ESTUDIO EVOLUTIVO DEL ORIGEN DEL ARROZ MALEZA DE URUGUAY

Tesis de Maestría

PEDECIBA Biología, Subárea Genética.

2012

Silvia Garaycochea Solsona

Tutor

Fernando Álvarez-Valín

Co-Tutor

Ing. Agr. Fabián Capdevielle.

Agradecimientos.

Sección Biomatemática, Facultad de Ciencias, UdelaR.

INIA, Programa Nacional Arroz, Estación Experimental del Este y Unidad Biotecnología.

ANII, Sistema Nacional de Becas y Beca Movilidad.

Lifesequencing SL., Valencia, España.

Fernando Álvarez-Valín.

Fabián Capdevielle.

Pablo Speranza.

Guillermo Lamolle.

Miguel Ponce de León.

Marco Dalla Rizza.

Victoria Bonnecarrère.

Mis compañeros de la Unidad de Biotecnología y del piso 4 de la Facultad de Ciencias.

Quiero agradecer especialmente a mi familia y amigos por su apoyo constante.

Martín e Iñaki: les dedico este trabajo, ya que sin su apoyo no lo hubiera logrado.

GRACIAS a todos!

INDICE

Resumen

El “arroz rojo” es una de las malezas que más perjudican el cultivo del arroz (*Oryza sativa* L.). Se denomina bajo este nombre genérico a una serie de biotipos de arroz salvaje muy emparentados con el arroz cultivado, cuya principal característica distintiva es la de poseer pericarpio de color rojizo, negro o marrón. El esclarecimiento del origen del arroz tipo maleza es complejo y aún no completamente resuelto; los procesos principales que podrían estar involucrados son: pre–adaptación, evolución e introgresión.

En este trabajo se planteó el estudio el origen evolutivo de arroz maleza de Uruguay a través de aproximaciones de genómica evolutiva/comparativa y se buscó la generación de nuevas herramientas que puedan contribuir al abordaje del problema arroz maleza en Uruguay conociendo la diversidad existente y su dinámica poblacional.

A partir de la secuenciación de 1/4 de placa de 454/Roche se obtuvo 96 Mb de secuencias del biotipos de arroz maleza AM356-8, obteniendo una cobertura de 0.1X del genoma nuclear, 10X del mitocondrial y 106X del cloroplástico. Aplicando estrategias de genómica comparativa se obtuvo un conjunto de lecturas que contenían el genoma completo del cloroplasto del biotipo de arroz maleza. El establecer un método para la obtención de secuencias de los genomas de los organelos a partir de la secuenciación del ADN total de una planta, representa la ventaja de permitir aprovechar dicha información que habitualmente se descarta por considerarla como “contaminante” y a su vez evita las dificultades que conllevan la separación de los ADN de los 3 genomas contenidos en una célula vegetal (nuclear, mitocondrial y cloroplástico). Disponer de la secuencia completa de los genomas de los organelos, cloroplastos y mitocondrias, proveen de información valiosa para estudiar la ecología y evolución molecular en plantas.

La obtención del genoma del cloroplasto de AM356-8 permitió realizar estudios comparativos con regiones variables del tipo tipo SNP e INDELs identificadas entre cuatro de los genomas cloroplásticos públicos más cercanos al biotipos secuenciado (*O. rufipogon*; *O. nivara*; *O. sativa ssp japonica*; *O. sativa ssp indica*). Los InDels de mayor largo determinados en este trabajo fueron utilizados para la comparación con el conjunto de lecturas del cloroplasto AM356-8. A través de esta comparación de logró determinar el haplotipo del arroz maleza en estudio, el que presentó el mismo haplotipo que el genoma cloroplástico de *O. sativa ssp japonica*.

Por otro lado fue posible realizar comparaciones entre los tres genomas de la célula vegetal y el conjunto de lecturas del cloroplasto de AM356-8 e identificar regiones de ADN cloroplástico inserto en el núcleo. En este sentido se resalta que se pudieron identificar unas 16 regiones del genoma de la maleza que contienen segmentos cloroplásticos no presentes en el núcleo de *O. sativa ssp japonica*. Esto nos indicaría que habrían ocurrido alrededor de 160 eventos de transferencia desde la separación entre la maleza AM356-8 y arroz tipo *japónica* desde su separación (no mas de 200 años atrás). Dichos resultados conjugados, tiempos estimados de aparición de la maleza AM356-8 y número estimados de eventos, nos permitieron realizar estimaciones sobre la frecuencia de ocurrencia de dichas transferencias.

Introducción

El arroz maleza y su origen.

Las especies vegetales que han sido sometidas a domesticación conforman en general grandes complejos de especies los que están compuestos por cultivos, malezas conespecíficas y especies silvestres relacionadas (Anderson, 1967). Durante la domesticación de los cultivos, un proceso evolutivo complejo, el hombre dirigió cambios que hoy se ven reflejados a nivel morfológico y fisiológico, distinguiendo a las especies domesticadas de sus ancestros silvestres (Hancock, 2012). Este conjunto de cambios es conocido como síndrome de domesticación (Harlan, 1992) e incluye como principales características la pérdida en el desgrane, cambios en el hábito de crecimiento, perdida de la dormancia, cambio en la pigmentación de las semillas entre otros.

Las malezas conespecíficas son morfológica y ecológicamente divergentes de las domesticadas y sus especies congéneres silvestres (De Wet and Harlan, 1975; Gressel, 2005). El éxito evolutivo de las malezas conspecíficas es muchas veces atribuido a la adquisición de características asociadas con las plantas silvestres tales como: el mayor desarrollo vegetativo, muy alta capacidad de dispersión de la semilla y dormancia. Así como también la adquisición de características típicas de plantas domesticadas tales como mayor potencial de autofecundación y rápido crecimiento. El poseer esta combinación de características les favorece su capacidad invasiva a los agroecosistemas actuales (Reagon et al., 2010). La existencia de flujo de genes entre plantas domesticadas, malezas conspecíficas y especies silvestres es un proceso continuo que afecta directamente la diversidad genética de las poblaciones de los cultivos y puede ser fuente para la aparición de nuevas combinaciones genéticas. Se ha demostrado que el flujo de genes actúa en ambas direcciones, tanto desde domesticadas a silvestres cómo en la dirección opuesta (Jarvis and Hodgkin, 1999; Ellstrand, 2003).

La hibridación natural y la introgresión entre plantas domesticadas y especies silvestres relacionadas juega un importante rol en la evolución de las plantas generando mayor diversidad genética y diferenciación (Xia et al., 2011). Entender cuáles son los mecanismos evolutivos que dirigen la aparición de malezas provenientes de los mismos complejos de especies que las plantas domesticadas se hace necesario para abordar el problema que éstas significan.

El género *Oryza*, tiene dos especies domesticadas *Oryza sativa* (asiática) y *Oryza glaberrima* (africana) y más de 20 especies silvestres distribuidas a lo largo de las regiones tropicales y subtropicales. Las poblaciones de *Oryza sativa* presentan una profunda estructura constituida por dos subespecies mayoritarias, *Oryza sativa japónica* y *Oryza sativa spp indica*, las que son diferenciables tanto por característica morfológicas, fisiológicas y genéticas (Oka, 1988a; Garris et al., 2005). Estas dos subespecies están asociadas a diferentes hábitat de crecimiento, la subespecie *indica* es habitualmente encontrada en tierras bajas de Asia tropical, mientras que la subespecie japónica en tierras altas de China, sureste de Asia e Indonesia (Londo et al., 2006).

El ancestro silvestre de la especie asiáticas habría sido *O. rufipogon*, originario de Asia y capaz de crecer en todo el rango de hábitat donde se ha encontrado arroz. Sin embargo el centro de origen de la domesticación y cómo surgen las dos subespecies del cultivo es aún tema de discusión. Diversas hipótesis han sido planteadas para explicar el proceso de domesticación de *O. sativa*, entre éstas, existen dos hipótesis principales y cuyos postulados son contrastantes. Una de las hipótesis plantea que el cultivo de arroz fue domesticado una sola vez en China y que se diferenciaron en las dos subespecies luego, a través de la selección (Ting, 1957; Oka, 1988b; Lu et al., 2002; Gao and Innan, 2008). Los hallazgos arqueológicos apoyan esta hipótesis, ya que sólo habían sido encontrados en China, sin embargo en la actualidad se han encontrado restos fósiles del cultivo de arroz también en el este de Asia (Chen, 1999).

La hipótesis alternativa, es la de la doble domesticación, proponiendo que japónica e *indica* se originaron a partir de dos eventos de domesticación geográficamente independientes desde un ancestro común. *Japonica* es encontrada predominantemente en el este de Asia, mientras que *indica* al sur de Asia. Existen estudios de marcadores moleculares tales como las isozimas, AFLPs, SSRs y SNPs, que claramente muestran diferencias moleculares entre ambas subespecies (Second, 1982; Glaszmann, 1987; Prashanth et al., 2002; Garris et al., 2005; Londo et al., 2006; Caicedo et al., 2007). La existencia de la división en las dos subespecies de *O. sativa* fue estimada en 100.000 años, fecha que precede ampliamente a la domesticación, por lo tanto, y sumado a las evidencias moleculares estos trabajos apoyan la hipótesis que los eventos de domesticación para estas dos subespecies habrían sido independientes partiendo de conjuntos de genes pre-diferenciados en el ancestro silvestre (Sweeney and McCouch, 2007).

Morfológica y genéticamente, las poblaciones de arroz maleza han sido reportada como altamente variables. Éstos parecen ser un intermedio entre los biotipos silvestres y los biotipos cultivados (Kelly Vaughan et al., 2001; Yu et al., 2005; Cao et al., 2006; Londo and Schaal, 2007). La distribución simpátrica con el cultivo durante largos períodos de tiempo, y la fácil incorporación de genes desde los cultivos a los biotipos maleza e introgresión, pueden haber sido los causantes de la similaridad encontrada entre las poblaciones de arroz maleza y el cultivo con el que coexisten. Este proceso puede promover la persistencia y mejor adaptación de los biotipos maleza a ambientes influenciados por el hombre(Harlan, 1965), en donde existe una alta presión de selección artificial, lo que explicaría la mimetización de los biotipos maleza con el arroz cultivado en su coexistencia (Xia et al., 2011).

El arroz maleza pertenece al mismo género y especie que el arroz cultivado, siendo un claro ejemplo de maleza conespecífica. Desde un punto de vista fenológico, los diferentes biotipos de arroz maleza que se encuentran bajo condiciones de siembra directa en las mayores áreas de cultivo de arroz -Latinoamérica y Norte América, Caribe, África, y algunas regiones del Sur y Sudeste de Asia- constituyen poblaciones anuales de *O. sativa L* [Cao, et al.,2006]. La existencia de híbridos entre diferentes formas de *O. sativa* y sus poblaciones derivadas por una lado, así como *O. sativa* con especies silvestres relacionadas por otro, pueden presentar diferentes

combinaciones de caracteres funcionales asociados con el proceso de domesticación. En la etapa vegetativa es de difícil identificación, ya que solo se dispone como característica distintiva, la altura de la planta, color y pubescencia, lo cual a su vez no es exclusivo del arroz maleza.

Además, se debe tener en cuenta las evidencias de posibles flujos génicos con los cultivares, que dificultan la diferenciación de algunos biotipos de arroz maleza, ya que presentan alto grado de fertilidad en los cruzamientos (Majumder et al., 1997; Gealy et al., 2002; Chen et al., 2004; Kuroda et al., 2005; Song et al., 2006). Es por lo tanto frecuente en zonas con alta contaminación de arroz rojo y muchos años de cultivo, encontrar diversidad en los biotipos, algunos de los cuales son muy parecidos a las variedades utilizadas en esas zonas, debido al proceso de "mimetización" (Valverde, 2005). Esta diversidad de formas encontradas en campos de cultivo de arroz de todo del mundo, han sido estudiadas por diferentes metodologías para entender mejor la complejidad genética y el origen del arroz maleza (Olsen et al., 2007).

Estudios sobre la diversidad genética dentro de poblaciones de arroz maleza han reportado que éstos biotipos muestran una fuerte estructura poblacional semejante a la identificada en el arroz cultivado. Algunos de estos estudios analizaron poblaciones de arroz maleza provenientes de muy diversos lugares geográficos donde el arroz es cultivado. En el trabajo de Tang y Morishima (1996) se analizó con marcadores morfológicos e izoenzimas poblaciones provenientes de Japón, Brasil, Estados Unidos y China (zona alta y zona baja del valle de Yangtze), Corea y Asia tropical, donde se identificaron tres grupos donde se separaron claramente los biotipos provenientes de Japón, Brasil, Estados Unidos y los de la zona alta del valle de Yangtze por un lado, los que compartían características con la subespecie *indica*. Por otro lado se agruparon los biotipos provenientes de Corea y la zona baja del valle de Yangtze semejantes a la subespecie *japonica* de auto-propagación y en el tercer grupo se encontraron los biotipos procedentes de Asia tropical, con semejanzas a la subespecie *indica* pero con características de los biotipos silvestres tales como auto-propagación con altos niveles de dormancia y desgrane, pudiendo ser el resultado del flujo genético entre el arroz silvestre naturalmente abundante de ésta región y el arroz cultivado (Tang and Morishima, 1996).

En estudios más recientes, donde se estudiaron poblaciones de arroz maleza provenientes de otras regiones del cultivo de arroz, utilizando marcadores moleculares más sensibles, se obtuvieron resultados semejantes a lo reportado por Tang y Morishima (1996). En el trabajo de Ferrero et al. (2001), donde se comparó biotipos encontrados en la región mediterránea con biotipos provenientes de Brasil (Ferrero, 2001) y en el trabajo de Federici et al (2001), donde se analizó una población de arroz maleza uruguaya (Federici et al., 2001), se identificaron grupos de biotipos que compartían características morfológicas y moleculares con una u otra subespecie de *O. sativa*.

En otro estudio realizado sobre las poblaciones de arroz maleza de EEUU, se identificaron dos grupos principales de biotipos que presentaron perfiles genéticos y morfológicos bien diferenciales coincidentes con lo reportado por Federici et al. (2001). El primer grupo compartía el perfil genético nuclear y el citotipo con la subespecie *indica*, sugiriendo que estos podrían haber surgido a partir de hibridaciones con variedades de tipo *indica*. El segundo grupo compartía un alto porcentaje de los loci nucleares y el mismo citotipo que la subespecie *japonica*. Estos trabajos indican que el arroz maleza habría evolucionado a partir de biotipos domesticados y que en la mayoría de los casos estudiados éste actuaría como padre. El flujo génico desde las malezas al cultivo podría ser una vía alternativa para la evolución de éstos biotipos (Reagon et al., 2010).

Por otro lado, algunos autores consideran que dadas las semejanza morfológicas y genéticas de algunos biotipos de arroz maleza con los biotipos silvestres, la introgresión de germoplasma de arroz asiático no domesticado en ambientes modificados por el hombre para el cultivo, podría haber dado origen al arroz maleza (De Wet and Harlan, 1975; Olsen et al., 2007).

En resumen, se han planteado diferentes hipótesis para explicar el origen del arroz maleza y su situación actual. Los principales procesos involucrados en el origen de una maleza son: 1) evolución de biotipos silvestres junto con los domesticados para su adaptación a nuevos ambientes modificados por el hombre; 2) Hibridación entre plantas cultivadas y especies silvestres cercanas reproductivamente compatibles y 3) cultivares abandonados que evolucionan hacia la forma silvestre como manera de supervivencia al no tener contacto son el hombre (De Wet and Harlan, 1975; Cao et al., 2006; Reagon et al., 2010).

Los trabajos antes mencionados sugieren que el origen del arroz maleza pudo haber sido producto de múltiples procesos, los que además tienen dependencia con el lugar donde se encuentran. En los centro de origen del cultivo, donde aún hoy existen arroces silvestres o salvajes como son Asia y África, los eventos de hibridación natural entre éstos y el arroz cultivado, y la posterior selección en las condiciones de cultivo, podrían haber dado lugar a diferentes eventos de origen del arroz maleza (Delouche et al., 2007). En lugares de cultivo de arroz donde no existen especies de *Oryza* nativas, como los es nuestro país, el arroz maleza pudo haber tenido origen a partir de la introgresión de germoplasma contaminado (figura 1). Posteriormente a su ingreso, los procesos de selección y re-hibridación pueden haber dado lugar a los diferentes biotipos encontrados en la actualidad.

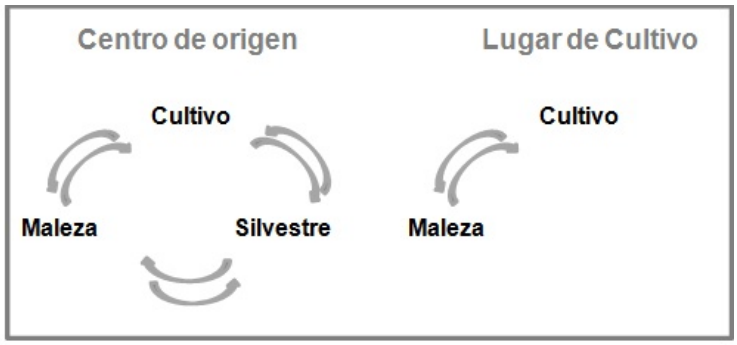

Figura 1: Esquema de los posibles flujos entre silvestre-maleza-cultivo en centro de origen y en lugares de cultivo.

En Uruguay se encuentran comúnmente dos tipos de arroz maleza que se diferencian por sus características fenotípicas, el "arroz negro" denominado así por el color de sus glumas con aristas largas (4-5 cm) y el "arroz rojo" de glumas ocres y de arista corta. Los efectos negativos del arroz maleza, se deben a su competencia con el arroz cultivado y su control se ve restringido a buenas prácticas de manejo. Sus efectos negativos desde el punto de vista agronómico, al igual que en otras malezas comunes, se deben a su competencia por diferentes factores productivos (nutrientes, agua, luz, etc.) con el arroz cultivado, con la particularidad de que el arroz maleza se ve favorecido por casi todas las labores que se realizan para el cultivo. Además existe el agravante de que no se dispone de controles químicos adecuados por tratarse del mismo género y especie. Los perjuicios de la infección con arroz maleza van más allá del campo, afectando no solamente la productividad del arroz cultivado sino también la calidad de los lotes cosechados. Esto último debido a que la coloración y textura del pericarpio afectan los procesos industriales de descascarado y pulido (Hoagland and Paul, 1978).

Como respuesta comercial para el control del arroz maleza y de reciente adopción en Uruguay, se liberó una variedad portadora de una mutación puntual en la enzima acetolactato sintasa (ASL) (tecnología Arroz Clearfield®) que le confiere resistencia a herbicidas del tipo imidazolinonas. Estas mutaciones pueden incorporarse por cruzamientos naturales a poblaciones de arroz malezas. La presión de selección por el uso del herbicida, selecciona híbridos entre maleza y cultivares portadores de la mutación, ya sea que hayan surgido por flujo génico o por eventuales mutaciones espontáneas, que generan poblaciones de maleza resistentes. La existencia de flujo de genes entre maleza y cultivares ha sido reportado por varios trabajos (Gealy et al., 2003; Kuroda et al., 2005; Shivrain et al., 2010), es por ello que se estudio una población de arroz maleza de Uruguay para evaluar el grado de flujo génico en campos con diferentes tiempos de utilización de dicha tecnología (Rosas, 2011). Rosas et al (2011) reportaron la existencia de arroz maleza portadores de dos alelos diferentes en el gen de la enzima ASL que le confieren resistencia, lo que indicaría la existencia de varios eventos de flujo génico en tan sólo 3 años de uso de la tecnología.

La intensificación del cultivo, sumado a las evidencias de flujo de genes entre maleza y cultivares en campos Uruguayos, hace imprescindible aumentar el conocimiento acerca de la diversidad genética y el comportamiento del arroz maleza. Comprender los procesos evolutivos que dieron origen a los biotipos malezas encontrados en Uruguay puede ser de ayuda para establecer estrategias efectivas para su control.

Los genomas de las plantas

Las células vegetales tienen la particularidad de contener tres tipos de genomas. Por un lado, el genoma nuclear organizado en cromosomas, en el caso del arroz es un genoma diploide de 400 Mb organizado en 12 cromosomas (2n=24). Por otro lado, los genomas de los organelos citoplasmáticos, la mitocondria y el cloroplasto. Estos genomas extranucleares presentan ciertas características en común, ambos son circulares y haploides. Su origen como demuestran muchos trabajos de evolución molecular fue por un proceso de endosimbiosis de bacterias. En el caso de las mitocondrias habrían surgido a partir del grupo de α-proteobacterium al igual que las mitocondrias animales y el cloroplasto a partir de las cianobacterias (figura 2) (Martin et al., 2002; Eguiarte et al., 2003; Bock and Timmis, 2008).

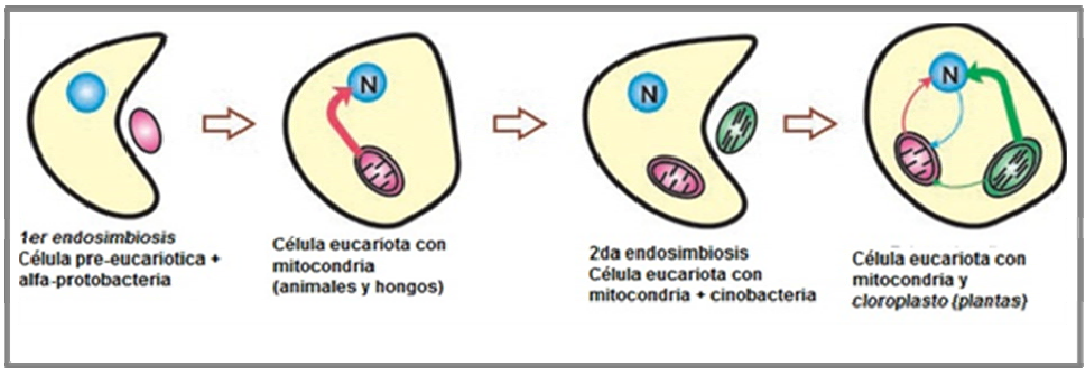

Figura 2: Endosimbiosis y transferencias de genes intracelulares entre genomas durante la evolución de la célula vegetal (Bock and Timmis, 2008). Las flechas *indica*n la dirección de la trasferencia. Los colores de las flechas corresponden al color del compartimento desde el que el material genético transferido se originó (rojo:mitocondria; azul:el núcleo; verde: cloroplasto.

