anotación de los elementos repetitivos en estos genomas representa un desafío ya que en la medida que baja la contigüidad, su descubrimiento por medio de programas bioinformático es más complejo, ya sea porque la alta fragmentación de su estructura no permite una clasificación precisa o a que directamente algunas de las secuencias repetitivas son anotadas como codificantes. Para los genomas tetraploides se adoptó una estrategia novedosa respecto de la utilizada en el capítulo anterior, básicamente debido a las diferentes características de los ensamblados. El porcentaje de nucleótidos cubiertos por repeticiones en Don Walter y Tanganyika INTA fue muy similar, 54,02% y 54,96%, respectivamente. A pesar de ello, la baja contigüidad de cada uno de estos genomas y la diferencia entre ellos en este parámetro, se vio reflejada en la cantidad de elementos no clasificados que representaron un 31,11% en Tanganyika INTA y 16,9% en Don Walter. Por otro lado, la proporción de elementos repetitivos fue aproximadamente el doble que en el genotipo diploide. En otras gramíneas como por ejemplo el trigo hexaploide *T. aestivum* (AABBDD;  $2n=6x=42$ ), que tiene 85% de elementos repetitivos (Appels et al., 2018) y su antecesor tetraploide *T. dicoccoides* (AABB;  $2n=4x=28$ ) un 82,2% (Avni et al., 2017), mientras que en los genomas diploides que le dieron su origen, *T. urartu* (AA,  $2n=2x=14$ ) y *Ae. tauschii* (DD,  $2n=2x=14$ ), el contenido de elementos repetitivos es de 66.8% (Ling et al., 2013) y 84.4% (Luo et al., 2017), respectivamente. Además, tomando como referencia lo observado en el genoma diploide, en el que la cuantificación de elementos repetitivos del ensamblado de las lecturas de PacBio cubrió 37,05% mientras que en el ensamblado final con el agregado de las lecturas de Chicago y Hi-C el porcentaje disminuyó al 28,7% y aumentó notablemente la contigüidad, se puede inferir que probablemente el elevado porcentaje de elementos repetitivos de las especies tetraploides sea un artefacto dado por la escasa contigüidad de sus ensamblados y por factores inherentes a la especie.

La anotación de genes de los genotipos tetraploides, al igual que en Victoria, parece estar sobreestimada principalmente en el cultivar Tanganyika INTA en el que, teniendo ensambladas menor cantidad de bases que Don Walter, se anotaron 38.314 genes más. Esto puede ser porque algunas regiones no pudieron colapsarse para formar el haplotipo en el ensamblado y por lo tanto algunos genes pudieron ser anotados dos o más veces. Por otro lado, como se dijo anteriormente, puede ser que algunos de los genes anotados en realidad correspondan a elementos repetitivos.

El análisis de genes ortólogos de los tres genomas indica que hay solo 5.527 genes específicos de Victoria y este número se eleva significativamente en Don Walter y Tanganyika INTA, con 36.201 y 42.461 genes específicos, respectivamente. Nuevamente, el factor contigüidad afecta estos números. De todos modos, el total de genes indispensables dentro de una especie, llamados “genes core”, varía según la especie. Por ejemplo, en un análisis de distintos cultivares de *O. sativa* se encontró que el 92,17% de los genes pertenecen al core y 7,83% son dispensables, es decir, generan variaciones específicas (Schatz et al., 2014). En *B. distachyon* y *Brassica oleracea* el porcentaje de genes core fue de 55% y 81,29%, respectivamente (Gordon et al., 2017, Golicz et al., 2016b). En el caso particular de *E. curvula*, además de la influencia en el número de genes específicos de la sobreestimación de genes anotados, puede pensarse que entre estos genes específicos de cada cultivar se encontrarían aquellos responsables de las características particulares distintivas, dentro de las cuales se encuentra el modo reproductivo.

Las nuevas facilidades de secuenciación también han permitido que se secuencien y anoten otras especies del género *Eragrostis* y de la subfamilia Chloridoideae. Gracias a esto se realizó un análisis filogenético de la subfamilia y se incorporaron genes de otros cultivares extraídos de análisis transcriptómicos previos (Garbus et al., 2017). Con este análisis se corroboró que Victoria es el cultivar más cercano a Tanganyika INTA y que los padres de la población de mapeo, Don Walter y OTA, serían más parecidos entre sí que lo que lo es Tanganyika USDA con respecto a ellos, pudiendo esta observación explicar el motivo por el cual este último no pudo ser cruzado en trabajos previos (Zappacosta et al., 2019).

Con toda la información surgida de todos los genomas, junto con el mapa de ligamiento de la especie fue posible delimitar una región sobre el genoma de Victoria en la que alinearon todos los marcadores ligados a la apomixis dándole otra fuente de validación al mapa (Zappacosta et al., 2019). Este análisis se realizó sobre el genoma de Victoria dado que es el que presenta mayor contigüidad correspondiendo los 7 contigs más largos a cromosomas completos. También valida una de las hipótesis de partida, en la que se proponía que este primer genoma diploide de la especie podría utilizarse como esqueleto para guiar el análisis de genomas más complejos de la especie. Luego se extrajeron los contigs sinténicos a esta región de Don Walter y Tanganyika INTA y se analizaron los genes ortólogos y no ortólogos de estos con los genes de la región en Victoria. Pudo observarse que existe un alto grado de sintenia entre los cultivares y que existen contigs en los individuos tetraploides con regiones sinténicas y regiones no sinténicas, es decir, hay contigs que tienen una fracción con genes ortólogos a la región delimitada por los marcadores en Victoria y otra fracción dentro del mismo contig que no es ortóloga. En otras especies apomícticas se ha encontrado una región vinculada a la apomixis que no tendría recombinación y que no

estaría en genotipos sexuales (Ozias-Akins et al., 1993, Sherwood et al., 1994, Ozias-Akins et al., 1998, Goel et al., 2003, Conner et al., 2008). Lamentablemente, no es posible analizar exhaustivamente y realizar comparaciones entre las regiones definidas en *E. curvula* y las establecidas en estas especies apomícticas, ya que sus genomas aún no se encuentran secuenciados. Los resultados obtenidos hasta aquí en *Eragrostis* indican que dentro de la región hay zonas comunes entre los genotipos apomícticos y sexuales y zonas particulares de los individuos apomícticos, por lo que la región podría ser más reducida que la encontrada a través del mapa de ligamiento o, alternativamente, dentro de la región coexistan zonas recombinantes y no recombinantes. El hecho de que hayan encontrado los marcadores en el genoma de Victoria indica que la resolución del mapa debería ser incrementada para identificar más precisamente las regiones específicas de los genomas tetraploides a través de la incorporación de mayor cantidad de híbridos al mapa de ligamiento.

Dentro de la región se realizó un análisis de expresión diferencial tomando información transcriptómica previamente secuenciada de individuos full apomícticos y full sexuales donde se encontraron genes sobreexpresados y reprimidos (Garbus et al., 2017). Por un lado se analizaron los genes reprimidos que podrían estar desregulando las vías sexuales donde se encontraron *CHX19*, *FER*, *H3-K9-HMTase 4*. La represión de estos genes podría tener un rol fundamental en el desarrollo reproductivo ya que la pérdida de función de *CHX19*, *FER* produce la fertilización de solo un núcleo polar y regulan la entrega del núcleo germinativos del polen y *H3-K9-HMTase 4* es responsable del inicio de la metilación de otros genes por lo que podría estar liberando su expresión. Dentro de los genes sobreexpresados los que se encontraron vinculados a la reproducción fueron *FAC1*, *PRL*, *EMB*, *PATL6*, *APC1* y *FAR1*. Este último actúa junto con un miRNA-159 el cual también fue encontrado diferencial para la apomixis en trabajos previos (Garbus et al., 2019).

Los resultados encontrados en este Capítulo de la Tesis, sustentan muchas de las hipótesis y líneas de investigación delineadas para la especie actualmente. El alto grado de heterosis de los tetraploides y el diploide, indica claramente que al igual que en otras especies (Paun et al., 2006, Šarhanová et al., 2017, Weber 2016), en *E. curvula* la apomixis está relacionada con la heterosis, probablemente como resultado de hibridaciones y mutaciones. Por otra parte, la identificación de una región en el mapa de ligamiento que coincide con el orden genómico en los genomas secuenciados da indicios de que alguno de los componentes de la apomixis estaría regulado por los genes dentro de la región delimitada al igual que ocurre en *Pennisetum* y *Cenchrus* (Conner et al., 2008). Además en esta región se encontraron genes involucrados en el control epigenético como *H3-K9-HMTase 4* y *FAR1* que respaldan los resultados encontrados en trabajos previos sobre microARN en *E.*

*curvula* y el rol importante que tiene la metilación en la regulación de la apomixis (Garbus et al., 2019, Podio et al., 2014, García-Aguilar et al., 2010). Todos estos hallazgos no solo respaldan las hipótesis formuladas previamente, sino que además representan un avance en términos del conocimiento de la regulación a través de la región, genes candidato, y posibles funciones de estos genes.

## Conclusión general

Por medio del trabajo que se presenta en esta Tesis se han podido secuenciar y ensamblar los genomas de tres cultivares de *E. curvula*, dos tetraploides apomícticos facultativos, Don Walter y Tanganyika INTA, y el cultivar diploide sexual Victoria, utilizando un abanico importante de tecnologías de secuenciación de última generación. Como postulamos en el comienzo, la disponibilidad de un ensamblado de alta calidad del genotipo diploide constituyó un andamiaje clave para el ensamblado de los genotipos tetraploides, logrados con menor contigüidad. A partir de estos tres ensamblados se pudieron anotar los genes y los elementos repetitivos de cada genotipo. Si bien los mecanismos de la apomixis están aún inconclusos, gracias a la secuenciación genómica se pudieron obtener una serie de candidatos que a priori parecen estar vinculados en su regulación, lo que constituye uno de los aportes más relevantes de este trabajo. Quedan abiertas incógnitas respecto a la región apomíctica que podrían cerrarse a través de la secuenciación de los genomas tetraploides presentados aquí con tecnologías complementarias, para mejorar sus N50 y sus contigüidades. Por otro lado la secuenciación de un genotipo tetraploide 100% sexual aportaría más precisión al mapa de ligamiento y a los análisis comparativos.

