UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Estudio de polimorfismos asociados a caracteres de interés agronómico en arroz (*Oryza sativa L.*) mediante técnicas de análisis genómico.

Tesis Doctoral

Presentada por

Juan Luis Reig Valiente

Para optar al grado de

Doctor en Biotecnología

Directora:

Dra. Concha Domingo Carrasco

Tutora:

Dra. Belen Picó Sirvent

Valencia, enero de 2019

Agradecimientos

En primer lugar me gustaría agradecer a dra. Concha Domingo haberme dado la oportunidad de trabajar en el IVIA, realizar el doctorado y a su gran dirección a lo largo de este tanto en el aspecto científico como humano. También en este sentido quisiera darle las gracias al dr. Manuel Talón, por su proximidad y disponibilidad a la hora de ayudar ante cualquier cuestión o problema que hayan surgido a lo largo de mi etapa en el IVIA. A Belen Picó me gustaría agradecerle haber accedido a ser mi tutora de tesis permitiéndome así realizar el doctorado en la Universidad Politècnica de València.

Además quiero darle las gracias a todos aquellos que me han ayudado en las diferentes tareas realizadas lo largo de la tesis: a Pilar Montero con los experimentos de campo, a Quico Tadeo por ayduarme con cualquier duda que surgiese, Javier Terol haber estado siempre disponible para cualquier problema de bioinformática, a Isabel Sanchís por ayudarme con los protocolos del laboratorio, a Álvaro García por lo cruces de las variedades de arroz, a Ángel Boix tanto por el trabajo de campo como de invernadero, a Toni Prieto por su ayuda en la medida de cloruros y pedidos, a Antonio López por el trabajo de invernadero, a Victoria Ibañez por su ayuda con las cuestiones de genómica, a Estela Pérez por su ayuda con el IGV y las extracciones de ADN nuclear, a Daniel Ventimilla por ayudarme con las RT-qPCRs y a Carles Borredá por haberme permitido emplear su programa “Allinone”. Finalmente quisiera agradecer a todo el personal del Departamento del Arroz, en Sueca tanto fijo como temporal, por su ayuda con el trabajo de campo y la recogida de datos.

Quiero agradecer a todo el mundo con el que he convivido en el IVIA haber hecho que esta época haya sido para mí realmente agradable y pienso que enriquecedora.

Para finalizar quiero agradecer a mi madre y a mi hermano su apoyo, cariño y comprensión durante estos años.

Abreviaturas:

ADN Acido desoxirribonucleico

ADNc ADN complementario

ARN Ácido ribonucleico

ARNm ARN mensajero

DH Tiempo de floración

EDTA Etil –diamino-tetraacetato

FDR Tasa de descubrimientos falsos, *False discovery rate*

GN Número de granos por panícula

GO Ontología génica

GWAS Estudios de asociación de genoma completo, *Genome Wide Association studies*

H Altura de la planta

Ha Hectáreas

IAEA/FAO Organización Internacional de la Energía Atómica/Organización de las Naciones Unidas para la Alimentación y la Agricultura

IRRI International Rice Ressearch Institute

IVIA Instituto Valenciano de Investigaciones Agrarias

QTL Locus de un carácter cuantitativo, Quantitative Trait Locus

LD Desequilibrio de ligamiento

MAF Frecuencia del alelo menos común, *Minimum Allele Frequency*

MLM Modelo linear mixto, *Mixed linear model*

MAPAMA Ministerio de Agricultura, Pesca y Alimentación

MQ Calidad de mapeo

ncRN ARN nucleolar

NGS Secuenciación de nueva generación

Nt Nucleótidos

PC Componente principal

PCA Análisis de componentes principales.

PCR Reacción en cadena de la polimerasa

PL Longitud de panícula

PN Número de panículas

RT-qPCR Reacción en cadena de la polimerasa cuantitativa, con transcriptasa inversa, *quantitative Retrotrasncription-Polymerae Chain Reaction*

SDS Sodio dodecil sulfato

SNP Polimorfismos de un solo nucleótido, *Single Nucleotide Polymorphism*

Tris 2-amino-2-hidroximetil-1,3-propanodiol

UTR *Untranslated Region*

Índice

Resumen

El arroz es uno de los cultivos más importantes para la humanidad siendo la principal fuente de calorías para una gran parte de la población mundial. Los continuos cambios en las demandas del sector arrocero y las predicciones de nuevas condiciones climáticas requieren la adaptación del cultivo a través de nuevas variedades. Los avances en investigación en arroz y las herramientas genómicas desarrolladas en los últimos años han permitido la modernización de la mejora de variedades, haciéndola dirigida y rápida. Los programas de mejora suelen realizarse de manera local ya que las plantas de arroz son muy sensibles a las condiciones ambientales, como el fotoperiodo. Por ello se utilizan habitualmente parentales ya adaptados a las regiones de cultivo locales, es decir, variedades que se cultivan en regiones de clima templado. La mayoría de las variedades cultivadas en las zonas templadas son poco o nada sensibles a este. Las regiones de clima templado donde se cultiva arroz albergan suficiente diversidad fenotípica y genotípica por explotar, y su conocimiento y caracterización es conveniente para poder facilitar los programas de mejora que llevan a cabo en estas regiones. En esta tesis se ha generado una colección de 193 variedades de arroz de tipo *japonica templada* representativas de la diversidad de esta región. También se ha desarrollado un panel de SNPs apto para el genotipado de variedades de tipo *japonica* con el cual se ha genotipado dicha colección. El análisis de la estructura poblacional de esta colección indica que las variedades cultivadas en clima templado pueden dividirse en cuatro grupos, basándose en el tipo de grano y el origen geográfico. Un grupo está formado por variedades de grano largo, mientras que las de grano medio se dividen en otros tres subgrupos compuestos por un primer grupo de variedades americanas y australianas, un segundo grupo formado por variedades italianas y finalmente un tercer grupo de variedades de origen asiático y variedades antiguas europeas. El análisis de las relaciones genéticas puso de manifiesto que la estructura genética observada está

ligada a la historia de la mejora del arroz. Los resultados del análisis de la estructura poblacional han sido de utilidad a la hora de realizar el estudio de asociación, ya que la fuerte estructura poblacional observada podría causar sesgos en las asociaciones llevando a falsos positivos. Se detectó una asociación de 43 SNPs, asociados a la variación en caracteres relacionados con el rendimiento y la floración.

Puesto que el fotoperiodo destaca como uno de los principales factores relacionado con la adaptación del cultivo a climas templados, como objetivo de la tesis se ha planteado la identificación de nuevos componentes reguladores de la floración. Con esta finalidad se ha generado, identificado y caracterizado fenotípica y genotípicamente una línea mutante de la variedad Gleva, ampliamente cultivada en nuestro territorio, que presenta un fenotipo de floración temprana. Mediante la combinación del uso de las técnicas Mutmap y Allinone, empleando la secuenciación de genoma completo de plantas en una generación F2 derivada del cruce entre el parental silvestre y el mutante, G123, se ha procedido además a la identificación de la mutación candidata responsable de la variación en el tiempo de floración que resultó ser una deleción en el cromosoma uno en la región en la que sitúa el gen PHOTOSENSITIVITY 13. De este modo se ha encontrado una variación interesante para la mejora y se ha desarrollado una metodología para la rápida identificación de mutaciones.

Así pues, durante el desarrollo de esta tesis se han empleado técnicas genómicas para el estudio de caracteres de interés agronómico y su aplicación en la mejora de las variedades cultivadas en nuestra región.

Resum

L'arròs és un dels cultius més importants per a la humanitat seient la principal font de calories per a gran part de la població mundial. Els continus canvis a les demandes del sector arrosser i les prediccions de noves condicions climàtiques requereixen l'adaptació del cultiu a través de noves varietats. Els avanços en la investigació en arròs i les ferramentes genòmiques desenvolupades els darrers anys han permés la modernització de la millora de les varietats, fent-la dirigida i ràpida. Els programes de millora solen realitzar-se de manera local, ja que les plantes d'arròs són molt sensibles a les condicions ambientals, com el fotoperíode. Per això habitualment s'utilitzen parentals ja adaptats a les regions de cultiu locals, és a dir, varietats que es cultiven a les regions de clima temperat. La majoria de les varietats cultivades a les zones temperades són poc sensibles a aquest. Les regions en clima temperat on es cultiva l'arròs albergen suficient diversitat fenotípica i genotípica per explotar, i el seu coneiximent i caracterització es convenient per tal de facilitar els programes de millora en aquestes regions. En aquesta tesi s'ha generat una col·lecció de 193 varietats d'arròs de tipus japonica temperat representativa de la diversitat d'aquesta regió. També s'ha generat un panel de SNPs apte per al genotipat de varietats tipus japonica amb el qual s'ha genotipat dita col·lecció. L'analisisl'estructura poblacional d'aquesta col·lecció indica que les varietats cultivades al clima temperat poden dividir-se en quatre grups, basa'n't-se en el tipus de gra i l'origen geogràfic. Un grup està format per varietats de gra llarg, mentres que les de gra mitja es divideixen en altres tres subgrups composats per un primer grup de varietats americanes i australianes un segon grup es format per varietats italianes i finalment un tercer grup de varietats d'origen asiàtic i varietats antigues europees. L'anàlisi de les relacions genètiques va posar de manifest que l'estructura genètica observada està lligada a l'història de la millora de l'arròs. Els resultats de l'anàlisi de l'estructura poblacional han sigut d'utilitat a l'hora de realizar l'estudi d'associació, ja que la fort estructura poblacional

observada podría causar biaixos a les associacions donant lloc a fals positius. Es detectà l'associació de 43 SNP, associats a les variacions en caràcters relacionats amb el rendiment i la floració.

Donat que el fotoperiode destaca com un dels principals factors relacionts amb la adaptació del cultiu al clima temperat com objectius de la tesis s'hi ha planteat la identificació de nous components reguladors de la floració para el qual s'ha generat, identificat i caracteritzat fenotípica i genotípicament una línia mutant de la varietat Gleva, àmpliament cultivada al nostre territori, que presenta un fenotip de floració primerenca. Mitjançant la combinació les tècniques Mutmap i Allione, emprant la seqüenciació del genoma sencer de les plantes en una generació F2 derivada del creuament entre els parentals salvatje i el mutant *G123* s'ha procedit a més a la identificació de la mutació candidata responsable de la variació en el temps de floració en una generació F2 derivada del encreuament entre el parental silvestre i el mutant, G123. D'aquest mode s'ha trobat una variació interessant per a la millora i s'hi ha desenvolupat una metodologia per a la ràpida identificació de mutacions.

Així doncs per al desenvolupament d'aquesta tesi s'hi ha empleat tècniques genòmiques per a l'estudi de caràcters d'interés agronòmic i la seua aplicació a la millora de les varietats cultivades a la nostra regió.

Abstract

Rice is one of the most important plants for mankind being the main source of calories for an extensive part of world population. The continuous changes at rice sector and previsions of new climatic conditions require the adaptation of the culture through new varieties. The advances in rice research and genomic tools developed in last past years have allowed the modernization of variety improvement, becoming fast and directed. Breeding programs are often carried on at local level because rice plants are very sensible to environmental conditions, as photoperiod. By this reason adapted to local cultivation regions parentals are usually employed, this means, varieties cultivated at temperate climate. Most of varieties cultivated in temperate areas are a little or non sensible to photoperiod. Temperate climate regions where rice is cultivated hold enough phenotypic and genotypic diversity to explode and it's knowledge and characterization is convenient in order to facilitate the breeding programs carried on this region. In this thesis a collection of 193 japonica temperate varieties representative of the diversity present at temperate regions was generated. Also a panel of SNPs valid to genotype *japonica* varieties was developed and used to genotype the collection. The structure analysis of the collection showed that the varieties cultivated in temperate climate can be divided in four groups, based on the type of grain and geographical origin. A group was composed of long grain varieties, whilst medium grain varieties are divided into other three subgroups a first group of American and Australian, a second group was formed by Italian varieties and a third group from Asian origin varieties and old European varieties. The analysis of genetic relationships showed that the genetic structure observed is linked to the rice breeding history. The results of the structure analysis were usefull when performing an association study, since the observed strong population structure could cause biased associations producing false positives. Association between 43 SNPs and the variation of traits related with yield and flowering was detected.

Since photoperiod stands out as one of the main factors related with cultive adaptation to temperate regions, as an objective of this thesis was the identification of new regulator components of flowering has been proposed. For this objective a mutant line from Gleva variety, widely cultivated across our territory, was generated, identified and characterized. Using a combination of Mutmap and Allione techniques, employing complete genome sequencing of plants in a F2 derivated from a cross between wild parental and the mutant, G123. As a result, an interesting variation for breeding has been found and a fast methodology for mutation identification has been developed.

For the development of this thesis genomic techniques for the study of traits of agronomic interest have been used and applied for the breeding of varieties cultivated in our region.

1. INTRODUCCIÓN

1.1. Importancia económica del arroz

El arroz está considerado uno de los cultivos más importantes para la humanidad puesto que aporta el 20% de las calorías a la población mundial (Garris, Tai, Coburn, Kresovich, & McCouch, 2005). En algunos lugares del sudeste asiático la población depende de este cultivo llegando a suponer más del 70% del aporte de calorías de la dieta (htttp://ricerp.org).

A nivel mundial, su producción se ha triplicado entre 1961 y 2016, pasando de 215 millones de toneladas a 740 millones, mientras que el área total destinada a su cultivo ha aumentado un 38% pasando de 115 a 159 millones de ha, lo que indica un aumento en el rendimiento de 2,5 veces (FAOSTAT, 2018). A día de hoy el arroz representa el 30% de la producción mundial de cereales y su mercado está valorado en 206.000 millones de dólares, el 13% del mercado de cultivos.

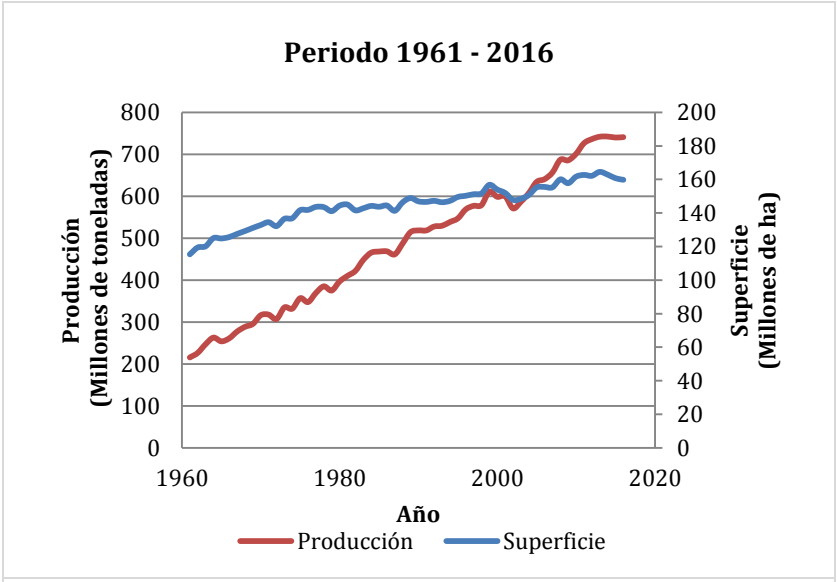


Figura 1: Producción de arroz en millones de toneladas frente a la superficie empleada para su cultivo en millones de hectáreas, fuente: http://www.fao.org.

En 2016 el mayor productor mundial de arroz fue China, seguido de India, Indonesia, Bangladesh, Vietnam, Myanmar, Tailandia, Filipinas, Brasil y Paquistán. En conjunto esto 10 países suman el 85% de la producción mundial, reflejando la localización del cultivo. España está entre los productores moderados a nivel mundial, produciendo tan solo el 0,12 % de la producción total de arroz.

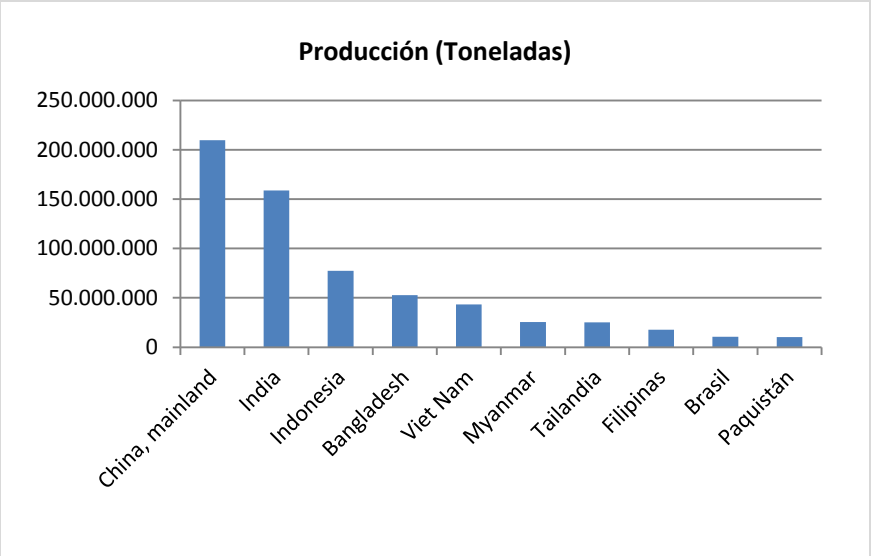


Figura 2: Producción de los 10 mayores productores de arroz en el año 2016, fuente: http://www.fao.org.

Si analizamos la producción de arroz por regiones en el año 2016 es evidente que la mayor parte de la producción mundial tiene lugar en Asia, produciéndose hasta el 90% del total, seguido muy de lejos de América con el 5%, África con el 4%, Europa con 1% y Oceanía no llegaría a representar ni el 1% de la producción total.

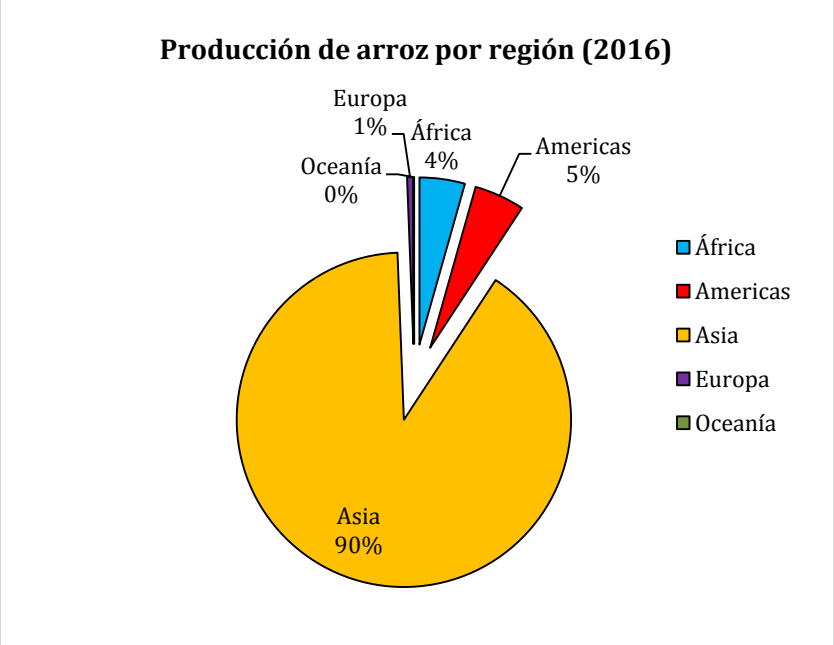


Figura 3: Producción de arroz por región geográfica, fuente: http://www.fao.org.

A nivel Europeo el principal productor es Italia con el 52% de la producción, seguido de España con un 28%, Grecia con 7,4%, Portugal con 5,5%, Francia con 2,6%, Bulgaria con 2,1%, Rumanía con 1,43% y Hungría con 0,3%.

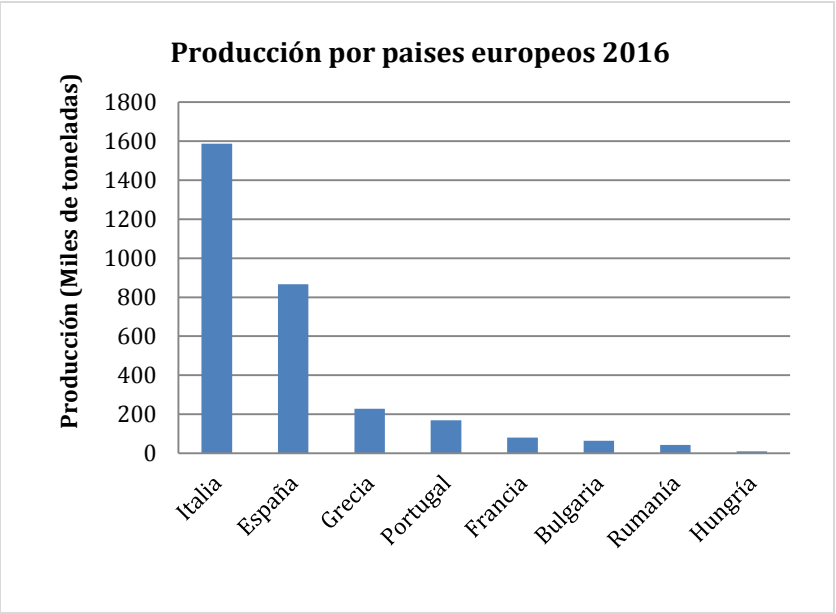


Figura 4: Países europeos productores de arroz, fuente http://www.fao.org.

A nivel mundial la cantidad de importaciones y exportaciones respecto a la producción total es muy baja (<9%) al existir una oferta y una demanda muy ajustadas. Existen pocos compradores y vendedores, siendo las exportaciones de India, Paquistán, Tailandia, EEUU y Vietnam el 80% del comercio mundial de arroz.

1.2. El cultivo en España

En España, las principales comunidades autónomas productoras de arroz son Andalucía, Extremadura, Cataluña, Comunidad Valenciana y Aragón (Tabla 1), aunque también se produce en el resto de territorios. La producción total según los datos del Ministerio de Agricultura, Pesca y Alimentación (www.mapama.gob.es) en el año 2016 fue de 835.400 toneladas, de las cuales 355.350 fueron de arroz de grano largo 480.050 de grano medio, en una superficie total de 109.272 ha, las cuales representan un 0,64 % de la superficie dedicada a cultivo.

Tabla 1: Datos MAPAMA por comunidad autónoma 2016 (http://www.mapama.gob.es).		
Comunidades Autónomas	**Superfice total (Hectáreas)**	**Producción total (toneladas)**
Andalucía	40.108	364.649
Extremadura	24.652	163.939
Cataluña	20.861	135.531
C. Valenciana	15.400	122.399
Aragón	5.485	30.564
Navarra	2.153	14.690
R. de Murcia	452	2.707
Castilla la mancha	133	865
Baleares	28	56
TOTAL	109.272	835.400

En el año 2013 se registraron cerca de 9.000 explotaciones de arroz, de las cuales el 37% se concentra en Valencia, el 23% en Cataluña, un 21% en Extremadura y el 11% en Andalucía. El 96,7% de las explotaciones tienen menos de 50 hectáreas de arroz y concentran el 62% de la superficie total de arroz. Pero hay 4 explotaciones que superan las 500 hectáreas las cuales se encuentran en Andalucía. El 81% de las hectáreas totales de arroz se concentra en explotaciones de menos de 100 hectáreas de arroz (MAPAMA).

La mayor parte del arroz que se comercializa lo hace a través de cooperativas. Siendo 7 las envasadoras las que facturan el 90% del mercado español.

España se puede considerar un país exportador de arroz, en las últimas en las últimas 5 campañas ha exportado 240.077 toneladas frente a un total de 79.449 toneladas importadas. La mayor parte del arroz de grano medio/redondo que se produce en España se destina a consumo interno, siendo el de grano largo el más exportado.

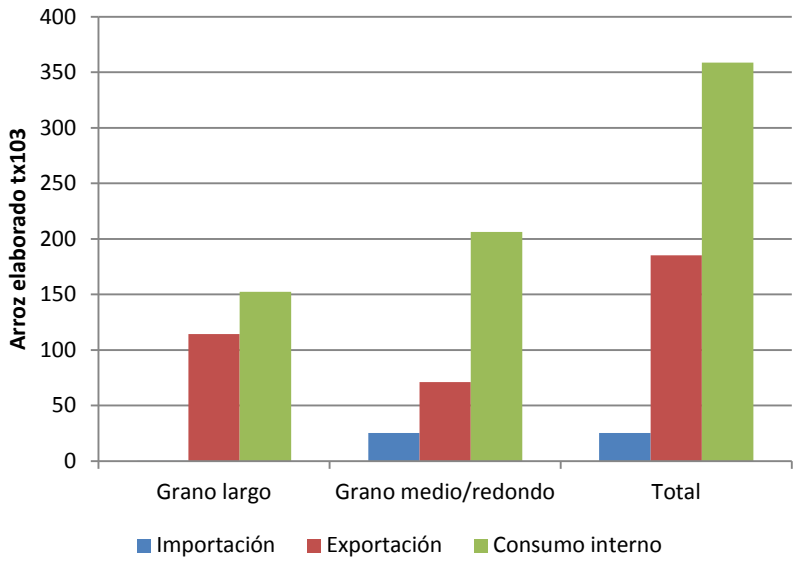


Figura 5: Datos provisionales de importación exportación y consumo interno de la temporada 2016/17 (https://www.mapama.gob.es).

1.3. Origen, domesticación y diversificación.

1.3.1. Origen, domesticación y diversificación.

El arroz más cultivado hoy en día tiene origen asiático y presenta una gran diversidad. Entre las subpoblaciones genéticas pueden encontrarse cinco grupos: *indica, aus, japónica tropical, japónica templada y aromatico* (Garris et al., 2005), de los cuales los más cultivados hoy en día son *indica y japonica* (Sweeney & McCouch, 2007). El origen de la domesticación del arroz asiático es incierto y hay un debate abierto a día de hoy. Los datos arqueológicos y botánicos sugieren que la subpoblación *japonica* fue domesticada primero, hace unos 7.000 años en la cuenca del rio Yangtze en China, mientras que las variedades *indica* fueron domesticadas posteriormente, hace unos 4000 años en las llanuras del Ganges (Fuller et al., 2011). De manera adicional,

estudios genómicos aportan nuevos datos que indican que cada grupo/supoblación de variedad de arroz asiático (aus, *indica* y *japonica*) provienen de distintas subpoblaciones de arroz silvestre (*O. nivara* o *O. rufipogon*) (Choi et al., 2017; Huang, Kurata, et al., 2012). A su vez, el arroz silvestre podría dividirse en 3 subpoblaciones principales Or-I, Or,II y Or-III según Huang *et al.*(2012). Filogenéticamente, las variedades *japonica* estarían más emparentada con el grupo Or-III, mientras que *aus* e *indica* con Or-I, pero cada una formaría un grupo monofilético con un subgrupo diferente de muestras Or-1 (Huang, Kurata, et al., 2012). A día de hoy se postulan dos modelos respecto a la domesticación (Choi & Purugganan, 2018). En el primer postulado, llamado "domesticación *de novo* con hibridación" por Huang *et al.* (2012), el arroz asiático se habría domesticado en una ocasión y las variedades que surgidas posteriormente se formaron a partir de introgresiones provenientes de progenitores silvestres. En el segundo modelo, propuesto por Civán *et al*. (2015), "modelo de domesticación múltiple" cada variedad fue domesticada de manera independiente en diferentes partes de Asia.

Independientemente del origen de la domesticación, la expansión de la especie, la adaptación, el aislamiento geográfico y la reducción del flujo génico debido a la reproducción autógama, han dado lugar a un aumento en las distancias genéticas entre los grupos, llegándose a generar barreras reproductivas (OKA, 2008) entre ellos, y, además, una fuerte estructura poblacional y una gran diversidad dentro de los subgrupos.

En Asia, las variedades *indica* son cultivadas predominantemente en la zona de India, Sri lanka, Tailandia, Malasia y los países adyacentes pero también son cultivadas en otras regiones del mundo con clima tropical. Son plantas altas, de tallo débil, sensibles a fotoperiodo, con dehiscencia fácil y hojas anchas y caídas. Este grupo hoy presenta una mayor diversidad que los otros. El motivo de esta diversidad parece deberse a la ausencia de cuellos de botella severos debido al flujo génico con variedades silvestres

o al haber tenido un tamaño poblacional más grande puesto que la principal ruta de dispersión fue por tierra. Las variedades *aus* posiblemente tienen un origen similar a las *indicas*, pero son plantas adaptadas al cultivo en zonas más septentrionales, principalmente en Bangladesh (Parsons, Newbury, Jackson, & Ford-Lloyd, 1999), a diferencia de las *indica* estas plantas han perdido la sensibilidad a fotoperiodo adquiriendo así la capacidad de florecer en condiciones de día largo en verano. Esta adaptación ha generado un aislamiento reproductivo temporal respecto a las *indica.* Respecto a las variedades *japonica* pese a dividirse en dos subgrupos, templadas y tropicales, es un grupo con menor diversidad que el *indica* (Glaszmann, 1987; Q. Zhang, Maroof, Lu, & Shen, 1992). En el caso de las variedades *japónica tropical,* esta diversidad menor puede deberse a que se cultivan mayormente en islas, indicativo del uso de una ruta marítima en su expansión, y una serie de cuellos de botella sucesivos dado su aislamiento. A su vez los datos moleculares parecen indicar que las variedades *japonica templadas* derivarían de las variedades del grupo *japonica tropicales*, que se habrían adaptado a la floración en los días largos y cálidos de verano a fin de evitar el frio de los inviernos en las zonas de clima templado. Las variedades *japonica templadas* se cultivan en Europa, Norteamérica y Australia. En Asia, se cultivan principalmente en la zona del valle del Yangtze de China, Korea y Japón. Tienen más hojas, menos tallos, son relativamente resistentes a la dehiscencia y al frío. Los granos son cortos y ricos en amilosa por lo que el arroz se vuelve pegajoso al cocinarlo. Finalmente el grupo aromático ha sido clasificado como un grupo intermedio entre *indica* y *japónica* (Ahuja, Panwar, Uma, & Gupta, 1995), pero también como grupo propio según diferente investigaciones (Jain, Jain, & McCouch, 2004). Tienen una relación próxima a las *japonica*. Y presentan una elevada proporción de loci monomórficos, lo que sugiere el efecto de un fuerte cuello de botella muy reciente (Garris et al., 2005; Nagaraju, Kathirvel, Kumar, Siddiq, & Hasnain, 2002).

Finalmente, hay que indicar que existe otro tipo de arroz, el arroz africano (*Oryza glaberrima*), que se originó de manera independiente al arroz asiático en el oeste de África a partir de cruces de *Oryza barthii* con otras especies silvestres de la zona (Cubry et al., 2018). Los llamados comúnmente arroz asiático y arroz africano constituyen especies distintas.

1.3.2. La regulación de la floración en el arroz y su papel en la diversificación.

La regulación del momento de floración en arroz ha tenido un papel tremendamente importante en la historia de su expansión y diversificación, siendo su principal mecanismo de adaptación al cultivo en regiones septentrionales a fin de evitar las bajas temperaturas del invierno. El origen de la domesticación del arroz tuvo lugar en una región de clima tropical en el que el día y la noche tienen más o menos la misma duración y hay pocas variaciones en temperatura y humedad. Es por ello que el arroz es considerado como una planta de día corto. La catalogación de una planta como de día largo o corto depende de su respuesta al fotoperiodo. El concepto de fotoperiodo fue definido por Garner y Allard en 1920 (Garner & Allard, 1920) como la longitud de día favorable para cada organismo y, puesto que el fenómeno se ha observado en animales y plantas, el fotoperiodismo es la respuesta de los organismos a la longitud relativa del día y la noche. Dependiendo de la cantidad de horas de luz que las plantas necesiten para florecer se clasifican en tres tipos: plantas de día largo, aquellas que florecen cuando las horas de luz superan el umbral de fotoperiodo, plantas de día corto, aquellas que florecen cuando las horas de luz son menores que las del umbral crítico y plantas de fotoperiodo neutro que florecen independientemente de la longitud del día. Como ya se ha dicho, el arroz es considerado una planta de día corto, y sin embargo se cultiva en verano en las zonas templadas, cuando el día es más largo que la noche. Este fenómeno es resultado de un proceso de adaptación y selección al cultivo a los días largos del verano, principalmente para evitar el frío invierno en estas

regiones. Como ya se ha indicado, la domesticación del arroz ocurrió en una región de clima tropical desde donde se expandió a otras regiones con diferente clima, cultivándose actualmente desde la latitud 55° N hasta 36°S (Khush, 1997). Este proceso de adaptación consistió básicamente en la pérdida paulatina de sensibilidad al fotoperiodo.

La transición de fase vegetativa a reproductiva en arroz está gobernada por la acción de dos genes máster, *Heading date 3a* (*Hd3a*) y *RICE FLOWERING LOCUS T 1* (*RFT1*), que codifican los llamados florígenos que señalizan el momento de floración. Una fina y compleja regulación de la expresión de estos dos genes gobierna la floración de la planta de arroz, siendo *Hd3a* el inductor de la floración en condiciones de día corto y *RTF1* en condiciones de día largo (Komiya et al, 2009). La expresión de *Hd3a* y *RFT1* está modulada mayormente por dos genes que representan dos rutas independientes, *Heading date 1* (*Hd1*), regulado por el ciclo circadiano, y *Early heading date 1* (*Ehd1*) que es un integrador de señales diferentes. Los niveles de expresión de *Hd1* están regulados por *Gigantea* (*OsGi*), un gen regulado por el ciclo circadiano que promueve la expresión de *Hd1* (Hayama, Yokoi, Tamaki, Yano, & Shimamoto, 2003). *Hd1* inhibe la floración, de manera que la sobreexpresión de *OsGi* resulta en un aumento de los niveles de *Hd1* resultando en la inhibición de la floración tanto en fotoperiodo corto como en fotoperiodo largo. El mismo efecto se observa al sobreexpresar *Hd1*. *Ehd1* induce la floración al activar la transcripción de *Hd3a* y *RFT1*. Hasta hace poco se postulaba que la expresión de *Ehd1* era independiente de *Hd1,* sin embargo, datos recientes indican que *Hd1* es capaz de regular los niveles de expresión de *Ehd1*(Goretti et al., 2017).

El gen *Ghd7 (Grain number plant height and heading date 7)* codifica una proteína con dominio CCT que se expresa en cantidades elevadas en condiciones de día largo. *Ghd7* modula negativamente la expresión de *Ehd1* ya que se ha observado que cuando los niveles de *Ghd7* son elevados, *Ehd1* se expresa poco (Xue et al., 2008).

Los estudios de la regulación de *Ehd1* y *Ghd7* parecen indicar que existe un sistema de ventanas de inducción y represión reguladas en primer lugar por *OsGI* (figura 6). *OsGI* induciría la expresión de *Ehd1* y *Ghd7* entorno al amanecer. En día corto la inducción de estos genes se da durante la noche y *Ehd1* induce la expresión de *Hd3a*. Sin embargo en condiciones de día largo el máximo de expresión de *Ghd7* se daría en el mismo momento, pero su inducción se mantendría al recibir luz del amanecer, llegando a presentar niveles superiores a los que se detectan en día corto, con lo que sería capaz de reprimir la expresión de *Ehd1* y de esta manera no es capaz de inducir a *Hd3a.* Estos datos sugieren que es necesaria la presencia de fitocromos activos para la correcta expresión de *Ghd7*, de manera que la inhibición de la floración por la luz roja se ejerce a través de *Ghd7*.

El arroz tiene tres fitocromos, PHYA, PHYB y PHYC (Makoto Takano et al., 2005), los mutantes *se5* presentan una actividad reducida de los tres fitocromos puesto que la enzima que codifica es la principal implicada en la conversión del grupo hemo a biliverdina IX α, pero no la única (M. Takano et al., 2009). Por lo tanto la síntesis del cromóforo, la fitocromobilina, en el mutante *se5* se ve reducida. Los análisis indican que los fitocromos no son necesarios para la determinación del inicio de la fase sensible para la expresión de *Ghd7.* Sin embargo, los homodímeros PhyA y los heterodímeros PhyB-PhyC independientemente son suficientes para activar la transcripción de *Ghd7* mientras que los PhyB pueden reprimirla (Osugi, Itoh, Ikeda-Kawakatsu, Takano, & Izawa, 2011).

Hd1 junto con Ghd7 reprimen la expresión de *Ehd1* únicamente por las mañanas en condiciones de fotoperiodo largo. Se ha demostrado que Hd1 y Ghd7 forman un complejo heterodimérico mediante el cual Ghd7 puede unirse al promotor de *Ehd1* a través de una interacción física con el dominio específico de monocotiledóneas CCT. Esto explicaría parcialmente la capacidad activadora o inhibidora de Hd1 según la duración del día (Nemoto, Nonoue, Yano, & Izawa, 2016). La actividad represora de

Hd1 sobre *Ehd1* necesita un *Ghd7* funcional bajo condiciones de fotoperiodo largo para llevarse a cabo, mientras que la función activadora se da por la noche y, por lo tanto, tanto en condiciones de fotoperiodo corto como de fotoperiodo largo. En ausencia de un *Ehd1* funcional, bajo condiciones de fotoperiodo largo, HD1 actúa como un fuerte represor (Doi et al., 2004; Endo-Higashi & Izawa, 2011). Por contra *Ghd7* es capaz de reprimir por si solo a *Ehd1, Hd3a* y *RFT1* durante la mañana bajo cualquier condición de fotoperiodo. Este hecho sugiere que posiblemente existan proteínas con las que Ghd7 interacciona para ejercer esta función. La expresión de *Hd1* y *Ghd7* está regulada prácticamente de forma independiente. Sin embargo la actividad represora del complejo Hd1-Ghd7 podría estar modificada por lo PHYB. Se ha observado que la represión de *Hd3a* debida a la interrupción de la noche es mucho mayor en las plantas portadoras del alelo funcional *Hd1* que en las que portan un alelo no funcional, sin embargo la represión de *Ehd1* debida a esta interrupción no parece verse claramente afectada por el alelo de Hd1 que se porte.

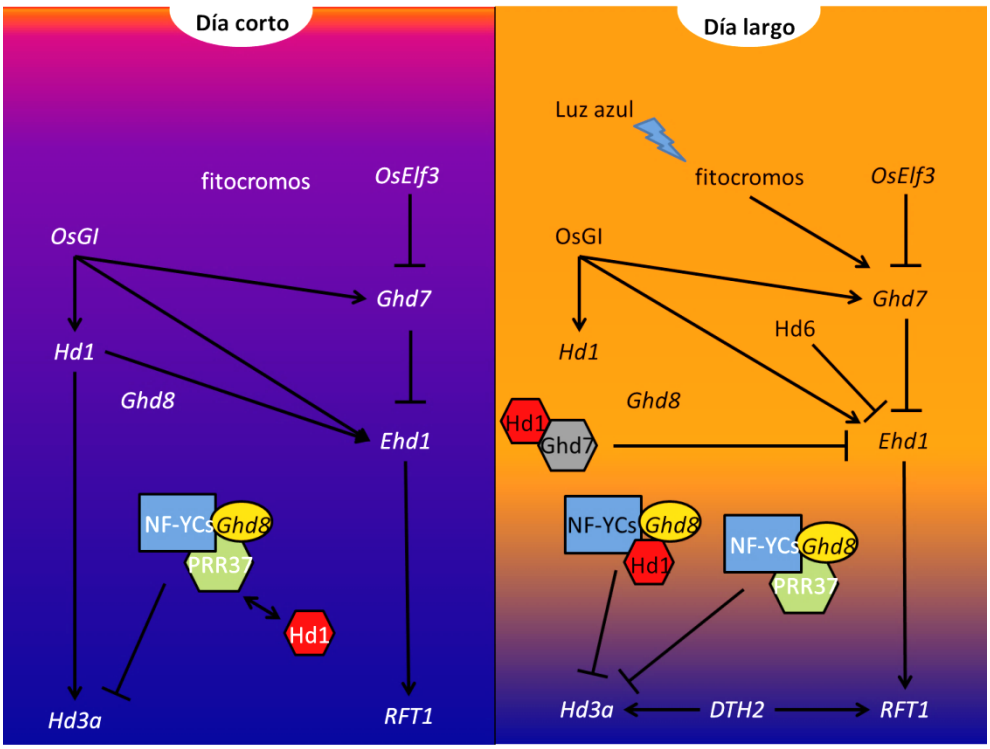


Figura 6: Modelo de regulación de la floración en arroz mediante fotoperiodo. Las flechas indican inducción de la transcripción mientras que las barras planas la represión. Los hexágonos indican dominios CCT de las proteínas, HD1 y Prr37 podrían competir por la unión al complejo heterotrimérico y su unión al promotor de *Hd3a*.

Hd17/ELF3 es un represor de *Ghd7* bajo condiciones día corto y día largo, y por lo tanto su expresión conlleva un aumento en los niveles de *Ehd1*. Sin embargo el efecto sobre la floración es indirecto ya que está relacionado con la regulación de la fase de crecimiento vegetativo sin afectar a la sensibilidad fotoperiodo lo cual sugiere que debe estar implicado en una ruta autónoma en arroz (Fu et al., 2009) afectando la función del reloj circadiano (Saito et al., 2012; Yang, Peng, Chen, Li, & Wu, 2013; J. Zhao et al., 2012). *HD5/DTH8/Ghd8* codifica una HEME ACTIVATOR PROTEIN 3 (HAP3) que es una subunidad del complejo del factor de transcripción CCAAT-box-binding.