Durante la evolución los genomas de las células eucariotas sufrieron grandes reestructuras, necesarias para su adaptación a los cambios generados por la proceso de endosimbiosis. Uno de los mayores cambios fue la compartimentalización del material genético entre el núcleo y los organelos. Si bien ambos organelos mantuvieron su ADN, la restructuración de las células eucariotas implicó la pérdida de genes prescindible o redundantes, así como la translocación masiva de genes desde los organelos ancestrales hacia el núcleo (Bock and Timmis, 2008).

En plantas con flores (angiospermas) el cloroplasto se hereda de manera casi exclusiva por vía materna, mientras que en coníferas (gimnospermas) usualmente los cloroplastos se heredan por vía paterna (Birky, 1991). Por tanto, mientras que en las angiospermas el cloroplasto describe la historia evolutiva femenina (patrones de dispersión de las semillas), en coníferas se relaciona con la función masculina (patrones de dispersión del polen) (Eguiarte et al., 2003). En el caso de las mitocondrias se heredan siempre por vía materna.

Estudios genéticos y bioquímicos (Dobberstein et al., 1977; Highfield and Ellis, 1978; Bedbrook, 1980), sugirieron que el genoma de los plastos tiende a la disminución de su tamaño. Éstos codifican entre 50 a 200 proteínas, mientras que las cianobacterias que le dieron origen codifican miles de proteínas (Butterfield, 2000; Martin, 2003). La disminución en el tamaño de los organelos se ha atribuido al ahorro de energía en la síntesis de ADN sobre todo en especies con genomas poliploides (Wolfe et al., 1991). Los genomas de los organelos, no presentan la información necesaria para codificar todas las proteínas que les permita cumplir con su función, haciendo que éstos sean dependientes de genes nucleares (Soll and Schleiff, 2004). El control nuclear sobre la forma y función de los organelos, tiene como resultado una compleja regulación de las actividad de los genes en los tres compartimentos.

El tamaño de los genomas cloroplásticos pueden variar desde 100 a 3000 kb, en arroz el tamaño promedio es de 134 kb. En las angiospermas el cloroplasto tiene una estructura muy conservada, presenta dos regiones de copia única, una larga (LSC, Long Single Copy) y otra corta (SSC, Short Single Copy), de aproximadamente unos 80-90 kb y 16-27 kb respectivamente, y una región duplicada invertida (IR, Inverted Repeat), formada por dos segmentos idénticos en sentidos opuestos, separados por la región SSC. Las regiones invertidas repetidas pueden tener un tamaño variable, de 12 a 25kb cada una (Yang et al., 2010) (Figura 3) El número de genes presente en el cloroplasto es también muy conservado, entre 110 a 113 genes.

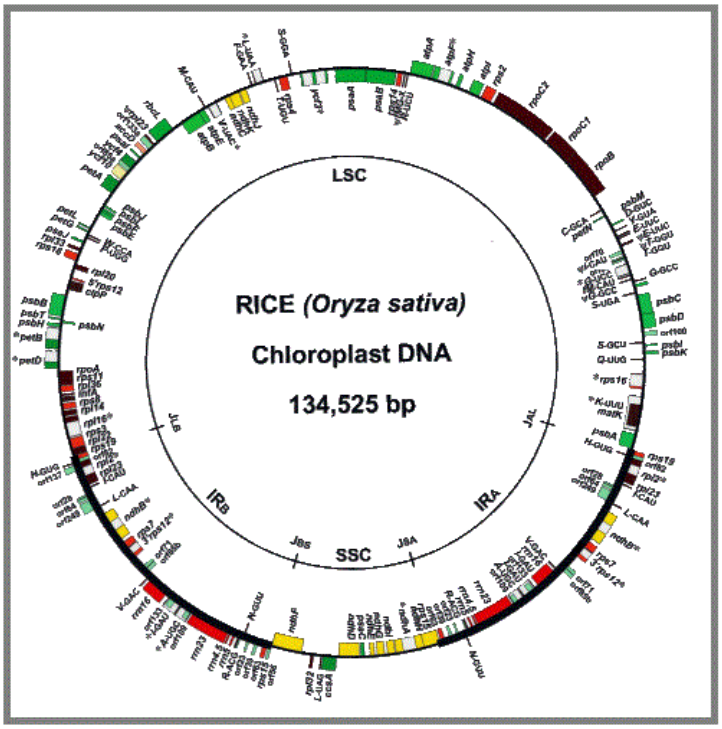

Figura 3: Esquema del genoma del cloroplasto de arroz.

Las características del genoma de los cloroplastos, tales como su pequeño tamaño, la herencia uniparental en estado haploide y el alto grado de conservación de su secuencia (Birky Jr, 1978), han facilitado la utilización de regiones cloroplásticas, ya sean regiones codificantes como no codificantes, para estudios poblacionales (Neale et al., 1988; Provan et al., 2001). La conservación en el orden de los genes, la perdida de heteroplasmia y de recombinación los hace muy útiles para estudios filogenéticos en plantas (Provan et al., 2001) y para resolución de las más distantes relaciones genealógicas (Graham and Olmstead, 2000; Shaw et al., 2007).

A pesar del grado de conservación de los genoma en estos organelos, algunas regiones contienen un grado de variabilidad suficiente como para realizar estudios de relaciones filogenéticas entre especies muy cercanas, así como también estudios evolutivos y de genética poblacional (Cronn et al., 2008). En arroz, la tasa evolutiva del ADN del cloroplasto es tres veces mayor que el de la mitocondria (Tian et al., 2006). Los tipos de mutaciones reportadas entre los genomas de los plastos son principalmente inversiones (Hiratsuka et al., 1989), translocaciones (Ogihara et al., 1988), e inserciones/deleciones (Ogihara et al., 1991; Kanno et al., 1993; Maier et al., 1995) así como también sustituciones de una única base (SNPs) (Morton and Clegg, 1995). Doyle et al. (1992) reportó la existencia de tres inversiones en el genoma de los cloroplastos de las gramíneas presentes en todos los miembros de dicha familia (Doyle et al., 1992). Éstas inversiones en el genoma del cloroplasto de arroz, en particular, generaron grandes reordenamiento de los genes especialmente en la región larga de copia única (LSC) (Shahid Masood et al., 2004). Éste tipo de mutaciones se han asociado a regiones de secuencias de ADN con repeticiones de segmentos cortos (zonas de baja complejidad). Éstas regiones son

propensas a la generación de errores en la síntesis de ADN debido al "corrimiento de hebra" de la polimerasa, produciendo así, mutaciones del tipo indel por recombinación intramolecular desfasada (Kelchner, 2000). Comparaciones de la secuencias completas de los genomas de los plastos entre tres cereales arroz, trigo y maíz, indicaron la existencia de regiones hot-spot para mutaciones tipo indel, las que presentaban como característica en común ser regiones de secuencias repetidas (Ogihara et al., 2002) .

El genoma mitocondrial de las plantas es de mayor tamaño que el cloroplástico, se encuentran entre 300 a 600 kb de tamaño. En algunas Cucurbitaceae y Malvaceae puede llegar a ser hasta de 2000kb (Ward et al., 1981; Soltis et al., 1992; Alverson et al., 2010). En arroz el genoma mitocondrial es de 490kb (Notsu et al., 2002). A diferencias del grado de conservación del genoma cloroplástico, el genoma de la mitocondria tiene un comportamiento evolutivo muy dinámico. Incorpora genes nucleares y del cloroplasto con facilidad, cambia la posición relativa de sus genes y en algunas especies puede ser encontrado en varias formas diferentes, como un gran cromosoma o varios pequeños. Si bien es un genoma con una estructura muy cambiante debido a los reordenamiento de las regiones no codificantes, la secuencia de los genes es muy conservada. (Graur and Li, 2000).

Dentro del género *Oryza* se encuentran publicados los genomas nucleares de dos de las subespecies mayoritarias de *Oryza sativa*, *O. sativa ssp japonica* y *O. sativa ssp indica*, cinco genomas completos de cloroplastos: *O. sativa ssp japonica*, *O. sativa ssp indica* (Tang et al., 2004), *O. nivara* (Shahid Masood et al., 2004b), *O. rufipogon*, *O. meridionalis*, (Waters et al., 2011) y un genoma mitocondrial de *O. sativa ssp japonica* (Notsu et al., 2002).

Existe entre los genomas de los organelos y el nuclear un continuo proceso de transferencia; el que se ha hecho más evidente en los últimos años debido a la disponibilidad de genomas completos nucleares y de cloroplastos de varias especies de plantas. Esto ha revelado que en el genoma nuclear existe gran cantidad de ADN proveniente de los organelos (Martin, 2003; Richly and Leister, 2004a, 2004b). Los cloroplastos y las mitocondrias están constantemente bombardeando al el núcleo, en un proceso de trasferencia lateral de genes, lo que le confiere dinámica al genoma nuclear. Los trabajos de Stegemann et al. (2003) y Huang et al. (2003), dos estudios independientes que estudiaron las transferencias entre el cloroplasto y el núcleo en una planta de tabaco, utilizaron la técnica de transformación cloroplástica con genes reporteros de resistencia a antibióticos para evaluar la existencia y frecuencias de dichas trasferencias. Ambos trabajos llegan a resultados semejantes, en el primero estimaron que 1 de cada 5.000.000 de células somáticas eran portadoras de una transferencia de ADN cloroplástico al núcleo. En el trabajo de Huang et al. (2003) determinaron que 1 de 16.000 plantas de tabaco, en tan sólo una generación, presentaban al menos una transferencia. Las tasas estimadas en éstos trabajos sólo consideraron una región (la portadora de los genes de resistencia) y los genes trasferidos de alta expresión por lo que estás tasas pueden ser aún mayores.

Como consecuencia, varios miles de genes funcionales han sido adquiridos por las plantas durante la evolución de los cloroplastos (Martin et al., 2002). Sin embargo, los organelos no sólo donan genes funcionales, se han encontrados fragmentos de ADN no codificantes de origen plastídico en muchos de los núcleos de las plantas (Yuan et al., 2002; Richly and Leister, 2004a).

Comparaciones filogenéticas de genes nucleares individuales distribuidos en todo el genoma de *A. thaliana* con genomas procariotas de organismos representativos, revelaron que 866 de las 9368 proteínas codificadas por el núcleo de *A. thaliana* eran suficientemente conservadas y mostraban alta similaridad con genes de cianobacterias (Martin et al., 2002). Extrapolando estos datos se concluyó que el 18% de los genes nucleares de *A. thaliana* fueron adquiridos a partir del ancestro procariota de los cloroplastos (Martin and Herrmann, 1998).

Los mecanismos por los que las regiones de ADN cloroplástico son transferidos al núcleo están aún en discusión. Hay tres posibles mecanismos propuestos: 1) grandes cantidades de ADN liberada del genoma del plasto y que recombina con el genoma nuclear, 2) mediado por mRNA, o 3) mediado por cDNA (posiblemente mediado por virus). En los trabajos de Stegemann et al. (2003) y Huang et al. (2003), mencionados antes, se propone que sería el continuo escape de grandes cantidades de ADN cloroplástico y su recombinación no homóloga con el ADN nuclear el mecanismo que mediaría estas trasferencias. Este mecanismo explica muy bien la composición del genoma nuclear de *A. thaliana*, donde se encotraron genes que codifican para t-RNAs cloroplásticos conteniendo intrones (Martin, 2003). Sin embargo el genoma del arroz muestra otras evidencias, la reciente incorporación de 33 kb de ADN cloroplástico en el cromosoma 10 del arroz parece haber sido mediada por cDNA (The Rice Chromosome 10 Sequencing Consortium).

La mayoría de los segmentos de ADN integrado a los genomas nucleares que han sido trasferidos desde los organelos citoplasmáticos tienen tamaños menores a 1kb de largo (Richly and Leister, 2004a). Sin embargo, se han identificado fragmentos mayores, como en el caso de la inserción de una región de ADN mitocondrial en el cromosoma 2 de *A. thaliana* de 620 kb (Stupar et al., 2001a) y la inserción de ADN cloroplástico en el cromosoma 10 de *O. sativa subespecie japonica* de 131 kb de largo (Yu et al., 2003).

En ambas especies, las inserciones de ADN de organelos citoplasmáticos en el núcleo no están distribuidos al azar, sino que están ligados entre sí. Se ha encontrado por estudios de homología, regiones de un organelo que se corresponden con la sumatoria de un determinado conjunto de insertos pequeños. Esto implicaría que las inserciones primarias fueron más grandes que las observadas hoy, pero que su tamaño decayó durante la evolución generando fragmentos más pequeños con secuencias divergentes a lo largo del tiempo (Richly and Leister, 2004b; Leister, 2005a).

La distribución y abundancia de los fragmentos cloroplásticos insertos en el genoma nuclear fue estudiada por Matsuo et al., 2005, reportando que los fragmentos de gran tamaño (> 10kb) se

encuentran insertos preferentemente en la región pericentromética de los cromosomas. Esta región es pobre en genes por lo que puede ser propicia para la integración de ADN foráneo. En este trabajo compararon el genoma del cloroplasto de *O. sativa ssp japonica* contra el genoma nuclear de la misma y determinaron que el 0.2% del genoma nuclear es de origen cloroplástico.

En el genoma de arroz en particular existe una dinámica de integración de fragmentos cloroplásticos a el núcleo muy alta. Estos fragmentos insertos pueden provenir de cualquier parte del genoma del cloroplasto, la transferencia y la integración en el genoma nuclear se produce casi igualmente en todo el genoma nuclear. (Matsuo et al., 2005). Sin duda que entender este proceso de plasticidad genómica es de gran interés tanto desde el punto de vista estrictamente académico por sus implicancias en genómica evolutiva, como también práctico pues los mismos pueden representar importantes marcadores génicos que puede ser aprovechados en estudios de variabilidad intraespecífica.

Secuenciación masiva para el estudio genómico en plantas.

La disponibilidad de genomas totalmente secuenciados de muchos de los cultivos importantes y la posibilidad de realizar secuenciación a gran escala a bajos costos provee de oportunidades para profundizar el estudio de la historia evolutiva de las plantas domesticadas.

La genómica comparativa en cultivos está siendo trasformada y una nueva generación de enfoques experimentales y computacionales están surgiendo a partir de la gran disponibilidad de datos de secuencias. El futuro del mejoramiento de cultivos puede estar centrado en la comparación de los genomas de plantas individuales, y algunos de las mejores oportunidades puede surgir de la combinación de nuevas estrategias de mapeo y análisis evolutivos para el descubrimiento y uso de la variabilidad genética (Morrell et al., 2011). En un artículo reciente se describe el potencial del uso de la resecuenciación para contribuir a la comprensión de la complejidad y diversidad asociada con la domesticación de los cultivos (Varshney et al., 2010).

Los primeros genomas secuenciados se obtuvieron utilizando el método "tradicional" de secuenciación. Para ello era necesario, generar bibliotecas a partir de segmentos individuales del ADN genómico. Éstos eran clonados en cromosomas artificiales bacterianos (BAC) y luego secuenciados por la técnica de Sanger (Rounsley et al., 2009).

El avance en las tecnologías de secuenciación de nueva generación (NGS), aceleró el proceso de obtención de genomas completos y permitió plantearse desafíos mayores. Las tecnologías NGS tienen en común la capacidad de generar datos en grandes volúmenes y de disminuir el error al no requerir del clonado para la obtención de las bibliotecas de secuenciación (Rounsley et al., 2009).

El contar con la secuencia de genomas completos de organelos (cloroplasto y mitocondrias) provee de recursos y de información muy valiosa para el estudio de la evolución de las plantas.

Con el desarrollo de las tecnologías de nueva generación, se está convirtiendo en una norma la obtención de genomas completos a través de la aplicación de éstas tecnologías (Zhang et al., 2011). Recientemente se ha visto un incremento acelerado en el número de genomas de organelos secuenciados. Hoy están disponibles 285 genomas completos de cloroplastos y 82 genomas completos de mitocondrias de plantas. http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html revisado 10/10/2012).

El sistema 454 fue la primera plataforma de secuenciación de nueva generación disponible como un producto comercial. Este sistema está basado en la técnica de pirosecuenciación del ADN, desarrollada inicialmente por Mostaza Rognaghi y colaboradores a finales de la década del '90 (Ronaghi et al., 1998).

La plataforma de secuenciación 454/Roche ha sido aplicada con diversos objetivos, secuenciación de novo de organismos no modelos, diagnóstico, secuenciación de trascriptomas, secuenciación de ADN antiguo, de BACs, amplicones y secuenciación de metagenomas entre otras. La pirosecuenciación ha sido perfeccionada para la secuenciación de genomas completos (Jarvie and Harkins, 2008) y es un método eficiente para la detección de SNPs y análisis de metilación.

La longitud de las lecturas obtenidas ha ido en aumento, desde 200 pb en la primer versión, hasta los 1000 pb obtenidos actualmente aplicando GS FLX+ System. La obtención de fragmentos largos, lo hace un método excelente para la secuenciación de especies no modelos para las que no se disponen de genomas de referencias y facilita la resolución del ensamblaje de genomas con abundante ADN repetitivo (Kircher and Kelso, 2010).

El importante desarrollo de las ciencias genómicas en los últimos años, debido a los avances en las técnicas de secuenciación masiva antes mencionado y al desarrollo de las capacidades bioinformáticas para el análisis de grandes cantidades de información de secuencia, posibilita abordajes prácticamente impensables hace pocos años atrás. Sin embargo a pesar de la posibilidad técnica de obtener con relativa facilidad y a costos bajos secuencias genómicas en grandes volúmenes, la principal limitante que existe en la actualidad es la capacidad de analizar e interpretar estos resultados. Por esta razón, otro de los objetivos de este trabajo de tesis de maestría es precisamente adquirir conocimientos en la integración de herramientas bioinformáticas y datos de secuenciación de nueva generación para analizar genomas de plantas abordando un tema de gran importancia tanto evolutivo como para su posible aplicación en el mejoramiento genético de una de las principales producciones agrícolas de Uruguay.

Objetivo general

Estudio del origen evolutivo del arroz maleza de Uruguay a través de aproximaciones de genómica evolutiva/comparativa para la generación de información y herramientas que permitan estudiar sus procesos evolutivos y diseñar estrategias de control.

Objetivos específicos

1) Desarrollo de una metodología para la obtención de las secuencias cloroplásticas a partir de datos de secuencias de ADN total.

2) Identificación de variantes entre los genomas cloroplásticos públicos para ser aplicadas como marcadores genéticos que permitan determinar el origen del arroz maleza en estudio y generar herramientas para profundizar en el conocimiento de la diversidad existente de las poblaciones de ésta maleza en el país, así como su dinámica poblacional..

3) Estudio de la dinámica de transferencias de segmentos genómicos entre el genoma del cloroplasto y el núcleo

Materiales y Métodos

Material vegetal.

El individuo de arroz maleza elegido para la secuenciación fue obtenido desde una chacra ubicada en el Departamento de Cerro Largo (manchón n° 356, planta nro. 8), de aquí en más AM356-8. Las semillas de AM356-8 fueron germinadas en maceta hasta estadio de 4 hojas en invernáculo.

Aislamiento de ADN.

El aislamiento de ADN genómico se realizó a partir del material verde obtenido bajo las condiciones antes descritas. El protocolo de extracción utilizado fue FAO/IAEA (Interregional Training Course on Mutant Germplasm Characterization), con modificaciones menores. Se molieron de 2 a 4 g de tejido vegetal en mortero con nitrógeno líquido, colocando el material finamente molido en tubos con 5 ml de buffer de extracción con 5 ml de buffer de extracción precalentado (65°C) (2% CTAB, 1,4 M NaCl, 20 mM ácido etilendiamintetracético [EDTA] pH 8, 100 mM Tris pH 8, 2% PVP – 40), agregando a cada tubo individual 6,25 ul β – mercaptoetanol. Luego de incubar las muestras a 65°C durante 20 minutos con agitación suave y periódica, se le agregó 1 volumen de Cloroformo-isoamílico (24:1), mezclando cuidadosamente. Seguido por una centrifugación a 12 000 rpm durante 20 minutos a 4°C. El sobrenadante fue transferido a otro tubo, donde se precipitó el ADN con 2/3 de volumen de isopropanol. Las muestras fueron centrifugadas a 5 000 rpm durante 20 minutos a temperatura ambiente (TA). El pellet obtenido fue lavado con etanol 70% y centrifugado a 5 000 rpm durante 5 minutos.

La muestra fue cuantificada por electroforesis en gel de Agarosa 0,9 %, revelado con tinción de bromuro de etidio. Se utilizó Mass Ruler High Range 42.2 ng/ul (Fermentas) como estándar de concentración y tamaño. La concentración y cálculo de relación ADN/Proteína fue realizada por medición espectrofotométrica en NanoDrop 8000 Spectrophotometer (Thermo Scientific).

Se ajusto la concentración de la muestra a 700 ng/ul con un ratio de ADN/proteína mayor 1.9, requerimientos mínimos para la secuenciación. La concentración fue media luego con Picogreen (Invitrogen) confirmando la concentración y calidad obtenida.

Secuenciación.

Preparación de librería genómica:

La librería genómica de secuenciación fue realizada con el kit " GS FLX Titanium Rapid Library MID Adaptors" (Roche), según los fundamentos descritos por Marguiles et al, 2005 y las modificaciones al protocolo descrito por Roche. Se partió de 5 µg de ADN genómico. El ADN fue fraccionado mecánicamente para obtener fragmentos de 400-1000 bp. Luego los extremos de estos fragmentos fueron reparados mediante una reacción catalizada por la T4 Polimerasa y la T4 polinucleótido quinasa. En la etapa siguiente fueron ligados los adaptadores A y B y el identificador para secuenciación de múltiples muestras (RLMID- MID5: ATCAGACACG), mediante la enzima T4 ligasa. Éstos adaptadores proporcionan las secuencias de hibridación para la posterior amplificación y secuenciación de los fragmentos de la librería. Los fragmentos menores a 350 pb fueron descartados utilizando el método AMPure Bean incluido en el kit y según protocolo descrito por Roche. La librería luego es inmovilizada, esto es la fijación de los fragmentos con los adaptadores ligados a perlas magnéticas. La unión es mediante el adaptador B. Este adaptador está biotinilado en su extremo 5' lo que permite la fijación de la librería a las perlas recubiertas de estreptoavidina. Luego de un paso de relleno de huecos, las hebras no biotiniladas son removidas de las perlas y es generada la librería molde de ADN de simple cadena (ssDNA). La librería fue cuantificada utilizando TBS 380 Fluorometer según protocolo descrito por Roche, para la determinación de la proporción óptima (moléculas de ADN:perlas) necesaria para la PCR en emulsión (emPCR). La calidad se evalúo mediante Agilent Bioanalyzer High Sensitivity DNA chip. Las alícuotas de trabajo fueron preparadas con una concentración final de $1x10^7$ moléculas/µl.