La disponibilidad de estos genomas representa una herramienta en sí misma, ya que a través de ellos se pueden identificar otros genes de interés agronómico además de los aquí reportados para ser utilizados en programas de mejoramiento, diseñar marcadores moleculares para su análisis, realizar estudios filogenéticos y evolutivos, constituyendo una contribución muy importante para la comunidad científica

## Referencias

Aida M, Beis D, Heidstra R, Willemsen V, Blilou I, Galinha C, Nussaume L, Noh YS, Amasino R, Scheres B. The PLETHORA genes mediate patterning of the Arabidopsis root stem cell niche. *Cell*. 2004 Oct 1;119(1):109-20.

Albertini E, Barcaccia G, Carman JG, Pupilli F. Did apomixis evolve from sex or was it the other way around?. *Journal of experimental botany*. 2019 May 15;70(11):2951-64.

Alonge M, Shumate A, Puiu D, Zimin A, Salzberg SL. Chromosome-scale assembly of the bread wheat genome, *Triticum aestivum*, reveals over 5700 new genes. *BioRxiv*. 2020 Jan 1.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of molecular biology*. 1990 Oct 5;215(3):403-10.

Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends in biochemical sciences*. 1998 Nov 1;23(11):444-7.

Alves-Ferreira M, Wellmer F, Banhara A, Kumar V, Riechmann JL, Meyerowitz EM. Global expression profiling applied to the analysis of Arabidopsis stamen development. *Plant Physiology*. 2007 Nov 1;145(3):747-62.

Amiteye S, Corral JM, Vogel H, Kuhlmann M, Mette MF, Sharbel TF. Novel microRNAs and microsatellite-like small RNAs in sexual and apomictic *Boechera* species. *MicroRNA*. 2013 Apr 1;2(1):46-63.

Anderson SN, Johnson CS, Chesnut J, Jones DS, Khanday I, Woodhouse M, Li C, Conrad LJ, Russell SD, Sundaresan V. The zygotic transition is initiated in unicellular plant zygotes with asymmetric activation of parental genomes. *Developmental cell*. 2017 Nov 6;43(3):349-58.

Anderson SN, Johnson CS, Jones DS, Conrad LJ, Gou X, Russell SD, Sundaresan V. Transcriptomes of isolated *Oryza sativa* gametes characterized by deep sequencing: evidence for distinct sex-dependent chromatin and epigenetic states before fertilization. *The Plant Journal*. 2013 Dec;76(5):729-41.

Appels R, Eversole K, Stein N, Feuillet C, Keller B, Rogers J, Pozniak CJ, Choulet F, Distelfeld A, Poland J, Ronen G. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*. 2018 Aug 17;361(6403).

Asker S, Jerling L. Apomixis in plants. CRC press; 1992 Mar 10.

Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K, Jordan KW. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science*. 2017 Jul 7;357(6346):93-7.

Bai G, Tefera H, Ayele M, Nguyen HT. A genetic linkage map of tef [*Eragrostis tef* (Zucc.) Trotter] based on amplified fragment length polymorphism. *Theoretical and Applied Genetics*. 1999 Aug 1;99(3-4):599-604.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*. 2009 Jul 1;37(suppl\_2):W202-8.

Balsamo RA, Willigen CV, Bauer AM, Farrant J. Drought tolerance of selected *Eragrostis* species correlates with leaf tensile properties. *Annals of botany*. 2006 Jun 1;97(6):985-91.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Pribelski AD, Pyshkin AV. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*. 2012 May 1;19(5):455-77.

Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna*. 2015 Dec 1;6(1):11.

Beck JB, Alexander PJ, Allphin L, Al-Shehbaz IA, Rushworth C, Bailey CD, Windham MD. Does hybridization drive the transition to asexuality in diploid *Boechera*?. *Evolution: International Journal of Organic Evolution*. 2012 Apr;66(4):985-95.

Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*. 2000 Mar;25(1):60-83.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM. Accurate whole human genome sequencing using reversible terminator chemistry. *nature*. 2008 Nov;456(7218):53-9.

Bonny S. Corporate concentration and technological change in the global seed industry. *Sustainability*. 2017 Sep;9(9):1632.

Book AJ, Smalle J, Lee KH, Yang P, Walker JM, Casper S, Holmes JH, Russo LA, Buzzinotti ZW, Jenik PD, Vierstra RD. The RPN5 subunit of the 26S proteasome is essential

for gametogenesis, sporophyte development, and complex assembly in Arabidopsis. *The Plant Cell*. 2009 Feb 1;21(2):460-78.

Boutilier K, Offringa R, Sharma VK, Kieft H, Ouellet T, Zhang L, Hattori J, Liu CM, van Lammeren AA, Miki BL, Custers JB. Ectopic expression of BABY BOOM triggers a conversion from vegetative to embryonic growth. *The Plant Cell*. 2002 Aug 1;14(8):1737-49.

Braslavsky I, Hebert B, Kartalov E, Quake SR. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences*. 2003 Apr 1;100(7):3960-4.

Bultosa G. Physicochemical characteristics of grain and flour in 13 tef [*Eragrostis tef* (Zucc.) Trotter] grain varieties. *Journal of Applied Sciences Research*. 2007;3(12):2042-50.

Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature biotechnology*. 2013 Dec;31(12):1119.

Calzada JP, Crane CF, Stelly DM. Apomixis--the asexual revolution. *Science*. 1996 Nov 22;274(5291):1322-3.

Campbell MS, Holt C, Moore B, Yandell M. Genome annotation and curation using MAKER and MAKER-P. *Current protocols in bioinformatics*. 2014 Dec;48(1):4-11.

Cannarozzi G, Plaza-Wüthrich S, Esfeld K, Larti S, Wilson YS, Girma D, de Castro E, Chanyalew S, Blösch R, Farinelli L, Lyons E. Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (*Eragrostis tef*). *BMC genomics*. 2014 Dec;15(1):581.

Cardone S, Polci P, Selva JP, Mecchia M, Pessino S, Hermann P, Cambi V, Voigt P, Spangenberg G, Echenique V. Novel genotypes of the subtropical grass *Eragrostis curvula* for the study of apomixis (diplospory). *Euphytica*. 2006 Sep 1;151(2):263-72.

Carman JG. Asynchronous expression of duplicate genes in angiosperms may cause apomixis, bispority, tetraspority, and polyembryony. *Biological Journal of the Linnean Society*. 1997 May 1;61(1):51-94.

Carman, J. C. "The gene effect: genome collisions and apomixis." 2001.

Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC genomics*. 2012 Dec 1;13(1):375.

Cervigni GD, Paniego N, Díaz M, Selva JP, Zappacosta D, Zanazzi D, Landerreche I, Martelotto L, Felitti S, Pessino S, Spangenberg G. Expressed sequence tag analysis and development of gene associated markers in a near-isogenic plant system of *Eragrostis curvula*. *Plant molecular biology*. 2008 May 1;67(1-2):1-0.

Chen L, Auh CK, Dowling P, Bell J, Chen F, Hopkins A, Dixon RA, Wang ZY. Improved forage digestibility of tall fescue (*Festuca arundinacea*) by transgenic down-regulation of cinnamyl alcohol dehydrogenase. *Plant Biotechnology Journal*. 2003 Nov;1(6):437-49.

Chen L, Miyazaki C, Kojimai A, Saito A, Adachi T. Isolation and characterization of a gene expressed during early embryo sac development in apomictic guinea grass (*Panicum maximum*). *Journal of plant physiology*. 1999 Jan 1;154(1):55-62.

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods*. 2013 Jun;10(6):563.

Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H, Lipscombe J. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome research*. 2017 May 1;27(5):885-96.

Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International journal of plant genomics*. 2008;2008.

Conner JA, Goel S, Gunawan G, Cordonnier-Pratt MM, Johnson VE, Liang C, Wang H, Pratt LH, Mullet JE, DeBarry J, Yang L. Sequence analysis of bacterial artificial chromosome clones from the apospory-specific genomic region of *Pennisetum* and *Cenchrus*. *Plant physiology*. 2008 Jul 1;147(3):1396-411.

Conner JA, Mookkan M, Huo H, Chae K, Ozias-Akins P. A parthenogenesis gene of apomict origin elicits embryo formation from unfertilized eggs in a sexual plant. *Proceedings of the National Academy of Sciences*. 2015 Sep 8;112(36):11205-10.

Conner JA, Podio M, Ozias-Akins P. Haploid embryo production in rice and maize induced by PsASGR-BBML transgenes. *Plant reproduction*. 2017 Mar 1;30(1):41-52.

Da Maia LC, Palmieri DA, De Souza VQ, Kopp MM, de Carvalho FI, Costa de Oliveira A. SSR locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *International journal of plant genomics*. 2008;2008.

Dante RA, Larkins BA, Sabelli PA. Cell cycle control and seed development. *Frontiers in Plant science*. 2014 Sep 23;5:493.

De Oliveira FA, Vigna BB, da Silva CC, Fávero AP, de Matta FP, Azevedo AL, de Souza AP. Coexpression and Transcriptome analyses identify active Apomixis-related genes in *Paspalum notatum* leaves. *BMC genomics*. 2020 Dec 1;21(1):78.