Actúa como represor de la floración bajo condiciones de día largo, retrasando la floración mediante la inhibición de la expresión de *Ehd1* y, consecuentemente, de *Hd3a* y *RFT1* (Wei et al., 2010). Por el contrario, en condiciones de SD, se ha observado que *Ghd8* induce la expresión de estos reguladores (Yan et al., 2011). La expresión de *Ghd8* no está afectada por *Ghd7* ni *Hd1*, lo que indica una ruta genética distinta en el control de la floración (Wei et al., 2010). *OsPrr37* es un gen implicado en el ciclo circadiano que presenta un dominio CCT con homología al CCT de HD1 y homología estructural a NF-YA (Gao et al., 2014; Koo et al., 2013; Petroni et al., 2012). *OsPrr37* reprime la floración mediante la inhibición de la expresión de *Hd3a*. Se ha demostrado que las proteínas Hd1, Ghd8 y OsPRR37 se ensamblan en complejos proteicos tipo NF-Y de un orden superior. HD1 se une a Ghd7 para modular *Ehd1,* al igual que HD1, Ghd8 y OsPRR37 actúan formando trímeros OsPrr37/NF-C1/Ghd8, OsPrr37/NF-YC7/Ghd8, Hd1/NF-YC1/Ghd8 y Hd1/NF-YC7/Ghd8, que se unen a la región promotora de *Hd3a*. El hecho de que existan heterodímeros diferentes indica la posible existencia de varias dianas de complejos OsNF-Y dentro de la red de regulación de la floración (Goretti et al., 2017). Otro gen implicado en la regulación de la floración es *Hd6,* que codifica una subunidad α de una proteína kinasa CK2, y está implicado en la sensibilidad a fotoperiodo (Takahashi, Shomura, Sasaki, & Yano, 2001; T. Yamamoto, Lin Hongxuan, Sasaki, & Yano, 2000). La subunidad Hd6 de CK2 requiere un gen *Hd1* funcional para realizar su función, actuando de manera independiente a los mecanismos del reloj circadiano (Ogiso, Takahashi, Sasaki, Yano, & Izawa, 2010). Finalmente, un gen especialmente interesante por su papel en la historia de la domesticación es *DTH2* (*Days to heading on chromosome 2*). Este gen, aunque tiene un papel menor en la regulación de la floración, promueve la floración en condiciones de fotoperiodo largo. *DTH2* codifica una proteína similar a Hd1 que induce la expresión de *Hd3a* y *RFT1* actuando independientemente de *Hd1* y *Ehd1*. Su expresión

está regulada por el reloj circadiano, teniendo sus picos de expresión en presencia de luz (Wu et al., 2013).

1.4. El arroz como planta modelo

Dada la importancia económica y social del arroz a nivel mundial, la investigación sobre arroz ha sido promovida en todos sus aspectos. Esto ha sido facilitado por el hecho de que la especie cuenta con unas características apropiadas para considerarla en investigación como una especie modelo en monocotiledóneas. Es una especie diploide con un genoma formado por 12 cromosomas (2n =24) y con un tamaño de 389 megabases aproximadamente (Matsumoto et al., 2005), menor que otros cereales con los que presenta sintenia, tiene un ciclo vegetativo relativamente corto y facilidad de manipulación de las plantas en el laboratorio (Shimamoto & Kyozuka, 2002) . Puesto que se trata de un cultivo que constituye el alimento de una gran población a nivel mundial, gran parte de la investigación se ha enfocado al aumento del rendimiento y la seguridad de las cosechas a través de la generación de nuevas variedades más productivas y resistentes a enfermedades y estreses abióticos. La mayor parte de la investigación en arroz está enfocada a la reducción de la pobreza y el hambre gracias a la financiación por varias organizaciones internacionales como Global Rice Science Partnership (GRiSP) del CGIAR (Consultative Group in international Agricultural Research), varias instituciones Chinas, de EE.UU, y otros gobiernos, compañías privadas como Syngenta y fundaciones filantrópicas como Melisa y Bill Gates Foundation o la fundación Rockefeller.

Gracias a la secuenciación del genoma del arroz en el año 2005 (Matsumoto et al., 2005) hoy sabemos que el genoma del arroz alberga un total de 66.338 transcritos (modelos de genes) de los cuales, según se recoge en la base datos del Rice Genome Annotation Project (http://rice.plantbiology.msu.edu), el 46,8% corresponde a genes putativos, el 23,6% se expresa, un 0,2% son genes hipotéticos conservados, un 3,3% hipotéticos y un 26% está relacionado con los elementos transponibles.

La existencia de una gran diversidad de arroz y el avance en las técnicas de secuenciación impulsaron en el año 2014 la realización del proyecto *3000 Genomes Rice Project* en el cual, se resecuenciaron más de 3000 variedades de arroz de todo el mundo (Access, 2014) identificando 18,9 millones de SNPs (Polimorfismo de un solo nucleótido, Single Nucleotide Polymorphism) que pueden caracterizar cualquier variedad existente en el mundo. Los datos obtenidos se recogieron en la base de datos Rice SNP-Seek Database (snp-seek.irri.org) en la que se puede consultar y comparar la información genómica de las 3000 variedades secuenciadas, junto con algunos de sus datos fenotípicos, siendo así una herramienta de un enorme valor que permite explorar la diversidad genética y fenotípica del arroz a nivel mundial. Estos datos son gratuitos y fácilmente accesibles. El grupo genético más representado en este proyecto es el *indica*, con 1174 variedades, puesto que la mayor parte del arroz cultivado pertenece a este grupo, mientras que el grupo *japonica templada*, tiene una representación menor, con datos genómicos de 288 variedades de las que sólo tres variedades son de origen español. Por tanto, sigue siendo necesario un estudio específico y detallado de las variedades cultivadas en las zonas templadas y específicamente en la zona del Mediterráneo que aporte conocimiento de las variedades que cultivamos y apoye la mejora de variedades.

El conocimiento de la diversidad genética y fenotípica da cuenta de la importancia del germoplasma a la hora de explotar las miles de variaciones genéticas aplicables a la mejora asistida por marcadores. A día de hoy existen varios bancos de germoplasma que permiten tener disponibles aproximadamente 780.000 accesiones de arroz. Entre ellos, el International Rice Genebank (IRRI, Filipinas, http://irri.org/our-work/research/genetic-diversity/international-rice-genebank), alberga 128.000 variedades y 4.647 especies silvestres emparentadas, el Rice Genetic Stock Center (USDA, USA-Arkansas, http://www.ars-grin.gov/) almacena 23.090 variedades o el OryzaBase (Japón, http://www.shigen.nig.ac.jp/rice/oryzabase/) que mantiene una

colección de 3.200 variedades de arroz, 1.400 especies emparentadas y 2.000 mutantes.

1.4.1. Herramientas genómicas.

Una de las consecuencias del establecimiento del arroz como una planta modelo es la rápida aplicación de las nuevas tecnologías de análisis genómico en su estudio. Como ejemplo de estas tecnologías está el desarrollo de chips de mejora (Tabla 3), que han permitido la elaboración de mapas genómicos altamente saturados. El uso de tecnologías Next Generation Sequencing ha permitido una aceleración en el desarrollo de técnicas de mapeo de mutaciones surgiendo varias estrategias basadas en el *bulk segregant analysis* como son el *Shoremap* (Schneeberger et al., 2009), QTLseq y el Mutmap (Abe, Kosugi, Yoshida, & Natsume, 2012) y técnicas de análisis transcriptómico como el RNA-seq. El desarrollo de una metodología para la selección genómica, el *Germplams fingerprint* o la predicción de los problemas derivados de las barreras de esterilidad entre variedades *indica* y *japonica*.

Tabla 2: Ejemplos de chips de genotipado enfocados a la mejora de arroz.

Tamaño	**Tecnología**	**Referencia**
5K	Illumina Infinium BeadChip	Ps et al., 2017
6K	Illumina Infinium BeadChip	(H. Yu, Xie, Li, Zhou, & Zhang, 2014)
6K	Illumina Infinium BeadChip	(Thomson et al., 2017)
44K	GeneChip (Affymetrix)	(Tung et al., 2010)
50K	Illumina Infinium BeadChip	(H. Chen et al., 2014)
50K	Axiom	(Singh et al., 2015)

Mutmap

El Mutmap es una técnica derivada del clásico *bulk segregant analysis* pero que se aprovecha de la reducción de costes de las tecnologías de secuenciación de nueva generación. Se desarrolló para la identificación de mutaciones recesivas inducidas en plantas.

En la generación de plantas F2 derivadas de un cruzamiento, se puede observar la segregación de la característica de interés en que difieren los dos parentales originales. Para detectar marcadores asociados con la variación del carácter de interés, la técnica del *Bulk segregant analysis* se basa en la comparación de los marcadores empleados para el mapeo en una muestra de ADN (ácido desoxirribonucleico) que contiene una mezcla, a partes iguales, del ADN de diferentes individuos que presentan el fenotipo alterado con una muestra de una planta con fenotipo silvestre. La comparación detecta polimorfismos genéticos presentes de manera uniforme en todas las secuencias derivadas de la muestra de mezclas de ADN y que difieren de la muestra procedente del silvestre, puesto que las alteraciones no relacionadas con el fenotipo buscado segregan de forma aleatoria en cada grupo y por lo tanto no serán uniformes. El gran salto que supone el uso de la secuenciación de nueva generación (NGS, next generation sequencing) es la secuenciación de todo el genoma en lugar de mapear empleando marcadores distribuidos a lo largo del genoma. Por tanto la detección es inmediata y específica, pudiéndose emplear directamente en una generación F2, reduciendo el tiempo y costes que suponen las técnicas de mapeo clásico que pueden llegar a necesitar hasta una generación F7.

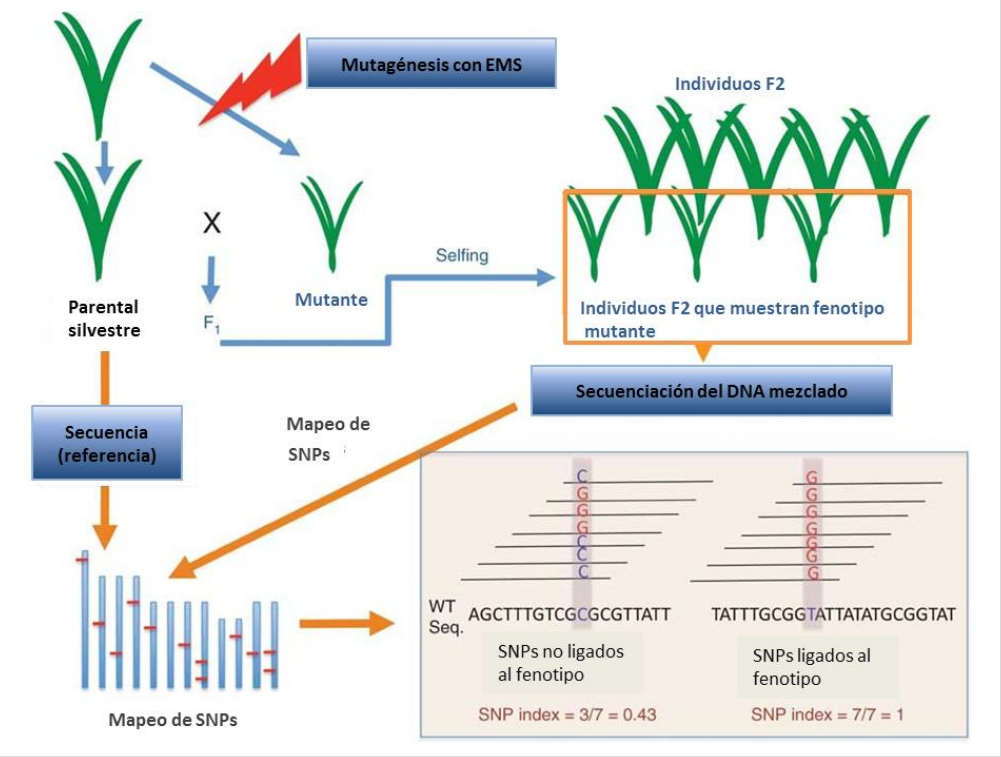


Figura 7: Esquema de la metodología de detección de mutaciones Mutmap.

RNA-seq

Otra técnica de las que han surgido recientemente gracias al desarrollo del NGS es el RNA-seq, que permite el estudio del transcriptoma. El transcriptoma es el conjunto de todos los ARNs (ácido ribonocleico) de una célula o conjunto de células. A diferencia del genoma que prácticamente no varía en un organismo, el transcriptoma varía en función de la célula, tejido, estado de desarrollo o estado fisiológico de la planta. Puesto que en el transcriptoma se incluyen los ARNms (ARN mensajero), su análisis muestra que genes que se están expresando en un momento específico.

Las técnicas para el estudio del transcriptoma han avanzado bastante desde sus inicios en 1990 con la aparición de las *Expressed Sequence Tags* (ESTs) (Adams et al., 1991; Sutcliffe, Milner, Bloom, & Lerner, 1982) que consisten en ADNcs (ADN

complementario) secuenciados derivados de ARNm. A día de hoy existen dos metodologías mayoritarias para su estudio, la hibridación de micromatrices o chips de ADN y la técnica de RNA-seq. Las micromatrices permiten cuantificar un set predeterminado de secuencias mediante la adhesión de sondas específicas a una superficie de vidrio, plástico o silicona. La medición de los niveles de expresión de los genes se basa en la hibridación entre las moléculas diana en la muestra y las sondas, que se suele cuantificar mediante fluorescencia y análisis de imagen. El RNA-seq es una metodología basada en secuenciación masiva que permite la cuantificación de todos los transcritos presentes en la muestra.

La metodología empleada para la realización de RNA-Seq suele basarse en la extracción del ARN (total o una fracción, ya sea eliminando el ribosómico o enriqueciendo en ARNm) con el que se genera una biblioteca de ADNc con adaptadores unidos a uno o ambos extremos. Posteriormente estas bibliotecas son secuenciadas mediante secuenciación masiva de modo que se obtienen lecturas cortas de uno de los extremos (single-end sequencing) o por ambos extremos (pair-end sequencing). Las lecturas suelen tener tamaño de 30 a 400 pares de bases, dependiendo de la tecnología de secuenciación, siendo las más típicas Illumina IG (Morin et al., 2008), Applid Biosystem SOLiD (Cloonan et al., 2008) y Roche 454 (Brad, J., D., Li, & S., 2007; Emrich, Barbazuk, Li, & Schnable, 2007). Una vez obtenidas las lecturas estas se mapean frente al genoma de referencia o al transcrito de referencia o se ensamblan *de novo* si no existe una secuencia referencia con lo que se puede obtener tanto la estructura transcripcional como el nivel de expresión de cada gen.

El RNA-Seq, a diferencia de las micromatrices, puede detectar cualquier tipo de transcrito ya que no hay una selección o set de detección. Esto es especialmente útil para aquellos organismos que no tienen un genoma de referencia con el que comparar. Además las lecturas cortas ofrecen la posibilidad de detectar diferentes formas de *splicing* alternativo. Además, tratándose de secuenciación masiva, permite

la detección de variaciones en las secuencias transcritas como por ejemplo SNPs (Cloonan et al., 2008; Morin et al., 2008). Finalmente el RNA-seq no presenta el problema del ruido de fondo, como en las micromatrices, ya que cada secuencia se mapea sin ambigüedad respecto a las regiones correspondientes del genoma y, además, tampoco presenta el problema de saturación.

La genética de asociación es un método poderoso para la relacionar las variaciones fenotípicas con los polimorfismos genéticos, facilitando la identificación de genes, con sus respectivas variantes, responsables de caracteres determinados así como de las variedades portadoras de los alelos óptimos (Ingvarsson & Street, 2011; J. Yu et al., 2006). Las nuevas técnicas de ultrasecuenciación permiten obtener de una forma rápida y eficiente la secuencia y la comparación de genomas (Lam et al., 2010). Junto con las nuevas plataformas de genotipado rápido y relativamente barato, posibilitan modernizar los métodos de mejora desplazando el tradicional análisis de QTLs (*Quantitative Trait Locus*) hacia la asociación basada en el desequilibrio de ligamiento (LD).

1.4.2. Bases de datos de arroz

El auge de las tecnologías de secuenciación masiva ha dado lugar a la necesidad de generar bases de datos que recojan los resultados obtenidos en los estudios genómicos a fin de recopilarlos y hacerlos fácilmente accesibles. Las dos bases principales que contienen datos del genoma de arroz y de su anotación son el Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/, Ouyang et al., 2007) o The Rice Annotation Project Database (https://rapdb.dna.affrc.go.jp/, Sakai et al., 2013). Entre las bases de datos que recopilan datos de transcriptomas está la Comprehensive Annotation of Rice Multi-Omics (CARMO, http://bioinfo.sibs.ac.cn/carmo/, J. Wang, Qi, Liu, & Zhang, 2015) en la que, además, encontramos información sobre conjuntos de datos transcriptómicos, sitios con

modificaciones epigenéticas y SNPs. La base de datos RiceXPro (http://ricexpro.dna.affrc.go.jp/, Sato et al., 2013) almacenan los perfiles de expresión obtenidos de análisis mediante hibridaciones con micromatrices de ARN de tejidos u órganos a lo largo del desarrollo de la planta en diferentes condiciones de cultivo y bajo distintos tratamientos. La Rice SNP-Seek Database (http://snp-seek.irri.org/, (Alexandrov et al., 2015) derivada del *3.000 Rice Genomes Project* citada anteriormente, almacena los SNPs derivados del proyecto, pero también datos fenotípicos de las variedades y resultados de distintos estudios de asociación de genoma completo (*GWAS*). La base de datos QTARO (http://qtaro.abr.affrc.go.jp/) alberga un repositorio de QTLs para diferentes caracteres y sus posiciones por distintos estudios.

En conjunto, todas estas bases de datos resultan una herramienta de gran utilidad a la hora de obtener información y contrastarla. El avance de las tecnologías y la intersección entre diferentes bases de datos ayudará a la compresión de los mecanismos que llevan desde el ADN que presenta las variedades hasta su fenotipo final, pasando por su transcriptoma, proteoma y metaboloma.

1.5. La mejora del arroz.

El crecimiento y la productividad de las variedades dependen en gran medida de la climatología y las prácticas de cultivo. Las variedades que se cultivan en España poseen unas buenas cualidades, fruto de la adaptación a las condiciones agroclimáticas a través de continuos programas de mejora desde hace muchos años. No obstante, tanto el cultivo como el sector asociado, incluido los agricultores, está sujetos a diversos factores fluctuantes que exigen constantes cambios. La mejora de variedades atiende a sus demandas aportando variedades adaptadas al clima, que son cada vez más productivas y resistentes a estreses abióticos y a patógenos. Por otro lado, las variedades deben adaptarse a las condiciones agroclimáticas cambiantes, que exigen nuevos tipos de plantas y marcan las pautas de hacía dónde dirigir la

mejora. Las previsiones del cambio climático auguran cambios muy desfavorables en las condiciones ambientales del cultivo.

Para entender las dificultades que conlleva un programa de mejora hay que tener en cuenta la propia naturaleza de la planta y su adaptación a la zona de cultivo. Los programas de mejora suelen realizarse de manera local ya que las plantas de arroz son muy sensibles a las condiciones ambientales, como el fotoperiodo, la temperatura o el tipo de suelo. Los dos grupos varietales de arroz mayoritarios, *indica* y *japonica*, tienen gran divergencia fisiológica que dificulta enormemente el flujo genético entre ellos y, de manera añadida, la endogamia producida por años de mejora varietal, con cruzamientos entre variedades emparentadas, dificulta en gran medida la identificación de parentales para los programas de mejora. Por otro lado, debido a la divergencia genética de ambos subgrupos, algunos caracteres agronómicos son propios de cada uno y, por consiguiente, el desarrollo de marcadores moleculares asociados a caracteres agronómicos puede ser válido únicamente dentro de cada subgrupo. Por este motivo, es importante obtener recursos genéticos que faciliten la mejora aportando parentales y material de análisis para la detección de QTLs, generar líneas de mejora más productivas adaptadas a esta zona y, además, obtener información de los genes responsables de caracteres agronómicos de las regiones mediterráneas.

1.5.1. La Mutagénesis en la mejora

La inducción de mutaciones en especies vegetales es una técnica útil para generar variabilidad genética y por eso ha sido muy utilizada en los programas de mejora en todo el mundo durante años. Dada la naturaleza física del agente mutagénico, esta técnica ofrece la ventaja de producir modificaciones genéticas permitiendo la comercialización posterior de las plantas resultantes. Desde los años 60 se han generado un gran número de líneas mutantes que abarcan un amplio espectro de

especies (Shu y Lagoda, 2007). Muchas de estas han sido utilizadas en la obtención de nuevas variedades y otras han servido para el estudio de diferentes aspectos en relación con la biología de las plantas. Las nuevas variedades que se han desarrollado provienen tanto de la explotación directa de las líneas mutantes obtenidas como del uso de estas como progenitores en programas de mejora genética. La base de datos de Variedades Mutantes (Mutant Varieties Database) de la IAEA/FAO (Atomic Energy Agency/Food and Agriculture Organization) contaba en 2007 con más 2.500 variedades oficialmente liberadas, de las cuales unas 1.600 fueron generadas directamente después del tratamiento de mutagénesis y el consecuente proceso de selección. Cabe destacar que 525 de estas nuevas variedades pertenecen a especies de arroz (Wu et al., 2004; Shu et al., 2007; https://mvd.iaea.org/), habiéndose obtenido la mayor parte mediante irradiación con rayos gamma. Se pueden citar varios ejemplos de mutantes de arroz cuya liberación ha tenido un gran impacto económico. El arroz mutante semienano Calrose 76, liberado en California, ha contribuido notablemente a la producción en Estados Unidos, al igual que lo hizo el arroz Basmati 370 en Pakistán. La variedad Zhefu 802, más temprana y productiva que la original, ha sido cultivada extensivamente en China llegando a ocupar 10,5 millones de hectáreas a principios de los noventa (Liu et al., 2004).

1.5.2. Caracteres de interés en la mejora

Las variedades de arroz que cultivamos hoy en día muestran una gran adaptación a nuestro entorno agroclimático y social, pero son susceptibles de mejora en algunos aspectos como el rendimiento, el acortamiento del ciclo vegetativo, la tolerancia a enfermedades y a estreses abióticos.

El aumento del rendimiento siempre es deseable ya que hace más rentable el cultivo y aumenta el beneficio de los agricultores. El rendimiento es un carácter complejo que

depende de factores ambientales, prácticas culturales y también de aspectos fisiológicos y morfológicos de la planta con un componente fuertemente genético. El rendimiento del grano está determinado genéticamente por tres factores: el número de tallos con panículas producidas por la planta, el número de granos de las panículas y el peso del grano.

Los principales factores abióticos que afectan al cultivo del arroz en nuestra zona son la salinidad, la sequía y la exposición a temperaturas extremas. En cuanto a enfermedades, la principal amenaza es la infección por *Magnaporthe oryzae*, un hongo cuyos daños disminuyen el rendimiento y que, en caso de infección grave, puede devastar una cosecha en pocos días.

Una forma eficiente de evitar las pérdidas debidas a los factores climáticos es la reducción del tiempo de cultivo mediante el adelanto de la floración, puesto que se reduce la probabilidad de exposición a fenómenos atmosféricos adversos, a enfermedades y se reduce el consumo de agua. En España las variedades que se cultivan tienen un ciclo de maduración corto o temprano si es de 135 a 145 días, medio de 145-155 y largo en aquellas cuyo ciclo dura 155 días o más.

La transición de la fase vegetativa a la reproductiva, ha sido y es objeto de estudio, dado su interés agronómico. Los estudios de las rutas de regulación de la floración han permitido la obtención de dianas génicas en la mejora para acortar y alargar los tiempos de desarrollo del cultivo mediante técnicas como la selección asistida por marcadores, o la generación de líneas mutantes para alguno de los genes implicados en la regulación.

2. OBJETIVOS

El arroz presenta una gran diversidad natural. Sin embargo, existen barreras genéticas entre distintos grupos poblacionales que dificultan la mejora y que han sido ocasionados durante la expansión del cultivo y su adaptación a otros climas. Por ello, la caracterización de la diversidad genética del arroz cultivada en regiones de clima templado y la identificación de marcadores asociados a los caracteres de interés dentro de esta población, como el rendimiento y la floración, puede acelerar la mejora de las variedades locales facilitando la introducción de caracteres de interés sin arrastrar caracteres indeseables derivados de la adaptación al cultivo a otros climas. En este mismo sentido, puesto que la respuesta al fotoperiodo es la principal barrera en la mejora entre las variedades cultivadas en los climas tropicales y templados, conocer los mecanismos implicados en la regulación de la floración mediante fotoperiodo es indispensable para poder hacer una mejora de manera dirigida.

Por ello el **objetivo** general de esta tesis es identificar marcadores asociados al rendimiento y a la floración en las regiones arroceras de clima templado mediante la caracterización de la diversidad natural existente en dichas regiones y el estudio de un mutante de inducción que presenta la regulación de la floración alterada.

Este objetivo principal se desglosa en varios objetivos parciales:

Objetivo 1: **Establecimiento de las bases genéticas para la asociación de caracteres en clima templado**

Este objetivo se divide en los siguientes apartados:

a. Generar un panel de SNPs de alta densidad que permita genotipar variedades de arroz *japonica*, adaptadas a climas templados y potenciales parentales en programas de mejora.
b. Anotación fenotípica de una colección de 193 variedades de arroz *japonica*, adaptadas a climas templados.

Objetivo 2: **Asociación de polimorfismos genómicos a caracteres agronómicos importantes en el cultivo del arroz: el rendimiento y la floración**

Objetivo 3: Identificación de nuevos componentes reguladores de la floración mediante el estudio de un mutante, generado por irradiación, con ciclo de floración temprano.

Este objetivo se divide en los siguientes apartados.

a) Obtención de líneas mutantes con fenotipo de floración alterado.
b) Caracterización de una línea mutante que presente floración temprana.
c) Identificación de la mutación responsable del adelanto de la floración.

3. CAPí TULOS

3.1. Diversidad genética y estructura poblacional de las variedades de arroz cultivadas en las regiones templadas.

La información presente en este capítulo se encuentra publicada en el artículo:

Reig-Valiente, J. L., Viruel, J., Sales, E., Marqués, L., Terol, J., Gut, M., … Domingo, C. (2016). Genetic Diversity and Population Structure of Rice Varieties Cultivated in Temperate Regions. *Rice*, *9*(1), 58. https://doi.org/10.1186/s12284-016-0130-5

3.1.1. Introducción

El arroz es uno de los principales cultivos, tiene un impacto económico enorme a nivel mundial y es cultivado ampliamente a lo largo de todo el planeta (Lu and Chang 1980). La domesticación del arroz moderno (*Oryza sativa L.*) tuvo lugar en una región situada entre los trópicos en el sureste de China y, simultáneamente a las migraciones humanas y al comercio, se expandió a lo largo de un amplio rango de regiones geográficas con climas diversos (Gross & Zhao, 2014). Como consecuencia, se generó una extensa y vasta gama de diversidad genética que cuenta con dos subgrupos principales cultivados hoy en día (Childs, 2004), los grupos varietales *indica* y *japonica*. Estos grupos genéticos presentan adaptaciones a climas específicos, de acuerdo a las condiciones agroecológicas donde son cultivadas. Los genotipos *indica* se cultivan exclusivamente en las latitudes tropicales, mientras que los *japonica* se encuentran tanto en climas tropicales como templados (Mackill & Lei, 1997).

La productividad del arroz está muy influida, entre otros factores, por las condiciones climáticas. Durante el proceso de expansión del arroz desde su lugar de domesticación, la adaptación a nuevos climas implicó la selección de plantas portadoras de características ventajosas en las condiciones de cultivo adversas a las que iban a enfrentarse y que fueran transferibles a generaciones posteriores. Durante la expansión hacia el norte, una vez cruzado el trópico, el cultivo se encontró con dos impedimentos, uno de ellos fue la diferencia de fotoperiodo y el otro la variación de temperatura. Mientras que el estrés producido por las bajas temperaturas se

mantuvo como una limitación infranqueable en la producción de arroz en las regiones templadas (Andaya & Tai, 2006), las variedades se adaptaron a nuevas condiciones de fotoperiodo largo, puesto que las temperaturas permisivas para el cultivo se daban en el verano del clima templado, con días largos y noches cortas. La aclimatación a las condiciones de día largo en los variedades de arroz de latitudes norte representa uno de los principales cuellos de botella a lo largo de la expansión, y la diferencia más evidente con las variedades que se mantuvieron en las latitudes tropicales (Takeshi Izawa, 2007). El cultivo y la mejora a lo largo de los siglos en diversas condiciones agroecológicas ha dado lugar a una miríada de variedades de arroz que muestran su potencial máximo en las regiones específicas dónde se han desarrollado. Las adaptaciones implicadas en este proceso son modificaciones en la regulación de los procesos metabólicos y fisiológicos que disminuyen el rendimiento y la productividad cuando son cultivadas fuera de sus condiciones de cultivo. En este contexto, los genes implicados en la regulación de la floración deberían mostrar diferencias alélicas a lo largo de los gradientes ambientales. Las frecuencias alélicas deberían reflejar los mecanismos de adaptación de las plantas a diferencias de longitud de día a lo largo de la geografía con un patrón correlacionado (Naranjo, Talón, & Domingo, 2014). En este sentido, una variación importante respecto a la tolerancia al estrés por frio podría haberse observado también conforme la zona de cultivo se aproxima al límite norte (Baruah et al., 2009) y, recientemente, Ma et al. (2015) han descrito el locus COLD1 el cual está implicado en la aclimatación al frio del arroz *japónica*. Como consecuencia de este intenso y largo proceso de mejora, las variedades de arroz se han adaptado a regiones específicas a lo largo del mundo estrechando de este modo el remanente genético, puesto que muchos caracteres fueron olvidados debido a la falta de interés en un momento dado. A día de hoy es laborioso recuperar estos caracteres porque la distancia genética y las diferencias fisiológicas han aumentado entre variedades de regiones tropicales y aquellas cultivadas en regiones templadas. El uso de variedades

no adaptadas en los programas de mejora es un desafío puesto que la incorporación de nuevos caracteres de interés por lo general viene acompañada de otros no deseables que no se adecuan a los requisitos de adaptación climática y a las preferencias de los consumidores.

A pesar de la estrechez del acervo genético provocado por la adaptación específica de las variedades a su entorno agroclimático, la región donde se cultivan variedades *japónica* es suficientemente amplia para albergar diversidad natural relevante capaz de cubrir un espectro amplio de variaciones morfológicas y fisiológicas. La caracterización de esta diversidad, especialmente la relacionada con los caracteres agronómicos de interés, constituye la base de los análisis de asociación genética. La identificación de loci responsables de la variación fenotípica es crucial para los mejoradores, puesto que ofrece oportunidades para incorporar nuevos caracteres de interés en las variedades locales al mismo tiempo que mantiene los caracteres responsables de la adaptación a fotoperiodo.

La caracterización de la diversidad del genoma puede ser realizada eficientemente mediante el uso de NGS, que permite la identificación masiva de SNPs. Estos polimorfismos son marcadores que ya han sido aplicados previamente con éxito a la hora de caracterizar poblaciones en muchos de los cultivos principales (p.e. Myles et al., 2010), incluyendo arroz (p.e. Reig-Valiente et al., 2016; Xu et al., 2012; Huang et al. 2012; Xu et al. 2011). Las bases de datos de SNPs desarrolladas a partir de la secuenciación de diferentes accesiones de arroz cultivado y arroz silvestre son múltiples. Entre las bases de datos de gran escala de SNPs destaca la 3KRPG, generada por el International Rice Informatics Consortium (IRIC), disponible en http://oryzasnp.org/iric-portal/, que ha sido elaborada a partir de los datos obtenidos de la re-secuenciación de 3000 genomas de arroz (The 3, 2014). Esta gran cantidad de información constituye una herramienta invalorable para los mejoradores. Sin embargo, las actividades de mejora conciernen a variedades locales adaptadas a

condiciones agro-climáticas específicas, y así pues, es importante acotar el estudio de la variabilidad genética a variedades que se cultiven en zonas con condiciones agroclimáticas de la zona a fin de identificar los alelos específicos que pueden introducir mejoras mediante combinación de manera fácil y directa. En este apartado, he llevado a cabo la caracterización genotípica de una colección de 217 variedades de arroz mediante la identificación de SNPs, empleando técnicas de NGS que permiten discernir entre ellas. La colección incluye variedades antiguas y modernas, algunas de ellas son variedades élite y, también, locales tradicionales, todas ellas se cultivan habitualmente en las condiciones de fotoperiodo largo que se dan en regiones de clima templado, exceptuando un pequeño grupo de variedades *indica* que se utilizarán como control. Además, algunas variedades, en particular las más antiguas, poseen caracteres que han sido dejados atrás debido a la continua presión de selección y a la adaptación a ambientes agronómicos desafiantes y a las demandas de los consumidores. Esta colección representa la diversidad genética disponible en regiones de clima templado y son de utilidad para los mejoradores de arroz.

3.1.2. Resultados

Selección de 14 variedades representativas de la diversidad genética de la colección

En primer lugar, se generó una colección de 217 variedades con el fin de analizar la estructura de la población de arroz cultivada bajo condiciones de fotoperiodo largo. La colección incluye variedades modernas así como variedades antiguas autóctonas para cubrir una diversidad genética mayor (Fichero adicional 1: Tabla S1, Fichero adicional 2: Figura S1). La colección está compuesta principalmente por variedades de tipo *japónica* con diferentes orígenes geográficos localizados en latitudes norte y con clima templado. La procedencia de las variedades comprende 26 países distintos, siendo un 52,5 % de origen europeo (Fichero adicional 1: Tabla S1). Un grupo de variedades de tipo *indica* fue también incluido en la colección a modo de referencia

para la divergencia genética. Dentro de la colección los tiempos de floración de las variedades abarcan desde los 48 días hasta los 107 (Figura 8) siendo los periodos más largos los correspondientes a variedades *indica*. Cabe señalar que una variedad *indica*, Nona Bokra, no llegó a florecer bajo nuestras condiciones de cultivo.

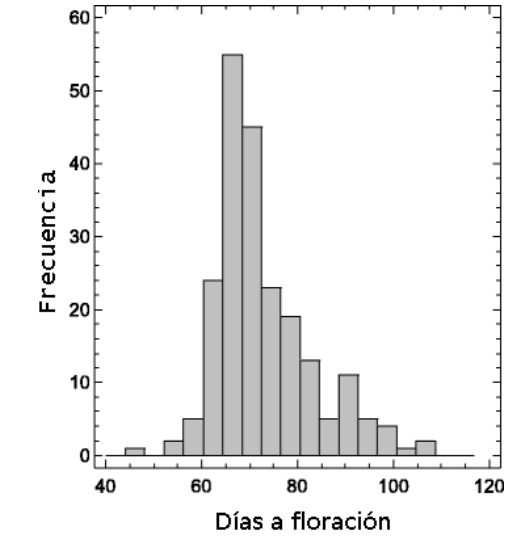


Figura 8: Distribución del tiempo medio de floración en las variedades de arroz *japonica* incluidas en la colección y cultivadas bajo condiciones de luz natural en verano.

De esta colección, se seleccionaron 14 variedades representativas que cubrían una diversidad genotípica neutral útil para la mejora (Tabla 4). Las variedades fueron escogidas de acuerdo a los datos genealógicos conocidos, a la temperatura, al fotoperiodo requerido en zonas de clima templado y al tipo de grano.

Tabla 3: Variedades de arroz, seleccionadas en este estudio, representativas de la colección. Se indica el país de origen, las principales características de interés para los mejoradores y datos de la secuenciación del genoma. Se muestra el número de lecturas, el porcentaje de lecturas únicas y duplicadas y la cobertura media resultante. El número total de SNPs (SNPs vcf) y el total de SNPs tras el filtrado de acuerdo a su significación.

Línea	Origen	Características	Número de lecturas($\times 10^3$)	% lecturas únicas	% Lecturas duplicadas	Cobertura media	SNPs vcf	SNPs Filtrados
Arroz da Terra	Portugal	Tolerancia al frio, floración temprana	80.492	78,9	0,14	39,9	465.830	117.170
Bahia	España	Parental, calidad de grano	78.446	78,2	0,20	38,3	341.451	84.321
Bomba	España	Variedad tradicional, calidad de grano	80.038	77,0	0,12	38,7	771.460	200.619
Gigante Vercelli	Italia	Resistencia a piricularia	69.031	78,1	0,09	33,8	592.763	111.351
Gleva	España	Tipo de grano	78.505	78,6	0,12	38,7	564.829	141.363
Italica Livorno	Italia	Tolerancia al frio, floración temprana	72.919	79,7	0,20	36,3	404.125	91.806
Kalao	Francia	Resistencia a piricularia	75.052	76,9	0,13	36,4	1.025.081	253.553
L202	EE.UU	Parental, tipo de grano	78.854	77,1	0,14	38,3	989.104	269.476
Loto	Italia	Floración temprana	71.612	78,2	0,19	35,1	527.474	106.014
LTH	China	Tolerancia al frio	77.667	77,7	0,10	37,8	573.944	149.950
M202	EE.UU	Parental	69.364	79,1	0,14	34,5	669.310	124.931
Pavlovski	Rusia	Floración temprana, tipo de grano	66.962	80,1	0,16	33,6	396.196	69.894
Puntal	España	Calidad de grano	69.581	77,6	0,17	34,0	928.794	182.272
Senia	España	Tipo de grano	81.376	79,7	0,16	40,6	358.900	100.783

Secuenciación de genoma e identificación de polimorfismos, panel de SNPs para la diversidad del arroz cultivado en regiones templadas

La secuenciación del genoma de un grupo de 14 variedades generó una media de 75 x 10^6 lecturas cortas por variedad que fueron mapeadas respecto al genoma de referencia Nipponbare (IRGSP-1.0). Aproximadamente el 78% de las lecturas correspondió a lecturas únicas (Tabla 4), resultando en una media de cobertura del genoma de x36. La comparación de las secuencias con el genoma de referencia aportó un número relativamente alto de polimorfismos entre los genomas analizados. Los datos fueron filtrados de acuerdo a diferentes criterios como la predicción de la significancia, el tipo y número de alelos, y la ausencia de secuencias repetitivas. Una media de 143.107 SNPs por genoma fueron identificados en los 14 genomas (Tabla 4), los cuales estaban distribuidos uniformemente a lo largo de los 12 cromosomas (Fichero adicional 3: tabla S2). El número de SNPs no redundantes fue 763.021.

Panel de SNPs empleado para analizar la diversidad genética del arroz japónica cultivado en las regiones templadas

Se diseñó un panel de genotipado de SNPs apropiado a la tecnología Infinium (Illumina) incluyendo 2.697 SNPs seleccionados a partir de los polimorfismos identificados en las 14 variedades de arroz secuenciadas. El panel de SNPs fue generado seleccionando polimorfismos bialélicos con una distribución uniforme a lo largo de los 12 cromosomas, pero evitando las regiones de los centrómeros donde la presencia de genes es escasa. Se dio preferencia a los SNPs presentes en más de una variedad. Los SNPs seleccionados fueron filtrados manualmente empleando el programa Integrative Genomics Viewer (IGV, Broad institute). La distancia de intervalo medio entre los SNPs adyacentes fue de 137.525 pb. El número de SNPs por cromosoma abarcó desde 163 en el cromosoma 9 (el más pequeño, con 23,0 Mb) a 321 SNPs en el cromosoma 1 (el de mayor tamaño, 43,2 Mb). La distribución de estos SNPs reflejó su no redundancia. El número de SNPs seleccionados varió entre cultivos,

abarcando desde 761 SNPs en L-202, a 502 en Bahía, con una media de 607 SNPs por genoma (Tabla 5).

Tabla 4: Número de SNPs por cromosoma seleccionados en cada variedad de arroz y número de SNPs no redundantes.

	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10	chr11	chr12	Total
Arroz da Terra	81	71	60	53	42	52	61	54	39	34	35	38	620
Bahia	96	54	49	34	45	37	27	28	16	46	32	38	502
Bomba	61	36	56	63	53	52	54	60	41	58	44	52	630
Gigante Vercelli	87	61	56	56	49	52	39	53	39	42	40	59	633
Gleva	76	53	34	35	54	52	25	59	19	43	27	55	532
Italica Livorno	89	62	54	42	50	69	48	49	38	31	31	49	612
Kalao	101	57	74	65	71	52	64	63	49	41	45	59	741
L-202	105	55	74	71	74	55	57	69	52	41	47	61	761
Loto	65	42	63	56	63	57	59	45	17	48	27	34	576
LTH	56	45	50	64	53	55	53	46	37	36	39	45	579
M-202	95	50	38	44	60	37	50	61	19	45	39	50	588
Pavlovski	58	60	44	45	50	48	46	28	27	35	38	24	503
Puntal	108	61	65	56	66	62	34	62	44	52	45	64	719
Senia	77	55	48	41	59	53	23	30	16	44	24	41	511
Total non-redundant SNPs	321	259	281	248	218	225	199	211	163	179	200	194	2.698

La proporción de SNPs localizados en regiones intergénicas (54,8%) fue mayor que aquellas situadas en regiones génicas (22,1%) (Fig. 9a). Los SNPs situados en regiones génicas fueron clasificados como exónicos (6,4%), intrónicos (11,2%) o UTR (4,5%) mientras que la proporción de SNPs distribuidos dentro de regiones promotoras fue similar a aquellos SNPs situados en regiones génicas (21,9%). Entre los SNPs codificantes, sustituciones no sinónimas (57,3%) fueron más frecuentes que las sustituciones sinónimas (41,8%) (Fig. 9b). Mientras, los SNPs causantes de efectos graves como los que afectan a la integridad de las proteínas codificadas fueron menos

frecuentes, y solo un 0,9% de los SNPs causantes de la ruptura del codón de parada fueron detectados.