Amplificación en emulsión (emPCR):

La amplificación en emulsión se realizó con el kit GS FLX Titanium SV emPCR Kit (Lib-L) (Roche), según protocolo descrito por Roche. A las alícuotas de trabajo de la librería genómica se les agrego la mezcla de amplificación (mezcla de enzimas para emPCR y PPiasa) y aceite. Mediante una agitación vigorosa en TissueLyser II (Qiagen) se emulsiona la mezcla generando los microrreactores de 50-100 µm de diámetro donde se realizó la reacción de amplificación. En la siguiente etapa fueron recuperadas las perlas con el producto de la amplificación (librería genómica doble cadena). Para la recuperación de las perlas con ADN fue necesario primero romper la emulsión químicamente. La recuperación y lavado de las perlas se realizó según protocolo descrito por Roche para volúmenes pequeños de emulsión (SVE- Small Volume Emulsions). Una vez recuperadas las perlas con ADN doble cadena, se hibridaron los cebadores para secuenciación. Finalizada ésta etapa, se realizó un enriquecimiento en perlas con ADN hibridado y con amplificación eficiente (perlas con ADN doble hebra) y luego se cuantifico mediante un contador de partículas Beckman-Coulter el porcentaje de perlas enriquecidas que se obtuvo. Esta cuantificación permitió evaluar la eficiencia de la reacción de amplificación y la condición de la muestra para continuar con el procedimiento.

Secuenciación:

La secuenciación se realizó con el kit GS Titanium Sequencing XLR70 (Roche), para secuenciar 1/4 de GS Titanium PicoTiterPlate (PTP) 70x75 en 454 Genome Sequencer FLX System (Roche). La técnica utilizada es la pirosecuenciación (Ronaghi et al., 1998), basada en la secuenciación por síntesis. En esta es acoplada la síntesis de ADN a una reacción quimioluminiscente. La librería ssDNA amplificada con el cebador de secuenciación unido se incubó con las enzimas DNA polimerasa, ATP sulfurilasa, luciferasa y apirasa, más los sustratos adenosina-5'-fosfosulfato (APS) y luciferina. Luego es añadido uno de los 4 dNTPs, donde la ADN polimerasa cataliza su incorporación si este es complementario al molde. Si existe incorporación es liberado PPi equivalente a la cantidad de dNTP incorporado. La ATP-sulfurilasa convierte cuantitativamente el PPi en ATP en presencia de APS. El ATP generado permite la conversión de la luciferina en oxiluciferina por acción de la luciferasa, generando luz visible en cantidades proporcionales a la cantidad de ATP presente. La luz emitida es detectada por una cámara CCD y es analizada por el software asociado al equipo. La intensidad de la señal emitida es proporcional a la cantidad de nuleótidos incorporados. Los nucleótidos no incorporados son degradados por enzima apirasa. La entrada de un nuevo flujo de nucleótidos comienza un nuevo ciclo. Los patrones lumínicos detectados por la cámara CCD son traducidos a secuencias por el programa GS Sequencer. Los datos son almacenados en un archivo binario (formato standard flowgram format, sff) donde se encuentra la información de secuencias de cada uno de los lecturas y sus respectivos datos de calidad

Análisis de datos.

Análisis primario y evaluación de calidad.

Los datos crudos se analizaron con programas desarrollados por el grupo y con programas proporcionados por Roche.

En el procesado primario de los datos fueron removidas las secuencias de los adaptadores y el código de múltiples muestras (MID5), sff_extract (ROCHE) y/o una aplicación del programa de ensamblado Newbler (ROCHE).

El cálculo del largo promedio y el porcentaje de lecturas duplicadas artificialmente se determinaron aplicando programas desarrollados por el laboratorio de Biomatemáticas. (*media_largos.pl* y *cutlecturas.pl*).

Representación de los genomas de la célula vegetal en los datos de secuencia.

Genoma nuclear: se comparó las lecturas de AM356-8 con el genoma nuclear de *O. sativa ssp. japonica cv Nipponbare*, utilizando el programa de ensamblado y mapeo Newbler (Roche). Las condiciones optimizadas para el mapeo de genomas eucariotas completos.

Genoma de mitocondria: se comparó las lecturas de AM356-8 con el genoma mitocondrial de *O. sativa ssp. japonica cv Nipponbare* (NC_011033.1), utilizando el programa blastn con los parámetros e-value: $1x10^{-10}$, -FF.

Genoma cloroplástico: se comparó las lecturas deAM356-8 con el genoma del cloroplasto de *O. sativa ssp japonica cv Nipponbare*, utilizando el programa blastn con los parámetros e-value: $1x10^{-10}$, -FF.

Identificación y clasificación de secuencias cloroplásticas.

Genomas públicos de referencia.

La secuencias de los genomas públicos fueron obtenidas desde NCBI-RefSeq. (http://www.ncbi.nlm.nih.gov/RefSeq). (Tabla 1)

Genoma cloroplasto	N° acceso	Tamaño (pb)	Publicación
Orza rufipogon	NC017835.1	134544	2012
Oryza nivara	AP006728	134494	2004
Oryza sativa ssp. indica	AY522329.1	134496	2004
Oryza sativa ssp. japonica	AY522330.1	134551	2004

Tabla 1: Genomas cloroplásticos públicos de referencia obtenidos desde NCBI

La identificación de las lecturas cloroplásticas de AM356-8 se realizó por comparación de las lecturas obtenidas en la secuenciación con los tres genomas de referencia disponibles a marzo 2011(genoma cloroplasto *O. nivara*, genoma cloroplasto *O. sativa ssp indica*, *O. sativa ssp japonica*). La comparación se realizó con el algoritmo BLAST [Altschul et al., 1999], opción bastn con e-value de $1x10^{-10}$ y -F F (mantención de secuencias de baja complejidad) para cada genoma de referencias por separado. Sobre el resultado de cada uno de los blast se seleccionó el conjunto de lecturas con similitud a cada genoma cloroplástico, generando tres conjuntos de lecturas RJ, RI, RN respectivos a cada genoma comparado.

Sobre el resultado de blast de los conjunto de lecturas RJ, RI, RN se aplicaron dos filtros. El primero fue sobre el largo del alineamiento de la lectura con el genoma con el que se comparó, se seleccionaron en los tres casos lecturas con un alineamiento mayor a 100 nucleótidos de largo. Se generaron de esta manera tres conjuntos RJ_100, RI_100, RN_100. El segundo filtro aplicado fue sobre el porcentaje de alineamiento de la lectura con el genoma con el que se comparó. El porcentaje de alineamiento o solapamiento (%s) se calculó como el largo el alineamiento/largo de la lectura. El largo de la lectura se obtuvo con el programa *fastaUtilis.pl*.

El conjunto llamado RC, se generó con las lecturas que presentaron un porcentaje de alineamiento con el cloroplasto de 99% (%s >=99%). El conjunto RI se generó con aquellas lecturas que presentaban un alineamiento menor al 99% de su largo (%s <99). Los mismos filtros fueron aplicados sobre los tres conjuntos de lecturas (RJ, RI, RN), generando 2 archivos por genoma de referencia *japonica*: RJC/RJI, *indica*: RIC/RII, *nivara*: RNC/RNI. El conjunto RIJ se subdividió generando el conjunto RIJa, donde se agruparon lecturas con un porcentaje de solapamiento (%s) menor a 90%.

Ensamblado DeNovo del cloroplasto AM356-8.

Los conjuntos de lecturas RCJ,RCI,RCN se utilizaron para el ensamblado de novo del cloroplasto de AM356-8.

a) El conjunto de lecturas RCJ fue ensamblado con el programa Newbler (Roche). El archivo de extensión sff conteniendo el subconjunto de lecturas deseado se obtuvo utilizando el programa *sfffile* (Roche). Los parámetros utilizados para el ensamblado fueron los recomendados para el ensamblado datos genómicos de organismos no complejos (largo mínimo de solapamiento=50 bases; mínima identidad de la región solapante= 90%).

b) Los conjuntos de lecturas RCI, RCN, se ensamblaron bajo las mismas condiciones que en a).

Mapeo contra genomas de referencia.

a) El mapeo del conjunto de lecturas RCJ se realizó utilizando MIRA 3 versión 3.4.0 (Chevreux, 2005) con los parámetros recomendados para mapeo contra genoma de referencia con datos genómicos de organismos eucariotas generados a partir de secuenciador 454/Roche. Los parámetros de mapeo utilizados fueron: job= mapping,accurate,454 ; AS:ard=yes; AS:urd= no; SK: mnr=no; nrr=10; mmhr=1). Se mapeo contra el genoma del cloroplasto de *O. sativa ssp japonica* (AY522330.1).

Búsqueda de regiones divergentes entre los genomas públicos.

La búsqueda de las regiones divergentes entre los genomas de cloroplastos de *O. nivara*, *O. rufipogon*, *O. sativa ssp indica*, *O. sativa ssp japonica* se realizó utilizando el paquete de herramientas para genómica comparativa Whole Genome VISTA Tools (Zambon et al., 2005), En este paquete está implementado el algoritmo LAGAN para alineamiento de genomas y MLAGAN para alineamientos múltiples de genomas (Brudno et al., 2003).

La búsqueda de las regiones variables entre los 4 genomas se realizó en ventanas de 600 bases recorriendo el alineamiento por completo (Kumagai et al., 2010). Las regiones en las cuales se encontraron diferencias en las secuencias del tipo SNPs y/o INDELs fueron extraídas. Confirmando lo observado mediante una nuevo alineamiento con el algoritmo ClustalW (Thompson et al., 1994) implementado en MEGA 4 (Kumar et al., 2008) confirmando la existencia de SNP y/o INDELS.

La detección de SNPs se realizó además con el algoritmo Muscle 3.6 implementado en el programa MAUVE (Darling et al., 2004). Como archivos de salida el programa reporta las regiones conservadas entre los genomas comparados, así como las regiones que presentan diferencias de una única base entre los genomas. Para cada sitio polimórfico identificado (SNP) en el alineamiento, el programa registra en el archivo de SNP los nucleótidos presentes en cada genoma en ese sitio, junto con las coordenadas del lugar en cada genoma de forma de verificar los encontrados por el procedimiento anterior.

Las secuencias conteniendo INDELs, fueron extraídas seleccionando 600 bases que contuvieran el INDEL y estás se compararon vía blastn (e-value=$11X10^{-10}$) con las lecturas del conjunto RCJ. Aquellas lecturas que mostraron identidad superior a 90%, con un largo de alineamiento mínimo de 100 nucleótidos y que contenían la región de interés fueron extraídos utilizando del programa *awk*. Las lecturas extraídas se alinearon con las secuencias correspondientes de referencia utilizando el algoritmo ClustalW, de esta manera se confirmó la existencia de la variante tipo INDEL evaluada. Para cada INDEL evaluado se realizó el mismo procedimiento. Las secuencias de referencia de cada región interrogada se obtuvieron a partir del archivo genebank (gbk) de cada genoma correspondiente, seleccionando la región de interés mediante el programa Artemis (Rutherford et al., 2000).

Identificación de trasferencias desde genoma cloroplástico hacia nuclear.

Para la identificación de las regiones transferidas desde el genoma cloroplástico al nuclear se comparó las lecturas cloroplásticas con el genoma nuclear. El conjunto de lecturas consideradas como de origen cloroplástico fueron los conjuntos RC generados anteriormente. La comparación del conjunto RCJ con el genoma nuclear de *O. sativa ssp japonica cv Nipponbare* y con el cloroplasto *O.sativa subsp. japónica* se realizó mediante el programa BLAST [Altschul et al., 1999] opción blastn, fijando un e-value de $1x10^{-13}$ Una vez obtenidos los resultados de BLAST contra ambos genomas se le agrego los largos de cada una de las lecturas. Los largos de las lecturas se calcularon utilizando el programa fastaUtilis.pl.

Se definieron tres tipos de trasferencias a ser identificadas:

Trasferencias Modernas.

Sobre el resultado de blast del conjunto de lecturas RIJa con el genoma cloroplástico de *japonica* se aplicaron los filtros: porcentaje de identidad con el cloroplasto menor al 90% y porcentaje de alineamiento/solapamiento con el cloroplasto mayor o igual a 99%. (%ID<90%; %s>=99%). Sobre el resultado de blast del conjunto de lecturas RIJa con el genoma nuclear de *japonica* se aplicaron los filtros: porcentaje de identidad mayor o igual a 99 % y el porcentaje de alineamiento/solapamiento mayor o igual a 99%. Las lecturas que cumplieron con los filtros aplicados sobre ambos resultados de blast se clasificaron como transferencias modernas.

Transferencia Antiguas.

Caso 1: Sobre el resultado de blast del conjunto de lecturas RCJ con el genoma cloroplástico de *japonica* se aplicaron los filtros: porcentaje de identidad con el cloroplasto menor al 98% y porcentaje de alineamiento/solapamiento con el cloroplasto mayor o igual a 99%. (%ID<98%; %s>=99%). Sobre el resultado de blast del conjunto de lecturas RCJ con el genoma nuclear de *japonica* se aplicaron dos criterios de filtrados: 1) porcentaje de identidad mayor o igual a 99 % y el porcentaje de alineamiento/solapamiento mayor o igual a 98%. 2) porcentaje de identidad mayor o igual a 99 % y el porcentaje de alineamiento/solapamiento mayor o igual a 99%.

Caso 2 (considera bordes o fragmentos insertos de menor tamaño que la propia lectura):

Sobre el resultado de blast del conjunto de lecturas RIJa con el genoma cloroplástico de *japonica* se aplicaron los filtros: porcentaje de identidad con el cloroplasto menor al 98% y porcentaje de alineamiento/solapamiento con el cloroplasto menor al 90%. (%ID<98%; %s<90%). Sobre el resultado de blast del conjunto de lecturas RIJa con el genoma nuclear de *japonica* se aplicaron dos criterios de filtrados: 1) porcentaje de identidad mayor o igual a 99 % y el porcentaje de alineamiento/solapamiento mayor o igual a 98%. 2) porcentaje de identidad mayor o igual a 99 % y el porcentaje de alineamiento/solapamiento mayor o igual a 99%.

Transferencia exclusivas del biotipo maleza.

Sobre el resultado de blast del conjunto de lecturas RIJa con el genoma cloroplástico de *japonica* se aplicó el filtro: porcentaje de alineamiento/solapamiento con el cloroplasto menor al 80% (%s<80%). Sobre el resultado de blast del conjunto de lecturas RIJa con el genoma nuclear de *japonica* se aplicó: el porcentaje de alineamiento/solapamiento menor a 80%. Para las lecturas que cumplían con ambos criterios de filtrados se evalúo las características del alineamiento contra cada genoma.

Diseño de cebadores para evaluación de las transferencias de maleza: Los cebadores fueron diseñados sobre la secuencia del genoma nuclear de *O. sativa ssp japonica* y la secuencia de las lecturas a interrogar: GCFF90V02JK097 y GCFF90V02GW6GW (llamadas K097 y W6GW respectivamente), utilizando el programa Primer3. se diseñaron dos pares de cebadores por cada lectura. La secuencia de cada uno se muestra en la Tabla 2.

Lectura	Cebador	Secuencia
GCFF90V02JK097	F1	aggtgtgttgcaaggtgttc
	R1	gaacaccttgcaacacacct
	F2	aatgagggagtaacgtggca
	R2	ccagaccgcgcgaagaag
GCFF90V02GW6GW	F1	actcggaatgctgccaaga
	R1	tgagcttatgtgaaacgccg
	F2	ccagcacacgaagaacatca
	R2	tggagcttaaatctgcctgag

Tabla 2: Secuencia de cebadores diseñados para la confirmación de trasferencias entre el genoma cloroplástico y el genoma nuclear de AM356-8 y O. sativa *japonica*.

Se utilizaron tres combinaciones de los cebadores para ambos casos: F1/R1; F2/R2; F2/R1.

La reacción en cadena de la polimerasa (PCR) para las reacciones de amplificación fue: Se utilizaron 10 ul de reacción conteniendo 10 mM de primers, 5 mM desoxirribonucleótidos (dNTPs), 2,5 mM MgCl2 (*Fermentas*), 10X buffer (*Fermentas*), 100 ng ADN molde, 0,25 unidades de *Taq* (*Fermentas*) polimerasa sujeto al siguiente ciclo de PCR: 95°C por 7 minutos, 32 ciclos de 94°C por 1 minuto, 55°C por 1 minuto, 72°C por 1 minuto, 72°C por 5 minutos, hold a 4°C.

Los resultados de las amplificaciones se visualizaron por geles de agarosa 2% revelado con tinción de bromuro de etidio, se utilizaron 1 µl de 100 pb (Fermentas) como marcador de peso molecular.

Resultados y Discusión

Secuenciado con 454/Roche.

Se obtuvieron 295.159 lecturas con un largo promedio de 277 nucleótidos a partir de la secuenciación de la muestra de ADN genómico del arroz maleza AM356-8. Esto corresponde a 96 Mb de datos. Las secuencias obtenidas fueron evaluados para determinar la calidad de los mismos, buscando detectar posibles errores inherentes a la técnica de secuenciación utilizada, así como posibles contaminaciones de la muestra con ADN de otros organismos.

Calidad de los datos.

Cómo parámetro de calidad de la secuenciación se evalúo el porcentaje de lecturas repetidas artificialmente identificando 28.427 lecturas, lo que representa 9.6% del total de datos de secuencias obtenidas. Este porcentaje representa un valor bajo de duplicación si lo comparamos con los reportados por otros trabajos donde es utilizada la pirosecuenciación. En la secuenciación de un genoma antiguo de una planta extinta se obtuvo un 12% de lecturas duplicadas (Zhang et al., 2011). Mientras que en un estudio reciente realizado sobre datos de metagenoma se reporta que en datos crudos provenientes de un secuenciador 454/Roche se espera desde un 11 a un 35% de lecturas duplicadas artificialmente (Gomez-Alvarez et al., 2009).

La aparición de lecturas repetidas artificialmente es un artefacto muy común encontrado en conjuntos de secuencias obtenidas a partir de la técnica de pirosecuenciación. El artefacto se genera en el paso de PCR en emulsión, donde un mismo segmento de ADN es secuenciando más de una vez. Esta repetición se debe a que un único fragmento de ADN junto a varias perlas magnéticas quedan dentro del mismo microrreactor y cómo resultado se obtiene la secuencia de ese fragmento tantas veces como perlas magnéticas haya en el microrreactor. En muestras de ADN genómico como lo es ésta, es muy poco probable encontrar dos lecturas diferentes que comiencen en el mismo lugar dada la forma en que fue fragmentado el genoma. Este criterio se utilizó para la identificación de las lecturas repetidas artificialmente, es decir aquellas que empezaran exactamente en el mismo sitio, compartiendo las n-primeras. En este caso se compararon las primeras 50 bases de todas las lecturas obtenidas.

Las células vegetales presentan tres genomas, el nuclear, el mitocondrial y el cloroplástico como se comentó anteriormente. Dado que la extracción de ADN realizada en este trabajo no incluyó ningún paso donde los genomas extranucleares fueran removidos, se evalúo la representación de estos 3 genomas en los datos de secuenciación. Para evaluar la representación de éstos genomas en los datos de secuencias de AM356-8, se comparo el conjunto de lecturas con los genomas de referencias nuclear, cloroplástico y mitocondrial de *O. sativa ssp japonica*

Genoma nuclear.

La cobertura del genoma nuclear obtenida en este trabajo fue de 0.13X (Tabla 1). La representación del genoma nuclear se evaluó aplicando técnicas de mapeo contra genoma de referencia, en este caso el genoma nuclear de *O. sativa ssp japonica* cv *Nipponbare* utilizando el programa Newbler (Roche).

El tamaño del genoma nuclear del arroz es de 400 Mb, dado que en este trabajo se utilizó un cuarto de placa de 454/Roche para la secuenciación, la cobertura esperada era de 0.25X, pues la cantidad de datos de secuencia esperado es aproximadamente 100 Mb (por cada placa de secuenciación de 454/Roche se obtiene entre 350-500 Mb). Sin embargo luego de realizado el mapeo fue menor.

Coberturas bajas de genomas nucleares del orden de 0.1X, como la obtenida en este trabajo, se han utilizado como estrategia para la obtención de secuencias GSS en especies no modelos (Rasmussen and Noor, 2009) y para el desarrollo de marcadores moleculares en las mismas. Dado el gran avance que han tenido en los últimos 5 años las tecnologías de secuenciación, lo que ha generado una reducción en el costo por base secuenciada en más de cien mil veces (Lander, 2011), abordajes del estilo del trabajo de Rasmussen et al. (2009) no son aplicados hoy en día.

Genomas de organelos.

La representación de los genomas de los organelos se evalúo con el programa BLAST (Altschul et al., 1990), opción blastn.

En la comparación con el genoma mitocondrial *O. sativa ssp japonica* se identificaron 17.003 lecturas con similaridad a dicho genoma, esto es el 5.7 % de los datos totales de secuencia. La cobertura obtenida con estos datos de secuencia para el genoma mitocondrial del arroz de 490 Kb de tamaño (Notsu et al., 2002) fue de 10X (Tabla 3)

Para el genoma cloroplástico de *O. sativa ssp japonica* (Tang et al., 2004), se identificaron 47.817 lecturas de AM356-8 con similaridad significativa al cloroplasto de referencia. Esta cantidad de lecturas representa el 16% del total de los datos lo que arroja una cobertura de 106X (Tabla 1), considerando que el tamaño de este genoma es de 134.551 pb.

La mayor representación del genoma del cloroplasto en estos datos de secuencias, puede deberse a la combinación de varios factores. En primer lugar, la abundancia de estos organelos en hojas jóvenes, tejido del que se obtuvo la muestra para la secuenciación. or cada célula vegetal existen múltiples copias del genoma cloroplástico, en células de hojas puede haber entre 400 a 1600 copias de este genoma, dependiendo de la especie (Pyke, 1999). Estos valores implicarían para el caso de arroz que el porcentaje de ADN de origen cloroplástico vaya desde 11% a 35%. Un segundo elemento que contribuye en forma significativa a incrementar el número de lecturas con similitud al genoma cloroplástico es que el genoma nuclear contiene varios

segmentos (que representan una parte importante del mismo), que son de origen cloroplástico. Esto es debido a transferencias entre ambos genomas relativamente recientes. Como veremos más adelante el estudio de estas transferencias representan un capítulo importante de esta tesis.