D'Erfurth I, Jolivet S, Froger N, Catrice O, Novatchkova M, Mercier R. Turning meiosis into mitosis. *PLoS biology*. 2009 Jun;7(6).

Di Fino LM, D'Ambrosio JM, Tejos R, van Wijk R, Lamattina L, Munnik T, Pagnussat GC, Laxalt AM. Arabidopsis phosphatidylinositol-phospholipase C2 (PLC2) is required for female gametogenesis and embryo development. *Planta*. 2017 Apr 1;245(4):717-28.

Díaz ML, Garbus I, Echenique V. Allele-specific expression of a weeping lovegrass gene from the lignin biosynthetic pathway, caffeoyl-coenzyme A 3-O-methyltransferase. *Molecular breeding*. 2010 Dec 1;26(4):627-37.

Du Y, Scheres B. PLETHORA transcription factors orchestrate de novo organ patterning during Arabidopsis lateral root outgrowth. *Proceedings of the National Academy of Sciences*. 2017 Oct 31;114(44):11709-14.

Duan Q, Liu MC, Kita D, Jordan SS, Yeh FL, Yvon R, Carpenter H, Federico AN, Garcia-Valencia LE, Eyles SJ, Wang CS. FERONIA controls pectin-and nitric oxide-mediated male-female interaction. *Nature*. 2020 Mar;579(7800):561-6.

Dubin MJ, Scheid OM, Becker C. Transposons: a blessing curse. *Current opinion in plant biology*. 2018 Apr 1;42:23-9.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*. 2004 Mar 1;32(5):1792-7.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010 Oct 1;26(19):2460-1.

Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, Smith RD, Teresi SJ, Nelson AD, Wai CM, Alger EI. Origin and evolution of the octoploid strawberry genome. *Nature genetics*. 2019 Mar;51(3):541-7.

- Egert A, Peters S, Guyot C, Stieger B, Keller F. An Arabidopsis T-DNA insertion mutant for galactokinase (AtGALK, At3g06580) hyperaccumulates free galactose and is insensitive to exogenous galactose. *Plant and Cell Physiology*. 2012 May 1;53(5):921-9.
- Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009 Jan 2;323(5910):133-8.
- Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics*. 2008 Dec 1;9(1):18.
- Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome biology*. 2015 Dec;16(1):157.
- Eulgem T, Somssich IE. Networks of WRKY transcription factors in defense signaling. *Current opinion in plant biology*. 2007 Aug 1;10(4):366-71.
- Fagerlind F. Die Terminologie der Apomixis-Prozesse. *Hereditas*. 1940 Jan;26(1-2):1-22.
- Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Jou WM, Molemans F, Raeymaekers A, Van den Berghe A, Volckaert G. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*. 1976 Apr;260(5551):500-7.
- Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztányi Z, El-Gebali S, Fraser M, Gough J. InterPro in 2017—beyond protein family and domain annotations. *Nucleic acids research*. 2017 Jan 4;45(D1):D190-9.
- Fraser CM, Rider LW, Chapple C. An expression and bioinformatics analysis of the Arabidopsis serine carboxypeptidase-like gene family. *Plant physiology*. 2005 Jun 1;138(2):1136-48.
- Fu H, Reis N, Lee Y, Glickman MH, Vierstra RD. Subunit interaction maps for the regulatory particle of the 26S proteasome and the COP9 signalosome. *The EMBO journal*. 2001 Dec 17;20(24):7096-107.
- Galbraith DW, Harkins KR, Maddox JM, Ayres NM, Sharma DP, Firoozabady E. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science*. 1983 Jun 3;220(4601):1049-51.

Garbus I, Romero JR, Selva JP, Pasten MC, Chinestra C, Carballo J, Zappacosta DC, Echenique V. De novo transcriptome sequencing and assembly from apomictic and sexual *Eragrostis curvula* genotypes. *PloS one*. 2017;12(11).

Garbus I, Selva JP, Pasten MC, Bellido AM, Carballo J, Albertini E, Echenique V. Characterization and discovery of miRNA and miRNA targets from apomictic and sexual genotypes of *Eragrostis curvula*. *BMC genomics*. 2019 Dec 1;20(1):839.

Garcia-Aguilar M, Michaud C, Leblanc O, Grimanelli D. Inactivation of a DNA methylation pathway in maize reproductive organs results in apomixis-like phenotypes. *The Plant Cell*. 2010 Oct 1;22(10):3249-67.

Gargano AO, Adúriz MA, Arelovich HM, Amela MI. Forage yield and nutritive value of *Eragrostis curvula* and *Digitaria eriantha* in central-south semi-arid Argentina. *Tropical Grasslands*. 2001 Sep 1;35(3):161-7.

Gaut BS, Doebley JF. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences*. 1997 Jun 24;94(13):6809-14.

Gebre YG, Bertolini E, Pè ME, Zuccolo A. Identification and characterization of abundant repetitive sequences in *Eragrostis tef* cv. Enatite genome. *BMC plant biology*. 2016 Dec;16(1):1-3.

Gilbert W, Maxam A. The nucleotide sequence of the lac operator. *Proceedings of the National Academy of Sciences*. 1973 Dec 1;70(12):3581-4.

Giordano A, Liu Z, Panter SN, Dimech AM, Shang Y, Wijesinghe H, Fulgueras K, Ran Y, Mouradov A, Rochfort S, Patron NJ. Reduced lignin content and altered lignin composition in the warm season forage grass *Paspalum dilatatum* by down-regulation of a Cinnamoyl CoA reductase gene. *Transgenic research*. 2014 Jun 1;23(3):503-17.

Goel S, Chen Z, Conner JA, Akiyama Y, Hanna WW, Ozias-Akins P. Delineation by fluorescence in situ hybridization of a single hemizygous chromosomal region associated with aposporous embryo sac formation in *Pennisetum squamulatum* and *Cenchrus ciliaris*. *Genetics*. 2003 Mar 1;163(3):1069-82.

Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CK, Severn-Ellis A, McCombie WR, Parkin IA, Paterson AH. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature communications*. 2016 Nov 11;7(1):1-8.

Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S, Stritt C, Roulin AC, Schackwitz W, Tyler L, Martin J. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature communications*. 2017 Dec 19;8(1):1-3.

Grimanelli D. Epigenetic regulation of reproductive development and the emergence of apomixis in angiosperms. *Current Opinion in Plant Biology*. 2012 Feb 1;15(1):57-62.

Grossniklaus U. A bright future for apomixis. *Trends Plant Sci*. 1998;3:415-6.

Grossniklaus U. From sexuality to apomixis: molecular and genetic approaches. The flowering of apomixis: From mechanisms to genetic engineering. 2001:168-211.

Gustine DL, Sherwood RT, Huff DR. Apospory-linked molecular markers in buffelgrass. *Crop science*. 1997;37(3):947-51.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*. 2013 Aug;8(8):1494-512.

Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular biology and evolution*. 2013 Aug 1;30(8):1987-97.

Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic acids research*. 2010 Sep 29;38(22):e199-.

Hojsgaard D, Klatt S, Baier R, Carman JG, Hörandl E. Taxonomy and biogeography of apomixis in angiosperms and associated biodiversity characteristics. *Critical Reviews in Plant Sciences*. 2014 Sep 3;33(5):414-27.

Hojsgaard D. Apomixis Technology: Separating the Wheat from the Chaff. *Genes*. 2020 Apr;11(4):411.

Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A. Structure of a ribonucleic acid. *Science*. 1965 Mar 19:1462-5.

Horstman A, Willemsen V, Boutilier K, Heidstra R. AINTEGUMENTA-LIKE proteins: hubs in a plethora of networks. *Trends in plant science*. 2014 Mar 1;19(3):146-57.

Ingram AL, Doyle JJ. The origin and evolution of *Eragrostis tef* (Poaceae) and related polyploids: evidence from nuclear waxy and plastid rps16. *American Journal of Botany*. 2003 Jan;90(1):116-22.

Jackson JP, Lindroth AM, Cao X, Jacobsen SE. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature*. 2002 Apr;416(6880):556-60.

Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJ, Ohyanagi H, Mineta K, Mitchell CT, Saber N, Kharbatia NM. The genome of *Chenopodium quinoa*. *Nature*. 2017 Feb;542(7641):307.

Jiang Y, Liang G, Yu D. Activated expression of WRKY57 confers drought tolerance in *Arabidopsis*. *Molecular plant*. 2012 Nov 1;5(6):1375-88.

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin CS, Guill K. Improved maize reference genome with single-molecule technologies. *Nature*. 2017 Jun;546(7659):524-7. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, De Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*. 2008 Dec 15;24(24):2938-9.

Jiao, W.B., Accinelli, G.G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., Willing, E.M., Piednoel, M., Woetzel, S., Madrid-Herrero, E. and Huettel, B., 2017. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome research*, 27(5), pp.778-786.

Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic acids research*. 2019 Jan 8;47(D1):D590-5.

Katz, Yarden, et al. "Analysis and design of RNA sequencing experiments for identifying isoform regulation." *Nature methods* 7.12 (2010): 1009-1015.

Kaul S, Koo HL, Jenkins J, Rizzo M, Rooney T, Tallon LJ, Feldblyum T, Nierman W, Benito MI, Lin X, Town CD. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *nature*. 2000 Dec 14;408(6814):796-815.

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, Childs KL. Improvement of the *Oryza sativa*

Nipponbare reference genome using next generation sequence and optical map data. *Rice*. 2013 Dec 1;6(1):4.

Kent F. Forrajeras cultivadas anuales y perennes más difundidas en la provincia de La Pampa. Ediciones INTA Anguil; 2019 Sep 26.