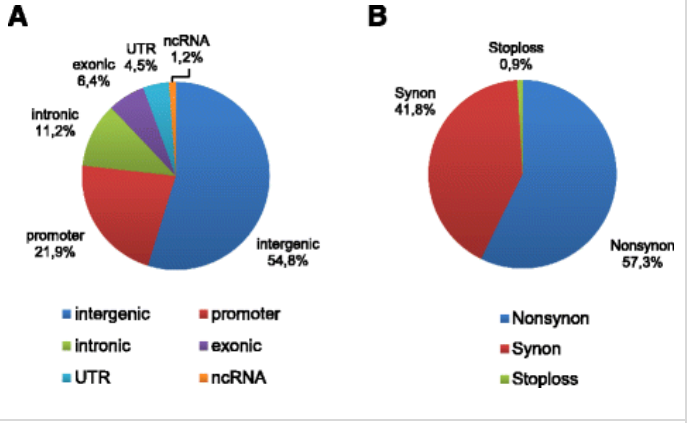


Figura 9: Distribución de SNPs de acuerdo a la anotación de genes según TIGR (The Institute for Genomic Research). Los SNPs génicos fueron clasificados como exónicos, intrónicos y UTR SNPs. Los SNPs localizados hasta 2.0 Kb aguas arriba de la región codificante fueron identificados como SNPs promotores. Los ARNs no codificantes (ncRNA) también se indican. **B** Distribución de los SNPs exónicos causantes de substituciones sinónimas (synon) o no-sinónimas (nonsynon) o responsables de cambios en los codones de parada (Stoploss).

Una vez obtenida la matriz de 2.697 SNPs, se utilizó para genotipar las 217 variedades de arroz de la colección, obteniéndose sus perfiles genéticos. La comparación entre perfiles permitió identificar alelos de baja frecuencia de los cuales se eliminaron aquellos con una frecuencia del alelo menos común (MAF) menor al 5%, originándose un panel de 1.713 SNPs uniformemente distribuidos a lo largo del genoma y con una distancia media de 215.223 pb. Empleando este panel se calculó la extensión del desequilibrio de ligamiento en la subpoblación *japónica* de la colección y se comparó con la calculada para la subpoblación *japónica* del 3KRPG (http://oryzasnp.org/iric-portal/). La diminución del desequilibrio de ligamiento fue calculada como la distancia cromosómica a la cual el coeficiente de correlación (r^2) entre dos SNPs disminuyó a la

mitad del máximo estimado. El LD estimado refleja una fuerte estructura poblacional de la subpoblación *japonica* en nuestra colección, puesto que r^2 cae a 0,23 cuando la distancia cromosómica alcanza las 368 kb, mientras que en la subpoblación del 3KRPG se aprecia un decaimiento mucho más rápido que se extiende hasta 174 kb para el mismo valor de coeficiente de correlación (Fig 10).

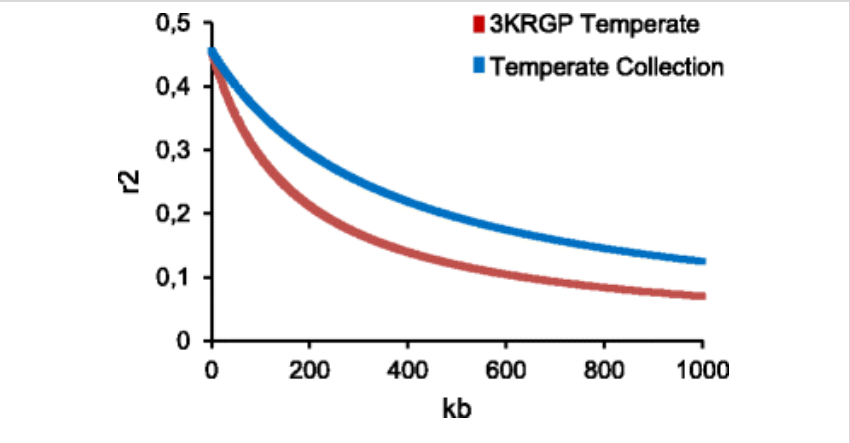


Figura 10: Caída del LD estimado para el subgrupo de variedades *japonica* dentro de la colección de variedades de clima templado en este estudio y la subpoblación *japonica* de la colección del 3KRPG, expresando el desequilibrio de ligamiento como r^2.

Para corroborar la adecuación del panel de SNPs generado para el análisis de la estructura población, examinamos el espectro de frecuencia en 2D entre los grupos varietales *japonica* e *indica* en nuestra colección y en la subpoblación *japónica* del 3KRPG (Alexandrov et al. 2015). Encontramos que la mayoría de SNPs con elevada frecuencia en *japonica* en nuestra colección se encuentran en bajas frecuencia en las variedades *indica* (Figura 4) mientras que la mayoría de SNPs que están en baja frecuencia en *japonica* en nuestra colección también se encuentran en baja frecuencia en la subpoblación de *japonica* del 3KRPG (Figura 11).

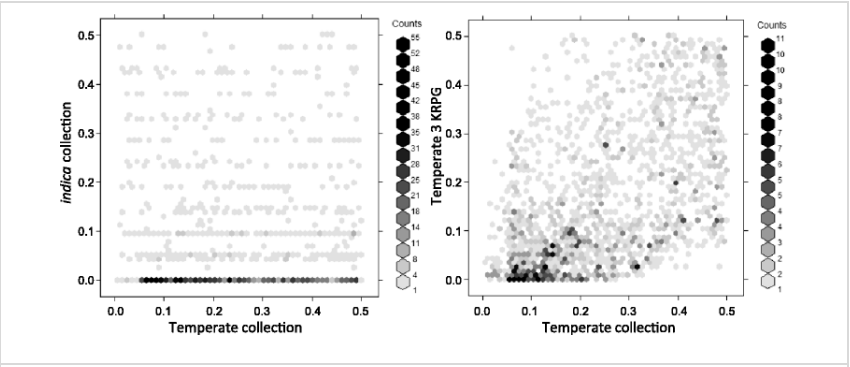


Figura 11: Espectro bidimensional de frecuencias de los loci. En el panel izquierdo se compara la MAF de la subpoblación *indica* frente a la subpoblación *japonica* de la colección utilizada en esta tesis. En el panel derecho se comparan las MAF de la subpoblación *japonica* de la colección del presente estudio frente a la subpoblación *japonica templada* de la colección del 3KRPG.

Estructura genética de la colección

Para determinar la estructura de la colección, se generó un panel de 948 SNPs basado en los resultados de disminución del desequilibrio de ligamiento obtenidos anteriormente y con un intervalo de distancia medio de 390.228 pb. El número más probable de subpoblaciones y de variedades incluidas en cada uno fue calculado mediante el programa STRUCTURE. De acuerdo a estos análisis ΔK mostró su valor máximo para K = 4 (ΔK = 101,2) indicando que el número óptimo de subpoblaciones era 4 (Figura. 12, Fichero adicional 1: Tabla S1). Las variedades de grano largo conformaron la subpoblación 2 y muestran orígenes geográficos distintos como Europa, América o Australia. Las variedades de grano medio se distribuyeron en distintas subpoblaciones de acuerdo a su origen geográfico. Las variedades del grupo 4 provenían mayoritariamente de Italia, mientras que las del grupo 3 eran originarias de América, España y Australia. La mayoría de las variedades españolas incluidas en el grupo 3 han sido generadas recientemente, mientras que algunas accesiones antiguas de España e Italia se dispusieron en el grupo 1 junto a variedades *japonica* asiáticas,

revelando de este modo un origen ancestral asiático de estas variedades europeas antiguas. La diferenciación de estas cuatro subpoblaciones fue corroborada mediante los valores elevados de Fst, obtenidos para cada grupo a pesar de que el cuarto grupo genético, compuesto principalmente por accesiones italianas, mostró los valores más bajos de diferenciación genética (Fichero adicional 4 : Tabla S3, Fichero adicional 5: Tabla S4).

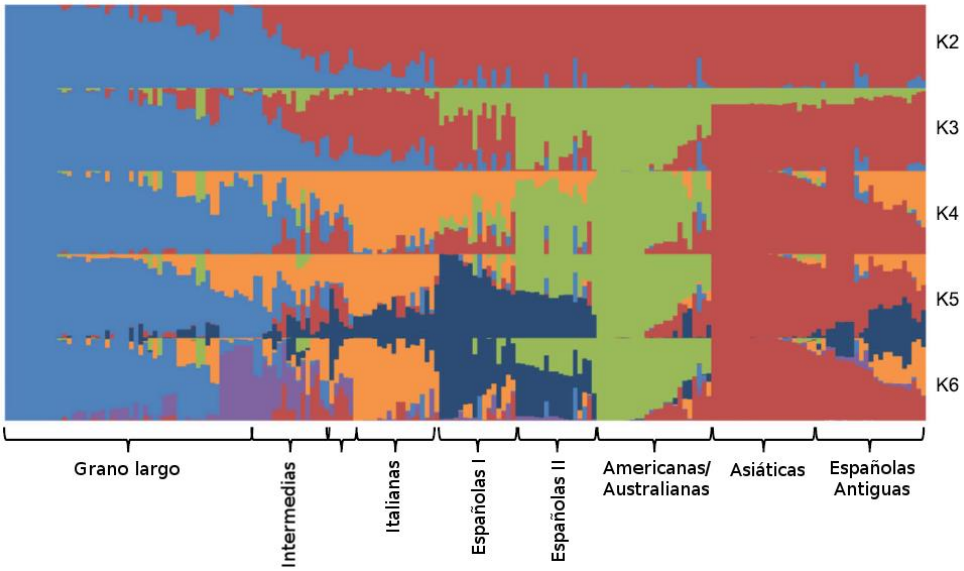


Figura 12: Estructura poblacional de las variedades *japonica* de la colección. Se muestra el valor calculado de probabilidad de pertenencia a los grupos asignados en la colección según una estructura desde 2 (K2) hasta 6 (K6) grupos. La longitud de la barra representa la probabilidad de cada variedad de pertenecer a diferentes subgrupos. Las variedades están ordenadas de acuerdo a su probabilidad de pertenencia a un grupo.

El siguiente valor máximo de ΔK se encontró para K = 5 (ΔK=1,8), el cual separó las variedades españolas en un nuevo subgrupo. Las accesiones que muestran valores más elevados del 80% de pertenencia a este grupo fueron generadas a mediados del siglo XX. Un total de 12 variedades fueron clasificadas como mestizas (*admixed*) y mostraban una pertenencia de un 50% a los grupos americano y español. Este grupo incluye accesiones originadas en España que fueron registradas aproximadamente al mismo tiempo, durante la segunda mitad del siglo XX. La proporción de contenido genético perteneciente a estos dos grupos claramente refleja un periodo de la historia de la mejora en España, puesto que estos cultivos fueron obtenidos cuando germoplasma americano fue introducido en los programas de mejora españoles a mediados del siglo XX.

La siguiente división que se observó al obtener K = 6 (ΔK=3,5) afectó al grupo de grano largo, la cual separaba en un nuevo subgrupo algunas variedades de grano largo, como Moroberekan, Agami y Honduras, cultivadas en regiones tropicales.

Con la finalidad de profundizar en los patrones de la estructura poblacional, se realizó un análisis de componentes principales (PCA) con un grupo de 1.713 SNPs seleccionados a partir de los datos obtenidos del análisis con Infinium. El primer componente principal, el cual abarca un 22,1% de la varianza, separó las variedades de acuerdo al tamaño de grano, lo cual coincide con los cuatro grupos óptimos descritos en los análisis de la estructura, puesto que las variedades de tipo grano largo incluidos en el grupo 2 están separados de los otros grupos (Figura 13).

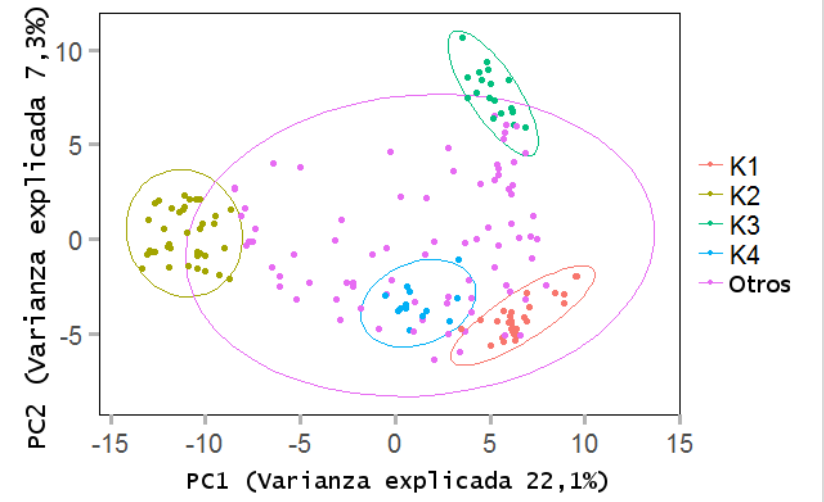


Figura 13: Gráfico del análisis de componentes principales (PCA) de los 1.713 SNPs de las variedades de arroz *japonica* de la colección. Se muestra el primer y segundo componente principal, el porcentaje de varianza explicado por cada uno de ellos se muestra entre paréntesis. Los colores se refieren a los K= 4 grupos genéticos en un porcentaje de pertenencia superior al 80% tal y como se muestra en la figura 12 (ver resultado del análisis de estructura bayesiano).

Relaciones genéticas dentro de la colección de arroz de zonas templadas

El panel de 1.713 SNPs fue empleado para determinar la relación de los cultivos dentro de la colección y para calcular las distancias genéticas entre ellos. Para ello, las 217 variedades fueron agrupadas en un dendograma en ramas de modo similar a la distribución obtenida en el análisis STRUTURE (Figura 14, Fichero adicional 6: Figura S2). La distribución de las variedades obtenida coincidió básicamente con el origen o el tipo de grano. En un extremo del árbol, las variedades *indica* aparecen en un grupo claramente separado. Próximo a este grupo de *indica,* las variedades de grano largo divergen en un grupo que incluye la mayoría de las variedades del grupo 2 del análisis de la estructura poblacional. Estas son variedades *japonica* de regiones tropicales (como Azucena o Moroberekan) y regiones templadas (Cormorán o Apolo).

Variedades de arroz aromático, por ejemplo, Fragrance o Giglio, pueden distinguirse también entre estas.

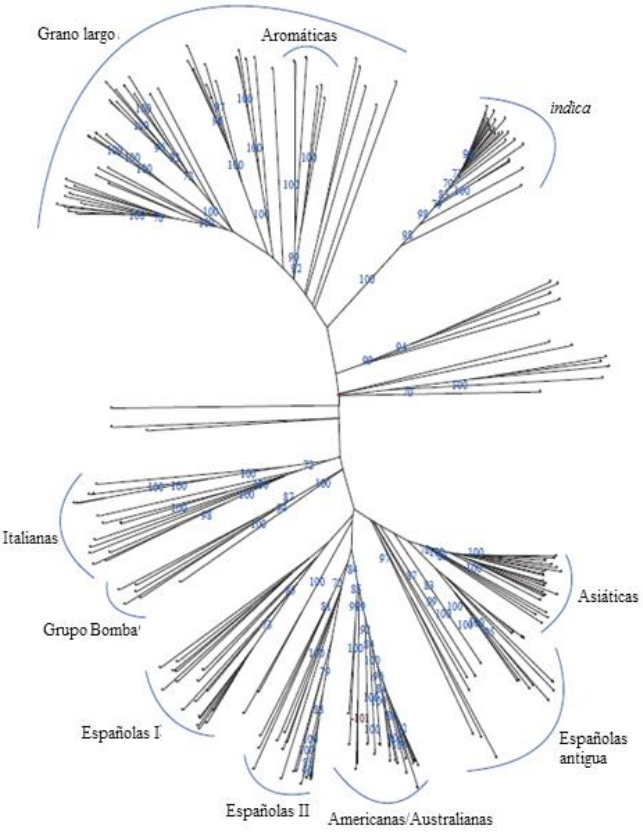


Figura 14: Dendograma de las 217 variedades de la colección generado mediante Neighbour-Joining basado en la distancia de Sokal y Michener implementado en el Darwin 6.0. Los valores de las ramas indican las convergencias del bootstrap (10.000 replicas).

El resto de variedades, mayoritariamente de grano medio, aparecen en el otro extremo del árbol de acuerdo a su distancia genética con respecto a las variedades de grano largo y los grupos *indica*. Las variedades de grano medio, a su vez, se distribuyeron en 7 grupos distintos en los que se aprecia una clara influencia de los

orígenes geográficos (figura 14). Varias variedades autóctonas europeas antiguas fueron recogidas en un grupo más amplio que incluye representantes asiáticos. El uso de un número más elevado de SNPs para construir el dendograma permitió la fragmentación de los cultivos europeos en varios subgrupos pequeños, con una distribución que claramente concuerda con su origen e incluso en algunas ocasiones con el momento de generación de las líneas. Las variedades asiáticas están claramente relacionadas con el grupo formado por las variedades antiguas españolas, mostrando la influencia del germoplasma asiático en los inicios de la historia del cultivo de arroz en España. Otros dos grupos están formados por variedades españolas, las cuales están muy relacionadas con el grupo americano/australiano, indicando la relevancia que tuvo la introducción del germoplasma americano en los programas de mejora españoles. Esta relación genética también ha sido observada en el análisis genético de la población. Finalmente, un cuarto grupo claramente diferenciado incluye cinco variedades autóctonas españolas con características culinarias específicas y similares entre ellas, que dominamos tipo Bomba y que pueden tener un origen común.

3.1.3. Discusión

La expansión del cultivo del arroz, tras su domesticación, a una amplia área geográfica trajo consigo una gran diversidad debido a la progresiva aclimatación de las plantas a las diferentes condiciones climáticas que iba encontrando en su avance. La selección de plantas con buen comportamiento agronómico y rendimiento superior condujo a la aparición de un amplio rango de variedades con fenotipos distintos. La diversidad genética es un recurso natural muy apreciado en la mejora de arroz ya que provee de parentales y, además, permite la asociación de rasgos complejos con los genes responsables de estos. Para realizar estudios de asociación y la identificación de genes funcionales relacionados con caracteres agronómicos es imprescindible entender la estructura poblacional y las variaciones en el genoma que son fuente de la diversidad

genética. El presente estudio aporta un panel de marcadores de tipo SNP con una buena cobertura del genoma de arroz. Este panel de 1.713 SNPs ha permitido caracterizar 217 variedades de arroz de una colección generada como muestra representativa de todos los grupos genéticos encontrados a lo largo de las regiones geográficas con clima templado, tal y como ha mostrado el análisis de la estructura de la población. La colección incluye variedades antiguas y modernas de tipo *japonica* en un intento por cubrir el máximo espectro de variabilidad. Esto nos ha permitido reconstruir la relación genética entre diversas variedades de arroz de clima templado procedentes de orígenes geográficos lejanos (Fichero adicional 1: Tabla S1) las cuales poseen una enorme variabilidad en caracteres agromorfológicos y fisiológicos.

El uso de marcadores tipo SNP, aportado gracias a la secuenciación del genoma de un grupo de variedades seleccionadas, ha permitido analizar en detalle el grupo de variedades *japonica* de clima templado, y en particular las relaciones genéticas entre algunas variedades españolas y otros grupos procedentes de distintos países. Esto nos ha permitido reconstruir la historia de la mejora del arroz en España. De acuerdo a estudios genéticos previos en arroz (Huang et al., 2010; Xu et al., 2011; K. Zhao et al., 2010) nuestro análisis mostró la existencia de una fuerte subestructura genética poblacional dentro de las variedades *japonica*. Este hecho es coherente con la lenta reducción del desequilibrio de ligamiento observada en nuestra colección en comparación a otras pequeñas subpoblaciones (Xu et al., 2011), que puede ser por el efecto de la metodología de muestreo y el tamaño muestral. Una de las observaciones relevantes de este análisis es la influencia del tipo de grano respecto a la clasificación de las variedades, sugiriendo que este carácter es un factor de importancia crítica en la estructura de la morfología de las variedades actuales y antiguas. Del mismo modo, las distancias mostradas en el dendograma establecen una clara diferencia entre las variedades de grano largo y medio con independencia de su origen (Figura 7). Este patrón también ha sido observado en los análisis de PCA,

los cuales separan ambos grupos en el primer componente principal (Figura 6). El origen de las variedades de grano largo no parece ser un factor discriminante dentro de ese grupo que ciertamente contiene muchas variedades de distintos continentes. Así pues, este grupo incluye variedades cultivadas en regiones tropicales (por ejemplo, Filipinas u Honduras), así como en países de climas templados como lo son los países europeos. Además, las variedades de la colección consideradas como *japonica* tropical, como Azucena, Hondura, Katy y Lemont (K. Zhao et al., 2010), se localizan dentro de los cultivos de grano largo.

Por otro lado, varias subpoblaciones de variedades *japonica* de grano largo pueden distinguirse en ramas secundarias. En este caso, las variedades geográficamente cercanas estaban más próximas a nivel de parentesco que aquellas provenientes de localizaciones distantes, mostrando unos patrones geográficos de la variación genética probablemente causados por la mejora y la selección asociada con las preferencias alimenticias locales.

Las variedades españolas, el grupo más numeroso en la colección, estaban organizadas en cuatro grupos dependiendo principalmente del momento en el que fueron liberadas, reflejando la historia de la mejora del arroz en España. La familia Bomba, conformada por las variedades autóctonas más antiguas de España y cuyos orígenes preceden a los registros disponibles, aparecen como un grupo distinto aislado en los dendogramas. Algunas de estas variedades, son muy apreciadas por los consumidores, siendo cultivadas a día de hoy debido a las características peculiares de su grano y a unas características agronómicas que las hacen adecuadas para un cultivo orgánico. Un segundo grupo de variedades españolas derivado de las actividades más tempranas de mejora en España permanece en un grupo distinto pero próximo al grupo de variedades asiáticas, indicando el origen de los primeros cultivos empleados como donantes en los primeros programas de mejora. En un tercer grupo se encuentran las variedades cultivadas en España durante la primera mitad del siglo XX.

Este grupo se sitúa muy próximo al nodo que separa los grupos americano-asiático. Las variedades españolas modernas están situadas en un grupo secundario anidado al grupo americano/australiano. Estos resultados, en conjunto con la mezcla genética sugerida por los análisis de estructura bayesianos respecto a los subgrupos americano y español cuando K = 5, sugieren la introducción de variedades de germoplasma americano en el programa de mejora español.

Los estudios genéticos previos apuntaban a una diversidad genética más reducida en las variedades *japonica* que en las *indica,* probablemente a causa de haber sufrido un cuello de botella más severo durante la domesticación (Garris et al., 2005). Además, las actividades de mejora durante el último siglo también han constreñido el remanente génico de las variedades seleccionadas para su producción en diferentes ambientes agroecológicos. En este sentido, la búsqueda de nuevos donantes potenciales para las actividades de mejora se ha convertido en una tarea complicada puesto que el uso de donantes no adaptados de otras latitudes para la introducción de caracteres particulares suele venir acompañada de caracteres agronómicos no apropiados para la región de cultivo. Este es el caso de la hibridación entre variedades *japonica* templadas y tropicales. Además, se ha observado una elevada esterilidad en los híbridos *indica-japonica* debido a las barreras reproductivas que han entorpecido el flujo génico entre estas (Jeung, Hwang, Moon, & Jena, 2006). Por el contrario, el flujo génico ocurrido entre las variedades producidas en las regiones templadas ha puesto de manifiesto la historia de la mejora en España a partir de nuestros análisis genéticos, mostrando la influencia de varios grupos genéticos en las variedades españoles. Al comienzo de la actividad de mejora en España, el arroz asiático tuvo una mayor influencia en los arroces de grano medio existiendo también conexiones con las variedades italianas. Esta influencia se redujo conforme se generaron las nuevas variedades mejoradas, habiendo un aumento en el porcentaje de germoplasma procedente de cultivos de origen americano, en detrimento del procedente de Italia.

La diversidad natural distribuida a lo largo de diferentes regiones geográficas durante la expansión del arroz constituye una fuente de variabilidad genética de gran valor para los programas de mejora a la hora de generar nuevas variedades que estén adaptadas a las condiciones climáticas locales. Los mejoradores necesitan los recursos genéticos naturales disponibles para recuperar caracteres agronómicos de interés que fueron abandonados durante el proceso de selección, sin embargo se desconocen las regiones cromosómicas responsables de la variación fenotípica. El mapeo de asociación y la selección genética son metodologías recientemente generadas que han demostrado ser efectivas para conectar las variaciones genéticas con las variaciones en el fenotipo, permitiendo así explotar aquellos alelos élite presentes en el germoplasma (p.e. Begum et al., 2015; P. Zhang, Liu, Tong, Lu, & Li, 2014). El conjunto de SNPs de alta variabilidad en las variedades de arroz templado que hemos diseñado constituye una herramienta de fácil acceso y rápida con la que explorar genes candidatos implicados en varios caracteres relacionados con la adaptación al clima templado como pueden ser el tiempo de floración o el rendimiento.

3.1.4. Conclusiones

El análisis de los perfiles genéticos de una colección de 217 variedades de arroz reveló la diversidad y la relación entre las variedades de arroz cultivadas en las regiones de clima templado. Las observaciones realizadas en este estudio indican que la contribución del tipo de grano a la estructura genética de la población de las variedades *japónica* templadas es más fuerte que la debida a los orígenes geográficos. El uso de marcadores de tipo SNP en este estudio ha revelado el flujo génico elevado así como mestizaje entre variedades cultivadas en regiones muy distantes, probablemente como consecuencia de las actividades de mejora locales. También se observó la influencia de germoplasma asiático al principio y americano más tarde en los programas de mejora españoles

3.1.5. Materiales y métodos

Material vegetal y condiciones de cultivo

La colección fue generada a partir de diferentes bancos de germoplasma: del IRRI, U.S. Natitonal Plant Germplasm System (NPGS, USA), Rice Genome Resource Center (RGRC, Japón) y el IVIA. Las semillas de diferentes accesiones fueron germinadas y cultivadas en macetas en invernadero (39° 28’ N) bajo condiciones controladas de temperatura (25 °C) y humedad relativa (50 % RH), y en condiciones de luz natural durante los veranos de 2013 y 2014. Las plántulas fueron trasplantadas manualmente en filas de 20 plantas en mayo y cosechadas en septiembre. Los campos fueron regados mediante inundación. El tiempo de floración fue anotado como el momento en el que el 50% de las panículas de cada fila habían emergido.

Secuenciación de genoma completo

Catorce variedades representativas de la subespecie *japonica* fueron seleccionadas previamente para llevar a cabo la caracterización de SNPs (Tabla 3). El ADN nuclear de hojas de plántulas cultivadas 7 días en presencia de luz seguido de dos días en oscuridad fue extraído mediante el empleo de un protocolo basado en el uso de CTAB (Schneeberger et al., 2009). La secuenciación del genoma fue realizada en el Centro Nacional de Análisis Genómico (Barcelona, España) como se indica a continuación: se generaron bibliotecas de inserto corto y extremos emparejados (paired-end) con un protocolo NO-PCR. Se empleó el kit TruSeq^{TM}DNA Sample Preparation Kit v2 (Illumina inc) y el kit KAPA Library Preparation (Kapa Biosystems). En resumen, 2,0 microgramos de ADN genómico fragmentado fue reparado por los extremos (end-repaired), adenilado y ligado a adaptadores emparejados por los extremos e indexados específicos de illumina. El ADN fue seleccionado de acuerdo a su tamaño empleando bolas AMPure XP (Agencourt, Beckman Coulter) con el objetivo de seleccionar tamaños de 220-550 pb. Las bibliotecas finales fueron cuantificadas mediante el Library Quantification Kit (Kapa Biosystems).

Las bibliotecas fueron secuenciadas empleando TrueSeq SBS Kit v3-HS (Illumina Inc.), en modo emparejado por los extremos (parired end), 2x101 pb, en ½ columna del secuenciador Hiseq200 flowcell v3 (Ilumina Inc.) de acuerdo al procedimiento de operación estándar de Illumina con una producción de > 14 Gb y una cobertura media de 33-41x. El análisis de datos preliminar, el análisis de imagen, la anotación de bases y la calificación de calidad del proceso fueron realizados empleando el software del fabricante Real Time Analysis (RTA 1.13.48) a lo que siguió la generación de ficheros FASTQ mediante CASAV. Los datos han sido depositados en el European Nucleotide Archive (ENA) en el European Bioinformatics Institute (EBI) con el número de acceso PRJEB13328.

Análisis de los datos secuenciados.

Las lecturas fueron recortadas de principio a fin para un valor de calidad Phred ≥10. Las lecturas con una longitud ≥40 nt fueron mapeadas respecto al genoma de referencia de la variedad de arroz Nipponbare (Os-Nipponbare-Reference-ITGSP-1.0 (IRGSP-1.0)) empleando el kit de herramientas GEM (versión 2) (Marco-Sola, Sammeth, Guigó, & Ribeca, 2012) permitiendo hasta 8 discordancias por lectura. Sólo fueron empleadas para los siguientes análisis los pares de lecturas mapeados no duplicados. Se empleó la suite SAMtools (versión 0.1.18) (H. Li et al., 2009) con valores por defecto para detectar SNVs y pequeños INDELS por variedad. Las anotaciones de genoma de http://rapdb.dna.affrc.go.jp/download/irgsp1.html, tales como los genes y los CDS, fueron añadidos a los ficheros VCF generados empleando vcftools (Danecek et al., 2011). Las variaciones detectadas en las regiones con baja cobertura, con un sesgo claro de la cadena con un p-valor <0,001 o un sesgo pronunciado en la cola con un p-value <0,05 fueron excluidas

Panel de SNPs y genotipado

Se generó un panel de 2.697 SNPs representativos de las variedades de arroz *japonica* cultivadas bajo condiciones de día largo mediante la selección de los polimorfismos

identificados en las 14 variedades cuyo genoma fue secuenciado. Se seleccionaron aquellos polimorfismos bialélicos de modo que se mantuviese una distribución uniforme a lo largo de los 12 cromosomas, evitando las regiones centroméricas y teloméricas y manteniendo una distancia media de 137,6 Kb entre ellos. Se dio prioridad a aquellos SNPs que estuviesen presentes en más de una accesión. Los SNPs fueron visualmente inspeccionados empleando el programa *Integrative Genomics Viewer* (IGV, Broad Institute). La adaptabilidad al sistema de detección Illumina iSelect de los SNPs y su secuencias colindantes fue valorada y se procedió a la selección de aquellos SNPs con una puntuación superior a 0,4 para el ensayo de Multiplexing Infinium (Illumina Inc.).

Las 217 variedades fueron genotipadas con el panel de 2.697 SNPs. La comparación de los genotipos detectados se llevó a cabo empleando el programa GenomeStudio (Illumina). Se consideraron alelos de baja frecuencia aquellos presentes en menos del 5% de las variedades y se eliminaron. Se consideraron marcadores fallidos aquellos que no daban señal en menos del 50% de las variedades.

Los análisis de variación genómica y clasificación de SNP de acuerdo a los modelos de genes de arroz de TIGR se llevó a cabo en la plataforma CARMO (Comprehensive Annotation of Rice multi-Omics data) (Wang, Qi, Liu, & Zhang, 2015, http:/bioinfo.sibs.ac.cn/carmo/).

Cálculo del desequilibrio de ligamiento

Los datos brutos fueron filtrados para seleccionar 2.066 válidos en al menos el 75% de las variedades y para eliminar aquellos marcadores monomórficos con un MAF inferior al 5%. El desequilibrio de ligamiento fue calculado empleando el programa Plink 1.07 (Purcell et al., 2007) de acuerdo a Zhao et al. (K. Zhao et al., 2011). Un total de 1.713 marcadores superaron el filtro y fueron empleados para calcular el LD utilizando el comando "-r2 –ld-window 99999 –ld-window-r2 0". La disminución del

equilibrio de ligamiento respecto a la distancia fue ajustada empleando el modelo esperado de Hill and Weir (Hill & Weir, 1988) para r^2 entre sitios adyacentes. Los cálculos fueron realizados de acuerdo a la ecuación de Remington (2001). Se empleó un ajuste de mínimos cuadrados no linear, implementado en el pack "nls" de R, para ajustar los datos a la ecuación tal como sugiere Marroni et al., (2011).

Análisis de componentes principales.

Se realizó un análisis de componentes principales empleando el comando *prcomp* en R (versión 3.2.3.). Se empleó el genotipo de los 1.713 SNPs del panel para el análisis del LD. Los datos fueron analizados empleando *ggbiplot* (versión 0.55)

Estimación de la estructura genética

La estructura genética de las variedades de la colección fue calculada empleando el método de agrupación bayesiano implementado en el programa STRUCTURE2.3.4 (Pritchard, Stephens, & Donnelly, 2000). Esta aproximación estimó el número óptimo de grupos (K) y la proporción de pertenencia de los cultivos a estos. Los análisis se basaron en el modelo ancestral de mestizaje para un rango de valores de K entre 1 y 15. Se realizaron 20 repeticiones para cada K y se eliminaron aquellas repeticiones cuyos valores de L(K) aparecían fuera de rango de acuerdo a Evanno *et al*., (2005). Cada repetición se realizó con un periodo de *burn-in* de 100.000 repeticiones seguido 1.000.000 repeticiones Monte Carlo vía Cadenas de Markov. El número óptimo de grupos K fue estimado con el parámetro ad hoc (ΔK) de Evano et al. (2005) usando el programa Structure Harvester (http://taylor0.biology.ucla.edu/structureHarvester/, Earl & vonHoldt, 2012). Calculamos el alineamiento óptimo para las 20 réplicas con el programa CLUMPP (Jakobsson & Rosenberg, 2007), empleando el algoritmo *greedy* con 10.000 repeticiones. Las accesiones fueron subdivididas en diferentes subgrupos de acuerdo a su máximo de probabilidad de pertenencia entre los subgrupos y el umbral de probabilidad de pertenencia de 0,8.

Las distancias genéticas entre variedades de arroz se calcularon tal y como implementa el programa Darwin 6. (Perrier et al. 2003), mediante el uso del índice de Sokal & Michener con 10.000 repeticiones *bootstrap*. Posteriormente se obtuvo un árbol *neighbour joining* (NJ), cuya robustez se evaluó de nuevo mediante un análisis de 10.000 réplicas *bootstrap*.

También se empleó el programa STRUCTURE para calcular los valores de Fst para cada grupo genético establecido, se calculó la media de las 20 réplicas. Finalmente los Fst entre las subpoblaciones definidas por el análisis de STRUCTURE fueron calculadas con el programa Arlequin (Excoffier & Lischer, 2010), con 10.000 permutaciones incluyendo únicamente aquellas variedades con valores de pertenencia superiores al 80%.

3.2. Estudio de asociación de polimorfismos genéticos a las variaciones en caracteres fenotípicos de interés agronómico

La información presente en este capítulo se encuentra en revisión para ser publicada bajo el título de:

"Genome-wide association study of agronomic traits in rice cultivated in temperate regions"

3.2.1. Introducción.

Uno de los principales desafíos para los mejoradores de arroz es desarrollar variedades de alto rendimiento. Puesto que el arroz es un alimento fundamental en la mayoría de países, una productividad elevada es necesaria para aportar alimento a una gran población creciente. También es deseable aumentar las ganancias de los agricultores y la rentabilidad de la cosecha.

Desde su domesticación, las variedades de arroz se han expandido hacia ambientes agroecológicos diferentes mediante la generación de nuevas variedades a partir de la selección de plantas adaptadas a las nuevas condiciones a través de una actividad de mejora intensa y continua (Huang, Kurata, et al., 2012). Una de las barreras que el arroz hubo de superar a la hora de alcanzar las regiones templadas fue la diferencia en la longitud del día que se convirtió en uno de los principales determinantes de la adaptación de la planta a nuevas regiones (Itoh et al., 2018; T. Izawa, 2007). Como consecuencia, la distancia genética entre las variedades de las regiones tropicales y aquellas cultivadas en regiones templadas aumentó lo suficiente como para que se generarán barreras reproductivas que estrechasen todavía más el flujo genético entre las variedades de clima tropical y templado (Jung et al. 2005) contribuyendo a la reducción de la reserva genética del cultivo y la emergencia de nuevas subpoblaciones. Los dos principales grupos varietales cultivados, *indica* y *japónica*, se caracterizan por adaptaciones a climas específicos de acuerdo a las condiciones agroecológicas donde son cultivados. Las variedades de tipo *indica* son cultivadas en

las latitudes tropicales, mientras que las variedades *japonica* pueden encontrarse tanto en zonas tropicales como templadas (Mackill & Lei, 1997). Debido a la adaptación de las variedades a climas específicos en particular al fotoperiodo local, el uso de variedades de diferentes subpoblaciones como donantes en los programas de mejora es un reto puesto que implica la introducción de caracteres no deseables y que pueden ser inadecuados para los requisitos climáticos específicos locales.

El área donde el tipo *japónica* se cultiva es suficientemente extensa como para retener una diversidad natural relevante para cubrir un amplio espectro de variaciones morfológicas y fisiológicas (Reig-Valiente et al., 2016). La caracterización de las bases genéticas de esta diversidad permitirá la identificación de loci responsables de esta variación fenotípica, especialmente aquellos referentes a los caracteres agronómicos con aplicación directa en la mejora. Este logro ofrecerá nuevas oportunidades para la selección de parentales apropiados para la incorporación de nuevos caracteres de interés en las variedades locales al mismo tiempo que se conservarán aquellos caracteres responsables de la adaptación a clima templado.

La productividad es un carácter complejo que implica múltiples caracteres y depende de varios factores genéticos y ambientales (Xing & Zhang, 2010). La adaptación de variedades a condiciones agroecológicas específicas es un requisito esencial para una productividad elevada, y en este sentido, es necesario un tiempo de floración óptimo. El número de granos que una planta puede producir también es determinante y puede variar de acuerdo al número de granos por panícula y el número de panículas, que son caracteres cuantitativos (Xing & Zhang, 2010). El número de panículas es un carácter difícil de estudiar ya que depende no sólo de factores genéticos sino también de las condiciones de crecimiento y el ambiente así como de la densidad de plantas en el campo.

Otro factor morfológico es la altura y la longitud de panícula, que también afectan a la productividad final.

Los estudios de asociación de genoma completos (GWAS) se han convertido en un análisis popular a la hora de identificar QTLs en poblaciones de plantas, al aprovechar la diversidad del arroz basándose en eventos de recombinación históricos y en el desequilibrio de ligamiento. Varios GWAS han sido realizados en busca de genes implicados en caracteres que afectan a la productividad. Estudios sobre la arquitectura de la panícula, el número a espiguillas, el tamaño de grano o el tiempo de floración (Begum et al., 2015; Rebolledo et al., 2016) . La mayoría de estos han sido realizados en poblaciones *indica* comparando variedades *indica* y *japónica* con otras especies silvestres (Huang, Zhao, et al., 2012; Rebolledo et al., 2016; Yonemaru et al., 2014). También se han realizado unos cuanto estudios para detectar asociaciones dentro de poblaciones *japonica* (Volante et al., 2017; M. Yano et al., 2000). Pese a ello, las bases genéticas de las variaciones fenotípicas entre las variedades templadas necesitan ser analizadas en mayor profundidad.

En un trabajo previo, generamos una colección de variedades de arroz representativas de la diversidad genética presente en las regiones templadas y que poseen una enorme variedad de caracteres agromorfológicos y fisiológicos. La colección está compuesta por 193 variedades de diferentes periodos incluyendo variedades antiguas y modernas, así como locales tradicionales y variedades élite. El estudio de la estructura poblacional y las relaciones génicas entre estas variedades evidenció la fuerte subestructura en la colección de variedades de la zona templada, predominantemente basadas en el tipo de grano y el origen geográfico de las variedades (Reig-Valiente et al., 2016). Observamos la existencia de un flujo génico relativamente elevado y de elevados ratios de mestizaje entre variedades cultivadas en regiones remotas, probablemente favorecidas por las actividades de mejora locales. Los resultados permiten los estudios de asociación de genoma completo para

caracteres complejos e investigaciones de genes funcionales entre las variedades aclimatadas a las regiones templadas evitando la generación de asociaciones espurias debidas a los efectos de la estructura poblacional y relaciones de parentesco desconocidas entre variedades.