Comparación con los genomas nuclear, mitocondrial y cloroplástico O. sativa *japonica*	
N° lecturas totales	295.159
N° lecturas incluidos	295.092
N° lecturas mapeados contra núcleo	177.920
Cobertura	0.13X
N° lecturas mapeados contra mitocondria	17003
Cobertura	10X
N° lecturas mapeados contra cloroplasto	47817
Cobertura	106X

Tabla 3: Resumen de resultados obtenidos en la comparación con los 3 genomas de la célula vegetal: Nuclear (mapeo con Newbler), mitocondrial y cloroplástico (Blast en ambos organelos) de las secuencias de AM356-8.

Una vez evaluada la representación de los 3 genomas de la célula vegetal en los datos de secuenciación y habiendo considerado la cobertura obtenida y las características de cada genoma se continuaron los análisis con el genoma del cloroplasto.

Identificación y clasificación de secuencias cloroplásticas.

La disponibilidad de datos de secuencias cloroplásticas de la misma especie que el biotipo secuenciado, como es el caso de las dos subespecies de *O. sativa* y de los dos posibles ancestros de ésta, *O. nivara* y *O. rufipogon*, hizo que se planteara una estrategia de genómica comparativa para la obtención y estudio de secuencias cloroplásticas a partir del conjunto de secuencias de AM356-8. La estrategia utilizada se esquematiza en las figuras 5 y 6.

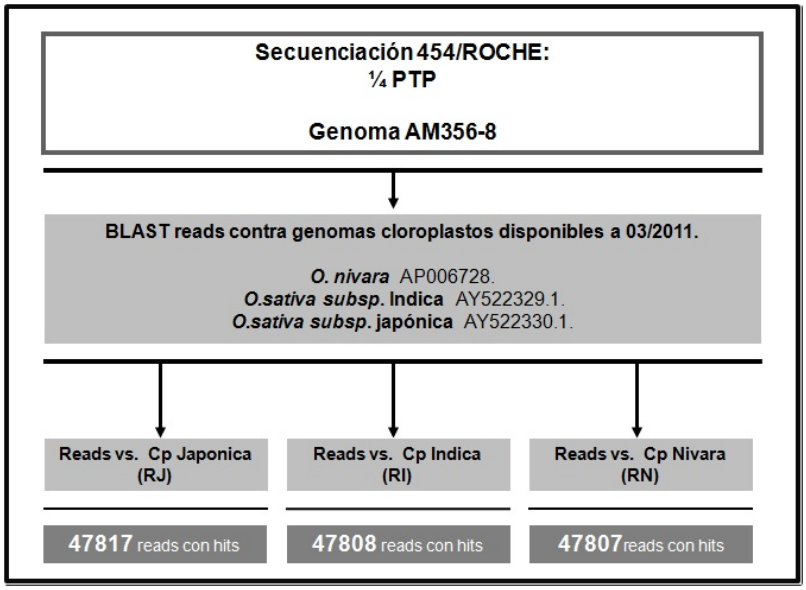

Figura 5: Estrategia utilizada para la identificación de las lecturas cloroplásticos de AM356-8.

Por comparación de las 295.159 lecturas de secuencia con los tres genomas cloroplásticos disponibles públicamente a marzo 2011, cloroplasto *O. sativa ssp. indica* (AY522329.1), cloroplasto *O. sativa ssp. japónica* (AY522330.1), cloroplasto de *O. nivara* (AP006728) (de aquí en más referidas como *indica* , *japonica* y *nivara* respectivamente) se identificaron en promedio . 47.800 lecturas con similaridad a los genomas cloroplásticos de referencia (Figura 5), esto representa el 16.2 % del total de los datos como había sido comentado anteriormente.

Los tres conjuntos de lecturas a los que llamamos RJ, RI, RN (Figura 5), se compararon entre sí para determinar cuántas lecturas diferentes existían entre éstos conjuntos. En la comparación de los conjuntos RJ y RI se identificaron 15 lecturas de diferencia, mientras que contra el conjunto de lecturas de RN fueron identificados 18 lecturas diferentes. Cuando se compararon los conjuntos RI y RN sólo se encontraron 3 lecturas de diferencia entre los dos conjuntos.

Las lecturas de cada uno de los tres conjuntos de secuencias obtenidos a partir de las comparaciones contra los genomas de referencia correspondientes (RJ, RI, RN), se clasificaron según las características del alineamiento de cada lectura contra cada genoma cloroplástico. Para ello, sobre los resultados de blastn independientemente se aplicaron filtros para descartar

lecturas que generan ruido en los análisis siguientes. El primer filtro que se aplicó fue sobre el largo del alineamiento, aquellas lecturas que presentaron un largo de alineamiento con el genoma cloroplástico menor a 100 nucleótidos no fueron considerados. De esta forma se generaron tres nuevos conjuntos de lecturas llamados R_100 (RJ_100, RI_100, RN_100, Figura 2), eliminando así, lecturas muy cortas (menores a 100 nucleótidos de largo) y/o lecturas que sólo alineaban en un fragmento del mismo menor a 100 nucleótidos, disminuyendo los errores provenientes de la secuenciación. En Tabla 4 se resumen las cantidades de lecturas de cada uno de los subconjuntos obtenidos hasta el momento.

Es necesario tener en cuenta que el conjunto de lecturas de AM356-8 usado en este trabajo es una mezcla de secuencias provenientes de los 3 genomas de la célula vegetal, cloroplasto, mitocondria y núcleo y que existen transferencias de segmentos de ADN entre éstos genomas (Ward et al., 1981; Huang et al., 2005; Noutsos et al., 2007; Bock and Timmis, 2008; Alverson et al., 2010). Esto implica que las lecturas del biotipo secuenciado representan una combinación compleja de material hereditario, siendo muchas veces difícil determinar el origen de algunas de éstas lecturas. Como criterio adicional de filtrado se consideró el siguiente aspecto, una vez transferidos los segmentos de ADN plástico al genoma nuclear, éstos sufren una muy rápida fragmentación, proceso mucho más veloz que la acumulación de mutaciones puntuales, como se discute más adelante en este trabajo. Teniendo en cuenta éstas características del proceso de evolución de los fragmentos insertos, se aplicó como filtro adicional que el porcentaje de solapamiento (%s) de la lectura sobre el genoma del cloroplasto fuera mayor a un determinado umbral, dado que solapamientos menores indicarían el origen nuclear del fragmento. Este porcentaje se calculó como el largo del alineamiento de la lectura contra el genoma sobre el largo de la lectura (largo aln / largo lectura). El filtro de %s se aplicó sobre el resultado de blastn de los conjuntos de lecturas RJ_100, RI_100 y RN_100 (Figura 6), siendo seleccionadas aquellas con un alineamiento casi completo contra el genoma del cloroplasto, esto es, lecturas que presentaron un porcentaje de solapamiento con el cloroplasto superior o igual a 99%. A este conjunto se lo llamó lecturas completas (RC). Por otra parte las lecturas que presentaron un %s menor a 99% se agruparon en el conjunto de lecturas incompletas (RI) (Figura 6) y tal como veremos más adelante estas últimas son informativas sobre los procesos de transferencia de segmentos genómicos entre plastos y núcleo.

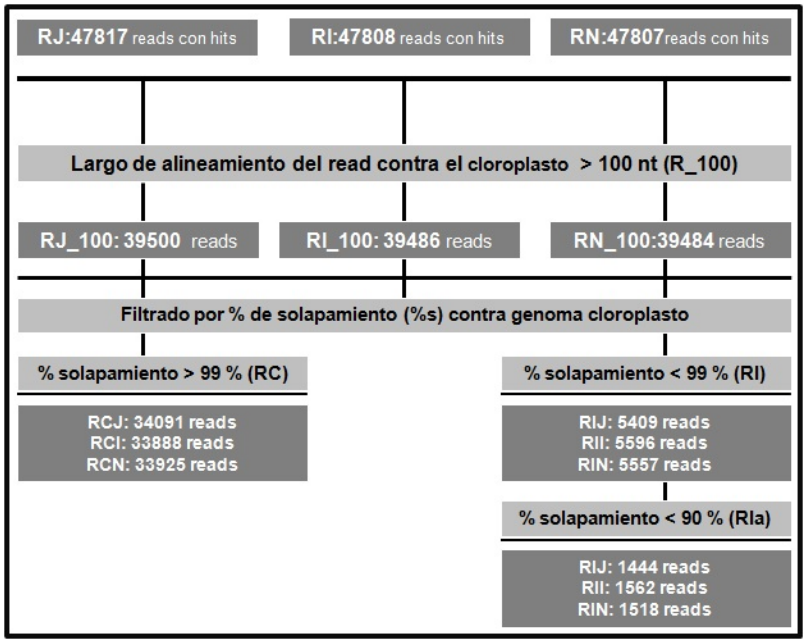

Figura 6: Estrategia utilizada para la identificación de las lecturas cloroplásticos de AM356-8; criterios de filtrado.

Las lecturas de los conjuntos RC fueron las consideradas como de origen cloroplásticos (a pesar que este grupo también puede contener lecturas de origen nuclear indistinguible de una cloroplástica), mientras que en los conjuntos RI habría una mezcla de lecturas de diferente origen o generados por errores en la secuenciación. Más adelante veremos con mayor detalle estos aspectos.

Genomas cloroplasto	***Japonica* (J)**	***Indica* (I)**	***Nivara* (N)**
Read BLAST (R)	47817	47807	47808
Read alineamiento > 100 nucleótidos (R_100)	39500	39486	39484
Lecturas Completos (RC)	34091	33888	33925
Lecturas Incompletos (RI)	1444	1562	1518

Tabla 4: Resumen de conjuntos obtenidos a partir de la aplicación de filtros sobre porcentaje de solapamiento para cada uno de los genomas cloroplásticos considerados

Ensamblado "de novo" del cloroplasto de AM356-8.

Las 34091 lecturas pertenecientes al conjunto de lecturas que alineó completo (99% de solapamiento) con el genoma del cloroplasto de *japonica* (RCJ), fueron utilizadas para realizar el

ensamblado de novo del genoma cloroplástico de AM356-8 utilizando el programa Newbler. Cómo resultado del ensamblado se obtuvieron 2 contigs, cuyas secuencias suman 114 kb, representando el 85% del genoma del cloroplasto. Así mismo, sólo el contig más largo obtenido en este ensamblado es de 101.363 pb lo que representa el 75% del genoma del cloroplasto. En la Tabla 5 se detallan los resultados del ensablado de novo obtenido.

Información de Ensamblado	
N° lecturas ensamblados:	34063
N° contigs:	2
Tamaño ensamblado	114kb
N50	101.363
Contigs más largo	101.363

Tabla 5: Detalle del ensamblado de novo del cloroplastos de 356-8 con Newbler.

Tendiendo en encuenta la estructura consevada de los genoma cloroplásticos, la cual se divide en cuatro regiones, región larga de simple copia (LSC), región corta de simple copia (SSC), y 2 regiones invertidas repetidas (IR) como se describió anteriormente y tomando en cuenta las caracteristicas de las secuencias de las IR, se consideró que en estos dos contigs estaba contenido el genoma del cloroplasto de AM356-8. El programa Newbler colapsa las lecturas provenientes de secuencias repetidas, por tanto las dos regiones IR fueron colapsadas en solo una, lo que determina la ausencia de un segmento de 20 Kb en nuestro ensamblado. Para confirmar si en estos dos contigs se contenía el genoma cloroplastico, éstos se alinearon con el genoma del cloroplasto de *japonica* mediante blastn. El alineamiento luego se visualizó con el programa ACT (Artemis Comparision Tool)(Carver et al., 2005). (Figura 7).

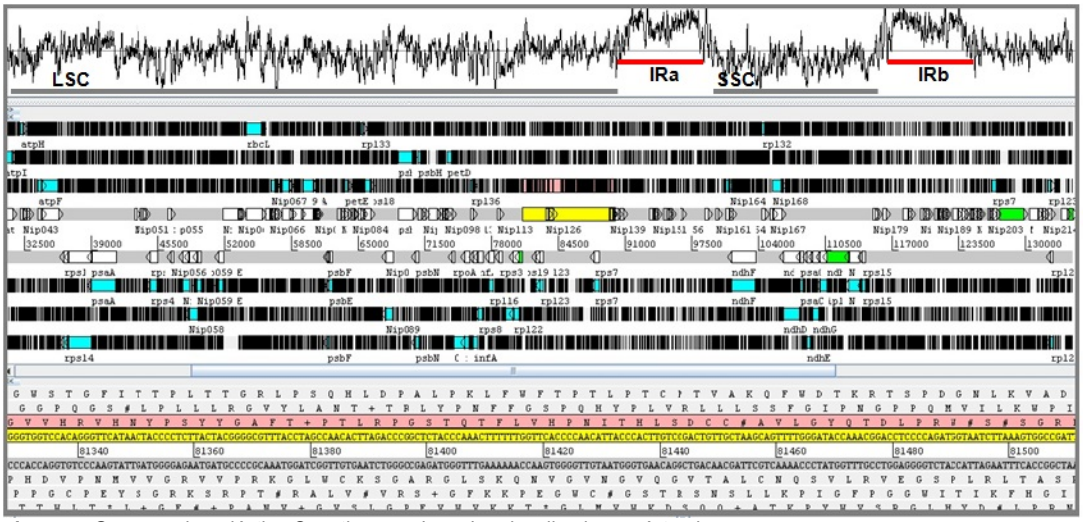

Figura 7: Genoma cloroplástico O. sativa ssp *japonica* visualizado con Artemis

Cómo se oberva en el resultado del alineamiento contra las regiones del cloroplasto de *japonica* en la figura 8, en el contig 1 está contenida la región larga de simple copia (LSC) y la región

invertida repetida. Ésta última sólo aparece una vez en los contigs debido a que fue colapsada por las caracteristicas de su secuencia como se explicó antes. En el contig 2 está contenida la región corta de simple copia.

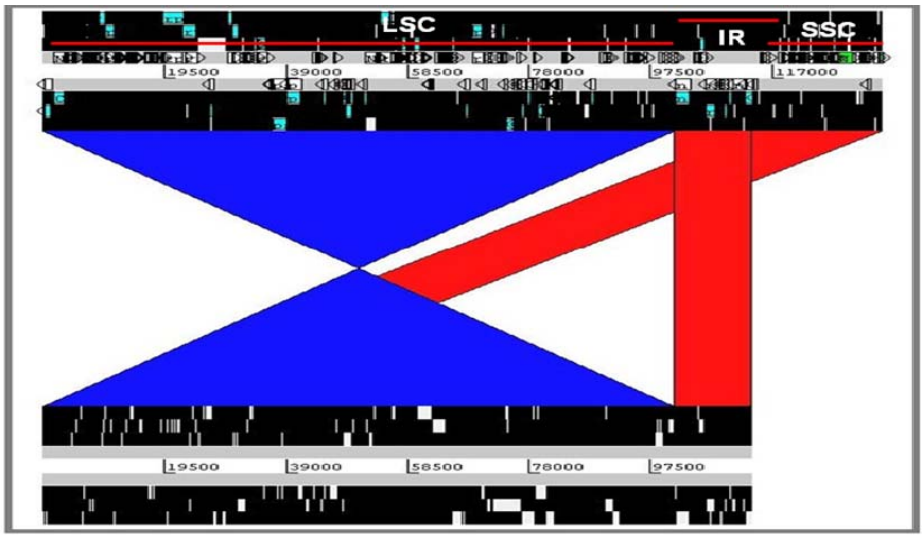

Figura 8: Alineamiento del genoma cloroplasto de *O. sativa japonica* y los 2 contigs obtenidos en el ensamblado de novo del cloroplasto de AM356-8, visualizado en ACT.

De igual manera se realizaron los ensamblados de novo con los conjuntos de lecturas que mapearon en forma completa con el genoma del cloroplasto de *indica* (RCI) y de *nivara* (RCN). Los resultados obtenidos de estos ensamblados fueron semejantes, aunque el ensamblado de mayor calidad se obtuvo a partir de RCJ. Para la comparación entre los tres ensamblajes se tomó al estadístico N50, (definido como la longitud de los contigs, tal que sumando el largo de contigs de igual o mayor tamaño se obtiene la mitad de las bases del genoma). El valor de N50 obtenido para el ensamblado con el conjunto RCN fue de 92.754 bases, mientras que el valor de N50 para el conjunto RCI fue de 83.599 bases, ambos menores al obtenido con el ensamblado del conjunto RCJ (ver Tabla 4).

En relación a esta sección, un último aspecto a resaltar es que incluso trabajando con un genoma pequeño de 134 kb, como lo es el genoma del cloroplasto de *O. sativa*, con una muy alta cobertura no fue posible obtener el genoma del cloroplasto de AM356-8 ensamblado completamente. Esto pone de manifiesto la complejidad intrínseca del proceso de obtención de un genoma totalmente ensamblado. A partir de estos resultados es posible considerar que en el conjunto de lecturas RCJ está contenido el genoma del cloroplasto del biotipo de arroz maleza. Para comprender mejor las causas que impiden obtener este genoma completamente ensamblado se realizó el mapeo de lecturas procurando evaluar la cobertura del genoma del cloroplasto de *japonica* por el conjunto de lecturas RCJ.

Mapeo contra genoma de referencia.

De manera de evaluar el nivel de cobertura de las distintas regiones del cloroplasto, una vez más se tomaron los conjuntos de lecturas que alinearon de forma completa con cada uno de los genomas cloroplásticos (RCJ, RCI, RCN) y fueron mapeados contra el genoma de referencia correspondiente.

En el caso de del mapeo de las lecturas del conjunto RCJ contra el genoma de cloroplasto de *japonica* fueron mapeados el 97.8% de las lecturas, lo cual implica que una cobertura de 81X. En los otros dos casos donde las referencias fueron el genoma cloroplástico de *indica* y de *nivara* respectivamente, los resultados fueron similares 97.9% de las lecturas fueron mapeadas obteniendo una cobertura de 80X en el mapeo contra el cloroplasto de *indica* y el 98% de las lecturas obteniendo una cobertura promedio igual a la anterior para el caso del genoma del cloroplasto de *nivara*. En Tabla 6 se muestran los detalles de cada uno de los mapeos.

El mapeo se realizó utilizando el programa MIRA3, el cual toma a el genoma de referencia como una lectura más y realiza un ensamblado. Para ello aplica una estrategia de múltiples pasos donde toma como base para agrupar las lecturas, regiones del alta confianza. Si es necesario vuelve sobre regiones de menor confianza para completar el ensamblado. MIRA3 no enmascara las regiones repetidas, ya que posee un editor automático que le permite analizar los alineamientos en profundidad, esta característica es una ventaja frente a otros programas ya que permite evaluar la representación en los datos de secuenciación de regiones repetidas como son las regiones invertidas repetidas de los cloroplastos.

MAPEO contra genoma de referencia

Genoma de referencia	*japonica*	*Indica*	*Nivara*
Lecturas Totales	34091	33888	33925
Lecturas Mapeados	33369	33208	33925
Lecturas no mapeados	21	21	21
Cobertura promedio	81x	80x	80x

Tabla 6: Detalle de los mapeos contra los genomas de referencia, cloroplasto *O.sativa ssp indica* y *ssp japonica* y *O. nivara* realizados con el programa MIRA3

Los tres mapeos realizados mostraron una alta cobertura de los genomas de referencia con los que se compararon los conjuntos de lecturas correspondientes. En la figura 9 se muestra el resultado del mapeo de las lecturas contra el genoma de *japonica* a modo de ejemplo.

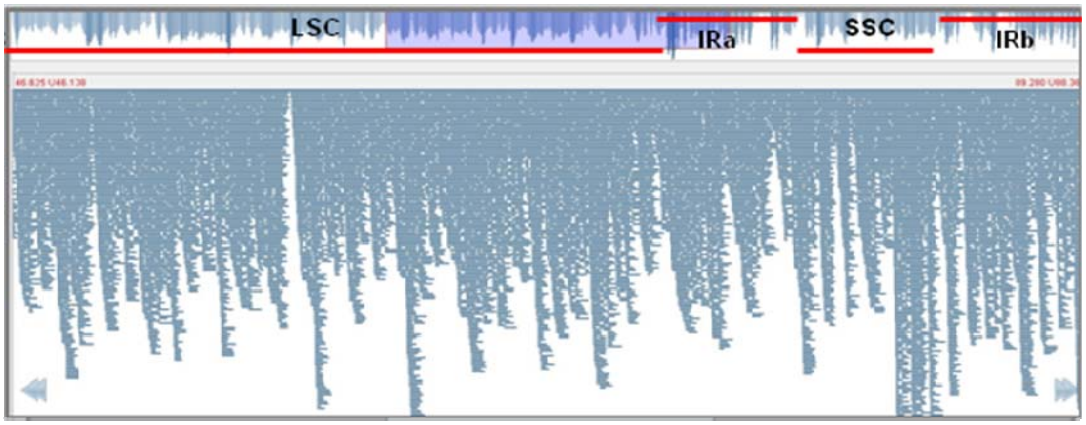

Figura 9: Resultado del mapeo con el programa Mira3 de RCJ contra el genoma de O. sativa ssp *japonica* visualizado en Tablet

Como se puede apreciar estos mapeos confirman que en el conjunto RCJ contiene el 100 % del genoma del cloroplasto de AM356-8 y no es ensamblado en un único contig debido a las características de la secuencia de las regiones repetidas invertidas. Sería necesario utilizar otra estrategia de secuenciación, tal como paired end para poder ensamblar el 100% del genoma del cloroplasto. La opción de secuenciación paired-end es una simple modificación en la preparación de las librerías de secuenciación que permite hacer la lectura de las dos hebras, hebra molde (forward) y la complementaria (reverse). Además de la información de secuencia está opción de secuenciado da información de posición ya que se conoce la distancia entre cada extremos del fragmento secuenciado.

Es posible afirmar que a través de este trabajo se obtuvo el primer genoma cloroplástico de un biotipo de arroz maleza, el cual podrá ser útil para eventuales estudios posteriores, algunos de los cuales se presentan a continuación.

Búsqueda de regiones divergentes entre los genomas de cloroplastos.

En este trabajo se compararon 4 de los genomas cloroplásticos del género *Oryza*: *O. sativa ssp japonica*, *O. sativa ssp indica*, *O. nivara*, *O. rufipogon* para la identificación de variaciones filogenéticamente informativa adicionales a las que descritas. Para este fin los cuatro genomas fueron alineados utilizando el paquete de herramientas para genómica comparativa r-vista (whole genome VISTA) (Zambon et al., 2005). Una vez obtenido el alineamiento de los 4 genomas, éste fue recorrido en ventanas de 600 bases, en búsqueda de sitios variables de una única base (SNP) y/o inserciones/deleciones (indels). En la figura 10 se muestra en forma esquemática la estrategia seguida.