Khanday I, Skinner D, Yang B, Mercier R, Sundaresan V. A male-expressed rice embryogenic trigger redirected for asexual propagation through seeds. *Nature*. 2019 Jan;565(7737):91-5.

Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. *Genome research*. 2011 Mar 1;21(3):487-93.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology*. 2019 Aug;37(8):907-15.

Kishimoto N, Higo H, Abe K, Arai S, Saito A, Higo K. Identification of the duplicated segments in rice chromosomes 1 and 5 by linkage analysis of cDNA markers of known functions. *Theoretical and Applied Genetics*. 1994 Aug 1;88(6-7):722-6.

Kitakura S, Vanneste S, Robert S, Löffke C, Teichmann T, Tanaka H, Friml J. Clathrin mediates endocytosis and polar distribution of PIN auxin transporters in *Arabidopsis*. *The Plant Cell*. 2011 May 1;23(5):1920-31.

Koch L. Technique: Chicago HighRise for genome scaffolding. *Nature Reviews Genetics*. 2016 Feb 29;17(4):194.

Koch MA, Dobeš C, Mitchell-Olds T. Multiple hybrid formation in natural populations: concerted evolution of the internal transcribed spacer of nuclear ribosomal DNA (ITS) in North American *Arabis divaricarpa* (Brassicaceae). *Molecular Biology and Evolution*. 2003 Mar 1;20(3):338-50.

Koltunow AM, Bicknell RA, Chaudhury AM. Apomixis: molecular strategies for the generation of genetically identical seeds without fertilization. *Plant Physiology*. 1995 Aug;108(4):1345.

Koltunow AM, Grossniklaus U. Apomixis: a developmental perspective. *Annual review of plant biology*. 2003 Jun;54(1):547-74.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*. 2017 May 1;27(5):722-36.

Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics*. 23, 1289–1291 (2007).

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*. 2018 Jun 1;35(6):1547-9.

Kumar S. Epigenetic control of apomixis: a new perspective of an old enigma. *Adv. Plants Agric. Res.* 2017;7(00243):10-5406.

Labombarda P, Busti A, Caceres ME, Pupilli F, Arcioni S. An AFLP marker tightly linked to apomixis reveals hemizyosity in a portion of the apomixis-controlling locus in *Paspalum simplex*. *Genome*. 2002 Jun 1;45(3):513-9.

Lambers H, Chapin III FS, Pons TL. *Plant physiological ecology*. Springer Science & Business Media; 2008 Oct 8.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods*. 2012 Apr;9(4):357.

Laspina NV, Vega T, Seijo JG, González AM, Martelotto LG, Stein J, Podio M, Ortiz JP, Echenique VC, Quarín CL, Pessino SC. Gene expression analysis at the onset of aposporous apomixis in *Paspalum notatum*. *Plant molecular biology*. 2008 Aug 1;67(6):615-28.

Lauric IA, Marinissen A, Marinozzi C. Curvas de producción y calidad de pasto llorón bajo condiciones reales de producción en el partido de bahía blanca. 2012.

Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. Software for computing and annotating genomic ranges. *PLoS Comput Biol*. 2013 Aug 8;9(8):e1003118.

Leblanc O, Grimaneli D, Gonzalez-de-Leon D, Savidan Y. Detection of the apomictic mode of reproduction in maize-Tripsacum hybrids using maize RFLP markers. *Theoretical and Applied Genetics*. 1995 Jun 1;90(7-8):1198-203.

Leigh JH. The relative palatability of various varieties of weeping lovegrass (*Eragrostis curvula* (Schrad) Nees). *Grass and Forage Science*. 1961 Jun;16(2):135-40.

León G, Holuigue L, Jordana X. Mitochondrial complex II is essential for gametophyte development in *Arabidopsis*. *Plant physiology*. 2007 Apr 1;143(4):1534-46.

Li L, He Y, Wang Y, Zhao S, Chen X, Ye T, Wu Y, Wu Y. Arabidopsis PLC 2 is involved in auxin-modulated reproductive development. *The Plant Journal*. 2015 Nov;84(3):504-15.

Li X, Chen W, Zhao Y, Xiang Y, Jiang H, Zhu S, Cheng B. Downregulation of caffeoyl-CoA O-methyltransferase (CCoAOMT) by RNA interference leads to reduced lignin production in maize straw. *Genetics and molecular biology*. 2013;36(4):540-6.

Lieberman-Aiden E, Van Berkum NL, Williams L, Imapkaev M, Ragozcy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*. 2009 Oct 9;326(5950):289-93.

Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y, Gao C. Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature*. 2013 Apr;496(7443):87-90.

Lippman Z, May B, Yordan C, Singer T, Martienssen R. Distinct mechanisms determine transposon inheritance and methylation via small interfering RNA and histone modification. *PLoS Biol*. 2003 Dec 22;1(3):e67.

Long JM, Liu Z, Wu XM, Fang YN, Jia HH, Xie ZZ, Deng XX, Guo WW. Genome-scale mRNA and small RNA transcriptomic insights into initiation of citrus apomixis. *Journal of experimental botany*. 2016 Oct 1;67(19):5743-56.

Lovell JT, Jenkins J, Lowry DB, Mamidi S, Sreedasyam A, Weng X, Barry K, Bonnette J, Campitelli B, Daum C, Gordon SP. The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nature communications*. 2018 Dec 6;9(1):1-0.

Lovell JT, Williamson RJ, Wright SI, McKay JK, Sharbel TF. Mutation accumulation in an asexual relative of *Arabidopsis*. *PLoS genetics*. 2017 Jan 9;13(1):e1006550.

Luo MC, Gu YQ, Puiu D, Wang H, Twardziok SO, Deal KR, Huo N, Zhu T, Wang L, Wang Y, McGuire PE. Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature*. 2017 Nov;551(7681):498-502.

Lynch M, Bürger R, Butcher D, Gabriel W. The mutational meltdown in asexual populations. *Journal of Heredity*. 1993 Sep 1;84(5):339-44.

MacRobert JF, Setimela PS, Gethi J, Regasa MW. Manual de producción de semilla de maíz híbrido.

- Maheshwari P. An introduction to the embryology of angiosperms. McGraw-Hill,; 1950.
- Malagnac F, Bartee L, Bender J. An Arabidopsis SET domain protein required for maintenance but not establishment of DNA methylation. *The EMBO journal*. 2002 Dec 16;21(24):6842-52.
- Matsumura H, Nasir KH, Yoshida K, Ito A, Kahl G, Krüger DH, Terauchi R. SuperSAGE array: the direct use of 26-base-pair transcript tags in oligonucleotide arrays. *Nature Methods*. 2006 Jun;3(6):469-74.
- Meier M, Zappacosta D, Selva JP, Pessino S, Echenique V. Evaluation of different methods for assessing the reproductive mode of weeping lovegrass plants, *Eragrostis curvula* (Schrad.) Nees. *Australian Journal of Botany*. 2011 May 30;59(3):253-61.
- Meinke DW. Genome-wide identification of EMBRYO-DEFECTIVE (EMB) genes required for growth and development in Arabidopsis. *New Phytologist*. 2020 Apr;226(2):306-25.
- Mieulet D, Jolivet S, Rivard M, Cromer L, Vernet A, Mayonove P, Pereira L, Droc G, Courtois B, Guiderdoni E, Mercier R. Turning rice meiosis into mitosis. *Cell research*. 2016 Nov;26(11):1242-54.
- Nogler GA, GA N. Genetics of apospory in *Ranunculus auricomus*. IV. Embryology of F3 and F4 backcross offspring. 1975
- Nogler GA. Gametophytic apomixis. In *Embryology of angiosperms 1984* (pp. 475-518). Springer, Berlin, Heidelberg.
- Nyrén P. Enzymatic method for continuous monitoring of DNA polymerase activity. *Analytical biochemistry*. 1987 Dec 1;167(2):235-8.
- Ochogavía A, Galla G, Seijo JG, González AM, Bellucci M, Pupilli F, Barcaccia G, Albertini E, Pessino S. Structure, target-specificity and expression of PN\_LNC\_N13, a long non-coding RNA differentially expressed in apomictic and sexual *Paspalum notatum*. *Plant molecular biology*. 2018 Jan 1;96(1-2):53-67.
- Olmedo-Monfil V, Durán-Figueroa N, Arteaga-Vázquez M, Demesa-Arévalo E, Autran D, Grimanelli D, Slotkin RK, Martienssen RA, Vielle-Calzada JP. Control of female gamete formation by a small RNA pathway in Arabidopsis. *Nature*. 2010 Mar;464(7288):628.
- Ortiz JP, Leblanc O, Rohr C, Grisolia M, Siena LA, Podio M, Colono C, Azzaro C, Pessino SC. Small RNA-seq reveals novel regulatory components for apomixis in *Paspalum notatum*. *BMC genomics*. 2019 Dec;20(1):487.

Ortiz JP, Revale S, Siena LA, Podio M, Delgado L, Stein J, Leblanc O, Pessino SC. A reference floral transcriptome of sexual and apomictic *Paspalum notatum*. *BMC genomics*. 2017 Dec 1;18(1):318.

Otto SP. The evolutionary consequences of polyploidy. *Cell*. 2007 Nov 2;131(3):452-62.

Ou S, Jiang N. LTR\_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant physiology*. 2018 Feb 1;176(2):1410-22.

Ou Y, Lu X, Zi Q, Xun Q, Zhang J, Wu Y, Shi H, Wei Z, Zhao B, Zhang X, He K. RGF1 INSENSITIVE 1 to 5, a group of LRR receptor-like kinases, are essential for the perception of root meristem growth factor 1 in *Arabidopsis thaliana*. *Cell research*. 2016 Jun;26(6):686-98.