En este apartado se han estudiado las bases genéticas responsables de caracteres agronómicos que contribuyen a un rendimiento elevado en el arroz *japonica* templado con el objetivo de facilitar la mejora del arroz en esas zonas. Hemos llevado a cabo un estudio de asociación de genoma completo y hemos identificado y localizado varios QTLs de los caracteres investigados

3.2.2. Resultados

Evaluación del fenotipo

Las plantas de una colección de 193 variedades *japónica* adaptadas a las regiones templadas fueron cultivadas y evaluadas durante dos años consecutivos en campo en condiciones de día largo durante la estación de verano. Todas las plantas fueron capaces de florecer bajo estas condiciones de fotoperiodo. El tiempo de floración (DH), la altura de la planta (H), el número de panículas (PN) y la longitud de panícula (PL) fueron anotados (Fichero adicional 7. Tabla S5). Los caracteres presentaron una gran variación fenotípica (Tabla 5) con unos coeficientes de variación que abarcaban desde 0,12, en el caso del tiempo de floración, a 0,41, para el número de granos. Las distribuciones de las frecuencias de los fenotipos mostraron una aproximación a una distribución normal (Figura 15). Únicamente las frecuencias para longitud de panícula presentaban una distribución normal. Los datos referentes al número de panículas en el campo de Malta durante 2016 fueron eliminados del análisis debido a diferencias en los sistemas de anotación entre las distintas localizaciones y años, mostrando valores más bajos que en otros ensayos (Fichero adicional 8: Figura S3). Puesto que algunas variedades presentaban dehiscencia, el número de granos por panícula (GN)

fue anotado en dos baterías de plantas cultivadas en invernadero bajo condiciones de día largo antes de llegar a la madurez completa. Todas las variables medidas mostraron unos valores de heredabilidad en sentido amplio elevada con valores superiores a 0,77 excepto PN que presentó un valor de 0,54 y altura que presento el valor más elevado, 0,92.

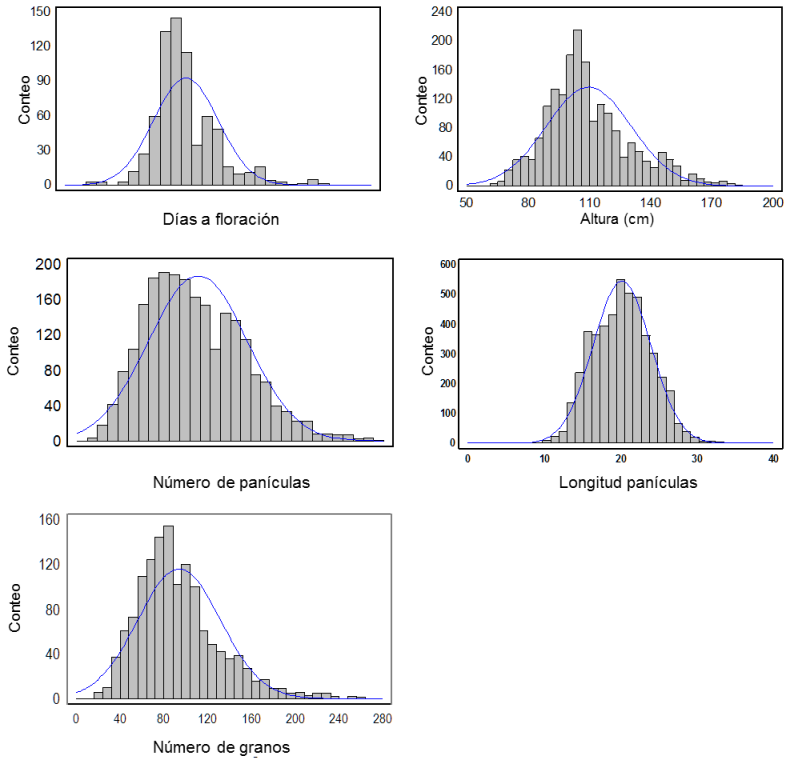


Figura 15: Distribución fenotípica para A) Tiempo de floración B) altura de la planta C) número de panícula D) longitud de panícula E) número de granos por panícula. En los histogramas se incluyen los valores tomados en todas las plantas de los ensayos. Las curvas mostradas se producen tras ajustar los datos a una distribución normal.

Tabla 5: Variación fenotípica de los caracteres agronómicos estudiados.						
Carácter	**Media**	**Mínimo**	**Máximo**	**SD**	**CV**	**H2**
Altura (cm)	109,7	62,0	182,0	20,8	0,19	0,92
Tiempo de floración	71,6	47,0	108,0	8,5	0,12	0,79
Longitud de panícula (cm)	20,1	9,6	34,4	3,8	0,19	0,81
Número de granos por panícula	93,8	22,0	264,0	37,9	0,40	0,77
Número de panículas	23,6	4,0	59,0	9,7	0,41	0,54

Estimación de la media, valores mínimos, máximos, desviación estándar (SD) y coeficiente de variación (CV) para los caracteres de altura, tiempo de floración, longitud de panícula, número de granos por panícula y número de panículas para todos los experimentos. La heredabilidad en sentido amplio (H2) se calculó para los experimentos en dos años.

Las correlaciones exhibidas por los caracteres agronómicos entre cada uno de ellos se muestran en la tabla 6. La longitud de panícula mostró una elevada correlación con la altura de la planta. El tiempo de floración mostró una correlación moderada con la altura y la longitud de panícula. La longitud de panícula presentaba una correlación baja con el número de granos. Otros caracteres mostraron unas correlaciones débiles o ninguna correlación entre ellos. Para identificar las relaciones entre los caracteres para analizar la distribución de los cultivos de acuerdo a sus características agronómicas, se realizó un análisis de componentes principales empleando las variables fenotípicas estudiadas en este trabajo. El primer componente principal (PC1), el cual comprendía el 45,9% de la varianza, separó a los cultivos de acuerdo a su longitud de panículas, número de granos por panícula y altura (Figura 16). En concordancia con la correlación observada para los caracteres mostrada en la tabla 2. Por otro lado, el componente principal 2, explicaba el 20,1 % de la varianza y distribuía las variedades según el número de panículas.

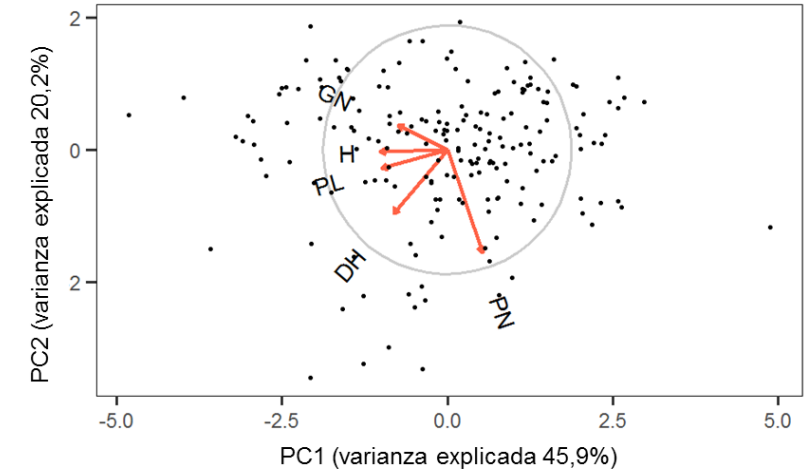


Figura 16: Gráfica del análisis de componentes principales (PCA). La proporción de varianza explicada por el primer componente principal (PC1) y el segundo componente principal (PC2) se indica en paréntesis. DH, tiempo de floración; H, altura de la planta; PN, número de panículas; PL, longitud de panícula y GN, número de granos por panícula.

Tabla 6: Coeficientes de correlación entre los caracteres.

Carácter	Altura	Tiempo de floración	Longitud de panícula	Granos por panícula
Días a floración	0,42**			
Longitud de panícula	0,61**	0,41**		
Número de granos por panícula	0,32**	0,29**	0,28**	
Número de panículas	-0,30**	-0,02	-0,21*	-0,24**

** significante a P < 0,001. * significante at P < 0,01

Análisis de asociación

Con el fin de detectar la asociación entre los marcadores SNP y las variaciones en los caracteres evaluados en las 193 variedades *japónica* de la colección, empleamos el set de datos genotípicos generado en un estudio previo, que consistía en un panel de 1.713 SNPs uniformemente distribuidos a lo largo del genoma, a excepción de las

regiones teloméricas y centroméricas, con una distancia media de 214 Kb entre cada uno (Reig-Valiente et al., 2016). De acuerdo al desequilibrio de ligamiento estimado para esta colección, que desaparece a una distancia media de 368 kb (Reig-Valiente et al., 2016) y el tamaño del genoma, 321 Mb (Kawahara et al., 2013). El número de SNPs en el panel es adecuado para detectar asociaciones a lo largo del genoma. Entre los diferentes modelos a la hora de analizar la asociación entre el fenotipo y el genotipo decidimos emplear el modelo linear mixto (Mixed Linear model, MLM) puesto que produce menos falsos positivos que otros métodos (J. Yu et al., 2006). Se ha sugerido que en el caso del arroz, el MLM reduce el número de falsos positivos pero por el contrario aumenta el número de falsos negativos, al sobre compensarlas correcciones debidas a la estructura de la población y al parentesco (K. Zhao et al., 2011). Hemos mostrado previamente que nuestra colección de arroz de zona templada está fuertemente estructurada mostrando varias subpoblaciones (Reig-Valiente et al., 2016). Así pues, para evitar falsas asociaciones debidas a la estructura genética de la población, he empleado dos aproximaciones, en primer lugar empleando los datos de la estructura de la población obtenidos mediante el programa STRUCTURE (Reig-Valiente et al., 2016) y en segundo lugar corrigiendo la estructura poblacional según los resultados de un PCA. Establecimos como valor umbral un p-valor <0,001 para considerar un SNP significativamente asociado a la variación de un carácter. Además, también se calcularon los q-valores. Las gráficas de cuantil-cuantil (QQ) y los gráficos de Manhattan para el análisis de los caracteres que fueron generados empleando la matriz Q obtenida con STRUCTURE se muestran en la figura 17. Los gráficos QQ para tiempo de floración, longitud de panícula y número de granos indican que el modelo se ajusta bien a los datos, ya que los p-valores observados se distribuyen uniformemente con desviaciones en valores elevados, respecto a los p-valores esperados.

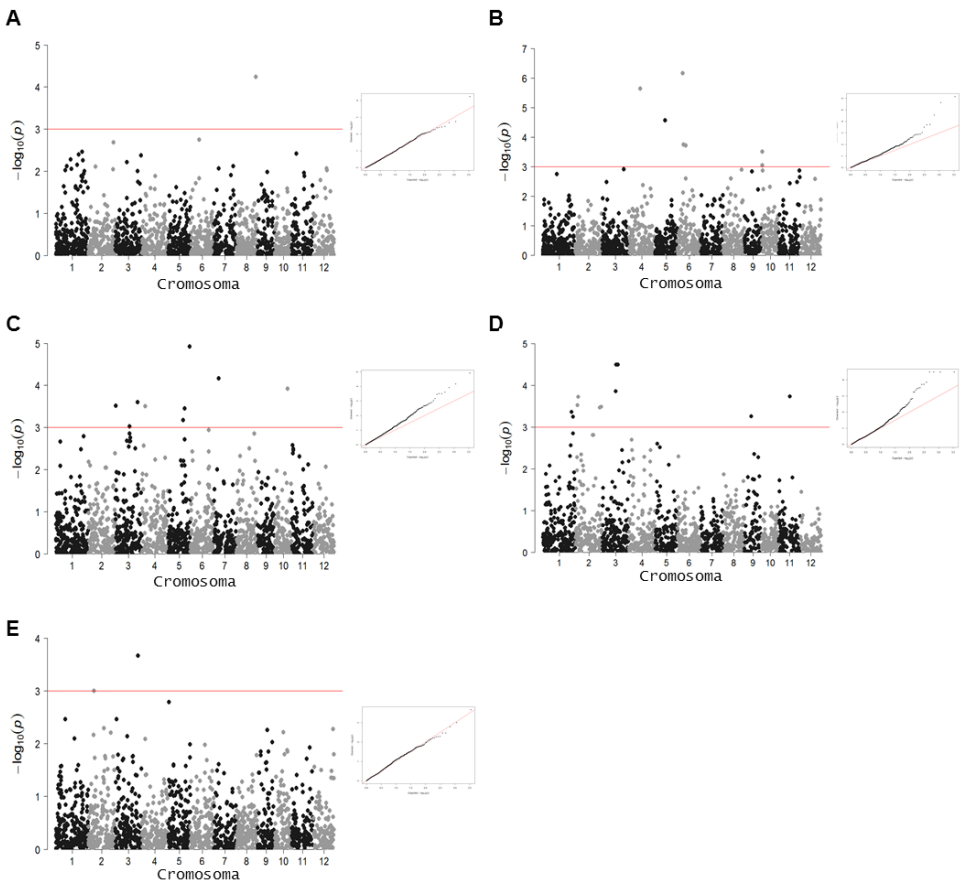


Figura 17: Gráficos Manhattan y cuantil-cuantil obtenidos en el mapeo asociativo de genoma completo para los caracteres de este estudio. (A) altura de la planta, (B) tiempo de floración, (C) longitud de panícula (D) número de granos por panícula y (E) número de panículas. La línea roja horizontal indica el umbral de significancia para las asociaciones de genoma completo

El análisis estadístico reveló 43 SNPs significantemente asociados con los caracteres estudiados (Tabla 7). Los SNPs identificados se distribuyeron a lo largo del genoma, a excepción del cromosoma 12, donde ningún SNP fue detectado (Figura 18). La varianza explicada, dada por el R^2, abarcó entre un 5,2% y 14,3%. Algunos SNPs identificados fueron corroborados mediante las dos aproximaciones, de acuerdo a la

estructura poblacional definida por el programa STRUCTURE y el PCA, ambas aproximaciones presentaron 18 SNPs en común con un p-valor más bajo que el umbral escogido para la significancia. Dado el LD estimado y la distancia entre los diferentes SNPs identificados, estos pueden agruparse en 33 loci asociados a los caracteres estudiados (Figura 18, Tabla 7).

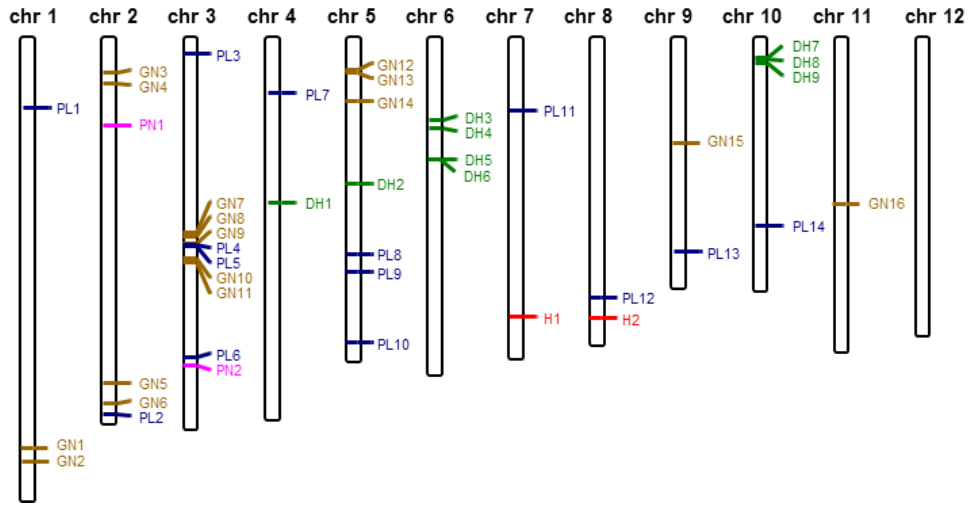


Figura 18: Posición física de los SNPs significativamente asociados detectados mediante GWAS. Los sitios identificados están anotados en verde oscuro para tiempo de floración (DH), en rojo para altura (H), en azul para longitud de panícula (PL) en naranja para el número de granos por panícula (GN) y en rosa para el número de panículas (PN).

Tabla 7. Lista de SNPs significativamente asociados con los caracteres estudiados.

				STRUCTURE			PCA		
SNP	**Número de loci**	**Chr**	**Posición**	**p-valor**	**q-valor**	**% varianza explicado**	**p-valor**	**q-valor**	**% varianza explicado**
H-1	1	7	26.125.810				8,53E-04	0,691	5,5
H-2	2	8	26.290.663	5,74E-05	0,098	8,0	1,54E-05	0,026	9,0
DH-1	3	4	15.171.082	2,33E-06	0,002	14,3	1,01E-05	0,009	12,5
DH-2	4	5	13.373.075	2,76E-05	0,014	8,9	6,75E-05	0,031	8,1
DH-3	5	6	7.249.825	6,91E-07	0,001	12,4	1,67E-07	0,001	13,5
DH-4	6	6	8.122.336	1,82E-04	0,058	7,2	7,18E-05	0,031	8,0
DH-5	7	6	11.002.220				7,12E-04	0,185	7,4
DH-6	7	6	11.018.229	1,94E-04	0,058	7,1	9,60E-05	0,033	7,7
DH-7	8	10	1.366.920	9,09E-04	0,181	7,2			
DH-8	8	10	1.500.932	3,10E-04	0,078	6,7			
DH-9	8	10	1.871.964				7,59E-04	0,185	5,8
PL-1	9	1	6.126.957				7,34E-05	0,054	7,7
PL-2	10	2	35.406.962				6,78E-04	0,119	5,8
PL-3	11	3	1.003.670	3,09E-04	0,081	6,7	6,48E-04	0,119	5,9
PL-4	12	3	19.250.700				6,99E-04	0,119	7,2
PL-5	12	3	19.349.943	9,37E-04	0,169	8,1	5,93E-04	0,119	8,5
PL-6	13	3	29.987.079	2,49E-04	0,081	6,8	8,55E-04	0,131	5,5
PL-7	14	4	4.758.113	3,15E-04	0,081	6,6	1,98E-04	0,084	6,9
PL-8	15	5	20.125.768	6,72E-04	0,136	5,9			
PL-9	16	5	21.838.857	3,52E-04	0,081	6,5			
PL-10	17	5	28.626.704	1,21E-05	0,020	11,3	1,43E-05	0,024	10,8
PL-11	18	7	6.375.823	6,80E-05	0,055	9,7			
PL-12	19	8	24.250.565				3,15E-04	0,108	8,0
PL-13	20	9	19.877.778				9,50E-05	0,054	7,5
PL-14	21	10	17.376.318	1,19E-04	0,064	9,2	4,78E-04	0,119	7,6
GN-1	22	1	38.011.958	4,33E-04	0,064	6,4			
GN-2	23	1	40.012.174	5,64E-04	0,070	6,1			
GN-3	24	2	2.720.687	2,99E-04	0,056	6,7			

				STRUCTURE			PCA		
SNP	**Número de loci**	**Chr**	**Posición**	**p-valor**	**q-valor**	**% varianza explicado**	**p-valor**	**q-valor**	**% varianza explicado**
GN-4	25	2	3.755.851	1,89E-04	0,044	7,1			
GN-5	26	2	32.500.801	3,46E-04	0,056	8,1			
GN-6	27	2	34.386.360	3,25E-04	0,056	6,6			
GN-7	28	3	17.999.538	1,39E-04	0,044	7,4			
GN-8	29	3	18.500.293	3,23E-05	0,013	8,7	9,63E-04	0,197	5,2
GN-9	30	3	19.125.149	3,23E-05	0,013	8,7	9,63E-04	0,197	5,2
GN-10	31	3	20.500.343	3,23E-05	0,013	8,7	9,63E-04	0,197	5,2
GN-11	31	3	20.875.010	3,23E-05	0,013	8,7	9,63E-04	0,197	5,2
GN-12	32	5	2.460.569				7,67E-04	0,197	5,2
GN-13	32	5	2.753.203				1,48E-04	0,197	6,6
GN-14	33	5	6.499.710				9,98E-04	0,197	6,5
GN-15	34	9	9.375.311	5,58E-04	0,070	6,2	9,06E-04	0,197	5,4
GN-16	35	11	15.123.344	1,87E-04	0,044	7,2			
PN-1	36	2	7.879.224	9,94E-04	0,710	5,5			
PN-2	37	3	30.752.398	2,15E-04	0,318	6,9	1,20E-04	0,171	6,5

SNPs asociados a la altura (H), tiempo de floración (HD), longitud de panícula (PL), número de granos por panícula (GN) y número de panículas (PN) de acuerdo a los análisis con corrección de estructura empleando la Q matriz del STRUCTURE o de PCA.

Entre los sitios identificados, 9 SNPs estaban significativamente asociados a la variación del tiempo de floración (Tabla 7) y 7 de ellos no han sido descritos previamente. La distribución alélica de algunos de los marcadores asociados en las variedades de la colección fueron prometedoras, particularmente DH-8 (Fichero adicional 10. Tabla S7). Las variedades portadoras de diferentes alelos del SNP DH-2, localizadas en la posición 1.337.075 del cromosoma 5, florecieron de media 17 días antes que las variedades portadoras del alelo alternativo (fichero adicional 9. Tabla S6). Los valores medios para el tiempo de floración de las variedades portadoras de diferentes alelos en el marcador DH-7, en la posición 1.366.920 el cromosoma 10 fue 7 días (Fichero adicional 9: Tabla S6). Además, para las 20 variedades más tempranas, se confirmó que 19 presentaban el alelo G en la posición DH-7, mientras que las 20 más tardías presentaron el alelo A. Con el fin de encontrar posibles genes relacionados con la floración en las cercanías de los SNPs identificados, se analizó las regiones colindantes en un intervalo de 368 kb de acuerdo al decaimiento del desequilibrio de ligamiento. DH-3 y DH-4 colocalizaron con genes o QTLs previamente descritos. Dos QTLs, *photoperiod sensitive phase 6 (qPSP-6)* y *qDTH6*, implicados en la regulación de la floración mediados por el fotoperiodo (Cai & Morishima, 1998; Fujino & Sekiguchi, 2005) mostraron una relación posicional con DH-3, localizado en la posición 7.249.825. Además, DH-4 está relativamente próximo, a menos de 0,6 Mb, a *Su-Se1* (Tabla 8), un QTL que también está implicado en la regulación de la floración mediante fotoperiodo (Yu et al., 2005). El valor medio para el tiempo de floración de las variedades portadoras de diferentes alelos de DH-3 mostró una diferencia de 5 días en la floración (fichero adicional 10: Tabla S6). Para las 20 variedades más tempranas, 13 fueron confirmadas para portar el alelo G de DH-3 mientras que 19 variedades de las 20 más tardías también presentaban el mismo alelo.

Tabla 8.- Lista de genes o QTLs relacionados con los caracteres estudiados en un interval de 0,4 Mb de distancia de los loci detectados mediante GWAS.

SNP	QTL / gen	QTL/Nombre del gen	Locus	chr	Posición	Referencia
H-1	*DEP2*	*cleistogamy gene*	Os07g0616000	7	26041969-26049689	Li et al 2010
DH-3	qPSP-6	photoperiod sensitive phase 6	-	6	6720901-8066362	Cai et al 1998
DH-4	qDTH-6	-	-	6	8054255-8066362	Fujino et al 2005
DH-4	Su-Se-1(t)		-	6	8054255-8066362	Yu et al 2005
PL-6	*rip-3*	*rice panicle 3* (α-tubulin)	Os03g0726100	3	30288045-30290829	Sheoran et al, 2014
PL-8	*sdg / gid1*	*semidwarf gene g / gibberellin insensitive dwarf1*	Os05g0407500	5	19875232-19878096	Ueguchi-Tanaka et al 2005; Zhang et al 2012
PL-13	*LGD1*	*Lagging Growth and Development 1*	Os09g0502100	9	20078670-20084674	Thangasamy et al 2012
PL-14	qssd10	spikelet setting density 10	-	10	12044545-19623828	Xiao et al 1996
GN-3, GN-4	*LRK1*	*leucine-rich repeat receptor-like kinase1*	Os02g0154200	2	2978853-2982354	Za et al 2009

El carácter de longitud de panícula mostró 14 marcadores asociados, distribuidos a lo largo de 9 cromosomas (Figura 18). Cinco de estos marcadores colocalizaban con genes implicados en la altura de la planta o la longitud de la panícula. PL-6, en el cromosoma 3 colocalizaba con *rice panicle 3 (rip-3)*, que codifica para una α tubulina que probablemente participa en la supresión de la elongación de la panícula durante el déficit hídrico (Sheoran, Koonjul, Attieh, & Saini, 2014). *Semidwarf gene g / gibberellin insensitive dwarf1* (*sdg / gid1*) colocaliza con PL-8, en el cromosoma 5 (Tabla 8). Los mutantes portadores del alelo nulo de *gid1* mostraron fenotipo enano (Sui et al., 2012; Ueguchi-Tanaka et al., 2005). El alelo A en la posición PL-8 estaba presente en 17 variedades de la 20 con las panículas más cortas. Por otro lado, el alelo G estaba presente en 17 de las 20 variedades con las panículas más largas (Fichero adicional 3: Tabla S2). De modo similar, el alelo G en PL-9, en el mismo locus que PL-8 estaba presente en 17 de las 20 variedades con la panícula más corta, mientras que el alelo A estaba presente en todas las variedades correspondientes a las 20 con las panículas más largas.

PL-13, asociado con la longitud de la panícula se encuentra en el cromosoma 9, colocaliza con *LAGGING GROWTH AND DEVELOPMENT 1(LGD1)* implicado en el crecimiento y en la formación de la panícula. Los mutantes *lgd1* presentan alteraciones en la arquitectura de la panícula (Thangasamy, Chen, Lai, Chen, & Jauh, 2012). Finalmente, el QTL *ssd10* (*spikelet setting density*) se encuentra en la misma posición que el marcador PL-14 en el cromosoma 10 (Xiao, Li, Yuan, & Tanksley, 1996).

Detectamos 16 marcadores significativamente asociados al número de granos por panícula. GN-3 en el cromosoma 2 colocaliza con *leucine-rich repeat receptor-like kinase 1 (LRK-1*) (Tabla 8). LRK1 es una proteína de membrana plasmática que, supuestamente, regula la ramificación mediante la estimulación de la proliferación celular (Zha et al., 2009). El marcador en el cromosoma 7 estaba asociado a la altura. H-1 colocalizó con *dense and erect panicle (DEP2*) (F. Li et al., 2010). *DEP2* está

implicado en la elongación del raquis y la ramificación primaria y secundaria en las panículas, las plantas portadoras de mutaciones en DEP2 muestran panículas más pequeñas pero también altura reducida (F. Li et al., 2010).Finalmente, dos marcadores más, en el cromosoma 2 y 3 se encuentran asociados con el número de panículas (Tabla 3). Sin embargo ningún gen candidato ni QTL se sitúa en las proximidades.

3.2.3. Discusión

Tras la domesticación, a pesar de las condiciones de floración no favorables, el área de cultivo del arroz se expandió a lo largo de las regiones de clima templado. La planta de arroz siempre ha sido considerada como una planta de día corto, es decir, su floración se induce en días de corta duración mientras que es inhibida cuando la duración del día es larga. Sin embrago las plantas de arroz que se cultivan en clima templado florecen en condiciones de fotoperiodo largo. Durante la expansión del cultivo, para alcanzar latitudes septentrionales, el arroz tuvo que modular la sensibilidad a fotoperiodo. Las elevadas temperaturas de verano, acompañadas de un mayor número de horas de luz con buena radiación solar constituyeron unas condiciones ambientales excelentes permitiendo la expansión y originando una gran diversidad que se ve reflejada en diferentes caracteres agronómicos, así como se muestra en este estudio. La diversidad de las plantas ya adaptadas a las condiciones de fotoperiodo es un valioso recurso para los mejoradores puesto que puede aportar parentales apropiados para las condiciones agronómicas locales. La mayoría de los análisis realizados previamente sobre la identificación de las variaciones en genes de interés agronómico han sido realizadas comparando las variedades *japónica* e *indica*. Así pues, los factores implicados en las variaciones fenotípicas dentro de la subpoblación *japónica* necesitan ser investigados.

En este estudio hemos empleado el GWAS como un método efectivo para detectar la asociación entre regiones genómicas y el fenotipo (Huang et al., 2010; McCouch et al., 2016) de las variedades que se cultivan a lo largo de las regiones de clima templado.

Hemos empleado una colección de variedades generada previamente, incluyendo cultivos de 23 países que representan la diversidad genética presente en las regiones de clima templado y constituye un recurso útil para la identificación de factores genéticos que gobiernan las variaciones en los caracteres agronómicos de esta región. De hecho, esta colección ha sido probada exitosamente a la hora de detectar polimorfismos asociados con la tolerancia al frío (Sales, Viruel, Domingo, & Marqués, 2017). Los análisis realizados en este estudio han tenido en cuenta la estructura de la población y las relaciones de parentesco entre las variedades, dos factores que pueden llevar a falsas asociaciones en los estudios de GWAS. Hemos considerado dos posibles aproximaciones para solventar estos problema, uno que la población está estructurada en 4 grupos principales, de acuerdo a los resultado obtenidos con el programa STRUCTURE (Reig-Valiente et al., 2016) y una segunda corrigiendo la estructura de acuerdo al resultado de un PCA. Los análisis realizados con ambas matrices Q dieron resultados similares y la mayoría de los SNPs asociados fueron identificados con cualquiera de las dos aproximaciones es el caso de la altura, la longitud de panícula y el número de panículas (Tabla 7). Hay que resaltar que para DH-7, DH-8 y DH-9, los cuales se sitúan en intervalos de 505 Kb, esta asociación fue detectada mediante una u otra aproximación.

La floración es un carácter de gran interés para los mejoradores. Una floración en el momento óptimo, apropiada a la longitud del día de cada región es necesaria para obtener un rendimiento máximo. El control de la floración a través del fotoperiodo ha sido, supuestamente, uno de los principales factores que han contribuido a la expansión del arroz a través de las regiones con condiciones de día largo. Así pues es un factor clave en la adaptación de los cultivos a cada condición local ya que determina la latitud a la que son cultivados. La regulación de la floración mediante fotoperiodo ha sido estudiada extensivamente, a día de hoy se sabe que la delicada regulación de *Hd3a* y *RFT1* promueven la floración bajo condiciones de día largo o

corto (Komiya, Yokoi, & Shimamoto, 2009). Mientras que *Hd3a* dicta la transición del arroz de la fase vegetativa a la reproductiva bajo condiciones de día corto, *RFT1* ha sido descrito como el principal activador de la floración en día largo. Las temperaturas frías del invierno en las regiones de clima templado restringen el cultivo del arroz, impidiendo dos cosechas por año, incluso en el caso de las variedades de floración temprana. Pero la reducción de la fase de crecimiento en algunos días sigue siendo una característica deseable por los agricultores dado que aumenta la seguridad de la cosecha al reducir el riesgo de ataques de patógenos o condiciones climáticas adversas, como las tormentas al final de la temporada de cultivo, que son frecuentes en la costa este de España. Todos los cultivos de nuestra colección de clima templado pudieron florecer bajo condiciones de día largo y los tiempos de floración abarcaron desde los 47 hasta los 106 días (Tabla 5). Seis loci aparecieron asociados con el tiempo de floración. Para este estudio esperábamos encontrar las fuentes de la variación en los tiempos de floración entre las variedades ya adaptadas a las condiciones de día largo, por tanto, no dependientes de fotoperiodo. Pero sorprendentemente, un marcador en el cromosoma 6, DH4 colocalizaba con *Su-Se-1* y qDTH6, QTLs que fueron identificados en estudios de regulación por fotoperiodo (Tabla 8). *Su-Se-1* es un gen supresor sensible a fotoperiodo dominante identificado en un cruce entre Asominori x IR24 (YU et al., 2005). qDTH6 fue identificado en un cruce entre Hayamasari, una variedad temprana de Japón, e Italica Livorno (Fujino & Sekiguchi, 2005). Estos autores sugirieron la posibilidad de que *qDTH6* pudiese ser *Hd1 un gen regulador clave en la regulación de* la floración por fotoperiodo (M. Yano et al., 2000). *Hd1* se localiza en el cromosoma 6 a una distancia de 1,2 Mb de DH-4, una distancia muy grande teniendo en cuenta la asociación entre ambas regiones genómicas. Se ha observado en algunos estudios que la heterogeneidad alélica y la compleja estructura genómica de *Hd1* causan falsas asociaciones así como un sesgo en la señal correspondiente a su posición (K. Yano et al., 2016). Hemos investigado previamente

la estructura de *Hd1* en 52 variedades incluidas en la colección y hemos descubierto 12 variantes que incluyen alelos no funcionales de *Hd1* (Naranjo et al., 2014). Hd1 es un represor de la floración bajo condiciones de día largo a través de la inhibición de la expresión de *Hd3a* (Kojima et al., 2002) y se ha sugerido que la falta de funcionalidad de HD1 ha sido crucial para la adaptación del arroz a condiciones de día largo en las regiones templadas (Goretti et al., 2017). Esta hipótesis se ve reforzada por el hecho de que la pérdida de función de alelos de HD1 es frecuente o común en las variedades de arroz cultivadas en las latitudes nortes (Fujino et al., 2010; Gómez-Ariza et al., 2015). Pero estudios previos han concluido que HD1 no está implicado en la regulación de la floración bajo condiciones de día largo, puesto que muchas variedades cultivadas en estas condiciones presentan tanto los alelos funcionales como los no funcionales de HD1 independientemente del origen geográfico de las variedades (Naranjo et al., 2014). La posibilidad de la presencia de otro factor implicado en la floración en las proximidades de *Hd1* no debería ser descartada y necesita más investigación. No se detectaron marcadores asociados al tiempo de floración en las proximidades de *Hd3a* (2.939.760-2.942.696) o *RFT1* (2.926.823-2.928.474) localizados en el cromosoma 6.

Algunos nuevos loci asociados a la fecha de floración que han sido observados en este estudio son prometedores de acuerdo a las diferencias observadas en los tiempos de floración de las variedades portadoras de uno u otro alelo. En este sentido, DH-7 y DH-8 en el cromosoma 10, que representan el 7,2% y el 6,7% de la varianza respectivamente (Tabla 7). Si estos loci están asociados a la regulación por fotoperiodo de la floración es a día de hoy desconocido siendo necesario su estudio.

Varios loci fueron identificados como asociados a la longitud de panícula. No encontramos ningún gen caracterizado funcional en las proximidades de estos loci, por lo que consideramos que representan nuevos sitios asociados a la variación en este carácter. Algunos de los SNPs detectados estaban próximos a genes que

participan en la elongación de la panícula pero que también afectan a otras partes de la planta. PL-8, que se encuentra a una distancia de 0,5 Mb de *SALT-RESPONSIVE ERF1 (SERF1)*, es un factor transcripcional que actúa como regulador negativo del llenado del grano. Los mutantes portadores del alelo nulo SERF1 presentan granos más largos y una longitud de panícula alterada (Schmidt et al., 2014). *OsEBS*, se encuentra próximo a PL-10, que produce un mayor número de espiguillas en las panículas aumentando así la productividad de número de granos total a pesar de que también afecta a la altura de la planta y al tamaño de la hoja (Dong et al., 2013).

Las líneas de introgresión portadoras del alelo *OsEBS* de *Oryza rufipogon* también producen panículas más largas que el tipo silvestre (Dong et al., 2013). Se han detectado variedades de arroz enano con sobreexpresión de genes inducidos por giberelinas (*OsDOG*) implicado en la homeostasis de giberelinas se situaba a 0,8 Mb d PL-12 y a menos de 1,2 Mb de H-2. OsDOG es una proteína tipo dedos de zinc A20/AN1 que regula negativamente la elongación celular regulada por giberelinas. Las plantas transgénicas que expresan *OsDOG* presentan fenotipos enanos y unas panículas más cortas que no llegan a emerger de la vaina debido a la deficiencia en la elongación celular (Liu et al., 2011). En un estudio reciente *OsDog* se asoció con la longitud de panícula en el arroz tipo *japónica* cultivado bajo condiciones de inundación permanente (Volante et al., 2017). *LAGGING GROWTH AND DEVELOPMENT 1 (LGD1)* también se encuentra próximo a PL-13. El mutante *lgd1* genera efectos pleiotrópicos en arroz, mostrando un crecimiento lento, reducido número de tallos y altura de planta, arquitectura de planta alterada y productividad de granos reducida (Thangasamy et al., 2012). Por otro lado, PL-6 está situado próximo a *rice panicle 3 (rip 3)* que codifica una α-tubulina putativa, cuya expresión se ha visto observada en todos los órganos reproductivos de la panícula de arroz, pero no en otras partes de la planta. Interesantemente, *rip-3* actúa, supuestamente, como supresor de la elongación en las regiones con elevado crecimiento y en periodos de

déficit hídrico (Sheoran et al., 2014). La longitud de la panícula está frecuentemente asociada con la productividad puesto que las panículas largas son capaces de producir un mayor número de espiguillas, y por tanto un mayor número de granos (S. Li et al., 2009). Pero en nuestra colección no se observó una elevada correlación entre la longitud de la panícula y el número de granos por panícula (Tabla 6). Resultado que coinciden con el hecho de que no se detectaron marcadores significativamente asociados en común cuando se analizan tanto la longitud de panícula como el número de granos por panícula.

3.2.4. Conclusiones

En este estudio de asociación hemos identificado marcadores moleculares relacionados con importantes caracteres de interés agronómico entre variedades adaptadas a condiciones de fotoperiodo templado, así pues, con una aplicación directa en los programas de mejora. Algunos de estos marcadores colocalizan con genes o QTLs ya conocidos, lo cual valida nuestra metodología, como han sido los casos de *OsEBs*, *OsDOG* o *LGD1*. Nuestros hallazgos también aportan nuevos marcadores moleculares que pueden ayudar a elucidar los complicados mecanismo genéticos que controlan importantes caracteres agronómicos en las variedades de arroz *japónica*, como es la floración bajo condiciones de día largo

3.2.5. Materiales y métodos

Material vegetal, condiciones de cultivo y fenotipado

Se empleó una colección de variedades de arroz *japonica* compuesta de 193 variedades obtenidas de diferentes fuentes (Reig-Valiente et al., 2016): IRRI, U.S. National Plant Germplasm System (NPGS, EE.UU), Rice Genome Resource Center (RGRC, Japón), Copsemar y el IVIA. Las semillas de diferentes variedades fueron cultivadas en dos localizaciones en 2015 y 2016 en verano, el Tancat de Malta y la Finca de Raga, en la zona de cultivo de arroz en Valencia, España. Las plántulas fueron

germinadas en macetas y trasplantadas manualmente a tierra en dobles filas de 20 plantas en mayo y cosechadas en septiembre. Los campos se mantuvieron en irrigación continua mediante inundación y fueron secados dos semanas antes de la cosecha. Dos baterías adicionales de plantas fueron cultivadas por separado durante el verano de 2017 en invernadero bajo condiciones de día natural

La altura y el tiempo de floración fueron anotados en el campo. Se consideró el momento de la floración cuando la panícula había emergido en el 50% de las plantas de cada variedad. Se recolectaron tres plantas por variedad en cada localización y se anotó el número de panículas y la longitud de las panículas, tres panículas por planta. El número de granos por panícula fue medido a partir de tres panículas por planta, de las plantas cultivadas en invernadero.

Análisis estadístico

El análisis de los datos fenotípicos (media, coeficiente de variación y los histogramas de distribución de frecuencias) fueron obtenidos empleando Statgraphics Plus (Version 16.1.03 (32 bits)). Las correlaciones entre los valores de los caracteres se analizaron empleando el test de coeficiente de correlación de Pearson implementado en el software R (versión 3.4.1) (http://www.R-project.org). La heredabilidad en sentido amplio fue calculada empleando la regresión lineal entre la media del valor para las variedades en el año 2015 y 2016 en el Tancat de Malta y entre los años 2015 y 2016 en la Finca de Raga, empleando el comando "lm" de R. La regresión para el número de granos se realizó entre las diferentes baterías cultivadas en el invernadero.

Análisis de asociación entre marcadores y caracteres.

Para el análisis de asociación entre marcadores y carácter, se empleó el panel de 1.713 SNPs obtenido en un estudio previo (Reig-Valiente et al., 2016). Este panel fue generado a partir de un panel de genotipado de SNPs Infinium (Illumina) de 2.697 SNPs específicos de variedades de arroz *japonica* originarias de países con clima

templado, tras la eliminación de los SNPs que no estuviesen presentes en al menos el 75% de las variedades de la colección de arroz, o mostrando una frecuencia del alelo menos común (MAF) menor a un 5%. La asociación entre los marcadores y los caracteres fue analizada empleando el programa Tassel (versión 5.2.32). Se empleó un modelo lineal mixto (MLM) con corrección para la estructura (Q) y las relaciones de parentesco (K) a fin de evitar falsas asociaciones (J. Yu et al., 2006). La matriz de parentesco se obtuvo empleando el programa Tassel, y se emplearon dos matrices Q. La primera matriz Q fue obtenida a partir de los resultados obtenidos con el programa STRUCTURE (v. 2.3.4) (Pritchard et al., 2000) reteniendo los valores de pertenencia a 3 de los 4 grupos en los que se divide la colección de forma óptima (Reig-valiente et al, 2016). Posteriormente se realizó un análisis de componentes principales empleando Tassel y mediante la visualización del gráfico de varianza por componente se seleccionaron 7 componentes principales para el análisis, los cuales retenían un 49% de la varianza total (fichero adicional 11: Tabla S8). El valor umbral para considerar significativa una asociación fenotipo marcador fue de $p<10^{-3}$ puesto que se ha indicado que el MLM reduce el número de falsos positivos pero a su vez genera un mayor número de falsos negativos. También se calcularon los Fdr (Tasa de descubrimientos falsos, *False discovery rate*) empleando el pack "qvalue" 1.99.1. Se generaron gráficos de Manhattan empleando el pack "qqman" (versión 0.1.4) para R con ligeras modificaciones.