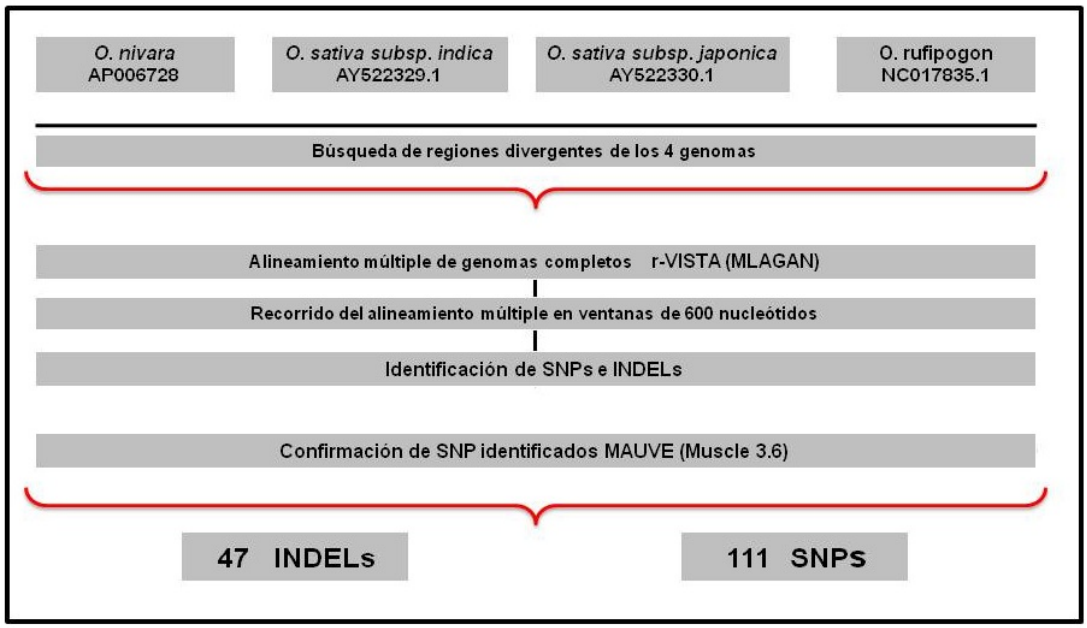

Figura 10: Esquema estrategia de búsqueda de regiones divergentes entre los 4 genomas del género Oryza comparados**.**

De esta manera se identificaron 64 SNPs y 47 INDELs (Tablas 1 y 2 anexo). Los SNPs hallados se confirmaron utilizando otro programa de alineamiento de genomas completos MAUVE (Darling et al., 2004), en este programa los genomas completos son alienados utilizando el algoritmo Muscle 3.6. Como archivos de salida además de alineamiento el programa reporta las regiones conservadas entre los genomas comparados, así como las regiones que presentan diferencias de una única base entre los genomas. Para cada sitio polimórfico identificado (SNP) en el alineamiento, el programa MAUVE registra en el archivo de SNP los nucleótidos presentes en cada genoma en ese sitio, junto con las coordenadas del lugar en cada genoma. De esta manera se obtuvo un listado de las variantes tipo SNP de manera automática. Este listado contenía 111 SNP, 47 más que los identificados en el alineamiento del r-Vista. Se realizó un nuevo recorrido de los genomas alineados por r-Vista utilizando como guía la lista de SNPs reportados por MAUVE para la confirmación de los mismos.

Los 111 SNP identificados automáticamente por MAUVE fueron confirmados en el recorrido por ventanas del alineamiento de r-Vista. (TABLA 1anexo).

El 57% de las variantes tipo SNP se encontraron en regiones codificantes de proteínas y de tRNAs, anotadas en al menos uno de los genomas comparados. La mayor densidad de SNP se observó principalmente en dos regiones coincidentes con las regiones de copia única, LSC y SSC separadas por una región donde no se identificó ningún SNPs, la que coincide con una de las IR (Figura 11). Los SNP identificados se encuentran hasta las coordenadas 109.000, mientras que el tamaño promedio de estos cuatro genomas es 134.000 bases. Anteriormente se determinó la ubicación de las IR utilizando el programa Artemis (Rutherford et al., 2000) (Figura 4), la región IRa comienza alrededor del nucleótido 89.900 y se extiende hasta la base 98.980

aproximadamente. La segunda región invertida repetida, IRb, comienza alrededor de la coordenada 116.214 y se extiende hasta la base 125.192. Por tanto podemos confirmar que los SNP identificados en la comparación realizada se ubican en las regiones de copia única.

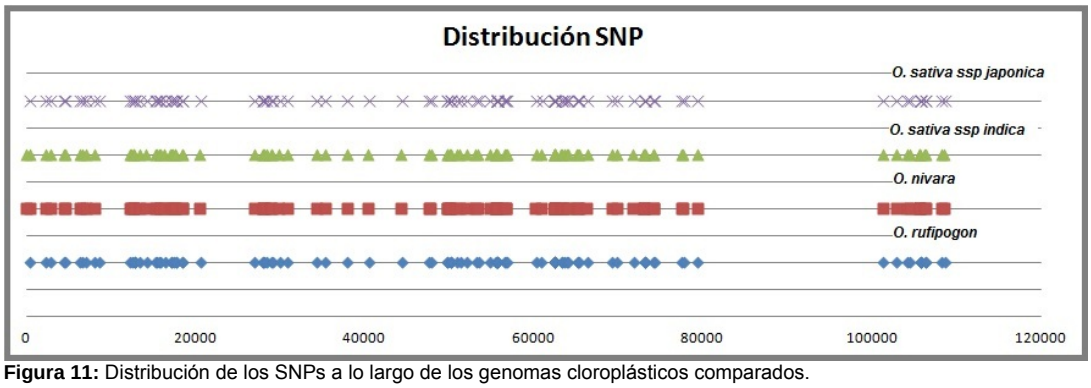

Figura 11: Distribución de los SNPs a lo largo de los genomas cloroplásticos comparados.

Los resultados obtenidos en cuanto a la distribución espacial de las variantes tipo SNP son compatibles con las tasas de divergencia reportadas para las diferentes regiones del genoma cloroplástico. Las secuencias ubicadas en las regiones IR divergen más lentamente que las secuencias LSC/SSC (Wolfe et al., 1987). La diferencia en la tasa de sustitución entre las IR y LSC/SSC puede ser atribuida a la reducida tasa mutacional de las IR o a mayor restricción funcional (Shahid Masood et al., 2004b).

Los 47 indel identificados en la comparación entre los cuatro genomas cloroplásticos mostraron una distribución particular. Se observaron 14 indels exclusivos del genoma del cloroplasto de *japonica*, otros 12 exclusivos del genoma del cloroplasto de *rufipogon* y tres más compartidos por ambos genomas. Para los genomas de los cloroplastos de *indica* y *nivara* se identificaron 7 y 9 indels exclusivos respectivamente y entre estos genomas se identificaron tres indels compartidos. No se observaron indels compartidos entre las subespecies de *sativa*, entre *indica* y *rufipogon*, ni tampoco entre *japonica* y *nivara* (Tabla 2 anexo).

Por tanto esto agruparía a *nivara* con *indica* y a *japonica* con *rufipogon*. Este agrupamiento observado, coincide con lo reportado Huang et al (2005), donde fueron comparados los genomas completos de los cloroplastos de *indica*, *japonica* y *nivara*, identificaron las substituciones sinónimas y no sinónimas reportando una divergencia del genoma del plasto de *japonica* anterior a la divergencia entre *indica* y *nivara* (Figura 12).

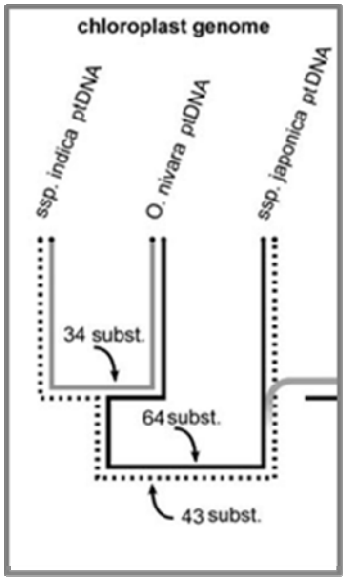

Figura 12: Esquema de la historia evolutiva de la separación entre los cloroplastos de *O. nivara*, *O. sativa indica* y *O. sativa japonica* (Huang et al., 2005)

La distribución espacial presentada por los 47 indels en cada uno de los genomas fue semejante a la mostrada por los SNPs, se observan dos regiones donde se concentran la mayor cantidad de indels. La mayoría de ellos se ubican en la región de copia única larga (LSC), dos indel en la región SSC y sólo uno en la IRa. Del total de indels identificados 23 se encontraban en regiones intergénicas y 24 en regiones codificantes. De éstos, el 42 % presentaron un motivo variante de un único nucleótido, el resto presentó motivos que van desde dinucleótidos hasta variantes más largas como indels de 32 y 69 pares de bases. Es de destacar que todos los indels de un único nucleótido identificados fueron variaciones de adenina o timina ubicados en las regiones de copia única. Si bien estos resultados estarían de acuerdo con lo reportado por otros estudios, donde encuentran que el contenido en GC de las regiones repetidas está muy por encima del promedio del resto del genoma cloroplástico (Tang et al., 2004), sugiriendo que puede existir un sesgo en el sistema de replicación y reparación del ADN del cloroplasto que genere la aparición de variantes de secuencias más frecuentemente en regiones de bajo contenido en GC como las regiones de copia única. Se debe tener en cuenta que al tratarse de variantes de un único nucleótido encontradas en la comparación de los 4 genomas cloroplásticos públicos, y no contando con los archivos de secuencia y calidad originales con los que éstos genomas fueron ensamblados, éste tipo de variante no serán consideradas por poderse tratar de errores de secuenciación.

Dos de los 47 indels identificados en este trabajo, habían sido reportados en trabajos anteriores. La región que llamamos 8k, que presenta una inserción de 69 pb en el genoma de *japonica* y *rufipogon* (Kanno and Hirai, 1993) y la inserción de 32 pb contenida en la región 18_b (ver Tabla 2 anexo) en el genoma cloroplástico de *indica* reportado en el trabajo de Tang et al (2004). En particular el indel identificado en la región 8k (inserción presente en *japonica* y *rufipogon*), define

dos haplotipos. Uno de éstos haplotipos compartido por *indica* y *nivara* (sin inserción) y el otro haplotipo ya mencionado compartido por *japonica* y *rufipogon*. Ambos haplotipos están ligados a cada una de las subespecies de *O. sativa* respectivamente (Chen et al., 1993; Kanno and Hirai, 1993; Tang et al., 2004).

En un trabajo previo donde se analizó la variabilidad intrapoblacional de *O*. *rufipogon* mediante la determinación de la presencia o ausencia de este indel de 69 pb se observaron los dos haplotipos, con inserción y sin inserción. Además se observó una muy alta correlación de estos dos haplotipos con biotipos que presentan ciclos de vida bien diferenciados y adaptados a diferentes ambientes. Específicamente los biotipos sin la inserción de 69 pb eran de ciclo anual (biotipos llamado por algunos autores como O. *nivara*), mientras que los perennes (O. *rufipogon*) presentaban el haplotipo con la inserción. Considerando que *O*. *rufipogon* es el ancestro más probable de *O. sativa*, estos resultados estarían apoyando la hipótesis de que en el genoma del ancestro ya existía una pre-diferenciación en dos biotipos. Biotipos que luego darían lugar a las dos subesepcies mayoritarias de *O. sativa* (Chen et al., 1993) a partir de dos procesos de domesticación independientes (Garris et al., 2005; Londo et al., 2006). Los resultados obtenidos en este trabajo coinciden con lo reportado por Chen et al (1993). Una interpretación alternativa es que la presencia de este segmento es ancestral a todas las variantes de arroz domesticado y que el mismo se perdió en el linaje que conduce a *indica* y *nivara* luego de que ambas se separaron del linaje que conduce a *O. sativa japonica*.

Las regiones conteniendo los sitios con variantes del tipo indels de al menos 2 nucleótidos de largo fueron tomadas para su comparación con las secuencias de AM356-8 con el objetivo de determinar el relacionamiento filogenético del biotipo que estamos analizando en este trabajo. En la figura 13 se muestra un esquema de la metodología seguida para la comparación de los indels identificados contra las lecturas de AM356-8.

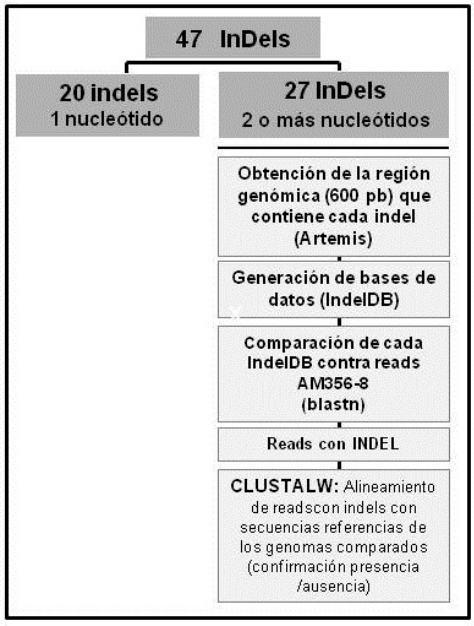

Figura 13: Metodología para la comparación de indels obtenidos en la comparación de los 4 genomas públicos con las lecturas de AM456-8

Las regiones genómicas incluidas en este análisis fueron extraídas a partir de los archivos gbk (genebank) de cada uno de los genomas cloroplásticos de referencia utilizando el programa Artemis (Rutherford et al., 2000). Las secuencias de cada región conteniendo el indel identificado, se tomaron como base de datos y se compararon mediante blastn con el conjunto de secuencias de AM356-8. El análisis fue realizado para cada una de las regiones con indel por separado. Aquellas lecturas que presentaron un porcentaje de identidad mayor a 90% en el alineamiento sobre la región de interés fueron separadas para su posterior análisis. En concreto, estas lecturas fueron luego alineados junto con las secuencias de referencias a través de ClustalW. En la figura 14 se muestra a modo de ejemplo el alineamiento de las lecturas extraídas desde el resultado de blastn contra el indel localizado en la región 8k (ver Tabla 7), la cual contiene una inserción que está presente en *O. sativa. japonica* y en las lecturas provenientes AM356-8 . De esta manera se determinó la presencia/ausencia de cada indel interrogado en las secuencias del cloroplasto de AM356-8. El detalle de los indels y los resultados de la comparación correspondiente se resume en Tabla 2 de anexo.

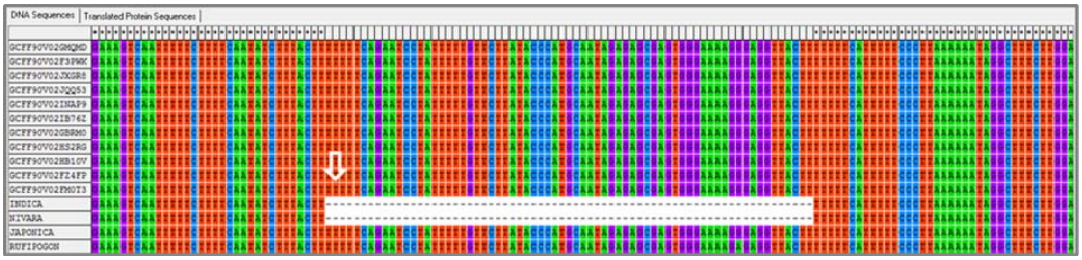

Figura 14: Alineamiento de lecturas de AM356-8 con secuencias referencias de región 8 kb**.**

Los 27 indels comparados con las secuencias de AM356-8 corresponden, 4 indels únicos del genoma de *indica*, 5 al genoma de *nivara* y 3 a ambos; 8 indels correspondían al genoma de *japonica*, 4 al de *rufipogon* y 3 a ambos. Los resultados obtenidos para los indels de dos y tres bases fueron en algunos casos ambiguos no permitiendo concluir su presencia o ausencia en las secuencias de la maleza. Sin embargo, en las comparaciones de regiones con motivos variantes de mayor largo fueron determinados claramente. En la figura 15 se muestra a modo de ejemplo, el alineamiento de lecturas conteniendo una variante del tipo indel de 2 bases en la que no se puede discriminar si es un errores de secuenciación o una variante existente en *japonica* y en el biotipo maleza. Este tipo de secuencias ricas en homopolímero son regiones de difícil resolución, siendo uno de los más comunes errores de la técnica de pirosecuenciación (Huse and Welch, 2011). Regiones con repetidos de un único nucleótido como el que se observa son interesantes para ser utilizados en estudios poblacionales, ya que estos han sido reportadas como entre las regiones de los genomas cloroplásticos que presentan mayor variabilidad(Diekmann et al., 2012). Para la resolución de este tipo de este problema sería necesario resecuenciar esta región con la metodología Sanger por ejemplo, para determinar si se trata de un polimorfismo o un artefacto de la técnica.

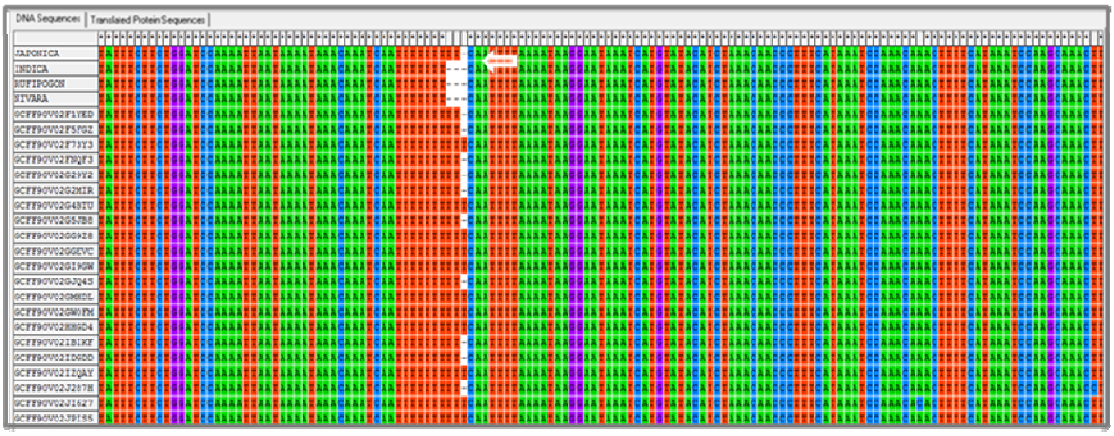

Figura 15: Alineamiento de secuencias referencias conteniendo un indel de mayor complejidad en la resolución y las lecturas de AM3536-8 extraídas a partir del resultado de blast.

En la comparación de los 27 indels con las secuencias de AM356-8 se observaron sólo 8 presentes en el biotipo de arroz secuenciado. De estos ocho, cinco eran indels identificados como exclusivos del genoma de *O. sativa ssp japonica* y los otros tres indels eran los compartidos por esta especie con *O. rufipogon* (Tabla 7).

Región INDEL	Genoma variante	AM-356-8 (n° lecturas)	Largo	Secuencia Variable	Anotación
8 kb	*japonica*/ *Rufipogon*	23	69	GAATCCTATTTTTGTTCTT ATACCCATGCAATAGAGA GGAGTGGGAAAAGGGAG GTTACTTTTTTTCA	ORF predicho sin anotación
12kb	*japonica*	14	4	AGGG	Intergénica
14kb	*japonica*	1		AC	Intergénica
46kb	*japonica*	8	5	TATAT	Intergénica
57kb	*japonica*/ *Rufipogon*	10	16	TTTTTTAGAATACTAA	Intergénica
60kb	*japonica*	32	5	D: TATTG	Intergénica
65kb	*japonica*	23	2	TT	Intergénica
77kb	*japonica*/ *Rufipogon*	43	3	D: TGG	Intergénica

Tabla 7: Resumen de resultados obtenidos en la comparación entre los indels identificados entre los 4 genomas de referencia y las secuencias de AM356-8. En Tabla se muestra los indels para los cuales se identificaron lecturas que los contenían, las regiones donde se ubicaron los INDELs en los genomas, y los genoma cloroplástico con la variante.

Las regiones del genoma cloroplástico con indels de las especies comparadas en este trabajo e identificados como presentes en el genoma del cloroplasto del biotipo de arroz maleza secuenciado, muestran una clara similitud de éste con el cloroplasto de la subespecie *japonica*.

Los indels identificados en este trabajo serán de utilidad tanto para estudios poblacionales dentro del género *Oryza*, así como para trazar el origen de una población de arroz maleza. Estos indels representan haplotipos que diferencian claramente entre las subespecies de *O. sativa*, y así mismo entre haplotipos de *nivara* y *rufipogon*. Aplicarlos en un estudio poblacional del complejo *Oryza* podría ayudar a resolver las relaciones filogenéticas de éstas especies aún en discusión, así como trazar el origen de los biotipos de arroz maleza encontrados en campos uruguayos.

Las hipótesis planteadas sobre los posibles orígenes del arroz maleza en sitios donde no hay especies silvestres emparentadas, como lo es nuestro país, plantean que éstos biotipos pueden haber sido introducidos junto con los cultivares en el comienzo del cultivo de arroz y que por procesos sucesivos de selección y re-hibridación se originaron los diversos biotipos que hoy existen en el país.

Considerando la historia del cultivo de arroz en Uruguay, la que puede dividirse en dos etapas: desde el comienzo y hasta la década del 80 donde las variedades plantadas eran en su totalidad del tipo *japonica* (primeras variedades italianas y brasilera, Bluebelle introducida desde EEUU más tarde), y actualmente donde la variedad mayoritariamente plantada es El Paso 144 de tipo *indica*. Sumada a la historia de cultivo del campo donde fue hallado el biotipo, campo donde se planta la subespecie *indica* (variedad INIA Olimar) no Clearfield, las características morfológicas de la planta del biotipo en estudio muy semejantes al fenotipo silvestre (pericarpio y cascara de color negro, más alta que las cultivadas y gran número de macollos) y la caracterización del cloroplasto cómo *japonica* realizada en este trabajo indicarían que este biotipo tuvo origen por

cruzamiento entre una maleza y un cultivar apoyando la hipótesis antes mencionada. Además, la caracterización del cloroplasto nos permite afirmar que en el cruzamiento que dio origen al biotipo secuenciado, el arroz maleza habría actuado como padre y un cultivar *japonica* como madre. Debido a que la variedad plantada es *indica* y no ha habido presión de selección por herbicida, es de esperar que el biotipo provenga de otros campos.