Ozias-Akins P, Lubbers EL, Hanna WW, McNay JW. Transmission of the apomictic mode of reproduction in *Pennisetum*: co-inheritance of the trait and molecular markers. *Theoretical and Applied Genetics*. 1993 Jan 1;85(5):632-8.

Ozias-Akins P, Roche D, Hanna WW. Tight clustering and hemizyosity of apomixis-linked molecular markers in *Pennisetum squamulatum* implies genetic control of apospory by a divergent locus that may have no allelic form in sexual genotypes. *Proceedings of the National Academy of Sciences*. 1998 Apr 28;95(9):5127-32.

Ozias-Akins P, van Dijk PJ. Mendelian genetics of apomixis in plants. *Annu. Rev. Genet.*. 2007 Dec 1;41:509-37.

Padmanaban S, Czerny DD, Levin KA, Leydon AR, Su RT, Mangel TK, Zou Y, Chanroj S, Cheung AY, Johnson MA, Sze H. Transporters involved in pH and K<sup>+</sup> homeostasis affect pollen wall formation, male fertility, and embryo development. *Journal of experimental botany*. 2017 Jun 1;68(12):3165-78.

Pardo J, Wai CM, Chay H, Madden CF, Hilhorst HW, Farrant JM, VanBuren R. Intertwined signatures of desiccation and drought tolerance in grasses. *Proceedings of the National Academy of Sciences*. 2020 Apr 23.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*. 2009 Jan;457(7229):551-6.

Paun O, Stuessy TF, Hörandl E. The role of hybridization, polyploidization and glaciation in the origin and evolution of the apomictic *Ranunculus cassubicus* complex. *New Phytologist*. 2006 Jul;171(1):223-36.

Pessino SC, Ortiz JP, Leblanc O, Do Valle CB, Evans C, Hayward MD. Identification of a maize linkage group related to apomixis in *Brachiaria*. *Theoretical and Applied Genetics*. 1997 Mar 1;94(3-4):439-44.

Peterson PM, Webster RD, Valdes-Reyna J. Genera of new world eragrostideae (Poaceae: Chloridoideae). *Smithsonian Contributions to Botany*. 1997.

Phillips SM. A numerical analysis of the Eragrostideae (Gramineae). *Kew Bulletin*. 1982 Jan 1:133-62.

Podio M, Cáceres ME, Samoluk SS, Seijo JG, Pessino SC, Ortiz JP, Pupilli F. Analysis of the methylation status of the apomixis-specific region in *Paspalum* spp. suggests an epigenetic control on parthenogenesis.

Poirier GC, Pyati J, Wan JS, Erlander MG. Screening differentially expressed cDNA clones obtained by differential display using amplified RNA. *Nucleic acids research*. 1997 Feb 15;25(4):913-4.

Poovaliah CR, Nageswara-Rao M, Soneji JR, Baxter HL, Stewart Jr CN. Altered lignin biosynthesis using biotechnology to improve lignocellulosic biofuel feedstocks. *Plant biotechnology journal*. 2014 Dec;12(9):1163-73.

Poverene M. Contribución citogenética y quimiosistemática a la taxonomía del pasto llorón, *Eragrostis curvula* (Schrad.) Nees s. lat (Doctoral dissertation, Ph. D. thesis, Universidad Nacional del Sur, Bahía Blanca). 1988.

Prasad V, Strömberg CA, Alimohammadian H, Sahni A. Dinosaur coprolites and the early evolution of grasses and grazers. *Science*. 2005 Nov 18;310(5751):1177-80.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, Haussler D. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome research*. 2016 Mar 1;26(3):342-50.

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouédraogo N. Real-time, portable genome sequencing for Ebola surveillance. *Nature*. 2016 Feb;530(7589):228-32.

Ramsey J, Schemske DW. Neopolyploidy in flowering plants. *Annual review of ecology and systematics*. 2002 Nov;33(1):589-639.

Ramsey J, Schemske DW. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual review of ecology and systematics*. 1998 Nov;29(1):467-501..

Ramulu KS, Sharma VK, Naumova TN, Dijkhuis P, van Lookeren Campagne MM. Apomixis for crop improvement. *Protoplasma*. 1999 Mar 1;208(1-4):196-205.

Ren R, Ray R, Li P, Xu J, Zhang M, Liu G, Yao X, Kilian A, Yang X. Construction of a high-density DArTseq SNP-based genetic map and identification of genomic regions with segregation distortion in a genetic population derived from a cross between feral and cultivated-type watermelon. *Molecular Genetics and Genomics*. 2015 Aug;290(4):1457-70.

Robinson M, McCarthy D, Chen Y, Smyth GK. edgeR: differential expression analysis of digital gene expression data User's Guide. *Most*. 2011 Nov 23;23(March):1-77.

Rodrigo JM, Zappacosta DC, Selva JP, Garbus I, Albertini E, Echenique V. Apomixis frequency under stress conditions in weeping lovegrass (*Eragrostis curvula*). *PLoS One*. 2017;12(4).

Rösti J, Barton CJ, Albrecht S, Dupree P, Pauly M, Findlay K, Roberts K, Seifert GJ. UDP-glucose 4-epimerase isoforms UGE2 and UGE4 cooperate in providing UDP-galactose for cell wall biosynthesis and growth of *Arabidopsis thaliana*. *The Plant Cell*. 2007 May 1;19(5):1565-79.

Sahu PP, Gupta S, Malaviya DR, Roy AK, Kaushal P, Prasad M. Transcriptome analysis of differentially expressed genes during embryo sac development in apomeiotic non-parthenogenetic interspecific hybrid of *Pennisetum glaucum*. *Molecular biotechnology*. 2012 Jul 1;51(3):262-71.

Sailer C, Schmid B, Grossniklaus U. Apomixis allows the transgenerational fixation of phenotypes in hybrid plants. *Current Biology*. 2016 Feb 8;26(3):331-7.

Saleme MD, Cesarino I, Vargas L, Kim H, Vanholme R, Goeminne G, Van Acker R, de Assis Fonseca FC, Pallidis A, Voorend W, Junior JN. Silencing CAFFEYOYL SHIKIMATE ESTERASE affects lignification and improves saccharification in poplar. *Plant physiology*. 2017 Nov 1;175(3):1040-57.

Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*. 1975 May 25;94(3):441-8.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*. 1977 Dec 1;74(12):5463-7.

Sanger F, Thompson EO. The amino-acid sequence in the glycol chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. *Biochemical Journal*. 1953 Feb;53(3):366.

Šarhanová P, Sharbel TF, Sochor M, Vašut RJ, Dančák M, Trávníček B. Hybridization drives evolution of apomicts in *Rubus* subgenus *Rubus*: evidence from microsatellite markers. *Annals of botany*. 2017 Aug 1;120(2):317-28

Savidan Y, Carman JG, Dresselhaus T. The flowering of apomixis: from mechanisms to genetic engineering. *Cimmyt*; 2001.

Savidan Y. Chromosomal and embryological analyses in sexual x apomictic hybrids of *Panicum maximum* Jacq. *Theoretical and Applied Genetics*. 1980 Mar 1;58(2):153-6.

Sawicki MP, Samara G, Hurwitz M, Passaro Jr E. Human genome project. *The American journal of surgery*. 1993 Feb 1;165(2):258-64.

Schalamun M, Schwessinger B. High molecular weight gDNA extraction after Mayjonade et al. optimised for eucalyptus for nanopore sequencing. *protocols*. io. 2017.

Schatz MC, Maron LG, Stein JC, Wences AH, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban E, Wright MH. Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. *Genome biology*. 2014 Nov 1;15(11):506.

Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK. Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science*. 1993 Oct 1;262(5130):110-4.

Selva JP, Siena L, Rodrigo JM, Garbus I, Zappacosta D, Romero JR, Ortiz JP, Pessino SC, Leblanc O, Echenique V. Temporal and spatial expression of genes involved in DNA methylation during reproductive development of sexual and apomictic *Eragrostis curvula*. *Scientific reports*. 2017 Nov 8;7(1):1-1.

Selva JP, Zappacosta D, Carballo J, Rodrigo JM, Bellido A, Gallo CA, Gallardo J, Echenique V. Genes Modulating the Increase in Sexuality in the Facultative Diplosporous Grass *Eragrostis curvula* under Water Stress Conditions. *Genes*. 2020 Sep;11(9):969.

Seppy M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. InGene Prediction 2019 (pp. 227-245). Humana, New York, NY.

Shen G, Kuppu S, Venkataramani S, Wang J, Yan J, Qiu X, Zhang H. ANKYRIN REPEAT-CONTAINING PROTEIN 2A is an essential molecular chaperone for peroxisomal membrane-bound ASCORBATE PEROXIDASE3 in Arabidopsis. *The Plant Cell*. 2010 Mar 1;22(3):811-31.

Sherwood RT, Berg CC, Young BA. Inheritance of apospory in buffelgrass. *Crop Science*. 1994 Nov;34(6):1490-4.

Singh M, Burson BL, Finlayson SA. Isolation of candidate genes for apomictic development in buffelgrass (*Pennisetum ciliare*). *Plant molecular biology*. 2007 Aug 1;64(6):673-82.

Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*. 2005 Dec 1;6(1):31.

Smit, A., Hubley, R. & Green, P. Open-1.0, <http://www.repeatmasker.org> (2008).

Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0, <http://www.repeatmasker.org> (2013).

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. Fluorescence detection in automated DNA sequence analysis. *Nature*. 1986 Jun;321(6071):674-9.

Soderlund C, Bomhoff M, Nelson WM. SyMAP v3. 4: a turnkey synteny system with application to plant genomes. *Nucleic acids research*. 2011 May 1;39(10):e68-.