Para la búsqueda de genes funcionales caracterizados situados en la proximidad de los SNPs detectados o dentro de un intervalo de 0,3 Mb se usó la base de datos de referencia de Nipponbare (Os-Nipponbare-Reference-IRGSP-1.0), se exploró en la base de datos OGRO (E. Yamamoto, Yonemaru, Yamamoto, & Yano, 2012) y, además, en la base de datos Q-TARO (Yonemaru et al., 2010) para detectar posibles QTLs para el mismo carácter.

3.3. Caracterización de un mutante de floración temprana e identificación de la mutación responsable del fenotipo alterado

3.3.1. Introducción

En este capítulo se describe la caracterización de la línea mutante de arroz *G123*, que presenta un ciclo vegetativo más corto que la variedad parental y además, *G123* es insensible al fotoperiodo, a diferencia de la línea parental.

La regulación de la floración en arroz es un mecanismo finamente ajustado que ha formado parte del proceso de adaptación a nuevas condiciones climáticas durante la expansión del cultivo, desde su domesticación en una región de clima tropical hasta alcanzar latitudes con clima templado. Se conocen dos rutas independientes de regulación de la floración que involucran a dos genes reguladores, *Heading date 1* (*Hd1*) y *Early heading date 1* (*Ehd1*). Estas dos rutas convergen en la modulación de la expresión de los dos genes máster inductores de la floración, *Heading date 3a* (*Hd3a*) y *Rice Flowering Locus T1* (*RFT1*), los cuales codifican el conocido florigen, la señal móvil que se produce en las hojas y se transporta hasta el meristemo apical para inducir el paso de estado vegetativo a reproductivo (Komiya et al., 2009; Tamaki, Matsuo, Wong, Yokoi, & Shimamoto, 2007; Yan et al., 2011). Hoy en día se conoce que mientras la expresión de *Hd3a* es la inductora de la floración en condiciones de día corto, en condiciones de día largo la inducción viene dada por la expresión de *RFT1* (Komiya et al, 2009). Al igual que en otras plantas, la apreciación de la longitud del día viene dada por los fitocromos. Estas moléculas son proteínas que actúan como fotorreceptoras gracias a un cromóforo que forma parte de su estructura. El papel regulador de los fitocromos en la floración se ha puesto de manifiesto en varios estudios que incluyen los realizados con los mutantes de pérdida de función de *SE5* (Andrés, Galbraith, Talón, & Domingo, 2009; Takeshi Izawa, Oikawa, Tokutomi, Okuno, & Shimamoto, 2000). El gen *SE5* codifica una hemo-oxigenasa implicada en la síntesis de la fitocromobilina, el cromóforo de los fitocromos. La falta de funcionalidad de *SE5* resulta en plantas con fitocromos inactivos por no disponer de cromóforo y, en consecuencia, son defectuosas en la detección de la luz. Puesto que

los fitocromos son inhibidores de la floración, de manera directa o indirecta a través de Hd1, las plantas defectuosas en *SE5* florecen de manera prematura. Es destacable que*, Hd1*, el ortólogo de *CONSTANS en Arabidopsis thaliana*, presenta una doble función en arroz, actuando como inductor y represor de la floración dependiendo del fotoperiodo en el que se cultive la planta (M. Yano et al., 2000). Inicialmente se pensó que las dos rutas de regulación de la floración eran independientes pero numerosos estudios recientes indican una interacción entre ellas. Por ejemplo, se ha observado que *Ehd1* es capaz de reprimir la expresión de *Hd1* en condiciones de día largo (Andrés et al 2009) y, por otro lado, *Hd1* es capaz de inhibir la expresión de *Ehd1* también en condiciones de día largo (Goretti et al., 2017). Este método de regulación parece llevarse a cabo mediante un sistema de andamiaje mediante la formación de dímeros y trímeros de diferentes elementos reguladores de la floración y relacionados con el ciclo circadiano, en la que un mismo elemento puede actuar como inductor o represor en función de con qué otros interactúe (Goreti et al,. 2017). También los fitocromos intervienen en la regulación puesto que actúan como inductores de genes cuyos productos génicos pueden formar parte de estos complejos, por ejemplo al regular la expresión de *Ghd7* y desplazar los picos de expresión en relación a los ciclos del ritmo circadiano (Lee, Yi, & An, 2016; Shrestha, Gómez-Ariza, Brambilla, & Fornara, 2014; Yoshitake et al., 2015). De manera adicional, se ha observado que su efecto sobre diferentes genes varía en función del momento de desarrollo de la planta, indicando que la inducción de la floración es un proceso regulado por múltiples elementos cuyo papel varía en función del estado fenológico de la planta.

El crecimiento y la producción de las planta de arroz dependen en gran medida de las condiciones agro-climáticas de la zonas de cultivo. La fecha de floración de las plantas de arroz es uno de los factores que afectan al rendimiento de las plantas. Las variedades adaptadas a una región concreta presentan un crecimiento pobre o un ciclo vegetativo inadecuado en otra, sin llegar a florecer en muchas ocasiones. Este

motivo justifica el hecho de que los programas de mejora se realicen de manera local con variedades adaptadas a cada región y, además, reduce la disponibilidad de parentales adecuados.

En este apartado se ha generado, detectado y caracterizado fenotípica y genotípicamente un mutante de floración temprana derivado de la variedad Gleva, ampliamente cultivada en España, con el objetivo de comprender los factores implicados en la regulación de la floración.

3.3.2. Resultados.

Obtención y caracterización fenotípica de un mutante con fenotipo de floración temprana.

Con el fin de obtener plantas con características agronómicas diferenciales a Gleva, se rastrearon plantas M2 provenientes de la irradiación de semillas con neutrones rápidos, cultivadas en campo. Se anotó la altura, el número de panículas, el peso de las panículas, el peso del grano y se observó la diferencia de floración con respecto a Gleva. Se seleccionaron plantas que mostraron diferencias en algunos de estos aspectos (tabla 9). Algunas líneas florecieron más de una semana antes o después de la línea silvestre Gleva y fueron seleccionadas con el fin de estudiar la regulación de la floración en arroz. Entre las líneas seleccionadas cuatro plantas provenientes de la familia M2 *G123* destacaron por florecer dos semanas antes que Gleva.

Las semillas de las plantas seleccionadas se sembraron en semilleros y se cultivaron en invernadero, con la finalidad de confirmar el fenotipo de floración temprana una vez más, bajo condiciones controladas.

Tabla 9: Datos de campo de las líneas mutantes para los caracteres de altura, número de panículas, peso de las panículas, peso del grano y floración respecto a Gleva, (A, anterior, P, posterior).

Línea	Altura (cm)	Nº panículas	Peso grano (g/planta)	Peso/panícula	Floración
G3.1	75	28	90	3,21	
G4.1		13	43,4	3,34	P
G5.1	60	9	22	2,44	A
G6.2	70	24	70,6	2,94	
G6.1	70	29	91	3,14	
G8.1	65	25	102	4,08	
G10.1		31	88,6	2,86	
G14.2	75	34			
G16.2	70	11	41,6	3,78	A
G16.3	66	21	50,9	2,42	
G16.4	70	13	58,1	4,47	A
G16.1		20	74	3,70	A
G35.1	72	23	79,5	3,46	
G35.2	70	24	83	3,46	
G40.1	68	36	104	2,89	
G63.1	70	24	67,2	2,80	
G63.2	68	33	103	3,12	
G88.1	71	28	102	3,64	
G91.3	70	22	81	3,68	
G97.1		30	114,2	3,81	
G107.1		16	67	4,19	A
G108.1	75	27	89	3,30	
G108.2	73	29	100	3,45	
G111.1	62	8	19	2,38	A
G112.1		10	15,4	1,54	A
G123.1	71	21	44	2,10	A
G123.2	65	34	50	1,47	A
G125.1		25	97,3	3,89	
G125.2		40	150,8	3,77	
G131.1	75	34	105	3,09	
Gleva. 2	77	14	51	3,64	
Gleva. 1	75	19	58,9	3,10	

Cuatro líneas de la familia *G123* presentaron un comportamiento idéntico al observado durante el rastreo respecto a las características fenotípicas. La línea *G123*.2.6 fue la seleccionada como representante de toda la familia para el resto del estudio y caracterización. Los siguientes dos años las semillas de *G123*.2.6 fueron

cultivadas en balsas en condiciones naturales de luz y temperatura (semillas M3 y M4).

Se procedió al análisis de la segregación del carácter con el fin de conocer si el fenotipo observado es debido a una mutación recesiva en un único gen. Para ello se realizó el cruzamiento de la línea mutante *G123* y su parental Gleva y, una vez obtenidas las semillas de la generación F2, se cultivaron 400 plantas F2 en macetas en invernadero, en junio de 2017, y se anotó la fecha de floración.

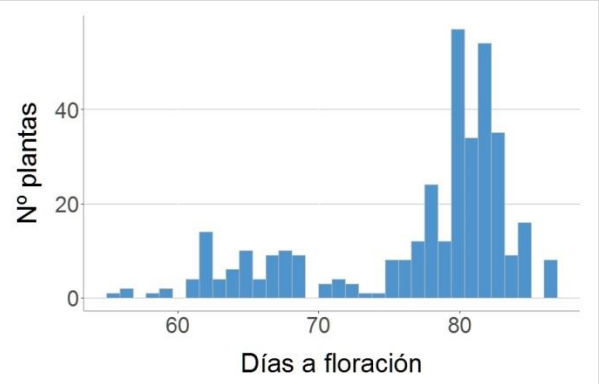


Figura 19: Tiempos de floración de las plantas F2 del cruce Gleva por *G123*.

Como se puede ver en la figura 19 las frecuencias de floración muestran una distribución bimodal. Para comprobar el modelo de un solo gen recesivo (segregación 3:1) consideramos como plantas recesivas homocigotas aquellas que presentaban un tiempo de floración más corto que 72 días tras la siembra. El modelo de 3:1 no pudo ser rechazado mediante un test de chi-cuadrado, dándose un p-valor de 0,18, corroborando el carácter recesivo de la mutación.

Figura 20: Siembra de las líneas mutantes en campo.

Caracterización fenotípica del mutante *G123*.

La floración en arroz está controlada por la duración del día. Con el objetivo de entender la variación en el tiempo de floración observada entre la línea mutante *G123*, y el parental Gleva se cultivaron plantas de ambas bajo condiciones de fotoperiodo largo (14 horas de luz y 10 horas de oscuridad) y de fotoperiodo corto (10 horas de luz y 14 de oscuridad). La intensidad de luz y la temperatura (27 °C) se mantuvieron constantes. En ambas condiciones las plantas *G123* florecieron tras 53 días desde la siembra. Mientras que las plantas Gleva florecieron tras 68 días bajo condiciones de fotoperiodo corto y 89 bajo condiciones de día largo.

El hecho de que el tiempo de floración de la línea mutante no se viese afectado por la cantidad de horas de luz al día pone de manifiesto que la línea *G123* es insensible a fotoperiodo.

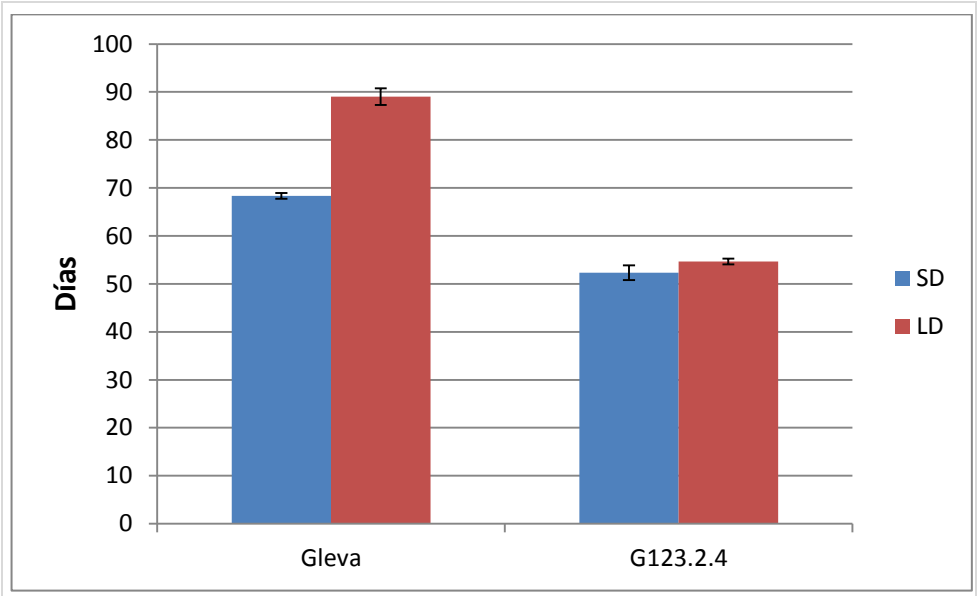


Figura 21: Tiempo de floración de las líneas Gleva y *G123* cultivadas en fitotrón bajo condiciones de fotoperiodo largo y corto a 27 °C e intensidad de luz constante.

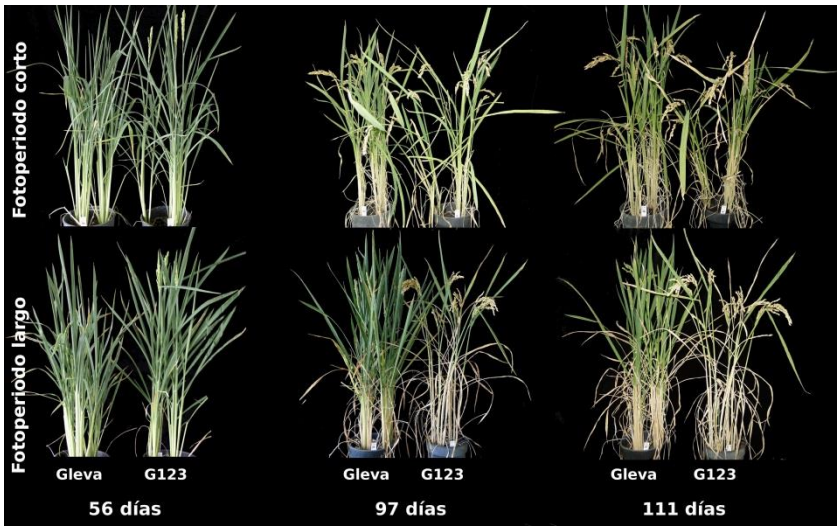


Figura 22: Variación en estadios de desarrollo de las líneas *G123* y Gleva cultivadas en fitotrón bajo condiciones de fotoperiodo corto o largo y temperatura (27°C) e

intensidad de luz constantes. Se indica el número de días tras la siembra.

Análisis de la expresión génica

RT-qPCR

Con el objetivo de analizar los niveles de expresión de los genes implicados en la ruta de regulación de la floración mediada por el fotoperiodo que pudieran aportar información acerca del fenotipo diferencial de la línea mutante, se sembraron plantas Gleva y *G123* que, una vez germinadas, fueron cultivadas cuatro semanas bajo fotoperiodo neutro (12 horas de luz y 12 de oscuridad). Durante la quinta semana la mitad de las plantas de cada línea se sometieron a condiciones de fotoperiodo largo y la otra mitad a fotoperiodo corto. Al finalizar la quinta semana se tomaron series temporales de las muestras a las 0, 4, 8, 12, 16 y 20 horas desde el encendido de la luz. Los niveles de expresión se analizaron mediante RT-qPCR (Figura 23).

Los análisis de la expresión de los dos florigenos, *RFT1* y *Hd3a*, así como de los dos genes reguladores *Hd1* y *Ehd1* de las dos rutas principales de floración mostraron los siguientes patrones:

RFT1 presenta la máxima diferencia entre ambas líneas bajo condiciones de día largo. Los niveles de expresión en Gleva aunque oscilan desde la 0 a las 16 horas, tras las 16 horas aumentan de modo que el máximo de diferencia se da a las 20 horas, al igual que con *Hd1* en fotoperiodo largo.

HD3a presenta un gran aumento de los niveles de expresión al poco tiempo de que las plantas comiencen a recibir luz. Su máximo se encuentra a las 4 horas en día corto en ambas línea y a las 8 horas en día largo para *G123*. A pesar de que en Gleva se observa un aumento en los niveles de expresión, el incremento es insignificante respecto a los observados en *G123*.

Ehd1: la expresión aumenta tanto en Gleva como en *G123* en oscuridad y se ve reprimida en presencia de luz. Bajo condiciones de día corto *G123* presenta niveles de expresión claramente más elevados que Gleva a las 4, 8 y 20 horas. Bajo condiciones de día largo Gleva presenta niveles de expresión similares a aquellos de *G123* excepto a las 4 horas del encendido de la luz, cuando *G123* presenta un pico de expresión.

***Hd1*:** los niveles de expresión de *Hd1* eran más bajos en la línea mutante que en Gleva tanto en día largo como en día corto. Encontrándose la mayor diferencia a las 20 horas bajo condiciones de fotoperiodo largo.

También se analizó la expresión de otros genes implicados en la regulación de la floración que modulan en alguna medida la expresión de *Hd1* y *Ehd1* (ver esquema en página 14):

OsGi: Tanto en la línea mutante como en Gleva los niveles de expresión comienzan a aumentar a partir de las 0 horas de luz y alcanzan el máximo en ambas variedades a las 8 horas en ciclo corto y a las 12 horas en ciclo largo. Los niveles de expresión son ligeramente superiores en Gleva.

***Ghd7*:** Su perfil de expresión es similar al de Ghd8 en la línea mutante y en Gleva.

Ghd8: En condiciones de día corto ambas líneas presentan niveles de expresión similares, a excepción del pico de 12 horas de *G123* que no está presente en Gleva. Siguiendo la misma tendencia, bajo condiciones de día largo, ambas líneas presentan niveles de expresión similares con la excepción de las 0 horas en el que Gleva presenta un pico que está ausente en *G123*.

DTH2: Los niveles de expresión tanto de Gleva como de *G123* parecen aumentar tras el apagado de las luces y tras 8 horas en oscuridad empiezan a reducirse. En los momentos de máxima expresión los niveles son más elevados en Gleva que en *G123*,

sin embargo cuando la expresión está inhibida, ambas líneas presentan niveles similares.

Prr37: En condiciones de día corto los niveles de expresión son similares en ambas líneas salvo un pico de expresión a las 12 horas en el mutante. En condiciones de día largo los niveles de expresión son ligeramente superiores en Gleva excepto en el pico de expresión de ambas líneas, que se produce también a las 12 h.

Hd6**:** Las plantas Gleva muestran los niveles de expresión máximos a las 0h en ambas condiciones, tras haber acabado el periodo de oscuridad. Se observa un segundo pico de menor altura a las 12 h en fotoperiodo corto o a las 8 en fotoperiodo largo. El mutante sin embargo en día corto presenta un pico de expresión a las 8 y 20 horas y el momento de menor expresión a las 12 horas, mientras que en día largo el mínimo se observa a las 0 horas y dos picos de expresión a las 4 y 12 horas. Bajo ambas condiciones el máximo de expresión en Gleva es superior al de *G123*.

OsELF3: La expresión de este gen se activa en ausencia de luz tanto en Gleva como en *G123*. Gleva presenta mayores niveles de expresión tanto bajo condiciones de fotoperiodo corto como largo en los momentos de máxima actividad.

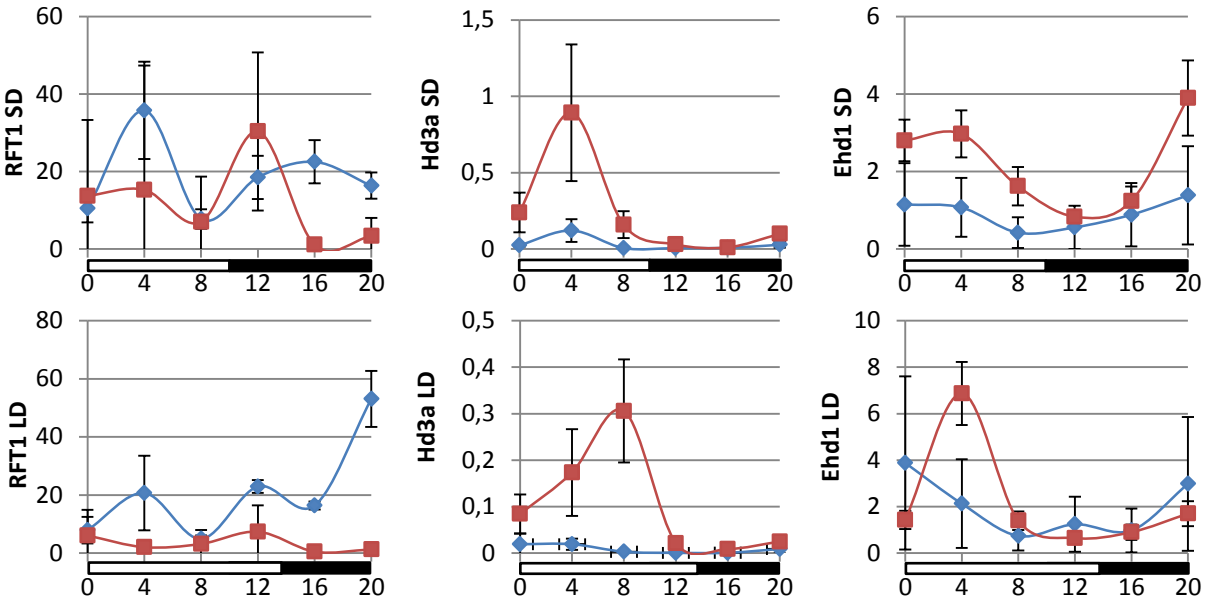
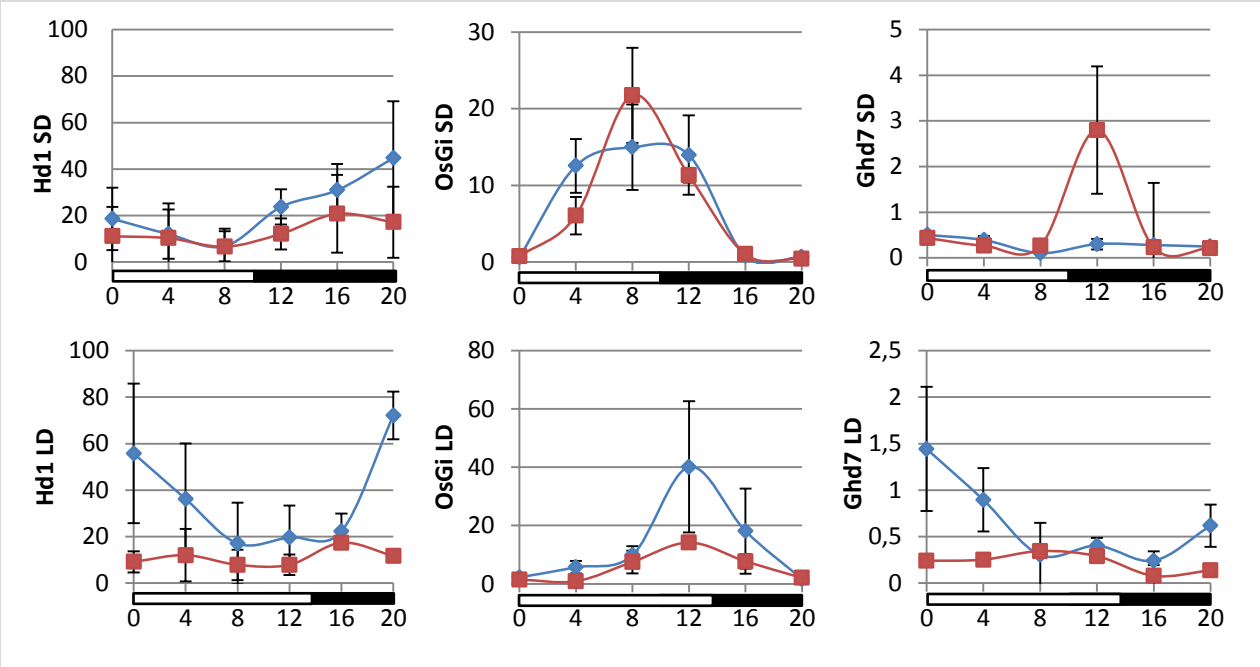
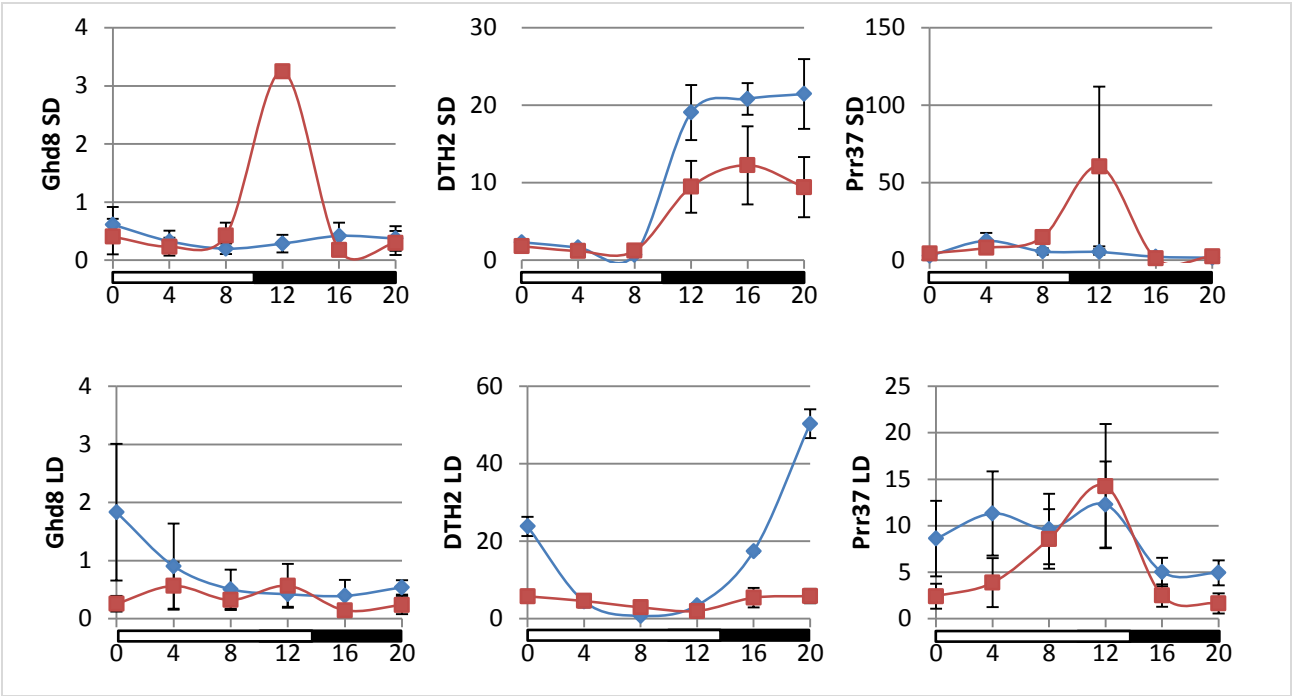
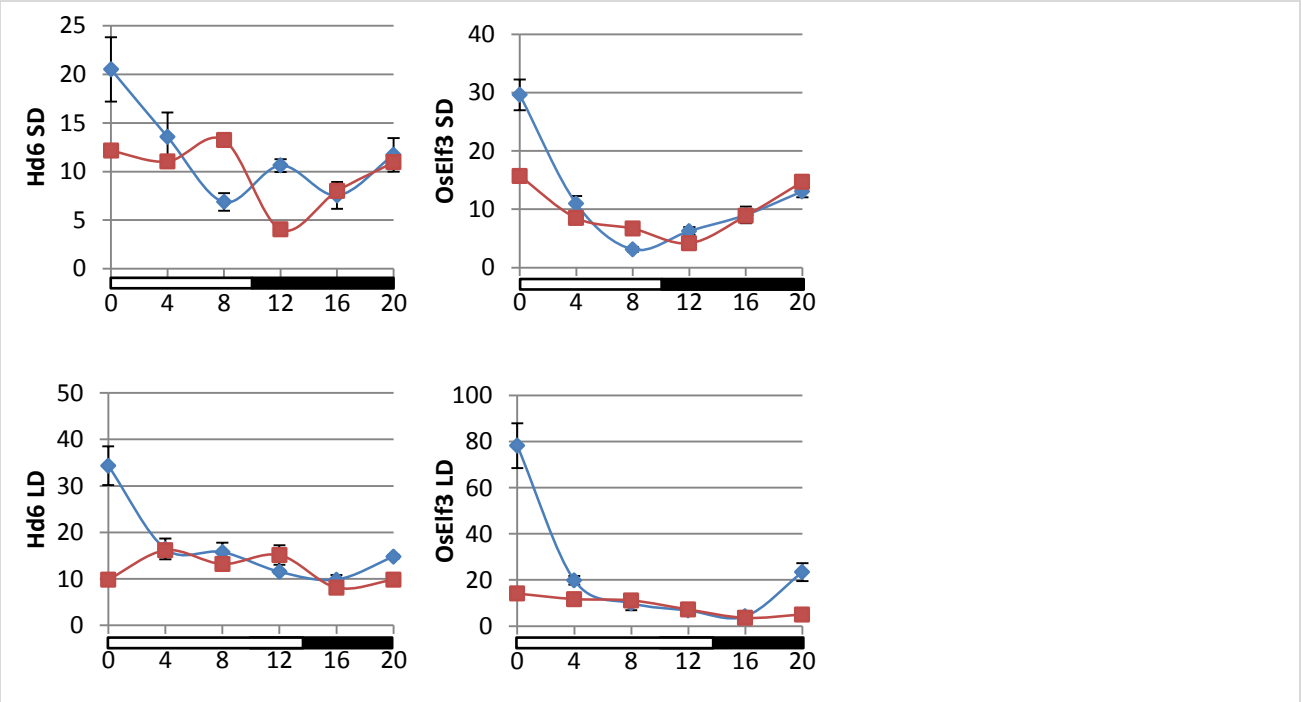


Figura 23: Serie temporal de RT-qPCR a las 0, 4, 8, 12, 16 y 20 horas desde el encendido de la luz de los principales genes implicados en la regulación mediante fotoperiodo en *arroz*, de las líneas Gleva (azul) y *G123* (rojo). En la base del eje horizontal la región blanca indica el periodo en el que las plantas reciben luz mientras que la negra cuando no.


Hd1 SD
OsGi SD
Ghd7 SD
Hd1 LD
OsGi LD
Ghd7 LD
0
4
8
12
16
20


Ghd8 SD
DTH2 SD
Prr37 SD
Ghd8 LD
DTH2 LD
Prr37 LD
0
4
8
12
16
20


Hd6 SD
25
20
15
10
5
0
0 4 8 12 16 20
OsElf3 SD
40
30
20
10
0
0 4 8 12 16 20
Hd6 LD
50
40
30
20
10
0
0 4 8 12 16 20
OsElf3 LD
100
80
60
40
20
0
0 4 8 12 16 20

RNA-seq

Con el objetivo de tener una visión global del transcriptoma de la línea mutante y observar el efecto causado por la mutación, realizamos un análisis de expresión génica diferencial entre *G123* y la variedad silvestre, Gleva, mediante un experimento de RNA-seq. Se compararon plantas de ambas, *G123* y Gleva, expuestas una semana a fotoperiodo largo tras cuatro semanas en fotoperiodo neutro y se tomaron tres muestras de plantas 20 horas después del inicio del día. Se extrajo el ARN de las muestras y se obtuvo su secuencia a través de a Novogen Bioinformatics Technology Co., Ltd. (Hong-Kong).

Las lecturas obtenidas mediante la secuenciación del ARN fueron filtradas según parámetros de calidad y se recortaron los primeros 15 pb del extremo 5', puesto que presentaban discrepancias en el porcentaje de bases. Las secuencias se mapearon frente al genoma de referencia Nipponbare (MSU V7) y se realizó un análisis numérico de la expresión diferencial de genes (*differential gene expression*) empleando las RPKM (*reads per kilobase million*) como método de normalización (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008). Se consideraron como genes con expresión significativamente diferencial aquellos con un FDR < 0,1. Siguiendo este criterio un total de 116 genes presentaron una expresión diferencial entre las muestras de Gleva y la línea mutante *G123*, de los cuales 62 estaban inducidos en *G123* respecto a Gleva y 54 estaban reprimidos (Fichero adicional 9, tabla S8).

La anotación funcional, así como la asignación a los 116 genes de las categorías funcionales en base a su Ontología Génica (GO) se realizó mediante la base de datos Comprehensive Annotation of Rice Multi-Omics (CARMO, http://bioinfo.sibs.ac.cn/carmo). Los 116 genes fueron anotados correctamente y clasificados respecto a las tres categorías GO establecidas: procesos biológicos,

componente celular y función molecular. Esta clasificación puede verse en la Fichero adicional 12, tablas S9, S10 y S11.

Con respecto a los 62 genes inducidos en *G123* con respecto a Gleva, dentro de la categoría de proceso biológico destacan los grupos correspondientes a transporte y a fotosíntesis que incluyen 14 y 11 genes respectivamente (Fichero adicional 13, tabla S12 y S13). Algunos de los genes incluidos en el grupo de transporte también aparecen en el grupo de genes implicados en la fotosíntesis. Otros grupos están relacionados con la respuesta a luz. De manera adicional, se puede observar la inducción de genes relacionados con la respuesta a luz como 8 genes que atañen a proteínas de unión a clorofila. De acuerdo a la clasificación por componente celular, los grupos más numerosos corresponden a la localización en plástidos y en membrana, destacando los cloroplastos como la ubicación mayoritaria. Según las categorías asignadas en cuanto a la función biológica, se puede apreciar grupos de genes con capacidad de unión a metales y función transportadora.

Con respecto a los 54 genes reprimidos en *G123* con respecto a Gleva, dentro de la categoría de proceso biológico destacan los grupos correspondientes a procesos metabólicos y respuesta a estímulos internos que incluyen 21 y 10 genes respectivamente.

Como conclusión de la clasificación de los genes diferencialmente expresados en *G123* respecto a Gleva según las categorías GO, se puede deducir que el mutante presenta alterada la fotosíntesis y procesos relacionados con la respuesta a la luz.

G123 es un mutante que presenta floración temprana, en este sentido cabría esperar la expresión diferencial de genes relacionados con la regulación de la floración. De hecho, se puede apreciar la inducción de *Hd3a* (LOC_Os06g06320), gen master inductor de la floración, que se encuentra 37,86 veces más expresado en *G123*. De igual forma, MADS14 (LOC_Os03g54160) y MADS18 (LOC_Os07g41370) presentan

niveles de expresión 20,5 y 3,8 veces mayores en *G123* (Fichero adicional 14, tabla S14).

Detección de la mutación

Con el fin de obtener información útil que facilite la identificación de posibles genes responsables de la variación en el fenotipo los datos obtenidos del RNA-seq se representaron en un gráfico de volcán (*volcano plot*). En estos gráficos se representa el nivel de significación de la expresión de cada gen, dado por el –log en base 10 del p-valor, en el eje de ordenadas y la variación de expresión de cada gen en el eje de abscisas. De esta manera, los genes con una diferencia significativa mayor (cuanto más pequeño es el p-valor) aparecen representados en la parte superior de la gráfica permitiendo la rápida identificación de aquellos genes que presentan una expresión diferencial significativamente mayor. En la figura 24 se muestra el gráfico de volcán realizado con los genes que presentaban un FDR < 0,1 en el análisis de expresión diferencial. Los genes que se encuentran más reprimidos respecto al control se muestran en la parte izquierda del gráfico de volcán (Figura 24). Tres de ellos, LOC_Os01g72170, LOC_Os01g72130 y LOC_Os01g72120, codifican proteínas con función glutation S-transferasa. Un cuarto gen, LOC_Os01g72100, codifica una proteína sensora de calcio relacionada con la calmodulina. El quinto gen más reprimido respecto al control es el LOC_Os01g72090 que codifica una fitocromobilin sintasa (PΦB) implicada en la biosíntesis de fitocromos y respuesta a fotoperiodo. Todos estos genes se encuentran muy próximos en el cromosoma 1, abarcando una región de 29.796 pb, entre la posición 41.825.087 y la 41.854.883, lo cual podría indicar una deleción en esa zona.

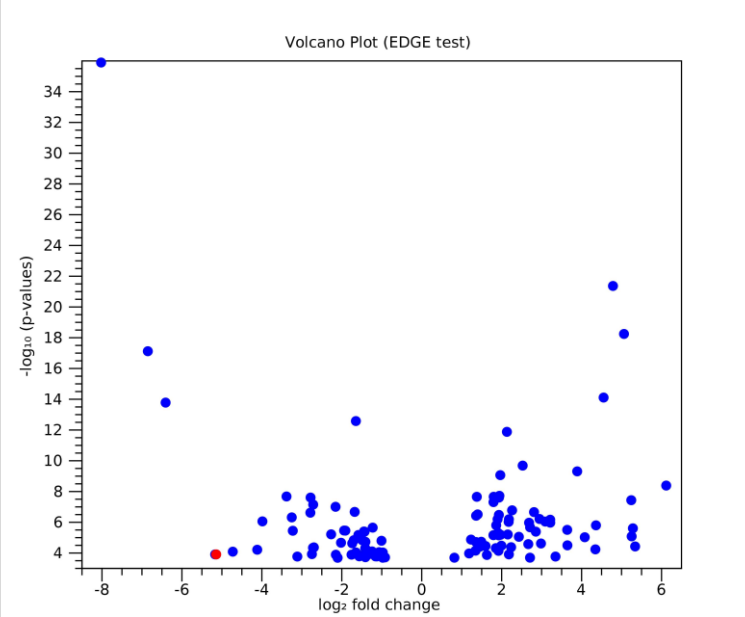


Figura 24: Gráfico de volcán de los genes que presentan un FDR < 0,1 en el análisis de expresión diferencial del RNA-seq de Gleva y *G123,* tras cuatro semanas bajo fotoperiodo neutro y una semana bajo fotoperiodo largo. Las muestras se tomaron a las 20h tras el encendido de la luz.

Mutmap

Con el objetivo de detectar la mutación responsable de las diferencias fenotípicas, se realizaron varias aproximaciones: un análisis tipo *Mutmap* (Abe, et al NatBiotech 2012), método desarrollado para la detección de mutaciones recesivas de tipo SNPs, y un análisis de variaciones estructurales utilizando el método Allinone, desarrollado en el departamento.

Una vez confirmado el carácter recesivo de la mutación, se procedió a la secuenciación del ADN nuclear de plantas del parental silvestre, Gleva, el parental mutante, *G123*, y el correspondiente a la mezcla de ADN, en igual proporción, de 20 plantas F2 con fenotipo temprano igual a *G123* (Epool). El ADN nuclear de las tres muestras fue secuenciado por Novogen Bioinformatics Technology Co., Ltd (Hong-Kong).

La secuenciación de los ADN produjo un total de 44,7 G de lecturas limpias. Las estadísticas del proceso de secuenciación y limpieza se encuentran resumidas en la tabla 10.

Tabla 10: Resumen del análisis estadístico de los datos de secuenciación

Muestra	Lecturas brutas	Datos brutos (G)	Datos limpios (G)	Eficiencia (%)	Error (%)	Q20 (%)	Q30 (%)	GC (%)
Gleva	50.039.992	15	15	99,74	0,01	96,6	92,317	43,36
Epool	53.165.697	15,9	15,9	99,7	0,01	96,84	92,63	43,11
G123	46.010.519	13,8	13,8	99,72	0,01	96,86	92,63	43,44

Lecturas brutas: Número de pares de lecturas, cuatro líneas son consideradas como una unidad de acuerdo al formato FASTQ.

Datos brutos (G): Volumen de datos original

Datos limpios (G): Volumen de datos calculado con las lecturas limpias.

Eficiencia (%): Relación de datos limpios respecto a datos brutos.

Error (%): Tasa de error total de las bases.

Q20 (%): Porcentaje de bases con valor Phred por encima de 20.

Q30(%): Porcentaje de bases con valor Phred por encima de 30.

GC(%): Porcentaje de G y C respecto al total de bases.

Previo al análisis de Mutmap, las lecturas limpias fueron mapeadas frente al genoma de referencia Nipponbare MSU v7 se detectaron las diferencias respecto a este. Las estadísticas del mapeo se encuentran resumidas en la tabla 11.

Tabla 11: Resumen del análisis estadístico de mapeo de las muestras Gleva, *G123* y Epool respecto al genoma de referencia Nipponbare MSU v7.