Identificación de trasferencias desde genoma cloroplástico hacia el genoma nuclear.

Las transferencias de fragmentos de ADN entre los genomas de los organelos citoplasmáticos y el genoma nuclear ha sido reportado por diferentes trabajos (Stupar et al., 2001b; Yu et al., 2003; Richly and Leister, 2004a, 2004b; Huang et al., 2005; Matsuo et al., 2005). Basados en éstos trabajos, se planteó una estrategia para la identificación de regiones transferidas entre el genoma cloroplástico y el genoma nuclear de *O. sativa spp japonica*. Siendo de particular interés la detección de dos tipos de regiones transferidas, las que en este trabajo llamamos contemporáneas, es decir que ocurrieron luego que la maleza se separó del arroz cultivado y aquellas regiones que estarían presentes en el genoma nuclear de *japonica*, pero no aparecen en el ensamblado del mismo.

En el estudio de las regiones cloroplásticas insertas en el genoma nuclear se tomó en cuenta la dinámica entre la frecuente integración de ADN de los organelos y la muy rápida eliminación de éstos fragmentos del genoma nuclear. Así mismo, aquellos fragmentos que son retenidos en el núcleo sufren un proceso de fragmentación muy rápido y acumulan mutaciones puntuales, pero con menor velocidad que su fragmentación. Es de esperar entonces que los fragmentos transferidos desde el cloroplasto hacia el núcleo tengan un grado de divergencia con el donador directamente proporcional al tiempo de inserción en el núcleo de ese fragmento siguiendo este patrón mencionado (Matsuo et al., 2005) (ver Figura 16).

El proceso de fragmentación de los insertos no está claro, pero podría deberse a inserciones de elementos transponibles en la región del fragmento cloroplástico en el núcleo (Noutsos et al., 2005). En efecto, la mayoría de los insertos cloroplásticos largos parecen comenzar a disminuir su tamaño dentro del primer millón de años (Matsuo et al., 2005). Se observó tanto para *A. thaliana* como para *O. sativa* la existencia de dos tipos de insertos, uno que es colinear con el del genoma de origen y otro tipo de insertos que son mosaicos de ADN de los organelos, a menudo ADN proveniente de ambos organelos (Noutsos et al., 2005)

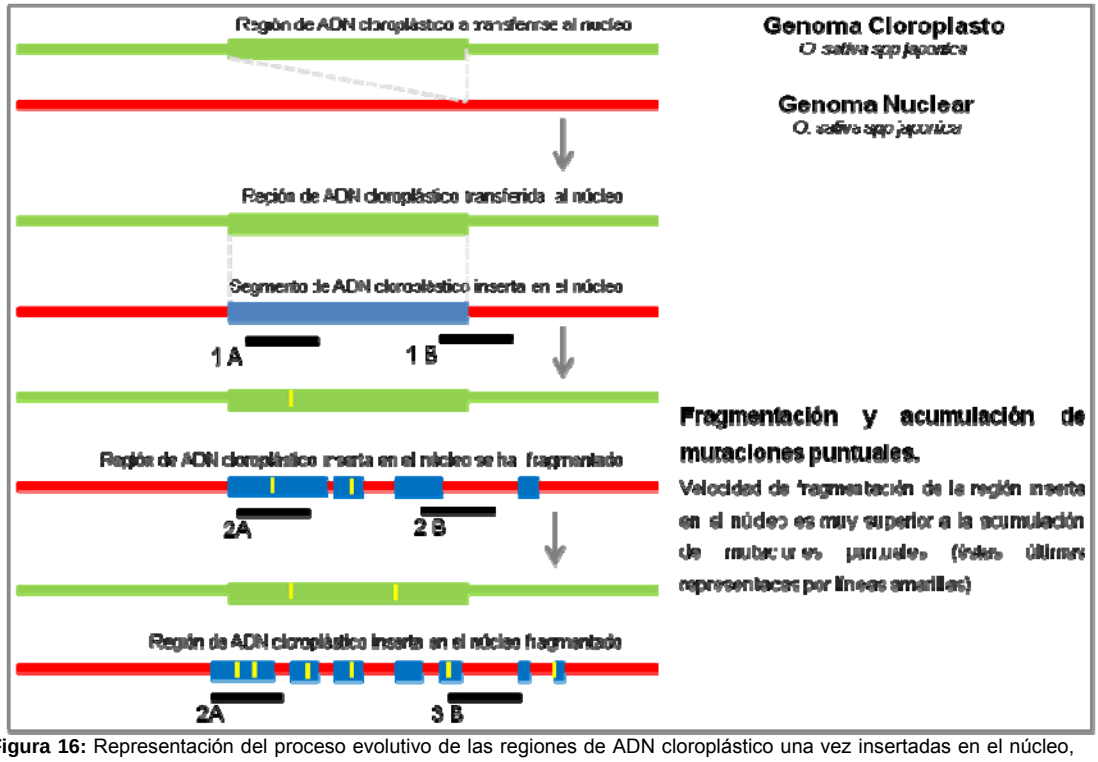

Figura 16: Representación del proceso evolutivo de las regiones de ADN cloroplástico una vez insertadas en el núcleo, el cual implica rápida fragmentación y acumulación de mutaciones puntuales a una tasa significativamente menor

Para poder interpretar los datos obtenidos por nuestra secuenciación debemos tener en cuenta que el conjunto de lecturas de AM356-8 usado en este trabajo es una muestra de fragmentos de secuencias provenientes de los 3 genomas (núcleo, cloroplasto y mitocondria) por lo que estas representan una combinación compleja de material hereditario, siendo muchas veces difícil de determinar el origen de algunas de éstas lecturas. La compresión del proceso evolutivo de los segmentos cloroplásticos una vez insertos en los cromosomas ayuda a determinar el origen de los mismos. En concreto, el esquema presentado en la figura 16 nos permite predecir cómo serían las lecturas provenientes de las distintas regiones genómicas. Por ejemplo, una lectura puede ser derivada de ADN nuclear pero presentar alta identidad con el genoma cloroplástico por tratarse precisamente una transferencia desde el cloroplasto al núcleo relativamente recientemente. Ésta situación representada en la figura 16 caso 1, comprendería lecturas con alta identidad tanto con el genoma nuclear como con el cloroplástico dificultando su asignación a un genoma en particular. Proporcional al tiempo de ocurridas estas transferencias, las regiones cloroplásticas insertas en el genoma nuclear comienzan a diferenciarse de las originales por acumulación de mutaciones puntuales y fragmentación. Estos casos están representados en la figura 16 como los casos 2 y caso 3. Las mutaciones puntuales generarían la divergencia de las secuencia entre las lecturas derivadas de estas regiones y el genoma cloroplástico de origen. La fragmentación se identificaría como lecturas que solo mapean parcialmente contra el genoma cloroplástico y en forma completa contra el genoma nuclear.

Basados en estas premisas se planteó una estrategia de búsqueda y clasificación de los fragmentos cloroplásticos insertos en el genoma nuclear.

Transferencias recientes.

El grupo de transferencias que se definieron como recientes, está compuesto por aquellos fragmentos que fueron insertos recientemente en el genoma nuclear de *japonica* y que en consecuencia mantienen una alta similitud, o incluso identidad completa con el genoma cloroplástico. Estos se representan en la figura 16 como 1A y 1B.

De las dos situaciones posibles esquematizadas en la figura 16, es evidente que en el caso representado como 1A no es posible determinar si una lectura proviene del genoma cloroplástico o del nuclear. Un aspecto importante es determinar qué proporción del genoma cloroplástico se clasifica dentro de esta categoría. Cabe aclarar, que si bien dicha estimación podría realizarse simplemente mapeando el genoma cloroplástico de *O. sativa spp japonica* sobre el genoma nuclear de la misma, la estimación estaría sesgada por la calidad del ensamblaje nuclear y cloroplástico disponible.

Para las estimación de las trasferencias recientes, se tomó el conjunto de lecturas que mapean en forma completa con el cloroplasto (RCJ), estas son lecturas que presentaron un porcentaje de solapamiento (%s) con el genoma cloroplástico cercano al 100% y que su vez presentaron un porcentaje de identidad (%ID) cercano a 100 con el mismo. Los filtros aplicados para seleccionar estas lecturas fueron: %ID con cp >= 99% y %s con cp >= 99%. De esta manera se identificaron 27731 lecturas que cumplen con dichos criterios. Estas lecturas luego, se mapearon en el genoma nuclear mediante blastn. Sobre este resultado de blast se aplicaron los mismo filtros que en la comparación con el cloroplasto, %ID con el núcleo >=99% y %s con el núcleo >= 99% identificando 27283 lecturas que cumplen con el criterio de filtrado, es decir casi el 100% de las lecturas. En otras palabras podemos decir que casi todo el genoma cloroplástico ha sido transferido recientemente al núcleo. Esto no es sorprendente pues cómo se indicó antes una copia completa de éste genoma se encuentra en el cromosoma 10 de *O. sativa japonica* (Yu et al., 2003).

La determinación de lecturas provenientes de las zonas límite entre cromosomas y segmentos de ADN cloroplástico en ellos insertos (situación 1B), por otra parte, nos permitirá estimar cuántos segmentos distintos de este tipo existen en el genoma nuclear de *O. sativa ssp japonica*. Una lectura proveniente de estas zonas se caracteriza por tener un alineamiento completo con el genoma nuclear y solo parcial con el cloroplástico y presentar una identidad de secuencia cercana a 100 % con ambos genomas (Figura 17). En este caso los filtros aplicados sobre ambos resultados de blast fueron: %ID con cloroplasto y núcleo >= 99%; % s con cloroplasto < 90%, %s con núcleo > 99%. Tal búsqueda permitió identificar 25 lecturas que cumplen con los criterios de filtrado aplicados en base a lo esperado para trasferencias modernas o muy recientes

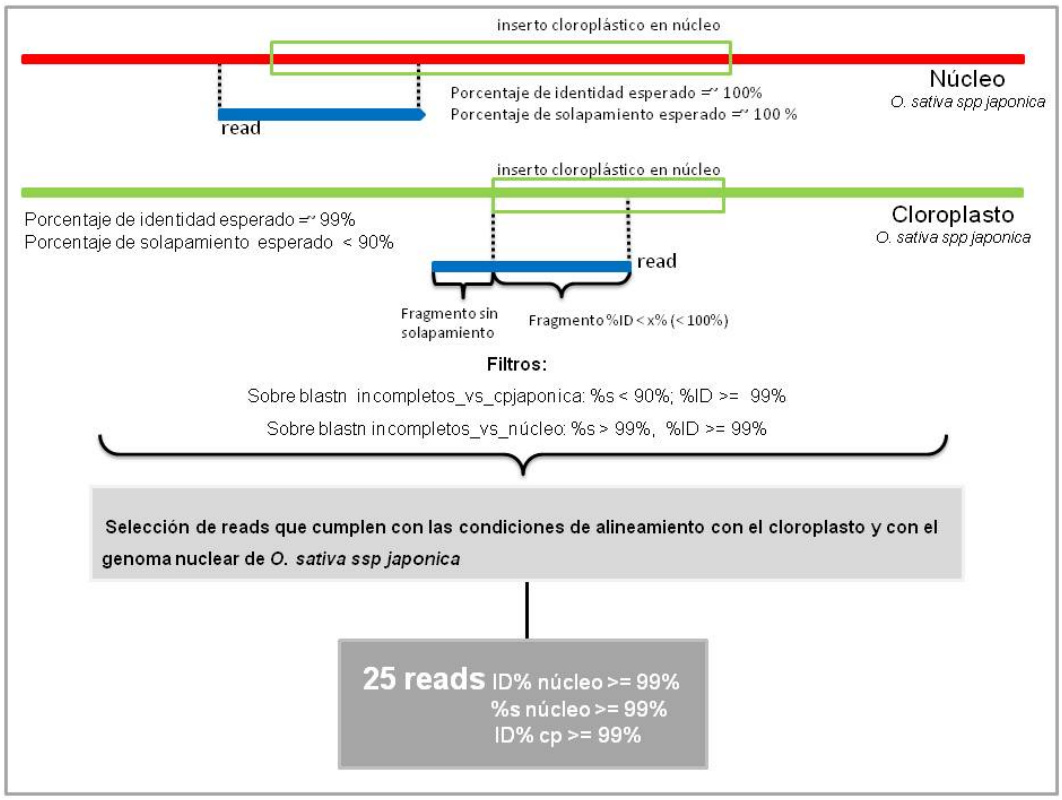

Figura 17: Estrategia de búsqueda de lecturas que representaba trasferencias modernas entre cloroplasto y núcleo de *O. sativa japonica* (situación 2 A)

Teniendo en cuenta que nuestra muestra tiene una representación de sólo el 10 % del genoma nuclear, se estima que habrá 250 lecturas totales representando regiones trasferidas. Esto correspondería a 125 eventos de transferencias recientes.

Transferencias antiguas

Las transferencias antiguas están representadas por lecturas que presentaron un alineamiento completo o casi completo (más de 99% de solapamiento) con el núcleo y un porcentaje de identidad significativamente mayor con el núcleo que con el cloroplasto. Esta diferencia en los porcentajes de identidad se debería a la acumulación de cambios en la secuencia desde la inserción del fragmento en el nuevo genoma, lo que llevó a éste a diferenciarse de su genoma de origen (figura 16 casos 3A y 3B). Por ello, para la identificación de lecturas que representaran estás transferencias se comparó las características de alineamiento de cada lectura con el genoma cloroplástico y con el genoma nuclear de *O. sativa ssp japonica*.

Ésta búsqueda se realizó con dos conjuntos de lecturas por separado, por un lado se utilizó el conjunto RCJ (lecturas que presentaron un alineamiento con el cloroplasto *japonica* con un %s > 99%) y luego el conjunto RIJ (lecturas que presentaron un alineamiento con el cloroplasto *japonica* con un %s < 99%).

Sobre el resultado de blastn del conjunto RCJ con el cloroplasto de *japonica* se buscaron lecturas que presentaran un porcentaje de identidad con el cloroplasto menor al 98%. De las 34091 lecturas del conjunto RCJ se identificaron 2094 lecturas que presentaban un %ID menor al 98% con dicho genoma. Estas 2094 lecturas se buscaron luego, en el resultado de blast de RCJ con el núcleo, aplicando una vez más filtros sobre el porcentaje de identidad y sobre el porcentaje de solapamiento (Figura 16 caso 3A). En esta instancia se aplicaron dos criterios de filtrado: con el primer criterio menos restrictivo, se seleccionaron de las 2094 lecturas, las que presentaron un %ID con el núcleo >= 98% y %s con el núcleo >=99%. De esta forma se identificaron 114 lecturas. En el segundo criterio de filtrado utilizado, más restrictivo, se seleccionó lecturas con un %ID >=99% y %s >= 99%. Con este segundo criterio de filtrado se identificaron 30 lecturas, éstas fueron consideradas como de origen nuclear. En la figura 18 se detalla la estrategia de búsqueda y un resumen de los resultados.

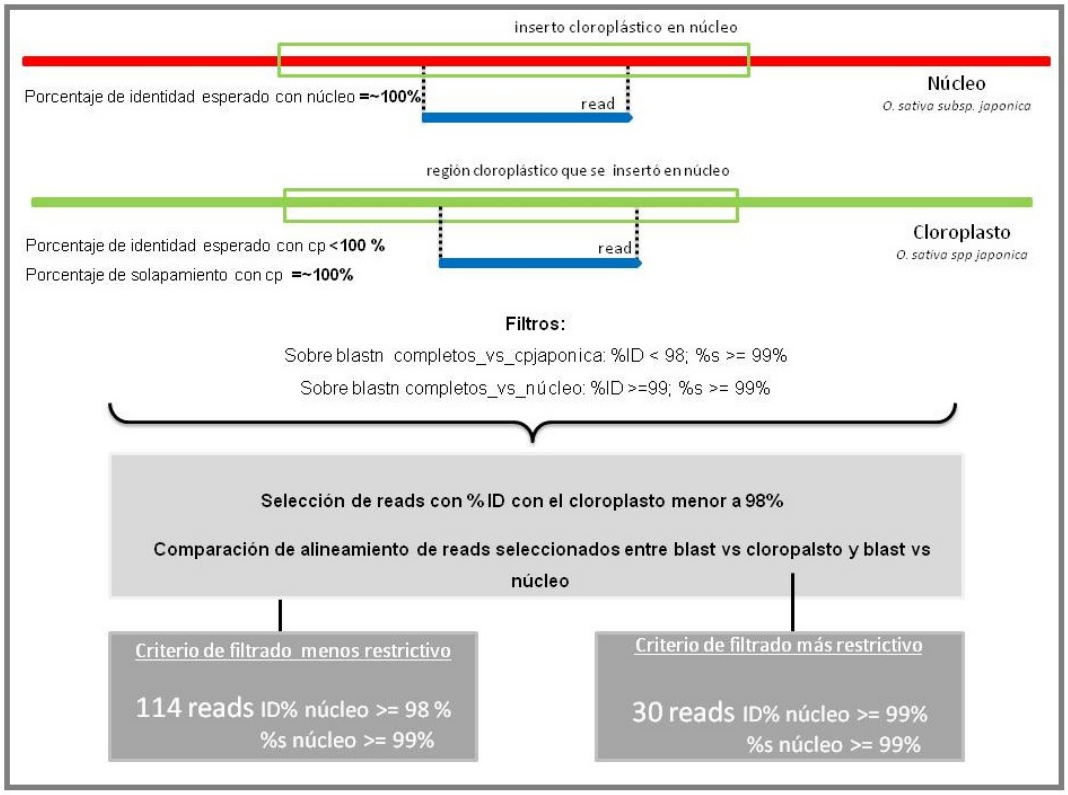

Figura 18: Estrategia de búsqueda de lecturas que representaba trasferencias antiguas entre cloroplasto y núcleo de *O. sativa japonica* (situaciones 2 A y 3 A).

El segundo conjunto de lecturas utilizado para la búsqueda de regiones transferidas entre el genoma cloroplástico y el nuclear fue RIJ (lecturas con un alineamiento incompleto con el cloroplasto de *japonica*, %s < 99%). Al utilizar este conjunto de lecturas se espera identificar bordes de los fragmentos cloroplásticos transferidos al núcleo, y/o regiones más cortas que las propias lecturas generados por la fragmentación de las regiones de ADN insertas durante el proceso evolutivo (Figura 16 1B, 2B, 3B).

La identificación de lecturas que representaran bordes de trasferencia o regiones cortas (fragmentadas) se hizo sobre un subconjunto de las lecturas RIJ. Este subconjunto nombrado como RIJa, está formado por 1444 lecturas que presentaron un solapamiento menor a 90% con el genoma del cloroplasto de *japonica*. En este caso se aplicaron también dos criterios de filtrado, uno menos restrictivo y otro más restrictivo. Para el primer criterio se exigió a las lecturas un alineamiento mayor o igual al 98% de solapamiento con el núcleo, a través del cual se identificaron 55 lecturas. Con el segundo criterio, más restrictivo se exigió que el %s con el núcleo fuera mayor o igual a 99%, identificando 50 lecturas que cumplían con dicho criterio. Éstas 55 lecturas que presentan un solapamiento parcial con el cloroplasto y completo con el núcleo pueden ser considerados como provenientes del genoma nuclear de zonas que contienen fragmentos de origen cloroplástico (figura 19).

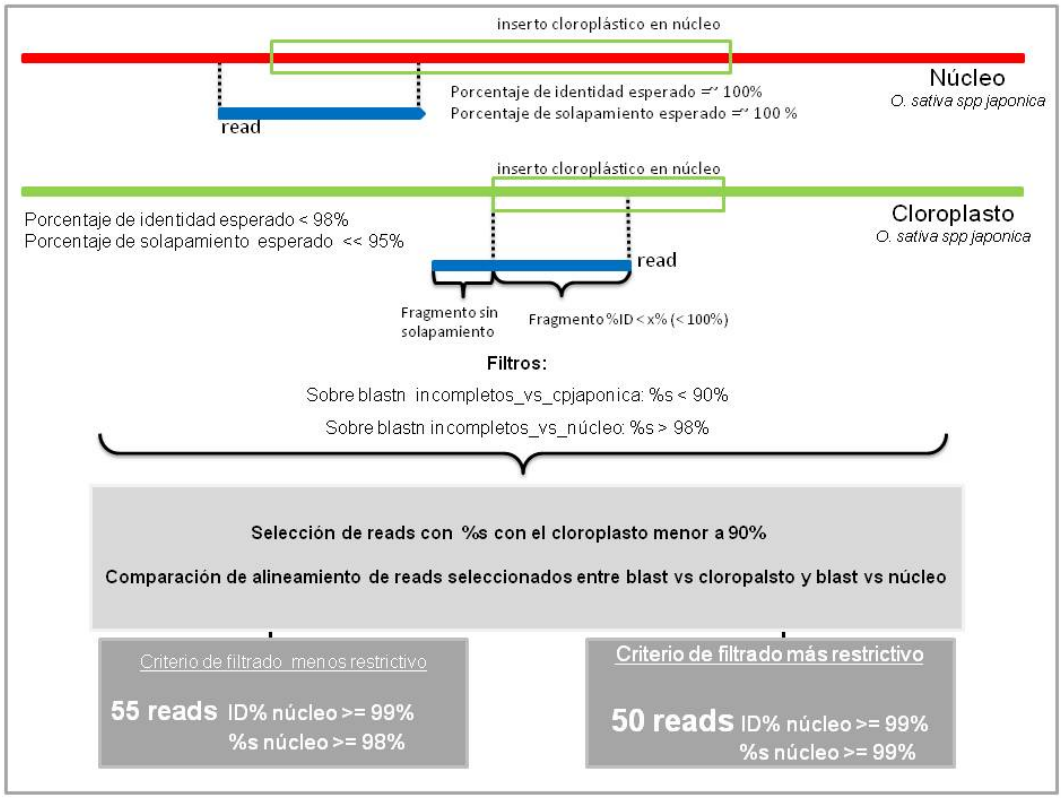

Figura 19: Estrategia para la detección de lecturas que representen transferencias antiguas desde el cloroplasto hacia el núcleo de *O. sativa subespecie japonica.* Situaciones 2B y 3B.

Comparando los resultados obtenidos en las dos búsquedas de regiones transferidas entre el genoma cloroplástico y el genoma nuclear se observa casi el doble de lecturas cuando se buscan bordes o regiones cortas. Este resultado está de acuerdo con el proceso de fragmentación reportado, ya que se espera que el largo de la región transferida decaiga con el tiempo (Figura 16 caso 3) (Leister, 2005b; Matsuo et al., 2005).

Transferencias exclusivas de la maleza.