Spillane C, Curtis MD, Grossniklaus U. Apomixis technology development—virgin births in farmers' fields?. *Nature biotechnology*. 2004 Jun;22(6):687-91.

Springer PS, Holding DR, Groover A, Yordan C, Martienssen RA. The essential MCM7 protein PROLIFERA is localized to the nucleus of dividing cells during the G (1) phase and is required maternally for early Arabidopsis development. *Development*. 2000 May 1;127(9):1815-22.

Staden R. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic acids research*. 1980 Aug 25;8(16):3673-94.

Staden R. A strategy of DNA sequencing employing computer programs. *Nucleic acids research*. 1979 Jun 11;6(7):2601-10.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research*. 2006 Jul 1;34(suppl\_2):W435-9.

Stein, N. "Synteny (Syntenic Genes)." (2013): 623-626.

Sze H, Padmanaban S, Cellier F, Honys D, Cheng NH, Bock KW, Conéjéro G, Li X, Twell D, Ward JM, Hirschi KD. Expression patterns of a novel AtCHX gene family highlight potential roles in osmotic adjustment and K<sup>+</sup> homeostasis in pollen development. *Plant Physiology*. 2004 Sep 1;136(1):2532-47.

Tadele Z. Orphan crops: their importance and the urgency of improvement. *Planta*. 2019 Jun 12:1-8.

Tanaka H, Hirakawa H, Kosugi S, Nakayama S, Ono A, Watanabe A, Hashiguchi M, Gondo T, Ishigaki G, Muguerza M, Shimizu K. Sequencing and comparative analyses of the genomes of zoysiagrasses. *DNA Research*. 2016 Apr 1;23(2):171-80.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and collinearity in plant genomes. *Science*. 2008 Apr 25;320(5875):486-8.

Teh BT, Lim K, Yong CH, Ng CC, Rao SR, Rajasegaran V, Lim WK, Ong CK, Chan K, Cheng VK, Soh PS. The draft genome of tropical fruit durian (*Durio zibethinus*). *Nature genetics*. 2017 Nov;49(11):1633-41.

Tejos R, Rodriguez-Furlán C, Adamowski M, Sauer M, Norambuena L, Friml J. PATELLINS are regulators of auxin-mediated PIN1 relocation and plant development in *Arabidopsis thaliana*. *Journal of cell science*. 2018 Jan 15;131(2).

Tester M, Langridge P. Breeding technologies to increase crop production in a changing world. *Science*. 2010 Feb 12;327(5967):818-22.

Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*. 2012 Jan;13(1):36-46.

Tu Y, Rochfort S, Liu Z, Ran Y, Griffith M, Badenhorst P, Louie GV, Bowman ME, Smith KF, Noel JP, Mouradov A. Functional analyses of caffeic acid O-methyltransferase and cinnamoyl-CoA-reductase genes from perennial ryegrass (*Lolium perenne*). *The Plant Cell*. 2010 Oct 1;22(10):3357-73.

van Dijk PJ, Bakx-Schotman JT. Formation of unreduced megaspores (diplospory) in apomictic dandelions (*Taraxacum officinale*, sl) is controlled by a sex-specific dominant locus. *Genetics*. 2004 Jan 1;166(1):483-92.

VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, Spittle K, Hall R, Gu J, Lyons E, Freeling M. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature*. 2015 Nov;527(7579):508-11.

VanBuren R, Wai CM, Wang X, Pardo J, Yocca AE, Wang H, Chaluvadi SR, Han G, Bryant D, Edger PP, Messing J. Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. *Nature communications*. 2020 Feb 14;11(1):1-1.

Vanholme R, Cesarino I, Rataj K, Xiao Y, Sundin L, Goeminne G, Kim H, Cross J, Morreel K, Araujo P, Welsh L. Caffeoyl shikimate esterase (CSE) is an enzyme in the lignin biosynthetic pathway in *Arabidopsis*. *Science*. 2013 Sep 6;341(6150):1103-6.

Vargas L, Cesarino I, Vanholme R, Voorend W, Saleme MD, Morreel K, Boerjan W. Improving total saccharification yield of *Arabidopsis* plants by vessel-specific complementation of caffeoyl shikimate esterase (cse) mutants. *Biotechnology for biofuels*. 2016 Dec 1;9(1):139.

Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research*. 2017 May 1;27(5):737-46.

Vera RR, Irazoqui H, Menvielle EE. THE NUTRITIVE VALUE OF WEEPING LOVEGRASS DURING THE SPRING SEASON: 1. Digestibility and yield of digestible nutrients. *Grass and Forage Science*. 1973 Sep;28(3):149-52.

Vielle JP, Burson BL, Bashaw EC, Hussey MA. Early fertilization events in the sexual and aposporous egg apparatus of *Pennisetum ciliare* (L.) Link. *The Plant Journal*. 1995 Aug;8(2):309-16.

Vogel, JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Lail K. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*. 463, 763–768 (2010).

Voigt PW, Bashaw EC. Apomixis and Sexuality in *Eragrostis curvula* 1. *Crop Science*. 1972 Nov;12(6):843-7.

Voigt PW, Bashaw EC. Facultative Apomixis in *Eragrostis curvula* 1. *Crop Science*. 1976 Nov;16(6):803-6.

Voigt PW. Discovery of Sexuality in *Eragrostis curvula* (Schrad.) Nees 1. *Crop Science*. 1971 May;11(3):424-5.

Voigt PW. Registration of OTA-S Weeping Lovegrass Germplasm1 (Reg. No. GP 8). *Crop Science*. 1976;16(6):886-.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS one*. 2014 Nov 19;9(11):e112963.

Wang K. Fixation of hybrid vigor in rice: synthetic apomixis generated by genome editing. *aBIOTECH*. 2020 Jan;1(1):15-20.

Wang X, Wang J, Jin D, Guo H, Lee TH, Liu T, Paterson AH. Genome alignment spanning major Poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. *Molecular plant*. 2015 Jun 1;8(6):885-98.

Wang X, Wang L. GMATA: an integrated software package for genome-scale SSR mining, marker development and viewing. *Frontiers in plant science*. 2016 Sep 13;7:1350.

Wang Y, Hou Y, Gu H, Kang D, Chen ZL, Liu J, Qu LJ. The Arabidopsis Anaphase-Promoting Complex/Cyclosome Subunit 1 is Critical for Both Female Gametogenesis and Embryogenesis. *Journal of integrative plant biology*. 2013 Jan;55(1):64-74.

Watson JD, Crick FH. Molecular structure of nucleic acids. *Nature*. 1953 Apr 25;171(4356):737-8.

Weber HE. Former and modern taxonomic treatment of the apomictic rubus complex. *Folia Geobotanica*. 1996 Dec 1;31(3):373-80.

Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang XJ, Buck D, Au KF. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*. 2017;6.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. *Genome research*. 2017 May 1;27(5):757-67.

Wetterstrand KA. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). (2019)

Winkler H. *Über Parthenogenesis und apogamie in Pflanzereich*. *Fortschritte der botanic*. 1908. pp 293–454

Wu R, Kaiser AD. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *Journal of molecular biology*. 1968 Jan 1;35(3):523-37.

Wu W, Yang YL, He WM, Rouard M, Li WM, Xu M, Roux N, Ge XJ. Whole genome sequencing of a banana wild relative *Musa itinerans* provides insights into lineage-specific diversification of the *Musa* genus. *Scientific reports*. 2016 Aug 17;6:31586.

Xie Y, Zhou Q, Zhao Y, Li Q, Liu Y, Ma M, Wang B, Shen R, Zheng Z, Wang H. FHY3 and FAR1 Integrate Light Signals with the miR156-SPL Module-Mediated Aging Pathway to Regulate Arabidopsis Flowering. *Molecular Plant*. 2020 Mar 2;13(3):483-98.

Xiong W, He L, Lai J, Dooner HK, Du C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proceedings of the National Academy of Sciences*. 2014 Jul 15;111(28):10263-8.

Xu J, Zhang HY, Xie CH, Xue HW, Dijkhuis P, Liu CM. EMBRYONIC FACTOR 1 encodes an AMP deaminase and is essential for the zygote to embryo transition in Arabidopsis. *The Plant Journal*. 2005 Jun;42(5):743-58.

Yang T, Bar-Peled L, Gebhart L, Lee SG, Bar-Peled M. Identification of galacturonic acid-1-phosphate kinase, a new member of the GHMP kinase superfamily in plants, and comparison with galactose-1-phosphate kinase. *Journal of Biological Chemistry*. 2009 Aug 7;284(32):21526-35.

Ye C, Hill CM, Wu S, Ruan J, Ma ZS. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Scientific reports*. 2016 Aug 30;6:31900.

Zappacosta D, Meier M, Carrera A, Pacheco G, Cardone S, Selva JP, Echenique V. Marcadores moleculares para el estudio de la variabilidad dentro del complejo *Eragrostis curvula*. *Phyton (Buenos Aires)*. 2011 Dec;80(2):211-20.

Zappacosta DC, Gallardo JA, Carballo J, Meier MS, Rodrigo JM, Gallo CA, Selva JP, Stein J, Ortiz JP, Albertini E, Echenique VC. A high-density linkage map of the forage grass *Eragrostis curvula* and localization of the diplospory locus. *Frontiers in plant science*. 2019;10:918.

Zhang G, Liu X, Quan Z, Cheng S, Xu X, Pan S, Xie M, Zeng P, Yue Z, Wang W, Tao Y. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nature biotechnology*. 2012 Jun;30(6):549-54.

Zhang L, Guo C. The important function of Mediator complex in controlling the developmental transitions in plants. *International journal of molecular sciences*. 2020 Jan;21(8):2733.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. *Bioinformatics*. 2013 Nov 1;29(21):2669-77.