Muestra	Lecturas Mapeadas	Total de lecturas	Ratio de mapeo (%)	Profundidad media (x)	Covertura de al menos 1x (%)	Covertura de al menos 4x (%)
Gleva	98.379.106	99.823.180	98,55	35,96	96,48	95,34
Epool	104.502.027	106.008.260	98,58	37,63	96,42	95,27
G123	90.477.791	91.764.892	98,60	33,18	96,38	95,21

También se realizó un análisis de los diferentes tipos de variaciones presentes respecto al genoma de referencia, incluyendo polimorfismos de un solo nucleótido, InDel, pequeñas variaciones y Variaciones en el número de copias.

Estadísticas de detección y anotación de SNPs

Los SNPs, son variaciones de un único nucleótido que puede darse en posiciones específicas del genoma, incluyendo transiciones y transversiones de un solo nucleótido. Los SNPs entre las tres muestras de ADN secuenciado fueron detectados usando SAMtools (H. Li et al., 2009) con el comando "mpileup –m 2 –F 0.002 –d 1000.

Con el fin de reducir el ratio de error en la detección de SNPs, los resultados fueron filtrados siguiendo los siguientes criterios:

1. El número de lecturas necesario para secundar cada SNP debe ser superior a 4.
2. La calidad del mapeo (MQ) de cada SNP debe ser superior a 20.

Para la anotación de los SNPS se empleó el software ANNOVAR (K. Wang, Li, & Hakonarson, 2010) puesto que presenta múltiples posibilidades, incluyendo la anotación basada en genes, regiones, filtros y otras muchas funcionalidades. Las estadísticas respecto a los SNPs detectados y sus anotaciones se encuentran resumidas en la tabla 12.

Tabla 12: Resumen del análisis estadístico de la detección y anotación de los SNPs de las muestras Gleva, *G123* y Epool tomando como referencia el genoma de Nipponbare MSU v7.

		Exónicos													
Muestra	**5'**	**Stop gain**	**Stop loss**	**Sinónimos**	**No sinónimos**	**Intronicos**	**Splicing**	**3'**	**5'/3'**	**Intergénico**	**ts**	**tv**	**ts/tv**	**Het (‰)**	**Total**
Gleva	114.113	4.053	481	79.052	101.196	136.382	1.056	96.774	13.479	329.312	630.894	262854	2,400	0,897	893.748
Epool	115.469	4.095	490	79.992	101.905	137.200	1.035	98.038	13.980	332.257	637.003	265426	2,400	0,921	902.429
G123	116.064	4.109	490	79.567	101.433	137.141	1.051	97.967	13.918	333.462	637.842	265195	2,405	0,931	903.037

(1) Muestra: Nombre de la muestra
(2)5': Número de SNPs localizados a menos 1 kb en 5' (del punto de inicio de la transcripción) del gen.
(3) Exónicos: SNPs localizados en regiones exónicas: No sinónima: SNPs que cambian la secuencia aminoacídica; Stop gain/loss: SNPs no sinónimos que conllevan la introducción/eliminación de un codón de para; Sinónimos: Mutaciones de un solo nucleótido que no cambian la secuencia aminoacídica.
(4) Intrónicos: SNPs localizados en regiones intrónicas;
(5) Splicing: SNPs localizados en puntos de splicing alternativo (dentro de un rango de 2 pb del límite intrón/exón).
(6) 3': SNPs localizados a menos de 1 kb en 3' (respecto al punto de fin de la transcripción) de la región génica.
(7)5'/3': SNPs localizados a menos de 2 kbs de una región intergénica, esto es en 1 kb en 5' o 3' de los genes.
(8) Intergénico: SNPs localizados dentro de región a mas de 2kb SNPs.
(9) ts: Transiciones, mutaciones puntuales que cambian un nucleótido de purina por otra purina (A ↔ G) o una pirimidina por otra pirimidina (C ↔ T). Aproximadamente dos de cada tres SNPs son transiciones.
(10) tv: Transversiones, la sustitución de purinas (de dos anillos) por pirimidinas (de un anillo) o *vice versa*.
(11) ts/tv: Ratio de transiciones transversiones.
(12) Het: Ratio de heterozigosidad a lo largo del genoma, se calcula como el ratio de SNPs heterocigotos respecto al total de bases del genoma.
(13) Total: Número total de SNPs.

Estadísticas de detección y anotación de Indel

InDel hace referencia a la inserción o deleción de secuencias ≤ 50 pb, para la detección de InDels se utilizó el programa SAMTOOLs con el parámetro "mpileup –m 2 –F 0.002 –d 1000" y para la anotación se usó ANNOVaR. Las condiciones para filtrar los InDels detectados fueron las mismas que las empleadas para los SNPs. Las estadísticas del mapeo y anotación de los InDel están resumidas en la tabla 13.

Tabla 13: Resumen del análisis estadístico de la detección y anotación de los InDel de las muestras Gleva, *G123* y Epool tomando como referencia el genoma de Nipponbare MSU v7.

		Exonicos														
Muestra	**5'**	**Stop gain**	**Stop loss**	**Frameshift deletion**	**Frameshift insertion**	**Non-frameshift deletion**	**Non-frameshift insertion**	**Intronicos**	**Splicing**	**3'**	**5'/3'**	**Intergénico**	**Inserción**	**Deleción**	**Het rate (‰)**	**Total**
Gleva	16146	84	9	1561	1188	1412	997	19628	96	13564	2064	44578	44477	48714	0,032	94579
Epool	16108	87	13	1578	1195	1394	982	19610	91	13637	2106	44533	44569	48802	0,032	94708
G123	15929	85	9	1576	1191	1479	992	19410	96	13370	2098	44109	44272	48686	0,033	94247

(1) Muestra: nombre de las muesrtas Sample names.
(2) 5': Número de InDel localizados a menos 1 kb en 5' (del punto de inicio de la transcripción) del gen.
(3) Exónicos: InDels localizados en regiones exónicas; Stop gain/loss: InDel que conllevan la introducción eliminación de codones de parada en el lugar de la variación; Frameshift deletion/insertion: InDel que cambian el marco de lectura con una deleción o inserción; Non-Frameshift delecion/insercion: InDel que no cambian el marco de lectura con la inserción o deleción de secuencias de un número de base 3 o múltiplo de 3.
(4) Intronicos: InDel localizados en una región intrónica.
(5) Splicing: InDel localizados en puntos de splicing alternativo (dentro de un rango de 2 pb del límite intron/exón).
(6) 3': InDel localizados a menos de 1 kb en 3' (respecto al punto de fin de la transcripción) de la región génica.
(7) 5'/3': SNPs Localizados a menos de 2 kbs de una región intergénica, esto es en 1 kb en 5' o 3' de los genes.
(8) Intergénico: InDel localizados dentro de regiones a más de 2kb.
(9) Het: Ratio de heterocigosidad a lo largo del genoma, se calcula como el ratio de InDel heterocigotos respecto al total de bases del genoma.
(10) Total: Número total de InDel.

Detección y anotación de SV

Las SV (variaciones estructurales) son variaciones genómicas con mutaciones de un tamaño relativamente grande (>50pb) incluyendo deleciones, duplicaciones, inserciones, inversiones y translocaciones. Para detectar las inserciones (INS) deleciones (DEL) inversiones (INV), las translocaciones intra-cromosómicas y las traslocaciones inter-cromosómicas (CTX) se empleó el programa Breakdancer (K. Chen et al., 2013), este software se basa en los resultados del mapeo respecto al genoma de referencia y el tamaño de los insertos detectados. Los SV detectados se filtran mediante la eliminación de aquellos con menos de 2 lecturas emparejadas, además las INS, DEL y INV fueron anotadas mediante el programa ANNOVAR. Las estadísticas delos SV detectados se encuentran resumidas en la tabla 14.

Tabla 14: Resumen del análisis estadístico de la detección y anotación de las SV de las muestras Gleva, *G123* y Epool tomando como referencia el genoma de Nipponbare MSU v7.

Muestra	5'	Exónicas	3'	Intrónicas	5'/3'	Intergénicas	Splicing	INS	DEL	INV	ITX	CTX	Total
Gleva	335	1987	274	207	65	880	1	27	3213	585	1425	4789	10039
Epool	329	1896	253	214	65	887	5	49	3087	574	1601	4636	9947
G123	298	1745	239	202	63	803	0	56	2845	504	1374	4337	9116

(1) Muestra: nombre de las muestras
(2) 5': Número de SV localizadas a menos de 1 kb en 5' (del punto de transcripción) de un gen.
(3) Exónicas: Número de SV localizados a menos 1 kb en 5' (del punto de inicio de la transcripción) del gen.
(4) Intrónicas: SVs InDel localizadas en una región intrónica.
(5) 3': localizados a menos de 1 kb en 3' (respecto al punto de fin de la transcripción) de la región génica.
(6) 5'/3': SVs localizados a menos de 2 kbs de una región intergénica, esto es en 1 kb en 5' o 3' de los genes.
(7) Intergénicas: SVs localizados dentro de regiones a más de 2kb.
(8) Splicing: SVs localizados en puntos de splicing alternativo (dentro de un rango de 2 pb del límite intrón/exón).
(9) INS: Inserción.
(10) DEL: Deleción.

(11) INV: Inversión.
(12) ITX: Translocaciones intracromosómicas.
(13) CTX: Translocaciones intercromosómicas.
(14) Total: Número total de SVs.

Detección y anotación de CNV

CNV son variaciones de número de copias (Copy-number variation) se trata de un tipo de variación estructural que muestra deleciones o duplicaciones en el genoma. Se detectan basándose en la profundidad de lecturas del genoma de referencia. Para detectar las CNV de posibles deleciones y duplicaciones se empleó el programa CNVnator (Abyzov, Urban, Snyder, & Gerstein, 2011) empleando el comando "-call 100". Las CNV detectadas fueron anotadas empleando el programa ANNOVAR. Las estadísticas de la detección y anotación de las CNV se encuentran resumidas en la tabla 15.

Tabla 15: Resumen del análisis estadístico de la detección y anotación de las CNV de las muestras Gleva, *G123* y Epool tomando como referencia el genoma de Nipponbare MSU v7.

Muestra	5'	Exónicas	Intrónicas	3'	5'/3'	Intergénicas	Duplicaciones	Deleciones	Longitud de las duplicaciones (bp)	Longitud de las deleciones (bp)	Total
GLeva	662	3628	300	608	146	1848	1056	6231	5729500	25139500	7287
Epool	685	3749	331	651	176	1939	1129	6504	6130600	24964300	7633
G123	528	3413	258	544	112	1589	1037	5500	5776400	24241700	6537

(1) Muestra: nombre de las muestras
(2) 5': Número de CNV localizadas a menos de 1 kb en 5' (del punto de transcripción) de un gen.
(3) Exonicas: CNVs localizadas en una región exónica
(4) Intronicas: NV localizadszas en una región intrónica.
(5)3': CNVs localizados a menos de 1 kb en 3' (respecto al punto de fin de la transcripción) de la región génica.
(6) 5'/3': CNVs localizados a menos de 2 kbs de una región intergénica, esto es en 1 kb en 5' o 3' de los genes.
(7) Intergenicas: CNVs localizados dentro de regiones a más de 2kb.

(8) Longitud de las duplicaciones: CNVs con un aumento del número de copias.
(9) Deleciones: CNVs con una disminución del número de copias.
(10) Longitud de las duplicaciones (bp): la longitud total del número de duplicaciones por CNV
(11) Longitud de las deleciones (bp): la longitud total de las deleciones de CNV.
(12) Total: número total de CNVs.

Presentación visual de las variaciones entre los genomas *G123*, Gleva y Epool

Las variaciones estructurales del genoma completo pueden ser visualizadas según los diferentes tipos de mutaciones mediante su representación con un programa circos en los que:

1. Para los SNP/InDel, se dibujó la distribución de densidad.
2. Para pequeñas variaciones/variación en el número de copias, se muestra la localización y el tamaño.

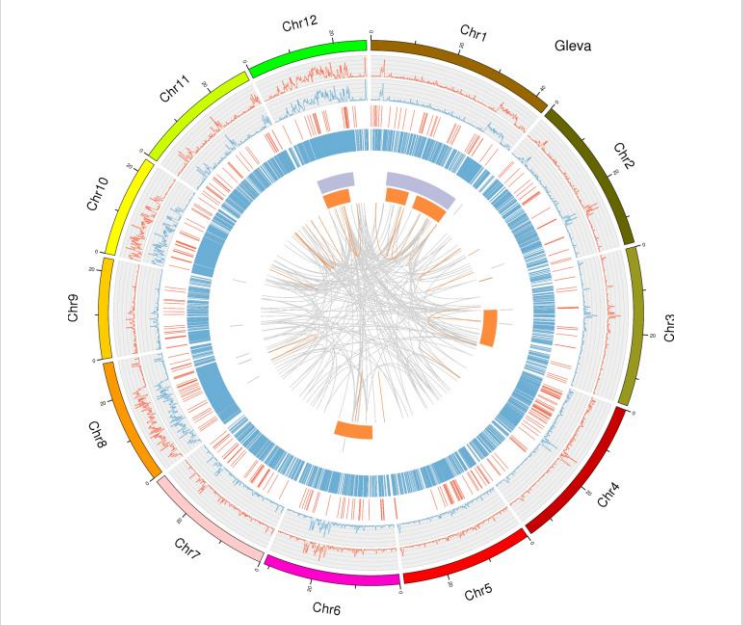
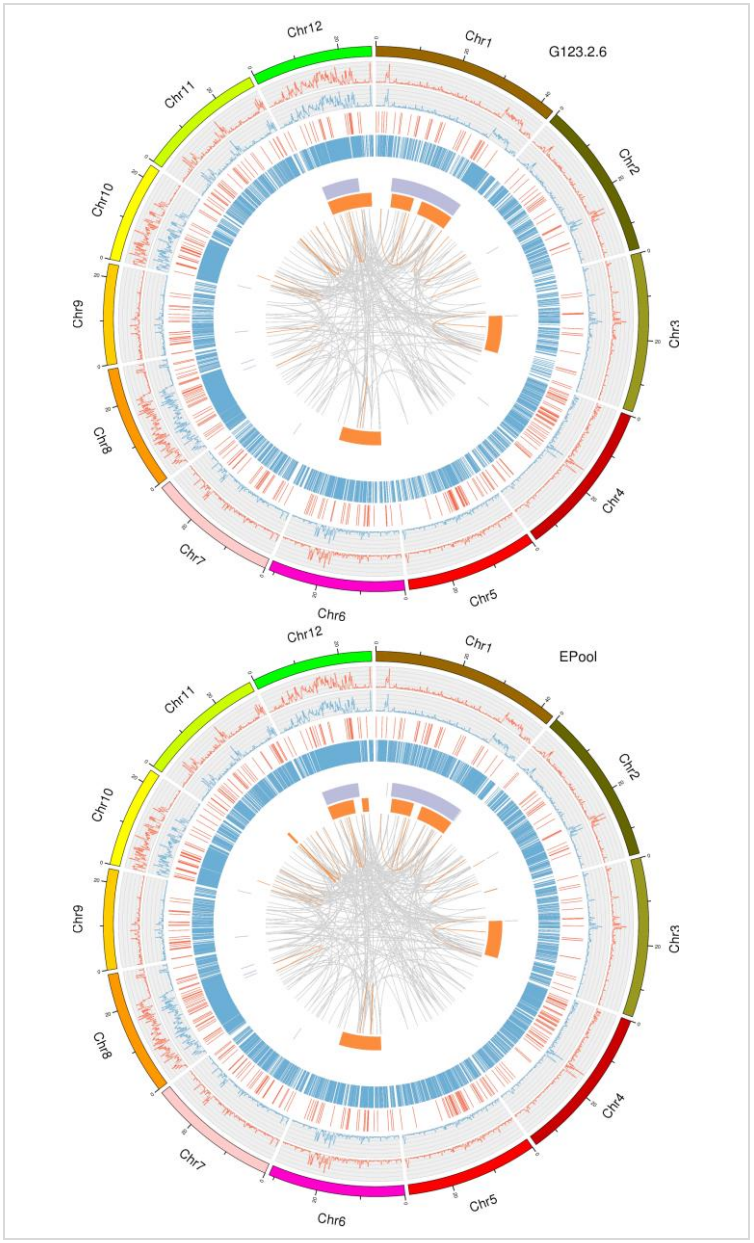


Figura 25: Representación circos del genoma de Gleva, *G123* y Epool. De fuera a dentro se muestran el cromosoma, SNP, InDel, duplicaciones, deleciones, inserciones, deleciones, inversiones, traslocaciones intracromosómicas y traslocaciones intercromosómicas.



Con el fin de identificar la mutación responsable del fenotipo de floración temprana el mutante *G123*, se procedió al análisis bioinformático *Mutmap*. Se utilizó la versión

1.4.4. del programa *Mutmap*, obtenida en http://genome-e.ibrc.or.jp/, y se usó sobre las secuencias filtradas de los ADN de Gleva y del Epool. El programa calcula el índice SNP, que indica el número de veces que un SNP está presente en la agrupación de plantas con fenotipo con respecto al parental silvestre. Cuanto más próximo a 1 es el índice de SNP, mayor es la probabilidad de que el SNP sea el responsable del fenotipo. La ruta de trabajo de este programa consta de tres fases, en la primera las secuencias son filtradas de modo que solo se emplean aquellas de mayor calidad. En la segunda se genera un "genoma de referencia" a partir de los datos del parental. Ya en la última fase se analizan las frecuencias de lo SNP-index calculados y se comparan con los valores esperados en el caso de que hubiese una asociación entre fenotipo y genotipo, en la que cada alelo del SNPs se encontraría entorno al 50% en las lecturas de la mezcla de ADN de las muestras de la F2 con fenotipo mutante. Para la identificación se comparan los valores de SNPs-index observados frente a los esperados en gráficas a lo largo de los cromosomas, buscándose la identificación de picos en los que el valor observado sea superior al intervalo de confianza para una segregación aleatoria.

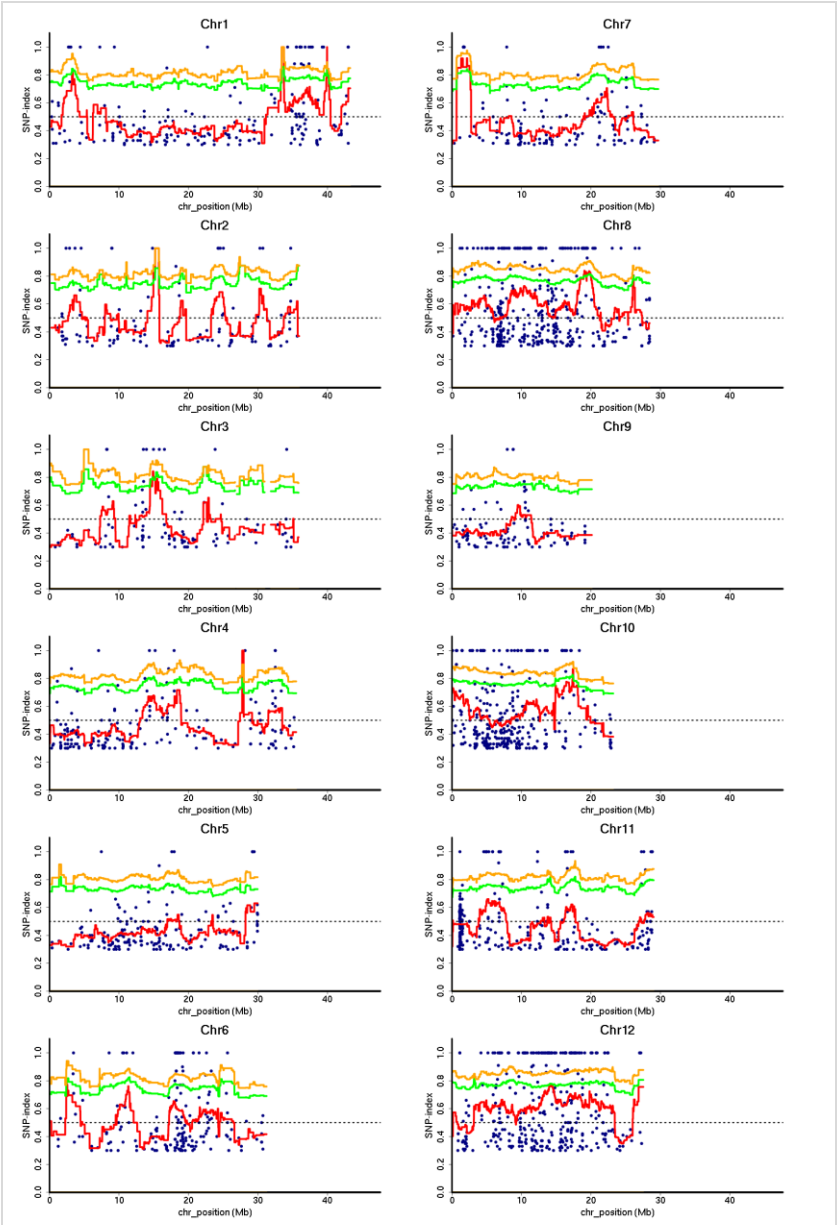


Figura 26: Gráfica de Mutmap para cov3, co7, con ventana deslizante de 2M. La gráfica indica tres regiones candidatas, dos en el cromosoma uno en posición próxima a 30 Mb y a 40 Mb, y una en el cromosoma cuatro próxima 30 Mb.

Las gráficas del Mutmap permiten una visualización rápida de regiones candidatas a contener SNPs responsables del fenotipo mutante. En nuestro estudio tres regiones aparecieron como candidatas, dos en el cromosoma uno y una en el cromosoma cuatro. Estas regiones presentaban tres SNPs con un SNP-index = 1, Chr01_39252250 (G→A), Chr01_39443284 (C→ T), Chr04_28150293 (G→A). Se procedió a detectar genes funcionales en la cercanía de estos SNPs candidatos usando la base de datos Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/). El SNP candidato localizado en el cromosoma 4 fue descartado por estar muy alejado de cualquier gen anotado. Por lo respecta a los dos SNPs candidatos localizados en el cromosoma 1, el SNP Chr1_39443284 se encuentra a 411 pb en 5' de gen LOC_Os01g67870, el cual codifica un péptido señal, con actividad transferasa y el SNP Chr1_39252250 está situado 733 pb en 5' del LOC_Os01g67530. LOC_Os01g67530 codifica una proteína similar a Acetil-Coa sintetasa.

Ambos genes no presentan expresión diferencial en los datos procedentes del análisis de RNA-seq, no obstante, se comprobaron los perfiles de expresión a lo largo del día bajo condiciones de día largo y corto mediante RT-qPCR (figura 27). Los perfiles de expresión mostraron que el LOC_Os01g67530 se expresa mucho más en Gleva que en *G123*, situándose el máximo de diferencia en niveles de expresión a las 0 horas desde el encendido de la luz. Puesto que su expresión se activa entre las 8 y las 12 horas independientemente de fotoperiodo. Ambos genes parecen tener un nivel de expresión basal en Gleva y *G123* en día corto, pero bajo condiciones de día largo la expresión de ambos genes en Gleva se ve inducida teniendo su máximo a las 0 horas de luz, mientras que en los mutantes parecen continuar en su nivel basal de día corto habiendo pocas discrepancias entre la línea mutante y la parental silvestre.

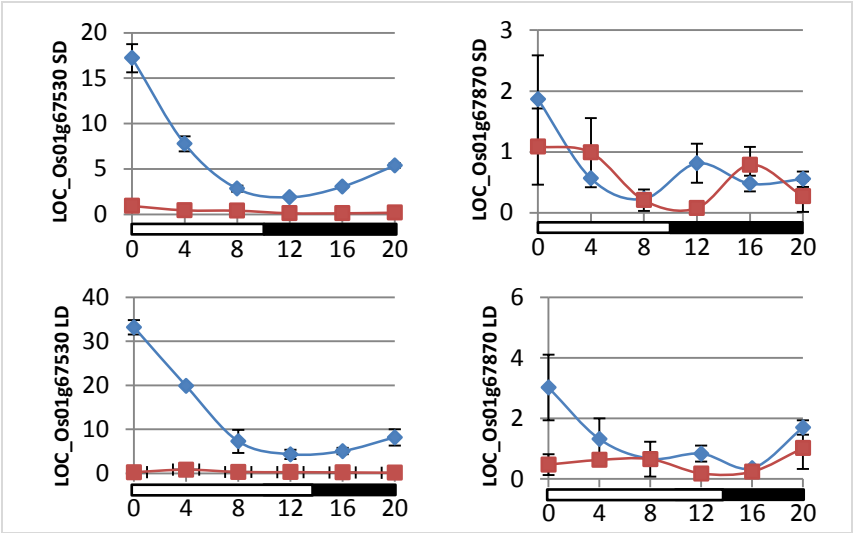


Figura 27: Perfiles de expresión a lo largo del día tras cuatro semanas en condiciones de fotoperiodo neutro seguida de una semana en fotoperiodo corto o largo de los genes próximos a los SNPs candidatos del análisis Mutmap, en azul Gleva, en rojo *G123*.

Finalmente para verificar lo SNPs candidatos se diseñaron cebadores para amplificar las zonas donde estos se encontraban y secuenciar los productos de PCR mediante tecnología Sanger. Inesperadamente los SNPs no se encontraban presentes en el mutante ni en la planta silvestre. Por tanto se asumió que los SNPs candidatos fueron detectados por un mal alineamiento de las lecturas o porque las regiones donde se encontraban presentaban baja cobertura. Así pues los descartamos. Nuestra siguiente hipótesis fue que quizás la mutación responsable del fenotipo pudiese ser otro tipo de mutación.

Análisis de las variaciones estructurales.

Puesto que los SNPs fueron descartados como responsables de la mutación que generaba el fenotipo, se decidió usar un programa desarrollado en el propio departamento (Carles et al, manuscrito en preparación). Este programa “Allinone”

combina varios programas con la finalidad de detectar y filtrar las variaciones estructurales entre dos genomas. De este modo en primer lugar se detectan todas las variaciones estructurales presentes en Gleva, *G123* y el Epool, respecto al genoma de referencia Nipponbare y en un segundo filtro se seleccionan solo aquellas que están presentes en *G123* y en el Epool pero no en Gleva. Las variaciones estructurales que pasaron ambos filtros se encuentran resumidas en la tabla 16.

Tabla 16: Variaciones estructurales apoyadas por el 95% de las lecturas de *G123* y Epool y ausentes en Gleva. "ID" es un valor identificativo de la mutación, en las tranlocaciones el mismo evento aparece con la misma ID pero en las distintas localizaciones se distinguen por el subíndice.

Cromosoma	Posición	ID	Tipo
Chr1	41822688	829	Deleción
Chr2	23819730	1184	Deleción
Chr8	2878658	3769_1	Traslocación
Chr8	5037462	3769_2	Traslocación
Chr8	8430256	4110	Inversión
Chr10	15910040	6541	Deleción
Chr10	16389174	6581	Deleción
Chr12	10873676	7899	Deleción
Chr5	29186488	14374_1	Traslocación
Chr9	21044854	14374_2	Traslocación
Chr10	3604760	17224_1	Traslocación
Chr11	13243899	17224_2	Traslocación
Chr10	15846229	17372_1	Traslocación
Chr12	10056201	17372_2	Traslocación

Un total de 14 variaciones estructurales permanecieron tras aplicar el filtro, las cuales consistían en 5 deleciones, una inversión y 4 eventos de translocación.

A continuación se procedió a la verificación manual de las 14 variaciones con el fin de descartar falsos positivos. Para el filtrado manual se usó el programa IGV mediante el cual únicamente se pudo verificar una variación: una deleción de 33.373 pb situada en

el cromosoma 1 entre la posición 41.822.688 y la 41.856.061. En esta región se encuentran los genes LOC_Os01g72100, LOC_Os01g72120, LOC_Os01g72130, LOC_Os01g72140, LOC_Os01g72150, LOC_Os01g72160, LOC_Os01g72170. Algunos de estos coinciden con los que presentaban mayor diferencia de expresión entre *G123* y Gleva en el análisis del transcriptoma.

Los datos respecto a estos genes se encuentran resumidos en la tabla 17.

Tabla 17: Genes presentes en la región de la deleción localizada en Chr1: 414822688 a 41856061.

MSU	Descripción	Longitud (pb)	CGSNL Gene Name	CGSNL Gene Symbol
LOC_Os01g72090	Similar to Phytochromobilin synthase precursor. (Os01t0949400-01); Similar to elongated mesocotyl1. (Os01t0949400-02)	2819	PHOTOSENSITIVITY 13	SE13
LOC_Os01g72100	Similar to Calmodulin (CaM). (Os01t0949500-01)	1046	CALMODULIN-LIKE PROTEIN 10	CML10
LOC_Os01g72120	Glutathione S-transferase, C-terminal domain containing protein. (Os01t0949700-01)	1489	TAU GLUTATHIONE S-TRANSFERASE 7	GSTU7
LOC_Os01g72130	Similar to Glutathione S-transferase GST 28 (Fragment). (Os01t0949750-00)	833	TAU GLUTATHIONE S-TRANSFERASE 35	GSTU35
LOC_Os01g72140	Similar to Glutathione S-transferase GST 28 (EC 2.5.1.18) (Fragment).	1294	TAU GLUTATHIONE S-TRANSFERASE 36	GSTU36
LOC_Os01g72150	Conserved hypothetical protein. (Os01t0949900-01)	917	TAU GLUTATHIONE S-TRANSFERASE 37	GSTU37
LOC_Os01g72160	Similar to Glutathione S-transferase GST 28 (EC 2.5.1.18) (Fragment). (Os01t0950000-01)	1373	TAU GLUTATHIONE S-TRANSFERASE 41	GSTU41
LOC_Os01g72170.1	Conserved hypothetical protein. (Os01t0950300-01)	861	TAU GLUTATHIONE S-TRANSFERASE 42	GSTU42

Entre estos genes destaca gen PHOTOSENSITIVITY 13 por estar relacionado con la respuesta de las plantas a la luz, de manera que parece ser el candidato más plausible para ser el responsable de la variación de fenotipo de la línea mutante *G123*.

3.3.3. Discusión

La línea de arroz mutante *G123* ha sido identificada en un rastreo de mutantes de irradiación en busca de plantas que florezcan de manera temprana. El acortamiento del ciclo vegetativo de una variedad de arroz es de interés agronómico ya que permite ahorrar agua de riego, favorece la seguridad de la cosecha al reducir el periodo de exposición a riesgos como las tormentas de final de verano o el ataque de piricularia y, además, el disponer de una gama de variedades con diferentes longitudes de ciclo permite escalonar la cosecha y, así, aprovechar de manera más eficiente la maquinaria e infraestructura relacionada con la siega.

Un análisis más detenido de la línea *G123* llevó a averiguar que, además de mostrar una floración más temprana, también presentaba insensibilidad a fotoperiodo. Este hecho es relevante en el análisis del mutante, ya que revela que la mutación producida por la irradiación está alterando la regulación de la floración a través de la mediación del fotoperiodo. La señal de floración en arroz viene dada por dos genes clave, *Hd3a*, que induce la floración en condiciones de día corto, y *RFT1*, que juega el papel inductor en condiciones de día largo (Komiya et al, 2009). Tanto *Hd3a* y *RFT1*, están modulados por dos rutas principales de regulación de la floración independientes, una está regida por el ciclo circadiano a través del gen *Hd1* y la segunda ruta por *Ehd1*, que es un integrador de señales y que está modulado por el fotoperiodo. En latitudes norte, dónde el clima es templado, es frecuente encontrar variedades con variantes de Hd1 no funcionales, ya que este gen es un inhibidor de la floración en condiciones de días largos y, por lo tanto, la floración está gobernada por *Ehd1*. No obstante, la variedad Gleva, parental de *G123* presenta un alelo funcional, por lo que ambas rutas son funcionales (Naranjo et al, 2014).

En estudios previos del laboratorio dónde se ha desarrollado esta tesis, se aisló el mutante *S73*, en esa ocasión por irradiación de la variedad Bahía, que también resultó ser insensible a fotoperiodo. La identificación del gen mutado, *se5*, llevó a sugerir que

los fitocromos inhiben la floración modulando negativamente la expresión de *Ehd1* como la acción inhibidora de Hd1 sobre *Hd3a* (Andrés et al, 2009). En el mutante *G123*, se observa que los niveles del gen *Hd3a* son más elevados que en el parental Gleva, indicando que en este caso es *Hd3a* es el inductor de la floración en *G123* en lugar de *RFT1*.

El análisis de las variaciones estructurales permitió postular al gen *Se13*/*OsHY2* como candidato más plausible a ser el responsable de la variación fenotípica de *G123*. *Se13* codifica una fitocromobilina (PΘB) sintetasa, enzima que participa en el último paso de la ruta de síntesis de fitocromobilina, cromóforo que forma parte de los fitocromos (Saito et al 2011).

El gen *Se13*/*OsHY2* fue descrito por primera vez por Saito (Saito et al., 2011) en una línea mutante de la variedad Gimbozu. Esta línea presentaba una deleción de un solo nucleótido en el primer exón que originaba un desplazamiento en el marco de lectura dando lugar a un codón de parada prematuro. Como principal consecuencia la variedad mutante, X61, presentaba bajo condiciones de luz natural en día largo un adelanto de floración respecto a la línea no mutada de 35 días, prácticamente los mismos días que presenta de adelanto en tiempo de floración el mutante *G123* frente a Gleva, bajo fotoperiodo largo, hecho que concuerda con unas horas de luz similares en ambas localizaciones durante el periodo de cultivo (Malta, 39°17'N; Kioto, 35°01'N). *Se13*/*OsHY2*, al participar en la síntesis fitocromos, actuaría como inhibidor de *Hd3a* bajo condiciones de día largo a través de Hd1 y, además, reprimiría la expresión de *Ehd1*, mediante *Ghd7* que sería incapaz de activar la síntesis de *RFT1* (Yoshitake et al., 2015). De esta forma, la falta de funcionalidad de *Se13*/*OsHY2* produce plantas defectuosas en el contenido de fitocromos y, por lo tanto, no se produce la inhibición de la floración por estos compuestos, floreciendo antes.

Al comparar los perfiles de expresión de los genes principales de la regulación de la floración en X61 y en *G123* en condiciones de día largo, observamos que *Hd3a* presentan niveles superiores en ambas líneas mutantes respecto a sus correspondientes líneas no mutadas, como era de esperar. De hecho en Gleva la expresión de *Hd3a* es prácticamente nula mientras que *G123* presenta unos niveles significativamente más elevados entre las 0 y las 8 horas. Sin embargo, los niveles de *RFT1* son más altos en Gleva que en *G123*, mientras que en X61 son mayores que en Gimbozu (Yoshitake et al., 2015). Esta discrepancia es llamativa, no obstante no se observa ninguna expresión de *Hd3a* ni *RFT1* por lo que posiblemente la inducción de la floración en Gimbozou no se haya producido en el momento del análisis de la expresión. De igual manera, la mutación en el gen *se5* (Andrés et al 2009; Izawa et al., 2000), que se encuentra en el paso previo en la ruta de síntesis de fitocromobilina, produciendo la hemoxigenasa, el sustrato de OsHY2, debería presentar alteraciones similares a *se13*. La expresión de *Hd3a* en el mutante *s73* presenta niveles muy superiores a los de la variedad parental no mutada, al igual que en *G123* y X61. De manera adicional, la expresión de *Ehd1*, gen inductor de *Hd3a*, presenta también su máximo, muy superior a la línea no mutada, a las 4 horas de comenzar a recibir luz en condiciones de día largo, hecho que concuerda con la falta de inhibición de *Ehd1* por los fitocromos. De esta manera, se refuerza la idea de que es *Hd3a* el gen máster que está induciendo la floración en el mutante *G123*.

Respecto a otros genes que participan en la ruta de regulación de la floración, *DTH2*, gen inductor de la floración bajo condiciones de fotoperiodo largo, se ha visto que sus niveles de expresión son más elevados en la variedad Gleva que en *G123*. *OsElf3/Hd17* es un represor de *Ghd7*, por lo tanto plantas defectuosas en este gen presentan mayores niveles de expresión de *Ehd1*, *RFT1* y *Hd3a* bajo condiciones de fotoperiodo largo principalmente (Yang et al., 2013). Mientras que en plantas silvestres la expresión de *Ghd7* se activa cuando se dan pulsos de luz, en los mutantes

elf7 el patrón de expresión de *Ghd7* no cambia, pero presenta niveles más elevados que la planta silvestre (Saito et al., 2012). En *Arabidopsis*, se ha podido demostrar que ELF3 interactúa directamente con PHYB y el complejo ELF3-PHYB tiene capacidad de regular la expresión génica de varios genes de la ruta de floración. Aunque la interacción ELF3-PHYB no se ha demostrado en arroz y *Ghd7* es específico de arroz, puesto que *OsElf3/Hd17* es un represor de *Ghd7*, cabría esperar que este modelo fuese válido en arroz. Gleva presenta niveles muy superiores de OsELF3 respecto a *G123* y menores de *Ghd7* por lo que se reforzaría la hipótesis de que la señalización por fitocromos está afectada en *G123*. *Ghd8*/*DTH8* (*DAYS TO HEADING 8*) codifica una proteína activadora de HAP3 (HEME ACTIVATOR PROTEIN 3) la cual es una subunidad del complejo factor de transcripción CCAAT-box-binding. Promueve la floración bajo condiciones de día corto pero la reprime durante condiciones de día largo actuando sobre *Ehd1* y *Hd3a*, su actividad no se ve afectada por Hd1 ni *Ghd7*. El mutante *G123* presenta un perfil de expresión de *Ghd8* prácticamente idéntico al de *Ghd7* tanto en día corto como en día largo presentando niveles más bajos que en Gleva. Dado que ambos genes presentan la activación de la expresión al recibir luz (Xue et al., 2008) se refuerza la idea de que la vía de regulación mediante fitocromos está alterada en *G123*. *Hd6* codifica una subunidad α de la protein kinasa CK2 (CK2α) (Takahashi et al., 2001; Wei et al., 2010)**.** *Hd6* actúa junto a Hd1 para reprimir la floración en día largo, su actividad estimula la represión pero no es necesaria para que esta se dé, es más *Hd6* necesita la presencia de Hd1 funcional, pese a no haberse detectado su acción directa sobre este ni sobre *Ghd7* (Ogiso et al., 2010)**.** En nuestro análisis se ha observado que Gleva presenta mayores niveles de expresión de *Hd6* en el momento máximo de *Hd1*, que se da a las 0 horas de luz tanto en día corto como en día largo, lo cual sería coherente con una floración más atrasada en Gleva que en el mutante.

Los resultados del análisis del trascriptoma de *G123*, obtenidos mediante RNA-seq, son coherentes con la mutación detectada en el gen *se13*. En ellos se observa la

inducción de genes relacionados con de fotosíntesis y complejo colector de luz. Estos resultados cobran sentido al tener en cuenta la ruta de síntesis de los fitocromos, o ruta de síntesis de los tetrapirroles. Esta ruta es común desde la introducción del glutamato hasta la síntesis de Protoporfirina IX, sin embargo, al llegar a este punto, este sustrato puede ser convertido por una Mg-quelatasa en Mg-protoporfirina IX y derivarse a la síntesis de clorofilas o mediante una ferroquelatasa puede ser convertido a hemo derivándose a la síntesis de las fitocromobilinas. Hay que tener en cuenta que los tetrapirroles libres pueden producir efectos fotoxidativos que darían lugar a daños en la planta, por lo que a fin de evitarlo se ha de regular la síntesis del ácido 5-aminolevulínico (5-ALA). Se sabe que el grupo hemo es un importante inhibidor en feedback de la síntesis de 5-ALA (Beale & Weinstein, 1991; Cornah, Terry, & Smith, 2003; Matthew J. Terry & Kendrick, 1999). Las lesiones en la ruta de degradación del grupo hemo (síntesis de PΦB) causarían la acumulación de este reduciendo la síntesis de ALA, y como consecuencia la reducción de la síntesis de clorofila. Por simple observación, los mutantes s73 presentan un fenotipo más débil que los *G123*, por lo que podría pensarse que son debidas a distintos fondos genéticos. Sin embargo, esta misma observación se ha realizado respecto a mutantes de Arabidopsis defectuosos en estas enzimas (M J Terry, 1997). El mutante *hy2*, que presenta el gen de la fitocromobilin sintasa alterado (Kohchi, 2001), presentaba un fenotipo más suave que el *hy1* (Muramoto, 1999), portador de una mutación en el gen que codifica la hemoxigenasa. Las diferencias observadas en ambos mutantes se darían por la acumulación de hemo. Ambos mutantes son incapaces de producir la fitocromobilina, pero en el caso del mutante *s73* además al no poder convertir el hemo en biliverdina IX inhibiría la ruta de síntesis de tetrapirroles, inhibiéndose también la síntesis de clorofilas. Mientras que *G123* está retroinhibición no se daría. De hecho, el mutante *s73* muestra una coloración verde pálido, mientras que el mutante *G123* es verde oscuro. Otra fuente para las discrepancias entre ambos

mutantes sería el hecho de que ambos son considerados de pérdida de función parcial (M. Takano et al., 2009) puesto que ambas enzimas pertenecen a una pequeña familia génica (Makoto Takano et al., 2005). Si tenemos en cuenta este efecto de fuga debido a la actividad de los otros miembros de las familias génicas de la hemoxigenasa y la fitocromobilin sintasa. Ambos mutantes presentarían problemas en las primeras fases de desarrollo debido a que en estos estadíos los fitocromos de tipo I, es decir el fitocromo A, tienen un papel más importante, puesto que son los principales reguladores de la germinación y el crecimiento fotomorfogénico. Estos fitocromos además se diferencias de los de tipo II por una rápida degradación tras su fotoconversión de la forma Pr a Pfr. Esta degradación de los fitocromos es un problema al no producirse la suficiente fitocromobilina a un ritmo adecuado. Sin embargo, en estadíos de desarrollo más avanzados en los que la regulación de la germinación y el desarrollo escotomorfogénico ya no son necesarios. Los fitocromos de tipo II, el fitocromo B y C, adquieren el papel protagonista al regular las respuestas a los ratios de luz Rojo/Rojo lejano. Estos fitocromo al no ser tan inestables permiten la fotoconversion Pr <-> Pfr varias veces, con lo que la misma fitocromobilina puede ser empleada varias veces y el bajo ratio de producción pasa a ser un problema menor respecto a los estadio iniciales. Así pues ambos mutantes, *s73* y *G123* deberían presentar fenotipo similar a la ausencia de fitocromos en los primeros estadios de desarrollo, pero estos deberían desaparecer en fases más avanzadas. Aun así, los efectos derivados de la acumulación de Hemo en la ruta de síntesis de tetrapirroles, es decir, la inhibición de la síntesis de clorofila, deberían ser más graves en el mutante *s73*.