Dentro del conjunto de lecturas definidas como incompletos RIJ, se podían encontrar lecturas de orígenes diferentes. Algunos de ellos, pueden haber sido generados por errores de secuenciación y es ese el motivo por el que no alinean en su totalidad con el cloroplasto. Mientras que otras lecturas de este grupo podrían estar representando diferentes eventos de trasferencias. Eventos tales como bordes o regiones cortas insertas hace algún tiempo en el genoma nuclear, como se mostró anteriormente. Otros podrían estar representando transferencias muy recientes.

Las transferencias muy recientes o exclusivas de la maleza comprenden lecturas provenientes de regiones del genoma nuclear que corresponden a inserciones del genoma cloroplástico ocurridas luego de la separación de la maleza con el arroz cultivado *japonica. A*lternativamente, éstas lecturas pueden no haber sido incluidas en el ensamblado del genoma nuclear de *O. sativa japonica* por su similitud con el genoma del cloroplasto. Para cualquiera de los dos casos posibles de origen de este grupo de lecturas, éstas deberían alinear en forma de mosaico. Es decir, un fragmento de la lectura alinearía con muy alta identidad y en único sitio con el núcleo y el fragmento complementario de la misma lectura alinearía, también con muy alta identidad con el cloroplasto de *japonica*. En la figura19 se muestra de forma esquemática el alineamiento esperado, la estrategia de búsqueda y el resumen de los resultados.

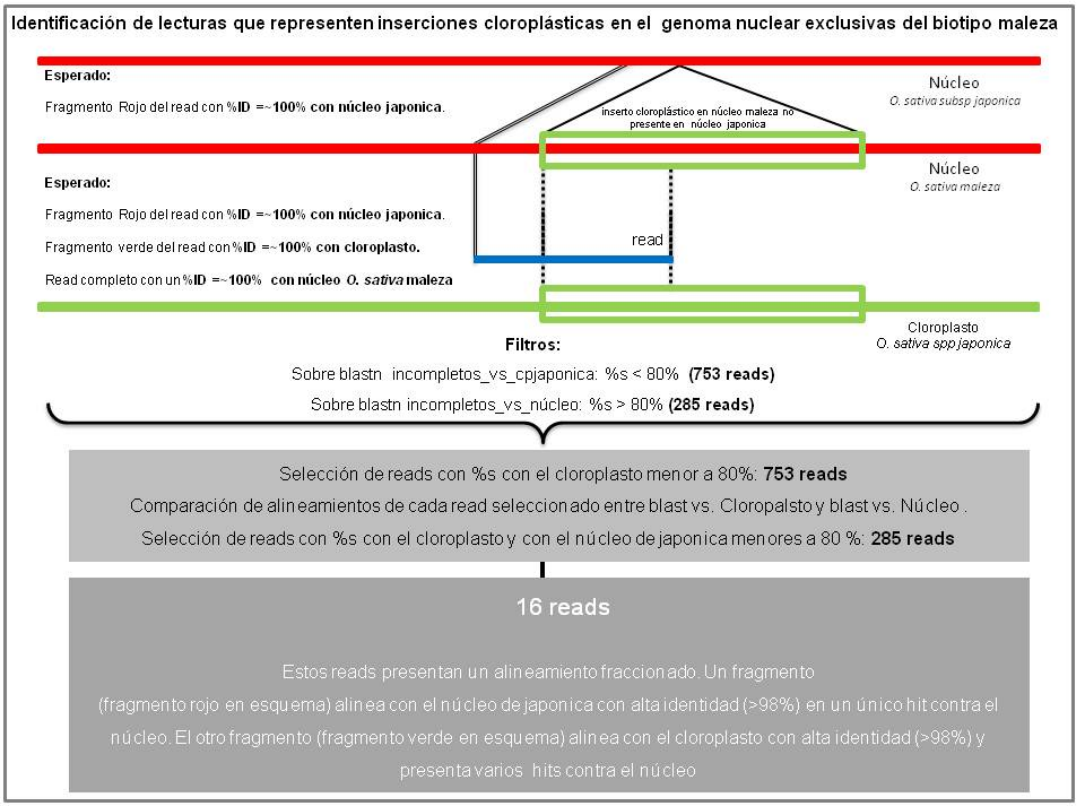

Figura 19: Estrategia para la identificación de lecturas que representan transferencias contemporáneas entre el cloroplasto y el núcleo de *O. sativa* maleza.

La identificación de lecturas con un alineamiento fraccionado con los genomas nuclear y cloroplástico se realizó de la siguiente forma. Se buscó sobre el resultado de blastn del conjunto de lecturas RIJa con el cloroplasto *japonica* aquellas que presentaran un porcentaje de alineamiento (%s) menor al 80%. Se identificaron 753 lecturas de las 1444 totales del conjunto RIJa. A éstas 753 lecturas se las buscó sobre el resultado de blastn de RIJa con el núcleo de *japonica*, sobre el cual se aplicó nuevamente el filtro de solapamiento de manera de seleccionar aquellas lecturas con solapamiento menor al 80% también con el núcleo. De esta manera se identificaron 285 lecturas. Son éstas últimas las lecturas con un alineamiento parcial con el genoma nuclear y con el cloroplástico. Sobre éstas 285 lecturas se realizó una búsqueda adicional para determinar cuáles de las mismas podían ser candidatas a representar trasferencias recientes exclusivas del biotipo secuenciado. Para ello se exigió que el segmento que alineaba con uno y otro genoma no se superpusieran y además que el alineamiento del fragmento con el núcleo fuera único (figura 19). Se identificaron 16 lecturas que cumplen con esta condición y por tanto son candidatas a representar transferencias recientes. El detalle del alineamiento de éstas lecturas con ambos genomas se muestra en la Tabla 6.

READ	Genoma	%ID	Largo aln	Coord. Inicio lecturas	Coord. Fin read	Coord inicio DB	Coord fin DB	e-value	Largo read
GCFF90V02GW6GW	AY522330.1	99.59	246	271	515	54081	54326	5,00E-135	517
GCFF90V02GW6GW	cromosoma 2	99.63	270	1	270	1879451	1879182	2,00E-148	517
GCFF90V02F9BJE	AY522330.1	100	116	267	382	111458	111573	2,00E-62	382
GCFF90V02F9BJE	cromosoma 2	98.13	268	1	268	15462658	15462391	9,00E-138	382
GCFF90V02IEQIE	AY522330.1	100	115	1	115	30704	30818	1,00E-61	516
GCFF90V02IEQIE	cromosoma 4	99.76	411	106	516	32858510	32858919	0.0	516
GCFF90V02JK097	AY522330.1	100	199	223	421	43044	42846	7,00E-112	421
GCFF90V02JK097	cromosoma 5	99.56	228	1	228	2460207	2460433	5,00E-121	421
GCFF90V02HC0XG	AY522330.1	100	131	307	437	45371	45241	3,00E-71	437
GCFF90V02HC0XG	cromosoma 5	98.04	306	1	306	3748044	3748344	5,00E-155	437
GCFF90V02GJK3F	AY522330.1	100	134	226	359	46482	46615	4,00E-73	359
GCFF90V02GJK3F	cromosoma 5	99.12	228	1	228	113635156	113634929	4,00E-121	359
GCFF90V02G6KGA	AY522330.1	98.70	154	198	351	67685	67533	6,00E-78	351
GCFF90V02G6KGA	cromosoma 5	100	205	1	205	26830841	26831045	4,00E-112	351
GCFF90V02H8EV7	AY522330.1	89.36	141	1	134	71978	71844	3,00E-37	488
GCFF90V02H8EV7	cromosoma 8	96.87	319	113	430	25340556	25340238	9,00E-154	488
GCFF90V02JD9YU	AY522330.1	100	185	1	185	88217	88401	1,00E-103	397
GCFF90V02JD9YU	cromosoma 12	100	218	180	397	23930784	23930567	7,00E-120	397
GCFF90V02GBG4O	AY522330.1	100	123	159	281	77439	77561	1E-66	281
GCFF90V02GBG4O	cromosoma 4	100.00	162	1	162	16565916	16566077	1,00E-87	281
GCFF90V02F55JZ	AY522330.1	100	185	1	185	88217	88401	1E-103	397
GCFF90V02F55JZ	cromosoma 12	100.00	218	180	397	23930784	23930567	6,00E-121	397

Tabla 8: Detalle de alineamiento contra genoma de cloroplasto y nuclear de *O. sativa ssp japonica* de lecturas candidatos a representar trasferencias recientes sólo presentes en núcleo de *O. sativa* maleza.

En base a estos resultados podemos hacer algunas estimaciones en cuanto al número de trasferencias que ocurrieron desde que se separó el arroz maleza del arroz *japonica* cultivado, así como su posible tasa de ocurrencia. Considerando la cobertura del genoma nuclear obtenida en este trabajo (0.13X) y las 16 regiones identificadas como posibles trasferencias en el genoma nuclear de la maleza (Tabla 8), se puede estimar que existen en el genoma nuclear del biotipo en estudio aproximadamente 150 fragmentos que representan trasferencias desde el genoma cloroplástico hacia el nuclear. Esto se correspondería con 75 nuevas inserciones ocurridas exclusivamente en el linaje que condujo al biotipo de arroz maleza.

Por otra parte, si se tiene en cuenta que el proceso de separación entre el biotipo cultivado y el biotipo maleza, asumiendo que este último surgió a partir de un proceso de selección y re-hibridación entre cultivares y arroz maleza, habría ocurrido hace 100 a 200 años atrás (Olsen et al., 2007). Es posible estimar una tasa de trasferencias entre el cloroplasto y el núcleo de una transferencia cada 1.5 a 3 años dependiendo de el tiempo exacto de origen del biotipo maleza.

Sin embargo debemos considerar la posibilidad que estas transferencias exclusivas de maleza que mapean en forma de mosaico con los genomas nucleares y cloroplástico sean en realidad artefactos de secuenciación. Si bien esto es un hecho relativamente improbable debido a la calidad de la secuenciación obtenida, tal como se mostró con los análisis de calidad realizados al comienzo del trabajo, tal posibilidad debe ser testada.

Por ello se diseñó un ensayo experimental que interroga la presencia de estas regiones en el genoma nuclear del biotipo maleza y su eventual ausencia en el del arroz *japonica*. Dos de estas inserciones fueron elegidas, las representadas por las lecturas GCFF90V02JK097 y GCFF90V02GW6GW. Para cada una de ellas se diseñaron dos pares de cebadores, de acuerdo a los esquemas que se presentan en las figuras 20 y 21. La combinación de estos cebadores permite amplificar solo (o en forma diferencial) en ausencia de la inserción cloroplástica o solo en presencia de la misma.

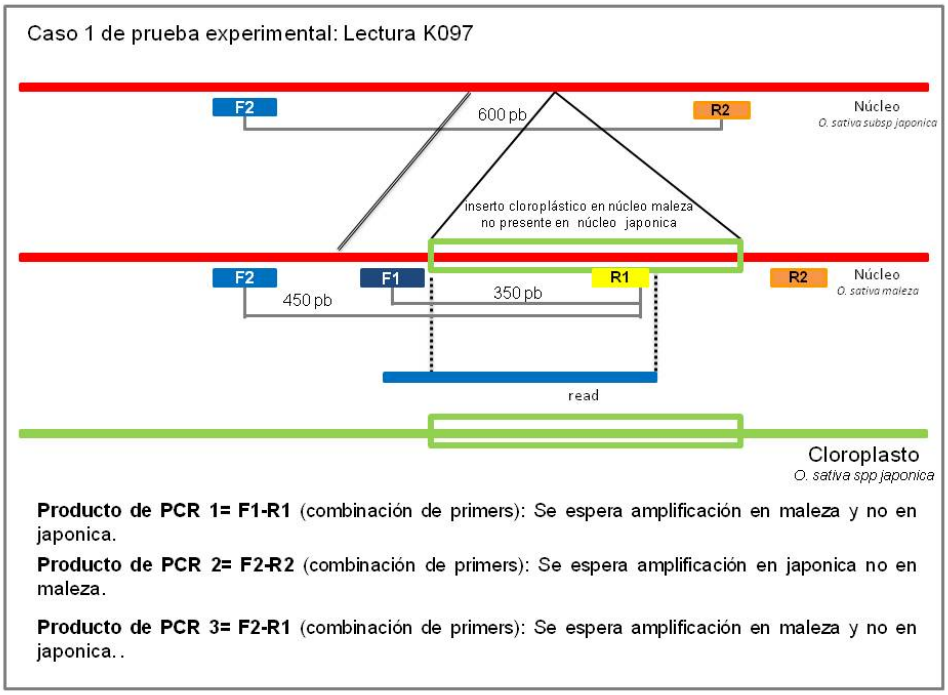

Figura 20: Esquema de sitios de hibridación de los cebadores diseñados para cada genoma incluido en el ensayo y resultado esperado para la región representada por la lectura K097

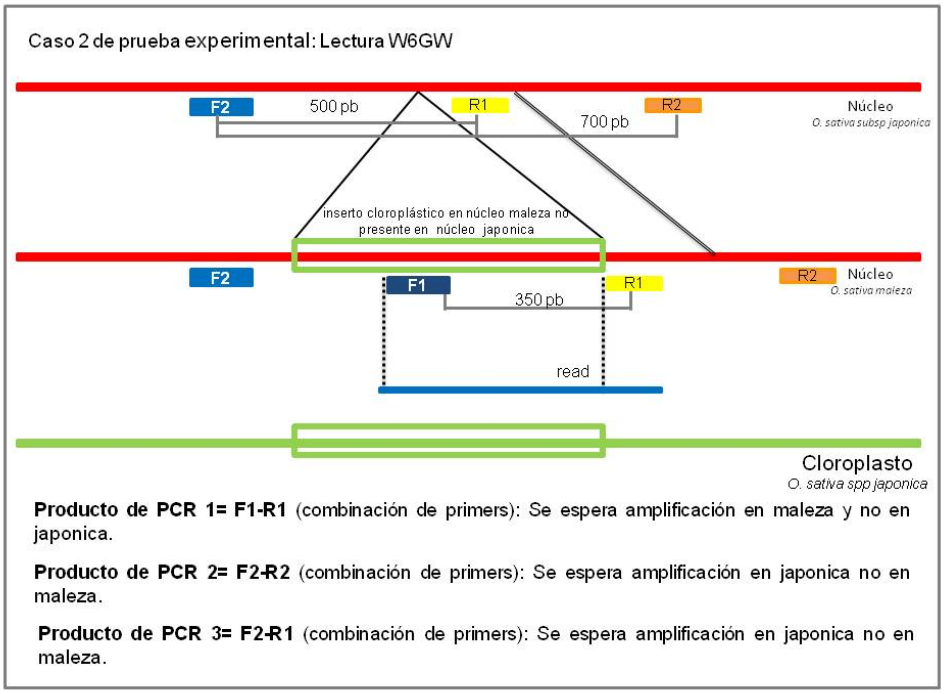

Figura 21: Esquema de sitios de hibridación de los cebadores diseñados para cada genoma incluido en el ensayo y resultado esperado para la región representada por W6GW.

Las amplificaciones se realizaron sobre el ADN de AM356-8 y cultivar *japonica* Tacuarí para cada una de las regiones evaluadas. Se utilizó una muestra de ADN bovino como control negativo de amplificación. Cómo control positivo de amplificación se utilizaron cebadores para los microsatélites RM215 y RM319 (Figura 22 d)

En el caso 1, representado por la lectura K097, se testo la combinación de cebadores F2-R2 para los que se obtuvo amplificación sólo en el biotipo cultivar *japonica* como era esperado (Figura 22a y c). Cuando fue testada la combinación F2-R1 se obtuvo amplificación para ambos biotipos, maleza y cultivar *japonica* (Figura 22b). En este caso no se esperaba amplificación en *japonica*, dado que uno de los cebadores se diseño para hibridar sobre la región cloroplástica inserta en el núcleo. A partir de estos resultados se puede concluir que en ambos biotipos está presente la inserción representada por la lectura K097. En la maleza esta región sería homocigota, mientras que en *japonica* sería heterocigota y es por ello que se obtuvo amplificación con ambas combinaciones de cebadores. INIA Tacuarí la variedad *japonica* utilizada para el ensayo, es una variedad comercial de una planta autógama, por lo que no se espera encontrar loci heterocigotos en su genoma. Luego de aproximadamente 12 generación de retrocruzamientos, necesarios para la liberación de una variedad, la heterocigosis se ve muy reducida, pero no se elimina por completo. Éste podría ser el caso del loci representado por la lectura K097.

Este caso además representa una región cloroplástica transferida desde el genoma del cloroplasto al núcleo que definitivamente está presente en el genoma nuclear de *japonica* pero no en el genoma ensamblado utilizado en este trabajo como referencia. Es decir que esta región fue excluida del ensamblaje del genoma nuclear de *japonica* probablemente por el estado heterocigoto de la misma.

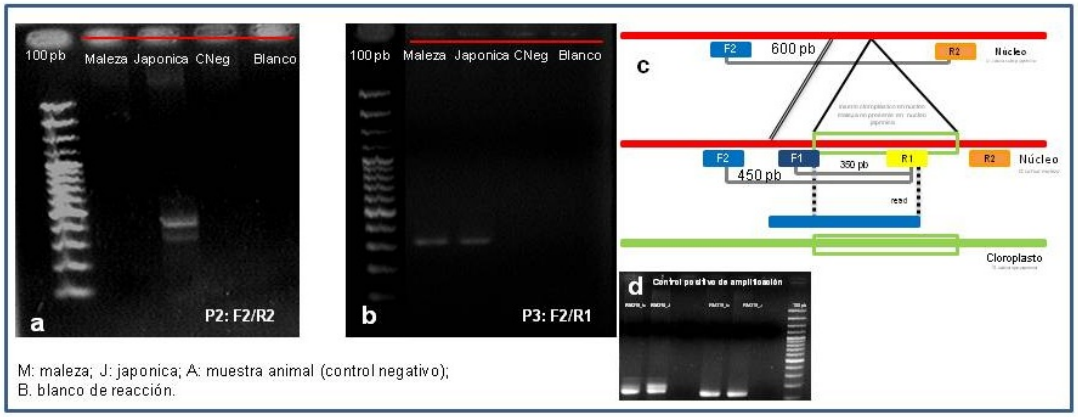

Figura 22: Resultado amplificaciones para el testado de la presencia/ausencia de inserción representada por la lectura K097. a)) Resultado de amplificación con combinación de cebadores F2/R2,visualizado en gel de agarosa 2% con tinción de bromuro de etidio; b) Resultado de amplificación con combinación de cebadores F2/R1, visualizado en gel de agarosa 2% con tinción de bromuro de etidio ; c) Esquema de sitios de hibridación de los cebadores en cada genoma testado y tamaños esperados; d) Control positivo de reacción**.**

En el caso 2, representado por la lectura GCFF90V02GW6GW, se testaron las combinaciones de cebadores F2-R2 y F2-R1 (Figura 23a y b) los cuales hibridan fuera de la región del inserto. En ambos casos se obtuvo amplificación sólo en *japonica* con los tamaños de producto amplificado esperados (Figura 23c). A partir de éstos resultados se podría decir que en el genoma nuclear de *japonica* no estaría presente la inserción cloroplástica representada por la lectura W6GW, pero si es probable que dicha inserción esté presente en el genoma de la maleza. Siendo la inserción el motivo por el que no hubo amplificación en el biotipo maleza, pues la presencia de la misma incrementaría la distancia entre los puntos de hibridación de los cebadores a punto tal que impide la amplificación.

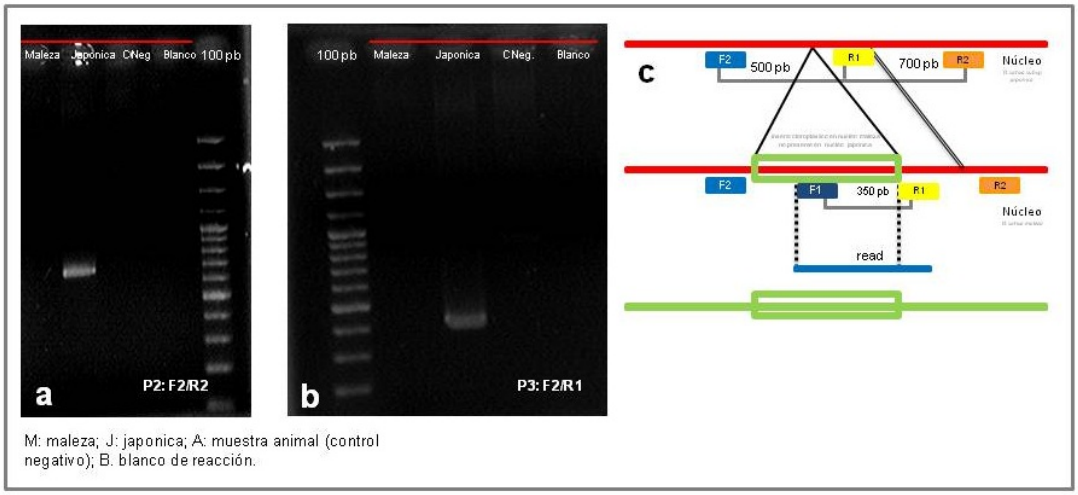

Figura 23: Resultado amplificaciones para el testado de la presencia/ausencia de inserción representada por la lectura W6GW. a) Resultado de amplificación con combinación de cebadores F2/R2, visualizado en gel de agarosa 2% con tinción de bromuro de etidio; b) Resultado de amplificación con combinación de cebadores F2/R1, visualizado en gel de agarosa 2% con tinción de bromuro de etidio; c) Esquema de sitios de hibridación de los cebadores en cada genoma testado y tamaños esperados.

En el presente trabajo se secuenció el genoma del arroz maleza obteniendo una cobertura de 0.13X, es decir que en los datos de secuencias está representado el 13% del genoma nuclear del biotipo. Considerando, que el genoma nuclear de arroz contiene aproximadamente un 0.2% del genoma cloroplástico (Matsuo et al., 2005), se realizaron algunas estimaciones de interés.

Si el 0.2% del ADN nuclear de arroz es de origen cloroplástico, el número esperado de lecturas que representen estas inserciones, con un 13% del total del genoma analizado sería 359 lecturas. En este trabajo identificamos 114 lecturas como transferencias antiguas, 55 lecturas representaban bordes de las mismas trasferencias antiguas o regiones insertas de menor longitud que la propia lectura. Además se identificaron 25 lecturas que representaban bordes de transferencias modernas y 16 lecturas representando las trasferencias más recientes encontradas exclusivamente en el genoma del biotipo maleza. En total identificamos 210 lecturas que podrían estar representando insertos cloroplásticos en el núcleo.