Zou H, Jakovlić I, Chen R, Zhang D, Zhang J, Li WX, Wang GT. The complete mitochondrial genome of parasitic nematode *Camallanus cotti*: extreme discontinuity in the rate of mitogenomic architecture evolution within the Chromadorea class. *BMC genomics*. 2017 Dec;18(1):1-7.

## ANEXO

Nombre	Primer Forward	Primer Reverse	Motivo SSR	Contig	Amplicon en Victoria (bp)		% identidad		% Cobertura	
					Predicho	Obtenido	Primer Forward	Primer Reverse	Primer Forward	Primer Reverse
Victoria-N1	GTCCCTGCAGTCAGAGTAGC	ACGAAGGGGAGGATGAAAGC	(CTG)5	Contig10	145	140	100	100	100	100
Victoria-N2	TTCCTTTGTTGTGGGGCCAT	TGAGTATCCACACGCAGCTG	(TGT)5	Contig10	235	240	100	100	100	100
Victoria-N3	TGCTATTGGACGGTGCAGTT	GAGCAATTCGGAGGGGAGAG	(TGT)5	Contig30	276	270	100	100	100	100
Victoria-N4	ATCCCTACGTCAACGATGGC	GGCTTTTGGCACTCGTCAAT	(TCT)12	Contig30	242	260	100	100	100	100
Victoria-N5	CCCATCTCCAAGACCAAGCA	TGCCGTAGTAGGGAGAAGGA	(CCTTT)5	Contig30	263	260	100	100	100	100
Victoria-N6	CCCACTGAAGACAAGACAGC	ACCCTGGGTGTGAATTAGCG	(GT)9(AT)31	Contig936	142	No específico	100	100	100	100
Victoria-N7	AGCATGGTCGAACGGTGAAT	AACTGAGGATGCCTAGCTGC	(CTGG)5	Contig38	227	230	100	100	100	100
Victoria-N8	GTCACAGCGCCAAAGAAAGG	TGGGATCAGGATGGGTAGGG	(AGAC)5	Contig38	159	150	100	100	100	100
Victoria-N9	CGTGAAGAGAAGGCGCAAC	CAGCAGCGGTGAGAAGAAGA	(GCCT)5	Contig28	251	250	100	100	100	100
Victoria-N10	CTTCACCAAGCTCCGGAAGT	ACCCCTAAGAGATCCAGCA	(TTTCT)5	Contig28	270	270	100	100	100	100
Victoria-N11	GCATGTTGACCTCCACTCCA	ATCTCAGTGACCCATGTGCC	(TTTTC)6	Contig8	260	260	100	100	100	100
Victoria-N12	CTCGAGCTAGAGCCAGATGC	TGGAATCGGCATCGGTGTAG	(GAGCCG)5	Contig38	270	270	100	100	100	100
Victoria-N13	GCCAACCCAGCACTCTTAT	GGAGAGGTGAAAAGGTGGG	(ATCTCT)5	Contig3	244	240	100	100	100	100
Victoria-N14	GGTGGTGGGGGTGAGTAAAG	AGAGTCAAAAAGCGTTCCCGT	(GGCAC)5	Contig190	136	No amplifico	100	100	100	100
Victoria-N15	AGACACCATCTCCAACAGCG	GTTCCCTAAGTCACGGCAA	(GCTCG)5	Contig4	218	230	100	100	100	100
isotig01946*	TCCATATCCATCTACGAACC	GTTGTTGTGAATTTTTGCT	(TCT)5		No encontrado	No amplifico	-	-	-	-
isotig06745*	GTGAAGGAGGAGAAGTCGAG	GGAGAGCGAGTCGTCC	(CCG)5	Contig30	291	300	100	100	100	100
isotig07764*	CCTGCTTCTTGCTCCC	AAAGAAGACGTCGTAAGCAG	(CCG)6	Contig1	186	170	100	100	100	85
isotig09593*	TGACATGTGACAGACTCCAT	CGTATATTGGCATCTCGTCT	(T)108		No encontrado	No amplifico	-	-	-	-
isotig10159*	TAGCCAGATGACCTCCAC	CCTGCTCTCCTCCGAC	(GGA)8	Contig50	244	310	100	100	100	100

isotig13242*	GCTCGACTCCGTCTCC	GAAACCGTCCAAGAAGAAG	(CTT)5	Contig25	187	No amplifico	100	100	100	100
isotig13665*	TGCGCTGGACCTCTACTA	ACTGCTTCAGCCTCATCTT	(CGG)6	Contig25	352	300	100	100	94.44	100
isotig13699*	G TTCAGCATGAGTAACCACA	CTCCTTCCTCTCCTCATTCT	(AGAAGG)6	Contig12	321	300	100	100	100	95
isotig13776*	TCCA ACTCATCAACCAGTAA	GTAGCTCTTGCCGAACC	(CCT)5	Contig33, Contig175	227	230	100	100	100	100
isotig15838*	AGTGGAGGAAGATGTAGCC	ACAGAGTTGAAGGAGCAGAG	(C)10	Contig50, Contig12, Contig6	241	250	100	100	100	100
isotig18079*	GACAGGACCCCTCTTCC	AGGACTCCAGCTCAGAT	(GGC)5	Contig705	153	150	100	100	100	100
isotig18488*	ACAGAAACCAACAGTGGAAC	GATGGTTCCGTAGATCGTAG	(CA)6	Contig685	233	260	100	100	100	100
isotig20053*	CCACCAACCAATTATCCTAA	GATCTCGCGACAACCC	(CAC)7	Contig38	236	250	100	100	100	100
isotig20164*	GA ACTACCTGCAGATCAAGG	TGCAC TTTGTTGTCATCTTC	(GAT)7	Contig38	184	200	100	100	100	100
isotig23517*	CTGCATCCAGTGGTTTTTC	ATG TACTCAAGGCACGACA	(GCC)6	Contig50	171	160	100	100	100	100
isotig23773*	ACCGCATTCTTTGATTTAGA	TGTGATTTCTTCGGTTCTTT	(GAG)5	Contig25	211	200	100	100	100	100
isotig24162*	TCATCTTCCTCTTGATCTGC	ATGGAAGCAGAGCACTACC	(TCG)5		No encontrado	No amplifico	-	-	-	-
isotig25156*	ACCACAGTACGGACTCTC	CGTACTCCTTGTAACGCTG	(GAG)6	Contig28	771	800	100	100	100	100
isotig25372*	GATGACAACCAATAAGCC	ATTTCTTCCACGACGAAGTT	(CCG)5	Contig50	163	160	100	100	100	100
isotig25706*	GGTGAGGAAGTCGAGGAT	ACTTCAAGGTGGCGTACA	(GTT)8	Contig28	245	Inespecifico	100	100	100	100
isotig27380*	G TTCGGCGTCTCCATC	GTATATGGGCCAGTGGTTC	(TCGCCG)5	Contig38	193	200	100	100	100	100
isotig27968*	ACTTGCAGAAATCACAAAGG	GGATCGTGTGATTGAAGTT	(A)12	Contig38	158	160	100	100	100	100
isotig28255*	GTCGTTGAACATGATCCCT	GAGCAGGAGAACATCAAGAG	(CGC)5	Contig30	340	300	100	100	100	100
isotig28834*	GCGTAGCTCTTCTCCATGA	CACTTCTTCGGGGTGTTTC	(CGG)5	Contig188, Contig3	306	300	100	100	100	100
isotig29328*	AGCATCATCACC AATTCT	GTAACGTAGGGTGGGAATC	(GCA)5	Contig30	370	No amplifico	100	100	73.68	100
isotig30246*	TCTTCTTCAGCCCCTTG	ACGTCGTCGTCATCCTC	(GTC)5	Contig3, Contig10	113-249	160	100	100	100	100
isotig32396*	CGCCATGAAGCATCTC	CAACTTTGTTTTCTCGTC	(GCC)5		No encontrado	No amplifico	-	-	-	-
isotig33661*	ATGGCATCTTCTCCTGTGAT	GAACCCCTTCGCAGC	(CGC)5	Contig12	152	150	100	100	100	100
isotig33732*	ATCCTGTTTTTGTGCTTGTT	GGGAAAAAGCTCCAGTAATC	(CCT)7		No encontrado	No amplifico	-	-	-	-
isotig33734*	GTCCATCCGATCCCAG	AACTCATCTTTCACGACTT	(CGC)5	Contig3	666	250	100	100	100	100

isotig34134*	TACTCATCGTCGTACCTCCT	GGCTGAGTACTTGGTCTCTC	(GCG)5	Contig10	168	170	100	100	100	100
isotig35118*	CATCATCATCCCCATTTATC	TGATCACAAGATGACCAAGA	(GAA)5	Contig1	284	300	100	100	100	95
isotig35144*	ACAAGTCATCGGACTCCTC	AGGGAGAGAGACGAGAAACT	(TCC)5		No encontrado	300	-	-	-	-
isotig38544*	AAAAAGCGGAAGCAGAAC	CGCTTGAATCAGCATAGG	(GGC)7	Contig155, Contig902	244	No amplifico	100	100	100	100
isotig42614*	GATTCAACCGTACTTCAGGA	AAAAACCTCATCTCCCCC	(AG)7		No encontrado	No amplifico	-	-	-	-

Tabla 1: Primers de SSR evaluados en los diferentes genotipos de *E. curvula*. Los primers indicados con un asterisco "\*" fueron diseñados en base a trabajos transcriptómicos previos (Garbus et al., 2019) y los primers que comienzan con "Victoria" fueron diseñados sobre el genoma. La predicción de los amplicones esperado fue determinada *in silico* y la obtenida mediante corrida electroforética en geles utilizando marcadores de peso molecular.