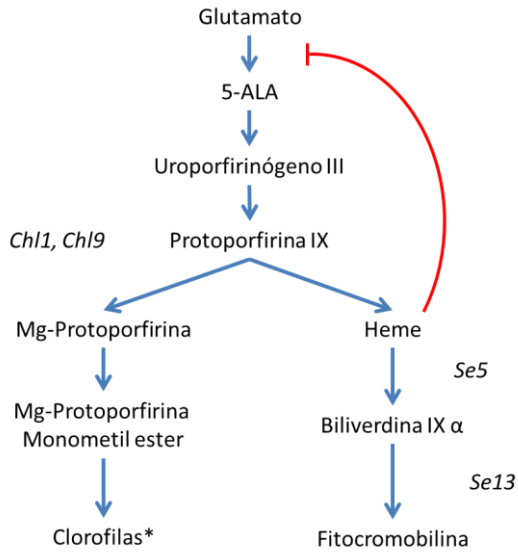


Figura 28: Esquema ruta de síntesis de los tetrapirroles en arroz, Yoshitake et al., 2015.

Los datos fenotípicos obtenidos en este estudio junto a los análisis del transcriptoma del mutante *G123* indican que la deleción detectada en se13/*OsHY2* es la responsable del fenotipo alterado de *G123*. Para confirmar esta hipótesis sería necesario producir la reversión del fenotipo del mutante mediante la sobre expresión de este gen en plantas transgénicas. *Se13*/*OsHY2* parece ser el único gen presente en la zona de la deleción común en las 20 plantas de la F2 con fenotipo mutante relacionado con la floración. Esta técnica empleada para su detección pertenece a las últimas metodologías empleadas en genómica que se aprovecha de la reducción de costes en la secuenciación de genoma completo, a diferencia de técnicas anteriores como eran rastreo de QTLs, o el GWAs, basado en marcadores intercalados a lo largo del genoma.

En los últimos años se han desarrollado metodologías basadas en las secuenciación de genoma completo que permiten detectar las mutaciones responsables de fenotipos

alterados mediante el análisis de plantas de una generación F2. Estas técnicas presentan como principal ventaja respecto al mapeo de QTLs la reducción en el tiempo empleado para realizar un mapeo fino que evita tener que cultivar varias generaciones de plantas con el tiempo que conlleva. Como ejemplos de estas técnicas recientemente desarrolladas se puede citar el SHOREMAP (Ossowski et al., 2008; Schneeberger et al., 2009), desarrollado en 2009, con el que se puede detectar la mutación mediante un cruce entre la línea mutante y la silvestre y posteriormente secuenciando una mezcla de 500 plantas F2 y comparándolas con el genoma de referencia. Otro método reciente es el Mutmap (Abe et al., 2012). Este sistema presenta la ventaja sobre el SHOREMAP de utilizar tan solo 20 plantas F2 con fenotipo recesivo para detectar la mutación responsable. Sin embargo, su uso se restringe a la detección mutaciones de tipo SNP. El sistema empleado en esta tesis ha permitido la detección de una mutación con alto potencial a ser la candidata responsable, pese a no ser de tipo SNP. Pero ha sido necesaria la combinación del Mutmap y de la metodología recientemente desarrollada en el laboratorio, para poder barrer rápidamente todo el genoma en busca de la mutación responsable del fenotipo mutante, cubriendo todo tipo de alteraciones. Esta nueva metodología es mencionable dado que existen pocos programas desarrollados actualmente capaces de detectar correctamente las mutaciones estructurales. Y el hecho de haber generado un pipeline capaz de descartar la mayoría de estas mutaciones generando pocos falsos positivos, pudiendo ser capaz de realizar un filtrado manual, abre la posibilidad a la reducción de tiempos en la investigación de mutaciones y a la aceleración de esta.

3.3.4. Conclusiones

En este capítulo se ha presentado la generación e identificación de una línea mutante con fenotipo temprano de floración respecto a la variedad silvestre. Se ha caracterizado tanto a nivel fenotípico como de expresión génica y se ha identificado la

mutación mediante una variación de la técnica de "*bulk segregant analysis*". La reducción de costes junto con la elevada capacidad de computación y los avances en los algoritmos de detección de mutaciones puntuales y estructurales han permitido identificar la mutación candidata a ser la responsable del fenotipo alterado. A falta de verificación, el gen *Se13/OsHY2* se postula como una variación interesante para su introducción en variedades comerciales al provocar un acortamiento del ciclo de las plantas de arroz y con pocos efectos laterales al situarse en un punto próximo al final de la ruta de síntesis de las fitocromobilinas.

3.3.5. Materiales y métodos

Obtención de una línea de floración temprana.

Se irradiaron semillas de la variedad Gleva con 25 Gy de neutrones rápidos. Las semillas irradiadas (M1) fueron cultivadas en macetas en invernadero. Una vez las plantas maduraron, se agruparon en familias de 5 plantas y se recolectaron sus semillas. Ciento veintidós plantas M2 de cada familia fueron cultivadas en filas en campo con un marco de cultivo de 20 x 20 cm y se rastreó en busca de plantas con una floración más temprana que el parental control (Gleva). Las plantas M3 fueron cultivadas en verano en balsas al aire libre asemejando las condiciones de cultivo del campo y se anotaron las fechas de floración.

Condiciones de cultivo en cámara.

Las plantas utilizadas en los experimentos de sensibilidad a fotoperiodo, análisis de la expresión génica, RNA-seq y detección de mutación fueron cultivadas en fitotrones (SANYO Mod. MLR350) equipadas con tubos fluorescentes de amplio espectro (400-700 nm) (GROLUX F36W/GRO-T8. Sylvania, Germany) con una intensidad lumínica de 250$\mu mol^{-2} \cdot s^{-1}$. Se consideró fotoperiodo largo 14 horas de luz más 10 horas de

oscuridad, fotoperiodo corto 10 horas de luz y 14 horas de oscuridad y fotoperiodo neutro 12 horas de luz y 12 de oscuridad. La temperatura se mantuvo constante a 27 °C en todos los experimentos.

Ensayos de sensibilidad a fotoperiodo

Se sembraron 3 semillas de cada variedad por maceta en dos macetas (6 semillas en total) de las variedades Gleva, como control, y *G123*. Tras cuatro semanas en condiciones de fotoperiodo neutro (12h luz: 12h oscuridad), una de las macetas de cada variedad se cambió a fotoperiodo largo y la otra a fotoperiodo corto. El tiempo de floración se anotó como aquel momento en el que la mitad de la primera panícula había emergido de la vaina.

Análisis del perfil de expresión de los principales genes de floración

Obtención material vegetal

Se sembraron 6 macetas con 3 semillas de cada variedad (18 semillas en total) de las variedades Gleva, como control, y *G123*. Tras cuatro semanas de cultivo en condiciones de fotoperiodo neutro, una de las macetas de cada variedad se cambió a fotoperiodo largo y la otra a fotoperiodo corto. Al finalizar la quinta semana se tomaron series temporales de las muestras tomando la penúltima hoja de tres plantas diferentes cada hora cada 4 horas (seis muestreos en total). Las muestras se congelaron con nitrógeno líquido y se almacenaron a -80 °C hasta la extracción de ARN. Se considera 0 horas al momento en el que las plantas comienzan a recibir luz.

Extracción de ARN.

La extracción de ARN se realizó con un método basado en la precipitación del ARN con LiCl. El material vegetal se congeló en nitrógeno líquido, se pulverizó con mortero y se recogió el polvo en tubos Eppendorf de 1,5 mL. A continuación se añadió 700 µL de tampón de extracción y 175 µL de una mezcla de fenol:cloroformo:alcohol isoamílico

(25:24:1) y se agitó con un vórtex. Después se centrifugó a 2.000 rpm 10 minutos a temperatura ambiente. Posteriormente la fase acuosa superior se transfirió a un tubo nuevo. A la fase acuosa se le añadieron 750 µL de fenol:cloroformo:alcohol isoamílico (25:24:1) y se agitó vigorosamente con vórtex. De nuevo se centrifugó a 2.000 rpm 10 minutos a temperatura ambiente. Se recuperó la fase acuosa superior en otro tubo Eppendorf y se añadió 12,5 µL de LiCL 8M (para obtener una concentración final de 2M de LiCl). Posteriormente se mezcló con vórtex y se incubó a 4 °C un mínimo 4 horas tras las cuales se mezcló por inversión de los tubos y se centrifugó a 4°C 10 minutos a 12.000 rpm. Mediante decantación se eliminó la fase superior y se añadió 500 µL de LiCL2 M. Tras mezclarlos se centrifugó a 4 °C 10 minutos 12.000 rpm. Tras eliminar la fase superior, se resuspendió el pellet en 500 µL de agua. Se añadieron 500 µL de cloroformo:isoamil alcohol (24:1) y se centrifugó 10 minutos a 4°C a 12.000 rpm. Se transfierió la fase acuosa a un tubo nuevo y se precipitó el ARN añadiendo 1/10 del volumen en acetato de sodio 3 M y 2 volúmenes de etanol absoluto. Se incubó 1 hora a -20°C. Se centrifugó a 4 °C 15 minutos a 12.000 rpm. Se eliminó la fase superior mediante decantación y se resuspendió el pellet en TE o agua.

Tampón de extracción de ARN

1. 0,1 M LiCl
2. 0,1 M Tris pH8
3. 1% SDS
4. 0,01 EDTA

La concentración de ARN de las muestras se midió con el Qubit™ RNA BT Assay Kit (Ref: Q10211), siguiendo las instrucciones del fabricante, y usando el Qubit® 2.0 Fluorometer (Life technologies, EE.UU).

RT-qPCR

Se diseñaron cebadores específicos para las reacciones de RT-qPCR empleando la herramienta PrimerQuest (https://eu.idtdna.com/PrimerQuest/Home/Index). Esta página web permite indicar los puntos de unión exón-exón del gen de interés de modo que facilita el diseño de cebadores que se superpongan a estas posiciones, con el fin de evitar la amplificación de productos derivados de una posible contaminación de ADN en las extracciones. Cuando no fue posible el diseño de cebadores de buena calidad que se situasen en la zona de unión exón-exón, se utilizó el programa "Primer3" (http://bioinfo.ut.ee/primer3-0.4.0/). Ambas herramientas permiten especificar las características de los cebadores deseados como el tamaño del amplicón, el GC%, la temperatura de desnaturalización, el máximo de autocomplementariedad y el máximo de autocomplementariedad en el extremo 3'.

Los análisis de expresión génica fueron llevados a cabo realizando RT-qPCR en un solo paso, estas se realizaron empleando una máquina LightCycler 2.0 de Roche (Roche Diagnostics GmbH, Mannheim, Alemania) provista del programa LightCycler 4.0. Las mezclas de las reacciones se prepararon siguiendo las instrucciones del fabricante del kit light Cycler® Fast Start DNA Master^{Plus} SYBR Green I (Applied Biosystems TM, Ref: 03515885001), añadiendo transcriptasa reversa M-MuLV Roche®, empleando 100 ng de ARN de las muestras y cebadores específicos para cada gen, con una concentración final de 250 nM en un volumen total de 10 µL. La solución final contenía 2,5 µL de muestra, 4,4 µL de vial 2 (H_2O calidad PCR), 2 µL de vial 1b+1a (Master Mix), 0,5 µL de cebador 5' (5 µM), 0,5 µL cebador 3' (5 µM), 0,05 µL de inhibidor de RNasa (ribonucleasa) (20 U/ µL; Applied Biosystems; Ref. N8080119). El vial "1a" contiene la enzima LightCycler®FastStart, el vial "1b" contiene la Taq ADN polimerasa FastStart, el buffer de reacción, $MgCl_2$, SYBR Green I dye y el mix de dNTP (con dUTP en lugar de DTTP). Los programas de reacción consistieron en 30 minutos de incubación a 48°C (retrotranscripción) 95°C 10 minutos (desnaturalización y activación de la FastStart

Taq ADN polimerasa), 45 ciclos de amplificación que consistían en 95°C dos segundos, 55-61°C 3-8 segundos, 72°C 8 segundos. A continuación 95°C 15 segundos, 42 °C un minuto y gradiente de temperatura desde los 42 °C a los 95°C con una rampa de 0,1 °C/s. La intensidad de la fluorescencia se tomó durante la extensión a 72°C y el gradiente de temperatura. La especificidad de la reacción se comprobó mediante la curva de desnaturalización obtenida durante el gradiente de temperatura.

Para convertir los valores de intensidad de fluorescencia en una medida relativa de los niveles de expresión génica, primero se obtuvo la eficiencia de cada par de cebadores mediante la obtención de una curva de calibrado usando diluciones seriadas de muestras de ARN a 80 ng/µL, 40 ng/µL, 20 ng/µL, 10 ng/µL y realizando tres réplicas para cada concentración. Una vez las curvas de eficiencia fueron obtenidas, la fluorescencia se normalizó respecto a LOC_Os03g13170, que codifica una la ubiquitina, la cual es considerada un gen expresado constitutivamente con poca variación en su expresión. Empleando la fórmula:

$$\frac{Efificiencia_{RG}^{CT_{RG}}}{Eficiencia_{GOI}^{CT_{GOI}}}$$

CT: ciclo umbral
GOI: gen de interés
RG: Gen referencia.

Tabla 18: Cebadores usados para las reacciones de RT-qPCR.				
Gen	Locus	Primer Sequence F/R (5'-3')	Temperatura de alineamiento (^{o}C)	Ciclos
DTH2	LOC_Os02g49230	GATTTCTGCAGGGAGCAAAG / TTCAAGACAACGGACTGCTG	60	45
Ehd1	LOC_Os10g32600	TCGGAGAAGACAAGGCAGTT /CCGTGTTTGCTTGTTGTTGG	59	45
ELF3	LOC_Os06g05060	ACTACTTCCCGCCTTTCAGC /ATCCACGACTGCTGCTCAAA	60	45
Ghd7	LOC_Os07g15770	TATTGTGGGAGCACGTTCAC / ATCTGAACCATTGTCCAAGC	57	45
Ghd8	LOC_Os08g07740	CGAAGGAGCAGGACAGGTTC / AGCTGATGAACTCCGACACG	62	45
Hd1	LOC_Os06g16370	CTTACACAGATTCCATCAGC / CATACGCCTTTCTTGTTTCA	55	45
Hd3a	LOC_Os06g06320	GATGCACCAAGCCCAAGT / GGAACAGCACGAACACCA	61	45
HD6	LOC_Os03g55389	GTTCAATGGGGTGAGCAGGA / CTTCACAGGCTTGAGTATCTTGA	59	45
OsGI	LOC_Os01g08700	GTGCCGTCTATCAACCACCA / AAGGACGGACATGCTGAGTG	60	45
PRR37	LOC_Os07g49460	CCTATGGCAGCATGTGTGGA / ACCATCGTCGTCATCATCGT	60	45
RFT1	LOC_Os06g06300	GGATTGAACGGCAGGAGATA /CGGCCATGTCAAATTAATAACC	60	45
Ubiquitin fusion protein	LOC_Os03g13170	GCTCCGTGGCGGTATCAT / CGGCAGTTGACAGCCCTAG	55	45

RNA-seq

Obtención de material vegetal

Las plantas fueron cultivadas en un fitotrón durante cuatro semanas bajo condiciones de fotoperiodo neutro a 27 °C. Transcurridas las cuatro semanas fueron cultivadas en condiciones de fotoperiodo largo durante una semana antes de la toma de muestras. Las muestras se tomaron de la penúltima hoja 20 horas tras el encendido de las luces y se mantuvieron en nitrógeno líquido hasta que fueron almacenadas a -80 antes de la extracción de ARN.

Extracción de ARN:

La extracción de ARN para el RNA-seq se realizó empleando el quit NucleSpin® RNA plant (Ref: 740949.50, MACHEREY-NAGEL, Alemania). Siguiendo las instrucciones del fabricante. La concentración final y la calidad del ARN fue comprobada con espectrofotómetro Nanodrop®.

Secuenciación de ARN

Las muestras de ARN se secuenciaron por la compañía Novogen Bioinformatics Technology Co., Ltd, Honkong tras pasar un control de calidad por parte de la compañía. La secuenciación consistió en la construcción de una biblioteca de 250~300 pb de tamaño de inserto, seguido de su secuenciación mediante lecturas *pair-end* de 150 pb.

La calidad de la muestra fue comprobada de nuevo mediante una cuantificación preliminar con un espectrofotómetro Nanodrop, una electroforesis en gel de agarosa para comprobar la degradación del ARN y las posibles contaminaciones, y, finalmente, mediante el bioanalizador Algilent® 2100 para comprobar la integridad y cantidad.

Tras el control de calidad, el ARNm fue enriquecido empleando bolas de oligo(dT). Primero, el ARNm es fragmentado aleatoriamente mediante la adición de un tampón de fragmentación, a continuación se sintetiza el ADNc empleando cebadores de hexámeros aleatorios, tras lo cual se sintetiza la cadena complementaria mediante la adición del tampón "second-strand synthesis buffer" (Illumina) que contiene dNTPs (desoxirribonucleótidos trifosfato), RNasa H y ADN polimerasa. Posteriormente se realizan unas series de reparaciones de los extremos terminales, ligación y ligación del adaptador de secuenciación. Finalmente, la biblioteca de ADNc de doble cadena se ve terminada al realizar una selección por tamaño mediante enriquecimiento por PCR.

El control de calidad de la biblioteca consta de tres pasos: primero se comprueba la concentración mediante Qubit 2.0, segundo con el programa Alginet 2100 se comprueba el tamaño del inserto, y finalmente la concentración efectiva de la biblioteca se cuantifica de forma precisa mediante Q-PCR.

Las bibliotecas que han pasado los controles de calidad se introducen en el secuenciador HiSeq/MiSeq tras mezclarlas según sus concentraciones efectivas y el volumen datos esperados. Las lecturas brutas son filtradas para eliminar las lecturas de baja calidad y los adaptadores. De esta manera, se eliminan las lecturas que: contienen secuencias de los adaptadores, las lecturas que contienen un 10% de sus bases indeterminadas y las lecturas con más el 50% de las bases de baja calidad (Qscore <= 5).

La secuencia de los adaptadores usada fue:

Adaptador 5'

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

Adaptador 3' (las seis bases subrayadas corresponden el índice):

5'-
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG
-3

Análisis de expresión diferencial

Previo al análisis de expresión diferencial se realizó un análisis de calidad extra empleando FstQC High Throughput Sequence QC Report (version:0.11.5, www.bioinformatics.babraham.ac.uk/projects/). Para el análisis de expresión diferencial se empleó el software CLC Genomics Workbench versión 7.5.2 (QIAGEN, Alemania). Las lecturas fueron importadas como "Paired end", y se eliminaron las lecturas fallidas, mediante una de las opciones del programa para las lecturas Illumina, y se incluyó la restricción de una distancia entre los pares entre 250 pb y 300 pb. Las lecturas fueron recortadas en base a su calidad con un límite de 0,049 y un número máximo de nucleótidos ambiguos de 2. Además los primeros 15 nucleótidos del extremo 5' de todas las lecturas fueron eliminados debido a que presentaban discrepancia el porcentaje de bases según los informes del FQC.

Para el mapeo de las lecturas, el CLC Genomics presenta una herramienta llamada RNA-seq, para la cual se selecciona la opción "Genome annotated with genes and transcripts" para introducir el genoma de referencia, esta opción, que es la recomendada para genomas de eucariotas, permite la introducción de un fichero con los datos del ARNm cuando estos se encuentran disponibles, también esta opción permite tener en cuenta las versiones alternativas de los genes. El genoma de referencia de arroz y los ficheros gff fueron descargados de la base de datos Rice Genome annotation Project (MSU versión http://rice.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dir/). Para el modo de mapeo se seleccionó la opción "intergenic", las opciones de mapeo y de alineamiento de secuencias fueron:

“Coste del discrepancia: 2”, “coste de inserción: 3”, “Coste de deleción: 3”,” fracción de lectura: 0.8”, “auto detectar distancia de emparejado”, “dirección específica: ambas”, “máximo número de mapeo por lectura: 10”. Los niveles de expresión se representaron como RPKM (*Reads Per Kilobase per Million mapped reads*) como método de normalización.

Para el análisis de expresión diferencial de genes se empleó la herramienta “DGE”. Esta herramienta implementa el “Exact test” para la comparación de dos grupos (Robinson & Smyth, 2008). Los parámetros para este test fueron*: “two group comparison (unpaired), expression values: set new expression values from samples, common dispersion cutoff: 5. Exact test comparison: against reference, add corrected p-values (FDR-corrected)”* el umbral para considerar a dos genes como diferencialmente expresados entre control y los mutantes fue FDR < 0.1.

Para el enriquecimiento de términos GO (*Gene Ontology*) de genes diferencialmente anotados se utilizó la plataforma CARMO (Comprehensive Annotation of Rice Multi-Omics data) (J. Wang, Qi, Liu, & Zhang, 2015, http://bioinfo.sibs.ac.cn/carmo/).

Detección de la mutación

Material vegetal, generación de una F2

Semillas del parental silvestre, Gleva, fueron sembradas con 14 días de antelación a las de la línea mutante, *G123*, para sincronizar la floración de ambos y posibilitar su cruzamiento. Flores de la línea mutante fueron emasculadas en el momento que sus panículas emergieron entre 5 y 10 cm de la hoja bandera. En primer lugar, se separó la panícula deseada del resto de los tallos, y se eliminaron manualmente tanto las espiguillas de la parte inferior en la que los ovarios son inmaduros, como las de la parte superior de la panícula que pudiesen haber sido autopolinizadas. La emasculación de las flores restantes se realizó sumergiendo la panícula en agua caliente (45°C) 5 minutos. Tras la emasculación las flores fueron cortadas ligeramente

por la mitad superior sin dañar el ovario, con el fin de eliminar las anteras que no hubiesen sido dañadas, y favorecer la polinización por parte del parental donante. Posteriormente la panícula emasculada y una panícula del parental masculino, cuyas flores contenían polen maduro, visible puesto que las anteras emergen de la flor, son introducidas conjuntamente en una bolsa de papel muy fina. Las raíces del tallo de la panícula receptora se introducen en un recipiente con agua mientras que la panícula masculina, a la que se le han cortado las raíces se ata a la femenina evitando que se caiga y manteniéndola junto a esta, ligeramente por encima. Las panículas se sacuden ligeramente con el objetivo de favorecer que el polen caiga sobre las flores emasculadas. Una semana más tarde los granos fertilizados son visibles. El tiempo de desarrollo de las semillas F1 es de unos 28 días. Es preferible realizar la emasculación antes de las 10:00 am y el cruce entre las 10:00 am y las 12:30 am puesto que en ese periodo es cuanto más polen se libera.

Se obtuvo semillas de la generación F2 mediante la autopolinización de las plantas F1. Cuatrocientas semillas de plantas F2 fueron sembradas en macetas, tres semillas por maceta, y cultivadas en invernadero bajo condiciones de luz natural, en verano de 2017. La fecha de floración de las plantas fue anotada al momento en el que la mitad de la primera panícula había emergido. Se consideraron plantas tempranas aquellas cuya floración resultó inferior a 72 días tras la siembra. Se empleó un test de chi cuadrado para comprobar la hipótesis de un solo gen recesivo.

Extracción de ADN nuclear

El ADN nuclear fue extraído de 20 plantas de la generación F2 de fenotipo temprano, de una planta del parental silvestre, Gleva, y de una de la línea mutante, *G123*.

Se partió de 2 gramos de material fresco (el material vegetal puede ser almacenado en nitrógeno y posteriormente guardado en un -80 °C hasta el momento de la

extracción, pero ha de ser descongelado a temperatura ambiente). Las hojas se cortaron en trocitos de 4 mm con un bisturí y se introdujeron en un tubo Falcon de 50 mL, al que se le añadió el doble de volumen de tampón de extracción 1x. Se homogenizó empleando un Polytron (manteniendo el tubo Falcon el hielo) y se filtró decantando en un nuevo tubo Falcon empleando un filtro de Miracloth®. Se añadió tampón de extracción 1 x hasta llegar a los 30 mL. A continuación se añadieron lentamente 1,24 mL de solución Triton X-100 25 % y agitando con el fin de evitar concentraciones elevadas de la solución en algún punto. Posteriormente se preparan las soluciones para los gradientes de Percoll®, consistentes en 6 mL de Percoll® 80% y 12 mL al 30 %. A continuación en un nuevo tubo Falcon se añaden 6 mL de Percoll® 30% y, empleando una pipeta de cristal de 25 mL, se añaden 6 mL de percoll 80% atravesando la capa de 30% y dejando que se deposite bajo esta lentamente para no perturbar la interfase. A continuación sobre la capa de Percoll® 30% se deposita el homogenizado de material vegetal, el tampón de extracción y Triton-X100 25% decantando lentamente contra la pared del tubo Falcon. Se centrifuga a 2.000 rcf 30 minutos a 4 °C en una centrífuga de rotor basculante. A continuación, la capa superior, situada encima de la capa de Percoll® 80% es eliminada empleando una pipeta de 25mL de cristal, lentamente a fin de evitar el reflujo (los núcleos se encuentran suspendidos sobre la capa de Percoll® 80% por lo que es importante no tocar la interfase). A continuación el tubo Falcon se rellena hasta los 10 mL con tampón de extracción 1x y justo debajo, empleando una pipeta de cristal de 5 mL, para pasar a través, se depositan 6 mL de Percoll®30%. El tubo es centrifugado a 2.000 rcf 10 minutos a 4°C en una centrífuga de rotor basculante. El sobrenadante es entonces decantado. El pellet depositado en el fondo del tubo Falcon contiene los núcleos (grises), junto a almidón (blanco) y restos celulares (verde). Posteriormente el pellet es resuspendido en 500 µL de tampón CTAB a 60°C y transferido a tubos Eppendorf de 2 mL. Se incuba 30 minutos a 60 °C y a continuación se añaden 750 µL

de cloroformo:alcohol isoamílico (24:1) y se mezcla por inversión hasta que se forma una emulsión. Esta mezcla es centrifugada 10 minutos a 7.000 rcf a 4°C en una centrífuga de rotor fijo. La fase superior se transfiere a un nuevo tubo Eppendorf de 1,5 mL evitando tocar la interfase. Después se añaden 0,4 mg de RNasa A y se incuba a 37°C 30 minutos. Luego se añade un volumen de fenol:cloroformo:alcohol isoamílico (25:24:1), se mezcla por inversión y se centrifuga a 12.000 rcf 5 minutos en una centrífuga de rotor fijo. A continuación se tranfiere el sobrenadante a un nuevo tubo de 1,5 mL. Se añade un volumen de isopropopanol, se mezcla por inversión, se incuba 2,5 minutos a temperatura ambiente y se centrifuga 15 minutos a 12.000 rcf en una centrifuga de ángulo fijo. Se decanta el sobrenadante y se lava el pellet con 500 µL de EtOH 70% (invertir una sola vez suavemente). Se centrifuga 10 minutos a velocidad máxima a 4 °C. Se decanta el sobrenadante y se deja secar a temperatura ambiente. Finalmente el pellet es resuspendido en 55µL de TE.

Tampones:

Tampón de extracción 5 x

- 10 M Hexilenglicol (2-methyl-2,4-pentandiol).
- 100 mM PIPES-KOH (pH 7)
- 50 mM MgCl2
- 25 mM β-Mercaptoetanol (añadir justo antes de usar para preparar el tampón de extracción 1x)

Tampón para gradiente 5x

- 2,5 M hexilenglicol

- 25 mM PIPES-KOH (pH 7)
- 50 mM MgCl2
- 5% Triton X-100
- 25 mM β-Mercaptoetanol (añadir justo antes de usar)

CTAB 2%

- 4g CTAB (bromuro de cetiltrimetilamonio)
- 16,36 g NaCl
- 8 mL 0,5 N EDTA pH 8
- 20 mL 1M Tris-HCl pH 8
- Enrasar con agua hasta llegar a los 200 mL.

Percoll 80%

- Percoll® 80%
- Tampón para gradiente 5x 20%

Percoll 30%

- Percoll® 30%
- Tampón para gradiente 5x 20%
- Agua 50%

La cantidad y calidad del ADN de las muestras fueron comprobadas con Qubit™ dsDNA BR Assay Kit (Ref: Q32853), siguiendo las instrucciones del fabricante y empleando Qubit® 2.0 Fluorometer (Life technologies, EE.UU).

Secuenciación.

Tal y como indican Abe (Abe et al., 2012) se preparó una mezcla de ADN nuclear de 20 individuos de la generación F2 que presentaban el mismo fenotipo que el mutante, en la misma proporción. La secuenciación de genoma completo de la muestra de la mezcla de los 20 individuos, junto con una muestra del parental silvestre y el parental mutante se obtuvo por medio de Novogen Bioinformatics Technology Co., Ltd

Para la construcción de la biblioteca de ADN genómico cada muestra fue troceada en fragmentos de unos 350 pb. Los fragmentos obtenidos se emplearon junto con el kit NEBNext® DNA Library Prep Kit, siguiendo las instrucciones del fabricante. A continuación se realizó la reparación de extremos, adición de colas dAMP (*dA-tailing*), y una ligación más con el adaptador NEBNext, los fragmentos requeridos (300-500 pb) fueron enriquecidos mediante oligos indexados P5 y P7. Tras la purificación y la comprobación de calidad, la biblioteca resultante se encontraba lista para la secuenciación.

El control de calidad de la biblioteca se realizó primero con un fluorímetro Qubit®2.0, que determinó la concentración de la biblioteca. Tras la dilución a 1ng/µL se empleó el bioanalizador Algilent® 2100 para verificar el tamaño del inserto. Finalmente se realiza una PCR en tiempo real (qPCR) para detectar la concentración efectiva de cada biblioteca. Las bibliotecas aptas fueron aquellas con un tamaño de inserto apropiado (~350 pb) y una concentración efectiva de más de 2 nM.

Las bibliotecas de ADN aptas fueron mezcladas según su concentración efectiva y la cantidad de datos a producir esperados. La secuenciación de lecturas pair-end se

realizó en la plataforma Illumina®sequencing con un tamaño de lectura PE150pb en cada extremo.

Los datos brutos obtenidos de la secuenciación fueron filtrados con el fin de descartar las lecturas emparejadas que presentasen contaminación con adaptadores o que lo nucleótidos indeterminados constituyesen más del 10% de la secuencia o nucleótidos con una calidad baja (calidad de las bases menor que 5, Q<5) constituyesen más el 50% de la lectura.

Mutmap

Para el análisis Mutmap se empleó el software Mutmap framework 1.4.4 (http://genome-e.ibrc.or.jp/home/bioinformatics-team/mutmap) con ligeras modificaciones. Este programa realiza el análisis en tres fases. En la primera fase, "Secuenciación del genoma completo y selección de lecturas de gran calidad" se seleccionan únicamente lecturas cortas de gran calidad, que por defecto presentaban un valor phred >30 en al menos el 90% de las bases de la lectura. En la segunda fase, se genera una secuencia referencia a partir de unos de las variedades o líneas empleadas en el cruce para la obtención de la F2. En este paso, en primer lugar se detectan los SNPs mediante alineamiento y el filtrado de las lecturas *paired-end* frente al genoma de referencia mediante el programa BWA (H. Li & Durbin, 2009). Como genoma de referencia se empleó el de Nipponbare MSU v7. En segundo lugar, se genera un nuevo genoma de referencia mediante la sustitución de los SNPs detectados respecto al genoma de referencia. En tercer lugar, las lecturas usadas para la generación del nuevo genoma de referencia son realineadas a la secuencia de referencia para detectar SNPs adicionales. Estos SNPs adicionales pueden darse debido a errores de alineamiento y es recomendable que se excluyan del análisis MutMap. En la última fase se realiza el análisis de frecuencias de SNPs (SNP-index) a lo largo de todo el genoma. En este paso, en primer lugar los SNPs son detectados mediante el alineamiento de las lecturas filtradas y la ecualización de cada ADN

mezclado mediante el programa BWA. En segundo lugar se filtran los SNPs detectados mediante el programa coval (Kosugi et al., 2013). Los SNPs detectados en el paso dos son excluidos. En tercer lugar se generan los intervalos de confianza para los valores de SNP-index mediante su cálculo en cada posición de SNP bajo la hipótesis nula de muestras aleatoriamente recogidas. En cuarto lugar se pueden delinear las regiones candidatas mediante las gráficas del SNP-index para todos los cromosomas. Para identificar la región candidata se realiza un análisis de "ventana deslizante" *(sliding window)* para ventanas de 2 y 4 Mb con intervalos de 10 kb.

Se descartaron aquellas regiones genómicas que presentaban un SNP-index < 0,9, los SNPs presentes en las regiones restantes fueron filtrados visualmente empleando el programa Integrative Genomics viewer (http://software.broadinstitute.org/software/igv/, Broad institute) empleando como genoma de referencia el obtenido a partir de la secuenciación del genoma de Gleva. Tras la filtración tres SNPs quedaron como posibles canidatos.

Con el fin de comprobar los SNP candidatos resultantes del análisis Mutmap, se diseñaron cebadores para amplificar la región donde se encontraban, empleando el programa Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/), especificando que el cebador 5' estuviese situado al menos a 90 pb del SNP candidato. El tamaño del amplicón producto de la PCR fue comprobado empleando geles de agarosa y electroforesis con tampón TAE 1,5%.

Tabla 19: Cebadores usados en las PCRs para la verificación de lo SNP candidatos productos del Mutmap.

Región de amplificación	**Secuencia de lo cebadores F/R (5' - 3')**	**Temperatura de annealing (°C)**	**Tamaño del producto (pb)**
Chr04:28150431_28150749	TGATAGCGGTTTCGTTGACA/ GGAAAATCTCGATGGCGTAA	60	318
Chr01:39252079_39252329	TTGGACCATCGGATATGCTT/ GCGTTGGTGACAGGAAATCT	55 *	250
Chr01:39443046_39443416	TGAATTGGAGCTGCTACACG/ CGCAATCTCCAGTCAGCATA	60	370

*amplificación específica empleando la polimerasa Phusion High-Fidelity DNA polymerase (Ref: F530S, ThermoFisher Scientific).

Una vez se verificaron los tamaños de los productos de PCR estos fueron secuenciados mediante tecnología Sanger por Secugen S.L (Madrid, Spain). Los cromatogramas fueron visualizados empleando el programa Chromas (v. 2.6.5, Technelysium Pty Ltd). Las secuencias fasta fuero alineadas al referencia para su visualización empleando el programa Mega 7.0.26 (Kumar, Stecher, & Tamura, 2016)

Análisis de las variaciones estructurales.

Para el análisis de las variaciones estructurales se empleó el software Allinone generado en el propio departamento de Genómica del Instituto Valenciano de Investigaciones Agrarias. Allinone combina varios programas informáticos generando una línea de trabajo que permite la detección y filtrado de variaciones estructurales entre el genoma de referencia y las muestras.

Los pasos que conforman este script son, en primer lugar filtrar las lecturas por calidad empleando el script de Python QC.py con los parámetros "minQF =30 y minpercQF = 70" (elimina aquellas lecturas con un valor phred < 30 en menos del 70%

de la lectura). En segundo lugar las lecturas que pasan el filtro de calidad son mapeadas frente al genoma de referencia, en este caso Nipponbare (MSU versión 7) empleando el software BWA, con el comando "mem –t 4 –r 1.2 –R $6". Como resultado se generan unos ficheros Sam que son convertidos a formato BAM empleando el programa samtools versión 1.4 (H. Li et al., 2009) empleando el comando "samtools view –b". Los ficheros BAM son ordenados empleando el comando "samtools sort" y se crea un índice empleando el comando "samtools index". Para la detección de las variaciones estructurales diferentes ficheros BAM deben ser creados, con el comando "samtools view –b –F 1294" | "samtools sort" las lecturas discordantes son seleccionadas para un nuevo fichero BAM. Con el comando "samtools view –h" se genera un nuevo fichero SAM y junto con el programa Lumpy (Layer, Chiang, Quinlan, & Hall, 2014) empleando el comando "extractSplitReads_BwaMem –i stdin | samtools view Sb - |samtools sort" se genera un fichero bam con las split reads.

En tercer lugar las variaciones estructurales fueron detectadas para cada una de las muestras empleando de nuevo Lumpy con el comando "lumppyexpress –B (fichero BAM) – D (LecturasDiscordantes.bam) –S (SplitReads.bam) –o (Salida.vcf). Con esto se genera un fichero VCF con las posiciones de las variaciones estructurales. A continuación se emplea el programa *svtyper* con el comando "–i –B –o" de modo que se genera el mismo BAM obtenido con el Lumpy pero con la información genética para cada variación estructural detectada.

Finalmente con el programa SnpSift.jar (versión 4.3) los VCFs anteriormente generados son filtrados para el número de lecturas que soporta cada variación en cada una de las muestras. En nuestro caso se filtró de modo que solo se retuviesen aquellas variaciones que fuesen soportadas por el 0,001 de las lecturas en Gleva y más del 0,95 en *G123* y la mezcla de F2 "GEN[gleva].AB<0.001 & GEN[*G123*].AB>0.95 & GEN[EPool_HCMC3CCXY_L2].AB>0.95">". Es decir aquellas variaciones ausentes en

Gleva pero presentes en el parental mutante y en los 20 individuos de la F2 con fenotipo mutante.

Finalmente las variaciones estructurales que pasaron el anterior filtro fueron manualmente verificadas empleando el programa IGV a modo de evitar falsos positivos.

4. Discusión general

Desde la publicación de la secuencia del genoma del arroz (Matsumoto et al., 2005) y la subsiguiente generación de bases de datos de miles de SNPs que pueden caracterizar cualquier variedad que se cultive hoy en día (Alexandrov et al., 2015), el uso de marcadores en los programas de mejora del arroz se ha hecho un procedimiento rutinario, acelerándolos y haciéndolos más dirigidos. En contrapartida, a nivel local existe escasa variabilidad genética por lo que para introducir nuevos caracteres de interés los mejoradores se ven obligados a hacer uso de parentales genéticamente distantes, adaptados a condiciones climáticas diferentes a las presentes en nuestro país y con una duración del ciclo vegetativo inapropiada. Las plantas resultantes del cruce entre las variedades locales y las exóticas suelen arrastrar junto con la característica de interés otras no aptas para el cultivo en la región, siendo necesarios varios retrocruzamientos hasta obtener una variedad superior. Como estrategia para evitar este problema, puesto que el principal factor limitante para el cultivo es el fotoperiodo, en esta tesis se ha estudiado la diversidad genética y fenotípica presente en las regiones de clima templado, en la que todas las variedades presentan escasa sensibilidad a fotoperiodo, en mayor o menor grado, pudiendo así ser cultivadas durante los largos días de verano. Para ello se ha generado y caracterizado una colección de 193 variedades representativas de los cultivos en esta área. Con el fin de genotipar esta colección, se ha desarrollado un panel de 1.793 SNPs apto para el genotipado de las variedades *japonica*. El desarrollo de este panel era necesario puesto que los chips de genotipado para arroz se han desarrollado principalmente a partir de variedades *indica* y son de poca utilidad en esta zona de cultivo. Empleando este panel de SNPs se analizó la estructura genética de la colección y de sus relaciones genéticas. Los resultados de este análisis han hecho evidente el importante papel que ha tenido la acción del hombre mediante la selección de las variedades y los programas de mejora en la estructura genética de la población. Se ha podido observar analizado la estructura poblacional, que las

variedades cultivadas en clima templado pueden dividirse en cuatro grupos, basándose en el tipo de grano y el origen geográfico. Uno de los grupos está formado por variedades de grano largo, mientras que el resto presentan grano medio. Entre estos, uno incluye variedades americanas y australianas, otro variedades italianas y, finalmente, un grupo de variedades de origen asiático que incluye variedades antiguas europeas. El análisis de las relaciones genéticas puso de manifiesto que la estructura genética observada de las variedades de grano medio está ligada a la historia de la mejora del arroz. Estos resultados han sido de utilidad a la hora de realizar el estudio de asociación, ya que la fuerte estructura poblacional observada podría causar sesgos en las asociaciones llevando a falsos positivos. Como resultado de este estudio se han encontrado 43 marcadores asociados a variación en los caracteres de rendimiento y en tiempo de floración, que podrán ser inmediatamente empleados en los programas de mejora incorporando los alelos de interés en variedades ya adaptadas al cultivo en nuestra región.