Esta subrepresentación puede deberse a la incapacidad de detección de lecturas tipo 1A de la figura 16, las que representan regiones cloroplásticas insertas tan recientemente que aún

mantienen completamente o casi completamente la identidad con el cloroplasto. O a la incapacidad de detección por no estar presentes en el genoma de referencia, como ya se comentó, estos fragmentos pueden haber sido eliminados del ensamblado del genoma nuclear de referencia debido a la alta similitud que mantiene con el genoma de origen.

Conclusiones finales

1) En el presente trabajo se obtuvieron vastos segmentos de secuencias genómicas de arroz maleza así como el genoma completo del cloroplasto con una cobertura de 106x. Esta información permitió realizar varios estudios de genómica comparativa los cuales a su vez nos aportaron datos de diversa índole. Además se refinó una metodología para la detección de secuencias cloroplásticas a partir de datos de secuencia de ADN total.

2) a - Se generaron herramientas para estudios de poblacionales dentro del género *Oryza.* Los InDels identificados en este trabajo representan haplotipos que diferencian claramente entre las subespecies de *O. sativa*, y así mismo entre haplotipos *nivara* y *rufipogon*. Aplicarlos en un estudio poblacional del complejo *Oryza* podría ayudar a resolver las relaciones filogenéticas de éstas especies aún en discusión, así como para trazar el origen de los biotipos de arroz maleza encontrados en campos uruguayos.

b- Se determinó a través técnicas de genómica comparativa el origen del arroz maleza secuenciado. Éste habría surgido por cruzamiento entre un biotipo maleza y un cultivar.

3) a- Se ajustó un procedimiento bioinformático que permite la detección de transferencias entre los diferentes genomas de una célula vegetal.

b- Se identificaron nuevas transferencias de ADN desde el genoma cloroplástico y nuclear en el arroz maleza estudiado. Estos resultados permitieron realizar estimaciones en la tasa de transferencia, las que muestran que existe una alta tasa de transferencia de material genético aún hoy entre el núcleo y los genomas de los organelos citoplasmáticos.

.

Bibliografía

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., and others (1990). Basic local alignment search tool. Journal of Molecular Biology *215*, 403–410.

Alverson, A.J., Wei, X., Rice, D.W., Stern, D.B., Barry, K., and Palmer, J.D. (2010). Insights into the Evolution of Mitochondrial Genome Size from Complete Sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae). Mol Biol Evol *27*, 1436–1448.

Bedbrook, J.R. (1980). Molecular cloning and sequencing of cDNA encoding the precursor to the small subunit of chloroplast ribulose-1, 5-bisphosphate carboxylase. Nature *287*, 692–697.

Birky, C.W. (1991). Evolution and population genetics of organelle genes: mechanisms and models. Evolution at the Molecular Level 112–134.

Birky Jr, C.W. (1978). Transmission genetics of mitochondria and chloroplasts. Annual Review of Genetics *12*, 471–512.

Bock, R., and Timmis, J.N. (2008). Reconstructing evolution: gene transfer from plastids to the nucleus. Bioessays *30*, 556–566.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., Green, E.D., Sidow, A., and Batzoglou, S. (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. *13*, 721–731.

Butterfield, N.J. (2000). Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiology *26*, 386–404.

Caicedo, A.L., Williamson, S.H., Hernandez, R.D., Boyko, A., Fledel-Alon, A., York, T.L., Polato, N.R., Olsen, K.M., Nielsen, R., McCouch, S.R., et al. (2007). Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice. PLoS Genet *3*, e163.

Cao, Q., LU, B.A.O.R., Xia, H.U.I., Rong, J., Sala, F., Spada, A., and Grassi, F. (2006). Genetic diversity and origin of weedy rice (Oryza sativa f. spontanea) populations found in north-eastern China revealed by simple sequence repeat (SSR) markers. Annals of Botany *98*, 1241–1252.

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G., and Parkhill, J. (2005). ACT: the Artemis comparison tool. Bioinformatics *21*, 3422–3423.

Chen, B. (1999). Origin of 8000-year-old cultivated rice in Henan's Jia Lake site. Agric Archaeol *1*, 55–57.

Chen, L.J., Lee, D.S., Song, Z.P., Suh, H.S., and Lu, B.-R. (2004). Gene Flow from Cultivated Rice (Oryza sativa) to Its Weedy and Wild Relatives. Ann Bot *93*, 67–73.

Chevreux, B. (2005). MIRA: an automated genome and EST assembler. Duisburg: Heidelberg.

Cronn, R., Liston, A., Parks, M., Gernandt, D.S., Shen, R., and Mockler, T. (2008). Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucl. Acids Res. *36*, e122–e122.

Darling, A.C.E., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Research *14*, 1394–1403.

Delouche, J.C., Burgos, N.R., Gealy, D.R., Zorrilla de San Martin, G., Labrada, R., Larinde, M., and Rosell, C. (2007). Arroces maleza-origen biología, ecología y control (Food & Agriculture Org.).

Diekmann, K., Hodkinson, T.R., and Barth, S. (2012). New chloroplast microsatellite markers suitable for assessing genetic diversity of Lolium perenne and other related grass species. Ann Bot *110*, 1327–1339.

Dobberstein, B., Blobel, G., and Chua, N.H. (1977). In vitro synthesis and processing of a putative precursor for the small subunit of ribulose-1, 5-bisphosphate carboxylase of Chlamydomonas reinhardtii. Proceedings of the National Academy of Sciences *74*, 1082–1085.

Doyle, J.J., Davis, J.I., Soreng, R.J., Garvin, D., and Anderson, M.J. (1992). Chloroplast DNA inversions and the origin of the grass family (Poaceae). Proceedings of the National Academy of Sciences *89*, 7722.

Eguiarte, L.E., Castillo, A., and Souza, V. (2003). Evolución molecular y genómica en angiospermas. Interciencia *28*, 141–147.

Ellstrand, N.C. (2003). Current Knowledge of Gene Flow in Plants: Implications for Transgene Flow. Phil. Trans. R. Soc. Lond. B *358*, 1163–1170.

Federici, M.T., Vaughan, D., Norihiko, T., Kaga, A., Xin, W.W., Koji, D., Francis, M., Zorrilla, G., and Saldain, N. (2001). Analysis of Uruguayan weedy rice genetic diversity using AFLP molecular markers. Electronic Journal of Biotechnology *4*, 5–6.

Ferrero, A. (2001). Biology and control of red rice (Oryza sativa L. var. sylvatica) infesting European rice fields. Medoyrzae—Interregional Cooperative Research Network on Rice in the Mediterranean Climate Areas. Montellier, France: FAO-CIHEAM-Institut Agronomique Mediterranéen 2–4.

Gao, L., and Innan, H. (2008). Nonindependent Domestication of the Two Rice Subspecies, Oryza sativa ssp. *indica* and ssp. *japonica*, Demonstrated by Multilocus Microsatellites. Genetics *179*, 965–976.

Garris, A.J., Tai, T.H., Coburn, J., Kresovich, S., and McCouch, S. (2005). Genetic structure and diversity in Oryza sativa L. Genetics *169*, 1631–1638.

Gealy, D.R., Mitten, D.H., and Rutger, J.N. (2003). Gene Flow Between Red Rice (Oryza sativa) and Herbicide-Resistant Rice (O. sativa): Implications for Weed Management 1. Weed Technology *17*, 627–645.

Gealy, D.R., Tai, T.H., and Sneller, C.H. (2002). Identification of red rice, rice, and hybrid populations using microsatellite markers. Weed Science *50*, 333–339.

Glaszmann, J.C. (1987). Isozymes and classification of Asian rice varieties. TAG Theoretical and Applied Genetics *74*, 21–30.

Gomez-Alvarez, V., Teal, T.K., and Schmidt, T.M. (2009). Systematic artifacts in metagenomes from complex microbial communities. The ISME Journal *3*, 1314–1317.

Graham, S.W., and Olmstead, R.G. (2000). Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. American Journal of Botany *87*, 1712–1730.

Graur, D., and Li, W.H. (2000). Fundamentals of molecular evolution (Sinauer Associates Sunderland, MA).

Hancock, J.F. (2012). Plant Evolution and the Origin of Crop Species (CABI).

Harlan, J.R. (1965). The possible role of weed races in the evolution of cultivated plants. Euphytica *14*, 173–176.

Harlan, J.R. (1992). Crops and Man, 2nd Edition (American Society of Agronomy-Crop Science Society).

Highfield, P.E., and Ellis, R.J. (1978). Synthesis and transport of the small subunit of chloroplast ribulose bisphosphate carboxylase.

Hiratsuka, J., Shimada, H., Whittier, R., Ishibashi, T., Sakamoto, M., Mori, M., Kondo, C., Honji, Y., Sun, C.R., Meng, B.Y., et al. (1989). The complete sequence of the rice (Oryza sativa) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Molecular and General Genetics MGG *217*, 185–194.

Hoagland, R.E., and Paul, R.N. (1978). A comparative SEM study of red rice and several commercial rice (Oryza sativa) varieties. Weed Science 619–625.

Huang, C.Y., Grünheit, N., Ahmadinejad, N., Timmis, J.N., and Martin, W. (2005). Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiology *138*, 1723–1733.

Huse, S.M., and Welch, D.B.M. (2011). Accuracy and quality of massively parallel DNA pyrosequencing. Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches 149–155.

Jarvie, T., and Harkins, T. (2008). 3K Long-Tag Paired End sequencing with the Genome Sequencer FLX System. Nature Methods *5*,.

Jarvis, D.I., and Hodgkin, T. (1999). Wild relatives and crop cultivars: detecting natural introgression and farmer selection of new genetic combinations in agroecosystems. Molecular Ecology *8*, S159–S173.

Kanno, A., Watanabe, N., Nakamura, I., and Hirai, A. (1993). Variations in chloroplast DNA from rice (Oryza sativa): differences between deletions mediated by short direct-repeat sequences within a single species. TAG Theoretical and Applied Genetics *86*, 579–584.

Kelchner, S.A. (2000). The evolution of non-coding chloroplast DNA and its application in plant systematics. Annals of the Missouri Botanical Garden 482–498.

Kelly Vaughan, L., Ottis, B.V., Prazak-Havey, A.M., Bormans, C.A., Sneller, C., Chandler, J.M., and Park, W.D. (2001). Is all red rice found in commercial rice really Oryza sativa? Weed Science *49*, 468–476.

Kircher, M., and Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. Bioessays *32*, 524–536.

Kumagai, M., Wang, L., and Ueda, S. (2010). Genetic diversity and evolutionary relationships in genus Oryza revealed by using highly variable regions of chloroplast DNA. Gene *462*, 44–51.

Kumar, S., Nei, M., Dudley, J., and Tamura, K. (2008). MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief. Bioinformatics *9*, 299–306.

Kuroda, Y., Sato, Y.I., Bounphanousay, C., Kono, Y., and Tanaka, K. (2005). Gene flow from cultivated rice (Oryza sativa L.) to wild Oryza species (O. *rufipogon* Griff. and O. *nivara* Sharma and Shastry) on the Vientiane plain of Laos. Euphytica *142*, 75–83.

Lander, E.S. (2011). Initial impact of the sequencing of the human genome. Nature *470*, 187–197.

Leister, D. (2005a). Origin, evolution and genetic effects of nuclear insertions of organelle DNA. TRENDS in Genetics *21*, 655–663.

Leister, D. (2005b). Origin, evolution and genetic effects of nuclear insertions of organelle DNA. TRENDS in Genetics *21*, 655–663.

Londo, J.P., Chiang, Y.-C., Hung, K.-H., Chiang, T.-Y., and Schaal, B.A. (2006). Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, Oryza sativa. Proc. Natl. Acad. Sci. U.S.A. *103*, 9578–9583.

Londo, J.P., and Schaal, B.A. (2007). Origins and population genetics of weedy red rice in the USA. Molecular Ecology *16*, 4523–4535.

Lu, H., Liu, Z., Wu, N., Berne, S., Saito, Y., Liu, B., and Wang, L. (2002). Rice domestication and climatic change: phytolith evidence from East China. Boreas *31*, 378–385.

Maier, R.M., Neckermann, K., Igloi, G.L., and Kössel, H. (1995). Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. Journal of Molecular Biology *251*, 614–628.

Majumder, N.D., Ram, T., and Sharma, A.C. (1997). Cytological and morphological variation in hybrid swarms and introgressed population of interspecific hybrids (*Oryza rufipogon* Griff.$\times$ Oryza sativa L.) and its impact on evolution of intermediate types. Euphytica *94*, 295–302.

Martin, W. (2003). Gene transfer from organelles to the nucleus: frequent and in big chunks. Proceedings of the National Academy of Sciences *100*, 8612–8614.

Martin, W., and Herrmann, R.G. (1998). Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiology *118*, 9–17.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., and Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proceedings of the National Academy of Sciences *99*, 12246.

Matsuo, M., Ito, Y., Yamauchi, R., and Obokata, J. (2005). The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast–nuclear DNA flux. The Plant Cell Online *17*, 665–675.

Morrell, P.L., Buckler, E.S., and Ross-Ibarra, J. (2011). Crop genomics: advances and applications. Nature Reviews Genetics *13*, 85–96.

Morton, B.R., and Clegg, M.T. (1995). Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. Journal of Molecular Evolution *41*, 597–603.

Neale, D.B., Saghai-Maroof, M.A., Allard, R.W., Zhang, Q., and Jorgensen, R.A. (1988). Chloroplast DNA diversity in populations of wild and cultivated barley. Genetics *120*, 1105–1110.

Notsu, Masood, Nishikawa, Kubo, Akiduki, Nakazono, Hirai, and Kadowaki (2002). The complete sequence of the rice (<SMALL>Oryza sativa</SMALL> L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. Molecular Genetics and Genomics *268*, 434–445.

Noutsos, C., Kleine, T., Armbruster, U., DalCorso, G., and Leister, D. (2007). Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. Trends in Genetics *23*, 597–601.

Noutsos, C., Richly, E., and Leister, D. (2005). Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. Genome Research *15*, 616–628.

Ogihara, Y., Isono, K., Kojima, T., Endo, A., Hanaoka, M., Shiina, T., Terachi, T., Utsugi, S., Murata, M., Mori, N., et al. (2002). Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. Molecular Genetics and Genomics *266*, 740–746.

Ogihara, Y., Terachi, T., and Sasakuma, T. (1988). Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. Proceedings of the National Academy of Sciences *85*, 8573.

Ogihara, Y., Terachi, T., and Sasakuma, T. (1991). Molecular analysis of the hot spot region related to length mutations in wheat chloroplast DNAs. I. Nucleotide divergence of genes and intergenic spacer regions located in the hot spot region. Genetics *129*, 873–884.

Oka, H.I. (1988a). Origin of cultivated rice.

Oka, H.I. (1988b). Origin of cultivated rice. (Japan Scientific Societies Press).

Olsen, K.M., Caicedo, A.L., and Jia, Y. (2007). Evolutionary genomics of weedy rice in the USA. Journal of Integrative Plant Biology *49*, 811–816.

Prashanth, S.R., Parani, M., Mohanty, B.P., Talame, V., Tuberosa, R., and Parida, A. (2002). Genetic diversity in cultivars and landraces of Oryza sativa subsp. *indica* as revealed by AFLP markers. Genome *45*, 451–459.

Provan, J., Powell, W., and Hollingsworth, P.M. (2001). Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends in Ecology & Evolution *16*, 142–147.

Pyke, K.A. (1999). Plastid division and development. The Plant Cell Online *11*, 549–556.

Rasmussen, D.A., and Noor, M.A. (2009). What can you do with 0.1× genome coverage? A case study based on a genome survey of the scuttle fly Megaselia scalaris (Phoridae). BMC Genomics *10*, 382.

Reagon, M., Thurber, C.S., Gross, B.L., Olsen, K.M., Jia, Y., and Caicedo, A.L. (2010). Genomic patterns of nucleotide diversity in divergent populations of U.S. weedy rice. BMC Evol. Biol. *10*, 180.

Richly, E., and Leister, D. (2004a). NUMTs in sequenced eukaryotic genomes. Molecular Biology and Evolution *21*, 1081–1084.

Richly, E., and Leister, D. (2004b). NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. Molecular Biology and Evolution *21*, 1972–1980.

Ronaghi, M., Uhlén, M., and Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. Science *281*, 363, 365.

Rounsley, S., Marri, P., Yu, Y., He, R., Sisneros, N., Goicoechea, J., Lee, S., Angelova, A., Kudrna, D., Luo, M., et al. (2009). De Novo Next Generation Sequencing of Plant Genomes. Rice *2*, 35–43.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics *16*, 944–945.

Second, G. (1982). Origin of the genic diversity of cultivated rice (Oryza spp.): study of the polymorphism scored at 40 isozyme loci. Jpn J Genet *57*, 25–57.

Shahid Masood, M., Nishikawa, T., Fukuoka, S., Njenga, P.K., Tsudzuki, T., and Kadowaki, K. (2004a). The complete nucleotide sequence of wild rice (< i> *Oryza nivara*</i>) chloroplast genome: first genome wide comparative sequence analysis of wild and cultivated rice. Gene *340*, 133–139.

Shahid Masood, M., Nishikawa, T., Fukuoka, S.-I., Njenga, P.K., Tsudzuki, T., and Kadowaki, K.-I. (2004b). The complete nucleotide sequence of wild rice (*Oryza nivara*) chloroplast genome: first genome wide comparative sequence analysis of wild and cultivated rice. Gene *340*, 133–139.

Shaw, J., Lickey, E.B., Schilling, E.E., and Small, R.L. (2007). Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. American Journal of Botany *94*, 275–288.

Shivrain, V.K., Burgos, N.R., Sales, M.A., and Kuk, Y.I. (2010). Polymorphisms in the< i> ALS</i> gene of weedy rice (< i> Oryza sativa</i> L.) accessions with differential tolerance to imazethapyr. Crop Protection *29*, 336–341.

Soll, J., and Schleiff, E. (2004). Protein import into chloroplasts. Nature Reviews Molecular Cell Biology *5*, 198–208.

Soltis, P.S., Soltis, D.E., and Doyle, J.J. (1992). Molecular Systematics of Plants (Springer).

Song, Z., Zhu, W., Rong, J., Xu, X., Chen, J., and Lu, B.R. (2006). Evidences of introgression from cultivated rice to *Oryza rufipogon* (Poaceae) populations based on SSR fingerprinting: implications for wild rice differentiation and conservation. Evolutionary Ecology *20*, 501–522.

Stupar, R.M., Lilly, J.W., Town, C.D., Cheng, Z., Kaul, S., Buell, C.R., and Jiang, J. (2001a). Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proceedings of the National Academy of Sciences *98*, 5099.

Stupar, R.M., Lilly, J.W., Town, C.D., Cheng, Z., Kaul, S., Buell, C.R., and Jiang, J. (2001b). Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proceedings of the National Academy of Sciences *98*, 5099.

Sweeney, M., and McCouch, S. (2007). The Complex History of the Domestication of Rice. Ann Bot *100*, 951–957.

Tang, J., Xia, H., Cao, M., Zhang, X., Zeng, W., Hu, S., Tong, W., Wang, J., Wang, J., Yu, J., et al. (2004). A comparison of rice chloroplast genomes. Plant Physiol. *135*, 412–420.

Tang, L.H., and Morishima, H. (1996). Genetics characteristics and origin of weedy rice. Paper on Origin and Dissemination of Cultivated Rice in China 211–218.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Tian, X., Zheng, J., Hu, S., and Yu, J. (2006). The rice mitochondrial genomes and their variations. Plant Physiology *140*, 401–410.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Reviews Genetics *5*, 123–135.

Ting, Y. (1957). The origin and evolution of cultivated rice in China. Acta Agron Sin *8*, 243–260.

Valverde, B.E. (2005). The damage by weedy rice-can feral rice remain undetected. Crop Ferality and Volunteerism 279–294.

Ward, B.L., Anderson, R.S., and Bendich, A.J. (1981). The mitochondrial genome is large and variable in a family of plants (Cucurbitaceae). Cell *25*, 793–803.

Waters, D.L.E., Nock, C.J., Ishikawa, R., Rice, N., and Henry, R.J. (2011). Chloroplast genome sequence confirms distinctness of Australian and Asian wild rice. Ecology and Evolution.

De Wet, J., and Harlan, J. (1975). Weeds and Domesticates: Evolution in the man-made habitat. Economic Botany *29*, 99–108.

Wolfe, K.H., Li, W.H., and Sharp, P.M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. PNAS *84*, 9054–9058.

Wolfe, K.H., Morden, C.W., and Palmer, J.D. (1991). Ins and outs of plastid genome evolution. Current Opinion in Genetics & Development *1*, 523–529.

Xia, H.-B., Wang, W., Xia, H., Zhao, W., and Lu, B.-R. (2011). Conspecific Crop-Weed Introgression Influences Evolution of Weedy Rice (Oryza sativa f. spontanea) across a Geographical Range. PLoS ONE *6*, e16189.

Yang, M., Zhang, X., Liu, G., Yin, Y., Chen, K., Yun, Q., Zhao, D., Al-Mssallem, I.S., and Yu, J. (2010). The complete chloroplast genome sequence of date palm (Phoenix dactylifera L.). PloS One *5*, e12762.

Yu, G., Bao, Y., Shi, C., Dong, C., and Ge, S. (2005). Genetic Diversity and Population Differentiation of Liaoning Weedy Rice Detected by RAPD and SSR Markers. Biochemical Genetics *43*, 261–270.

Yu, Y., Rambo, T., Currie, J., Saski, C., Kim, H.R., Collura, K., Thompson, S., Simmons, J., Yang, T.J., Nah, G., et al. (2003). In-depth view of structure, activity, and evolution of rice chromosome 10. Science *300*, 1566–1569.

Yuan, Q., Hill, J., Hsiao, J., Moffat, K., Ouyang, S., Cheng, Z., Jiang, J., and Buell, C. (2002). Genome sequencing of a 239-kb region of rice chromosome 10L reveals a high frequency of gene duplication and a large chloroplast DNA insertion. Molecular Genetics and Genomics *267*, 713–720.

Zambon, A.C., Zhang, L., Minovitsky, S., Kanter, J.R., Prabhakar, S., Salomonis, N., Vranizan, K., Dubchak, I., Conklin, B.R., and Insel, P.A. (2005). Gene expression patterns define key transcriptional events in cell-cycle regulation by cAMP and protein kinase A. PNAS *102*, 8561–8566.

Zhang, T., Zhang, X., Hu, S., and Yu, J. (2011). An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. Plant Methods *7*, 38.

Anexo.