Especie	Query ID	Target ID	% Ident	% Cob	Longitud	Mm	Gap	Query start	Query end	Target start	Target end	E-val
<b><i>E. curvula</i></b>												
	E. curvula Class II	Contig6	100	100	1014	0	0	1	1,014	10,797,008	10,798,021	0
	E. curvula Class I	Contig10	100	100	1023	0	0	1	1,023	52,957,363	52,956,341	0
<b><i>E. curvula</i> transcriptome</b>												
	E. curvula Class II	isotig39171	98.915	100	1,014	11	0	1	1,014	1,079	66	0
	E. curvula Class I	isotig40238	96.752	84	862	20	6	165	1,023	1,023	167	0
<b><i>O. thomaeum</i></b>												
	E. curvula Class II	gi 7	88.889	100	1,017	100	4	1	1,014	4,574,595	4,575,601	0
	E. curvula Class I	gi 4	88.583	96	762	80	4	223	977	4,063,536	4,064,297	0
	E. curvula Class I	gi 4	87.946	96	224	27	0	1	224	4,058,612	4,058,835	1.98E-74
<b><i>E. tef</i></b>												
	E. curvula Class II	CL3332Contig4	94.655	100	1,029	37	5	1	1,014	210	1,235	0
<b><i>P. halli</i></b>												
	E. curvula Class II	ref NC_038042.1	88.287	100	1,033	98	6	1	1,023	5,724,226	5,723,207	0
	E. curvula Class I	ref NC_038043.1	91.735	96	980	66	3	37	1,013	52,404,597	52,403,630	0

Espece	Query ID	Target ID	% Ident	% Cob	Longitud	Mm	Gap	Query start	Query end	Target start	Target end	E-val
<b><i>O. sativa</i></b>												
	E. curvula Class II	2	87.087	95	968	125	0	19	986	6,039,452	6,,038,485	0
	E. curvula Class I	7	87.29	93	952	98	7	52	999	22,694,322	22,693,390	0
<b>Especies solo con CSE II</b>												
<b><i>Z. mays</i></b>												
	E. curvula Class II	2	89.98	96	978	80	11	26	998	212,556,872	212,555,908	0
<b><i>S. bicolor</i></b>												
	E. curvula Class II	gb CM000761.3	90.419	98	1,002	75	11	21	1,013	70,832,179	70,831,190	0
<b><i>S. italica</i></b>												
	E. curvula Class II	II	91.179	96	975	71	7	26	997	43,454,946	43,453,984	0
<b><i>B. distachyon</i></b>												
	E. curvula Class II	1	87.187	98	999	106	9	19	1,007	19,776,425	19,777,411	0
<b><i>T. aestivum</i></b>												
	E. curvula Class II	2D	86.735	95	980	109	5	47	1,014	163,020,312	163,019,342	0
	E. curvula Class II	2A	85.944	97	996	113	6	34	1,014	176670726	176,671,709	0
	E. curvula Class II	2B	86.27	95	976	110	6	47	1,007	224,649,275	224,650,241	0

Tabla 2: Especies que contienen ambas CSE y especies que contienen solo CSE II

Primer	Primer forward	Primer reverse	Contig	Forward start	Reverse end	Amplicon	Coding
<b><i>E.curvula-CSE-Class-I</i></b>	GAACATCGACTGCACCGAAC	AGGGGCACTGTGCTGATTTA	Contig10	52,957,431	52,956,315	1,116	1,014
<b><i>E.curvula-CSE-Class-II</i></b>	AACCCAAGTCTGATCCCCCG	GATAGACTTGTTGGTCAGGTTTAC	Contig6	10,758,145	10,759,319	1,174	1,023

<b><i>E.tef</i>-CSE-Class-I</b>	TCTTCACCCACTCCTTCCA	GACTTGCCCACCATCTTCTT	Contig10	52,957,233	52,956,732	501	
---------------------------------	---------------------	----------------------	----------	------------	------------	-----	--

Tabla 3: primers utilizados para la validación de CSE diseñados a partir del genoma de *E. curvula*.

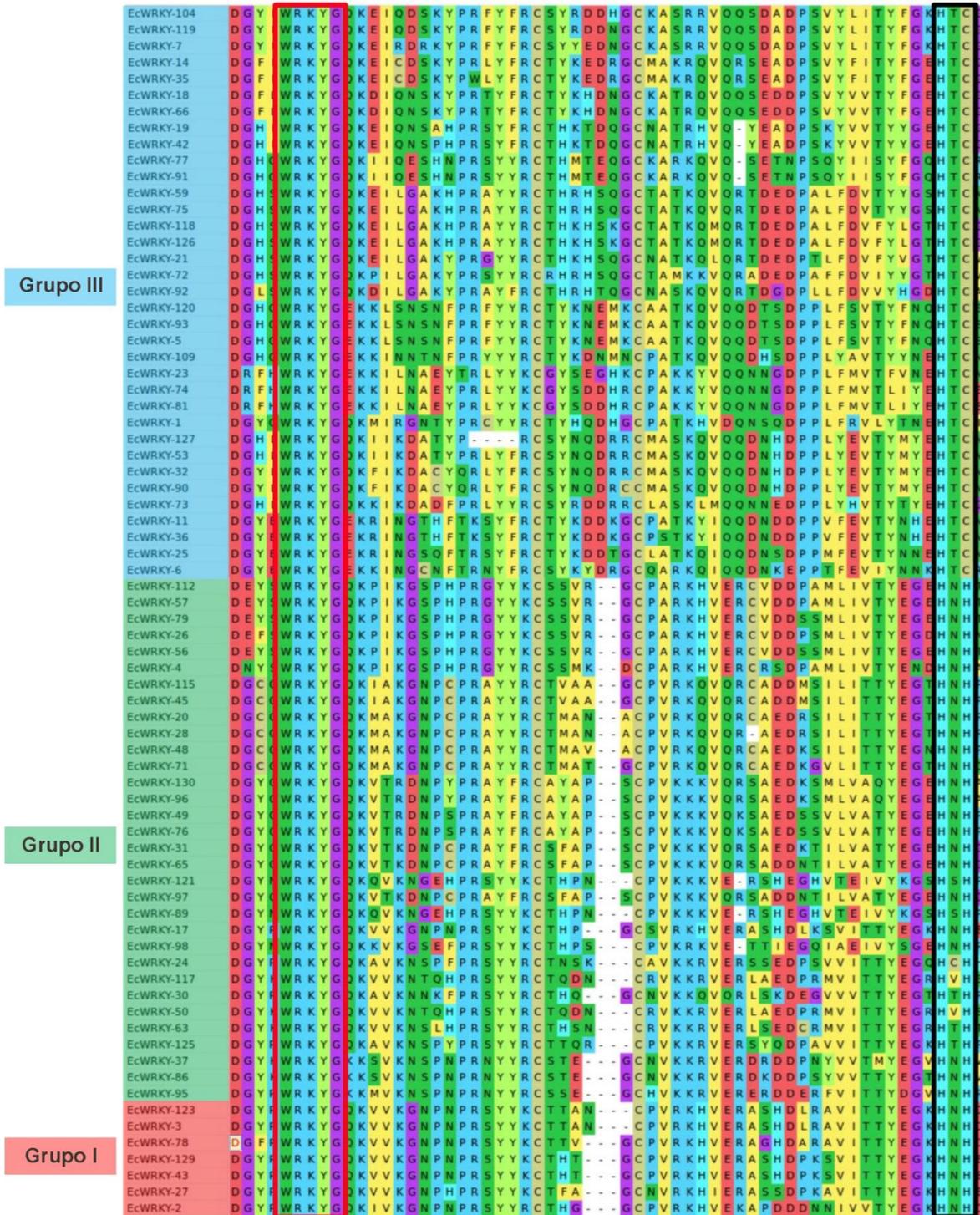


Figura 1: Alineamiento múltiple de los motivos endontrados en los factores de transcripción WRKY.

	Victoria	Don Walter	Tanganyika INTA	OTA	Tanganyika USDA	E. tef A	E. tef B	E. nindensis	O. thomaeum	Z. japonica	Z. mays
Genes	56.469	103.961	142.275	17.808	16.503	34.032	32.255	22.340	28.446	59.271	131.496
Genes en ortogrupos	52.881	82.129	114.105	16.957	15.583	31.632	30.199	19.563	23.489	35.85	106.508
Genes no asignados	3.588	21.832	28.170	851	920	2.400	2.056	2.777	4.957	23.421	24.988
Porcentaje de genes en ortogrupos	93,6	79,00	80,20	95,2	94,4	92,9	93,6	87,6	82,6	60,5	81
Porcentaje de genes no asignados	6,4	21,00	19,80	4,8	5,6	7,1	6,4	12,4	17,4	39,5	19
Ortogrupos contenidos en la especie	23.439	32.551	33.430	8.430	7.874	18.683	18.454	10.720	15.826	16.99	17.001
Porcentaje de ortogrupos contenido en la especie	60,9	84,6	86,9	21,9	20,5	48,6	48	27,9	41,1	44,2	44,2
Especie especifico ortogrupos	15	15	40	4	8	1	3	15	26	41	124
Genes en ortogrupos especie especifico	36	34	421	8	21	2	8	46	72	1.077	1.383
Porcentaje de genes especie especificos	0,06	0,03	0,30	0,04	0,13	0,01	0,02	0,21	0,25	1,82	1,05

Tabla 4: resultados obtenidos con el software OrthoFinder sobre la anotación de las especies chloridoideae secuenciadas hasta el momento y los transcritos traducidos a proteínas de Tanganyika USDA y OTA. El análisis se hizo sobre el total de los genes incluidos en la primer fila los cuales se agruparon en ortogrupos.