La reducción del tiempo de cultivo disminuye la probabilidad de una posible exposición a estreses abióticos, como son las lluvias a finales del periodo de cosecha, y bióticos, como la piricularía, y el gasto en recursos tanto hídricos como agroquímicos. La comprensión de los elementos responsables de la regulación del momento de floración es de gran importancia. En esta tesis se ha procedido a la generación de una variedad mutante de la variedad Gleva, ampliamente cultivada en España, y se ha realizado su caracterización fenotípica y genotípica. Este mutante resultó ser insensible a fotoperiodo, tardando los mismos días en florecer bajo condiciones de fotoperiodo largo como de fotoperiodo corto. Los análisis de la expresión de los principales genes de floración mostraron unos aumentos anormales de los niveles de expresión de *Hd3a*, principal gen máster de la floración, cuya expresión suele estar inhibida en condiciones de día largo. Para la identificación de la mutación candidata se emplearon técnicas de análisis genómico analizando plantas de

una generación F2 derivadas del cruce entre el parental silvestre y la línea mutante. La identificación de la mutación candidata en una generación F2 supone una notable reducción en tiempo y trabajo respecto a otros métodos de mapeo, como la identificación de QTLs, en los que la determinación de la variación responsable puede necesitar de hasta una generación F7. La mutación candidata ha resultado ser una deleción en la que se encuentra el gen *Se13* (Saito et al., 2011) que actúa en el último paso de la ruta de síntesis de la fitocromobilina, el cromóforo de los fitocromos. Al actuar en el último paso en la ruta de síntesis resulta una mutación interesante para la mejora puesto que presenta pocos efectos indeseados que hubiesen podido haber ocurrido en puntos superiores como le sucede al mutante *Se5* que codifica una hemoxigenasa (Andrés et al., 2009; Takeshi Izawa et al., 2000) y presenta un aspecto mucho más endeble que las plantas *Se13*.

Para finalizar, el uso de técnicas de análisis genómico ha permitido la localización de varios marcadores asociados a la variación en caracteres implicados en el rendimiento y el tiempo de floración de variedades de arroz *japonica* cultivadas en clima templado, y que podrán ser inmediatamente empleados por los mejoradores. También se ha caracterizado un mutante de floración temprana e identificado una mutación interesante a la hora de reducir el ciclo vegetativo de las variedades de arroz utilizando una metodología para la detección de mutaciones en una F2 independientemente de su carácter recesivo o dominante.

5. Conclusiones generales

De los resultados obtenidos en esta tesis se extraen las siguientes conclusiones:

- La colección de 193 variedades *japonica templada* generada es representativa de la diversidad genética y fenotípica que presentan las variedades cultivadas en regiones con clima templado.
- Se ha desarrollado un panel de SNP apto para el genotipado de las variedades *japonica*, con el que se ha obtenido el perfil genético de la colección de variedades *japonica templadas*.
- Las variedades cultivadas en clima templado pueden dividirse en cuatro grupos, basándose en el tipo de grano y el origen geográfico. Las variedades de grano largo se concentran en un grupo mientras que las variedades de grano medio se distribuyen en varios grupos, uno formado por variedades americanas y australianas, otro grupo formado por variedades italianas y, finalmente, un grupo de variedades de origen asiático que incluye variedades antiguas europeas. El análisis de las relaciones genéticas puso de manifiesto que la estructura genética observada está ligada a la historia de la mejora del arroz.
- Un total de 43 SNPs están asociados a caracteres agronómicos asociados al rendimiento y al tiempo de floración en las variedades *japonica templadas.*
- El mutante G123, generado mediante irradiación, presenta una deleción en el gen *se13*/*OsHY2* que, dado el estudio del fenotipo y los análisis del transcriptoma, posiblemente sea la causa del fenotipo de adelanto de la floración e insensibilidad al fotoperiodo de sus plantas.
- El método Allinone, desarrollado recientemente en el laboratorio dónde se ha desarrollado esta tesis es válido para la detección de variaciones estructurales en el genoma y la detección de mutaciones.

- La línea mutante *G123* presenta unas características agronómicas interesantes y puede constituir una variedad de arroz con un ciclo vegetativo más corto que las cultivadas actualmente.

6. Referencias bibliográficas

Abe, A., Kosugi, S., Yoshida, K., & Natsume, S. (2012). Genome sequencing reveals agronomically important loci in rice using MutMap. *Nature …*, *30*(2), 174–178. https://doi.org/10.1038/nbt.2095

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, *21*(6), 974–984. https://doi.org/10.1101/gr.114876.110

Access, O. (2014). The 3,000 rice genomes project. *GigaScience*, *3*(1), 7. https://doi.org/10.1186/2047-217X-3-7

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., … et, al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, *252*(5013), 1651 LP-1656. Retrieved from http://science.sciencemag.org/content/252/5013/1651.abstract

Ahuja, S. C., Panwar, D. V. S., Uma, A., & Gupta, K. R. (1995). Basmati rice: the scented pearl. *Basmati Rice: The Scented Pearl.*

Alexandrov, N., Tai, S., Wang, W., Mansueto, L., Palis, K., Fuentes, R. R., … McNally, K. L. (2015). SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Research*, *43*(D1), D1023–D1027. https://doi.org/10.1093/nar/gku1039

Andaya, V. C., & Tai, T. H. (2006). Fine mapping of the qCTS12 locus, a major QTL for seedling cold tolerance in rice. *Theoretical and Applied Genetics*, *113*(3), 467–475. https://doi.org/10.1007/s00122-006-0311-5

Andrés, F., Galbraith, D. W., Talón, M., & Domingo, C. (2009). Analysis of PHOTOPERIOD SENSITIVITY5 sheds light on the role of phytochromes in photoperiodic flowering in rice. *Plant Physiol*, *151*(2), 681–690. https://doi.org/10.1104/pp.109.139097

Baruah, A. R., Ishigo-Oka, N., Adachi, M., Oguma, Y., Tokizono, Y., Onishi, K., & Sano, Y. (2009). Cold tolerance at the early growth stage in wild and cultivated rice. *Euphytica*, *165*(3), 459–470. https://doi.org/10.1007/s10681-008-9753-y

Beale, S. I., & Weinstein, J. D. (1991). Biochemistry and regulation of

photosynthetic pigment formation in plants and algae. In *New comprehensive biochemistry* (Vol. 19, pp. 155–235). Elsevier.

Begum, H., Spindel, J. E., Lalusin, A., Borromeo, T., Gregorio, G., Hernandez, J., … McCouch, S. R. (2015). Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (Oryza sativa). *PLoS ONE*, *10*(3), 1–19. https://doi.org/10.1371/journal.pone.0119873

Brad, B. W., J., E. S., D., C. H., Li, L., & S., S. P. (2007). SNP discovery via 454 transcriptome sequencing. *The Plant Journal*, *51*(5), 910–918. https://doi.org/10.1111/j.1365-313X.2007.03193.x

Cai, H., & Morishima, H. (1998). Mapping QTLs for heading behavior using RI population derived from a cross between wild and cultivated rice strains. *Rice Genetics Newsletter*, *15*, 144–146.

Chen, H., Xie, W., He, H., Yu, H., Chen, W., Li, J., … Zhang, Q. (2014). A High-Density SNP Genotyping Array for Rice Biology and Molecular Breeding. *Molecular Plant*, *7*(3), 541–553. https://doi.org/10.1093/MP/SST135

Chen, K., Wallis, J. W., Mclellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., … Elaine, R. (2013). BreaDancer - An algorithm for high resolution mapping of genomic structure variation. *Nature Methods*, *6*(9), 677–681. https://doi.org/10.1038/nmeth.1363.BreakDancer

Childs, N. W. (2004). Production and utilization of rice. *Rice: Chemistry and Technology*, *3*, 1–23.

Choi, J. Y., Platts, A. E., Fuller, D. Q., Hsing, Y. I., Wing, R. A., Purugganan, M. D., & Kim, Y. (2017). The rice paradox: Multiple origins but single domestication in Asian Rice. *Molecular Biology and Evolution*, *34*(4), 969–979. https://doi.org/10.1093/molbev/msx049

Choi, J. Y., & Purugganan, M. D. (2018). Multiple Origin but Single Domestication Led to *Oryza sativa*. *G3: Genes|Genomes|Genetics*, g3.300334.2017. https://doi.org/10.1534/g3.117.300334

Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., … Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, *5*, 613. Retrieved from

http://dx.doi.org/10.1038/nmeth.1223

Cornah, J. E., Terry, M. J., & Smith, A. G. (2003). Green or red: What stops the traffic in the tetrapyrrole pathway? *Trends in Plant Science*, *8*(5), 224–230. https://doi.org/10.1016/S1360-1385(03)00064-5

Cubry, P., Tranchant-Dubreuil, C., Thuillet, A.-C., Monat, C., Ndjiondjop, M.-N., Labadie, K., … Vigouroux, Y. (2018). The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes. *Current Biology*, *28*(14), 2274–2282.e6. https://doi.org/10.1016/J.CUB.2018.05.066

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Doi, K., Izawa, T., Fuse, T., Yamanouchi, U., Kubo, T., Shimatani, Z., … Yoshimura, A. (2004). Ehd1, a B-type response regulator in rice, confers short-day promotion of flowering and controls FT-like gene expression independently of Hd1. *Genes & Development*, *18*(8), 926–36. https://doi.org/10.1101/gad.1189604

Dong, X., Wang, X., Zhang, L., Yang, Z., Xin, X., Wu, S., … Luo, X. (2013). Identification and characterization of *OsEBS* , a gene involved in enhanced plant biomass and spikelet number in rice. *Plant Biotechnology Journal*, *11*(9), 1044–1057. https://doi.org/10.1111/pbi.12097

Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*(2), 359–361. https://doi.org/10.1007/s12686-011-9548-7

Emrich, S. J., Barbazuk, W. B., Li, L., & Schnable, P. S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, *17*(1), 69–73. https://doi.org/10.1101/gr.5145806

Endo-Higashi, N., & Izawa, T. (2011). Flowering Time Genes Heading date 1 and Early heading date 1 Together Control Panicle Development in Rice. *Plant and Cell Physiology*, *52*(6), 1083–1094. https://doi.org/10.1093/pcp/pcr059

EVANNO, G., REGNAUT, S., & GOUDET, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular*

Ecology, *14*(8), 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

EXCOFFIER, L., & LISCHER, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, *10*(3), 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x

Fu, C., Yang, X. O., Chen, X., Chen, W., Ma, Y., Hu, J., & Li, S. (2009). OsEF3, a homologous gene of Arabidopsis ELF3, has pleiotropic effects in rice. *Plant Biology (Stuttgart, Germany)*, *11*(5), 751–757. https://doi.org/10.1111/j.1438-8677.2008.00156.x

Fujino, K., & Sekiguchi, H. (2005). Identification of QTLs Conferring Genetic Variation for Heading Date among Rice Varieties at the Northern-limit of Rice Cultivation. *Breeding Science*, *55*(2), 141–146. https://doi.org/10.1270/jsbbs.55.141

Fujino, K., Wu, J., Sekiguchi, H., Ito, T., Izawa, T., & Matsumoto, T. (2010). Multiple introgression events surrounding the Hd1 flowering-time gene in cultivated rice, Oryza sativa L. *Molecular Genetics and Genomics*, *284*(2), 137–146. https://doi.org/10.1007/s00438-010-0555-2

Fuller, D. Q., van Etten, J., Manning, K., Castillo, C., Kingwell-Banham, E., Weisskopf, A., … Hijmans, R. J. (2011). The contribution of rice agriculture and livestock pastoralism to prehistoric methane levels. *The Holocene*, *21*(5), 743–759. https://doi.org/10.1177/0959683611398052

Gao, H., Jin, M., Zheng, X.-M., Chen, J., Yuan, D., Xin, Y., … Wan, J. (2014). Days to heading 7, a major quantitative locus determining photoperiod sensitivity and regional adaptation in rice. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(46), 16337–42. https://doi.org/10.1073/pnas.1418204111

Garner, W. W., & Allard, H. A. (1920). Effect of relative length of day and night and other factors of the environment on growth and reproductionin plants. *Monthly Weather Review*, *48*(7), 415–415. https://doi.org/10.1175/1520-0493(1920)48<415b:EOTRLO>2.0.CO;2

Garris, A. J., Tai, T. H., Coburn, J., Kresovich, S., & McCouch, S. (2005). Genetic structure and diversity in Oryza sativa L. *Genetics*, *169*(3), 1631–1638.

https://doi.org/10.1534/genetics.104.035642

Glaszmann, J. C. (1987). Isozymes and classification of Asian rice varieties. *Theoretical and Applied Genetics*, *74*(1), 21–30. https://doi.org/10.1007/BF00290078

Gómez-Ariza, J., Galbiati, F., Goretti, D., Brambilla, V., Shrestha, R., Pappolla, A., … Fornara, F. (2015). Loss of floral repressor function adapts rice to higher latitudes in Europe. *Journal of Experimental Botany*, *66*(7), 2027–2039. https://doi.org/10.1093/jxb/erv004

Goretti, D., Martignago, D., Landini, M., Brambilla, V., Gómez-Ariza, J., Gnesutta, N., … Fornara, F. (2017). Transcriptional and Post-transcriptional Mechanisms Limit Heading Date 1 (Hd1) Function to Adapt Rice to High Latitudes. *PLoS Genetics*, *13*(1), 1–22. https://doi.org/10.1371/journal.pgen.1006530

Gross, B. L., & Zhao, Z. (2014). Archaeological and genetic insights into the origins of domesticated rice. *Proceedings of the National Academy of Sciences*, *111*(17), 6190–6197. https://doi.org/10.1073/pnas.1308942110

Hayama, R., Yokoi, S., Tamaki, S., Yano, M., & Shimamoto, K. (2003). Adaptation of photoperiodic control pathways produces short-day flowering in rice. *Nature*, *422*(6933), 719–722. https://doi.org/10.1038/nature01549

Hill, W. G., & Weir, B. S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*, *33*(1), 54–78. https://doi.org/10.1016/0040-5809(88)90004-4

Huang, X., Kurata, N., Wei, X., Wang, Z.-X., Wang, A., Zhao, Q., … Han, B. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature*, *490*(7421), 497–501. https://doi.org/10.1038/nature11532

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., … Han, B. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics*, *42*(11), 961–967. https://doi.org/10.1038/ng.695

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., … Han, B. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nature Genetics*, *44*(1), 32–39. https://doi.org/10.1038/ng.1018

Ingvarsson, P. K., & Street, N. R. (2011). Association genetics of complex traits in plants. *New Phytologist*, *189*(4), 909–922. https://doi.org/10.1111/j.1469-8137.2010.03593.x

Itoh, H., Wada, K. C., Sakai, H., Shibasaki, K., Fukuoka, S., Wu, J., … Izawa, T. (2018). Genomic adaptation of flowering-time genes during the expansion of rice cultivation area. *The Plant Journal*, *94*(5), 895–909. https://doi.org/10.1111/tpj.13906

Izawa, T. (2007). Adaptation of flowering-time by natural and artificial selection in Arabidopsis and rice. *Journal of Experimental Botany*, *58*(12), 3091–3097. https://doi.org/10.1093/jxb/erm159

Izawa, T. (2007). Daylength Measurements by Rice Plants in Photoperiodic Short-Day Flowering. *International Review of Cytology*, *256*(07), 191–222. https://doi.org/10.1016/S0074-7696(07)56006-7

Izawa, T., Oikawa, T., Tokutomi, S., Okuno, K., & Shimamoto, K. (2000). Phytochromes confer the photoperiodic control of flowering in rice (a short-day plant). *Plant Journal*, *22*(5), 391–399. https://doi.org/10.1046/j.1365-313X.2000.00753.x

Jain, S., Jain, R. K., & McCouch, S. R. (2004). Genetic analysis of Indian aromatic and quality rice (Oryza sativa L.) germplasm using panels of fluorescently-labeled microsatellite markers. *Theoretical and Applied Genetics*, *109*(5), 965–977. https://doi.org/10.1007/s00122-004-1700-2

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*(14), 1801–1806. https://doi.org/10.1093/bioinformatics/btm233

Jeung, J. U., Hwang, H. G., Moon, H. P., & Jena, K. K. (2006). Fingerprinting temperate japonica and tropical indica rice genotypes by comparative analysis of DNA markers. *Euphytica*, *146*(3), 239–251. https://doi.org/10.1007/s10681-005-9022-2

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., … Matsumoto, T. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice*, *6*(1), 4. https://doi.org/10.1186/1939-8433-6-4

Khush, G. S. (1997). Origin, dispersal, cultivation and variation of rice. *Oryza: From Molecule to Plant*, 25–34. https://doi.org/10.1007/978-94-011-5794-0_3

Kohchi, T. (2001). The Arabidopsis HY2 Gene Encodes Phytochromobilin Synthase, a Ferredoxin-Dependent Biliverdin Reductase. *The Plant Cell Online*, *13*(2), 425–436. https://doi.org/10.1105/tpc.13.2.425

Kojima, S., Takahashi, Y., Kobayashi, Y., Monna, L., Sasaki, T., Araki, T., & Yano, M. (2002). Hd3a, a Rice Ortholog of the Arabidopsis FT Gene, Promotes Transition to Flowering Downstream of Hd1 under Short-Day Conditions. *Plant and Cell Physiology*, *43*(10), 1096–1105. https://doi.org/10.1093/pcp/pcf156

Komiya, R., Yokoi, S., & Shimamoto, K. (2009). A gene network for long-day flowering activates RFT1 encoding a mobile flowering signal in rice. *Development*, *136*(20), 3443–3450. https://doi.org/10.1242/dev.040170

Koo, B. H., Yoo, S. C., Park, J. W., Kwon, C. T., Lee, B. D., An, G., … Paek, N. C. (2013). Natural variation in OsPRR37 regulates heading date and contributes to rice cultivation at a wide range of latitudes. *Molecular Plant*, *6*(6), 1877–1888. https://doi.org/10.1093/mp/sst088

Kosugi, S., Natsume, S., Yoshida, K., MacLean, D., Cano, L., Kamoun, S., & Terauchi, R. (2013). Coval: Improving Alignment Quality and Variant Calling Accuracy for Next-Generation Sequencing Data. *PLoS ONE*, *8*(10), e75402. https://doi.org/10.1371/journal.pone.0075402

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, *33*(7), 1870–1874. https://doi.org/10.1093/molbev/msw054

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., … Zhang, G. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics*, *42*(12), 1053–1059. https://doi.org/10.1038/ng.715

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), 1–19. https://doi.org/10.1186/gb-2014-15-6-r84

Lee, Y. S., Yi, J., & An, G. (2016). OsPhyA modulates rice flowering time mainly

through OsGI under short days and Ghd7 under long days in the absence of phytochrome B. *Plant Molecular Biology*, *91*(4–5), 413–427. https://doi.org/10.1007/s11103-016-0474-7

Li, F., Liu, W., Tang, J., Chen, J., Tong, H., Hu, B., … Chu, C. (2010). Rice DENSE AND ERECT PANICLE 2 is essential for determining panicle outgrowth and elongation. *Cell Research*, *20*(7), 838–849. https://doi.org/10.1038/cr.2010.69

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, S., Qian, Q., Fu, Z., Zeng, D., Meng, X., Kyozuka, J., … Wang, Y. (2009). *Short panicle1* encodes a putative PTR family transporter and determines rice panicle size. *The Plant Journal*, *58*(4), 592–605. https://doi.org/10.1111/j.1365-313X.2009.03799.x

Liu, Y., Xu, Y., Xiao, J., Ma, Q., Li, D., Xue, Z., & Chong, K. (2011). OsDOG, a gibberellin-induced A20/AN1 zinc-finger protein, negatively regulates gibberellin-mediated cell elongation in rice. *Journal of Plant Physiology*, *168*(10), 1098–1105. https://doi.org/10.1016/J.JPLPH.2010.12.013

Ma, Y., Dai, X., Xu, Y., Luo, W., Zheng, X., Zeng, D., … Chong, K. (2015). COLD1 Confers Chilling Tolerance in Rice. *Cell*, *160*(6), 1209–1221. https://doi.org/10.1016/J.CELL.2015.01.046

Mackill, D. J., & Lei, X. (1997). Genetic Variation for Traits Related to Temperate Adaptation of Rice Cultivars. *Crop Science*, *37*(4), 1340. https://doi.org/10.2135/cropsci1997.0011183X003700040051x

Marco-Sola, S., Sammeth, M., Guigó, R., & Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, *9*(12), 1185–1188. https://doi.org/10.1038/nmeth.2221

Marroni, F., Pinosio, S., Zaina, G., Fogolari, F., Felice, N., Cattonaro, F., & Morgante, M. (2011). Nucleotide diversity and linkage disequilibrium in

Populus nigra cinnamyl alcohol dehydrogenase (CAD4) gene. *Tree Genetics & Genomes*, *7*(5), 1011–1023. https://doi.org/10.1007/s11295-011-0391-5

Matsumoto, T., Wu, J., Kanamori, H., Katayose, Y., Fujisawa, M., Namiki, N., … Burr, B. (2005). The map-based sequence of the rice genome. *Nature*, *436*(7052), 793–800. https://doi.org/10.1038/nature03895

McCouch, S. R., Wright, M. H., Tung, C.-W., Maron, L. G., McNally, K. L., Fitzgerald, M., … Mezey, J. (2016). Open access resources for genome-wide association mapping in rice. *Nature Communications*, *7*, 10532. https://doi.org/10.1038/ncomms10532

Morin, R. D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., … Marra, M. A. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques*, *45*(1), 81–94. https://doi.org/10.2144/000112900

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628.

Muramoto, T. (1999). The Arabidopsis Photomorphogenic Mutant hy1 Is Deficient in Phytochrome Chromophore Biosynthesis as a Result of a Mutation in a Plastid Heme Oxygenase. *The Plant Cell Online*, *11*(3), 335–348. https://doi.org/10.1105/tpc.11.3.335

Myles, S., Chia, J. M., Hurwitz, B., Simon, C., Zhong, G. Y., Buckler, E., & Ware, D. (2010). Rapid genomic characterization of the genus Vitis. *PLoS ONE*, *5*(1). https://doi.org/10.1371/journal.pone.0008219

Nagaraju, J., Kathirvel, M., Kumar, R. R., Siddiq, E. A., & Hasnain, S. E. (2002). Genetic analysis of traditional and evolved Basmati and non-Basmati rice varieties by using fluorescence-based ISSR-PCR and SSR markers. *Proceedings of the National Academy of Sciences*, *99*(9), 5836–5841.

Naranjo, L., Talón, M., & Domingo, C. (2014). Diversity of floral regulatory genes of japonica rice cultivated at northern latitudes. *BMC Genomics*, *15*(1), 101. https://doi.org/10.1186/1471-2164-15-101

Nemoto, Y., Nonoue, Y., Yano, M., & Izawa, T. (2016). Hd1,a CONSTANS ortholog in rice, functions as an Ehd1 repressor through interaction with monocot-

specific CCT-domain protein Ghd7. *The Plant Journal : For Cell and Molecular Biology*, *86*(3), 221–233. https://doi.org/10.1111/tpj.13168

Ogiso, E., Takahashi, Y., Sasaki, T., Yano, M., & Izawa, T. (2010). The Role of Casein Kinase II in Flowering Time Regulation Has Diversified during Evolution. *Plant Physiology*, *152*(2), 808–820. https://doi.org/10.1104/pp.109.148908

OKA, H.-I. (2008). CONSIDERATIONS ON THE GENETIC BASIS OF INTERVARIETAL STERILITY IN *ORYZA SATIVA*. In *Rice Genetics and Cytogenetics* (Vol. Volume 6, pp. 158–174). World Scientific Publishing Company. https://doi.org/doi:10.1142/9789812814302_0019

Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N., & Weigel, D. (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.*, *18*, 2024–2033. https://doi.org/10.1101/gr.080200.108.

Osugi, A., Itoh, H., Ikeda-Kawakatsu, K., Takano, M., & Izawa, T. (2011). Molecular Dissection of the Roles of Phytochrome in Photoperiodic Flowering in Rice. *Plant Physiology*, *157*(3), 1128–1137. https://doi.org/10.1104/pp.111.181792

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., … Buell, C. R. (2007). The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Research*, *35*(Database issue), D883-7. https://doi.org/10.1093/nar/gkl976

Parsons, B. J., Newbury, H. J., Jackson, M. T., & Ford-Lloyd, B. V. (1999). The genetic structure and conservation of aus, aman and boro rices from Bangladesh. *Genetic Resources and Crop Evolution*, *46*(6), 587–598. https://doi.org/10.1023/A:1008749532171

Petroni, K., Kumimoto, R. W., Gnesutta, N., Calvenzani, V., Fornari, M., Tonelli, C., … Mantovani, R. (2012). The promiscuous life of plant NUCLEAR FACTOR Y transcription factors. *The Plant Cell*, *24*(12), 4777–92. https://doi.org/10.1105/tpc.112.105734

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Ps, S., Sv, A. M., Prakash, C., Mk, R., Tiwari, R., Mohapatra, T., & Singh, N. K. (2017). High Resolution Mapping of QTLs for Heat Tolerance in Rice Using a

5K SNP Array. *Rice (New York, N.Y.)*, *10*(1), 28. https://doi.org/10.1186/s12284-017-0167-0

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Rebolledo, M. C., Peña, A. L., Duitama, J., Cruz, D. F., Dingkuhn, M., Grenier, C., & Tohme, J. (2016). Combining Image Analysis, Genome Wide Association Studies and Different Field Trials to Reveal Stable Genetic Regions Related to Panicle Architecture and the Number of Spikelets per Panicle in Rice. *Frontiers in Plant Science*, *7*, 1384. https://doi.org/10.3389/fpls.2016.01384

Reig-Valiente, J. L., Viruel, J., Sales, E., Marqués, L., Terol, J., Gut, M., … Domingo, C. (2016). Genetic Diversity and Population Structure of Rice Varieties Cultivated in Temperate Regions. *Rice*, *9*(1), 58. https://doi.org/10.1186/s12284-016-0130-5

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., … Buckler, E. S. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(20), 11479–84. https://doi.org/10.1073/pnas.201394398

Robinson, M. D., & Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, *9*(2), 321–332. https://doi.org/https://doi.org/10.1093/biostatistics/kxm030

Saito, H., Ogiso-Tanaka, E., Okumoto, Y., Yoshitake, Y., Izumi, H., Yokoo, T., … Tanisaka, T. (2012). Ef7 encodes an ELF3-like protein and promotes rice flowering by negatively regulating the floral repressor gene Ghd7 under both short-and long-day conditions. *Plant and Cell Physiology*, *53*(4), 717–728. https://doi.org/10.1093/pcp/pcs029

Saito, H., Okumoto, Y., Yoshitake, Y., Inoue, H., Yuan, Q., Teraishi, M., … Tanisaka, T. (2011). Complete loss of photoperiodic response in the rice mutant line X61 is caused by deficiency of phytochrome chromophore biosynthesis gene. *Theoretical and Applied Genetics*, *122*(1), 109–118. https://doi.org/10.1007/s00122-010-1426-2

Sakai, H., Lee, S. S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., … Itoh, T. (2013). Rice Annotation Project Database (RAP-DB): An Integrative and Interactive Database for Rice Genomics. *Plant and Cell Physiology*, *54*(2), e6–e6. https://doi.org/10.1093/pcp/pcs183

Sales, E., Viruel, J., Domingo, C., & Marqués, L. (2017). Genome wide association analysis of cold tolerance at germination in temperate japonica rice (Oryza sativa L.) varieties. *PLOS ONE*, *12*(8), e0183416. https://doi.org/10.1371/journal.pone.0183416

Sato, Y., Takehisa, H., Kamatsuki, K., Minami, H., Namiki, N., Ikawa, H., … Nagamura, Y. (2013). RiceXPro Version 3.0: expanding the informatics resource for rice transcriptome. *Nucleic Acids Research*, *41*(D1), D1206–D1213. https://doi.org/10.1093/nar/gks1125

Schmidt, R., Schippers, J. H. M., Mieulet, D., Watanabe, M., Hoefgen, R., Guiderdoni, E., & Mueller-Roeber, B. (2014). SALT-RESPONSIVE ERF1 Is a Negative Regulator of Grain Filling and Gibberellin-Mediated Seedling Establishment in Rice. *Molecular Plant*, *7*(2), 404–421. https://doi.org/10.1093/MP/SST131

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., … Andersen, S. U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature Methods*, *6*(8), 550–551. https://doi.org/10.1038/nmeth0809-550

Sheoran, I. S., Koonjul, P., Attieh, J., & Saini, H. S. (2014). Water-stress-induced inhibition of α-tubulin gene expression during growth, and its implications for reproductive success in rice. *Plant Physiology and Biochemistry*, *80*, 291–299. https://doi.org/10.1016/J.PLAPHY.2014.04.011

Shimamoto, K., & Kyozuka, J. (2002). RICE AS A MODEL FOR COMPARATIVE GENOMICS OF PLANTS. *Annual Review of Plant Biology*, *53*(1), 399–419. https://doi.org/10.1146/annurev.arplant.53.092401.134447

Shrestha, R., Gómez-Ariza, J., Brambilla, V., & Fornara, F. (2014). Molecular control of seasonal flowering in rice, arabidopsis and temperate cereals. *Annals of Botany*, *114*(7), 1445–1458. https://doi.org/10.1093/aob/mcu032

Singh, N., Jayaswal, P. K., Panda, K., Mandal, P., Kumar, V., Singh, B., … Singh, N. K. (2015). Single-copy gene based 50 K SNP chip for genetic studies and

molecular breeding in rice. *Scientific Reports*, *5*(1), 11600. https://doi.org/10.1038/srep11600

Sui, J.-M., Guo, B.-T., Wang, J.-S., Qiao, L.-X., Zhou, Y., Zhang, H.-G., … Liang, G.-H. (2012). A New GA-Insensitive Semidwarf Mutant of Rice (Oryza sativa L.) with a Missense Mutation in the SDG Gene. *Plant Molecular Biology Reporter*, *30*(1), 187–194. https://doi.org/10.1007/s11105-011-0321-6

Sutcliffe, J. G., Milner, R. J., Bloom, F. E., & Lerner, R. A. (1982). Common 82-nucleotide sequence unique to brain RNA. *Proceedings of the National Academy of Sciences*, *79*(16), 4942 LP-4946. Retrieved from http://www.pnas.org/content/79/16/4942.abstract

Sweeney, M., & McCouch, S. (2007). The complex history of the domestication of rice. *Annals of Botany*, *100*(5), 951–957. https://doi.org/10.1093/aob/mcm128

Takahashi, Y., Shomura, A., Sasaki, T., & Yano, M. (2001). Hd6, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the alpha subunit of protein kinase CK2. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(14), 7922–7927. https://doi.org/10.1073/pnas.111136798

Takano, M., Inagaki, N., Xie, X., Kiyota, S., Baba-Kasai, A., Tanabata, T., & Shinomura, T. (2009). Phytochromes are the sole photoreceptors for perceiving red/far-red light in rice. *Proceedings of the National Academy of Sciences*, *106*(34), 14705–14710. https://doi.org/10.1073/pnas.0907378106

Takano, M., Inagaki, N., Xie, X., Yuzurihara, N., Hihara, F., Ishizuka, T., … Shinomura, T. (2005). Distinct and cooperative functions of phytochromes A, B, and C in the control of deetiolation and flowering in rice. *The Plant Cell*, *17*(December), 3311–3325. https://doi.org/DOI 10.1105/tpc.105.035899

Tamaki, S., Matsuo, S., Wong, H. L., Yokoi, S., & Shimamoto, K. (2007). Hd3a Protein Is a Mobile Flowering Signal in Rice. *Science*, *316*(5827), 1033–1036. https://doi.org/10.1126/science.1141753

Terry, M. J. (1997). Phytochrome chromophore-deficient mutants. *Plant Cell And Environment*, *20*(6), 740–745. https://doi.org/10.1046/j.1365-3040.1997.d01-102.x

Terry, M. J., & Kendrick, R. E. (1999). Feedback Inhibition of Chlorophyll Synthesis in the Phytochrome Chromophore-Deficient *aurea* and *yellow-green-2* Mutants of Tomato. *Plant Physiology*, *119*(1), 143–152. https://doi.org/10.1104/pp.119.1.143

Thangasamy, S., Chen, P.-W., Lai, M.-H., Chen, J., & Jauh, G.-Y. (2012). Rice LGD1 containing RNA binding activity affects growth and development through alternative promoters. *The Plant Journal*, *71*(2), 288–302. https://doi.org/10.1111/j.1365-313X.2012.04989.x

The 3, 000 Rice Genomes Project. (2014). The 3,000 rice genomes project. *GigaScience*, *3*(1), 7. https://doi.org/10.1186/2047-217X-3-7

Thomson, M. J., Singh, N., Dwiyanti, M. S., Wang, D. R., Wright, M. H., Perez, F. A., … McCouch, S. R. (2017). Large-scale deployment of a rice 6 K SNP array for genetics and breeding applications. *Rice*, *10*(1), 40. https://doi.org/10.1186/s12284-017-0181-2

Tung, C.-W., Zhao, K., Wright, M. H., Ali, M. L., Jung, J., Kimball, J., … McCouch, S. R. (2010). Development of a Research Platform for Dissecting Phenotype–Genotype Associations in Rice (Oryza spp.). *Rice*, *3*(4), 205–217. https://doi.org/10.1007/s12284-010-9056-5

Ueguchi-Tanaka, M., Ashikari, M., Nakajima, M., Itoh, H., Katoh, E., Kobayashi, M., … Matsuoka, M. (2005). GIBBERELLIN INSENSITIVE DWARF1 encodes a soluble receptor for gibberellin. *Nature*, *437*(7059), 693–698. https://doi.org/10.1038/nature04028

Volante, A., Desiderio, F., Tondelli, A., Perrini, R., Orasen, G., Biselli, C., … Valè, G. (2017). Genome-Wide Analysis of japonica Rice Performance under Limited Water and Permanent Flooding Conditions. *Frontiers in Plant Science*, *8*, 1862. https://doi.org/10.3389/fpls.2017.01862

Wang, J., Qi, M., Liu, J., & Zhang, Y. (2015). CARMO: a comprehensive annotation platform for functional exploration of rice multi-omics data. *The Plant Journal*, *83*(2), 359–374. https://doi.org/10.1111/tpj.12894

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), 1–7. https://doi.org/10.1093/nar/gkq603

Wei, X., Xu, J., Guo, H., Jiang, L., Chen, S., Yu, C., … Wan, J. (2010). DTH8 Suppresses Flowering in Rice, Influencing Plant Height and Yield Potential Simultaneously. *Plant Physiology*, *153*(4), 1747–1758. https://doi.org/10.1104/pp.110.156943

Wu, W., Zheng, X.-M., Lu, G., Zhong, Z., Gao, H., Chen, L., … Wan, J. (2013). Association of functional nucleotide polymorphisms at DTH2 with the northward expansion of rice cultivation in Asia. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(8), 2775–80. https://doi.org/10.1073/pnas.1213962110

Xiao, J., Li, J., Yuan, L., & Tanksley, S. D. (1996). Identification of QTLs affecting traits of agronomic importance in a recombinant inbred population derived from a subspecific rice cross. *Theoretical and Applied Genetics*, *92*(2), 230–244. https://doi.org/10.1007/BF00223380

Xing, Y., & Zhang, Q. (2010). Genetic and Molecular Bases of Rice Yield. *Annual Review of Plant Biology*, *61*(1), 421–442. https://doi.org/10.1146/annurev-arplant-042809-112209

Xu, X., Liu, X., Ge, S., Jensen, J. D., Hu, F., Li, X., … Wang, W. (2011). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology*, *30*(1), 105–111. https://doi.org/10.1038/nbt.2050

Xue, W., Xing, Y., Weng, X., Zhao, Y., Tang, W., Wang, L., … Zhang, Q. (2008). Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. *Nature Genetics*, *40*(6), 761–767. https://doi.org/10.1038/ng.143

Yamamoto, E., Yonemaru, J., Yamamoto, T., & Yano, M. (2012). OGRO: The Overview of functionally characterized Genes in Rice online database. *Rice*, *5*(1), 26. https://doi.org/10.1186/1939-8433-5-26

Yamamoto, T., Lin Hongxuan, Sasaki, T., & Yano, M. (2000). Identification of heading date quantitative trait locus Hd6 and characterization of its epistatic interactions with Hd2 in rice using advanced backcross progeny. *Genetics*, *154*(2), 885–891.

Yan, W. H., Wang, P., Chen, H. X., Zhou, H. J., Li, Q. P., Wang, C. R., … Zhang, Q. F. (2011). A major QTL, Ghd8, plays pleiotropic roles in regulating grain

productivity, plant height, and heading date in rice. *Molecular Plant*, *4*(2), 319–330. https://doi.org/10.1093/mp/ssq070

Yang, Y., Peng, Q., Chen, G. X., Li, X. H., & Wu, C. Y. (2013). OsELF3 Is involved in circadian clock regulation for promoting flowering under long-day conditions in rice. *Molecular Plant*, *6*(1), 202–215. https://doi.org/10.1093/mp/sss062

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P., Hu, L., … Matsuoka, M. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature Genetics*, *48*(8), 927–934. https://doi.org/10.1038/ng.3596

Yano, M., Katayose, Y., Ashikari, M., Yamanouchi, U., Monna, L., Fuse, T., … Sasaki, T. (2000). Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS. *The Plant Cell*, *12*(12), 2473–2484. https://doi.org/10.1105/tpc.12.12.2473

Yonemaru, J., Mizobuchi, R., Kato, H., Yamamoto, T., Yamamoto, E., Matsubara, K., … Yano, M. (2014). Genomic regions involved in yield potential detected by genome-wide association analysis in Japanese high-yielding rice cultivars. *BMC Genomics*, *15*(1), 346. https://doi.org/10.1186/1471-2164-15-346

Yonemaru, J., Yamamoto, T., Fukuoka, S., Uga, Y., Hori, K., & Yano, M. (2010). Q-TARO: QTL Annotation Rice Online Database. *Rice*, *3*(2–3), 194–203. https://doi.org/10.1007/s12284-010-9041-z

Yoshitake, Y., Yokoo, T., Saito, H., Tsukiyama, T., Quan, X., Zikihara, K., … Tanisaka, T. (2015). The effects of phytochrome-mediated light signals on the developmental acquisition of photoperiod sensitivity in rice. *Scientific Reports*, *5*(Ld), 16–18. https://doi.org/10.1038/srep07709

YU, C. Y., WEI, X. J., CHEN, L. M., JIANG, L., ZHAI, H. Q., & WAN, J. M. (2005). 16. Identification of a dominant suppressor of photoperiod-sensitive gene using indica/japonica backcrossed progenies in rice (Oryza sativa L.). *Rice Genetics Newsletter*, *22*, 54.

Yu, H., Xie, W., Li, J., Zhou, F., & Zhang, Q. (2014). A whole-genome SNP array (RICE6K) for genomic breeding in rice. *Plant Biotechnology Journal*, *12*(1), 28–37. https://doi.org/10.1111/pbi.12113

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., …

Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208. https://doi.org/10.1038/ng1702

Zha, X., Luo, X., Qian, X., He, G., Yang, M., Li, Y., & Yang, J. (2009). Over-expression of the rice *LRK1* gene improves quantitative yield components. *Plant Biotechnology Journal*, *7*(7), 611–620. https://doi.org/10.1111/j.1467-7652.2009.00428.x

Zhang, P., Liu, X., Tong, H., Lu, Y., & Li, J. (2014). Association Mapping for Important Agronomic Traits in Core Collection of Rice (Oryza sativa L.) with SSR Markers. *PLoS ONE*, *9*(10), e111508. https://doi.org/10.1371/journal.pone.0111508

Zhang, Q., Maroof, M. A. S., Lu, T. Y., & Shen, B. Z. (1992). Genetic diversity and differentiation of indica and japonica rice detected by RFLP analysis. *Theoretical and Applied Genetics*, *83*(4), 495–499. https://doi.org/10.1007/BF00226539

Zhao, J., Huang, X., Ouyang, X., Chen, W., Du, A., Zhu, L., … Li, S. (2012). OsELF3-1, an Ortholog of Arabidopsis EARLY FLOWERING 3, Regulates Rice Circadian Rhythm and Photoperiodic Flowering. *PLoS ONE*, *7*(8), e43705. https://doi.org/10.1371/journal.pone.0043705

Zhao, K., Tung, C.-W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., … McCouch, S. R. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nature Communications*, *2*(1), 467. https://doi.org/10.1038/ncomms1467

Zhao, K., Wright, M., Kimball, J., Eizenga, G., McClung, A., Kovach, M., … McCouch, S. R. (2010). Genomic Diversity and Introgression in O. sativa Reveal the Impact of Domestication and Breeding on the Rice Genome. *PLoS ONE*, *5*(5), e10780. https://doi.org/10.1371/journal.pone.0010780